RADIATION EFFECTS IN SEMICONDUCTORS

Devices, Circuits, and Systems

Series Editor
Krzysztof Iniewski
CMOS Emerging Technologies Inc., Vancouver, British Columbia, Canada

Internet Networks: Wired, Wireless, and Optical Technologies
Krzysztof Iniewski

Semiconductor Radiation Detection Systems
Krzysztof Iniewski

Electronics for Radiation Detection
Krzysztof Iniewski

Radiation Effects in Semiconductors
Krzysztof Iniewski

FORTHCOMING

Radio Frequency Integrated Circuit Design
Sebastian Magierowski

Semiconductors: Integrated Circuit Design for Manufacturability
Artur Balasinki

Electrical Solitons: Theory, Applications, and Extensions in High Speed Electronics
David Ricketts

Integrated Microsystems: Materials, MEMs, Photonics, Bio Interfaces
Krzysztof Iniewski

RADIATION EFFECTS IN SEMICONDUCTORS

Edited by

Krzysztof Iniewski

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4398-2694-2 (Hardback)

Library of Congress Cataloging-in-Publication Data

Radiation effects in semiconductors / editor, Krzysztof Iniewski.
p. cm. -- (Devices, circuits, and systems)
Includes bibliographical references and index.
ISBN 978-1-4398-2694-2 (hardcover : alk. paper)
1. Semiconductors--Effect of radiation on. 2. Photon emission. I. Iniewski, Krzysztof. II. Title. III. Series.

TK7871.85.R317 2010
621.3815'2--dc22 2010011811

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

SECTION I Devices

SECTION II Circuits and Systems

Preface

The need to understand radiation effects and to combat potential radiation damage in semiconductor devices and circuits has been growing in recent years. Space applications, nuclear physics, and military operations in radiation environments are obvious areas where radiation damage can have serious consequences. In addition, modern silicon processing techniques introduce radiation damage. Last, but not least, radiation is used heavily in medical imaging, from simple X-ray examinations to complex CT or SPECT/PET scanning procedures.

The interaction of radiation with matter is a very broad and complex topic. In this book we try to analyze the problem, with the aim of explaining the most important aspects for comprehending the degradation effects observed in semiconductor devices, circuits, and systems when they are irradiated. The manner in which radiation interacts with solid materials depends on the type of incident particle and on the atomic number and density of the target material. Photons interact with matter in three different ways.

Photoelectric effect: The incident photon ionizes the target atom and is completely absorbed. A photoelectric electron is emitted and an electron in an outer orbit of the atom falls into the spot vacated by the photoelectron, causing a low-energy photoelectric photon to be emitted.

Compton effect: An electron of the target atom is set free and a photon is emitted. The energy of the incident photon is divided between the two products of the interaction.

Creation of electron-positron pairs: The incident photon is completely annihilated. This phenomenon never happens when the energy of the incident photon is less than 1.024 MeV.

The probability of these three effects occurring changes with the energy of the incident photon and also depends on the atomic number of the target.

The effects of photons on matter can be grouped into two classes: ionization effects and nuclear displacement. These phenomena may be caused directly by the incident particle or from secondary phenomena induced by it. Ionization in a semiconductor or insulating material creates electron-hole pairs. The number of pairs created is proportional to the quantity of energy deposited in the material that is expressed through the total absorbed dose.

This book is a must-read for anyone serious about understanding radiation effects in the electronics industry. It is aimed at post-graduate researchers, semiconductor engineers, and nuclear and space engineers with some electronics background. Emerging detector technologies, circuit design techniques, new materials, and innovative system approaches are explored by top-notch international experts in industry and academia. The book can be used as recommended reading and supplementary material in a graduate course curriculum.

All MATLAB® files found in the book are available for download from the publisher's Web site. MATLAB® is a registered trademark of The MathWorks, Inc. For product information please contact

The MathWorks, Inc.
3 Apple Drive
Natick, MA 01760-2098 USA
Tel: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Kris Iniewski,
Coquitlam, BC, Canada

About the Editor

Krzysztof (Kris) Iniewski is manager of research and development at Redlen Technologies Inc., a start-up company in Vancouver, Canada. Redlen's revolutionary production process for advanced semiconductor materials enables a new generation of more accurate, all-digital, radiation-based imaging solutions. Kris is also an executive director of CMOS Emerging Technologies (www.cmoset.com), a series of high-tech events covering communications, microsystems, optoelectronics, and sensors.

During his career Dr. Iniewski has held numerous faculty and management positions at the University of Toronto, the University of Alberta, SFU, and PMC-Sierra Inc. He has published over 100 research papers in international journals and conferences, and he holds 18 international patents granted in the USA, Canada, France, Germany, and Japan. He is a frequently invited speaker and has consulted for multiple organizations internationally. He has written and edited several books for Wiley & Sons, CRC Press, McGraw Hill, Artech House, and Springer.

His personal goal is to contribute to sustainability through innovative engineering solutions. He can be reached at kris.iniewski@gmail.com.

List of Contributors

Anne-Johan Annema
National Institute for Subatomic Physics (Nikhef)
Amsterdam, the Netherlands

Jean-Luc Autran
Université de Provence
Marseille, France

Gregory Avenier
STMicroelectronics
Crolles, France

Marco Bellini
ABB
Zürich, Switzerland

Bharat L. Bhuva
Vanderbilt University
Nashville, Tennessee

Michael Caffrey
Los Alamos National Laboratory
Los Alamos, New Mexico

Lucio Carrara
ESPROS Photonics
Baar, Switzerland

James Carroll
Brigham Young University
Provo, Utah

Gianluigi Casse
University of Liverpool
Liverpool, UK

Andrea Cester
Università degli Studi di Padova
Padova, Italy

Alain Chantre
STMicroelectronics
Crolles, France

Edoardo Charbon
TU Delft
Delft, the Netherlands

Peng Cheng
Georgia Institute of Technology
Atlanta, Georgia

Pascal Chevalier
STMicroelectronics
Crolles, France

Lawrence T. Clark
Arizona State University
Phoenix, Arizona

John D. Cressler
Georgia Institute of Technology
Atlanta, Georgia

Ryan M. Diestelhorst
Georgia Institute of Technology
Atlanta, Georgia

Daniel M. Fleetwood
Vanderbilt University
Nashville, Tennessee

Gilles Gasiot
STMicroelectronics
Crolles, France

Cosimo Gerardi
Numonyx
Catania, Italy

Derrick Gibelyou
Brigham Young University
Provo, Utah

Paul Graham
Los Alamos National Laboratory
Los Alamos, New Mexico

Vladimir Gromov
National Institute for Subatomic Physics (Nikhef)
Amsterdam, the Netherlands

William Timothy Holman
Vanderbilt University
Nashville, Tennessee

William Howes
Brigham Young University
Provo, Utah

Jonathan Johnson
Brigham Young University
Provo, Utah

Jim Krone
Los Alamos National Laboratory
Los Alamos, New Mexico

Salvatore Lombardo
CNR - IMM
Catania, Italy

Kevin Lundgreen
Brigham Young University
Provo, Utah

Juan Antonio Maestro
Nebrija Universidad
Madrid, Spain

Paul W. Marshall
NASA
Washington, DC

Lloyd W. Massengill
Vanderbilt University
Nashville, Tennessee

Daniel McMurtrey
Brigham Young University
Provo, Utah

Keith S. Morgan
Los Alamos National Laboratory
Los Alamos, New Mexico

Daniela Munteanu
Université de Provence
Marseille, France

Balaji Narasimham
Broadcom Corporation
Irvine, California

Cristiano Niclass
EPFL
Lausanne, Switzerland

Patrick Ostler
Brigham Young University
Provo, Utah

Stanley D. Phillips
Georgia Institute of Technology
Atlanta, Georgia

Rosario Portoghese
Numonyx
Catania, Italy

Brian Pratt
Brigham Young University
Provo, Utah

Heather Quinn
Los Alamos National Laboratory
Los Alamos, New Mexico

Philippe Roche
STMicroelectronics
Crolles, France

Sébastien Sauze
Université de Provence
Marseille, France

Ronald D. Schrimpf
Vanderbilt University
Nashville, Tennessee

Noémy Scheidegger
Oerlikon Space
Zurich, Switzerland

Herbert Shea
EPFL
Lausanne, Switzerland

Robert L. Shuler, Jr.
NASA Johnson Space Center
Houston, Texas

Marek Turowski
CFD Research Corporation
Huntsville, Alabama

Pedro Reviriego Vasallo
Nebrija Universidad
Madrid, Spain

Massimo Violante
Politecnico di Torino
Torino, Italy

Michael Wirthlin
Brigham Young University
Provo, Utah

Arthur F. Witulski
Vanderbilt University
Nashville, Tennessee

Nicola Wrachien
Università degli Studi di Padova
Padova, Italy

Xing J. Zhou
IBM Semiconductor Research and Development Center
Hopewell Junction, New York

Section I

Devices

1 Radiation Damage in Silicon

Gianluigi Casse

CONTENTS

1.1 INTRODUCTION

The operations of silicon detectors are progressively degraded by radiation, ultimately leading to their failure. The radiation damage mechanism in the sensors can be divided in two classes: surface and bulk damage.

1.1.1 Surface Damage

The passage of ionizing radiation causes the build-up of trapped positive charge in the dielectric layer (usually SiO_2) that covers the silicon detectors. The ionized e-h pairs either recombine or move in the oxide electric field: the electrons toward the SiO_2-Si interface and the holes toward the metallic contact. The higher mobile electrons escaped from the recombination are injected into the silicon bulk in a typical time of ≈ 100 ps. The less mobile holes can be trapped at the SiO_2-Si interface. This trapping results in an increase of the oxide positive space charge with degradation of its quality. This charge build-up saturates at a value close to 2×10^{12} cm^{-3} after a dose of about 100 kRad [1]. In addition to the trapped charge, the ionizing radiation also produces new energy levels in the band gap at the SiO_2-Si interface [2]. These levels can be occupied by electrons or holes, depending on the position of the Fermi level at the interface, and the corresponding charge can be added or subtracted to the oxide charge. The effects of radiation on the oxide and the silicon surface depend on the specific detector design. The most relevant aspect for the operations of modern sensors is the formation of a conductive layer of electrons attracted to the interface by the positive oxide charge. This aspect will be briefly discussed later when presenting p-side readout segmented detectors.

1.1.2 Bulk Damage

The radiation impinging onto the silicon crystal can cause point-like defects (a single silicon atom displaced from its lattice position) or *cluster* defects (a high concentration of damaged crystal in a volume with radius between 10 nm and 200 nm [3]), depending on the particle energy and type. It is known that protons produce more point-like and less cluster defects than neutrons, due to the different nature of the scattering with the crystal caused by the electromagnetic interaction. Also, the same particles with different energies produce different damage to the silicon crystal. The radiation-induced defects can be electrically neutral or active: in the second case they can act as generation-recombination or trapping centers of the charge carriers. In particular, they cause considerable changes in parameters relevant to the operation of silicon sensors, like the full depletion voltage (V_{FD}), reverse current (I_R), and the signal height (S).

To compare the effects of the various particles to the silicon lattice structure, the radiation damage is scaled using the *nonionizing energy loss* (NIEL) [4]. This quantity evaluates the energy deposited in the crystal by interactions that cannot be described by the reversible process of ionization. This scaling allows the macroscopic electrical properties of the silicon devices to be compared and is performed by folding the energy spectra of the particles of a given radiation field with the appropriate NIEL factor to express the fluence in terms of a reference monochromatic particle. The common reference particles are 1 MeV neutrons that have therefore a NIEL normalization factor k = 1. The NIEL normalized fluence is called the 1 MeV neutron equivalent (n_{eq}). The radiation-induced changes in the reverse current have been found to scale well with the calculated NIEL normalization factors. As an example, the changes of the reverse current of identical silicon detectors irradiated with 24 GeV/c protons and 1 MeV neutrons match very well if the proton fluence is multiplied by $k = 0.62$.

After irradiation, the defects can interact with other mobile impurities in silicon (e.g., hydrogen, carbon or oxygen, interstitial silicon) and form permanent complexes with a possibly different electrical nature from the original ones. This *annealing* process is a function of time and temperature and changes again the electrical properties of the detectors.

The studies carried out with simple pad diodes have allowed the parameterization of the changes of the full depletion voltage and the reverse current as a function of irradiation and time after irradiation. Extensive literature is available on this subject; for a summary, see, for example, [5].

The V_{FD} is proportional to the effective space charge density (N_{eff}):

$$V_{FD} = \frac{ew^2 \left|N_{eff}\right|}{2\varepsilon_0 \varepsilon_{Si}} \tag{1.1}$$

The changes of N_{eff} as a function of fluence are described by the following equation:

$$N_{eff}(\phi) = N_D e^{-c\phi} - N_A e^{-d\phi} + (\beta_D - \beta_A)\phi \tag{1.2}$$

where N_D is the initial donor concentration, N_A is the initial acceptor concentration, c and d are removal constants, and β_A and β_D are the parameters accounting for the introduction of acceptor- or donor-like defects. The first two terms describe the exponential removal of the initial doping. In n-type silicon (p-type) the initial acceptor (donor) concentration N_A (N_D) can be neglected. The second factor of Equation (1.2) describes the linear changes of N_{eff} with the radiation fluence. In fact, the radiation introduces defects acting as both types of doping, but the acceptor-like defects appear to be dominant, at least in high-resistivity detector-grade float-zone (FZ) silicon. The factor ($\beta_D - \beta_A$) is in general negative and can be replaced with an effective introduction of acceptor defects $-\beta_{Aeff}$. In this case, the initial (positive) doping concentration of n-type silicon decreases exponentially until conductivity-type inversion [6] to effective p-type and then becomes more negative linearly with fluence (Figure 1.1), at least up to the maximum doses where direct measurements have been performed (a few times 10^{15} n_{eq} cm^{-2}). No space charge inversion has been measured with p-type bulk devices. Although this is true for the majority of the silicon crystal types studied, exceptions (depending on the impinging particle type) will be described as follows.

The reverse current I_R is proportional to the detector volume V and increases linearly with fluence:

$$\Delta I_R(\phi) = I_R(\phi, T) - I_R(\phi = 0, T) = \alpha_{T,t}\phi V \tag{1.3}$$

The *damage factor* $\alpha_{T,t}$ depends on temperature and time after irradiation (annealing). The current of irradiated sensors depends exponentially on temperature and is described by the following expression:

$$I_R \propto T^2 \exp\left(-\frac{E_A}{k_B(T + 273.2)}\right) \tag{1.4}$$

where the *activation energy* E_A is 1.12 eV. In practice, the reverse current decreases by a factor of ~2 by reducing the temperature by 9°C.

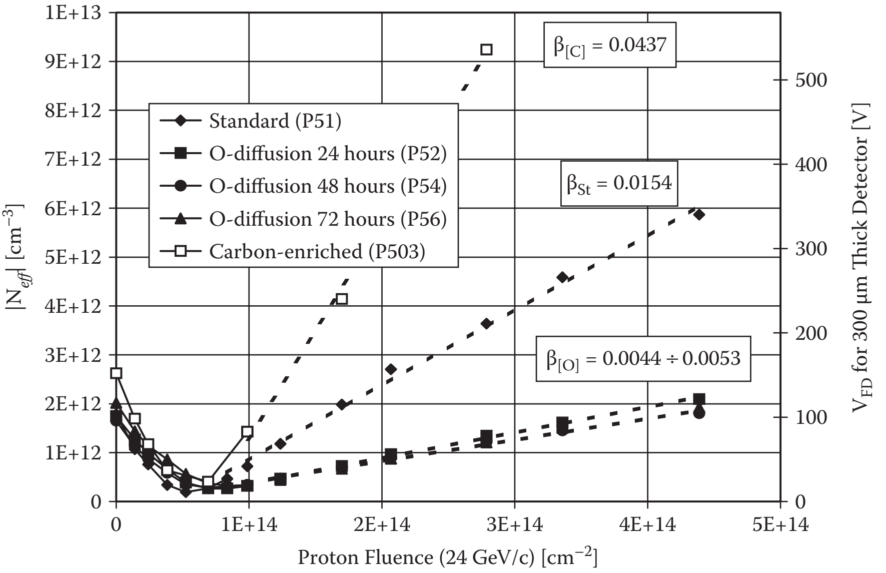


FIGURE 1.1 Absolute value of N_{eff} (V_{FD}) as a function of the proton fluence for standard, oxygen enriched and carbon enriched silicon. It is visible the initial exponential decay followed by the linear increase of the absolute value of the effective space charge, being in fact negative after type inversion. It is evident the advantage of oxygen enrichment and the opposite effect of carbon on the changes of these parameters [5]. (G. Lindstrom et al., "Radiation hard silicon detectors developments by the RD48 (ROSE) Collaboration," *Nucl. Instrum. Meth.* Vol. 466, Issue 2, July 1, 2001, pp. 308–326. With permission from Elsevier.)

1.2 ANNEALING OF I_R AND N_{eff}

Both the full depletion voltage and the reverse current change with time after irradiation (*annealing*) [5,7–9]. The current decreases with a series of time constants, τ_i, and can be parameterized as follows:

$$\Delta I_{Vol}(t,\Phi) = \Delta I_{Vol}(0,\Phi)\sum_i a_i \exp\left(-\frac{t}{\tau_i}\right), \quad \sum_i a_i \equiv 1 \tag{1.5}$$

where the parameters' a_i's weigh the various contributions to the reverse current with different annealing times. Different parameterizations exist (see, e.g., [5,8]), but they also introduce some arbitrary factor to reproduce the annealing of the current. Figure 1.2 shows an example of fits to the annealing of the reverse current after different levels of irradiation using Equation (1.5).

The annealing rate can be slowed down (accelerated) by lowering (increasing) the temperature. In fact, the time constants are an exponential function of the temperature through the following Arrhenius relationship [5,7]:

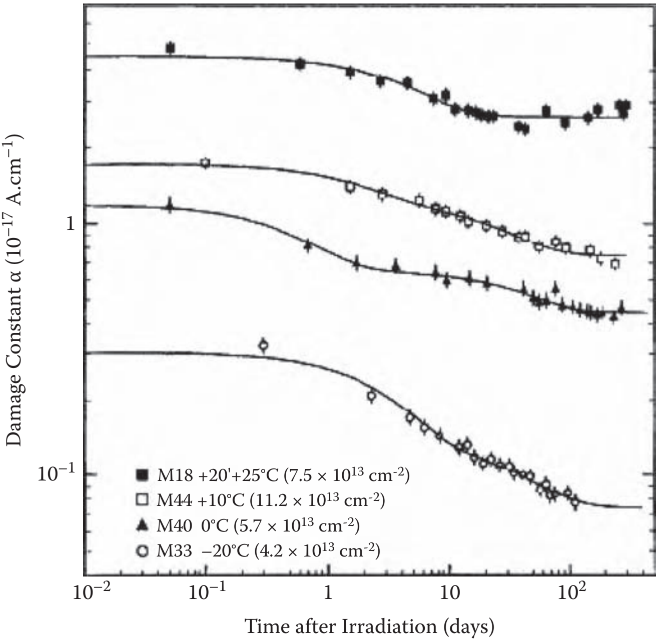


FIGURE 1.2 Dependence of the reverse current damage factor on the time after irradiation at the indicated temperatures [9]. (F. Lemeilleur, S. J. Bates, A. Chilingarov, C. Furetta, M. Glaser, E. H. M., et al., "Study of characteristics of silicon detectors irradiated with 24GeV/c protons between –20°C and +20°C," *Nucl. Instrum. Meth.* Volume 360, Issues 1–2, June 1, 1995, pp.438–444. With permission from Elsevier.)

$$\frac{1}{\tau_i} = k_{0,i} \exp\left(-\frac{E_A}{k_B T_A}\right) \tag{1.6}$$

where E_A is the silicon energy gap. Due to the strong dependence of the current on temperature and annealing, when giving a value for the constant $\alpha_{T,t}$ in Equation (1.3), it is necessary to specify these quantities. The value for $\alpha_{T,t}$, given in literature after an annealing time of 80 minutes at 60°C and at the reference temperature T_{REF} = 20°C corresponds to 3.99×10^{-17} A cm^{-1}.

The dependence of N_{eff} (V_{FD}) on the time after irradiation is shown in Figure 1.3 in the case of an inverted n-type detector. N_{eff} is found to decrease during 80 minutes at 60°C (or about 30 days at T_{REF}). After this *beneficial short-term* annealing, N_{eff} starts to increase again over many years at T_{REF} up to an apparent saturation. This second phase of the N_{eff} dependence on the time after irradiation is called *reverse annealing.* The first-phase (short-term) annealing is described by

$$N_{ST} = g_{ST} \exp\left(-\frac{t}{\tau_{ST}(T)}\right)\phi \tag{1.7}$$

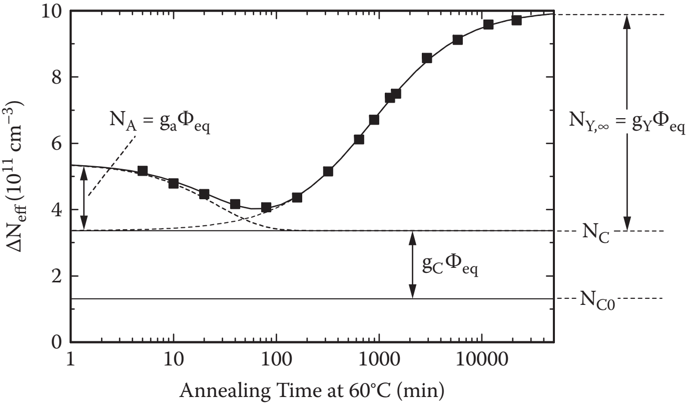


FIGURE 1.3 Evolution of N_{eff} with time after irradiation for a conductivity type inverted n-type silicon detector. The various factors of the parameterisation are visualized. The minimum of this curve occurs at about 80 minutes at 60°C, corresponding to about 30 days at 20°C (the acceleration factor being about 540) [5]. (G. Lindstrom et al., "Radiation hard silicon detectors developments by the RD48 (ROSE) Collaboration," *Nucl. Instrum. Meth.* Vol. 466, Issue 2, July 1, 2001, pp. 308–326. With permission from Elsevier.)

with the usual Arrhenius dependence of the inverse of the short-term annealing rate τ_{ST}:

$$\frac{1}{\tau_{ST}} = k_{ST} \exp\left(-\frac{E_{ST}}{k_B T_{ST}}\right)$$

with $k_{ST} = (2.4 \pm 1) \times 10^{13}\ s^{-1}$ and $E_{ST} = 1.09 \pm 0.03$ eV.

The reverse annealing data (for $t > 10$ days) are well fitted by the form [5,8]

$$N_Y(\phi,t,T) = N_{C0}(\phi) + N_{Y\infty}(\phi)\left[1 - \frac{1}{\left(1 + \frac{t}{\tau_{LT}}\right)}\right] \tag{1.8}$$

where $N_{C0}(\phi)$ is called the stable damage component, $N_{Y\infty}(\phi)$ is the saturation value at the end of the process, and t_{LT} is the reverse annealing rate constant. $N_{C0}(\phi)$ defines the minimum value of V_{FD} after the short-term annealing and exhibits a certain variability for different materials. The amplitude of the reverse annealing can be expressed as a function of the fluence as

$$N_{Y\infty}(\phi) = g_Y \phi \tag{1.9}$$

with $g_Y = 5.16 \pm 0.09 \times 10^{-2}\ cm^{-1}$.

The dependence of the τ_{LT} on temperature is again of the type

$$\frac{1}{\tau_{LT}} = k_{LT} \exp\left(-\frac{E_{LT}}{k_B T}\right) \tag{1.10}$$

with $k_{LT} \cong 1.5 \times 10^{15}$ s^{-1} and $E_{LT} = 1.33 \pm 0.03$ eV.

The parametric functions shown here for the changes of N_{eff} and I_R have been determined experimentally by V_{FD} and I_R measurements of silicon diodes and not derived from fundamental kinetics of the defects and semiconductor generation-recombination statistics. The microscopic explanation for the annealing behavior is still not satisfactory, and the correlation between the identified radiation-induced defects and the annealing of the leakage current is not available.

1.2.1 Impurities in Silicon

Two possible approaches to the radiation hardening of the silicon sensors are the engineering of the impurity content of the silicon substrate and the optimization of the geometry of the devices. The first approach has been extensively studied, especially by the dedicated CERN-RD48 research and development (R&D) project [10]. It has been shown [11] that the variation of the content of carbon and oxygen in the silicon crystal can have an effect on the degradation of some of the electrical properties with irradiation. In particular, a relatively high concentration of oxygen can reduce the degradation rate of N_{eff} (V_{FD}) as a function of charged hadron fluence. Figure 1.1 shows a comparison of the changes of N_{eff} as a function of 24 GeV/c proton fluence for standard and carbon- and oxygen-enriched n-type FZ silicon sensors. A reduction of the β_{Aeff} (with a value of 0.0044 cm^{-1} if measured as a function of 24 GeV/c proton, or 0.007 cm^{-1} for 1 MeV neutron equivalent fluence) is clearly seen in oxygen-enriched (to [O]~2×10^{17} cm^{-3}) wafers with respect to the standard ($\beta_{Aeff} = 0.0154$ cm^{-1} with proton corresponding to 0.025 cm^{-1} with 1 MeV n_{eq} for a [O] < 1×10^{16} cm^{-3}) and carbon-enriched ones ($\beta_{Aeff} = 0.0437$ cm^{-1} with protons or 0.07 cm^{-1} with n_{eq}). Unlike this result after proton irradiation, no significant effects have been measured with oxygen-enriched silicon sensors after neutron irradiation [5,11]. Also, no effect on changing the degradation rate of the reverse current (both after charged hadron or neutron irradiations) has been measured with any type of impurity-enriched silicon sensor with respect to the standard high-purity detector-grade FZ silicon (Figure 1.2). Although these electrical parameters are important for defining the properties of irradiated silicon, they don't give any direct indication of the expected performances of the detectors.

The evolution of V_{FD} has often been used as a qualifying parameter, because in nonirradiated detectors the capacitance versus voltage characteristic saturates at the same value of the charge collection. The ability of biasing the detector at voltages above V_{FD} was considered a criterion for assessing the functionality of the device. It can be extrapolated from Figure 1.1 that V_{FD} exceeds 1,000 V (for a 300 μm detector) above a fluence of about 2×10^{15} n_{eq} cm^{-2} in the more favorable case (oxygenated silicon irradiated with proton). If V_{FD} was the qualifying parameter it is evident that silicon detectors could hardly be used after this fluence due to the practical limitation to provide bias voltages ≥ 1,000 V in any large detector system.

1.2.2 Charge Trapping and Charge Collection

The radiation-induced defects also contribute to the reduction of the signal height by means of trapping of the signal charge carrier for a time exceeding the charge collection time. The density of the defects increases linearly with fluence; therefore, the trapping is assumed to have the same trend. In fact, it can be described by a parameter called the trapping time, τ_{tR}, which is inversely proportional to the concentration of trapping centers and therefore to the fluence:

$$\frac{1}{\tau_{tr}(\phi)} = \frac{1}{\tau_{tr}(\phi = 0)} + \beta_{tr}\phi \tag{1.11}$$

The trapping for nonirradiated detectors is order of magnitudes higher than τ_{tr} even after moderate doses of hadron irradiation, and the first term of Equation (1.11) can be neglected. The measured values of β_{tR} range from 0.41×10^{-6} cm^2 s^{-1} and 0.62×10^{-6} cm^2 s^{-1} for electrons and holes, respectively [12].

The signal is influenced to a different extent depending on the readout geometry of the irradiated detectors. The silicon detectors are made by high-density doping implants (n^+ and p^+) on low doped bulk to form the diode junction and the ohmic contact. P-in-n sensors have a p^+ diode side and n-in-p sensors an n^+ diode side. As shown in Figure 1.4, due to the introduction of defects acting as acceptor doping with the radiation damage, the silicon bulk is effectively *p*-type after a few 10^{13} cm^{-2} of hadron irradiation, irrespective of the original space charge sign. The electric field of inverted n-type silicon detectors is stronger under the implanted n^+ contact [13], like in *p*-type bulk detectors, although a narrow region with high field is present also on

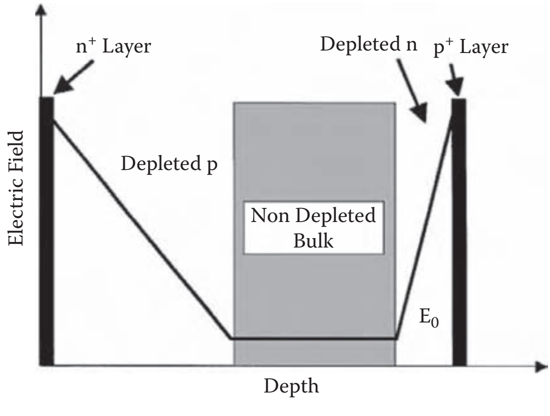


FIGURE 1.4 Schematics of the electric field profile of an irradiated and inverted n-type (or a *p*-type) silicon detector. The stronger electric field has shifted from the original junction side (under the n^+ implant) to the former ohmic side (n+ implant).

the p$^+$ contact (this effect is usually called double junction)[13,14]. Due to the profile of the electric field, it is advantageous in terms of charge collection to segment and read-out the detectors from the n-side. In fact, the signal deficit caused by trapping of the charge carriers at radiation-induced defect centers can be described by the following equation:

$$N_{e,h}(t) = N_{e,h}(0)\exp\left(-\frac{t_c}{\tau_{tr,e,h}}\right) \tag{1.12}$$

where $N_{\mathrm{e,h}}$ is the number of collected charges (electron or holes, respectively), $N_{\mathrm{e,h}}(0)$ is the number of ionized electrons and holes, t_{c} is the collection time, and $\tau_{\mathrm{tr,e,h}}$ is the electron and hole effective trapping time.

From Equation (1.12) it is clear that a shorter t_{c} provides substantial advantages in terms of collected charge. If segmented n-type silicon devices are read out from the n$^+$ side (n-in-n detectors), the signal will mainly be formed by the electron current and will benefit from a shorter t_{c}, with respect to the standard p-in-n, due to the faster electron collection (higher mobility carriers moving in the higher electric field with respect to holes) [15,16]. The n-side readout is, however, more readily achieved with *p*-type substrates (n-in-p), where no inversion takes place with irradiation and the main electric field is always located on the original n$^+$-p junction side. A difference between n-in-n and n-in-p detectors is that in the first case double-sided processing is needed to implant edge protection structures on the back side (guard rings) with significant impact on the complexity and cost (up to 50% higher) of the processing with respect to *p*-type substrates that require only front-side guard ring implants [17]. This turns out to be a very important factor for experiments where a large coverage area is required, due to the cost reduction and easier handling of single-sided devices. The expected improvement of the charge collection properties and the simpler processing of the *p*-type substrate n-side readout sensors motivated the proposal of this type of devices as competitive radiation tolerant detectors [18,19].

1.3 ASSESSING THE RADIATION HARDNESS OF SILICON DETECTORS

A possible definition of radiation tolerance is the maximum fluence at which the devices are fully functional. This definition is certainly setup and device dependent. A considerable difference is found between *pad diodes* and the more complex finely segmented detectors. These latter typically have one or both the electrodes segmented in individual diode structures with at least one dimension comparable to or smaller than the thickness of the sensor: microstrip detectors have the readout contact realized in the strip with a typical pitch size from under 40 μm to a few hundred μm, and pixel sensors can be made in a 50 μm × 50 μm size or larger. The typical thickness of silicon detectors varies from 300 μm to over 1,000 μm. The segmented sensors are used in fast and precise tracking systems thanks to their speed and resolution. The degradation of their functionality implies quantities like the reduction of the signal size, the increase of the electronics noise, and the possible deterioration of the resolution. In the following we neglect the latter because it is a convolution of many effects, also involving the electronics readout.

The increase of the electronics noise is dependent on the performance of the readout electronics, the geometrical parameters of the detector (e.g., through the input capacitance of the individual channels), and the operating conditions (e.g., the temperature, determining the intensity of the reverse current and its contribution to the noise). All these parameters are known in the design phase of the detector system. The main parameter to be determined for predicting the functionality of the irradiated sensors is the reduction of the signal with fluence. A smaller signal yields a reduced efficiency. The concept of acceptable efficiency varies in different applications (e.g., 50%–80% in the case of gamma rays imaging in space experiments [20] to close to 100% in particle physics ones [21,22]). Besides, the efficiency after irradiation is very dependent on the applied bias voltage (V_B), and different systems can tolerate very diverse ranges of applicable V_B's: a limited range with low maximum bias voltage is acceptable for experiments in space, while voltages up to 1,000 V are possible in high-energy or nuclear physics accelerator experiments. In the following we discuss the radiation tolerance of finely segmented detectors designed for high-energy physics (HEP) as a representative example. This application is particularly demanding for it requires accurate position sensitivity and tracking efficiencies close to 100% for particles at the minimum ionization energy (mips). These particles yield a small signal (the ionized charge is about 79 electrons per micron of silicon crossed by a mip); its degradation as a function of fluence could lead to an early failure of the tracking system.

1.3.1 Silicon Detectors and High-Energy Physics Experiments: A Success Story

Since planar processing was developed around 1980 ([23]), finely segmented (pixel and microstrip) silicon detectors have been used to cover ever larger areas in high-energy physics experiments (see, e.g., [24] for a review). The reason for this success is their low mass, high speed, and resolution, which allow for fast three-dimensional (3-D) reconstruction of charged particle tracks in very high multiplicity collision events. All the more recent HEP experiments use mainly silicon sensors for tracking and vertexing. In particular, all four major experiments hosted by the Large Hadron Collider (LHC) [25] at the CERN accelerator complex in Geneva use an unprecedented number of silicon detectors [21,22,26,27]. The LHC is presently the frontier machine, with 14 TeV center of mass energy interactions of proton-proton bunches colliding at 40 MHz, for a design luminosity of 10^{34} cm^{-2} s^{-1}. The radiation environment created by the multitude of relativistic particles emerging from the p-p collisions and the intense flux of neutrons (with average energy of about 1 MeV) backscattered from the calorimeter regions of the experiments are extremely challenging. The segmented silicon detectors (microstrip and pixel sensors) developed for this machine have been designed to survive 10 years worth of experiments, with a maximum final fluence of about 2×10^{15} n_{eq} cm^{-2}. Their radiation tolerance requirements have stimulated a significant effort in radiation hardening of silicon detectors for well over the 10 years preceding their installation in 2007. Upgrading this machine (the Super-LHC, SLHC) presents a new, more severe challenge to a factor of 10 higher luminosity [28], with a similar increase in the expected radiation tolerance: the SLHC sensors will have to operate up to about 2×10^{16} n_{eq} cm^{-2}.

An efficient way of defining the radiation tolerance of the detectors for this application is the signal-over-noise (S/N) ratio. An S/N of 10 is considered a safe limit for high-efficiency (percentage of identified tracks) and high-purity (rejection of fake events) mip tracking in HEP experiments (where over 98% track hit efficiency and a few 10^{-5} noise occupancy hits are required). To have an idea of the expected noise, the present electronics for microstrip detectors has typically

$$ENC = 400 + 45 \times C_{load} \tag{1.13}$$

where the equivalent noise charge (*ENC*) is expressed in electrons, and C_{load} is the input capacitance to the individual electronics channels in picofarads (pF). A typical value of C_{load} for 80 μm pitch silicon strips on a 300 μm thick silicon detector is about 0.7 pF/cm. Another contribution to the *ENC,* which is to be added in quadrature to Equation (1.12), comes from the shot noise induced by the fluctuations of the reverse current:

$$ENC = \sqrt{12 I_R t_{SH}} \tag{1.14}$$

with I_R in nA and t_{SH} in ns. The estimate of the shot noise contribution is performed by evaluating the reverse current at the operating temperature per electronics channel after the appropriate dose of irradiation. The missing ingredient to determine the functionality of the sensors is the degradation of the signal with fluence. This parameter is largely independent of the particular readout system; thus, its evaluation allows the prediction of the radiation tolerance of every silicon detector system. To give a specific example of radiation environment we can use the layout of the upgraded ATLAS experiment at the SLHC [29]. This detector has cylindrical geometry around the axis of the colliding proton beams and comprehends four layers of pixel detectors from 3.8 cm to 21 cm and five layers of microstrip detectors from 38 cm to 100 cm cm radii from the beam line. The most demanding radiation tolerance is for the pixel sensors, with the expected fluences ranging from 2×10^{16} n_{eq} cm^{-2} (~1 Grad) to 3×10^{15} n_{eq} cm^{-2} (150 Mrad) from the inner to the outer layer. The innermost microstrip sensors will receive a final fluence of 1×10^{15} n_{eq} cm^{-2} (50 Mrad).

The evaluation of the signal as a function of the hadron fluence and the estimate of the noise on the basis of the known geometry and the measured reverse current after irradiation allow the determination of the achievable S/N ratio at the different radii at the end of the physics program of the experiment.

1.3.2 Radiation Hardening of Silicon Detectors

As previously mentioned, the radiation hardening of silicon sensors can be pursued by exploiting different geometry options and by engineering the silicon crystal bulk with impurities. Figure 1.1 showed some advantages offered by oxygen-enriched silicon after proton irradiation. Some possible enhancements of the charge collection of detectors made with silicon crystals other than the high-purity FZ (i.e., the standard for detector applications) are discussed in this section, but the more important results in terms of radiation hardening of silicon sensors were obtained by changing the geometry of the readout segmented electrodes. It has been previously discussed that the electron signal (n-side readout) allows for a shorter collection time with predicted benefits in terms of reduction of the charge trapping. Nonetheless, tracking planar silicon sensors have been traditionally produced by implanting p^+ segmented structures on n-type high-resistivity silicon. This is due to the simplicity of this technology, because no special care for isolating the implanted structures has to be taken, unlike with segmented n^+ implants. A low-resistivity connection between n-implants is caused by an electron accumulation layer at the interface Si-SiO_2 created by the positively charged oxide. This positive charge increases with irradiation (until saturation), as briefly mentioned before. The electron layer can be interrupted, and the n-electrodes (strips, pixels) isolated, by means of dedicated structures: p-spray (a shallow homogeneous p-implantation over the whole wafer [30]); p-stop implants (photo-lithographically defined p-implants between the readout strips [31]); or a combination of the two methods [32].

Using one of these methods, n-side readout devices can be produced. The extra processing complication has actually a large payout in terms of performances after irradiation.

Figure 1.5 shows the comparison of the charge collected by n-in-n, n-in-p, and a p-in-n sensor after 3 and 12 × 10^{14} n_{eq} cm^{-2}. After the lower dose, comparable performances (although the n-side readout shows superior charge collection at low bias voltages) are measured, while after the higher fluence the p-side readout is considerably lower performing. The remarkable difference in signal between the two readout sides is measured despite a similar value of V_{FD} after the same dose of irradiation. It is apparent that the effect of the n-side readout is not dependent (for radiation doses at which the *n*-type silicon is inverted) on the substrate, and n- and p-silicon substrates perform equally well with this type of readout. The use of *p*-type substrates is a simpler and cheaper processing for implementing the n-side readout. The extra complexity and cost involved in processing n-in-n single-sided sensors would be justified only by a substantial advantage in terms of radiation hardness, but no appreciable difference is measured between n-side readout sensors made with the two types of FZ substrates.

1.3.3 Radiation Tolerance of n-Side Readout Sensors

Having proved that irradiated n-side readout sensors deliver a considerably superior signal than p-side readout ones, this paragraph aims to establish in absolute terms the signal degradation with fast hadron fluence with this type of detector made with high-resistivity FZ materials (both n and p conductivity type). The results reported here were obtained after irradiations with charged and neutral hadrons to investigate possible differences in the charge collected as a function of the bias voltage (CC(V)) due to the different nature of the damage. The fluences are all given in 1 MeV n_{eq}

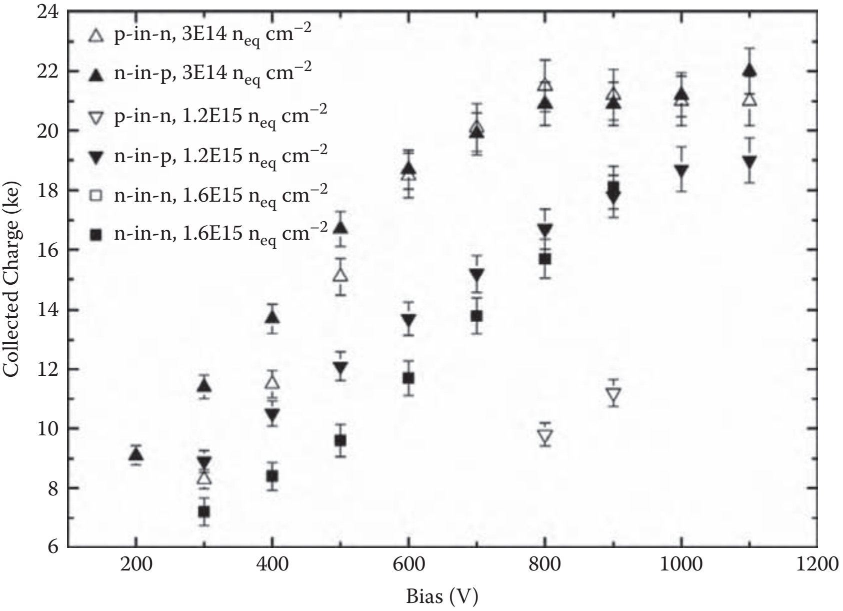


FIGURE 1.5 Signal as a function of the bias voltage for n-in-n and n-in-p silicon microstrip sensors after 24GeV/c proton irradiation to 3 and 12 × 10^{14} n_{eq} cm^{-2}. Also an n-in-n sensor irradiated to 16 × 10^{14} n_{eq} cm^{-2} is shown. Taking into account the difference in the irradiation dose, the n-in-n and n-in-p detectors behave very similarly. The significantly smaller signal delivered by the p-in-n sensors is well noticeable.

using the standard NIEL normalization. The detectors used for the measurements presented here are ~1 × 1 cm^2, 80 μm pitch silicon sensors (128 channels) coupled with 25 ns shaping time electronics, 40 MHz clock rate (SCT128 [33]). The measurements were performed at about –25°C to control the reverse current. The minimum ionizing particle signal is mimicked by fast electrons from a ^{90}Sr, and the signal is expressed in number of collected electrons at the most probable value of the distribution of the charge ionized in silicon by a passing mip (Figure 1.6).

Figure 1.7 shows the CC(V) of n-in-p sensors after various reactor neutron [33,34] doses, up to 2 × 10^{16} n_{eq} cm^{-2} (about 1 Grad). This dose is well in the range for qualifying the sensors to the most demanding radiation tolerance for the aforementioned high-energy physics application (SLHC). The signal degradation with fluence is clear. Nonetheless, approximately 5,000 e^- are still collected at 1,000 V after the highest fluence.

Figure 1.8 shows the CC(V) of n-in-p sensors after various 26 MeV and 24 GeV/c proton [34,35] doses, up to 2.2 × 10^{16} n_{eq} cm^{-2}. A similar discussion as for the neutron irradiations also holds for these measurements. A charge of about 5,000 e^- is collected here at 1,000 V after 1.6 × 10^{16} n_{eq} cm^{-2} and about 4,000 e^- after 2.2 × 10^{16} n_{eq} cm^{-2}. It can be noted that after corresponding NIEL normalized fluences, the CC(V) of detectors irradiated with both energy protons are well comparable, indicating a similar damage from the relatively low energy and the relativistic protons.

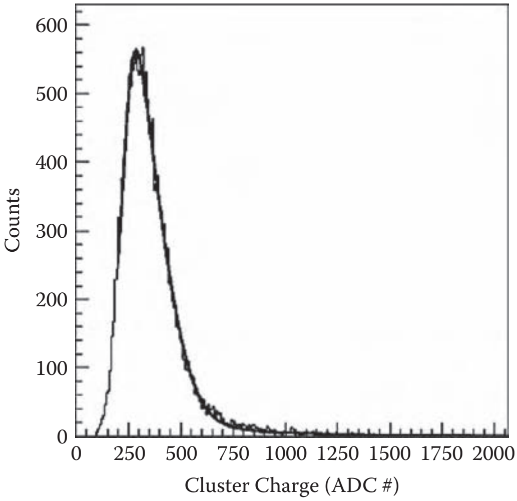


FIGURE 1.6 Charge distribution deposited by a minimum ionizing particle in silicon. The most probable value of this distribution corresponds to ~24000 electrons in 300 μm silicon.

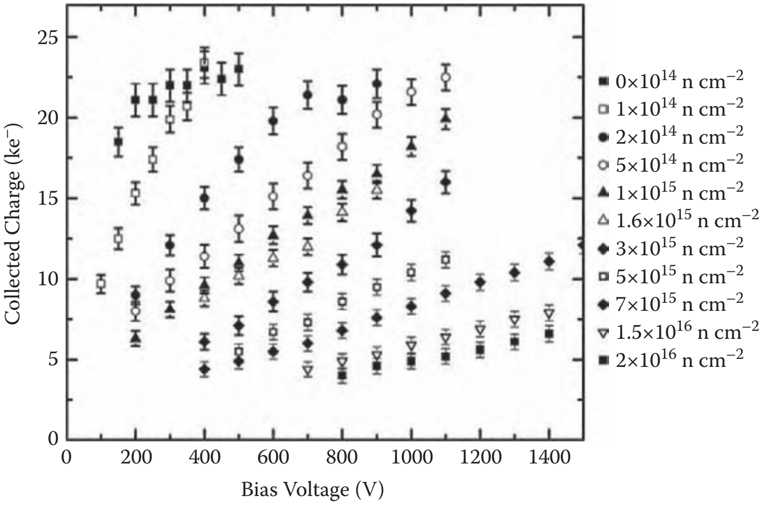


FIGURE 1.7 Degradation of the signal as a function of the bias voltage for n-in-p silicon microstrip sensors irradiated with neutrons to various doses up to 2×10^{16} n_{eq} cm^{-2}.

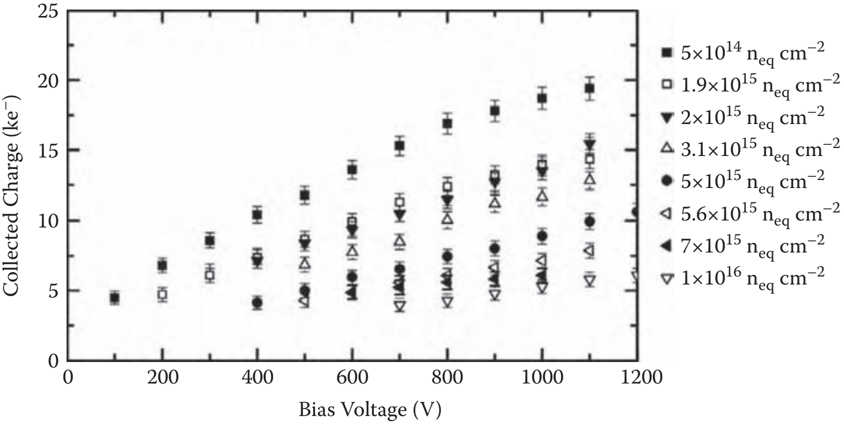


FIGURE 1.8 Degradation of the signal as a function of the bias voltage for n-in-p silicon microstrip sensors irradiated with 26 MeV and 24GeV/c protons to various doses up to 2.5 × 10^{16} n_{eq} cm^{-2}.

A summary of the degradation of the signal measured at 500 V and 900 V as a function of neutrons, and protons with different energies and 280 MeV pions [37] is shown in Figure 1.9. The decrease of the signal after the various types of irradiation shows a good degree of agreement within the experimental errors on the fluence and on the measured charged.

It is important to stress that the collected charge is well above the expectations derived by the measurement of the increase of the charge-trapping probability. Using the parameterization in [12], the expected charge collection distance after 1 × 10^{16}

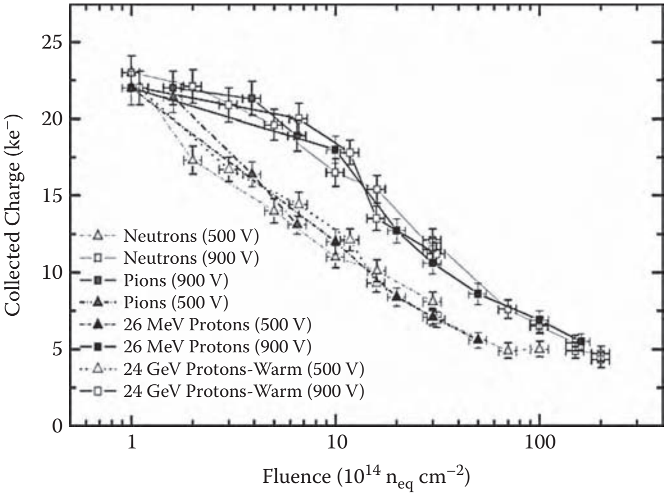


FIGURE 1.9 Degradation of the signal of n-in-p silicon microstrip sensors at 500V and 900V as a function of the n_{eq} fluence of various energies protons, pions and neutrons.

n_{eq} cm^{-2} would be less than 30 μm, for less than 2,400e$^-$ maximum collected charge. The much bigger size of the signal at high doses indicates that some mechanism is enhancing the charge collection after heavy irradiation. Possible explanations are a nonlinear dependence of the charge trapping after high doses, a possible field-enhanced fast detrapping, or a controlled charge multiplication at high electric fields in irradiated devices [38].

1.3.4 Effect of Varying the Detector Thickness

The signal generated by a mip crossing a silicon detector is proportional to the path length of the particle in the sensor, and the ionized charge is about 79 e/μm. This charge is entirely collected by a nonirradiated detector when the device is biased at or above full depletion voltage, thanks to the presence of a strong electric field in the entire volume of the sensor. After irradiation, two effects contribute to reduce the collected charge: (1) the charge carrier trapping at radiation-induced defect centers in the crystal; and (2) a lower electric field (for the same bias voltage) due to the increased effective space charge. In particular, after heavy hadron irradiation, a substantial electric field might occupy a volume smaller than the detector (on the junction side) even for very high applied voltages. In this situation the effective charge collection distance (CD), influenced by these two effects, is shorter than the detector thickness and progressively reduced by further irradiation. At high doses the CD defines the sensitive volume of the sensor. Thinner detectors could have an advantage with respect to the standard thickness (300 μm) when the CD is shorter than this value, due to a possible higher average electric field over the active volume of the irradiated sensor for the same applied voltage. To explore this possibility, the charge collection properties of standard (300 μm) and thinned (140 μm) n-in-p readout sensors have been studied after various doses of hadron irradiation up to 2 × 10^{16} n_{eq} cm^{-2} [39,40]. Different behaviors of the thin and thick devices are reported at different fluences. At the lower doses investigated (up to 3 × 10^{15} n_{eq} cm^{-2}), all the devices collect a similar charge with increasing bias voltage, until the charge collected by the thinner devices saturates (see an example in Figure 1.10). It should be noticed that the saturation value is very close to the maximum preirradiation charge (12,000 e$^-$), so to the entire ionization charge released by a mip. After 7 × 10^{15} n_{eq} cm^{-2}, the thin device is slightly more efficient in collecting charge at corresponding bias voltages up to a maximum of 1,100 V (Figure 1.11). Above this voltage, the thick sensor delivers a higher signal (indicating a CD larger than 140 μm), up to the maximum applicable voltage of 1,900 V. This remarkably high bias voltage can be applied only after the device has been irradiated (breakdown voltages of less than 500 V are usually measured for this type of nonirradiated device). The slight improvement exhibited by the thinner sensors in the lowest part of the CC(V) curve is confirmed after ever higher fluences.

Figure 1.12 shows the CC(V) of 140 μm and 300 μm thick n-in-p sensors after extremely high doses: 1, 1.5, and 2 × 10^{16} n_{eq} cm^{-2}. The data show a better charge collection at every applicable bias voltage (before thermal runaway) from the thin devices. A higher value ranging from about 10% after the lowest to over 25% after the two higher fluences is measured with the 140 μm thick *p*-type sensors. The highest measured charge corresponds to a mip ionization distance of 110 μm at 900 V after 1

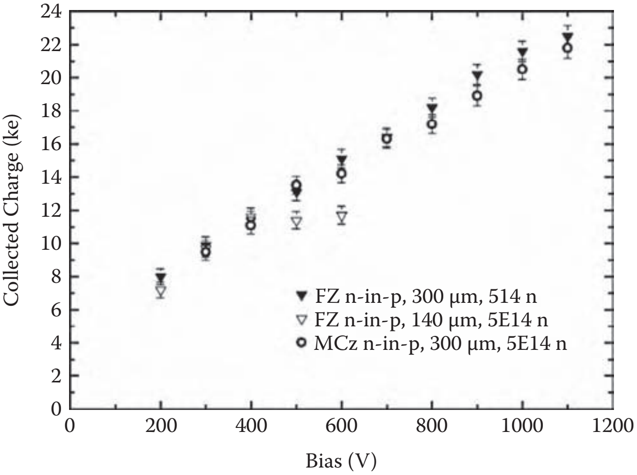


FIGURE 1.10 Collected charge as a function of the bias voltage for thin (140 μm) and standard (300 μm) n-in-p detectors irradiated to 5 × 10^{14} n_{eq} cm^{-2}. The charge collected by the thin sensor saturates just under 12,000 electrons (the pre-irradiation value for this detector thickness).

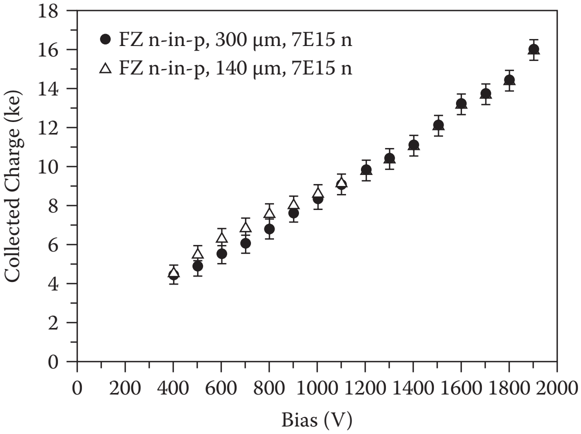


FIGURE 1.11 Collected charge as a function of the bias voltage for thin (140 μm) and standard (300 μm) n-in-p detectors irradiated to 7 × 10^{15} n_{eq} cm^{-2} (left) and 1 × 10^{16} n_{eq} cm^{-2} (right).

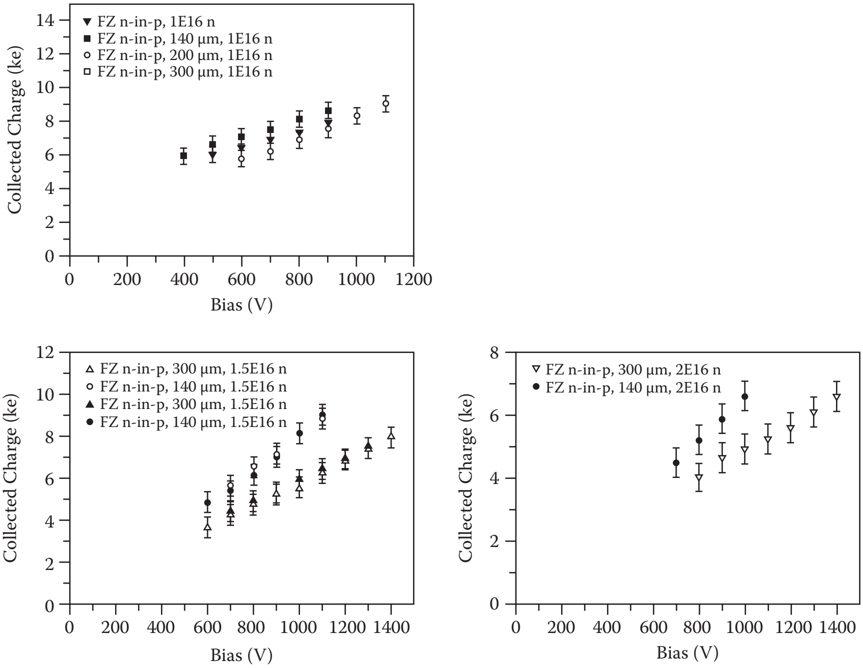


FIGURE 1.12 Collected charge as a function of the bias voltage for thin (140 μm) and standard (300 μm) n-in-p detectors irradiated to 1, 1.5 and 2 × 10^{16} n_{eq} cm^{-2}.

× 10^{16} n_{eq} cm^{-2}, 110 μm at 1,100 V after 1.5 × 10^{16} n_{eq} cm^{-2}, and 75 μm at 900 V after 2 × 10^{16} n_{eq} cm^{-2}, if no trapping effects are taken into consideration. On the other hand, thin devices can tolerate a lower applied bias voltage than the thicker ones. In the case of the sensors irradiated to the two highest fluences the 300 μm sensors could be measured up to 1,400 V, while no more than 1,100 V could be applied to the thin ones.

1.3.5 Reverse Current in Heavily Irradiated Thin and Standard Silicon Sensors

As said before, the reverse current of irradiated silicon detectors increases linearly with the irradiation fluence and is proportional to the volume of the silicon sensor. The increase of the reverse current is an important limiting factor to the operation of the irradiated sensors. As shown before, the charge collected by irradiated sensors keeps increasing as a function of the applied bias voltage. If an arbitrarily high voltage could be applied to heavily damaged devices, a significant amount of charge could be recovered and the lifetime of the detectors notably extended. Excluding possible practical limits of real detector systems (where routing voltages higher than 1,000 V could raise technical difficulties), the limitation to the maximum applicable voltage comes from the thermal runaway of the reverse current. In fact, it is experimentally verified that a destructive breakdown of the junction does not take place in irradiated devices, because it is preceded by the rapid increase of the reverse current

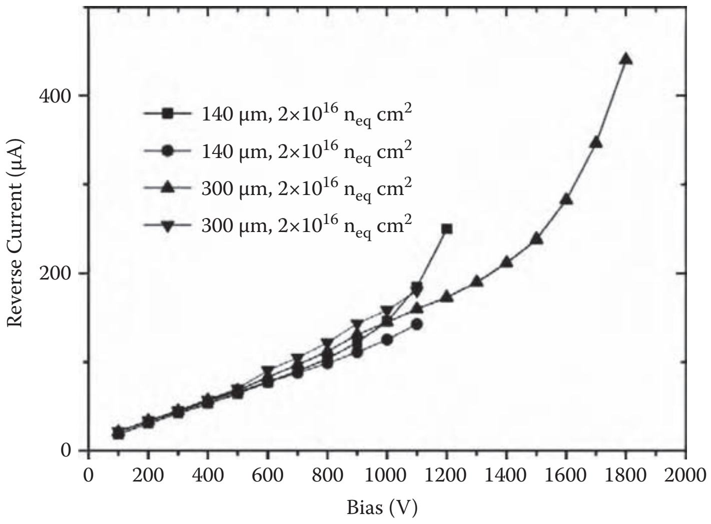


FIGURE 1.13 Reverse current as a function of the bias voltage for 140 and 300 μm thick silicon microstrip detectors irradiated to 7 × 10^{15} n$_{eq}$ cm^{-2}.

with time. The mechanism is due to the increased current drawn by a warmer detector. This excess current in turn generates heat, further increasing the reverse current. This thermal runaway takes place when the cooling is not capable of removing all the heat generated within the sensor. The failure to provide adequate bias for collecting the minimum signal due to the excess reverse current could be the first failure mode for irradiated sensors. The control of I_R is therefore a very important parameter for detector systems under heavy radiation. The use of low temperatures to control the current is the first method to limit I_R. On the other hand, the use of thin silicon could be envisaged to this purpose because I_R is proportional to the volume. This idea has been tested with 140 μm and 300 μm thick sensors [39], and the typical results are shown in Figure 1.13. Up to medium-high doses (aound 7 × 10^{15} n$_{eq}$ cm^{-2}) the reverse current is essentially equal for both thicknesses of sensors up to the point that the sensitive volume of the thicker device becomes larger than the thin detector. But after the very high doses the differences between the current of the two thickness are somewhat unexpected (Figure 1.14). The current of the thin devices is always higher, and a thermal runaway occurs at 1,100 V, preventing stable operation of the devices. The thermal runaway also takes place for the thicker sensors, but at a higher value of the bias voltage, although at a similar value of the reverse current in both cases. It can be concluded that 140 μm thick devices have similar performances (or they display the same CC(V) and reverse currents) to standard thickness sensors at lower bias voltages (about 400 V lower) for hadron irradiation fluences above 1 × 10^{16} n$_{eq}$ cm^{-2}.

1.3.6 Radiation Tolerance of Different Single Crystal Silicon

As suggested already, the different content of some impurities (mainly oxygen and carbon) in the silicon crystal can affect the degradation rate of some of the electrical

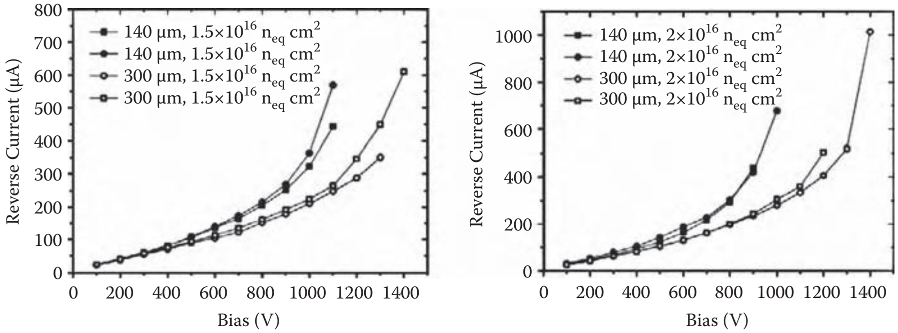


FIGURE 1.14 Reverse current as a function of the bias voltage for 140 and 300 μm thick silicon microstrip detectors irradiated to 1.5 and 2 × 10^{16} n_{eq} cm^{-2}.

properties of silicon diodes. In terms of performances of tracking detectors, it is important to investigate if the advantages are seen also with charge collection measurements with segmented sensors.

1.3.6.1 MCz Silicon

The high-purity FZ method yields high-resistivity wafers with very low concentrations of O (~10^{16} cm^{-3}) and other impurities. A much higher O concentration is found in wafers produced with the Cz method (~10^{18} cm^{-3}). This type of single crystal is usually employed by the microelectronics industry, but its very low resistivity makes it unsuitable for particle detectors where low V_{FD} (therefore high resistivity) values are required. A novel production technique that combines the classical Cz method with a magnetic field (MCz) to reduce the concentration of impurities, allowing higher resistivity in the single-crystal ingot, has been made available by industry in recent years [41]. The O concentration in MCz silicon is about 4 to 5 × 10^{17} cm^{-3}. n- and *p*-type single-crystal ingots with relatively high resistivity (1–2 kΩ cm) have been produced with this method and proposed as a possibly radiation harder substrate for silicon detectors (see, e.g., [42]). The degradation rate of the V_{FD} for this type of sensor irradiated with neutrons is sensibly lower than for standard and oxygen-enriched FZ silicon [43]. Also, CC(V) measurements confirm the enhanced radiation tolerance of this material after neutron irradiations. Figure 1.15 shows the comparison of the CC(V) of n-side readout detectors made on n- and *p*-type substrates grown with both FZ and MCz methods and irradiated with neutrons to 5 and 10 × 10^{14} n_{eq} cm^{-2} [44]. The nMCz detector exhibits the fastest rising signal with bias voltage, resulting therefore in the most performance after these doses. It can be noticed that the maximum collected charge is essentially the same for all materials. This indicates that the charge trapping is material independent and the improvement found with the nMCz crystal is due to a lower degradation of V_{FD}.

In high-energy physics applications, and, namely, in the SLHC environment, the radiation field is mainly composed of backscattered neutrons and charged particles emerging from the interactions. Their ratio will depend on the radial distance from the beam axis, with the charged hadron dominating at low radii and the 1 MeV

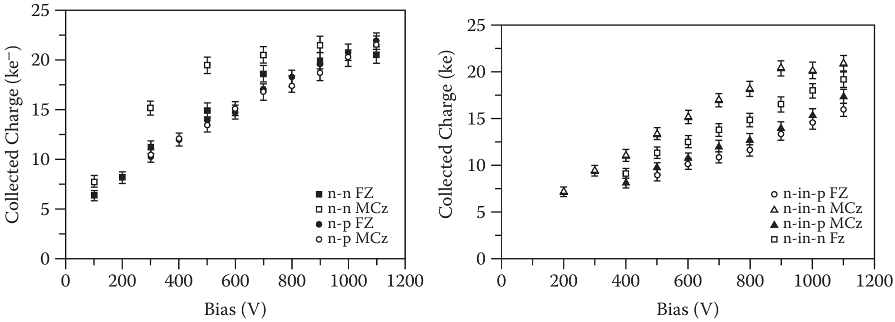


FIGURE 1.15 Collected charge as a function of bias with the four type of detectors used for this study after 1 and 5 × 10^{14} neutrons cm^{-2}. All detectors are configured to read-out on segmented n-implant electrodes.

neutrons being the main cause of damage at outer ones. The radius at which the two components are equal is about 25 cm. It has been found that these two types of radiation introduce a different ratio of donor (n-type) or acceptor (*p*-type) defects on different silicon crystals. In the case of FZ sensors, both radiations introduce predominantly *p*-type defects. In the case of n-MCz, the neutrons introduce mainly *p*-type defects, while charged particles mainly n-type defects. This effect was measured by capacitance-voltage measurements on diodes irradiated with one or the other type of radiation [45]. This particular feature of the n-MCz silicon could be advantageous for detectors exposed to radiation fields composed of a comparable mix of neutrons and charged hadrons because the n- and *p*-type radiation-induced defects can partially compensate. To test this effect, n-in-p FZ and n-in-n MCz detectors have been irradiated with neutrons only and n-in-n FZ and n-in-n MCz detectors with an equal mix (5 × 10^{14} n_{eq} cm^{-2}) of neutrons and 26 MeV/c protons for a total dose of 1 × 10^{15} n_{eq} cm^{-2} for every sensor. Figure 1.16 shows the CC(V) measurements of these devices and confirms the compensation effect. The two FZ detectors exhibit almost identical CC(V) characteristics after both the neutron and mixed irradiation, while the n-MCz shows a faster rise of the CC(V) in the case of mixed irradiation when compared with an identical detector irradiated with neutrons only [46]. This feature of the n-MCz material can efficiently extend the lifetime of detectors in applications where silicon tracker sensors have to operate after very high doses of a mixed irradiation field.

1.3.6.2 Epitaxial Silicon

Epitaxial grown silicon layers (Epi) on low-resistivity Cz substrates are commonly used in the complementary metal-oxide semiconductor (CMOS) electronics industry. The typical thickness of the Epi layer is < 10 μm. The use of thicker layers of this material for processing enhanced radiation hard detectors was proposed in [47]. Encouraging results have been obtained in terms of reduced degradation of the full depletion voltage with fluence. The main disadvantage with this material is the cost and the difficulty of growing thick (for epitaxial growth standards) layers of single-crystal, relatively high-resistivity (> 500 Ω cm) silicon. The control of the growth-

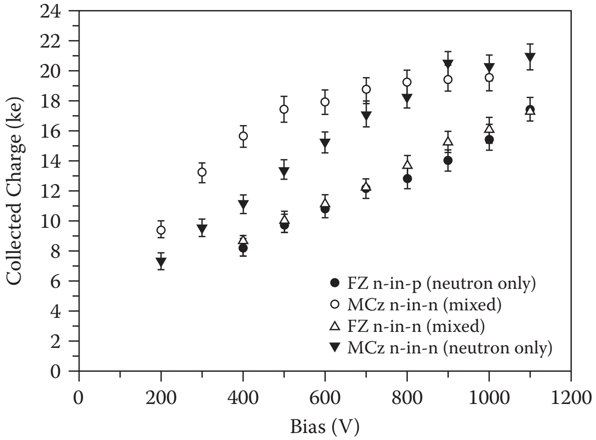


FIGURE 1.16 CC(V) of n-in-n and n-in-p FZ and n-in-n MCz detectors irradiated with neutron only or with an equal dose of neutrons and 26 MeV protons to the same total dose of 1 × 10^{15} n_{eq} cm^{-2}. The compensation effect of acceptor and donor-like defects introduced by the two different types of radiation in the nMCz substrate is visible in the faster rise of the CC(V) in the case of mixed irradiation. (G. Casse, A. Affolder, P.P. Allport, "Studies on Charge Collection Efficiencies for Planar Silicon Detectors after Doses up to 1016 Neq/cm2 and the Effect of Varying Substrate Thickness," 2008 Nuclear Science Symposium, 19–25 October 2008 Dresden, Germany, http://www.nss-mic.org/2008/Program/ListProgram.asp?session=N54, to be published in *IEEE Trans. Nucl. Sci.*)

processing parameters becomes problematic, and the yield and reproducibility of the Epi wafers can be affected. Nevertheless, *p*- and *n*-type Epi silicon layers up to 150 μm, 500 Ω cm have been obtained on 4-inch Cz wafers. Some of these wafers have been processed to produce 1 × 1 cm^2, 80 μm pitch microstrip sensors. A few devices were selected and irradiated to various doses of reactor neutrons [40]. In general, high reverse currents were measured with the Epi irradiated detectors, but it was possible to select a few samples to measure the charge collection properties of n-in-p detectors up to 8 × 10^{15} n_{eq} cm^{-2} and p-in-n detectors up to 3 × 10^{15} n_{eq} cm^{-2} (Figure 1.17).

The Epi sensors are able to collect almost the same charge as before irradiation up to 3 × 10^{15} n_{eq} cm^{-2}, though at increasing bias voltages. The degradation of the p-in-n devices is faster than the n-in-p, as expected. The p-in-n Epi have, however, a much reduced degradation rate with respect to p-in-n FZ sensors. Also, the n-in-p Epi exhibit a reduced degradation rate with respect to FZ detectors of comparable thickness (Figure 1.18).

The epitaxial grown substrates display better relative performances after irradiation than any other silicon substrate. Despite these good results, the cost and the reproducibility of the performances with thick Epi layers make these substrates unlikely to be competitive for high-energy physics applications where the sensitive volume is required to be over 100 μm thick. Also, epitaxial detectors are expected to demonstrate enhanced performance if operated in a mixed (charged and neutral particles) irradiation field due to the opposite sign of the effective space charge induced

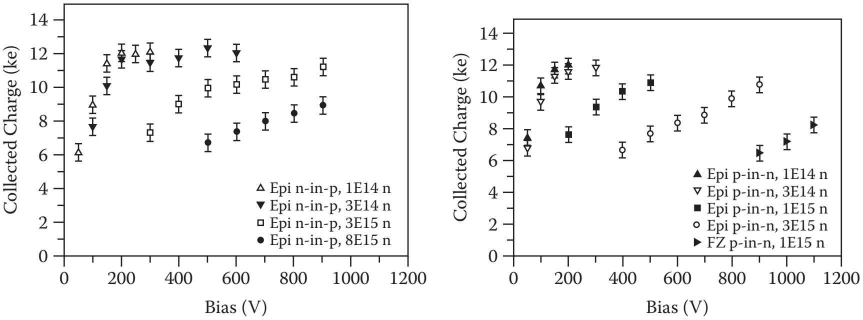


FIGURE 1.17 Collected charge as a function of the bias voltage for Epi (150μm) n-in-p detectors irradiated to various doses up to 8 × 10^{15} n_{eq} cm^{-2} and p-in-n detectors to 3 × 10^{15} n_{eq} cm^{-2} [40]. (With permission from G. Casse, A. Affolder, P. P. Allport, "Charge collection Efficiency Measurements for Segmented Silicon Detectors Irradiated," *IEEE Trans. Nucl. Sci.*, vol. 55, Issue 2, 2008, pp1695–1699.

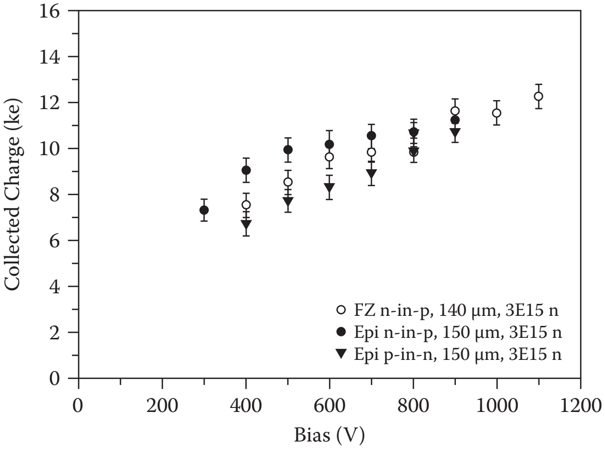


FIGURE 1.18 Comparison of n-in-p Epi and FZ sensors with similar active thickness after 3 × 10^{15} n_{eq} cm^{-2}.

by the two types of radiation [45], as for the nMCz substrate, although CC(V) measurements of this effect are not available.

1.4 ANNEALING EFFECTS

After Equation (1.5), the reverse current of the irradiated silicon sensors could be reduced by annealing. The annealing is a strong function of temperature and can be effectively used to suppress or accelerate this effect. Figure 1.19 shows the measured changes of the reverse current as a function of the annealing time. The data are shown in terms of equivalent annealing time at 20°C, but the measurements have been taken after accelerated annealing steps at 60°C (acceleration factor about 540) and 80°C (acceleration factor about 7,400) [49,50].

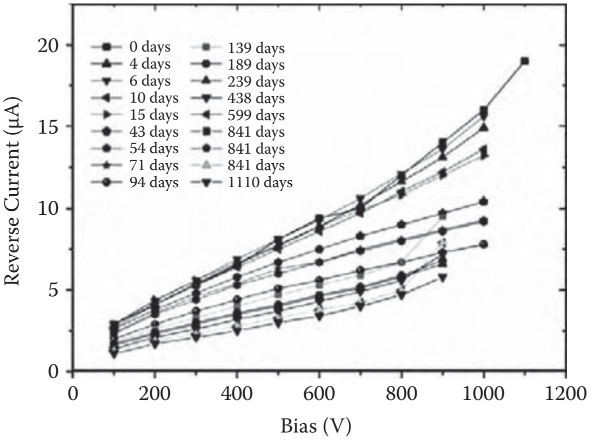


FIGURE 1.19 Changes of the reverse current as a function of the bias voltage, measured at –25°C, of a silicon detector irradiated to 1 × 10^{15} n_{eq} cm^{-2} with different annealing time after irradiation. The sensor was annealed at 60°C but the time are rescaled to equivalent time at T_{REF}.

As shown, a substantial reduction of the current is achieved (between 30% to 40% at high voltage after ~50 days at 20°C). In certain applications (e.g., high-energy physics experiments) the reduction of the current by means of annealing is not exploited because of the reverse annealing of the V_{FD}. This has always been considered detrimental to the sensor operations because of the increase of V_{FD} after a short-term decrease for about 30 days at T_{REF} (Figure 1.3). It was assumed that this increase would entail a corresponding decrease of the charge collected at a given voltage below V_{FD}. But studying the charge collected at a given voltage as a function of the accelerated (at 60°C and 80°C) annealing time with n-side readout detectors, a different picture emerged. The collected charge at any given bias voltage increased to reach a maximum after about 40 days (20°C equivalent). Up to a 30% increase in the collected charge has been verified (Figure 1.20). The signal stays > 20% higher than the starting value for about 300 days and returns to its initial value after about 1,100 days. The results of these measurements have changed the way the reverse annealing is regarded in high-energy physics experiments. In fact, the annealing can be used to achieve two important goals: (1) recovering some fraction of the signal height; and (2) reducing the reverse current.

1.5 CONCLUSIONS: THE ATLAS EXAMPLE CASE

As a conclusion, we can anticipate the S/N performance for the ATLAS upgrade at the SLHC. The pitch of the innermost strips is 75 μm for a length of 2.5 cm. The pixels are 50 μm wide and 250 μm long. As mentioned before, the target doses are 1 × 10^{15} n_{eq} cm^{-2} (innermost strips), 3 × 10^{15} n_{eq} cm^{-2} (outermost pixels), and 2 × 10^{16} n_{eq} cm^{-2} (innermost pixels). Assuming an operation temperature of –25°C, the noises in the three cases, calculated using Equations (1.12) and (1.13) for the microstrip sensors with the measured currents (before annealing) after the corresponding

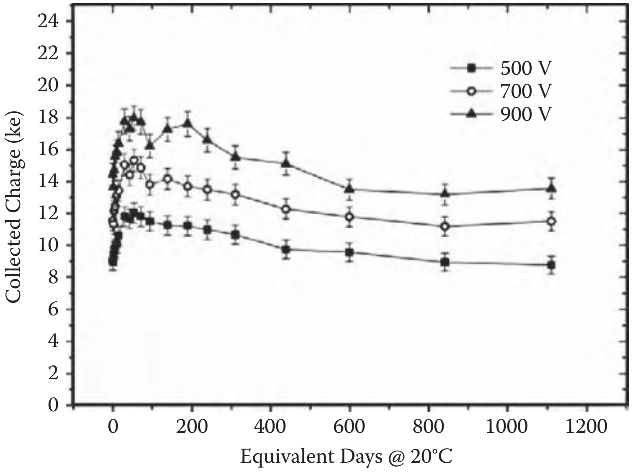


FIGURE 1.20 Changes of the signal collected by a silicon detector irradiated to 1x10^{15} n$_{eq}$ cm^{-2} with time after irradiation. The sensor was annealed at 60°C but the time axis has been rescaled to equivalent time at T$_{REF}$

radiation doses, and the present estimate for the pixel sensor noise (with a different parameterization from Equation (1.12)) are ~650 ENC, < 300 ENC, and < 400 ENC, respectively. Looking at the signal after the corresponding radiation doses, the S/N expected for the innermost microstrip layer is about 17 at 500 V and 25 at 900 V. For the two pixel layers, the S/N is about 23 at 500 V and 30 at 900 V for the outer pixels and 11.2 and 12.7 for the inner ones. As already stated, these values are for unannealed detectors and can be considered worst immediately after irradiation. A significant improvement is expected if a controlled annealing at 20°C (corresponding to between 100 to 300 days during the lifetime of the experiment). Nonetheless, the S/N values demonstrate that these detectors are in principle capable of operating fully efficiently after the highest doses of the most demanding application in high-energy physics.

REFERENCES

1. H.E. Boesch et al., *IEEE Trans. Nucl. Sci.* vol. NS-3, no. 6, 1981, pp. 1191–1197.
2. H.E. Boesch and T.L. Taylor, "Charge and interface state generation in field oxides," *IEEE Trans. Nucl. Sci.* vol. NS-31, no. 6, 1984, p. 1273.
3. K. Gill et al., *J. Appl. Phys.* vol. 82, no. 1, July 1997.
4. G.P. Summers et al., *IEEE Trans. Nucl. Sci.* vol. 34, 1987, p. 1134.
5. G. Lindstrom et al., "Radiation hard silicon detectors developments by the RD48 (ROSE) Collaboration," *Nucl. Instrum. Meth.* vol. A466, 2001, pp. 308–326.
6. F. Lemeilleur, M. Glaser, E.H.M. Heijne, P. Jarron, C. Soave, C. Leroy, et al., "Neutron, proton, and gamma irradiations of silicon detectors," *IEEE Trans. Nucl. Sci.* vol. 41, no. 3, 1994, pp. 425–431.
7. R. Wunstorf, "Systematische Untersuchangen zur Strahlenresistenz von Silizium-Detektoren für die Verwendung in Hochenenergiephysik-Experimenten," PhD thesis, Hamburg University, 1992 (see also DESY FH1K-92-01, 1992).

8. M. Moll, “Radiation damage in silicon particle detectors—microscopic defects and macroscopic properties,” PhD thesis, DESY-*THESIS*-1999-040, December 1999.
9. F. Lemeilleur, S.J. Bates, A. Chilingarov, C. Furetta, M. Glaser, E.H.M. Heijne, et al., “Study of characteristics of silicon detectors irradiated with 24 GeV/c protons between −20°C and +20°C,” *Nucl. Instrum. Meth.* vol. A360, no. 1–2, June 1, 1995, pp. 438–444.
10. CERN-RD48, “Research and development on silicon for future experiments,” http://rd48.web.cern.ch/RD48.
11. A. Ruzin et al., “Radiation effects in silicon detectors processed on carbon and oxygen-rich substrates,” *Mater. Sci. Semicond. Proc.* vol. 3, 2000, p. 257.
12. G. Kramberger, V. Cindro, I. Mandic, M. Mikuz, and M. Zavrtanik, “Effective trapping time of electrons and holes in different silicon materials irradiated with neutrons, protons and pions,” *Nucl. Instr. Meth. A* vol. 481, no. 1–3, 2002, pp. 297–305.
13. L.J. Beattie et al., “The electric field in irradiated silicon detectors,” *Nucl. Instr. Meth. A* vol. 418, no. 2, December 1998, pp. 314–321.
14. G. Casse, M. Glaser, E. Grigoriev, G. Casse, M. Glaser, and E. Grigoriev, “Study of evolution of active volume in irradiated silicon detectors,” *Nucl. Instr. Meth. A* vol. 426, 1999, pp. 140–146.
15. T. Dubbs et al., *Nucl. Instr. and Meth. A* vol. 383, 1996, p. 174.
16. S. Martí i García et al., *Nucl. Instr. and Meth. A* vol. 426, 1999, p. 24.
17. P. Holl, J. Kemmer, U. Prechtel, T. Ziemann, D. Hauff, G. Lutz, et al., “A double-sided silicon strip detector with capacitive readout and a new method of integrated bias coupling,” *IEEE Trans. Nucl. Sci.* vol. 36, February 1989.
18. G. Casse, P.P. Allport, and M. Hanlon, “Improving the radiation hardness properties of silicon detectors using oxygenated n-type and p-type silicon,” *IEEE Trans. Nucl. Sci.* vol. 47, no. 3, June 2000, pp. 527–532.
19. G. Casse, P.P. Allport, T.J.V. Bowcock, A. Greenall, M. Hanlon, and J.N. Jackson, “First results on the charge collection properties of segmented detectors made with p-type bulk silicon,” *Nucl. Instr. Meth. A* vol. 487, no. 3, July 2002, pp. 465–470.
20. C. Pittori and M. Tavani, “Gamma-ray imaging by silicon detectors in space: the AREM method,” *Nucl. Instr. Meth. A* vol. 488, 2002, pp. 295–306.
21. ATLAS Inner Detector TDR, CERN/LHCC/1997-17.
22. CMS Tracker at the SLHC, http://cmsdoc.cern.ch/cms/Tracker/html/Tracker2005/TKSLHC/index.html.
23. J. Kemmer, “Fabrication of a low-noise silicon radiation detector by the planar process,” *Nucl. Instrum. Meth. A* vol. 169, 1980, p. 499.
24. H.F.W. Sadrozinski, “Applications of silicon detectors,” *IEEE Trans. Nucl. Sci.* vol. 48, no. 4, 2001, pp. 933–940.
25. The LHC Conceptual Design Report—The Yellow Book CERN/AC/95-05 (LHC).
26. LHCb Technical Proposal, CERN/LHCC 98-4, February 20, 1998.
27. ALICE TDR 9, CERN/LHCC 2001-021, October 3, 2001.
28. LHCC 1/4/2008—Roland Garoby, http://indico.cern.ch/conferenceDisplay.py?confId=36149.
29. The ATLAS experiment, ID layout, http://atlas.web.cern.ch/Atlas/GROUPS/UPGRADES/layout.html.
30. G. Pellegrini et al., “Technology development of p-type microstrip detectors with radiation hard p-spray isolation,” *Nucl. Instr. and Meth. A* vol. 566, no. 2, October 15, 2006, pp. 360–365.
31. R.H. Richter et al., *Nucl. Instr. and Meth. A* vol. 377, 1996, pp. 412.
32. J. Kemmer et al., U.S. Patent N. US 6, 184, 562 B 1, 2001.
33. F. Anghinolfi et al., “SCTA-a rad-hard BiCMOS analogue readout ASIC for the ATLAS Semiconductor Tracker,” *IEEE Trans. Nucl. Sci.* vol. 44, no. 3, June 1997, pp. 298–302.

34. G. Casse, P.P. Allport, and A. Greenall, “Response to minimum ionising particles of p-type substrate silicon microstrip detectors irradiated with neutrons to LHC upgrade doses,” *Nucl. Instr. and Meth. A* vol. 581, 2007, pp. 318–321.
35. A. Affolder, P. Allport, and G. Casse, “Studies of charge collection efficiencies for p-type planar silicon detectors after reactor neutron and 26 MeV proton doses up to 2x10^{16} n_{eq}/cm^2,” 7th International Conference on Radiation Effects on Semiconductor Materials Detectors and Devices, October 15–17, 2008, Florence, Italy.
36. P.P. Allport, G. Casse, M. Lozano, P. Sutcliffe, J.J. Velthuis, and J. Vossebeld, “Performance of p-type micro-strip detectors after irradiation to 7.5 × 10^{15} p cm^2,” *IEEE Trans. Nucl. Sci.* vol. 52, no. 5 III, 2005, pp. 1903–1906.
37. A. Affolder, P.P. Allport, and G. Casse, “Charge collection efficiencies of planar silicon detectors after reactor neutron, pion and proton doses up to 2.5×10^{16} neq cm^{-2},” 1st International Conference on Technology and Instrumentation in Particle Physics, March 12–17, 2009, Tsukuba, Japan.
38. G. Casse et al., 14th *RD50* Workshop, June 3–5, 2009, Freiburg, Germany, http://indico.cern.ch/conferenceOtherViews.py?view=cdsagenda&confId=52883, and “Enhanced efficiency of segmented silicon detectors of different thicknesses after hadron irradiations up to 2×10^{16} n_{eq} cm^2,” 11th European Symposium on Semiconductor Detectors, June 7–11, 2009, Wildbad Kreuth, Germany.
39. G. Casse, A. Affolder, and P.P. Allport, “Charge collection efficiency measurements for segmented silicon detectors irradiated to 1x10^{16} n cm^{-2},” *IEEE Trans. Nucl. Sci.* vol. 55, no. 3, pt. 3, June 2008.
40. G. Casse, A. Affolder, and P.P. Allport, “Studies on charge collection efficiencies for planar silicon detectors after doses up to 10^{16} N_{eq}/cm^2 and the effect of varying substrate thickness,” 2008 Nuclear Science Symposium, October 19–25, 2008, Dresden, Germany, http://www.nss-mic.org/2008/Program/ListProgram.asp?session=N54.
41. V. Savolainen, et al., “Simulation of large-scale silicon melt flow in magnetic Czochralski growth,” *J. Cryst. Growth* vol. 243, no. 2, 2002, p. 243.
42. J. Härkönen, E. Tuovinen, P. Luukka, H.K. Nordlund, and E. Tuominen, “Magnetic Czochralski silicon as detector material,” *Nucl. Instr. and Meth. A* vol. 579, 2007, pp. 648–652.
43. G. Kramberger, “Charge collection measurements on MICRON RD50 detectors,” ATLAS Tracker Upgrade Workshop, December 11–14, 2007, Valencia, http://ific.uv.es/slhc/ATLASUpgrade/.
44. G. Casse, A. Affolder, P.P. Allport, and M. Wormald, “Study of the response to minimum ionising particles of microstrip detectors made with float zone and magnetic Czochralski silicon after neutron irradiation,” *Nucl. Instr. and Meth A* vol. 598, no. 3, January 21, 2009, pp. 671–674.
45. E. Fretwurst et al., “First results on 24 GeV/c proton irradiated thin silicon detectors,” 11th RD50 Workshop, CERN, November 2007.
46. G. Casse, A. Affolder, P. Allport, and M. Wormald, “CCE studies in silicon,” 17th International Workshop on Vertex Detectors, July 28–August 1, 2008, Utö Island, Sweden, PoS(VERTEX 2008)036.
47. B. Dezillie et al., “Radiation hardness of silicon detectors manufactured on wafers from various sources,” *Nucl. Instr. and Meth. A* vol. 388, no. 3, 2007, pp. 314–317.
48. E. Fretwurst, F. Hönniger, G. Kramberger, G. Lindström, I. Pintilie, and R. Röder, “Radiation tolerant epitaxial silicon detectors of different thickness,” 6th RD50—Workshop on Radiation Hard Semiconductor Devices for Very High Luminosity Colliders, June 2–4, 2005, Helsinki.
49. G. Casse, P.P. Allport, and A. Watson, “Effects of accelerated annealing on p-type silicon micro-strip detectors after very high doses of proton irradiation,” *Nucl. Instr. and Meth. A* vol. 568, 2006, pp. 46–50.

50. A. Affolder, P. Allport, and G. Casse, "Charge collection efficiency measurements of irradiated segmented n-in-p and p-in-n silicon detectors for use at the super-LHC," *IEEE TNS* vol. 56, no. 3, June 2009, pp. 765–770.

2 Radiation-Tolerant CMOS Single-Photon Imagers for Multiradiation Detection

Edoardo Charbon, Lucio Carrara, Cristiano Niclass, Noémy Scheidegger, and Herbert Shea

CONTENTS

2.1 INTRODUCTION

Sensors capable of detecting a few photons or even a single photon have been available for several decades now, albeit in a single point of detection. In these sensors photoelectrons are generally multiplied by micropressure cavities over relatively large distances. Imaging small areas usually requires highly optimized micromechanical scanning. Imaging large areas can, on the contrary, be done in parallel using arrays of photomultipliers but at steep costs.

Imaging photon-starved scenes has become more affordable with the introduction of highly miniaturized solid-state avalanche photodiodes (APDs) that can be arranged in relatively large arrays and tessellated in even larger areas. APDs may operate both in linear and in Geiger mode, thus demonstrating their usefulness in photon counting, photon energy evaluation, and single-photon time-of-arrival detection and time-resolved imaging.

More recently, the demonstration of these devices implemented in CMOS has opened the way to the creation of large arrays of single-photon detectors on one hand and of advanced functionality on the other. Researchers have shown several uses for this technology and applications today range from multiphoton microscopy [1] to voltage sensitive dye (VSD)-based imaging [2,3], particle image velocimetry (PIV) [4], and instantaneous gas imaging [5]. Fluorescence-based imaging (both single- and multiphoton, lifetime- and correlation-based) is the research direction that has perhaps most influenced the development of fast and sensitive optical detectors [6–8].

Other uses for single- and few-photon detectors include a variety of time-resolved imaging used in astronomical surveys, positron emission tomography, and single-molecule imaging, whereby single-event coincidences are routinely used to solve inverse problems or to reduce the impact of random noise originating from background radiation or inherent sensor imperfections. Accurate photon time of arrival over a large array of pixels is gaining traction in disciplines that have used this technique for decades but so far have lacked the tools for large-scale surveys. Applications of this type include Hanbury Brown Twiss (HBT) interferometry for the analysis of light micro and macro light sources as well as for stellar bodies [9–11], X-ray imagers, and three-dimensional (3-D) vision [12–15], just to name a few.

The next frontier for CMOS APDs is space, where one of the most important requirements is radiation hardness. Other hostile environments, such as the human body and heavy B-field cavities, are also of interest to researchers, where sensors could be used for monitoring purposes, for example, in radiation therapy and biomedical imaging science.

In the last four decades, solid-state APDs have gradually evolved from relatively crude devices to the sophistication of today. Almost every imaging technology has one such device, and the range of implementations is quite wide. There are two main lines of research in silicon APDs: one advocates the use of highly optimized processes to boost performance; and the other proposes to adapt design to existing processes to reduce cost and to maximize miniaturization.

In this chapter, we focus on the latter approach and discuss the latest advances in CMOS arrays designed for radiation tolerance. We also discuss how advanced processes can ensure in-pixel and on-chip processing of ultra-high-speed signals that are typical of single-photon detectors. In this context we study one such architecture and show how the high-speed characteristics enabled us to trade off speed with a number of performance constraints while keeping pixel pitch low. This chapter presents results achieved recently in the field of tolerance to high-dose gamma radiation, proton bombardment, and X-rays.

2.2 SOLID-STATE SINGLE-PHOTON-DETECTING PIXELS

Devices for single-photon detection are realizable in many solid-state and nonsolid-state implementations [16]. To contextualize our discussion on single-photon detection, we mention here two classes of detectors that have been thus far the solution of choice in many applications: multichannel or microchannel plates (MCPs) and photomultiplier tubes (PMTs) [17]. A number of solid-state solutions have been proposed as a replacement of MCPs and PMTs using conventional imaging processes. The challenge, though, has been to meet both single-photon sensitivity and low timing uncertainty. To address the sensitivity problem, cooled, intensified, and electron multiplier charge-coupled devices (EMCCD) [18] as well as ultra-low-noise CMOS APS architectures [19] have been proposed. Multiplication of photogenerated charges by impact ionization has also been used in conventional CCDs both off pixel [20] and on pixel [21]. Matching the picosecond timing uncertainty of PMTs, however, to the best of our knowledge has not been possible in CCD/CMOS imagers, even though uncertainties as low as 1 microsecond in CCD [22] and a few nanoseconds in CMOS APS [23] have been demonstrated. While CCD streak cameras can achieve a resolution of a few picoseconds, they require a two-dimensional (2-D) pixel array to resolve a string of photon arrivals. Moreover, long acquisition latency and the added complexity to form and deflect the photoelectron beam make this device unsuitable for miniaturization and low-cost operation.

Solid-state sensors based on APDs were proposed decades ago to simultaneously achieve high sensitivity and dynamic range and low timing uncertainty [24]. In APDs, carriers generated by the absorption of a photon in the p-n junction are multiplied by impact ionization, thereby producing an avalanche. APDs can reach timing uncertainties as low as a few tens of picoseconds thanks to the speed at which an avalanche evolves from the initial carrier pair forming in the multiplication region. An APD is implemented as a photodiode is reverse biased near breakdown, where it exhibits optical gains greater than 1. An APD is *proportional* or *linear* when it is biased below breakdown. It can be used to detect clusters of photons and to determine their energy. When biased above breakdown, the optical gain becomes virtually infinite. Therefore, with relatively simple ancillary electronics, the APD becomes capable of detecting single photons. The APD operating in this regime, known as Geiger mode of operation, is called *single-photon avalanche diode* (SPAD).

APDs—and especially SPADs—have recently evolved toward more and more compact devices following Moore's Law. As a result, researchers have shown functional devices in 0.8 μm [12,25], 0.35 μm [26–30], 0.18 μm [31,32], and 130 nm [33–35] CMOS processes. With the availability of SPADs in deep-submicron CMOS processes, it has become possible to implement more and more complex functionality in proximity to the detector [36–38]. At the same time, smaller pixels are now feasible but with low or no functionality implemented per pixel.

Unlike in conventional pixel arrays, where the intensity of light is coded in terms of a voltage that is stored in a parasitic capacitance, in SPAD arrays the state of photon counts is not statically available on pixel. Photon counts can be stored on pixel only when a counter and a memory are integrated *in situ*, with the consequence of increasing the pitch. In addition, while a large memory is beneficial in increasing the counting resolution—and thus the dynamic range for a given readout speed—it further absorbs area and power, increases the pitch, and decreases the size of a feasible array. An alternative to this approach is the use of a memory of minimum size: 1 bit. The effect is to minimize pitch while requiring more frequent readouts. However, with the reduction of feature size, high speeds can be easily achieved at a cost of higher power consumption, which in turn can be traded off for resolution.

2.3 APDS AND SPADS FABRICATED IN CMOS PROCESSES

2.3.1 Basic Structure Design

There exist at least two main implementation styles for APDs and SPADs. In the first style, known as reach-through APD (RAPD), one builds a p+-π-p-n structure [39].

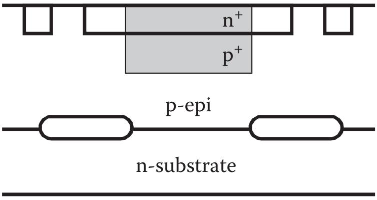


FIGURE 2.1 Cross-section of APDs that can be fabricated in a planar process. (Reprinted with permission from Edoardo Charbon. "Towards large scale CMOS single-photon detector arrays for lab-on-chip applications." *J. Phys. D*: *Appl. Phys.*, 41, 094010 (9pp) 2008.)

When reverse-biased, the depletion region extends from the cathode to the anode. Accordingly, the multiplication region is deep in the p/n+ junction. Due to the depth of the multiplication region, this device is indicated for absorption of red and near-infrared (NIR) photons up 1.1 μm (for silicon). Since the photoelectrons drift until the multiplication region, a larger timing uncertainty is generally observed.

The second implementation style is compatible with planar CMOS processes, and it involves shallow or medium-depth p- or n-layers to form high-voltage pn junctions. Cova and others have investigated devices designed in this style since the 1970s, yielding a number of structures [40]. All these structures have in common a p-n junction and a zone designed to prevent premature edge breakdown (PEB). An example of the early structures is shown in Figure 2.1. In [41], n+/p+ enrichment in p-substrate was used, while PEB was prevented by confining p+ enrichment in the center of the APD.

More recently, many authors have developed APDs both in linear and Geiger mode using *dedicated* planar and nonplanar processes, achieving superior performance in terms of sensitivity and noise. A good example is the work of Kindt [42]. The main disadvantage of using dedicated processes is generally the lack of libraries that can support complex functionalities and deep-submicron feature sizes, thus limiting array sizes.

An interesting alternative is the use of a hybrid approach whereby the APD array and ancillary electronics are implemented in two different processes, each optimized for APD performance and speed, respectively [43]. If the ancillary electronics is implemented in CMOS, high degrees of miniaturization are possible. The price to pay is increased fabrication complexity and higher cost.

In 2003 the integration of linear and Geiger mode APDs in a low-cost CMOS process became feasible [44]. In planar processes, one of the main challenges is PEB prevention. This is done by design forcing the electric field everywhere to be lower than that on the planar multiplication region, where it should be uniform. Figure 2.2 shows some of the most popular structures. In (a) the n+ layer maximizes the electric field in the middle of the diode. In (b) the lightly doped p-implant reduces the electric field at the edge of the p+ implant. In (c) a floating p-implant locally increases the breakdown voltage. With a polysilicon gate one can further extend the depletion region (gray line in the figure). Finally, in a process with trenches it is possible to decrease the electric field using the geometry of solution (d).

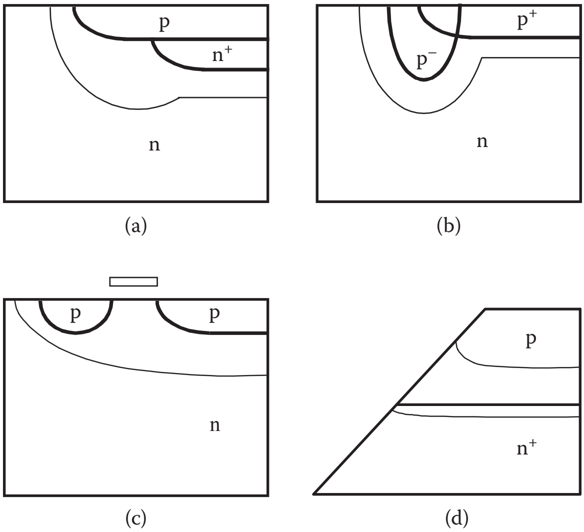


FIGURE 2.2 Techniques for prevention of Premature Edge Breakdown (PEB) in planar processes. (Reprinted with permission from Edoardo Chabon. "Towards large scale CMOS single-photon detector arrays for lab-on-chip applications." *J. Phys. D.: Appl. Phys.* 41 094010 (9pp) 2008.)

In our work we have selected solution (b) as the standard implementation in most of our designs. This implementation assumes a p-substrate and an n-well isolation. There are several advantages to using an n-well. First, photocharges generated in a given pixel cannot cause avalanches in neighboring pixels, therefore minimizing electrical cross talk. Second, only photocharges relatively near the multiplication region can trigger an avalanche, thus reducing timing uncertainty. The main disadvantage is a set of tighter separation rules, thus lessening pixel packing potential of a given technology, fill factor, and, ultimately, pixel pitch. Modern imaging processes provide several lightly doped implants at three or more depths. So, an optimal layer combination (p+/p-/n-well) generally exists that can yield a good trade-off between timing uncertainty and noise. However, care should be used to avoid full depletion of the well and punch-throughs between shallow wells and substrate. Buried layers should also be used with care to prevent punch-through across the n-well.

More recently, to reduce pixel pitch while achieving a reasonably effective PEB prevention capability, some authors have suggested the use of p-STI structures instead of full-blown lightly doped implants [34,35]. These structures are robust and could greatly enhance miniaturization.

2.3.2 Quenching and Recharge

Linear APDs are multiphoton detectors when used as charge accumulators. In this case, the charges generated at each avalanche are integrated, and subsequent amplification may not be needed. In single-photon detection mode, fast amplifiers should

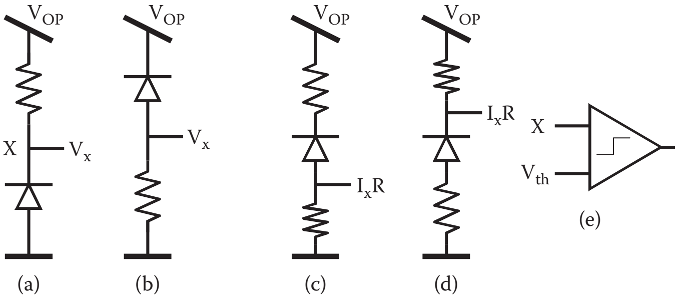


FIGURE 2.3 Passive quenching variants. Voltage detection mode (a,b); Current detection mode (c,d). Node X may be connected to a comparator (e) or a simple inverter. (Reprinted with permission from Edoardo Chabon. "Towards large scale CMMOS single-photon detector arrays fo lab-on-chop applications." *J. Phys. D: Appl. Phy.* 41 094010 (9pp) 2008.)

be used, adding to jitter and dark noise. SPADs, on the contrary, can operate only in single-photon mode. As mentioned earlier, this mode of operation is achieved biasing the diode *above* breakdown by a voltage known as *excess bias voltage*. Upon photon absorption, an avalanche may be triggered involving a sufficient number of charges to be easily detected, requiring no further amplification. However, the avalanche needs be quenched.

There exist two main quenching mechanisms, known as passive and active methods. In passive quenching the avalanche current itself is used to drop the voltage across the diode. This is generally accomplished via a ballast resistor placed on the anode or the cathode of the diode, as shown in Figure 2.3. The detection of the avalanche can be accomplished by measuring the voltage across the ballast resistance (Figures 2.3a, 2.3b) or the current across a low- or zero-resistivity path (Figures 2.3c, 2.3d). Pulse shaping may be performed using a comparator (Figure 2.3e).

Excess bias voltage, V_E, satisfies the equality $V_E = |V_{OP}| - |V_{bd}|$, where V_{bd} is the true breakdown voltage, and V_{OP} is the overall bias voltage shown in the figure. The ballast and sense resistances can be implemented in polysilicon [44] or using the nonlinear characteristics of a PMOS or NMOS biased in weak [26] or strong inversion [45,46]. In active quenching mode, the avalanche current is used to actively stop the avalanche. The literature on active quenching is extensive. In [44,47], for example, some of the existing schemes can be found. Other authors in the imaging community have recently revisited the issue [48].

After quenching, the device enters another phase known as *recharge*. During this phase the photodiode bias voltage must return to the pre-avalanche state as quickly as possible. Again, there are passive and active schemes to achieve recharge. The simplest approach is shown in Figure 2.3. The diode will automatically recharge to V_{OP} via the ballast resistance. The recharge, in this case, follows the *RC* exponential, where *R* is the equivalent quenching resistance, and *C* is the total parasitic capacitance at node X. In active recharge schemes, the photodiode is forced to the initial state generally via a fast switch controlled by a current sense amplifier. Even though these schemes are attractive, they usually require extra complexity to a pixel, thus potentially hindering miniaturization unless in-pixel feedback is used. In addition, active recharge may result in increased afterpulsing probability if the recharge time is lower than the device intrinsic relaxation time. Recently, an active recharge scheme was introduced to achieve a different goal (i.e., to fix the recharge time to a predefined value, possibly above intrinsic relaxation) to keep afterpulsing within predefined boundaries [45,46].

The quenching and recharge times are collectively known as *dead time*. Dead time in passive quenching/recharge methods is potentially longer than that of their active counterparts. However, the advantage of a reduced dead time in large array may be reduced by limited speeds of pixel readout schemes.

2.3.3 The Importance of Miniaturization

The first SPAD implementations in 0.35 μm CMOS technology have demonstrated fully scalable pixels at a pitch of 25 μm. However, for a realistic Mpixel sensor realization, this limit should be further reduced. Pixel miniaturization has other benefits too. The reduction of anode and cathode areas in SPADs generally reduces the dark count rate (DCR) (i.e., the average frequency of spurious pulses in the dark) [44]. It also reduces the parasitic capacitance at node X in Figure 2.3 and therefore possibly reduces dead time. In addition, the number of carriers involved in an avalanche is also reduced, hence decreasing the probability of carrier trapping and, consequently, of afterpulsing. Finally, fewer carriers involved in impact ionization will cause smaller photon emission during the avalanche and also less optical cross talk.

2.4 BUILDING AND TESTING RADIATION-HARDENED SPADS

The justification for the use of photon counting in imaging is two-fold. First, photon counting enables a quantitative approach to imaging. This may be important in applications where a calibration phase is not desirable or possible, such as in space and in other hostile environments like human implants. Second, assuming nearly 100% fill factor, using, for example, optical means [49], photon counters enable very high sensitivities, which are useful whenever the analysis of low-light illumination scenes is needed. In addition, when SPADs are used, the imager can be made resilient to a variety of radiation types, from cosmic rays to high-energy proton beams to X-rays to gamma radiation and even strong magnetic fields.

The effects of radiation are quite different, depending on its nature, the energy of its quanta, and the materials it traverses; the literature on the subject is extensive (see, e.g., [50,51] for a review). Techniques to maximize sensor tolerance have also been developed for a number of years, and several imagers resistant to up to 30Mrad (Si) of gamma radiation have been reported. These sensors have several shortcomings: either significant noise performance degradation, up to several orders of magnitude [51], or unacceptably high preradiation noise levels [52]. In addition, many radiation-tolerant sensors reported in the literature use dedicated processes, thus possibly limiting their suitability for mass-market applications [53].

In what follows we describe a CMOS photon-counting imager that was designed to detect Earth's airglow, the atmospheric oxygen emission at 762 nm due to oxygen recombination. Airglow occurs day and night and enables geostationary and orbiting satellites to infer their position referred to the earth's center for attitude determination purposes [54,55]. The goal of this project was to develop a sensor that could dramatically reduce the requirements on weight and size of the navigational telescope optics to be mounted on ultra-low-cost satellites.

Figure 2.4 shows the imager concept and examples of airglow emission under distinct conditions. The sensor, operating in photon-counting mode, consists of an array of 32 × 32 pixels. Each pixel comprises a single-photon detector, a 1-bit counter, and fast-readout circuitry. Radiation hardness was achieved with a combination of schematic and physical design techniques while keeping pixel complexity to a minimum and spreading all its components uniformly. This was done to minimize the probability of radiation particles damaging a vital section of the chip. Moreover, contrary to the general trend in the literature, the fill factor was kept low. This has the advantage of reducing the probability that the detector is hit by radiation; in addition, light is reclaimed using optical concentrators. Finally, due to the digital nature of the detector, a higher resilience to charge injection due to irradiation was built into the sensor.

The block diagram of the sensor is shown in Figure 2.5. The timing diagram of the sensor readout is shown in the figure inset. The array is read out in rolling shutter mode via the high-speed row decoder and may be reset after each read operation or read out nondestructively. The column decoder is used to issue a readout control signal, while the signal conditioning synchronizes it with the clock. All 32 columns are read in parallel, thus enabling a complete 1,024-pixel frame readout in $T_{\min}$ = 1.2 μs with 1-bit depth. To achieve a higher number of gray levels we accumulate N frames, thus reaching an intensity resolution of $\log_2(N)$ bits at the expense of lower frame rates. The saturation count rate is $1/T_{\min}$; $SNR_{\max}$ for integration time t_{int} is computed as

$$SNR_{\max} = 20\log\left(\frac{t_{\text{int}}}{T_{\min}}\right) - 10\log\left(\frac{t_{\text{int}}}{T_{\min}} + \text{Var}[DC]\right) \tag{2.1}$$

where the noise power is given by the sum of Poisson noise power and Var[DC], that is, the variance of the stochastic process underlying dark count generation. The latter is approximated by the average of dark counts during integration, or $DCR\ t_{\text{int}}$, where DCR is the dark count rate of the detector. We use median DCR. We believe that this figure is a better representation of the noise performance of the chip as it represents the DCR upper bound for 50% of the pixels.

The schematic of the pixel is shown in Figure 2.6. The detector is implemented as a SPAD, the counter as a latch, and the readout as a pull-down transistor. The SPAD is a p-n junction biased above breakdown to operate in Geiger mode. In this design, the avalanche voltage is sensed by M2 that forces the latch to logic level "1." Transistor M7 acts as pull-down of the column line that is kept high by resistor R_{PU}, while M6 is the row selection switch, controlled by RowSEL. When the column is pulled down, a buffer (not shown in the schematic) controls a pad, and the output of the chip for that column is interpreted as a photon detected in the previous interval of time. Transistors M4 and M5 are controlled, respectively, by column line (ColSET)

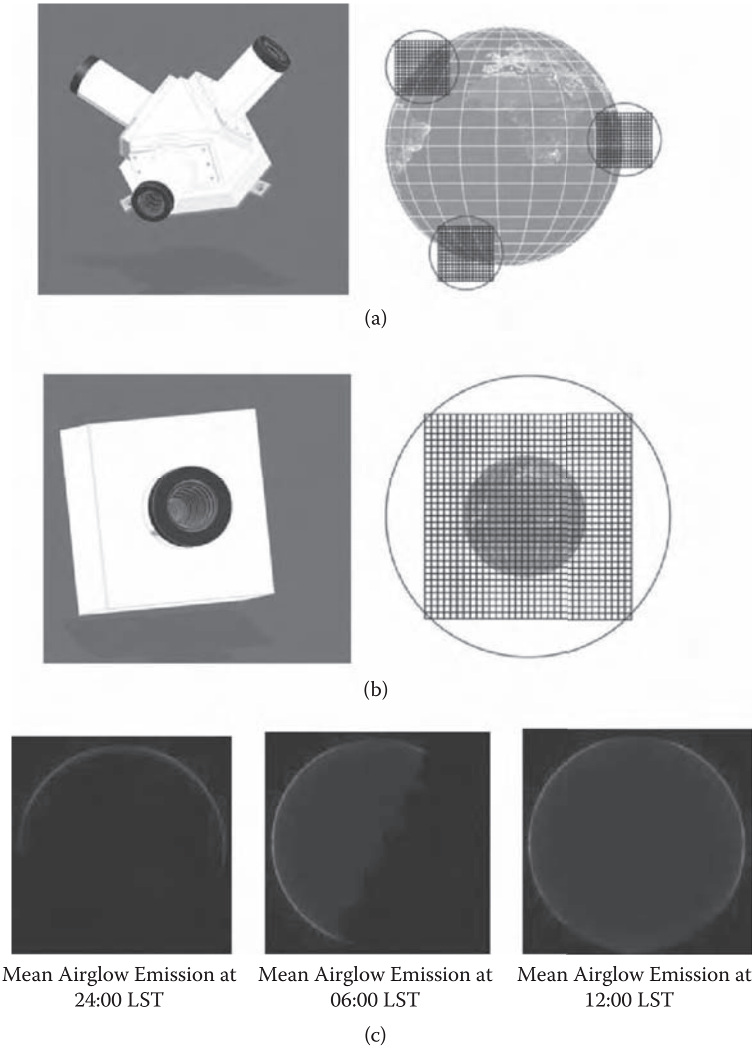

(a)
(b)
Mean Airglow Emission at 24:00 LST
Mean Airglow Emission at 06:00 LST
Mean Airglow Emission at 12:00 LST
(c)

FIGURE 2.4 Airglow imaging from orbit for attitude determination purposes: (a) in orbiting satellites (altitude: 2,000 km), using a triple sensor; (b) in geostationary satellites (altitude: 36,000 km) using a single sensor. The grids indicate the observation window of each sensor and the lateral resolution with a FOV of 20 degrees. (c) The center wavelength of airglow emission is 762 nm with a minimum estimated photon flux of 6400 counts/s/pixel. The figures show examples of airglow emission measured by an earlier mission. (Reprinted with permission from, Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications." Solid-State Circuits Conference – *Digest of Technical Papers*, ISSCC 2009. *IEEE International* February 8–12, 2009, pp. 40–41, 41a.

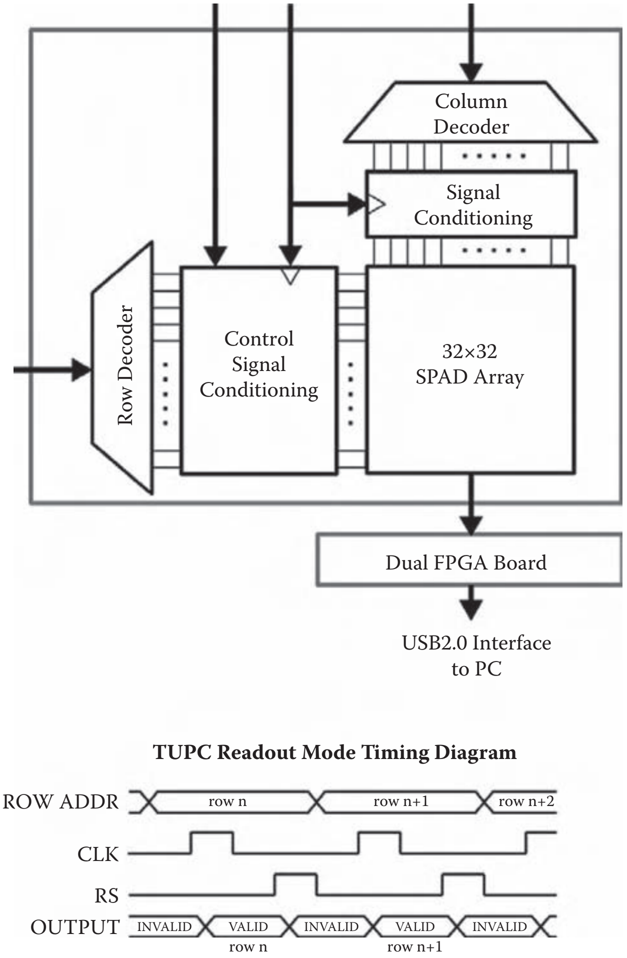


FIGURE 2.5 Block diagram of the sensor system. The inset shows the basic timing diagram for time-uncorrelated photon counting (TUPC) mode. Rows are read out in rolling shutter mode with or without row-wise reset (RS). The clock (CLK) determines the minimum integration time. Nonsequential rows or array subsets may also be read for frame rate increase and power dissipation reduction. (Reprinted with permission fron Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications." Solid-State Circuits Conference – *Digest of Technical Papers*, ISSCC 2009, *IEEE International* February 8–12, 2009, pp. 40–41, 41a.)

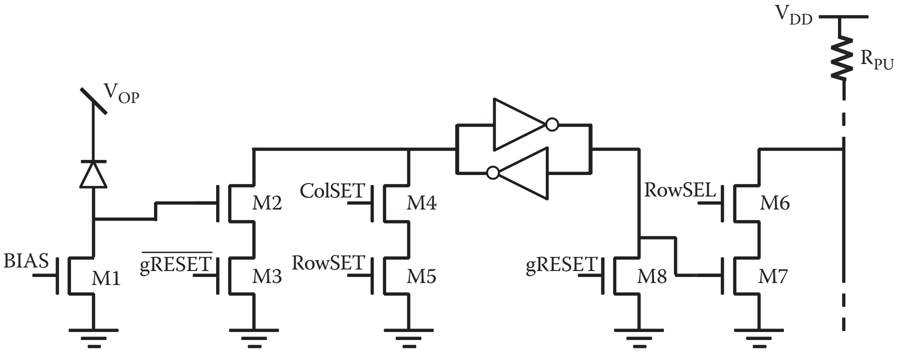


FIGURE 2.6 Pixel schematics. The pixel comprises 10 NMOS and 2PMOS transistors. The most critical devices were implemented in source-surrounded-by-gate style. The pull-up resistor was external to the pixel. (Reprinted with permission from Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications." Solid-State Circuits Conference – *Digest of Technical Papers*, ISSCC 2009. *IEEE International*, February 8–12, 2009, pp. 20–41, 41a.)

and row line (RowSET) to force the static memory of a specific pixel to logic level "1," regardless of the SPAD state, for testing purposes. M8 is used to operate a global- or row-based reset via signal gRESET, whereas M3 prevents memory conflicts in case of a SPAD firing during reset. SPAD quenching and recharge are performed by M1 that can be adjusted globally via signal BIAS, so as to select a proper trade-off between dead time and afterpulsing probability [27]. The pixel comprises a total of 12 MOS transistors, 10 NMOS, and only 2 PMOS transistors, thus enabling minimization of NWELL surface and ensuring a pitch of 30 µm.

To sustain massive doses of radiation, measures were taken at the layout level as well. Most of the NMOS transistors were implemented in source-surrounded-by-gate style to minimize defect-induced leakage. Other radiation-hardening techniques included the increase of certain design rules, extensive use of contacts to minimize potential latch-up, and the implementation of n+ and p+ trenches at well boundaries. The layout of the pixel is shown in Figure 2.7. The SPAD was implemented as a p+/p-well/deep n-well junction; its cross section is shown in Figure 2.8. The breakdown voltage, V_{bd}, of the SPAD in this design is 17.7 V. At its cathode, a bias voltage of 21 V was applied to operate with an excess bias voltage, V_E, of 3.3 V. Thanks to this configuration, a lower capacitance at the sensing node was achieved, thus reducing the charges involved in an avalanche and also reducing optical cross talk and afterpulsing at given dead time.

Figure 2.9 shows a micrograph of the chip, whose total surface is 2.00 × 2.35 mm^2. The sensor was first tested for speed and functionality. For this test, we used a breadboard system based on a dual Xilinx Virtex II Pro FPGA board similar to [56]. In the current firmware implementation, the minimum integration time is 2.6 µs, limited by a clock frequency of 48 MHz. The chip was also tested for sensitivity, signal uniformity, and noise performance. The results of the full characterization of the chip are reported in the table of Figure 2.12. The radiation testing was performed

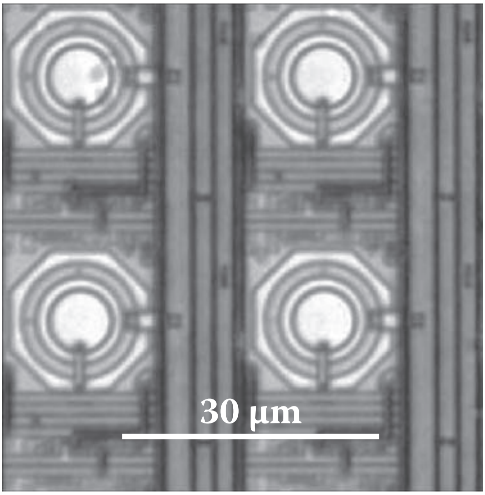


FIGURE 2.7 Photomicrograph of the pixel conceived for radiation hardness. Most NMOS transistors were implemented in source-surrounded-by-gate style. The active area of the detector was minimized with heavy use of contacts to minimize the chance of latch up.

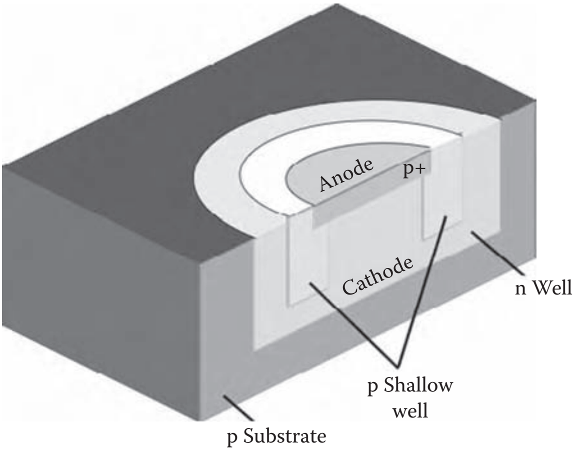


FIGURE 2.8 Cross-section of the SPAD used in this chip. The device comprises a circular sensitive area surrounded by PEB prevention guard rings. (Reprinted with permission from Edoardo Chabon. "Towards large scale CMOS single-photon detector arrays for lab-on-chip applications." *J. Phys. D: Appl. Phys.* 41, 094010, (9pp) 2008.)

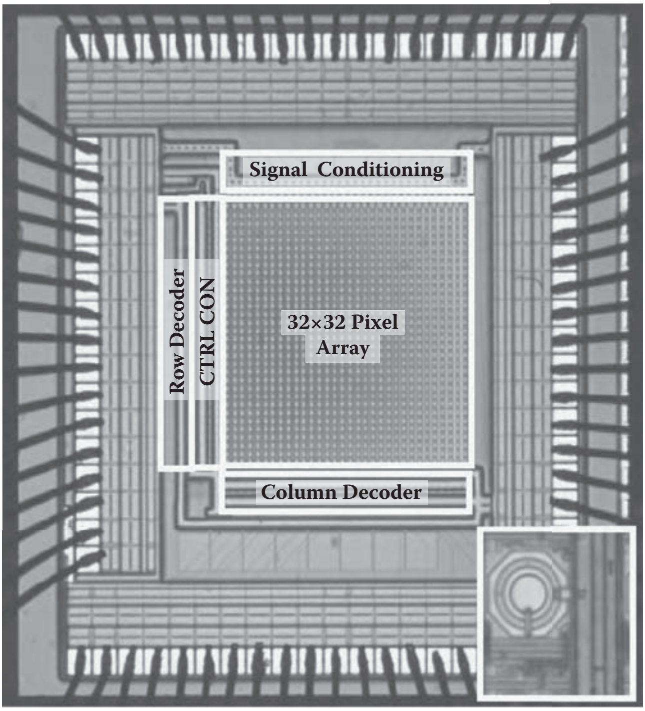
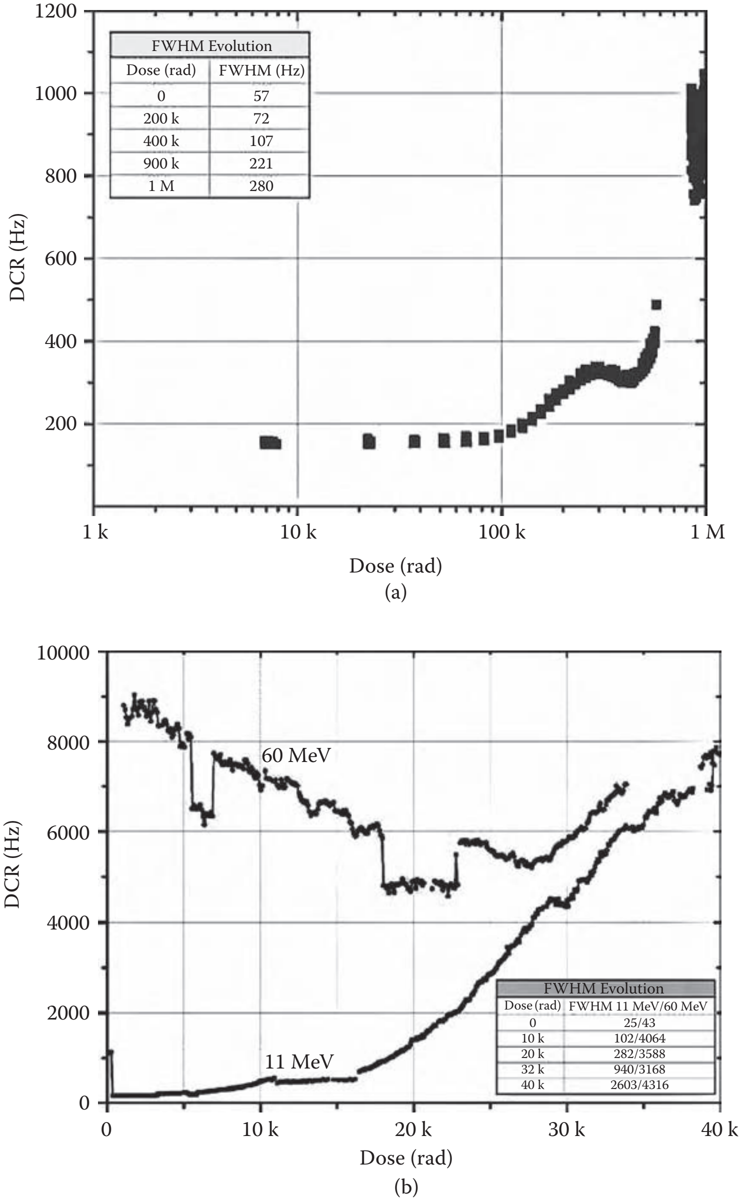


FIGURE 2.9 Photomicrograph of the sensor chip. The circuit, fabricated in 0.35 μm CMOS technology, has a surface of 2.0 × 2.35 sqmm. The pixel, in the inset, has a pitch of 30 μm. (Reprinted with permission from Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications. Solid-State Circuits conference – *Digest of Technical Papers*, ISSCC, 2009. *IEEE International*, February 8–12, 2009, pp. 40–41, 41a.)

in three separate measurement campaigns. Gamma radiation was performed at ESA-ESTEC in Noordwijk (The Netherlands) using a standard Co60 source. The sensor received a total dose of 1 Mrad (Si), followed by 172 h of annealing at 80°C. The results are summarized in Figure 2.13. The median DCR measured during the experiment is reported in Figure 2.10a. We also exposed a sensor with identical detectors but different pixel and system electronics, not optimized for radiation hardness [56]. The sensor sustained a catastrophic failure at 2 kRad, and no recovery was possible after annealing.

In the second experiment, the sensor was exposed to two separate proton beams at a constant energy of 11 MeV and 60 MeV, respectively. The experiment was performed at the Paul Scherrer Institute in Villigen (Switzerland). Figure 2.10b shows the median DCR versus dose for a maximum of 40 krad. The evolution of the DCR distribution over the array for the gamma irradiations is shown in Figure 2.11.

In the third experiment, the chip was exposed to a massive X-ray dose at the University Institute for Radiation Physics in Lausanne (Switzerland). The X beam, generated by a bipolar metal-ceramic tube Comet-Yxlon TU 320-D03, achieved



FWHM Evolution

Dose (rad)	FWHM (Hz)
0	57
200 k	72
400 k	107
900 k	221
1 M	280

(a)

FWHM Evolution

Dose (rad)	FWHM 11 MeV/60 MeV
0	25/43
10 k	102/4064
20 k	282/3588
32 k	940/3168
40 k	2603/4316

(b)

FIGURE 2.10 Median DCR evolution during three irradiation experiments before annealing: (a) gamma irradiation; (b) proton irradiation (11 MeV; 60 MeV). The graphs include median and FWHM values of DCR (in inset). All measurements were conducted at room temperature. (Reprinted with permission from Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications. Solid-State Circuits conference – *Digest of Technical Papers*, ISSCC, 2009. *IEEE International*, February 8–12, 2009, pp. 40–41, 41a.)

fluence and total dose levels reported in Figure 2.13. The table also lists the results of the DCR change. Preliminary irradiations were performed without any filtering and using a large collimation (27 mm). A series of irradiations at 15 kV, 120 kV, and 200 kV showed negligible impact on DCR, PDP, and afterpulsing.

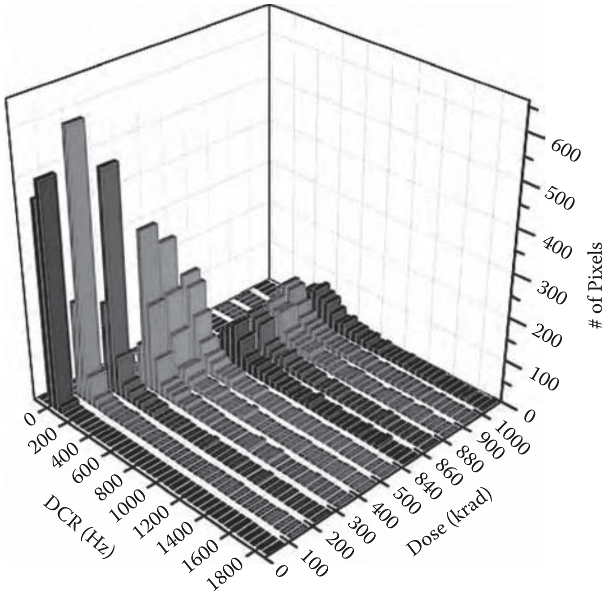
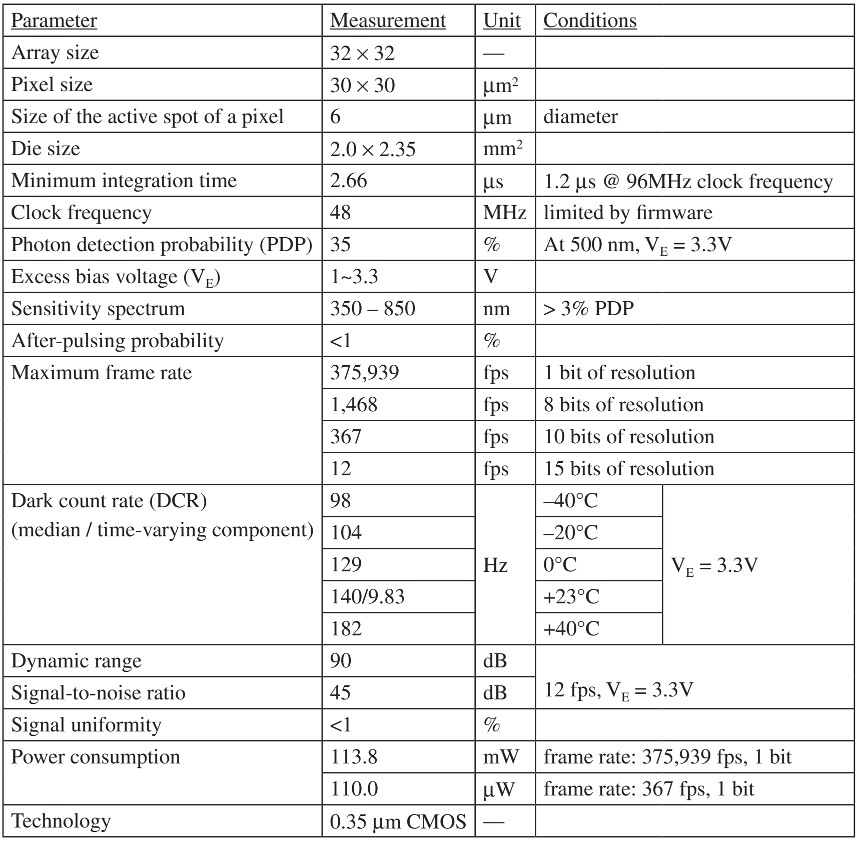


FIGURE 2.11 Evolution of DCR distribution as a function of dose during gamma irradiation. All measurements were conducted at room temperature. (Reprinted with permission from Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications. Solid-State Circuits conference – *Digest of Technical Papers*, ISSCC, 2009. *IEEE International*, February 8–12, 2009, pp. 40–41, 41a.)



Parameter	Measurement	Unit	Conditions	
Array size	32 × 32	—		
Pixel size	30 × 30	μm^2		
Size of the active spot of a pixel	6	μm	diameter	
Die size	2.0 × 2.35	mm^2		
Minimum integration time	2.66	μs	1.2 μs @ 96MHz clock frequency	
Clock frequency	48	MHz	limited by firmware	
Photon detection probability (PDP)	35	%	At 500 nm, V_E = 3.3V	
Excess bias voltage (V_E)	1~3.3	V		
Sensitivity spectrum	350 – 850	nm	> 3% PDP	
After-pulsing probability	<1	%		
Maximum frame rate	375,939	fps	1 bit of resolution	
	1,468	fps	8 bits of resolution	
	367	fps	10 bits of resolution	
	12	fps	15 bits of resolution	
Dark count rate (DCR) (median / time-varying component)	98	Hz	–40°C	V_E = 3.3V
	104		–20°C	
	129		0°C	
	140/9.83		+23°C	
	182		+40°C	
Dynamic range	90	dB	12 fps, V_E = 3.3V	
Signal-to-noise ratio	45	dB		
Signal uniformity	<1	%		
Power consumption	113.8	mW	frame rate: 375,939 fps, 1 bit	
	110.0	μW	frame rate: 367 fps, 1 bit	
Technology	0.35 μm CMOS	—		

FIGURE 2.12 Image sensor performance summary. All measurements were performed at room temperature when not otherwise indicated. (Reprinted with permission from Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications. Solid-State Circuits conference – *Digest of Technical Papers*, ISSCC, 2009. *IEEE International*, February 8–12, 2009, pp. 40–41, 41a.)

Irradiation type	Source	Fluence/ Flux	Dose (Si)	Initial DCR	Final DCR	DCR after Annealing (anneal time)
Gamma	Co60	10.46 rad/min	1.0 Mrad	153	569	276 (172 h)
X	Comet-Yxlon TU320-D03	4.3 AsV^2	0.25 mGy	540	545	540 (1 min)
		324 AsV^2	0.25 mGy	540	640	540 (1 min)
		900 AsV^2	0.5 mGy	540	701	540 (1 min)
Proton	Accelerator	1.8 × 10^7p/cm^2/s (11MeV)	40.0 krad	140	6298	3884 (10 d)
		8.3 × 10^7p/cm^2/s (60 MeV)	40.0 krad	142	6290	1299 (21 d)

FIGURE 2.13 Irradiation experiment summary. DCR is reported in Hz at room temperature. PDP, afterpulsing probability, and maximum frame rate remained unchanged after all three types of irradiation. (Reprinted with permission from Carrara, L., Niclass, C., Scheidegger, N., Shea, H., Charbon, E. "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications. Solid-State Circuits conference – *Digest of Technical Papers*, ISSCC, 2009. *IEEE International*, February 8–12, 2009, pp. 40–41, 41a.)

REFERENCES

1. W. Denk, J.H. Stricker, and W.W. Webb, "2-Photon Laser Scanning Fluorescence Microscopy," *Science*, Vol. 248, pp. 73–76, 1990.
2. A. Grinvald et al., "*In-Vivo* Optical Imaging of Cortical Architecture and Dynamics," *Modern Techniques in Neuroscience Research*, U. Windhorst and H. Johansson (Eds.), Springer, 2001.
3. J. Fisher et al., "In Vivo Fluorescence Microscopy of Neuronal Activity in Three Dimensions by Use of Voltage-Sensitive Dyes," *Optics Letters*, Vol. 29, No. 1, pp. 71–73, January 2004.
4. S. Eisenberg et al., "Visualization and PIV Measurements of High-Speed Flows and Other Phenomena with Novel Ultra-High-Speed CCD Camera," *Proceedings of SPIE*, Vol. 4948, pp. 671–676, 2002.
5. W. Reckers et al., "Investigation of Flame Propagation and Cyclic Combustion Variations in a DISI Engine using Synchronous High-Speed Visualization and Cylinder Pressure Analysis," *International Symposium für Verbrennungdiagnostik*, pp. 27–32, 2002.
6. W. Becker, A. Bergmann, E. Haustein, Z. Petrasek, P. Schwille, C. Biskup, et al., "Fluorescence Lifetime Images and Correlation Spectra Obtained by Multidimensional Time-Correlated Single Photon Counting," *Microscopy Research and Technique*, Vol. 69, pp. 186–195, 2006.
7. W. Becker, "Advanced Time-Correlated Single Photon Counting Techniques," *Springer Series in Chemical Physics*, 2005.
8. J.R. Lakowicz, *Principles of Fluorescence Spectroscopy,* 2d ed., Kluwer Academic/Plenum Publishers, 1999.
9. R. Hanbury Brown and R.Q. Twiss, "Correlation between Photons in Two Coherent Beams of Light," *Nature*, Vol. 177, pp. 27–29, 1956.
10. R.J. Glauber, "Coherent and Incoherent States of the Radiation Field," *Phys. Rev.*, Vol. 131, pp. 2766–2788, 1963.
11. D.L. Boiko, N.J. Gunther, N. Brauer, M. Sergio, C. Niclass, G. B. Beretta, et al., "A Quantum Imager for Intensity Correlated Photons," *New Journal of Physics*, Vol. 10, No. 11, November 2008.
12. C. Niclass, A. Rochas, P.A. Besse, and E. Charbon, "A CMOS Single Photon Avalanche Diode Array for 3D Imaging," *IEEE International Solid-State Circuits Conference* (ISSCC), pp. 120–121, February 2004.

13. C. Niclass and E. Charbon, "A Single Photon Detector Array with 64x64 Resolution and Millimetric Depth Accuracy for 3D Imaging," *IEEE International Solid-State Circuits Conference* (ISSCC), pp. 364–365, February 2005.
14. C. Niclass, A. Rochas, P.A. Besse, and E. Charbon, "Design and Characterization of a CMOS 3-D Image Sensor Based on Single Photon Avalanche Diodes," *IEEE Journal of Solid-State Circuits*, Vol. 40, No. 9, September 2005.
15. D. Stoppa et al., "A CMOS 3-D Imager Based on Single Photon Avalanche Diode," *Transactions on Circuits and Systems I*, 2007.
16. E. Charbon, "Will CMOS Imagers Ever Need Ultra-High Speed?" *IEEE International Conference on Solid-State and IC Technology*, Vol. 3, pp. 1975–1980, October 2004.
17. J. McPhate, J. Vallerga, A. Tremsin, O. Siegmund, B. Mikulec, and A. Clark, "Noiseless Kilohertz-frame-rate Imaging Detector based on Microchannel Plates Readout with Medipix2 CMOS Pixel Chip," *Proceedings of SPIE*, Vol. 5881, pp. 88–97, 2004.
18. T.G. Etoh et al., "Design of the PC-ISIS: Photon-Counting *In-Situ* Storage Image Sensor," *IEEE Workshop on CCDs and Advanced Image Sensors*, pp. 113–116, June 2005.
19. N. Kawai and S. Kawahito, "A Low-Noise Signal Readout Circuit Using Double-Stage Noise Cancelling Architecture for CMOS Image Sensors," *IEEE Workshop on CCDs and Advanced Image Sensors*, pp. 27–30, June 2005.
20. J. Hynecek, "Impactron—A New Solid State Image Intensifier," *IEEE Transactions on Electron Devices*, Vol. 48, No. 10, pp. 2238–2241, October 2001.
21. R. Shimizu et al., "A Charge-Multiplication CMOS Image Sensor Suitable for Low-Light-Level Imaging," *IEEE International Solid-State Circuits Conference* (ISSCC), pp. 50–51, February 2009.
22. T.G. Etoh et al., "An Image Sensor Which Captures 100 Consecutive Frames at 1,000,000 Frames/s," *IEEE Transactions on Electron Devices*, Vol. 50, No. 1, pp. 144–151, January 2003.
23. G. Patounakis, K. Shepard, and R. Levicky, "Active CMOS Biochip for Time-Resolved Fluorescence Detection," *IEEE Symposium on VLSI*, pp. 68–71, June 2005.
24. R.H. Haitz, "Studies on Optical Coupling between Silicon p-n Junctions," *Solid-State Electronics*, Vol. 8, pp. 417–425, 1965.
25. C. Niclass, M. Sergio, and E. Charbon, "A 64x48 Single Photon Avalanche Diode Array with Event-Driven Readout," *IEEE European Solid-State Circuits Conference* (ESSCIRC), pp. 556–559, September 2006.
26. C. Niclass, M. Sergio, and E. Charbon, "A Single Photon Avalanche Diode Array Fabricated in Deep-Submicron CMOS Technology," *Design Automation & Test in Europe* (DATE), pp. 79–84, March 2006.
27. C. Niclass, M. Sergio, and E. Charbon, "A Single Photon Avalanche Diode Array Fabricated in 0.35 μm CMOS and Based on an Event-Driven Readout for TCSPC Experiments," *Adv. Photon Counting Tech. Meeting* (Boston, MA), ed. W Becker Proc. SPIE 6372 paper 637205, pp. V216–V227, 2006.
28. D. Mosconi et al., "CMOS Single-photon Avalanche Diode Array for Time-Resolved Fluorescence Detection," *IEEE European Solid-State Circuit Conference* (ESSCIRC), pp. 564–567, September 2006.
29. B. Rae et al., "A Microsystem for Time-Resolved Fluorescence Analysis Using CMOS Single-Photon Avalanche Diodes and Micro-LEDs," *IEEE International Solid-State Circuits Conference* (ISSCC), pp. 166–167, February 2008.
30. S. Tisa, F. Guerrieri, A. Tosi, and F. Zappa, "100kframe/s 8 Bit Monolithic Single-Photon Imagers," *IEEE European Solid-State Device Conference* (ESSDERC), pp. 274–277, September 2008.

31. M.A. Marwick and A.G. Andreou, "Single Photon Avalanche Photodetector with Integrated Quenching Fabricated in TSMC 0.18μm CMOS Process," *Electronics Letters*, Vol. 44, Issue 21, No. 10, pp. 1284, October 2008.
32. N. Faramarzpour, M.J. Deen, S. Shirani, and Q. Fang, "Fully Integrated Single Photon Avalanche Diode Detector in Standard CMOS 0.18-μm Technology," *IEEE Transactions on Electron Devices*, Vol. 55, No. 3, pp. 760–767, March 2008.
33. C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implementation in 130-nm CMOS Technology," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 13, pp. 863–869, 2007.
34. M. Gersbach, C. Niclass, J. Richardson, R. Henderson, L. Grant, and E. Charbon, "A Single-Photon Detector Implemented in a 130nm CMOS Imaging Process," *IEEE European Solid-State Device Conference* (ESSDERC), pp. 270–273, September 2008.
35. M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, et al., "A Low-Noise Single-Photon Detector Implemented in a 130nm CMOS Imaging Process," *Solid-State Electronics*, Vol. 53, No. 7, pp. 803–808, July 2009.
36. J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, et al., "A 32x32 50ps Resolution 10 Bit Time to Digital Converter Array in 130nm CMOS for Time Correlated Imaging," *Custom Integrated Circuits Conference*, September 2009.
37. M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, et al., "A Parallel 32x32 Time-to-Digital Converter Array Fabricated in a 130nm Imaging CMOS Technology," *IEEE European Solid-State Circuits Conference* (ESSCIRC), September 2009.
38. D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R.K. Henderson, et al., "A 32x32-Pixel Array with In-Pixel Photon Counting and Arrival Time Measurement in the Analog Domain," *IEEE European Solid-State Circuits Conference* (ESSCIRC), September 2009.
39. R.J. McIntyre, "Recent Developments in Silicon Avalanche Photodiodes," *Measurement*, Vol. 3, No. 4, pp. 146–152, 1985.
40. S. Cova, A. Longoni, and A. Andreoni, "Towards Picosecond Resolution with Single-Photon Avalanche Diodes," *Rev. Sci. Instr.,* Vol. 52, No. 3, pp. 408–412, 1981.
41. F. Zappa et al., "Integrated Array of Avalanche Photodiodes for Single-Photon Counting," *IEEE European Solid-State Circuits Conference* (ESSCIRC), pp. 600–603, 1997.
42. W.J. Kindt, *Geiger Mode Avalanche Photodiode Arrays for Spatially Resolved Single Photon Counting,* Delft University Press, 1999.
43. B. Aull et al., "Geiger-Mode Avalanche Photodiodes for Three-Dimensional Imaging," *Lincoln Laboratory Journal*, Vol. 13, No. 2, pp. 335–350, 2002.
44. A. Rochas, "Single Photon Avalanche Diodes in CMOS Technology," Ph.D. Thesis, Lausanne, 2003.
45. C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, "Single-Photon Synchronous Detection," *IEEE European Solid-State Circuits Conference* (ESSCIRC), pp. 114–117, September 2008.
46. C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, "Single-Photon Synchronous Detection," *IEEE Journal of Solid-State Circuits*, Vol. 44, No. 7, pp. 1977–1989, July 2009.
47. S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche Photodiodes Quenching Circuits for Single-Photon Detection," *Applied Optics*, Vol. 35, No. 12, pp. 1956–1976, 1996.
48. J. Richardson, R. Henderson, and D. Renshaw, "Dynamic Quenching for Single Photon Avalanche Diode Arrays," International Imaging Sensor Workshop, June 2007.
49. Süss MicroOptics, http://www.suss-microoptics.com

50. G.R. Hopkinson, "Cobalt60 and Proton Radiation Effects on Large Format, 2-D, CCD Arrays for an Earth Imaging Application," *IEEE Transactions on Nuclear Science*, Vol. 39, pp. 2018–2025, December 1992.
51. E.-S. Eid, T.Y. Chan, E.R. Fossum, R.H. Tsai, R. Spagnuolo, J. Deily, et al., "Design and Characterization of Ionizing Radiation-Tolerant APS Image Sensors up to pp. 1796–1806, 30 Mrd (Si) Total Dose," *IEEE Transactions on Nuclear Science*, Vol. 48, No. 6, December 2001.
52. B.R. Hancock et al., "Multi-megarad (Si) Radiation Tolerant Integrated CMOS Imager," *SPIE Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications II*, Vol. 4306, pp. 147–155, 2001.
53. J. Bogaerts, B. Dierickx, and C. Van Hoof, "Radiation-Induced Dark Current Increase in CMOS Active Pixel Sensors," *SPIE Photonics for Space Environments VII*, Vol. 4134, pp. 105–114, 2000.
54. N. Scheidegger, R. Krpoun, H. Shea, C. Niclass, and E. Charbon, "A New Concept for a Low-Cost Earth Sensor: Imaging Oxygen Nightglow with Arrays of Single Photon Detectors," *30th Annual AAS Guidance and Control Conference*, pp. 501–517, February 2007.
55. L. Carrara, C. Niclass, N. Scheidegger, H. Shea, and E. Charbon, "A Gamma, X-ray and High Energy Proton Radiation-Tolerant CIS for Space Applications," *IEEE International Solid-State Circuits Conference* (ISSCC), pp. 40–41, February 2009.
56. C. Niclass, C. Favi, T.H. Kluter, M. Gersbach, and E. Charbon, "A 128x128 Single-Photon Imager with on-Chip Column-Level 10b Time-to-Digital-Converter Array Capable of 97ps Resolution," *IEEE International Solid-State Circuits Conference* (ISSCC), pp. 44–45, February 2008.

3 Effects of Hydrogen on the Radiation Response of Field-Oxide Field-Effect Transistors and High-*K* Dielectrics

Xing J. Zhou, Daniel M. Fleetwood,
and Ronald D. Schrimpf

CONTENTS

3.1 INTRODUCTION

Hydrogen can strongly affect the radiation response, long-term aging, and reliability of microelectronic devices and integrated circuits (ICs) [1,2]. Hydrogenous species and moisture exist in the oxide and surrounding materials of devices and ICs, especially for nonhermetically sealed IC packages where water can diffuse into critical gate and field oxides depending on device type, processing conditions, and/or storage conditions. Hydrogen can increase the oxide and interface-trap charge in the gate oxides of metal-oxide semiconductor (MOS) devices and ICs, especially in a radiation environment [1]. Moreover, hydrogenous species and radiation exposure can increase the low-frequency noise of MOS devices.

Low-frequency noise measurements are commonly used to characterize defect densities and energy distributions in the near-interfacial gate oxides of devices that are irradiated or exposed to high-field stress [3-14]. Noise measurements are seldom applied to evaluate defects in parasitic field oxides. However, the charge trapping in MOS field oxides more often limits the radiation response of modern complementary metal-oxide semiconductor (CMOS) devices than does the gate oxide response, especially for devices with SiO_2 or oxynitride gate dielectrics, owing to the continuing reduction in gate oxide thickness with technology scaling [15-17].

In a recent study, it was shown that low-frequency noise measurements can provide insight into the effects of moisture exposure on radiation-induced charge buildup in parasitic field oxides [18]. The test structures used in this work were parasitic field-oxide field-effect transistors (FOXFETs) built in a 130-nm CMOS technology that is used in high-energy physics applications [18-23]. Some of the devices were exposed to moisture after irradiation and annealing to help understand the potential roles of hydrogen and water in defect buildup and annealing. The effects of hydrogen and radiation exposure can be different in field oxide structures from what is found in typical MOS gate oxides [24-28].

In addition, hydrogen plays an important role in the radiation response and long-term reliability of high-K dielectric layers, which are important in advanced CMOS IC technologies, and are becoming especially crucial elements in sub-45 nm gate stacks. So this chapter also describes the effects of hydrogen on the irradiation and annealing responses of MOS devices with HfO_2 gate dielectrics. Again, different kinds of responses are observed in many cases from what is found for CMOS devices built in technologies using SiO_2 or oxynitride gate dielectrics [24-30]. These results emphasize the need to continue to investigate the effects of hydrogen and radiation response on MOS gate and field-oxide dielectric responses, especially as IC technologies continue to employ more complex gate stacks and surrounding materials.

3.2 BACKGROUND ON 1/*f* NOISE

Before describing the results of the noise measurements that were performed on the FOXFETs, it is useful to understand how radiation exposure and the consequent buildup of defects in dielectric layers and at dielectric-to-semiconductor interfaces affect MOS 1/*f* noise [31,32]. After irradiation, MOS device structures typically exhibit both an increase in the fixed-charge density within the oxide and an increase in the interface-trap concentration, resulting in a reduction of the transconductance and a change in the threshold voltage. Additionally, the low-frequency noise increases [11,33-36]. This increase typically has a strong correlation with oxide-trapped charge but not usually with interface-trap charge [11,33], leading to the conclusion that oxide traps within a few nm of the Si-SiO_2 interface, defined as border traps, are responsible for 1/*f* noise in MOS devices [36]. The usual, first-order number fluctuation model assumes that the 1/*f* noise is due primarily to charge trapping and emission [11,33]. In particular, density-functional theory and 1/*f* noise measurements suggest that the 1/*f* noise of n-channel MOS devices is caused by the capture and emission of electrons at oxygen vacancy defects near the Si/SiO_2 interface [37,38]. These processes likely are accompanied by significant SiO_2 network relaxation, involving an oxygen vacancy defect in SiO_2 that is either initially charged positively or neutral. In order for the trap centers to become a 1/*f* noise source, there needs to be a large number of traps available at suitable energy levels or locations. This is typically the case in MOS gate dielectrics and is certainly the case in parasitic field oxides, which are much thicker and inferior in quality to MOS gate oxides.

A variety of models have been proposed to explain 1/*f* noise in MOSFETs [39-48]. Fluctuations in the oxide-trap charge couple to the channel, both directly through fluctuations in the numbers of inversion layer charges and indirectly through fluctuations in scattering rates that are associated with changes in trap occupancy. Data from narrow-channel MOSFETs confirm that both effects can be important [49]. In general, noise studies on n-channel MOSFETs tend to follow a number fluctuation model, at least to the first order. In p-channel devices, the noise is often attributed to both number and mobility fluctuations [33], although there is significant evidence that much pMOS noise may also be dominated by number fluctuations [10].

3.3 EXPERIMENTAL DETAILS

The FOXFET devices in this study are from a commercial 130 nm CMOS technology using the n-shallow trench isolation (STI) oxide as the gate dielectric for the test structure. The cross section is shown schematically in Figure 3.1. The gate is poly-crystalline Si. The source and drain electrodes are made via extended n-well contacts. The channel is formed at the bottom of the isolated STI oxide. The channel length, *L,* is 1.48 µm or 0.92 µm, and the channel width, *W,* is 200 µm. These structures are useful for evaluating defects in the isolation oxide [18-21]. Additional details of the device fabrication are provided in [22].

The FOXFETs were irradiated at room temperature with 10 keV X-rays to 300 krad(SiO_2) at a dose rate of 31 krad(SiO_2)/min. During the irradiation, all terminals of the devices were grounded except the gate, which was biased at 2.5 V. The current-voltage (I_d-V_g) characteristics and the excess noise power spectral density, S_v (corrected for background noise), were monitored as a function of frequency, *f,* pre- and post-irradiation, using the circuit shown in Figure 3.2. All noise measurements

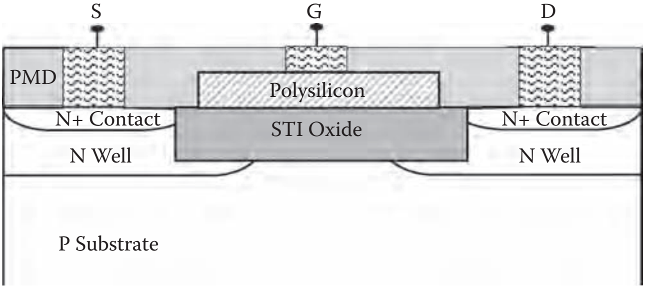


FIGURE 3.1 Schematic diagram of n-well FOXFET structure. (After X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, F. Faccio, and L. Gonella, *IEEE Trans. Nucl. Sci.*, 55, 2975–2980, 2008.)

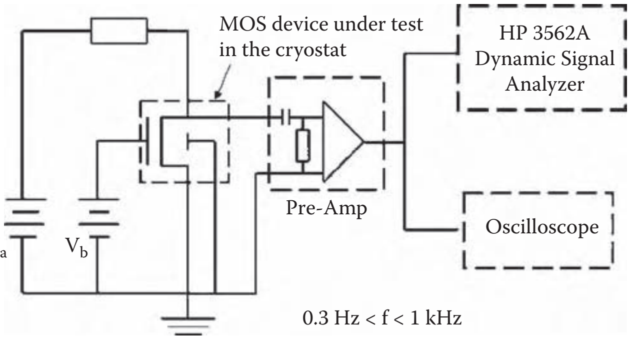


FIGURE 3.2 1/*f* noise measurement circuit diagram. The box in series with V_a is a variable resistor, which was typically set to ~20 kΩ for the noise measurements reported here. The oscilloscope was not connected to the circuit during the actual noise measurements. (After H. D. Xiong, D. M. Fleetwood, B. K. Choi, and A. L. Sternberg, *IEEE Trans. Nucl. Sci.,* 49, 2718–2723, 2002.)

reported here were taken at room temperature. The results shown are representative of the responses of several devices, which showed similar responses. The resistor shown in series with the transistor channel limits the current and controls the bias point. The drain to source voltage noise was amplified by a Standard Research SR560 low-noise preamplifier in the 1 Hz to 1 kHz frequency range. The output of the preamplifier was connected to an HP 3562A dynamic signal analyzer, which recorded the noise measurements, and was under computer control for data storage purposes. For the following FOXFET experimental results, we measure the noise power spectral density of the drain-source voltage. The n-type MOSFET device is operated in the linear region in strong inversion, so the number fluctuation model should apply to describe the resulting noise [3-14,18]. In this regime, electrical and device properties such as electrical field, channel carrier density, and depletion length are assumed to be roughly constant along the channel. Deviations from this assumption can lead to difficulties in relating noise magnitudes to the underlying defect densities. For all noise measurements, background noise measurements were made at each gate bias with zero channel current [33,34]. The background noise is mostly composed of preamplifier noise and thermal noise. All the following low-frequency noise data reflect noise spectra after background noise subtraction.

After noise measurements, the devices were annealed at room temperature at 0 V for 48 hours to allow their threshold voltage and noise to stabilize. Some devices were exposed to 85% relative humidity at 130°C for 72 hours after noise measurement and irradiation to determine whether the combination of high temperature and moisture might efficiently passivate or activate defects in the dielectric layers and at the isolation oxide to Si interface. During the noise measurements, the drain voltage, V_d, was kept at 500 mV, and the noise spectral density, S_v, was measured as a function of gate bias. The gate bias was usually 2 to 12 V above the threshold to ensure the device was operating in the linear regime [50]. Threshold voltage shifts due to oxide and interface-trap density were calculated by the midgap charge separation technique of Winokur et al. [24].

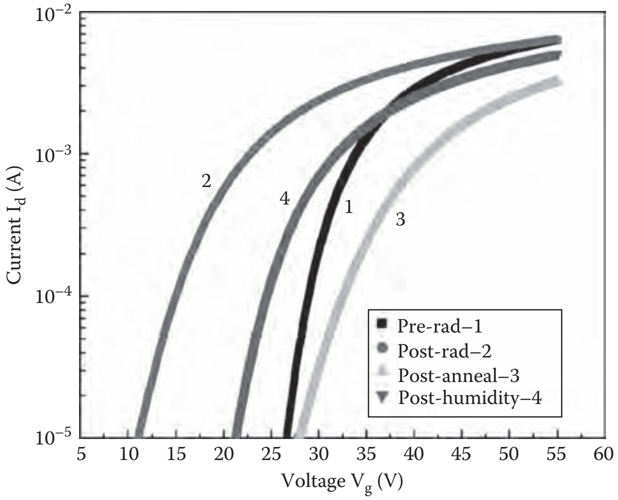


FIGURE 3.3 I_d-V_g characteristics of FOXFETs (1) before irradiation, (2) after 300 krad(SiO_2) irradiation at 2.5 V bias, (3) after irradiation and annealing at room temperature for two days at 0 V bias, and (4) after subsequent moisture exposure for three days at 130°C at 0 V. (After X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, F. Faccio, and L. Gonella, *IEEE Trans. Nucl. Sci.*, 55, 2975–2980, 2008.)

3.4 RESULTS AND DISCUSSION

3.4.1 Electrical Measurements

Figure 3.3 shows I_d-V_g characteristics of FOXFETs (1) before irradiation, (2) post-irradiation, (3) post-room-temperature annealing, and (4) post-humidity exposure. These operations were done in sequence, so the post-humidity devices previously experienced both irradiation and room-temperature annealing before the humidity testing was performed. From linear I_d-V_g curves (not shown here) using standard extrapolation techniques, the threshold voltage for the devices is ~30 V before irradiation. After irradiation to 300 krad(SiO_2) with 10 keV X-rays, the threshold voltage shifts from ~30 V to ~18 V. Threshold voltages after the post-irradiation anneal and then after the subsequent humidity exposure are 34 V and 28 V, respectively.

To separate MOS oxide and interface-trap charge, we use the midgap charge separation technique developed by Winokur et al. [24], which assumes that interface traps are net charge neutral at midgap. Thus, the voltage shifts at midgap are primarily due to oxide-trap charge buildup. Hence [24, 51],

$$\Delta N_{\text{ot}} = -C_{\text{ox}}\Delta V_{\text{mg}} / qA \tag{3.1}$$

$$\Delta N_{\text{it}} = C_{\text{ox}}(\Delta V_{\text{fb}} - \Delta V_{\text{mg}}) / qA \tag{3.2}$$

where C_{ox} is the oxide capacitance, $-q$ is the electronic charge, A is the area, ΔV_{mg} is the midgap voltage shift, and ΔV_{fb} is the flatband voltage shift. Threshold-voltage

TABLE 3.1
Threshold Voltage Shifts Due to Oxide- and Interface-Trap Charges for the Devices and Conditions Employed in This Study

	ΔV_{th} (V)	ΔV_{it} (V)	ΔV_{ot} (V)
Post-radiation	−12.5	~0.0	~−12.5
Post-anneal	5.0	13.1	−8.1
Post-humidity	−3.3	5.2	−8.5

Source: L.Gonella, F. Faccio, M. Silvestri, S. Gerardin, D. Pantano, V. Re, M. Marighisoni, L. Ratti, and A. Ranieri, "Total ionizing dose effects in 130-nm commercial CMOS technologies for HEP experiments," *Nucl. Inst. Meth. Phys. Res. A.*, vol. 582, no. 3, pp. 750–754, Dec. 2007.

shifts due to net oxide-trap and interface-trap charge are plotted as a function of the different experimental conditions in Table 3.1.

After irradiation, the threshold voltage shift is dominated primarily by hole trapping in these devices, with no measurable contribution to the threshold-voltage shift due to interface traps at early times following irradiation [52-54]. The effective areal density of oxide-trap charge ΔN_{ot}, projected to the interface, is ~1.7 × 10^{11} cm^{-2} immediately after irradiation. During the post-irradiation annealing, ΔV_{ot} decreases by ~40%, which remains approximately constant during the moisture exposure. The interface-trap density ΔN_{it} increased by ~1.9 × 10^{11} cm^{-2} during the room-temperature annealing and decreases dramatically during moisture exposure [18,21].

3.4.2 Noise Measurements

The excess low-frequency noise power spectral density, S_v, is the difference between the noise measured with drain current flowing and the background noise at the same gate voltage, with $V_d = 0$ V. Figure 3.4 shows S_v as a function of frequency for the devices before irradiation (squares), after irradiation (circles), after irradiation and annealing (triangles pointing up), and after irradiation, annealing, and humidity exposure (triangles pointing down). The drain voltage was kept at 500 mV during the measurement, and the applied gate voltage is 8 V above threshold to maintain linear operation of devices. Typical noise measurements take ~10 minutes for each gate bias in these cases; some annealing will inevitably occur in the "post-irradiation" case, owing to the large positive bias [18]. Significant $1/f^{\gamma}$ noise is observed; the frequency exponent, γ, is very close to unity ($0.8 \leq \gamma \leq 1.1$) for these devices and measurement conditions. The noise increases after irradiation, drops significantly after annealing,

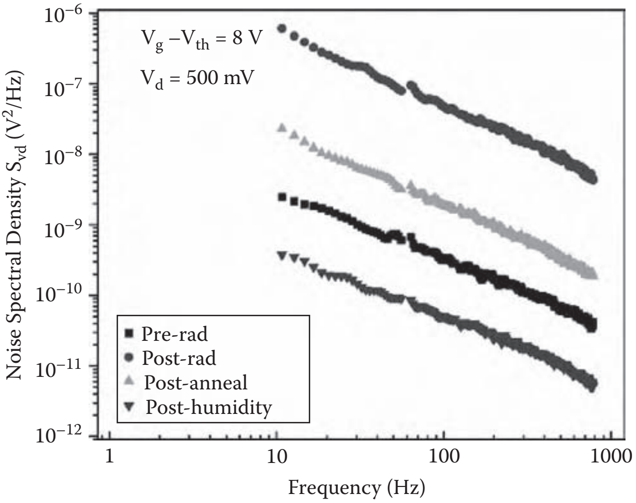


FIGURE 3.4 FOXFET noise for the devices and experimental conditions of Figure 3.3 and Table 3.1. (After X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, F. Faccio, and L. Gonella, *IEEE Trans. Nucl. Sci.*, 55, 2975–2980, 2008.)

and falls even more significantly after humidity exposure. The posthumidity noise is much lower than the preirradiation noise.

Figure 3.5 shows the gate voltage dependence of the noise as a function of frequency, after irradiation, and annealing. Qualitatively consistent with previous studies of low-frequency noise in MOS gate oxides [3-14], the noise decreases significantly with increasing gate voltage. Figure 3.6 shows the normalized noise magnitude, *K*, as a function of $V_g - V_{th}$. Here

$$K = S_{v_d} \frac{f^{\gamma}(V_g - V_{th})^2}{V_d^2} \tag{3.3}$$

where V_g, V_{th}, and V_d are the gate voltage, the threshold voltage, and the drain voltage [8-11], respectively. If the FOXFET noise is due primarily to number fluctuations (i.e., conductivity fluctuations due to carrier trapping and emission) and if the defects responsible for the noise are distributed approximately evenly in energy in the SiO_2 band gap, then one would expect *K* to be approximately constant with $V_g - V_{th}$. In all cases, except immediately after irradiation, this expectation is fulfilled [18], which suggests that the effective border trap density is approximately constant over the energy range covered by the noise measurements except for the curves measured just following the irradiation. Hence, the noise measurements suggest that the effective border trap densities increase dramatically with irradiation, decrease significantly with room temperature annealing, and then decrease even further to levels below pre-irradiation densities with the subsequent elevated temperature and moisture exposure [18].

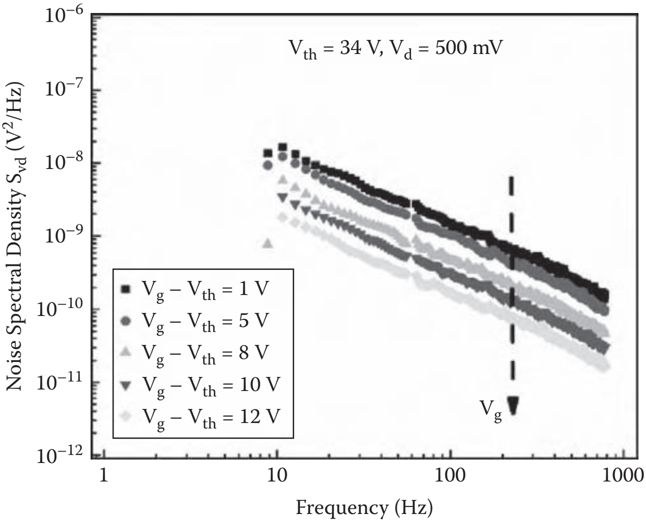


FIGURE 3.5 Noise spectral density S_v as a function of frequency at different gate biases for the irradiated and room-temperature annealed devices of Figure 3.4. (After X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, F. Faccio, and L. Gonella, *IEEE Trans. Nucl. Sci.*, 55, 2975–2980, 2008.)

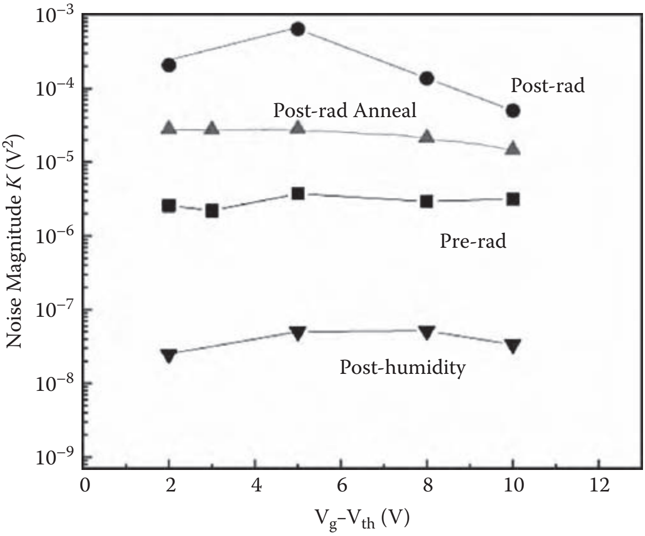


FIGURE 3.6 Normalized noise magnitude K as a function of gate bias for the devices and experimental conditions of Figures 3.3 through 3.5. (After X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, F. Faccio, and L. Gonella, *IEEE Trans. Nucl. Sci.*, 55, 2975–2980, 2008.)

The relative stability of ΔN_{ot} and the significant reductions in ΔN_{it} and noise are consistent with the passivation of process-induced and radiation-induced interface and border traps by water. The results of Figures 3.4 through 3.6 thus suggest that reactions of H_2O at and near the FOXFET channel may explain the reductions in ΔN_{it} and noise with humidity exposure [18]. These passivation reactions will occur in parallel with defect-inducing interactions of hydrogenous species that have been noted in studies of aging and moisture effects on MOS devices [55-58] and may mitigate some effects of aging on the long-term device performance, reliability, and radiation response of MOS devices [59].

3.5 HIGH-*K* DIELECTRICS

Radiation effects studies performed on high-*K* MOS devices following X-ray irradiation also show an important role for hydrogen in many cases. Here we illustrate this via a detailed study of the effects of switched-bias annealing after radiation exposure [60]. Capacitors were fabricated on *p*-type Si (100) wafers, and high-κ layers (HfO_2) were deposited by atomic layer deposition at 300°C. There is a thin oxynitride layer (SiO_xN_y) between the HfO_2 gate dielectric and the *p*-type Si substrate for these devices to improve the interface properties of the devices [61]. The physical thickness of the HfO_2 layer was 6.8 nm, as measured ellipsometrically; the interfacial oxynitride layer was 1.0 nm. The relative dielectric constants, κ, of the HfO_2 and interfacial oxynitride layer (SiO_xN_y) are ~20 and ~4, resulting in an equivalent oxide thickness (EOT) of 2.1 nm. Al was deposited to form gate electrodes.

MOS capacitors were irradiated with 10 keV X-rays to 1 Mrad(SiO_2) at a dose rate of 517 rad(SiO_2)/s with an oxide electric field of 2 MV/cm. Alternating negative and positive bias-temperature annealing at ±2 MV/cm was performed at temperatures from 50°C to 150°C after the irradiation exposure. Flatband-voltage shifts due to net oxide-trap charge ΔV_{ot} and interface-trap charge ΔV_{it} were estimated from high-frequency (1 MHz) capacitance-voltage (C-V) measurements via the midgap charge separation method [24] after cooling the devices to room temperature.

Figure 3.7 shows irradiation and annealing results for Al/HfO_2 + SiO_xN_y/Si pMOS capacitors irradiated to 1 Mrad(SiO_2) at 0.3 V. Values of ΔV_{ot} and ΔV_{it} increase in magnitude for negative bias-temperature stress (NBTS) and decrease in magnitude for positive bias-temperature stress (PBTS). A significant fraction of this reversibility in ΔV_{ot} during the annealing periods in Figure 3.7a is similar to switched-bias experiments for irradiated thermal SiO_2 [29,36,62-67]. This occurs because both positive and negative charges are trapped during the irradiation. Then compensating electrons are released during NBTS and are captured during PBTS [60,68]. Different defects participate in the reversibility of charge trapping for HfO_2 dielectrics from that for thermal SiO_2, but O vacancies almost certainly play a key role case [60,62-68].

It is quite interesting that the reversibility in ΔV_{it} after irradiation in Figure 3.7b is more pronounced than is typical for similar irradiation and annealing sequences for thermal SiO_2 [62-67]. This cannot be explained easily by the two-stage buildup of interface traps associated with the release of protons in the gate dielectric and their subsequent transport under bias and reactions at the Si/dielectric interface

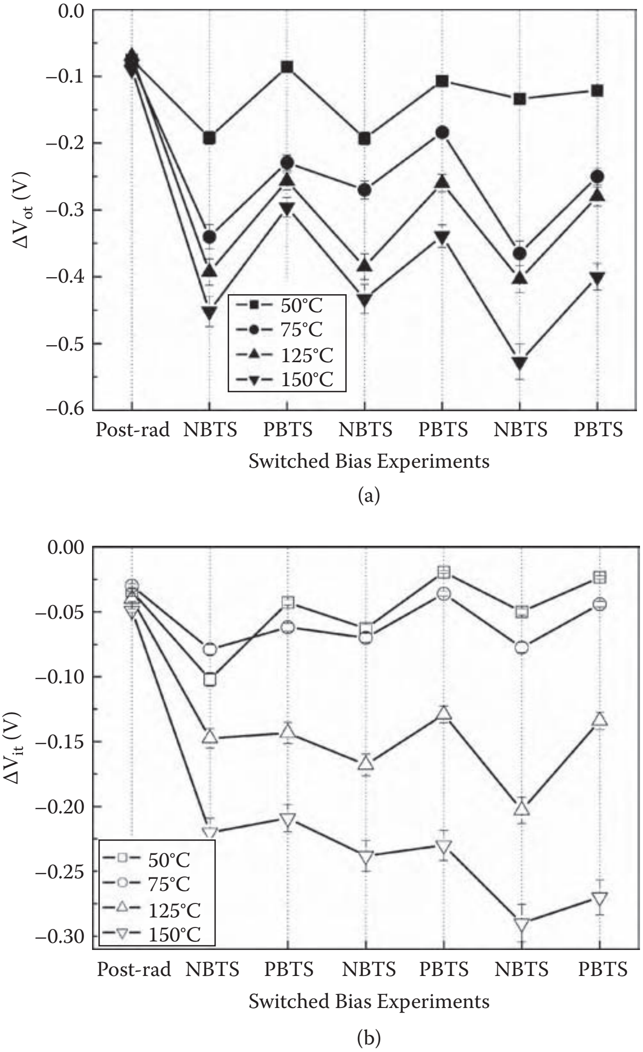


FIGURE 3.7 (a) Induced ΔV_{ot} and (b) ΔV_{it} for Al/HfO$_2$ + SiO$_x$N$_y$/Si pMOS capacitors irradiated to 1.0 Mrad(SiO$_2$) with 10 keV X-rays, followed by switched bias anneals at 50 to 150°C. The gate bias for irradiation is 0.3 V. The switched bias anneals are ±0.3 V (PBTS, NBTS), and the stress time for both is 600 s. (After X. J. Zhou, D. M. Fleetwood, L. Tsetseris, R. D. Schrimpf, and S. T. Pantelides, *IEEE Trans. Nucl. Sci.*, 53, 3636–3643, 2006.)

[1,28,69,70]. For this case, interface traps typically increase in magnitude during positive bias annealing and stay approximately constant during negative-bias annealing. Interface-trap reversibility has been seen in switched-bias annealing experiments performed on SiO_2 devices at elevated temperatures [71] when significant densities of positive oxide-trap charge and hydrogenous species are simultaneously present in MOS devices. Similar reversibility in ΔV_{ot} and ΔV_{it} is also observed after constant-voltage stress [60].

The reversibility in ΔV_{it} (as well as some variability in ΔV_{ot}) in Figure 3.7 evidently is associated with the motion, trapping, and reactions of protons near the Si/dielectric interface. During the negative bias annealing, H^+ drift to the interface from the oxide is inhibited by the applied electric field, and Si dangling bonds at the dielectric-to-Si interface are positively charged. In this case, passivation of dangling bonds by H (illustrated in reaction (3.4)) is suppressed, since both species are of the same charge.

$$H^+ + Si^- \rightarrow Si\text{–}H \tag{3.4}$$

However, the depassivation of passivated dangling bonds can still occur via reaction (3.5) [72]:

$$Si\text{–}H + H^+ \rightarrow Si^+ + H_2 \tag{3.5}$$

This reaction can lead to an increase of ΔV_{it} in magnitude during the negative bias anneal, if there is a source of hydrogen either at the dielectric to dielectric layer interface or in the Si substrate. Possible sources of hydrogen in *p*-type Si are B-H complexes [72] or oxygen protrusions [51], as identified in previous work on negative bias-temperature instabilities. Depassivation of a Si-H bond and the formation of an interface trap via reaction (3.5) [73,74] are illustrated schematically as mechanism (2) in Figure 3.8a.

Other mechanisms breaking Si-H bonds may contribute to effects observed in Figure 3.7, as also shown schematically in Figure 3.8a (mechanisms (3) and (4)). The release of a proton can lead to the formation of an interface trap and a proton that can be trapped in the dielectric layer. Simple thermally assisted Si–H bond breaking is highly improbable for a passivated dangling bond under normal device operating conditions [72]. But density functional theory calculations show that the presence of an impurity Hf atom in the near-interfacial oxynitride (mechanism (3) in Figure 3.8) can facilitate the "shuttling" of a proton between the Hf atom and the interface [75]. These Hf atoms can be incorporated into the SiO_2 interlayer between the HfO_2 gate dielectric and the Si substrate during rapid thermal annealing, as evidenced by scanning transmission electron microscopy, and can result in additional mobility degradation [76-78]. Density functional theory calculations show that Hf assisted proton shuttling likely dominates over suboxide bond assisted shuttling (mechanism (4) in Figure 3.8), owing to the reduced barrier for proton motion in the presence of Hf [75].

Under positive bias (Figure 3.8b), the decrease in ΔV_{it} occurs because of the passivation of negatively charged Si dangling bonds by protons. The protons either can be released and transported from the oxide or can be released from Hf-H or suboxide

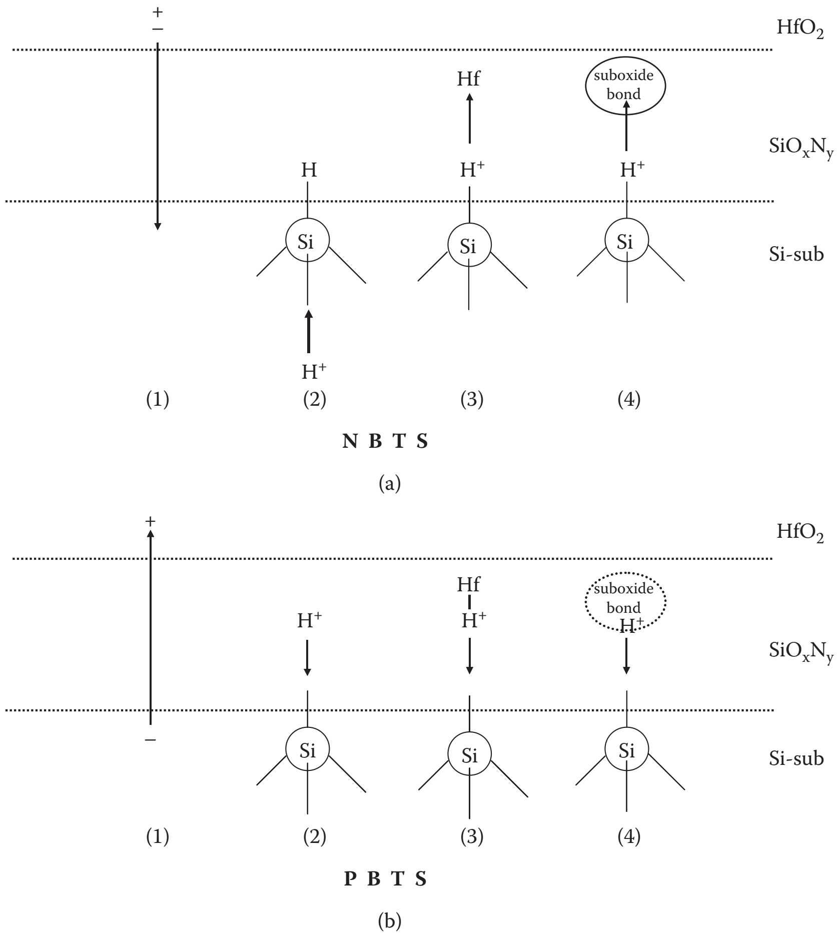


FIGURE 3.8 Schematic diagram of the processes that can lead to oxide- and interface-trap charge reversibility in HfO$_2$ based high-*K* dielectrics, after ionizing radiation exposure, (a) during negative-bias annealing, and (b) during positive-bias annealing. (After X. J. Zhou, D. M. Fleetwood, L. Tsetseris, R. D. Schrimpf, and S. T. Pantelides, *IEEE Trans. Nucl. Sci.*, 53, 3636–3643, 2006.)

bonds, illustrated by mechanisms (2) through (4) in Figure 3.8b. There is a relatively large initial density of Si dangling bonds and a relatively high concentration of protons in the near-interfacial SiO$_2$, since HfO$_2$ is a weak diffusion barrier for hydrogenous species [79]. The protons can passivate a preexisting defect via reaction (3.2). A similar mechanism has been observed to lead to a decrease in ΔV_{it} during irradiation for some high-κ dielectrics [80], emphasizing the significance of these hydrogen effects in high-*K* devices. Once the defect is passivated by hydrogen, it no longer is an interface trap, therefore reducing ΔV_{it} in magnitude. This leads to an increase in magnitude of both ΔV_{ot} and ΔV_{it} during NBTS (more trapped protons in the oxide; more unpassivated dangling bonds) and to a decrease in magnitude of ΔV_{ot} and ΔV_{it}

during PBTS (fewer trapped protons in the oxide; fewer dangling bonds), consistent with the trends in the data of Figure 3.7 [60,75]. This emphasizes the differences in hydrogen reactions in high-*K* materials compared with thermal SiO_2. The degree to which hydrogen shuttling is observed in high-K gate dielectrics can vary and is strongly affected by the detailed processing conditions [61,68,81-83].

3.6 SUMMARY AND CONCLUSIONS

We have evaluated the radiation response and low-frequency (1/*f*) noise of FOXFETs before and after irradiation, after post-irradiation annealing at room temperature, and after moisture exposure at elevated temperatures. The noise magnitude increases after irradiation and decreases after post-irradiation annealing and humidity exposure. The noise level after humidity exposure is well below the preirradiation noise magnitude. Significant passivation of interface and border traps is observed in these devices upon moisture exposure of an irradiated FOXFET. These interface-trap and border-trap passivation processes will mitigate some kinds of aging and moisture effects for MOS devices in nonhermetic radiation environments.

In addition, it was shown that hydrogen-related defects can increase the post-irradiation reversibility in interface- and oxide-trap charge densities in high-K gate dielectrics. The presence of Hf atoms in the near-interfacial gate dielectric layer facilitates the shuttling of protons between a charged state in the dielectric layer and the interface. The degree to which hydrogen affects the radiation response and long-term reliability of high-*K* gate dielectrics is a strong function of the materials employed and the particular device processing conditions. The effects can differ dramatically from those typically observed in SiO_2- or oxynitride-based Si devices.

Taken together, these results illustrate that hydrogen continues to play a strong role in the radiation response and reliability of advanced microelectronic devices. Hydrogen and moisture can alter field-oxide and gate dielectric response, often in surprising ways that are difficult to predict in advance of detailed characterization studies. This emphasizes the continuing need to explore the role of hydrogen in defect creation and passivation both experimentally and theoretically and the need to develop refined radiation and reliability test methods for sub 45 nm MOS technologies that will incorporate an ever-increasing number of new materials and more and more complex device structures.

ACKNOWLEDGMENTS

We thank S. A. Francis, A. Dasgupta, F. Faccio, L. Gonella, M. R. Shaneyfelt, J. R. Schwank, I. G. Batyrev, B. R. Tuttle, and S. T. Pantelides for stimulating discussions.

REFERENCES

1. D. M. Fleetwood, “Effects of hydrogen transport and reactions on microelectronics radiation response and reliability,” *Microelectron. Reliab.*, vol. 42, no. 4–5, pp. 523–541, Jul. 2002.

2. S. T. Pantelides, L. Tsetseris, S. N. Rashkeev, X. J. Zhou, D. M. Fleetwood, and R. D. Schrimpf, "Hydrogen in MOSFETs—a primary agent of reliability issues," *Microelectron. Reliab.*, vol. 47, no. 7, pp. 903–911, Jul. 2007.
3. Z. H. Fang, S. Cristoloveanu, and A. Chovet, "Analysis of hot-carrier-induced aging from 1/*f* noise in short-channel MOSFETs," *IEEE Electron Dev. Lett.*, vol. 7, no. 6, pp. 371–373, Jun. 1986.
4. D. M. Fleetwood and J. H. Scofield, "Evidence that similar point defects cause 1/*f* noise and radiation-induced-hole trapping in MOS devices," *Phys. Rev. Lett.*, vol. 64, pp. 579–582, Jan. 1990.
5. D. M. Fleetwood, W. L. Warren, M. R. Shaneyfelt, R. A. B. Devine, and J. H. Scofield, "Enhanced MOS 1/*f* noise due to near-interfacial oxygen deficiency," *J. Non-Crystalline Solids*, vol. 187, pp. 199–205, Dec. 1995.
6. J. A. Babcock, J. L. Titus, R. D. Schrimpf, and K. F. Galloway, "Effects of ionizing radiation on the noise properties of DMOS power transistors," *IEEE Trans. Nucl. Sci.,* vol. 38, no. 6, pp. 1304–1309, Dec. 1991.
7. M. H. Tsai and T. P. Ma, "Effect of radiation-induced interface traps on 1/*f* noise in MOSFETs," *IEEE Trans. Nucl. Sci.,* vol. 39, no. 6, pp. 2178–2185, Dec. 1992.
8. D. M. Fleetwood, M. R. Shaneyfelt, W. L. Warren, J. R. Schwank, T. L. Meisenheimer, and P. S. Winokur, "Border traps: issues for MOS radiation response and long-term reliability," *Microelectron. Reliab.*, vol. 35, pp. 403–428, Mar. 1995.
9. D. M. Fleetwood, M. R. Shaneyfelt, and J. R. Schwank, "Estimating oxide, interface, and border-trap densities in MOS Transistors," *Appl. Phys. Lett.,* vol. 64, no. 15, pp. 1965–1967, Apr. 1994.
10. J. H. Scofield, N. Borland, and D. M. Fleetwood, "Reconciliation of different gate-voltage dependencies of 1/*f* noise in n-MOS and p-MOS transistors," *IEEE Trans. Electron Dev.*, vol. 41, no. 11, pp. 1946–1952, Nov. 1994.
11. D. M. Fleetwood, T. L. Meisenheimer, and J. H. Scofield, "1/*f* noise and radiation effects in MOS devices," *IEEE Trans. Electron Dev.*, vol. 41, no. 11, pp. 1953–1964, Nov. 1994.
12. E. Simoen and C. Claeys, "The low-frequency noise behavior of SOI technologies," *Solid-St. Electron.*, vol. 39, no. 7, pp. 949–960, Jul. 1996.
13. E. Simoen and C. Claeys, "On the flicker noise in submicron silicon MOSFETs," *Solid-St. Electon.*, vol. 43, no. 5, pp. 865–882, May 1999.
14. D. M. Fleetwood, H. D. Xiong, Z.-Y. Lu, C. J. Nicklaw, J. A. Felix, R. D. Schrimpf, et al., "Unified model of hole trapping, 1/*f* noise, and thermally stimulated current in MOS devices," *IEEE Trans. Nucl. Sci.,* vol. 49, no. 6, pp. 2674–2683, Dec. 2002.
15. J. M. Terrell, T. R. Oldham, A. J. Lelis, and J. M. Benedetto, "Time dependent annealing of radiation-induced leakage currents in MOS devices," *IEEE Trans. Nucl. Sci.*, vol. 36, no. 6, pp. 2205–2211, Dec. 1989.
16. M. R. Shaneyfelt, P. E. Dodd, B. L. Draper, and R. S. Flores, "Challenges in hardening technologies using shallow-trench isolation," *IEEE Trans. Nucl. Sci.*, vol. 45, no. 6, pp. 2584–2592, Dec. 1998.
17. H. L. Hughes and J. M. Benedetto, "Radiation effects and hardening of MOS technology: devices and circuits," *IEEE Trans. Nucl. Sci.*, vol. 50, no. 3, pp. 500–521, Jun. 2003.
18. X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, F. Faccio, and L. Gonella, "Radiation effects on the 1/*f* noise of field oxide field effect transistors," *IEEE Trans. Nucl. Sci.*, vol. 55, no. 6, pp. 2975–2980, Dec. 2008.
19. F. Faccio and G. Cervelli, "Radiation-induced edge effects in deep submicron CMOS transistors," *IEEE Trans. Nucl. Sci.*, vol. 52, no. 6, pp. 2413–2420, Dec. 2005.
20. V. Re, M. Manghisoni, L. Ratti, V. Speziali, and G. Traversi, "Total ionizing dose effects on the noise performances of a 0.13 μm CMOS technology," *IEEE Trans. Nucl. Sci.*, vol. 53, no. 3, pp. 1599–1606, Jun. 2006.

21. L. Gonella, F. Faccio, M. Silvestri, S. Gerardin, D. Pantano, V. Re, et al., "Total ionizing dose effects in 130-nm commercial CMOS technologies for HEP experiments," *Nucl. Inst. Meth. Phys. Res. A.*, vol. 582, no. 3, pp. 750–754, Dec. 2007.
22. F. Faccio, H. J. Barnaby, X. J. Chen, D. M. Fleetwood, L. Gonella, M. McLain, and R. D. Schrimpf, "Total ionizing dose effects in shallow trench isolation oxides," *Microelectron. Reliab.*, vol 48, no. 7, pp. 1000–1007, Jul. 2008.
23. M. Silvestri, S. Gerardin, A. Paccagnella, F. Faccio, "Degradation induced by X-ray irradiation and channel hot carrier stresses in 130-nm NMOSFETs with enclosed layout," *IEEE Trans. Nucl. Sci.*, vol. 55, no. 6, pp. 3216–3223, Dec. 2006.
24. P. S. Winokur, J. R. Schwank, P. J. McWhorter, P. V. Dressendorfer, and D. C. Turpin, "Correlating the radiation response of MOS capacitors and transistors," *IEEE Trans. Nucl. Sci.,* vol. 31, no. 6, pp. 1453–1460, Dec. 1984.
25. C. M. Svensson, "The defect structure of the Si-SiO$_2$ interface, a model based on trivalent silicon and its hydrogen compounds," in *The physics of SiO$_2$ and its interface*, edited by S. T. Pantelides (New York: Pergamon Press, 1978), pp. 328–332.
26. A. G. Revesz, "Chemical and structural aspects of the irradiation behavior of SiO$_2$ films on silicon," *IEEE Trans. Nucl. Sci.,* vol. 24, no. 6, pp. 2102–2107, Dec. 1977.
27. A. G. Revesz, "Hydrogen in SiO$_2$ films on silicon," in *The physics of SiO$_2$ and its interface*, edited by S. T. Pantelides, New York: Pergamon Press, 1978, pp. 222–226.
28. F. B. McLean, "A framework for understanding radiation-induced interface states in SiO$_2$ MOS structures," *IEEE Trans. Nucl. Sci.,* vol. 27, no. 6, pp. 1651–1657, Dec. 1980.
29. J. R. Schwank, P. S. Winokur, P. J. McWhorter, F. W. Sexton, P. V. Dressendorfer, and D. C. Turpin, "Physical mechanisms contributing to device rebound," *IEEE Trans. Nucl. Sci.,* vol. 31, no. 6, pp. 1434–1438, Dec. 1984.
30. D. M. Fleetwood and H. A. Eisen, "Total-dose radiation hardness assurance," *IEEE Trans. Nucl. Sci.*, vol. 50, no. 3, pp. 552–564, June 2003.
31. A. van der Ziel, *Noise in solid state devices and circuits*, New York: John Wiley & Sons, 1986.
32. A. A. Balandin, *Noise and fluctuations control in electronic devices*, American Scientific Publishers, 2002.
33. T. L. Meiseheimer and D. M. Fleetwood, "Effect of radiation-induced charge on 1/*f* noise in MOS devices," *IEEE Trans. Nucl. Sci.,* vol. 37, pp. 1696–1702, 1990.
34. J. H. Scofield, T. P. Doerr, and D. M. Fleetwood, "Correlation between preirradiation 1/*f* noise and postirradiation oxide-trapped charge in MOS transistors," *IEEE Trans. Nucl. Sci.,* vol. 36, pp. 1946–1953, 1989.
35. T. L. Meisenheimer, D. M. Fleetwood, M. R. Shaneyfelt, and L. R. Riewe, "1/*f* noise in n-channel and p-channel MOS devices through irradiation and annealing," *IEEE Trans. Nucl. Sci.,* vol. 38, pp. 1297–1303, 1991.
36. D. M. Fleetwood, P. S. Winokur, R. A. Reber, T. L. Meisenheimer, J. R. Schwank, M. R. Shaneyfelt, L. C. Riewe, "Effects of oxide traps, interface traps, and border traps on metal-oxide-semiconductor devices," *J. Appl. Phys.*, vol. 73, pp. 5058–5074, 1993.
37. H. D. Xiong, D. M. Fleetwood, B. K. Choi, and A. L. Sternberg, "Temperature dependence and irradiation response of 1/*f* noise in MOSFETs," *IEEE Trans. Nucl. Sci.,* vol. 49, pp. 2718–2723, 2002.
38. D. M. Fleetwood, M. J. Johnson, T. L. Meisenheimer, P. S. Winokur, W. L. Warren, and S. C. Witczak, "1/*f* noise, hydrogen transport, and latent interface-trap buildup in irradiated MOS devices," *IEEE Trans. Nucl. Sci.,* vol. 44, no. 6, pp. 1810–1817, Dec. 1997.
39. J. J. Simonne, G. Blasquez, and G. Barbottin, "1/*f* noise in MOSFETs," in *Instabilities in silicon devices: silicon passivation and related instabilities*, vol. 2 (Elsevier, Amsterdam, 1989), pp. 639–657.

40. S. Christensson, I. Lundstorm, and C. Svensson, "Low frequency noise in MOS transistors," *Solid-State Electronics*, vol. 11, pp. 797–812, 1968.
41. F. Berz, "Theory of low frequency noise in Si MOSTs," *Solid- State Electronics*, vol. 13, pp. 631–647, 1970.
42. S. T. Hsu, "Surface state related 1/*f* noise in MOS transistors," *Solid- State Electronics*, vol. 13, pp. 1451–1459, 1970.
43. A. van der Ziel, "Flicker noise in electronic devices," in *Advances in electronics and electron physics,* vol. 49, edited by Martin and Martin (Academic Press, New York, 1979), pp. 225–297.
44. G. Blasquez and A. Boukabache, "Origins of 1/*f* noise in MOS transistors," in *Noise in physical systems and 1/f noise*, edited by H. Savelli, G. Lecoy, and J. P. Nougier (Elsevier, Amsterdam, 1983), pp. 303–306.
45. Z. Celik and T. Y. Hsiang, "Study of 1/f noise in n-MOSFETs: linear regime," *IEEE Trans. Electron. Dev.,* vol. 32, pp. 2797–2801, 1985.
46. O. Jantsch and B. Borchert, "Determination of interface state density especially at the band edges by noise measurements on MOSFETs," *Solid-State Electronics*, vol. 30, pp. 1013–1015, 1987.
47. K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A unified model for the flicker noise in metal-oxide semiconductor field-effect transistors," *IEEE Trans. Electron. Dev.,* vol. 37, pp. 654–665, 1990.
48. K. K. Hung, P. K. Ko, C. Hu, and Y. C. Cheng, "A physical-based MOSFET noise model for circuit simulators," *IEEE Trans. Electron. Dev.,* vol. 37, pp. 1323–1333, 1990.
49. M. J. Kirton and M. J. Uren, "Noise in solid-state microstructures – a new perspective on individual defects, interface states and low-frequency (1/*f*) noise," *Adv. Phys.*, vol. 38, pp. 367–468, 1989.
50. J. H. Scofield and D. M. Fleetwood, "Physical basis for nondestructive tests of MOS radiation hardness," *IEEE Trans. Nucl. Sci.,* vol. 38, no. 6, pp. 1567–1577, Dec. 1991.
51. X. J. Zhou, L. Tsetseris, S. N. Rashkeev, D. M. Fleetwood, R. D. Schrimpf, S. T. Pantelides, et al., "Negative bias-temperature instabilities in metal-oxide-silicon devices with SiO_2 and SiO_xN_y/HfO_2 gate dielectrics," *Appl. Phys. Lett.,* vol. 84, no. 22, pp. 4394–4396, May 2004.
52. H. E. Boesch, Jr., "Interface-state generation in thick SiO_2 layers," *IEEE Trans. Nucl. Sci.*, vol. 29, no. 6, pp. 1446–1451, Dec. 1982.
53. H. E. Boesch, Jr., "Charge and interface state generation in field oxides," *IEEE Trans. Nucl. Sci.*, vol. 31, no. 6, pp. 1273–1279, Dec. 1984.
54. R. L. Pease, D. Emily, and H. E. Boesch, Jr., "Total dose induced hole trapping and interface state generation in bipolar recessed field oxides," *IEEE Trans. Nucl. Sci.*, vol. 32, no. 6, pp. 3946–3952, Dec. 1985.
55. I. G. Batyrev, B. Tuttle, D. M. Fleetwood, R. D. Schrimpf, L. Tsetseris, and S. T. Pantelides, "Reactions of water molecules in silica-based network glasses," *Phys. Rev. Lett.*, vol. 100, article no. 105503, Mar. 2008.
56. M. P. Rodgers, D. M. Fleetwood, R. D. Schrimpf, I. G. Batyrev, S. Wang, and S. T. Pantelides, "The effects of aging on MOS irradiation and annealing response," *IEEE Trans. Nucl. Sci.*, vol. 52, no. 6, pp. 2642–2648, Dec. 2005.
57. D. M. Fleetwood, M. P. Rodgers, L. Tsetseris, X. J. Zhou, I. Batyrev, S. Wang, et al., "Effects of device aging on microelectronics radiation response and reliability," *Microelectron. Reliab.*, vol. 47, no. 7, pp. 1075–1085, Jul. 2007.
58. J. R. Schwank, M. R. Shaneyfelt, A. Dasgupta, S. A. Francis, X. J. Zhou, D. M. Fleetwood, et al., "Effects of moisture and hydrogen exposure on radiation-induced MOS device degradation and its implications for long-term aging," *IEEE Trans. Nucl. Sci.,* vol. 55, no. 6, pp. 3206–3215, Dec. 2008.

59. D. M. Fleetwood, S. A. Francis, A. Dasgupta, X. J. Zhou, R. D. Schrimpf, M. R. Shaneyfelt, et al., “Moisture effects on the 1/*f* noise of MOS devices,” *Transactions of the 215th ECS Meeting, Vol. 19(2), Silicon Nitride, Silicon Dioxide, and Emerging Dielectrics 10*, edited by R. Ekwal Sah, J. Zhang, J. Deen, J. Yota, and A. Toriumi (San Francisco, CA, May 24–29, 2009), pp. 363–377.
60. X. J. Zhou, D. M. Fleetwood, L. Tsetseris, R. D. Schrimpf, and S. T. Pantelides, “Effects of switched-bias annealing on charge trapping in HfO_2 gate dielectrics,” *IEEE Trans. Nucl. Sci.*, vol. 53, no. 6, pp. 3636–3643, Dec. 2006.
61. E. P. Gusev, E. Cartier, D. A. Buchanan, M. Gribelyuk, M. Copel, H. Okorn-Schmidt, et al., “Ultrathin high-k metal oxides on silicon: processing, characterization and integration issues,” *Microelectron. Engrg.*, vol. 59, no. 1–4, pp. 341–349, 2001.
62. A. J. Lelis, H. E. Boesch, Jr., T. R. Oldham, and F. B. McLean, “Reversibility of trapped hole annealing,” *IEEE Trans. Nucl. Sci.*, vol. 35, no. 6, pp. 1186–1191, Dec. 1988.
63. A. J. Lelis, T. R. Oldham, H. E. Boesch, Jr., and F. B. McLean, “The nature of the trapped hole annealing process,” *IEEE Trans. Nucl. Sci.,* vol. 36, no. 6, pp. 1808–1815, Dec. 1989.
64. D. M. Fleetwood, M. R. Shaneyfelt, L. C. Riewe, P. S. Winokur, and R. A. Reber, Jr., “The role of border traps in MOS high-temperature postirradiation annealing response,” *IEEE Trans. Nucl. Sci.*, vol. 40, no. 6, pp. 1323–1334, Dec. 1993.
65. R. K. Freitag, D. B. Brown, and C. M. Dozier, “Experimental evidence of two species of radiation-induced trapped positive charge,” *IEEE Trans. Nucl. Sci.*, vol. 40, no. 6, pp. 1316–1322, Dec. 1993.
66. A. J. Lelis and T. R. Oldham, “Time dependence of switching oxide traps,” *IEEE Trans. Nucl. Sci.,* vol. 41, no. 6, pp. 1835–1843, Dec. 1994.
67. J. F. Conley, P. M. Lenahan, A. J. Lelis, and T. R. Oldham, “Electron spin resonance evidence that E’(gamma) centers can behave as switching oxide traps,” *IEEE Trans. Nucl. Sci.*, vol. 42, no. 6, pp. 1744–1749, Dec. 1995.
68. X. J. Zhou, D. M. Fleetwood, J. A. Felix, E. P. Gusev, and C. D’Emic, “Bias-temperature instabilities and radiation effects in MOS devices,” *IEEE Trans. Nucl. Sci.,* vol. 52, no. 6, pp. 2231–2238, Dec. 2005.
69. P. S. Winokur, H. E. Boesch, Jr., J. M. McGarrity, and F. B. McLean, “Two-stage process for buildup of radiation-induced interface states,” *J. Appl. Phys.*, vol. 50, pp. 3492–3494, 1979.
70. D. B. Brown and N. S. Saks, “Time dependence of radiation-induced interface-trap formation in MOS devices as a function of oxide thickness and applied field,” *J. Appl. Phys.*, vol. 70, pp. 3734–3747, 1991.
71. D. M. Fleetwood, W. L. Warren, J. R. Schwank, P. S. Winokur, M. R. Shaneyfelt, and L. C. Riewe, “Effects of interface traps and border traps on MOS postirradiation annealing response,” *IEEE Trans. Nucl. Sci.*, vol. 42, no. 6, pp. 1698–1707, Dec. 1995.
72. L. Tsetseris, X. J. Zhou, D. M. Fleetwood, R. D. Schrimpf, and S. T. Pantelides, “Physical mechanisms of negative-bias temperature instability,” *Appl. Phys. Lett.*, vol. 86, article no. 142103, 2005.
73. S. N. Rashkeev, D. M. Fleetwood, R. D. Schrimpf, and S. T. Pantelides, “Defect generation by hydrogen at the Si-SiO2 interface,” *Phys. Rev. Lett.*, vol. 87, article no. 16, no. 165506, Oct. 2001.
74. S. N. Rashkeev, D. M. Fleetwood, R. D. Schrimpf, and S. T. Pantelides, “Proton-induced defect generation at the Si-SiO_2 interface,” *IEEE Trans. Nucl. Sci.*, vol. 48, no. 6, pp. 2086–2092, Dec. 2001.
75. A. G. Marinopoulos, I. Batyrev, X. J. Zhou, R. D. Schrimpf, D. M. Fleetwood, and S. T. Pantelides, “Hydrogen shuttling near Hf-defect complexes in Si/SiO_2/HfO_2 structures,” *Appl. Phys. Lett.*, vol. 91, article no. 233503, 2007.

76. S. N. Rashkeev, K. van Benthem, S. T. Pantelides, and S. J. Pennycook, "Single Hf atoms inside the ultrathin SiO_2 interlayer between a HfO_2 dielectric film and the Si substrate: how do they modify the interface?" *Microelectron. Engrg.*, vol. 80, pp. 416–419, 2005.
77. K. van Benthem, A. R. Lupini, M. Kim, H. S. Baik, S. J. Doh, J. H. Lee, et al., "Three-dimensional imaging of individual hafnium atoms inside a semiconductor device," *Appl. Phys. Lett.*, vol. 87, article no. 034104, 2005.
78. M. H. Evans, M. Caussanel, R. D. Schrimpf, and S. T. Pantelides, "First-principles modeling of double-gate UTSOI MOSFETs," *IEEE IEDM Tech. Digest*, pp. 597–600, Dec. 2005.
79. C. Driemeier, E. P. Gusev, and I. J. R. Baumvol, "Room temperature interactions of water vapor with HfO_2 films on Si," *Appl. Phys. Lett.*, vol. 88, article no. 201901, 2006.
80. J. A. Felix, M. R. Shaneyfelt, D. M. Fleetwood, T. L. Meisenheimer, J. R. Schwank, R. D. Schrimpf, et al., "Radiation-induced charge trapping in thin $Al_2O_3/SiO_xN_y/Si(100)$ gate dielectric stacks," *IEEE Trans. Nucl. Sci.*, vol. 50, no. 6, pp. 1910–1918, Dec. 2003.
81. S. K. Dixit, X. J. Zhou, R. D. Schrimpf, D. M. Fleetwood, S. T. Pantelides, R. Choi, et al., "Radiation induced charge trapping in ultrathin HfO_2-based MOSFETs," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 6, pp. 1883–1890, Dec. 2007.
82. D. K. Chen, F. E. Mamouni, X. J. Zhou, R. D. Schrimpf, D. M. Fleetwood, K. F. Galloway, et al., "Total dose and bias temperature stress effects for HfSiON on Si MOS capacitors," *IEEE Trans. Nucl. Sci.*, vol. 54, no. 6, pp. 1931–1937, Dec. 2007.
83. H. Park, S. K. Dixit, Y. S. Choi, R. D. Schrimpf, D. M. Fleetwood, T. Nishida, et al., "Total ionizing dose effects on strained HfO_2-based MOSFETs," *IEEE Trans. Nucl. Sci.*, vol. 55, no. 6, pp. 2981–2985, Dec. 2008.

4 Novel Total Dose and Heavy-Ion Charge Collection Phenomena in a New SiGe HBT on Thin-Film SOI Technology

Grégory Avenier, Marco Bellini, Alain Chantre, Peng Cheng, Pascal Chevalier, John D. Cressler, Ryan M. Diestelhorst, Paul W. Marshall, Stanley D. Phillips, and Marek Turowski

CONTENTS

4.1 INTRODUCTION

Silicon-germanium heterojunction bipolar transistor (SiGe HBT) technology has recently achieved significant success in analog and mixed signals and in radio-frequency (RF) through mm-wave integrated circuits (ICs) because of its excellent frequency response, low noise, high gain, and capability to support high levels of integration [1,2]. Meanwhile, silicon-on-insulator (SOI) technology has increasingly received commercial attention because it exhibits improved device isolation and cross talk and reduces parasitics and leakage [3]. Removing the substrate junction results in lower capacitances and in elimination of substrate leakage, which facilitates high-temperature operation and provides immunity to latchup [3,4]. The smaller parasitic capacitances, the absence of leakage, the reduced power consumption, and the increased current drive in metal-oxide semiconductor (MOS) transistors are particularly attractive to the complementary metal-oxide semiconductor (CMOS) digital logic market: according to [5], SOI wafers accounted for more than one third of the total revenues of the 300 mm wafer logic market in 2007.

Considering the increasing commercial interest in SOI CMOS and the large popularity of BiCMOS platforms, it becomes natural to investigate the feasibility of BiCMOS-on-SOI devices that combine the advantages of both technologies [6,7]. As also mentioned in [8], SOI is a possible logical next step in the evolution of the bipolar device, after the optimization of the emitter with polysilicon and of the base with SiGe [8]. From these perspectives the recent demonstrations of SiGe HBTs fabricated on CMOS-compatible SOI substrates [9-11] appear to be an attractive path for future SiGe BiCMOS scaling.

Importantly, in the radiation context, SiGe HBT-on-SOI potentially offers significant built-in radiation hardness from both a total ionizing dose (TID) and a single-event upset (SEU) perspective.

In fact, silicon-on-insulator technology was developed to reduce vulnerability to single-event effects (SEEs) [12]. As is widely known, a heavy-ion strike on a semiconductor generates a very large number of electron-hole pairs [13]. The generated carriers separate because of the drift mechanism (supported by the electric field in the semiconductor) and the diffusion mechanism (due to the large concentration of holes and electrons along the path of the strike). As the carriers move toward the terminals of the device, they induce large current pulses that can significantly disturb the behavior of a circuit. For instance, the charge stored in a capacitive node may be altered, leaving a digital circuit in an incorrect logic state. Also, ion strikes induce current pulses of significant amplitude and duration that can modify a sequence of bits in a shift register. These disruptions of circuit functions are called single-event upsets and are a common soft (i.e., recoverable) error. But an ion strike can also generate currents large enough to trigger a single-event latchup (SEL), causing the complete destruction of the device [12].

Although soft errors do not threaten the integrity of devices, they can undermine the reliability of circuits in environments with a high fluence of heavy ions. To increase SEU hardness, a number of techniques are used at layout level (e.g., introducing auxiliary junctions to mitigate charge collection [14,15]) or at the circuit level (e.g., spatial redundancy: triplicating the vulnerable circuit and introducing resistive majority voting [16]). These techniques usually are costly in terms of area, power consumption, and system complexity.

In the context of SEE hardening, SOI devices possess a tremendous advantage over traditional bulk devices because in general the amount of electron-hole pairs generated is directly proportional to the silicon volume of the device. Therefore, SOI substrates enable a dramatic reduction in collected charge because the silicon layer thickness is of the order of hundreds of nm versus hundreds of μm for bulk substrates [12]. Previous studies demonstrate a clear reduction of collected charge in HBTs fabricated on a 1 μm SOI layer compared with bulk devices [14].

Also, bulk SiGe HBTs show considerable built-in total dose radiation tolerance because of vertical and lateral scaling and the high base doping [17]. Therefore, HBT-on-SOI technology is expected to exhibit the same TID hardness of bulk HBTs combined with significant improvements to SEU immunity without any need for additional process hardening, eliminating a well-known weakness of bulk SiGe HBT technologies. Moreover, SiGe HBTs-on-SOI are characterized by the same excellent cryogenic performance demonstrated by bulk SiGe HBTs that results from the germanium-induced bandgap narrowing [18]. Consequently, SiGe HBTs-on-SOI have the unrivaled potential to improve the performance and reliability of orbital electronics, systems for planetary and space missions, cryogenically cooled radiation detectors, and semiconductor-superconductor systems [17].

This chapter investigates the impact of 63 MeV proton and 10 keV X-ray radiation (up to a total dose of 2 Mrad(SiO_2)) on the *ac* and *dc* characteristics of a new high-performance SiGe HBT-on-SOI technology from STMicroelectronics [11]. The charge collection response of the devices is investigated through technology computer-aided design (TCAD) simulations. The results presented are based on studies published in [19,20]. The radiation response of this SiGe HBT-on-SOI is compared with that of a bulk SiGe HBT fabricated with an identical emitter-base (EB) structure. In other words, the devices differ only in terms of substrates (bulk vs. SOI) and collector dopings. Although SOI devices exhibit comparable degradation in the forward mode, their *ac* performance and the breakdown voltage B_{VCEO} increase as a result of the deposition of positive charge in the buried oxide.

In addition, the devices under study feature an innovative C_BE^BC layout (with off-plane base contacts, as shown in the inset of Figure 4.1), introduced to enhance the *ac* performance. Experimental data and calibrated three-dimensional (3-D) TCAD simulations demonstrate that, in the inverse mode, the current flow in proximity of the large oxide surface between the collector and the base can be pushed toward the buried oxide (BOX) by substrate bias V_S, reducing the radiation-induced leakage.

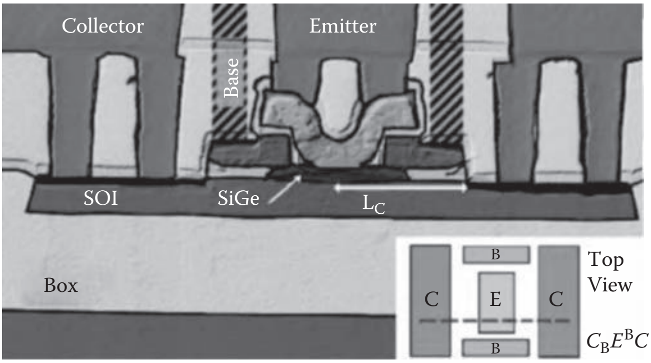


FIGURE 4.1 Cross-sectional transmission electron microscopy micrograph of the SiGe HBT-on-SOI with C_BE^BC layout. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

Also, the thermal resistances, R_{TH}, of bulk and SOI HBTs have been compared before and after irradiation over the temperature range from 300 K to 390 K, demonstrating that exposure to radiation causes the R_{TH} of SOI devices to increase. To conclude, calibrated 3-D TCAD simulations of heavy-ion strikes in the center of the emitter indicate that the adoption of the $C_B E^B C$ layout introduces novel charge collection phenomena.

4.2 DEVICE STRUCTURE AND BASIC OPERATION

A conventional (bulk) SiGe HBT is essentially a vertical device: the current flows from the top of the emitter, through the base and the collector to the highly doped, thick subcollector region, which is fundamentally a low resistivity path between the actual collector terminal on the top of the device and the collector-base (CB) junction. As such, the subcollector plays a very minor role in the *dc* and *ac* performance of the device. In fact, in most TCAD simulations, with obviously the exception of SEU simulations, the subcollector structure is drastically simplified, and the boundary condition corresponding to the electrical collector contact is placed at the bottom of the device mesh without any loss of accuracy.

However, fabricating a SiGe HBT on thin-film SOI layer is an especially challenging task since it is not possible to use the same structure of a bulk SiGe HBT. The 0.1–0.2 μm SOI layer is too thin to accommodate the thick, heavily doped subcollector that is essential for high-speed devices.

Recently, the new “folded” SiGe HBT structure shown in Figure 4.1 has been demonstrated [9-11,21,22]. This device is characterized by emitter and base profiles comparable to those of second-generation bulk SiGe HBTs, but the subcollector is replaced by either a fully or a partially depleted collector (according to the doping N_C).

Interestingly, these radical changes in the device structure introduce physical phenomena not observed in bulk devices: the voltage bias applied to the SOI substrate strongly affects the current flow and electric field, significantly altering the *dc* and *ac* performance. These effects are also very significant from a radiation hardness perspective: the positive charge deposited in the BOX by irradiation is electrically equivalent to increasing the substrate bias and therefore affects device performance and reliability concerns in the same way. In the remainder of this section, these phenomena will be introduced and explained with the aid of TCAD.

Initially, the analysis will focus on devices with conventional top layouts (CBEBC) because they can be completely described with two-dimensional (2-D) simulations that are easier to visualize and understand. Then, the effects of adopting the novel $C_B E^B C$ top layout, investigated in [20] with the assistance of 3-D TCAD simulations, will be briefly explained. The impact of these phenomena on device and circuit behavior will be quantified and discussed in the following section.

The most important phenomenon in SiGe HBTs-on-SOI is the change of the current flow in the collector with substrate bias or irradiation. The collector doping of a fully depleted device is carefully chosen so that the built-in voltage depletes the whole collector area, when the substrate is floating or grounded. As demonstrated by the TCAD simulations in Figure 4.2, the voltage, V_S, applied to the SOI substrate alters the current flow within the collector, affecting f_T, f_{max}, and collector resistance,

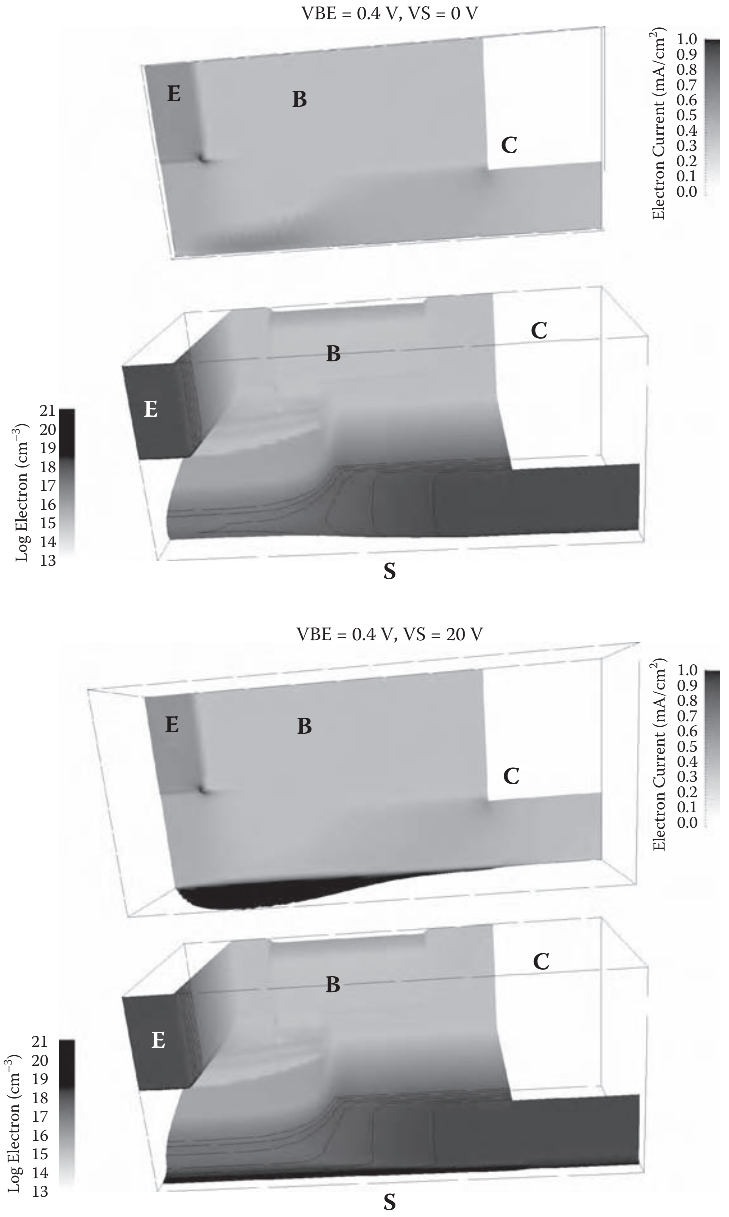


FIGURE 4.2 TCAD simulations of electron current density (top row) and electron density (bottom row) row in a SiGe HBT-on-SOI with V_{BE} = 0.4 V, V_{CB} = 0 V, and V_S biased at 0 V (left column) and 20 V (right column). (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

R_C [9,23]. As shown on the left side of Figure 4.2, when the substrate is grounded the current flows in the center of the SOI layer. Conversely, positive substrate bias, V_S, creates an accumulation layer at the SOI/BOX interface, which acts as a low resistivity path to the top collector contact. This preferential path along the SOI/BOX interface reduces the collector resistance, R_C, and hence the quasi-saturation effect at high currents—the forward biasing of the CB junction caused by the voltage drop on R_C [23].

Substrate bias also has a dramatic impact on the electric field in the collector region, affecting impact ionization in the device and hence $M - 1$ (i.e., the avalanche multiplication factor).

At low V_S, the collector-base voltage completely depletes the collector region under the emitter, exposing the positive fixed charge, N_C^+. No further extension of depletion is possible in the vertical direction; therefore, the electric field in this region is pinned and cannot become larger. As V_{CB} increases, the depletion layer widens in the lateral direction toward the collector contact region, leading to an increase of the electric field in the lateral region.

Conversely, at high V_S the electron accumulation layer forms at the SOI/BOX interface, providing a low resistivity path on the bottom of the device, under the space charge region. Consequently, the voltage drop between the collector electrode and the bottom of the SOI layer becomes negligible, which means that the entire V_{CB} is applied to the vertical region under the emitter. The distribution of the electric field in the collector changes dramatically: the peak field moves from the lateral region to the vertical region under the emitter. The following section discusses how this phenomenon produces significant changes in $M - 1$ and breakdown voltage with radiation or with substrate bias.

Finally, radiation or substrate bias also affects the *ac* performance of SiGe HBTs-on-SOI. One of main physical phenomena limiting the speed of SiGe HBTs (on bulk or SOI substrates) is the heterojunction barrier effect (HBE), which is triggered when the device operates at high currents, in the high injection condition [24-26].

When the current flowing in an Si bipolar junction transistor (BJT) or in an SiGe HBT results in amount of carriers in the collector comparable to N_C, high injection effects such as the Webster-Rittner or the Kirk effect are triggered [1], limiting the device performance. For example, as the collector current density approaches the Kirk current density, J_{KIRK}, the mobile carriers in the CB space charge region compensate the fixed charge and collapse the electric field [27]. This causes a base pushout in the Si BJT and a decrease in *ac* performance, but the impact on SiGe HBTs is even more dramatic.

Most importantly in SiGe HBTs, the germanium grading from the SiGe base to the Si collector induces a barrier for holes in the valence band. During normal operation, the barrier is hidden by the reverse voltage applied to the CB junction and has no impact on device performance. However, at high currents the collapse of the CB-junction space charge region exposes the valence-band barrier. The holes pile up at this location, inducing a conduction-band barrier for the electrons and thereby reducing collector current, I_C, and gain, b [1]. Since the HBE is triggered at J_{KIRK}, which is proportional to the collector doping, N_C, HBTs-on-SOI are particularly vulnerable because of the low doping depleted collector design. Substrate bias (and also the positive charge in the BOX due to radiation) retards the HBE, consequently improving the *ac* performance. At high injection an equal number of holes and electrons flood the CB junction, reaching concentrations higher than N_C. However, increasing the substrate voltage enhances the vertical electric field sweeping carriers away from the CB junction and reducing HBE. As demonstrated in the following section, this effect significantly enhances f_T and f_{max} with positive substrate bias or with exposure to radiation.

All these physical phenomena characterize every SiGe HBT fabricated on thin-film SOI, regardless of the top geometry. This work, however, focuses on HBTs-on-SOI with an innovative layout. As mentioned before, adapting the vertical structure of SiGe HBTs to an SOI substrate comes at the expense of *ac* performance. In fact, the emitter–collector distance limits the *ac* performance of SiGe HBTs-on-SOI because of the length of the drift path in the case of the fully depleted device or because of the $R \times C_{CJC}$ delay time in the case of the partially depleted devices [9,10]. Both these factors are reduced, minimizing the distance, L_C, between the emitter and collector contact. The novel C_BE^BC layout proposed in [10,28] (in contrast to the more conventional CBEBC layout employed in [9]) places the base contact out of the plane defined by the emitter and the collector, thereby minimizing L_C (as shown in the inset of Figure 4.1) and increasing *ac* performance.

These HBTs have been developed by STMicroelectronics with the addition of only four-mask layers on top of a 130 nm SOI CMOS process and feature a 150 nm SOI layer on top of a 400 nm SiO_2 BOX [20]. The reduction of L_C to 0.4 μm results in the figures of merit shown in Table 4.1 [10,29].

The optimized layout, however, significantly alters the current flow inside the device. As mentioned before, the current flow in a bulk SiGe HBT is essentially one-dimensional (1-D), vertical under the emitter, while it is 2-D in a SiGe HBT-on-SOI with a conventional CBEBC layout, initially vertical under the emitter and then horizontal along the SOI/BOX interface [9]. But the current flow in a SiGe HBT-on-SOI with C_BE^BC layout is intrinsically 3-D in nature. At V_S = 0 V most of the

TABLE 4.1
Figures of Merit of SiGe HBTs on Thin-Film SOI with C_BE^BC Layout

Figure of Merit	(300 K)
b	390
f_T (V_{CB} = 0.5 V)	35 GHz
f_{max} (V_{CB} = 0.5 V)	134 GHz
B_{VCEO}	5.4 V
B_{VCBO}	15 V

Source: After Avenier, G., Schwartzmann, T., Chevalier, P., Vandelle, B., Rubaldo, L., Dutartre, D., et al., Proceedings of the Bi-Polar/Bi CMOS Circuits and Technology Meeting, pp. 128-131, 2005.

current flow in the vertical direction occurs in the center of the SOI layer, as in the CBEBC device. Conversely, the current flow in the horizontal plane (normal to the vertical direction) is confined to a narrow region between the emitter and the collector contact. However, at $V_S = 20$ V the formation of the accumulation layer results in a downward shift in the vertical direction of the current flow, closer to the SOI/BOX interface, as is demonstrated in Figure 4.2 for the CBEBC device. Interestingly, the increased vertical electric field results also in a much larger spread of the current on the horizontal plane toward the base contact [20].

4.3 IRRADIATION

The radiation response of both bulk SiGe HBTs and fully depleted SiGe HBTs-on-SOI was assessed, exposing samples to 63.3 MeV protons and 10 keV X-rays up to a total dose of 2 Mrad(SiO_2). The devices (of effective emitter areas, A_E, of $12 \times (0.17 \times 0.5)$ μm^2 and $7 \times (0.17 \times 0.85)$ μm^2) were irradiated in delidded packages with grounded pins and immediately measured in situ. Passive exposure (terminals floating) of *ac* test structures for both kinds of devices to a total dose of 4.2 Mrad(SiO_2) was used to quantify the impact of radiation on the *ac* performance.

The bulk and an SOI SiGe HBT show very similar forward Gummel plots because they share the same EB structure. However, at $V_{BE} = 1$ V the characteristics of the SOI device reveal a slight decrease of I_C and increase of I_B that possibly result from a stronger quasi-saturation effect (due to the lower N_C). As expected, the proton-induced degradation of the forward mode base current I_B with increasing proton dose is similar for the SOI and bulk devices.

Figure 4.3 compares the normalized excess base current in forward and inverse mode for proton and X-ray irradiation at cumulative dose steps of 100, 300, 600, 1,000, and 2,000 krad(SiO_2). The radiation-induced degradation in the forward mode for both devices is similar and is expected because the transistors have identical emitter-base structures.

The leakage, $\Delta I_B/I_{B0}$, measured in the inverse mode of the SOI device (measured with emitter and collector swapped) is much larger compared with the forward mode. This is explained by the differences between the composition of the EB oxide and the pedestal oxide (used to separate the collector and the base), by the different emitter and collector doping, and by the different geometrical dimensions of the Si/SiO_2 interfaces in both forward and inverse mode. Moreover, Figure 4.4 demonstrates that it is possible to reduce the base leakage during inverse mode operation applying a positive substrate voltage, V_S, to the SOI device after irradiation. The NanoTCAD 3-D TCAD simulation package (previously used to investigate the radiation effects on other advanced devices [15,30]) has been used to provide calibrated analyses of both the SOI and the bulk device in forward and inverse mode.

First a model of the device before radiation was calibrated to the *dc* and *ac* characteristics using measured secondary ion mass spectroscopy (SIMS) profiles and then a trap concentration of roughly 10^{10} cm^{-2} (as suggested in [31]) was introduced at the pedestal oxide/SOI interface to reproduce the nonideal base current in the inverse mode. The resulting simulations correctly capture the impact of V_S on the inverse Gummel. TCAD simulations suggest that the current is pushed away from

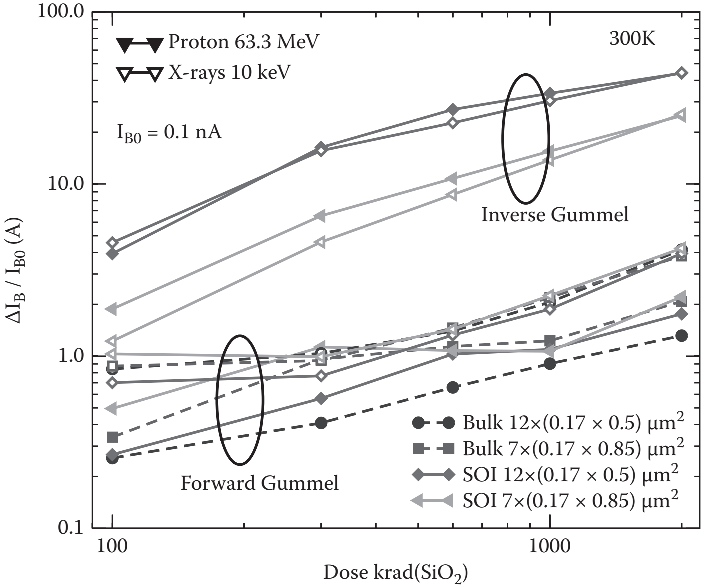


FIGURE 4.3 Excess normalized base current, $\Delta I_B/I_{B0}$, versus total radiation dose in krad(SiO_2), in forward and inverse mode for devices irradiated with 63 MeV protons or 10 keV X-rays. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

the Si/SiO_2 interface by the applied electric field, leading to a decrease of generation/recombination (G/R)-induced leakage, as shown in Figure 4.5.

Figure 4.6 shows the multiplication factor $M - 1$ before and after irradiation for a bulk HBT and an HBT-on-SOI. As explained in Section 4.2, in SOI devices the electric field peak shifts from the region under the emitter to the lateral path because of the voltage perturbation introduced by substrate bias or positive charge in the BOX. Before irradiation and at $V_S = 0$ V, the electric field under the emitter is pinned; consequently, $M - 1$ saturates, as shown on the right side of Figure 4.6. When the V_{CB} surpasses 4 V, the voltage drop on the lateral path significantly enhances the field, causing an increase of $M - 1$. $M - 1$ of a bulk device is much larger and exhibits the characteristic "arc" shape because the peak of the electric field is under the emitter, at the CB junction. When substrate bias increases in the SOI device, the peak field also shifts under the emitter, approaching a configuration similar to the bulk transistor, causing $M - 1$ to increase and to assume a more rounded shape. As clearly demonstrated in Figure 4.6, radiation causes no change in bulk devices but significantly reduces $M - 1$ for SOI devices, especially at high V_S and high V_{CB}. This is again caused by a sheet of positive charge deposited at the SOI/BOX, which contributes to increase the vertical field in Si but reduces the field in the BOX at high V_S, shielding part of the applied substrate bias [32,33]. This mechanism also explains the slight increase of $M - 1$ at $V_S = 0$ V and $V_{CB} \geq 4$ V [32].

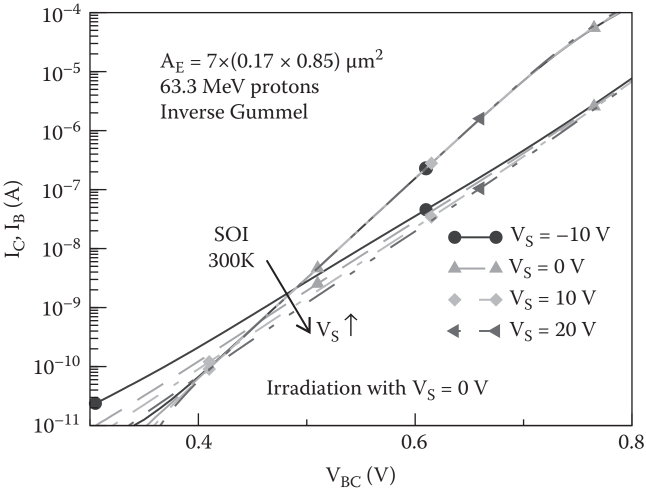


FIGURE 4.4 Inverse Gummel of a SiGe HBT-on-SOI after a proton dose of 2 Mrad(SiO_2) as a function of V_S during post-irradiation measurements. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

The impact of radiation on the breakdown voltage, BV_{CEO}, is shown in Figure 4.6 by the crosses indicating the base current reversal points. Breakdown voltage increases slightly in the bulk device even if $M - 1$ does not vary because of the decrease of current gain, *b*. For the SOI device, the combination of decrease of $M - 1$ and of *b* explains the noticeable increase in breakdown voltage with irradiation. In general, a modest increase of breakdown voltage with irradiation is not a concern for circuit operation, whereas a degradation would be much less desirable. However, any change of device performance should be carefully quantified and modeled, especially at high doses.

Since the performance of SiGe HBTs is strongly affected by temperature, a partially depleted device is measured before and after irradiation at a proton dose of 4.2 Mrad(SiO_2) in the range of temperatures between 30 K and 300 K, as shown in Figure 4.7. As expected from the presence of germanium in the base, the preirradiation current gain increases significantly as temperature decreases: from 250 at 300 K to more than 1,500 at 77 K. Importantly, even in this wide temperature range and after the large 4.2 Mrad(SiO_2) dose the gain degradation is less than 10% at peak current and is negligible at currents employed in most IC applications. The ideality factor, *n,* of the excess base current, ΔI_B, increases significantly (from about 2 at room temperature to more than 40 at 30 K), implying that a trap-assisted tunneling mechanism is dominant at low temperatures [34].

Interestingly, the inverse Gummel characteristics show a large amount of leakage both at high and low temperatures. However, the ideality factor, *n,* of the excess base

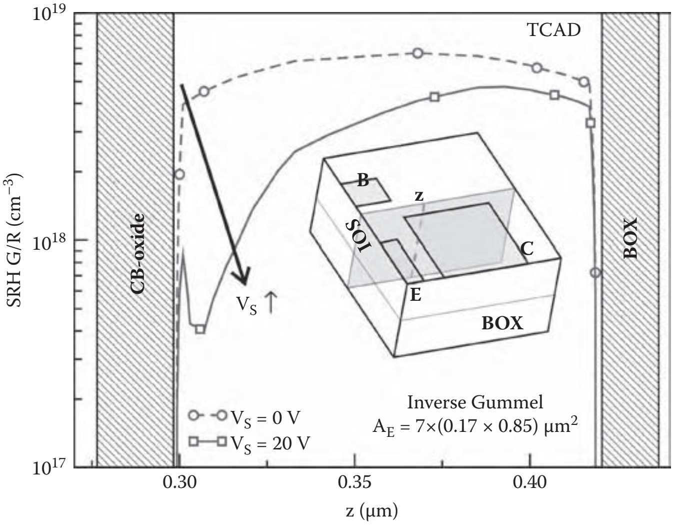


FIGURE 4.5 One-dimensional (1-D) cut of a Shockley-Read-Hall (SRH) generation-recombination rate for $V_S = 0$ V and $V_S = 20$ V in the region between the base and the collector, as indicated in the inset. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

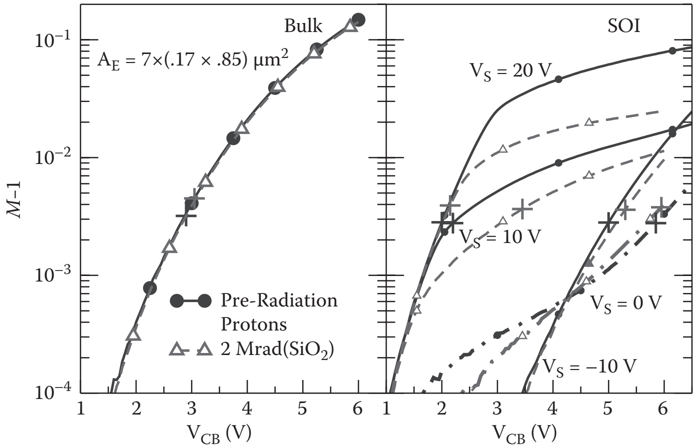


FIGURE 4.6 $M - 1$ for a bulk HBT (left) and a partially depleted SiGe HBT-on-SOI (right) before and after a proton dose of 4.2 Mrad(SiO$_2$). The crosses indicate the base current reversal point. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

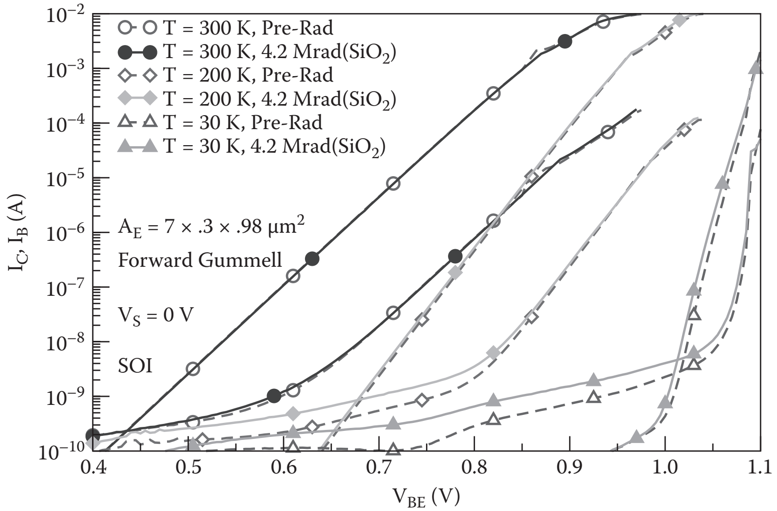


FIGURE 4.7 Forward Gummel characteristics of a partially depleted SiGe HBT-on-SOI as a function of temperature before and after a proton dose of 4.2 Mrad(SiO$_2$). (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

for inverse mode base operation increases much less than as forward mode operation, from 1.4 at 300 K to 6 at 30 K. The ideality factor was also extracted from the I_C currents before and after irradiation to validate the accuracy of the temperature sensor of the cyrosystem. The values of n agree within 1% and slightly increase from 1.03 at 300 K to 2.3 at 30 K, possibly due to nonequilibrium transport phenomena, as reported in the literature [1].

As far as *ac* performance is concerned, the bulk devices show no change in f_T and f_{max}. Conversely, Figure 4.8 shows a reproducible enhancement of the *ac* performance of the fully depleted SiGe HBT-on-SOI after irradiation, in agreement with previous findings [23,32]. Radiation creates positive charge at the SOI/BOX interface, delaying the onset of the Kirk effect and thereby increasing f_T and f_{max} [23]. This is electrically equivalent to applying a higher substrate voltage, V_S, as shown in Figure 4.8.

Figure 4.9 compares the C_{BC} capacitance of the bulk and SOI devices with A_E = 5 × (0.17 × 1.2) μm^2 and L_C = 0.72 μm. The observed hump in the C_{BC} characteristic for the SOI device with V_S = 0 V has been reported in [35] and explained by the combined expansion of the space charge region in both the vertical and the horizontal directions. Interestingly, the application of substrate bias V_S affects the direction of expansion of the depletion region. As shown in Figure 4.9, positive V_S results in a predominance of the vertical component (as in the bulk HBTs), making the C_{BC} of the SOI device similar to that of a bulk device. The capacitance, C_{BC}, of the bulk HBT after irradiation shows a negligible change, consistent with the observed small

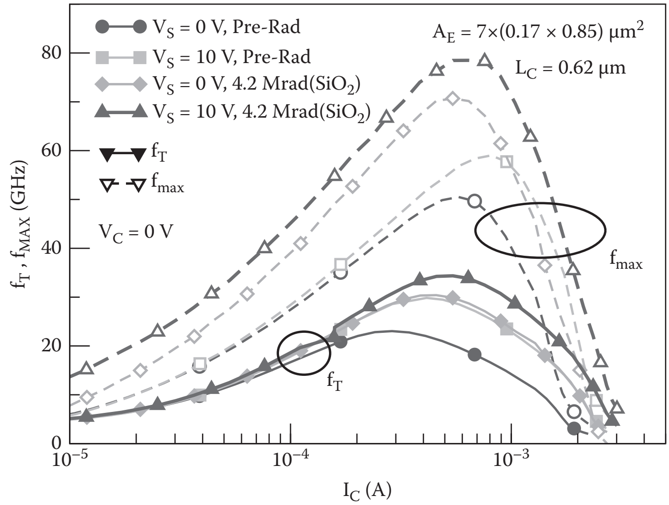


FIGURE 4.8 f_T and f_{max} for fully depleted SiGe HBT-on-SOI, before and after a 4.2 Mrad(SiO$_2$) dose, at different V_{CB} and V_S. The emitter area A_E is 7 × (0.17 × 0.85) µm^2. L_C is 0.62 µm. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

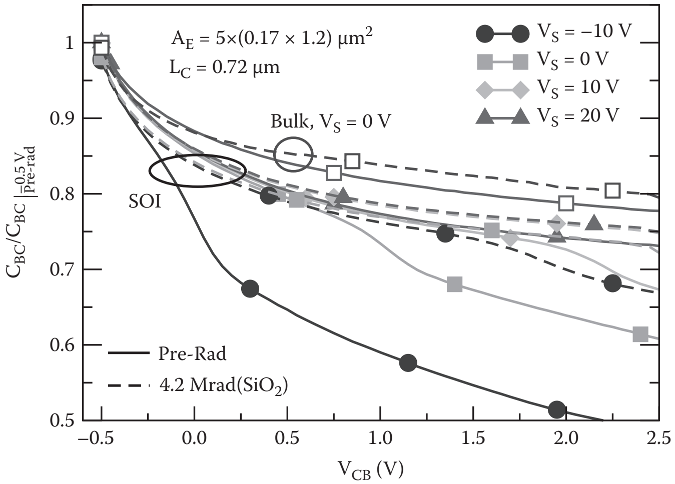


FIGURE 4.9 Normalized C_{BC} versus V_{CB} for both a bulk device (V_S = 0 V) and an HBT-on-SOI (V_S ranging from –10 V to 20 V, in 10 V steps), before and after a 4.2 Mrad(SiO$_2$) dose. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

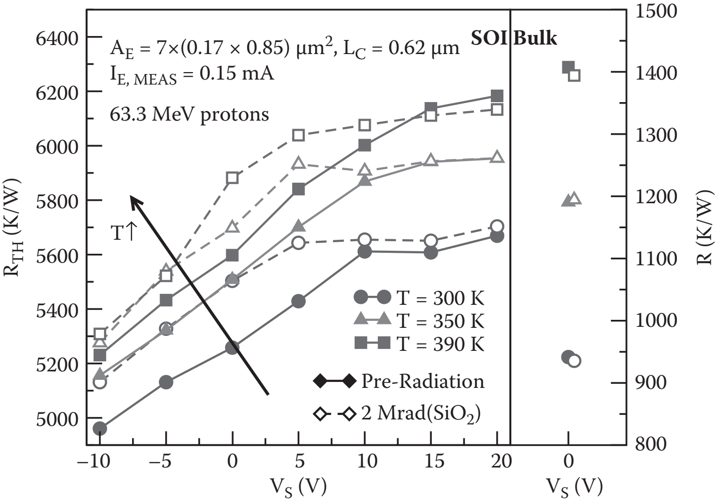


FIGURE 4.10 Thermal resistance, R_{TH}, of HBTs fabricated on bulk and SOI substrates as a function of V_S, before and after a 2 Mrad(SiO_2) dose, at temperatures of 300 K, 350 K, and 390 K. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

change of f_{max}. Conversely, proton irradiation of the SOI device leads to a dominance of the vertical component of the electric field in the depleted collector.

The impact of irradiation on the thermal resistance, R_{TH}, of both bulk and SOI HBTs is also examined using the technique described in [36]. Figure 4.10 shows that, while bulk devices exhibit negligible change in thermal resistance, radiation in SOI devices has the same impact on R_{TH} as an increase in V_S. TCAD simulations have been used to compare the power density distributions in the SiGe HBT-on-SOI biased at V_{BE} = 0.7 V and V_{CB} = 2 V for substrate voltages of 0 V and 20 V. Figure 4.11 shows 1-D cuts of the power density P along the line z under the emitter for V_S = 0 V and 20 V.

Since the thermal conductivity of SiO_2 is lower than for Si, the heat generated in the transistor flows mainly through the Si layer and through the top contacts rather than through the SiO_2 BOX, as reported in [37]. Therefore, at V_S = 20 V the additional power dissipated at the SOI/BOX interface, as shown in Figure 4.11, will flow through the whole SOI layer, resulting in a noticeable increase of the thermal resistance. Since large radiation doses result in larger thermal resistance, the potential increase of self-heating can be a reliability concern for devices operating at large collector-base voltages in radiation environments.

In conclusion, the impact of 63.3 MeV protons on SiGe HBTs on both SOI and bulk substrates fabricated with identical emitter-base structures is assessed by comparing the *dc* and *ac* performance and the thermal resistance. Although SOI devices exhibit larger degradation in the inverse mode than in the forward mode, the excess

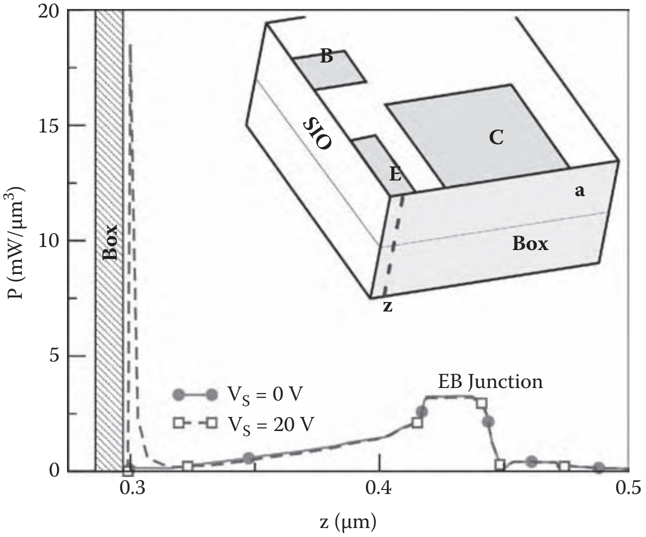


FIGURE 4.11 1-D plots of power density in a SiGe HBT on SOI for $V_S = 0$ V and $V_S = 20$ V, along the line *z* indicated in the inset. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

leakage can be reduced by increasing the substrate bias. Radiation also alters the current flow in the device, increasing R_{TH}. This constitutes a possible reliability concern for devices operating at large collector-base voltages.

4.4 SIMULATION STUDY OF SINGLE-EVENT UPSET RESPONSE

Although bulk SiGe HBTs exhibit considerable total ionizing dose hardness, their vulnerability to single-event upset is considered to be the Achilles' heel of the technology [1]. In fact, because of the bulk Si substrate, vertical devices can be affected by SEU even when the strike occurs outside the deep trenches due to diffusion of carriers [15]. HBTs-on-SOI are obviously immune from this problem because the silicon active volume is completely surrounded by oxide and no diffusion of charge from strikes outside the active area can happen.

However, the introduction of advanced layouts could potentially trigger effects of significance for circuit operation, such as strong dependence of the SEU response on the location of the strike. Also, especially in thin-film devices, the doping of the depleted collector and the substrate voltage significantly alters the electric fields in the device and could affect the SEU response.

3-D TCAD simulations used to reproduce SEU transients in bulk devices can be very time-consuming because the mesh needs to be large enough to capture the complete ion strike event without introducing unphysical approximations. Not only should the vertical extension of the mesh be in the order of tens of microns to

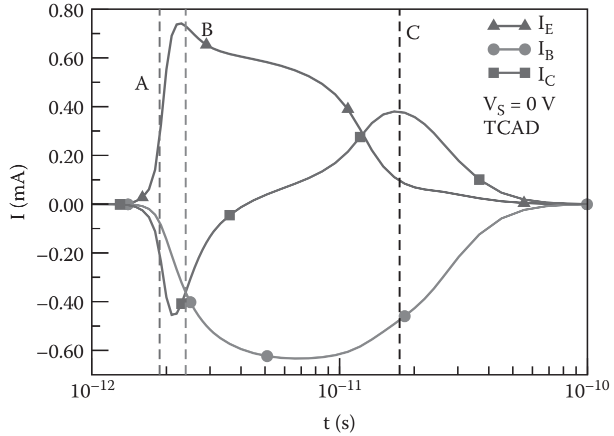


FIGURE 4.12 Collected charge at the terminals for an ion strike in the center of the emitter of a SiGe HBT-on-SOI. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

accommodate the penetration of a heavy ion, but the lateral boundaries should also be far enough that no artificial reflection of charge at the boundaries is experienced. These requirements result in meshes with an extremely large number of elements and consequently simulation times in the order of several days [2]. Luckily, the active region of SOI devices is small and completely surrounded by SiO_2. Therefore, it is possible to accurately describe these transistors with a small number of mesh elements, significantly reducing computational times without any loss in accuracy.

This section describes calibrated 3-D ion strike simulations of the SiGe HBT-on-SOI device with C_BE^BC layout [10] illustrated in Section 4.2. Figure 4.12 shows the simulated SEU currents resulting from an ion strike in the center of the emitter of the device. The total collected charge is less than 0.025 pC, in contrast with about 1 pC for a comparable bulk device, suggesting a significant reduction in vulnerability to SEU [15].

Interestingly, the shape of the current pulses is remarkably different from an ion strike in the center of the emitter of a bulk device with a conventional CBEBC layout (with base contacts placed in-plane between emitter and collector). TCAD simulations indicate that at first I_B is negligible and that I_E and I_C have opposite signs. Also, I_C is characterized by two pulses of opposite polarity. These phenomena can be explained as follows: initially, the negative I_B pulse is caused by excess holes leaving through the base, and the positive I_E pulse is due to electrons leaving through the emitter, as shown by the arrows in Figure 4.13. Then, the change of sign of the I_C pulse is caused by two distinct phenomena occurring at the times marked by *A* and *B* in Figure 4.12.

At time *A*, the ion strike creates a large number of electron-hole pairs, causing the SOI layer to leave equilibrium and resulting in a sharp increase in carrier

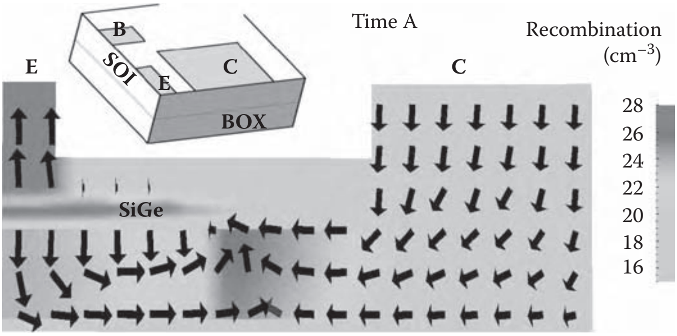


FIGURE 4.13 2-D plots of SRH recombination rate for the ion strike in the center of the emitter at time *A*, as indicated in Figure 4.12. The arrows visualize the electron flow. The inset shows the 2-D cut plane. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *Nuclear Science*, IEEE Transactions on Volume 55 Issue 6, pp. 3197–3201, 2009.)

recombination, as shown in Figure 4.13. The recombination peaks at the extremity of the high doping n-region, which is used to lower the collector resistance, R_C, and which is characterized by the shortest carrier lifetimes. The sudden increase of recombination triggers a large current flow from the collector contact, which results in the negative I_C pulse.

Then at time *B*, the potential modulation induced by the ion strike significantly perturbs the electrostatic potential in the base of the device forward biasing the EB and CB junctions, as shown in Figure 4.14. The positive I_C current component originating from the forward-biased CB junction overcomes the negative component due to recombination and results in a net positive I_C pulse at time *C*. At the same time, the forward biasing of the EB junction decreases the total emitter current I_E, as shown in Figure 4.12. The transistor operates in the saturation region, as shown by the large I_B current supporting both I_C and I_E. Eventually, the strike-induced charge is removed from the device, and the SEU-induced transient pulses end.

The exact shapes and magnitude of strike-induced currents depend on the doping of the region affected, by the proximity of the contacts and by the geometrical layout. Since the C_BE^BC layout creates a significant asymmetry in the device geometry, it is reasonable to expect an increase of the variability of the SEU response with strike position.

This is confirmed in Figure 4.15, which shows the currents generated by an ion strike between the emitter and base, as indicated in the inset. In this case, most of the electrons flow directly to the collector—the closest n-type contact. TCAD simulations suggest that an ion strike in this region is not able to significantly turn on the device, explaining why there is no change of sign in the strike-induced currents.

This analysis proves that studies of the effects of SEU on circuits featuring devices with C_BE^BC layout require accurate 3-D TCAD simulations to correctly model the shape of current pulses resulting from heavy-ion strikes.

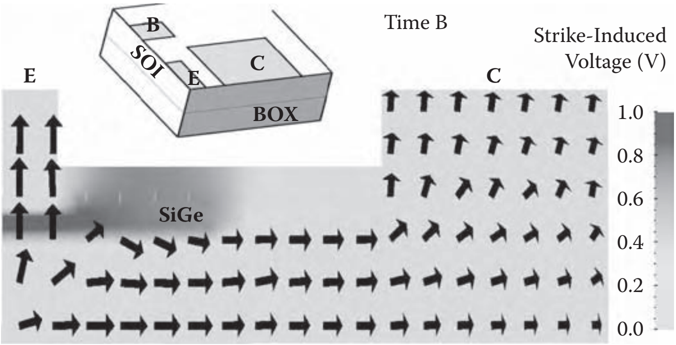


FIGURE 4.14 2-D plots of ion-strike induced electric potential for the strike in the center at time *B*, as indicated in Figure 4.12. The arrows visualize the electron flow. The inset shows the 2-D cut plane. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *IEEE Transactions on Nuclear Science*, Volume 55 Issue 6, pp. 3197–3201, 2009.)

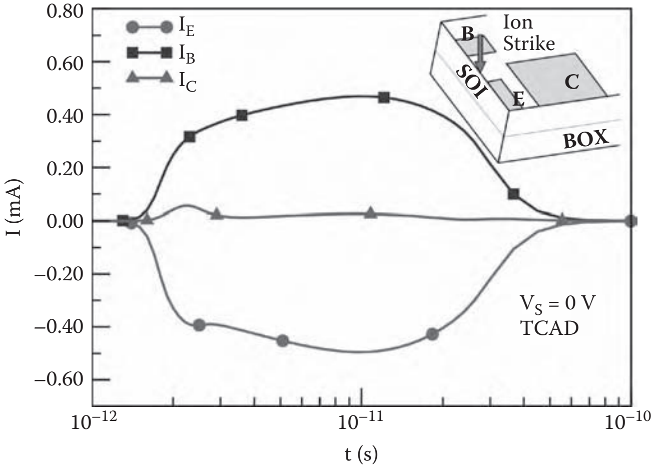


FIGURE 4.15 Collected charge at the terminals for an ion strike between the emitter and the base of a HBT-on-SOI, as shown by the inset. (Reprinted with permission from Bellini, M., Phillips, S. D., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., Turowski, M., Avenier, G., Chantre, A., Chevalier, P. "Novel total dose and heavy-ion charge collection phenomena in a SiGe HBT on thin film SOI technology." *IEEE Transactions on Nuclear Science*, Volume 55 Issue 6, pp. 3197–3201, 2009.)

4.5 CONCLUSIONS

The impact of 63.3 MeV protons and 10 keV X-rays on SiGe HBTs on both SOI and bulk substrates fabricated with identical emitter-base structures is assessed by comparing the *dc* and *ac* performance and the thermal resistance before and after irradiation. Degradation in the forward mode is substantially identical for both types of substrates and for both radiation sources. Although SOI devices exhibit larger degradation in the inverse mode than in the forward mode, the excess leakage can be reduced by increasing the substrate bias. Radiation also lowers the breakdown voltage due to reduction of current gain *b* and to shielding effect on the substrate bias. The positive charge introduced in the BOX also lowers the C_{BC} capacitance and enhances the *ac* performance. Finally, radiation also alters the current flow in the device, increasing R_{TH}. This constitutes a possible reliability concern for devices operating at large collector-base voltages.

To conclude, 3-D TCAD simulations indicate that the novel C_BE^BC layout used in these SiGe-on-SOI devices significantly affects the shape of the current pulses induced by ion strikes, potentially altering their SEU immunity.

REFERENCES

1. Cressler, J. D. and Niu, G., *Silicon-germanium heterojunction bipolar transistors*. Boston: Artech House, 2003.
2. Cressler, J. D., *Silicon heterostructure handbook*. Boca Raton, FL: CRC Press, 2006.
3. Mitrovic, I. Z., Buiu, O., Hall, S., Bagnall, D. M., and Ashburn, P., “Review of SiGe HBTs on SOI,” *Solid-State Electronics*, vol. 49, no. 9, pp. 1556–1567, 2005.
4. Shahidi, G., Ajmera, A., Assaderaghi, F., Bolam, R., Bryant, A., Coffey, M., et al., “Mainstreaming of the SOI technology,” *Proceedings of IEEE SOI Conference*, pp. 1–4, 1999.
5. “SOI enables new generations of lower-power consumption devices,” *SiGe News Review,* vol. 60, p. 5, 2007.
6. Washio, K., Ohue, E., Shimamoto, H., Oda, K., Hayami, R., Kiyota, Y., et al., “A 0.2 μm 180 GHz f_{max} 6.7 ps ECL SOI/HRS self-aligned SEG SiGe HBT/C-MOS technology for microwave and high-speed digital applications,” *IEEE Transactions on Electron Devices*, vol. 49, no. 2, pp. 271–278, 2002.
7. Sato, F., Hashimoto, T., Tezuka, H., Soda, M., Suzaki, T., Tatsumi, T., et al., “A 60-GHz f_T super self-aligned selectively grown SiGe-base (SSSB) bipolar transistor with trench isolation fabricated on SOI substrate and its application to 20-Gb/s optical transmitter ICs,” *IEEE Transactions on Electron Devices*, vol. 46, no. 7, pp. 1332–1338, 1999.
8. Ning, T. H., “Why BiCMOS and SOI BiCMOS?” *IBM Journal of Research and Development*, vol. 46, no. 2, pp. 181–186, 2002.
9. Cai, J., Kumar, M., Steigerwalt, M., Ho, H., Schonenberg, K., Stein, K., et al., “Vertical SiGe-base bipolar transistors on CMOS-compatible SOI substrate,” in *Proceedings of IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 215–218, 2003.
10. Avenier, G., Schwartzmann, T., Chevalier, P., Vandelle, B., Rubaldo, L., Dutartre, D., et al., “A self-aligned vertical HBT for thin SOI SiGeC BiCMOS,” in *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 128–131, 2005.
11. Avenier, G., Fregonese, S., Chevalier, P., Bustos, J., Saguin, F., Schwartzmann, T., et al., “Electrical behavior and technology optimization of Si/SiGeC HBTs on thin-film SOI,” *IEEE Transactions on Electron Devices*, vol. 55, no. 2, pp. 585–593, 2008.

12. Schwank, J. R., “Advantages and limitations of silicon-on-insulator technology in radiation environments,” *Microelectronic Engineering*, vol. 36, no. 1, pp. 335–342, 1997.
13. Sexton, F. W., *1992 IEEE Nuclear Space and Radiation Effects Conference Short Course*. New Orleans, LA: IEEE, 1992.
14. Pellish, J. A., Reed, R. A., Schrimpf, R. D., Alles, M. L., Varadharajaperumal, M., Niu, G., et al., “Substrate engineering concepts to mitigate charge collection in deep trench isolation technologies,” *IEEE Transactions on Nuclear Science*, vol. 53, no. 6, pp. 3298–3305, 2006.
15. Sutton, A. K., Bellini, M., Cressler, J. D., Pellish, J., Reed, R., Marshall, P. W., et al., “An evaluation of transistor-layout RHBD techniques for SEE mitigation in SiGe HBTs,” *IEEE Transactions on Nuclear Science*, vol. 54, no. 6, pp. 2044–2052, 2007.
16. Krithivasan, R., Marshall, P. W., Nayeem, M., Sutton, A. K., Kuo, W., Haugerud, B. M., et al., “Application of RHBD techniques to SEU hardening of third-generation SiGe HBT logic circuits,” *IEEE Transactions on Nuclear Science*, vol. 53, no. 6, pp. 3400–3407, 2006.
17. Cressler, J. D., “On the potential of SiGe HBTs for extreme environment electronics,” *Proceedings of the IEEE*, vol. 93, no. 9, pp. 1559–1582, 2005.
18. Bellini, M., Chen, T., Cressler, J. D., and Cai, J., “Cryogenic operation of SiGe HBTs on CMOS-compatible thin-film SOI substrates,” in *International Workshop on Low Temperature Electronics—WOLTE 2006*, vol. WPP-264 (Noordwijk, The Netherlands), pp. 87–92, 2006.
19. Bellini, M., Jun, S. D. P., Diestelhorst, R. M., Cheng, P., Cressler, J. D., Marshall, P. W., et al., “Novel total dose and heavy-ion charge collection phenomena in a new SiGe HBT on thin-film SOI technology,” *IEEE Transactions on Nuclear Science*, vol. 55, no. 6, pp. 3197–3201, 2008.
20. Bellini, M., Cressler, J. D., Turowski, M., Avenier, G., Chantre, A., and Chevalier, P., “3-D regional transit time analysis of SiGe HBTs on thin-film SOI,” in *Electrochemical Society Symposium*, vol. 16 (Waikiki, HI), pp. 1079–1088, 2008.
21. Cai, J., Ajmera, A., Ouyang, C., Oldiges, P., Steigerwalt, M., Stein, K., et al., “Fully-depleted-collector polysilicon-emitter SiGe-base vertical bipolar transistor on SOI,” *Digest of Technical Papers of the Symposium on VLSI Technology*, pp. 172–173, 2002.
22. Ouyang, Q. C., Cai, J., Ning, T., Oldiges, P., and Johnson, J. B., “A simulation study on thin SOI bipolar transistors with fully or partially depleted collector,” *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 28–31, 2002.
23. Chen, T., Bellini, M., Zhao, E., Comeau, J. P., Sutton, A. K., Grens, C. M., et al., “Substrate bias effects in vertical SiGe HBTs fabricated on CMOS-compatible thin film SOI,” in *Proceedings of IEEE Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 256–259, 2005.
24. Tiwari, S., “A new effect at high currents in heterostructure bipolar transistors,” *Electron Device Letters, IEEE*, vol. 9, no. 3, pp. 142–144, 1988.
25. Mazhari, B. and Morkoc, H., “Effect of collector-base valence-band discontinuity on Kirk effect in double-heterojunction bipolar transistors,” *Applied Physics Letters*, vol. 59, pp. 2162–2164, Oct. 1991.
26. Roenker, K. P. and Mushini, P., “Dynamic formation of a parasitic barrier to electron flow in SiGe HBTs operating at high current densities,” *Microelectronics Journal*, vol. 31, pp. 353–358, May 2000.
27. Kirk, C., “A theory of transistor cutoff frequency (fT) falloff at high current densities,” *IRE Transactions on Electron Devices*, vol. 9, no. 2, pp. 164–174, 1962.
28. Avenier, G., Chevalier, P., Troillard, G., Vandelle, B., Brossard, F., Depoyan, L., et al., “0.13 μm SiGe BiCMOS technology for mm-wave applications,” in *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 89–92, 2008.

29. Avenier, G., Chevalier, P., Vandelle, B., Lenoble, D., Saguin, F., Fregonese, S., et al., “Investigation of fully- and partially-depleted self-aligned SiGeC HBTs on thin film SOI,” in *ESSDERC 2005 European Solid-State Device Research Conference*, pp. 133–136, 2005.
30. *NanoTCAD Software User Manual, Version 2006*. Huntsville, AL: CFD Research Corp., Sept. 2006. Available at: http://www.cfdrc.com
31. Li, Y., Radiation Effects and Temperature Effects of SOI CMOS Technology, Ph.D. dissertation, Auburn University, Department of Electrical and Computer Engineering, June–Aug. 2003.
32. Chen, T., Sutton, A. K., Bellini, M., Haugerud, B. M., Comeau, J. P., Liang, Q., et al., “Proton radiation effects in vertical SiGe HBTs fabricated on CMOS-compatible SOI,” *IEEE Transactions on Nuclear Science*, vol. 52, no. 6, pp. 2353–2357, 2005.
33. Bellini, M., Jun, B., Chen, T., Cressler, J. D., Marshall, P. W., Chen, D., et al., “X-ray irradiation and bias effects in fully-depleted and partially-depleted SiGe HBTs fabricated on CMOS-compatible SOI,” *IEEE Transactions on Nuclear Science*, vol. 53, no. 6, pp. 3182–3186, 2006.
34. Woo, J. C. S., Plummer, J. D., and Stork, J. M. C., “Non-ideal base current in bipolar transistors at low temperatures,” *IEEE Transactions on Electron Devices*, vol. 34, no. 1, pp. 130–138, 1987.
35. Fregonese, S., Avenier, G., Maneux, C., Chantre, A., and Zimmer, T., “A transit time model for thin SOI Si/SiGe HBT,” in *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 184–187, 2005.
36. Vanhoucke, T. and Hurkx, G. A. M., “Simultaneous extraction of the base and thermal resistances of bipolar transistors,” *IEEE Transactions on Electron Devices*, vol. 52, no. 8, pp. 1887–1892, 2005.
37. Palestri, P., Pacelli, A., and Mastrapasqua, M., “Thermal resistance in $Si_{1-x}Ge_x$ HBTs on bulk-Si and SOI substrates,” in *Proceedings of the Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 98–101, 2001.

5 Radiation-Hard Voltage and Current References in Standard CMOS Technologies

Vladimir Gromov and Anne-Johan Annema

CONTENTS

5.1 INTRODUCTION

Particle accelerators are used to get insight into the basic constituents of matter by providing data on particle interaction. This information is gathered from data acquired by the particle detectors inside the particle accelerator setup.

In the Large Hadron Collider (LHC) particle accelerator experiments, very high radiation levels are attained, especially close to the particle interaction point, where many particle detectors are located. For this reason the on-detector readout electronics must be resistant to radiation up to the level of hundreds of Mrad.

Reference voltage generating circuits and current generating circuits with low sensitivity to the temperature variation and the power supply variations are commonly used in analog blocks such as voltage regulators, analog-to-digital (A/D) and digital-to-analog (D/A) converters and hence are used throughout the particle detector readout circuitry. In applications like circuits for the LHC experiments, there is an additional requirement to deliver a stable voltage and current even when operating in ionizing radiation environments.

Historically, radiation-hard application-specific integrated circuits (ASICs) for military and space applications were fabricated in silicon-on-insulator (SOI) or silicon-on-sapphire (SOS) technologies [1,2]. Compared with mainstream silicon technologies, SOI reduces the radiation sensitive volume by isolating the entire device from the bulk substrate with the help of the buried oxide layer. This makes SOI highly resistant to single-event upsets (SEUs). Furthermore, because SOI has no wells in the substrate, irradiation-triggered single-event latchup (SEL) cannot occur. However, commercial SOI is still sensitive to total ionizing dose (TID) effects; the TIDs originate from a charge induced by gamma rays or X-rays that gets trapped in the buried oxide. The accumulated charge causes a major performance degradation of the analog blocks through mainly a shift of the threshold voltage in metal-oxide semiconductor (MOS) structures [3,4]. Therefore, SOI requires special technological hardening steps to achieve a sufficient level of robustness to TID [5].

Recently, however, ASICs fabricated in standard deep-submicron complementary metal-oxide semiconductor (CMOS) technologies have demonstrated robustness to SEL and TID especially when dedicated design topologies like enclosed (edgeless) transistor geometry and guard rings are used [3,6,7]. Without using a buried oxide, deep-submicron CMOS technologies have inherently a high tolerance to TID due to the reduced thickness of the gate oxide ($t_{ox} = 2.2$ nm).

This chapter focuses on basic aspects of design of high-quality voltage reference circuits and current references in standard commercial 130 nm CMOS technologies that are capable of operating in harsh radiation environments.

5.2 RADIATION-TOLERANT LAYOUT APPROACH FOR BANDGAP REFERENCE CIRCUITS

The bandgap reference circuit [8,9] is commonly used to implement a reference voltage generator. The operation of this type of circuit relies on the properties of the forward-biased p-n junction (diodes). However, with steady progress in downscaling of CMOS technologies, the use of bandgap reference circuits with conventional diodes in radiation-hard environments has two distinct disadvantages. First, the low supply voltage in modern CMOS technologies significantly complicates bandgap reference circuit design when conventional diodes are used [10-12]; a suitable approach using conventional diodes was introduced by Banba [14]. Second, it has been found that bandgap references featuring conventional diodes are rather vulnerable to TID effect [14]. Detailed analysis of the behavior of conventional bandgap references in deep-submicron CMOS technology indicates that radiation damage in diodes is the main cause of reference voltage shifts [15]. A short discussion of this follows.

In conventional bandgap reference circuits in CMOS, the diodes are usually implemented using a p-diffusion in (grounded) n-well (Figure 5.1). A shallow trench isolation (STI) field-oxide layer surrounds the p-diffusion area. The field oxide is usually thick ($t_{ox} >> 10$ nm); therefore, irradiation-induced holes get easily trapped and accumulated in the body of field oxide near the SiO_2-Si interface [16]. This trapped charge induces an excess concentration of electrons in n-well close to the field-oxide border, forming a parasitic side p-n+ junction (see also Figure 5.1) that has different properties than the principal p-n junction.

The parasitic side p-n+ junction is in parallel to the principal p-n junction, while its effective doping concentration depends heavily on the accumulated radiation

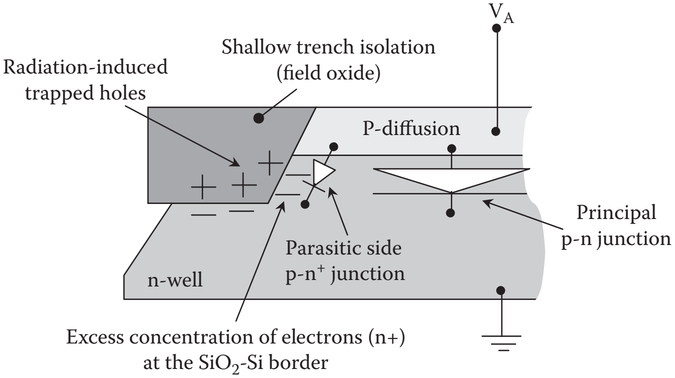


FIGURE 5.1 A conventional diode in the 130 nm CMOS: p-diffusion in a grounded n-well.

dose. Therefore, the voltage-to-current characteristic of the total device shifts considerably when it is operating in a radiation environment. This shift effect due to the charging surrounding STI appears to be the dominant factor of instability of the output voltage. For radiation doses up to 79 Mrad about 4% shift in the reference voltages, due to this effect, has been found [15].

A possible solution of this problem is the replacing of the (thick and radiation-intolerant) field oxide next to the p-diffusion by thin radiation-tolerant gate oxide. In this way two structures can be obtained: the gated diode and the conventional pMOS transistor shown in Figure 5.2. The gated diode has been proposed to be used for the assessment of radiation damage [17]. However, this device cannot be implemented in any CMOS technology because of design-rule limitations.

The second way to avoid field oxide adjacent to a p-n junction is using an MOS transistor layout in which the source-well junction is used as a diode. To get conventional diode-like behavior the effect of the gate must be either minimized or well defined. One possibility is tieing the gate to a high voltage, which is not a simple solution in low-voltage CMOS technologies. The other possibility is tieing the gate to the p-diffusion (drain) to obtain a constant effect of the gate on the diode's behavior. The corresponding device is shown in Figure 5.2. The obtained

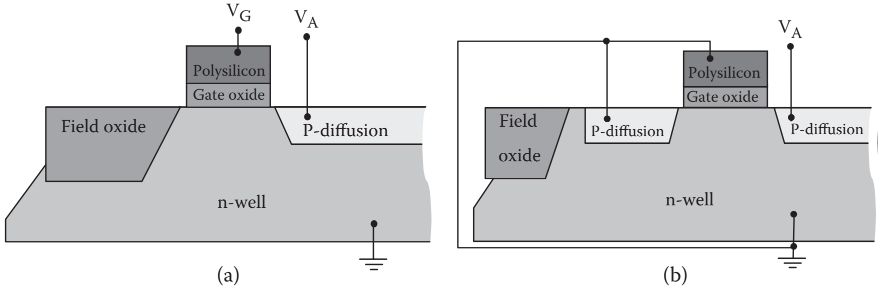


FIGURE 5.2 (a) Gated diode. (b) DTMOST.

structure is called a dynamic-threshold MOS transistor (DTMOST) [18,19]. This device can be operated as a diode with a low effective bandgap. In our design we used a P-channel DTMOS diode that can be realized in any twin-well (p-bulk) CMOS process.

The internal p-diffusion area (source) of the DTMOST can be surrounded by a gate oxide to form an enclosed layout geometry. In this way, the device is inherently radiation hard due to the absence of any thick oxide near the p-n junction.

5.3 TYPICAL CMOS BANDGAP VOLTAGE SUMMING REFERENCE

A typical CMOS bandgap reference circuit is shown in Figure 5.3.

In this type of circuit, the reference voltage depends heavily on the characteristics of the diodes. The current-to-voltage characteristic of a p-n junction is [20]:

$$I(V) = I_0 \cdot \left(e^{\frac{qV_{pn}}{kT}} - 1 \right)$$
$$I_0 = Const \cdot A \cdot T^{3+\gamma/2} \cdot e^{\frac{-E_g(T)}{kT}} \tag{5.1}$$

In this relation, q is the electron charge, k is Boltzmann's constant, T is the absolute temperature of the junction, A is the junction area, $E_g(T)$ is the material bandgap, and γ is a constant that is related to the temperature dependence of the mobility and the diffusion coefficients of the minority carriers. Rewriting this relation yields

$$V_{pn}(T) = V_g(T) + \frac{kT}{q} \cdot \ln\left(\frac{I}{Const \cdot A \cdot T^{3+\gamma/2}} \right) \tag{5.2}$$

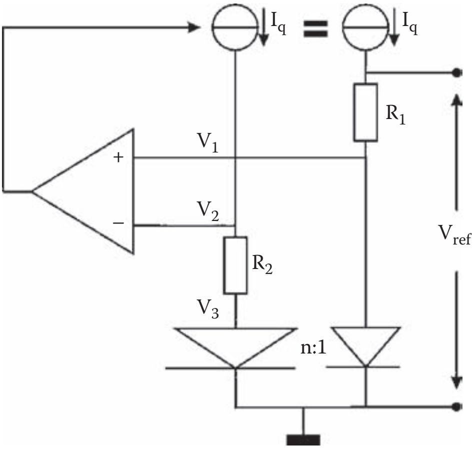


FIGURE 5.3 A typical CMOS bandgap voltage reference circuit.

In this relation, $V_g(T) = E_g(T)/q$ is the bandgap voltage. It is important to note that the voltage across a p-n junction is about conversely proportional to absolute temperature (CTAT). For $T \rightarrow 0$ the function $V_{pn}(I,T)$ tends to the value $V_g(T = 0)$ regardless of the current; in silicon $V_g(T = 0)$ is 1.12 V.

The operating current, I_q, in the circuit in Figure 5.3 is set by the feedback loop including the operational amp, which forces $V_1 = V_2$:

$$I_q = \frac{V_1 - V_3}{R_2} \tag{5.3}$$

where V_1 and V_3 are the voltages across the rightmost diode, D_1, and the leftmost diode, D_2, which is n times as big as D_1. Assuming that the diodes D_1 and D_2 differ only in size and taking into account (5.2), the operating current is

$$I_q = \frac{\frac{kT}{q}\ln(n)}{R_2} \tag{5.4}$$

Note that this current $I_q(T)$ is proportional to absolute temperature (PTAT).

The reference voltage as delivered by this type of circuit is the sum of the CTAT voltage across one of the diodes and a scaled up version of the PTAT voltage across resistor R_2. The easiest way to create such a voltage is shown in Figure 5.3, using one resistor, R_1. The total output voltage is then

$$V_{ref}(T) = V_g(T) + \frac{kT}{q}\cdot\ln\left(\frac{I}{Const\cdot A\cdot T^{3+\gamma/2}}\right) + \frac{R_1}{R_2}\frac{kT}{q}\ln(n) \tag{5.5}$$

which is temperature-independent (in first order) for reference voltages just a little higher than the material bandgap extrapolated to 0 K. The typical CMOS bandgap voltage reference circuit in Figure 5.3 generates an output voltage close to 1.22 V. In 130 nm CMOS technology the nominal power supply voltage is as low as 1.2 V, which is clearly insufficient to accommodate this type of bandgap reference circuit.

5.4 RADIATION-HARD VOLTAGE REFERENCES

The aim of this chapter is to present basics of the design of the radiation-tolerant bandgap reference circuits. As previously discussed, the DTMOST in gate-enclosed geometry is inherently robust to radiation effects and has a diode-like current-voltage relation. Therefore, we have chosen a DTMOST-based architecture for the design of the radiation-hard voltage reference circuit. Originally, the DTMOS transistor was proposed for ultra-low-voltage operation [19].

Figure 5.4a represents a MOS structure with the gate and the n-type substrate contacts connected together [18]. The built-in potential for the heavily doped *p*-type

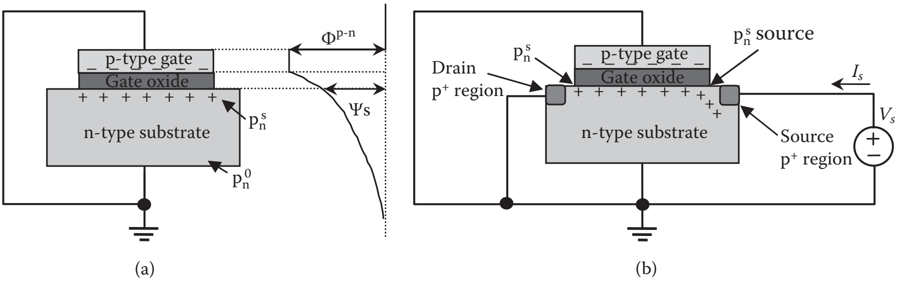


FIGURE 5.4 (a) MOS structure with the gate tied to the n-type substrate. (b) The DTMOST diode: MOS-transistor with the gate tied to the n-well and the drain.

gate and the n-type substrate, Φ^{p-n}, is about –1 V when the substrate doping concentration, N_D, is about 10^{17} cm^{-3} [20]. This built-in voltage partly drops in the substrate, making a potential Ψ_S on its surface. In the depletion and weak inversion region Ψ_S is a fraction of Φ^{p-n}:

$$\Psi_S = \Phi^{p-n} \big/ n \tag{5.6}$$

where $n = 1.2\ldots1.6$ (process dependent); therefore,

$$\Psi_S \approx -0.8 \cdots -0.6V \tag{5.7}$$

Due to the built-in potential the surface concentration of holes, p_n^s, exceeds the equilibrium concentration of holes, p_n^0, in the bulk of the substrate as follows:

$$p_n^s = p_n^0 e^{\frac{q|\Phi_S|}{kT}} \tag{5.8}$$

$$p_n^0 = \frac{n_i^2}{N_D} = \frac{Const_1 \cdot T^{3+\gamma/2} e^{\frac{-E_g(T)}{kT}}}{N_D} \tag{5.9}$$

where n_i is the intrinsic carrier concentration, and N_D is the doping concentration in the substrate.

$$p_n^s = Const_2 \cdot T^{3+\gamma/2} e^{\frac{q(-V_g+|\Psi_s|)}{kT}} \tag{5.10}$$

Expression (5.10) demonstrates that due to the effect of the built-in potential, the surface concentration of minority carriers increases, and the effective bandgap voltage is lowered:

$$V_g^{eff} = V_t\left(T\right) - \left|\Psi_S\right| \approx 0.3 \cdots 0.5V \tag{5.11}$$

The DTMOST diode is in fact a PMOS transistor with gate, drain, and substrate contacts connected together (Figure 5.4b). We restrict the analysis of the device's operation to the weak inversion region. In this region the source current, I_s, is caused by the diffusion of the inverse charge on the surface as follows [21]:

$$I_S = \frac{W}{L} \cdot \mu \cdot \frac{kT}{q}\left(Q'_{I,source} - Q'_{I,drain}\right) \tag{5.12}$$

where W and L are the width and the length of the device, respectively, μ is the surface mobility of holes, and Q_{I}' is the inversion charge per unit area, which is proportional to the surface concentration of holes, p_n^s. Combining the previous relations yields the following voltage–current relation for the DTMOS diode:

$$\begin{gathered} I_S = I_{S0} \cdot \left(e^{\frac{qV_s}{kT}} - 1 \right) \\ I_{S0} = Const_3 \frac{W}{L} \cdot T^{4+\gamma/2} e^{\frac{-E_g(T) - q|\Psi_s|}{kT}} \end{gathered} \tag{5.13}$$

Comparing (5.13) with (5.1) shows that a conventional p-n junction and the DTMOST diode (in the restricted region of weak inversion) demonstrate identical exponential current-to-voltage characteristics (Figure 5.5). However, the saturation current is much higher for the DTMOST diode due to the built-in potential Ψ_S.

Figure 5.5a shows measured *I-V* characteristics of a DTMOS diode and a conventional diode, at room temperature. These measurements nicely illustrate the exponential behavior of the DTMOS diode. At higher currents the *I-V* characteristic starts to deviate from the ideal exponential relation because there the weak inversion assumption is not satisfied anymore. Figure 5.5b shows the voltage across the DTMOST as a function of temperature at a few typical current settings that are nicely within the exponential region; clearly the diode voltage is conversely proportional to absolute temperature. By extrapolating the $V_s(T)$ curves at various bias currents to $T = 0$ K, the effective bandgap voltage is estimated to be 410 mV, with a temperature gradient (at constant current) of about 0.8 mV/°C.

The exponential character of the current-to-voltage relation of the DTMOST diode in weak inversion (5.13) enables the construction of a PTAT voltage source using the approach described for typical CMOS bandgap voltage reference. Due to the effect of lowering the bandgap voltage (5.11), the reference voltage of the present circuit will be much lower than that for the typical CMOS bandgap voltage reference. A bandgap reference circuit using DTMOST diodes (Figure 5.6) may be used to implement a low-voltage and radiation-tolerant voltage reference in standard CMOS technology.

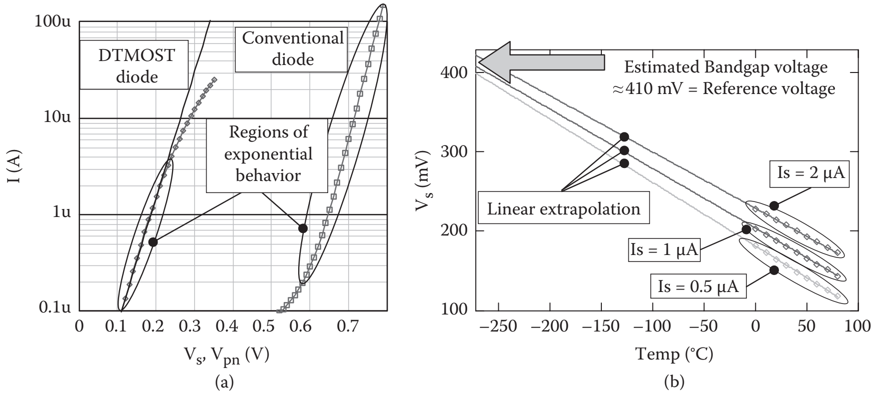


FIGURE 5.5 Current-to-voltage characteristics for both DTMOST configuration and conventional diode configuration. Voltage across the DTMOST at various currents as a function of temperature.

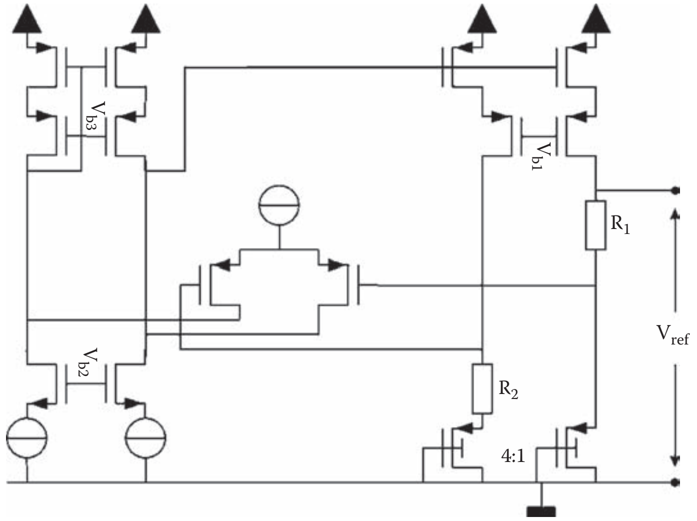


FIGURE 5.6 Schematic of the radiation-hard voltage bandgap reference.

The designed bandgap voltage reference circuit [22] is a straightforward circuit (Figure 5.6), consisting of two DTMOS transistors, a pair of cascoded current sources, and a two-stage operational amplifier. All MOS devices of the circuit are designed in the gate-enclosed geometry [3] with guard rings to guarantee radiation tolerance. The circuit was fabricated in a standard 130 nm CMOS process and occupies 0.064 mm^2.

The circuit generates V_{ref} of about 405 mV at a supply voltage down to 0.85 V, with a supply current of 170 µA. The spread of the reference diode is dominated by the threshold spread of the DTMOS diodes that directly affects the reference voltage. Being a differential circuit, spread effects cancel in first order for the remaining circuitry. In some applications not only the stability of the reference voltage but also its absolute value are important; this absolute value differs from chip to chip and is caused by the process variation and mismatch. The quadratic mean value of statistical spread of the reference voltage has been estimated as low as 6 mV.

To be able to compensate for process spread and model inaccuracies, our first implementation included a trimming possibility: resistor R_1, which generates the PTAT voltage, was built in multiple sections that can be bypassed externally. In this way the slope of the PTAT voltage can be trimmed to the slope of the CTAT voltage to get the minimum temperature coefficient of the reference voltage. Under this condition the reference voltage to temperature relation is a parabola with maximum deviation around 1 mV within the range from 0°C up to 80°C. Without trimming, the temperature coefficient of the reference voltage ranges from –0.1 mV/°C to +0.2 mV/°C, which is again due to the spread in the threshold voltage of the DTMOS diodes.

We used X-ray (10 keV) facility for the irradiation of the chips. The change of the reference voltage as a function of the radiation dose is shown in Figure 5.8 for six (unselected) samples.

Measurements show that due to the effect of the radiation the reference voltage fluctuates in the range less than 1% for doses up to 40 Mrad. This change is much lower than the typical 4% change at 79 Mrad for bandgap references using conventional diodes [15].

5.5 RADIATION-HARD CURRENT REFERENCES

The previous section showed a voltage reference that can also be used to create a proportional-to-absolute temperature current. However, for some applications a constant current is required; this section shows the design of a radiation-hard current reference in standard CMOS. Following the approach proposed by Banba [13] a current-summing current reference circuit can be designed. The circuit consists of DTMOST devices used as diodes, a pair of cascaded current sources, and a two-stage operational amplifier (Figure 5.7).

The core of the circuit is very similar to that of the voltage reference presented in the previous section. The main difference is the addition of two resistors, R_2, and using the current through the parallel combination of the diode and R_2 as (scalable) output current. The voltage across the DTMOST is CTAT; therefore, the current through resistor R_2 is also CTAT. On the other hand, the current in the DTMOS diodes is PTAT. After appropriate adjustment, superposition of the PTAT and the CTAT currents results in a temperature-independent reference current, I_{ref}.

The value of the reference current will vary in the range ±15% due to process spread of the resistors. At the same time temperature dependence of the value of I_{ref} is influenced by only the mismatch of the resistors, which is quite good in modern CMOS technologies. Note that any value of reference current can be generated with proper sizing of the rightmost branch in the PMOS mirror.

The circuit was made in a standard 130 nm CMOS technology, occupying 0.025 mm^2. The minimum supply voltage is 0.8 V, and the circuit is dimensioned to generate an I_{ref} of about 45 μA. When properly adjusted, the current-to-temperature

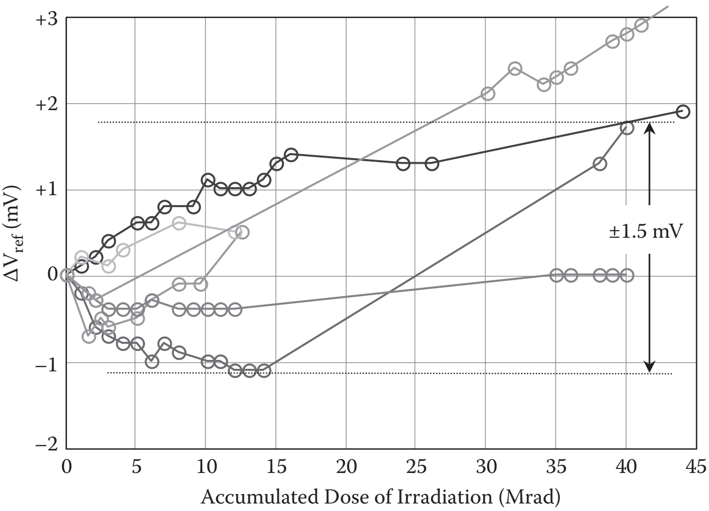


FIGURE 5.7 Measured shift of the value of the reference voltage during irradiation for six prototype chips.

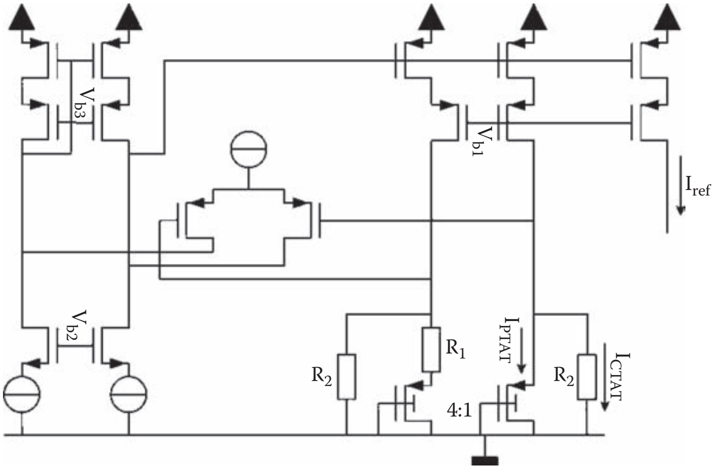


FIGURE 5.8 Schematic of the radiation-hard current bandgap reference.

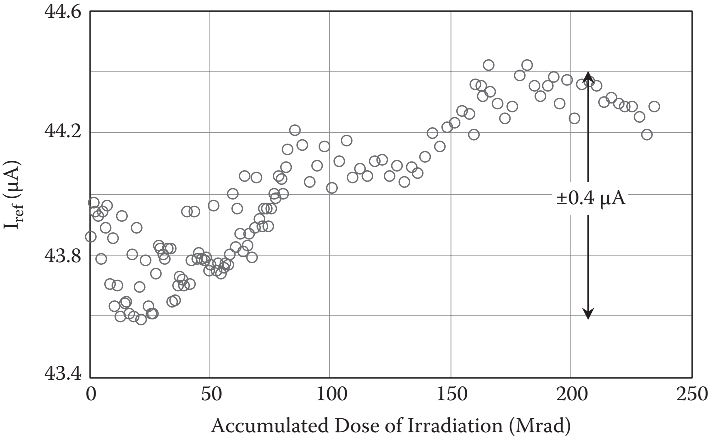


FIGURE 5.9 Measured shift of the value of the reference current during irradiation.

relation is a parabolic function with a maximum deviation of less than 0.2 μA (0.5%) in the range from 0°C up to 50°C.

Irradiation results in a shift of the reference current; Figure 5.9 shows the reference current as a function of irradiation. The figure demonstrates that the shift is relatively small: only ±0.4 μA (0.9%) after it has been irradiated with a dose as high as 200 Mrad.

5.6 CONCLUSIONS

With ongoing CMOS evolution, the gate-oxide thickness steadily decreases, resulting in an increased radiation tolerance of MOS transistors. Combined with special layout techniques, this yields circuits with a high inherent robustness against X-rays and other ionizing radiation. In bandgap voltage and current references, the dominant radiation susceptibility is then no longer associated with the MOS transistors but is dominated by the diodes. This chapter presents a few solutions to realize radiation-hard voltage/current reference circuits in standard low-voltage CMOS technologies using DTMOS diodes as radiation-tolerant diodes.

REFERENCES

1. E. Sall and M. Vesterbacka, “Design of a comparator in CMOS SOI,” in *Proc. IEEE 4th Int. Workshop on System-on-Chip for Real-Time Application,* pp. 229–232, 2004.
2. C.F. Edwards et al., “A multibit Σ Δ modulator in floating-body SOS/SOI CMOS for extreme radiation environment,” *IEEE J. Solid-State Circuits,* vol. 34, no. 7, pp. 937–948, 1999.
3. G. Anelli et al., “Radiation tolerant VLSI circuits in standard deep submicron CMOS technologies for the LHC experiments: practical design aspects,” *IEEE Trans. Nucl. Sci.,* vol. 46, no.6, 1999.
4. D.R. Alexander et al., “Design issues for radiation tolerant microcircuits in space,” in *Proc. 1996 IEEE NSREC Short Course,* pp. V-1–V-54, 1996.

5. M. Alles et al., "Evaluating manufacturability of radiation-hardened SOI substrates," *SOI Conference, 2001 IEEE International,* pp.131–132.
6. R.C. Lacoe et al., "Application of hardness-by-design methodology to radiation-tolerant ASIC technologies," *IEEE Trans. Nucl. Sci.,* vol. 47, pp. 2334–2341, 2000.
7. F. Faccio, K. Kloukinas, and A. Marchioro, "Single event effects in static and dynamic registers in a 0.25 µm CMOS technology," *IEEE Trans. Nucl. Sci.,* vol. 46, pp. 1434–1439, 1999.
8. R.J. Widlar, "New developments in IC voltage regulators," *IEEE J. Solid-State Circuits,* vol. 6, pp. 2–7, 1971.
9. K.E. Kuijk, "A precision reference voltage source," *IEEE J. Solid-State Circuits,* vol. 8, pp. 222–226, 1973.
10. A. Boni, "Op-amps and startup circuits for CMOS bandgap references with near 1-V supply," *IEEE J. Solid-State Circuits,* vol. 37, pp. 1339–1343, 2002.
11. J. Doyle et al., "A CMOS subbandgap reference circuit with 1-V power supply voltage," *IEEE J. Solid-State Circuits,* vol. 39, pp. 252–255, 2004.
12. J. Yueming and L. Edward, "Design of low-voltage bandgap reference using transimpedance amplifier," *IEEE TCAS II*, vol. 47, pp. 552–555, 2000.
13. H. Banba et al., "A CMOS bandgap reference circuit with sub-1-V operation," *IEEE J. Solid-State Circuits,* vol. 34, no. 5, pp. 670–674, 1999.
14. P. Moreira, "Radiation effects on the "CERN_Bandgap" circuit, private communication, 2004.
15. P. Moreira, "130 nm bandgap design review," CERN, private communication, 2005.
16. T.R. Oldham et al., "Post-irradiation effects in field-oxide isolation structures," *IEEE Trans. Nucl. Sci.,* vol. 34, no. 6, pp. 1184–1189, 1987.
17. A. Czerwinski et al., "Gated-diode study of the corner and peripheral leakage current in high-energy neutron irradiated silicon p-n junctions," *IEEE Trans. Nucl. Sci.,* vol. 50, pp. 278–287, 2003.
18. A.J. Annema, "Low-power bandgap references featuring DTMOSTs," *IEEE J. Solid-State Circuits,* vol. 34, pp. 949–955, 1999.
19. S.M. Sze, *Physics of semiconductor devices*. John Wiley & Sons, New York, 1981.
20. F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P.K. Ko, and C. Hu, "A dynamic threshold voltage MOSFET (DTMOST) for ultra-low voltage operation," in *Proc. IEDM '94,* pp. 809–812, 1994.
21. Y.P. Tsividis, *Operation and modeling of the MOS transistor,* McGraw-Hill Book Company, 1987.
22. V. Gromov, A.J. Annema, R. Kluit, et al., "A radiation hard bandgap reference circuit in a standard 0.13 µm CMOS technology," *IEEE Transactions on Nuclear Science,* vol. 54, no. 6, pp. 2727–2733, Dec. 2007.

6 Nanocrystal Memories: An Evolutionary Approach to Flash Memory Scaling and a Class of Radiation-Tolerant Devices

Cosimo Gerardi, Andrea Cester, Salvatore Lombardo, Rosario Portoghese, and Nicola Wrachien

CONTENTS

6.1 INTRODUCTION

The flash memory was conceived as a functional improvement of the erasable programmable read only memory (EPROM), which was invented in the 1980s from the initial idea of Frohman-Bentchkowsky [1]. The EPROM memory electrically programmed and erased by ultraviolet (UV) irradiation became the most important nonvolatile memory (NVM) application in the 1980s. The flash memory, called flash because the whole memory array is erased at the same time, introduced the advantages of the electrical erase as well as the possibility to reprogram the read only memory in situ, with no need of removing it from the system [2,3].

Over the years, flash memory has widely been accepted as the NVM of choice for many applications, and today the large majority of nonvolatile memories are based on flash technology. In the last decade, the flash market has grown faster due to the large diffusion of portable and low power consumption multimedia applications, which require an extensive use of nonvolatile function. The flash NAND has become the most scaled memory and hence the driver of the memory technology (both volatile and nonvolatile). Nonetheless, the continuous scaling of nonvolatile memories has pushed the technology of flash toward its limits [4]. Several constraints, mainly owing to electrical and reliability reasons, threaten the future scaling of the flash, forcing the memory research community to investigate new nonvolatile memory concepts.

Nonvolatile memories based on natural traps in dielectrics (e.g., SONOS) or on floating nanocrystals, artificially embedded in dielectrics, offer an interesting scaling alternative to the conventional floating-gate cell because of several potential advantages associated with the discrete nature of the storage [5-7]. These memories are considered an evolution of the flash concept, because the monolithic floating-gate (FG) is eliminated from the cell and is replaced by a number of discrete nodes. The discrete storage nodes make devices immune to the stored charge leakage caused by localized oxide defects, allowing for a very aggressive scaling of the tunnel oxide and hence of the cell area, by keeping good performance and reliability characteristics.

Today, charge trap memories have found an important field of application in embedded systems, where the nonvolatile memory is hosted into a logic system. Embedded applications are of such great interest mainly because of the ease of their process: a very thin storage layer can be implemented instead of the thick polysilicon floating-gate, and lower voltages can be used as well. Some semiconductor companies have announced that they have started production of embedded memories based on silicon nanocrystals. In particular, nanocrystals memory (NCM) has shown a superior endurance to high temperature than its counterpart SONOS. Recently NCMs have also shown a promising route toward radiation tolerant application. Actually, as information is stored in discrete centers, they are expected to exhibit a higher tolerance to radiation effects such as total ionizing dose effects (TID) and single-event effects (SEE).

This chapter is divided in two main sections. In the first part of this chapter, we will present an overview of the NCM technology as a candidate as an alternative to the conventional flash NVM, by comparing it to the mainstream technology. The discussion focuses on the scalability of the device as well as on performances and reliability. In the second part of the chapter, we will address the application of NCMs as radiation tolerant devices, for applications in fields such as avionics, nuclear power stations, nuclear waste disposal sites, medical, space, and military. In particular, we will compare NCMs radiation hardness characteristics with those of flash memories.

6.2 FLASH MEMORIES

6.2.1 Flash Memories: An Overview

Products using flash memories like cell phones, music players, memory cards, and universal serial bus (USB) drives are ubiquitous in everyday life. For this reason, the business of flash memories has grown very much in the last decade, exceeding US$22B in 2007. In the last year, the decrease in the semiconductor global market, related to the global economy turmoil as well as the erosion of memory unit prices, has reduced of more than 15% the incomes of flash market, and a recovery of the market is forecast starting from 2010.

Almost all flash memories are based on one of two architectures: NOR and NAND. NOR is the technology preferred by cellular handset makers since it provides fast reads. The NAND device reads data slowly but has fast write speeds and desirable features for storing digital photos and for MP3 audio, GPS, and other multimedia products. The exponential growth of many multimedia applications has driven the exponential growth of flash memories in particular for NAND devices, which have surpassed DRAMs (Dynamic Random Access Memory) in terms of scaled technology. The next challenge for flash memory will be the solid-state disk application in notebooks, which will compete with the current hard disk drive technology starting from 2010 [8].

As for other semiconductor devices, several technology innovations have been the driving force of a continuous cost reduction since the 1980s, with lithography improvement being the fundamental of these. In addition, innovations in flash have been boosted by the use of new architectures and designs. Self-aligned technologies, such as the one that aligns the floating-gate to the cell active area by using chemical mechanical polishing, are cardinal in flash technology because they reduce the space without the need of additional lithographic layers [5]. The introduction of NAND flash [9] has led to a further area scaling with respect to the NOR, thus allowing the introduction of a circuitry able to manage error detection and correction algorithms (compatible in terms of requested times with most applications such as memory card, USB drive, and MP3). Above all, the introduction of multilevel cells, where additional logical states are introduced by exploiting a precise control of the programmed or erased threshold voltage distributions, has allowed the memory storing capability to be increased by a factor of two or more [10] without any additional dimensional scaling, hence demolishing the paradigm of Moore's Law.

In the next few years, it will be increasingly difficult to scale the flash technology by following Moore's Law at the same pace, but through design improvements, new materials, and the introduction of algorithms, flash memories will be pushed beyond the 40 nm lithography node [9]. A number of complexities in maintaining a good trade-off between dielectrics scaling (tunnel and inter-poly) and nonvolatility will enhance electrical and reliability issues. The consequent scaling limitations will likely force cell architectural changes—by using discrete storage nodes in place of the continuous floating-gate and moving toward tridimensional cells [11,12], by using tridimensional process integration [13], or even by moving over new roads with the use of new emerging materials and data-storage concepts (including phase-change and resistive switching materials).

6.2.2 Basics of Flash Operations

A flash cell is a floating-gate metal-oxide semiconductor field-effect transistor (MOSFET; Figure 6.1)—that is, a transistor with a gate, the floating-gate, which is fully surrounded by dielectrics and hence isolated from the external. An external control gate (CG) drives the internal FG by means of capacitance coupling.

Being electrically isolated, the FG acts as a potential well, storing the injected electrons that screen the cell channel modulating the threshold voltage. The dielectrics are chosen to be a suitable trade-off between programming and erase performances and nonvolatility of data. Adequate dielectric thickness and quality ensure that stored electrons do not escape pushed out by the internal electric fields, generated by the stored charge itself. On the other hand, the dielectric thickness is chosen to allow the injection or ejection of electrons into or from the FG, induced by high electric fields generated by externally applied electrical pulses.

Usually, the gate dielectric in a floating-gate transistor is in the range of 7 ÷ 10 nm and is called *tunnel oxide* because, for some memory operations, electrons traverse the dielectric by a tunneling effect. The dielectric that separates the FG from the CG is defined as *interpoly dielectric* (IPD) or control dielectric; typically it consists of a triple layer of oxide–nitride–oxide (ONO). The ONO has an equivalent oxide thickness (EOT) in the range of 14–20 nm. A fully deposited triple layer has been introduced to improve the tunnel oxide quality, because the use of a thermal oxide over polysilicon implies growth at high temperature, affecting the underneath tunnel oxide [10]. The introduction of a nitride layer, sandwiched between two oxide layers, reduces the EOT of the dielectric with respect to a full SiO_2 layer but preserves an adequate tunneling barrier, thus improving the electrical performances of the cell without degrading retention characteristics.

As mentioned before, the most used architectures for flash memories are called NOR and NAND. The common element of both architectures is the unit cell (FG transistor). The programming and erase operations change the threshold voltage of the cell, due to the presence or absence of charge stored in the FG. The neutral state is associated with the logical state "1," and the negatively charged state, corresponding to electrons stored in the FG, is associated with the logical "0" [14].

The "NOR" cells are arranged through rows and columns as in a NOR-like logical gate [15]. Cells sharing the same gate form the word-line (WL), while those

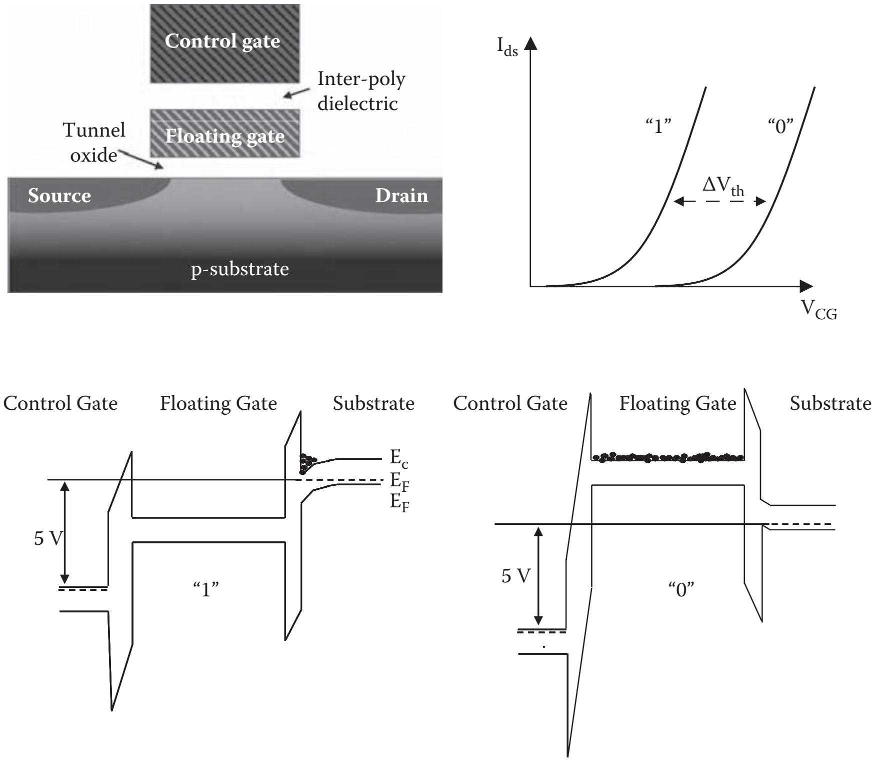


FIGURE 6.1 Schematic cross-section of a Flash cell. The floating-gate structure is common to all the nonvolatile memory cells based on the floating-gate MOS transistor. The other pictures show the band configuration of substrate, floating-gate and control-gate when the cell is erased (no electrons in the FG, "1") and when it is programmed (electrons stored in the floating-gate, "0"). Electrons are injected into the floating-gate, through either hot carrier injection or tunneling through the oxide potential barrier. They are ejected from the floating-gate by tunneling through the oxide potential barrier.

sharing the same drain electrode (one contact common to two bit-cells) constitute the bit-line (BL). In this array, the source electrode is common to all of the cells. Figure 6.2a shows the typical architecture of a NOR flash (the unit cell is enclosed in the rectangle); the figure also shows an electron microscopy cross section along the bit-line direction.

In NAND devices, the memory cells are arranged in series, with 16 or 32 memory cells connected to the bit-line and source line through two select transistors (selectors). This serial approach leads to a remarkable reduction of cell size with a consequent lower die cost compared with the NOR case. Typically, if F is the minimum design rule of the technology the NOR flash single-bit cell area is ~10 F^2 while the NAND single-bit cell area is ~ 4.5 F^2. Figure 6.2.b shows the typical architecture of a NAND flash device and an electron microscopy cross section along the bit-line direction. Note that in multilevel cell applications the effective cell areas reduce to ~5 F^2 for NOR and to ~2.2 F^2 for NAND.

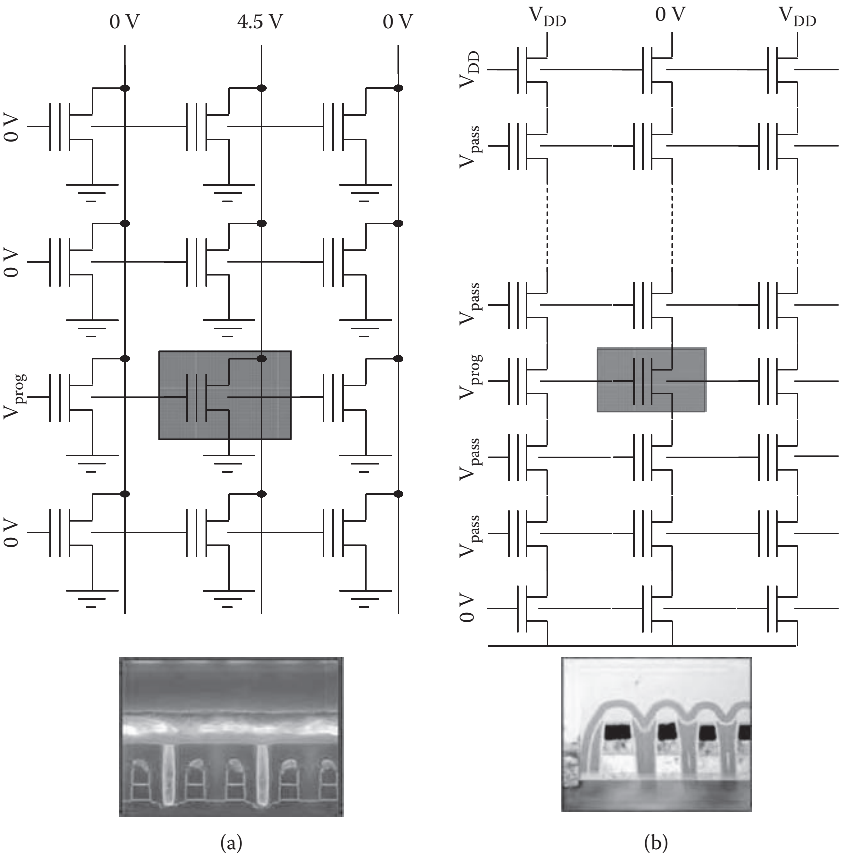


FIGURE 6.2 (a) NOR cell addressing; the cell under program is highlighted by the rectangle. (b) NAND cell addressing during the program operation; the cell under program is indicated by the rectangle.

6.2.2.1 The Read Operation in NOR and NAND

In NOR flash memories, the information stored in a cell of the array is sensed in a differential way—that is, by comparing the current of the read cell to that of a reference cell, physically identical to the matrix cell and biased with the same voltages [14]. In the case of single-level memories (1 bit per cell), the write (FG charged) and erase (FG discharged) logic states are well separated, as shown in Figure 6.3a. Since the read voltage, V_R, applied to the control gate is the same, the written cell ("0") current is lower than that of the erased cell ("1"). To distinguish between the two characteristics it is necessary to position the threshold voltage of the reference cell between the erased and the written cell characteristics. If the read current, Ids_R, is higher than that of the reference the cell is in the erase state (logic "1") if it is lower it is programmed (logic "0").

In the NAND architecture two selection transistors, placed at the beginning and at the end of the string, ensure the connections to ground (GND) and to the bit-

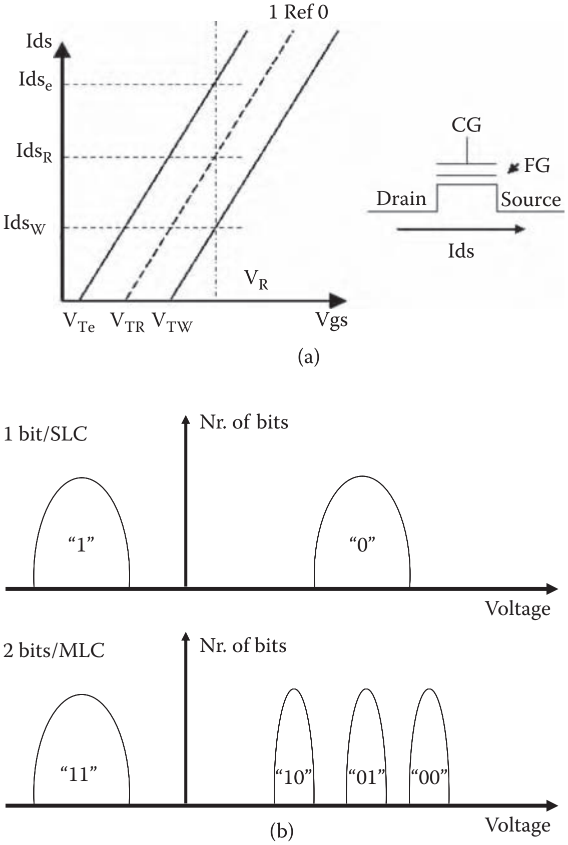


FIGURE 6.3 (a) Current/voltage characteristics as a function of the threshold voltage of a Flash cell. (b) Distributions of the erased VTH ("1") and of the written VTH ("0"), in a single level and in a multilevel Flash. In the multilevel cell, two intermediate levels ("10" and "01") with very tight and controlled distributions are introduced between the most erased ("11") and the most programmed distributions ("00").

line. When a cell is read, its control gate is set to GND, while the other gates are biased with a voltage typically of ~5 V. Therefore, they act as pass transistors. The threshold voltage of an erased NAND cell is negative, whereas a programmed cell has a positive threshold voltage. Driving the selected gate with a voltage close to zero, a current will flow through the series of all the cells if the addressed one is in the erase state. Conversely, a negligible current will flow if it is in the programmed state. Contrary to the NOR case, the sensed current in the serial string is very low, typically of 200–300 nA (tens of μA in NOR). Hence, for a NAND it is not practical to detect such a low current in a differential way. Therefore, the read operation is performed by charge integration, using the parasitic capacity of the bit-line. This is initially charged to a value of ~1 V; then, if the cell is erased the current flowing through the cell will discharge the bit-line, and if it is programmed the bit-line will remain charged.

The two reading modes in NOR and in NAND architectures have a different impact on the reading time, which turns out to be of a few tens of nanoseconds for NOR architecture and of a few tens of microseconds for NAND.

6.2.2.2 "Program–Erase" Operation and Reliability

The "program operation" is achieved by transferring the electrons from the substrate of the cell into its floating-gate, resulting in an increase of the cell threshold voltage. Due to the different programming mechanisms, the number of programmed bits per second is greater for a NAND memory than for a NOR.

The information in a NOR cell is written by the channel hot electron (CHE) mechanism [16]. The application of a voltage difference between the source and the drain generates a strong longitudinal electric field. Channel electrons acquire energy higher than the thermal equilibrium energy in the lattice, so they are called *hot*. Most of the hot electrons acquire the high energy in the depletion region close to the drain (where the field is higher). Under these conditions, some electrons gain enough energy to overcome the potential barrier at the Si channel–silicon oxide interface (3.15 eV). Applying a voltage difference between the control gate and the substrate introduces an additional transversal electric field that favors the injection of electrons from the channel into the floating-gate. As time elapses, the injection of electrons saturates, because as the negative charge accumulated in the FG increases it generates an opposite electric field that blocks the electron injection.

The CHE programming operation is quite fast, but the current necessary for programming is high: the larger the number of cells to be quickly programmed, the bigger the current consumption. Consequently, CHE programming is a very expensive operation, especially in terms of the area devoted to peripheral circuits for high current generation. Typical voltages applied to the cell during programming operation are 4.5 V on drain and ~10 V between gate and substrate (in some cases the substrate is biased at a low negative bias to enhance the CHE by secondary hot electron generation) [16]. Actually, the choice of voltages is determined by several factors, such as programming speed; reliability against parasitic effects such as drain turn-on; snap-back; and disturbs related to write, erase, and read operations of the adjacent cells.

A distribution of threshold voltages results from programming or from erasing operations, carried out on a large number of cells, as shown in Figure 6.3b. The spread in each distribution is due to different factors such as process variation, power-supply variation, and source and drain modulation. Despite all these effects, it is important that the distributions have well-controlled width to correctly place them in the memory cell working window (the available program–erase window of the cell).

The requirement is very stringent in multilevel memories—that is, memories in which there is more than one bit per cell (see the lower part of Figure 6.3b). As the number of bits to be stored in the memory cell increases, the number of distributions to be incorporated in the working window increases as well. Since the cell operating window width is limited by physical process and tends to degrade with increasing the number of program–erase cycling, the read window between neighboring states becomes smaller. The consequence is a requirement for smaller V_{TH} distribution per state, resulting in very strong constraints on reliability margins [10].

If a cell does not need to be programmed, the simultaneous application of high voltages on both gate and drain must be avoided, inhibiting the addressing of the cell. Due to the device architecture, there will be some cells with a high voltage on the gate but with 0 V on drain and some other cells with a high voltage on drain but 0 V on gate (Figure 6.2a). The cells sharing the same bit-line of the cell that has to be programmed can be affected by a "drain disturb"—that is, due to the voltage applied on the drain, the already programmed cells (with a negative charge in the floating-gate) may lose charge. On the contrary, the erased cells sharing the same word-line may be subject to V_{TH} increase (programming) due to a tunneling effect through the tunnel oxide. In addition, the programmed cells sharing the word-line with the cell under programming may suffer from electron ejection from the FG toward the control gate, with a consequent lowering of the threshold voltage.

As already mentioned, the write operation in NAND memories occurs with a different mechanism based on the tunneling of electrons in the presence of a high electric field. An intense electric field across the tunnel oxide tends to modify the tunneling barrier, which assumes a triangular shape. The tunneling across a triangular potential barrier is known as Fowler-Nordheim (FN) tunneling [17]. During programming, a high voltage difference is applied between cell gate and substrate. Therefore, electrons are injected from the cell channel into the floating-gate, traversing by tunneling effect the oxide barrier, which is made thinner because of the triangular shape induced by the high electric field. To improve the programming performance it is necessary to increase the tunneling efficiency and hence the applied electric fields [16]. This requirement has the heavy consequence of degrading the oxide by generation of traps. A reduction of tunnel thickness leads to an improvement in injection efficiency. However, the tunnel oxide thickness is strongly constrained by a stress-induced leakage effect (SILC)—that is, the tunneling assisted by traps forming in the oxide layer during the stress. The SILC effect becomes prominent in thinner oxides limiting the tunnel oxide thickness to ~7 nm [10,16]. Another disadvantage of the FN tunneling is the time needed for programming, typically much longer than in the CHE. On the other hand, a big advantage is the very low programming current (~nA per cell). This feature renders the FN method suitable for parallel programming a very big number of cells.

As mentioned, in a NAND memory a cell is a part of a string (Figure 6.2b), which can be addressed by drain and source selectors. As an example, if a specific cell must be programmed, the drain selector must be biased to VDD (drain supply voltage), the cells of the strings that should not be programmed are placed at ~9 V, the gate of the source selector is GND, and the bit-line is biased at 0 V. The gate of the cell under programming ranges from 15 to 20 V; therefore, it is fundamental to prevent erased cells, which share the word-line of the cell under programming, from being subject to programming. A method is to bias the drain of the erased cells with high voltages so that the channel potential increases. In this way, the electric field between the substrate and the FG is reduced, inhibiting the electron FN tunneling into the floating-gate. This approach is quite expensive because a large device area has to be devoted to high-voltage circuitry. In recent devices the consumption of area circuitry is avoided by the *self-boosting* [18] mechanism. The self-boost, which uses a dynamical substrate voltage buildup by capacitance coupling, has been of great benefit for NAND development, favoring a remarkable device area reduction.

The FN tunneling is also used to exploit the erase operation in both NOR and in NAND. To originate the tunneling, a high voltage across the tunnel oxide is applied to remove the electrons from the FG. To avoid deleterious junction voltage breakdown and band-to-band tunneling effects, the erase bias is applied between the control gate and the cell channel. The memory cell is placed in an insulated triple-well to allow high-voltage biasing of the p-well. The electrons are extracted along the channel, thus eliminating parasitic source junction leakage with a consequent strong reduction in current consumption [19].

Both NOR and NAND flash memories are erased per blocks, so a large number of cells is erased in a short time [10]. Due to the architectural and technical specification differences, the sector erase time in NOR is higher than in NAND architecture [14]. In NOR flash memories the erase algorithm is much more complex than the simple application of a voltage high enough to enable the tunneling; it has to manage the voltage values applied both to the addressed cells and to the unselected cells as well as their transients. The erase distribution must be low enough to ensure a good sensing margin; however, it cannot be too low because the more erased (overerased) bits of the distribution can become negative. This occurrence is not acceptable in NOR architecture because it can lead to spurious current consumption from the unaddressed cells. An erase verify algorithm has to be applied, controlling that the cell current is high enough to distinguish between erased and programmed cells.

In NAND flash the erase is performed by biasing the isolated p-well with a quite high voltage with the word-lines at 0 V. Contrary to NOR, in NAND architecture it is possible to locate the erased distribution into the negative V_{TH} half-plane. In fact, the cell threshold voltage can be negative because the cells must act as pass transistors to activate the read operation. In NAND the leakage due to the subthreshold current is not a concern, because the selectors of the unselected bit-lines prevent them from injecting any spurious current. In NAND, the erased distribution width is broad but does not significantly affect the string series resistance during the read. Indeed, during read algorithm, all the cells of the string except the addressed one are biased with a sufficiently high gate voltage (~4÷5 V). Though in NAND great precision in placing the erased distribution is not necessary, an adequate margin with respect to the read condition is mandatory.

A further requirement is that the erased distribution has enough margins to contain the degradation due to extensive program–erase cycling. Figure 6.4 shows the "cycling window" (endurance) of a NAND cell as a function of the number of program–erase cycles, obtained with fixed pulses and without correction algorithms. The threshold voltages of both the erased and the programmed cells increase with the number of cycles. This phenomenon is ascribed to charge trapped in the oxide and cell gain degradation. Since the erased V_{TH} tends to increase with cycling, a suitable margin has to be defined at the beginning of the cell life to account for such a shift.

The cycling degradation also affects the NOR device, but in this case it is possible to apply further erase pulses. In other words, the cycling degradation results in an increase of erase time, but the NOR specifications can account for that. In NAND architecture, the specifications do not allow room for further recovery erase pulses

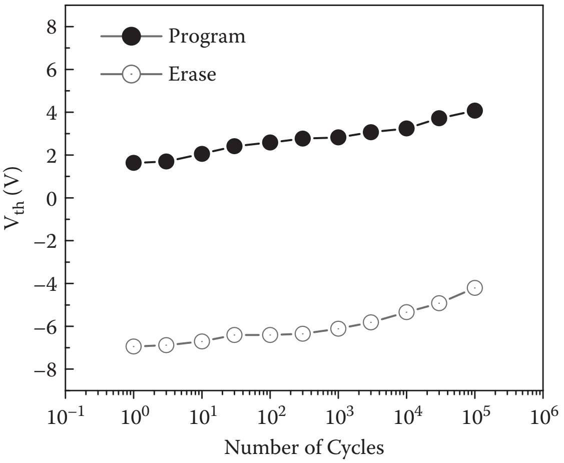


FIGURE 6.4 Cycling window in NAND Flash memories; since the erased level shifts toward positive voltages, a proper threshold margin at the beginning of the operating life is required.

TABLE 6.1
Comparison between NOR and NAND Flash Memories

	NOR	NAND
Architecture	1 contact every 2 cells	1 contact every 16 or 32 cells
Max memory size	1 G	32 Gb
Cell size (F^2)	10	4.5
Read access	Random/fast (~50 ns)	Serial/slow (10–30 µs)
Programming mechanism/throughput	CHE/0.5 MBs^{-1}	FN tunneling/8–10 MBs^{-1} (Parallel programming)
Erase time	1 s/sector	1 ms/sector
Function	Code storage	Data storage

if the first one has not been effective; for this reason, the erase pulse level and width must be calibrated very carefully.

Table 6.1 reports a comparison of the main characteristics of NOR and NAND flash memories.

6.3 NANOCRYSTAL MEMORIES

6.3.1 Nanocrystal Memories: An Overview

The nanocrystal memory cell consists of a single MOSFET device where a few electrons are stored in a layer of randomly distributed floating nanocrystals (dots), artificially obtained by different techniques. Figure 6.5 shows a schematic representation

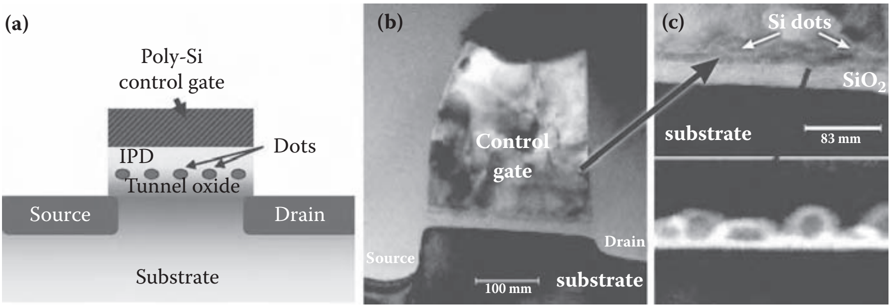


FIGURE 6.5 (a) Schematic of a nanocrystal memory cell, with nanocrystals (dots) in place of the continuous floating-gate of a standard Flash. (b) TEM micrographs showing the device structure and the gate stack with the dots. (c) The figure on the bottom right is an energy filtered TEM cross-section, which shows the Si dots.

of a nanocrystal memory cell, where the continuous floating-gate layer, typical of flash memories, is replaced by the nanocrystals. This structure gained a certain interest in the mid-1990s mainly as capacitorless DRAM [20]. The new concept, particularly appealing for mass production of NVMs, was proposed for the first time at the 2000 IEEE International Electron Devices Meeting [21].

More recently, significant results with multimegabit demonstrators have been reported in literature, further underlying the interest in NCMs for NVM applications [22,23]. Such results demonstrate the use of memory cells which reversibly store charge in isolated silicon nanocrystals, embedded in the dielectrics, enabling reduction of the program–erase voltages (due to tunnel oxide scaling). Moreover, due to the area savings from memory module peripheral voltage scaling and the reduction in mask count over conventional floating-gate technology, silicon nanocrystal NVM technology could substantially reduce the cost of flash devices in future technology nodes.

Experimental and theoretical analysis of the threshold voltage distributions of silicon nanocrystal memories—addressing the main issues of NCM scaling limits inherent in their process formation—has indicated that, particularly for embedded applications, this technology has serious potential to push the scaling of flash at least down to the 32 nm node [24,25].

A large number of literature articles report on both semiconductor and metal nanocrystals, which have been extensively studied with respect to their ability to store charge. The major advantages of metal nanocrystals over their semiconductor counterparts include a higher density of states around the Fermi level, an improved capacitance coupling with the conduction channel, a wide range of available work functions, and smaller energy perturbations due to carrier confinement. Metal nanocrystals also provide higher size scalability.

Moreover, to enable single-electron or few-electron memories by the Coulomb blockade effect, smaller nanocrystals should be preferred. In a semiconductor nanocrystal, the band-gap of nanocrystals is widened compared with that of bulk material due to the multidimensional carrier confinement that reduces the effective depth of the potential well, compromising the retention time. This effect is much smaller in metal dots because of the presence of thousands of conduction-band electrons in the nanocrystal, even in charge neutral state. As a result, the increase of Fermi level is minimal in metal nanocrystals [26,27].

However, despite the high potentialities of metal dots, their integration in a complementary metal-oxide semiconductor (CMOS) front end is still a limitation, and so far only simple demonstrators based on a single memory cell or capacitor have been reported. In fact, the possible contaminations arising from the implementation of metals in a CMOS front end raise serious reliability problems. Consequently, the semiconductor nanocrystal technology has progressed faster, and several prototypes of nanocrystal-based multi-Mb devices have been reported in the literature [22,23].

6.3.2 Si Nanocrystals Realization

The key technology for silicon nanocrystal memories is how to obtain the nanometer scale dots embedded in a dielectric layer. In fact, high nanocrystal density, nanometer size, good uniformity both in size and in shape, lateral isolation, planarity on the tunnel oxide, and background charge minimization are all required in a reliable fabrication process. Several methods to realize the Si dots have been proposed and investigated in the last several years. The most used methods reported in the literature are ion implantation [28], aerosol [29], self-assembling photo-resist technology [30], and chemical vapor deposition (CVD) [25,31,32]. Specifically, the use of CVD methods has proven to be a very convenient technique because of its immediate implementation in a CMOS processing and because of the excellent control it provides on the deposition parameters. Si nanocrystals size and density are driven by pressure and temperature conditions as well as by the substrate nature. Actually, the physical properties such as stress, roughness, or defect as well as the chemical state of the substrate can play an important role in the dots nucleation. Several results have been reported on the effects of surface treatment, and very high Si dot density (larger than 10^{12} dots/cm²) has been obtained on highly hydroxilated SiO_2 [25].

To study the kinetics of the dots formation, a suitable characterization technique is required. From this point of view, transmission electron microscopy (TEM) associated with an energy filtering system (EF–TEM) couples the high spatial resolution typical of the TEM analysis (of the order of the angstrom) to the compositional information obtainable by electron energy loss spectroscopy (EELS) [33].

Through this technique, it is possible to monitor the evolution of a thin Si film on top of an SiO_2 layer, gradually following the growth, from nucleation to coalescence (see Figures 6.6a–6.6f). By accurately controlling CVD process parameters it is possible to drive the Si nuclei growth up to the formation of islands with a diameter of a few nanometers. Figure 6.6g shows an example of the dependence of both Si coverage and dot density as a function of the deposition time for a set of isothermal process conditions.

As Si dots are formed by island growth during CVD, a remarkable concern for memory applications is the impact of dot density fluctuations from cell to cell. Nonetheless, the physics and chemistry inherent in the formation process lead to some benefits. The silicon nucleation and growth during CVD, in fact, is not completely random but takes advantage from a quasi self-ordering induced by the process itself.

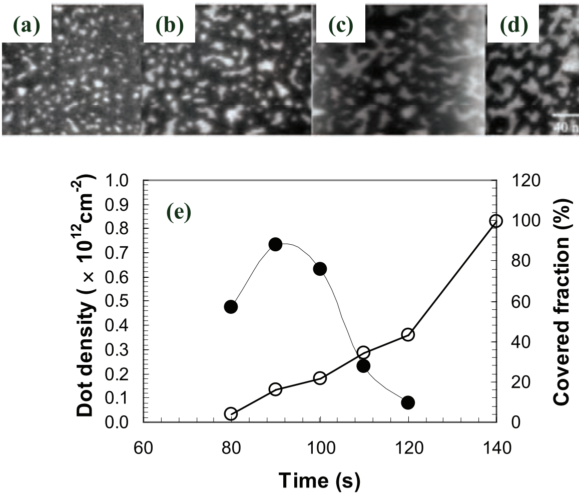


FIGURE 6.6 A sequence of EF–TEM micrographs of Si dots on a SiO2 layer for isothermal processes at 550°C and times of: (a) 90 s, (b) 100 s, (c) 110 s and (d) 120 s, respectively. The white regions are Si dots. As the deposition time elapses, the Si coverage increases since the nucleated dots grow and coalesce. At the same time, nucleation of new dots takes place at any stage of the deposition, even very close to the complete coverage with Si. (e) Si dot density and surface fraction covered by the dots as a function of deposition time for a set of isothermal processes. (Reprinted from S. Lombardo, B. Salvo, C. Gerardi, T. Baron. "Silicon nanocrystal memories." *Microelectronic Engineering*, vol. 72, Issues 1–4, 2004, pp. 388–394, with permission from Elsevier.)

These characteristics arise from the formation of denuded regions around each stable island, which prevents the growth of new nuclei. These regions result in a depletion of adatoms in the proximity of a stable nano-island. As a result, neighboring nuclei are well separated from each other [24].

Figure 6.7a shows a plan-view EFTEM micrograph of silicon dots grown with CVD. The nanocrystal size and separation distributions are reported in Figures 6.7c and 6.7d, respectively. The figures show that the distribution of separations exhibits a peak indicating a spatially nonrandom nucleation process (Figure 6.7c), which cannot be explained by a pure random (Poisson) nucleation model (Figure 6.7d) [34].

6.3.3 Nanocrystal Memory Cell

Using a simplified model, the threshold voltage shift of the memory cell after electron injection in the dots, ΔV_{TH}, can be approximated by [24]

$$\Delta V_{TH} = \frac{qn}{\varepsilon_{ox}}\left(t_{ipd} + \frac{1}{2}\frac{\varepsilon_{ox}}{\varepsilon_{Si}} t_{NC} \right) \tag{6.1}$$

where ΔV_{TH} is the threshold voltage shift, t_{ipd} is the thickness of the interpoly dielectric, t_{NC} is the size of the nanocrystal, ε_{ox} and ε_{Si} are the permittivities of SiO_2 and Si,

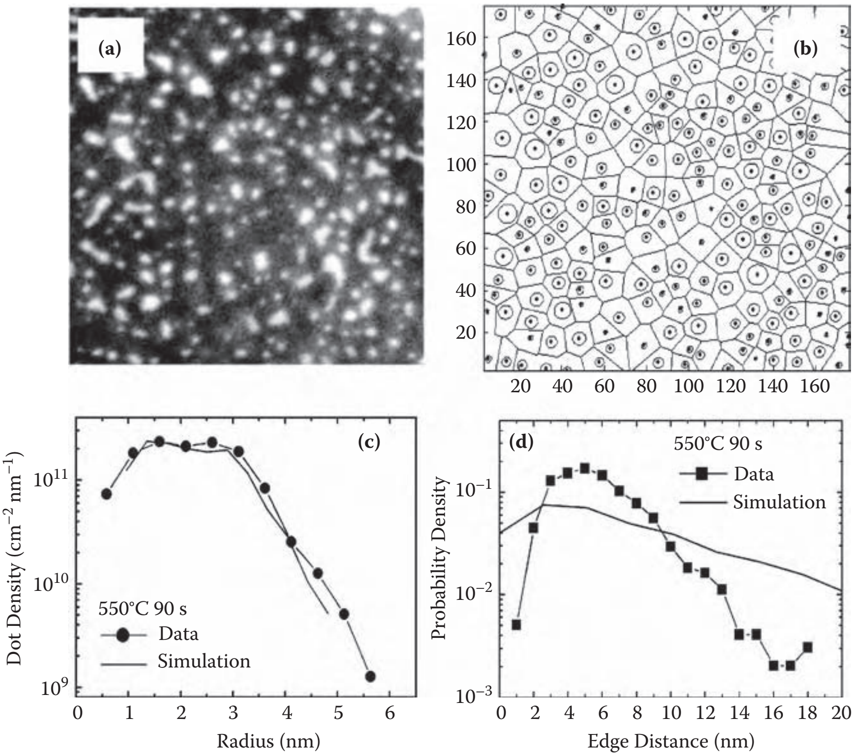


FIGURE 6.7 (a) EFTEM micrographs in plan view of Si nanocrystals (white spots) deposited at 550°C for 90 s. (b) Representation of the Si stable nuclei with their exclusion zones. Measured nanocrystal size distribution (c) and edge to edge separation, (d) extracted from experimental data such as the plan view TEM image shown in (a). The solid curve in both figures represents simulation based on a random nucleation model (Monte Carlo simulation). A random nucleation model cannot explain the peak in the nanocrystal separation distribution function (d).

respectively, q is the electronic charge, and n is the density of nanocrystals. We can further approximate the charge stored in the nanocrystals as an ideal sheet of charge located at a distance t_{ipd} from the gate of the device. In this case, the threshold voltage shift reduces to

$$\Delta V_{TH} = \frac{qn}{\varepsilon_{ox}} t_{ipd} \tag{6.2}$$

From this equation, it is clear that the maximum threshold voltage shift, which is a measure of the memory working window, is directly proportional to the number of Si dots. A nanocrystal density of 10^{12} dots/cm^2 is a good trade-off between adequate device sensing and interdot separation, which ensures charge lateral localization. This value is obtained by taking into account dot diameters of ~5 nm that must be separated at least ~4÷5 nm to avoid interdot electron direct tunneling. Such a

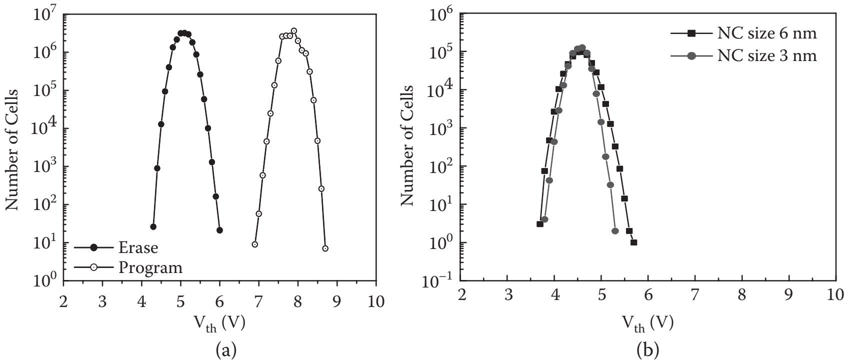


FIGURE 6.8 (a) Typical threshold voltage distributions (for erased and written states) for a 16 Mb nanocrystal array with dot size of ~ 6 nm and density of ~1012 cm^{-2}. (b) Comparison of the erased distributions for dots with average sizes of 3 and 6 nm, respectively. Square symbols represent the cells with bigger dot sizes and more dispersed distribution. [From C. Gerardi et al. " Nanocrystal memory cell integration in a stand-alone 16-Mb nor flash device," *IEEE Transactions on Electron Devices*, Vol. 54, Issue 6, pp. 1376–1383 (2007)].

separation is necessary to preserve the immunity against SILC, due the localization of charge in discrete nodes. A density much lower than 10^{12} dots/cm^2 implies a serious device failure because of a too small program–erase window and very reduced sensing margins. For a nanocrystal density of 10^{12} dots/cm^2, a tunnel oxide of 5 nm, and a control oxide thickness of 10 nm, the threshold shift is nearly 0.5 V if one electron per dot is considered.

In multimegabit or gigabit devices, the distributions of erased and programmed threshold voltages exhibit a typical spread, which in conventional flash devices is related to technological parameters such as fluctuations in the cell dimensions as well as in the dielectrics thickness. Typically, the distribution widths are of 1 V or even less; therefore, to have enough sensing margin the separation between the least erased bit (erased distribution tail) and the least programmed bit (write distribution tail) must be at least ~1 V. Hence, to match the program–erase window with a suitable sensing functionality, ΔV_{TH} has to be higher than 2 V. Figure 6.8a shows the write and erase threshold voltage distributions of a 16 Mb nanocrystal memory array in NOR architecture (CHE write, FN erase): a separation of ~1 V between the distributions is enough to distinguish the logical states.

To obtain a ΔV_{TH} higher than 2 V, we have to consider a number of electrons per nanocrystals higher than four; this is typically obtained in NCMs [35]. When several electrons are stored in a single dot, charge confinement or Coulomb blockade effect, which raises electron energy levels in the dot, becomes evident. If we consider a dot of 5 nm in size, containing five stored electrons, the single particle energy level of the fifth electron is approximately 0.5 eV higher than the silicon conduction-band edge, which effectively reduces the oxide barrier from 3.15 eV to ~2.6 eV.

CHE programming speed and threshold voltage saturation depend on the dot average size and density. In each case, programming saturation occurs when the rate of electrons injected into the dot is in balance with the electron removal rate due to Fowler-

Nordheim tunneling through the control dielectric. As the dot mean size is reduced, decreased capture cross section and Coulomb blockade effects cause a slower programming speed and a lower saturation V_{TH}. The higher energy level of a smaller dot results in a faster erase due to the Coulomb blockade effect. Figure 6.8b shows a comparison between the erased distributions corresponding to the dots with average sizes of 3 nm and 6 nm, respectively; the distribution with higher Si dot sizes has the larger width. A better control of the erase can be obtained on smaller dots due to Coulomb blockade: the higher energy level of the smaller dot results in faster erase [20].

Active dielectrics are fundamental for NVM not only because they have to ensure the functionality of the MOSFETs but also because they have to be good tunneling barriers, allowing charge transfer from the channel to the storage medium during programming or erase operations and at the same time ensuring the charge storage. In conventional flash memory the tunnel oxide has a thickness higher than 9 nm for NOR architecture and higher than 7 nm for NAND architecture. In NCMs, the tunnel thickness can be scaled down to 4÷5 nm because of the nanocrystals, immunity to SILC degradation. The interpoly dielectric can also be scaled from the conventional 15 nm down to 10÷12 nm.

It is worth noting that the discrete charge storage enables NCM cells to operate with the so-called dual-bit mode. This mechanism differs from the multilevel mechanism; in fact, in the dual-bit case the charge is stored in two different locations of the cell resulting in two stored bits per cell. During CHE programming, electron injection occurs mainly near the drain; electrons are mainly stored in the nanocrystals localized over the channel region in proximity to the drain. Consequently, the charged dots will screen only the portion of channel near the drain. So the memory cell will exhibit two different threshold voltages, depending on the way it is read: collecting the channel current from the source (reverse read) or from the drain (forward read). By means of an adequate circuit, it is possible to enable the alternate reading from drain or from source allowing the dual-bit cell mode. Figure 6.9 schematizes the dual-bit mechanism in an NCM, also showing an excellent robustness to lateral charge migration at high temperatures [25].

The process integration flow defined for nanocrystal memories is similar to the conventional process used in state-of-the-art flash memory technology. The main changes are the replacement of the continuous monolithic floating-gate electrode with the nanocrystals and the use and optimization of thinner tunnel and control dielectrics. Details of the cell are shown through TEM cross section in Figure 6.5 (along the bit-line) and in Figure 6.10 (along the word-line).

Eliminating the floating-gate improves the aspect ratio of the NCM cell by reducing its height and facilitating the integration in a complex CMOS process. Figures 6.10a and 6.10b show a comparison of a flash cell with an NCM cell, both realized with the same design rules corresponding to a technology node of 90 nm with shallow trench isolation technology.

As mentioned before, conventional control dielectric in a flash memory cell is the multilayer ONO where bottom- and top-layer oxides are obtained by CVD high-temperature oxide (HTO), in which the typical process temperature is ~800°C. In view of process integration, the dots must be passivated to prevent oxidation during HTO layer deposition. This is a fundamental aspect for the functionality of the device.

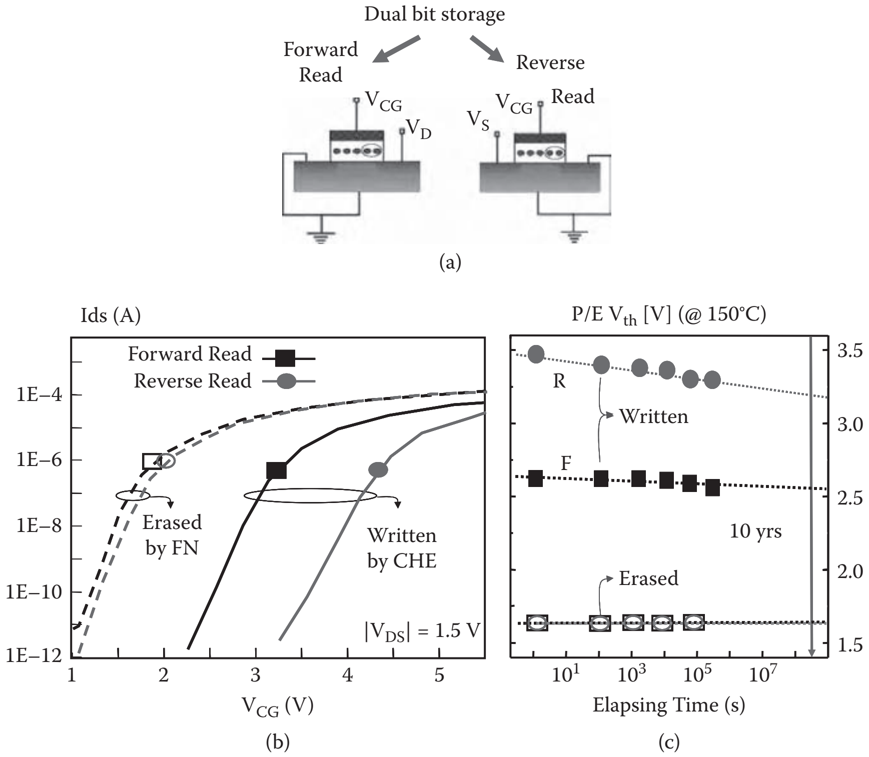


FIGURE 6.9 (a) Schematics of forward and reverse reading operations. (b) Cell transfer characteristics in the erased state and after CHE programming while the reading voltage is applied either to the source (reverse) or to the drain (forward). (c) Data-retention at 150°C for the erase level and for the programmed level read in forward and reverse modes, the high temperature data shows that there is poor lateral migration of the electrons, which remain mainly localized near the drain region after the CHE injection. [From B. DeSalvo et al. "Performance and reliability features of advanced nonvolatile memories based on discrete traps (silicon nanocrystals, SONOS). *IEEE Transactions on Device and Materials Reliability*, Sept. 2004, vol, 4, Issue 3, pp. 377–389].

Si dot passivation can be obtained by using a rapid thermal process in a nitridation environment, allowing the dots to be protected by a thin shell of Si_3N_4. Figure 6.10c shows the N1s X-ray photoelectron spectra (XPS) comparing the as-deposited dot with the nitridation. In the latter case, the XPS peak witnesses the formation of the nitride shell. Besides the advantages of preserving the dot from oxidation, the nitridation process neutralizes the interface states between the dots and oxide, thus neutralizing the charge trapping [36].

6.3.4 Nanocrystal Process Integration in a Multimegabit Array

Several studies have addressed the integration and compatibility of such memory cells with complex circuitry, containing high-voltage and low-voltage transistors and

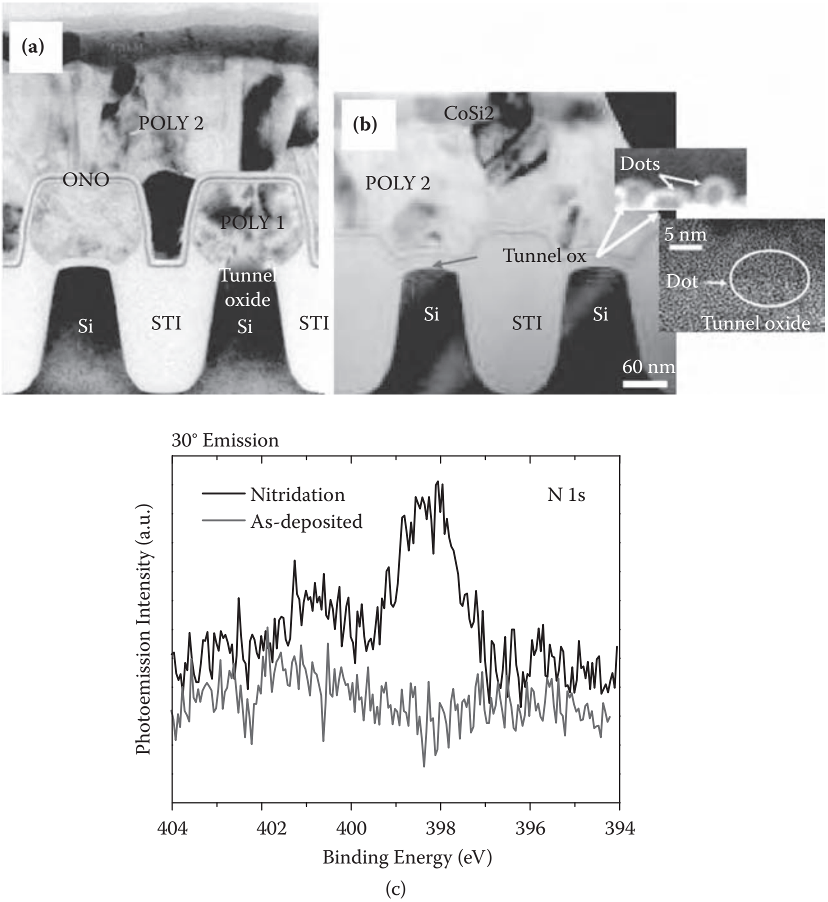


FIGURE 6.10 Cross-section comparison of a conventional 90 nm node Floating-Gate cell (a) with a nanocrystal memory processed with the same technology node (b). The cross-sections have been performed along the word-line direction. The insets in (b) are TEM magnifications showing the nanocrystals embedded between the substrate and the control gate. Typical dot size here is of 5 nm. (c) N1s XPS spectra comparison between the as-deposited dots subjected to passivation by nitridation. (Reprinted with permission from, Crupi, I., Corso, D., Ammendola, G., Lombardo, S., Gerardi, C. DeSalvo, B., Ghibaudo, G., Rimini, E., Melanotte, M. “Peculiar aspects of nanocrystal memory cells: data and extrapolations. *IEEE Transactions on Nanotechnology*, Volume 2, Issue 4, pp. 319–323.)

charge pump capacitors. Poor control on the size distribution of a CVD process has also been found to induce several reliability issues on the memory cells [23]. This drawback is minimized by controlling the dispersion in dot size and number as much as possible. The more advanced NCM technology reported so far is realized with a 90 nm lithography node. Figure 6.11a shows the cross section of a 90 nm node cell along the word-line direction; the cell has a width of 70 nm, and the nanocrystals are shown by EFTEM microscopy [37]. Figure 6.11b shows the write and erase threshold voltage distributions of the 4 Mb NOR device. The 90 nm node technology inte-

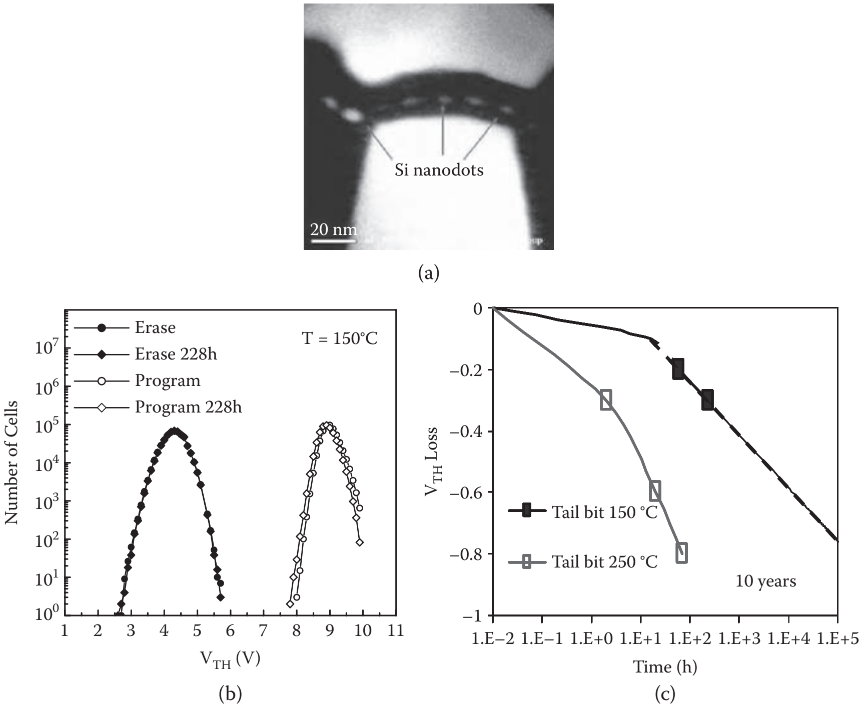


FIGURE 6.11 (a) Memory cell cross-section along the word-line direction of the 90 nm technology node in a 4 Mb Si-NC NOR Flash array. (b) Program-erase threshold voltage distributions (CHE program with 9 V between gate and substrate and 4.4 V for 10 μs to the drain; erase with 15 V for 10 ms). No correction algorithms are used (so the distribution widths can be further reduced). The figure shows program-erase distributions at 150°C of the programmed and erased distributions of a 512 kb sector. (c) Evolution of the programmed tail-bits with time at 150°C and 250°C, respectively; the erased tail-bit does not show significant changes in V_{TH} during all the experiment duration. The extrapolation at 10 years of the programmed tail-bit V_{TH} shows enough lifetime margin (higher than 1 V of spacing between erased and programmed tail-bit). (Reprinted with permission from, Gerardi, C., Ancarani, V., Portoghese, R., Giuffrida, S., Bileci, M., Bimbo, G., Brafa, O., Mello, D., Ammendola, G., Tripiciano, E., Puglisi, R., Lombardo, S.A., "Nanocrystal memory cell integration in a stand-alone 16-Mb nor flash device." on *IEEE Transactions on Electron Devices*, June 2007, Volume 54, Issue 6, pp. 1376–1383.)

gration includes shallow trench isolation, self-aligned silicided junctions, and three copper metallization levels.

In Figure 6.11b the programming and erase level distributions are recorded at 150°C and monitored as a function of time to assess ability of data retention. Measurements performed by analyzing the high-temperature V_{TH} evolution of the tail bits (Figure 6.11c) show that, extrapolating to 10 years, the useful working window is still higher than 1 V, which ensures a quite good read margin [37].

The NCM shows a very high robustness against disturbs as an example Figure 6.12a shows the high resistance of an erased cell to the drain disturb during the programming of an adjacent cell sharing the same bit-line. We have programmed one cell with pulses of 1 μs, $Vd = 4.5$ V and $Vg = 9$ V. The erase V_{TH} of the adjacent cell in the same bit-line does not shift significantly up to 10^6 pulses.

The resistance of the memory cell against degradation, caused by extensive program–erase cycling (endurance), is typically an issue that must be addressed with care in modern scaled flash technologies. Typically, NCMs show very good robustness because of their intrinsic immunity to stress-induced charge leakage. Excellent program–erase endurance (up to 10^5 cycles) is shown by the memory cells written by CHE (Figure 6.12b) or by FN (Figure 6.12c).

The program–erase window is well open and does not exhibit significant shifts due to parasitic charge trapping in the dielectrics. These results demonstrate that the NCM working window has enough margin even after long (100 kcycles) program–erase cycling.

So far, NCMs have proven to be a viable technology for scaled memory devices, showing good robustness and lower voltage operations. The high robustness, lower required power, lower number of masks to be added to a conventional CMOS logic process, and ease of technology make NCMs very appealing for integration into a host logic device to form an embedded memory system.

6.4 RADIATION EFFECTS ON NONVOLATILE MEMORIES

It is well known that exposure to ionizing radiation degrades the electrical properties of solid-state electronics. The effects of, for example, γ-rays, X-rays, electrons, protons, neutrons, and heavy ions on MOS device characteristics has been the topic of several works, books, and review articles over the past two decades [38-41]. For instance, the reliability of electronic systems employed in space and satellite applications can be severely endangered by charged particles coming from the sun or by galactic cosmic rays drifting into solar system (see, e.g., [42] and references cited therein). We may mention the events of November 2003, when on the sun one of the largest solar flares ever recorded occurred, knocking down satellites and cellular communications [43]. The high-energy charged particles from space can also enter the atmosphere and generate a cascade of secondary particles, giving rise to an appreciable neutron flux at ground level that seriously threatens electronic circuits.

Generally, when considering the ionizing radiation effects in electronic devices, we distinguish between two main classes of phenomena: (1) the total ionizing dose (TID) effects; and (2) the single-event effects (SEEs). TID effects come from the radiation-induced electron-hole pair ionization and the progressive charge buildup

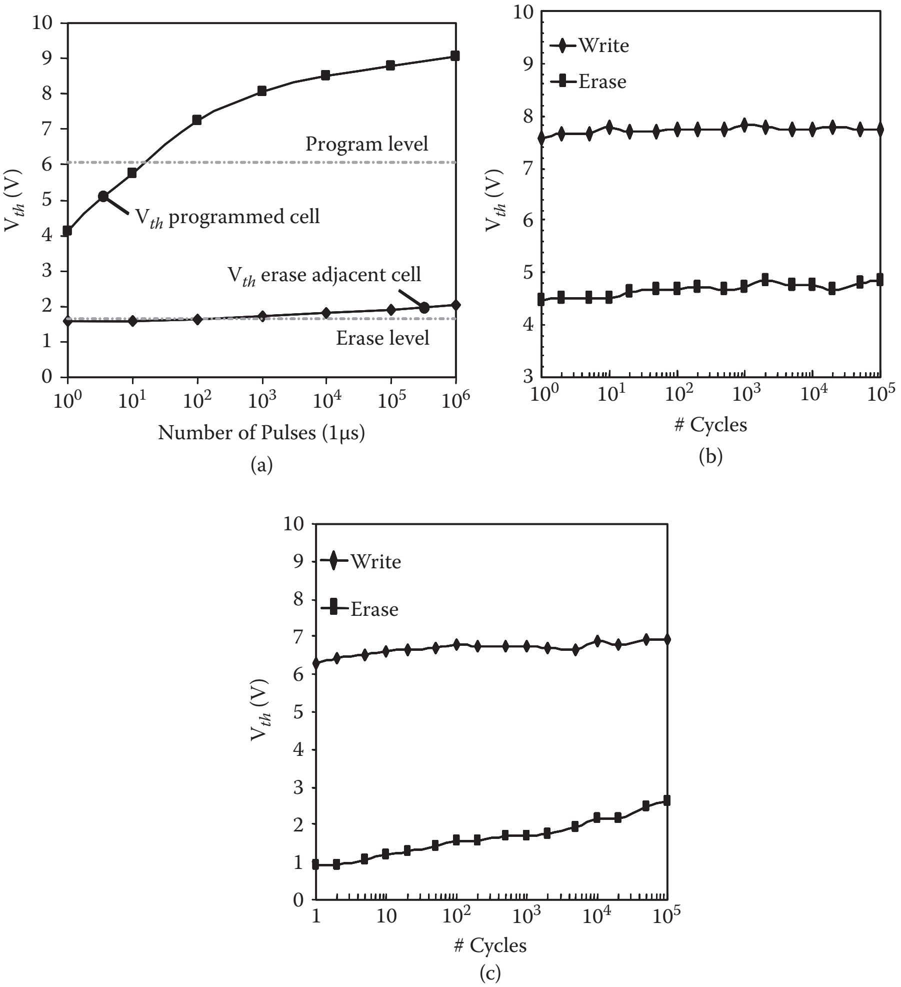


FIGURE 6.12 (a) Program disturb on an erased bit-cell during CHE programming of the adjacent cell sharing the same bit-line. Each pulse is of 1us with Vg = 9 V, and Vd = 4.5 V. (b) (a) Endurance characteristic of a Si-nc bitcell by using CHE programming (Vg = 8 V, Vb = 1.2 V, Vd = 4.4 V, 1 μs) and FN tunneling erase (Vg = –15 V, 10 ms). (c) Endurance characteristics of a Si-nc bitcell by using FN programming (Vg = +16 V, 1 ms) and FN tunneling erase (Vg = –16 V, 10 ms. (Reprinted with permission from, Gerardi, C., Ancarani, V., Portoghese, R., Giuffrida, S., Bileci, M., Bimbo, G., Brafa, O., Mello, D., Ammendola, G., Tripiciano, E., Puglisi, R., Lombardo, S.A., "Nanocrystal memory cell integration in a stand-alone 16-Mb nor flash device." *IEEE Transactions on Electron Devices*, June 2007, Volume 54, Issue 6, pp. 1376–1383.)

in the device. Several radiation sources can produce TID (e.g., X-rays, γ-rays emitted by ^{60}Co, high-energy electrons, protons). Due to the very low density of radiation-induced electron-hole pairs, immediately after their generation the two carriers tend to recombine in times as short as a few picoseconds [44], according to the geminative recombination model [45]—that is, each carrier recombines with its own partner. Carrier ionization and recombination can produce trapped charge, bulk defects, or interface defects. The energy released by the recombination process might generate defects in the bulk oxide or at the semiconductor/dielectric interfaces [38,39,45,46]. Such traps are responsible for the majority of degradation mechanisms in MOS devices, such as the radiation-induced leakage current (RILC) [47,48], which consists of a parasitic leakage current due to a tunneling process assisted by the radiation-induced neutral traps in the oxide. When traps are close to the silicon/oxide interface, they might affect the subthreshold region of a MOSFET, increasing the subthreshold swing and the threshold voltage. The recombination process is influenced by the electric field, which tends to separate the pair, avoiding the recombination and also producing trapped charge [38,39], which in turn changes the MOSFET threshold voltage.

Single-event effects depend on the energy released by a single and localized high-energy particle, typically a heavy ion, passing through the sensible area of a circuit node or a device. For each ion passing through a material, the amount of energy lost in ionization processes per unit length is defined as linear energy transfer (LET), measured in MeVcm^2/mg, which is directly proportional to the number of electron-hole pairs generated per unit length. From the physical point of view, the energy of the impinging particle is transferred to the lattice, generating electron-hole pairs, photons, and phonons. The recombination process is far more complicated than for TID effects. In SiO_2, the recombination follows the columnar recombination model [49,50], which can produce clusters of defects in the dielectric, possibly creating localized leakage paths or localized damaged regions. From the electrical viewpoint, the SEEs range over a number of different phenomena, for example, single-event gate rupture (SEGR) [51], radiation soft breakdown [52-55], single-event transient (SET) [56], single-event upset (SEU) [57], and single-event latchup (SEL) [58].

A comprehensive description of the interaction between matter and radiation is out of the scope of this chapter. The interested reader may refer to a number of books and journal papers in the literature (see, e.g., [38-41] and the references cited therein).

6.4.1 Radiation Effects on NVM: An Overview

As mentioned in this chapter's introduction, the majority of flash memories are based on the FG MOSFET and can be subject to both SEE and TID effects, which interact with the dielectric layers and may corrupt the stored information. In addition to all the radiation effects observed in the conventional MOS devices, flash NVMs present some peculiar radiation effects, due to the presence of the storage medium, being either the conventional floating-gate or the nanocrystal layer. Among them, the most important issues are the prompt charge loss after irradiation and the long-term data retention degradation, which may hamper the correct functioning of a flash cell also at low doses. At higher radiation doses, the permanent radiation effects on the electrical characteristics also become a concern, because they can produce a permanent shift of the cell threshold voltage similar to the conventional MOS devices. This section discusses the most important radiation effects on NVMs, mostly focusing on the data retention and the prompt charge loss.

6.4.1.1 Prompt Charge Loss Due to TID

Figure 6.13 summarizes the effects of the charge loss due to the TID taken from results reported in the literature on irradiation of floating-gate memories with ^{60}Co γ-rays [59]. In particular, Figure 6.13a shows the threshold voltage probability as a Weibull plot for FG arrays programmed in the "0" and "1" state before and after different γ-rays TID levels. During irradiation, the FGs are progressively losing their charges. Consequently, the threshold voltage of all FGs programmed at high V_{TH} value uniformly moves toward lower V_{TH} due to the loss of negative charge. The low V_{TH} distribution features the opposite behavior due to the loss of positive charges. The progressive closure of the programming windows as a function of the TID is shown for the same device in Figure 6.13b [59]. Similar results have been reported for other TID sources, such as X-rays and protons [60,61].

Two important mechanisms contribute to the charge loss:

1. Neutralization of the FG stored charge due to the electron-hole pairs generated by ionizing radiation in the tunnel oxide or IPD, which are in turn injected through the oxide. Irradiation generates electron-hole pairs in all oxides surrounding the FG [38,39]. Part of these carriers suddenly recombines, depending on the oxide electric field [38,39]; due to their high mobility, the electrons that survive the prompt recombination quickly thermalize and are swept away from the oxide [62]. Instead, the holes slowly

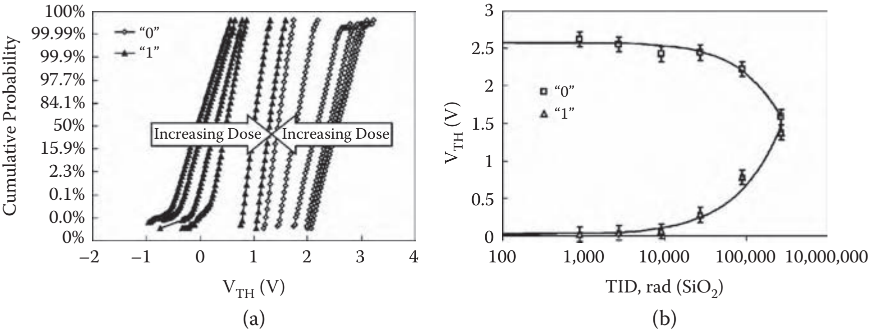


FIGURE 6.13 (a) Cumulative threshold voltage distributions of floating-gate memories programmed at "0" (squares) and "1" (triangles), for different total ionizing doses: fresh, 9 krad(SiO_2), 27 krad(SiO_2), 90 krad(SiO_2), 270 krad(SiO_2), 900 krad(SiO_2) [From: G. Cellere, et al. *IEEE – Trans. Nucl.* Sci. (51) 2004–© 2004]. (b) Evolution of the average threshold voltage of floating-gate memory cells programmed at "0" (squares) and "1" (triangles), as a function of the total ionizing dose [From: G. Cellere, et al. *IEEE – Trans. Nucl. Sci.* (51) 2004–© 2004]

move across the oxide by drift or diffusion and may be trapped in the bulk oxide or silicon/oxide interface or may move toward the FG [61]. The fraction of holes reaching the FG recombines part of the stored charge.

2. Photoemission. The incoming radiation can directly interact with electron-holes stored in the floating-gate, transferring enough energy to the carrier, which may jump over the oxide barrier. In addition, photoemission may occur also in the substrate and in the control gate. Part of the electron-holes generated by the ionizing radiation in the substrate or control gate can jump the barrier and reach the FG, neutralizing the stored charge. The balance among substrate photoemission, control gate photoemission, and FG photoemission depends on the applied electric field and the polarity of the charge stored in the FG.

6.4.1.2 Prompt Charge Loss Due to SEE

SEEs are generally produced by heavy ions, which produce a dense electron-hole pair track around their hit positions. Figure 6.14a summarizes the effect of heavy-ion irradiation on the threshold voltage distribution of an FG memory (taken from [63]). The chip was irradiated with 2 × 10^7 iodine ions/cm^2. Before irradiation, the V_{TH} distribution has the expected Gaussian shape. After irradiation, the V_{TH} distribution exhibits a secondary peak around 6 V due to the cells that experienced a charge loss after the ion hit. The amount of charge loss depends on the programming status of the cell, the impinging ion LET, and the technology node [64-66]. For instance, Figure 6.14b shows the number of errors as a function of ion fluence for different

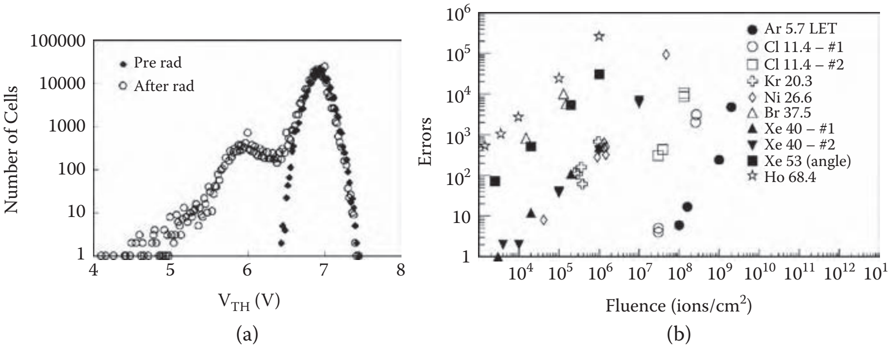


FIGURE 6.14 (a) Threshold voltage density distribution of floating-gate memory cells before (diamonds, filled) and after 2 × 10^7 iodine ions/cm^2. Very large threshold voltage variation (as high as 3V) are observed. [From: G. Cellere, et al. *IEEE – Trans. Nucl. Sci.* (48) 2001–© 2001]; (b) Number of errors as a function of the ion fluence and LET coefficient. For high-LET ions almost 100% of hit cells fail. (Reprinted with permission from, Cellere, G., Pellati, P., Chimenton, A. ,Wyss, J., Modelli, A., Larcher, L., Paccagnella, A., "Radiation effects on floating-gate memory cells," *IEEE Transactions on Nuclear Science*, Dec. 2001, Volume 48, Issue 6, pp. 2222–2228 Figure 6.14B reprinted with permission from, Guertin, S. M., Nguyen, D. M., Patterson, J. D., "Microdose induced data loss on floating-gate memories," *IEEE Trans.Nuclear Science*, Volume 53, Issue 6, 2006, pp. 3518–3524.)

LET values (from [64]). For high-LET ion irradiation, almost 100% of the hit cells fail. As Moore's Law proceeds, the shrinking transistor sizes, featuring smaller and smaller FGs, make them more and more sensitive to the impact of a single ion. A single ion strike may produce even multiple bit flips, as soon as the cell size and spacing become smaller than the ion track size [66].

Even though the formation of the secondary peak is not unexpected, because of the high energy released by the ion hit, its physical origin is a source of some controversy. In principle, the appearance of the secondary peak in the V_{TH} distribution can be explained similarly to the rigid shift of the V_{TH} after TID, with the additional consideration that the heavy ion releases a huge quantity of energy in a very small volume (the *microdose effect*) and only in a small percentage of cells. Neutralization and photoemission may occur locally, leading to the complete or the partial discharge of the hit FGs. Based on the columnar recombination model, several thousands of electron-hole pairs should be generated in the tunnel oxide by a single ion, depending on its LET coefficient, but only a small fraction of these pairs (on the order of some tens) survives the prompt recombination [50]. The surviving electrons are quickly swept toward the substrate, thanks to their high mobility [62], whereas holes slowly move toward the FG, where they recombine with part of the stored negative charge. The same could happen in the IPD. Nonetheless, this estimation is not in agreement with the number of charges, which are stored in the floating-gate. In fact, assuming a floating-gate capacitance of 1 fF, a 1 V shift of the cell threshold voltage corresponds to a charge loss of more than 6,000 electrons.

A model that tentatively explains the prompt charge loss is based on the formation of a transient conductive path around the ion hit position [67]. Following this model, the dense track of electron-hole pairs forms a conductive path, which shorts the floating-gate with the substrate. It is assumed in [67] that the resistance of such a path depends on the oxide thickness (i.e., the length of the path) and on the ion LET coefficient (i.e., the amount of ionized charges). If the floating-gate cell is considered as equivalent to the series of two capacitors—one between FG and substrate, source, and drain junctions and one between FG and control gate—the formation of a resistive path across the tunnel oxide can discharge the floating-gate. The amount of charge loss depends on the time constant of the equivalent resistor-capacitor circuit and the time needed to shut down the ion-strike-induced conductive path. Such time has been estimated based on considerations on the times needed for carrier recombination and on electron mobility [67] in about 10 fs, in agreement with several published works on related topics [50,68-70]. Even though this phenomenological model can fit the experimental data, recent works questioned the validity of the transient conductive path [71] and the lack of physical details of the mechanisms governing the path resistance and oxide barrier lowering [65].

A more consistent model points to photoemission of ion-induced hot electron-hole pairs from the substrate and the polysilicon control gate, which are in turn injected across the tunnel oxide and IPD though the FG, neutralizing a fraction of the FG charge. The results of this model agree with recent simulation results by Dodd [72], which reported that the ion track size in silicon is much larger than in the oxide and is on the order of several tens of nanometers.

6.4.1.3 Long-Term Retention Capability

One of the most important aspects of a nonvolatile memory is its retention capability, which is typically at least 10 years [73]. Ionizing radiation can severely compromise the retention of FG memories. The irradiation effects may vary, depending on the type of irradiation: heavy ion or TID.

Several works in the literature [74-77] reported retention experiments carried out on FG memories irradiated with heavy ions. All these works highlight a very poor retention on cells hit by at least one ion, whereas the retention of nonhit cells was unchanged. For instance, Figure 6.15a shows a typical V_{TH} distribution as a Weibull plot of an irradiated flash memory array (from [76]), reprogrammed after irradiation and measured immediately after programming, after 1.5 hours, 48 hours, and 164 hours. Even though immediately after programming the whole cell distribution resembles that of a fresh (not irradiated) device, just after 1.5 hours a large tail appears, indicating a slow charge loss only from the hit cells. Such a tail is a signature of the formation of a permanent leakage path across the dielectrics. Further characterizations, performed at longer times after the reprogramming, showed a broadening of the tail, indicating that there were other FG cells, which had smaller leakage currents.

A few works in the literature [61,77] showed also the TID effects on FG cell arrays. Some results are shown in Figure 6.15b (data taken from [77]). The retention

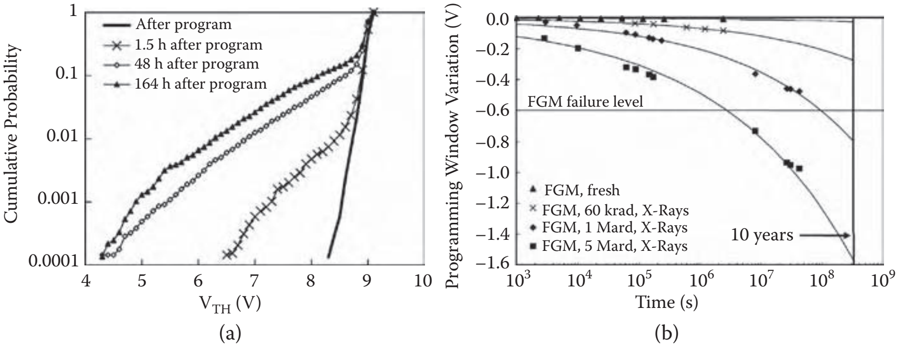


FIGURE 6.15 (a) Retention test performed on floating-gate cells, which were identified as hit by a single iodine ion. Large threshold voltage variations as high as 5 V indicate that the retention can be severely compromised by heavy ion hits. (From G. Cellere et al., *IEEE Transactions on Nuclear Science,* 52, 2005. With permission.) (b) Retention test performed on floating-gate memory arrays irradiated at different doses with X-rays. (Data taken from D.N. Nguyen, C.I. Lee, and A.H. Johnston, of the 1998 IEEE *Proceedings of the Radiation Effects Data Workshop*, 100–103. With permission.) The failure level is taken as the 20% of charge loss closure (R. Bez et al., *Proceedings of the IEEE*, vol. 91, 489–502, 2003). The evolutions show that even after only 60 krad (SiO2) the retention is strongly modified. The device is expected to fail before 10-years if subjected to 1 Mrad (SiO2). The retention fails after four months (about 107 seconds) if the device is irradiated with 5 Mrad (SiO2). Reprinted with permission from Cellere, G., Larcher, L., Paccagnella, A., Visconti, A., Bonanomi, M., “Radiation induced leakage current in floating-gate memory cells.” *IEEE Transactions on Nuclear Science*, Volume 52 Issue 6, pp. 2144 – 2152.)

of an irradiated FG cell array is appreciably modified with respect to the nonirradiated devices, even after only 60 krad(SiO_2), while variations as large as 1 V are expected after 10 years if it is irradiated with 1 Mrad(SiO_2). From the application point of view, TID effects on the retention appear somewhat less concerning, at least at irradiation doses below 100 krad(SiO_2). In fact, at levels in the 100 krad(SiO_2) range, the flash memory chip starts failing due to excessive peripheral circuitry degradation, which has been identified as the weak point of a commercial device [78,79]. Still, employing radiation hardening techniques on the peripheral circuitry can bring this failure level to higher values, and data retention might become a more concerning issue even after TID.

The progressive cell threshold voltage variation on irradiated devices derives from the formation of oxide neutral traps (after TID) or cluster of defects (after heavy-ion irradiation) in the tunnel oxide. Such neutral traps are responsible for the onset of the well-known RILC [47,48] and radiation soft breakdown [52-55], which slowly discharge the FGs. It is worth noting that even a very small RILC value (below the aA range) should be enough to discharge the FG in approximately 1 hour. The physical nature of the oxide traps that lead to the RILC was investigated in several works, and was similar to the traps responsible for SILC, which affects the thin gate oxide after electrical stress. The interested reader may refer to the numerous studies in the literature for more details (see, e.g., [80-83] and the references cited therein).

6.4.2 Radiation Effects on Nanocrystal Memory Cells

To overcome the scaling limits of FG flash memories, some solutions have been explored, like phase change memories, ferroelectric memories, and magnetic memories. However, these solutions require the employment of uncommon materials and need nonstandard production processes. Nanocrystal memory represents a feasible alternative solution. This memory does not need the employment of any uncommon material, and the production process is simple. The structure of an NCM is very similar to that of an FG memory and may suffer, in principle, the same radiation effects as the FG cell. However, the strength of the NCM approach lies just in the fact that the electric charge is stored within a layer of discrete nano-dots rather than in a monolithic FG. A single defect or a cluster of defects in the tunnel oxide is expected to discharge only the few neighboring nanocrystals. Nanocrystals located far from those defects cannot interact with the defects, and the cell V_{TH} is expected to be almost unchanged. The advantages of NCM technology over FG technology have already been discussed. Thanks to their structure, it is reasonable also to expect high robustness of NCMs against ionizing radiation compared with FG memories, in particular concerning SEEs. Heavy ions can produce defects in the tunnel oxide of an FG cell through which leakage currents can flow, discharging the FG. NCMs exhibit a higher robustness to RILC thanks to the discreteness of the storing nodes.

In the last two years, some contributions [71,76,84-89] have investigated the radiation tolerance of NCMs in terms of both TID and SEE. In [84] the authors focused on the total dose effect on the cell electrical characteristics. However, in that work nanocrystals were fabricated by low-energy silicon implantation, which is not the best candidate technology for nanocrystal memory mass production because of the large dispersion in both nanocrystal size and depth. The first work on a prototype NCM intended to evaluate different cell designs and processes [85] presented results on nanocrystal memories, favorably comparing with results on currently available commercial flash technology and indicating a promising future for NCM even for space and military applications. In [77,87] more comprehensive studies have been carried out, including also the retention time and the permanent effects on the memory cells.

Some efforts also have been made for studying SEEs due to heavy-ion irradiation on NCM arrays considering important issues, such as data retention characteristics and endurance [71,85,86,88]. Oldham et al. [85] reported better resistance of NCMs under heavy ion irradiation compared with FG memories, without any degradation of the NCM cell retention characteristic. In [86,88] the authors studied in detail the effects of heavy-ion irradiation on addressable arrays of NCM cells, mainly focusing on the prompt charge loss from a single cell during irradiation and the retention properties of the irradiated cells.

The following sections summarize the most relevant achievements in NCM radiation tolerance, separating the discussion between SEE and TID effects.

6.4.2.1 Total Ionizing Dose Effects in Nanocrystal Memory

6.4.2.1.1 Prompt Charge Loss

As previously discussed, TID can produce both prompt charge loss and degradation of the retention properties in NVM. Being the NCM based on the same structure than FG memories, they might suffer, in principle, from both these phenomena. Due to the prompt charge loss, the NCM V_{TH} moves toward its intrinsic value, corresponding to zero net charge in the nanocrystals. The phenomena responsible for the threshold voltage variation are the same as in FG memories: the *neutralization* and the *photoemission* of the stored charge. However, both contributions are somewhat reduced in NCM with respect to FG memories. In particular, the reduction of the tunnel oxide thickness with respect to FG memories results in a smaller amount of radiation-generated charge in the oxide that can neutralize the stored charge. Photoemission current, on the other hand, has a more minor impact in the NCM prompt charge loss than in FG memories, because the area covered by nanocrystals is only a fraction of the total gate area. Hence, most of the energy is released in regions where the nanocrystals are not present. This is also in agreement with the observation that flash memory cells with smaller FG area lose a smaller quantity of charge during irradiation due to the reduction of the photoemission contribution [90]. All these considerations lead to the conclusion that NCMs are expected to be less sensitive to the photoemission and charge neutralization than FG memories. Experimental results in the literature confirm this expectation. Figure 6.16 shows the threshold voltage evolutions of negatively charged NCM and FG cells fabricated with the same technology as a function of the irradiation doses. The threshold voltage was measured just after each irradiation dose, without reprogramming the cells. A direct comparison between Figures 6.16a and 6.16b highlights the faster programming window closure of FG cells with respect to NCM. For instance, after 200 krad(SiO_2), NCM and FG cells feature the same residual programming window, even though the

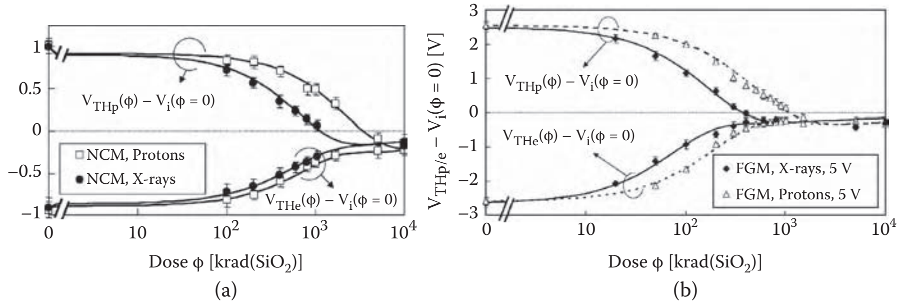


FIGURE 6.16 (a) Threshold voltages of nanocrystal memory arrays irradiated with protons and X-rays at different doses. The lines represent the model presented in N. Wrachien et al., *IEEE Transactions on Nuclear Sci*ence, 55, 3000–3008, 2008. (b) Threshold voltages of floating-gate memory arrays irradiated with X-rays and protons at different doses. The lines represent the model presented in Wrachien., N. (Reprinted with permission from, Wrachien, N., Cester, A., Portoghese, R., Gerardi, C., "Investigation of proton and x-ray irradiation effects on nanocrystal and floating-gate memory cell arrays." *IEEE Transactions on Nuclear Science*, Volume 55 Issue 6 pp. 3000 – 3008.)

FG had an initial programming window that was more than two times larger with respect to NCM. Considering the percentage of charge loss, the FG cells lost about 75% of their stored charge, whereas the NCM lost less than 40% of the stored charge. For comparison, the required dose to achieve a 40% charge loss in FG cells is just 50 krad(SiO_2). This means that the NCM approach increases the radiation tolerance (from the charge loss viewpoint) at least by a factor of three.

Several models have been proposed to predict and quantify the prompt charge loss [77,91,92]. A first-order model can provide good approximation in a useful range of irradiation doses: by exposing the cell to an infinitesimal irradiation dose, $d\varphi$, the amount of charge loss, dQ, is proportional to the stored charge, Q, and to the infinitesimal irradiation dose, $d\varphi$. This results in a simple exponential law, which can describe the stored charge at various doses.

$$Q(\phi) - Q(0) = Q(0)[1 - \exp(-C \cdot \phi)] \tag{6.3}$$

The amount of stored charge is proportional to the difference of the threshold voltage of the cell and its intrinsic value (i.e., the threshold voltage of the neutral cell); hence, the threshold voltage evolution can be predicted. Of course, the intrinsic threshold voltage value is not constant during irradiation because of permanent irradiation effects such as the subthreshold degradation and the positive charge trapping in the dielectrics. These have to be modeled separately. The fitting lines in Figure 6.16 represent the complete model (including the permanent degradation). The interested reader may refer to [77] for more details.

Some efforts have been also made to investigate the effects of different radiation sources, such as X-rays and protons. It has been demonstrated that X-rays have much stronger effects than proton irradiation in both NCM and FG cells, despite the same nominal SiO_2 dose. This derives from two main contributions. First, X-rays feature different dose rates in Si and SiO_2, whereas proton irradiation features almost the same dose rate in the two materials. In particular, the 10-keV X-rays used in [77] feature a dose rate in Si that is double with respect to the dose rate in SiO_2. In other words, for a given photon fluence, the deposited energy in silicon is almost twice the energy adsorbed by SiO_2. Consequently, for the same SiO_2 dose, the X-ray irradiation deposited much more energy in the silicon storage sites (the nanocrystals or the floating-gate) with respect to proton irradiation, leading to an enhanced photoemission during the X-ray irradiation. The second contribution is the *dose-enhancement effects*, which derive from the presence of high-Z metals in silicides, and to the photoelectric adsorption, especially at the Si/SiO_2 interfaces [93], which further enhances the charge loss from nanocrystals/FG.

6.4.2.1.2 Data Retention

TID irradiation typically induces on data retention smaller effects than heavy-ion irradiation. Likely for this reason, few works in the literature addressed this point after TID [77,89], reporting some data on the retention of NCM cell arrays, and even fewer works provide a direct comparison between the NCM and the FG memories after TID irradiation.

Figures 6.17a and 6.17b, taken from [77] and [89], respectively, show the retention behavior of two NCM technologies, manufactured with different processes low pressure chemical vapor deposition (LPCVD and silicon ion implantation, respectively). The results show that very high irradiation doses are required to induce appreciable variations on the retention characteristics.

Figure 6.17a also shows the direct comparison of NCMs and FG memories fabricated with the same technology, emphasizing the much stronger robustness of NCM with respect FG memories. After 10 Mrad(SiO_2) the retention kinetics do not exhibit appreciable changes with respect to the kinetics of the fresh device, and the device is not expected to fail in 10 years. Conversely, the same irradiation dose severely compromised the retention of the FG array, which exhibited more than 20% of window closure in nine months. Furthermore, in FG memory the retention is compromised even after a 1 Mrad(SiO_2) proton irradiation, and the device is expected to fail before the 10-year limit: in other words, the NCM approach brings a significant improvement by a factor of 10 in the retention robustness against TID.

The improvement in radiation tolerance of NCM derives mostly from two factors. First, the discrete nature of the storage sites allows some sort of redundancy: if one or a few nanocrystals are discharged for a leakage path, the others remain charged without any charge redistribution (at least assuming negligible lateral tunneling). The overall cell threshold voltage is negligibly affected if the number of nanocrystals is large enough. In contrast, in an FG memory even a very tiny leakage path due to a single weak spot can entirely discharge the floating-gate. Second, an additional improvement comes from the limited coverage area of the nanocrystal. In fact, nanocrystals do not completely cover the whole channel area because they are isolated and separated from each other by the surrounding oxide. For instance, in the samples tested in Figure 6.17a the average nanocrystal mutual distance is 12 nm. Nanocrystals could be discharged only when defects are located underneath them.

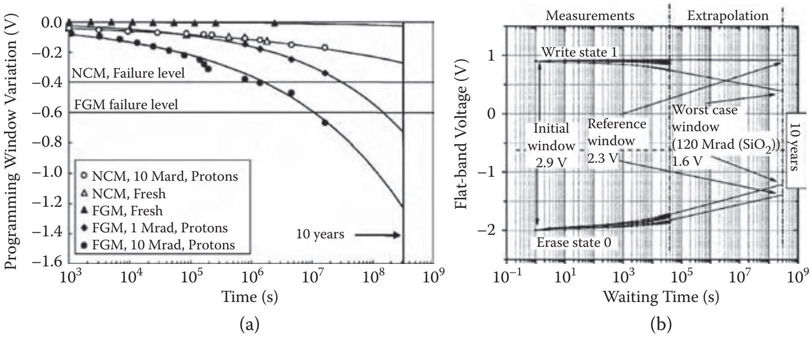


FIGURE 6.17 (a) Comparison between the retention tests performed on nanocrystal memories (manufactured by CVD) and floating-gate memories. The retention characteristics of nanocrystal memories is almost unaffected by proton irradiation (up to the dose of 10Mrad (SiO2), while it is strongly compromised on floating-gate memories after 1Mrad (SiO2) [From: N. Wrachien, et al. *IEEE – Transactions on Nuclear Sci*ence, (55) 2008 - © 2008]. The failure levels, corresponding to the 20% charge loss (as [10]) are indicated, and they show that nanorcrystal memories can sustain dose levels more than 10x higher than floating-gate memories, without failing. (b) Retention tests performed on nanocrystal memories manufactured using silicon ion implantation. Extrapolation shows a 0.6V programming window after 10 years for the device irradiated at 75Mrad (SiO2). (Reprinted with permission from, Wrachien, N., Cester, A., Portoghese, R., Gerardi, C., "Investigation of proton and x-ray irradiation effects on nanocrystal and floating-gate memory cell arrays." *IEEE Transactions on Nuclear Science,* Volume 55 Issue 6, pp. 3000 – 3008. Figure 6.17b: Reprinted with permission from, Verrelli, E., Tsoukalas, D., Kokkoris, M., Vlastou, R., Dimitrakis, P., Normand, P., "Proton radiation effects on nanocrystal non-volatile memories." *IEEE Transactions on Nuclear Science*. Volume 54, Issue 4, pp. 975 – 981.)

Instead, a defect far from any nanocrystals is expected not to significantly contribute in NC discharge.

6.4.2.2 Single-Event Effects in Nanocrystal Memory

The most impressive improvements of the NCM technology over the conventional FG memories come from the immunity to SEE. We previously cited that the most important issues related to SEE in NVM are the prompt charge loss after irradiation and the long-term data retention characteristics. The electron-hole pairs generated along the ion track are responsible for the former process. The onset of oxide leakage current is strictly correlated with the second issue, since it may produce an abnormal and premature charge loss even a short time after programming. NCM technology suppresses the majority of the mechanisms involved in prompt and long-term data corruption.

6.4.2.2.1 Prompt Charge Loss

In NCM, only a moderate charge loss occurs in a small percentage of irradiated cells, which consequently appears not to be a problem for this technology. An example is shown in Figure 6.18a. These data refer to a 16 Mbit array irradiated with 1.67 × 10^9 Cu ions/cm^2 (surface LET = 33.5 MeV × cm^2 × mg^{-1}, energy = 50 MeV) [71]. In that case, the ion beam was focused in a small portion of the array to avoid any damage

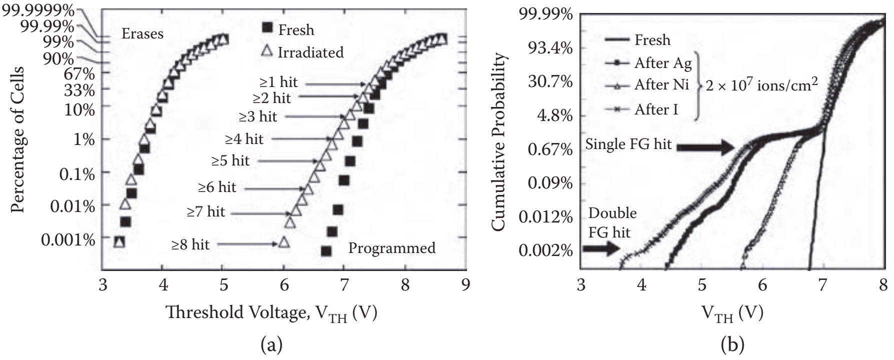


FIGURE 6.18 (a) Cumulative threshold voltage distributions of an NCM array, programmed with a checkerboard pattern before and after irradiation with 1.67 × 109 Cu ions/cm2 (LET = 33.5 MeV×cm2×mg–1). (From A. Cester et al., *IEEE Transactions on Nuclear Sci*ence, 55, 2008. With permission.) (b) Cumulative distributions of floating-gate memory cell array before and after irradiation with Ni, Ag, and I. Fluence = 2×107 ions/cm2 (corresponding to 0.8% of nominally hit FGs), LET = 29.3 MeV×cm2×mg–1 (Ni); 57.3 MeV×cm2×mg–1 (Ag); 64.2 MeV×cm2×mg–1 (I). The two arrows indicate the probability of a single and a double ion hit, using the model presented in A. Cester, et al., *IEEE Transactions on Nuclear Science*, 54, 2196–2203, 2007. (From L. Larcher et al., *IEEE Transactions on Nuclear Science,* 50, 2003. With permission.)

on the peripheral circuitry. With such ion fluence, 100% of the cells should have nominally received one ion hit within the irradiated area, but a remarkable number of multiple hits have likely occurred. The eight arrows in Figure 6.18a represent the cumulative probability that at least 1, 2, 3,…, 8 ions hit a single cell, calculated with the model proposed in [88], also taking into account that only a fraction of the sector is actually irradiated. Even though a small tail appears in the V_{TH} distribution, the charge stored in nanocrystals is not completely neutralized, despite several thousands of cells experiencing a multiple hit (see markers in Figure 6.18a). By comparing these data with those reported in the literature on the conventional floating-gate memories (see, e.g., Figure 6.18b taken from [74]), the tail of the threshold voltage distribution may be larger than 1 V after Ni irradiation at only 2 × 10^7 ions/cm^2 (LET = 29.3 MeV × cm^2 × mg^{-1}). The two arrows in Figure 6.18b have been added to indicate the probability of single and double hits on the same cell by using the model in [88]. Still, all hit FG cells showed a significant charge loss, whereas the NCM V_{TH} distribution tail is less than 1 V at the eight-hit probability.

The NCM technology strongly reduces the effects of all the major contributions to the prompt charge loss:

1. The formation of a transient conductive path (if any) along the ion track, which may partially or totally discharge the stored charge. In [71] it has been demonstrated that the discrete storage approach strongly reduces the effects of any potential transient conductive path, whose radius is as small

as about 10 nanometers, because only those nanocrystals within the ion track may be partially or completely discharged, and the remaining nanocrystal charge does not redistribute.

2. The carriers' photoemission from the substrate/polysilicon that in turn is injected through the FG/nanocrystals. These microdose effects may still occur in an NCM cell but are limited to the nanocrystals, which lie within a region as large as the ion track in silicon that may approach several tens of nanometers in radius [65,71,72].
3. The neutralization of the stored charge with the holes generated in the IPD and tunnel oxide by the ionizing radiation. The NCM approach improves the robustness to this microdose effect in two ways. First, only a small percentage of nanocrystals are involved in neutralization, and the charge does not redistribute in all the nanocrystals. Second, the reduction of the tunnel oxide thickness with respect to floating-gate flash memories (5 nm in the samples of Figure 6.18a vs. 10 nm typical of conventional NOR flash) results also in a smaller quantity of charge produced by the ion along its track in the oxide.
4. The electron photoemission from the FG/nanocrystals, which is reduced due to the small coverage area of the nanocrystal layer and the absence of charge redistribution.

The reports in the literature clearly show that after a single ion hit an NCM cell does not fail. Some interesting questions are as follows: How many nanocrystals are discharged by a single ion hit? How many ion hits are needed to change V_{TH} appreciably? How do the ion hit positions impact V_{TH} variation?

It is not easy to find a relation between the charge loss and the threshold voltage of a nanocrystal memory cell. In fact, whereas in a conventional flash cell the charge in the floating-gate is always uniformly distributed, in the case of an NCM cell the charges are stored in discrete locations. When some nanocrystals have lost part of the stored electrons, the remaining stored charges do not rearrange themselves. This gives rise to a local variation of the charge density over the nanocrystal layer, which induces a local variation of the potential at the silicon/oxide interface. Consequently the channel starts forming earlier in those regions, where the nanocrystals have lost some of their electrons. This means that the effective threshold voltage shift of an NCM cell is a function not only of the total charge lost but also of the position of the discharged nanocrystals. To achieve an appreciable drain current and to read a premature cell-turn-on, a large amount of nanocrystals must lose part of their charge so that a conductive path can be formed along the MOSFET channel from source to drain.

A charge loss model has been developed in [71] based on a statistical description of the ion hit events producing the threshold voltage shift. Such a model permitted us to analyze the average behavior of a large number of cells under heavy-ion irradiation, to extrapolate from these results the expected average behavior of a single cell, and to estimate some interesting parameters, such as the size of the ion hit impact region as well as the amount of charge loss per hit.

This model is summarized in Figure 6.19 and assumes that each ion hit partially discharges the nanocrystals within its track, projecting over the channel a spot, where

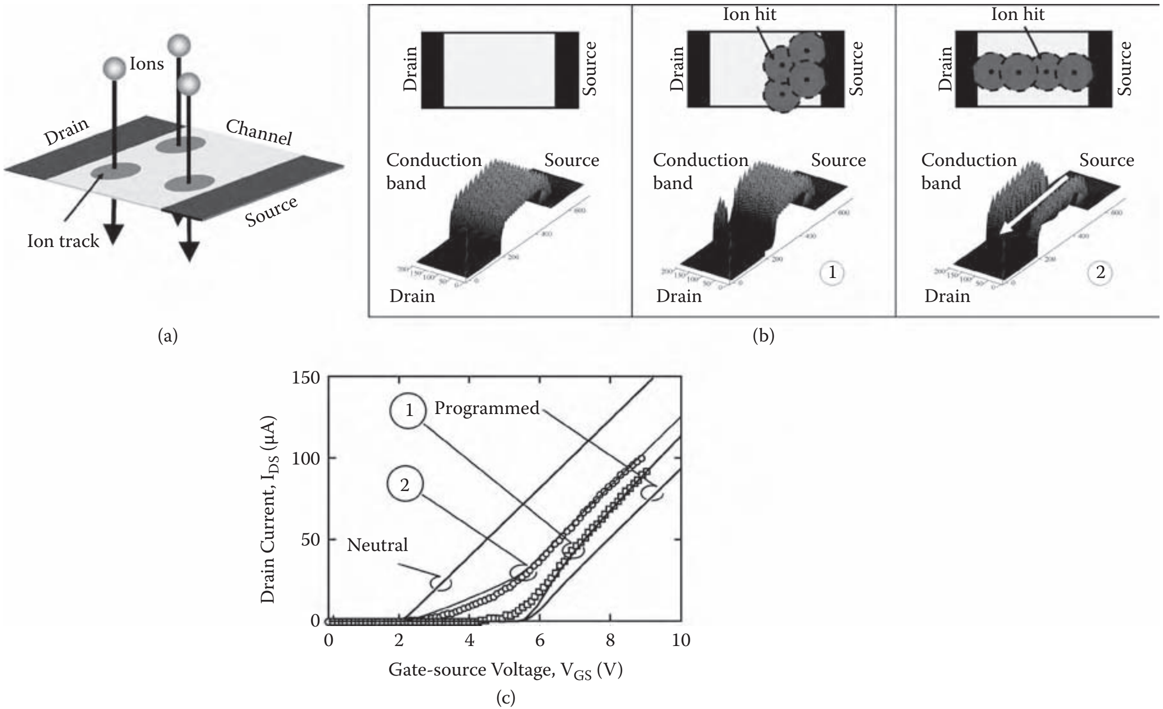


FIGURE 6.19 (a) Overview of the model proposed in A. Cester, et al., *IEEE Transactions on Nuclear Science*, 55, 2895–2903, 2008. Each ion hit generates a region of discharged nanocrystals with circular shape. The ion hits are randomly distributed over the gate area. (b) Comparison between a fresh device and two possible quadruple hit patterns (the effective track size is supposedly S = 100 nm). Hit pattern #1: four hits close to the source. Hit pattern #2: four hits perfectly aligned along the channel and simulated channel conductance of the hit pattern. (b) Simulated IDS-VGS curves of the neutral, programmed, and hit cells with pattern #1 and #2, respectively. (c) Two experimental examples of hit cells are also shown for comparison.

the threshold voltage is locally decreased (Figure 6.19a). Given the ion fluence, the number of hits per cell as well as the hit positions were statistically modeled by means of the Poisson processes.

For instance, Figure 6.19b shows three simulations of the potential barrier between source and drain for a programmed device before and after two opposite cases of quadruple hit patterns: (1) four hits localized near the source; and (2) four hits aligned along the channel. The resulting $I_{DS} - V_{GS}$ curves are shown in Figure 6.19c for a programmed device and the two cells that were hit by four ions. For reference, in the same plot there is also the simulation of a device where the nanocrystals are all empty (neutral).

The way the ion hits modify the cell $I_{DS} - V_{GS}$ is very different and strongly depends on the ion hit positions. In the case of ion hit pattern 2, a percolation path exists between source and drain and the cell turns on earlier than the fresh programmed cell. Instead, in pattern 1 the ion hit positions are very close to each other and are not able to generate any percolation path. The cell V_{TH} slightly reduces but is less affected than in case 2. Such difference of behavior cannot occur on conventional floating-gate cells, because the V_{TH} variation mostly depends on the charges lost from the FG. In fact, after each ion hit the residual floating-gate charge uniformly redistributes, independent of the hit positions. The interested reader may refer to [71] for more details about this statistical model.

By means of this model, an ion track size of 85 nm (in diameter) for 50 MeV Cu ions has been estimated. These results are comparable with the size of the interface physically damaged region observed in MOSFETs after heavy-ion irradiation [94]. Such results also permit some further insight on what happens inside the ion track and what the charge loss mechanisms are. In fact, the 85 nm wide ion track size well correlates with the results reported by Dodd [72] and by Butt et al. [65]. In [72] Dodd proposed an elaborated simulation, which permitted him to calculate the ion track size in silicon as a function of the ion LET and energy. In [65] Butt et al. simulated the charge loss in FG memories by photoemission from the substrate/polysilicon, evaluating also the number of electrons per hit lost by the FG as a function of the LET coefficient. The good correlation between the data reported in [65,71,72] suggest that the ionized electron-hole pairs in the polysilicon gate and in the substrate, which are in turn injected through the oxide, are most likely responsible for the neutralization of the charge stored in the nanocrystals/FG.

6.4.2.2.2 Data Retention

Moving from a floating-gate MOSFET typical of contemporary flash memories, with a relatively thick tunnel oxide (8 to 10 nm), to the novel nanocrystal technology, with thinner gate oxide (4 to 5 nm), the oxide leakage currents should become an even bigger issue, at least in principle. In fact, numerous studies have demonstrated that the radiation-induced oxide leakage currents quickly increase as the oxide thickness is reduced below 6 nm [47,48,81]. These currents are either radiation-induced leakage current [47] due to the formation of single-oxide neutral defects [48] or radiation soft breakdown [52-55] due to the formation of clusters of defects allowing a large current flow. Some studies (see, e.g., [86]) reported a steady-state leakage current after I ion irradiation of NCM as large as hundreds of picoamps, which reveals a similarity

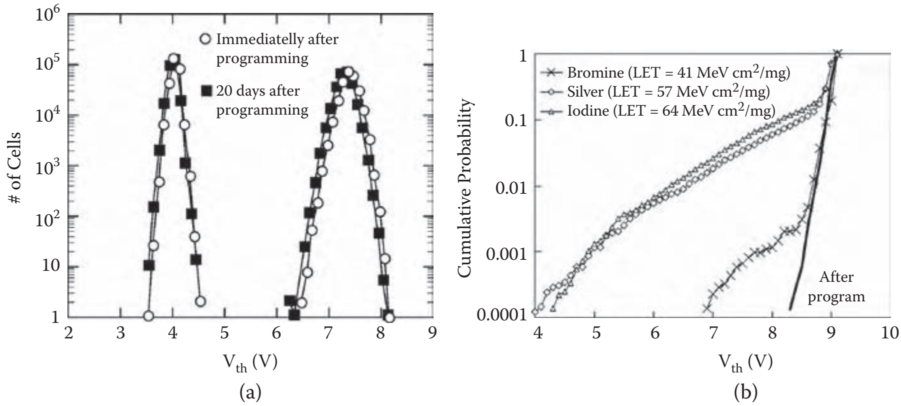


FIGURE 6.20 (a) V_{TH} distributions of a NCM array, irradiated with $5{\times}10^8$ Br ions/cm^2 and reprogrammed with the checkerboard pattern (open symbols). The filled symbols are the V_{TH} distribution measured after 20-days retention experiment. (Reprinted with permission from Cester, A., Wrachien, N., Gasperin, A., Paccagnella, A., Portoghese, R., Gerardi, C., "Radiation tolerance of nanocrystal-based flash memory arrays against heavy ion irradiation." *IEEE Transactions on Nuclear Science,* 2005, Volume 54 Issue 6, pp. 2196 – 2203.) (b) Comparison between the threshold voltage distributions of floating-gate memory cells hit by a single ion (Bromine, Silver, Iodine) measured just after program, and 164h after program. Large threshold voltage variations indicate that heavy ion irradiation severely compromises the retention of floating-gate cells. (Reprinted with permission Cellere, G., Larcher, L., Paccagnella, A., Visconti, A.,Bonamomi, M., "Radiation induced leakage current in floating-gate memory cells." *IEEE Transactions on Nuclear Science*, Volume 52 Issue 6, pp.2144-2152.)

with that shown by Candelori et al. in [95]. This oxide leakage was attributed to a multitrap-assisted conduction through a defect cluster generated along overlapping ion tracks across the overall dielectric stack from the control gate to the substrate. Even though this leakage current can easily discharge the FG capacitance of a conventional flash cell in times as short as a few tens of microseconds, in NCMs no tail of failing cells has been observed due to the discharging of the nanocrystal layer. The good retention of NCM is shown in Figure 6.20a. Twenty days after programming, no tail is observed in NCM, indicating that no critical charge loss occurs in the irradiated cells upon retention test. In contrast, a tail larger than 1 V has been reported in FG memories irradiated with Br even after only seven days [76]. Again, the improved robustness of NCM to heavy-ion irradiation derives from the discrete storage effect, limiting the impact of the radiation-induced oxide leakage currents. In fact, an oxide trap or a defect cluster, generated by the impinging ions and located close to a storage node, should discharge only one or a few nanocrystals in proximity to the cluster of traps.

6.4.3 Radiation Tolerance of Nanocrystal Memory versus Floating-Gate Memories

Several reports demonstrated that the nanocrystal approach brings a significant improvement in all fields. The most impressive improvement is the immunity to leakage currents up to 10 Mrad(SiO_2) total ionizing dose irradiation reported in [77] and shown in Figure 6.17. In fact, after 10 Mrad(SiO_2) proton irradiation, NCMs show the same retention characteristics as a fresh NCM. The retention of flash memories irradiated with protons at only 1 Mrad(SiO_2) is even worse than NCM after 10 Mrad(SiO_2) irradiation. If we take as typical failure criteria the 20% reduction of the programming window [10], from the reports in [77], nanocrystal memories irradiated at 10 Mrad(SiO_2) with protons are expected to remain within this specification even after 10 years, while FG memories are widely out of the specifications for doses as low as 1 Mrad(SiO_2). This indicates that *the improvement factor of NCM versus FGM is greater than 10*, even though the NCM tunnel oxide thickness is approximately half that in flash.

Another very significant improvement is the charge loss after TID. A direct comparison between FG and NCM approaches reveals that the nanocrystal technology increases the charge loss robustness by a factor of 3 and 4.5 for proton and X-ray irradiations, respectively (compare Figures 6.16a and 6.16b). In fact, by considering the percentage of the window closure in Figure 6.16a, after 200 krad(SiO_2) X-ray irradiation, NCMs keep 63% of the initial programming window. In contrast, the FG memories, programming window reduces to 58% of its initial value after only 50 krad(SiO_2).

This large improvement of NCM charge loss moves the reliability issues to the peripheral circuitry. In fact, the works reported in [78,79] showed that the most sensitive part of the flash memory chip is the peripheral circuitry, which may fail at doses as low as 10 krad. This is more than one decade below the dose required to reach the 20% charge loss in NCMs. However, the thinner oxides of NCMs allow the reduction of the programming voltages with respect to FG memories. It is well known from many studies in the literature that thin-oxide MOS technologies are more radiation tolerant from several points of view, such as radiation-induced charge trapping. Hence, if the programming voltage is reduced in NCM, charge pumps and the peripheral circuitry MOSFETs might be designed with thinner oxides, potentially increasing the immunity against trapped charge, which can sensibly change the threshold voltage.

In addition, the SEE experiments with heavy ions, compared with those reported on the conventional floating-gate memories, highlight an outstanding improvement of nanocrystal technology over the floating-gate device. Threshold voltage shift values as large as 3÷4 V are reported after heavy-ion irradiation of FG memories [63], and very large bitflip probability values (close to 100%) have been often observed [64,74]. Contrary to floating-gate memories, the NCM approach highlights a moderate threshold voltage shift (1 V at most) immediately after irradiation, even after several hits in the same cell.

In light of these recent achievements on NCM, some questions arise also on the radiation sensitivity of NCMs as the size of the devices is scaling down. In fact, even a single ion hit may easily mark the failure of a very small-size cell. Some predictions have been done by means of the model proposed in [71] and shown in Figure 6.21. If the cell dimension is scaled below the size of the ion track and the number of nanocrystals per cell may be as small as 10, a single ion hit may be enough to discharge (almost) all the nanocrystals. Nevertheless, NCMs are expected

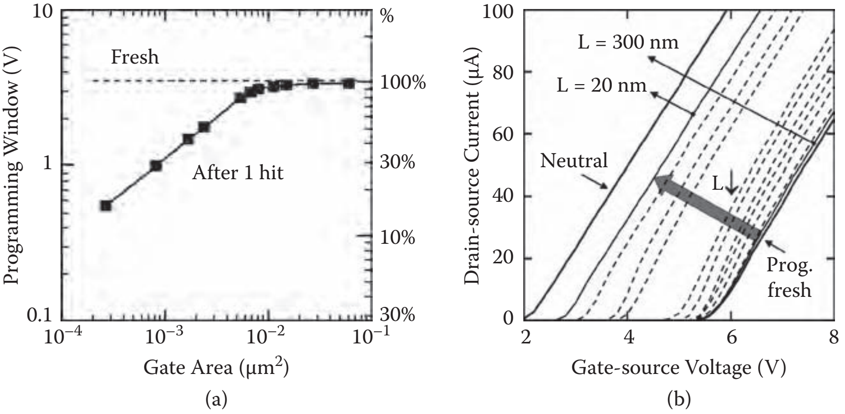


FIGURE 6.21 (a) Predicted threshold voltage shift after a single ion hit as a function of the cell gate area (the ion hit has been supposed in the center of the channel). b) Simulated I_{DS}-V_{GS} curves after a single hit (in the same condition of Figure (a) as a function of the gate area. (Reprinted with permission Cester, A., Wrachien, N., Schwank, J.R., Vizkelethy, G., Portoghese, R., Gerardi, C., "Modeling of heavy ion induced charge loss mechanisms in nanocrystal memory cells." *IEEE Transactions on Nuclear Science*, Volume 55 Issue 6, pp.2895–2903.)

to be still more robust than conventional flash memories for several reasons. In fact, multiple flips have been reported in floating-gate devices starting from the 90 nm technology node (i.e., when the FG area scale below 10^{-2} μm²) due to the charge collection at multiple nodes [66]. For comparison, the simulations in [71] predict that NCMs as small as 10^{-3} μm² still preserve an appreciable programming window (in the range of 1 V) after a single ion hit, indicating that the nanocrystals are not completely discharged yet and that the residual nanocrystal charge is still able to sustain an appreciable electric field, which prevents the channel formation.

6.5 CONCLUSIONS

Over the last 25 years, flash memories have become the nonvolatile memory of choice for most of the commercial portable devices that are widely present in everyday life. They have shown an exponential growth because of the use of very innovative technology and design solutions, which have made flash technology the most scaled among the CMOS technologies. Flash memories are entering the 40 nm lithography node, and their scaling has become even more challenging because of several electrical and reliability issues. The problems are mainly related to the very high electric fields that the thin tunnel and interpoly dielectrics have to sustain without losing their endurance and ability to retain data for more than 10 years. Overcoming these issues will require substantial innovations in cell structures and materials. The use of floating trapping nodes such as nanocrystals can provide some relief in terms of electrical and reliability drawbacks, even though the integration of nanocrystals in ultra-scaled technologies is very challenging because of the high control needed on dot density fluctuations. In this chapter, we have reported the state-of-the-art of silicon-based nanocrystal memory technologies, showing their functionality in multimegabit memory arrays. Because of their low voltage functionality and ease of integration with CMOS technology, nanocrystal memories have shown their advantages, especially in embedded flash memories (i.e., in applications where the nonvolatile memory is integrated into a host logic device to help accomplish intended system functions).

We have discussed how the information stored in NVMs can be severely endangered by ionizing radiation, due to the transient or permanent damage produced in the dielectric layers surrounding the FG. TID sources such as protons or X-rays result in progressive charge loss on the entire memory array. SEE because of heavy-ion exposure results in a dramatic shift of V_{TH} in hit FG cells due to the sudden discharge of the FG charge, mainly from photoemission from substrate and polysilicon. SEE is becoming a more and more challenging issue in the modern technologies, where the device sizes are reduced following Moore's Law. The discrete storage approach, employed in nanocrystal memories technology, mitigates most of those effects. Models and experimental results, reported in the literature, show that nanocrystal memories feature improved radiation robustness against total ionizing dose. Nanocrystal memories can withstand a radiation dose three to ten times larger than floating-gate memories in terms of charge loss and data retention, respectively. Several factors contribute to improve the NCM radiation tolerance: the presence of discrete storage sites, the smaller nanocrystal coverage area, and the thinner dielectrics. On the other hand, SEE experiments show that at least three to four ion hits are required to observe an appreciable threshold voltage shift, but despite several cells experiencing multiple hits they are still functional after the irradiation, showing no changes in the retention characteristics.

All the recent developments highlight an outstanding improvement of the nanocrystal technology over the conventional floating-gate memories in terms of radiation tolerance, which is encouraging for a potential application in radiation harsh environments and projects the NCM technology as a new class of radiation-tolerant devices.

REFERENCES

1. D. Frohman-Bentchkowsky, "Memory behavior in a floating-gate avalanche-injection MOS (FAMOS) structure," *Appl. Phys. Lett.,* vol. 18, pp. 332–334, 1971.
2. F. Masuoka, M. Asano, H. Iwahashi, T. Komuro, and S. Tanaka, "A new flash E2PROM cell using triple polysilicon technology," *Proceedings of the International Electron Devices Meeting — IEDM,* vol. 30, pp. 464–467, 1984.
3. V. N. Kynett, A. Baker, M. Fandrich, G. Hoekstra, O. Jungroth, J. Kreifels, et al., "An in-system reprogrammable 256K CMOS flash memory," Proceedings of the *IEEE International Solid-State Circuits Conference—ISSCC*, pp.132–133, 1988.
4. S. Lai, "Non-volatile memory technologies: the quest for ever lower cost," *Proceedings of the International Electron Devices Meeting — IEDM*, pp. 121-126, 2008.
5. G. Atwood, "Future directions and challenges for ETox flash memory scaling," *IEEE Trans. Device Material Rel.,* vol. 4, no. 3, pp. 301–305, 2004.

6. K. Kim, "Future memory technology: challenges and opportunities," *Proceedings of the International* Symposium on VLSI Technology, Systems and Applications — VLSI-TSA*, pp.5-9,* 2008.
7. S. Lai, *J. IBM J. Res. & Dev.,* vol. 52, pp. 529–535, 2008.
8. Source: IC Insights 2008.
9. K. Kim and G. Jeong, "Memory technologies for sub-40nm node," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM*, pp. 27–30, 2007.
10. R. Bez, E. Camerlenghi, A. Modelli, A. Visconti, "Introduction to flash memory," *Proc. IEEE,* vol. 91, pp. 489–502, 2003.
11. S. Lombardo, C. Gerardi, L. Breuil, C. Jahan, L. Perniola, G. Cina,D. Corso, E. Tripiciano, V. Ancarani, G. Iannaccone, G. Iacono, C. Bongiorno, J. Razafindramora, C. Garozzo, P. Barbera, E. Nowak, R. Puglisi, G. Costa, C. Coccorese, M. Vecchio, E. Rimini, J. Van Houdt, B. De Salvo, M. Melanotte, "Advantages of the FinFET architecture in SONOS and nanocrystal memory devices," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp.921–924, 2007.
12. D. Kwak, J. Park, K. Kim, Y. Yim, S. Ahn, Y. Park, et al., "Integration technology of 30nm generation multi-level NAND flash for 64Gb NAND flash memory," *Symposium on VLSI Technology, Digest of Technical Papers,* Kyoto, Japan, pp. 12–13, 2007.
13. H. Tanaka, M. Kido, K. Yahashi, M. Oomura, R. Katsumata, M. Kito, Y. Fukuzumi, M. Sato, Y. Nagata, Y. Matsuoka, Y. Iwata, H. Aochi, A. Nitayama, "Bit cost scalable technology with punch and plug process for ultra high density flash memory," *Symposium on VLSI Technology, Digest of Technical Papers,* pp. 14–15, 2007.
14. G. Campardo, R. Micheloni, and D. Novosel, *VLSIDdesign of Nonvolatile Memories,* Springer Series in Advanced Microelectronics, 2005.
15. R. Feynman, *Lectures on Computation,* T. Hey and R.W. Allen (Eds.), Westview, University of Southampton, England, p. 30, 1996.
16. L. Selmi and C. Fiegna, "Physical aspects of cell operation and reliability," in *Flash Memories,* P. Cappelletti et al. (Eds.), Kluwer Academic Publishers, Boston, pp. 153–239, 1999.
17. R.H. Fowler and L. Nordheim, "Electron emission in intense electric fields," *Proc. Royal Society London Series A,* vol. 119, p. 173, 1928.
18. T.S. Jung, "A 3.3-V 128-Mb multilevel NAND flash memory for mass storage applications," *Proceedings of the IEEE International Solid-State Circuits Conference—ISSCC,* pp. 32–33. 1996.
19. S. Kenney et al., "Complete transient simulation of flash EEPROM devices," *IEEE Trans. Electron Dev,* vol. 39, pp. 2750–2757, 1992.
20. S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chan, and D. Buchanan, "Volatile and non-volatile memories with nano-crystal storage," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp. 521, 1995.
21. T. Ishii et al., "Engineering variations: toward practical single-electron (few-electron) memory," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp. 305, 2000.
22. R. Muralidhar, R.F. Steimle, M. Sadd, R. Rao, R.C.T. Swift, E.J. Prinz, J. Yater, L. Grieve, K. Harber, B. Hradsky, S. Straub, B. Acred, W. Paulson, W. Chen, L. Parker, S.G.H. Anderson, M. Rossow, T. Merchant, M. Paransky, T. Huynh, D. Hadad, Ko-Min Chang; B.E.White Jr., "A 6V embedded 90nm silicon nanocrystal nonvolatile memory," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp. 601–604, 2003.
23. C. Gerardi, V. Ancarani, R. Portoghese, S. Giuffrida, M. Bileci, G. Bimbo, O. Brafa, D. Mello, G. Ammendola, E. Tripiciano, R. Puglisi, S.A. Lombardo, "Nanocrystal memory cell integration in a stand-alone 16-Mb NOR flash device," *IEEE Trans. Electron Devices,* vol. 54, p. 1376, 2007.

24. S. Lombardo, R.A. Puglisi, I. Crupi, D.Corso, G.Nicotra, L.Perniola, B.DeSalvo, C.Gerardi, "Distribution of the threshold voltage window in nanocrystal memories with Si dots formed by chemical vapor deposition: effect of partial self-ordering," *Proceedings of the Non-Volatile Semiconductor Memory Workshop — 20th NVSM Workshop,* pp. 69–79, 2004.
25. B. De Salvo, C. Gerardi, S. Lombardo, T. Baron, L. Perniola, D. Mariolle, P. Mur, A. Toffoli, M. Gely, M.N. Semeria, S. Deleonibus, G. Ammendola, V. Ancarani, M. Melanotte, R. Bez, L. Baldi, D. Corso, I. Crupi, R.A. Puglisi, G. Nicotra, E. Rimini, F. Mazen, G. Ghibaudo, G. Pananakakis, C.M. Compagnoni, Ielmini, A. Lacaita, A. Spinelli, Y.M. Wan, K. van der Jeugd, K., "How far will silicon nanocrystal push the scaling limits of NVM's technology?", *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp. 597, 2003.
26. Z. Liu, C. Lee, V. Narayanan, G. Pei, and E.C. Kan, "Metal nanocrystal memories—part II: electrical characteristics," *IEEE Transactions on Electron Devices,* vol. 49, p. 1614, 2002.
27. J.J. Lee and D.-L. Kwong, "Metal nanocrystal memory with high-k tunneling barrier for improved data retention," *IEEE Transactions on Electron Devices,* vol. 52, p. 507, 2005.
28. C. Bonafos et al., "Manipulation of two-dimensional arrays of Si nanocrystals embedded in thin SiO2 layers by low energy ion implantation," *Journal of Applied Physics,* vol. 95, p. 5696, 2004.
29. J. De Blauwe, "A novel aerosol nanocrystal floating-gate me device for nonvolatile memory application," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* p. 683, 2000.
30. K.W. Guarini, C.T. Black, Y. Zhang, I.V. Babich, E.M. Sikorski, and L.M. Gignac, "Low voltage, scalable nanocrystal flash memory fabricated by templated self assembly," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp. 22.2.1–22.2.4, 2003.
31. S. Jacob et al., "Investigation of reliability characteristics of Si nanocrystal NOR memory arrays," *Proceedings of the Non-Volatile Semiconductor Memory Workshop — 23rd NVSM Workshop,* pp. 71–72, 2007.
32. G. Ammendola, M. Vulpio, M. Bileci, N. Nastasi, C. Gerardi, G. Renna, et al., "Nanocrystal metal-oxide-semiconductor memories obtained by chemical vapor deposition of Si nanocrystals," *J. Vac. Sci. Technol.,* vol. B20, p. 2075, 2002.
33. G. Nicotra, R.A. Puglisi, S. Lombardo, C. Spinella, M. Vulpio, G. Ammendola, et al., "Nucleation kinetics of Si quantum dots on SiO2," *J. Appl. Phys,* vol. 95, p. 2049, 2004.
34. R.A. Puglisi, G. Nicotra, S. Lombardo, C. Spinella, G. Ammendola, and C. Gerardi, "Partial self-ordering observed in silicon nanoclusters deposited on silicon oxide substrates by chemical vapor deposition," *Physical Review,* B71, 125322, 2005.
35. B. Hradsky, R. Muralidhar, R. Rao, B. Steimle, S. Straub, B.E. White Jr., M. Sadd, S.G.H. Anderson, J.A. Yater, C.T. Swift,B. Acred, J. Peschke, E.J. Prinz, E.J., K.M. Chang, "Nanocrystal physical attributes influencing non-volatile memory performance," *IEEE — Device Research Conference Digest,* vol. 1, pp. 37–38, 2005.
36. I. Crupi et al., "Peculiar aspects of nanocrystal memory cells: data and extrapolations," *IEEE Trans. Nanotechnology,* vol. 2, p. 319, 2003.
37. C. Gerardi, G. Molas, G. Albini, E. Tripiciano, M.Gely, A. Emmi, O. Fiore, E.Nowak, D. Mello, M. Vecchio, L. Masarotto, R. Portoghese, B. De Salvo, S. Deleonibus, A. Maurelli, "Performance and reliability of a 4Mb Si nanocrystal NOR flash memory with optimized 1T memory cells," *Proceedings of the International Electron Devices Meeting — IEEE-IEDM,* pp. 821–824, 2008.

38. T.P. Ma and P.V. Dressendorfer, *Ionizing Radiation Eeffects in MOS Devices and Circuits,* Wiley, New York, 1989.
39. A.G. Holmes-Siedle and L. Adams, *Handbook of Radiation Effects,* Oxford University Press, 2002.
40. T.R. Oldham, *Ionizing Radiation Effects in MOS Oxides,* World Scientific, 2000.
41. *IEEE Trans. Nuclear Science, special issue on single-event effects and space radiation environment,* April 1996.
42. H. Mavromichalaki, A. Papaioannou, G. Mariatos, M. Papailiou, A. Belov, E. Eroshenko, et al., "Cosmic ray radiation effects on space environment associated to intense solar and geomagnetic activity," *IEEE Trans. Nucl. Sci.,* vol. 54, pp. 1089–1096, Aug. 2007.
43. S. Moore, "Extreme solar storm strikes Earth," *IEEE Spectrum,* vol. 40, pp. 15–16, Dec. 2003.
44. J.M. Benedetto and H.E. Boesch, *IEEE Trans. Nucl. Sci.,* vol. 33, no. 6, p. 131, 1986.
45. L. Osanger, "Initial recombination of ions," *Phys. Rev.,* vol. 54, pp. 554–557, 1938.
46. T.R. Oldham and J.M. McGarrity, "Comparison of 60Co response and 10 KeV x-ray response in MOS capacitors," *IEEE Trans. Nucl. Sci.,* vol. 30, pp. 4377–4381, Dec. 1983.
47. A. Scarpa, A. Paccagnella, F. Montera, G. Ghibaudo, G. Pananakakis, G. Ghidini, et al., "Ionizing radiation induced leakage current on ultra-thin gate oxides," *IEEE Trans. Nucl. Sci.,* vol. 44, no. 6, pp. 1818–1825, Dec. 1997.
48. M. Ceschia, A. Paccagnella, A. Cester, A. Scarpa, and G. Ghidini, "Radiation induced leakage current and stress induced leakage current in ultra-thin gate oxides," *IEEE* — *Trans. Nucl. Sci.,* vol. 45, pp. 2375–2382, Dec. 1998.
49. J.N. Bradford, "Clusters in ionization tracks of electrons in silicon dioxide," *IEEE Trans. Nucl. Sci.,* vol. 33, pp. 1271–1275, Dec. 1986.
50. T.R. Oldham, "Recombination along the tracks of heavy Charged particles in SiO films," *J. Appl. Phys.,* vol. 57, p. 2695, 1985.
51. F.W. Sexton, D.M. Fleetwood, M.R. Shaneyfelt, P.E. Dodd, and G.L. Has, "Single event gate rupture in thin gate oxides," *IEEE Trans. Nucl. Sci,* vol. 44, no. 6, pp. 2345–2352, 1997.
52. M. Ceschia, A. Paccagnella, M. Turrini, A. Candelori, G. Ghidini, and J. Wyss, "Heavy ion irradiation of thin oxides," *IEEE Trans. Nucl. Sci.,* vol. 47, pp. 2648–2655, Dec.2000.
53. J.F. Conley, Jr., J. S. Suehle, A. H. Johnston, B. Wang, T. Miyahara, E. M. Vogel, et al.,"Heavy ion induced soft breakdown of thin gate oxides," *IEEE Trans. Nucl. Sci.,* vol. 48, no. 6, pp. 1913–1916, Dec. 2001.
54. A. Cester, L. Bandiera, M. Ceschia, G. Ghidini, and A. Paccagnella, "Noise characteristics of radiation-induced soft breakdown current in ultrathin oxides," *IEEE Trans. Nucl Sci.,* vol. 48, p. 2093, 2001.
55. L.W. Massengill, B.K. Choi, D.M. Fleetwood, R.D. Schrimpf, K.F. Galloway, M.R.Shaneyfelt, et al., "Heavy-ion-induced breakdown in ultra-thin gate oxides and high-k dielectrics," *IEEE Trans. Nucl. Sci.,* vol. 48, pp. 1904–1912, 2001.
56. A.L. Sternberg, L.W. Massengill, S. Buchner, R.L. Pease, and Y. Boulghassoul, "Sensitivity classification of analog single-event transients," *Proceedings of RADECS,* vol. 85, 2002.
57. S.C Moss, S.D. LaLumondiere, J.R. Scarpulla, K.P. MacWilliams, W.R. Crain, and R.Koga, "Correlation of picosecond laser-induced latchup and energetic particle-induced latchup in CMOS test structures," *IEEE Trans. Nucl. Sci.*, vol. 42, pp. 1948–1956, Dec.1995.
58. T.J. O'Gorman, "The effect of cosmic rays on the soft error rate of a DRAM at ground level," *IEEE Trans. Nucl. Sci.*, vol. 41, pp. 553–557, Apr. 1994.

59. G. Cellere, A. Paccagnella, S. Lora, A. Pozza, G. Tao, and A. Scarpa, "Charge loss after ^{60}Co irradiation of flash arrays," *IEEE Trans. Nucl. Sci.,* vol. 51, pp. 2912–2916, Oct. 2004.
60. G. Cellere, A. Paccagnella, A. Visconti, M. Bonanomi, A. Candelori, and S. Lora, "Effect of different total ionizing dose sources on charge loss from programmed floating gate cells," *IEEE Trans. Nucl. Sci.,* vol. 52, pp. 2372–2377, Dec. 2005.
61. E.S. Snyder, P.J. McWhorter, T.A. Dellin, and J.D. Sweetman, "Radiation response of floating gate EEPROM memory cells," *IEEE Trans. Nucl. Sci.*, vol. 36, pp. 2131–2139, 1989.
62. R.C. Hughes, "Charge-carrier transport phenomena in amorphous SiO2: direct measurement of the drift mobility and lifetime," *Phys. Rev. Lett.,* vol. 30, pp. 1333–1336,1973.
63. G. Cellere, P. Pellati, A. Chimenton, J. Wyss, A. Modelli, L. Larcher, et al., "Radiation effects on floating-gate memory cells," *IEEE Trans. Nucl. Sci.,* vol. 48, pp. 2222–2228, Dec. 2006.
64. S.M. Guertin, D.M. Nguyen, and J.D. Patterson, "Microdose induced data loss on floating gate memories," *IEEE Trans. Nucl. Sci.,* vol. 53, pp. 3518–3524, Dec. 2006.
65. N.Z. Butt and M.A. Alam, "Single event upsets in floating gate memory cells," *Proc. Of International Reliability Physics Symposium,* pp. 547–555, Apr. 2008.
66. G. Cellere, A. Paccagnella, A. Visconti, M. Bonanomi, R. Harboe-Sørensen, and A.Virtanen, "Angular dependence of heavy ion effects in floating gate memory arrays," *IEEE Trans. Nucl. Sci.,* vol. 54, pp. 2371–2378, Dec. 2007.
67. G. Cellere, A. Paccagnella, A. Viconti, M. Bonanomi, and A. Candelori, "Transient conductive path induced by a single ion in 10 nm SiO2 layer," *IEEE Trans. Nucl. Sci.,* vol. 51, pp. 3304–3311, Dec. 2004.
68. E.J. Yoffa, "Dynamics of dense laser-induced plasmas," *Phys. Rev. B,* vol. 21, pp. 2415–2425, 1980.
69. A. Meftah, F. Brisard, J.M. Costantini, E. Dooryhee, M. Hage-Ali, H. Herviu, "Track formation in SiO2 quartz and the thermal-spike mechanism," *Phys. Rev. B*, vol.49, no. 18, pp. 12457–12463, 1994.
70. M. Toulemonde, C. Dufour, and E. Paumier, "Transient thermal process after a high-energy heavy-ion irradiation of amorphous metals and semiconductors," *Phys. Rev. B*, vol. 46, no. 22, pp. 14362–14369, 1992.
71. A. Cester, N. Wrachien, J. Schwank, G. Vizkelethy, R. Portoghese, and C. Gerardi, "Modeling of heavy ion induced charge loss mechanisms in nanocrystal memory cell", *IEEE Trans. Nucl. Sci.,* vol. 55, no. 6, pp. 2895–2903, 2008.
72. P.E. Dodd, "Physics-based simulation of single-event effects," *IEEE Trans. Dev. Mat. Reliab.,* vol. 5, pp. 343–357, Sep. 2005.
73. http://public.itrs.net
74. L. Larcher, G. Cellere, A. Paccagnella, A Chimenton, A. Candelori, and A. Modelli, "Data Retention after heavy ion exposure of floating gate memory: analysis and simulation," *IEEE Trans. Nucl. Sci.,* vol. 50, pp. 2176–2183, Dec. 2003.
75. G. Cellere and A. Paccagnella, "A review of ionizing radiation effects in floating gate memories," *IEEE Trans. Device and Material Reliability*, vol. 4, pp. 359–370, Sept. 2004.
76. G. Cellere, L. Larcher, A. Paccagnella, A. Visconti, and M. Bonanomi, "Radiation induced leakage current in floating gate memory cells," *IEEE Trans. Nucl. Sci.,* vol. 52, pp. 2144–2152, Dec. 2005.
77. N. Wrachien, A. Cester, R. Portoghese, and C. Gerardi, "Investigation of proton and x-ray irradiation effects on nanocrystal and floating gate memory cell arrays," *IEEE Trans. Nucl. Sci.,* vol. 55, no. 6, pp. 3000–3008, 2008.
78. D.N. Nguyen, C.I. Lee, and A.H. Johnston, "Total ionizing dose effects on flash memories," in *Proc. 1998 IEEE Radiation Effects Data Workshop*, pp. 100–103.

79. D.N. Nguyen, S.M. Guertin, G.M. Swift, and A.H. Johnston, "Radiation effects on advanced flash memories," *IEEE Trans. Nucl. Sci.,* vol. 46, pp. 1744–1750, Dec. 1999.
80. D.M. Fleetwood and N.S. Saks, "Oxide, interface, and border traps in thermal, N_2O, and N_2O-nitrided oxide," *J. Appl. Phys.,* vol. 79, no. 3, 1996.
81. E.F. Runnion, S.M. Glastone, R.S. Scott, D.J. Dumin, L. Lie, and J.C. Mitros, "Thickness dependence of stress-induced leakage currents in silicon oxide," *IEEE Trans. Electron Devices*, vol. 44, pp. 993–1001, 1997.
82. K. Sakakibara, N. Ajika, M. Hatanaka, H. Miyoshi, and A. Yasuoka, "Identification of stress-induced leakage current components and the corresponding trap models in SiO2 films," *IEEE Trans. Electron Dev.*, vol. 44, pp. 986–992, 1997.
83. D.J. DiMaria and E. Cartier, "Mechanism for stress-induced leakage currents in thin silicon dioxide films," *J. Appl. Phys*, vol. 78, pp. 3883–3894, 1995.
84. M.P. Petkov, L.D. Bell, and H.A. Atwater, "High total dose tolerance of prototype silicon nanocrystal non-volatile memory cells," *IEEE Trans Nucl. Sci.,* vol. 51, pp. 3822–3826, Dec. 2004.
85. T.R. Oldham, M. Suhail, P. Kuhn, E. Prinz, H. Kim, and K.A. LaBel, "Effect of heavy ion exposure on nanocrystal non-volatile memory," *IEEE Trans. Nucl. Sci.,* vol. 52, Dec. 2005.
86. A. Cester, A. Gasperin, N. Wrachien, A. Paccagnella, V. Ancarani, and C. Gerardi, "Impact of heavy-ion strikes on nanocrystal non volatile memory cell arrays," *IEEE Trans. Nucl. Sci.,* vol. 53, pp. 3195–3202, Dec. 2006.
87. A. Gasperin, A. Cester, N. Wrachien, A. Paccagnella, V. Ancarani, and C. Gerardi, "Radiation-induced modifications of the electrical characteristics of nanocrystal memory cells and arrays," *IEEE Trans. Nucl. Sci.,* vol. 53, pp. 3693–3700, Dec. 2006.
88. A. Cester, N. Wrachien, A. Gasperin, A. Paccagnella, R. Portoghese, and C. Gerardi, "Radiation tolerance of nanocrystal-based flash memory arrays against heavy ion irradiation," *IEEE Trans. Nucl. Sci.,* vol. 54, pp. 2196–2203, Dec. 2007.
89. E. Verrelli, D. Tsoukalas, M. Kokkoris, R. Vlastou, P. Dimitrakis, and P. Normand, "Proton radiation effects on nanocrystal non-volatile memories," *IEEE Trans. Nucl. Sci.,* vol. 54, pp. 975–981, Aug. 2007.
90. G. Cellere, A. Paccagnella, A. Viconti, M. Bonanomi, P. Caprara, and S. Lora, "A model for TID effects on floating gate memory cells," *IEEE Trans. Nucl. Sci.,* vol. 51, pp. 3753–3758, Dec. 2004.
91. L.Z. Scheick, P.J. McNulty, and D.R. Roth, "Measurement of the effective sensitive volume of FAMOS cells of an ultraviolet erasable programmable read-only memory," *IEEE Trans. Nucl. Sci.,* vol. 47, pp. 2428–2434, Dec. 2000.
92. P.J. McNulty, L.Z. Scheick, D.R. Roth, M.G. Davis, and M.R.S. Tortora, "First failure predictions for EPROMs of the type flown on the MPTB satellite," *IEEE Trans. Nucl. Sci.,* vol. 47, pp. 2237–2243, Dec. 2000.
93. D.M. Fleetwood, D.E. Beutler, L.J. Lorence, D.B. Brown, B.L. Draper, L.C. Riewe, et al., "Comparison of enhanced device response and predicted x-ray dose enhancement effects in MOS oxides," *IEEE Trans. Nucl. Sci.,* vol. 35, no. 6, pp. 1265–1271, Dec. 1988.
94. A. Cester, S. Gerardin, A. Paccagnella, J.R. Schwank, G. Vizkelethy, A. Candelori, et al., "Drain current decrease in MOSFETs after heavy ion irradiation," *IEEE Trans. Nucl. Sci.,* vol. 51, pp. 3150–3157, Dec. 2004.
95. A. Candelori, M. Ceschia, A. Paccagnella, J. Wyss, D. Bisello, and G. Ghidini, "Thinoxide degradation after high energy ion irradiation," *IEEE Trans. Nucl. Sci.,* vol. 48, pp.1735–1743, Oct. 2001.

7 Radiation Hardened by Design SRAM Strategies for TID and SEE Mitigation

Lawrence T. Clark

CONTENTS

7.1 CHAPTER OVERVIEW

7.1.1 Embedded SRAMs in Integrated Circuit Design

Static random access memory (SRAM) is ubiquitous in modern system-on-a-chip (SOC) integrated circuits (ICs). Due to its value in programmable systems by providing fast scratchpad memory in embedded and real-time applications as well as space for large working sets in microprocessor designs, IC SRAM content continues to grow. As ICs surpass 1 billion transistors, and given the high relative design and power efficiency of memory arrays compared with random logic, SRAM is projected to comprise 90% of the total die area by 2013 [1]. For instance, the Itanium processor has progressed from 6 MB and 9 MB L3 caches on 130 nm fabrication processes to 24 MB caches on the 65 nm technology generation [2-4]. The Xeon processors include 16 MB caches [5]. Consequently, ICs designed for space and other radiation environments require robust SRAM designs if they are to track the size and performance of commercial ICs.

7.1.2 The Radiation Space Environment and Effects

The earth's radiation environment consists of electrons, protons, and heavy ions. The former two are trapped by the earth's magnetic field where they follow the field lines, where these particle fluxes are highest. A total of 85% of galactic cosmic ray particles are protons, with the rest composed of heavy ions [6]. Cosmic ray flux is essentially omnidirectional, so microelectronics may be affected by particles impinging at any angle. Importantly, this means that ions can transit an IC parallel to the device surface, since there is no practical level of shielding that can stop all protons and heavy ions. Solar cycles also strongly affect the radiation environment. Ordinarily the helium ions in the solar emitted particle fluxes comprise 5–10%, and heavier ion fluxes are very small, well below the galactic background. During major solar events, some heavy-ion fluxes may increase by up to four orders of magnitude above the galactic background, for as long as days at a time.

The dominant radiation effects on microcircuits in space are due to deposited charge from ionization tracks produced by single particles. These produce two primary effects. First, collected charge from a single particle can upset circuit state, referred to as a single-event effect (SEE). Second, changes in the charge state of dielectrics due to total accumulated ionization can alter device characteristics, referred to as total ionizing dose (TID) effects [7].

Both protons and heavy ions can deposit charge that can upset the circuit state. Upsetting a feedback (state storage) node such as a memory bit is defined as a single-event upset (SEU). Heavy ions affect the circuit state through direct ionization due to columbic interaction with the substrate material, producing about 10 fC of charge per μm of track length per linear energy transfer (LET). Memory cells are often characterized for SEU by the total charge Q_{crit} that is required to upset their state. Charge that temporarily disrupts a logic node results in an incorrect voltage transient of a magnitude and duration determined by the node capacitance and the driving circuit's ability to remove the charge. These are referred to as a single-event transient (SET). An SET can affect the IC architectural state (the state that is visible to the surrounding system) only if sampled by a latch whose output is subsequently used.

Protons interact with the silicon through multiple mechanisms, predominantly by direct ionization but also through secondary nuclear particle emission due to Si recoil. The former generates relatively small amounts of charge, but the latter can upset circuits hardened to high LET. Approximately 1 in 100,000 protons impinging will produce a nuclear reaction. Moreover, the multiple secondary particles may interact with the circuit after moving in multiple directions. A single particle produces charge in linear tracks. Charge is collected by diffusion and by drift, with the latter due to the device depletion regions. Charge collection is enhanced by "funneling," which is a third field driven collection mechanism that extends the field driven collection by the redistribution of the deposited carriers. Parasitic bipolar action can also increase the current collected at a specific node, greatly increasing the upset rate and extent.

Impinging particles can also permanently disable the microcircuit by excessive displacement damage or by rupturing the gates. Such permanent effects are not pertinent to the discussions in this chapter.

7.1.3 Chapter Outline

This chapter focuses on SRAM design using radiation-hardening-by-design (RHBD) techniques. Both TID and SEE hardening are covered. The latter approaches described assume that error detection and correction (EDAC) is used to mitigate individual SEU, as RHBD hardened cell approaches have diminishing value in modern highly scaled fabrication processes. Small, dense geometries make simultaneous upset of multiple circuit nodes from a single particle strike increasingly likely. A primary focus, therefore, is on mitigating SETs that can cause upsets that confound the EDAC or otherwise cause incorrect SRAM operation. All of the approaches examined in this chapter have been fabricated and tested—measurements quantifying their effectiveness are also described and discussed.

The last section briefly outlined the space radiation environment. Subsequent sections include a discussion of basic SRAM cell design, which is tutorial in nature. Test structures to characterize SRAM cells are then described. This is important, particularly for RHBD SRAM cells, which do not undergo the same rigorous testing and validation during the fabrication process development that the foundry provided cells do. The TID response of SRAM cells hardened by various techniques and that of an unhardened version are examined, as are the trade-offs in cell size and hardness for various TID hardening approaches. Heavy-ion beam testing results show the importance of multiple-bit upset (MBU) and SET response. The design of an SET hardened SRAM is then described, as well as its response in ion beam testing, which is compared with that of an unhardened device. We then briefly summarize the results to conclude the chapter.

7.2 RADIATION HARDENING

All hardened designs should mitigate four issues: (1) single-event latchup (SEL) due to ion-strike-induced substrate currents; (2) single-event logic upset due to the capture of SETs in sequential circuits (e.g., latches and flip-flops); (3) single-event upsets (SEUs) of storage nodes, which includes storage latches in registers and SRAM memories; and (4) TID, which can affect the individual device and isolation characteristics. These device changes in turn may deleteriously affect the circuit behavior.

There are two basic approaches to fabricate radiation tolerant ICs: hardening by process [8] and hardening by design [9]. Hardening by process uses a specialized fabrication process that has features specifically added to mitigate radiation effects, such as silicon-on-insulator (SOI) substrates, special body ties, and dense, high-value resistors [10-12]. RHBD allows radiation-tolerant circuits to be fabricated on commercially available state-of-the-art complementary metal-oxide semiconductor (CMOS) manufacturing processes [9,13] to reduce cost and to improve circuit performance. It relies exclusively on special circuit topologies and layouts rather than specialized process features and devices to provide hardening. For example, *p*-type guard rings around n-channel metal-oxide semiconductor (NMOS) diffusions, similar to those used for input/output (I/O) electrostatic discharge (ESD) protection, provide increased SEL immunity. Of course, actual designs may use a combination of approaches. For instance, SOI substrates are available on commercial unhardened processes. Furthermore, specific radiation-hard (rad-hard) circuits and layouts are still required when using rad-hard fabrication processes.

7.2.1 Total Ionizing Dose Effects

In modern processes with sub-3 nm thick gate oxides, TID primarily increases leakage under isolation oxides and at the gate edges, that is, at the thin gate oxide to isolation oxide interfaces. This slowly increases leakage from a parasitic transistor at the transistor edge as its threshold voltage, V_{th}, decreases with TID. Since the trapped charge is positive, only NMOS transistors suffer from increased leakage due to these parasitic devices along the gate edges. Similarly, leakage between n-type diffusions (e.g., between the n-well and NMOS drains) can be increased by reduction of the field oxide V_{th} [14,15]. These increases in leakage are manifest in a given IC as increased I_{DD} measured in the quiescent state, commonly referred to as standby current, I_{SB}. TID has been shown to cause functionality loss in SRAMs [16]. Increased leakage currents can interfere with proper precharging or small swing bit-line signal development. Leakage within the cell can also affect the read stability by changing the cell static noise margin [17].

TID is mitigated by higher doping at the oxide interfaces or, when using RHBD approaches, by using annular or edgeless NMOS transistor gates. The standard RHBD technique for mitigating TID increased leakage in the parasitic edge transistors is to use "edgeless" or annular transistor geometries. The annular transistor fully encloses the drain or source, so the same potential is at both sides of the transistor edge to isolation oxide interface.

7.2.2 Single-Event Effects in SRAMs

SRAMs are prone to both SEU and SET generated errors. The former have been addressed by the use of the dual interlocked storage cell (DICE) or other approaches that add transistors [18]. The DICE circuit adds redundant storage nodes and a self-correction mechanism, which allows three correct storage nodes to correct one incorrect node. Other approaches generally attempt to limit the ability of collected charge to affect the latch feedback state. It is important to realize that at any critical node separation errors may occur in space, where particles can be incident at any angle. In any of these approaches, the cell storage (critical) nodes must be spaced sufficiently far apart to minimize the probability, by increasing the incident ionizing radiation particle track angle required to upset multiple latch nodes with a single ionizing particle strike [19,20]. This becomes more difficult as process dimensions are scaled down, as this naturally places critical nodes closer together. Temporal latches mitigate both types of upset [21] but with so much added circuit area and path delay that they are impractical for high-speed, high-density memory arrays.

Easily the most common approach in hardened processes is the addition of resistors in the SRAM cell latch feedback path [22,23]. Since charge can be collected only at a p-n junction, the series RC produces a delay that allows the collected charge to be removed by the latch feedback transistors before the feedback node can transition, thus restoring the original cell state. Since undoped polysilicon conductivity is constant, it becomes increasingly difficult to produce resistors providing sufficiently long time constants as fabrication processes scale. Thus, more modern versions use special high resistivity vias or layers [12]. Hardened processes often use SOI substrates or special implants to limit the track length of the collected charge from impinging ionizing radiation particles [24]. Limiting the track length reduces the required RC time constant by attenuating the collected charge and thus the duration of an upset. Similarly, it mitigates SET durations [12].

SRAMs are not only susceptible to cell state upset (SEU). It has also been long known that an SET on word-line (WL) signals [25,26] can cause improper operation by asserting the wrong or more than one of the normally one-hot WLs high. Similarly, any SET in the control, clocking, or decode paths may cause the wrong operation or the wrong address to be accessed. In the worst case (e.g., when the wrong memory address is read), the parity or error correction code (ECC) may be correct. Referring to Figure 7.1a, WLs act as selects that allow a row of SRAM cells to discharge the appropriate precharged bit-lines (BL and BLN) in each column. In the event of a WL SET a number of circuit-level behaviors may occur. If two WLs are asserted high simultaneously during a read operation (e.g., WLn and WL0 in Figure 7.1a), then the BLs will logically OR the values, as a BL is a dynamic NOR

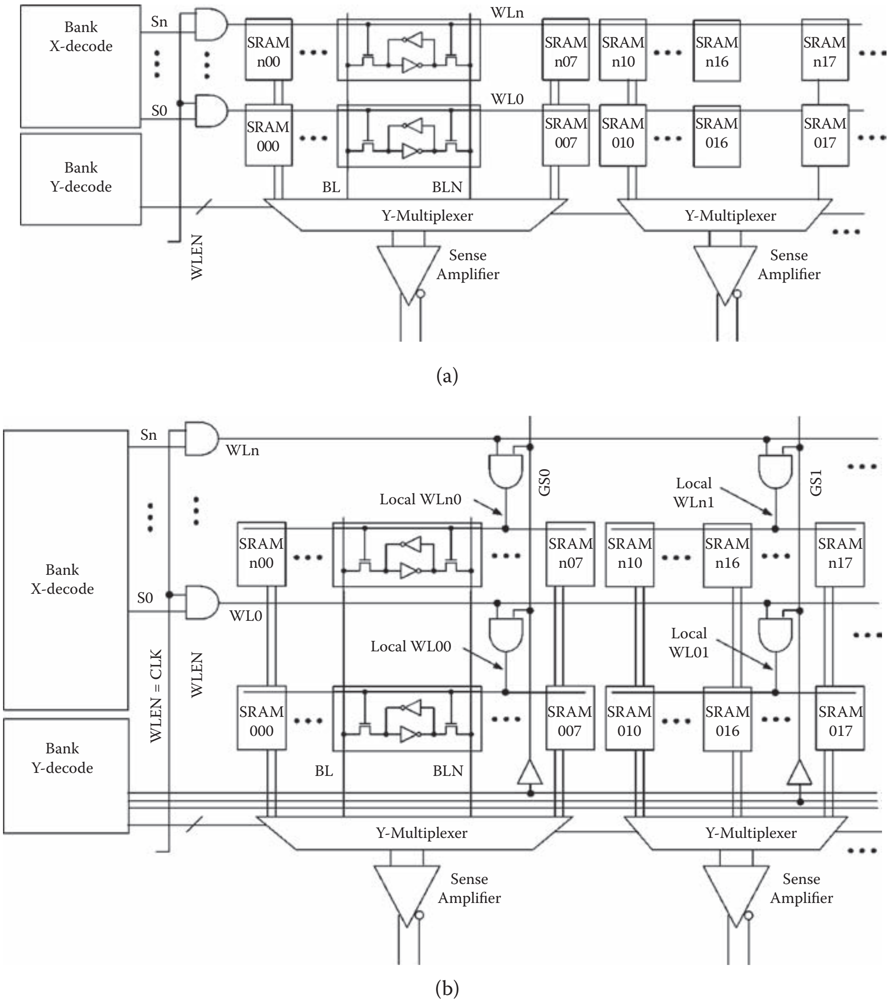


FIGURE 7.1 Conventional SRAM bank word line architecture (a) and divided word line architecture (b). An SET may affect all or a number of local WLxx by incorrectly asserting the global WLx attached to each of them.

multiplexer. If the SET occurs late in the read or write phase, and one of the bit-line pairs in each column has discharged to V_{SS}, a subsequent WL misassertion may write the BL state into the row controlled by that misasserted WL. When this value inadvertently enters the IC architectural state and is subsequently read, this undetected error is termed a silent data corruption (SDC).

The "column group," which is the basic design unit that can read or write one bit, generally contains many SRAM cell columns. There are eight, labeled SRAM000 to SRAM007 in the leftmost group in the examples in Figure 7.1, sharing one sense amplifier and associated write circuitry through the column or "Y" multiplexer. By convention WLs are the X multiplexer selects. Multiple SRAM cell columns per sense/write circuit are required primarily by the fact that the former are large. Thus, the layout is eased by not trying to fit large sense circuits into the tight SRAM cell pitch. It also forces spacing between individual cells containing data from the same word, assuming a given word is read in one cycle. This separation due to Y multiplexing makes it less likely that an MBU will upset multiple bits in a single protected codeword, as has been common knowledge in the SRAM design community since the 1980s. Commercial designs have tended to use at least four SRAM columns per group, but that may need to increase in the future as SRAM cells scale to smaller dimensions [27].

Figure 7.1b shows a technique that has been employed to mitigate such control and WL SET-induced errors. Each column group again contains multiple SRAM cell columns, but each WL (and control line, not shown) is individually buffered. Thus, an SET on the local WL (e.g., LocalWLn0 in Figure 7.1b) will affect only that local column group. This scheme was applied and errors due to local WL ion strikes were recorded [28]. Since only one bit can be read at a time from a column group, EDAC can correct such an error, whether on a read or write operation. However, the global WL (WL0 to WLn in Figure 7.1b) are not so protected. Sufficient WL capacitance and drive will provide some SET immunity, but in general large array sizes are required to raise the threshold LET sufficiently [29].

One approach that has been put forward to mitigate this issue is the "bit per array" architecture, where each bit in an EDAC protected word is stored in a separate SRAM bank. Conventionally a "bank" is a stand-alone unit containing clocking, control, decode, and array circuits. However, this in itself is insufficient to protect against such errors in all cases. An example of this case is shown in Figure 7.2a. Here, all bits in an EDAC protected codeword reside in separate SRAM banks, providing excellent critical node separation and greatly limiting the probability of a single ionizing particle strike upsetting multiple nodes in a single EDAC word. Referring to Figure 7.2, note that all addresses and control signals fan out from single registers, which we may assume are protected against errors on the inputs or clocks (e.g., by the use of temporal or other techniques) [18,21]. However, the output node or one of the inverters that provides fan up to drive the heavily loaded address bus is not

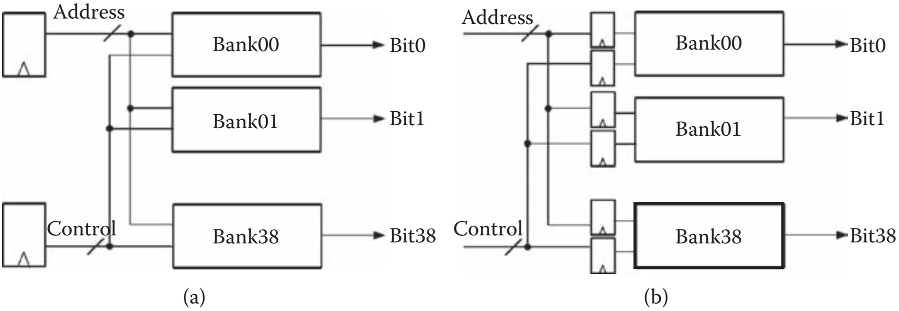


FIGURE 7.2 SRAM bit per bank architecture (a). A SET may still affect all banks by mis-asserting a the address or control signals unless mitigating latches are placed at each bank input as in (b).

so protected. Consequently, an SET on one of these nodes propagates to all arrays, which can manifest as an SDC by having all the arrays perform the wrong operation or access the wrong memory location. The way to protect against this, albeit small cross section, failure scenario is to place the SET mitigating latches in each memory array or subarray, as shown in Figure 7.2b. However, the address setup time, measured at the array input, increases commensurately, as it must be greater than 2· t_{SET}, where t_{SET} is the SET duration [30]. Consequently, such retiming must be comprehended by the overall micro-architecture. Power dissipation obviously increases with the bit per array architecture, as many SRAM banks must be activated in a single clock cycle, where otherwise only one might be. The SRAM caches in embedded commercial microprocessors account for as much as 43% of the total power dissipation [31]. This can be driven down to 15% by fine-grained clock gating—that is, activating only the necessary, smaller banks [32]. Thus, the designer is faced with a number of circuit and micro-architectural challenges that may profoundly affect the speed and power dissipation, when implementing radiation-hard SRAMs.

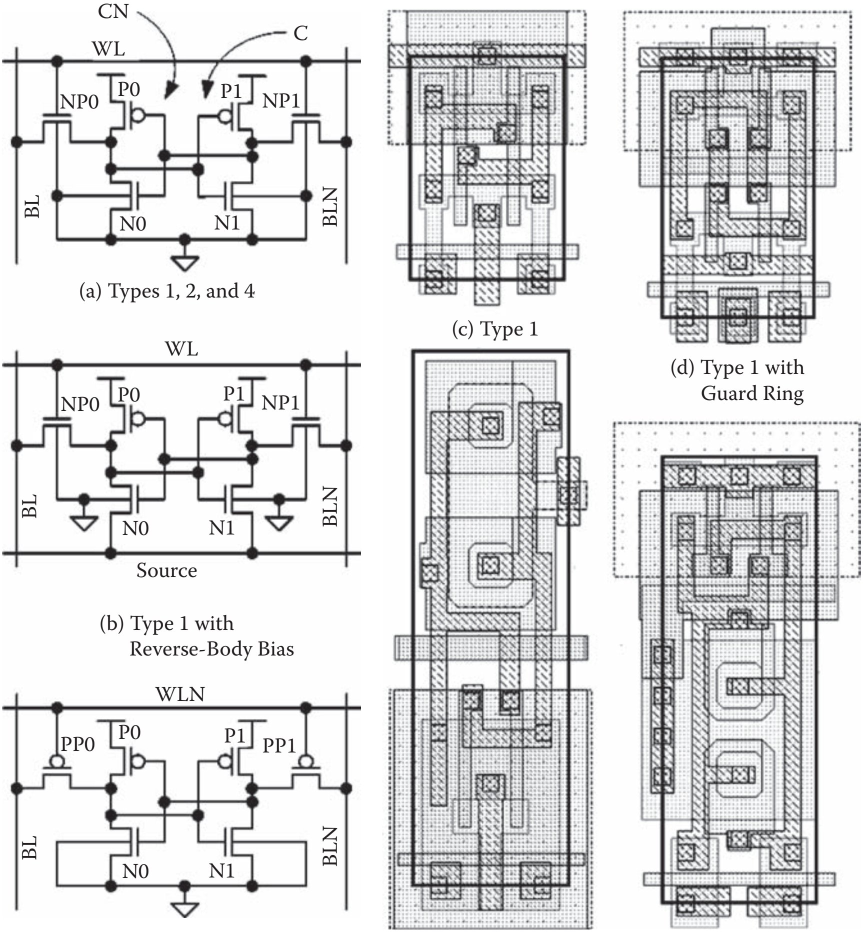
7.3 RADIATION HARDENING BY DESIGN IN SRAMS

As mentioned already, edgeless NMOS transistors effectively mitigate a major TID-induced leakage current increase component. The drawback of using annular NMOS transistor gates is that the topology imposes a relatively large minimum transistor width, since a contact must be placed within the edgeless gate. The necessary gate to contact spacing sets the inside perimeter, while the gate length, which must usually include extra margin on the 45° angle gate edges, sets the outer perimeter. For the 130 and 90 nm processes used in most examples in this chapter, the minimum widths are increased by 11.1× and 9.1×, respectively. For high drive gates (e.g., inverters driving clock or large control nodes), the RHBD penalty is negligible [33].

In SRAM cells, RHBD using conventional techniques imposes a significant increase in size. To avoid a significant area impact for RHBD SRAMs more clever TID mitigation techniques must be found. Furthermore, commercial foundries offer smaller SRAM cells that violate the standard process layout design rules. These tighter SRAM cell layouts are optimized by fabricating large numbers of cells and optimizing the required "array rules" during the process technology development. Essentially, which rules can be "cheated" for these highly regular SRAM layouts is determined experimentally. This makes the RHBD impact even greater, since no RHBD design will be able to similarly validate the use of such aggressive design rules.

7.3.1 SRAM Cell Read and Write Margins

The commonly used six transistor SRAM cell is shown in Figure 7.3a. The figure also illustrates the layout of the SRAM cells, designed on a 130 nm foundry technology. The key SRAM cell design requirements are to ensure writeability and read margin. To this end, the SRAM cell device sizes are a compromise between those that result in the smallest cell but still provide adequate read and write margins. When sizing the transistors, all process, voltage, and temperature (PVT) corners must be considered. The SRAM transistors are small enough that random dopant fluctuations have



(a) Types 1, 2, and 4

(c) Type 1

(d) Type 1 with Guard Ring

(b) Type 1 with Reverse-Body Bias

(e) Type 3

(f) Type 3

(g) Type 4

FIGURE 7.3 Hardened and unhardened 130 nm SRAM cell designs. 4-NMOS, 2-PMOS standard SRAM design (a) and with body bias capability (b). Layout type 1 and type 1 with guard ring (c) and (d) reach near foundry densities, but using annular NMOS pull downs (g) is much larger. Using 4-PMOS and 2-NMOS transistors (e) does not save area since the PMOS devices must be very large to overpower the edgeless NMOS transistors to write the cell (f). (Reprinted with permission from Clark, L.T., Mohr, K.C., Holbert, K.E., Xiaoyin Yao, Knudsen, J., Shah, H., "Optimizing radiation hard by design SRAM cells." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036.)

a considerable effect on the actual cell margins [34]. The large on-die memory and cache sizes noted in Section 7.1.1 mandate that 10–14 sigma manufacturing variability must be considered. To comprehend the increasing variability and diminishing cell margins, a number of statistical methodologies for SRAM cell and array design have been proposed [35,36]. The difficulty in designing commercial SRAM cells with adequate margins points to the importance of advanced techniques for RHBD memory cells, since there are fewer validation resources available for the small rad-hard market and the impact of more difficult to pattern and fabricate annular gate geometries must be comprehended.

The cell is written differentially, where one BL is at the high precharge potential and the other (BLN) is driven low (or vice versa). During a write, the NMOS access transistor (NP1) must overpower the PMOS pull up transistor—the cell is a ratioed circuit during writes. Adequate write margin requires that the access transistor NP1 in Figure 7.1 be stronger than the pull-up device P1. The write margin is typically defined in one of two ways. The DC approach is to measure the BL voltage required to flip the SRAM cell state, by keeping BLN high and lowering BL from V_{DD} toward V_{SS} until the cell state is flipped. Alternatively, the delay to write the cell when the BL is driven to V_{SS} may be measured [37].

When the SRAM is read, the low storage node rises due to the voltage divider composed of the two series NMOS transistors in the read current path (N0 and NP0 in Figure 7.3a). The storage node C is between them, rising above V_{SS} during a read. This reduces the SRAM cell static noise margin (SNM) as measured by the smallest side of the square with largest diagonal that can fit in the small side of the static voltage curves [38]. The worst-case SNM, shown in Figure 7.4, is usually determined by Monte Carlo simulation or response surface models [39] based on the measured process variation

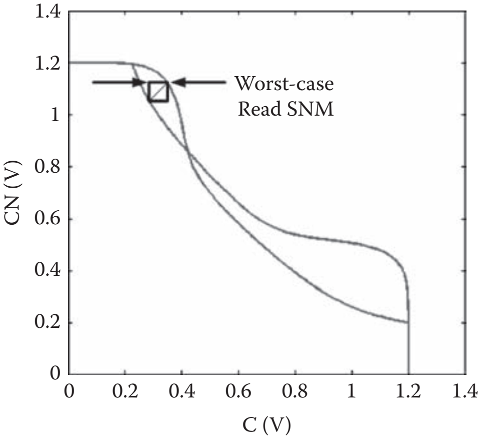


FIGURE 7.4 SRAM static noise margin (SNM) plot. The line depicts the worst-case found in a Monte Carlo simulation to five sigma variations. Note the severe asymmetry. (Reprinted with permission from Clark, L.T.; Mohr, K.C., Holbert, K.E., Xiaoyin Yao, Knudsen, J., Shah, H., "Optimizing radiation hard by design SRAM cells." *IEEE Transaction. on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036.)

parameters. For the unhardened cell simulated here, the SNM in Figure 7.4 is quite small at 58 mV, even at the nominal V_{DD} of 1.2 V. The large transistor mismatch due to both systematic (die-to-die) and random (within-die) variation causes asymmetry in the SNM plot. Other noise margin definitions have been developed, based on imbalance created across the cell by disturb voltages [37]. Read margin is also ensured by transistor sizing. Typically, the pull-down NMOS transistors are drawn wider than the access transistors. The access transistor NP0 is also frequently drawn with a longer channel than that of the pull-down N0 and pull-up P0. Consequently, there is limited design latitude—the PMOS pull-up must be weaker than the NMOS access device, which in turn must be weaker than the NMOS pull-down transistor.

7.3.2 Reverse-Body Bias

Reverse-body bias (RBB) with and without simultaneous power supply collapse has been used in commercial integrated circuits for full-chip [40,41] and memory [42,43] standby power reduction. RBB is presently used on commercial ICs for low standby power (LSP) state retention on processes varying from 250 nm through 65 nm process generations [40,42,44,45]. These modes often combine RBB and supply collapse, here termed RBB + SC, since the latter helps to mitigate emerging transistor leakage paths such as direct band-to-band tunneling through the gate oxide [45]. RBB electrically increases the transistor threshold voltage by the body effect, which is also applied to the parasitic edge and field-oxide field-effect transistor (FET). Its use for TID mitigation has been demonstrated at the transistor level on low V_{th} 0.35 μm bulk CMOS transistors [46]. RBB mitigation of TID on advanced fabrication process technologies allows the use of the smallest foundry optimized cells and thus eliminates RHBD SRAM size penalties.

7.3.3 RHBD SRAM Cell Design

A number of RHBD SRAM cell designs have been investigated on 130 nm and 90 nm technologies [47,48]. Since, as mentioned, any SRAM cell must have adequate write margin and read stability for a fair analysis these are constrained to be as good or better than that of the baseline two-edge foundry cells when evaluating potential circuit topologies and layouts. A number of potential SRAM cell schematics and layouts are shown in Figure 7.3, which were used in a 130 nm study [48].

The type 1 cell (see Figures 7.3a and 7.3c) is essentially a conventional (commercial nonrad-hard) design with two-edge NMOS pull-down transistors and two-edge NMOS access devices. Figure 7.3b shows a variation that separates the NMOS source from the substrate taps (i.e., allowing RBB). Since in SRAM arrays the well and substrate taps are placed in special tap rows (or columns for vertical well designs) RBB support does not add size to the cell, as shown in Figure 7.3c. Note, however, that in an RHBD IC these tap spacings must be considerably smaller to avoid SEL.

The type 2 SRAM cell employs annular NMOS pull down transistors and annular NMOS pass gates. The type 3 SRAM cell uses edgeless NMOS pull-downs and two-edge PMOS access transistors (Figures 7.3e and 7.3f). The gate bias dependence of NMOS transistor TID degradation, where the leakage current increase is suppressed when the gate is biased at 0 V so the electric fields do not repel the positive trapped charge toward the oxide/Si interface, suggests the type 4 SRAM cell (see Figures 7.3a and 7.3g). It has annular NMOS pull-down transistors and two-edge NMOS access transistors. This variation relies on that fact that most of the time all but one SRAM WL are deasserted at 0 V.

The type 2 through type 4 cells include PMOS guard rings between the NMOS transistor drains and the n-well to limit TID-induced leakage between the two. No guard rings are used to limit leakage between the NMOS drains at different potentials in any of the cell designs investigated. For instance, in the type 4 cell shown in Figure 7.3g, the NMOS pull-down sources, at V_{SS}, are near the access transistor drains and not separated by a *p*-type guard ring. If these guard rings are necessary, the cells must grow to accommodate them.

For each of the designs in Figure 7.3, simulations were used to determine write margins and read stability. Since write margin can be increased at the expense of read stability, the designs are optimized by minimizing total transistor width at similar write margin but forcing read stability to meet the baseline set by the foundry SRAM cell. This analysis assumes that the total cell size is proportional to the required transistor widths.

7.3.3.1 Conventional Two-Edged Transistor Cell (Type 1)

In a conventional SRAM design where all transistors are two-edged, all devices can be drawn at or near minimum width as in the commercial foundry cell. The large NMOS to PMOS mobility ratio and use of minimum width PMOS pull-ups provide adequate write margin. Adding a guard ring to mitigate SEL and TID-induced NMOS drain to n-well leakage increases the cell size by about 20%, as evident in Figure 7.3d. If guard rings are unnecessary, standard production SRAMs employing even tighter design rules and smaller cell size can be used—the foundry supplied 130 nm cell is 27% smaller than the cell in Figure 7.1c, which is drawn to the logic layout rules.

7.3.3.2 Annular NMOS Based SRAM Cell (Type 2)

The simple analysis shows that using a conventional cell with four annular NMOS and two PMOS transistors results in a total transistor width approximately 7× that of the conventional two-edge cell at the same write margin and read stability. While annular NMOS layout eliminates the source-to-drain leakage path formed at the shallow trench isolation (STI) to channel interface, their greater minimum size increases the preirradiation cell leakage commensurately. One potential design is shown in Figure 7.5. This cell, implemented on a 90 nm foundry bulk CMOS technology, is 5.1 times the size of the foundry cell, which uses tighter SRAM design rules, and 3.6 times the size of a cell drawn to the same (90 nm) logic design rules. Of this, about 20% of the size is attributable to providing portability between process versions, which have different gate lengths. Thus, the cell could have been 20% smaller if portability were not a requirement. Another 20% is attributable to the guard rings, similar to the impact on the 130 nm cells previously described. The aspect ratio helps this, since the *p*-type guard rings are oriented vertically. The wide cell increases the critical node spacing in the key horizontal dimension, making a column group wider, with the same n:1 Y multiplexing.

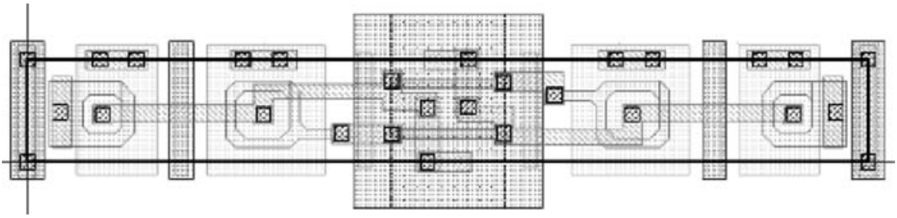


FIGURE 7.5 Full edgeless 90 nm NMOS SRAM cell layout. This cell is 5.1x larger than the smallest available foundry cell on this 90 nm process, due to compatibility with the LSP version with longer gate length, *p*-type guard rings, and annular NMOS gates.

The particular RHBD NMOS layout choice also affects the transistor leakage. For the "ring" topology used in the NMOS transistors in Figures 7.3f, 7.3g, and 7.5, placing the transistor drain inside the ring and source outside minimizes the area used as the sources can be shared between cells or transistors within the cell. Extra guard rings are then unnecessary. Transistor I_{DS} versus V_{DS} measurements of ring gate transistors show 3× greater I_{OFF} and reduced I_{DSAT} with the source inside the ring [48]. The former is due to much higher drain-induced barrier lowering (DIBL) that is a function of the drain/gate versus source/gate interface widths. It is therefore important when determining stability to account for the specific gate geometries used and their particular I_{DS} versus bias characteristics. Standard foundry transistor models do not provide this level of modeling detail, particularly for the edgeless gate geometries, so appropriate test transistor arrays must be used for modeling and validation.

7.3.3.3 PMOS Access Transistor SRAM Cell (Type 3)

The TID immunity of two-edge PMOS transistors on modern processes suggests their use as the SRAM access transistors. However, to write the cell, the PMOS pass transistors must overpower the NMOS pull-downs, which is made difficult by the large NMOS/PMOS mobility difference. Consequently, using annular NMOS pull-downs implies wide PMOS access transistors as shown in Figure 7.3f. In this cell a NMOS long-channel "ring" gate pull-down transistor is used to make it as weak as possible, as the channel comprises only one edge of the NMOS transistor and has a long channel length, *L*, and narrow width, *W*, for a low *WL* ratio. The SRAM cell area is still large. One advantage of the large NMOS gate area is increased capacitance and thus increased Q_{crit}. For the conventional bulk CMOS processes used here, this cell is very large and has sufficient drawbacks that further analysis is unnecessary. Future processes may incorporate so-called hybrid orientation transistors (HOTs) that promise similar PMOS and NMOS mobilities [49]. For such a fabrication process, this cell topology may be very good.

7.3.3.4 Two-Edged NMOS Access Transistor SRAM Cell with Annular Pull-Down Transistors (Type 4)

Using two-edged NMOS access transistors (NP0 and NP1 in Figure 7.3a) and annular pull-down transistors N0 and N1 is an interesting alternative since the *WL* is high for only the row being accessed. A low gate voltage minimizes TID-induced leakage at the transistor edges, and this is the bias condition on access transistors NP0 and NP1 over 99% of the time. This cell has good stability and write margin even for narrow access transistor widths. Cell size is reduced, as shown in Figure 7.3g, but is still significantly larger than the conventional SRAM cell. A drawback to this configuration is that, depending on the layout, the access transistor source and drain nodes may be difficult to shield from the other cell transistors using guard rings. Fortunately, on sub-150 nm processes, this TID component may be tolerable or mostly mitigated by avoiding polysilicon crossing from n+ to n-well diffusions [50].

7.4 SNM TEST STRUCTURE

Since read stability is so important and since it is unlikely that an RHBD IC designer will be able to include RHBD SRAM cells on the process development test chips as the standard foundry cells are, appropriate test structures to rapidly determine the cell quality are essential. A test structure that allows direct measurement of the static noise margin of individual SRAM cells is shown schematically in Figure 7.6. It is based on the standard SRAM array and can be integrated with a production design. To allow accurate analog signal propagation, two supply voltages, V_{DD} and V_{DD_CELL}, are used. V_{DD} is independent of V_{DD_CELL} allowing the gate overdrive of the cell and access devices to be controlled independently. By applying $V_{DD} > V_{DD\_CELL}$, the resistance of the analog signal path multiplexers is reduced, limiting their affect on the measurements. During the test, nodes WL and the access multiplexer enables are asserted high. The high WL voltage allows single-ended writes of the SRAM cell, unlike the normal operating condition, where writes must be differential. The test is DC, so there is no time dependence in the measurement. Consequently, the BL and BLN voltages, when used as outputs, accurately represent those of the SRAM cell storage nodes C and CN, respectively. Thus the circuit allows direct measurement of the as-fabricated SRAM cell p-n ratios through observation of the switching points when driving the BL (or BLN) high or low.

An analog multiplexer under software control is used to connect the test structure to a digital to analog converter. The multiplexer allows switching either the BL or BLN attached to the FPGA driver, with its complement BLN or BL attached to the analog to digital converter to measure the cell state. The measurements can be made with the device under test (DUT) inside the Co-60 irradiator so measurements versus TID can be made in situ, allowing determination of the TID impact on the individual SRAM cell read and write characteristics. Measuring the DUT in situ, that is, while being irradiated, avoids relaxation of the TID effects that would occur when removing the device from the irradiator to make a measurement. Measured TID results from unhardened and hardened SRAM cells are presented in the following section.

7.5 EXPERIMENTAL TID TESTING RESULTS

Test die were fabricated on both 130 nm and 90 nm CMOS bulk processes at the same foundry. The test die included both SRAM arrays and transistor test structures.

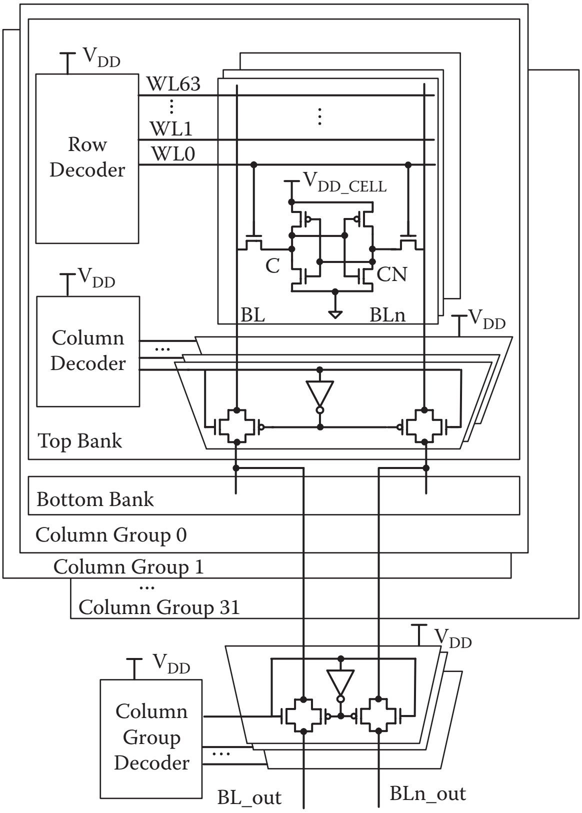


FIGURE 7.6 90 nm test array allowing measurement of individual cell margins. (Reprinted with permission from Xiaoyin Yao, Hindman, N., Clark, L.T,; Holbert, K.E., Alexander, D.R., Shedd, W.M., "The impact of total ionizing dose on unhardened SRAM cell margins". *IEEE Transactions on Nuclear Science*, Volume: 55 Issue: 6, 3280–3287.)

The latter are laid out in the SRAM cell layouts, so the results are representative of the device responses that would occur in the actual SRAM arrays. SRAM test structures with annular pull-down transistors (type 4) were also fabricated.

A 90 nm 1.2 Mb SRAM design fabricated on a low leakage (low standby power) variation of the process can provide a baseline for the discussion [17]. This SRAM exhibited a 131 times increase in I_{SB} after irradiation to 1 Mrad(Si), as shown in Figure 7.7. Note, however, that significant leakage increase does not occur until after 300 krad(Si), which may be sufficient for many spaceborne IC applications. The design is fully functional despite this large leakage increase after a 1 Mrad(Si) dose. While prior SRAM designs have failed at relatively low TID levels [16] careful circuit design can avoid this. In general, since size scaling requires increasing doping levels and V_{DD} scaling requires lower V_{th} to maintain gate overdrive, smaller geometry processes exhibit less TID-induced I_{SB} increases. The higher I_{OFF} and I_{gate} leakage currents in more highly scaled processes tend to mask what TID-induced increase there is until higher doses.

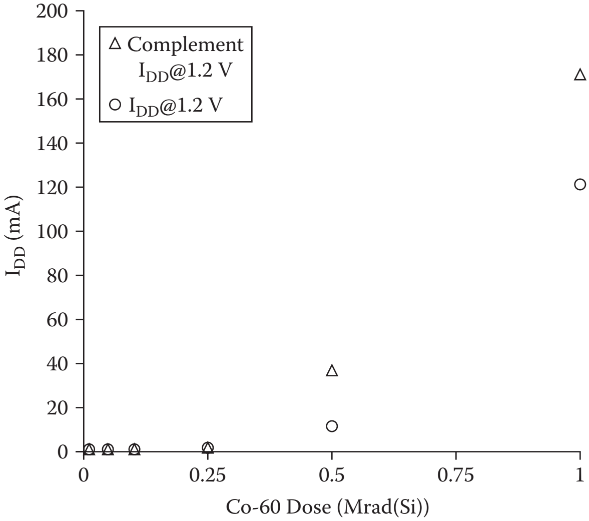


FIGURE 7.7 90 nm 1.2 Mb conventional (non-hardened) SRAM TID results. The increase at 1 Mrad(Si) for the array in the state opposite that when irradiated, shows a 131x increase in standby I_{DD} (I_{SB}) for this device fabricated on the LSP process version. (Reprinted with permission from Xiaoyin Yao, Hindman, N., Clark, L.T., Holbert, K.E., Alexander, D.R., Shedd, W.M., "The impact of total ionizing dose on unhardened SRAM cell margins." *IEEE Transactions on Nuclear Science*, Volume: 55 Issue: 6, 3280–3287.)

It is important to know the exact SRAM cell organization to apply the worst-case TID conditions. In particular, horizontally adjacent cells may have adjacent NMOS diffusions biased the same or differently with a solid or checkerboard pattern. This is not just a matter of geographic cell location but also a function of whether the BL and BLN are stepped or folded in the layout. For example, the 90 nm SRAM uses a pattern BL0 BLN0, BL1 BLN1,…BL7 BLN7. However, the 130 nm design uses BL0 BLN0, BLN1 BL1,…BLN7 BL7. In the former case, a solid array pattern of all 1's or 0's is the worst case for TID leakage increase, while in the latter a *physical* checkerboard is. Finally, the physical and logical organization can be quite different, so knowledge of the physical layout is critical here, as it is in choosing appropriate production SRAM test patterns.

7.5.1 Impact of V_{DD} Bias on TID Response

The 1.2 Mb 90 nm SRAM, fabricated on the foundry LSP process version, irradiation results were already described. Additionally, 5 kB SRAM test arrays were fabricated on the standard process version that supports the shorter gate length. The test SRAMs include an array without RBB and an array with RBB capability, with the latter configured with node SOURCE (see Figure 7.3b) biased at V_{SS} during these initial irradiations. Two bias conditions, V_{DD} = 1.0 V and 1.3 V, were used (see Figure 7.8). The I_{SB} is normalized to the initial values for each die, which exhibit substantial (and expected) die to die variations. The V_{DD} = 1.3 V TID-induced I_{SB}

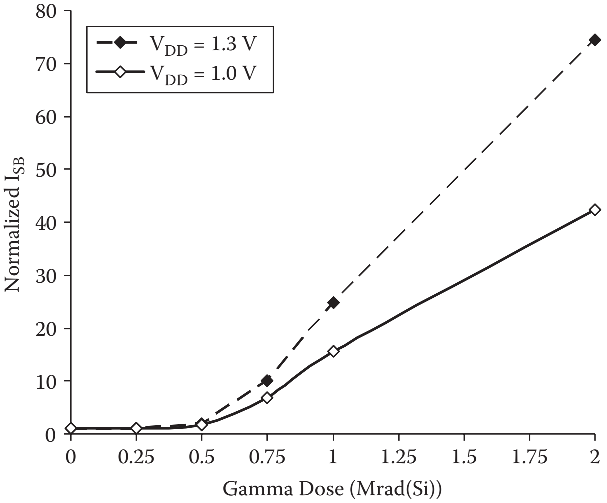


FIGURE 7.8 Measured effect of V_{DD} on the TID induced standby I_{DD} (I_{SB}) increase in 90 nm 5kB SRAM array fabricated on the standard process. The 2 Mrad(Si) value is 75x the initial I_{SB}. (Reprinted with permission from Clark, L.T.; Mohr, K.C., Holbert, K.E., Xiaoyin Ya, Knudsen, J., Shah, H,. "Optimizing radiation hard by design SRAM cells." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036.)

increase of 75x is 1.8x times the TID-induced increase at V_{DD} = 1.0 V. This indicates that if sufficient SEU tolerance can be provided, there is substantial TID response benefit to low V_{DD} operation. The TID leakage increase at 1 Mrad(Si) is larger on the LSP process is in whole or part due to the substantially lower original leakage.

7.5.2 Impact of TID on Cell Margins

Using the previously described test structure in Section 3.4 (see Figure 7.6) a 90 nm LSP 4 kB test array of unhardened SRAM cells was exposed to Co-60 radiation at an approximate rate of 20 rad/sec [17]. During this exposure, the BL switch points were measured continuously, and during irradiation (after each measurement switched the cells) the cells were rewritten to the state where node C is 0 V and node CN is 1 V. The BL switch point response versus the applied dose is plotted in Figure 7.9, where the SRAM node is being pulled high by the BL. It is unaffected until about 300 krad(Si), where TID-induced leakage becomes significant compared with the inherent leakage components, consistent with the full SRAM results in Figure 7.7. The TID impact on the measured switching voltage and hence cell write margins saturates near 1.5 Mrad(Si).

At doses of 1.5 Mrad(Si) and 3.0 Mrad(Si) both the BL and BLN switch points were measured, indicating a strong downward shift in the BLN switch point across the cells, again indicative of a shift in the cell effective p-n ratios due to the irradiation. The cells were irradiated with the gate of transistor N0 high, and as expected this device exhibits the most degradation, that is, increased leakage due to TID. The SRAM access transistors, NP0 and NP1, are assumed to be largely unaffected, since

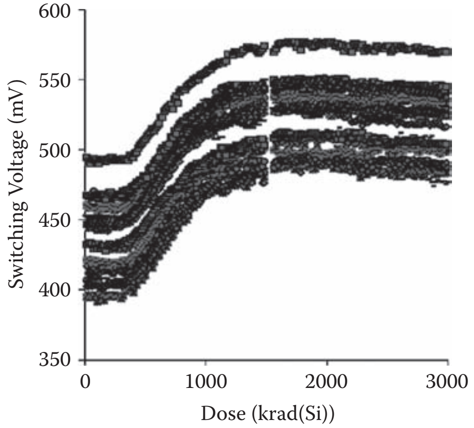


FIGURE 7.9 Behavior of sample 2-edge (unhardened) SRAM cell trip points vs TID. (Reprinted with permission from, Xiaoyin Yao, Hindman, N., Clark, L.T., Holbert, K.E., Alexander, D.R., Shedd, W.M., "The impact of total ionizing dose on unhardened SRAM cell margins." *IEEE Transactions on Nuclear Science*, Volume: 55 Issue: 6, 3280–3287.)

they have 0 V gate bias most of the time. One key observation is a lower BLN voltage required to write the cell state, indicating diminishing write margin—that greater drive is required to write the cell in this direction over time. This result, consistent with an increase in the drive of transistor N0, presents a possible failure mechanism due to TID.

The impact on the SRAM cell read margins is shown in Figure 7.10, which compares the pre- and post-irradiation SNM as simulated by changing the leakage of transistor N0 to match the TID measurement results. The test structure does not allow direct measurement of the SRAM read SNM, which must be inferred from the write margin measurement results. To determine the impact of TID on the read margin response, the NMOS response was modeled from transistor TID measurements on the same process. Two responses with the degradation on each NMOS pull-down transistor were simulated independently. Immediately evident is the closing of the larger "eye" post-TID. In one case (the dashed lines), the initially weaker NMOS pull-down transistor is made slightly stronger by the TID-induced leakage, and the worst-case read SNM is slightly improved (from 58 to 59 mV). The read SNM on the other node is diminished to 53.7 mV (note the smaller "eye" at the top outlined by the thin lines) when the TID-induced increased leakage is on the initially stronger NMOS pull-down transistor. Whether TID mitigation is necessary to maintain SRAM cell read margins is thus determined by the initial as fabricated margins—a larger cell with large margins may still be smaller than that required by annular transistor layout and guard rings, as well as the TID environment expected.

The TID switching point response versus irradiation dose of the RHBD cell of Figure 7.5 is shown in Figure 7.11. Since the leakage currents are completely mitigated, as indicated by I_{SB} measurements versus TID, the switch points are stable up to 2 Mrad(Si). Clearly, allowing sufficient cell size, the RHBD techniques are effective.

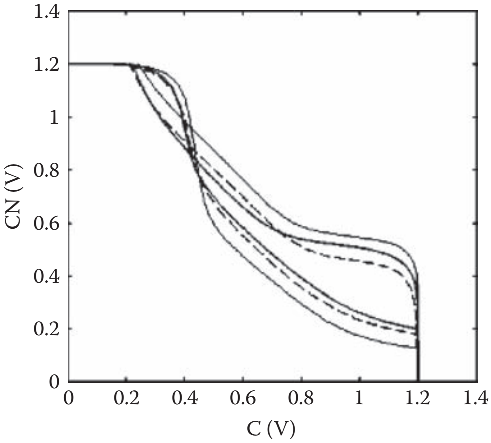


FIGURE 7.10 Simulated worst-case Monte Carlo derived read SNM pre and post-irradiation. The thin solid and thin dashed lines show the post-irradiation SNM, while the thick grey lines show the pre-irradiation response. (Reprinted with permission from Xiaoyin Yao, Hindman, N.; Clark, L.T., Holbert, K.E., Alexander, D.R., Shedd, W.M., "The impact of total ionizing dose on unhardened SRAM cell margins." *IEEE Transactions on Nuclear Science*, Volume: 55 Issue: 6, 3280–3287.)

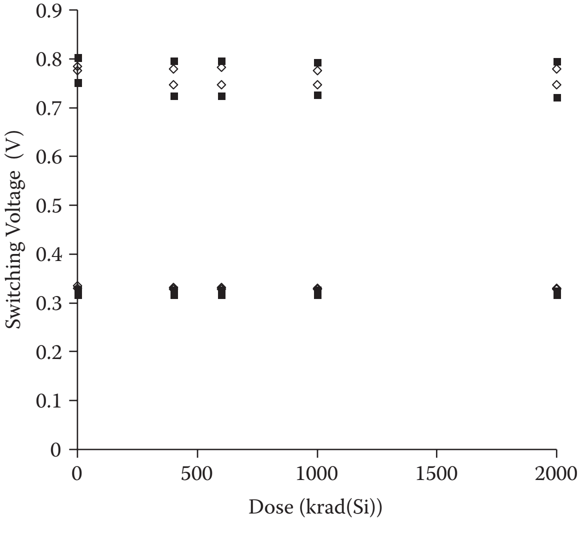


FIGURE 7.11 Sample SRAM cell trip points vs TID for the all annular 90 nm NMOS SRAM cell. No variation due to irradiation is measured for this cell.

7.5.3 Type 4 Cell

In general, a worst-case experiment uses a high NMOS gate voltage to maximize TID effects. However, this condition simply cannot occur on the NMOS access transistors in an SRAM, as the WL decoder ensures that only one can be active and then only during one clock phase. To validate that WL bias at 0 V for all unaccessed cells is sufficient to mitigate TID-induced leakage in the NMOS pass devices NP0 and NP1, experiments were made using test transistors fabricated on both the 130 nm and 90 nm processes. On the former, four 0.28 μm width, minimum length two-edge NMOS transistors connected in parallel with $V_{DS} = 1.2$ V and $V_{GS} = 0$ V, the relevant access transistor bias, exhibit TID-induced increase in I_{OFF} less than 3 times after exposure to 500 krad(Si). A 1.5 μm effective width annular transistor (the minimum for this geometry) has 5 times greater I_{OFF} current preirradiation than this narrower two-edge device. These experiments indicate that 130 nm SRAM cells using annular pull-down NMOS and two-edge pass transistors exhibit less total leakage current after exposure to 500 krad(Si), with less area than a cell using PMOS access transistors. The same experiments were carried out on transistors fabricated on a 90 nm foundry process to 1 Mrad(Si) and indicate that for the off-state bias condition, the TID-induced I_{OFF} increase is less than 2 times [47].

7.5.4 Type 1 Cell with RBB—Array Design Considerations

The low WL activity factor makes annular gates less important for the SRAM cell access transistors, as already shown experimentally, but TID-induced leakage remains an issue if two-edge NMOS pull-down transistors are used. This can be mitigated by applying RBB or RBB + SC. By setting V_{DD} to 1.2 V and the external V_{SOURCE} to 0.7 V, the SRAM cells have 0.5 V V_{DS} storing the SRAM cell state. The SRAM cell supply voltage can be varied by raising the NMOS sources (cell node SOURCE in Figure 7.3b) and row by row (SOURCE0 to SOURCE63 in Figure 7.12a) while maintaining the NMOS transistor bulk connection at 0 V. This allows RBB + SC to reduce the NMOS transistor leakage in the pull-down transistors N0 and N1 as well as significantly reducing I_{OFF} in transistors NP0 and NP1 through negative V_{GS}. Alternatively, since the channel surface potential is pinned, that is, the gate fields are unaffected by the bulk potential, V_{DD} and V_{SOURCE} can be raised to provide reduction in TID-induced leakage without affecting the V_{DS} and hence cell Q_{crit}.

Low cell voltages during operation reduce TID effects, as already shown, but modern SRAM cells are not read stable at low voltages. Consequently, to employ reduced biases for TID mitigation the cell bias must be changed dynamically to full V_{DD} during reads. The circuits providing this ability are shown in Figure 7.12. By driving the row SOURCE node to 0 V dynamically before the WL is selected, the SRAM cells in that row can be read without upset that might otherwise occur since SNM diminishes rapidly with decreasing V_{DD}. Sufficient address setup time ensures that the row SOURCE node is driven to 0 V before the WL is asserted. The raised source structure was chosen since it can apply RBB with power supply collapse. This is applied dynamically to allow full read stability in the selected row. This configuration can also simulate a triple-well SRAM by using the appropriate bias conditions,

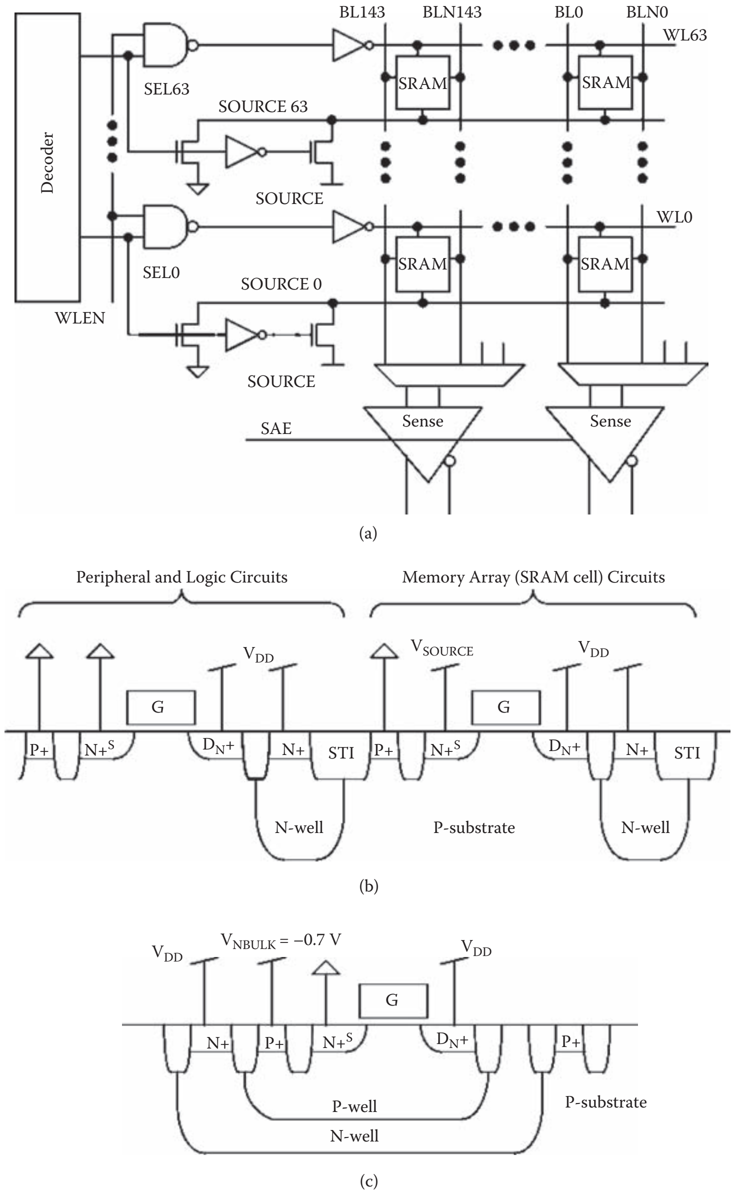


FIGURE 7.12 The circuits providing dynamic RBB in the 130 nm and 90 nm SRAMs for NMOS transistor TID mitigation (a). The connections of the well and substrates for the periphery are shown in (b). A triple-well process allows continuous negative NMOS transistor bulk bias (c). These well bias conditions were simulated in the arrays by applying high V_{DD} to the peripheral circuits and raising V_{SS} above the bulk voltage in some of the TID experiments. (Reprinted with permission from Clark, L.T., Mohr, K.C., Holbert, K.E., Xiaoyin Yao, Knudsen, J., Shah, H., "Optimizing radiation hard by design SRAM cells." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036.)

as shown in Figures 7.12b and 7.12c, by holding the bulk (p-substrate) at 0 V and making V_{SS} and V_{DD} 0.7 V and 1.7 V, respectively.

7.5.5 Type 1 Cell with RBB—Transistor Level Measurements

Four 0.28 μm wide 130 nm minimum length NMOS transistors connected in parallel were irradiated with $V_{DS} = V_{GS} = 1.2$ V ($V_{DD} = 1.9$ V), V_S (the transistor source) = 0.7 V, and $V_{bulk} = 0$ V, the worst-case condition for the raised V_{SS} TID mitigation scheme. The I_{OFF} was measured preirradiation and after a total dose of 750 krad(Si). An operational SRAM will have data that change, so this was applied as 250 krad(Si) biased on ($V_G = V_{DD}$), 250 krad(Si) biased off ($V_G = V_S$), and then 250 krad(Si) biased on ($V_G = V_{DD}$). The 130 nm NMOS transistors exhibit an 8× increase in TID-induced leakage, as opposed to over 200× increase for $V_{DD} = 1.2$ V and no RBB applied, as shown in Figure 7.13. The pull-down transistors account for about 1/3 of the total leakage, so when this result is combined with the access transistor TID response, the SRAM cell will exhibit about 3× I_{OFF} increase after exposure to 750 krad(Si). The post-irradiation I_{OFF} with $V_{SB} = 0.7$ V and $V_{GS} = V_{DS} = 0.5$ V is nearly one order of magnitude less than the preirradiation I_{OFF} at $V_{SB} = 0$ V and $V_{GS} = V_{DS} = 1.2$ V. By using dynamic source biasing, a conventional SRAM cell will exhibit less post-TID leakage than cells employing annular gates preirradiation. The I_{SB} for this condition is lower at 1 Mrad(Si) than the preirradiation annular (type 4) SRAM. The voltage-collapsed SRAM with RBB applied has lower I_{SB} at 1 Mrad(Si) than the conventional two-edge transistor SRAM (type 1) cell has preirradiation. We attribute the annular (type 4) SRAM cell I_{SB} increase to leakage under the field oxide at the two-edge access transistors, since, as mentioned, no cells had guard rings between adjacent n-type source and drain diffusions.

7.5.6 Test SRAM Designs and Experiments

The 5 kB RHBD test SRAMs implemented in 0.13 μm bulk CMOS contain 4 kB for data and 1 kB for EDAC parity bits. A single 40-bit read and write port is organized as 32 data bits plus 8 EDAC bits. EDAC protects against SEU in the array while dual redundant control lines are used to detect and prevent SET data corruption due to word line SET, described as follows. The 90 nm SRAM design uses similar circuits to apply RBB and supports the same (40, 32) single error correct, double error detect EDAC. Both interleave the storage bits by 8 cells to avoid multiple bit error upsets in the same EDAC code word. The 90 nm design does not mitigate SET-induced errors.

In commercial designs, the most important leakage-induced failure is the case where during a read operation, a single bit is driving the BL high, but leakage on the other cells is driving the same BL low, significantly slowing the BLN – BL voltage signal development. If the total BL leakage approaches the cell read current, small signal differential sensing may fail. This is even more important in memory designs that use single-ended sensing, since the output high bit-line node may register as a logic 0 after being discharged by leakage within the timing window. This failure mechanism is avoided in the designs here by using full swing differential sensing and relatively short BLs with 64 cells attached as well as the high I_{ON}/I_{OFF} ratio of the foundry process.

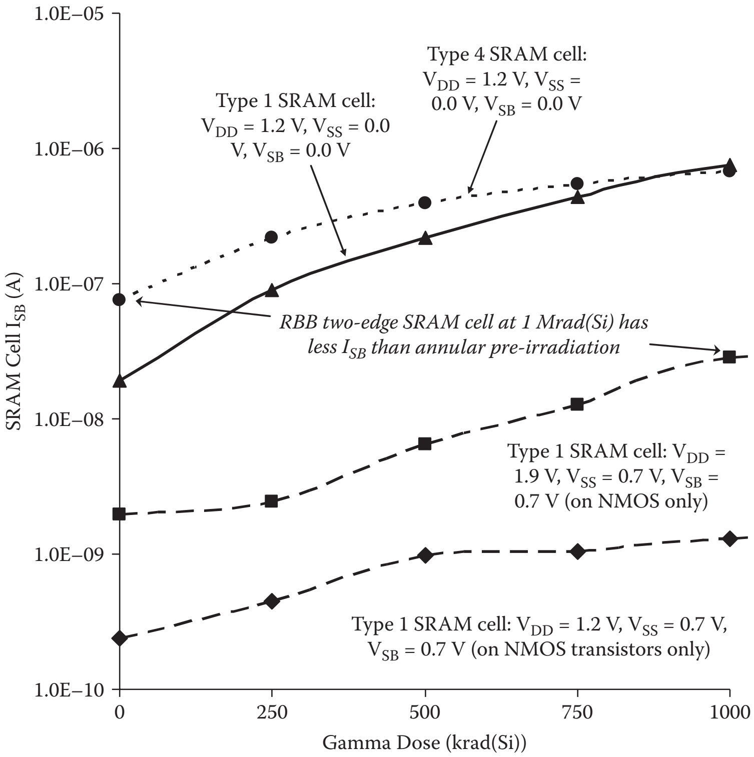


FIGURE 7.13 Co-60 irradiation results of different SRAM cell I_{SB} responses. Note that the type 4 cell, which has annular NMOS pull down transistors, has higher I_{SB} pre-irradiation than the type 1 with RBB and full V_{DS} post-irradiation. The higher pre-irradiation leakage of the type 4 eliminates much of its advantage over the type 1 (unhardened) cell. The type 4 cell has leakage increase attributable to the lack of guard rings blocking STI leakage paths. The type 1 with RBB+SC has little leakage increase and reduced I_{SB} at all dose levels. (Reprinted with permission from Clark, L.T.; Mohr, K.C., Holbert, K.E., Xiaoyin Yao, Knudsen, J., Shah, H., "Optimizing radiation hard by design SRAM cells." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036.)

The 90 nm SRAM design uses cross-coupled NAND gate set-reset latches to sense. This allows a robust timer-free sense, eliminating one timing signal that could otherwise be upset by a SET. The SRAM cell high node provides current through the on pass transistor (NP0 in Figure 7.3) clamping the BL reading a logic "1" at approximately $V_{DD} - V_{tN}$. Meanwhile, the BLN reading a logic '0' discharges completely to V_{SS}. The 130 nm design uses bit-line keepers to hold the non-discharging bit-line to V_{DD}.

7.5.7 Type 1 Cell with RBB—SRAM Measurements

Triple well processes, which are common, allow RBB to be applied statically without collapsing V_{DS} (see Figure 7.12c). Full cell V_{DS} maintains the cell Q_{crit}, allowing better

SEU response. The effect of this bias condition on TID response was investigated in an experiment comparing the 130 nm SRAM TID performance on a nontriple-well process with V_{DD} = 1.9 V, V_{DS} = 1.2 V, and V_{SB} = 0.7 V (referring to Figure 7.12, V_{DD} = 1.2 V and V_{SOURCE} = –0.7 V) to the case without RBB. The experimentally measured SRAM array I_{SB} versus irradiation level on the 130 nm test chip (Figure 7.14) also clearly shows that RBB is a viable approach to mitigate TID-induced increase in I_{SB} up to 1 Mrad(Si). Since measuring I_{SB} encompasses all leakage components, both I_{OFF} and conduction under the STI field oxide is mitigated by RBB. This allows the use of the foundry cells. Both RBB and RBB + SC improve I_{SB} compared with the standard bias and RBB + SC reduces leakage sufficiently such that the total I_{SB} of the array drops below that of the SRAM control logic, indicated by the dashed line at the bottom of Figure 7.14. The control logic was not reverse-body biased. Consequently, further decreases in SRAM array I_{SB} will not provide significant improvements in the overall SRAM static power unless all circuits have RBB applied.

Note that using RBB + SC allows the post-irradiation leakage to be less than the preirradiation leakage without it. On this 130 nm SRAM, some single-bit failures were observed for the standard bias condition starting at 725 krads(Si), indicating that increased leakage currents had destabilized some of the SRAM cells, presumably

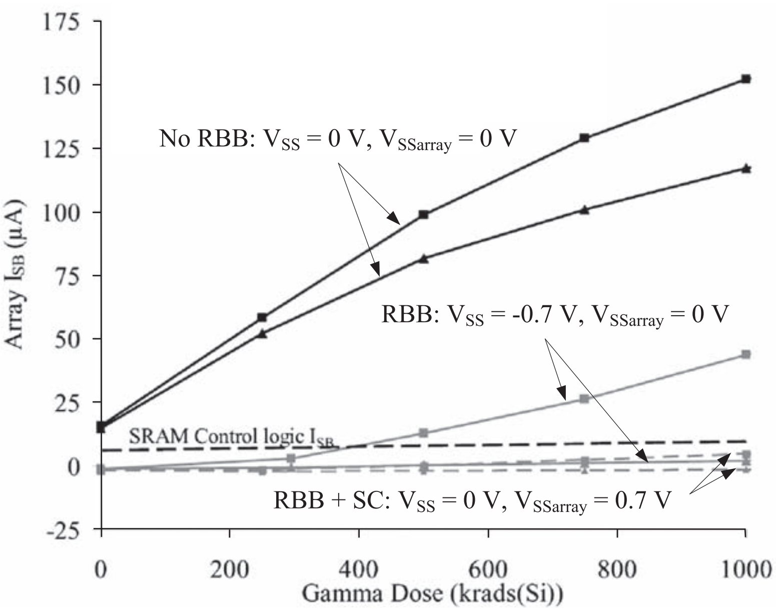


FIGURE 7.14 130 nm SRAM array I_{SB} vs. TID and irradiation bias for chips irradiated with and without RBB, and with RBB+SC. I_{SB} was measured at the irradiation bias. Irradiation was performed with the array programmed to a checkerboard pattern, and I_{SB} measured with the same [triangles] and opposite [squares] patterns. (Reprinted with permission from, Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

those that were least stable to begin with. No failures were observed for the RBB + SC bias up to 1 Mrad(Si).

This TID mitigation approach was also investigated on a 90 nm bulk CMOS 5 kB SRAM using two-edged transistor cells. Test SRAMs were irradiated with V_{DD} = 0.9 V (the non-RBB case, i.e., with V_{SB} = 0 V) and with V_{DD} = 1.6 V and node SOURCE—the source of the SRAM NMOS pull-down transistors (N0 and N1 in Figure 7.3b)—at 0.7 V and the *p*-type bulk at 0 V (the RBB case). This latter bias condition applies a 0.7 V NMOS RBB with the cell V_{DS} = 0.9 V, the same as the non-RBB case. The SRAM array SOURCE node (see Figure 7.12a) current was measured with all 1's and with all 0's stored in the array to determine the NMOS pull-down transistor I_{OFF} pre- and post-irradiation for both cases. The SRAM was written with all 1's during irradiation. The measured results are shown in Figure 7.15 for TID up to 1 Mrad(Si) [47]. While the non-RBB SRAM I_{SOURCE} increases by 10× for array data opposite to that during irradiation, no increase is observed in the RBB I_{SOURCE} at that irradiation level. Between 0 and 500 krad(Si), the high intrinsic leakage delays the onset of noticeable TID impact to higher irradiation.

The same measurements with irradiation at V_{DD} = 1.6 V and node SOURCE = 0.6 V are shown in Figure 7.16 [47]. I_{SOURCE} with RBB of 0.5 V and 0.6 V shows the

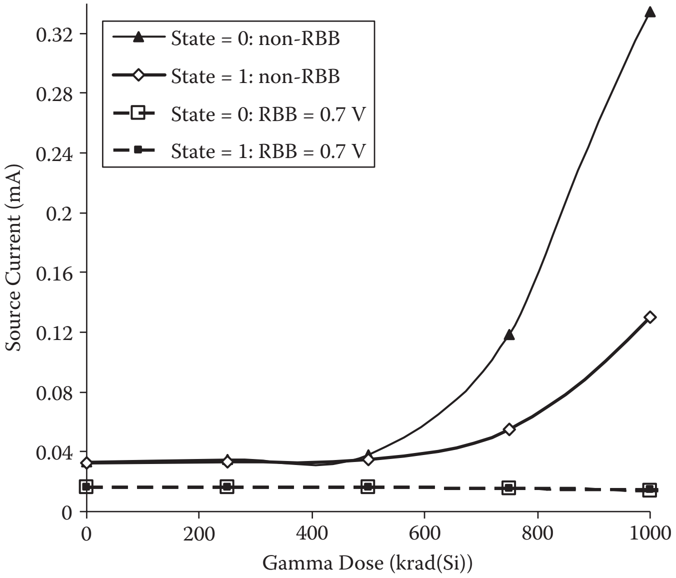


FIGURE 7.15 Measured current on the RBB two-edge 90 nm SRAM SOURCE node (I_{OFF} through the pull down transistors) after Co-60 irradiation to 1 Mrad(Si) with V_{GS} = 1.0 V and RBB using V_{SB} = 0.7 V in the State = 1 condition. Measurement biases are given in the legend. A large increase in I_{OFF} is evident, exacerbated when the measurement is in the opposite state. Application of RBB fully mitigates the I_{OFF} increase. (Reprinted with permission from Clark, L.T., Mohr, K.C., Holbert, K.E., Xiaoyin Yao Knudsen, J., Shah, H., "Optimizing radiation hard by design SRAM cells". *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036)

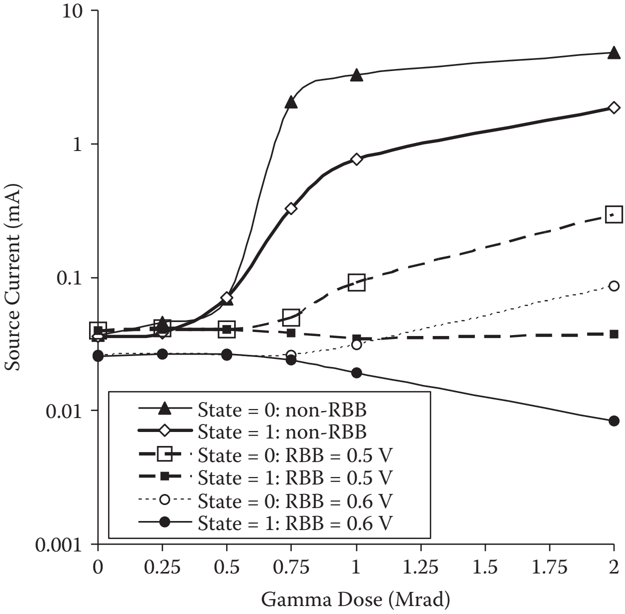


FIGURE 7.16 Measured SRAM I_{SOURCE} on the 90 nm 5 kB SRAM. The part was irradiated with $V_{SOURCE} - V_{SS} = 0.6$ V and $V_{DD} = 1.6$ V (SRAM cell $V_{DS} = 1.0$ V). (Reprinted with permission from, Clark, L.T., Mohr, K.C., Holbert, K.E., Xiaoyin Yao, Knudsen, J., Shah, H., "Optimizing radiation hard by design SRAM cells." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2028–2036.)

sensitivity to the amount of RBB bias applied. The cell $V_{DS} = 1$ V for both cases. Two key points are evident. First, the applied RBB strongly affects the measured I_{OFF}. Thus, RBB does not mitigate the trapped charge or TID-induced interface traps; it merely raises the parasitic edge transistor V_{th} sufficiently to alleviate the increased leakage. Second, the current required by node SOURCE can decrease with TID. This suggests that current can be delivered by another path separate from the V_{SOURCE} node. This current path is clearly dependent on the SRAM cell state during irradiation, despite its symmetrical nature.

The reduction in I_{SOURCE} at high TID suggests that the cell bias current is being fed from a path other than the test chip pin (see Figure 7.16). This is surmised to be due to leakage under the STI from a V_{SS} node to the SOURCE node. The small magnitude of the current difference between the two cell states suggests that this will not limit chip level I_{SB}. Previous work has reported that the SRAM source can be completely floated without the SRAM losing state [51]. This was proven on our 90 nm SRAM as well—gate leakage is sufficient to maintain the bias in a standby mode. I_{SB} measurements of irradiated NMOS transistors suggest that much of the large leakage increase in the 5 kB arrays is under the STI between n-type diffusions, such as between NMOS drains and the n-well. This further validates the conclusion that the RBB is effective in mitigating this leakage component in the 90 nm process and that this under STI component is responsible for the slight I_{SOURCE} at high TID.

Figure 7.16 also clearly shows higher I_{SB} for arrays measured without RBB but irradiated with it. This response, similar to that of the 130 nm SRAM, indicates that the RBB does not mitigate STI interface charging or trap formation but that the net effect is again mitigated by the RBB application.

7.5.8 90 nm Transistor-Level Response

Two-edged transistor arrays were also measured pre- and post-irradiation to help determine details of the TID leakage increase in the SRAMs. These arrays have the same narrow width NMOS transistors as the SRAM cells. The NMOS transistor arrays were irradiated with 0.7 V RBB applied and $V_{DS} = 0$ V and with no RBB and $V_{DS} = 0$ V with high transistor gate voltage applied (i.e., $V_{GS} = 1.0$ V) to determine the worst-case response. The results of measurements with RBB applied during irradiation but with $V_{SB} = 0$ V during the measurement sweeps showed that application of RBB during irradiation subtly enhances the increase in transistor I_{OFF} due to TID. Since only one SRAM row is accessed at a time, the impact of a higher I_{OFF} in the non-RBB condition is negligible and RBB application reduces the leakage by much more than the actual STI oxide degradation.

When irradiated, the measured SRAM cell leakage increase is greater than the NMOS transistor I_{OFF} increase experimentally measured on transistors. If the primary SRAM TID effect was increased parasitic NMOS drain to source leakage increase, the transistor increase would be higher than that of the SRAM. Since no *p*-type guard rings were used (all SRAM cells use the layout in Figure 7.3c), this suggests that leakage under the STI field oxides is a significant contributor. Subsequent work has shown that the field-oxide FET formed by the polysilicon bridging from the NMOS transistors to the n-well are a dominant TID-induced leakage path [50].

When used to reduce circuit standby leakage, the actual leakage improvement can be limited by both gate leakage and by drain-to-source leakage at the drain edge, either I_{GIDL} or I_{ZENER} [52]. All are direct band-to-band tunneling effects—the former through the thin oxide and the latter two due to sharp band bending caused by the steep doping profile at the drain-to-bulk transition region. The steep doping profiles are from halo implants used to control short channel effects [45,53]. Since the RBB bias creates higher drain-to-bulk bias conditions, it is important to determine if this leakage component will limit the available improvement that can be provided by using RBB. For example, if the I_{ZENER} increase with RBB is larger than I_{OFF}, RBB application will actually only mask TID-induced leakage by increasing the baseline value. Experiments on the 90 nm foundry process showed that this was not the case. Since doping increases exponentially as processes are scaled and the precise fields are process dependent, RBB TID mitigation on future processes may be limited by this mechanism.

7.6 SINGLE-EVENT EFFECTS IN UNHARDENED SRAM

SEE was investigated on the 1.2 Mb unhardened SRAM using the ion beam at the SEE facility at Texas A&M University. Since no SET mitigation techniques are used, this design provides a baseline for comparison with a SET mitigated design. This design uses small signal sensing and conventional circuits commonly used in commercial SRAMs, with two exceptions. First, tighter well and substrate tap spacing to avoid SEL in the beam testing was used. Second, RHBD I/O was used to avoid test failures due to the TID-induced I/O failure before the core TID effects could be seen.

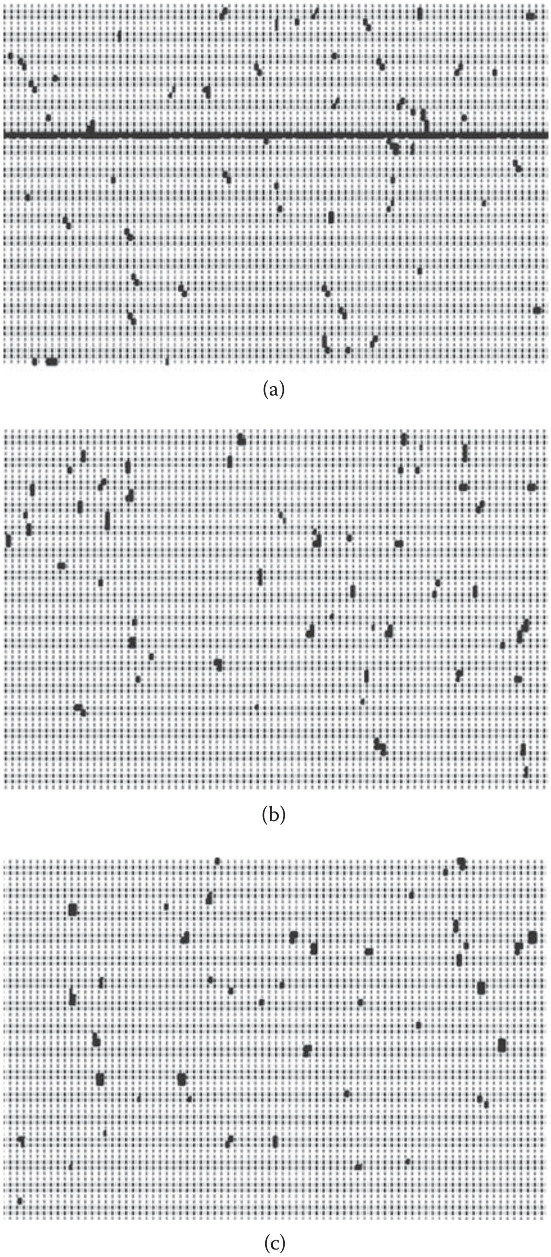
Figure 7.17 shows some measured SEU patterns for the LET = 59 MeV-cm^2/mg beam normal to the surface of the die (0° angle) with different stored patterns. Each strike is noted by one of eight different colors. The MBU detection algorithm assumes that bits more than eight cells apart are from different strikes, as the cycle count may not be indicative because multiple particles may strike between read sweeps through the array. Note the prevalence of MBUs. Figure 7.17a shows that for checkerboard patterns diagonal upsets predominate. This is due to the cell nodes storing the same logic one or zero being on the diagonal—for a hit between them, both can collect charge and be upset by one particle strike. Note that this can occur for both PMOS and NMOS collection, but referring to Figure 7.3c for strikes at opposite ends of the cells. The figure also clearly shows one row completely upset. This is due to an SET-induced WL assertion, which wrote the BL values into the cells. Luckily, the test pattern had alternating rows of 010101…. and 101010… (since it is a checkerboard) so this error could be caught. Figure 7.17b shows the pattern measured with vertical stripes programmed into the array, while Figure 7.17c shows the resulting patterns from a solid 0's pattern. In the latter, since the bit-lines alternate as BL, BLN, BLN, BL, up to four adjacent diffusions may collect charge simultaneously, as evident. Stripes predominate in the stripes case. The likelihood that a strike generates an MBU is thus dependent on the stored data pattern, with the likelihood of a four-bit upset rising from under 5% for the checkerboard and stripe patterns to over 16% for the solid pattern. No. 5 bit upsets were detected for the two former cases, but they comprised over 15% of the strikes in the solid pattern 0° angle tests at the same LET. Again, knowing the exact physical organization of the array down to whether cells are tiled or folded in the horizontal direction is critical to accurate SEU analysis.

How many bits are upset by a given strike versus beam incident angle with the checkerboard pattern, again at LET = 59 MeV-cm^2/mg, is shown in Figure 7.18. Here, one- and two-bit upsets predominate at normal incidence, while at 42° the majority of hits upset two or more cells. At the higher angles (see Figure 7.18d) most hits are two-bit MBU, with up to five bits upset by one particle strike.

Figure 7.19 shows the measured SRAM cell cross section versus effective LET (LET_{eff}) at different V_{DD} voltages. As expected, since increasing V_{DD} raises the cell Q_{crit}, the cross section diminishes with increasing V_{DD} at low LET. However, at high V_{DD}, as indicated by the points connected by lines, the cross section rises considerably at LET_{eff} above 70 MeV-cm^2/mg. This is due to enhanced charge generation from amplification by parasitic bipolar transistor action, which can cause very large MBU extents, particularly down SRAM wells [54].

7.7 SINGLE-EVENT EFFECTS MITIGATION

In this section, the SEE mitigation circuits implemented in the 130 nm design and their operation are described.



(a)

(b)

(c)

FIGURE 7.17 MBU extent vs. stored memory patterns observed in testing the unhardened 1.2Mb 90 nm SRAM at normal beam incidence for checkerboard (a) stripes (b) and all zeros (c) patterns, respectively. Note the entire row disrupted, presumably by a SET that in turn asserted a WL, which overwrote the contents.

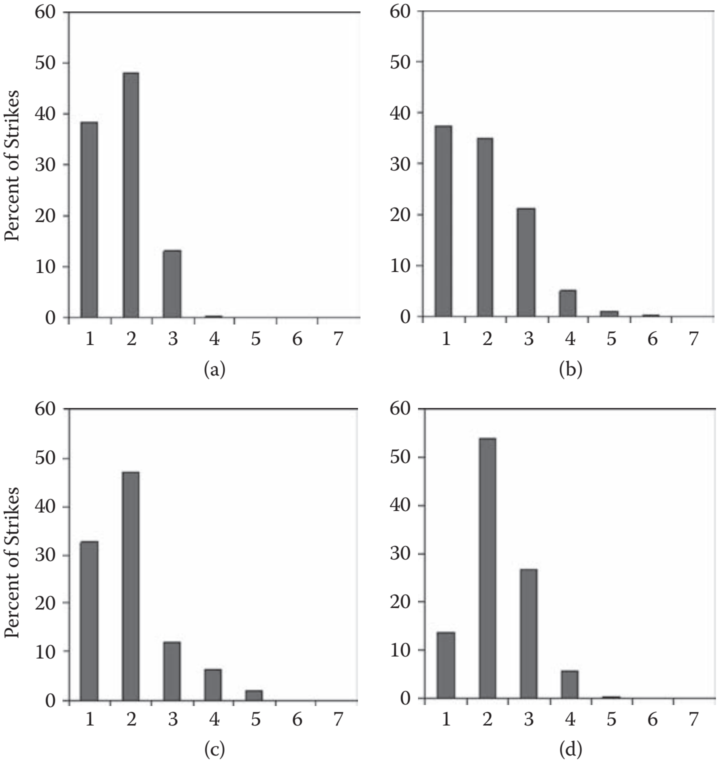


FIGURE 7.18 The effect of particle angle on MBUs in the unhardened 90 nm 1.2 Mb SRAM. At 0° one and two bit upsets predominate, with few three and four bit upsets (a). As the angle increases to 42°, a larger number of three bit upsets occurs (b), until at 53° (c) and 65° (d), MBUs predominate.

7.7.1 130 nm SRAM Design with RBB + SC Support and SEE Mitigation

As mentioned already, the 130 nm 5 kB RHBD SRAM used dual redundant control lines to detect and prevent SET data corruption due to control or word-line SET. The dual redundant control logic SET mitigation used here imposes no absolute maximum clock frequency limits and allows operating frequencies above 500 MHz in this implementation.

Each row in the SRAM array is controlled by one of 128 dual redundant WL drivers labeled L for left and R for right (Figure 7.20). Separate decoders provide SelLx and SelRx with timing set by WL enable signals RowENLx and RowENRx, respectively. Each left dual redundant driver for row x controls the WLLNx, and, in part, the SOURCEx signals for one row. The redundant drivers are spatially separated to reduce the probability of a single ionizing particle strike affecting both of them. The test array uses RBB + SC (optionally RBB) to determine its value for TID mitigation on this 130 nm process.

As previously described, when a row is inactive it is biased to the higher V_{SOURCE} voltage that applies the RBB as the NMOS bulks are all at $V_{SS} = 0$ V. During a read

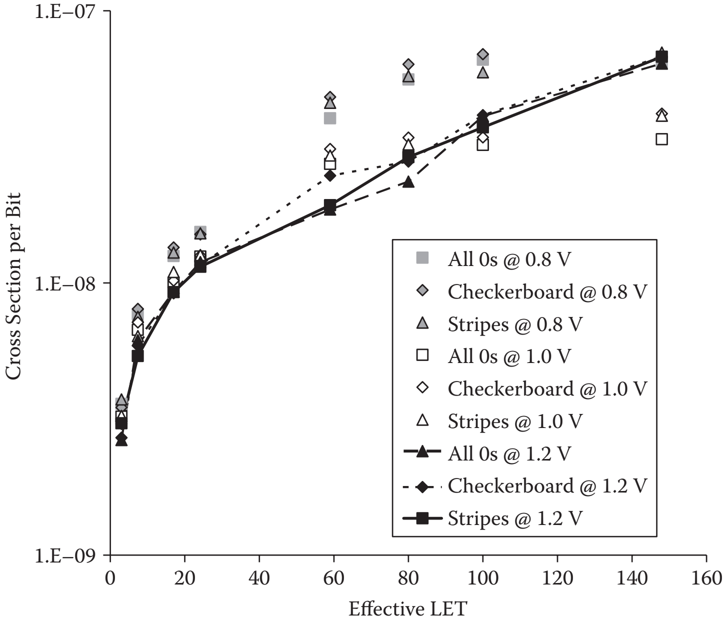


FIGURE 7.19 90 nm 1.2 Mb SRAM cross-section vs. LET$_{\text{eff}}$. Note that at low LET, the cross-section is higher at lower V$_{\text{DD}}$. At very high LET, higher V$_{\text{DD}}$ increases the cross section.

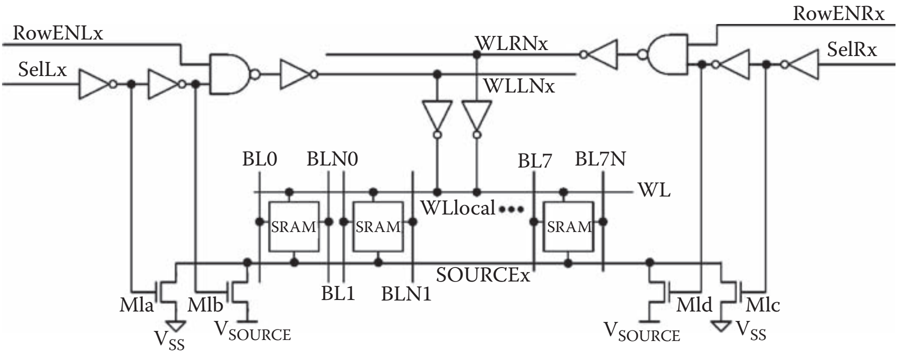


FIGURE 7.20 SRAM row design. Each subgroup of 8 bits is driven by two local WL drivers. The left side row driver (not shown) is an exact mirror image of this one, but drives WLbarL.

or write cycle, SOURCEx is driven to V_{SS} through transistor M1a to ensure read stability and fast writes. The row drivers must be immune to an SET that could cause the SOURCEx node voltage to rise above V_{SOURCE}, which would collapse the SRAM cell supply voltage, potentially upsetting the stored state in the entire row. SETs that drive the SOURCE node voltage low are not a concern as they momentarily increase that row's SRAM cell supply voltage magnitude. To avoid a strike raising a row's SOURCE bias node SOURCEx, it is connected to only n-type diffusions. These can collect only ionizing radiation-deposited electrons and thus drive only SOURCEx to a lower potential. Strikes on upstream logic that controls the bias are also a concern. For example, referring to Figure 7.20, enabling transistor M1a in one of the two redundant row drivers and M1d in the other creates contention between them. In this case, SOURCEx takes on an intermediate voltage between V_{SS} and V_{SOURCE}, reducing read speed and margin. Two approaches are taken to ensure that this does not affect functionality. First, the M1a and M1c transistors are sized wider than M1b and M1d, keeping the contention voltage low. Second, this SOURCEx contention condition bias value is used as the worst-case when determining read margin and speed.

To avoid an ionizing particle strike asserting an inactive WL signal during a read operation, as discussed in Section 2.1.2, the design uses two redundant active low WL signals per row, labeled WLLNx and WLRNx, which in turn control 40 local WL (labeled WLlocal) signals in each row (see Figure 7.20). WLRNx is driven by the right row driver and signal WLLNx by the left row driver. These two redundant signals are combined locally every eight bits, driving the WLlocal signals, which control eight SRAM cells. An SET assertion of a single WLlocal signal corrupts at most eight bits. Since each of the eight bits resides in a different EDAC code word, all eight-bit errors are correctable. Unlike the design in Figure 7.1b, which could still activate all local WLs by incorrectly asserting the global WL due to an SET, if either the WLbarL or WLbarR signal is corrupted due to an SET, a contention condition is created in the local WL driver inverters. The transistor sizing in the local WL drivers ensures that under contention the WLlocal signals will not rise to a high enough voltage to write the SRAM cells they control. Consequently, an SET enabled WL signal in a row that is not active cannot cause a false write. This local WL driver circuit was chosen over an AND gate because it is smaller. An erroneously disabled WL signal is detected, as described in this section, allowing the write or read operation to be repeated. Of course the controlling circuitry micro-architecture must comprehend this condition by buffering write data for a retry and by appropriately rerunning the operation as needed.

A SET-induced WL assertion is detected by two additional columns connected only to WLLN or WLRN signals rather than the local word-lines. These cells always discharge the BLs in their column and are read on both read and write operations. One of WLbarL or WLbarR incorrectly asserted is indicated by either being incorrect. In this case, the cycle is repeated. Of course a "false positive" error can be induced by an ionizing particle strike on the BL itself, in which case the data is correct, but rewritten nonetheless. BL development is much faster during a write, owing to the stronger write drivers, than during a read. This ensures that a write has been successfully completed if both WLchk outputs are valid.

7.7.2 SRAM Column Circuits

Similar to the WL protection, dual redundant column decoders generate the Y multiplexer control signals ColEnLWy and ColEnRWy for each of 16 words y (eight above and eight below the column sense and write circuits). Two decoders—one on the left and one on the right—form a dual redundant pair, with one pair for each of the top and bottom subarrays. There are sense and write circuits in each column of this design. The ColEnL and ColEnR signals are combined as shown in Figure 7.21. If either control signal is corrupted, an incorrect write cannot be generated, but a valid write may be aborted. Such a write abort is detected since the write data are monitored by the read sense circuits as they are being driven onto the bit-lines. These data are sampled at the end of the operation clock phase and compared with the data to be written. Any difference triggers an error, allowing the write to be repeated.

The read sensing is single ended using full-swing bit-lines. Both BL and BLN signals are sensed by high skew tristate inverters as shown in Figure 7.21. The outputs of these tristate inverters form the column (Y) output multiplexer. A SET that asserts one inadvertently will upset at most one output bit, which the EDAC will

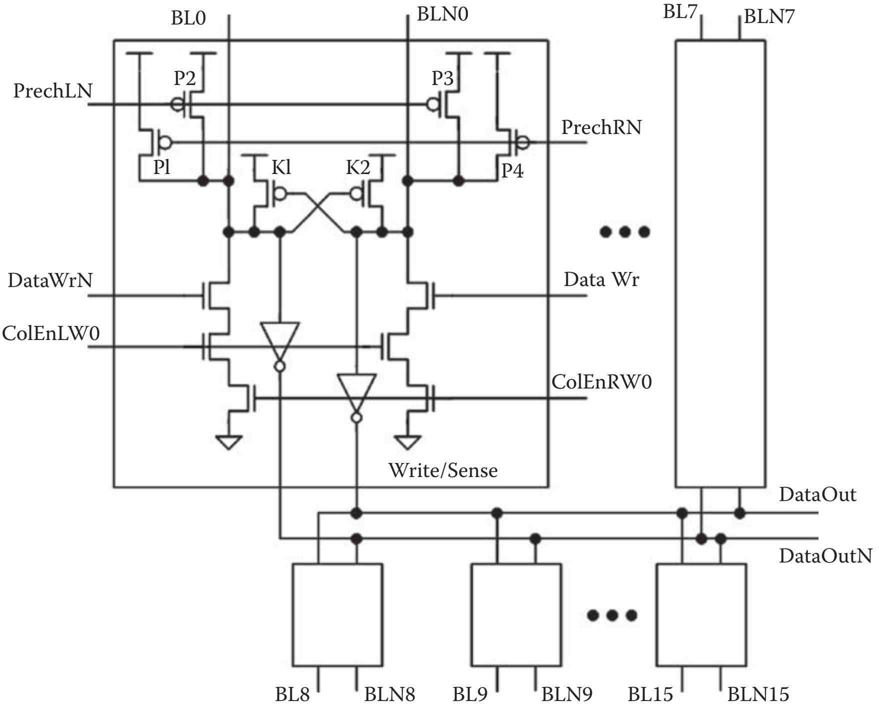


FIGURE 7.21 BL redundant pre-charge transistors, cross-coupled keeper PMOS transistors and sixteen-to-one 1-bit column multiplexer with redundant select and write control. If a precharge fails to turn off due to an SET, this condition is detected by the read/write detection columns.

correct. Dual redundant signals ColEnLW0 and ColEnRW0 enable DATAOUTN and DATAOUT, respectively, if word 0 is being read. If either of the ColEnLW0 or ColEnRW0 signals is corrupted due to an SET, the DataOut and DataOutN signals in a column will match (i.e., one not reflecting the discharged BL as appropriate), signaling that the read must be repeated. EDAC may also detect this but cannot be guaranteed to do so. Other dynamic BL read errors caused by SETs (e.g., on the precharge signals PrechLN and PrechRN) are detected similarly.

Cross-coupled keeper transistors K1 and K2 shown in Figure 7.21 ensure that once the bit cell discharges either of the two bit-lines, the opposite line maintains a full rail logic "1." Dual redundant precharge circuits preclude SET events from deasserting both PrechLN and PrechRN during BL precharge cycles.

7.7.3 SRAM Operation with RBB + SC

Figure 7.22 shows a simulated read cycle followed by a write cycle. The SRAM reads and writes in the low clock phase, with BLs precharged in the high clock phase. At about 1.5 ns node C voltage of the storage cell node holding a logic 0 transitions from the elevated SOURCE voltage to V_{SS} in preparation for the read cycle. Note that this is controlled by the address input, and at lower clock frequencies this may occur much earlier. At 2 ns node C rises—this is due to the read current, which reduces the cell stability during a read, as discussed in Section 7.3.1. The first falling clock signal CLK edge initiates a read of a stored logic 0 resulting in each of the 40 WLlocal nodes being asserted high, followed by the BL discharging and a logic

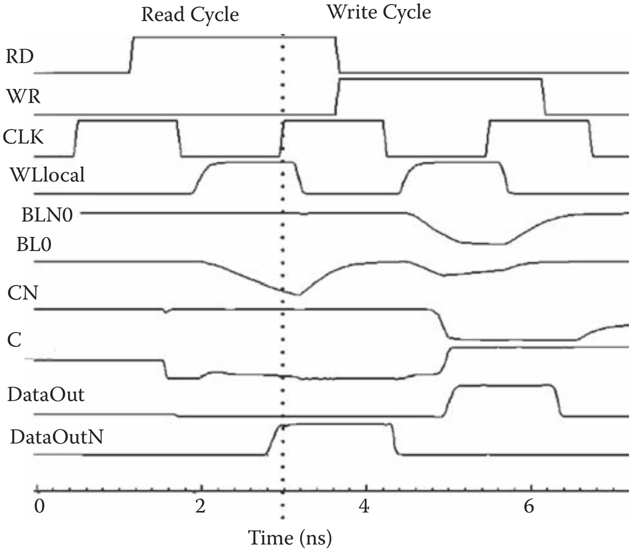


FIGURE 7.22 Simulation results showing the SRAM read and write cycles. Note the large BL read swing in the read cycle and in the write cycle, the BL discharge begins to discharge until the cell is written, whereupon it is restored by the BL keeper transistors.

1 driven out on DataOoutN. The second falling CLK edge at about 4.2 ns initiates a write cycle. Here a logic 1 is written into the SRAM cell, inverting the C and CN storage node signals.

7.7.4 Experimental SEE Measurements

While both RBB and RBB + SC improve standby leakage, they also reduce the SRAM cell drive strength, reducing Q_{crit} and making the cell more susceptible to SEU. The SEU impact of varying the V_{SS} and V_{SOURCE} potentials was quantified by cyclotron measurements.

Figure 7.23 compares the standard bias, V_{SOURCE} = 0.0 V, SRAM cell cross sections with those with RBB with V_{SS} voltages of –0.4, –0.6, and –0.8 V. No increase in cross section is observed for V_{SS} = –0.4 V. Measurements with V_{SS} = –0.6 and –0.8 V exhibit up to a 15% SRAM cell cross section increase at high effective LET. This is expected due to higher NMOS V_{th} as well as extended depletion regions, which improve funneling efficiency [6]. However, most of the change should occur as RBB is applied, that is, from 0.0 to –0.4 V where the V_{th} and depletion depth is most affected by the applied back bias, since it increases with the square root of V_{BS}. The MBU patterns produced by these tests, where no RBB is applied, are shown in

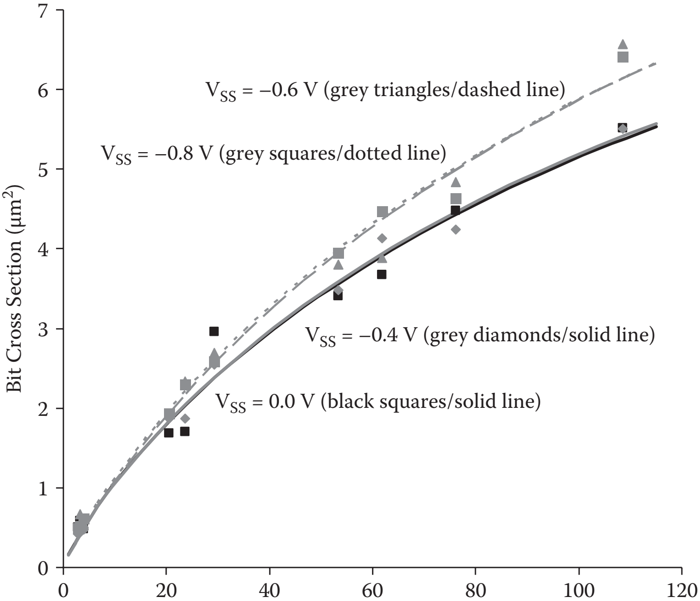


FIGURE 7.23 Measured RBB effect on SRAM cross-section vs. effective LET and bias: V_{DD} = 1.2 V, V_{SOURCE} = 0.0 V, V_{SS} = 0.0, –0.4, –0.6, and –0.8 V. (Reprinted with permission from Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

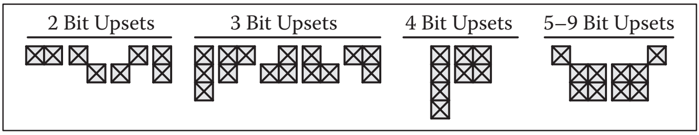


FIGURE 7.24 MBU patterns vs. at the standard bias $V_{DD} = 1.2$ V, $V_{SOURCE} = 0.0$ V, $V_{SS} = 0.0$ V, i.e., no RBB applied. (Reprinted with permission from, Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

Figure 7.24. At these angles, limited by shielding of the package to less than 60°, the MBU extent is limited.

Figure 7.25 shows the effect of RBB + SC on the measured SRAM cell cross sections at V_{SOURCE} biases of 0.4, 0.6, and 0.8 V relative to the standard bias of V_{SOURCE} = 0.0 V. RBB + SC has a significant effect on bit cell cross section. Cross section increases less than 60% for a V_{SOURCE} potential of 0.4 volts, while at V_{SOURCE} of 0.8 V the SEU cross section triples. This is easily attributable to the lower V_{GS} and V_{DS} of the transistors maintaining the SRAM cell state in these bias conditions, which significantly reduce Q_{crit}. Additionally, due to multibit errors (MBE) at these biases, the cell cross section is considerably larger than the physical extent of one SRAM cell.

The bias dependence of MBUs was examined to ensure that increases in the bit cell cross section due to changes in V_{SOURCE} bias can be effectively mitigated by increasing the EDAC scrub frequency. Required scrub frequencies are quite low [55] so small cross section increases can be easily dealt with. MBUs whose extent spans

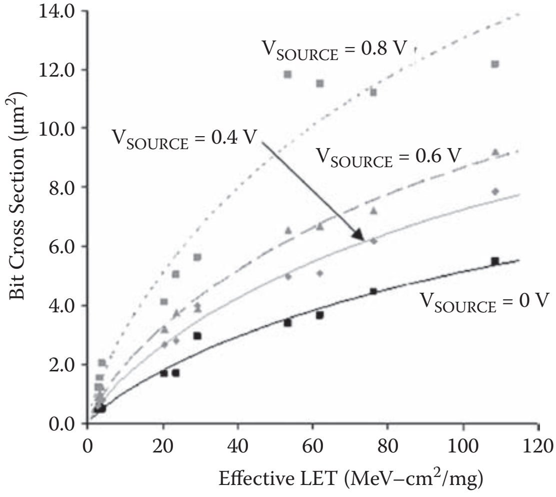


FIGURE 7.25 Effects of RBB+SC on bit cell cross section vs. effective LET and biases of $V_{DD} = 1.2$ V, $V_{SS} = 0.0$ V, with $V_{SOURCE} = 0.0$, 0.4, 0.6, and 0.8 V. (Reprinted with permission from, Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

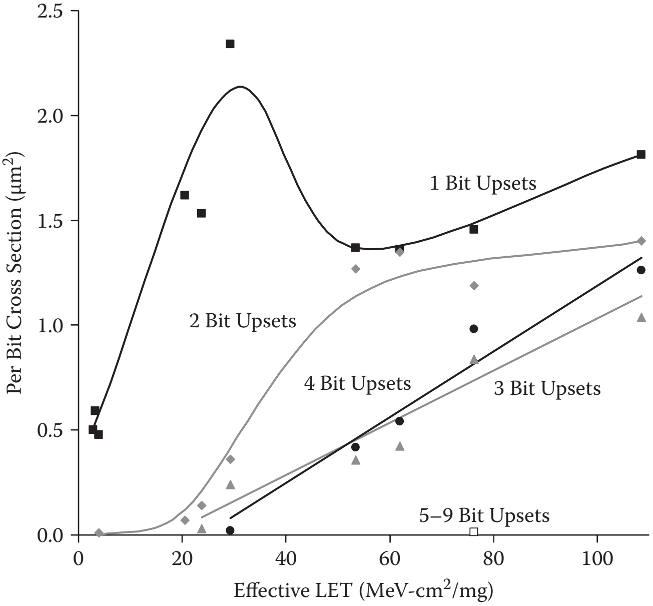


FIGURE 7.26 MBU per bit cross section vs. effective LET and number of upsets per particle strike at the V_{DD} = 1.2 V, no RBB applied. (Reprinted with permission from, Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

across EDAC code words at these low incident ion angles would result in an increase in the uncorrected error rate to unacceptable levels—recall that in space any angle is possible, so limiting the angles that can produce large MBU extent is the goal. The error cross section with error pattern size as a parameter for the standard bias (V_{DD} = 1.2 V, V_{SS} = V_{SOURCE} = 0.0 V) is shown in Figure 7.26. Single-bit errors dominate at low LET. Above LET = 30 MeV-cm²/mg strikes are more likely to result in MBUs causing the frequency of single-bit upsets to drop and two, three, and four bit upsets to increase as shown. The eight-SRAM bit cell codeword interleaving used in this SRAM appears sufficient at the incident ion angles below 60°. The incident ion angles in this experiment were limited by the IC package.

Using RBB + SC produces new MBU phenomena. Figure 7.27 shows that this significantly increases the frequency of MBUs with more than five bit upsets. Raising the V_{SOURCE} voltage to only 0.4 V increases the SRAM cell cross section MBU extent and frequency, allowing very large MBUs not observed without RBB + SC. The largest MBU observed at this bias, with 0.4 V RBB applied and 0.8 V across the SRAM transistors, was 11 bits, as shown in Figure 7.28. The large MBUs tend to be long and slender and oriented in the BL direction. Since words on different rows of this SRAM are always in different EDAC codewords, all codewords are still correctable as only one upset bit resides in each. Additionally these upsets cross many n-well boundaries. The n-wells should provide favorably biased charge collection nodes that collect deposited charge and thus mitigate upsets. We believe

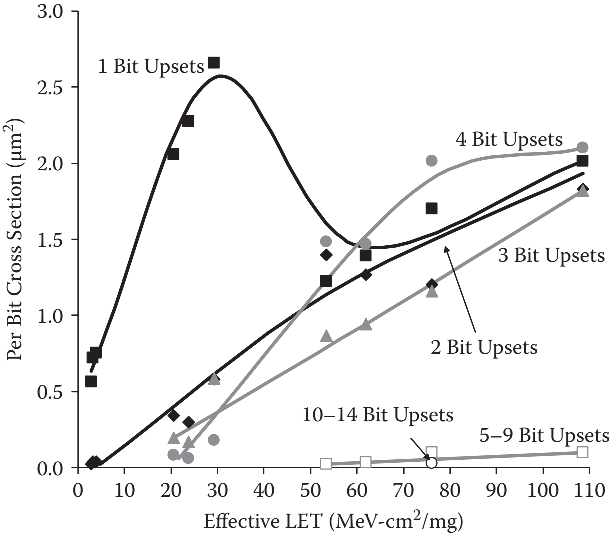


FIGURE 7.27 Effects of RBB+SC on MBU per bit cross section vs. effective LET and number of cells upset per particle strike for V_{DD} = 1.2 V, V_{SOURCE} = 0.4 V, V_{SS} = 0.0 V. The cross-section of five or more bits upset becomes noticeable at LET_{eff} > 50 MeV-cm²/mg. (Reprinted with permission from, Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

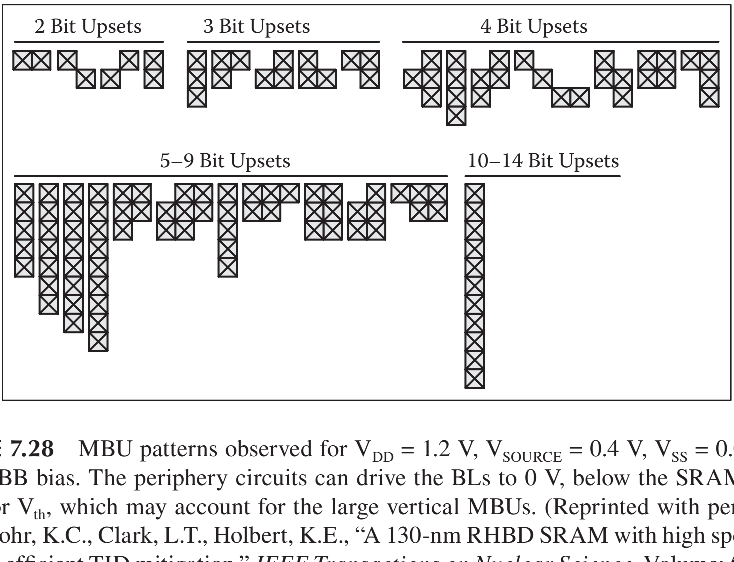


FIGURE 7.28 MBU patterns observed for V_{DD} = 1.2 V, V_{SOURCE} = 0.4 V, V_{SS} = 0.0 V, i.e., 0.4 V RBB bias. The periphery circuits can drive the BLs to 0 V, below the SRAM access transistor V_{th}, which may account for the large vertical MBUs. (Reprinted with permission from, Mohr, K.C., Clark, L.T., Holbert, K.E., "A 130-nm RHBD SRAM with high speed SET and area efficient TID mitigation." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2092–2099.)

that the source of such long slender errors is ion strikes on the BL or write driver diffusions, causing the BL to glitch below the WL voltage. This is a classical signaling noise scenario that causes the access transistors of multiple cells to be asserted on as their gates are at V_{SS} = 0.4 V and sources glitch to 0 V or less, writing multiple cells in the same column.

Further reductions in cell storage voltage by raising V_{SOURCE} increase the SEU cross section of the SRAM cell and affect the size and shape of single-strike MBUs. The trend continues and when V_{SOURCE} is raised to 0.8 V, very long thin errors are observed extending over all 64 bits in a column (the entire BL length) were observed. Consequently, while very effective at mitigating TID, RBB + SC must be used with caution as it can allow large MBU increases. RBB alone, however, appears to be effective at mitigating TID and does not appreciably affect the SEU rate.

7.8 SUMMARY AND CONCLUSIONS

The SRAM cell designs presented here were designed with similar read and write margins to ensure a fair comparison of the size impact of RHBD. Test layouts of the more promising designs show that RHBD SRAM cells using annular NMOS or PMOS access transistors are at least three times, and potentially greater than five times, larger in area than the foundry optimized unhardened cells. It is worth noting that many trade-offs are related. For instance, the use of two-edged access transistors may make the use of *p*-type guard rings superfluous, as the guard ring cannot mitigate flow between the two-edged transistor source and drain. 130 nm SRAM arrays showed 75 times I_{SB} increases at 2 Mrad(Si) and a 90 nm SRAM fabricated on an LSP process, with a fully commercial design style, exhibited 131 times I_{SB} increase after 1 Mrad(Si) in Co-60 accelerated TID experiments.

These clearly indicate that some form of mitigation is necessary to limit I_{SB} increase at high (e.g., Mrad-level) doses. Of course, the narrow SRAM transistors provide a worst case, and spaceborne ICs with low memory content may find such large I_{SB} increases acceptable. Additionally, most satellite requirements are met with lower specifications (e.g., 300 krad(Si)), which modern sub-100 nm processes may provide intrinsically. However, such a choice must be made cautiously. Recall from Section 7.3.5.7 that experiments showed that the cell stability can be affected by TID, where the 130 nm unhardened SRAM arrays had TID-induced bit failures starting at 750 krad(Si). Using a novel test structure, the switching points of SRAM cells using two-edged transistors were shown to change considerably as they were dosed to 1.6 Mrad(Si). Conversely, no cell switch point changes or I_{SB} increase was observed for a fully annular NMOS, *p*-type guard ringed 90 nm SRAM cell design. However, such larger hardened cells do have naturally higher SRAM leakage as a consequence of their wider transistors.

Measurements of fabricated 130 nm and 90 nm transistors and SRAM cells before and after TID irradiation indicate that two-edged NMOS access transistor cells are superior to PMOS access transistor designs at low radiation levels, that is, those below 500 krad(Si), and are probably adequate to higher doses. Experiments show increasing SRAM V_{SS} to apply RBB reduces post-irradiation leakage at 1 Mrad(Si) in conventional cells below the preirradiation leakage for the annular pull-down cell.

At full V_{DD} = 1.2 V, SEE is not adversely affected, and, at 1 Mrad(Si) at the same bias conditions, the post-irradiation two-edge 130 nm SRAM cell I_{SB} is below the preirradiation level for the annular design. The measurements show that relatively low (i.e., on the order of 500 mV) RBB is sufficient to mitigate TID-induced I_{SB} increase up to 1 Mrad(Si) on a 130 nm process. RBB + SC was shown to introduce a novel SEE failure where all cells on a single BL are upset. While the bits are all in separate EDAC words in any rational array organization, this may still be problematic, as large MBUs affect the required scrub rates. However, the RBB scheme appears effective at essentially no SEE penalty.

TID measurements of an operating 90 nm 5 kB SRAM show that RBB is effective in limiting I_{OFF} increase at least to 1 Mrad(Si) and probably at higher doses, where the current increases were still small. High intrinsic I_{OFF} and gate leakage on the 90 nm process limit the overall current savings. Concurrently, the high leakage floor masks TID-induced I_{OFF} increase until higher (i.e., greater than 500 krad(Si)) irradiation levels, suggesting that two-edged cells without RBB may be acceptable for many hardened systems. If RBB is used, the experiments presented here show that there is latitude in the choice of V_{DS} magnitude when RBB is applied. This will be a function of the required SEE hardness and the impact of V_{DS} on the cell Q_{crit}. The experiments presented here have also shown that band-to-band tunneling at the junction edge does not limit the use of RBB.

Fabricated line widths have moved considerably beyond the lithographic generation. For example, 193 nm lithography is used to fabricate 35 nm polysilicon gates [56] in production by using resolution enhancement techniques and phase shift masks. Consequently, support for the polysilicon shapes required for RHBD enclosed geometry gates on future deep submicron processes can be expected to diminish and probably vanish completely, at least for the core logic transistors. RBB applied to the NMOS transistors promises a potential RHBD approach that is compatible with such highly scaled fabrication processes, not just for SRAM but for logic as well.

Simulation studies have shown that the logic delay and active power increase over an unhardened design when using NMOS RBB are less than when using enclosed geometry transistors [33]. The former causes less than 5% increase in logic delay, at reduced leakage and similar active power, compared with a commercial two-edged only design. Nearly all modern ICs use higher I/O voltages for compatibility and lower-scaled V_{DD} in the core logic transistors. For example, V_{DDIO} = 1.8 V to 2.5 V, V_{DD} = 1.2 V (the core V_{DD}), and V_{SS} = 0 V (gnd) are common. RBB can easily be applied to an entire IC to provide NMOS transistor TID hardness by placing the core circuits in a domain between V_{DDIO} = 1.8 V and $V_{SS(CORE)}$ = 0.6 V. The resulting IC is still three power supplies as before. It is straightforward to convert the I/O circuit level shifters, which presently convert the upper rail between V_{DD} and V_{DDIO}, to convert the lower rail between $V_{SS(CORE)}$ and V_{SS}.

REFERENCES

1. International Technology Roadmap for Semiconductors, 2003. Available at: http://www.itrs.org

2. S. Rusu, J. Stinson, S. Tam, J. Leung, H. Muljono, and B. Cherkauer, "A 1.5-GHz 130-nm Itanium-2 Processor with 6-MB on-die L3 Cache," *IEEE J. Solid-State Circuits,* vol. 38, no. 11, Nov. 2003, pp.1887–1895.
3. J. Chang, S. Rusu, J. Shoemaker, S. Tam, H. Ming, M. Haque, et al., "A 130-nm Triple-Vt 9-MB Third-Level On-Die Cache for the 1.7-GHz Itanium-2 Processor," *IEEE J. Solid-State Circuits,* vol. 40, no. 1, Jan. 2005, pp. 195–203.
4. J. Wuu et al., "The Asynchronous 24MB On-Chip Level-3 Cache for a Dual-Core Itanium Family Processor," *Proc. Int. Solid-State Circuits Conf.,* 2005, pp. 488–489.
5. S. Rusu et al., "A Dual Core Multi Threaded Xeon Processor with 16MB L3 Cache," *IEEE Int. Solid-State Circuits Conf.,* 2006, pp. 315–324.
6. J. Srour and J. McGarrity, "Radiation Effects on Microelectronics in Space," *Proc. of the IEEE,* vol. 76, no. 11, Nov. 1988, pp. 1443–1469.
7. E. Stassinopoulos and J. Raymond, "The Space Radiation Environment for Electronics," *Proc. of the IEEE*, vol. 76, no. 11, Nov. 1988, pp. 1423–1442.
8. S. Kerns, B. Shafer, L. Rockett, J. Pridmore, D. Berndt, N. van Vonno, et al., "The Design of Radiation-Hardened ICs for Space: A Compendium of Approaches," *Proc. of the IEEE,* vol. 76, no. 11, Nov. 1988, pp. 1470–1509.
9. G. Anelli, M. Campbell, M. Delmastro, F. Faccio, S. Floria, A. Giraldo, et al., "Radiation Tolerant VLSI Circuits in Standard Deep Submicron CMOS Technologies for the LHC Experiments: Practical Design Aspects," *IEEE Trans. Nuc. Sci.,* vol. 46, no. 6, Dec. 1999, pp. 1690–1696.
10. H. Weaver, C. Axness, J. McBrayer, J. Browning, J. Fu, A. Ochoa, et al., "An SEU Tolerant Memory Cell Derived from Fundamental Studies of SEU Mechanisms in SRAM," *IEEE Trans. Nuc. Sci.,* vol. 34, no. 6, Dec. 1987, pp. 1281–1286.
11. J. Schwank, M. Shaneyfelt, B. Draper, and P. Dodd, "BUSFET—A Radiation-Hardened SOI Transistor," *IEEE Trans. Nuc. Sci.,* vol. 46, no. 6, Dec. 1999, pp. 1809–1817.
12. S. Liu, D. Nelson, J. Tsang, K. Golke, P. Fechner, W. Heikkila, et al., "The Effect of Active Delay Element Resistance on Limiting Heavy Ion SEU Upset Cross-Sections of SOI ADE/SRAMs," *IEEE Trans. Nuc. Sci.,* vol. 54, no. 6, Dec. 2007, pp. 2480–2487.
13. R. Lacoe, J. Osborne, R. Koga, and D. Mayer, "Application of Hardness-by-Design Methodology to Radiation-Tolerant ASIC Technologies," *IEEE Trans. Nuc. Sci.,* vol. 47, no. 6, Dec. 2000, pp. 2334–2341.
14. T. Oldham and F. McLean, "Total Ionizing Dose Effects in MOS Oxides and Devices," *IEEE Trans. Nuc. Sci.,* vol. 50, no. 3, June 2003, pp. 483–499.
15. H. Barnaby, "Total-Ionizing-Dose Effects in Modern CMOS Technologies," *IEEE Trans. Nuc. Sci.,* vol. 53, no. 6, Dec. 2006, pp. 3103–3121.
16. J. Felix, P. Dodd, M. Shaneyfelt, J. Schwank, and G. Hash, "Radiation Response and Variability of Advanced Commercial Foundry Technologies," *IEEE Trans. Nuc. Sci.,* vol. 53, no. 6, Dec. 2006, pp. 3187–3194.
17. X. Yao, N. Hindman, L. Clark, K. Holbert, D. Alexander, and W. Shedd, "The Impact of Total Ionizing Dose on Unhardened SRAM Cell Margins," *IEEE Trans. Nuc. Sci.,* vol. 55, no. 6, Dec. 2008, pp. 3280–3287.
18. T. Calin, M. Nicolaidis, and R. Velazco, "Upset Hardened Memory Design for Submicron CMOS Technology," *IEEE Trans. Nuc. Sci.,* vol. 43, no. 6, Dec. 1996, pp. 2874–2878.
19. R. Koga, K. Crawford, P. Grant, W. Kolasinski, D. Leung, T. Lie, et al., "Single Ion Induced Multiple-Bit Upset in IDT 256K SRAMs," *Proc. RADECS,* Sept. 1993, pp. 485–489.
20. C. Underwood, R. Ecoffet, S. Duzeffier, and D. Faguere, "Observations of Single-event Upset And Multiple-Bit Upset in Non-hardened High-Density SRAMs in the TOPEX/Poseidon Orbit," *IEEE Rad. Effects Data Workshop,* July 1993, pp. 85–92.
21. D. Mavis and P. Eaton, "Soft Error Rate Mitigation Techniques for Modern Microcircuits," *Proc. IRPS,* 2002, pp. 216–225.

22. H. Weaver, C. Axness, J. McBrayer, J. Browning, A. Fu, A. Ochoa, et al., “An SEU Tolerant Memory Cell Derived from Fundamental Studies of SEU Mechanisms in SRAM,” *IEEE Trans. Nuc. Sci.,* vol. 34, no. 6, Dec. 1987, pp. 1281–1286.
23. W. Jenkins, R. Martin, and H. Hughes, “Characterization of an Ultra-Hard CMOS 64K Static RAM,” *IEEE Trans. Nuc. Sci.,* vol. 34, no. 6, part I, Dec. 1987, pp. 1455–1459.
24. Z. Kai, L. Zhongli, Y. Fang, X. Zhiqiang, and H. Genshen, “Radiation Hardened 128K PDSOI CMOS Static RAM,” *Proc. ICSICT,* 2006, pp. 1922–1924.
25. P. McDonald, W. Stapor, A. Campbell, and L. Massengill, “Non-random Single Event Upset Trends,” *IEEE Trans. Nuc. Sci,* vol. 36, no. 6, Dec. 1989, pp. 2324–2329.
26. L Jacunski et al., “SEU Immunity: The Effects of Scaling on the Peripheral Circuits of SRAMs,” *IEEE Trans. Nuc. Sci,* vol. 41, no. 6, Dec. 1994, 2272–2276.
27. J. Maiz, S. Hareland, K. Zhang, and P. Armstrong, “Characterization of Multi-bit Soft Error Events in Advanced SRAMs,” *IEDM Tech. Dig.,* Dec. 2003, pp. 21.4.1–21.4.4.
28. D. Mavis et al., “Multiple Bit Upsets and Error Mitigation in Ultra-Deep Submicron SRAMs,” *IEEE Trans. Nuc. Sci.,* vol. 55, no. 6, Dec. 2008, pp. 3288–3294.
29. K. Mohr and L. Clark, “Experimental Characterization and Application of Circuit Architecture Level Single Event Transient Mitigation,” *Proc. IRPS,* Apr. 2007, pp. 312–317.
30. J. Knudsen and L. Clark, “An Area and Power Efficient Radiation Hardened by Design Flip-Flop,” *IEEE Trans. Nuc. Sci.,* vol. 53, no. 6, Dec. 2006, pp. 3392–3399.
31. J. Montonarro et al., “A 160MHz, 32b 0.5W CMOS RISC Microprocessor,” *IEEE J. of Solid-state Circ.,* vol. 31, no. 11, Nov. 1996, pp. 1703–1714.
32. L. Clark, E. Hoffman, J. Miller, M. Biyani, Y. Liao, S. Strazdus, et al., “An Embedded Microprocessor Core for High Performance and Low Power Applications,” *IEEE J. of Solid-State Circ.,* vol. 36, no. 11, Nov. 2001, pp. 1599–1608.
33. L. Clark, K. Mohr, and K. Holbert, “Reverse-Body Biasing for Radiation-Hard by Design Logic Gates,” *Proc. IRPS,* April 2007, pp. 582–583.
34. A. Bhavnagarwala, T. Xinghai, and J. Meindl, “The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability,” *IEEE J. Solid-State Circ.,* vol. 36, no. 4, Apr. 2001, pp. 658–665.
35. E. Grossar, M. Stucchi, K. Maex, and W. Dehaene, “Statistically Aware SRAM Memory Array Design,” *Proc. ISQED,* March 2006.
36. R. Heald and P. Wang, “Variability in Sub-100nm SRAM Designs,” *Proc. ICCAD-2004,* Nov. 2004, pp. 347–352.
37. N. Gierczynski, B. Borot, N. Planes, and H. Brut, “A New Combined Methodology for Write-Margin Extraction of Advanced SRAM,” *IEEE Int. Conf. on Microelectronic Test Structures,* March 2007, pp. 97–100.
38. E. Seevinck, F. List, and J. Lohstroh, “Static-Noise Margin Analysis of MOS SRAM Cells,” *IEEE J. Solid-State Circ.,* vol. SC-22, no. 5, Oct. 1987, pp. 748–754.
39. R. Myers and D. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2d ed., Wiley, New York, 2002.
40. L. Clark, N. Deutscher, F. Ricci, and S. Demmons, “Standby Power Management for a 0.18 μm Microprocessor,” *Proc. ISLPED*, Aug. 2002, pp. 7–12.
41. H. Mizuno and T. Nagano, “Driving Source-Line Cell Architecture for Sub-1V High-Speed Low-Power Applications,” *IEEE J. Solid-State Circ.,* vol. 31, no. 4, Apr. 1996, pp. 552–557.
42. H. Mizuno et al., “An 18-μA Standby Current 1.8-V, 200-MHz Microprocessor with Self-Substrate-Biased Data-Retention Mode,” *IEEE J. Solid-State Circ.,* vol. 34, no. 11, Nov. 1999, pp. 1492–1500.
43. N. Kim, K. Flautner, D. Blaauw, and T. Mudge, “Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power,” *IEEE Trans. VLSI Sys.,* vol. 12, no. 2, Feb. 2004, pp. 167–184.

44. PXA27x Processor Family Power Requirements Application Note. Available at: http://www.marvel.com
45. S. Zhao et al., “Transistor Optimization for Leakage Power Management in a 65 nm CMOS Technology for Wireless and Mobile Applications,” *IEEE Symp. VLSI Tech. Dig. Tech. Papers,* June 2004, pp. 14–15.
46. M. Xapsos, G. Summers, and E. Jackson, “Enhanced Total Ionizing Dose Tolerance of Bulk CMOS Transistors Fabricated for Ultra-Low Power Applications,” *IEEE Trans. Nuc. Sci.,* vol. 46, no. 6, Dec. 1999, pp. 1697–1701.
47. L. Clark, K. Mohr, K. Holbert, X. Yao, J. Knudsen, and H. Shah, “Optimizing Radiation Hard by Design SRAM Cells,” *IEEE Trans. Nuc. Sci.,* vol. 54, no. 6, Dec. 2007, pp. 2028–2036.
48. K. Mohr, L. Clark, and K. Holbert, “A 130-nm RHBD SRAM with High Speed SET and Area Efficient TID Mitigation,” *IEEE Trans. Nucl. Sci.,* vol. 54, no. 6, December 2007, pp. 2092–2099.
49. C. Yang, K. Chan, L. Shi, D. Fried, J. Stathis, A. Chou, et al., “Hybrid-Orientation Technology (HOT): Opportunities and Challenges,” *IEEE Trans. Electron Dev.,* vol. 53, no. 5, May 2006, pp. 965–978.
50. H. Barnaby, M. Mclain, I. Esqueda, and X. Chen, “Modeling Ionizing Radiation Effects in Solid State Materials and CMOS Devices,” *Proc. IEEE CICC,* Sept. 2008 pp. 273–280.
51. A. Agarwal, L. Hai, and K. Roy, “A Single-V_t Low-Leakage Gated-Ground Cache for Deep Submicron,” *IEEE J. Solid-State Circ.,* vol. 38, no. 2, Feb. 2003, pp. 319–328.
52. S. Wolf, *Silicon Processing for the VLSI Era, Volume 3—The Submicron MOSFET,* Lattice Press, Long Beach, CA, 1995.
53. B. Haugerud et al., “The Impact of Substrate Bias on Proton Damage in 130 nm CMOS Technology,” *Radiation Effects Data Workshop Record,* July 2005, pp. 117–121.
54. J. Black et al., “Characterizing SRAM Single Event Upset in Terms of Single and Multiple Node Charge Collection,” *IEEE Trans. Nuc. Sci.,* vol. 55, no. 6, Dec. 2008, pp. 2943–2947.
55. K. Mohr and L. Clark, “Delay and Area Efficient First-Level Cache Soft Error Detection and Correction,” *ICCD Proc.,* Oct. 2006, pp. 88–92.
56. C. Bencher, H. Dai, and Y. Chen, “Gridded Design Rule Scaling: Taking the CPU towards the 16 nm Node,” *Proc. SPIE,* vol. 7274, 2009, pp. 0G-1–0G-10.

8 A Complete Guide to Multiple Upsets in SRAMs Processed in Decananometric CMOS Technologies

Gilles Gasiot and Phillippe Roche

CONTENTS

8.1 INTRODUCTION

Susceptibility to radiation environment of advanced electronic devices is often responsible for the highest failure rate of all reliability concerns (e.g., electromigration, gate rupture, negative bias-temperature instability [NBTI]). In modern static random access memories (SRAMs) the two predominant single-event effects (SEEs) are the single-event upset (SEU) and multiple upsets (MUs). Multiple upsets are topological errors in neighboring cells. If the cells belong to the same logical word they are named multiple-bit upsets (MBUs); otherwise they are labeled as multiple-cell upsets (MCUs). Multiple upsets have received increased scrutiny in recent years [1-8] because MBUs are uncorrectable by a simple error correction code (ECC) scheme and therefore threaten the efficiency of error detection and correction (EDAC).

As technologies scale down, the amount of transistors per mm^2 doubles at each generation, while the radioactive feature size (ion track diameter) is constant. This is illustrated in Figure 8.1 with three-dimensional (3-D) technology computer-aided design (TCAD) simulation showing an ion impacting a single cell in 130 nm, while

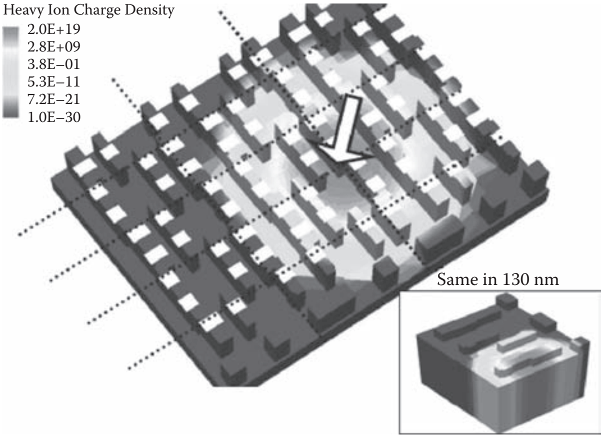


FIGURE 8.1 3D TCAD simulation of ion impact (single LET) in a single SRAM bitcell in 130nm and 12 SRAM bit cells in 45nm.

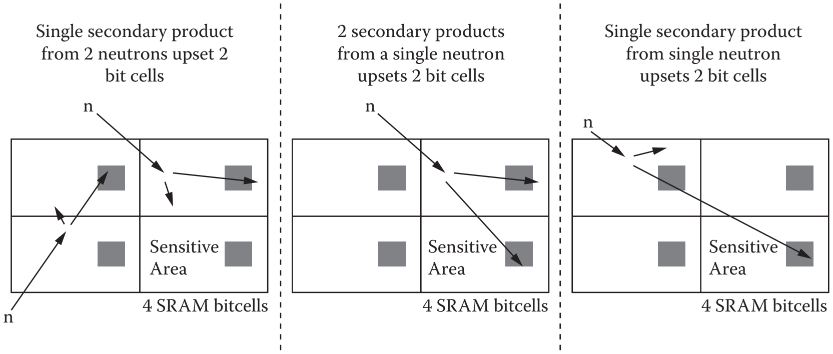

Single secondary product from 2 neutrons upset 2 bit cells
n
n
Sensitive Area
4 SRAM bitcells
2 secondary products from a single neutron upsets 2 bit cells
n
Sensitive Area
4 SRAM bitcells
Single secondary product from single neutron upsets 2 bit cells
n
Sensitive Area
4 SRAM bitcells

FIGURE 8.2 Scheme of neutron interaction that can cause Multiple Cell Upset in SRAM array. (From F. Wrobel et al. "Simulation of Nucleon-Induced nuclear reactions in a simplified SRAM structure: Scaling effects on SEU and MBU cross sections." *IEEE Transactions on Nuclear Science*, Volume: 48, Issue: 6, 1946–1952, December 2001.)

several are impacted in 45 nm. Moreover, the SRAM ability to store electrical data (critical charge) is reduced as technology feature size and power supply are jointly decreased. The probability that a particle upsets more than a single cell is therefore increased [9-11].

The mechanism for MCU occurrence in SRAM arrays is more than "enough energy was deposited to upset 2 cells" and depends on the radiation used. Directly ionizing radiation from single particles (e.g., alpha particles, ions) deposits charges diffusing in wells that can be collected by several bit cells. This phenomenon is enhanced by using tilted particles either naturally (alpha particles whose emission angle is random from the radioactive atom) or artificially (heavy ions can be chosen during experimental tests from 0° to 60°). Nonionizing radiation such as neutrons and protons can have different MCU occurrence mechanisms (Figure 8.2). A nonionizing particle can produce one or more secondary products. Several cases have to be considered: two secondary ions from two nucleons upset two or more bit cells; two secondary ions from a single nucleon upset two or more bit cells; and a single secondary ion from a single nucleon upsets two or more bit cells (in this case the phenomenon is close to the previously described direct ionizing mechanism). It has been shown that type 1 mechanism was negligible but that type 2 and 3 mechanisms coexist [12]. However, the proportion of MCUs due to these two mechanisms has never been precisely assessed.

One of the first experimental evidences of MBU was reported in 1984 in a 16 × 16 bit bipolar RAM under heavy-ion irradiation [13]. It is noteworthy that as many as 16 bit errors in columns from a single ion strike were detected. This means that 6% of the entire memory array was in error from a single particle strike. Since this first experimental evidence, multiple-bit errors were detected in several device types such as DRAM [14], polysilicon load SRAM [15], and antifuse-based field-programmable gate array (FPGA) [16] and under various radiation types, such as protons [17], neutrons [18], and laser [19].

The goal of this chapter is first to experimentally quantify MCU occurrence as a function of several parameters such as radiation type, test conditions (e.g., temperature, voltage), and SRAM architecture. These results will be used to sort by order of importance the parameters driving the MCU susceptibility. Second, 3-D TCAD simulations will be used to investigate the mechanisms leading to MCU occurrence and to determine the most sensitive location to trigger a two-bit MCU as well as the cartography of MCU sensitive areas.

8.2 DETAILS ON THE EXPERIMENTAL SETUP

The experiment design included different test patterns and supply voltages. The test procedure is compliant with the JEDEC soft error rate (SER) test standard JESD89 [20] for alpha and neutrons and European Space Agency (ESA) test standard n°22900 for heavy ions and protons [21].

8.2.1 Note on the Importance of Test Algorithm for Counting Multiple Upsets

When experimentally measuring MCUs, it is mandatory to distinguish (1) multiple independent failures from a cluster of nearest neighbor upset from a single multicell upset caused by a single energetic particle and (2) signature of errors due to a hit in redundancy latch or sense amplifier that may upset an entire row or column from an MCU signature. A test algorithm allows separating independent events due to multiple-particle hits from single events that upset multiple cells. Dynamic testing of memory usually involves writing once and then reading continuously at a specified operating frequency at which events are recorded one at a time. This gives a real insight on MCU shapes and occurrence. However, with static testing of memory, a test pattern is written once and stored for an extended period before reading the pattern back out. The result is a failure bit mapping in which events due to multiple-particle hits and single events that upset multiple cells cannot be distinguished. However, statistical tools can be applied to quantify the rate of neighboring upsets due to several ions [22,23]. One of these tools is described in detail in Annex 1 (Section 8.6).

8.2.2 Test Facility

8.2.2.1 Alpha Source

The tests were performed with an alpha source, which is a thin foil of Americium 241 with an active diameter of 1.1 cm. The source activity was 3.7 MBq as measured February 1, 2002. The alpha particle flux was precisely measured in March 2003 with an Si detector that was placed at 1 mm from the source surface. Since the atomic half-life of Am241 is 432 years, the activity and flux figures are still very accurate. During SER experiments, the Americium source lies above the chip package in the open air.

8.2.2.2 Neutron Facilities

Neutron experiments were carried out with the continuous neutron source available at the Los Alamos Neutron Science Center (LANSCE) and Tri University Meson

Facility (TRIUMF) in Vancouver. The neutron spectrums closely match the terrestrial environment for energies ranging from 10 MeV up to 500 MeV and 800 MeV for TRIUMF and LANSCE, respectively. The neutron fluence is measured with a uranium fission chamber. The total number of produced neutrons is obtained by counting fissions and applying a proportionality coefficient.

8.2.2.3 Heavy-Ion Facilities

The heavy-ion tests were conducted at the RADEF [24] (RADiation Effect Facility) cyclotrons. The RADEF facility is located in the Accelerator Laboratory at the University of Jyväskylä, Finland (JYFL). The facility includes beam lines dedicated to proton and heavy-ion irradiation studies of semiconductor materials and devices. The heavy-ion line consists of a vacuum chamber with component movement apparatus inside and ion diagnostic equipment for real-time analysis of beam quality and intensity. The cyclotron used at JYFL is a versatile, sector-focused accelerator for producing beams from hydrogen to xenon. The accelerator is equipped with three external ion sources. There are two electron cyclotron resonance (ECR) ion sources designed for high-charge-state heavy ions. Heavy ions used at the RADEF facility have stopping ranges in silicon much larger than the whole stack of back-end metallization and passivation layers (~10 µm).

8.2.2.4 Proton Facility

Proton irradiations were performed at the Proton Irradiation Facility (PIF) at Paul Scherrer Institute (PSI). This institute was constructed for testing of spacecraft components. The PIF main features are that irradiation takes place in air, that the flux/dosimetry is about 5% absolute accuracy, and that beam uniformity is higher than 90%. The experiments have used the low-energy PIF line whose energy range is 6 to 71 MeV and maximum proton flux is 5E8 p/cm^2/sec.

8.2.3 Tested Devices

Most of the data presented in this work were obtained using a single test chip (Figure 8.3). This test chip embeds three different bit cell architectures, two single-port (SP) and one dual-port (DP). It was manufactured in a 65 nm commercial complementary metal-oxide semiconductor (CMOS) technology with low-power (LP) process option. The main features of tested devices are summarized in Figure 8.4. Each bit cell was processed with and without the triple-well (TW) process option.

The triple-well layer consists of either an n+ or p+ buried layer in, respectively, a p- or n-doped substrate. As most devices are processed in a p-substrate, triple-wells are often referred to as deep n-well or n+ buried layer (Figure 8.5). For years TW layers have been used to electrically isolate the p-well and to reduce the electronic noise from the substrate. The TW is biased through the n-well contacts and ties connected to V_{DD}, while the p-wells are grounded. The well ties are regularly distributed along the SRAM cell array, as depicted in Figure 8.6. The triple-well process option has two main effects on the radiation susceptibility. First, it allows for decreasing the single-event latchup (SEL) sensitivity since the p-n-p base resistance is strongly

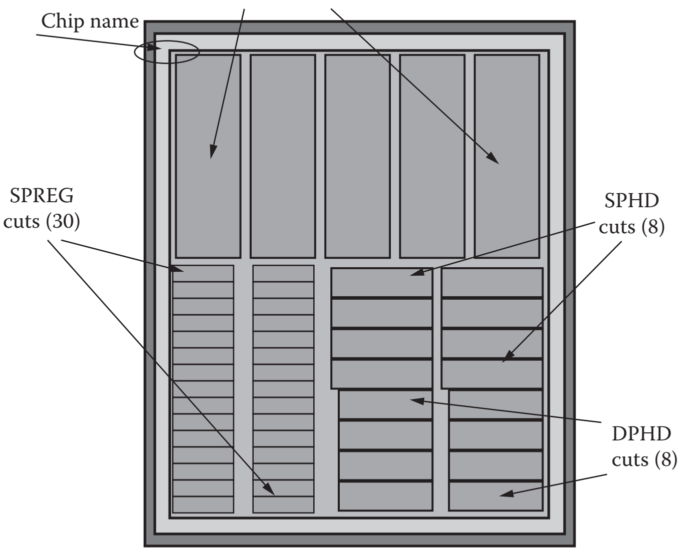


FIGURE 8.3 Floorplan of the test vehicle designed and manufactured in a 65nm CMOS technology.

Bit cell	Bit cell Area	Capacity	DNW
Single Port SRAM High Density	0.52 µm^2	2 Mb	No
Single Port SRAM High Density	0.52 µm^2	2 Mb	Yes
Single Port SRAM Standard Density	0.62 µm^2	2 Mb	No
Single Port SRAM Standard Density	0.62 µm^2	2 Mb	Yes
Dual Port SRAM High Density	0.98 µm^2	1 Mb	No
Dual Port SRAM High Density	0.98 µm^2	1 Mb	Yes

FIGURE 8.4 Content of the test vehicle. Three different bitcell architectures were embedded. Every bitcell is processed with and without triple well layer.

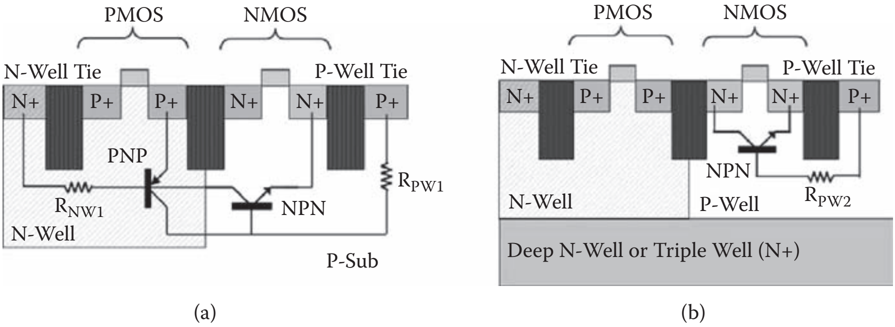


FIGURE 8.5 Schematic cross section of a CMOS inverter (a) without triple well and (b) with triple well. The PNP base resistance R_{NW1} is lowered by the TW: the PNP cannot be triggered. Conversely, the TW layer pinches the P-Well and increases the NPN base resistance R_{PW2}: the NPN triggering is facilitated. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering" *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

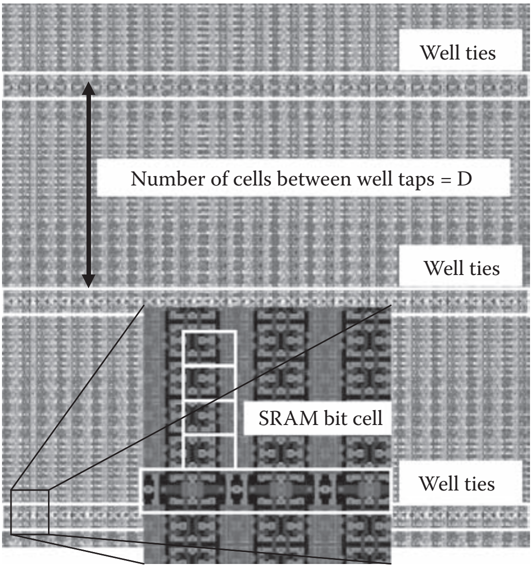


FIGURE 8.6 Layout of an SRAM cell array showing the periodical distribution of the well tie rows every 32 cells. (Reprinted with permission from, Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

reduced (Figure 8.1). TW makes the latchup thyristor more difficult to trigger on. In the literature, full latchup immunity is reported even under extreme conditions (high voltage, high temperature, and high linear energy transfer [LET]) [25,26]. Second, this buried layer allows for concurrently decreasing the single-event upset/soft error rate (SEU/SER) sensitivity since the electrons generated deep inside the substrate are collected by the TW layer and then evacuated through the n-well ties. The improvement of the SER using TW is reported in several papers [27-29]. However, other research teams have published an increased SER sensitivity due to the TW in a commercial CMOS 0.15 µm technology [30,31].

8.3 EXPERIMENTAL RESULTS

Multiple-cell upsets were recorded during the SER experiments on the 65 nm SRAM, but no MBU was ever detected as the tested memory uses bit interleaving or scrambling. All the MCU percentages reported in this work were computed by dividing the number of upsets from MCU by the total number of upsets (single-bit upsets [SBUs] plus MCUs). Note that in the literature, events are sometimes used instead of upsets [31]; the MCU percentages are in this case significantly underestimated. Unless otherwise specified, tests were performed at room temperature, in dynamic mode with checkerboard and uniform test patterns. In addition to the usual MCU percentages, we report in this work the failure rates due to MCU (also called MCU rate). MCU rates allow comparing quantitatively MCU occurrence between different technologies and test conditions.

8.3.1 MCU as a Function of Radiation Source

The four radiation sources have a different interaction mode, which is either directly ionizing (alpha and heavy ions) or nonionizing (neutron and protons). However, it is of interest to compare the MCU percentage from these radiations on the same test vehicle. The test vehicle chosen is a single-port SRAM of standard density processed without triple well. MCU percentages are synthesized in Table 8.1, which shows that alpha particles lead to the lower MCU occurrence. Moreover, heavy ions lead to the higher MCU percentages while neutrons and protons are similar. Heavy ions are the harshest radiation MCU-wise.

8.3.2 MCU as a Function of Well Engineering: Triple-Well Usage

Table 8.2 synthesizes and compares MCU rates and percentage for the standard density SP SRAMs processed with and without triple well. Table 8.3 indicates first that the usage of TW increases the MCU rate by a decade and the MCU percentage by a factor of × 3.6. Usage of MCU rate is mandatory since MCU percentages can lead to incomplete information. As presented in Figure 8.7, devices without TW have a lower number of bits involved per MCU event (≤ 8) compared with those with TW. This figure also indicates that for SRAMs with triple-well three-bit and four-bit MCU events are more likely than two-bit events.

TABLE 8.1
Percentage of MCU for the Same Single Port SRAM under Several Radiation Sources

Radiation Source	Single Port SRAM Standard Density CKB pattern no triple well
Alpha	0.5%
Neutron	21% @ LANSCE
Proton	4% @ 10 MeV
	20% @ 40 MeV
	25% @ 60 MeV
Heavy Ions	0% @ 5.85 MeV/cm2.mg
	87% @ 19.9 MeV/cm2.mg
	99.8% @ 48 MeV/cm2.mg

TABLE 8.2
MCU Rates and Percentages of a Single Port SRAM Processed with and without Triple Well

	MCU Rate	% MCU
SP SRAM standard Density No triple well	100 (norm)	21
SP SRAM standard Density Triple well	1000	76

Note: MCU rate is normalized to its value without triple well.

TABLE 8.3
MCU Percentages and Rates after Neutron Irradiation at Nominal Voltage and Room Temperature for Two Different Test Patterns

		CKB Pattern	
Technology	**Bitcell Area**	**MCU %**	**MCU Rate (au)**
Bulk	2.5 μm^2	16.90	100
SOI	2.5 μm^2	2.10	10

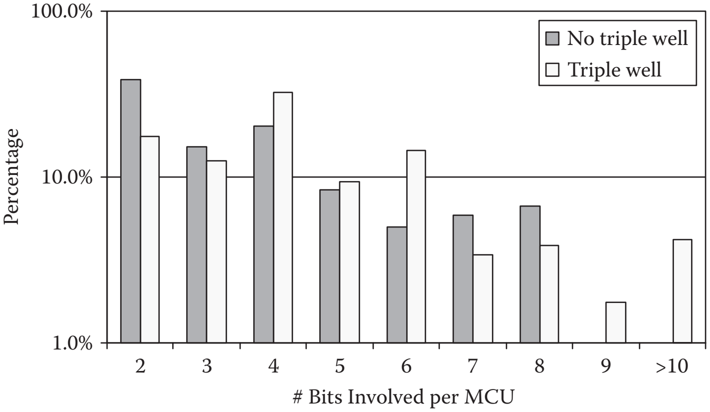


FIGURE 8.7 Number of bits involved in MCU events for high density SP-SRAM under neutron irradiation. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering". *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

The effect of the triple-well layer on MCU percentages under heavy ions is reported in Figure 8.8. The SRAM under test is a high-density (HD) SP SRAM. For the smallest LET MCUs represent 90% of the events with TW but less than 1% without TW. For LET_{eff} higher than 5.85 MeV.cm²/mg there is no SBU in the SRAM with TW. For LET higher than 14.1, all the MCU events induce more than five errors with TW. With TW, the significant increase in MCU amount and order causes an increase in the error cross section.

Whatever the radiation source, the usage of triple well strongly increases the occurrence of MCU. This increase is so high that it can be seen in the total bit error rate for neutrons and error cross section for heavy ions.

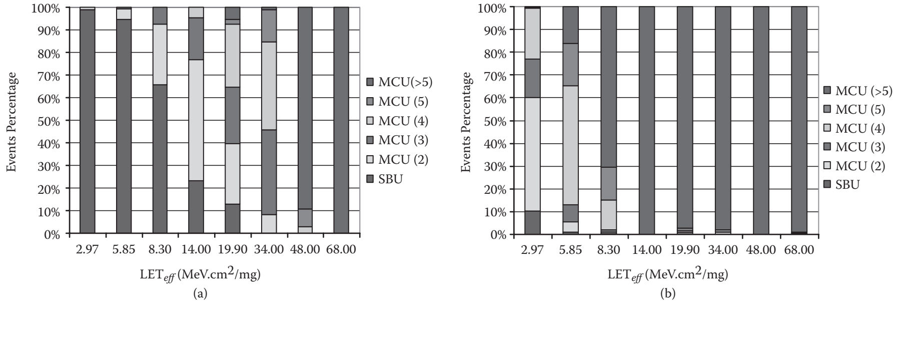


FIGURE 8.8 Proportion for Single and Multiple Event for (a) high density SP SRAM without triple well option (b) high density SP SRAM with triple well option. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

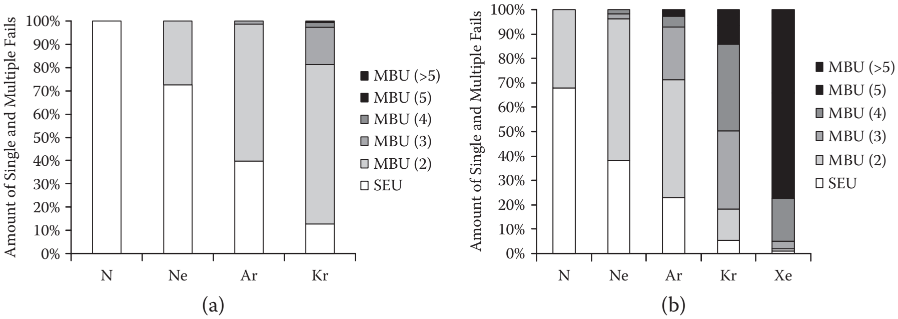


FIGURE 8.9 Amount of bitfails due to single and multiple events in 90nm SP-SRAM: (a) with heavy ion beam not tilted and (b) with heavy ion beam tilted at 60°. (Reprinted with permission from Giot, D., Roche, P., Gasiot, G., Harboe-Sorensen, R., "Multiple-bit upset analysis in 90 nm SRAMs: Heavy ions testing and 3D simulations". *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 4, 904–911.)

8.3.3 MCU as a Function of Tilt Angle during Heavy-Ion Experiments

Figure 8.7 shows, respectively, the amount of single- and multiple-bit failures induced by a given ion specie (N, Ne, Ar, Kr) whose tilt angle is either vertical (Figure 8.9a) or tilted by 60° (Figure 8.9b). Tilt angle from 0° to 60° increases the MBU percentages for each ion's species. For nitrogen, the MBU is increased from 0% to 30% with a tilt = 60°. For neon and argon, the amount of MBU failures is doubled at 60° compared with vertical incidence. For krypton, the increase of MBU with the tilt is less pronounced (+10% from 0° to 60°) because of the progressive substitution of low-order MBUs (order 2, order 3) by higher-order MBUs (order 5, order > 5).

On average the amount of bit failures due to MBU is doubled for 60° tilt compared with normal incidence [33].

8.3.4 MCU as a Function of Technology Feature Size

Figure 8.10 shows the experimental neutron MCU percentages as a function of technology feature size and compares data from this work with data from the literature. These data show that technologies with triple well have MCU percentages higher than 50% while technologies without have MCU percentages lower than 20%. Data from the literature fit either our set of data with triple well or without triple well. Consequently, Figure 8.8 suggests that MCU percentages can be sorted with a criterion of triple-well usage. Moreover, the MCU percentages increase with and without TW when the technologies scale down. This slope being higher without TW since for old technologies MCU percentages were very low (~1% in 150 nm).

8.3.5 MCU as a Function of Design: Well Tie Density

TCAD simulations on 3-D structures built from the layout of the tested SRAMs have been performed as shown in Section 8.4. Simulation results for the ratio between

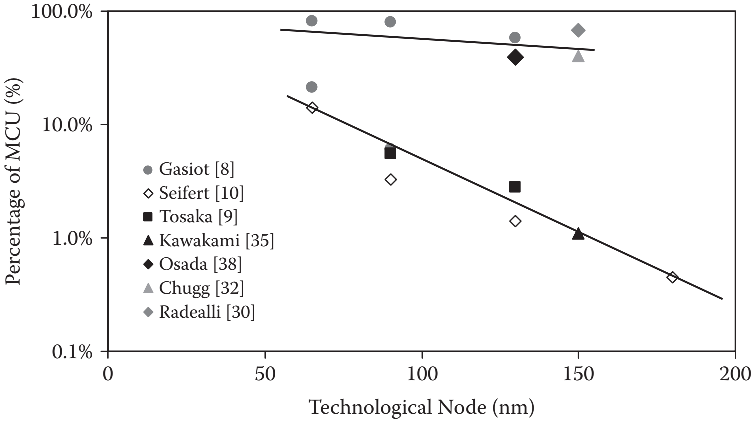


FIGURE 8.10 Neutron-induced MCU percentages as a function of technological node from this work and from the literature. Triple well usage is not indicated in the data from the literature. (Reprinted with permission from, Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

drain collected charge with and without triple well are plotted in Figure 8.11. This figure indicates first that the collected charge with triple well is higher than without for the three well tie frequencies that were simulated. Second, the charge collection increase ranges from ×2.5 to ×7 for the highest and the lowest well tie frequency, respectively. This demonstrates that when triple well is used, increasing the well tie frequency mitigates the bipolar effect and therefore the MCU rate and SER.

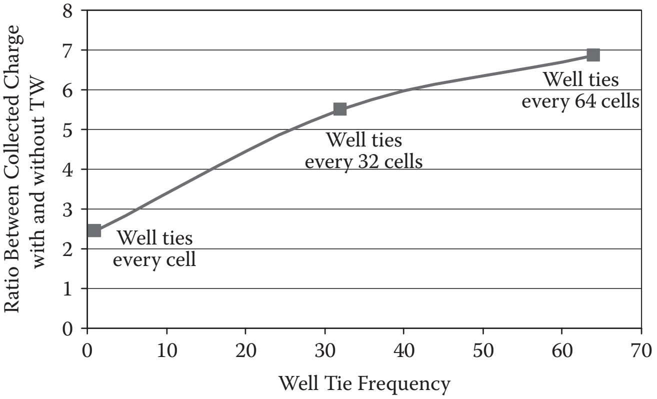


FIGURE 8.11 Simulation results for the ratio between collected charge by the N-Off drain with and without triple well. This ratio is plotted as a function of well ties frequency. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

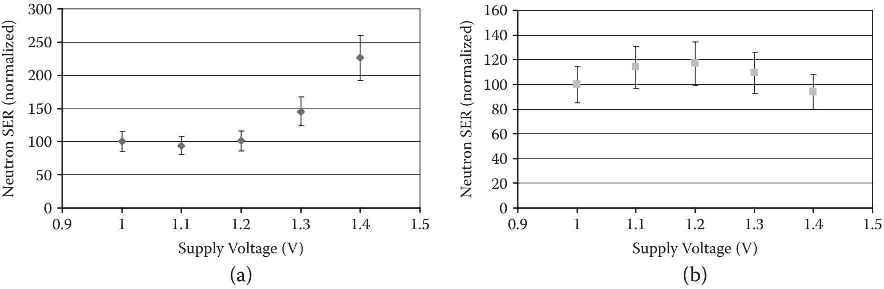


FIGURE 8.12 MCU rate as a function of supply voltage for the HD SRAM processed (a) without triple well and (b) with triple well process option. MCU rate are normalized to their value at 1V.

8.3.6 MCU as a Function of Supply Voltage

The effect of supply voltage on the radiation susceptibility is well known: the higher the voltage, the lower the susceptibility since the charge storing the information is increased proportionally to the supply voltage. However, the effect of the supply voltage on the MCU rate is not documented. Experimental measurements were performed at LANSCE on an HD SRAM processed with and without triple-well option at different supply voltages ranging from 1 V to 1.4 V. Results are synthesized in Figure 8.12. It shows that when the supply voltage is increased the device with triple-well MCU rate remains constant within the experimental uncertainty. However, a different trend is observed for the device without triple-well layer. When the supply voltage is increased the MCU rate is constant from 1.0 V to 1.2 V and then increases from 1.3 V to 1.4 V. The MCU rate increase is 220% for V_{DD} equal to 1.4 V.

8.3.7 MCU as a Function of Temperature

A high-temperature constraint is associated with high-reliability applications such as automotive. Some papers have quantified the temperature effect on SER or heavy-ion susceptibility [37,38]. At the time of this writing no reference can be found in the literature experimentally measuring the temperature effect on the MCU rate. Experimental measurements were performed at LANSCE on an HD SRAM processed with and without triple-well option at room temperature and 125°C. Results are synthesized in Figure 8.13. It demonstrates that the MCU rate increases by 65% for the device without triple well and by 45% for the device with triple well. Note that the usage of MCU percentage would have been misleading since the MCU percentage is constant between room temperature and 125°C for the device with triple well.

8.3.8 MCU as a Function of Bit Cell Architecture

Figure 8.14 synthesizes MCU rates for high-density and standard-density (SD) single-port SRAMs as well as a dual-port SRAM (eight transistors). These SRAMs were processed without triple well. Figure 8.14 indicates that the higher the density the

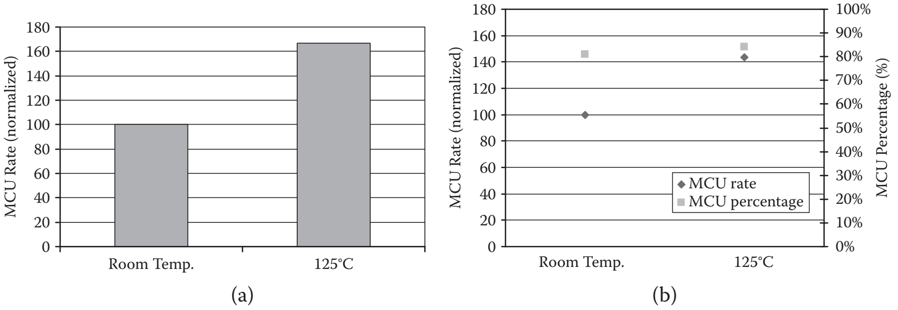


FIGURE 8.13 MCU rate as a function of temperature for the HD SRAM processed (a) without triple well and (b) with triple well process option. MCU rate are normalized to their value at room temperature. Figure xb also displays the MCU percentages.

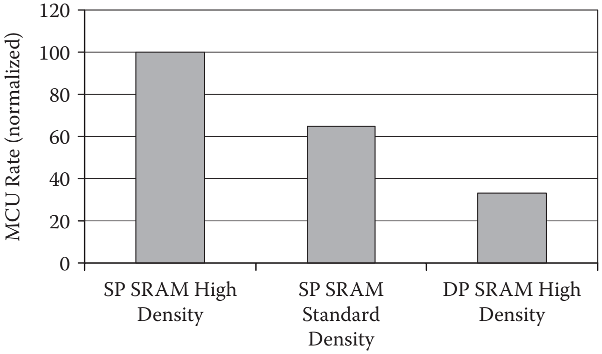


FIGURE 8.14 MCU rate comparison for several bitcell architectures. SP stands for Single Port, DP for Dual Port (8Transistor SRAM). The devices under test were processed without triple well.

higher the MCU rate. A decrease in the bit cell area by a factor of 2 (HD SP SRAM compared with DP SRAM) induces a decrease in the MCU rate by a factor of 3.

The effect of bit cell architecture on MCU percentages under heavy ions is reported in Figure 8.15. The devices under test are HD SP SRAMs (Figure 8.15a) and SD SP SRAMs (Figure 8.15b). Figures 8.5a and 8.5b show the respective amount of SBU and MCU events for experimental ion LET ranging from 2.97 to 68 MeV/cm².mg. For the HD SRAM, the first MCU occurs below 2.97 MeV/cm².mg, while for the SD SRAM it occurs between 5.85 and 8.30. For higher LET, the amount and the order of the MCU events increase while the proportion of SBU decreases. For every LET, the SBU component is the highest for the lowest density memory (SD SRAM) while the MCU component is the highest for the highest density SRAM (HD SRAM) [32].

8.3.9 MCU as a Function of Test Location LANSCE versus TRIUMF

Several facilities around the world provide white neutron beams for SER characterization. An exhaustive list of these facilities can be found in the JEDEC test standard [20].

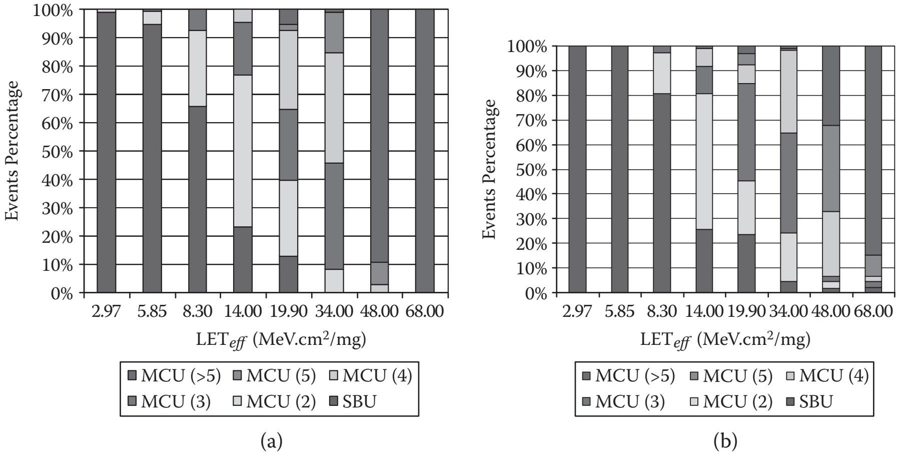


FIGURE 8.15 Amount of bitfails due to single and multiple upsets: (a) for high density SP-SRAM, (b) for standard density SP-SRAM. (Reprinted with permission from Giot, D., Roche, P., Gasiot, G., Autran, J.-L., Harboe-Sorensen, R., "Heavy ion testing and 3-D simulations of multiple cell upset in 65 nm standard SRAMs." *IEEE Transactions on Nuclear Science*, Volume: 5 Issue: 4, 2048–2054.)

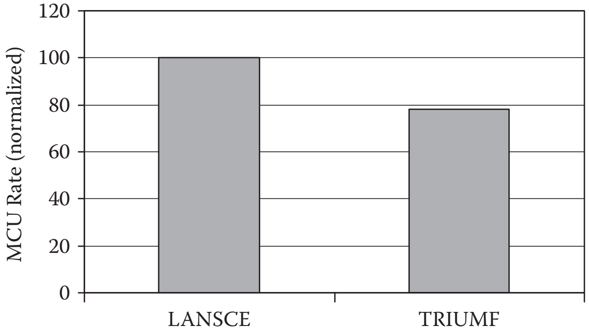


FIGURE 8.16 MCU rate comparison between LANSCE and TRIUMF white neutron beam sources. The device under test is a high density SRAM processed with triple well.

The best known facilities are LANSCE and TRIUMF. Experimental measurements on the same test chip embedding an HD SP SRAM processed with triple-well option were performed at these two facilities. The MCU percentages were perfectly equal to 76% for both facilities. The MCU rates are reported in Figure 8.16. It shows that the MCU rate decrease by 22% at TRIUMF compared with LANSCE. This can be explained by the energy cut-off, which is 800 MeV at LANSCE while it is 500 MeV at TRIUMF.

8.3.10 MCU as a Function of Substrate: Bulk versus SOI

SRAMs were manufactured with a CMOS 130 nm commercial technology either bulk or silicon-on-insulator (SOI). For comparison purposes both SRAM designs are

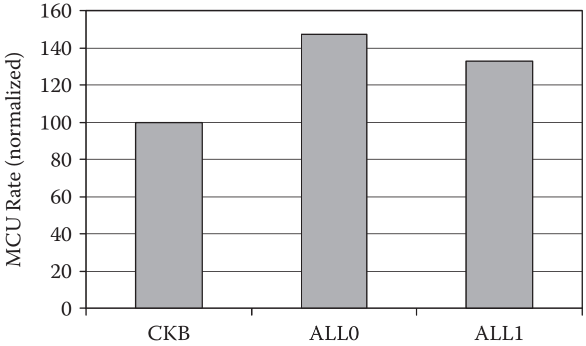


FIGURE 8.17 MCU rate comparison for several test patterns. CKB stand for Checkerboard, ALL0 and ALL1 for uniform of 0 and 1 respectively. Note that test patterns are physical. The device under test is a high density SRAM processed without triple well.

strictly identical. The test chip contains 4 Mb of single-port SRAMs in which two different bit cell designs were embedded. In this work only the SD SRAM will be reported. The bulk technology was processed without a triple-well layer. Table 8.3 therefore synthesizes the failure rates due to MCU (also called the MCU rate) and the MCU percentage for a single test pattern (CKB). It is noteworthy from Table 8.2 that SOI SRAMs have a much lower MCU rate and percentage compared with bulk. More parameters (pattern, bit cell area, supply voltage) were studied in [39].

8.3.11 MCU as a Function of Test Pattern

An HD SRAM was measured at LANSCE with several test patterns using a dynamic test algorithm. Results are synthesized in Figure 8.17, which shows that uniform patterns have a higher MCU rate than the CKB. To understand the reason for this discrepancy, it is necessary to plot the topological shape of experimental two-bit MCU events as a function of pattern filling the memory during the testings (Figures 8.18a and 8.18b). The prevailing shape for two-bit MCU and a checkerboard pattern is "diagonal adjacent," while it is "column adjacent" with uniform pattern (as observed in [36]). 3-D TCAD simulations have shown that two-bit MCU threshold LET is the lowest for two bit cells in the column (see [33] and Section 8.4.2). It is therefore consistent that uniform patterns have a higher MCU rate since their error clusters are the easiest to trigger.

It is also noteworthy from Figures 8.18a and 8.18b that triple-well usage did not modify the prevailing shape of MCU for a checkerboard or for a uniform pattern.

8.4 3-D TCAD MODELING OF MCU OCCURRENCE

The previous section clearly highlighted the importance of triple well in the MCU response. In this section 3-D TCAD simulations are set up to analyze the increased MCU occurrence when triple well is used. All 3-D SRAM structures in this section were built using a methodology described in [40] and the Tool Suite v10.0 of

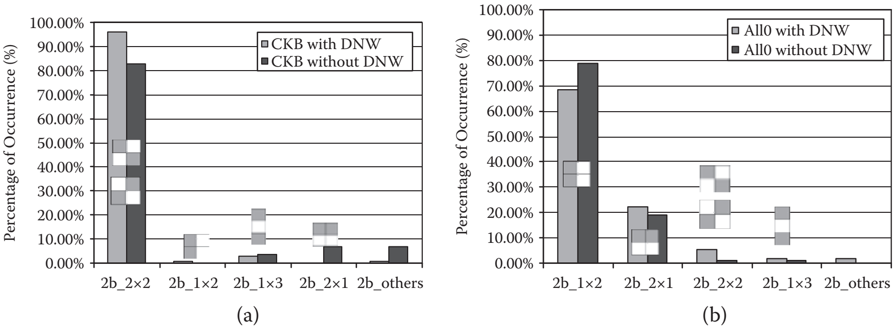


FIGURE 8.18 2-bits MCU cluster shape on high density SP SRAM processed with or without triple well after neutron irradiation when the test pattern is (a) a checkerboard or (b) a uniform pattern. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

the Sentaurus Synopsys package [41]. Cell boundaries are defined from the computer-aided design (CAD) layout and technological process steps. One-dimensional (1-D) doping profiles are precisely modeled from secondary ion mass spectrometry (SIMS) profiles. Cell boundaries are defined from the CAD layout and technological process steps. 1-D doping profiles are included to define n-well, p-well (with a 4 μm epi layer thickness), and active regions of transistors (Figure 8.19). Mesh refinements are included in regions of interest: channels, lightly doped drain (LDD), junction

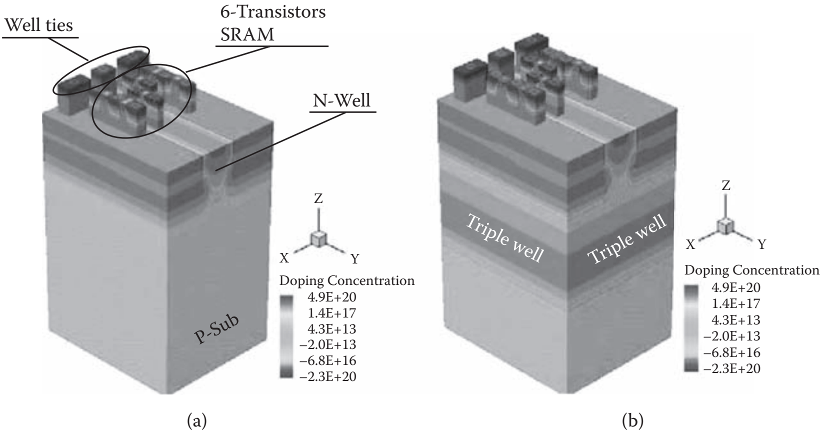


FIGURE 8.19 Full 3-D structures of the 65nm 6T SRAM located as close as possible to the well ties (a) without triple well and (b) with triple well. Two NMOS are embedded per PWell (one is a part of the inverter, the other is an access transistor). (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering" *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

boundaries (to tackle short channel effects), and a round ion track (to allow accurate generation of carriers in silicon). Wire connections between the different electrodes of the cell are modeled in the SPICE domain (mixed-mode TCAD simulations) to reduce the CPU burden. The parasitic circuit capacitances due to metallization layers are also taken into account.

Device simulations with ion impacts are performed using the Sentaurus device simulator. For this purpose, several physical models are activated: drift diffusion for carriers' transport; Shockley-Read-Hall and Auger for recombination; electric field and doping-dependent models for mobility; and heavy-ion module for carrier deposition along the particles track. The heavy-ion generation model uses a Gaussian radial distribution of charges with a fixed characteristic radius of 0.1 μm and a Gaussian time distribution centered at 1 ps. An additional assumption consists of taking a constant LET along the track because of the low diffusion depth of transistor active areas (~0.2 μm). Properties of boundaries are defined by the Neumann reflective conditions [40,41].

8.4.1 Bipolar Effect in Technologies with Triple Well

For an in-depth analysis of the MCU phenomenon, 3-D device simulations were performed on full SRAM bit cells. Ion strikes were located in the most sensitive MCU location (source of the SRAM) for different distances from the well taps, with and without triple well. It is noteworthy that Osada et al. [40] already tried to model the effect of the parasitic bipolar amplification on the MCU. A more simple mix of device (2-D uniformly extended) and circuit simulations was used but not for the worst sensitive location for MCU occurrence [33].

8.4.1.1 Structures Whose Well Ties Are Located Close to the SRAM

Figure 8.19 presents the 3-D SRAM bit cell made up of six transistors (6T), two p-wells, one n-well, and three well ties. The well ties are as close as possible to transistors. The simulation results of these structures are presented in Figure 8.20, which compares source and drain currents after an ion impact in the source at 1 ps. The charge collected at the n-off drain is slightly higher with triple well when well ties are located close to the SRAM transistors. With triple well a limited bipolar effect (see the next section for details on bipolar triggering) is observed for structures close to the ties. These simulation results are consistent with the experimental results presented in [22,30], which have shown that MCU occurrence is less likely close to well ties.

8.4.1.2 Structures Whose Well Ties Are Located Far from the SRAM

A second set of 3-D structures were built to model the effect of the spacing between well ties and SRAM cells with and without the TW doping profiles. Figures 8.21a and 8.21b illustrate four structures dedicated to well tie frequency modeling. The simulation results are presented in Figure 8.22 for ion features (LET and strike location) identical to those used in Figure 8.20. The charge injected by the source and the charge collected at the n-off drain are much higher with triple well when well ties are located away from the SRAM transistors.

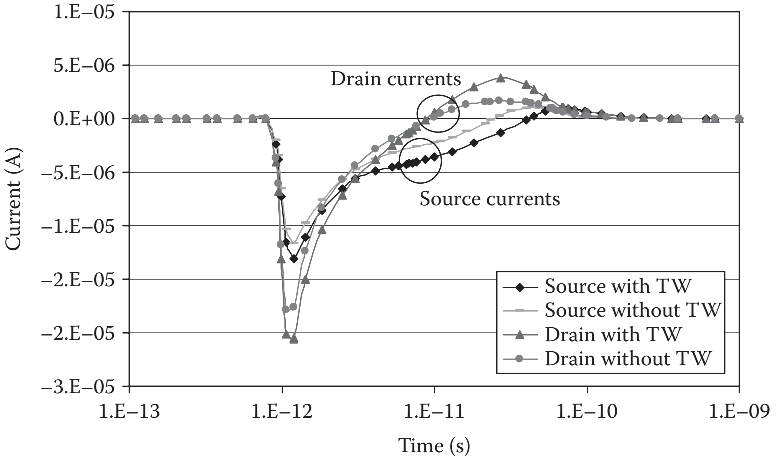


FIGURE 8.20 Full 3-D TCAD simulations results on the structure presented in Figure 8.9 (6T SRAM very close to the well taps) show a limited bipolar effect due to the presence of the triple well layer. Heavy ion LET is 5.5fC/μm. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

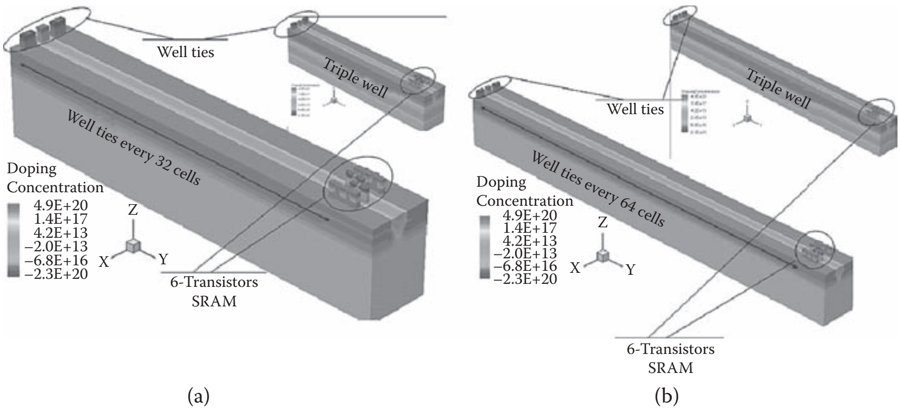


FIGURE 8.21 Full 3-D structures of the 65nm 6T SRAM whose well ties are located (a) 32 cells and (b) 64 cells away from the well taps without triple well. Same structures with triple well are shown in the upper right inserts. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

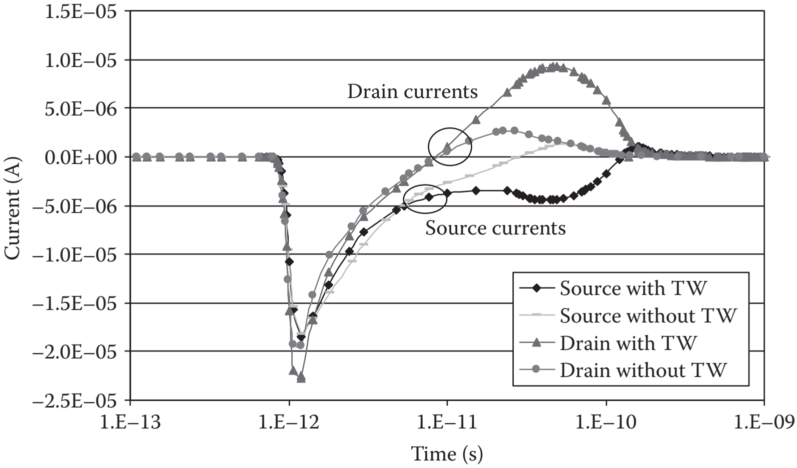


FIGURE 8.22 Full 3-D TCAD simulations results on the structure presented in Figure 8.11a. Source current shows a strong bipolar effect due to the presence of the triple well layer. Heavy ion LET is 5.5fC/μm. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOS SRAMs and its dependence on well engineering." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2468–2473.)

Injected carriers by the source are forerunners of the bipolar transistor triggering. Ion deposited majority carriers flow toward the well ties. The well resistance causes a voltage drop beneath source diffusion. If enough carriers are deposited or if there is enough distance between well ties and ion impact (the higher the distance the higher the voltage drop) the source-well junction will therefore be turned on, and additional carriers will be injected in the well (Figure 8.23). Most of these additional carriers will be collected at the drain junction and will thus increase the collected charge at the drain. The additional charge collection due to the source injection and to the parasitic bipolar action is responsible for the bit cell upset. Moreover, voltage drop in the well can turn on several sources along the well that will upset several bit cells and be responsible for the MCU pattern experimentally reported in Section 8.3.11.

The simulations have shown that with triple well a strong bipolar effect (electron injection from the sources) is observed for structures away from the ties. These simulation results are consistent with the experimental results presented in [22,30], which have shown that MCU occurrence is more likely away from well ties.

8.4.2 A Refined Sensitive Area for Advanced Technologies

This section aims to show by means of 3-D TCAD simulations that the bit cell SEE sensitive area is not restricted to the area of reverse-biased junctions. Figure 8.24 shows the 3-D TCAD final structures of two SP bit cells arranged in "column" (a) and "row" (b). These continuous TCAD domains include, respectively, 710,000 and 580,000 elements. The double bit cell structures are dedicated to double MBU

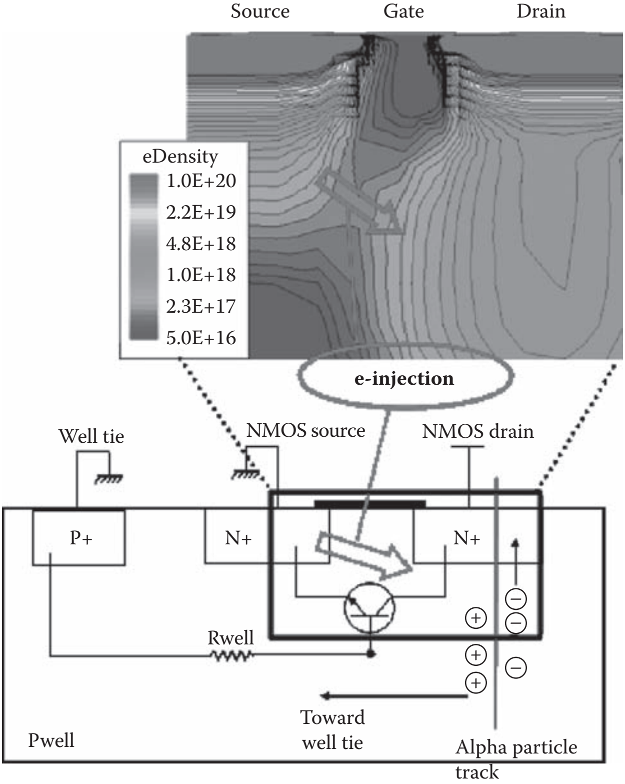


FIGURE 8.23 Illustration of the carrier injected by the source and triggering of the parasitic bipolar transistor after an alpha particle strike in the drain. Insert is from device simulation of the 65nm 3D structure. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Alpha-induced multiple cell upsets in standard and radiation hardened SRAMs manufactured in a 65 nm CMOS technology." *IEEE Transactions on Nuclear Science*, Volume: 53 Issue: 6, 3479–3486.)

studies. CPU burden is around one week to simulate a double SRAM structure with up-to-date high-performance workstations.

Figure 8.25 shows an area of four SP bit cells. Two bit cells of the same column share the sources of their MOS transistors, whereas two bit cells of the same row do not share a p-n junction and are isolated with shallow trench isolation (STI). At first order, an MBU of two adjacent cells is horizontal, vertical, or diagonal (configuration 1, 2, and 3 in Figure 8.25). The third case of diagonal double MBU was not simulated. Indeed, diagonal MBU would provide a higher MBU LET than one computed for row MBU because of the longer distance between the adjacent SEU sensitive areas (both are separated with STI).

8.4.2.1 Simulation of Two SRAM Bit Cells in a Row

The most efficient memory pattern to trigger a double row MBU is to reverse-bias neighboring drains. This is obtained with the logical pattern "01" (Figure 8.26).

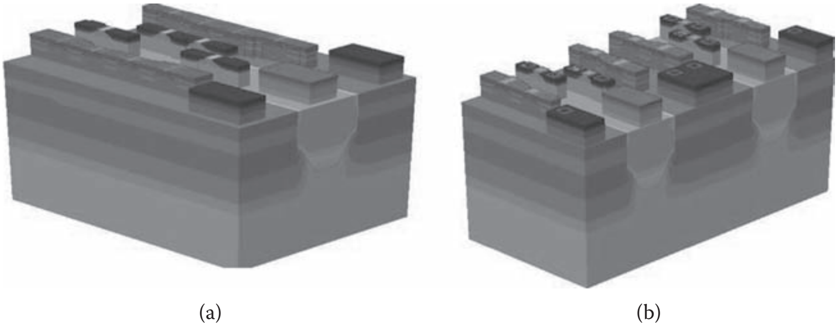


FIGURE 8.24 SRAM 3D structures (STI not displayed for clearness): (a) Double 6T bitcells in column, (b) Double 6T bitcells in row. (Reprinted with permission from Giot, D., Roche, P., Gasiot, G., Harboe-Sorensen, R. Multiple-bit upset analysis in 90 nm SRAMs: Heavy ions testing and 3D simulations. *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 4, 904–911.)

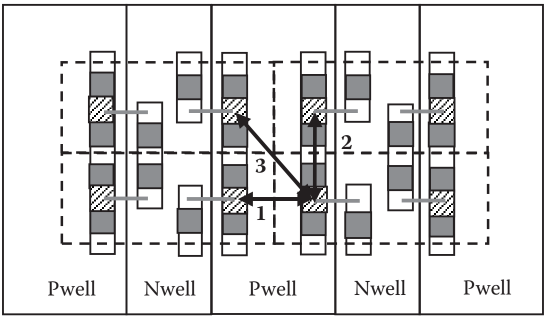


FIGURE 8.25 Four contiguous SRAM bitcells: dashed rectangles are bitcells. Connected striped and white squares are respectively drains of NMOS and PMOS transistors. Single grey and white squares are gates and sources of NMOS and PMO. (Reprinted with permission from Giot, D., Roche, P., Gasiot, G., Harboe-Sorensen, R., "Multiple-bit upset analysis in 90 nm SRAMs: Heavy ions testing and 3D simulations." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 4, 904–911.)

In row configuration PMOS cannot trigger MCU since they are separated by two reverse-biased n-well/p-well junctions. MCU threshold LET were computed for two ion locations shown in Figure 8.26. Table 8.5 synthesizes these LET and shows that an ion crossing an NMOS drain requires at least an LET of 13.5 MeV.cm²/mg to create an MCU while an ion at mid-distance between two NMOS drains requires a lower LET (8.5 MeV.cm²/mg). The gray area in Figure 8.26 shows the extrapolated spread out of the sensitive area for row MBU until an LET of 13.5 MeV.cm²/mg.

8.4.2.2 Simulation of Two SRAM Bit Cells in a Column

For the configuration depicted in Figure 8.27, the most efficient memory pattern to induce MBU is "11" or "00" because the transistors of adjacent bit cells (particularly

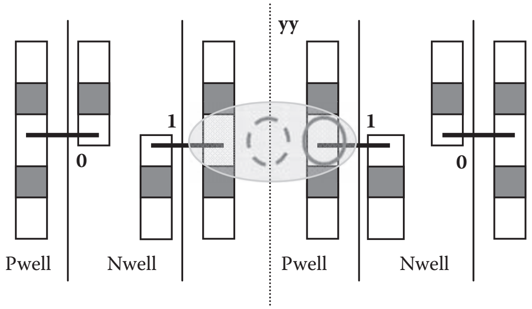


FIGURE 8.26 Scheme of the layout for 2 SRAMs bitcells arranged in row. Plain circle is an ion impact in the NMOS drain (most sensitive Single Bit Upset location) while dashed circles is an impact at mid-distance between two NMOS drains. Grey area is the spread of MCU sensitive area at a LET of 13.5 MeV.cm²/mg. (Reprinted with permission from Giot, D., Roche, P., Gasiot, G., Harboe-Sorensen, R., "Multiple-bit upset analysis in 90 nm SRAMs: Heavy ions testing and 3D simulations." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 4, 904–911.)

TABLE 8.4
Simulated MCU Threshold LET for Two Single Port SRAMs Arranged in Row and in Column

TCAD Structure	Ion Location	LETth (MeV.cm²/mg)
Double Row MBU	NMOS drain	13.5 ± 0.5
	Mid-distance between NMOS drains	8.5 ± 0.5
Double Column MBU	NMOS drain	11.5 ± 0.5
	Mid-distance between NMOS drains	3.75 ± 0.25
	Mid-distance between PMOS drains	5.25 ± 0.25

TABLE 8.5
Relative Neutron MCU Rate Variation as a Function of Several Parameters

Parameter	Details in Section	Relative MCU Rate
SOI Substrate[a]	3.10	10
Bitcell architecture	3.8	30
Reference 65nm Single Port SRAM without triple well	—	100
Test location	3.9	125
Test Pattern	3.11	145
Temperature	3.7	165
Supply Voltage	3.6	230
Triple well usage	3.2	1000

[a] Experimental results in 130 nm technology.

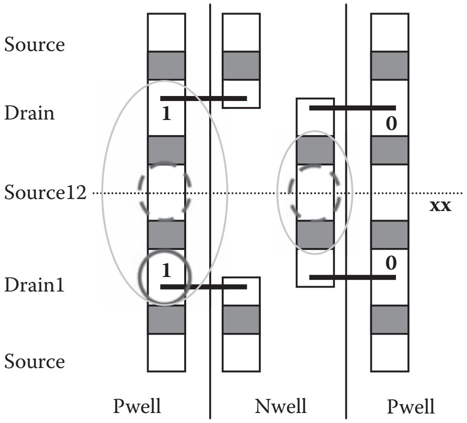


FIGURE 8.27 Scheme of the layout for 2 SRAMs bitcells arranged in column. Solid line circle is an ion impact in the NMOS drain (most sensitive Single Bit Upset location) while dashed circle is an impact at mid-distance between two NMOS or PMOS drains. Grey area is the spread of MCU sensitive area at a LET of 11.5 MeV.cm²/mg. (Reprinted with permission from Giot, D., Roche, P., Gasiot, G., Harboe-Sorensen, R., "Multiple-bit upset analysis in 90 nm SRAMs: Heavy ions testing and 3D simulations." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 4, 904–911.)

SEU sensitive areas) share the same well region and are separated by the same distance. Note that MCU can be triggered by NMOS as well as PMOS.

MCU threshold LET were computed for three ion locations schematized in Figure 8.27. MCU LET values are synthesized in Table 8.5. As already observed for row configuration, the lowest LET is obtained for an ion impact at mid-distance between NMOS drains (3.75 MeV.cm²/mg). MCU LET for an impact at mid-distance between PMOS drains is, however, slightly higher (5.25 MeV.cm²/mg). The gray areas in Figure 8.10(b) show the extrapolated spread out of the sensitive area for column MCU until an LET of 11.5 MeV.cm²/mg.

8.4.2.3 Conclusions and SRAM Sensitive Area Cartography

Despite a smaller distance between two adjacent SEU sensitive areas, the row MCU LET is twice as high as the column MBU LET. This is explained by the incidence of the ion that crosses through 0.3 μm of STI in the first case (dashed circle in Figure 8.26) whereas it directly strikes the active area of NMOS transistor in the second case (dashed circle in Figure 8.27). As a consequence, there is less silicon volume for the carriers' deposition in the case of row MBU. Row and column LET show that the layout of the memory cells (STI regions, silicon regions) strongly impacts their sensitive area.

SEE sensitive area cartography as a function of ion LET can be drawn from TCAD results shown in Sections 8.4.2.1 and 8.4.2.2. This cartography is shown in Figure 8.28. It is noteworthy that the double MBU sensitive area extends beyond a single bit cell area.

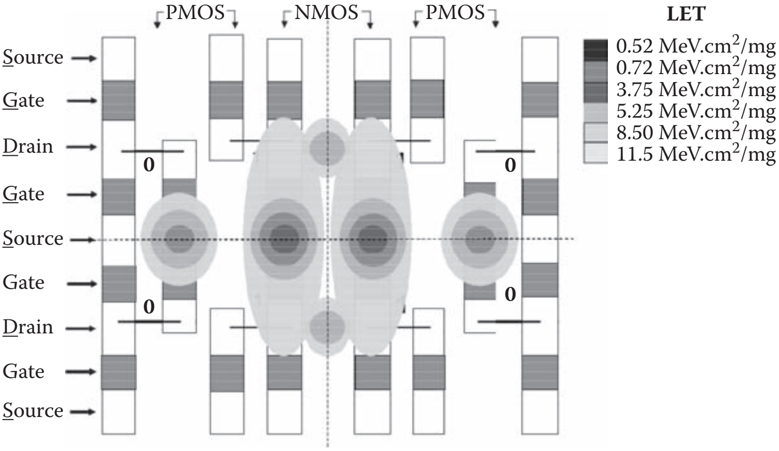


FIGURE 8.28 SEE sensitive area cartography as a function of ion LET.

8.5 GENERAL CONCLUSION: SORTING OF PARAMETERS DRIVING MCU SENSITIVITY

SEE testing carried out with alpha, neutrons, heavy ions, and protons on several SRAMs is reported in this chapter. These SRAMs were processed by STMicroelectronics in a CMOS 65 nm technology and were embedded in several test vehicles. MCU percentages and MCU rates were given as a function of a dozen parameters. These parameters are either technological (e.g., feature size, process option) or design (e.g., bit cell architecture, well tie density) or related to experimental test conditions (e.g., supply voltage, temperature, test pattern). Table 8.5 synthesizes the relative neutron MCU rate variations as a function of these parameters. It is noteworthy that the use of SOI substrate is the solution that will decrease the most MCU rate by taking advantage of its fully isolated transistors. Parameter which increases the most, the MCU rate is the use of triple well layer process option.

8.5.1 Experimental Results in 130 nm Technology

Full 3-D structures were built from a layout of 65 nm SRAM bit cells. The use of TCAD structures whose SRAM bit cells are located away from the well ties was mandatory to confirm that the bipolar effect enhances the collected charge with triple well. The simulations have additionally confirmed that the bipolar effect is reduced by increasing the well tie frequency and therefore efficiently mitigates MCU and SER.

Other 3-D structures embedding two SRAM bit cells were built. Bit cells were arranged either in a column or in a row to reproduce an actual SRAM array. Simulation of these structures has allowed a SEE sensitive area cartography to be built as a function of ion LET. This cartography shows that the sensitive area extends beyond a single bit cell area.

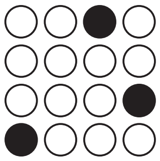
	Cell spacing criterion	MCU detected
	k = 1	no MCU
	k = 2	1 MCU of 2 cells
	k = 3	1 MCU of 3 cells

FIGURE 8.29 Illustration of the impact of cell spacing criterion on the MCU detection efficiency. (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Alpha-induced multiple cell upsets in standard and radiation hardened SRAMs manufactured in a 65 nm CMOS Technology." *IEEE Transactions on Nuclear Science*, Volume: 53 Issue: 6, 3479–3486.)

8.6 ANNEX 1

After radiation testing with a static algorithm, bitmap error can have thousands of SEUs. With such a high density of SEUs the key question is therefore how many upsets are "true" MCU (i.e., several SEUs simultaneously created by a single ion), and how many are "false" MCU (i.e., sequentially created in the same vicinity by different ion strikes)?

MCU rates and shapes depend on the test pattern filling the memory. It was experimentally verified that checkerboard, All1, and All0 test patterns have similar MCU rates. The following analyses and MCU counting are given for the CKB pattern. A cell-spacing (CS) criterion (k) is chosen when analyzing a post-irradiation error bitmap for MCU detection. This criterion corresponds to the upset-to-upset spacing (maximum number of cells between two SEUs in the X and Y directions to count an MCU). The effect of this criterion on the number of counted MCU is illustrated in Figure 8.29. This figure points out that the MCU number (zero or one bitflip) and type (2 or 3 cells) is a function of the CS criterion value: the larger this value (5, 6…), the higher the MCU number. However, a large k value would lead to count as an MCU two single SEU in neighboring cells created by two different events (i.e., not simultaneously generated). This would lead to a large overestimation of the MCU rates.

For this reason, formula (1) is proposed for quantifying the rates of "false" MCU to correct raw experimental data to count only the "true" MCUs. We believe this result is useful in hardness assurance processes. For example, it helps deciding the total number of fails to obtain during radiation experiments and also for the choice of radiation source intensity (here a radioactive alpha source.)

$$\text{false MCU }\% = 1 - e^{-E_{SRP} \times \frac{AdjCell}{Nbit}} \tag{8.1}$$

where E_{SRP} is the number of SEU recorded after irradiation, *AdjCell* is the number of cells around each SEU that are inspected to detect an MCU; and *Nbit* is the size of the memory array.

TABLE 8.6
Number of Adjacent Cells Inspected for MCU around Each Seus as a Function of the Cell Spacing Criterion

Cell spacing criterion	k = 1	k = 3	k = 5	k = 8
# of adjacent cells = AdjCell	8	48	120	288

The probability of counting a "false" MCU is given by

$$P = E_{SRP} \times \frac{AdjCell}{Nbit} \tag{8.2}$$

where E_{SRP} is the number of SEU recorded after irradiation (from a single readout period), *AdjCell* is the number of cells around each SEU that are inspected to detect an MCU (this number is function of the CS criterion (Table 8.6); and *Nbit* is the total number of bits in the memory array.

The probability that an MCU occurred is the complementary probability that no MCU occurred ($n = 0$) and is given using the cumulative Poisson probability by

$$MCU_{proba} = 1 - \sum_{i=0}^{n} \frac{e^{-P} \times P^{i}}{i!} = 1 - e^{-P} \quad \text{for } n = 0 \tag{8.3}$$

Multiplying this probability by the total number of SEU gives the number of multiple-cell upsets. This number divided by the total number of SEU is the percentage of MCU. Using Equations (8.1) and (8.2), the percentage of "false" MCU (SEUs from two different events are counted as an MCU) is

$$\text{false MCU } \% = 1 - e^{-E_{SRP} \times \frac{AdjCell}{Nbit}} \tag{8.4}$$

To double-check the relevance of this model, MCU percentages obtained from Equation (8.1) are compared to MCU percentages counted from randomly generated error bipmaps (Figure 8.30). This figure shows that whatever the CS criterion, the MCU percentages match perfectly.

Equation (8.1) is very convenient as it is easy to use, and it can be used for different devices (e.g., SRAM, DRAM) and many radiation sources (e.g., alpha, neutron, heavy ions).

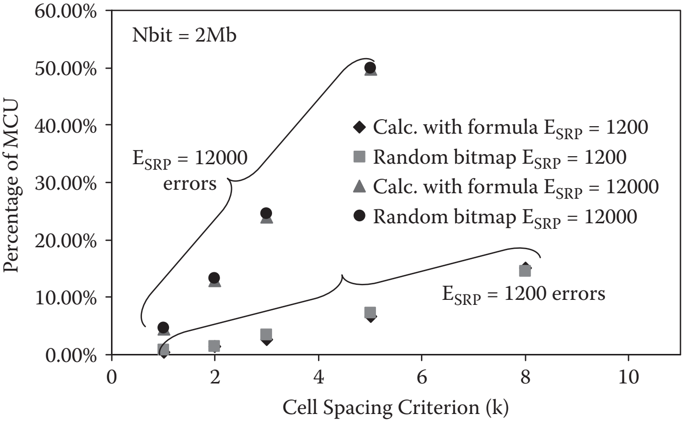


FIGURE 8.30 Comparison of MCU percentages obtained from either a randomly generated bitmap or from formula (1) for a 2Mb memory array (Nbit=2Mb). (Reprinted with permission from Gasiot, G., Giot, D., Roche, P., "Alpha-induced multiple cell upsets in standard and radiation hardened SRAMs manufactured in a 65 nm CMOS technology." *IEEE Transactions on Nuclear Science*, Volume: 53 Issue: 6, 3479–3486.)

REFERENCES

1. X. Zhu, X. Deng, R. Baumann, and S. Krishnan, "A Quantitative Assessment of Charge Collection Efficiency of N+ and P+ Diffusion Areas in Terrestrial Neutron Environment," *IEEE Transactions on Nuclear Science,* Volume 54, Issue 6, pp. 2156–2161, Part 1, Dec. 2007.
2. A.D. Tipton et al., "Device-Orientation Effects on Multiple-Bit Upset in 65 nm SRAMs," *IEEE Transactions on Nuclear Science,* Volume 55, Issue 6, Part 1, pp. 2880–2885, Dec. 2008.
3. V. Correas et al., "Simulations of Heavy Ion Cross-Sections in a 130nm CMOS SRAM," *IEEE Transactions on Nuclear Science,* Volume 54, Issue 6, Part 1, pp. 2413–2418, Dec. 2007.
4. D.G. Mavis et al., "Multiple Bit Upsets and Error Mitigation in Ultra-Deep Submicron SRAMS," *IEEE Transactions on Nuclear Science,* Volume 55, Issue 6, Part 1, pp. 3288–3294, Dec. 2008.
5. F.X. Ruckerbauer and G. Georgakos, "Soft Error Rates in 65nm SRAMs—Analysis of new Phenomena," presented at 13th IEEE International On-Line Testing Symposium, IOLTS 2007.
6. D. Heidel et al., "Single-Event-Upset and Multiple-Bit-Upset on a 45nm SOI SRAM," presented at IEEE International Conference NSREC, Québec City, July 20-24, 2009.
7. S. Uznanski, G. Gasiot, P. Roche, J.-L. Autran, and R. Harboe-Sørensen, "Single Event Upset and Multiple Cell Upset Modeling in a Commercial CMOS 65nm SRAMs," presented at IEEE international RADECS Conference, Bruges, September 14–18, 2009.
8. G. Gasiot, D. Giot, and P. Roche, "Multiple Cell Upsets as the Key Contribution to the Total SER of 65nm CMOS SRAMs and its Dependence on Well Engineering," presented at the 44th Annual International NSREC 2007, Honolulu, HI, July 2007.
9. T. Merelle et al., "Monte-Carlo Simulations to Quantify Neutron-Induced Multiple Bit Upsets in Advanced SRAMs," *IEEE Transactions on Nuclear Science,* Volume 52, Issue 5, pp. 1538–1544, Oct. 2005.

10. Y. Tosaka et al., “Comprehensive Study of Soft Errors in Advanced CMOS Circuits with 90/130 nm Technology,” IEEE International Electron Devices Meeting IEDM Conference, Technical Digest, 2004.
11. N. Seifert et al., “Radiation-Induced Soft Error Rates of Advanced CMOS Bulk Devices,” presented at IRPS Conference, San Jose, 2005.
12. F. Wrobel et al., “Simulation of Nucleon-Induced Nuclear Reactions in a Simplified SRAM Structure: Scaling Effects on SEU and MBU Cross Sections,” *IEEE Transactions on Nuclear Science,* Volume 48, Issue 6, pp. 1946–1952, Dec. 2001.
13. T.L. Criswell, P.R. Measel, and K.L. Wahlin, “Single Event Upset Testing with Relativistic Heavy Ions,” *IEEE Transactions on Nuclear Science,* Volume NS-31, Issue 6, pp. 1559–1561, Dec. 1984.
14. J.A. Zoutendyk, H.R. Schwartz, and R.K. Watson, “Single-Event Upset (SEU) in a DRAM with On-Chip Error Correction,” *IEEE Transactions on Nuclear Science,* Volume NS-34, Issue 6, pp. 1310–1315, Dec. 1987.
15. Y. Song, K.N. Vu, J.S. Cable, A.A. Witteles, W.A. Kolasinski, R. Koga, et al., “Experimental and Analytical Investigation of Single Event Multiple Bit Upsets in Polysilicon Load 64k NMOS SRAMs,” *IEEE Transactions on Nuclear Science,* Volume NS-35, Issue 6, 1673, 1988.
16. J.J. Wang et al., “Single Event Upset and Hardening in 0.15 μm Antifuse-Based Field Programmable Gate Array,” *IEEE Transactions on Nuclear Science,* Volume 50, Issue 6, Dec. 2003.
17. R.A. Reed et al., “Heavy Ion and Proton-Induced Single Event Multiple Upset,” *IEEE Transactions on Nuclear Science,* Volume 44, Issue 6, Part 1, pp. 2224–2229, Dec. 1997.
18. N. Seifert, B. Gill, K. Foley, and P. Relangi, “Multi-Cell Upset Probabilities of 45nm High-k + Metal Gate SRAM Devices in Terrestrial and Space Environments,” *IEEE International Reliability Physics Symposium IRPS,* April 27–May 1, 2008, pp. 181–186.
19. O. Musseau et al., “Analysis of Multiple Bit Upsets (MBU) in CMOS SRAM,” *IEEE Transactions on Nuclear Science,* Volume 43, Issue 6, Part 1, pp. 2879–2888, Dec. 1996.
20. JEDEC standard No. JESD 89, “Measurement and Reporting of Alpha Particles and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices,” August 2001.
21. European Space Agency (ESA), “Single Event Effects Test Method and Guidelines,” ESA/SCC Basic Specification No. 22900, 1995.
22. G. Gasiot, D. Giot, and P. Roche, “Alpha-Induced Multiple Cell Upsets in Standard and Radiation Hardened SRAMs Manufactured in 65nm CMOS Technology,” *IEEE Transactions on Nuclear Science,* Volume 53, Issue 6, pp. 3479–3486, Dec. 2006.
23. E.H. Cannon, M.S. Gordon, D.F. Heidel, A.J. KleinOsowski, P. Oldiges, K.P. Rodbell, et al., “Multi-Bit Upsets in 65nm SOI SRAMs,” presented at the IRPS Conference, Phoenix, AZ, May 2008.
24. A. Virtanen, R. Harboe-Sorensen, H. Koivisto, S. Pirojenko, and K. Rantilla, “High Penetration Heavy Ions at the RADEF Test Site,” presented at the RADECS Conference, 2003.
25. H. Puchner, R. Kapre, S. Sharifzadeh, J. Majjiga, R. Chao, D. Radaelli, et al., “Elimination of Single Event Latchup in 90nm SRAM Technologies,” *IEEE International Reliability Physics Symposium Proceedings,* pp. 721–722, March 2006.
26. P. Roche and R. Harboe-Sorensen, “Radiation Evaluation of ST Test Structures in Commercial 130nm CMOS Bulk and SOI, in Commercial 90nm CMOS Bulk and in Commercial 65nm CMOS Bulk and SOI,” European Space Agency QCA Workshop, January 2007.
27. T. Kishimoto et al., “Suppression of Ion-Induced Charge Collection against Soft-Error,” in *Proc. 11th Int. Conf. Ion Implantation Technology,* Austin, TX, Jun. 16–21, 1996, pp. 9–12.

28. D. Burnett et al., "Soft-Error-Rate Improvement in Advanced BiCMOS SRAMs," in *Proc. 31st Annual Int. Reliability Physics Symp.,* Atlanta, GA, Mar. 23–25, 1993, pp. 156–160.
29. P. Roche and G. Gasiot, "Impacts of Front-End and Middle-End Process Modifications on Terrestrial Soft Error Rate," *IEEE Transactions on Device and Materials Reliability,* Volume 5, Issue 3, pp. 382–396, Sept. 2005.
30. H. Puchner et al., "Alpha-Particle SEU Performance of SRAM with Triple Well," *IEEE Trans. Nucl. Sci.,* Volume 51, Issue 6, pp. 3525–3528, Dec. 2004.
31. D. Radaelli et al., "Investigation of Multi-Bit Upsets in a 150 nm Technology SRAM Device," *IEEE Transactions on Nuclear Science,* Volume 52, Issue 6, pp. 2433–2437, Dec. 2005.
32. D. Giot, P. Roche, G. Gasiot, J.-L. Autran, and R. Harboe-Sørensen, "Heavy Ion Testing and 3D Simulations of Multiple Cell Upset in 65nm Standard SRAMs," *IEEE Transactions on Nuclear Science,* 2007.
33. A.M. Chugg, "A Statistical Technique to Measure the Proportion of MBU's in SEE Testing," *IEEE Transactions on Nuclear Science,* Volume 53, Issue 6, pp. 3139–3144, Dec. 2006.
34. D. Tryen, J. Boch, B. Sagnes, N. Renaud, E. Leduc, S. Arnal, et al., "Temperature Effect on Heavy-Ion Induced Parasitic Current on SRAM by Device Simulation: Effect on SEU Sensitivity," *IEEE Transactions on Nuclear Science,* Volume 54, Issue 4, pp. 1025–1029, 2007.
35. M. Bagatin, S. Gerardin, A. Pacagnella, C. Andreani, G. Gorini, A. Pietropaolo, et al., "Factors Impacting the Temperature Dependence of Soft Errors in Commercial SRAMs," *IEEE Transactions on Nuclear Science,* Volume 55, Issue 6, pp. 3302–3308, 2008.
36. G. Gasiot, P. Roche, and P. Flatresse, "Comparison of Multiple Cell Upset Response of BULK and SOI 130nm Technologies in the Terrestrial Environment," presented at the IRPS Conference, Phoenix, AZ, May 2008.
37. Y. Kawakami et al., "Investigation of Soft Error Rate Including Multi-Bit Upsets in Advanced SRAM Using Neutron Irradiation Test and 3D Mixed-Mode Device Simulation," *International Electron Devices Meeting,* 2004.
38. Ph. Roche et al., "SEU Response of an Entire SRAM Cell Simulated as One Contiguous Three Dimensional Device Domain," *IEEE Transactions on Nuclear Science,* Volume 45, Issue 6, pp. 2534–2543, Dec. 1998.
39. Synopsys Sentaurus TCAD tools, Available at: http://www.synopsys.com/products/tcad/tcad.html
40. K. Osada et al., "Cosmic-Ray Multi-Error Immunity for SRAM, Based on Analysis of the Parasitic Bipolar Effect," *2003 Symposium on VLSI Circuit Digest of Technical Papers.*
41. D. Giot, G. Gasiot, and P. Roche, "Multiple Bit Upset Analysis in 90nm SRAMs: Heavy Ions Testing and 3D Simulations," presented at the RADECS Conference, Athens, Greece, Sept. 2006.

9 Real-Time Soft Error Rate Characterization of Advanced SRAMs

Jean-Luc Autran, Gilles Gasiot, Daniela Munteanu, Philippe Roche, and Sébastien Sauze

CONTENTS

9.1 INTRODUCTION

Since cosmic rays and on-chip radioactive impurities have been identified to be at the origin of soft errors in modern integrated circuits, the estimation of the soft error rate (SER) is rapidly becoming a major consideration for reliability aspects at device, circuit, and system levels—not only to investigate and understand technology sensitivity but also to extrapolate the trends for future generations of circuits [1]. Different experimental and simulation approaches are known to estimate SER: accelerated testing using alpha, neutron, or proton source/beam, real-time (i.e., life) testing under natural environments, modeling and software simulation at device or circuit level, combination of experimental/simulation approaches [1-6]. In contrast with accelerated testing, which is relatively easy to conduct, cheaper, and fast (a few hours per day is generally sufficient to obtain confident results), real-time testing is clearly time-consuming and expensive. But it appears to be the unique experimental solution to accurately estimate SER, ensuring that the test does not introduce artificial results due, for example, to beam uniformity/fluctuations, dosimetry errors, chip disorientation or difference in spectrum (largely introduced by the cut-off energy of the accelerator that is always well below cosmic ray energies). Real-time testing can also address SER at the system level for complex electronic solutions and, installed in an underground site, can provide an efficient method of monitoring for radioactive contamination. On the contrary, when based at an altitude to increase the flux of atmospheric particles (primarily neutrons, but also pions and protons), SER by life testing can be accelerated by a factor of ~2–15 depending on the geographical coordinates and altitude of the test site.

With the downscaling of complementary metal-oxide semiconductor (CMOS) technologies, natural radiation is inducing one of the highest failure rates of all reliability concerns for devices and circuits in the area of nano-electronics [7-8]. This sensitivity is a direct consequence of the reduction of device dimensions and spacing within memory cells combined with the reduction of supply voltage and node capacitance, resulting in a decrease of both the critical charge (i.e., the minimum amount of charge required to induce the flipping of the logic state) and the sensitive area (i.e., the minimum collection area inside which a given particle can deposit enough charge to induce the flipping of the cell) [5-8]. Because the response and sensitivity of a given technology to cosmic rays or (internal) residual radioactivity have not necessarily the same magnitude (depending on several design and process key parameters, such as the three-dimensional [3-D] cell architecture, the circuit layout, and the internal contamination level of chip materials and package), their impact on the SER must be separately evaluated in terms of fail occurrence (with distinguishing single upsets from multiple cell upsets) and failure-in-time (FIT) for both neutrons and alpha particles [8-13].

In this context and since 2005, we have developed a research program on the impact of radiation effects at ground level on components, circuits, and systems-on-chips. An initial objective of this work was to install permanent test facilities, both in altitude and underground, to perform long-term and real-time SER characterization of CMOS technologies. The altitude location was chosen to strengthen natural neutron irradiation; the cave environment allows the atmospheric neutron contribution to be completely screened and the remaining alpha SER directly induced by the presence of radioactive impurities in the chip materials to be quantified. We first launched in 2005 the Altitude SEE Test European Platform (ASTEP) [14,15] and installed in 2007 permanent test equipment at the Modane Underground Laboratory (LSM) [16,17].

This chapter briefly discusses different aspects of this research program, including the description of the two test platforms and their radiation environment, the real-time setups, and a synthesis of more than one cumulative year of real-time characterization concerning two generations of static random access memory (SRAM) circuits manufactured in 130 and 65 nm CMOS technologies.

9.2 TEST PLATFORMS AND ENVIRONMENTS

9.2.1 The ASTEP Platform

ASTEP is a dual academic research and research and development (R&D) platform (permanent facility) founded by STMicroelectronics, JB R&D, and L2MP-CNRS in 2004 [14]. The current platform, operated by IM2NP-CNRS (formerly L2MP), is dedicated to real-time SER testing of semiconductor circuits and systems. Located in the French Alps on the deserted Plateau de Bure at 2,552 m in a low electromagnetic noise environment, the platform is hosted by the Institute for Radio-astronomy at Millimeter Wavelengths (IRAM). ASTEP has been fully operational since March 2006. The main environment characteristics of the ASTEP platform are summarized in Table 9.1. Since 2006, this test location has been referenced in the latest release of JEDEC Standard JESD89A [18]. The data from Table 9.1 correspond to the values in Table A3.B in [18].

Figures 9.1a and 9.1b show a general view of the Plateau de Bure (IRAM Observatory) and an external view of the ASTEP building, respectively. The building extension (first floor) was finished in 2008 and has been occupied since July 2008 by the Plateau de Bure Neutron Monitor (PdBNM), a super 3-NM64 neutron monitor composed of three high-pressure (2,280 Torr) cylindrical He3 detectors (model LND 253109) surrounded by coaxial polyethylene and lead rings (thickness of 25 mm

TABLE 9.1
Location Parameters and Main Environment Characteristics Related to the ASTEP Platform

ASTEP, Plateau de Bure, France	
Latitude (°N)	44.6
Longitude (°E)	5.9
Elevation (m)	2,550
Atm. depth (g/cm^2)	757
Cutoff rigidity (GV)	5.0
Relative neutron flux	
Active Sun low	5.76
Quiet Sun peak	6.66
Average	**6.21**

Source: Reprinted with permission from Autran, J. L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D., Loiaza, P., Zampaolo, M., Borel, J.," Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science,* Volume: 56 Issue: 4, 2258–2266.)

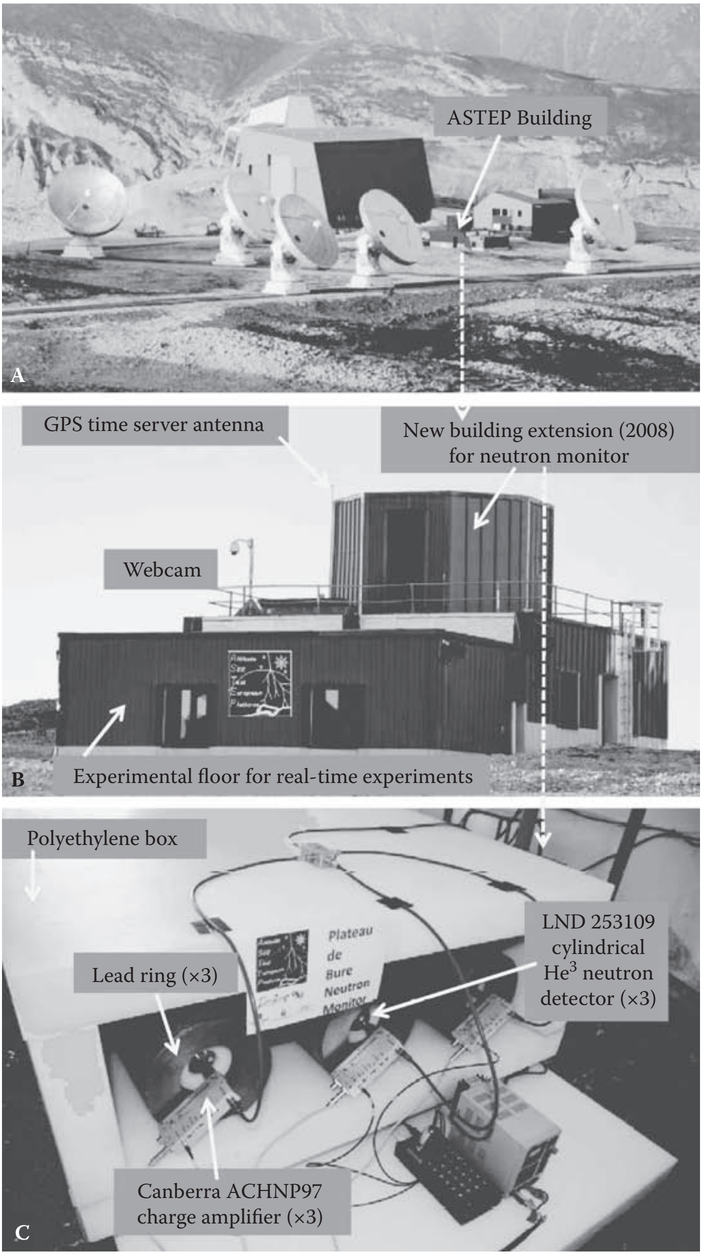


FIGURE 9.1 (a) General view of the Plateau de Bure. (b) External view of the ASTEP building showing the new extension (first floor) designed to host the Plateau de Bure Neutron Monitor (PdBNM). (c) Partial view of the PdBNM showing the extremities of the cylindrical neutron detector tubes connected to the charge amplifiers and to the acquisition module (electronic counters).

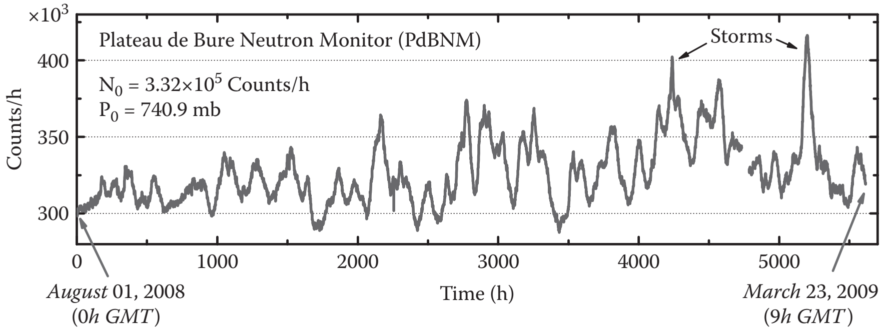


FIGURE 9.2 Plateau de Bure Neutron Monitor response recorded from August 1, 2008, to March 23, 2009. Data are uncorrected from atmospheric pressure and averaged over one hour. ~30% variations in neutron flux are evidenced, with two peaks (> 4 × 10^5 counts/h) corresponding to the passage of two severe atmospheric depressions (the first peak corresponds to the Klaus storm on January 25, 2009). (Reprinted with permission from Autran, J.L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D., Loaiza, P., Zampaolo, M., Borel, J., Rozov, S., Yakushev, E., "Combined altitude and underground real-time SER characterization of CMOS technologies on the ASTEP-LSM platform." *IC Design and Technology*, 2009. IEEE International Conference on ICICDT '09. 18–20 May, 2009, pp. 113– 120.)

each) inside a polyethylene box (wall thickness equal to 80 mm). A Canberra electronic detection chain (composed of three charge amplifiers model ACHNP97 and a high-voltage source 3200D) was chosen in complement to a Keithley KUSB3116 acquisition module for interfacing the neutron monitor with the control PC. The design and the construction of the PdBNM followed the recommendations published in [19,20] for the optimization of the apparatus response. Figure 9.2 shows the PdBNM averaged response (one point per hour) from August 1, 2008, to March 23, 2009. This uncorrected response from atmospheric pressure directly gives an accurate description of the neutron flux variation at the ASTEP location, evidencing ~30% variations of this averaged flux at ground level essentially due to atmospheric pressure variations.

During its installation, the PdBNM was used to experimentally determine the acceleration factor (AF) of the ASTEP location with respect to sea level, as explained in the following. Assembled and previously operated in Marseille during the years 2007–2008, the PdBNM was transported and installed on the Plateau de Bure in July 2008. With strictly the same setup, two series of data were thus recorded in Marseille and on the Plateau de Bure.

Figure 9.3 shows the barometric response of the PdBNM, that is, the variation of the counting rate as a function of the atmospheric pressure [21]. The difference between the counting rates of the two clouds of experimental points (~700 hourly data, which corresponds to one month of monitoring) directly gives the value of the acceleration factor of ASTEP with respect to the Marseille location, here estimated to 6.7. Taking into account latitude, longitude, and altitude corrections for the Marseille location with respect to the reference one (i.e., New York City), the final value of the acceleration factor is AF = 6.7 × 0.94 ≈ 6.3. This value is close to 6.2, the

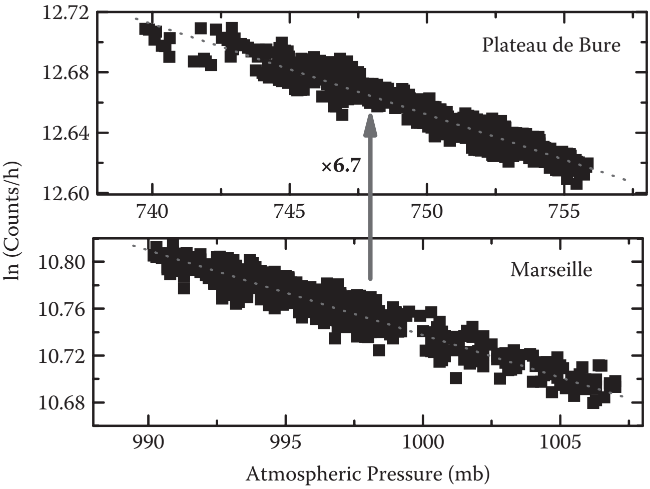


FIGURE 9.3 Experimental determination of the ASTEP acceleration factor (AF) from the barometric response of the neutron monitor successively installed in Marseille (2007–2008) and on the Plateau de Bure since July 2008. Experimental clouds correspond to one month recording (one point per hour). (Reprinted with permission from Autran, J.L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D., Loaiza, P., Zampaolo, M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science*, Volume: 56 Issue: 4, 2258–2266.)

average acceleration factor reported in Annex A of the JEDEC standard JESD89A [18,22], and is close to 5.9, the value given by the Qinetic Radiation Atmospheric Model (QARM) [23,24]. In the following, we will use the experimental value AF = 6.3 as the acceleration factor for the ASTEP location.

9.2.2 The LSM Laboratory

In October 2007, we installed the first real-time SER experimental setup in the underground LSM. This laboratory is located about 1,700 m under the top of the Fréjus Mountain (4,800 meters water equivalent), near the middle of the Fréjus highway tunnel connecting France and Italy [25]. It was created in 1983 to conduct particle physics and astrophysics experiments in a strongly reduced cosmic ray background environment. Due to the depth of the LSM, the average particle flux inside the laboratory is extremely reduced:

- ~4 muons/m^2/day corresponding to a two million reduction factor compared with the flux at sea level.
- A few 10^3 fast neutrons/m^2/day (depending on the neutron energy and the measurement location in the laboratory) emitted by natural radioactivity from the rock (see spectrum in Figure 9.4), the neutron component of cosmic rays being totally eliminated at this depth.

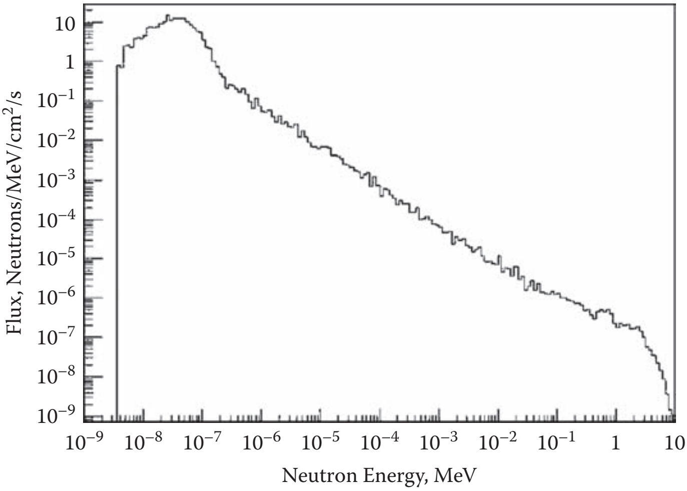


FIGURE 9.4 Neutron energy spectrum from rock activity at LSM simulated with GEANT4. (Reprinted with permission from Baumann, R., Hossain, T., Murata, S., Kitagawa, H., "Boron compounds as a dominant source of alpha particles in semiconductor devices." Reliablility Physics Symposium, 1995. 33rd Annual Proceedings, Volume 560, Issue 2, 10 May 2006, Pages 454–459. *IEEE Publicatio*n Date: 4-6 April 1995 on pages 297–302.)

In addition, the radon in the laboratory is maintained at a very low rate of ~20 Bq/m^3 owing to an air purification system that totally renews the volume of the air inside the laboratory twice an hour. Recent fast and thermal neutron measurements have been performed by E. Yakushev and co-workers [26] at immediate proximity to our setups; these results have then been modeled and reproduced within a few percent by calibrated GEANT4 Monte Carlo simulations. They give a flux of fast neutrons with E > 0.3 eV (cadmium threshold) of ~3 × 10^{-6} neutron/cm^2/s. Measurements of thermal neutrons at the same place with a bare He^3 filled proportional counter gave ~2 × 10^{-6} neutron/cm^2/s. Knowing that flux of fast and thermal neutrons are connected, the Monte Carlo predicted coefficient for really fast neutrons with E > 0.5 MeV to thermal neutrons (E < 0.3 eV) is about 0.64–0.66 (depends slightly on rock and concrete). Thus, we can estimate the number of such neutrons (E > 0.5 MeV) at places of SER experiments from this as ~1.2 × 10^{-6} neutron/cm^2/s. These measurements thus confirm the residual background value of only a few 10^3 fast neutrons/m^2/day inside the experimental room deduced from experimental measurements and resulting from simulation work [27].

9.3 EXPERIMENTAL DETAILS

9.3.1 Tested Devices

Real-time measurements have been performed on bulk SRAMs fabricated by STMicroelectronics using commercial CMOS processes in 130 nm (200 mm wafers) and 65 nm (300 mm wafers) technologies. These processes are based on a boro-

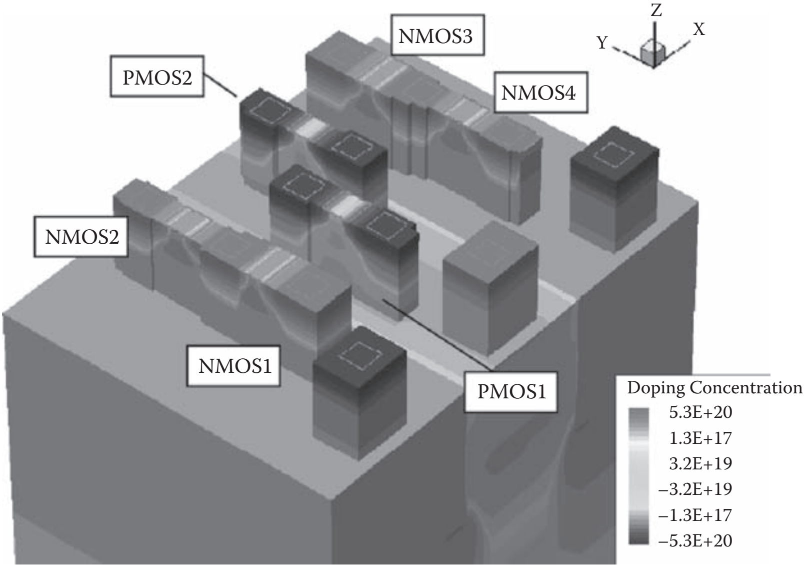


FIGURE 9.5 Full 3-D structure of the bulk 65 nm SP SRAM (six transistors, bit cell area of 0.525 μm^2). Shallow trench isolation (STI) structures have been removed to better show silicon regions below the transistor active area. The TCAD simulation domain contains more than 300,000 mesh elements. (Reprinted with permission from Autran, J.L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D., Loaiza, P., Zampaolo, M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science*, Volume: 56 Issue: 4, 2258–2266.)

phospho-silicate glass (BPSG)-free back-end-of-line (BEOL), which eliminates the major source of ^{10}B in the circuits and drastically reduces the possible interaction between ^{10}B and low energy neutrons (in the thermal range and below) [1,28,29]. The test vehicle for the 130 nm technology is composed of 4 MBit single-port SRAM (SP SRAM) with a bit cell area of 2.50 μm^2. A total of 3,664 MBit was considered for real-time experiments, both in altitude (ASTEP) and underground (LSM). For the 65 nm technology, the test chip contains 8.5 Mbit of SP SRAM (bit cell area of 0.525 μm^2) and 1 Mbit of dual-port SRAM (DP SRAM, bit cell area of 0.98 μm^2).

The SP SRAM bit cell, shown in Figure 9.5 for the 65 nm technology node, corresponds to the standard six-transistor SRAM designed with one access transistor on each internal node. DP SRAM has the same electrical schematic with two additional access transistors, one on each side of the memory, giving the ability to simultaneously read and write different memory cells at different addresses. No deep n-well [11] was used in both 130 and 65 nm devices tested in the present work. Both 130 nm and 65 nm bit cells were fully modeled (Figure 9.5) with 3-D technology computer-aided design (TCAD) tools (Sentaurus Synopsys package [30]) to evaluate their sensitivity to heavy ions and to determine the single-event unit (SEU)/single-bit upset (SBU) and multiple-bit upset (MBU)/multiple-cell upset (MCU) occurrences as a function of ion parameter [11,13]. In complement to TCAD work, numerous experimental studies were conducted these four last years to characterize the different test chips from an accelerated-test point of view with neutrons at the Los Alamos Neutron Science Center (LANSCE), as well as with an intense Am^{241} alpha source at STMicroelectronics. This point is detailed in [14,17].

9.3.2 Hardware Setups

Two different types of SER test equipment, specially designed for the study, were developed and assembled by Bertin Technologies (Aix-en-Provence, France) for the 130 nm devices [14-16] and by iRoC Technologies (Grenoble, France) for the 65 nm ones [17], respectively. Figure 9.6 shows a general view of the test equipment currently deployed at LSM. The 130 nm setup was successively installed on ASTEP during the

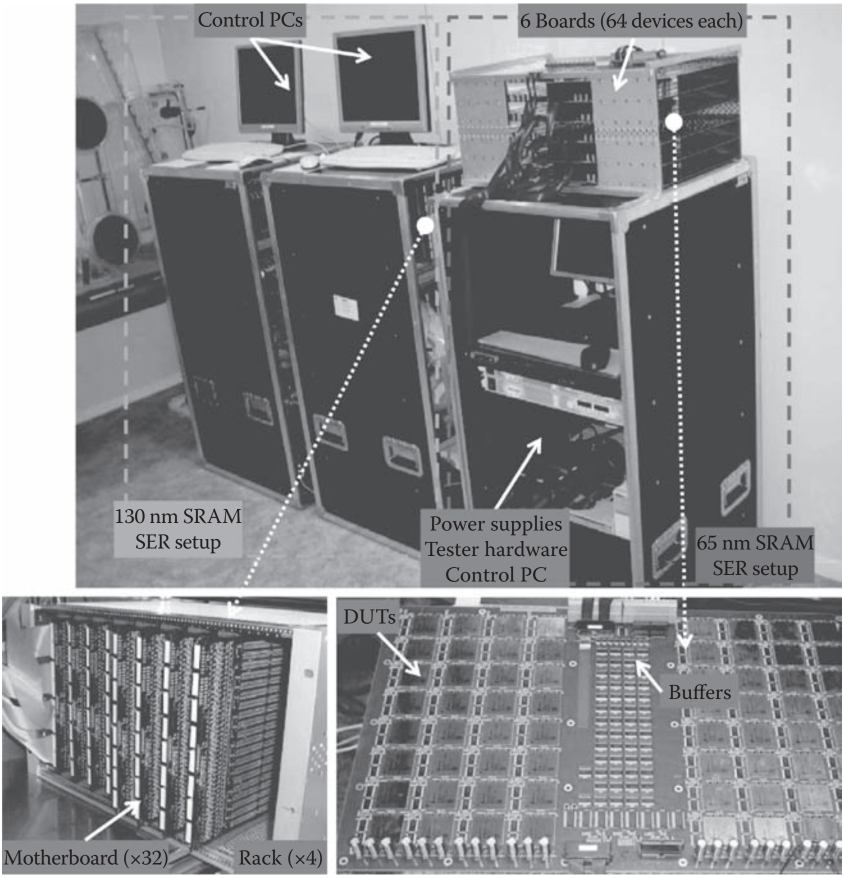


FIGURE 9.6 General view of the automatic test equipment installed in the microelectronics experimental room at LSM. The two insets show detailed views at the motherboard level for the two setups. (Reprinted with permission from Autran, J.L., Roche, P., Sauze, S.,Gasiot, G., Munteanu, D., Loaiza, P., Zampaolo, M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science,* Volume: 56 Issue: 4, 2258–2266.)

period March 31, 2006 to November 6, 2006, and then was transported to the LSM and installed on October 16, 2007. For the 65 nm experiment, two identical setups were constructed and were installed at ASTEP on January 21, 2008 and at LSM on April 11, 2008. The three setups have been working since these respective dates.

In the present configuration, each system is capable of monitoring several hundred of chips (1,280 chips for the 130 nm setup and 384 chips for the 65 nm one) and performing all requested operations such as writing/reading data to the chips, comparing the output data to the written data, and recording details on the different detected errors in SRAM chips. The different hardware and software components have been designed to strictly follow all the specifications of the JEDEC Standard JESD89A [18]. In particular, the design of the setup ensures that all detected errors come from the devices under test, not from external or system noise, by respecting the following guidelines [18,31]:

1. The test operation is such that once failing data is detected, the data is read again a given number of times before data is rewritten. Consistency of failed data over these read cycles ensures that the failure is a soft error in the device under test (DUT).
2. The tester implements high reliability system techniques: redundancy in the logic interface with the DUT and watchdog for periodic reinitialization of the tester.
3. The power supplies are designed for uninterruptible operation and very low noise. The power supply voltages are permanently monitored. The voltage and current drawn by the DUT during the standby mode are logged periodically.
4. The DUT boards are properly designed for very low internal noise. The boards are multilayered with alternated signal and ground planes for high immunity to electromagnetic interference (EMI). They are also designed with controlled impedance for maintaining signal integrity with a relatively high number of circuits in a bus.
5. Finally, the tester and array of DUT boards are properly shielded against EMI.

9.3.3 Test Procedure

The test algorithm used for SRAM testing is schematically shown in Figure 9.7. It allows detection of SBU, MCU, single-event functional interrupt (SEFI), or single-event latchup (SEL) events. Current consumption of all power lines provided by the tester is monitored and logged during the test. The user can see in real time the errors on the monitor of the tester. This test algorithm has dead time, but with the considered conditions of real-time SER (very low error rate) it is negligible.

9.4 EXPERIMENTAL RESULTS

This section summarizes our most recent results obtained for the 65 nm technology from real-time (i.e., life-testing) and accelerated tests (for both neutrons and alphas).

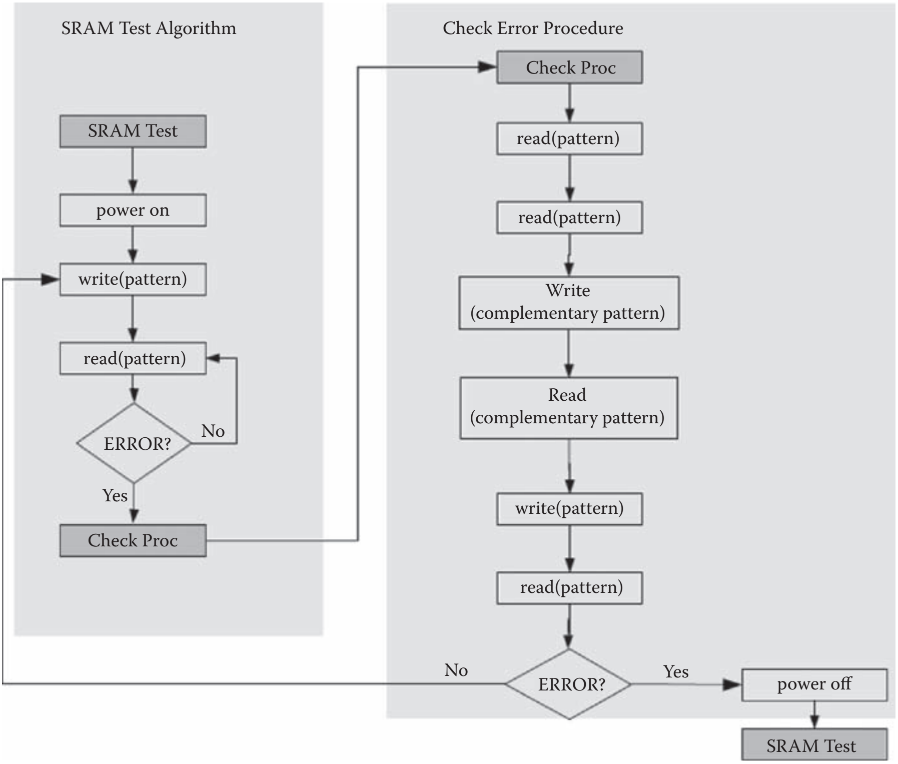


FIGURE 9.7 Flowchart of the test procedure implemented in the automatic test equipment for real-time SER testing both 130 and 65 nm SRAM technologies. (Reprinted with permission from Autran, J. L., Roche, P. Sauze, S. Gassiot, G., Munteanu, D., Loaiza, P., Zampaolo, M. Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science,* Volume: 56 Issue: 4, 2258-2266.)

Comparison with real-time data obtained for the 130 nm technology (SP SRAM) is also reported. In the following, all numerical results have been normalized by a common arbitrary scaling factor, set lower than 3×. The real order of magnitude for the reported data is thus not significantly altered.

9.4.1 Real-Time Measurements

Figure 9.8 shows the cumulative number of fails detected in SP SRAMs versus test duration (expressed in MBit × h) for both altitude and cave experiments. Because the two experiments started at different dates, the number of MBit × h accumulated in altitude (2.42×10^7 MBit × h) is higher than the value reached in the cave experiment (1.78×10^7 MBit × h).

A data analysis summarized in Table 9.2 shows that for the altitude experiment a total of 44 events (involving a total of 71 bitflips) was detected for SP SRAM, including 33 SBU and 11 MCU. These MCUs involve a total of 38 bitflips, which are

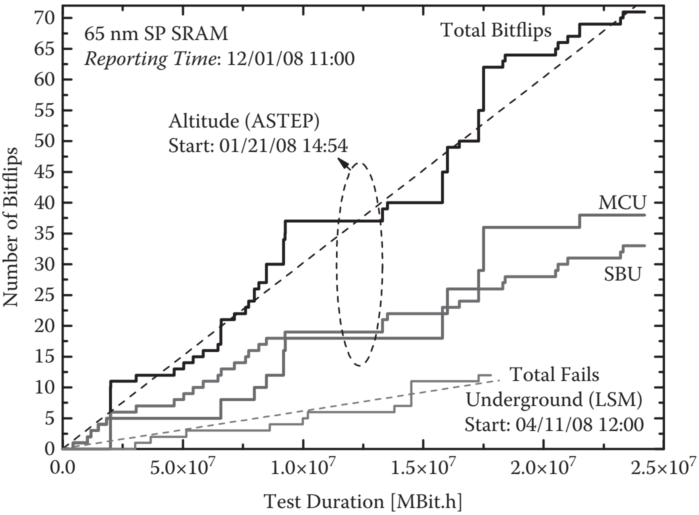


FIGURE 9.8 Cumulative total fails, SBU and MCU flips versus test duration for the 65 nm SP SRAM during both altitude and underground experiments. Tests were conducted under nominal conditions: V_{DD} = 1.2 V, room temperature, standard checkerboard test pattern. (Reprinted with permission from Autran, J. L., Roche, P. Gasiot, G., Munteanu, D. Loaiza, P., Zampaolo, M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science*, Volume: 56 Issue: 4, 2258–2266.)

physical adjacent bit cells in all cases with a multiplicity ranging from 2 to 7. The distribution of these MCUs is given in Figure 9.9. Note that this MCU contribution represents 11/44 = 25% of the detected events and 38/71 = 53.5% of the total number of detected bitflips, confirming via a real-time experiment the importance of MCU mechanisms in such a deep submicron technology. For the cave experiment, only 12 fails were recorded, corresponding to 10 SBU and 1 MCU (involving two adjacent cells). The fraction of MCU is reduced in this case to 16.7% of the total number of detected bitflips. Experimental data consistency (i.e., compliant with a "random process") was checked for the altitude experiment (the number of bitflips is statistically representative) in terms of statistical distribution of 0→1 and 1→0 bitflips and error bitmap: the frequency of bitflips is found close to 50% for each transition, which is randomly distributed in the memory plan.

From the data in Figure 9.8 for both experiments, we estimated the real-time SER at the test location, reported in Table 9.2, using the following expression:

$$\mathrm{SER} = \frac{N_r}{\Sigma_r} \times 10^9 \ \ (\mathrm{FIT/MBit}) \tag{9.1}$$

where N_r is the number of bitflips (for flip SER), SBU (for SBU SER), or MCU events (for MCU SER) observed at time T_r, and Σ_r is the number of MBit × h cumulated at time T_r.

TABLE 9.2
Summary and Key Values for the Real-Time 65 nm Experiment

	SP SRAM	DP SRAM
Altitude Experiment		
Starting date	01/21/08 14:54	
Reporting date	12/01/08 11:00	
Cumulative number of Mbit.h	2.42×10^7	2.84×10^6
Total number of events/bitflips	44/71	8/10
Number of SBU	33	6
Number of MCU/MCU flips	11/38	2/2
SBU SER on ASTEP (FIT/Mbit)	1,364	2,113
MCU SER on ASTEP (FIT/Mbit)	455	704
Total flip SER on ASTEP (FIT/MBit)	**2,934**	**3,521**
Lower and upper confidence limits for the total flip SER (FIT/Mbit)	*2,423* *3,574*	*2,172* *5,973*
Underground Experiment		
Starting date	04/11/08 12:00	
Reporting date	12/01/08 11:00	
Cumulated number of Mbit.h	1.78×10^7	2.10×10^6
Total number of events/bitflips	11/12	2/2
Number of SBU	10	2
Number of MCU/MCU flips	1/2	0/0
SBU SER (FIT/Mbit)	562	952
MCU SER (FIT/Mbit)	56	0
Total flip SER (FIT/Mbit)	**674**	**952**
Lower and upper confidence limits for the total flip SER (FIT/Mbit)	*432* *1,092*	*389* *2,998*

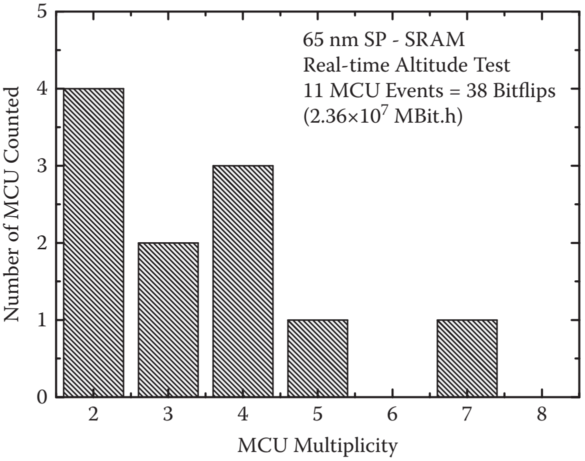


FIGURE 9.9 Distribution of the MCU multiplicity (i.e., number of bitflips per MCU event) for the 11 MCU events detected during the altitude test (SP SRAM). These MCUs involve a total of 38 bitflips that correspond to physical adjacent bit cells (in the memory plan) in all cases. (Reprinted with permission from Autran, J. L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D. iza, P., Zampaolo, M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science*, Volume: 56 Issue: 4, 2258–2266.)

We also reported in Table 9.2 the upper and lower confidence intervals at the 90% level based on the χ^2 distribution to estimate the experimental error margins [18]. We verified that the convergence of SER versus test hours is asymptotically reached within ~3,000 h of experiment (which corresponds to ~10^7 Mbit × h). Beyond this duration, the total flip SER remains constant around 2,934 FIT/Mbit for the altitude experiment and around 674 FIT/Mbit for the underground test.

The calculation of the normalized neutron real-time SER at the reference location of New York City (NYC) is obtained from the following expression, assuming that the fail rate due to alpha particles is identical to the alpha SER experimentally deduced from the underground experiment:

$$\begin{cases} neutron\text{-}SER\big|_{NYC} = \dfrac{SER\big|_{ASTEP} - SER\big|_{LSM}}{AF} \\ alpha\text{-}SER\big|_{NYC} = SER\big|_{LSM} \end{cases} \tag{9.2}$$

In Equation (9.2), the value of the acceleration factor of the ASTEP site is taken equal to AF = 6.3, the experimental value determined from data collected using the Plateau de Bure neutron monitor. The normalized neutron SER is then equal to (2934 – 674)/6.3 = 359 FIT/MBit and the total flip SER for both alpha and neutron contributions is equal to 359 + 674 = 1,033 FIT/MBit for the 65 nm SP SRAM.

For DP SRAM and because the test circuit contains only 1 MBit per chip (against 8.5 Mbit for SP SRAM), the statistics are not yet totally satisfactory, as illustrated by the very large confidence interval reported in Table 9.1. Therefore, a first estimation gives for the altitude test a value of ~3,521 FIT/Mbit and ~952 FIT/Mbit for the cave experiments, resulting in normalized (NYC) neutron SER = 407 FIT/MBit. SER values for DP SRAM will be consolidated in a future work. In the following, we will consider only results related to SP SRAM for comparison with 130 nm and discussion.

9.4.2 Accelerated Tests

Table 9.3 reports the results of neutron and alpha accelerated test results performed on 65 nm chips. The test procedures were fully compliant with the JEDEC test standard JESD89A [18]. Alpha SER was evaluated from accelerated measurements using an intense Am^{241} alpha source. The tests were performed in a characterization lab at STMicroelectronics. The alpha source is a thin foil of Am^{241} with an active diameter of 1.1 cm. The source activity was 3.7 MBq, as measured on February 1, 2002. The alpha particle flux was precisely measured in March 2003 with an Si detector that was placed at 1 mm from the source surface. This calibration measurement gave an alpha flux of 1.05 × 10^6 alpha/cm²/s. Since the atomic half-life of Am^{241} is 432 years, the activity and flux values recalculated for each experimental session are still very accurate. The reported SER values have been extrapolated to a nominal alpha flux of 0.001 alpha/cm²/h. This value emulates the alpha emissivity rate for the semiconductor processing and packaging materials with an "ultra low alpha" grade. During the SER experiments, the Americium source lies above the chip package. Therefore, the distance source-die is minimum and approximately equal to 1 mm (same distance as for the calibration). All experiments were performed at room temperature (25°C). Experimental measurements (Table 9.3) led to an accelerated alpha SER = 605 FIT/Mbit for SP SRAM and 790 FIT/Mbit for DP SRAM.

Accelerated neutron SER evaluation was conducted at the LANSCE WNR facility at Los Alamos in August 2006. The neutron flux available at LANSCE during the experiment was 1.6 × 10^5 n/cm^2/sec, which is 40% of the LANSCE maximum flux (this is a limitation at the LANSCE facility for the whole year 2006). The beam

TABLE 9.3
Summary of Flip SER Value

		SP SRAM	DP SRAM
Alpha SER (FIT/Mbit)	Total flip SER	605	798
Neutron SER (FIT/MBit)	SBU SER	353	461
	MCU (event) SER	38	21
	Total flip SER	470	535

Note: Normalized, that is, extrapolated to 0.001 alpha/cm^2/h and 13 neutrons/cm²/h. Obtained from accelerated tests (V_{DD} = 1.2 V, room temperature, standard checkerboard test pattern).

is collimated and uniform over an ~8 cm diameter. The neutron fluence was measured by the LANSCE uranium fission chamber. The total number of produced neutrons is obtained by counting fissions and applying a proportionality coefficient. The reported SER values, shown in Table 9.3, have been extrapolated to the reference (NYC) neutron integrated flux of 13 neutrons/cm²/h: we obtained 470 FIT/Mbit and 535 FIT/Mbit for SP SRAM and DP-SRAM, respectively. The MCU contribution represents ~31% of the total detected events in these accelerated tests.

9.5 DATA ANALYSIS AND DISCUSSION

In this last section, we analyze and discuss real-time data reported in Section 9.4. We also report comparison experimental data related to the 130 nm STMicroelectronics technology previously characterized [14,15]. Additional results deduced from accelerated SER tests and from wafer-level measurements are presented.

9.5.1 Real-Time versus Accelerated Tests for 65 nm

The direct comparison of real-time and accelerated SER values, reported in Section 9.4.2 for 65 nm SP SRAMs, shows a very reasonable agreement between the two sets of data, especially for alpha SER. We measured 674 FIT/MBit (real-time) and 605 (accelerated) FIT/MBit for alpha SER, resulting in a difference of only 11%, typically within the experimental error margins; we also measured 359 FIT/Mbit (real-time) and 470 FIT/Mbit (accelerated) for neutron SER, showing in this case an agreement within 30% margins. This very good agreement observed for the alpha SER tests first suggests that the accelerated alpha emission setup was properly designed and the test operation accurately conducted. In addition, the real-time SER value suggests that the alpha emission rate for the semiconductor processing and packaging materials is very close to the value of 0.001 alpha/cm^2/h initially assumed to calculate the accelerated SER value reported in Table 9.3.

For neutron SER, a discrepancy of 30% between the two approaches remains very acceptable with respect to dosimetry errors or statistical dispersions from sample-to-sample, lot-to-lot and error intervals on the knowledge of some physical, technological, and electrical key parameters (manufacturing variability) [1,32]. Moreover, this result could be explained by possible differences between the neutron beam and the real atmospheric neutron spectra, largely introduced by the cut-off energy of the accelerator that is always well below cosmic ray energies. This could also be confirmed by the relatively important difference in the percentages of bitflips involved in MCU events for the two experiments: 53.5% for real-time and 31% for accelerated tests.

Our recent results concerning heavy-ion testing and 3-D simulations of MCU occurrence in 65nm SRAMs [13] are consistent with this observation: the contribution of MCU to the total number of upsets strongly increases with the linear energy transfer (LET) of the incident ion, suggesting that high-energy neutrons (indirectly inducing a nonnegligible fraction of high LET ions) play a major role in the occurrence of large size MCUs effectively observed in real-time experiments. This MCU aspect will be consolidated in the future by increasing the experiment duration to significantly improve the statistics on MCUs (Figure 9.9).

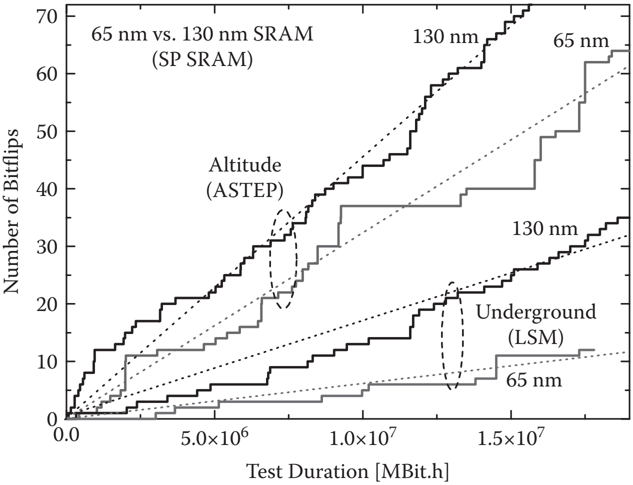


FIGURE 9.10 Cumulative total fails versus test duration for both 130 nm and 65 nm SP SRAMs detected in altitude and underground. The test was conducted under nominal conditions for both technologies: $V_{DD} = 1.2$ V, room temperature, standard checkerboard test pattern. For 130 nm data, experiment periods are March 31, 2006 to November 6, 2006 for the altitude test and October 16, 2007 to November 24, 2008 for the underground test. (Reprinted with permission from Autran, J. L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D., Loiaza, P., Zampaolo, M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science*, Volume: 56 Issue: 4: 2258–2266.)

9.5.2 65 nm versus 130 nm Technologies

Figure 9.10 shows a direct comparison of the total bitflip distributions versus test duration for the two technologies. For the 130 nm, a total of 72 bitflips was detected after 1.55 × 10^7 MBit × h in altitude, 35 fails after 1.9 × 10^7 MBit × h during the underground test. The analysis of Figure 9.10 indicates that, for both test locations, the 130 nm technology exhibits a higher soft error rate (directly linked to the slope of the curves) than the 65 nm one.

In addition, for the altitude test, five MCU events, each involving two physical adjacent bit cells, were recorded; no MCU event was detected for the cave experiment. This difference in MCU occurrence for the two technologies is clearly highlighted by the "staircase shape" of the curves: the 130 nm distribution has very regular stairs (each stair corresponding to a single bitflip); on the other hand, the 65 nm curve (especially for the altitude test) exhibits irregular and marked stairs, which corresponds to a kind of "visual signature" of MCU events.

These SER values related to the 130 nm technology are reported in Figure 9.11. We used Equation (9.2) to separate alpha from neutron contributions to the total normalized SER value. Figure 9.11 also summarizes the key values of experimental real-time SER for both 130 and 65 nm technologies (SP SRAM). Alpha SER is found to decrease by a factor of 2.3 for the 65 nm technology with respect to the 130 nm one and neutron SER by a factor of 1.4, resulting in a net improvement of the total SER by a factor of ~2.

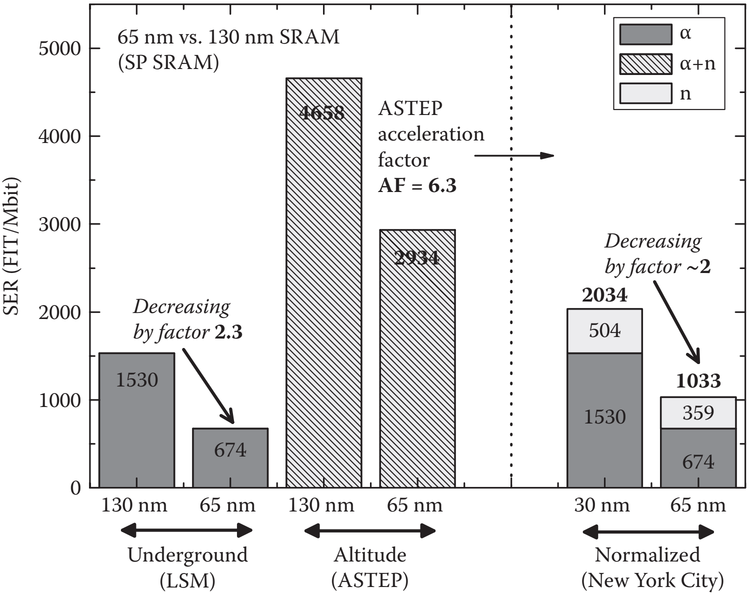


FIGURE 9.11 Synthesis of experimental real-time SER values obtained for both 130 nm and 65 nm SP SRAM from altitude and underground experiments and normalization of the SER at the reference flux of New York City (sea level) taking into account (i) the alpha contribution for the altitude test (fixed to the value measured at LSM) and (ii) the ASTEP acceleration factor AF = 6.3 for the neutron flux in altitude (experimentally measured using the Plateau de Bure Neutron Monitor). (Reprinted with permission from Autran, J. L., Roche, P., Sauze, S., Gasiot, G., Munteanu, D., Loaiza, P., Zampaolo,M., Borel, J., "Altitude and underground real-time SER characterization of CMOS 65 nm SRAM." *IEEE Transactions on Nuclear Science*, Volume: 56 Issue: 4, 2258–2266.)

9.5.3 Estimation of the Alpha Particle Emission Rates for 65 nm and 130 nm Technologies

Combining real-time and accelerated alpha SER values, for a given technology, allows us to estimate the alpha particle emission rate for the semiconductor processing and packaging materials. Because accelerated values are extrapolated (i.e., normalized) to the reference value of 0.001 alpha/cm^2/h (which corresponds to an "ultra low alpha" grade [5]), the real alpha particle emission rate is simply given by this value multiplied by a factor corresponding to the ratio of the real-time SER by the accelerated SER. We thus obtain for the 65 nm technology 0.001 × (674/605) ≈ 1.1 × 10^{-3} alpha/cm^2/h. For the 130 nm technology, a similar calculation from real-time SER values given in Figure 9.11 and considering the accelerated value of 380 FIT/Mbit reported in [16] gives 0.001 × (380/1530) ≈ 4.0 × 10^{-3} alpha/cm^2/h.

In addition to this indirect extraction of the alpha particle emissivity via SER tests, the alpha emission rates for both the tested wafers and packages (mold compound) were accurately characterized using an ultra low alpha background counter (gas flow type). The tests were performed in a dedicated characterization lab at

STMicroelectronics. A high purity in terms of radioactive contaminants was confirmed, around (0.9 ± 0.3) × 10^{-3} alpha/cm^2/h for the 65 nm technology. In parallel, the same measurement procedure was applied for the characterization of wafers and packages of the 130 nm technology. A value of (2.3 ± 0.2) × 10^{-3} alpha/cm^2/h was obtained, confirming with the same order of magnitude the reduction of the alpha emitter contamination for the 65 nm technology with respect to the 130 nm one. The discrepancy between SER-based and direct counting measurements is small and very acceptable with respect to the experimental uncertainties for the alpha counting, the SER testing, and the sample-to-sample/lot-to-lot variations for the trace amounts of alpha contaminants.

9.5.4 Synthesis and SER Trends

In this last section, we would like to summarize in Table 9.4 all the values related to the two technologies characterized in the framework of this study. We also indicate in the last column of the table the evolution factor between the 130 nm and 65 nm technologies. On one hand, a simple scaling of the sensitive area of the 130 nm SRAM versus the 65 nm SRAM should produce approximately a ÷4 reduction factor in the FIT/Mbit rates. On the other hand, as the cell size decreases, it holds less charge, causing the critical charge to decrease, typically by a factor of ~3 [33]. This

TABLE 9.4
65 nm versus 130 nm Technologies (Single-Port SRAM): Synthesis of Key Values for Real-Time and Accelerated Tests

		130 nm (SP)	65 nm (SP)	130→65 nm Variation
Bit cell area (μm^2)		2.50	0.525	÷4.76
Sensitive volume (μm^3) [28]		0.025	0.0035	÷7.14
Critical charge (fC) [28]		2.5	0.8	÷3.13
Nominal V_{DD} (V)		1.2	1.2	unchanged
	Experimental Results			
Accelerated SER (FIT/Mbit)	Alphas	380*	605*	×1.6
	Neutrons	665*	470*	÷1.4
	Total	**1045**	**1075**	~unchanged
Real-time SER (FIT/Mbit)	Alphas	1530	674	÷2.27
	Neutrons	504	359	÷1.4
	Total	**2034**	**1033**	**÷ ~2**
Alpha particle emission level measured at wafer level (alphas/cm²/h)		(2.3±0.2) ×10^{-3}	(0.9±0.3) ×10^{-3}	÷2.5
Alpha particle emission level deduced by combining accelerated and real-time data (alphas/cm²/h)		4×10^{-3}	1.1×10^{-3}	÷3.6

* Extrapolated to 0.001 alpha/cm^2/h and 13 neutrons/cm²/h.

makes it easier for the cell to be upset due a neutron-induced or an alpha particle strike. Since we only observe (from real-time data) a global ÷2 reduction factor, one could attribute this reduction to a combination of these two opposite trends, resulting in the sensitive area change with a small increase in the sensitivity of the basic cell [34]. The fact that the reduction factor is experimentally found (life-testing) more important for alpha SER (÷2.27) than for neutrons (÷1.4) could be explained by the additional impact of the alpha particle emission rate (decreasing by a factor of at least > 2.5) on this scaling factor. Because such an impact is relatively complex to quantify on the final SER value characterizing a given technology, the in-depth analysis of the SER evolution from the 130 nm to the 65 nm technologies would require a specific dedicated work (including material-level and technological options, memory cell, and circuits layout considerations [35]) that goes far beyond the present study.

9.6 CONCLUSION

In summary, this chapter presents a complete study dedicated to the real-time SER characterization of CMOS SRAM memories in both altitude and underground environments. Neutron and alpha particle SERs have been compared with data obtained from accelerated tests for two different technology nodes (i.e., 130 nm and 65 nm). By comparing accelerated and real-time measurements, it is noteworthy that for alpha particles both measurements are extremely close. This confirms that the alpha emission rate from the dice (wafer + package) is extremely close to the "ultra low alpha" extrapolated level, corresponding to the value of 0.001 alpha/cm²/h. For neutrons, accelerated and real-time measurements are also in good agreement (~30%) with respect to error margins in such experimental approaches. Data also show and quantify via a real-time experiment in a natural environment the importance of MCU mechanisms in such a deep submicron technology. Direct comparison with real-time measurements in 130 nm technology, tested in the same locations with a similar setup, clearly shows that the reduction of the neutron SER is related to the technologies (same trend as that of accelerated measurements), while for alpha it is related to a more complex evolution, combining the two opposite trends given by the technology (sensitive area, critical charge) with the observed decrease in the alpha emission rate for the semiconductor processing and packaging materials (trend opposite to that of accelerated measurements extrapolated to a given emission rate value).

In parallel with this characterization work, the installation of a neutron monitor on the altitude site (ASTEP) allowed us to precisely determine the acceleration factor related to the neutron flux and to validate previous results obtained for the 130 nm technology. This neutron monitor will be used in future work to follow in parallel the time evolution of the neutron flux and the time distribution of fails observed in microelectronic circuits.

ACKNOWLEDGMENTS

This work was supported in part by the MEDEA+ Project #2A704 ROBIN, the Institut Carnot STAR (Marseille), and the French Ministry of Economy, Finances and Industry under research conventions #052930215 and #062930210. This work was also performed in the framework of a scientific collaboration between the Altitude SEE Test European Platform (ASTEP platform, http://www.astep.eu) and the Laboratoire Souterrain de Modane (LSM, joint CEA-CNRS research unit, http://www-lsm.in2p3.fr). The authors would like to thank D. Gautier and M. Derbey from iRoC Technologies (Grenoble, France) for their important technical contribution to this work and the LSM staff, M. Zampaolo, P. Loaiza, and F. Picquemal for their efficient technical assistance. The logistical support of the Institute for Radioastronomy at Millimeter Wavelengths (IRAM) is also gratefully acknowledged. Special thanks are finally due to J.P. Schoellkopf (STMicroelectronics) and J. Borel (JB R&D) for their strong support and encouragement during all these years.

REFERENCES

1. J.F. Ziegler, H. Puchner, *SER—History, Trends and Challenges,* Cypress Semiconductor, 2004. See also references therein.
2. E. Normand, "Single Event Upset at Ground Level," *IEEE Transactions on Nuclear Science,* Volume NS-43, No. 6, pp. 2742–2750, 1996. See also references therein.
3. T.J. O'Gorman, J.M. Ross, A.H. Taber, J.F. Ziegler, H.P. Muhlfeld, I.C. J. Montrose, et al., "Field Testing for Cosmic Ray Soft Errors in Semiconductor Memories," *IBM Journal of Research and Development,* Volume 40, No. 1, pp. 41-50, 1996.
4. P.E. Dodd, "Device Simulation of Charge Collection and Single-Event Upset," *IEEE Transactions on Nuclear Science,* Volume NS-43, No. 2, pp. 561–574, 1996.
5. R.C. Baumann, "Radiation-Induced Soft Errors in Advanced Semiconductor Technologies," *IEEE Transactions on Device and Material Reliability,* Volume 5, No. 3, pp. 305–316, 2005.
6. P. Roche, "Year-in-Review on Radiation-Induced Soft Error Rate," tutorial at IEEE International Reliability Physics Symposium, San Jose, CA, March 2006.
7. 2009 International Technology Roadmap for Semiconductors, Available at: http://public.itrs.net/
8. S. Mitra, P. Sanda, and N. Seifert, "Soft Errors: Technology Trends, System Effects and Protection Techniques," IEEE VLSI Test Symposium, 2008.
9. N. Seifert, B. Gill, K. Foley, and P. Relangi, "Multi-cell Upset Probabilities of 45nm High-k + Metal Gate SRAM Devices in Terrestrial and Space Environments," *IEEE International Reliability Physics Symposium* (IRPS), 2008, pp. 181–186.
10. G. Gasiot et al., "Alpha-Induced Multiple Cell Upsets in Standard and Radiation Hardened SRAMs Manufactured in a 65 nm CMOS Technology," *IEEE Transactions on Nuclear Science,* Volume 53, No. 6, Part 1, pp. 3479–3486, 2006.
11. G. Gasiot, D. Giot, and P. Roche, "Multiple Cell Upsets as the Key Contribution to the total SER of 65nm CMOS SRAMs and its Dependence on Well Engineering," *IEEE Transactions on Nuclear Science,* Volume 54, No. 6, Part 1, pp. 3479–3486, 2006.
12. D. Giot, P. Roche, G. Gasiot, and R. Harboe-Sørensen, "Multiple Bit Upset Analysis in 90nm SRAMs: Heavy Ions Testing and 3D Simulations," *IEEE Transactions on Nuclear Science,* Volume 54, No. 4, pp. 904–911, 2007.
13. D. Giot, P. Roche, G. Gasiot, J.-L. Autran, and R. Harboe-Sørensen, "Ion Testing and 3D Simulations of Multiple Cell Upset in 65nm Standard SRAMs," *IEEE Transactions on Nuclear Science,* Volume 55, No. 4, pp. 2048–2054, 2008.
14. J.L. Autran, P. Roche, J. Borel, C. Sudre, K. Castellani-Coulié, D. Munteanu, et al., "Altitude SEE Test European Platform (ASTEP) and First Results in CMOS 130nm SRAM," *IEEE Transactions on Nuclear Science,* Volume 54, No. 4, pp. 1002–1009, 2007.

15. J.L. Autran, P. Roche, G. Gasiot, T. Parrassin, J.P. Schoellkopf, and J. Borel, "Real-time Soft-Error Rate Testing of Semiconductor Memories on the European Test Platform ASTEP," *Proceedings of the 2nd International Conference on Memory Technology and Design (ICMTD 2007),* Giens, France, pp. 161–164, May 7–10, 2007.
16. J.L. Autran, P. Roche, S. Sauze, G. Gasiot, D. Munteanu, P. Loaiza, et al., "Real-Time Neutron and Alpha Soft-Error Rate Testing of CMOS 130nm SRAM: Altitude versus Underground Measurements," IEEE Proceedings of the International Conference on IC Design and Technology, June 2–4, 2008, Grenoble, France.
17. J.L. Autran, P. Roche, S. Sauze, G. Gasiot, D. Munteanu, P. Loaiza, et al., "Altitude and Underground Real-Time SER Characterization of CMOS 65nm SRAM," *IEEE Transactions on Nuclear Science,* Volume 56, No. 4, pp. 2258–2266, 2009.
18. JEDEC Standard Measurement and Reporting of Alpha Particles and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices, JESD89 Arlington, VA: JEDEC Solid State Technology Association. Available at: http://www.jedec.org/download/search/JESD89A.pdf
19. P.H. Stoker, L.I. Dorman, and J.M. Clem, "Neutron Monitor Design Improvements," *Space Science Reviews,* Volume 93, pp. 361–380, 2000.
20. J.M. Clem and L.I. Dorman, "Neutron Monitor Response Functions," *Space Science Review,* Volume 93, pp. 335–359, 2000.
21. L.I. Dorman, *Cosmic Rays in the Earth's Atmosphere and Underground,* Kluwer Academic Publishers, 2004, chapter 6.
22. http://www.seutest.com/FluxCalculation.htm
23. F. Lei, S. Clucas, C. Dyer, and P. Truscott, "An Atmospheric Radiation Model Based on Response Matrices Generated by Detailed Monte Carlo Simulations of Cosmic Ray Interactions," *IEEE Transactions on Nuclear Science*, Volume 51, pp. 3442–3451, 2004.
24. F. Lei, A. Hands, S. Clucas, C. Dyer, and P. Truscott, "Improvement to and Validations of the QinetiQ Atmospheric Radiation Model (QARM)," *IEEE Transactions on Nuclear Science,* Volume 53, pp. 1851–1858, 2006.
25. V. Chazal, R. Brissot, J.F. Cavaignac, B. Chambon, M. De Jésus, D. Drain, et al., "Neutron Background Measurements in the Underground Laboratory of Modane," *Astroparticle Physics,* Volume 9, pp. 163–172, 1998.
26. E. Yakushev, "Neutron Flux Measurements in the LSM," LSM internal report. Available at: http://www-lsm.in2p3.fr
27. R. Lemrani, M. Robinson, V.A. Kudryavtsev, M. De Jesus, G. Gerbier, and N.J.C. Spooner, "Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors, and Associated Equipment," Volume 560, Issue 2, pp. 454–459, 2006.
28. R.C. Baumann, T. Hossain, S. Murata, and H. Kitagawa, "Boron Compounds as a Dominant Source of Alpha Particles in Semiconductor Devices," *Proc. IEEE International Reliability Physics Symposium (IRPS),* pp. 297–302, 1995.
29. R.C. Baumann and E.B. Smith, "Neutron-Induced Boron Fission as a Major Source of Soft Errors in Deep Submicron SRAM Devices," *Proc. IEEE Int. Reliab. Phys. Symp.,* 2000, pp. 152–157.
30. http://www.synopsys.com/products/tcad/tcad.html
31. iRoC Technologies, http://www.iroctech.com/sol_test_112.html
32. H. Kobayashi, H. Usuki, K. Shiraishi, H. Tsuchiya, N. Kawamoto, G. Merchant, et al., "Comparison between Neutron-Induced System-SER and Accelerated-SER in SRAMs," *Proceedings of the IEEE International Reliability Physics Symposium,* Phoenix, AZ, pp. 288–293, 2004.

33. P. Roche, G. Gasiot, K. Forbes, V. O'Sullivan, and V. Ferlet, "Comparisons of Soft Error Rate for SRAMs in Commercial SOI and Bulk Below the 130-nm Technology Node," *IEEE Transactions on Nuclear Science*, Volume 50, No. 6, pp. 2046–2054, 2003.
34. D. Munteanu and J.L. Autran, "Modeling of Digital Devices and ICs Submitted to Transient Irradiations," *IEEE Transactions on Nuclear Science,* Volume 55, No. 4, pp. 1854–1878, 2008.
35. S. Mukherjee, *Architecture Design for Soft Errors*, Elsevier, Inc., 2008.

10 Fault Tolerance Techniques and Reliability Modeling for SRAM-Based FPGAs

Keith S. Morgan, Michael Caffrey, James Carroll, Derrick Gibelyou, Paul Graham, William Howes, Jonathan Johnson, Daniel McMurtrey, Patrick Ostler, Brian Pratt, Heather Quinn, and Michael Wirthlin

CONTENTS

10.1 INTRODUCTION

Static random access memory (SRAM)-based field-programmable gate arrays (FPGAs) offer a lower cost alternative to application-specific integrated circuit (ASIC) technologies for custom hardware. Although FPGAs cannot provide the same level of performance as an ASIC, they boast at least an order of magnitude computational efficiency benefits over general-purpose processors and in many cases much higher performance. FPGAs also offer on-demand reconfiguration useful for in-field bug fixes, upgrades, or entirely new applications.

There is growing interest in using FPGAs for the data-intensive signal processing applications often used in space-based systems. In addition to their low nonrecurring engineering (NRE) costs and flexibility, the programmable logic and on-chip memories are well suited to complex high-throughput applications, particularly those used in signal processing. Furthermore, their on-demand reconfiguration capability supports the use of multiple applications on the same chip through time multiplexing. In addition, new applications can be “uploaded” on mission, and existing applications can be fixed if bugs are found or modifications are desired.

A variety of projects have demonstrated the benefits of using FPGAs in spacecraft [1,2]. Specific examples include the Mars rovers, which use FPGAs for motor control and landing pyrotechnics [3], and the Los Alamos National Laboratory satellite CFESat, which uses nine FPGAs as part of its high-performance computing payload [4,5].

SRAM-based FPGAs, however, are very sensitive to the space radiation environment—particularly radiation-induced single-event upsets (SEUs). SEUs are particularly detrimental to FPGAs because they can change not only the state of user flip-flops and internal block memory but also the contents of the configuration memory, which can alter the behavior of the user circuit. This is much different from ASICs. In an ASIC, the routing and logic are considered insensitive to SEUs, so only the latches need to be protected. In an FPGA, on the other hand, the latches, logic, and routing must all be protected. As a result, the safe use of FPGAs in space requires careful design considerations and the use of well-proven SEU mitigation techniques.

10.1.1 Organization

Section 10.2 of this chapter contains a brief introduction to radiation effects relevant to SRAM-based FPGAs, particularly the nondestructive single-event radiation effects. In Section 10.3, two techniques are introduced for *detecting* and *correcting* single-event effects in an SRAM-based FPGA. Section 10.4 introduces a laundry list of techniques for *mitigating* the errors induced by single-event effects. In most systems, complementary techniques are used for (1) detecting and correcting single-event effects and (2) mitigating their effects. Finally, in Section 10.5 a reliability model is introduced for estimating mean time to failure (MTTF) for an SRAM-based FPGA. Section 10.5 concludes with a case study for a hypothetical scenario.

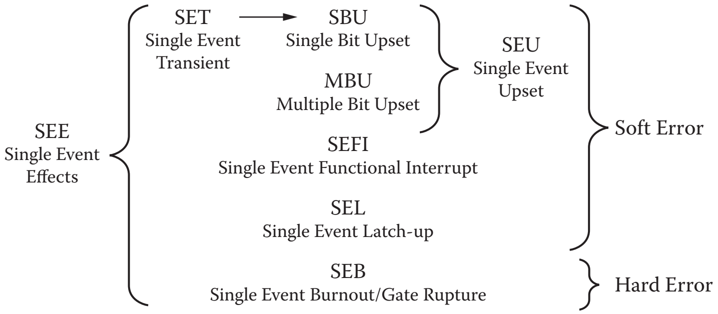


FIGURE 10.1 The different classes of soft and hard errors collectively known as single-event effects. (From K. Morgan, "SEU-induced persistent error propagation in SRAM-based FPGAs." Brigham Young University Masters Thesis, 2006). Cross-reference: R. Baumann, Single-event effects in advanced CMOS technology, in 2005 *IEEE NSREC Short Course*, Seattle, WA, July 2005, pp. II–1– II–59.)

10.2 FPGA RADIATION EFFECTS

Energetic particles, for example, protons trapped in the Van Allen radiation belts, can deposit unwanted charge in a microelectronic device. Excess charge can cause transient faults or even permanent damage. Figure 10.1 lists the different types of transient and permanent faults, commonly called soft and hard errors, respectively. The set of all soft and hard errors collectively is known as single-event effects (SEEs). Note that the name *single-event effect* implies that typically a single particle causes an effect at a specific instant in time. Other time-integrated effects also occur as the result of exposure to energetic particles over longer periods. This chapter focuses exclusively on single-event effects.

10.2.1 Destructive Single-Event Effects

The most common destructive single-event effect is single-event latchup (SEL). SEL is an unwanted short circuit caused by ionizing radiation that can destroy a device from the resulting overcurrent situation if the device is not power cycled. Other destructive SEEs include single-event gate rupture (SEGR) and single-event burnout (SEB).

10.2.2 Nondestructive Single-Event Effects

10.2.2.1 Single-Event Upsets

A single-event upset occurs when deposited charge directly causes a change of state in dynamic circuit memory elements (e.g., flip-flop, latch). In other words, an SEU occurs when a charged particle changes the stored value in a memory element from logic "1" to logic "0", or vice versa.

The change in state of one element is a single-bit upset (SBU). The change in state of more than one element is a multiple-bit upset (MBU) [6]. Both SBUs and MBUs are caused by a *single* particle. Coincident SEUs (either SBU or MBU) that occur in the same device within a relatively short amount of time are sometimes referred to as multiple independent upsets (MIUs) [7].

10.2.2.2 Single-Event Transients

A single-event transient (SET) occurs when deposited charge causes a dynamic circuit memory element (e.g., flip-flop, latch) to latch an incorrect value. For example, charge deposited by a particle into a combinational circuit can cause an undesired transient pulse that, if registered by a flip-flop, results in an incorrect value stored by the flip-flop.

SETs cause a problem only if the resulting pulse is latched by a memory element. In a flip-flop, for example, this occurs only if the pulse coincides with the clock edge that registers a new value. Consequently, the rate of faults caused by SETs typically varies with clock frequency. As clock frequency increases, the period decreases, and the likelihood of a pulse coinciding with a clock edge goes up.

10.2.2.3 Single-Event Functional Interrupts

Single-event functional interrupts (SEFIs) are a special class of SEUs. Electronic devices often contain system control circuitry that, if upset, causes system-wide failure of the device. In an FPGA, there can be several SEFI mechanisms. For example, an SEU in the power-on-reset (POR) control circuitry causes the entire FPGA to reset, including its configuration storage. Since SEFIs are a nondestructive effect a hard reset or power cycle typically can restore a device after an SEFI.

10.2.3 SEEs in FPGAs

Devices that contain dense arrays of memory cells are especially sensitive to SETs and SEUs due to the large amount of memory state within a relatively small amount of circuit area. Much like SRAM and DRAM, SRAM-based FPGAs contain large amounts of memory cells within a device and are especially sensitive to radiation-induced single-event effects. Most modern FPGAs contain tens of millions of bits for device configuration, internal block memory, user flip-flops, and so forth. For example, the newest Xilinx Virtex-6 LX760 FPGA contains over 176 million configuration bits.

In an FPGA the bits that define the operation of the user-designed circuit are the largest component of the configuration memory. These configuration memory cells define the operation of the configurable logic blocks, routing resources, input/output (I/O) blocks, and other programmable FPGA resources. Like all other (nonradiation-hardened) memory, the configuration memory is susceptible to SEUs. Upsets within the configuration memory are especially troublesome as they may change the operation of the look-up tables, routing, I/O, and other device resources.

The susceptibility of an FPGA's configuration memory to SEUs means that design reliability techniques for FPGAs must mitigate upsets in the configuration (e.g., logic, routing, I/O) in addition to the user circuit flip-flops. This is much different than soft error mitigation approaches for ASICs and other custom technologies. In these technologies, the routing and logic are usually considered insensitive to soft errors. As a result, soft error mitigation techniques usually address only the latches within the circuit.* Consequently, the overhead for constructing fault-tolerant latches within custom circuits is much lower than the overhead required to mitigate failures in the logic, routing, and I/O of FPGAs.

10.3 SEU DETECTION AND CORRECTION TECHNIQUES

A fault-tolerant device, by definition, tolerates occasional faults. A commercial SRAM-based FPGA cannot prevent faults so it must tolerate faults and, depending on the desired reliability, possibly mitigate the effects of the faults. In most practical SRAM-based FPGA systems the faults are repaired as quickly as possible to limit their effects. Two methods for detecting faults are discussed in this section. Methods for mitigating the effects of faults are discussed in Section 10.4.

10.3.1 Scrubbing

Memory scrubbing has been used for many years to increase the reliability of memory within radiation environments [8]. Memory scrubbing requires the ability to detect upsets within a memory and to use this information to repair the memory. Although there are many variations of memory scrubbing, most memory scrubbing systems use a simple single-error correction, double-error detection (SECDED) code embedded within the data to detect and correct upsets [8,9]. A memory scrubber begins by reading an error correction code (ECC) encoded codeword from the memory, determines if any errors exist in that codeword, and then repairs the codeword in memory if an error was found. Most memory scrubbers operate by continuously sequencing through the memory array, checking for and fixing SEUs.

The process of memory scrubbing can be applied to SRAM-based FPGAs to improve FPGA reliability in the presence of configuration SEUs. Like memory scrubbing, FPGA configuration scrubbing requires the ability to detect configuration upsets through readback as well as the ability to repair SEUs through reconfiguration. Most FPGA scrubbing techniques require some external hardware including external memory for configuration data storage. Like memory scrubbing, there are a variety of ways to implement configuration scrubbing in FPGAs [10,11]. One popular method speeds up the check-for-upsets step by using a checksum or CRC on each frame.

Configuration memory scrubbing is the most popular SEU detection and correction technique for SRAM FPGAs for several reasons. First, it is relatively straightforward to implement. Second, it guarantees that SEUs will be corrected within some bound. Third, and most important, it can run in the background without interrupting operation of the user circuit.

Configuration memory scrubbing also has several drawbacks. First, it is unable to detect errors that occur in dynamic user-defined memories (i.e., flip-flops or RAMs); only errors in the configuration bitstream are corrected. Second, there is a delay from the time an upset occurs to the time when it is detected by readback. The worst-case delay is the time it takes for a full readback cycle to occur (this can be over hundreds

* This is not always true as upsets within logic and routing may generate transient errors that are latched within the sequential circuitry.

of milliseconds). Third, systems that implement scrubbing require additional external circuitry that is generally implemented in dedicated hardware.

10.3.2 Duplication with Comparison

Duplication with comparison (DWC) is an alternative error detection technique for SRAM-based FPGAs. DWC is a simple hardware redundancy technique that detects errors in a circuit caused by an SEU rather than directly detecting the SEU itself. DWC is implemented within the user circuit rather external hardware. It uses two identical copies of a circuit and compares the outputs of these circuit copies to determine if an SEU has occurred. The comparator circuit detects differences in the operation of the two circuits and signals the system with an SEU flag [12].

Although DWC is not often used, it has many attractive benefits. First, it is relatively easy to apply to any circuit. Automated design tools exist for applying DWC to an arbitrary circuit. Second, it can be used to detect more than just SEUs, including transient errors and upsets within user flip-flops. Third, it can detect errors immediately and potentially allow the system to respond more quickly to SEUs. Fourth, it requires limited external hardware support.

DWC has one primary drawback that keeps it from being used in practice. DWC requires at least 50% of the FPGA's resources to be kept in reserve for the duplicate copy. Those resources could have been used for other functions or more processing power. Since application designers typically try to squeeze every last ounce of performance out of an FPGA, they are typically loath to give up resources if not absolutely necessary. However, DWC would be an excellent choice, for example, in a system in a benign low Earth orbit (LEO) that expects SEUs to rarely occur. System hardware complexity could be kept to a minimum by simply reconfiguring the entire device anytime the SEU flag was asserted.

10.4 SEU-INDUCED ERROR MITIGATION TECHNIQUES

Several methods for mitigating the errors caused by SEUs in an SRAM-based FPGA are discussed in this section.

10.4.1 Triple Modular Redundancy

Triple modular redundancy (TMR) is a well-known fault mitigation technique that uses redundant hardware to tolerate faults. John von Neumann laid the groundwork for this concept in 1956 [13]. He proposed a technique of independently computing a signal and using "restoring organs" to repair defects in the defective logical "organs." In this work, von Neumann proved mathematically that multiple-line redundancy can improve the reliability of a system composed of unreliable components. Since this seminal paper, numerous studies have introduced variations of this technique and proved various properties of redundant hardware systems [14].

As shown in Figure 10.2, a circuit protected by TMR has three redundant copies of the original circuit and a majority voter. Each copy of the circuit is often called a domain. A single fault in any of the domains will not produce an error at the output

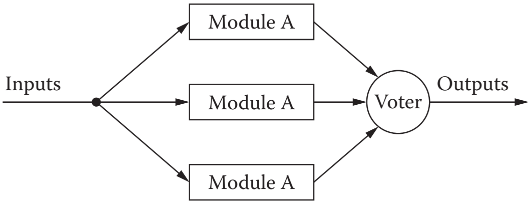


FIGURE 10.2 TMR fault masking. (From K. Morgan, "A comparison of TMR with alternative fault tolerant design techniques for FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2065–2072.)

because the majority voter will select the correct result from the two other working modules. TMR has been used in many systems as a straightforward way to mitigate single faults within a system.

TMR is used extensively in SRAM-based FPGAs to mitigate SEUs. Although TMR is conceptually simple to implement, there are several caveats with respect to implementation in an SRAM-based FPGA. First, TMR must be combined with some form of configuration scrubbing. TMR guarantees uninterrupted operation of the system only as long as two of the three domains are error-free. Without scrubbing, the accumulation of SEUs would eventually cause errors in multiple domains and break the redundancy. Second, circuits with feedback (i.e., state) must have voters inserted somewhere in the feedback loop to restore the state of a corrupted domain. Otherwise, once SEUs corrupted the state of two domains the redundancy would fail. Third, since the voters are implemented with normal FPGA resources susceptible to SEUs, the voters are points of failure. To remove these points of failure, the voters themselves must also be triplicated, as shown in Figure 10.3.

Numerous radiation and fault-injection experiments have demonstrated the improvements in SRAM FPGA reliability using TMR combined with scrubbing [15]. Furthermore, design tools have been created for automating the application of

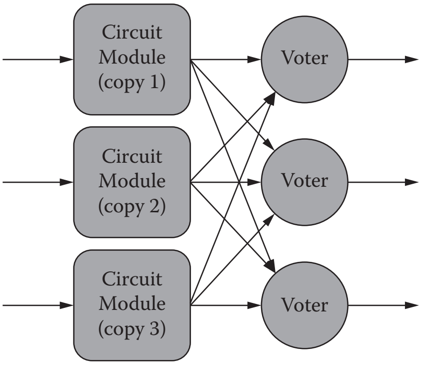


FIGURE 10.3 TMR is implemented with triplicated voters in an SRAM-based FPGA to eliminate single points of failure. (From Morgan, K. et al., Comparison of Time with Alternative Fault Tolerant Design Techniques for FPGAs, *Transactions on Nuclear Science,* 2007.)

TMR on FPGAs to simplify the design process [16,17]. These design tools automatically triplicate design resources, insert voters, and apply voting in circuit feedback paths to ensure sequential structures are resynchronized [18].

10.4.1.1 Partial TMR

Although TMR is straightforward and well accepted, unfortunately it comes at great cost. At a minimum, full TMR of a design requires three times the hardware to implement three identical copies of a given circuit. Additional hardware is also required to perform the majority voting on the three circuit modules. In the worst case, when voting after each look-up table (LUT) within a design, full TMR can require up to *six* times the area of the original circuit [19].

In some cases it may not be possible to sacrifice the resources necessary to fully TMR a circuit. In other cases TMR may provide more reliability than the system's reliability requirements dictate. In situations where full TMR is not possible or unnecessary, partial mitigation is an attractive alternative.

Partial mitigation can increase the overall reliability of the design at a lower cost than a comprehensive approach. Partial mitigation, of course, cannot provide the same level of reliability as full mitigation. A partial mitigation method, therefore, must focus on the components that will increase the reliability of the design the most.

The goal of using partial mitigation is to relax the amount of mitigation applied as compared to full TMR to reduce the area overhead cost with a minimal loss in reliability. Several methods have been proposed for selecting the most critical circuit structures, thus trading hardware cost for SEU immunity. Samudrala et al. proposed a method called *selective triple modular redundancy* (STMR), which uses signal probabilities to find the SEU sensitive subcircuits of a design [20]. Chandrasekhar et al. proposed a modified version of this method, which operates on LUTs rather than logic gates [21]. Pratt et al. proposed a partial mitigation method based on the concept of *persistence* [22] called *partial TMR* [17].

Partial TMR uses the concept of persistence, defined in detail in [22], as a first level of prioritization. A *persistent* error is caused by an SEU that corrupts the internal state of the circuit. While *nonpersistent* errors are corrected simply by repairing the FPGA configuration after an SEU (i.e., scrubbing), persistent errors remain even after the configuration is repaired. Partial TMR gives priority to the circuit components which are susceptible to persistent errors and applies TMR to them first. A software tool is available that automatically implements this technique [17]. It applies TMR using this prioritization scheme until a desired reliability level is reached or a maximum resource count is reached.

10.4.2 Temporal Redundancy

Unlike TMR, which uses spatial redundancy, temporal redundancy, as its name implies, uses redundancy in time. A computation is repeated on the same hardware at three different times. Many studies have shown the utility of temporal redundancy in custom circuit technologies [23,24].

The simplest method to implement temporal redundancy is to repeat the exact same computation on the same hardware module three times. This method, however,

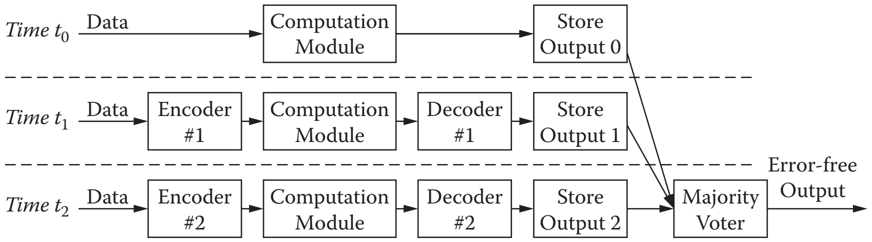


FIGURE 10.4 Each row represents the computations performed at time steps t_0, t_1, and t_2, respectively, in temporal redundancy. (Reprinted with permission from (K. Morgan et al., "A comparison of TMR with alternative fault tolerant design techniques for FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007, Volume: 54 Issue: 6, 2065–2072.)

is inferior to TMR since it corrects only transient errors. A permanent fault in a module would produce incorrect results all three times (except when the fault first manifests itself, in which case one, two, or all three results would be incorrect).

Other forms of temporal redundancy correct both transient faults and permanent faults (e.g., upsets in the FPGA's configuration memory). Figure 10.4 shows the basic method. At time t_0 the computation is performed. At time t_1 the inputs are encoded, run through the same computation module, and then decoded. At time t_2 the inputs are encoded with a different algorithm, run through the same computation module, and then decoded with a different algorithm. By uniquely encoding and decoding the inputs on the second and third pass, two of the three executions will correctly compute the output even if the computation module has a single permanent fault. Care has to be taken to ensure that the encoding scheme can handle permanent faults that manifest themselves *during* one of the three computations. With two correct executions a final majority voter can select the correct output.

Hsu and Swartzlander devised an alternative method of temporal redundancy for arithmetic computations they called time-shared TMR (TSTMR) [23]. For example, TSTMR splits up an addition module as well as the addition operation into three parts [23]. This way, each of the three partial sums is computed simultaneously on three separate hardware modules, as shown in Figure 10.5. The addition operation is performed in thirds, starting with the least significant bits and simultaneously performed on three addition modules. This approach has a relatively low hardware cost since the three partitioned addition modules roughly equal the size of the original module. The only hardware overhead comes in the form of control logic and storage for intermediate results. This method has been demonstrated with both adders and multipliers.

Townsend et al. developed a slight variation of TSTMR they called quadruple time redundancy (QTR) [24]. This method has a slightly lower hardware cost than TSTMR but a greater cost in terms of latency.

One challenge with temporal redundancy is finding appropriate encoding and decoding blocks. A simple set of encoders and decoders has not been found even for all trivial arithmetic operations. A second challenge is that, in an FPGA, the overhead logic (e.g., encoders) is susceptible to SEUs and can potentially add more unreliability

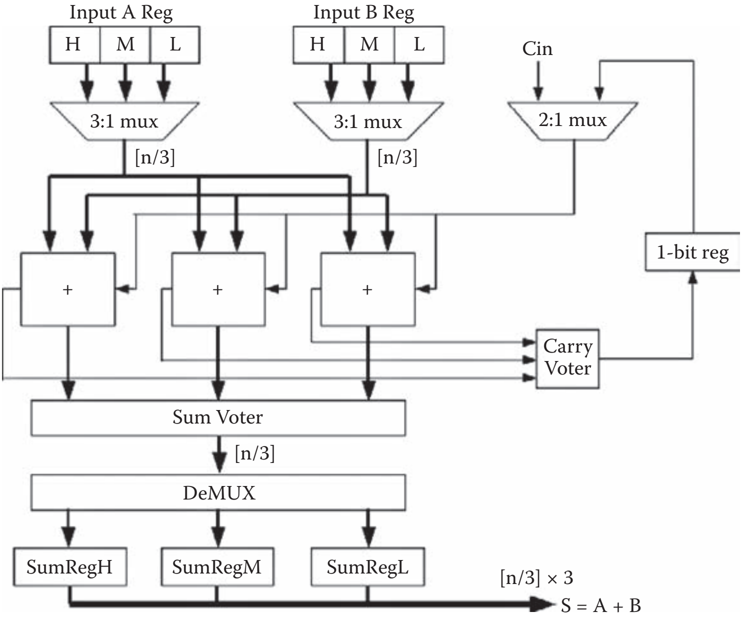


FIGURE 10.5 Time-shared triple modular redundancy (TSTMR) error-correcting adder. (Reprinted with permission from (K. Morgan et al., "A comparison of TMR with alternative fault tolerant design techniques for FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007, Volume: 54 Issue: 6, 2065–2072.) (Cross-reference: Time Redundant Error Correcting Adders and Multipliers, 1992).

than the reliability it adds to the original circuit. A third challenge is that temporal redundancy can alter the timing of the circuit. In addition, temporal redundancy requires three times or more clock cycles to complete as the original computation. Further, the clock period may be reduced, depending on the implementation used. Despite these challenges, temporal redundancy methods use fewer resources than TMR, which, in some situations, may be more important than timing considerations.

10.4.3 State Machine Encoding

State machine encoding is another SEU-induced error mitigation technique studied for SRAM-based FPGAS. The state machine encoding technique leverages ECC to protect finite state machines (FSMs). Protecting FSMs with an ECC is a well-studied problem for custom circuit technologies. In independent efforts Rochet and Kumar compared single-error ECCs to TMR in custom ASIC architectures [25,26]. Although most circuits do not consist entirely of FSMs, this technique may be applicable in situations where only the FSMs need to be protected to achieve the minimum required reliability.

The following techniques were studied by Morgan et al. for protecting FSMs in SRAM-based FPGAs: explicit error correction (EEC) [27], implicit error

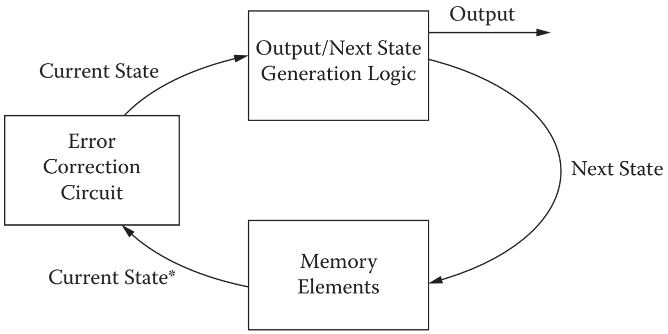


FIGURE 10.6 Finite state machine implementation with explicit error correction. (Reprinted with permission from (From K. Morgan et al., "A comparison of TMR with alternative fault tolerant design techniques for FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007, Volume: 54 Issue: 6, 2065–2072.)

correction (IEC), [27], and a technique proposed by Armstrong [28]. Unlike TMR where an FSM is simply replicated entirely, ECCs redundantly encode the FSM's state variable.

- Explicit error correction: For explicit error correction the state variables are encoded, and additional circuitry is added to detect and correct errors in the encoded state variable. Figure 10.6 shows a block diagram of an FSM protected with EEC. The error correction circuitry is placed between the state storage flip-flops and the next state and output forming logic. This circuitry detects and corrects errors in the state bits, providing a major advantage since the correct codeword is then available to the next state and output forming logic.
- Implicit error correction: Unlike explicit error correction, IEC does not use additional hardware as error correction circuitry. Instead the next state logic is expanded to include all the "erroneous" states that are a Hamming code distance of one away from valid states. The major advantage of this technique is there is no need for additional error correcting logic. However, with the added states, the next state forming logic is more obtuse than in its original unencoded form. "Don't cares" in the state logic are reduced because the set of invalid and valid codewords must be handled instead of just the valid codewords. The same principles must be applied to the output forming logic to ensure the output is correct.
- Armstrong's error correction technique: Armstrong's FSM encoding method can be seen in Figure 10.7 [28]. The outputs of the excitation circuit (including the next state codeword and the system outputs) are divided into *r* subunits. Each subunit generates p-bits of the total set of outputs. The input to each subunit is the only actual state variable. Breaking the next state logic and output forming logic into subunits reduces the chance that an SEU will affect the logic that forms more than one output or next state bit.

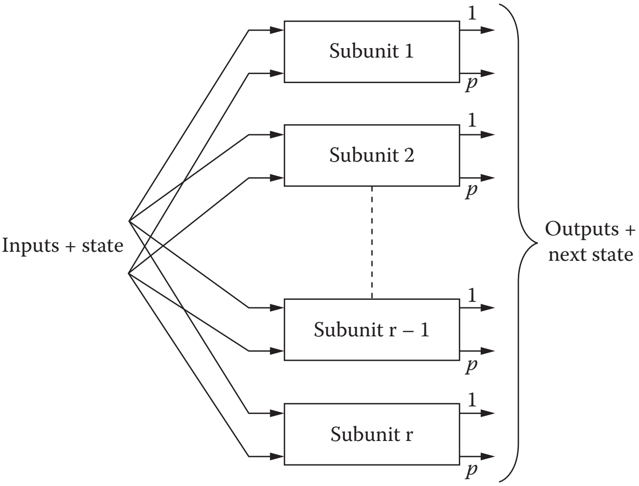


FIGURE 10.7 Armstrong's proposed error correction method. (Reprinted with permission from K. Morgan et al., "A comparison of TMR with alternative fault tolerant design techniques for FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007, Volume: 54 Issue: 6, 2065–2072).

Similar to the EEC technique, the inputs to the subunits must be correct to form the correct outputs and next state. Error correcting circuitry is placed before the inputs of the subunits. The advantage of this system is that the subunits function independently of one another. Only the actual state bits need to be corrected while errors in the check bits need not be corrected as the check bits are not used to generate the next state or output of the system. The check bits are used only to indicate which, if any, of the state bits is in error. This method has the same disadvantages of error correction circuitry that the EEC technique has.

A major advantage of ECCs is that the protected circuit is not 3× larger than the unmitigated circuit as with TMR. ECCs save resources by using minimal redundancy and no voters, although some ECCs do require additional overhead for an error correction circuit. As is the case with temporal redundancy, in an FPGA the overhead logic can potentially add more unreliability than the reliability it adds to the original circuit. Furthermore, like voters, the overhead circuitry can negatively affect timing.

10.4.4 Quadded Logic

Quadded logic is another error mitigation technique studied by Morgan et al. for SRAM-based FPGAs. Tryon [29] and Pierce [30] independently developed quadded logic in the early 1960s shortly after von Neumann [13] published his groundbreaking work on reliable computing. Pierce actually called his more generic theory interwoven logic. Quadded logic is the minimally redundant version of interwoven logic.

Quadded logic is built on four main concepts [29]:

1. The network of logic gates is a plane of alternating AND and OR stages.
2. Each logic gate is quadruplicated.
3. An error is corrected within two levels of logic from the place of origin.
4. The uncorrupted wires in the redundant digital signal mask the error.

In [29], Tryon outlined the steps required to apply these four principles to a circuit. First, the circuit must be specified as a network of alternating AND and OR stages. Second, each logic gate is replicated four times. Since each gate is quadruplicated, there are also four copies all signals. Each gate takes two copies of every input signal in the original circuit. Tryon developed a regular pattern for selecting two of the four signals, which Pierce later generalized. Figure 10.8 illustrates how a section of a circuit is modified to be protected by quadded logic.

Unlike TMR, which masks errors with voters, quadded logic includes error correction in the same hardware that performs the intended function. Figure 10.9 illustrates how this works. In Figure 10.9, AND gate A4 erroneously outputs a logical one. This error spreads to OR gates C3 and C4 to which the output of AND gate A4 is connected. This caused gates C3 and C4 to also erroneously output a logic one. However, these two errors were masked in the next level of logic. The zero output of gates C1 and C2 forced AND gates E1, E2, E3, and E4 to all output the correct value since the outputs of AND gates C3 and C4 do not combine anywhere in the next level of AND gates.

Since quadded logic has error correction embedded within the functional logic it does not incur the overhead of voters like TMR. Since voters require extra hardware and negatively impact timing, quadded logic is a potentially attractive solution for FPGAs. However, quadded logic requires 4× more gates and 2× more inputs at each gate. This input count growth, in particular, does not scale well in an LUT-based FPGA because LUT memory size grows exponentially with the number of inputs.

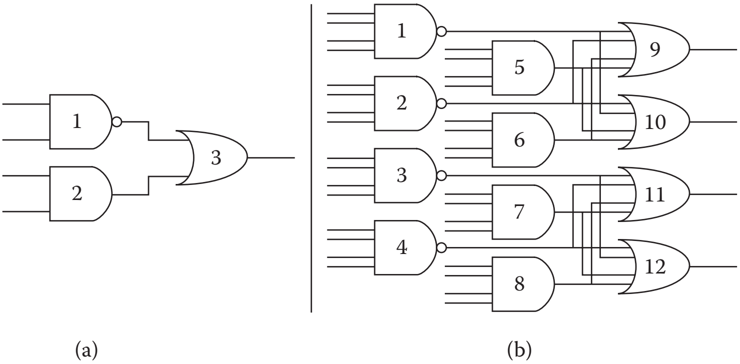


FIGURE 10.8 (a) An example network of two-input logic gates. (b) The resulting circuit after quadded logic is applied to the network in (a). (Reprinted with permission from K. Morgan et al., “A comparison of TMR with alternative fault tolerant design techniques for FPGAs.” *IEEE Transactions on Nuclear Science*, Dec. 2007, Volume: 54 Issue: 6, 2065–2072).

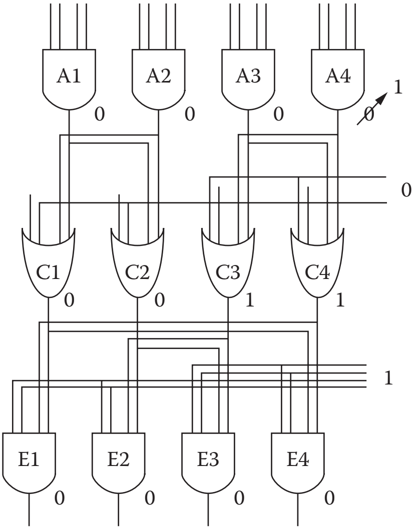


FIGURE 10.9 Erroneous logic “1” from an AND gate being corrected in the next level of AND gates. (Reprinted with permission from K. Morgan et al., “A comparison of TMR with alternative fault tolerant design techniques for FPGAs.” *IEEE Transactions on Nuclear Science*, Dec. 2007, Volume 54 Issue: 6, 2065–2072.)

10.5 RELIABILITY MODEL

Ostler et al. introduced a reliability model that estimates the reliability of FPGA designs protected by TMR and configuration scrubbing [7]. This model is based on R_s, a design-specific parameter for the reliability of an FPGA design during a single configuration scrubbing period. A related parameter, Q_s, is the unreliability of an FPGA design during a single configuration scrubbing period, where $Q_s = 1 - R_s = P(F_s)$. $P(F_s)$ is the probability that the design will fail during a single scrub cycle.* These related parameters are design specific as each FPGA design uses different logic, routing, and other FPGA resources. These parameters may vary widely from design to design, as some designs are very dense and use most FPGA resources while others consume few FPGA resources. An essential part of this reliability model is accurately estimating the parameter Q_s.

Q_s is estimated by computing $Q_{s,i}$ for multiple values of *i*. $Q_{s,i}$ is the joint probability that the circuit fails during a single scrub cycle *and i* upsets occur. $Q_{s,i}$ is computed with the following equation:

$$Q_{s,i} = P(F_s \cap A_i) = P(F_s \mid A_i)P(A_i) \tag{10.1}$$

* The model presented by Ostler et al. [7] is primarily based on the probability of failure rather than reliability. Therefore the design unreliability parameter, Q_s, is used instead of R_s.

where event F_s is a design failure during a single scrub cycle, and event A_i is i SEUs during a single scrub cycle. For example, the probability of both one SEU and design failure during a single scrub cycle can be computed as follows:

$$Q_{s,1} = P(F_s \cap A_1) = P(F_s | A_1)P(A_1) \tag{10.2}$$

Computing $Q_{s,i}$ requires $P(F_s|A_i)$, the conditional probability of failure given i upsets. This can be estimated for various values of i using fault injection or accelerator experiments (see Section 10.5.2). The equation for $Q_{s,i}$ also requires $P(A_i)$, the probability of i SEUs within a scrubbing period. This probability is computed as a Poisson distribution of the upset rate of a specific orbit/condition (see Section 10.5.1).

To determine the unconditional probability of failure during a single scrub cycle, Q_s, the joint probabilities of failure for all i are summed using the following equation:

$$Q_s = Q_{s,i>0} = \sum_{i=0}^{\infty} Q_{s,i} = \sum_{i=0}^{\infty} P(F_s \cap A_i) = \sum_{i=0}^{\infty} P(F_s | A_i)P(A_i) \tag{10.3}$$

Once the model parameter Q_s is known, the failure rate λ (failures in time) of the circuit can be estimated as follows:

$$\lambda = \frac{Q_s}{t_s} \tag{10.4}$$

where t_s is the period of a single scrub cycle. The mean time to failure, *MTTF*, is calculated from λ as follows:

$$MTTF = \frac{1}{\lambda} = \frac{t_s}{Q_s} \tag{10.5}$$

Because an FPGA system will operate in a variety of orbit conditions, it is helpful to estimate a composite failure rate , λ_c, and composite mean time to failure, $MTTF_c$, which incorporate the failure rate during each orbit condition and the probability of operating in that orbit condition. A composite, single-parameter failure rate, λ_c, can be calculated for an interval that spans multiple orbit conditions by obtaining the failure rate during each orbit condition, λ_i, and estimating the probability of being in that orbit condition, ρ_i:

$$\lambda_c = \rho_1\lambda_1 + \rho_2\lambda_2 + \ldots + \rho_n\lambda_n \tag{10.6}$$

where

$$\sum_{i=1}^{n} \rho_i = 1$$

since the FPGA must always operate in one of the *n* orbit conditions (thus the sum of all ρ_i must be 1). A composite mean time to failure, $MTTF_\mathrm{c}$, can be obtained using the equation:

$$MTTF_c = \frac{1}{\lambda_c} = \frac{1}{\rho_1\lambda_1 + \rho_2\lambda_2 + \ldots + \rho_n\lambda_n} \tag{10.7}$$

In Section 10.5.3 a case study is presented with proposed values for ρ_i for several orbit conditions during a solar cycle (e.g., solar min, solar max, worst week, worst day, and peak five minutes).

10.5.1 Estimating Upsets per Scrub Cycle, *P*(*A*$_\mathrm{I}$)

The first parameter needed to determine Q_s is $P(A_\mathrm{i})$, the probability that *i* upsets will occur during a single scrub cycle. This can be modeled with a Poisson distribution:

$$P(A_i) = e^{-\nu}\frac{\nu^i}{i!} \tag{10.8}$$

where ν is the average number of SEUs per scrub period. The parameter ν is calculated by multiplying the orbit-averaged upset rate (SEUs per time), μ, by the scrub period, t_s, as follows:

$$\nu = \mu \times t_s \tag{10.9}$$

The parameter μ can be estimated using modeling tools such as Cosmic Ray Effects on Micro Electronics (CREME96) from the Naval Research Laboratory [31]. CREME96 requires static cross section data for a particular device. The reader is referred to the literature for cross section data for particular FPGAs or for more information on how to collect cross section data. The Xilinx Radiation Test Consortium (XRTC) regularly publishes cross section data for Xilinx FPGAs [32].

10.5.2 Estimating Probability of Design Failure, *P*(*F*$_s$|*A*$_\mathrm{I}$)

The second parameter needed to determine Q_s is $P(F_\mathrm{s}|A_i)$, the conditional probability of design failure during one scrub cycle given *i* SEUs occurred during that scrub cycle. This parameter is design specific and must be estimated for each FPGA design that is to be considered. The parameter can be estimated using either fault injection experiments or accelerator testing.

A generic algorithm for measuring $P(F_s|A_i)$ using fault injection is as follows. Begin by selecting *i* random bits from the configuration bitstream of the design under test (DUT). Toggle each bit, and inject it back into the bitstream to emulate *i* SEUs in the bitstream. Compare the output signals of the DUT against a golden copy of the circuit to check for circuit errors. If a disparity exists between the output signals of the DUT and the output signals of the golden design, record a failure event. Repeat the process *m* times to estimate $P(F_s|A_i)$ or the probability that *i* configuration upsets will cause a design to fail during a single scrub cycle. The following is psuedo code for this general fault-injection algorithm:

```
do {
generate 'i' random bits to upset
inject 'i' upsets into bitstream
check for output error
fix upset bits
reset device
record data to output file
} while number of trials < 'm'
```

A generic algorithm for measuring $P(F_s|A_i)$ using accelerator testing is as follows. For each cycle of the loop pause for time *t* to allow upsets to accumulate on the DUT. Read back each frame and compare against a golden copy to count SEUs. Record and repair any SEUs. If SEUs were found, the device is checked for output errors, and the results are recorded. The algorithm continues until the trial length of time *T* is complete.

```
do {
sleep for time 't'
get and fix upset frames
if upsets found
record upsets to file
check for output error
end if
reset DUT
} while time < 'T'
```

10.5.3 Case Study

To illustrate how the reliability model is used, Ostler et al. performed a case study to demonstrate how to compute $MTTF_c$ for a hypothetical digital signal processing (DSP) circuit in a Xilinx Virtex-4 XQR4VSX55 SRAM FPGA on a satellite in geosynchronous orbit over an 11-year mission [7]. The DSP design is fully protected by TMR. The system detects and corrects SEUs with continuous event-driven scrubbing.

First the following data were collected:

TABLE 10.1
CREME96 Orbit-Averaged SEU Rates μ (SEUs/Devices)

Orbit	Altitude (km)	Inclination (deg.)	Solar Max	Worst Week	Worst Day	Peak 5 Min.
GEO	35,786	0	1.6E-5	1.7E-2	8.8E-2	3.3E-1

Source: Reprinted with permission from P. Ostler et al., "SRAM FPGA reliability analysis for harsh radiation environments." IEEE Transactions on Nuclear Science, Dec. 2009. Volume: 56 Issue: 6, 3519–3526.

- The proton and heavy-ion static cross section estimates, σ_p and σ_h, respectively, for the Xilinx Virtex-4 family of FPGAs. In this case the data were collected from a report published by the Xilinx Radiation Test Consortium (XRTC) [32].
- The combined (proton and heavy-ion) SEU rate, μ, for the XQR4VSX55 FPGA in geosynchronous (GEO) orbit. Since an 11-year mission will cover all orbit conditions, μ must be computed separately for all orbit conditions of interest. In this case estimates of μ were made for solar maximum, flare-enhanced worst week, flare-enhanced worst day, and flare-enhanced peak five minutes. The conditions are described in more detail later in this section. The four values of μ reported in Table 10.1 were estimated using the online CREME96 tool.
- The scrubbing period, t_s, for the scrubbing hardware/software system. For this case study we assume a period of 15 milliseconds.
- The probability of circuit failure given i upsets, $P(F_s|A_i)$, for the digital signal processing circuit implemented in the XQR4VSX55 FPGA. $P(F_s|A_i)$ was measured using fault injection and validated with accelerator testing. The $P(F_s|A_i)$ data are plotted in Figure 10.10.

Next, the collected data and Equation (10.8) were used to estimate the probability of i upsets during a single scrub period for each of the four orbit conditions. For example, the $P(A_i)$ values for the normal solar max conditions in GEO are plotted in Figure 10.11.

With $P(A_i)$ and $P(F_s|A_i)$, Equation (10.1) was used to estimate the joint probability that during a single scrub cycle the circuit fails and i upsets occur. $Q_{s,i}$ values for the flare-enhanced peak five minute orbit conditions are plotted in Figure 10.12 for i = 1 to i = 5.

The unconditional probability of failure during a single period was estimated using Equation (10.3). Q_s is simply the inner product of $P(A_i)$ and $P(F_s|A_i)$, or in other words, the sum of $Q_{s,i}$ over all i. Values of Q_s for each of the four orbit conditions are reported in Table 10.2. The values of Q_s were used to compute λ for each of the four orbit conditions. The values are reported in Table 10.3. *MTTF* values for each orbit condition were computed from λ and are reported in Table 10.2.

Finally, the composite mean time to failure was computed using Equation (10.7). The values of ρ_i were obtained by estimating the amount of time spent in each of the

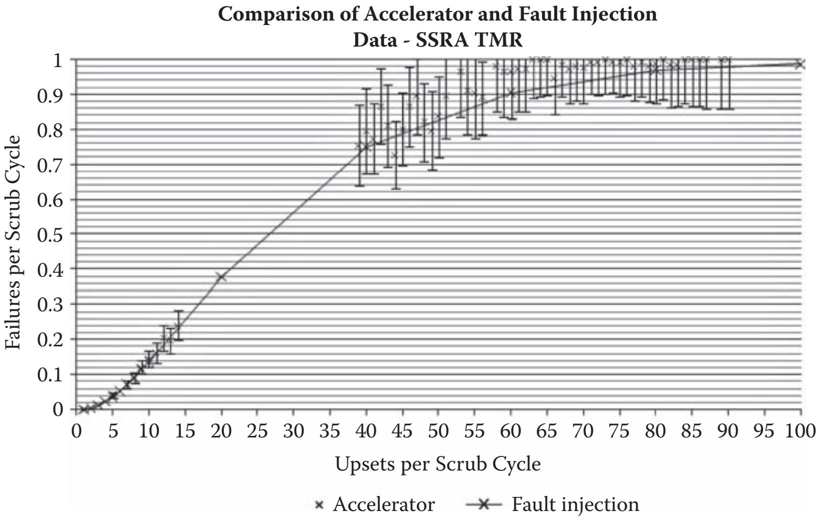


FIGURE 10.10 $P(F_s|A_i)$ measurement data from fault injection for the example DSP circuit protected by TMR. Accelerator validation data are overlaid with error bars. (From P. Ostler et al., "SRAM FPGA reliability analysis for harsh radiation environments." *IEEE Transactions on Nuclear Science*, Dec. 2009, Volume: 56 Issue: 6, 3519–3526.)

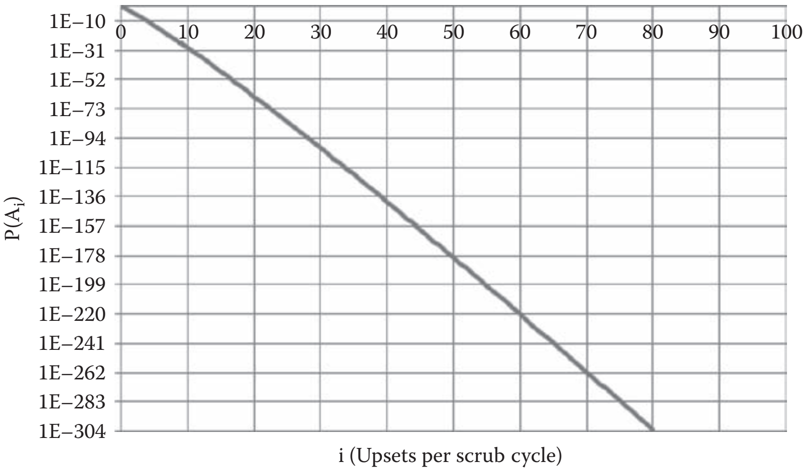


FIGURE 10.11 Probability of *i* upsets per scrub cycle for a Xilinx Virtex-4 XQR4VSX55 FPGA in solar max conditions in GEO with a 15 ms scrub period. (Reprinted with permission from P. Ostler et al., "SRAM FPGA reliability analysis for harsh radiation environments". *IEEE Transactions on Nuclear Science*, Dec. 2009 Volume 56 Issue: 6, 3519–3526.)

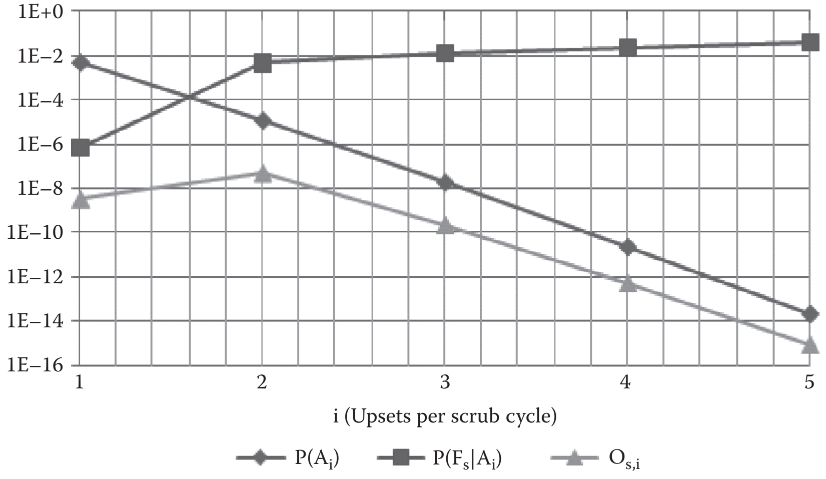


FIGURE 10.12 Plot of $P(A_i)$, $P(F_s|A_i)$, and $Q_{s,i}$ for the SSRA TMR design during peak five-minute conditions in GEO. (Reprinted with permission from P. Ostler et al., "SRAM FPGA reliability analysis for harsh radiation environments." *IEEE Transactions on Nuclear Science*, Dec. 2009, Volume: 56 Issue: 6, 3519–3526.)

TABLE 10.2
MTTF of the DSP Kernel Design for Orbit Conditions

	Solar Max		Worst Week		Worst Day		Peak 5 Minutes	
Orbit	Q_s	***MTTF* (s)**	Q_s	***MTTF* (s)**	Q_s	***MTTF* (s)**	Q_s	***MTTF* (s)**
GEO	1.7E-13	8.9E10 (2810 years)	3.7E-10	4.1E7 (1.3 years)	6.2E-9	2.4E6 (28 days)	7.8E-8	1.9E5 (2.2 days)

Source: Reprinted with permission from P. Ostler et al, "SRAM FPGA reliability analysis for harsh radiation environments." IEEE Transactions on Nuclear Science, Dec. 2009. Volume: 56 Issue: 6, 3519–3526.

following four orbit conditions: solar max, worst week, worst day, and peak five minutes. The amount of time estimated in each orbit condition is summarized as follows:

1. Peak five minutes: The CREME96 peak five minute flux model is based on the peak five-minute averaged fluxes observed on GOES in October 1989 [36]. For this case study, Ostler et al. assume each SEP event results in one peak five-minute orbit condition (300 seconds).
2. Worst day: The CREME96 worst-day model is based on SEP fluxes averaged over 18 hours beginning at 1,300 UT on October 20, 1989. This period was the single largest flux enhancement in October 1989 [34]. For this case study, Ostler et al. assume that each SEP results in one worst-day orbit condition for 18 hours minus the five minutes spent in a peak five-minute condition (6.5E4 seconds).

TABLE 10.3
Probability of GEO Orbit Conditions within an 11-Year Solar Cycle and Composite Failure Rate for the DSP Kernel Design

Condition	Time (s)	ρ	λ	ρλ
Solar max	2.97E8	.856	1.1E-11	9.7E-12
Worst week	4.49E7	.129	2.8E-8	3.7E-9
Worst day	4.97E6	.014	5.2E-7	7.5E-9
Peak 5 min	2.31E4	.000067	6.7E-6	4.4E-10
Composite	3.5E8	1.00	λ_c = 1.4E-8	

Source: Reprinted with permission from P. Ostler et al., "SRAM FPGA reliability analysis for harsh radiation environments." IEEE Transactions on Nuclear Science, Dec. 2009, Volume: 56 Issue: 6, 3519–3526.

3. Worst week: The CREME96 worst-week model is based on SEP fluxes averaged over 180 hours (7.5 days) beginning at 1,300 UT on October 19, 1989. This week was the most severe SEP environment observed in the last two solar maxima [34]. For this case study, Ostler et al. assume that each SEP results in the worst-week orbit condition for 7.5 days minus the time spent in a worst day and peak five-minute condition (5.8E5 seconds).
4. Solar max: For this case study, Ostler et al. [7] assume the remainder of the time is under normal, solar max conditions. For the purpose of this model we do not distinguish between solar min and solar max conditions as their flux levels are orders of magnitude lower than the flare-enhanced conditions.

For the purposes of this case study Ostler et al. assume seven SEP events* per year regardless of position in the solar cycle for a total of 77 SEP events during an 11-year solar cycle [7,36]. They also make the pessimistic assumption that each SEP event results in a worst week, worst day, and peak five-minute flux. In other words, they make the very pessimistic assumption that every SEP event is as bad as the October 1989 event.

The time spent in each orbit condition during an 11-year solar cycle is listed in Table 10.3. The total time is 3.5E8 seconds, the number of seconds in 11 years. The probability of operating in each of the four orbit conditions, ρ_i, is determined by dividing the amount of time spent in each orbit condition by the time in a full solar cycle. Assuming seven SEP events per year, 85.6% of the time involves normal conditions, 12.9% of the time involves worst-week conditions, 1.4% of the time involves worst-day conditions, and a very small amount of time is spent in the worst five-minute peak conditions.

* Each SEP event is assumed to produce 106 protons with energy greater than 30 MeV [36].

The values in Table 10.3 were applied to the equation for λ_c to estimate a composite failure rate, also reported in Table 10.3. The composite failure rate is used to compute a composite *MTTF* for the design over all orbit conditions. For example, using the results in Table 10.3, $MTTF_c$ of the digital signal processing circuit in GEO orbit is 1.1E8 seconds. In other words, the mean time to failure of this circuit in GEO orbit, protected by TMR and scrubbing, is 3.4 years.

10.6 CONCLUSION

SRAM-based FPGAs provide much higher performance than general-purpose processors, at a lower cost than ASIC technologies, plus on-demand reconfiguration. These benefits make FPGAs an attractive alternative to radiation hardened processors for the signal processing applications often used in space-based systems. However, SRAM-based FPGAs are only fault tolerant. They are particularly radiation-induced SEUs. As a result, the safe use of FPGAs in space requires careful design considerations and the use of well-proven SEU mitigation techniques.

Commercial SRAM-based FPGAs cannot prevent faults. Thus, to limit the effects of faults, they must be repaired as quickly as possible. Two methods were introduced for detecting and repairing SEUs: (1) configuration scrubbing; and (2) duplication with comparison. Configuration scrubbing uses external hardware to directly monitor and refresh the FPGA's configuration memory. DWC uses FPGA logic to indirectly monitor the FPGA's configuration memory by detecting errors induced by upsets. The configuration memory is optionally refreshed when an error is detected.

Since commercial SRAM-based FPGA systems can only limit the effects of SEUs with scrubbing or DWC, they must also use some form of SEU mitigation if it is desired to mask those effects. Several mitigation methods were introduced including triple modular redundancy, partial TMR, temporal redundancy, state machine encoding, and quadded logic. Each method has advantages and disadvantages. By far the most popular method is TMR since it is the easiest to implement and can provide excellent reliability. Partial TMR is also gaining traction since it can sometimes provide sufficient reliability at lower costs than full TMR.

Ostler et al.'s reliability model was also introduced for estimating the failure rate and MTTF of an SRAM-based FPGA design [7]. The model can be used to estimate a composite failure rate or composite MTTF for missions that span several orbit conditions.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy through the National Nuclear Security Administration's Office of Nonproliferation Research & Development (NA-22). NA-22 funding was received through the SNDD Processing and Communications Mission Sustainment Project, the Sensor-Oriented Processing and Networking Project, and the Cibola Flight Experiment.

The author would like to acknowledge the many contributions of the following individuals: James Carroll (Brigham Young University; BYU), Michael Caffrey (Los Alamos National Laboratory), Derrick Gibelyou (BYU), Paul Graham (LANL), William

Howes (BYU), Jonathan Johnson (BYU), Daniel McMurtrey (BYU), Patrick Ostler (BYU), Brian Pratt (BYU), Heather Quinn LANL, and Michael Wirthlin (BYU).

This work was approved for public release; distribution is unlimited. The release number is LA-UR-09-06447.

REFERENCES

1. Weigand, D. and Harlacher, M., A Radiation-tolerant Low-power Transceiver Design for Reconfigurable Applications, ITT Industries Advanced Engineering and Sciences Division, Earth Science Technology Conference (ESTC), Paper A1P2, 2002.
2. Morris, K., FPGAs in Space, *FPGA and Structured ASIC Journal,* August 2004.
3. Ratter, D., FPGAs on Mars, *Xilinx xCell Journal*, vol. 50, 2004.
4. Caffrey, M., A Space-Based Reconfigurable Radio, In: *International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA),* Las Vegas: CSREA Press, pp. 49–53, June 2002.
5. Caffrey, M. et al., *On-Orbit Flight Results from the Reconfigurable Cibola Flight Experiment Satellite (CFESat),* IEEE Computer Society Press, IEEE Symposium on FPGAs for Custom Computing Machines (FCCM), April 5–7, Napa, CA.
6. Quinn, H. et al., Radiation-Induced Multi-Bit Upsets in SRAM-Based FPGAs, *IEEE Transactions on Nuclear Science,* Vol. 52, Issue 6, Part 1, pp. 2455–2461.
7. Ostler, P. et al., SRAM FPGA Reliability Analysis for Harsh Radiation Environments, *IEEE Transactions on Nuclear Science,* Volume 56, Issue 6, Part, pp. 3519–3526.
8. Saleh, A., Serrano, J., and Patel, J., Reliability of Scrubbing Recovery Techniques, *IEEE Transactions on Reliability,* Volume 39, Issue 1, April, 1990, pp. 114–122.
9. LaBel, K. et al., Solid State Tape Recorders: Spaceflight SEU data for SAMPEX and TOMS/meteor-3, *IEEE Radiation Effects Data Workshop,* July 21, 1993. pp. 77-84.
10. Heiner, J. et al., FPGA Partial Reconfiguration via Configuration Scrubbing, 11th Internation Workshop, August 2009 pp. 99–104 FPL.
11. Berg, M.L., The NASA Goddard Space Flight Center Radiation Effects and Analysis Group Virtex-4 Scrubber, Annual Xilinx Radiation Test Consortium Meeting, 2007.
12. Johnson, J. et al., Using Duplication with Compare for On-line Error Detection in FPGA-based Designs, *IEEE Aerospace Conference,* Big Sky, *IEEE Aerospace* paper 1255, pp. 1–11, 2008.
13. von Neumann, J., *Probable Logics and the Synthesis of Reliable Organisms from Unreliable Components,* Princeton, NJ: Princeton University Press, Automata Studies (Annals of Match Studies No. 34), 1956.
14. Short, R., *The Attainment of Reliable Digital Systems through the Use of Redundancy—A Survey*, IEEE Computer Group News, pp. 2–17, 1968.
15. Pratt, B. et al., Improving FPGA Design Robustness with Partial TMR, IRPS Conference, April 2006. pp. 226–237
16. Xilinx, Xilinx TMR Tool Product Brief, 2006.
17. Pratt, B. et al., Fine-Grain SEU Mitigation for FPGAs Using Partial TMR, *IEEE Transactions on Nuclear Science,* Vol. 55, No. 4, August 2008. pp. 2274–2280, 2008.
18. Carmichael, C., Triple Module Redundancy Design Techniques for Virtex FPGAs, Xilinx Corporation Technical Report XAPP197, November 1, 2001.
19. Wirthlin, M. et al., Hardness by Design Techniques for Field-programmable Gate Arrays, *11th Annual NASA Symposium on VLSI Design,* Coeur d'Alene, pp. WA11.1–WA11.6, 2003.
20. Samudrala, P., Ramos, J., and Katkoori, S., Selective Triple Modular Redundancy (STMR) Based Single-Event Upset SEU Tolerant Synthesis for FPGAs, *IEEE Transactions on Nuclear Science,* Vol. 51, Issue 5, Part 4, October 2004. pp. 2957–2969, 2004.

21. Veezhinathan, K. et al., *Reduced Triple Modular Redundancy for Tolerating SEUs in SRAM-Based FPGAs,* Washington, DC: MAPLD Conference, 2005.
22. Morgan, K. et al., SEU-Induced Persistent Error Propagation in FPGAs, *IEEE Transactions on Nuclear Science,* Volume 52, Issue 6, Part 1, December 2005.
23. Swartzlnader, E., Hsu, Y., and Earl, J., Time Redundant Error Correcting Adders and Multipliers, *Defect and Fault Tolerance in VLSI Systems,* November 1992 *IEEE Workshop*.
24. Townsend, W., Abraham, J. and Swartzlander, E., Quadruple Time Redundancy Adders, *Defect and Fault Tolerance in VLSI Systems, 18th IEEE International Symposium* p. 250, 2003.
25. Rochet, R., Leveugle, R., and Saucier, G., Analysis and Comparison of Fault Tolerant FSM Architectures Based on SEC Codes, *Defect and Fault Tolerance in VLSI Systems,* p. 9, October 27–29 1993 *IEEE International Workshop*.
26. Zacher, D. and Kumar, N., *Automated FSM Error Correction for SEUs,* Washington, DC: MAPLD, 2004.
27. Frenzel, J. and Niranjan, S., A Comparison of Fault-Tolerant State Machine Architectures for Space-Borne Electronics, *IEEE Transactions on Reliability,* March 1996, Vol. 45 Issue 1 pp. 109–113.
28. Armstrong, D., A General Method of Applying Error Correction to Synchronous Digital Systems, *Bell System Technical Journal,* pp. 577–593, 1961.
29. Mann, W. and Wilcox, R. (eds.) *Redundancy Techniques for Computer Systems*, Spartan Books, 1962. J. G. Tryon, "Quadded Logic," pp. 205–228.
30. Pierce, W., *Fault-Tolerant Computer Design*, Academic Press, 1965.
31. Tylka, A. et al., CREME96: A Revision of the Cosmic Ray Effects on Micro-Electronics Code, *IEEE Transactions on Nuclear Science,* pp. 2150–2160, December Vol. 44 Issue 6 Part 1 1997.
32. Allen, G., Swift, G., and Carmichael, C., *Virtex-4QV Static SEU Characterization Summary,* Pasadena: Jet Propulsion Laboratory, 2008.
33. Violante, M. and Sterone, L., A New Reliability-Oriented Place and Route Algorithm for SRAM-based FPGAs, *IEEE Transactions on Computers,* June 2006 pp. 732–744, 2006.
34. NRL, CREME96. Available at: https://creme96.nrl.navy.mil/
35. Morgan, K. et al., A Comparison of TMR with Alternative Fault Tolerant Design Techniques for FPGAs, *Transactions on Nuclear Science,* December 2007 Vol. 54, Issue 6 Part 1.
36. Nymmik, R., Relationships among Solar Activity, SEP Occurrence Frequency, and Solar Energetic Particle Event Distribution Function, *Proceedings of the 25th ICRC,* Edited by B. H. Dingus, D. B. Kieda, and M. H. Salomon, Vol. 6, University of Utah, Salt Lake City, UT 1999, pp. 280–284.

Section II

Circuits and Systems

11 Assuring Robust Triple Modular Redundancy Protected Circuits in SRAM-Based FPGAs

Michael Caffrey, Paul Graham,
Jim Krone, Kevin Lundgreen, Keith S. Morgan,
Brian Pratt, and Heather Quinn

CONTENTS

11.1 INTRODUCTION

Field-programmable gate arrays (FPGAs) with volatile programming memory, such as the Xilinx Virtex families, have made inroads into space-based processing tasks [1]. These devices are attractive for a number of reasons. Static random access memory (SRAM)-based FPGAs can provide custom hardware implementations of applications that are often faster than traditional microprocessor implementations without the cost of manufacturing application-specific integrated circuits (ASICs). Furthermore, using commercial off-the-shelf (COTS) devices with available, mature design tools should reduce the cost of designing space-based systems. Finally, reprogrammability also allows designers to reconfigure the device while deployed with either new applications or new implementations of existing applications, which should increase the usable lifetime of the entire system.

In this chapter, we will focus on only the Xilinx Virtex reconfigurable SRAM-based FPGAs. Unlike most SRAM-based FPGAs, Xilinx has published several reports verifying latchup immunity [2,3], which have made them the preferred choice for space usage. This family of devices implements logic in look-up tables (LUTs), where logic is reduced from gates to a four input and one output equation that is stored in configuration memory. Furthermore, the wiring is programmable so that design flow tools can determine how best to optimize routing signals from one LUT to another LUT through routing switches. Therefore, unlike traditional SRAM devices that store data, in an SRAM-based FPGA much of the stored data defines the user circuit, including whether particular routes or LUTs are used. In the more modern devices embedded cores, such as multipliers and microprocessors, have become more common. On-chip SRAM, called BlockRAM, for storing intermediate processing values is increasing in size for each generation of device.

Single-event upsets (SEUs) caused by ionizing particles are a problem for these devices, as SEUs change values stored in SRAM. For an SRAM-based FPGA, SEUs could cause changes in the programmable logic and routing, which could potentially cause the user's circuit to malfunction. To this end, most FPGA-based systems attempt to mask SEUs by protecting the user's circuit with triple modular redundancy (TMR) [4-6]. The current suggestion for space-based FPGA designs is to triplicate all logic (modules and voters) and all signals (inputs, outputs, clock, and reset). The viability of TMR-protected circuits, particularly on a single chip, remains an open question. While our past research [6] on the Virtex-I has demonstrated through fault injection and accelerated testing that logic-level TMR with programming data scrubbing effectively mitigates single-bit upsets (SBUs), other researchers have shown analytically that SBUs can defeat TMR on the Virtex-I [7]. Our later research on the Virtex-II has shown that when the aforementioned guidelines are followed it is possible to completely remove all unprotected cross section [8], and the design will be susceptible only to multiple-bit upsets (MBUs), multiple independent upsets, and single-event functional interrupts. As discussed later in this chapter, multiple-bit upsets can be problematic for mitigated circuits. As the occurrence of single-bit SEUs dominates events on these devices, we believe that most designs using these TMR criteria should be adequately protected on orbit, if implemented properly.

Unfortunately, applying TMR techniques to users' circuits can be error-prone, leading to unprotected cross section in the protected circuits. Designers are not necessarily at fault in these scenarios. In particular, a number of problems can be tied to the design flow tools, which can be circumvented entirely only by avoiding many of the design automation tools—a choice most designers will not make. In other cases, designers are forced to apply TMR only partially to a design to meet device or resource constraints. In those scenarios, the unprotected cross section will be only partially removed.

In all of these scenarios, hardness assurance issues with TMR-protected circuits can be very difficult to ascertain, especially in complex systems. In the past, we have used fault injection [6] to estimate hardness assurance issues that might exist in FPGA designs. Unfortunately, designers cannot always perform fault injection effectively on their designs due to flight system limitations or the limitations of hardware prototypes amenable to fault injection. In these cases, a nonhardware method for estimating the unprotected cross section and for finding design flaws is necessary.

In this chapter, we will discuss the ability to use TMR to protect radiation-induced faults and the efficacy of modeling tools to determine design-level problems with the application of TMR. In Section 11.2 we will provide an overview of the sensitivity of Xilinx SRAM-based FPGAs to radiation-induced upsets. In Sections 11.3 and 11.4 we will discuss the use of TMR to protect FPGA user circuits. Finally, in Section 11.5 we will introduce modeling tools that might be helpful in determining problems with the application of TMR.

11.2 OVERVIEW OF SEU AND MBU DATA FOR FPGAS

Before continuing our discussion, we would like to discuss static testing results that we have collected that show the sensitivity of the devices to ionizing radiation. We have performed accelerator tests on five generations of Xilinx Virtex devices. We have used a similar test fixture, as shown in Figure 11.1, for static testing of the Virtex-II, Virtex-4, and Virtex-5 devices. The fixture consists of both hardware and software components. The hardware test fixture provides support for reading (*readback*) and writing (*programming*) the configuration data in the SRAM-based FPGA. The software test fixture controls the programming and reading back the FPGA.

The hardware test fixture the authors used for the Virtex-5 results is shown in Figure 11.1. It uses two Xilinx AFX series development boards (one Virtex-II and one Virtex-5) biased nominally. A third, smaller board contains a USB connection to a host computer that allows the computer to control the operation of the test fixture. The hardware test fixture uses custom software that performs programming, differencing, and readback, as well as keeping the Graphical User Interface (GUI) updated with minimal statistics to help the testers determine whether the test fixture remains operational. The FPGA is completely reprogrammed and error locations, called *the differential bitstream*, saved to the host computer's hard drive every second in this scheme, which allows us to test continuously at high fluences without accumulating too many upsets per readback. With this scheme we can collect approximately 3,600 differential readbacks per hour. Custom software is used to analyze the differential

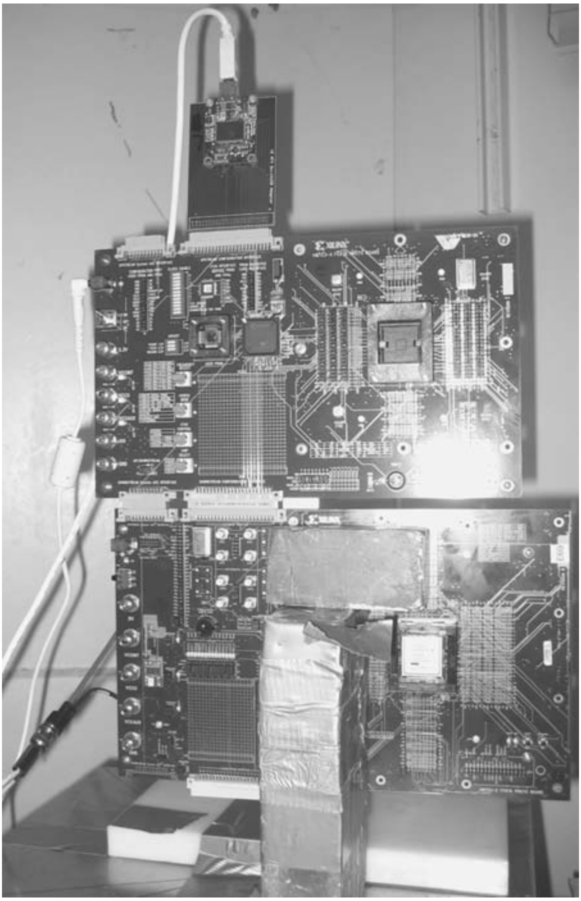


FIGURE 11.1 Hardware test fixture for the Xilinx Virtex-5 device. (Reprinted with permission from Heather Quinn, Keith Morgan, Paul Graham, Jim Krone, Michael Caffrey, "Static proton and heavy ion testing of the Xilinx Virtex-5 device." *Nuclear Space Radiation Effects Conference's Data Workshop* July 17–20, 2007, Honululu HI.)

readbacks for the device's sensitivity to errors and to categorize the errors by size and location.

From testing the Xilinx Virtex devices, we have been able to observe several trends [9]. The devices' bit cross sections (Figure 11.2)* have been within an order of magnitude over 10 years, but the percentage of MBUs for the entire device (Figure 11.3) has rapidly increased. Table 11.1 lists the frequency of MBUs in protons, and Figure 11.3 shows the frequency of MBUs for heavy ions for Virtex family devices. Both of the proton and heavy ion data sets have shown that MBUs have become more frequent in the newer devices. Figure 11.3 also shows how MBU frequency increases with energy. At the highest tested LETs there are 21% MBUs on

* The Virtex-I device is a combination of normal incidence and angular data, but the rest of the curves are solely from normal incidence data.

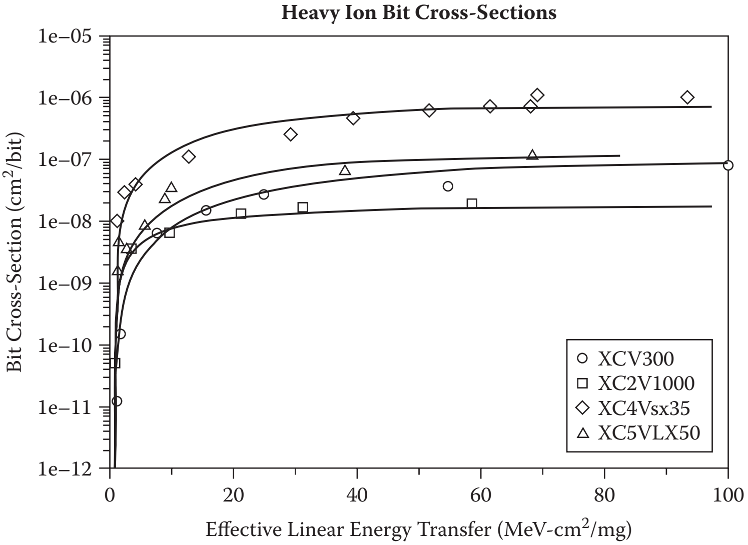


FIGURE 11.2 Heavy-ion bit cross sections for Virtex family devices. (From: H. Quinn, P. Graham, J. Krone, M. Caffrey, and S. Rezgui, *IEEE Transactions on Nuclear Science,* Vol. 52, No. 6, pp. 2455–2461, December 2005. With permission.)

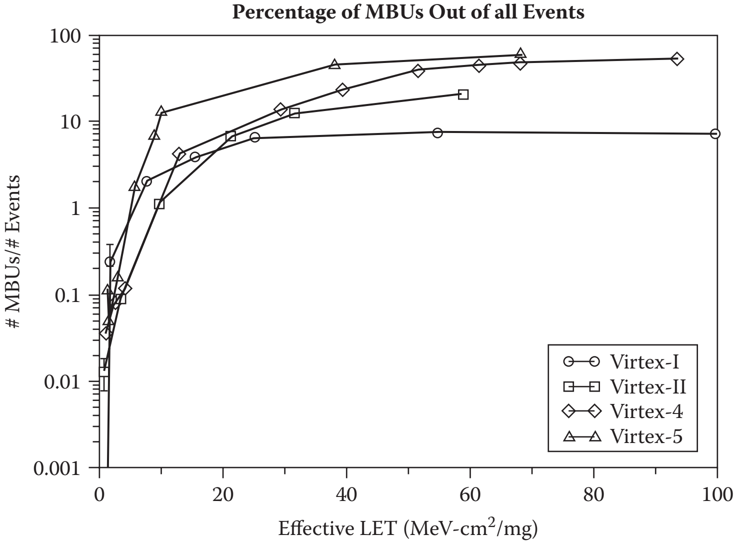


FIGURE 11.3 Percent of MBU events out of all events induced by heavy-ion radiation for four Xilinx FPGAs. (Reprinted with permission from Heather Quinn, Paul Graham, Keith Morgan, Jim Krone, Michael Caffrey, Michael Wirthlin, "An introduction to radiation-induced failure modes and related mitigation methods for Xilinx SRAM FPGAs." In the proceedings of the International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA) 2008.)

TABLE 11.1
Frequency of Upset Events and Percent of Total Events Induced by Proton Radiation (63.3 and 65 MeV) for Five Xilinx FPGAs

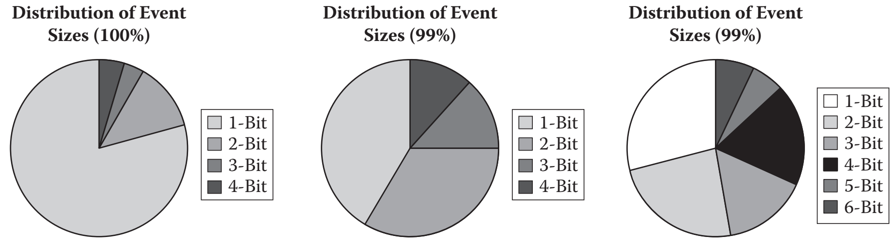
Family	Total Events	One-Bit Events	Two-Bit Events	Three-Bit Events	Four-Bit Events
Virtex	241,166	241,070 (99.96%)	96 (0.04%)	0 (0%)	0 (0%)
Virtex-II	541,823	523,280 (98.42%)	6,293 (1.16%)	56 (0.01%)	3 (0.001%)
Virtex-II Pro	10,430	10,292 (98.68%)	136 (1.30%)	2 (0.02%)	0 (0%)
Virtex-4	152,577	147,902 (96.44%)	4,567 (2.99%)	78 (0.05%)	8 (0.005%)
Virtex-5 (65 MeV)	2,963	2,792 (94.23%)	161 (5.43%)	9 (0.30%)	1 (0.03%)
Virtex-5 (200 MeV)	35,324	31,741 (89.86%)	3,105 (8.79%)	325 (0.92%)	110 (0.43%)

Source: Reprinted with permission from Heather Quinn, Paul Graham, Keith Morgan, Jim Krone, Michael Caffrey, Michael Wirthlin, "An introduction to radiation-induced failure modes and related mitigation methods for Xilinx SRAM FPGAs" in the proceedings of the International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA) 2008. Reprinted with permission from, Quinn, H., Graham, P,. Krone, J., Caffrey, M., Rezgui, S., "Radiation-induced multi-bit upsets in SRAM-based FPGAs," IEEE Transactions on Nuclear Science, Dec. 2005 Volume: 52 Issue: 6, 2455–2461.

TABLE 11.2
Bit Cross Section for SEUs for Protons for Several Xilinx FPGAs

Device	Energy (MeV)	σ_{bit}(cm^2/bit)
XCV1000	63.3	$1.32 \times 10^{-14} \pm 2.69 \times 10^{-17}$
XC2V1000	63.3	$2.10 \times 10^{-14} \pm 4.64 \times 10^{-17}$
XC4VLX25	63.3	$1.08 \times 10^{-14} \pm 2.71 \times 10^{-17}$
XC5VLX50	65.0	$7.57 \times 10^{-14} \pm 1.35 \times 10^{-15}$
XC5VLX50	200.0	$1.07 \times 10^{-13} \pm 5.37 \times 10^{-16}$

Source: Reprinted with permission from Heather Quinn, Paul Graham, Keith Morgan, Jim Krone, Michael Caffrey, Michael Wirthlin, "An introduction to radiation-induced failure modes and related mitigation methods for Xilinx SRAM FPGAs" in the proceedings of the International Conference on Engineering of Reconfigurable Systems and Algorithms (ERSA) 2008. Reprinted with permission from, Quinn, H., Graham, P,. Krone, J., Caffrey, M., Rezgui, S., "Radiation-induced multi-bit upsets in SRAM-based FPGAs," IEEE Transactions on Nuclear Science, Dec. 2005 Volume: 52 Issue: 6, 2455–2461.



(a) 2V1000, Normal Incidence, 58.7 MeV-cm²/mg

(b) 5VLX50, Normal Incidence, 68.3 MeV-cm²/mg

(c) 5VLX50, 60-Degree Angle, 72.8 MeV-cm²/mg

FIGURE 11.4 Worst-case distribution of heavy-ion event sizes. Reprinted with permission from, Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

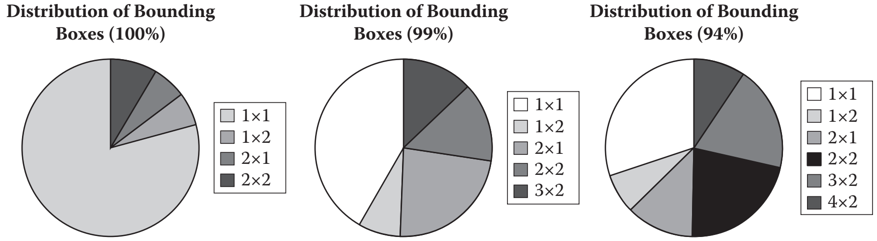
the Virtex-II (XC2V1000) at 58.7 MeV-cm²/mg, and there are 59% MBUs on the Virtex-5 (XC5VLX50) at 68.3 MeV-cm²/mg. While most of the events on both the Virtex-II and the Virtex-5 involve four or fewer bits, the distribution of event sizes changed. As shown in Figure 11.4, the dominant SEU sizes for the 150 nm Virtex-II are one-bit and two-bit events (Figure 11.4a), whereas three- and four-bit events total 25% of all events (Figure 11.4b) for the Virtex-5. This phenomenon has caused the average SEU event size to increase from 1.3 bits in the Virtex-II at 58.7 MeV-cm²/mg to 2.6 bits in the Virtex-5 at 68.3 MeV-cm²/mg at normal incidence.

We have also been watching the trend for MBU shapes, since it indicates the amount of spacing that would be needed to correct TMR defeats. With the diversity of possible MBU shapes, we report shapes as bounding boxes.* As shown in Figure 11.5, most of the events on the Virtex-II can be confined within two rows and two columns (Figure 11.5a), whereas most of the Virtex-5 events are confined by three rows and two columns (Figure 11.5b).

As shown in Figures 11.4c and 11.5c, both MBUs and bounding boxes worsen for nonnormal incidence radiation strikes. At an LET of 72.8 MeV-cm²/mg, striking the device with Kr at a 60 degree angle has a 72% probability of an MBU, and the average SEU size is 4.2. These figures show an increase in the percentage of larger MBUs, including a 13% probability of five- and six-bit events, and it is also visible from the figure that only 94% of MBUs are confined between four rows and two columns.

While much of these data might appear dire, SRAM-based FPGAs are less likely to experience MBUs than traditional SRAM devices. Gasiot [11] shows that multiple-bit upsets could comprise 23–81% of all events that occur on the device in neutron radiation, depending on the well design. Furthermore, Tosaka [12] reports that two-bit upsets occur at approximately 10% of the frequency of single-bit upsets in neutron radiation. Tosaka also noted that MBUs occurred more frequently in smaller feature-sized devices than larger feature-sized devices. As neutron and proton

* A bounding box is the number of rows and columns that completely cover an MBU. A discussion of bounding boxes can be found in [10].



(a) 2V1000, Normal Incidence, 58.7 MeV-cm^2/mg

(b) 5VLX50, Normal Incidence, 68.3 MeV-cm^2/mg

(c) 5VLX50, 60-Degree Angle, 72.8 MeV-cm^2/mg

FIGURE 11.5 Worst-case distribution of heavy-ion bounding boxes. (Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2037– 2043.)

radiation cause similar reactions in complementary metal-oxide semiconductor (CMOS) devices, these two articles indicate that that MBUs are 3 to 27 times more likely in traditional SRAM devices than SRAM-based FPGAs. As the structure of SRAM-based FPGAs is more heterogeneous in layout and the memory structures are not optimized for area like traditional SRAM devices, these devices are less likely to have MBUs than traditional SRAM devices.

11.3 TMR PROTECTION OF FPGA CIRCUITS

While the device is inherently radiation-tolerant and, therefore, SEU-sensitive while on orbit, using TMR to protect the circuit should mask the effects of many SEUs as long as there is only one error in the system at a time. Even still, there are a number of ways either the design could be flawed or the design implementation toolset could render the final implementation of the design flawed. Furthermore, there might be design constraints placed on the circuit (e.g., not enough input/output pins for full triplication) that affect the reliability of the design. The potential reliability issues for TMR-protected designs for these devices are three-fold: (1) problems with the circuit design; (2) design constraints; and (3) architectural influences on the circuit design. These issues are presented in the following sections.

11.3.1 Circuit Design Problems

The first issue is the design of the TMR-protected circuit. Many FPGA circuit designers use a hardware description language (HDL), such as VHDL or Verilog, to describe the FPGA circuit. The circuit description is then optimized for area and translated to an industry standard circuit representation, called Electronic Design Interchange Format (EDIF), using circuit synthesis tools, such as Synplify or Synopsys. Even the most careful descriptions of TMR-protected circuits are often undermined by the synthesis tools. As FPGA synthesis and implementation tools are designed to remove redundant logic to optimize the circuit for area and speed, these tools usually

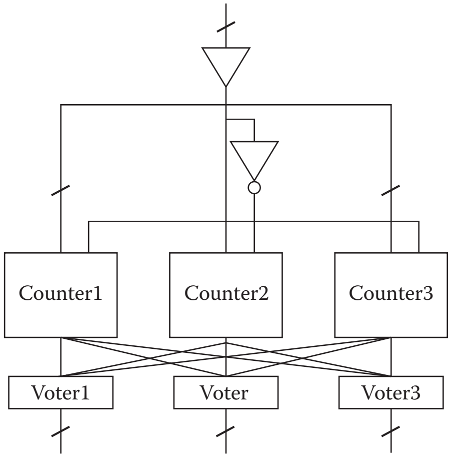


FIGURE 11.6 An example of a TMR-protected counter design with a number of design flaws. (Reprinted with permission from Quinn, H., Graham, P., Pratt, B., "An automated approach to estimating hardness assurance issues in triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Volume: 55 Issue:6, 3070–3076.)

recognize and remove the functional redundancy intended to improve reliability. More subtly, though, sometimes the redundant modules remain but are no longer functionally equivalent or independent. In this case, part of the redundant logic is reduced to a single implementation in one module that is shared by all three modules. This problem is shown in Figure 11.6. In this situation, the inverter that is used for the least significant bit in the counter has been removed from all three modules, and the inverted data are shared by all three counter modules. While the circuit is still functionally equivalent to a correctly TMR-protected design, untriplicated logic now exists in the circuit. In large circuit designs, detecting this issue is difficult.

Figure 11.6 also highlights a common problem in TMR-protected circuits with feedback loops. Feedback loops in TMR-protected systems are also sensitive to *persistent errors* [13] and need to use triplication and voters to break the feedback loops. The counter in Figure 11.6 shows a feedback loop that has not been cut properly, and the counters will not be able to autonomously resynchronize after the SEU is removed. In this scenario, while the first SEU in one feedback loop will be masked by the voters, a second SEU in another feedback loop is not guaranteed to be masked. To fix the counter design, the output of the voters will need to be fed back to the input of counters to remove the persistent cross section.

To circumvent issues with the synthesis tools, the recommended approach for applying TMR is to apply TMR to the EDIF circuit descriptions. While this can be done in a text editor for small designs, the authors suggest using one of the two automated tools (BL-TMR [13] and TMRTool [14]). As these tools work with the postsynthesis circuit representation, the synthesis tools are able to optimize the basic circuit without affecting the application of TMR. The optimization of the circuit after synthesis is usually limited to removing signals that do not route to output pins. Therefore, optimization of the redundant modules is unlikely. Also, these tools have been built with an understanding of persistence issues so that feedback loops are properly protected by TMR.

11.3.2 Device Constraint Problems

The second issue regards design constraints. Since these devices can be pin- and area-constrained, designers are sometimes unable to implement a fully triplicated design. In particular, not being able to triplicate input, output, clock, or reset signals is common, and SEUs in the input/output blocks, routing, global clock network, and flip-flops could cause errors to manifest across all three logic modules. The counter in Figure 11.6 shows that the three counters are sharing the same inputs. While this design is not uncommon in cases where the data stream originates from a single sensor, unprotected cross section exists between the input pins and the inputs of the counters. Furthermore, we have found that, when not using automated tools to apply TMR to a design, the optimization by the synthesis tools of the TMR-protected circuit with shared inputs is more likely to remove most of the reliability-based redundancy. While it is possible to triplicate some of these signals internally on the device,* an unprotected cross section still exists in the system between the input pins and the triplicated flip-flops responsible for splitting the signal.

Designers might also find themselves constrained by the device's size and are unable to fully triplicate the circuit logic. The BL-TMR tool addresses this problem by balancing the need to protect the most essential parts of the design and meeting area constraints by applying TMR partially to the circuit. BL-TMR gives highest priority to subcircuits that may reach a persistent error state due to feedback, since error recovery may require external intervention. In cases where TMR has been only partially applied to the circuit, there exists an unprotected cross section. The effect of this unprotected cross section can be hard to quantify.

11.3.3 Circuit Implementation and Architectural Problems

The third issue is the implementation of the circuit on the architecture. Several problems are directly tied to the placement of the circuit onto the device, such as domain crossing errors and logical constants. These devices are very complex and have a number of architectural components (e.g., the resources for fast carry-chains, shift registers, and embedded arithmetic functions) to improve the speed, power, and silicon use of user circuits. As an artifact of translating a design to the specific resources available on the FPGA, sometimes the inputs to carry-chains and multipliers need to be tied to a ground, as when the multiplication is using fewer inputs than the embedded multipliers have. These grounds are tied to a logical constant on the power network, called the *global logic network*. The power network for the Virtex-I and Virtex-II is a virtual network of grounds and V_{CC}s that use constant LUTs. Since the power network is load balanced by the design flow tools, redundant logic could share the same power network, introducing potential single points of failure into the design.

* Clocks should only be triplicated using the global clock buffers, and skew should be carefully monitored.

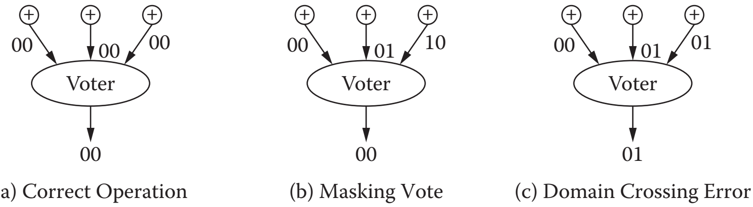
Further complicating the issue, the power network is implemented in SEU-sensitive logic, which could translate to unprotected cross section in the design. Since the load balancing affects the number of constant LUTs that are used, the exact quantity of single points of failure caused by them cannot be determined until after the design is placed. Much of this problem is minimized in the later devices, as the architecture has been modified to include ties to the ground plane throughout the device.

Both BL-TMR and TMRTool tools address this issue by extracting the half-latches and the constant LUTs to input/output pins to provide these constant logic values in a TMR domain-aware manner. Since this solution elevates logical constants to a global signal, like the clock tree, the input/output pins used for the logical constants will need to be triplicated.

The final reliability problem involves the placement of the design on the device. Since many of the tools involved in converting a designer's circuit description to a bitstream are attempting to minimize the implemented circuit's area and maximize the clock speed, redundant logic can be placed in proximity. We have shown in the Virtex-II that, when area and timing constraints cause the device to be highly used, there is a chance an MBU can defeat TMR by introducing errors into multiple redundant modules, a situation referred to as a domain crossing error (DCE) [8]. Given the complexity of DCEs, we will discuss them in greater detail in the following section.

11.4 DOMAIN CROSSING ERRORS

A DCE occurs when two or more redundant copies (domains) of the TMR circuit are corrupted such that the voter selects the wrong value. As shown in Figure 11.7c, the ionized particle would need to change at least two TMR domains to the same wrong value to cause a DCE. Since two domains have matching answers, the system does not detect the incorrect operation. Therefore, unless erroneous output data can be detected or locations on the device that have known DCE issues are accounted for, these errors could remain undetected. As MBUs can manifest in the system as independent errors, it is more likely that an MBU could trigger this condition than a single-bit SEU.



(a) Correct Operation (b) Masking Vote (c) Domain Crossing Error

FIGURE 11.7 An example of a domain crossing error in a two-bit adder with TMR and bit-wise voting. (Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K,. "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*. Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

Many factors can make systems both more or less vulnerable to DCEs, such as the robustness of the design, voter design, device use, and sensitivity of the system to errors. The most robust implementation of TMR has triplicated voters, data signals, and control signals, since untriplicated they would become single points of failure. When properly triplicated, identical failures in two domains are needed to propagate the error. Likewise, bitwise voting can mask many potential TMR vulnerabilities, as shown in Figure 11.7b, since failures affecting different bits would vote out. Designs that use most of the device could potentially heighten the risk of a DCE since there appears to be a correlation between high device use and DCEs. Finally, the sensitivity of the circuit to errors can be a factor in whether errors can propagate in a system. For example, logic masking lowers the probability that a DCE manifests as an observable output error. All of these situations will be discussed in greater detail in the results section.

The rest of this section focuses on our test methodology, the test results, analysis of results, and a simple probability model for determining the likelihood of occurrence.

11.4.1 Test Methodology and Setup

In this section, we present our test methodology for both our fault injection test fixture and our accelerator test fixture, as well as an overview of the test circuits used for this study. While the hardware aspect of the test fixtures is the same, the experimental approach and the software test fixtures are different.

11.4.1.1 Test Circuits

The test circuits implemented are listed in Table 11.3. All of these circuits were designed for the Virtex-II XC2V1000 device. These designs intentionally represent the worst-case scenarios for TMR limitations. While synthetic in nature, these circuits are representative of "corner cases" for circuits that can be used as part of a larger design. As noted in Table 11.3, there is a mixture of feed forward and feedback circuits within the complete set of circuits.

Each circuit has two TMR implementations, except the linear feedback shift register (LFSR). One TMR version has triplicated voters interspersed frequently in the design under test (DUT) design, and the other only votes once off-chip. The LFSR test circuit was made from an intellectual property module made available by Xilinx, and the TMR implementation only votes off-chip. The TMR implementations were designed with the accepted best practices for creating FPGA TMR circuits with triplicated data, control signals, and voters.

Each circuit was designed such that most of the device is used. In the case of the off-chip voting circuits, the design is functionally the same as the frequent voting circuits, so the device use is lower for these designs. We also explicitly created designs that used the special features of the Virtex-II device, such as the fast carry-chains and the embedded multipliers. The OR tree and AND tree circuits exclusively use LUTs to implement logic. While the Virtex-II device has BlockRAM resources for on-device temporary data storage, the cross section and the mitigation methods for the BlockRAM are substantially different from the reconfigurable fabric. Given space limitations, circuits using BlockRAM resources are not highlighted in this study.

TABLE 11.3
Circuit Resource Use

Circuit	Type	Voting	Flip-Flop (%)	LUT (%)	Slice (%)
Shift register	Feed Forward	Frequent	96	97	97
		Off-Chip	96	0	96
Adder tree	Feed Forward	Frequent	44	48	71
		Off-Chip	44	22	46
Divider tree	Feedback	Frequent	81	33	98
		Off-Chip	81	27	97
AND tree	Feed Forward	Frequent	45	90	100
		Off-Chip	45	45	45
OR tree	Feed Forward	Frequent	45	90	100
		Off-Chip	45	45	45
LFSR	Feedback	Off-Chip	89	2	100
Pseudo LFSR	Feedback	Frequent	50	99	99
		Off-Chip	50	49	50

Source: Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs ." IEEE Transactions on Nuclear Science, Dec. 2007 Volume: 54 Issue: 6, 2037–2043.

11.4.1.2 Fault Inject Test Methodology

The test circuits were fault injected using a Virtex-II SEU emulator we have used for previous studies [15]. Figure 11.8 shows a picture of the hardware test fixture. The SEU emulator operates two Xilinx Virtex-II AFX demonstration boards in lock step with a USB interface to a host computer. One AFX board has the "golden" device and the other has the DUT. While both devices have a copy of the same circuit that is being tested, the golden board has additional compution. This computation supplies the input vectors to both test circuits, receives the output vectors from both test circuits, and determines if there is mismatch between both sets of output vectors. Any mismatches are relayed to the host PC for logging.

The SEU emulator software test fixture is designed to inject faults across the entire device with user-specified patterns (e.g., one-bit upsets, two-bit vertical upsets). In this manner, it is easier with fault injection to gain complete coverage of the entire device than with accelerator testing. Still, since logic masking can play a strong role in whether errors propagate to outputs, it is necessary to run the SEU emulator multiple times for each test circuit and each test pattern to get a representative set of DCEs.

The software aspect of the SEU emulator injects faults in the following manner. First, a fault (or faults in the case of an MBU) is injected into the DUT at a specified location through programming data (i.e., bitstream) manipulation and partial

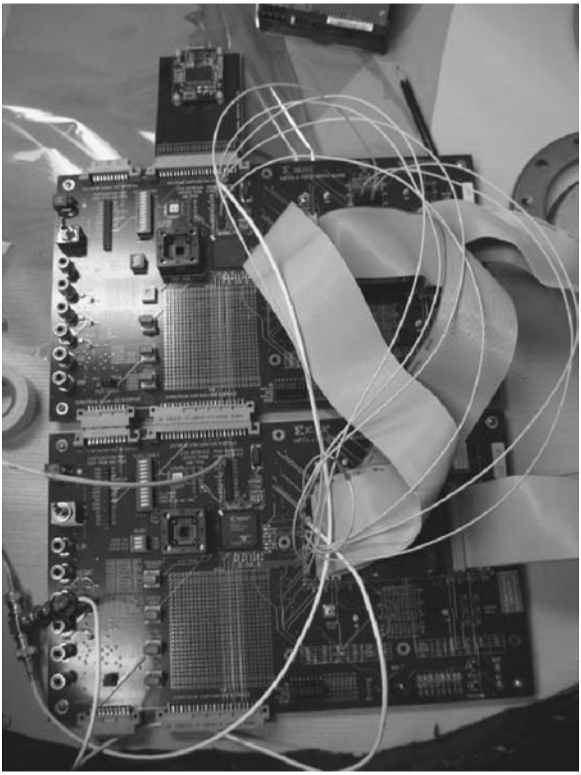


FIGURE 11.8 Fault injection and accelerator hardware test fixture for the Virtex-II. (Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science* Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

reconfiguration, which simulates the most common SEU on FPGAs. Once a fault is injected, the two boards are reset to synchronize the design and to clear the state of the device so that each fault injection trial is independent from previous trials. The designs operate in lockstep for many cycles to allow errors to propagate to the outputs. During this time period approximately 250,000 randomly generated test vectors are sent through both boards. Next, the software test fixture checks the golden board to determine if a miscompare has occurred and records the result. The software test fixture then removes the fault from the programming data through partial reconfiguration, and the boards are resynchronized to make certain the DUT returns to normal operation. Once this process is completed, the next fault can be injected.

To constrain the fault injection tests we injected the patterns that occurred most frequently in accelerator testing. Our Virtex-II heavy-ion accelerator data indicate that 99% of SEUs at 58.7 MeV-cm^2/mg can be classified as follows: one-bit upsets (79%), two-bit vertical events (6%), two-bit horizontal events (6%), three-bit corner events (4%), and four-bit squares (5%). As this data point is the highest tested heavy-ion LET, these percentages indicate a worst-case scenario for MBUs. While lower heavy-ion LETs or protons have fewer MBUs in frequency, the MBUs can be still be classified as one of the shapes from the previous list. Since these shapes represent most of the events that will occur on the Virtex-II device, simulating these patterns across the entire device will provide good coverage of Virtex-II DCEs.

11.4.1.3 Accelerator Test Methodology

The same hardware test fixture is used at the accelerator to validate the fault injection results. The software aspect of the test fixture is different, though. At the accelerator, the software test fixture performs a readback of the device's bitstream, compares the readback with a reference bitstream to determine the upset locations, records the upset locations, records the result of polling the golden for miscompares, performs a partial reconfiguration of the device to remove the faults in the upset locations, and resynchronizes the two boards through a design reset. All of these actions are performed while the part is being irradiated to simulate what would be done on orbit. Flux is deliberately kept low to minimize the number of DCEs due to uncorrelated upsets. We were recently able to conduct some preliminary accelerator testing at Indiana University Cyclotron Facility. We tested for a total fluence of 6.6 × 10^{11} protons/cm^2 in a little over two hours with two XC2V1000 parts. We also rotated the test fixture to a 45° angle to increase the MBU cross section. With this setup we average 1–3 upsets/readback cycle.

11.4.2 Fault Injection and Accelerator Test Results

The fault injection results can be found in Table 11.4. The number of DCEs are listed in two forms: (1) the raw number from fault injection; and (2) the analyzed version that represents only DCEs created by that shape. The analyzed data are in parentheses. The voting circuit column indicates whether the design votes frequently (Freq) or once off-chip (OC).

We were able to gather some preliminary accelerator data on DCEs using one design (adder tree, frequent voter) at IUCF using 200 MeV protons and the device angled at 45°. The intent of this test was to prove that DCEs would occur with both radiation-induced and fault injection methods. During a two-hour test we were able to observe 31 DCEs for a cross section of 6.6#10^{-11} ± 3.8#10^{-13} cm^2/device. With limited analysis we have been able to correlate 42% of these DCEs to known fault injection DCEs. In the future, we hope to correlate more of the DCEs to the fault injection data.

In comparison, during the same test we also observed 19 SEFIs for a cross section of 4.1#10^{-11} ± 4.9#10^{-13} cm^2/device. While the MBU cross section is several orders of magnitude larger than the SEFI cross section in the Virtex-II, the DCE cross section may be on the same order of magnitude of the SEFI cross section for this design due to the fraction of MBUs that cause DCEs. In fact, 1% of the device is affected by DCEs, and the DCE cross section is approximately 15 times smaller than the MBU cross section. Finally, the analogy to SEFIs is a useful one in how to approach DCEs. SEFIs, while possible, are not the first-order effect for these devices and can be approached as a manageable problem.

TABLE 11.4
Fault Injection Results

Circuit	Voting	All Pairs (without Overlap)	One Three-Bit Corner (without Overlap)	Four-Bit 2#2 Square (without Overlap)
Shift register	Freq	6355 (6355)	4545 (539)	9186 (489)
	OC	2185 (2185)	1364 (253)	2352 (489)
Adder tree	Freq	18733 (16843)	11264 (1116)	19464 (1783)
	OC	1166 (1104)	715 (101)	1310 (213)
Divider tree	Freq	1556 (1556)	1056 (259)	1966 (0)
	OC	1276 (1226)	767 (169)	1335 (274)
AND tree	Freq	0 (0)	2 (2)	0 (0)
	OC	0 (0)	0 (0)	0 (0)
OR tree	Freq	1784 (1784)	1645 (202)	3333 (814)
	OC	5 (5)	2 (0)	5 (1)
LFSR	Freq	4966 (4966)	2711 (297)	4709 (803)
Pseudo LFSR	Freq	26105 (26105)	18023 (2194)	31606 (4092)
	OC	44 (44)	28 (6)	58 (7)

Source: Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J,. Caffrey, M, Lundgreen, K., “Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs “. IEEE Transactions on Nuclear Science, Dec. 2007 Volume: 54 Issue: 6, 2037– 2043.

11.4.3 Discussion of Results

While we have both fault injector and accelerator results, complete test coverage is easier with fault injection. Therefore, much of our focus in this section will be on the fault injection results. Our most important result from our testing is that we were able to observe DCEs in both fault injection and accelerator testing. The discussion in this remaining section will cover circuit design and architectural characteristics that might be causing DCEs to manifest in the Virtex-II designs we tested.

11.4.3.1 DCE Characteristics

While Sterpone and Violante [7] analytically showed that SBU DCEs existed, no SBU DCEs occurred in our Virtex-I fault injection and accelerator testing. However, in the Virtex-II SBU DCEs manifested only for the adder tree and multiplier tree circuits. All of the SBU DCEs were tied to the global logic network that was providing constant zeros to the designs. Most of these SBU DCEs had similar characteristics, where a multiplexor that routes data internally in a slice from a LUT to the slice’s output signal was altered by the SEU. In these cases, further analysis showed that the slice’s output was receiving signals from the wrong multiplexor input. There was one case, though, where the multiplexor that is attached to the slice’s output signal was corrupted by an SEU. In this singular case, analysis indicated that SEU caused the output multiplexor to use a different slice’s output signal. This one case indicates

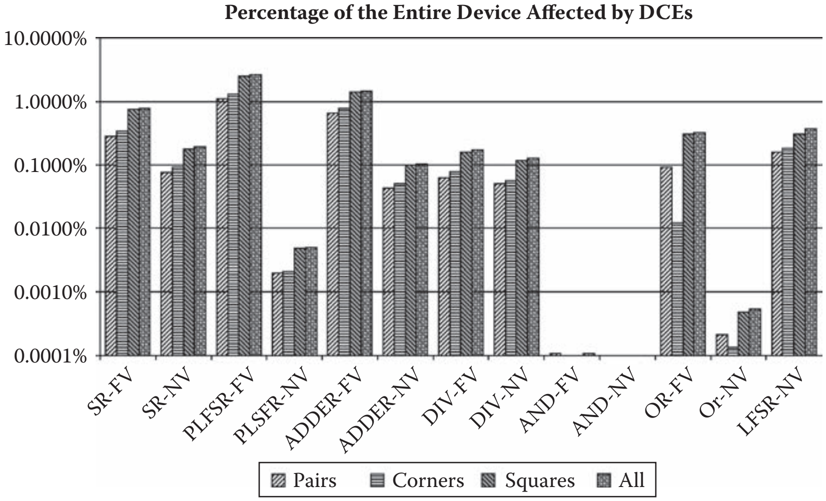


FIGURE 11.9 Percentage of the entire device affected by domain crossing errors. (Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

a possibility that a single-bit SEU could cause similar DCE corruption, but so far it appears to be rare.

Most of our test circuits were vulnerable to a number of MBU-induced DCEs, except in the case of the two AND tree implementations. We believe, as discussed later in this section, that the AND tree is particularly insensitive to errors, which caused the minimal response from the circuit implementations. As shown in Figure 11.9, of the circuits that exhibited DCEs between 0.0001% and 2.6% of the device is affected by DCEs. On average, only 0.9% of the bitstream was involved in MBU-induced DCEs for the frequent voting circuits compared with only 0.1% for the off-chip voting circuits.

11.4.3.2 Architectural Concerns

Approximately 99% of the MBU DCEs happened in the configurable logic blocks (CLBs). The remaining DCEs occurred in the input/output blocks, the global clock, and the BlockRAM interconnect (used for routing). On average 75% of CLB DCEs occurred in the CLB routing network, 22% spanned the routing network and the LUT region, and 2% occurred in the LUT region. Of the ones that occur solely in the LUT region, 80% span multiple frames, CLBs, or slices.

The CLB routing network is a concern, since it is the single largest resource type for the entire device with 53% of the configuration data in the XC2V1000. In accelerator testing, 95% of CLB SEUs and 48% of all SEUs involve the routing network. A schematic of one routing switch with its attached CLB for the Virtex-II is shown in Figure 11.10. Each CLB in the Virtex-II consists of four slices. Each slice has two LUTs, two flip-flops, and a number of bits that define the mode of the slice. Every routing switch has two main functions: (1) routing data and control signals to the

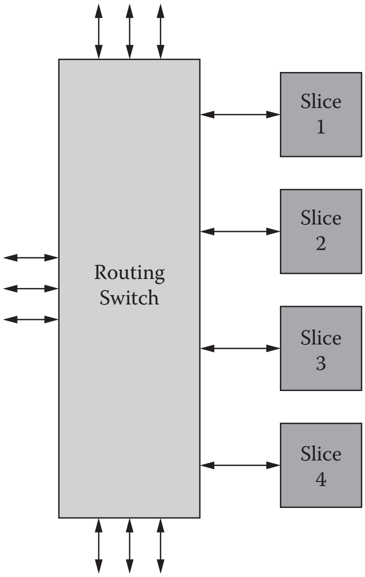


FIGURE 11.10 Routing switch and attached CLB. (Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

attached CLB; and (2) routing data and control signals to other routing switches. Even with multiple options for switch-to-switch communication, with a four-to-one ratio of slices to routing switches, routing can become congested. Furthermore, each of the four slices attached to a single switch can have four separate sets of data and control signals. Therefore, in the case of TMR circuits, it is possible that all three domains are placed in one CLB and routed through one routing switch matrix. In this scheme, routing switches becomes single points of failure.

In further analysis, we looked at what was being corrupted by the MBUs that caused DCEs. Many of the DCEs were caused by changes in global signals. We found many instances where the clocks for two different domains were switched, which could introduce subtle timing problems in the affected domains. We also found many instances where one domain's clock signal and another domain's reset signal would be switched, causing one domain's flip-flops to not be clocked and the other domain's flip-flops to be reset every clock cycle.

To reduce the impact of MBUs on the circuit, the placement and routing tools need to place the three separate domains far enough apart to not be affected by MBUs. Even simple changes, such as not allowing domains to share a CLB, could decrease the number of DCEs but might lead to poor device use.

11.4.3.3 Voting and Device Use

Our first attempt to analyze the fault injection results was to correlate the data to the device use statistics for each circuit. This analysis was fruitless, though. The worst design for DCEs, pseudo-LFSR, used approximately the same number of LUTs, slices, and PIPs as the shift register design that had 73% fewer DCEs. While the design with the smallest device use has one of the lowest numbers of DCEs, the design with no DCEs, the AND tree, uses nearly all of the slices and LUTs. Furthermore, AND tree's twin, OR tree, has all of the same use statistics with different LUT functionality and has DCEs. Therefore, the obvious device use statistics do not play a clear role in the number of DCEs.

We found increasing voting increased the number of DCEs, but a simple correlation cannot be made. Many circuits saw a large drop in DCEs in the off-chip voter implementations, such as the pseudo-LFSR circuit where the off-chip voting implementation had 1% as many DCEs as the frequent voter implementation. On the other hand, there were a few circuits that saw only a limited improvement in DCEs by voting off-chip, such as divider tree where the off-chip voting version of the circuit had 68% of the DCEs as the frequent voter implementation. Apparently, many factors influence the role of voters in these results.

The most obvious cause would be spacing, since MBUs are problematic only to closely spaced logic. The off-chip implementations all have lower device uses, which would allow the design tools to place the domains farther apart or keep large blocks of the domain together while still meeting timing requirements. Voting also causes the three domains to converge at LUTs to be voted, forcing the synthesis tool to place the domains closer together to meet timing requirements. Therefore, the voters force the domains to be proximately located. Since single-bit voters can be implemented in one LUT and each slice has two LUTs, it is possible that up to eight different voters are attached to one single routing switch. Therefore, voting frequently not only uses more resources but also causes the placement of the circuit on the device to become congested and entangled.

11.4.3.4 Design Sensitivity

The OR tree and AND tree circuits were designed to closely mimic each other in terms of layout and device use. The only difference is the logic realized in the LUTs. While they are essentially the same circuit, their DCE characteristics are different. In fact, the OR tree circuit differed from most of the other designs, as only 44% of DCEs are solely in the routing network and 53% are spanning LUTs and routing switches. Therefore, while the routing network is more suspect in most circuits, the OR tree circuit is more vulnerable to SEUs changing the routing switch and the LUT simultaneously. Likely, these errors are manifesting as multiple independent errors in the system. There are also many potential scenarios, such as the inputs to two LUTs getting stuck at zero, that would cause the OR tree to have many observable errors, while the AND tree could logically mask most of the same errors. Therefore, the OR tree is possibly more sensitive to how the LUT/routing network MBUs manifest, but the AND tree is not.

11.4.3.5 Probability of DCEs

We found that the number of DCEs increases with SEU size. The DCE space for each SEU size is composed of two parts: (1) DCEs caused by smaller events; and (2) DCEs unique to the event shape. For example, in Figure 11.11, the two-bit vertical

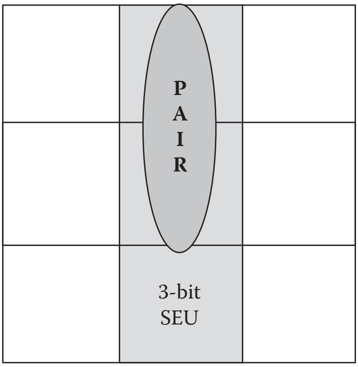


FIGURE 11.11 A three-bit MBU that overlaps with a two-bit DCE. (Reprinted with permission from Quinn, H., Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

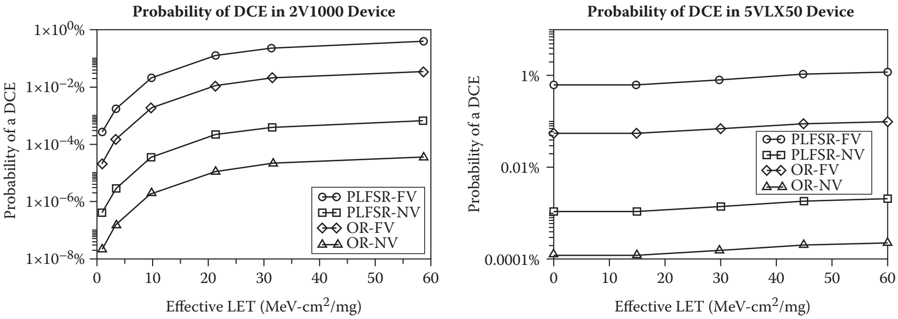
DCE is overlapped with a three-bit vertical MBU. Therefore, the event space for three-bit vertical MBU DCEs can be partitioned into DCEs caused by overlapping the two-bit vertical DCE locations and DCEs uniquely triggered by the three-bit vertical. Furthermore, the two-bit vertical DCE event space occurs twice in the three-bit vertical event space since each two-bit vertical DCE is triggered by two three-bit vertical MBUs. Therefore, the number of DCEs for a given SEU size is larger than or equal to the number of DCEs for all of the smaller SEU sizes that it overlaps. Table 11.4 indicates the number of unique DCEs for each shape in parenthesis. For three-bit and larger DCEs the event space is dominated by the smaller-sized DCEs.

As the SEU size increases the probability that a DCE is triggered goes to 1. Fortunately, while the probability of a DCE approaches 1, the probability that the event occurs goes to 0. Since the probability of a five-bit or larger MBU for the Virtex-II device is small, the probability of a DCE is dominated by the probability of one- to four-bit MBUs.

We have created a simple model for estimating the probability of a DCE occurring. The model is based on this reasoning:

$$\begin{aligned} P(DCE) &= \sum_{i=1}^{max} P(upset_i)P(DCE|upset_i) \\ &= \sum_{i=1}^{max} P(upset_i)\frac{N(DCE_i)}{C_i} \end{aligned} \tag{11.1}$$

where $P(upset_i)$ is the probability that an upset of *i* bits occurs based on accelerator data, $N(DCE_i)$ is the number of DCEs triggered by an SEU of size *i*, and C_i is the number of combinations for an SEU of size *i*. We used this model to determine the probability of a Virtex-II DCE from our results by using our normal incident



(a) 2V1000 Fault Injection

(b) 5VLX50 for a Range of Incident Angles (Projected based on data using Kr from LBNL's 10-MeV/nucleon cocktail)

FIGURE 11.12 Probability of DCEs from a heavy-ion event, where each line represents a tested design (FV = frequent voter and NV = no voter). (Reprinted with permission from Quinn, H, Morgan, K., Graham, P., Krone, J., Caffrey, M., Lundgreen, K., "Domain crossing errors: Limitations on single device triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science*, Dec. 2007 Volume: 54 Issue: 6, 2037–2043.)

heavy-ion static test data collected at Lawrence Berkeley National Laboratory for the $P(upset_{\mathrm{i}})$ values and the fault injection data for the $N(DCE_{\mathrm{i}})$ values. As shown in Figure 11.12a, the probability of a DCE in one of the Virtex-II test designs is still fairly low with a worst-case probability of 0.36%. We then extended our model to include the Virtex-5 using the Virtex-II fault injection DCE data for the $N(DCE_{\mathrm{i}})$ data and the Virtex-5 accelerator data collected at Lawrence Berkeley National Laboratory for Kr at five different angles. These projections, shown in Figure 11.12b, are for a range of angles based on data taken using Kr from LBNL's 10 MeV/nucleon cocktail with an LET of 36.4 MeV-cm^2/mg at normal incidence. This model predicts that DCEs could be up to 1.2% of all one- to four-bit events. Since we cannot account for the larger MBUs that occur on the Virtex-5 device, the real probability of a DCE is likely higher. Given these probabilities we determined the space rate for TMR defeats for a global positioning system (GPS) orbit of 20,200 km and 55 degrees. In the best-case scenario (solar maximum) only 0.6 upsets will happen per device/day, and in the worst-case scenario (peak) 3,700 upsets will occur per device/day. For the Virtex-II this translates to between 0.003 and 19 DCEs per device/day.

11.5 DETECTION OF SINGLE-BIT UPSETS, MULTIPLE-BIT UPSETS, AND DESIGN PROBLEMS

As many organizations would like to use these devices in critical space applications, methods for easily determining whether designs have been mitigated properly are necessary. Reliability modeling tools are attractive under these scenarios, because they are not hardware-dependent like fault injection. The Scalable Tool for the Analysis of Reliable Circuits (STARC) was designed to address not only the limitations of traditional reliability modeling tools in modeling user circuits for

FPGAs but also domain-specific issues with implementing TMR in FPGA circuits. In the past, this tool has been used to model both the reliability of supercomputers in the presence of neutron radiation [16] and nanoscale electronics in the presence of permanent yielding defects [17]. The main drivers for STARC are usability, computational complexity, scalability, and modularity. STARC addresses these limitations with these solutions:

- Usability: the industry-standard EDIF circuit representation is used for the input model, and input vector sets are not used. STARC was also designed to assess domain-specific problems of applying TMR to FPGA user circuits and can detect imbalances between the modules, find untriplicated logic, estimate unprotected cross section, and detect logical constant usage.
- Computational complexity: memoization of reliability values reduces recomputation of similar components, and the use of combinatorial reasonings simplifies the reliability calculation.
- Scalability: without input vectors the state space scales linearly with the circuit size.
- Modularity: the architectural and fault models that provide the basis of the reliability calculation are inputs to STARC and can be replaced with user-specified architectures and fault models.

By using the EDIF circuit representation, the designer can assess the reliability of a circuit during the design process, even if the design is not complete, the design does not work, or the hardware is not available. Without the use of input vector sets reliability is determined through the probability of device or input failure and is not dependent on specific input data sets. Without input data sets, the reliability of components is determined by type, such as a two-bit adder, and can be memoized for reuse. In this manner, large-scale circuits are analyzed in a fraction of the time and memory required by traditional approaches, making design exploration more worthwhile.

As STARC can estimate the hardness assurance of FPGA user circuits within minutes, STARC can also be used for designers facing area and resource constraints. Under these circumstances, it is possible to generate a range of designs in BL-TMR with different balances of unprotected cross sections and resource use. In this manner, STARC can help designers choose among a range of possible design choices by quantifying the remaining unprotected cross section for each.

There are a few disadvantages to this approach. First, since EDIF does not contain information about the routing, information regarding placement and routing is absent from the calculation. As routing can have a large impact on the protected and unprotected cross sections, the routing cross section is estimated statistically based on an analysis we did of several designs using JBits [18]. The point of the statistical model is to provide a good estimate of the single-bit cross section, as the only way to fix unprotected configuration bits in the routing is to mitigate the unprotected logic. Furthermore, currently there is no way to assess placement-related issues, such as MBU-induced TMR defeats. We are currently working on a solution for this limitation for designs that have completed the design flow. Second, without input vector sets, logic masking cannot be taken into account, and STARC estimates the worst-case failure rate. While this value may be lower than the value determined by other tools [19], STARC provides a useful lower bound on the circuit's reliability.

11.5.1 Related Work

Traditionally, circuit reliability has been determined using purely analytical approaches [20,21] or techniques that model Boolean networks as probabilistic systems [22-25]. These modeling techniques represent circuits as probabilistic transfer matrices, stochastic Petri-nets, Markov chains, or Bayesian networks. The combinatorics-based analytical approaches have been found to be error-prone and computationally complex for the analysis of large designs. Similarly, a number of limitations have been identified for many modeling-based approaches. First of all, model creation and input data sets greatly increase the time commitment of using these tools. Transforming circuits into intermediate probabilistic system models is an additional, computationally complex task. Within an analytical tool a state space is generated from the input model and input data vector set. The state space encodes all of the possible failure states in the circuit and grows exponentially with circuit size. The exception to these problems is the SETRA tool [26] that directly addresses the state space issues as well as automated model generation. Attempts at reducing computational complexity through circuit partitioning and hierarchical modeling of large circuits requires additional modeling effort. These limitations lead to the STARC tool, which uses EDIF circuit representations, no input data vectors, and simpler combinatorial reasoning to decrease the time commitment for the designer and reduce computational complexity in the tool.

Besides these differences between STARC and traditional reliability analysis tools, the tool methodologies differ greatly. Two distinct methods [27] can be used to analyze the reliability of circuits: (1) generalized; or (2) instance-based. The *generalized* approach entails the combinatorial modeling of circuits without considering specific failure distributions of the inputs, gates, and interconnects. A circuit's output's probability distribution is computed through combinatorics under the assumption that each gate can fail independently. Thus, the reliability is evaluated in stages using conditional probabilities. Generalized techniques to compute the reliability of large circuits require complex combinatorial reasonings. Reusing subcircuit analysis to reduce the combinatorial complexity in the analysis of a larger circuit is difficult. Since specific input probability distributions are not considered during analysis, the generalized approach determines either the circuit's lower or upper bound on reliability.

Several *instance-based* methodologies have been proposed recently [19,24,25]. Instance-based reliability circuit analysis uses probability distributions on the primary inputs as well as gate and interconnect failure probabilities to develop an instance of the circuit. Each instance is then transformed into probabilistic circuit models. This method computes the exact reliability of the circuit for the input distribution. The main drawback of these tools is that several instances of the circuit need to be analyzed to predict performance trends, which can be computationally expensive. Therefore, the input vector set needs to be limited to bound the computational cost yet to provide enough intuition on the circuit's reliability.

The STARC methodology is a hybrid of the two approaches. STARC, as with other generalized approaches, is independent of specific input vectors and their probability distributions yet uses specific gate distribution instances. Hence, this approach avoids the complex combinatorial reasonings that cause bottlenecks in generalized approaches and also bounds the computational complexity that affects instance-based methods. STARC computes a lower bound on reliability. When we compared STARC with a purely instance-based approach based on PRISM [25,28], the results of our comparison of STARC and PRISM were favorable. We tested four different designs with two probability-of-failure models based on estimated yield defects on a Dell Linux machine with 4 GB of RAM and dual 3.4 GHz Xeon microprocessors. We then compared the calculated reliability values and execution times. The ratio of the two calculated reliability values indicated that STARC was within three to seven digits of significance to PRISM. STARC also executes faster that PRISM and for several designs was more than nine times faster.

It should be noted that a reliability analysis tool, called SEUper_fast [29], designed by Boeing in the 1990s, uses many of the same reasonings as probability transfer matrix tools. This tool approached the problem far more generally than STARC and was hindered by solving a much more complex reliability equation than STARC uses. While currently not as generally applicable as SEUper_fast, we believe we will be able to generalize this technique to different problems without having to employ the more complex reliability analysis technique.

11.5.2 STARC Overview

In this section, we will provide an overview of STARC. The reliability of the circuit is determined from dependency graphs of the circuit that are created during a hierarchical exploration of the circuit. By using the EDIF circuit representation, the hierarchy in the circuit should be preserved. Since designers tend to create complex circuits by creating less complex components or subcircuits, maintaining this structure can be very useful in calculating the reliability. In particular, STARC can determine the reliability of a circuit hierarchically. STARC navigates through the layers of the circuit hierarchy to determine the smallest circuit component that needs to have its reliability calculated. Once an entire layer of the circuit hierarchy is completed, these values can be used to determine the reliability of the next higher layer. This hierarchical nature allows circuits to be examined at the highest level of abstraction or the most minute level of detail. STARC automatically determines the appropriate level of the hierarchy that needs to be explored.

Since input vectors are not used in the reliability calculation, the reliability is determined by component type. For example, one component type might be a two-bit adder. The first time a two-bit adder is found during hierarchical exploration these three steps are executed:

1. A dependency graph is determined.
2. The reliability of the dependency graph is calculated.
3. The reliability value of the dependency graph is memoized.

The next time another two-bit adder is found in a design, the memoized value is used, and the first two steps of the process are eliminated. It is in this way the state space of the circuit grows with circuit size, since the state space is limited to the unique number of components in the circuit. Even if a circuit has very little component reuse, the state space will never grow larger than the number of components in the circuit. Since the size of the state space has a first-order effect on the speed of computation, STARC is able to analyze the reliability of a circuit in polynomial time instead of the exponential time necessary for most traditional reliability tools. Therefore, STARC should be able to compute the reliability of circuits with thousands of components in the design in a matter of minutes.

As stated already, during hierarchical exploration dependency graphs are determined for each unique component. For maximum reuse, dependency graphs for each primary output at each level of the hierarchy are determined. These dependency graphs indicate all of the components that exist in the path between a particular output and the reachable inputs. Since not all logic or inputs are reachable from every output, this technique removes unrelated logic from the dependency graph and, hence, the reliability calculation.

Once the dependency graph for an output is determined, the reliability can be calculated. In unmitigated designs, the cross section is the total area of the dependency graph:

$$A(O) = \sum_{i=0}^{m} A(C_i) \tag{11.2}$$

where $A(X)$ is the sensitive area of X (where X is either a wire or a component), and $C = \{C_0, \ldots, C_\mathrm{m}\}$ is the set of components that can be reached from output wire O. STARC also applies a modular approach to the fault model and the architectural model. Since reliability is determined hierarchically in STARC, the only devices that need to be precalculated are the primitives for the given architecture. Figure 11.13 shows our methodology for library characterization. The primitives for a hardware platform are defined in an architectural model. Fault models for transient and permanent defects are combined with the architectural models to create the characterized primitive library. Traditional probability of failure equations are also available to calculate the reliability of defect-based architecture models. Our automation framework is designed so that users can define primitive libraries for their own architectural models or use our models for basic logic and the Xilinx architecture. To be used in our methodology, user-defined libraries have to be characterized for specific fault models to define their reliability.

In this manner, STARC was designed to be architecturally independent. While this chapter focuses on reliability as it relates to Xilinx FPGAs, STARC is modular in nature and the Xilinx cross section model is an input to this system. The tool has also been used for probability of failure calculations for nanoscale electronics based on yield estimates. In the future, we would like to expand into models for probability

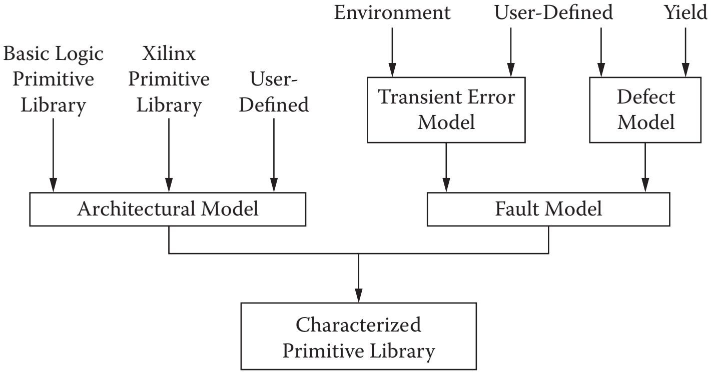


FIGURE 11.13 Library characterization. (Reprinted with permission from Quinn, H., Graham, P., Pratt, B., "An automated approach to estimating hardness assurance issues in triple-modular redundancy circuits in Xilinx FPGAs." *IEEE Transactions on Nuclear Science,* Volume: 55 Issue:6, 3070–3076.)

of failure and cross section models for structured ASICs, as these devices are frequently being used in space-based systems as well.

Finally, STARC was also designed to help designers find problems in the application of TMR. For mitigated circuits, the sensitive area is confined to the part of the design that is not triplicated, as triplication will mask errors as long as there is one voter for each redundant module. STARC also checks to make certain the modules have equivalent components. Any logical elements that might be shared by two or more TMR domains are considered unprotected cross section, even if the elements reside within one of the modules. STARC also checks to make sure the feedback loops are properly triplicated and cut. If persistent cross section is found, a warning is displayed to inform the designer that a particular component has not had TMR applied correctly.

Recently, we have been adding support for placement-related information in STARC to provide DCE predictions. For designers who are further along in the design process, it is possible to get placement-related information from the Xilinx Design Language (XDL) representation of the circuit. Like EDIF, XDL provides a human-readable circuit representation. Unlike EDIF, the component names from the circuit are slightly obscured. In our initial attempts, though, we have been able to map the XDL circuit representation onto the EDIF circuit representation. Because our analysis showed that the most common problems with DCEs were caused by domains sharing the same CLB, our initial attempt also includes the ability to gauge how many CLBs are populated with more than one domain.

In all of these cases, STARC provides warnings and information about the design to the designer. The output of the tool provides the designer a list of subcircuits that are untriplicated, a quantity for the unprotected cross section, and warnings about potential single points of failures from functionally nonequivalent modules and logical constants. Since EDIF is tightly coupled to the circuit design, the designer should be able to directly use STARC's output to find and fix the design flaws in the user circuit.

11.5.3 Case Study: Tradespace of Reliability Issues under Area Constraints

In this section we present a case study of two image processing algorithms that use STARC to explore the tradespace of reliability issues under an area-constrained design process. The two image processing algorithms we examined are an edge detection algorithm and a noise filtering algorithm. The edge detection algorithm uses the Sobel convolution masks [30] as the computational basis. These convolution masks are well matched to FPGA implementation, since the multiplication can be reduced to shifts. The noise filtering algorithm breaks the image into a series of small windows. The pixel in the center of the window is replaced with the minimum pixel value in the window. Both of these circuits are feed forward and, therefore, do not have error persistence issues. Since both algorithms use nine eight-bit pixels as input, the algorithms both use the same data input circuit.

Several implementations of these circuits were developed: without TMR, with full TMR, and two partial TMR approaches. To avoid design issues with applying TMR, BL-TMR was used. It should be noted that STARC has been modified to automatically recognize designs that have been mitigated through BL-TMR and the Xilinx TMRTool. Logical constants were also extracted to input pins. For the partial TMR approaches, we had BL-TMR triplicate the logic in both implementations for both algorithms and varied how the input and output signals were handled. In the partial TMR 1 implementations we had BL-TMR not triplicate any input or output signals, and in the partial TMR 2 implementations we had BL-TMR triplicate only the reset, logical constant, and clock input signals.

STARC was used to determine the unprotected cross section of all of the implementations, as shown in Table 11.5. The first thing to note from these values is that applying TMR to just the design's logic (partial TMR 1) provided little improvement for the noise filter and actually increased the cross section for the edge detection algorithm. When we looked through the STARC results we found the large unprotected cross section in the partial TMR 1 versions were due to the unmitigated signals. As Table 11.5 shows, all of the unprotected cross sections for these implementations are in the routing network, indicating that the logic was properly triplicated. Since the triplicated logic has three times as many flip-flops, the untriplicated clock, reset, and logical constant trees now have to route to three times as many locations. In a heavily pipelined design, like the edge detection algorithm, this decision was disastrous. When we went back to BL-TMR and chose to triplicate the logic and the global signals, the unprotected cross section for both designs was 99.8% smaller than the unprotected cross section in the unmitigated design. When full TMR is applied to both algorithms, there was no unprotected cross section.

Finally, STARC was able to find the hardness assurance issues that existed in the implementations without TMR and with partial TMR. In both algorithms the implementations without TMR used the device-provided logical zeros and STARC correctly identified this as a potential problem. Also, the implementation of the two algorithms with partial TMR had input signals, a voter, and input/output registers that were not triplicated. STARC was able to find these untriplicated signals and logic, to report them, and to properly calculate the cross section for them.

TABLE 11.5
STARC Results for Two Image Processing Algorithms

Design	Implementation	Total Unprotected Cross Section (Bits)	Unprotected Logic (Bits)	Unprotected Routing (Bits)	Number of Components	Time to Calculate (sec)
Edge detection	No TMR	15,418	3,641	11,777	1,356	56
	Partial TMR (1)	21,800	19	21,781	3,787	426
	Partial TMR (2)	24	16	8	3,793	401
	Full TMR	0	0	0	3,799	230
Noise filter	No TMR	14,914	4,522	10,392	1,603	95
	Partial TMR (1)	14,332	19	14,313	4,273	785
	Partial TMR (2)	24	16	8	4,279	565
	Full TMR	0	0	0	4,285	309

Source: Reprinted with permission from Quinn, H. Graham, P., Pratt, B., “An automated approach to estimating hardness assurance issues in triple-modular redundancy circuits in Xilinx FPGAs.” IEEE Transactions on Nuclear Science, Volume: 55 Issue: 6, 3070 – 3076.

TABLE 11.6
STARC Validation Results for the Unmitigated Implementation of Two Image Processing Algorithms

Design	Total Unprotected Cross Section (Bits)	Unprotected Logic (Bits)	Unprotected Routing (Bits)
Edge detection	14,461	2,291	12,170
Noise filter	9,507	1,462	8,045

Source: Reprinted with permission from Quinn, H., Graham, P., Pratt, B., “An automated approach to astimating hardness assurance issues in triple-modular redundancy circuits in Xilinx FPGAs.” IEEE Transactions on Nuclear Science, Volume: 55 Issue: 6, 3070–3076.

We have recently begun validation of the STARC tool. Table 11.6 shows some results from fault injection of the unmitigated implementations of the two image processing algorithms. While the edge detection algorithm is within 93.8% of the STARC-predicted cross section, the noise filtering algorithm is not as close at 63.8%. When looking at the numbers more closely, for both designs the routing estimates look reasonable, but the logic is overestimated in both cases. We believe that the reason there is such a gap in the logic values is due to logical masking on the fault injection hardware. In particular, we found that the outputs of the edge detection algorithm are much more sensitive to data changes than the noise filter. In examining the execution times we found that the tool was able to complete on average 12 components/second. Note that the execution time tripled from the unmitigated implementations to the mitigated implementations. As BL-TMR flattens the circuit hierarchy while applying TMR, the entire circuit’s state space must be analyzed to determine the reliability of the circuit.

11.6 CONCLUSIONS

In this chapter, we provided an overview of a number of topics regarding assuring the robustness of TMR-protected user circuits in Xilinx FPGAs. We have presented a number of hardness assurance, including redundant modules that share logic, the inability to fully triplicate designs, device-provided logical constants, and domain crossing errors. Our studies into domain crossing errors show that the CLB routing network has proven to be fragile in TMR applications with highly used and congested routing scenarios. We also introduced a tool called the Scalable Tool for the Analysis of Reliable Circuits that automates the process for identifying hardness assurance issues with TMR-protected circuits for Xilinx FPGAs as well as estimating their unprotected SEU cross-sections. As an illustration, we used STARC to analyze four implementations of two different image processing algorithms with different approaches to TMR. These results showed that full TMR provided a 100% reduction in cross section, and that triplicating just the logic, clock, and reset could reduce the unprotected cross section by 99.8%.

REFERENCES

1. E. Fuller, M. Caffrey, P. Blain, C. Carmichael, N. Khalsa, and A. Salazar, "Radiation test results of the Virtex FPGA and ZBT SRAM for space based reconfigurable computing," in *Proceedings of the Military and Aerospace Programmable Logic Devices International Conference (MAPLD),* Laurel, MD, September 1999.
2. G. M. Swift, "Virtex-II static SEU characterization," Xilinx Radiation Test Consortium, Tech. Rep. 1, 2004.
3. G. Allen, G. Swift, and C. Carmichael, "Virtex-4VQ static SEU characterization summary," Xilinx Radiation Test Consortium, Tech. Rep. 1, 2008.
4. C. Carmichael, "Triple Module Redundancy Design Techniques for Virtex FPGAs," Xilinx Corporation, Tech. Rep., November 1, 2001, XAPP197 (v1.0).
5. F. Lima, C. Carmichael, J. Fabula, R. Padovani, and R. Reis, "A fault injection analysis of Virtex FPGA TMR design methodology," in *Proceedings of the 6th European Conference on Radiation and Its Effects on Components and Systems (RADECS 2001),* 2001.
6. N. Rollins, M. Wirthlin, M. Caffrey, and P. Graham, "Evaluating TMR techniques in the presence of single event upsets," in *Proceedings of the 6th Annual International Conference on Military and Aerospace Programmable Logic Devices (MAPLD),* Washington, DC: NASA Office of Logic Design, AIAA, September 2003, p. P63.
7. L. Sterpone and M. Violante, "A new analytical approach to estimate the effects of SEUs in TMR architectures implemented through SRAM-based FPGAs," *IEEE Transactions on Nuclear Science,* Vol. 52, No. 6, pp. 2217–2223, 2005.
8. H. Quinn, K. Morgan, P. Graham, J. Krone, M. Caffrey, and K. Lundgreen, "Domain crossing errors: limitations on single device triple-modular redundancy circuits in Xilinx FPGAs," *IEEE Transactions on Nuclear Science,* Vol. 54, No. 6, pp. 2037–2043, 2007.
9. H. Quinn, P. Graham, J. Krone, M. Caffrey, and S. Rezgui, "Radiation-induced multi-bit upsets in SRAM-based FPGAs," *IEEE Transactions on Nuclear Science,* Vol. 52, No. 6, pp. 2455–2461, December 2005.
10. G. Gasiot, D. Giot, and P. Roche, "Multiple cell upsets as the key contribution to the total SER of 65 nm CMOs SRAMs and its dependence on well engineering," *IEEE Transactions on Nuclear Science,* Vol. 54, No. 6, pp. 2468–2473, 2007. Available at: http://dx.doi.org/10.1109/TNS.2007.908147
11. H. Quinn, K. Morgan, P. Graham, J. Krone, and M. Caffrey, "Static proton and heavy ion testing of the Xilinx Virtex-5 device," in *Proceedings of Data Workshop for Nuclear and Space Radiation Effects Conference,* July 2007.
12. Y. Tosaka, H. Ehara, M. Igeta, T. Uemura, H. Oka, N. Matsuoka, et al., "Comprehensive study of soft errors in advanced CMOs circuits with 90/130 nm technology," in *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International,* 2004, pp. 941–944. Available at: http://dx.doi.org/10.1109/IEDM.2004.1419339
13. K. Morgan, M. Caffrey, P. Graham, E. Johnson, B. Pratt, and M. Wirthlin, "SEU-induced persistent error propagation in FPGAs," *IEEE Transactions on Nuclear Science,* Vol. 52, No. 6, pp. 2438–2445, 2005.
14. "Xilinx TMRTool User Guide," Available at: http://www.xilinx.com/products/milaero/ug156.pdf
15. M. French, M. Wirthlin, and P. Graham, "Reducing power consumption of radiation mitigated designs for FPGAs," in *Proceedings of the 9th Annual International Conference on Military and Aerospace Programmable Logic Devices (MAPLD),* September 2006.
16. H. Quinn, D. Bhaduri, C. Teuscher, P. Graham, and M. Gohkale, "The STAR systems toolset for analyzing reconfigurable system cross-section," in *Military and Aerospace Programmable Logic Devices,* 2006, p. 162.

17. H. Quinn, D. Bhaduri, C. Teuscher, P. Graham, and M. Gohkale, “The STARC truth: analyzing reconfigurable supercomputing reliability,” in *Field-Programmable Custom Computing Machines,* 2005.
18. http://www.xilinx.com/labs/projects/jbits/
19. D. Bhaduri and S. Shukla, “NANOLAB—a tool for evaluating reliability of defect-tolerant nanoarchitectures,” *IEEE Transactions on Nanotechnology,* Vol. 4, No. 4, pp. 381–394, 2005.
20. J. A. Abraham, “A combinatorial solution to the reliability of interwoven redundant logic networks,” *IEEE Transactions on Computers,* Vol. 24, No. 6, pp. 578–584, May 1975.
21. J. A. Abraham and D.P. Siewiorek, “An algorithm for the accurate reliability evaluation of triple modular redundancy networks,” *IEEE Transactions on Computers,* Vol. 23, No. 7, pp. 682–692, July 1974.
22. C. Hirel, R. Sahner, X. Zang, and K. Trivedi, “Reliability and performability using SHARPE 2000,” in *11th Int’l Conf. on Computer Performance Evaluation: Modeling Techniques and Tools,* Vol. 1786, 2000, pp. 345–349.
23. F. V Jensen, *Bayesian Networks and Decision Graphs.* New York: Springer-Verlag, 2001.
24. S. Krishnaswamy, G.F. Viamontes, I.L. Markov, and J.P. Hayes, “Accurate reliability evaluation and enhancement via probabilistic transfer matrices,” in *Design, Automation and Test in Europe (DATE’05),* Vol. 1. New York: ACM Press, 2005, pp. 282–287.
25. G. Norman, D. Parker, M. Kwiatkowska, and S. Shukla, “Evaluating the reliability of NAND multiplexing with PRISM,” *IEEE Transactions on CAD,* Vol. 24, No. 10, pp. 1629–1637, 2005.
26. D. Bhaduri, S.K. Shukla, P.S. Graham, and M.B. Gokhale, “Reliability analysis of large circuits using scalable techniques and tools,” *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications,* Vol. 54, No. 11, pp. 2447–2460, November 2007.
27. D. Bhaduri, S.K. Shukla, P. Graham, and M. Gokhale, “Comparing reliability-redundancy trade-offs for two von Neumann multiplexing architectures,” *IEEE Transactions on Nanotechnology,* 2006.
28. D. Bhaduri and S. Shukla, “NANOPRISM: a tool for evaluating granularity vs. reliability trade-offs in nano-architectures,” in *14th GLSVLSI,* Boston, MA: ACM, April 2004, pp. 109–112.
29. M. Baze, S. Buchner, W. Bartholet, and T. Dao, “An SEU analysis approach for error propagation in digital VLSI CMOS ASICs,” *IEEE Transactions on Nuclear Science,* Vol. 42, No. 6, pp. 1863–1869, December 1995.
30. A.K. Jain, *Fundamentals of Digital Image Processing,* Prentice Hall Information and System Sciences Series, 1989.

12 SEU/SET Tolerant Phase-Locked Loops

Robert L. Shuler, Jr.

CONTENTS

12.1 INTRODUCTION

The phase-locked loop (PLL) is an old and widely used circuit for frequency and phase demodulation, carrier and clock recovery, and frequency synthesis [1]. Its implementations range from discrete components to fully integrated circuits and even to firmware or software. Often the PLL is a highly critical component of a system, as for example when it is used to derive the on-chip clock, but as of this writing no definitive single-event upset (SET)/single-event transient (SET) tolerant PLL circuit has been described. This chapter hopes to rectify that situation, at least in regard to PLLs that are used to generate clocks.

Older literature on fault-tolerant PLLs deals with detection of a hard failure, which is recovered by replacement, repair, or manual restart of discrete component systems [2,3]. Several patents exist along these lines (6349391, 6272647, and 7089442). A newer approach is to harden the parts of a PLL system, to one degree or another, such as by using a voltage-based charge pump [4,5] or a voted triple modular redundant (TMR) voltage-controlled oscillator (VCO) [6]. A more comprehensive approach is to harden by triplication and voting (TMR) all the digital pieces (primarily the divider) of a frequency synthesis PLL [7], but this still leaves room for errors in the VCO and the loop filter.

Instead of hardening or voting pieces of a system, such as a frequency synthesis system (i.e., clock multiplier), we will show how the entire system can be voted. There are two main ways of doing this, each with advantages and drawbacks. We will show how each has advantages in certain areas, depending on the lock acquisition and tracking characteristics of the PLL. Because of this dependency on PLL characteristics, we will briefly revisit the theory of PLLs. But first we will describe the characteristics of voters and their correct application, as some literature does not follow the voting procedure that guarantees elimination of errors. Additionally, we will find that voting clocks is a bit trickier than voting data where an infallible clock is assumed. It is our job here to produce (or recover) that assumed infallible clock!

12.2 VOTING ASYNCHRONOUS SIGNALS

When voting synchronous signals, data are latched according to a clock edge and can be unambiguously voted either before or after the clock edge. There are two common ways of doing this, according to what is to be protected. Figure 12.1 shows the two methods.

On the left we have a triplicated functional unit (usually memory, but could be anything). A single voter removes errors introduced within any one of the units but does not protect against errors in the voter itself or in anything prior to the triplicated units. On the right everything is triplicated, including the voters, so all errors are removed. As long as two of the three strings have a correct result, processing will continue correctly.

When voting asynchronous signals, those such as clocks that are not synchronized by some other signal, it is possible to have two correct signals and still get an incorrect result. Suppose, for example, that "correct" means a signal of a given frequency, such as the output of a clock multiplier. In Figure 12.2, signals X and Y are correct, and signal Z is in error. But because X and Y are not perfectly in phase, Z is able to influence the vote this way and that, producing anomalous results for the voting result (***maj*** – top signal), which is of incorrect frequency and has some transitions that are much too fast. These extra transitions will violate timing constraints and produce unpredictable errors if they occur on a clock signal.

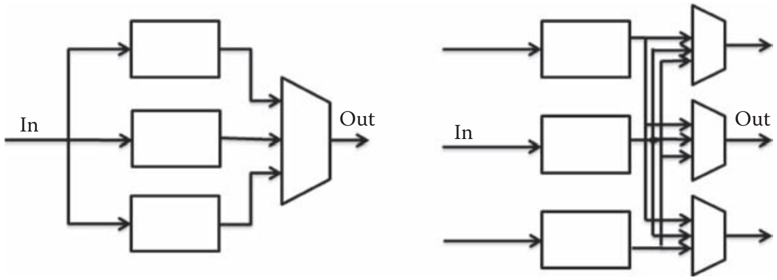


FIGURE 12.1 Single versus triple voters.

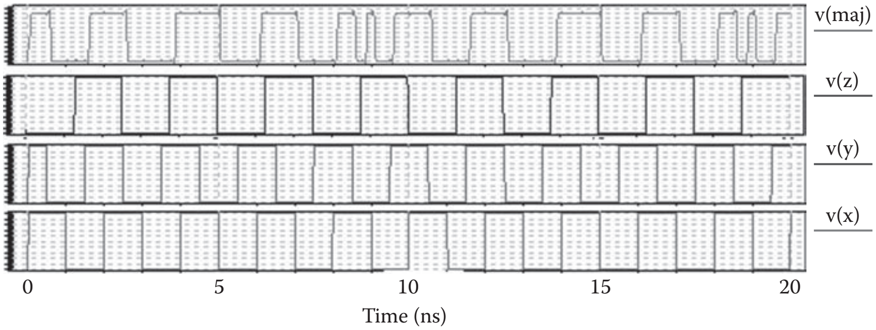


FIGURE 12.2 Phase-induced voting error.

While this example has been exaggerated for illustration, such an error can occur even for a small phase difference between the two correct signals. To guarantee that extra-transition errors (phase-induced voting error) will not occur, the phase difference between the two correct signals has to be smaller than the minimum pulse width to which the voters will respond!

The voting guidelines we have so far may be summarized as (1) use a triple voter configuration to protect against errors even in the voters, and (2) design your PLL so that redundant units will operate closely enough in phase that phase-induced voting errors will be avoided, as for example during the period when one unit is recovering from an SEU/SET and running temporarily at a different frequency or with glitches in its output. One more guideline is needed. What do you do if ultimately you wish to get one reliable output, such as one system master clock? In this case you must eventually rely on a single voter (though you can still use triple voters internal to your PLL system). The only type of voter that does not have a single SEU/SET susceptible point of failure is a heavily overdriven force voter, or conflict voter, that uses many gates to drive a single node. These gates should be spread out so that one single event will not strike several of them, and they should be driven from independent sources, such as a triple of prevoters. A rather elaborate example I have used is shown in Figure 12.3.

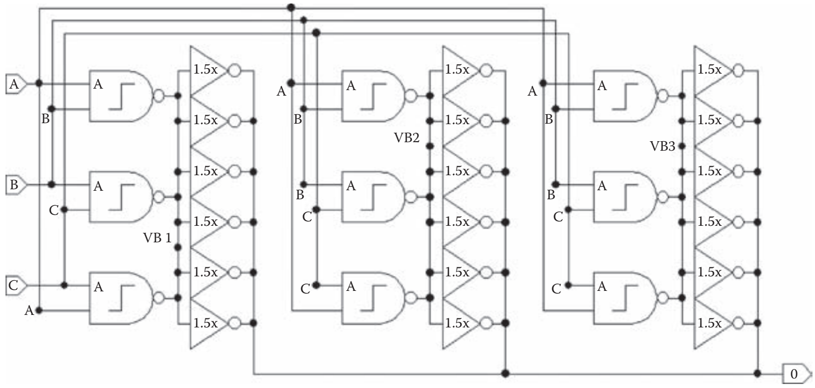


FIGURE 12.3 Force voter for consolidating triple to single string.

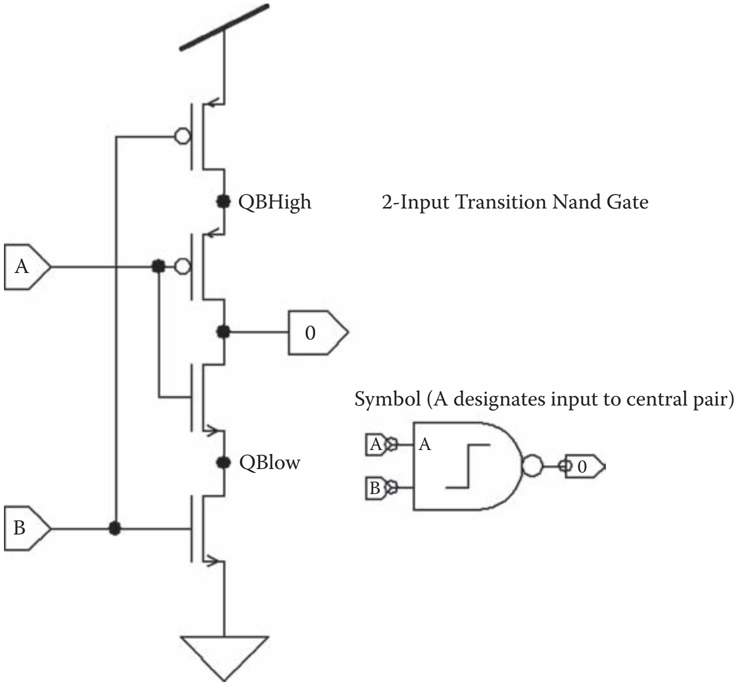

QBHigh
2-Input Transition Nand Gate
A
0
Symbol (A designates input to central pair)
A
A
B
0
QBlow
B

FIGURE 12.4 Transition nAnd gate.

The special symbol in Figure 12.3 is not a NAND gate. It is a transition nAnd gate (TAG) [8,9], also sometimes called a guard gate [10,11], broadly useful in radiation-hardening-by-design (RHBD) technology. The circuit for it is shown in Figure 12.4.

12.3 STABLE PLLS THAT MINIMIZE PHASE-INDUCED VOTING ERROR

A PLL is a feedback circuit that measures the phase of a signal compared with some reference and attempts to correct the phase of a local oscillator to match the reference. The local oscillator is frequently a VCO but may also be a numerically controlled oscillator, and in signal processing applications the whole PLL may be implemented mathematically in firmware rather than using actual components. But when a PLL must operate very fast and produce the clock on which digital logic depends, there is no alternative but to implement it directly in hardware.

An analysis of charge-pump PLLs by Gardner [12] points out several stability issues. First, the continuous-time approximation used for the analysis is not valid if the PLL loop bandwidth is high, and this introduces stability problems. For the fastest recovery from SEU/SET tolerance we will want the highest bandwidth practical. For clock generator PLL applications, it turns out it is very practical to increase loop bandwidth. For frequency synthesizer PLL applications, used to generate channel frequencies for communications systems, higher loop bandwidth

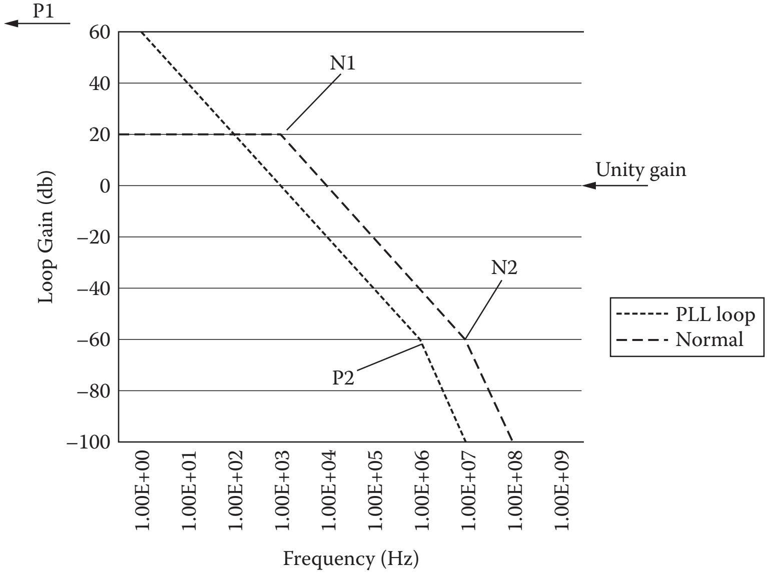


FIGURE 12.5 PLL versus normal feedback loop Bode plot.

is not so practical. We will examine redundant PLL architectures appropriate to both situations.

Second, frequency synthesis PLLs need to have second- or third- or higher-order loop filters to reduce ripple on the VCO control voltage, V_{ctl}, caused by charge-pump operation and by the workings of fractional-N and sigma-delta frequency dividers and associated compensation systems. Whereas second-order analog PLLs are unconditionally stable, a second-order charge-pump-based PLL is a sampled data system and is unstable with high loop gain. The characteristics of VCOs currently preferred for high speed PLLs, which will be described in the next section, virtually guarantee excessively high gain unless the designer takes careful steps to ensure otherwise. With a third-order loop the situation is even worse. We will consider the most effective ways to manage loop gain.

Because of the additional pressure toward instability due to the requirements of an SEU/SET tolerant PLL, we will briefly review PLL stability and introduce nonlinear circuit considerations.

For any feedback loop to be stable and not oscillate, the feedback must be negative at all frequencies for which the feedback loop gain is equal to or greater than unity. PLLs are rarely completely stable. Their residual instability, the reasons for which we will explore herein, shows up as continual oscillation in frequency, or phase, of the local oscillator. This residual instability is a source of excess phase jitter and can impair the ability to synchronize redundant PLLs in a fault tolerant circuit, and the design techniques often used to combat it (lower bandwidth loop filters) can slow and interfere with recovery from an SET or SEU.

Stability of a feedback circuit is often understood by means of a Bode plot, such as in Figure 12.5. Loop gain is asymptotically plotted as a function of frequency.

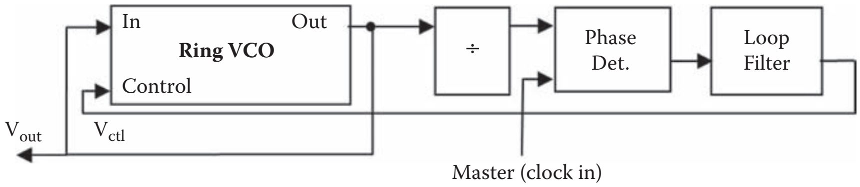


FIGURE 12.6 PLL block diagram with divider for frequency synthesis.

In the case of the PLL, this is the frequency with which the loop control voltage varies, not the frequency of oscillation of the VCO. The two are related of course by the transfer function of the VCO, which in the circuits we will be using is highly nonlinear.

First consider a normal operational amp (op amp) feedback loop, with loop gain represented by the dashed line. It is shown with two example poles in the loop response, N1 and N2. A 90 degree phase shift is associated with each pole. As long as no more than one pole is above the unity gain line, the circuit should be stable. If not naturally the case, this is often arranged by use of a Miller effect equalization capacitor to move one pole lower in frequency than any others.

A PLL, on the other hand, does not control the same thing it measures. It measures phase, but through the VCO it controls frequency. Phase is the integral of frequency. Therefore, every PLL has an unavoidable pole at zero frequency. You cannot move any other pole below it! This is illustrated by the dotted line, with poles P1 (infinitely off to the left on this logarithmic frequency scale plot) and P2. We have optimistically shown P2 below unity gain, but that is often difficult to arrange.

Unfortunately, determining gain for a highly nonlinear circuit is problematic. The best way is usually to run a transient simulation in Spice to see if the loop is stable. The point of this discussion is that one must be careful in designing the loop filter in a PLL. Figure 12.6 shows a high-level block diagram level view of a PLL clock multiplier.

Phase or phase-frequency detectors usually output a series of pulses, not a nicely behaved analog signal, so the loop filter must be introduced to smooth this signal. Otherwise, the VCO would vary between some very high and very low frequency, and its output would be unusable. The loop filter determines the bandwidth of the PLL [1]. Here we get conflicting requirements. For recovery from SETs and SEUs, the loop filter should have a high bandwidth so that the PLL performs like a tracking PLL and rapidly resynchronizes after an upset. But in clock or synthesis applications, a narrow frequency range is desired, meaning a low filter bandwidth. This is so that the VCO deviation over the counter cycle of the frequency divider will be small. In a fractional-N or sigma-delta PLL, the VCO deviation over several counter cycles must be small. A low filter bandwidth slows acquisition, either initially or after an upset. Additionally, the loop filter forms a second pole in the loop gain, and a low bandwidth loop filter moves that pole to the left on a Bode plot, toward a position of higher gain on the PLL's unavoidable pole-at-zero loop response characteristic.

Sometimes PLLs are designed with a tight (low bandwidth) loop for normal operation, and a separate means for initial "acquisition" of the target signal (the master clock, in the case of Figure 12.6). We do not recommend this for SEU/SET tolerant PLLs, because an SET or SEU can cause an unplanned reacquisition at any time. Having a separate means for acquisition might be possible but is difficult to verify for every case.

The requirement for unplanned reacquisition at any time implies that we should use a phase-frequency detector (PFD), not a pure phase detector (e.g., XOR gate, traditional frequency multiplier). Phase-only detectors often are tricky to design for initial or reacquisition, especially when using tight loop filters. The reason is that without intrinsic frequency information, the phase-only detector can slip phase repeatedly, and a tight loop filter will average the varying output of the phase-only detector and produce false lock. The tighter the loop filter, the closer the false lock frequency can be to the desired frequency. It does not have to be at a harmonic of the desired frequency.

A frequency synthesis or clock generator PLL can easily find itself operating in the unstable region. The tight loop filter needed to control the frequency of the synthesized signal, coupled with the inherently high gain of some high-speed VCOs (which we will describe in the next section), places their loop filter pole above unity gain. If the instability is small, they work anyway. A PFD is not a pure phase detector, so to some extent the unfortunate pole-at-zero is eliminated, but not completely. But over the region where the PFD functions as a phase detector, basically when the PLL is "locked" and tracking its input, the PLL control loop oscillates, producing unwanted phase jitter. If this source of jitter can be reduced below the cycle driven jitter (from alternating pulses out of the charge pump), it is no longer a concern. But changing the loop parameters to make a PLL more SEU/SET tolerant can increase instability. Figure 12.7 shows a plot of this instability in an original case and with some modifications that a designer might use to minimize the jitter.

A quick way to understand what is going on in a PLL is to examine the loop control voltage (V_{ctl} in Figure 12.6), that is, the input to the VCO. In Figure 12.7 this is the dark wavy line. Ideally the line would be flat, indicating no variation in the frequency the VCO is requested to produce. A common type of VCO used in high-speed complementary metal-oxide semiconductor (CMOS) PLLs is the current-starved inverter loop, such as shown in Figure 12.8.

The transfer characteristic of this type of VCO is such that when V_{ctl} is near the threshold voltage, a very wide variation in frequency occurs for tiny changes in V_{ctl}. In other words, the voltage-to-frequency gain (often denoted K_{vco}) is very high near $V_{\mathrm{threshold}}$. A small K_{vco} is beneficial in achieving low phase noise [15].

K_{vco} is not only high near $V_{\mathrm{threshold}}$; it is rapidly varying (i.e., highly nonlinear), which in itself can cause stability problems [14]. The plots of Figures 12.7a, 12.7b, and 12.7c were obtained using the schematic of a well-designed PLL, but operating it at a lower than designed master input clock frequency so that V_{ctl} would be too low and the loop gain very high. The result is Figure 12.7a, in which a steady oscillation in V_{ctl} produces a frequency instability that is unacceptable.

An alternative to the current-starved inverter VCO is a tank circuit. These typically have smaller K_{vco} but also more narrow tuning ranges. And the necessity of having an inductor makes them less desirable for fully integrated or redundant applications.

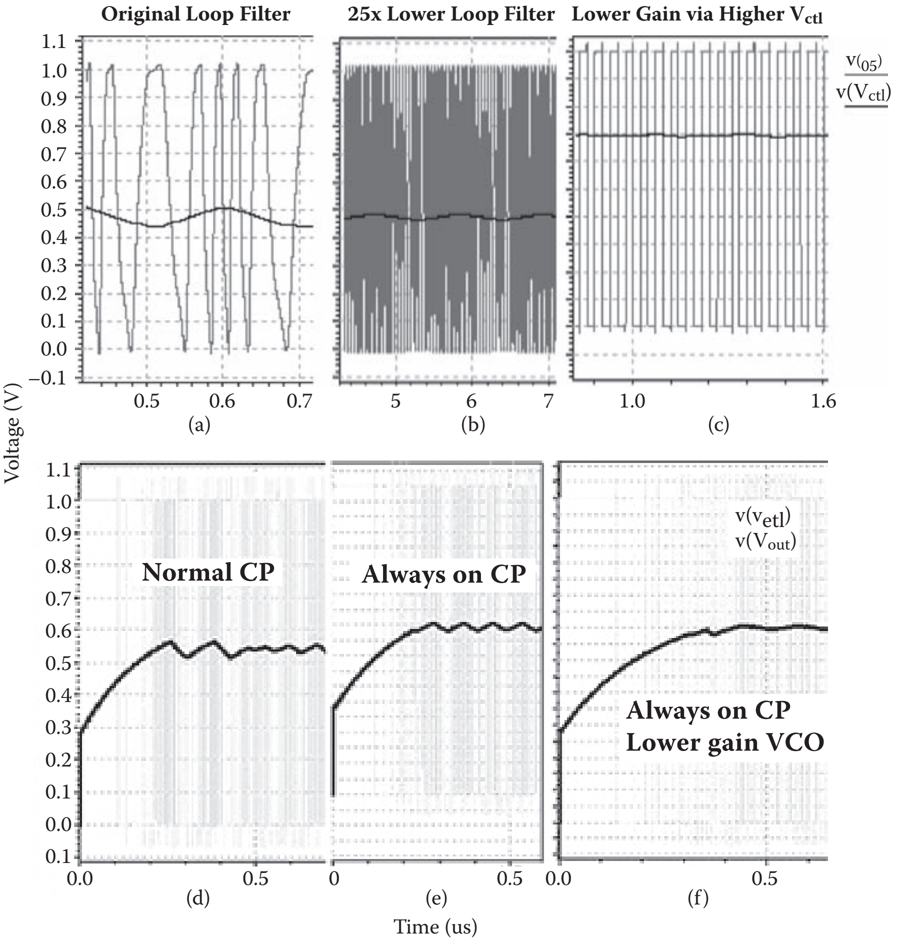
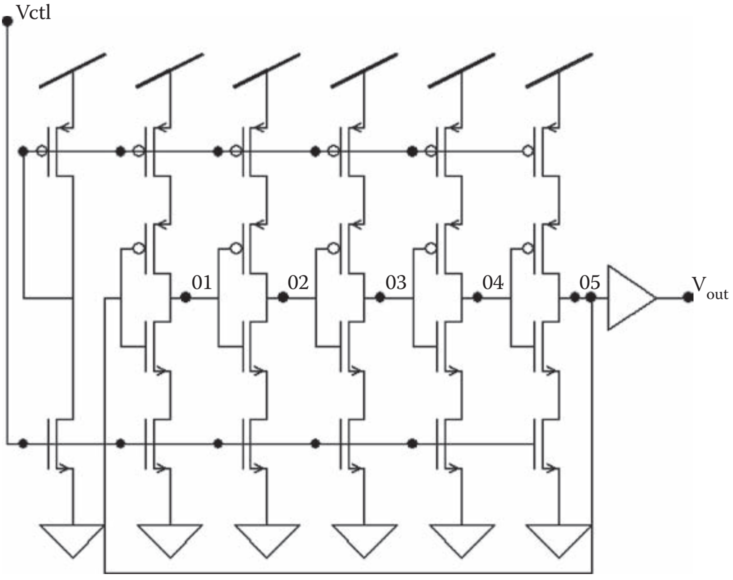


FIGURE 12.7 PLL instability management, comparing approaches.

Figure 12.7b shows what happens if one uses the instinctive solution feedback circuit designers would apply, that is, to "compensate" or dramatically lower the pole that the designer can control, the loop filter cutoff frequency. This is lowered by a factor of 25 in the middle figure. It appears to help slightly, perhaps by a factor of two, but the acquisition and tracking characteristics are dramatically reduced, in fact by a factor of 25! So while this sort of works, it might conflict with SEU/SET recovery performance requirements. Why doesn't this technique work better? It is because the PFD is highly nonlinear. A large frequency deviation can have approximately the same PFD output as a modest deviation. This makes the loop less sensitive than it should be to changes in the loop bandwidth.

Figure 12.7b shows what happens if, instead of tampering with the loop filter, we lower the VCO gain. Since the gain, if we are using a current-starved inverter VCO, is determined by the bias of V_{ctl}, we must figure how to raise V_{ctl}. This can be done by



(oscillator feedback)

FIGURE 12.8 Current-starved inverter loop VCO.

making the VCO slower, so a higher V_{ctl} is needed to operate at the desired frequency. To achieve this, 2.5 pF of capacitance was added to nodes O2 and O3. Modifying at least two nodes, and an even number of them, assures a symmetric waveform within the VCO, and that one node is not operated far beyond its cutoff frequency. One could modify all nodes if one wished. Adding stages to the VCO is not a particularly effective alternative for raising V_{ctl}, because very many stages are required. Making the circuit larger also would increase its SEU/SET error cross section.

Adding capacitance raised V_{ctl} from around 0.46 V to 0.8 V, and the output frequency stability was vastly improved. This was accomplished without significant degradation of the acquisition and tracking characteristics of the PLL! It might seem to an experienced feedback loop designer that changing the loop filter or loop gain to have a given stability effect should be more equivalent. But due to the nonlinear nature of the PFD, changes to the VCO predominantly affect stability, not acquisition. When there is a frequency mismatch between the VCO and the reference input, the PFD outputs either low or high, with no indication of how low or how high. So, excess gain in the VCO increases the frequency error without increasing the tracking "force."

Figure 12.7d shows a charge-pump PLL in acquisition mode. Notice that V_{ctl} jogs around in an irregular way, due to the nonlinear interactions among the VCO, PFD, and loop filter. It may take a long time for this to stabilize. During acquisition, a frequency deviation is produced by the PLL to bring phase into a matching condition. At match, the PFD produces no output. The phase momentarily matches. But the frequency deviation persists until enough phase error accumulates to drive the loop back the other way. For stability, one must guarantee this process eventually damps out, which may require a very slow (low bandwidth) loop, not what we'd like for fast SEU/SET recovery. There are four parameters by which such a "slow" PLL might be judged: acquisition time, jitter, spectral purity, and phase error (with regard to the input reference). In the slow design, both acquisition time and phase error are traded for jitter and spectral purity. A small and varying phase error would be of no consequence in clock generator applications of a PLL. But in a redundant voted PLL, if the three component PLLs have independent phase errors, then phase-induced voting error can result.

Figure 12.7e shows a PLL in which the PFD is modified to always produce an output. This is done by presuming that if the phase is not ahead, it is behind, so the PFD is always outputting a signal, or "always on." Such an architecture reintroduces the drawbacks of older more linear PLLs (the phase error bias), but notice that it also has a more linear behavior without the chaotic jogs of the classic three-state PFD. It is also free of dead-zone nonlinearity near-zero phase error and is easier to analyze. Figure 12.7f combines VCO gain reduction with the always-on PFD to produce a very well-behaved loop. A clever designer could match the V_{ctl} of the desired operating frequency of the VCO to the 50% duty cycle of the always-on PFD to create a PLL that would have no static phase error, low jitter, and also quick recovery from SEU/SET.

The faster a PLL acquires, the faster it will recover from an SEU/SET. Recovery time is important in a voting arrangement, because while one module is recovering, a second SEU/SET will cause an output error. Tight phase tracking is important to prevent phase-induced voting error. The key to making a good redundant fault-tolerant PLL is to start with a fast acquisition, low phase error, and single-string PLL design.

12.4 SEU/SET CHARACTERISTICS OF PLL BUILDING BLOCKS

Figure 12.6 shows five building blocks in a basic PLL for frequency synthesis. We describe the ring VCO in connection with Figure 12.8. The other four blocks can in principle be anything the designer chooses, but we discuss a representative design of each block here for purposes of understanding how their SEU/SET characteristics might affect our overall design.

12.4.1 Ring VCO

The SEU/SET characteristic of the ring VCO is straightforward and somewhat unfortunate. It comes to an erroneous phase, from which it does not return on its own. Each stage of the VCO has four transistors, about as many as a typical digital logic gate. There are usually at least 5 stages, and sometimes 10 or more, so the cross section of the VCO rivals that of any other part of the PLL, such as the frequency divider. To make matters worse, the VCO is operated in current starved mode, which means that less current is available for charge clearing after an ion strike than would be the case in a high drive digital circuit.

Furthermore, an SET in any part of the VCO causes a phase displacement, whereas in a digital circuit only half the circuit is susceptible most of the time. This at least doubles the error cross section of the VCO. There are three effects in play. In a digital circuit with multiple input gates, the state of the logic ignores many of the inputs. For example, a NAND gate with one input low effectively ignores the second input and any error on the second input. The VCO stages do have multiple inputs (the signal and the control voltage), but neither of them is ever ignored.

Second, digital logic transistors that are in the ON or conducting state do not experience a state change when an ion strikes, because ion strikes only increase conduction and do not decrease it. If the excess charge is cleared by the time the state changes, no effect is noticed. Timing of the digital logic is important only insofar as it meets minimum timing requirements. However in the VCO timing of the delay through each stage is always critical. Even a slight change in timing due to extra time required to clear charge from an ion strike will result in clock jitter and possible timing violations in the target circuit served by the clock.

Third, digital logic is sampled by the clock, and errors are counted only if they persist through a clock edge. But the clock, and thus the VCO, is not sampled by anything, and errors occurring at any time may result in system errors.

The first goal of an SEU/SET tolerant design is to quickly return the VCO to a correct state. This should be done regardless of the method of eliminating errors. If, for example, one of three VCOs lingers in an incorrect state, the chances it will cause phase-induced voting error increase. If it lingers long enough, there may even be a second SEU/SET in another part of the circuit, causing an error.

There are several ways of quickly returning a VCO to a correct state, and in later sections we will explore two of them in detail. One method is to use PLL parameters that produce fast acquisition and tracking. Presumably this will also result in fast correction. Fast correction will not prevent an error on the output and so must be used in combination with some other scheme for eliminating errors. But it will prevent the VCO from lingering in an incorrect state and thus minimize the probability of phase-induced voting error, or accumulation of a second error. The problem with fast acquisition and tracking is that, as described already, it is often at odds with frequency and phase stability, or tightness of tracking.

Another method of quickly returning a VCO to a correct state is to have three VCOs and vote them [13]. This works so quickly that it also eliminates errors. However, it eliminates only errors from the output of the VCO, if taken from the voter output, not from other parts of the PLL. Still another method is to vote only the output of the entire PLL (with two other identical PLLs) and to allow the feedback loop to resynchronize any PLL that experiences an SEU/SET induced error. We will explore both of these in a subsequent section.

12.4.2 Frequency Divider

The frequency divider, needed when the PLL is to provide frequency synthesis, is a digital state machine, and an SEU putting it in a different state is probably the most disruptive of any SEU/SET effect in a PLL. It is possible of course to vote every bit in the state machine [7]. Or one can allow the feedback loop of the PLL to eventually resynchronize the divider by reacquiring lock on the master clock input. In either case, a divider that minimizes SEU susceptibility is a good idea, such as a fully synchronous design.

It is very important to note that if the frequency divider is protected by voting, this cannot be done using the same components that otherwise participate in some other PLL voting scheme. For example, if three complete PLLs are voted, differences in the frequency dividers are essential to allow a failed PLL to resynchronize itself with the others.

12.4.3 Sigma-Delta Fractional-N Frequency Dividers

Sigma-delta or fractional-N frequency dividers use a dithering scheme to divide the VCO frequency by a sequence of integers that averages to a noninteger value. This creates two problems for the designer of a fault tolerant PLL. First, there is a lot more logic in the frequency divider, which must be protected from SEU/SET. This is inconvenient but not conceptually difficult. If voting is used, care must be taken that no internal state is left unprotected and allows an error to persist.

Second, either the variability of the output of the dithered frequency divider must be averaged over a longer period, implying a lower bandwidth loop filter and leading to problems we have already discussed, or compensation circuitry must be used to tune out the expected variations. Compensation circuitry can be analog in nature, getting involved with the charge pump, and can be both more susceptible to SET and more difficult to protect. A strategy of protecting an entire PLL is advantageous for such a situation. In the case where this circuit is protected by voting, care must be taken to vote every internal state variable so that there are no persistent errors.

12.4.4 Phase-Frequency Detector

There are many types of phase detectors and phase-frequency detectors [1,14]. For clock recovery PLLs, which examine data transitions and synthesize the implied clock, a PFD that tolerates missing transitions is required. For frequency synthesis, we are already in the position that the PLL is generating many more transitions than are in the master clock and, so for synthesis, PFDs are desired that use every clock edge, both leading and trailing, and produce immediate correction signals if the edge is leading or lagging. In the interest of setting a manageable scope for the current discussion, we limit ourselves to the second type. These are usually three-state or higher-logic circuits. While better acquisition performance can be obtained with complex higher-state PFDs, the additional states also increase SEU susceptibility and increase the difficulty of resynchronization. An ordinary three-state PFD, as shown in Figure 12.9, always resets itself when a clock edge has occurred on both inputs.

For most radiation-tolerant applications, we would use fully synchronous flip-flops. However, this PFD circuit works only with an asynchronous reset. The circuit of Figure 12.9 updates on leading edges of master and slave signals. One source of phase jitter in a frequency synthesizer is the PFD update cycle, because V_{ctl} will typically vary from some minimum to maximum value between PFD updates. The update cycle can be cut in half, reducing phase jitter, by using two PFD circuits and negating the inputs to the second one so that it updates on trailing (falling) edges.

As soon as both flip-flops have triggered (i.e., an edge is detected on both input signals) the flip-flops are reset, clearing any SEU condition.

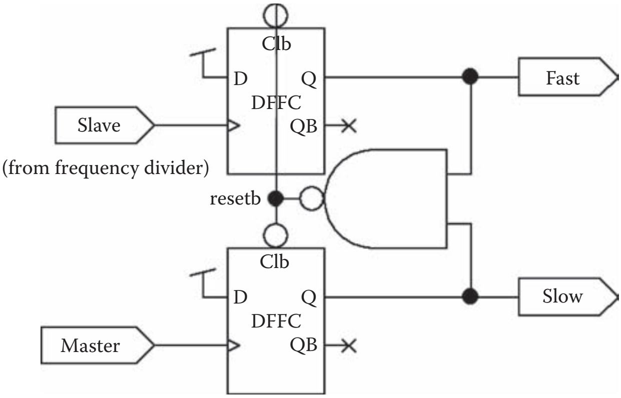


FIGURE 12.9 Single-edge three-state PFD.

12.4.5 Charge Pumps

There is a bit of an overlap when both the signals "fast" and "slow" are high, because of the time it takes the reset to operate. In a high-speed PLL this can be a significant error factor. If perfect charge pumps are used, in theory the "fast" and "slow" signals each results in a fixed current pulse into the loop filter and cancel out. But current-oriented charge pumps are relatively more susceptible to SEU/SET than a voltage-oriented charge pump (voltage with a high impedance switch) [4]. The larger number of transistors in a near-ideal current pump, and their lower drive strengths, increase both the exposure to SETs and the time required for recovery from SETs.

In the case of the voltage-oriented charge pump, the overlapping "fast" and "slow" signals do not exactly cancel, and phase error is produced. This can be eliminated by using pulse trimmers to reduce the length of these two signals by exactly the amount of the reset delay. Figure 12.10 shows a trimmer circuit that uses the same flip-flop reset to time the amount of trimming. Figure 12.11 shows the complete dual-edge PFD with trimmers and voltage-oriented charge pumps (switches are minimum-size pass gates).

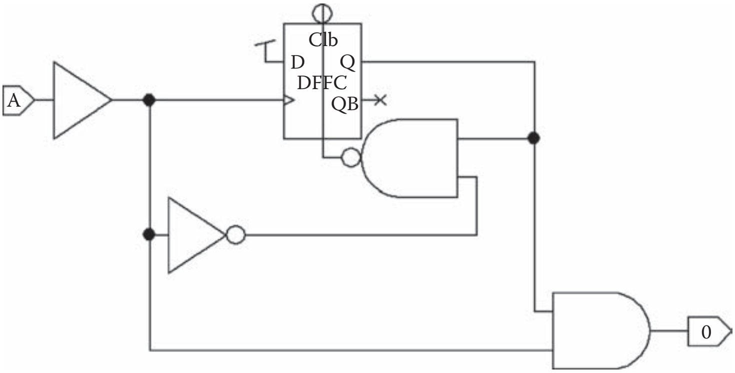


FIGURE 12.10 Reset pulse trimming circuit.

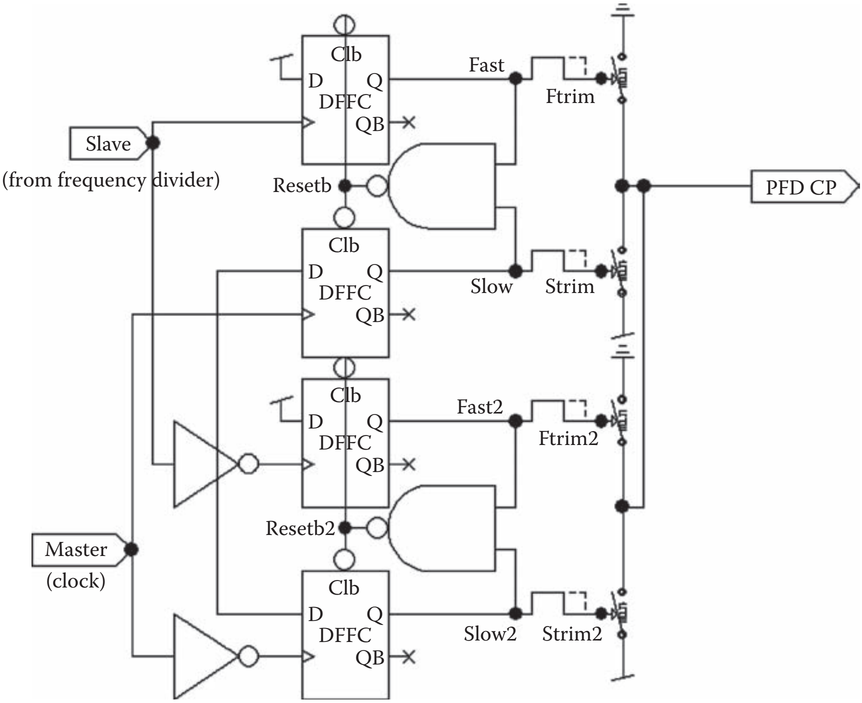


FIGURE 12.11 Dual-edge PFD with voltage-based charge pump.

A good bit of the literature on precision PLL design, such as frequency synthesizers for communication circuits, depends on a sophisticated current-based charge pump with precisely matched currents. There are a couple of approaches for dealing with this situation. One is to simply use the strategy of protecting the entire PLL, as we have been advocating, rather than its parts. SET recovery time will not be fast, but there will be errors only if a second SET occurs before the recovery from the first is complete. While this would be disastrous in a clock circuit, it is probably acceptable in a communication circuit where the communication protocol provides other means of handling errors. In other words, SEU/SET performance of frequency synthesizers is less critical than for clock generator circuits.

A second strategy is to address issues with the voltage-based charge pump. Unbalanced charge injection can be addressed by V_{ctl} tuning as already described, although this might be hard to make process-independent. Power supply noise is also a commonly voiced concern for voltage-based charge pumps (though also for current-based pumps). Lee and Wang [16] have shown significant benefits from using separate regulators for the VCO and the charge pump.

12.4.6 Loop Filter

The last block of the PLL is the loop filter. As emphasized, the ideal loop filter would not be anything more than a single-pole RC filter. Second- and third-order loop filters will lead to longer acquisition time and also longer recovery times from SET/SET errors. With the description in Section 12.3 of how PLLs can be stabilized, one may be able to solve phase jitter problems by modifying the VCO gain. If necessary, the always-on PFD technique could be used. However, as with the charge pump, a large amount of technical literature would have to be disregarded to follow our most aggressive recommendation. We caution only that, when using a higher-order filter, make sure to examine the SET recovery time. If an SET disturbance causes milliseconds of thrashing around, consider reducing the VCO gain.

Another consideration is simulation time. If you are designing a nonredundant PLL, it is sufficient to get one correct simulation for each frequency of operation, and you are done. But when designing a fault-tolerant PLL, it is necessary to consider the response to a variety of faults, increasing the number of simulations. In each simulation, one must wait for acquisition, inject a fault, and wait for it to settle, so each simulation is longer. It is tempting to incompletely verify the design in such a situation. The techniques we have outlined to reduce acquisition time will greatly speed up verification by allowing shorter simulations.

Ideally the RC loop filter would be just resistors and capacitors. In practice, in a CMOS circuit you can obtain an approximate RC filter by using resistor-connected field-effect transistors (FETs) for the resistor and the gates of FETs for the capacitor. This has the advantage of being process-independent and relatively scalable, possibly requiring no change when moving to new processes.

Figure 12.12 shows such a loop filter. It is important to use a symmetric pair of resistor connected FETs connected in opposite directions to avoid a nonlinear preference for charging the filter in one direction or another. This circuit uses a width/length (W/L) of 3/100, giving quite a high impedance. Because of threshold voltages, the circuit will not quite charge to either supply rail. For the capacitor,

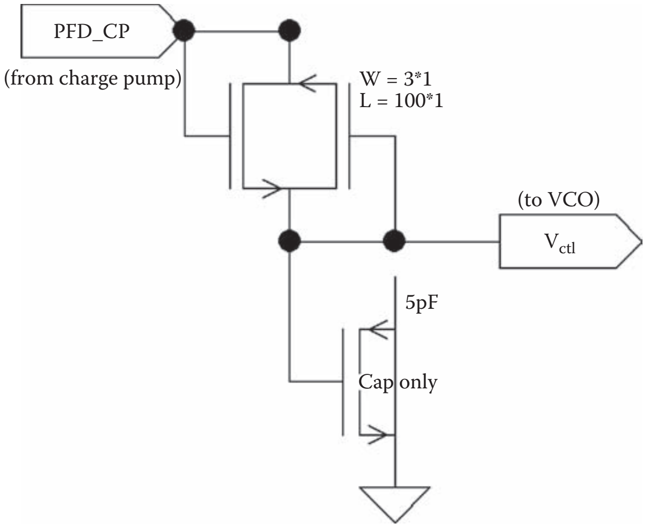


FIGURE 12.12 Loop filter.

enough FETs with large area gates are connected in parallel to make whatever value is needed.

It would seem that an SET in the loop filter might be significant, but in practice it might be of less consequence. If there is a lot of charge stored on the capacitor compared with the charge generated by an ion strike, an SET has less effect. Any effect it does have will be eliminated by the redundant voting techniques that will be used to protect other parts of the circuit.

12.5 APPLYING REDUNDANCY TO PLLS

It is possible to improve or vote or otherwise mitigate SEU/SET for the individual components of PLLs. This usually leaves some component, such as the loop filter with an analog output, unprotected. It is the purpose of this chapter to propose comprehensive treatments. These can be simpler, since a good PLL design is merely repeated three times and voted, with a single voter instead of many voters. But there are tricks and considerations to such an arrangement that require a little more examination.

12.5.1 Output-Only Voting Method

The first and simplest arrangement, shown in Figure 12.13, is to vote only the output. This arrangement relies on the individual PLLs to resynchronize themselves with the master input clock after an upset condition. This is a function they are already designed to do when they are turned on. No PLL ever looks at what is going on in another PLL, but only at the input signal. So if it works at power up, it will also work following an SEU/SET. The question with this design is whether it will work well enough to avoid phase-induced voting errors. And that will largely depend on whether you have designed a good PLL!

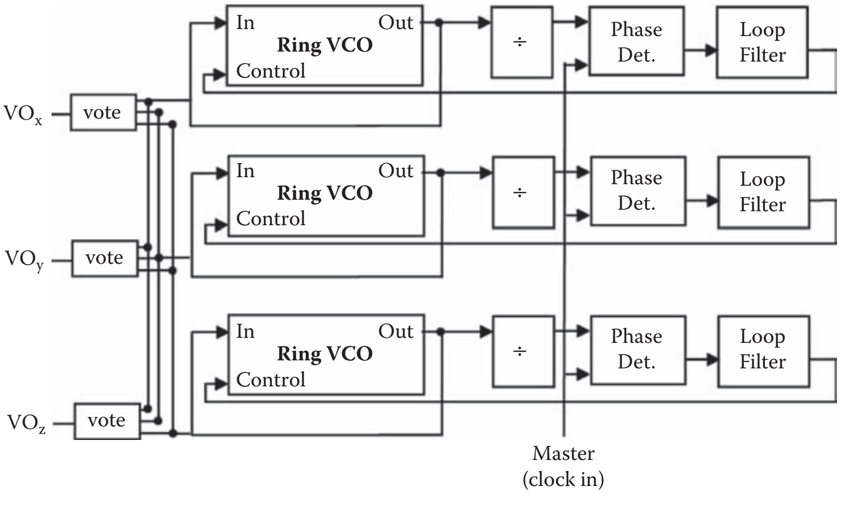


FIGURE 12.13 Output voted PLL.

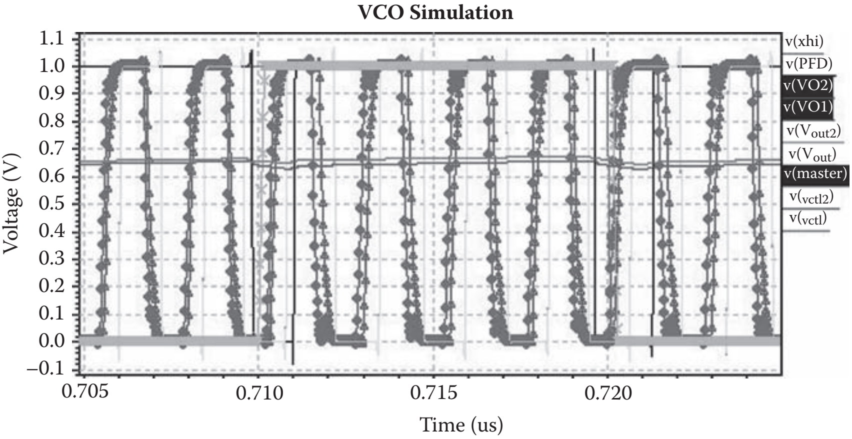


FIGURE 12.14 Comparing outputs of identical PLLs after disturbance.

When you run a SPICE simulation on three identical PLLs with identical starting conditions, the results will appear identical. But real PLLs will not be identical. However, when you disturb one of the PLLs by injecting a simulated SET, then as it recovers you will see what the relative tracking of your PLL design might be in the real circuit. Figure 12.14 shows a plot of two PLLs.

The light gray line is the master clock, and the darker highlighted waveforms at 8× frequency are the two PLL outputs. The two lines at about 0.65 V are the two V_{ctl}'s. Each horizontal axis tick mark is about 1 ns, so the difference in the two PLL outputs is a small fraction of a nanosecond, probably around 100–150 ps. A large voter such as shown in Figure 12.3 will be slow enough in this technology (90 nm) that voting these signals should not produce any phase-induced errors. However, if it gets any worse, it won't work.

The thing that might make it get worse is trying to fix a stability problem by dramatically lowering the loop filter bandwidth. This makes the control loop very slow, and it could be a long time, if ever, before the outputs line up again. Whether you fix stability problems by lowering the VCO gain or with the loop filter, you should carefully check the worst-case disturbances to make sure your PLL outputs will line up again. You can also check for the effect of process variation by changing the W/L of a bias transistor in one of the ring VCOs by an amount of half a lambda, or half the minimum feature size of your process. This model approximates the worst-case process parameter variation that you would normally see.

12.5.2 The VCO Voting Method

As mentioned earlier, the VCO itself can be voted, which guarantees that all three VCOs quickly return to lock step. So what do you then do about the rest of the PLL?

As far as the PFD, charge pumps, and loop filter are concerned, whether you have to do anything depends on the amount of jitter caused by an SEU/SET on these components. The effect of an SEU/SET on any of these components is some delta to V_{ctl}, which will result in an erroneous delta to the VCO frequency, but not any discontinuity in the output. It must be determined whether the worst case delta to VCO frequency, and resultant phase deviation, is within desired operating limits of the PLL.

The PFD will be cleared out every master (input) clock cycle. The worst case is the delta frequency caused by one master clock cycle. Adjustments can be made by changing charge-pump current, loop filter time constant, or VCO gain. While the PFD could be protected by a complicated voting scheme, it is likely these other adjustments will suffice.

An SET in the charge pump could last longer than a master clock cycle, since tiny high impedance transistors are used. This is difficult to determine directly by heavy-ion testing. It can be estimated by using laser testing or SPICE analysis, but in either case a calibration based on the transistor sizes used is advisable. Once the longest error state is determined, analysis and adjustment proceed in the same manner as for the PFD.

An SET in the loop filter makes a small change in the charge stored in the loop filter. Erroneous charge on the loop filter capacitor will eventually be eliminated by operation of the PLL control loop. It will produce a small change in V_{ctl} and, in turn, a small change in VCO frequency. The problem is that it is difficult to quantify.

So with the PFD, the charge pump, and the loop filter we have three progressively more difficult to quantify effects on V_{ctl}. Here is the dilemma. The voted VCO behaves like a single VCO, not three independent ones as in the voted output PLL. If we are to have a single V_{ctl} to control it, this V_{ctl} will either be subject to SET errors, or we will have to design a complex analog voter, which will itself be extremely challenging. If we replicate the PFD, charge pump, and loop filter and produce three V_{ctl} signals and separately drive the three VCOs, then there is no independent correction of errors by loop action, because with the voted VCO it is just one loop. Errors in the PFD will clear in one cycle. SETs in the charge pump will be removed by charge collection processes. Errors in the three loop filters will eventually decay away if the filters are passive RC circuits. So replicating these components and producing three V_{ctl} signals might be a feasible option. It seems to lack the positive assured error removal of the voted output PLL design, but it is not particularly susceptible to phase-induced voting error.

The problem with the frequency divider is different. An SEU on the frequency divider will never be cleared unless you do something special to clear it. Because the VCO is voted internally, the control loop of the affected PLL will never be allowed to correct the divider. The PLL control loop and the VCO voting will be fighting each other. This condition will persist, and the PFD of the failed PLL will constantly be trying to adjust its loop filter to “make up” the lost phase, stored in the divider, but it will never be able to. With only two good remaining PLLs, the next time one of them experiences an SET/SEU the circuit output will be in error.

So to make the VCO voting method work, you have to provide some means of correcting the dividers. The most obvious method is to replicate and vote the dividers. The internal signals of the divider must be voted so that the voting corrects their internal state. If only the outputs are voted, an erroneous count will simply be perpetuated. It is possible to dream up other schemes, such as forcing the three dividers to reset all at once, but these would become complicated. A simple scheme is to drive each divider with one of the three VCO outputs. The output of the dividers can either be from a triple output voter (preferred), driving replicated PFDs and so forth as previously described, or a single output driving a single-string PFD, charge pump, and loop filter if it can be determined that an SET impact does not disturb the VCO frequency by more than the desired design specification.

The advantage of the VCO voting method is not having to worry about phase-induced voting error. The VCO voting method can be used whenever it is too difficult to design robust PLLs that will remain in lock in spite of process parameter variations and will return to lock after a disturbance. The drawback of the VCO voting method is the difficulty of verifying that SETs in the PFD, charge pump, and loop filter are within tolerance, or the ambiguity of the lack of positive removal of errors if these components are replicated without independent feedback correction. Simulations suggest that errors in replicated charge pumps and loop filters decay rather quickly, in tens or at most hundreds of nanoseconds, but this has not been confirmed by heavy-ion testing.

12.6 CONCLUSIONS

We have seen that to design a fault-tolerant PLL, we must start with a good PLL design that will remain in tight lock to avoid phase-induced voting errors and will reacquire quickly so that the vulnerable time after one PLL has been upset is minimized. To design a good PLL has required a revisit of PLL stability theory and consideration of nonlinear factors, such as the differing effects of stabilizing by changing VCO gain versus changing the loop bandwidth.

We then examined two redundancy topologies that produce an entirely redundant and fault-tolerant PLL. The simpler one votes only the output and relies on the PLL feedback to resynchronize the failed PLL but is subject to phase-induced voting errors if the PLLs do not sync up tightly.

The more complicated method keeps the VCOs in lock step but requires other mitigation, especially of the frequency divider, to avoid a fight over control of the loop, which would result in making a transient error permanent.

For frequency synthesizer applications, the required phase error is likely already so small that phase-induced voting errors will not be an issue. These applications also are more likely to require a complex frequency divider of the fractional-N or sigma-delta type, which is tedious to mitigate for SEU/SET. So for frequency synthesizer applications, the output voted PLL is highly recommended.

For clock generator applications, the phase error should be carefully reviewed for potential voting problems. If there are any, they can probably be eliminated by the techniques we have given, such as VCO gain reduction. In that case, again the simpler output voted PLL is recommended.

If you have a PLL design that you simply must use but that does not control the phase tightly enough for the simpler output voting method or that is sensitive to device parameter or other variations, then the more complicated VCO voting method should work, provided all state variables in the frequency divider are voted, and the remaining components have no source of persistent error following an SET.

This analysis has been done for a frequency synthesis PLL of the sort needed for an on-chip clock multiplier. Similar principles can be applied to other types of PLLs.

The output voted PLL has been fabricated and tested by the author. Components of the VCO voted PLL have been investigated by Loveless at Vanderbilt [5,6,12], and a VCO voted PLL has been simulated by the author.

REFERENCES

1. Wolaver, D.H., *Phase-Locked Loop Circuit Design*, Prentice Hall, Englewood Cliffs, NJ, 1991.
2. Van Alen, D.J. and Somani, A.K., "An All Digital Phase Locked Loop Fault Tolerant Clock," *IEEE International Symposium on Circuits and Systems*, vol. 5, pp. 3170–3173, 1991.
3. Kessels, J.L.W., "Two Designs of a Fault-Tolerant Clocking System," *IEEE Transactions on Computers*, vol. C-33, no. 10, pp. 912–919, Oct. 1984.
4. Loveless, T. D., Massengill, L. W., Bhuva, B. L., Holman, W. T., Witulski, A. F., and Boulghassoul, Y., "A Hardened-by-Design Technique for RF Digital Phase-Locked Loops," *IEEE Transactions on Nuclear Science,* vol. 53, no. 6, pp. 3432–3438, Dec. 2006.
5. Loveless, T.D., Massengill, L.W., Bhuva, B.L., Holman, W.T., Reed, R.A., McMorrow, D. et al., "A Single-Event-Hardened Phase-Locked Loop Fabricated in 130 nm CMOS," *IEEE Transactions on Nuclear Science,* vol. 54, no. 6, pp. 2012–2020, Dec. 2007.
6. Loveless, T.D., Massengill, L.W., Holman, W.T., and Bhuva, B.L., "Modeling and Mitigating Single-Event Transients in Voltage-Controlled Oscillators," *IEEE Transactions on Nuclear Science*, vol. 54, no. 6, pp. 2561–2567, Dec. 2007.
7. Nemmani, Anantha, N., "Design Techniques for Radiation Hardened Phase-Locked Loops," master's thesis, Oregon State University, A874174, August 2005.
8. Shuler, R.L., Kouba, C., and O'Neill, P.M., "SEU Performance of TAG Based Flip-Flops," *IEEE Transactions on Nuclear Science*, vol. 52, no. 6, pp. 2550–2553, Dec. 2005.
9. Shuler, R.L., Balasubramanian, A., Narasimham, B., Bhuva, B.L., O' Neill, P.M., and Kouba, C., "The Effectiveness of TAG or Guard-Gates in SET Suppression Using Delay and Dual-Rail Configurations at 0.35 μm," *IEEE Transactions on Nuclear Science*, vol. 53, no. 6, pp. 3428–3431, Dec. 2006.
10. Mongkolkachit, P. and Bhuva, B., "Design Technique for Mitigation of Alpha-Particle-Induced Single-Event Transients in Combinational Logic," *IEEE Transactions on Device and Materials Reliability*, vol. 3, no. 3, pp. 89–92, Sept. 2003.
11. Balasubramanian, A., Bhuva, B.L., Black, J.D., and Massengill, L.W., "RHBD Techniques for Mitigating Effects of Single-Event Hits Using Guard-Gates," *IEEE Transactions on Nuclear Science*, vol. 52, no. 6, pp. 2531–2535, Dec. 2005.
12. Gardner, F., "Charge-Pump Phase-Lock Loops," *IEEE Transactions on Communications*, vol. 28, no. 11, pp. 1849–1858, Nov. 1980.
13. Loveless, T.D., Massengill, L.W., Bhuva, B.L., Holman, W.T., Casey, M.C., Reed, R.A. et al., "A Probabilistic Analysis Technique Applied to a Radiation-Hardened-by-Design Voltage-Controlled Oscillator for Mixed-Signal Phase-Locked Loops," *IEEE Transactions on Nuclear Science*, vol. 55, no. 6, pp. 3447–3455, Dec. 2008.
14. Abramovitch, D., "Phase-Locked Loops: A Control Centric Tutorial," *Proceedings of the 2002 American Control Conference*, vol. 1, pp. 1–15, 2002.
15. Ti, C.-L., Liu, Y.-H., and Lin, T.-H., "A 2.4-GHz Fractional-N PLL with a PFD/CP Linearization and an Improved CP Circuit," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp.1728–1731, May 18–21, 2008.
16. Lee, T.-J. and Wang, C.-C., "A Phase-Locked Loop with 30% Jitter Reduction Using Separate Regulators," *VLSI Design*, vol. 2008, Article ID 512946, 2008.

13 Autonomous Detection and Characterization of Radiation-Induced Transients in Semiconductor Integrated Circuits

Balaji Narasimham, Bharat L. Bhuva, Ronald D. Schrimpf, Lloyd W. Massengill, William Timothy Holman, and Arthur F. Witulski

CONTENTS

13.1 INTRODUCTION

As the device dimensions and operating voltages of integrated circuits (ICs) are shrunk to satisfy an ever-increasing demand for lower power and higher speed, integrated circuit sensitivity to radiation may increase significantly [1-3]. Deep submicron devices show increased susceptibility to single-event effects (SEEs), which constitute a particular category of radiation effects [2] compared with previous technology generations. A single event (SE) occurs when an energetic particle, such as a heavy ion or neutron, strikes a device and causes a change in the device's normal operation.

The category of SEEs encompasses a multitude of phenomena that have, as a common cause, the passage of an energetic particle through the semiconducting or insulating materials used in the manufacture of integrated circuits. The common sources of SEEs are cosmic rays and heavy ions for space applications and neutrons (which produce SEEs indirectly through secondary particles emitted as a result of nuclear interactions) and alpha particles for terrestrial applications. As an energetic particle passes through the IC, it excites electrons from the valence band and leaves behind a track of electrons and holes (Figure 13.1). If the track passes through or near a reverse-biased semiconductor p-n junction, the high electric field present in the region can efficiently separate the particle-induced electrons and holes. Carriers thus separated may be collected by a circuit node due to the nodal voltages present in the circuit, generating a current at the terminals of the semiconductor device. Deposited carriers can also diffuse from the bulk or substrate of the semiconductor into the vicinity of the depletion-region field where they may be collected by the circuit node, adding to the total charge collected. Charge generated along the particle track can locally extend the junction electric field due to the highly conductive nature of the charge track, leading to a field funnel region [4]. This funneling effect can increase charge collection at the struck node by extending the junction electric field further into the substrate, allowing charges deposited away from the junction to be collected efficiently through drift. In advanced complementary metal-oxide-semiconductor (CMOS) processes when electrons or holes released by a particle strike are confined within the well region in which a transistor exists, charge collection may be enhanced

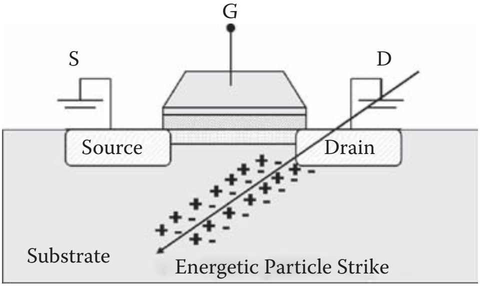


FIGURE 13.1 Generation of electron-hole pairs due to an energetic particle strike.

by a parasitic bipolar effect [3]. For example, for a *p*-type metal-oxide semiconductor field-effect transistor (PMOSFET) in an n-well process, holes induced by the particle strike may be collected at the drain or substrate junctions. However, electrons left behind in the well region lower the well potential. This lowers the source-well potential barrier and may result in injection of holes into the well from the source, which can then be collected at the drain. This adds to the original particle-induced current, and the effect is described as parasitic-bipolar charge collection. In sub-100 nanometer technologies the charge cloud from the ion strike can encompass multiple devices and well contacts and result in very complex charge collection behavior [5].

13.1.1 Soft Errors

Some types of SEEs are also referred to as soft errors in the commercial domain. Soft errors are the primary radiation concern for commercial terrestrial applications, as opposed to parametric degradation and hard errors, which are significant concerns in space and military environments [1]. A soft error occurs when a radiation event deposits enough charge to reverse or flip the data state of a memory cell, register, latch, or flip-flop. The error is "soft" because the circuit/device itself is not permanently damaged by the radiation and the error can be corrected by writing new data [1]. In contrast, a "hard" error is manifested when the device is physically damaged and the operation loss is permanent. Soft error rate (SER) is the rate at which a device encounters or is predicted to encounter soft errors. It is typically expressed as number of failures-in-time (FIT). One FIT equals one error per billion hours of device operation.

There are different types of effects that result in soft errors; the most important types are the single-event upset (SEU) and the single-event transient (SET). An SEU is a static upset in storage cells such as static random access memory (SRAM) cells, latches, and flip-flops. The upset rate due to such an event is independent of the clock frequency [1]. For sequential CMOS ICs, an energetic particle strike may cause a transient voltage perturbation, called an SET, which propagates through the circuit and may become stored as incorrect data, causing disruption of the circuit operation. An SET will result in an error if the SET pulse arrives at a storage node to get latched. For example, for a flip-flop, if an SET pulse arrives during the setup-and-hold time of the master latch, it will result in an error. Thus, upset rates due to SETs depend on the pulse width of the SET and the clock frequency [1,2]. With increasing clock frequency, there are more latching clock edges to capture an SET [1,2]. With decreasing feature sizes, the charge required to represent a logic HIGH state decreases and hence may result in increased susceptibility to SETs [2]. The width of the SET is a function of a multitude of factors including the CMOS restoring device drive strength (drain current magnitude) as well as the charge collection kinematics [2].

For advanced technologies, a large fraction of observed soft failures are estimated to be related to latched SET events. Precise knowledge of the radiation-induced transient pulse widths is thus important for determining error rates and for the design of hardening techniques to mitigate the effect of these transients.

This chapter presents an autonomous pulse characterization circuit technique to measure the distribution of SET pulse widths for different radiation environments.

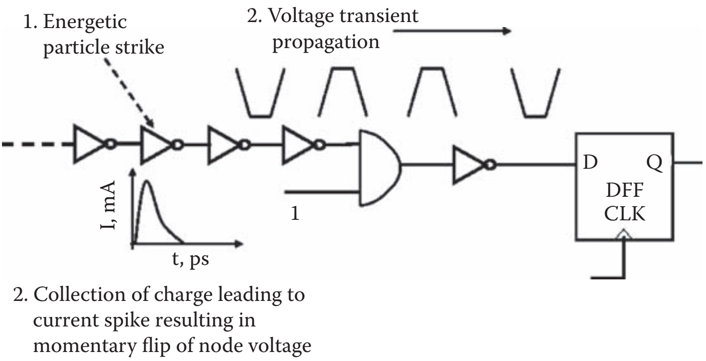

1. Energetic particle strike
2. Voltage transient propagation
I, mA
t, ps
1
D
Q
DFF
CLK
2. Collection of charge leading to current spike resulting in momentary flip of node voltage

FIGURE 13.2 Single-event transient propagation through a combinational logic chain.

The pulse characterization technique has been implemented in a range of CMOS technologies, and test chips have been used to measure the distribution of SET pulse widths for heavy ions, neutrons, and alpha particles.

13.2 SINGLE-EVENT TRANSIENTS AND LOGIC SOFT ERRORS

13.2.1 Single Events in Logic Circuits

In a combinational logic circuit, charge collection due to a single-event strike on a particular node will generate a low-to-high or high-to-low voltage transition or a transient. From a circuit analysis point of view, collection of charge induced by an ion first results in a current pulse on the node of interest. This current pulse is usually modeled using a double exponential current source in simulators or as a double exponential with a current plateau. This current pulse may momentarily flip the state of the output node of a logic cell, thus causing a “glitch” or transient to propagate along the combinational logic chain. The ability of this undesired pulse to propagate depends not only on its magnitude but also on the active logic paths from the struck node existing at that instant in time. An example of this is shown in Figure 13.2.

In Figure 13.2, a single-event strike generates a voltage transition on a node of this circuit. The possible propagation of this pulse to a latch (storage) element depends on several factors. First, the active combinational paths at that instant in time depend on the dynamic state of the logic. Second, assuming that an active path exists for the propagation of the pulse, the pulse will be shaped and phase delayed as it propagates through the intervening gates en route to a latch. Third, the temporal characteristics of the pulse as it arrives at a latch are important. The pulse must arrive within the setup-and-hold time of the latch element to be captured. The clocking characteristics of the latch and the previous state of the latch also affect the error rate. Depending on all three factors previously mentioned, the SE-generated noise pulse will be captured by the latch as erroneous information.

As long as an active path exists for the propagation of the single-event transient pulse, its capture as an error by a latch depends on the width of the transient, its arrival relative to the setup-and-hold time window of the clock, and on the clock frequency.

An error in this context is defined as latching an incorrect logic value. Depending on the magnitude of charge collected, the width of this transient voltage pulse varies. Thus, once an IC is manufactured, the pulse width of the transient (along with clock frequency) determines the vulnerability of the circuit to SETs [3,6].

Single-event strikes on control logic circuitry have also been identified as a significant contributor to the overall chip-level SER [7]. Specifically SETs created in the global and local clock buffers can result in clock jitter and race conditions.

For older technologies the SET could not propagate through a large number of logic gates since the ion strike and charge collection usually did not produce a full output swing (due to higher nodal capacitances) and was quickly attenuated due to large load capacitances and large propagation delays [1]. In advanced technologies with lower propagation delays and higher clock frequencies, the SET can more easily traverse many logic gates, and the probability that it is latched increases [1].

13.2.2 Logic Soft Errors—Scaling Trends

Previous work has shown that combinational-logic soft errors caused by latching SETs increase with technology scaling [6,8]. Figure 13.3 shows the SER per logic gate for different types of circuits. For memory and latch circuits, the per-bit SER decreases slightly with technology scaling and can be attributed to the faster scaling in the device cross sectional area than the critical charge of the cell. However, for logic circuits, the SER is predicted to increase with technology scaling and can be attributed to the

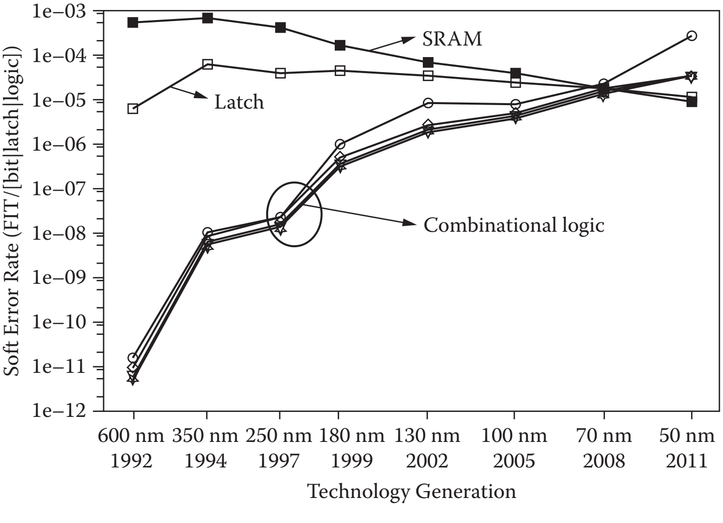


FIGURE 13.3 Soft error rate per logic gate for SRAM, latch, and combinational logic circuits indicating an increase in the SER of logic circuits with technology scaling. (Reprinted with permission from Shivakumar, P., Kistler, M., Keckler, S.W., Burger, D., Alvisi, L., "Modeling the effect of technology trends on the soft error rate of combinational logic." *Dependable Systems and Networks*, 2002. Proceedings of the International Conference on DSN: 2002: 389–398.)

decrease in critical charge combined with an increase in the operating frequency with technology scaling. For technologies beyond 65 nm it is projected that combinational logic SER may dominate SER from memory and latch circuits [8].

Some researchers indicate that logic soft errors may not scale up as rapidly as expected for advanced technologies or may even decrease with scaling for sub-100 nm technologies, due to a flattening of the scaling curve of supply voltage and clock frequencies [9]. While the per-gate logic SER may reduce slightly depending on frequency and supply voltage scaling trends, the overall contribution to the chip-level SER may still be significant especially with higher packing densities. Moreover, as pointed out in [9], the alpha particle contribution to logic SER may increase significantly with reduction in the critical charge with scaling; hence, it is important to understand and characterize scaling in SETs.

As stated earlier, the probability that a SET will result in an error depends on the propagation distance through the combinational logic circuit and the arrival time of the SET at the latch input [2,3,6,10-13]. Wider pulses have a greater probability of being present at the latching edge of the clock. Thus, characterizing transient pulse width is of paramount importance in both determining and mitigating single-event effects for advanced technologies.

Moreover, while error correction codes and latch-hardening designs have been developed to mitigate the effect of SEUs in memory elements, system-wide protection against SETs is quite difficult and involves considerable performance penalties [14]. A more manageable approach is to design limited protection against SETs through a targeted performance trade-off. Such trade-off decisions require detailed knowledge of the SET mechanisms and attributes, specifically SET pulse widths [14].

Transient pulse width is influenced by the nature of the ionizing particle, the technology used, location of the strike, and incident angle [2,15-18]. Modern submicron ICs are vulnerable to ionizing alpha particle and heavy-ion strikes and also to terrestrial neutrons that deposit charge through indirect ionization. Different ionizing particles interact differently with the silicon to deposit charge. Alpha particles that come from the radioactive decay of packages used for ICs have been a source of SETs through direct ionization in silicon. Energetic neutrons and protons can produce SETs indirectly through elastic scattering or a nuclear reaction in silicon. Low-energy neutrons can also interact with the boron (specifically, boron-10) in a semiconductor device, producing reaction products that can cause an SET. Cosmic ray heavy ions are also a source of SETs. The charge deposited by the different ionizing particles varies greatly and may affect the transient pulse width. For example, the charge deposited by the products of neutron-induced reactions (25–150 fC/μm) is much greater in magnitude than that deposited by alpha particles (4–16 fC/μm) and hence may pose a greater threat [1]. Likewise, the angle of the incident ionizing particle also significantly affects the charge collected and hence the SE pulse width.

13.2.3 Previous SET Characterization

Through the use of mixed-mode simulations, Dodd et al. characterized scaling trends in SET pulse widths for bulk silicon and silicon-on-insulator (SOI) technologies for processes ranging from 0.25 μm to 0.1 μm [19]. Their results indicate transients

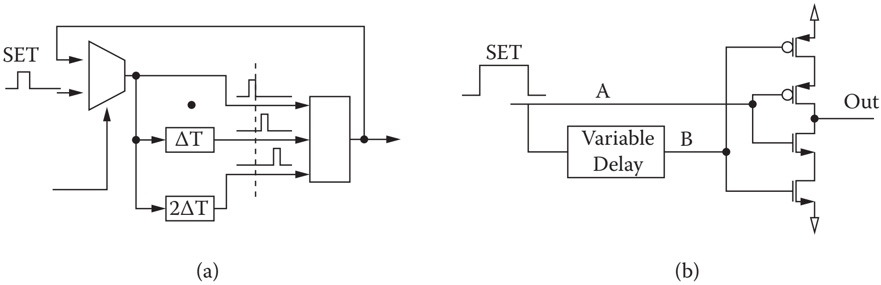


FIGURE 13.4 (a) Variable temporal latch technique. (b) Guard gate based technique for characterizing the width of SET pulses. In such techniques a delay element is tuned to match the width of the SET pulse. (Reprinted with permission from Eaton, P., Benedetto, J., Mavis, D., Avery, K., Sibley, M., Gadlage, M., Turflinger, T., "Single event transient pulse width measurements using a variable temporal latch technique." *IEEE Transactions on Nuclear Science*, Volume: 51 Issue: 6, 3365–3368. Reprinted with permission from, Baze, M. P., Wert, J., Clement, J. W., Hubert, M. G., Witulski, A., Amusan, O. A., Massengill, L., McMorrow, D., "Propagating SET characterization technique for digital CMOS libraries." *IEEE Transactions on Nuclear Science,* Volume: 53 Issue: 6, 3472–3478.)

of the order of 1 ns for CMOS bulk technologies at linear energy transfers (LETs) greater than about 50 MeV-cm^2/mg. The simulation results presented in [19] also suggest the presence of significant transients at LETs as low as 2 MeV-cm^2/mg at the 100 nm bulk process, and the authors predict an increase in susceptibility to alpha particles with technology scaling beyond 100 nm.

Researchers have also experimentally characterized transient pulse widths using multiple latches with delayed signal paths [20] or delayed clock signals (Figure 13.4a). Guard-gate-based techniques have also been used to measure SET pulse widths, as shown in Figure 13.4b [21]. In such techniques a delay element is tuned to a certain value, and the circuit measures all SETs longer than the delay. The techniques thus measure the event cross section for SETs greater than a certain threshold. Event cross section is defined as the ratio of number of SETs to the particle fluence used for the test. In such techniques the design of the delay element is critical, and any variations in the delay value can affect the measurement. Moreover, there has been little agreement on the range of SET pulse widths measured using the different techniques. For example, at the 130 nm technology node, pulse widths from a few hundred picoseconds [21] to several nanoseconds [22] have been reported. Baze et al. concluded that the majority of transients are 500 ps or shorter (with very few transients greater than 1 ns) in the 130 nm process based on SET measurements using the guard gate technique [21]. Benedetto et al. have observed transients greater than 2.5 ns long in the 130 nm process based on measurements using the variable temporal-latch technique. Furthermore, Benedetto et al. have predicted an increase in transient pulse width with technology scaling [22].

Another approach for SET pulse width measurement that has been previously reported is the use of a chain of cell copies that are monitored by latches to characterize the pulse width in terms of multiples of the individual cell delay, as shown in Figure 13.5 [14]. In this approach, the latches are clocked continuously to obtain

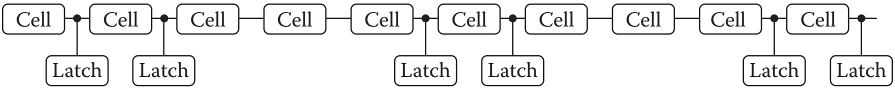


FIGURE 13.5 Chain of cell copies monitored by latches that are clocked continuously to capture information on the width of an SET pulse. (Reprinted with permission from Nicolaidis, M. Perez, R., "Measuring the width of transient pulses induced by ionising radiation." Reliability Physics Symposium Proceedings, 2003. 41st Annual. 2003 *IEEE International* Publication Date: 2003 On page(s): 56–59.)

information about the state of the cells. Since there are limitations to the maximum clock frequency that can be applied, it can be difficult to capture a very fast SET pulse using this approach. Transient current pulses have also been measured directly using oscilloscopes [23,24]. Such direct measurements are difficult to perform because of pulse distortion due to the capacitance of the measurement equipment and require costly experimental setup. While such techniques are suitable for measurement of laser-induced pulses, they are much harder to use with heavy ions. This is because such measurements are made on single transistors and are best suited for measurements where the ion-strike location is known a priori. It is difficult to make these measurements with heavy ions due to the random nature of the ion strikes.

The observable pulse width is a function not only of the base technology but also of the circuit topology through which it propagates [25] and the circuit operating parameters (e.g., supply voltage [22]). The measurement circuit used may even influence the measurement itself. Even if the influence of these parameters is eliminated, a fairly large statistical distribution of the collected charge has been observed based on the random nature of strike location relative to the affected node [26]. Previous SET pulse-width measurements, however, have not been able to capture the statistical distribution precisely.

The focus of this chapter is a test circuit that can characterize the width of SET pulses without the need for an external trigger or multiple laser strikes. The basic principle of operation of this circuit is similar to the one proposed in [14] but incorporates a *self-triggering* mechanism that does not require an outside signal to determine the presence of an SET pulse. This test circuit captures the SET pulse in a series of latches, which can be easily read out to determine the width of the pulse. This circuit technique can be used in CMOS and BiCMOS (integration of bipolar junction transistors and CMOS) processes (including SOI technologies) regardless of feature size or operating speed and can also be used for characterizing other spurious signals such as noise or cross talk pulses. This circuit has been implemented in 1.5 μm, 0.35 μm, 180 nm, 130 nm, and 90 nm bulk CMOS processes and in a 180 nm SOI process and has been tested with different energetic particles.

13.3 AUTONOMOUS PULSE-WIDTH CHARACTERIZATION

13.3.1 Propagation of a Transient through a Series of Inverters

A common parameter for specifying the performance of a digital IC is the propagation delay associated with an inverter, designated as one inverter delay. The test

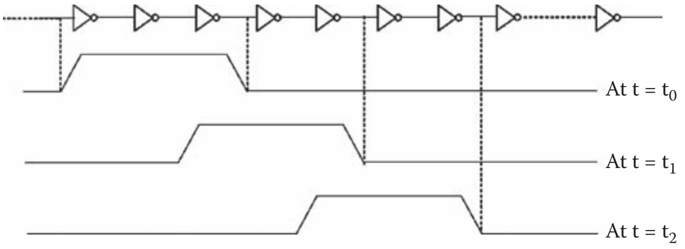


FIGURE 13.6 Pulse propagation through a series of inverters. Time instances t_0, t_1, and t_2 are two inverter delays apart.

circuit described here characterizes the SET pulse width in units of inverter delays. Pulse width is defined as the width of the voltage pulse measured at the inverter threshold ($V_{dd}/2$). If an SET pulse of sufficient duration excites an inverter chain, it will propagate through each inverter after a specific time delay (e.g., it will reach the third inverter after two inverter delays, the fifth inverter after four inverter delays). This is shown in Figure 13.6 where the leading edge of the transient pulse is shown to reach the inputs of inverters in a chain at different instances of time. As time progresses, this transient propagates through a series of inverters. Thus, at any instant of time, a certain number of inverters have their outputs affected/switched. This number of affected inverters is proportional to the transient pulse width. For extremely short pulses, the pulse gets attenuated as it propagates through logic gates. As discussed in [27], pulses wider than the sum of the logic transition times (rise and fall) of a gate propagate through the gate without attenuation, while pulses shorter than this transition time propagate with varying attenuation. The minimum pulse width required for propagation through multiple levels of logic is discussed in more detail later in this chapter.

13.3.2 Self-Triggered Transient Capture

Figure 13.7 illustrates an example of pulse propagation through a series of inverters when the SET pulse is three inverter delays long. The pulse affects three inverter outputs as it propagates through the chain. If the number of such inverters whose outputs

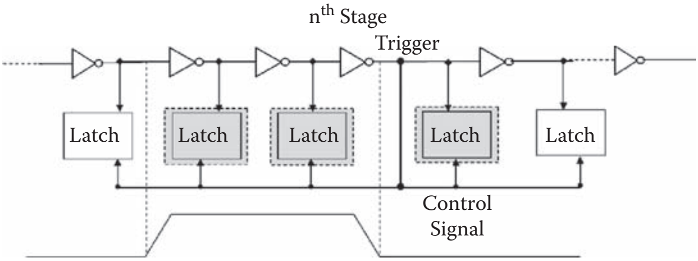


FIGURE 13.7 The output of the *n*-th stage can be used to provide hold signal for latches to freeze the data and the SET pulse.

are affected by the SET pulse can be determined at any instant, the pulse width can be estimated as a multiple of inverter delays. The number of inverters affected by the SET pulse is determined by the ratio of the SET pulse width to the individual stage delay. Simulations showed that for all pulse widths between [(n – 0.5) × stage delay] and [(n + 0.5) × stage delay], the number of affected stages is n. Thus the pulse width determined will be accurate to within ±half the propagation delay of an individual stage.

To measure the SET pulse originating from, for example, a target combinational logic circuit, it should first be fed to the measurement circuit composed of a chain of inverters. Next, the number of inverter stages affected by this SET pulse at any given instant of time must be measured. This can be accomplished if the SET pulse is frozen when it is within the measurement chain of inverters. Latches can be used to freeze the state of the inverter outputs at any given instant. Thus, to capture the affected outputs from a chain of inverters, the output of every inverter is connected to an asynchronous latch, as shown in Figure 13.7. As the SET pulse propagates through an inverter, the data stored in its respective latch will change. However, once the SET pulse passes, the inverter output and latch data will revert to their original states. (Note that the additional loading due to the latch at the inverter output will alter the pulse characteristics. Hence, capacitance at the latch input must be minimized and accounted for in the inverter delay for accurate measurement of pulse width.) If the latches are placed in a *hold* mode while the SET pulse is within the inverter chain, each latch will retain the logic state of its respective inverter.

For laser tests, the exact instant when the hit takes place is known, and the latches can be placed on *hold* after a certain delay, such that the SET pulse is guaranteed to be present within the inverter chain. However, for heavy-ion testing, information regarding the hit time and hit node are usually not available. To address autonomous operation in such cases, the output of an inverter stage in the measurement chain can be used as a trigger signal. This would cause additional loading at this inverter stage. However, as will be discussed later in this section, latches were used for both propagating the SET pulse and for its capture, and the trigger signal can be obtained from an inverter of the latch not directly in the path of the propagating SET pulse to minimize loading of the SET pulse. To make this circuit *self-triggering*, a transition at the output of the n-th stage (due to SET) can be used to trigger the latches to hold the states of the inverters, as shown in Figure 13.7. As the output of the n-th stage triggers the *hold* signal internally, precise information regarding the hit time (or location) is unnecessary. Any hit on stages beyond the trigger stage does not affect the trigger stage output. Thus, to latch an SET pulse, a hit must take place on a stage before the trigger stage.

The instant when the SET pulse is latched, the initial hit stage may or may not have recovered fully. If the initial stage has recovered fully when the pulse is latched, the pulse width measured is the actual pulse width (to within the accuracy of the measurement). However, if the initial stage has not recovered, it is possible that the charge collection is still continuing and the actual pulse width could be longer than the one measured. For heavy-ion tests, the hit stage is not identifiable, and hence it cannot be ascertained whether the hit stage has fully recovered. To address this uncertainty, a delay is introduced in the trigger signal. In addition, more inverter stages beyond the trigger stage are added to allow the SET pulse to propagate further.

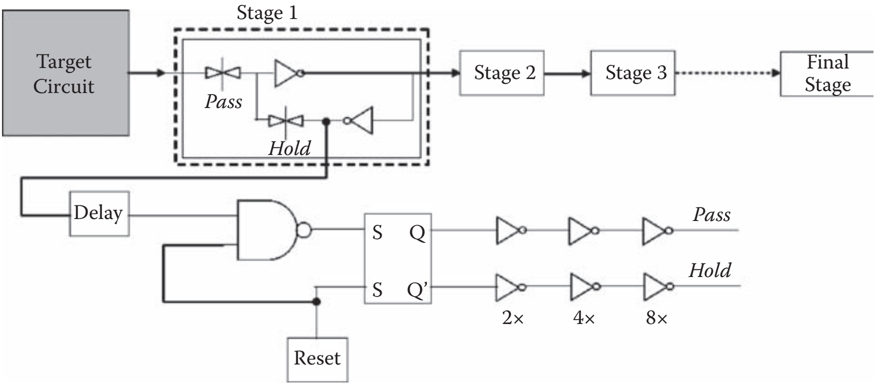


FIGURE 13.8 Test structure showing individual stages along with the trigger/reset circuit. Highlighted region shows the internal circuit of individual stages. (Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H. Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2506–2511.)

Thus, the delay on the trigger signal allows the SET pulse to propagate beyond the trigger stage. When the delayed trigger signal latches the SET pulse, the SET pulse may have propagated beyond the trigger stage. How far the SET pulse travels along the inverter chain is determined by the delay in the trigger signal. The delay in the trigger signal should be equal to the maximum SET pulse width expected for measurement. If the SET pulse has moved beyond the trigger stage, one can safely say that the estimated pulse width is the actual pulse width (within the accuracy of the measurement) irrespective of the hit node. This is because a hit on a stage beyond the trigger stage cannot initiate a latching process.

To increase the probability that an SET will be created in a given test environment, an array of target circuits, which functions as the source of SETs, precedes the measurement circuit, as shown in Figure 13.8. The target circuit also allows the trigger signal for the latches to be taken from the first stage of the measurement circuit and delayed in time to allow the SET to propagate completely into the measurement chain of inverters. Depending on the designer's requirement, the target circuit can be composed of any combinational logic network. However, the design of the target circuit should ensure that the target circuit does not affect the pulse shape or modify the generated SET pulse. Generally, a minimum drive-strength inverter chain should yield SETs similar to those in standard ICs and such a chain should propagate SETs with little attenuation.

13.3.3 Pulse Capture Circuit Design

A design of the complete test circuit (composed of the target circuit and the pulse measurement circuit) is shown in Figure 13.8. (To simplify the circuit and reduce loading effects, the individual inverter stages in the measurement circuit may be implemented using conventional CMOS pass-gate latches.) The operation of the test circuit is straightforward. An energetic particle hit in the target circuit creates an SET pulse that propagates to the measurement circuit. The measurement circuit essentially forms a series of latches that freeze the SET pulse for measurement (Figure 13.8). The latches in the measurement circuit are initially in the SET-propagate phase. During the SET-propagate phase, the pass signal is ON and the hold signal is OFF, which allows a pulse to propagate through the measurement chain of inverters and pass-gates. When the leading edge of the SET pulse reaches the first stage of the measurement circuit, it creates a trigger signal that is delayed in time, and hence the SET pulse continues to propagate through the inverters and pass-gates. Note that the trigger signal is obtained from an inverter in the latch that is not directly in the SET propagation path to minimize loading of the SET pulse. When the trigger signal reaches the SR flip-flop, it turns off all pass-gates by inverting the pass signal and freezing the data in the latches by turning on the hold signal. The SET pulse width is directly proportional to the number of latches whose output is affected. Once the latch outputs have been read out, a reset signal may be used to initialize the pass and hold signals and make the circuit ready for measuring the next pulse.

13.3.4 Illustration of Pulse Capture

Figure 13.9 illustrates measurement of an SET pulse. The chain of inverters shown in Figure 13.9 represents the inverters in pulse capture latch stages that propagate the SET. An SET pulse that is about four times the propagation delay of an inverter stage is input to this circuit. As soon as the leading edge of the SET arrives at the output of the first stage, it triggers a control signal. Since the control signal is delayed in time, it allows the SET to propagate beyond the first stage. Finally when the control signal triggers the pulse capture latches, it causes the SET to freeze somewhere within the chain of inverters. The waveforms in Figure 13.9 indicate that the SET pulse is between stages 16 and 19, as their outputs have a flipped state. The output of stage 15 returns back to its original state, and the SET pulse has not reached stage 20. In this case the SET pulse width would be estimated as four times the propagation delay of a single latch stage, which is the width of the input SET.

Latch upsets due to direct ion hits on the latches can corrupt the measurement. The total sensitive area of the target circuit is significantly larger than the total sensitive area of the latches in the data path. This, along with the fact that only latch upsets occurring after a SET event gets captured but before the data are read out (and the circuit is reset) are a concern, implies that latch errors can be neglected. This is because the latches themselves act like a chain of inverters that propagate the SET pulse until a trigger signal causes it to latch the data. Once data are latched, they are immediately read out, and a reset pulse causes the latch to return to the pulse propagating phase during which time a latch upset would result in only an SET. Since the frequency of operation is much higher than the rate at which events are created in this process, the probability of a latch upset during the time interval when an SET is captured and before the circuit is reset is very low. Also, most of these latch upsets can be identified by looking at the data pattern and hence can be discarded. The data pattern without any SET will be a string of alternating 1's and

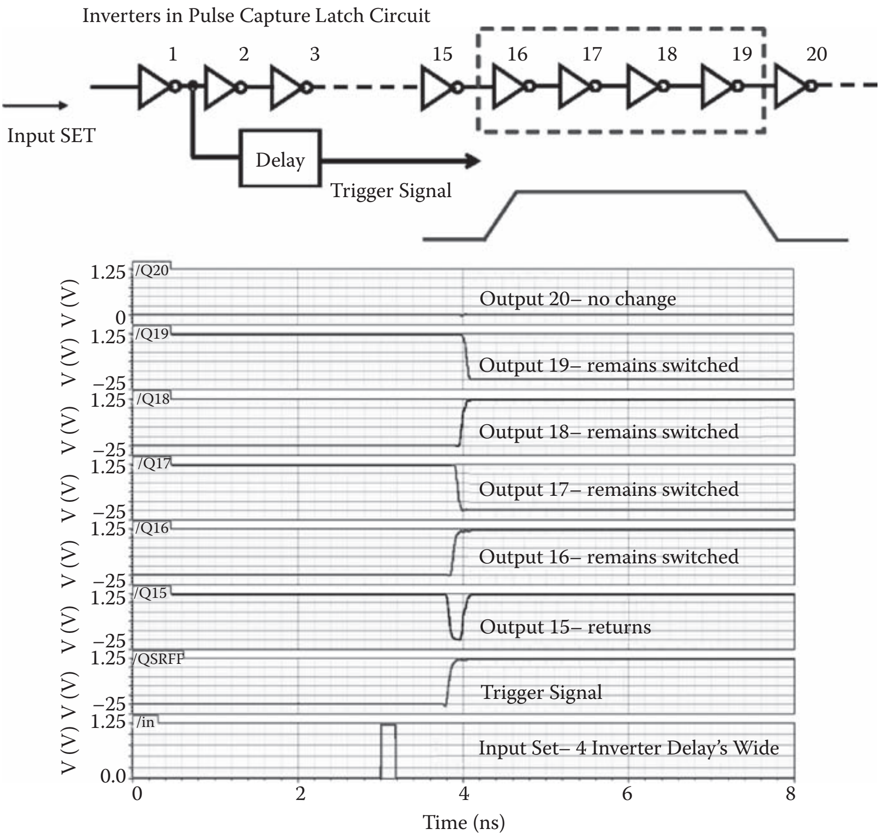


FIGURE 13.9 Simulation results illustrating capture of an SET pulse. SET pulse width is proportional to number of latches with a flipped state.

0's: "101010101010…". If an SET pulse that is five stages wide is created, then the output looks like "101011010110…". The start and end of an SET pulse is marked by two consecutive stages having the same value (except when the pulse is only a single stage wide). Latch upsets that cause the output data to read differently can easily be identified—for example, if the second latch is upset in the previous example during the data-read phase, then the data would read "111011010110…"—which does not correspond with a normal SET event, and hence such data can be discarded. For the experimental measurements, no erroneous data patterns of this type were observed for any of the multitude of experiments conducted for any of the technologies.

The latches in the data path are asynchronous and do not use a clock signal. Rather they are controlled or triggered by the SET pulse. The control logic consists of an SR flip-flop that provides the trigger signal to the latches and is also controlled by the SET pulse. A direct strike on this would result only in triggering a measurement. However, if no SET event has occurred, this would result in measuring the standard sequence of alternate 1's and 0's corresponding to outputs of a chain of inverters and would indicate a false measurement.

Finally, a parallel-in-serial-out shift register is used to serially output the data stored in the pulse capture latches. This shift register operates on the negative edge of an external clock signal. This shift register is also sensitive to strikes only during the time interval an SET event is captured but before the data are read out. Any strikes on the latches or on the clock buffers during such a time interval can affect the data that is read out. As explained earlier, by examining the data pattern most of these errors can also be identified.

13.3.5 Test Chip Designs

Integrated circuits with the aforementioned test structure were designed and fabricated in 1.5 μm, 0.35 μm, 180 nm, 130 nm, and 90 nm bulk CMOS processes and in an 180 nm SOI process. These designs are similar except for the number of inverters in the target circuit and the number of stages in the measurement circuit. Results from the 130 nm and 90 nm bulk processes are discussed here. Propagation and attenuation of SET pulses through the target and measurement circuits were analyzed using the Cadence Spectre simulator [28]. The parasitic resistances and capacitances were extracted from the layout and were included in the circuit simulations. The delay of a single latch stage was found to be about 65 ps in the 130 nm process and about 55 ps in the 90 nm process based on circuit simulations. The delay of a single-inverter stage is about 25 ps in the 130 nm process and about 21 ps in the 90 nm process. This indicates that additional loading has considerably increased the delay of the pulse-measurement latch stages.

As mentioned earlier, Massengill et al. have identified the minimum pulse width for infinite propagation to be the sum of the characteristic rise and fall time of a logic gate [27]. A minimum pulse width equal to the sum of the logic transition times is required to ensure the full rail-to-rail swing that is needed for unattenuated propagation [27]. As the ratio of the logic transition time to the propagation delay of a gate is a constant, the minimum pulse width can also be expressed in units of the propagation delay time. Since ring oscillator measurements can be used to obtain the propagation delay of a logic gate, the analysis presented next can be used to identify the minimum pulse width that would propagate through the SET pulse capture circuits. The rise and fall times for an inverter for logic swing between 10% and 90% of the supply voltage and the propagation delay times (low-to-high and high-to-low) can be expressed using the following first-order equations [29].

$$t_{rise} = R_P \times C_L \times \ln(9)$$

$$t_{fall} = R_N \times C_L \times \ln(9)$$

$$t_{plh} = R_P \times C_L \times \ln(2)$$

$$t_{phl} = R_N \times C_L \times \ln(2)$$

where

t_{rise} and t_{fall} = rise, fall times

t_{plh} and t_{phl} = propagation delay low-to-high and high-to-low

R_P = equivalent PMOS resistance

R_N = equivalent NMOS resistance

C_L = load capacitance

For a symmetric design, $R_N = R_P$ and hence $t_{rise} = t_{fall}$ and $t_{plh} = t_{phl}$. From the previous equations, the ratio of the logic transition time to the propagation delay time is ln(9)/ln(2), which is about 3.1. Circuit simulations with higher-order effects included indicate that the ratio of the logic transition time to the propagation delay time is about 2 for minimum-sized inverters designed in the 130 nm and 90 nm processes. Thus the minimum pulse required for unattenuated propagation in these processes can also be expressed as follows:

$$\begin{aligned}\text{Min Pulse Width} &= t_{rise} + t_{fall}\\ &= 2\times(t_{plh} + t_{phl})\\ &= 4\times t_p,\ \text{assuming } t_{plh} = t_{phl} = t_p\end{aligned}$$

Thus, the minimum pulse width for unattenuated propagation through an infinite number of logic gates is about four times the propagation delay of a single logic gate. The propagation delay through the pulse-capture latches was found to be about two and a half times that of the propagation delay through the inverter stages due to additional loading in the pulse-capture latches. Thus, the minimum pulse width for propagation is determined by the pulse-capture latch stages in the SET measurement circuit. Simulations showed that SET pulses greater than approximately three times the propagation delay of a single measurement latch stage propagated with less than about 10% attenuation through the 32 measurement latch stages in both the 130 nm and 90 nm processes. Thus, for such SETs the measured width, within the accuracy of measurement, is equal to the actual SET width. SET pulses less than this width are attenuated by greater amounts, depending on the initial pulse width. Figure 13.10 shows a plot of the transient pulse width normalized to an individual measurement stage delay as a function of the logic stage number for propagation through 50 identical stages.

A ring oscillator consisting of pulse-measurement circuit latch stages was fabricated to obtain the precise delay of an individual latch stage, as shown in Figure 13.11. The design of the ring oscillator and its output waveform are shown in Figures 13.11a and 13.11b. This delay was measured to be about 120 ps for the 130 nm process when operating at the nominal supply of 1.2 V. For the 90 nm process, the individual stage delay was found to be about 100 ps. The nominal operating voltage for the 90 nm process is also 1.2 V. The measured delays are about a factor of two longer than the values obtained through circuit simulations. Since parasitic resistances and capacitances were included in the circuit simulations, lower drive currents for the fabricated devices may be responsible for the observed longer delays.

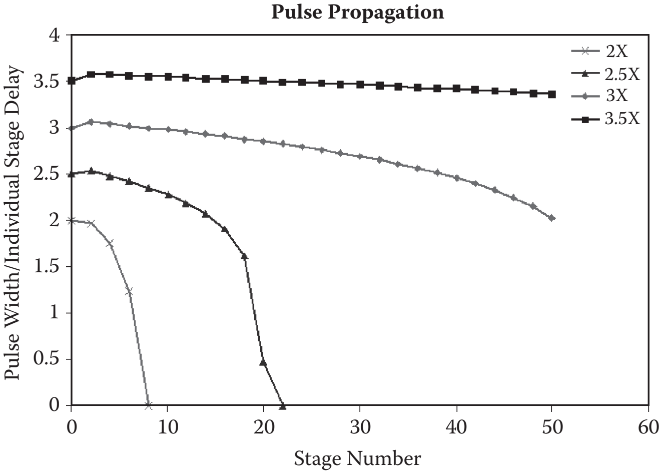


FIGURE 13.10 Propagation of a transient pulse through a long chain of identical logic gates. The transient pulse width normalized to an individual gate delay is plotted as a function of the logic stage number.

When the trigger signal was enabled in the measurement circuit, the leading edge of the pulse was latched at the 22nd stage. This enabled pulse widths to be measured from 120 ps (1 stage) to about 2,520 ps (21 stages, excluding the first stage) for the 130 nm process. The measurement range for the 90 nm process is from 100 ps to about 2.1 ns. The accuracy of measurement is about ±½ the individual latch stage delay.

13.4 HEAVY-ION TEST RESULTS

A particle accelerator, such as a cyclotron, is a reliable way to characterize space radiation effects on ICs using terrestrial experiments. A practical ion test uses a medium-energy particle accelerator to simulate galactic cosmic rays in space-radiation environments. The ability of an ionized particle to interact with materials is a function of its LET value. LET is essentially the measure of ionizing energy deposited in a material per distance traveled, generally in units of MeV-cm^2/mg. For particles in space, the range of LET varies primarily from a few hundredths to just under 100 MeV-cm^2/mg. Particles with low LET values are far more abundant than particles with high LET.

The SET test circuits fabricated in the 130 nm and 90 nm process were tested with heavy ions at different cyclotron facilities. Figure 13.12 shows a picture of the heavy-ion test setup at one of the facilities. Test results showing the distribution of SET pulse widths for various ions are discussed in this section.

13.4.1 Heavy-Ion Tests, 130 nm

The 130 nm ICs were tested with heavy ions at the cyclotron facility at Lawrence Berkeley National Laboratory [30]. The circuit was tested with ions at various

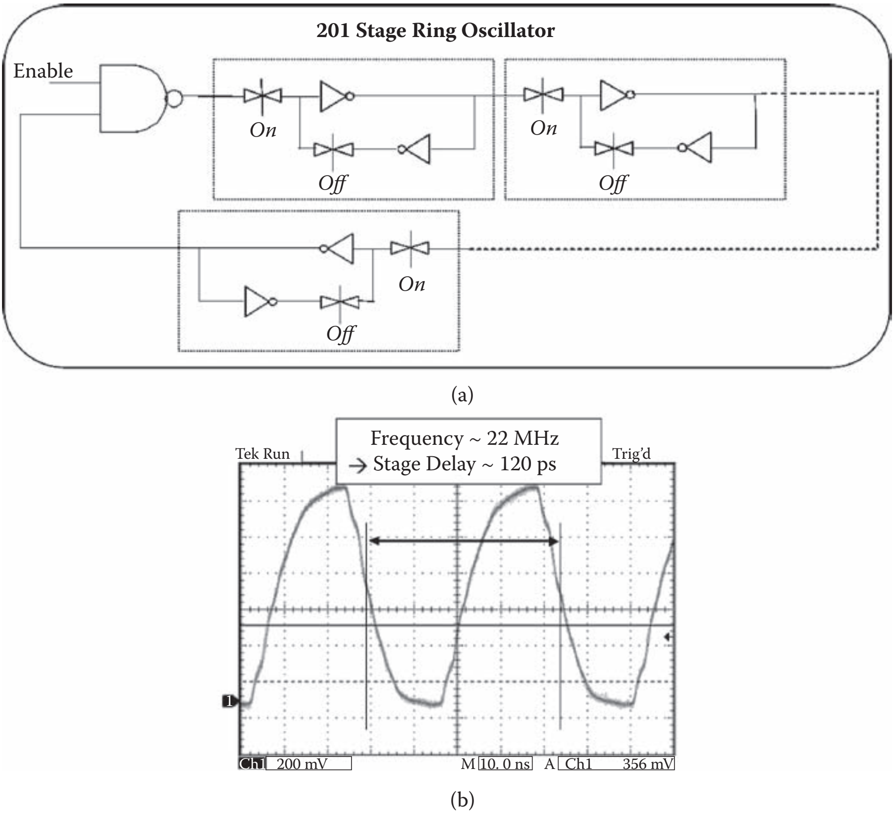


FIGURE 13.11 (a) Ring oscillator design composed of the pulse measurement circuit latch stages. (b) Output of the ring oscillator designed in the 130 nm process measured using an oscilloscope. It indicates that the delay of a single latch stage is about 120 ps.

angles to achieve an effective LET range from about 3.5 to 100 MeV-cm^2/mg (see Table 13.1). Effective LET is calculated as the ratio of the actual LET to the cosine of the angle of ion incidence relative to the perpendicular to the die surface. At each LET, the IC was tested to a fluence of 1 × 10^8 ions/cm^2. At an LET of 3.5 MeV-cm^2/mg no SET events were measured, and at 7 MeV-cm^2/mg only a statistically insignificant number of events were recorded. Figure 13.13a is a box plot representing the average SET width, the standard deviation, and the minimum and maximum SET widths for a range of LETs. The standard deviation data from Figure 13.13a clearly show that most of the SET pulses created are below 1 ns. Figure 13.13b shows plots of the number of SETs measured at each LET and the total SET cross section per inverter, which is the ratio of the total number of SETs measured at each LET to the fluence divided by the number of target inverters.

Figure 13.14 shows a histogram of the distribution of the event cross section per inverter as a function of LET. Event cross section per inverter is defined here as the ratio of the number of measured SET pulses with a given width to the total fluence divided by the number of target inverters. The SET pulse width for a given effective

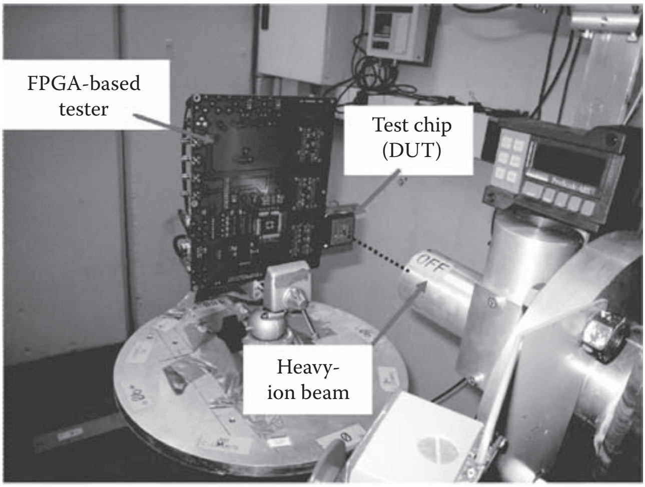


FIGURE 13.12 Picture of the heavy-ion test setup showing the DUT and FPGA-based tester boards. The DUT board is placed in line with the ion beam.

TABLE 13.1
Details of the Heavy-Ion Test, 130 nm

Ion	Angle (deg)	Effective LET (MeV-cm^2/mg)	Ion Energy (MeV)
Ne	0	3.45	216
Ne	60.5	7	216
Ar	0	9.7	400
Ar	60.9	20	400
Kr	0	31.2	886
Kr	49.3	48	886
Xe	0	58.7	1403
Xe	38.5	75	1403
Xe	54	100	1403

Source: Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." IEEE Transactions on Nuclear Science, Volume: 54 Issue: 6, 2506–2511.

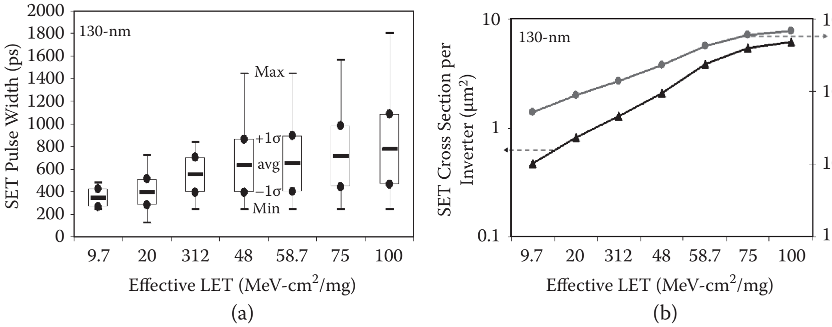


FIGURE 13.13 (a) Box plot indicating the average, ±1 standard deviation, minimum, and maximum SET pulse width as a function of LET for the 130 nm process. (b) Total SET cross-section per inverter and the number of events measured as a function of effective LET. (Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2506–2511.)

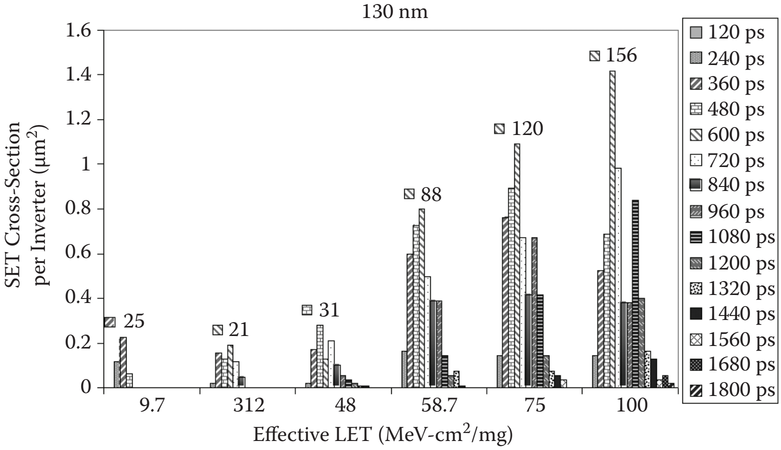


FIGURE 13.14 Distribution of event cross-section per inverter as a function of LET for the 130 nm process. The data labels on the chart indicate the maximum number of SET pulses of a given width measured at each LET. (Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2506–2511.)

LET is not a constant and varies over a wide range. This is because the pulse width created depends on the collected charge, which varies depending on the location of the strike with respect to the sensitive drain.

Previous work indicates that the distribution of the collected charge follows a Gaussian profile [26,31]. In [26] the collected charge by a circuit node after an ion strike was measured directly. In [31], Monte Carlo based simulations were used to show the distribution of amplitude and duration of the current transient due to charge collection for 63 MeV neutron interactions in silicon. The collected charge was then computed using the current transient waveforms and the distribution of the collected charge looks similar to a Gaussian profile. The distribution of collected charge correlates directly with the SET distribution observed in this work. The collected charge distribution indicates that strikes within the drain region lead to higher amounts of charge being collected and that strikes farther away from the drain lead to lower charge collection. Thus, one may expect to see many short transients since the area around the drain region where strikes lead to relatively low amounts of charge being collected is expected to be larger than the drain region. The reason for not observing a higher number of short transients may be due to the fact that many such transients are attenuated completely before measurement. As discussed earlier, transients that are less than about three times the propagation delay of a single latch stage, which corresponds to less than about 350 ps for the 130 nm process, may be partially or completely attenuated as they propagate through the measurement-latch stages. Hence, the lower end of the SET pulse-width distribution may extend farther and may contain more events. Based on the work of DasGupta et al., it is also likely that parasitic-bipolar charge collection may be an issue for the considered fabrication process [5]. Such a charge-collection mechanism may increase the amount of charge collected, leading to an increase in the number of wider transients.

The use of the autonomous pulse-capture technique enables most of the created SET pulses greater than about two to three times the delay of a single latch stage to be measured (except the ones that are created when reading the data, which are negligible as the frequency of operation was much higher than the rate at which SETs were created). Moreover, while temporal latch-based or guard gate-based techniques count all SET pulses greater than a certain width, the autonomous SET characterization measures the individual SET width and thus enables the precise estimation of event cross section for individual pulse widths greater than three times the delay of a latch stage at each LET.

13.4.2 Heavy-Ion Tests, 90 nm

The 90 nm ICs were tested at the cyclotron facility at Texas A&M University [30]. Extensive tests with different ions, all at normal incidence, were carried out. As the ion energy and LET change with the distance traversed in a material, the LET of an ion can be modified by adding degraders in the path of the beam before it reaches the test IC. As listed in Table 13.2, two different LETs—one obtained without using degraders and the other by using a degrader in the path of the beam—were obtained for each ion. The ICs were tested to a fluence of 1 × 10^8 ions/cm^2 at each LET.

TABLE 13.2
Details of the Heavy-Ion Test, 90 nm

Ion	Angle (deg)	LET (MeV-cm^2/mg)	Ion Energy (MeV)
Ne	0	1.8	526
Ne*	0	3	263
Ar	0	5.7	929
Ar*	0	9	468
Kr	0	20.6	1858
Kr*	0	30	860
Xe	0	40.7	2758
Xe*	0	59	824

Source: Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M.,; Eaton, P.H,. "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." IEEE Transactions on Nuclear Science, Volume: 54 Issue: 6, 2506–2511.

* Degrades used in the path of the ion beam to vary ion LET.

Figure 13.15a is a box plot representing the average SET width, the standard deviation of the SET width population, and the minimum and maximum SET widths for a range of ion LETs in the 90 nm technology. From Figure 13.15a, it can be seen that the threshold for SET events in the 90 nm process is less than 2 MeV-cm^2/mg compared with about 7 MeV-cm^2/mg for the 130 nm process. While the maximum SET width shows a slight dependence on the LET, for the most part the range of SET widths shows little dependence on the ion LET. The likely reasons for the observed distribution of SET pulse widths are discussed in the next section. Figure 13.15b shows plots of the number of events measured and the total SET cross section per inverter as a function of LET. Similar to the results for the 130 nm technology, the number of SET events measured strongly depends on the ion LET. As discussed earlier, since the individual latch delay is about 100 ps for the 90 nm process, transients less than about 250 ps to 300 ps may have been partially or completed attenuated. Thus, the lower end of the distribution may contain more events than those that are captured by the pulse-measurement circuit.

Figure 13.16 shows a histogram of the distribution of the event cross section per inverter as a function of LET. As stated earlier, event cross section per inverter is defined here as the ratio of number of measured SET pulses with a given width to the total fluence divided by the number of target inverters. The histogram of the SET distribution shows that the odd-numbered bars (e.g., 100, 300, 500) contain more

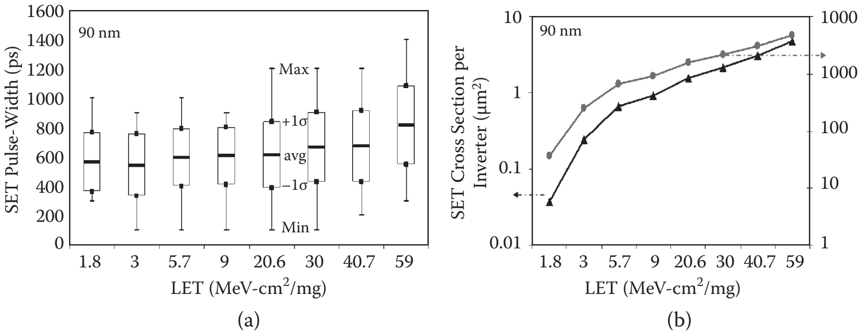


FIGURE 13.15 (a) Box plot indicating the average, ±1 standard deviation, minimum, and maximum SET pulse-width as a function of LET for the 90 nm process. (b) Total SET cross section per inverter and the number of events measured as a function of effective LET. (Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2506–2511.)

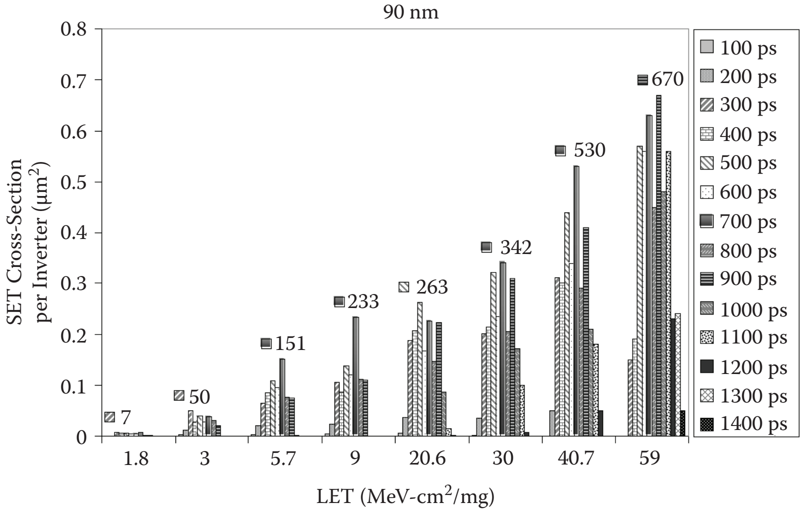


FIGURE 13.16 Distribution of event cross section per inverter as a function of LET for the 90 nm process. The data labels on the chart indicate the maximum number of SET pulses of a given width measured at each LET. (Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2506–2511.)

events than the even-numbered bars (e.g., 200, 400, 600). This shows that the number of stages affected by the SET has a slightly higher probability of being odd than even. The total number of events in the even-numbered bars was found to be about 20% to 30% lower than the number of events in the odd-numbered bars. The pulse widths for strikes on devices within the well region—that is, *p*-type metal-oxide semiconductor field-effect transistors (PMOSFETs)—have been shown to be longer than the pulse width for strikes on n-type MOSFETs (NMOSFETs) [32]. Thus, a convolution of the SET width distributions for the NMOSFETs and PMOSFETs can lead to a double peaked distribution with some variations in the number of events for the bars in the middle.

Furthermore, the periodicity in the measured distributions can be created by drive current variations between the PMOSFETs and NMOSFETs. The leading edge of the SET pulse that initiates the trigger signal was always latched at the 21st measurement latch stage. The SET pulse width should then determine the number of stages before stage 21 that have a flipped state. Due to the initial state of the circuit, the outputs of odd-numbered latch stages are initially high and vice versa. If the PMOSFETs have a lower drive current than the NMOSFETs, then it takes longer to transition from low to high than it does from high to low. If the SET pulse is, say, 3.55 times the individual stage delay, then it should, in theory, affect stages 21, 20, 19, and 18. In this case stage 18 is starting to recover back to its nominal state—that is, starting to transition from high to low. If this transition time is faster than expected, then the output of stage 18 can cross the threshold, resulting in the SET affecting only three stages. Similarly, it can be argued that a slower PMOSFET can result in an SET pulse that is 4.45 times the individual stage delay to be measured as five stages wide. Simulations with different PMOSFET and NMOSFET drive strengths also concur with the previously given explanations. Thus, variations in the drive currents can result in the observed periodicity in the number of measured SETs. However, such variations do not significantly affect the average and range of the SET pulse widths measured. One additional factor that can contribute to the variations in the drive currents is the amount of dose accumulated with testing the devices with heavy ions. For these tests, a total dose of a few hundred krads was accumulated in the tested devices. The parametric degradation associated with the total dose could lead to the types of drive imbalances already described.

13.4.3 Technology Scaling Trends Based on Heavy-Ion Experimental Results

A comparison of heavy-ion induced SET widths in 130 nm and 90 nm processes as a function of LET is shown in Figure 13.17. While the ion energies at the two test facilities used for these experiments may not be identical, it is still reasonable to compare the results based on LET as direct ionization events dominate over spallation secondary reaction events. Only data for normal incidence are plotted in Figure 13.17. For low to moderate LETs, the range of SET pulse widths in the 90 nm process is significantly larger than that of the 130 nm process, while they are comparable at higher LETs. A comparison of Figures 13.14 and 13.16 indicates that, in the 90 nm

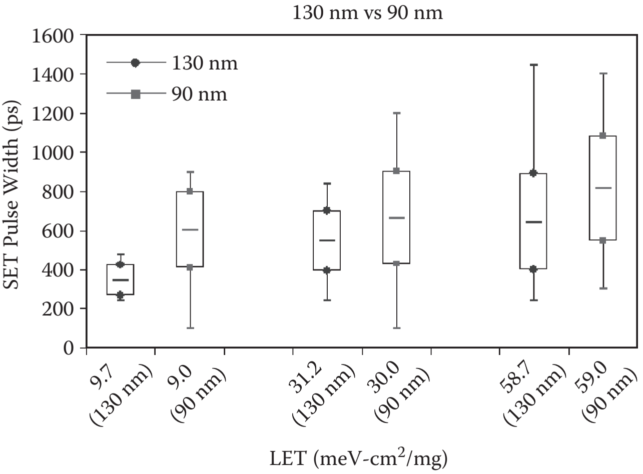


FIGURE 13.17 Box plot indicating the average, ±1 standard deviation, minimum, and maximum SET pulse-width in 130 nm and 90 nm process as a function of LET. Only data for normal incidence angle are plotted. (Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies." *IEEE Transactions on Nuclear Science*, Volume: 54 Issue: 6, 2506–2511.)

process, the SETs with the highest event section (~400 ps to ~900 ps) are wider than the most common events in the 130 nm process (~300 ps to ~700 ps). Assuming that the ions and LETs used for these experiments are comparable, it is evident that for this 90 nm technology the dominant SET pulse widths have increased compared with the 130 nm technology. The 130 nm and 90 nm processes used in this study have the same operating voltage of 1.2 V, and the circuits tested were identical except that their sizes were proportionately scaled from the 130 nm process to the 90 nm process. Since the combinational-logic soft errors may increase with increasing SET pulse widths, the increase in the number of wider transients with scaling suggests higher vulnerability for future technologies.

With the use of mixed-mode three-dimensional (3-D) technology computer-aided design (TCAD) simulations, the factors affecting SET pulse width were analyzed, and a brief discussion of the key points is presented here. Based on the simulations, reduction in the device drive current and a reduction in the minimum well contact size were found to increase the SET pulse width. This is understandable, since a reduction in the drive current translates to a reduction in the restoring drive capability, which in turn translates to longer time durations for neutralizing the charge deposited by the single event, leading to an increase in the SET width. The contact size reduction amplifies the well-potential-collapse effect, which increases parasitic-bipolar charge collection, leading to an increase in SET width. Finally, with scaling, the node capacitance also reduces. While the nodal capacitance affects the rise and fall times of the SET and the threshold or critical charge needed to create an SET, it has little effect on the charge neutralization process to restore the node back to its original state. Thus, the duration of the SET is primarily governed by charge-collection characteristics, and hence the capacitance does not have a major effect on the SET width.

The distribution of SET widths for the 90 nm technology indicates the presence of some long transients at low LETs compared to the 130 nm SET results. SET pulse-width distribution can be impacted by many factors, including the mechanism of charge deposition (whether direct or indirect ionization through secondary reaction products), initial ion track diameter, and circuit level effects such as drive currents. High-LET secondary reaction products lead to increased charge deposition, leading to wider transients. However, the probability of such events is relatively small. Initial ion track diameter has been shown to affect charge collection [33]. The drive currents for the fabricated device were found to be lower than the simulated drive currents by about a factor of two, and this could lead to longer restoring times, which in turn translate to wider SETs.

Finally, recent experimental evidence suggests the presence of propagation-induced pulse broadening due to a hysteresis or body-bias effect [27,34]. The sources of such an effect, if any, for well-contacted bulk devices are still not well understood. However, the presence of any pulse broadening through the target circuit will impact the measured distribution of pulse widths. It should be noted that the measurement circuit itself does not affect the measurement as the length of the measurement chain of gates is relatively small compared with the target circuit. A judicious design of the target circuit can thus easily avoid this problem.

13.5 NEUTRON AND ALPHA PARTICLE INDUCED TRANSIENTS

Atmospheric neutrons and alpha particles are the primary radiation sources that are of concern for terrestrial applications. Cosmic particles colliding with atoms in the atmosphere create cascades of neutrons, which in turn may interact with electronics, resulting in single events. Integrated circuits are also affected by alpha particles as a result of material contaminants in the packaging material such as uranium and thorium.

13.5.1 Neutron Induced SET Pulse Widths

The 90 nm test chips were tested separately with neutrons and alpha particles [35]. Accelerated high-energy neutron tests were performed at the Weapon Neutron Research (WNR) test facility at Los Alamos Neutron Science Center (LANSCE). This neutron energy spectrum, plotted in Figure 13.18, closely resembles the sea-level neutron spectrum for energies from 10 MeV to 500 MeV. Three circuit boards with two SET test chips per board were placed one behind another and normal to the path of the neutron beam. The center-to-center distance of the two SET test chips in each board was less than 2 inches, and the boards were placed such that both chips were covered by the neutron beam, which was 3 inches in diameter. Because of the low probability of interaction, the neutron beam penetrates through the circuit boards with minimum loss of flux, which enables testing of multiple chips at the same time. The total test time was about 102 hours, which resulted in

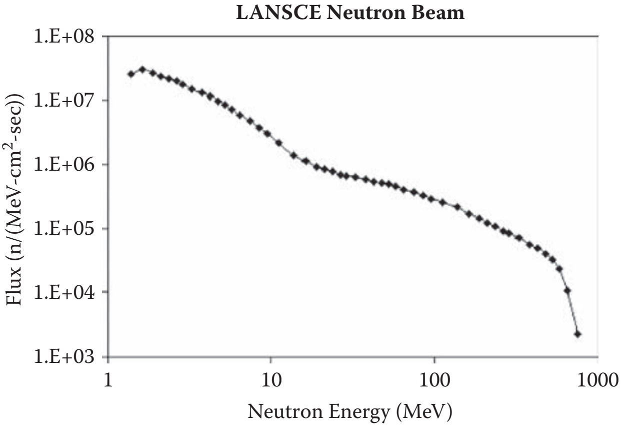


FIGURE 13.18 Energy spectrum of the LANSCE neutron beam. This spectrum closely resembles the energy spectrum of terrestrial neutrons.

a neutron fluence of 1.33 × 10^{11}/cm^2 based on the integration of neutron flux over the range of 10 MeV to 500 MeV. A total of 20 SET events ranging from about 300 ps to about 1.4 ns were measured during this test time. This converts to a neutron SET cross section of 1.6 × 10^{-6} μm^2/inverter. Based on the layout, the sensitive area of an inverter used in this design is about 0.75 μm^2. The low event rate is attributed to the small area of the target circuits and to the fact that neutrons ionize indirectly through secondary reaction products. Figure 13.19 shows the distribution of neutron-induced SET pulse widths.

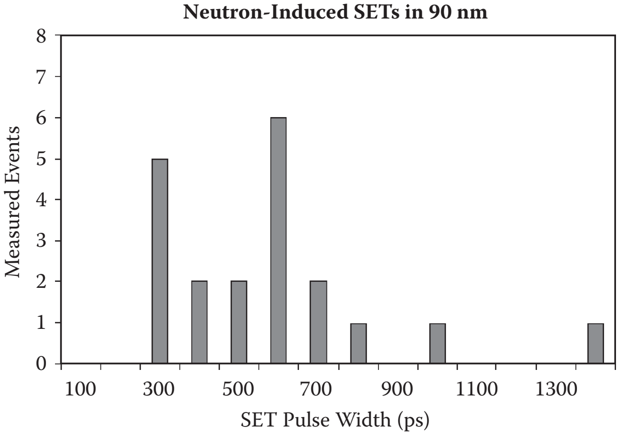


FIGURE 13.19 Neutron-induced distribution of SET pulses. (Reprinted with permission from Narasimham, B., Gadlage, M.J., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Holman, W.T, Witulski, A.F., Reed, R.A., Weller, R.A., Xiaowei Zhu, "Characterization of neutron- and alpha-particle-induced transients leading to soft errors in 90-nm CMOS technology." *IEEE Transactions on Device and Materials Reliability*, Volume: 9 Issue: 2, 325–333.)

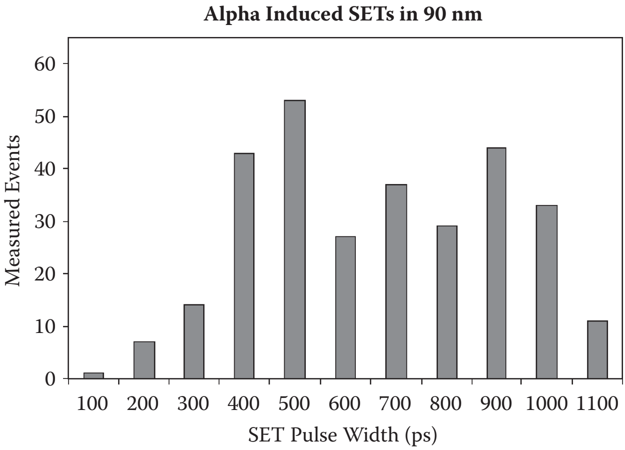


FIGURE 13.20 Alpha particle-induced distribution of SET pulses. (Reprinted with permission from Narasimham, B., Gadlage, M.J., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Holman, W.T., Witulski, A.F., Reed, R.A., Weller, R.A., Xiaowei Zhu, "Characterization of neutron- and alpha-particle-induced transients leading to soft errors in 90-nm CMOS technology." *IEEE Transactions on Device and Materials Reliability*, Volume: 9 Issue: 2: 325–333.)

13.5.2 Alpha Particle Induced SET Pulse Widths

Accelerated alpha particle tests were carried out at Texas Instruments using a foil of Americium-241 as the alpha source. The Americium-241 source was placed directly on top of the die while the device was operating and the transient pulses were recorded. The average energy of the alpha particles from this source is about 5.5 MeV. The total fluence of the alpha particles was estimated to be about 4.45 × 10^{10}/cm^2 and approximately 300 SET events were measured, which converts to an alpha particle SET cross section of about 6.74 × 10^{-4} μm^2/inverter. Figure 13.20 shows the distribution of alpha particle induced SET pulses.

The distribution of SETs clearly indicates that neutron and alpha particles can induce transients that are wide enough to be mistaken as valid logic or clock signals in the 90 nm node. These results imply that, as technology is scaled to lower voltages and higher operating frequencies, SETs may become a serious reliability problem.

A comparison of neutron, alpha, and heavy-ion SET pulse-width distributions is plotted in Figure 13.21. The LET values for the ions are specified in Figure 13.21. While the number of events varies with the particle type, the distribution of pulse widths was found to be similar for the different particle types. As expected, neutrons have the lowest event cross section as they ionize indirectly through secondary reaction products.

13.5.3 Neutron and Alpha FIT Rates

From the experimental SET cross section data, the FIT rate per inverter for this technology was estimated. The formula for computing the FIT rates is given by

$$\text{FIT/inverter} = \frac{\text{number of SETs measured} \times \text{particle flux} \times 10^9 \text{ hr}}{\text{total fluence} \times \text{number of target inverter cells}} \tag{13.1}$$

In this computation the number of SET events measured is derated to account for latch window and logical masking effects. This is because in a practical logic circuit design masking effects result in only a fraction of the SET events that are created being latched as errors, and hence the FIT rate computation needs to account for masking effects to be more accurate. Since SETs originating from a single chain of a target inverter circuit were measured in this work, the measurement does not account for logical masking. Similarly, all SETs that are not electrically attenuated are recorded, and hence the measurement does not account for latch-window masking effects. The measurement of the SET pulse widths, however, accounts for some electrical masking effects, and hence no additional electrical derating is accounted for. The probability of latching an SET pulse that is wider than the setup-and-hold time of a latch is simply given by the ratio of the SET pulse width to the clock period [36]. For this computation we assume a clock period of 1 GHz and that all SET pulses are wider than the latch setup-and-hold times. Logical masking may reduce the number of SET events that propagate to the latch element. Logical masking varies from circuit to circuit, and even for a given circuit

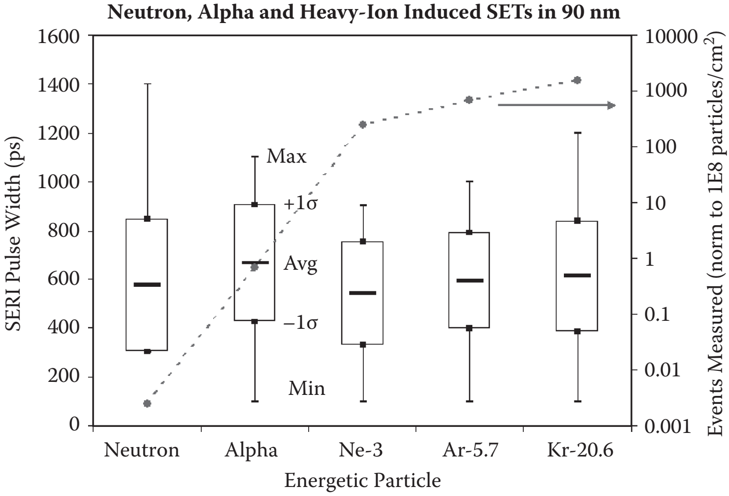


FIGURE 13.21 Box plot representing the average, minimum, maximum, and ±1 standard deviation in neutron, alpha, and heavy-ion induced SET pulse widths along with the number of measured SET events normalized to 1 × 10^8 particles/cm^2. The ion energies for Ne, Ar, and Kr are 263 MeV, 929 MeV, and 1,858 MeV, respectively. The ion's linear LETs are 3, 5.7, and 20.6 MeV-cm^2/mg for Ne, Ar, and Kr, respectively. (Reprinted with permission from Narasimham, B., Gadlage, M.J., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Holman, W.T., Witulski, A.F., Reed, R.A., Weller, R.A., Xiaowei Zhu, "Characterization of neutron- and alpha-particle-induced transients leading to soft errors in 90-nm CMOS technology." *IEEE Transactions on Device and Materials Reliability*, Volume: 9 Issue: 2, 325–333.)

TABLE 13.3
Alpha and Neutron FIT per Inverter

Particle	**Total Fluence (particles/cm^2)**	**Derated FIT/Inverter**
Alpha (from package)	4.45×10^{10}	1.1×10^{-5}
Neutron (sea level)	1.33×10^{11}	$4.4 \ 10^{-5}$

Source: Reprinted with permission from Narasimham, B., Bhuva, B.L., Schrimpf, R.D., Massengill, L.W., Gadlage, M.J., Amusan, O.A., Holman, W.T., Witulski, A.F., Robinson, W.H., Black, J.D., Benedetto, J.M., Eaton, P.H., "Characterization of neutron- and alpha-particle-induced transients leading to soft errors in 90-nm CMOS technology." IEEE Transactions on Device and Materials Reliability, Volume: 9 Issue: 2 : 325–333.

it depends on the inputs to the circuit. In [37], logical masking is analyzed in detail for different circuit types, and for sample computations the authors use logical masking values ranging from about 0.2 to 0.5. For this computation we chose a value of 0.5 for the logical masking.

The drain area of the target inverter circuit designed in this work was increased to increase the probability of creating SETs for measurement. The drain area was increased by a factor of about 3× compared with a minimum drain area layout. Thus, for standard layout practices, the SET cross section should be lower. The increase in the drain area is also accounted for in the computation of the FIT rates.

The average neutron flux is about 13 n/cm^2/hr at sea level [38], and average alpha particle flux from package impurities is of the order of 0.01 alpha/cm^2/hr. Using the derated number of SET events measured and the average flux values for neutrons and alpha particles, the FIT/inverter was calculated. Table 13.3 shows the FIT rate per inverter in this technology at sea level for neutrons and alpha particles. The reason the alpha and neutron FIT rates are similar despite the much higher counts in the alpha experiment is that the alpha flux is much lower in practice than the neutron flux. The neutron FIT/inverter value computed in this work is slightly higher but comparable to the projections made in [8]. In [8], the authors simulated the critical charge and charge collection efficiency for logic and memory cells in different technology nodes and used an empirical model to calculate the SER based on the critical charge and charge collection efficiency. Shivakumar et al. projected the neutron FIT/logic gate to be about 10^{-5} for a 100 nm technology node.

In this work, error rates have not been measured for memory or latch circuits. However, simulation-based projections made by Shivakumar et al. [8] indicate that the FIT/SRAM is of the order of 4×10^{-5} for a similar technology node. This is very similar to the FIT/inverter value computed in this work and indicates that the chip-level SER resulting from single-event transients may be a significant concern for some logic circuits. Other researchers have also characterized FIT rates for memory and latch circuits. However, most reported values are based on normalized units and hence cannot be directly compared with the FIT rates estimated in this work.

13.6 SUMMARY

With combinational logic soft errors projected to dominate the reliability issues of advanced semiconductor ICs, it is important to estimate the distribution of SET pulses for accurate determination of error rates and for developing appropriate mitigation techniques. These pulse widths are required inputs for detailed error-rate prediction methods.

This chapter presents an autonomous pulse detection and characterization technique for single-event transient pulse width measurements. The technique measures pulse width of individual SETs and results in characterizing the distribution of SET pulse widths for a given radiation environment. Circuit designs were implemented in IBM 130 nm and 90 nm bulk CMOS processes. Heavy-ion SET measurements show a reduction in the threshold for a measurable SET from about 7 MeV-cm^2/mg for 130 nm to less than 2 MeV-cm^2/mg for 90 nm, indicating the increase in vulnerability to single events with scaling. SET pulse widths ranging from about 100 ps to over 1 ns were measured in 130 nm and 90 nm processes, and the pulse widths were found to increase when scaling from 130 nm to 90 nm.

Neutron and alpha SET measurements in the 90 nm process show most such SETs to be of the order of hundreds of picoseconds. Neutron and alpha FIT rates were found to be about 10^{-5} FIT/inverter. The per-device FIT rates computed in this work correspond well with simulation-based projections made in [8] and are also comparable to the simulation estimates of per-bit SRAM and latch FIT rates in [8] for a similar technology node, indicating that logic SER will be an issue for certain terrestrial applications.

REFERENCES

1. R. C. Baumann, “Single event effects in advanced CMOS technology,” in *Proc. IEEE Nuclear and Space Radiation Effects Conf. Short Course Text*, 2005.
2. S. Buchner and M. Baze, “Single-event transients in fast electronic circuits,” in *Proc. IEEE Nuclear and Space Radiation Effects Conf. Short Course Text*, 2001.
3. P. E. Dodd and L. W. Massengill, “Basic mechanisms and modeling of single-event upset in digital microelectronics,” *IEEE Trans. Nucl. Sci.,* vol. 50, pp. 583–602, 2003.
4. C. M. Hsieh, P. C. Murley, and R. R. O’Brien, “A field-funneling effect on the collection of alpha-particle-generated carriers in silicon devices,” *IEEE Elec. Dev. Let.*, vol. 2, pp. 103–105, Dec. 1981.
5. S. DasGupta, A. F. Witulski, B. L. Bhuva, M. L. Alles, R. A. Reed, O. A. Amusan, et al., “Effect of well and substrate potential modulation on single event pulse shape in deep submicron CMOS,” *IEEE Trans. Nucl. Sci.*, vol. 54, pp. 2407–2412, Dec. 2007.
6. S. Buchner, M. Baze, D. Brown, D. McMorrow, and J. Melinger, “Comparison of error rates in combinatorial and sequential logic,” *IEEE Trans. Nucl. Sci.*, vol. 44, pp. 2209–2216, Dec. 1997.
7. N. Seifert, P. Shipley, M. D. Pant, V. Ambrose, and B. Gill, “Radiation induced clock jitter and race,” *IEEE Intl. Rel. Phy. Sym. Proceedings*, pp. 215–222, 2005.

8. P. Shivakumar, M. Kistler, S. W. Keckler, D. Burger, and L. Alvisi, “Modeling the effect of technology trends on the soft error rate of combinational logic,” *IEEE Intl. Conf. on Dep. Sys. Net. Proc.*, pp. 389–398, 2002.
9. N. Seifert, P. Slankard, M. Kirsch, B. Narasimham, V. Zia, C. Brookreson, et al., “Radiation-induced soft error rates of advanced CMOS bulk devices,” *IEEE Intl. Rel. Phy. Sym. Proceedings*, pp. 217–225, 2006.
10. L. W. Massengill, A. Baranski, D. Van Nort, J. Meng, and B. Bhuva, “Analysis of single-event effects in combinational logic—Simulation of the AM2901 bitslice processor,” *IEEE Trans. Nucl. Sci.*, vol. 47, pp. 2609–2615, Dec. 2000.
11. N. Seifert, X. Zhu, D. Moyer, R. Mueller, R. Hokinson, N. Leland, et al., “Frequency dependence of soft error rates for deep sub-micron CMOS technologies,” *IEDM Technical Digest,* pp. 14.4.1–14.4.4, 2001.
12. D. G. Mavis and P. H. Eaton, “Soft error rate mitigation techniques for modern microcircuits,” *IEEE Intl. Rel. Phys. Symp. Proc.*, pp. 216–225, 2002.
13. J. Benedetto P. Eaton, K. Avery, D. Mavis, M. Gadlage, T. Turflinger, et al., “Heavy ion induced digital single event transients in deep submicron processes,” *IEEE Trans. Nucl. Sci.*, vol. 51, pp. 3480–3485, Dec. 2004.
14. M. Nicolaidis and R. Perez, “Measuring the width of transient pulses induced by ionizing radiation,” *IEEE Intl. Rel. Phy. Sym. Proceedings*, pp. 56–59, 2003.
15. L. W. Massengill, “SEU modeling and prediction techniques,” in *Proc. IEEE Nuclear and Space Radiation Effects Conf. Short Course Text*, 1993.
16. P. E. Dodd, “Basic mechanisms for single-event effects,” in *Proc. IEEE Nuclear and Space Radiation Effects Conf. Short Course Text*, 1999.
17. A. H. Johnston, T. Miyahira, G. Swift, S. Guertin, and L. Edmonds, “Angular and energy dependence of proton upset in optocouplers,” *IEEE Trans. Nucl. Sci.*, vol. 46, pp. 1335–1341, Dec. 1999.
18. R. A. Reed, P. J. McNulty, and W. G. Abdel-Kader, “Implications of angle of incidence in SEU testing of modern circuits,” *IEEE Trans. Nucl. Sci.*, vol. 41, pp. 2049–2054, Dec. 1994.
19. P. E. Dodd, M. R. Shaneyfelt, J. A. Felix, and J. R. Schwank, “Production and propagation of single-event transients in high-speed digital logic ICs,” *IEEE Trans. Nucl. Sci*, vol. 51, pp. 3278–3284, Dec. 2004.
20. P. Eaton, J. Benedetto, D. Mavis, K. Avery, M. Sibley, M. Gadlage, et al., “Single event transient pulsewidth measurements using a variable temporal latch technique,” *IEEE Trans. Nucl. Sci.*, vol. 51, pp. 3365–3368, Dec. 2004.
21. M. P. Baze, J. Wert, J. W. Clement, M. G. Hubert, A. Witulski, O. A. Amusan, et al., “Propagating SET characterization technique for digital CMOS libraries,” *IEEE Trans. Nucl. Sci*, vol. 53, no. 6, pp. 3472–3478, Dec. 2006.
22. J. M. Benedetto, P. H. Eaton, D. G. Mavis, M. Gadlage, and T. Turflinger, “Digital single event transient trends with technology node scaling,” *IEEE Trans. Nucl. Sci*, vol. 53, no. 6, pp. 3462–3465, Dec. 2006.
23. H. Schone. D. S. Walsh, F. W. Sexton, B. L. Doyle, P. E. Dodd, J. F. Aurand, et al., “Time-resolved ion beam induced charge collection (TRIBICC) in micro-electronics,” *IEEE Trans. Nucl. Sci.*, vol 45, pp. 2544–2549, Dec. 1998.
24. V. Ferlet-Cavrois, P. Paillet, A. Torres, M. Gaillardin, D. McMorrow, J. S. Melinger, et al., “Direct measurement of transient pulses induced by laser and heavy ion irradiation in deca-nanometer SOI devices,” *IEEE Trans. Nucl. Sci.*, vol. 52, pp. 2104–2113, Dec. 2005.
25. M. Gadlage, R. Schrimpf, J. Benedetto, P. Eaton, D. Mavis, M. Sibley, et al., “Single event transient pulsewidths in digital microcircuits,” *IEEE Trans. Nucl. Sci.*, vol. 51, pp. 3285–3290, Dec. 2004.

26. V. Ferlet-Cavrois, P. Paillet, M. Gaillardin, D. Lambert, J. Baggio, J. R. Schwank, et al., "Statistical analysis of the charge collected in SOI and bulk devices under heavy lon and proton irradiation—implications for digital SETs," *IEEE Trans. Nucl. Sci.*, vol. 53, no. 6, pp. 3242–3252, Dec. 2006.
27. L. W. Massengill and P. W. Tuinenga, "Single-event transient pulse propagation in digital CMOS," *IEEE Trans. Nucl. Sci*., vol. 55, pp. 2861–2871, Dec. 2008.
28. *Cadence Spectre® Circuit Simulator User Guide,* Sept. 2003.
29. J. P. Uyemura, *CMOS logia circuit design,* Springer, 1999.
30. B. Narasimham, B. L. Bhuva, R. D. Schrimpf, L. W. Massengill, M. J. Gadlage, O. A. Amusan, et al., "Characterization of digital single event transient pulse widths in 130 nm and 90 nm CMOS," *IEEE Trans. on Nucl. Sci*., vol. 54, pp. 2506–2511, Dec. 2007.
31. G. Hubert, A. Bougerol, F. Miller, N. Buard, L. Anghel, T. Carriere, et al., "Prediction of transient induced by neutron/proton in CMOS combinational logic cells," *IEEE On-line Testing Symposium*, 2006.
32. B. D. Olson, O. A. Amusan, S. Dasgupta, L. W. Massengill, A. F. Witulski, B. L. Bhuva et al., "Analysis of parasitic PNP bipolar transistor mitigation using well contacts in 130 nm and 90 nm CMOS technology," *IEEE Trans. Nucl. Sci*., pp. 894–897, 2007.
33. W. J. Stapor, P. T. McDonald, A. R. Knudson, A. B. Campbell, and B. G. Glagola, "Charge collection in silicon for ions of different energy but same linear energy transfer," *IEEE Trans. Nucl. Sci*., vol. 35, pp. 1585–1590, Dec. 1988.
34. V. Ferlet-Cavrois, V. Pouget, D. McMorrow, J. R. Schwank, N. Fel, F. Essely, et al., "Investigation of the propagation induced pulse broadening (PIPB) effect on single event transients in SOI and bulk inverter chains," *IEEE Trans. Nucl. Sci.*, vol. 55, pp 2842–2853, Dec. 2008.
35. B. Narasimham, M. J. Gadlage, B. L. Bhuva, R. D. Schrimpf, L. W. Massengill, W. T. Holman, et al., "Characterization of neutron and alpha particle-induced transients leading to soft errors in 90-nm CMOS technology," *IEEE Trans. on Dev. and Mat. Rel.,* vol. 9, pp. 325–333, June 2009.
36. Y. Yanagawa, D. Kobayashi, K. Hirose, T. Makino, H. Saito, H, Ikeda, et al., "Experimental verification of scan-architecture-based evaluation technique of SET and SEU soft error rates at each flip-flop in logic VLSI systems," *IEEE Trans. Nucl. Sci.,* vol. 56, pp. 1958–1963, Aug. 2009.
37. H. T. Nguyen, Y. Yagil, N. Seifert, and M. Reitsma, "Chip-level soft error estimation method," *IEEE Trans. on Dev. and Mat. Rel.,* vol. 5, pp. 365–381, 2005.
38. M. S. Gordon P. Goldhagen, K. P. Rodbell, T. H. Zabel, H. K. Tang, J. M. Clem, et al., "Measurement of the flux and energy spectrum of cosmic-ray induced neutrons on the ground," *IEEE Transactions on Nuclear Science*, pp. 3427–3434, 2004.

14 Soft Errors in Digital Circuits: Overview and Protection Techniques for Digital Filters

Pedro Reviriego Vasallo and
Juan Antonio Maestro

CONTENTS

14.1 INTRODUCTION

This chapter covers the issue of soft errors on digital circuits focusing on ad hoc protection techniques that are illustrated using digital filters as a case study. It starts with a review of soft errors, covering how they are produced. Then different types of errors, both destructive and nondestructive, are reviewed. In the second part of the chapter, several methodologies to detect errors and measure the system reliability are described, together with the most usual techniques to protect digital circuits against soft errors. As a continuation to these, some ad hoc techniques for digital filters are presented as a case study in the last part of the chapter. These techniques show that efficient protection can be achieved with low overhead by exploiting the filter structure and the application requirements in terms of fault tolerance.

14.2 RADIATION EFFECTS ON ELECTRONIC DEVICES

The interaction of particles with integrated circuits, which can have different natures, can be generally classified as *single-event phenomena* (SEP) [1-4]. These phenomena must be isolated and random, both in time and space. High-ionizing particles that collide with the material may produce a large density of electron-hole pairs along their trajectory, which can alter the charge in the device.

This interaction is usually characterized by the "cross section" parameter. This is a macroscopic parameter that determines the number of errors in a particular device under radiation, given a certain fluence. The fluence is defined as the number of particles per unit of area that arrive at the material. The cross section depends not only on fluence but also on physical parameters, such as the angle of incidence of the particles. If the interaction occurs in a sensitive node of the circuit, it will likely produce a transient current pulse, thus altering the state of the circuit.

The effects on the device produced by SEP are called single-event effects (SEEs) [5-12].

These effects, depending on how they affect the circuit, may be classified into the following types:

- Nondestructive. These are temporary effects that affect the behavior of the system, which is usually recovered after some time or through a reset.
- Destructive. These effects produce permanent damage in the system, which leads to the inability of operation in a partial or total way.
- Cumulative. These are long-term effects, produced by long exposure time, due to the accumulation of numerous impacts.

14.2.1 Nondestructive Failures

These are failures that affect the system temporarily, which produce a misbehavior in part (or in the totality) of the circuit. The system will recover from these types of failures and will continue with its normal operation. Depending on how critical the affected zone is, the error will have more or less relevance. There may be applications for which not providing the right output during some cycles is not an issue, since they are noncritical. However, there may be cases in which just some cycles with a wrong behavior may lead to a disaster.

Although this type of failure is temporary, it does not mean that the effect will disappear spontaneously. For example, if it has affected a storage element, then a reset could be needed to bring the circuit back to normal operation. Therefore, it is very important to count with mechanisms able to detect when a wrong behavior is happening.

Several different effects may be differentiated, depending on their nature:

- Single-event error (also called soft error): the general term to describe nonpermanent effects.
- Single-event upset (SEU): happens when the error affects a storage element, thus potentially staying in the system for several cycles. This may affect both memory cells and registers.
- Multiple-bit upset (MBU): produced when an isolated event disturbs more than one storage element simultaneously. An increasing number of MBUs is being detected in digital systems, according to recent research works. They are potentially more dangerous than SEUs, since they affect several registers and can prevent usual protection techniques (e.g., triplication) from working.
- Single-event transient (SET): happens when the event produces a voltage pulse (i.e., glitch) to propagate through the circuit. If, as a result of this, an incorrect value is latched in a storage element, it is then considered an SEU.
- Single-event failure interrupt (SEFI): produced when the failure affects certain sensitive parts of the circuit (e.g., a state machine), making the system operate in a wrong (or unreachable) state. This usually produces an abnormal operation, as a restart or an exception.

14.2.2 Destructive Failures

These are failures that produce permanent damage to the circuit, destroying part of it. This will imply that this part would become inoperative and therefore will be useless for the rest of the circuit life. This may be an important issue for those systems operating in unreachable locations (e.g., space) since no on-site maintenance can be performed.

There are several types of effects in this category:

- Single-event burnout (SEB): usually happens in power transistors (metal-oxide semiconductor field-effect transistor [MOSFET], bipolar junction transistor [BJT], insulated-gate bipolar transistor [IGBT], and power diodes). The impact of heavy ions can activate a parasitic n-p-n structure, which can produce a positive feedback while in cutoff, thus permanently damaging the device.
- Single-event gate rupture (SEGR): happens in MOSFET power devices, where a heavy ion impacts the transistor channel region. A rupture may appear in the oxide that may produce destructive currents for the device.

- Single-event latchup (SEL): happens when a low-impedance path between the power supply rails of a MOSFET circuit is created, triggering a parasitic structure that disrupts proper functioning of the part and possibly leading to its destruction due to overcurrent.

14.2.3 Cumulative Failures

These are failures produced by the overall radiation received over time by the circuit, called the total ionizing dose (TID):

- Effects on the oxide: long-term effects related to circuit degradation produced by charge accumulation on the oxide, induced by the ionizing particles. Usually, it occurs when the TID reaches a certain level.
- Effects on the crystalline structure: can produce defects due to the crystalline structure displacement when the particles hit the device. This would induce permanent failures on the semiconductor.

14.3 METHODOLOGIES TO PREDICT THE BEHAVIOR OF INTEGRATED CIRCUITS IN THE PRESENCE OF SOFT ERRORS

One of the main concerns of integrated circuit (IC) designers is fault tolerance to soft errors, so that reliability and availability are guaranteed. It is indispensable to know beforehand how a certain circuit will behave in a radiation environment to see if it will meet the expected reliability constraints. Since many of the applications working in radiation environments (e.g., in the space industry) are costly, the information about reliability and fault tolerance has to be as accurate as possible.

Several magnitudes can be defined as a measure of reliability. The ones most often used by the industry are the following:

- Mean time to failure (MTTF): average time that the system works until the first failure happens.
- Mean time between failures (MTBF): average time that the system works after a recovery until the next failure happens.
- Mean time to repair (MTTR): average time needed to recover the system once a failure has been detected.

The most straightforward way to achieve this is performing direct measures through field studies—that is, to analyze results obtained through real test experiments, basically in radiation facilities, or with in-flight experiments (to test natural radiation). However, this approach has two main drawbacks.

The first one is that to conduct it, at least a prototype of the circuit is needed to apply real radiation. This means that this kind of experiment is feasible only in the later production stages. However, most of the time it is recommended to characterize circuits in earlier stages, since the obtained results may modify some design decisions. The second one is that this approach is quite costly. Radiation facilities are usually expensive, and their continuous use is feasible only for large corporations. Also, performing the test itself requires a high level of expertise; therefore, high-skilled technicians are also needed.

To avoid these problems, simulation processes are used to emulate the effect of real radiation on circuits. Simulations can be performed in very different ways. One alternative is to use another physical phenomenon to emulate the behavior of real radiation. This is the case of using a laser to produce bitflips in circuits, a technique that has been addressed recently. The advantage of this approach is that the use and maintenance of a laser device is much cheaper than a radiation facility. However, since both phenomena are essentially different, it is difficult to recreate real scenarios; therefore, conclusions cannot always be generalized.

As an alternative, simulation or emulation methods based on the "fault injection" concept are frequently used to reduce costs. The principle of these techniques is that artificial errors are forced in different positions and time instants as a simulation of real errors created by physical radiation phenomena. One of the challenges when using these techniques is the creation of environments as close as possible to real radiation scenarios.

To measure the quality of the different alternatives, the following indicators are frequently used:

- Intrusion: This is defined as the difference in behavior between the natural system operation and the operation when a fault injection has occurred, leaving aside the effect of the fault itself. This means that injecting implies altering in some way the system environment; therefore, side effects may occur. Therefore, methods should be as unintrusive as possible.
- Speed: A fault-injection campaign usually implies a large number of experiments, typically hundreds of thousands or even more than a million. This is because applications that work in a radiation environment are usually critical, and the more experiments to characterize them, the better. In this way, a good method should be able to perform large campaigns at a high speed.
- Precision: This parameter implies the capability to inject faults in a particular place (in a well-determined part of the system) and at a selected time instant. The higher the precision, the more realistic scenarios can be recreated.
- Cost: In this category, both the resources needed to perform the experiments and the time needed to set up the whole environment are considered. Obviously, a lower cost makes a technique more interesting.

In the following sections, some of the most commonly used test techniques to simulate radiation effects will be put in perspective.

14.3.1 Simulation-Based Fault Injection (SBFI)

This type of injection relies on assessing the system behavior through a high-level description in VHDL or Verilog, using simulation tools. First, once a test bench has been created, its behavior is simulated free of errors, called the golden simulation. This is how the system should behave in a normal operation and will be used to compare against and to detect any possible wrong behavior due to soft errors. After this, several bitflips are scheduled at particular storage elements and time instants that would simulate the effect of real radiation. There are several techniques to achieve this, but whichever one is selected should be as unobtrusive as possible. Then, the system is simulated again, and its behavior is compared with the golden data. The differences would represent errors that have been propagated to the output, thus producing a failure.

The main advantages of this method are its low cost, high precision, and the fact that no physical prototype is needed for the injection since it is performed on the VHDL description. This allows evaluating the circuit in earlier design phases, which can help to refine it in later stages.

On the contrary, the most usual drawback is usually its low speed. Software simulation is slow by nature, and this can be a problem if the injection campaign is very large.

Among the existent SBFI platforms, the following can be highlighted:

- MEFISTO, developed at Chalmers University of Goteborg (Sweden) and the LAAS Research Centre of Toulouse (France) [13].
- AMATISTA, developed under a project with the same name by Politecnico di Torino (Italy), Carlos III University in Madrid (Spain), and Thales Alenia Space (former Alcatel) [14].
- DEPEND, developed at Illinois University [15].
- SEU Simulation Tool (SST), developed by the European Space Agency and enhanced by Universidad Antonio de Nebrija (Spain) [16]. This tool will be explained in more detail in this chapter.

Especially interesting is the idea of extending these techniques to analog or mixed circuits, as was proposed with the VERIFY platform [17,18]. However, this approach has not been properly explored because the hardware description languages oriented to the analog part are not as mature as in the digital case.

14.3.2 Hardware Fault Injection (HWFI)

This approach implies the development of a platform with additional hardware that actually allows injecting faults and analyzing the results. HWFI techniques refer to a wide range of methodologies with very different characteristics. Most of them are based on “stuck-at” operations, in which pins are forced to a certain value to create a soft error. This approach is called pin-level injection. There are other alternatives, which are less common but also used, for example, to produce soft errors using electromagnetic interference or altering the power supply.

The advantages of this method are that the design under test corresponds to the real system and that the process can be performed in a very fast way. However, this approach also has some drawbacks, such as the potential high cost to set up the experiment, the limited number of points to perform the injection (pins and buses), and a low observability of the whole process.

The following are examples of platforms that follow this methodology:

- MESSALINE, developed at the LAAS Research Centre of Toulouse (France) [19].
- RIFLE, developed at Coimbra University (Portugal) [20].

14.3.3 Software Implementation Fault Injection (SWIFI)

These techniques apply only to microprocessor systems. They are based on the modification of the code executed in the processor to alter the system (simulating the impact of a particle). It is possible to inject failures in several abstraction levels (e.g., register transfer level [RTL], failures in memories). A commonly used methodology [21] adds code-emulated upsets (CEUs), which introduce a failure in a certain memory address when they are executed. The injection mechanism is usually linked to an interruption, which is triggered through a timer or a dedicated pin. This solution takes advantage of the trace mode, available in most processors, to stop normal execution to inject failures and to observe the results.

The problem with this approach is that it is highly intrusive and, in addition, quite slow, since the processor is constantly forced into the trace mode. On the other hand, it is an inexpensive solution, with a high controllability and observability.

Some examples of systems that follow this methodology are as follows:

- FlexFI, developed at Politecnico di Torino (Italy) [22].
- FERRARI, developed at the University of Texas [23].
- XCEPTION, developed at Coimbra University (Portugal), which has been commercialized [24].

14.3.4 Techniques Based on Hybrid Models: Hardware Emulation

Hardware emulation combines advantages of both SBFI and HWFI. The process consists of simulating a soft error by altering one or more bits of a field-programmable gate array (FPGA) bitstream. Then, this bitstream is "injected" in the FPGA, and the behavior is then emulated. The outputs of the system are compared with an error-free implementation (golden) to see how the behavior has been affected. This technique achieves high performance (hardware injection) together with high flexibility.

Examples of this methodology are as follows:

- FT-UNSHADES, implemented at the University of Seville (Spain) [25].
- Studies performed in [26,27] at Politecnico di Torino (Italy).

14.4 TECHNIQUES TO PROVIDE FAULT TOLERANCE IN ELECTRONIC DEVICES: RADIATION HARDENING

Once the techniques to predict the behavior of the circuits affected by soft errors have been presented, the next step is to provide techniques that can avoid these errors or mitigate their effects. The set of techniques whose goal is to make electronic components and systems more resilient to the effect of radiation is generically called radiation hardening.

TABLE 14.1
Fault Tolerance Classification and Unavailability Limits

Systems Type	Unavailability (min/year)	% Availability	Class
Unmanaged	52,560	90%	1
Managed	5,256	99%	2
Well managed	526	99.9%	3
Fault tolerant	53	99.999%	4
High availability	5	99.9999%	5
Very high availability	0.5	99.99999%	6
Ultra availability	0.05	99.999999%	7

Electronic circuits that will work in environments with a high level of radiation, like space or nuclear plants, are subject to important design restrictions to guarantee their correct operation. To deliver an appropriate level of reliability, the manufacturers of these integrated circuits, whose target applications are usually in the aerospace and military industries, apply to their processes these radiation-hardened (or rad-hard) techniques [28].

At the end, circuits should provide a good level of fault tolerance, as the capacity to recover from a failure with a minimum (or even nonexistent) effect on the behavior.

There have been some efforts to standardize fault-tolerance measurements, depending on the criticality of the application. For example, in Table 14.1 [29], a classification is proposed together with a suggested availability. This would be the fault-tolerance target when designing that particular system. In the following, a brief outline of the most frequently used radiation hardening techniques will be offered.

Radiation hardening by process (RHBP) encompasses techniques that seek protection of circuits through the manufacture process itself; this is different from the traditional processes of noncritical devices. Since rad-hard components have a quite small market (even residual), and considering the very costly manufacturing techniques that are required, the financial feasibility of this approach is most of the times doubtful.

For this reason, and although high reliability levels are achieved when using these techniques, new approaches are actively sought that can offer the same quality but with more reduced costs [28].

As reported in [30], the number of rad-hard foundries has declined drastically in the last years. The following are representative cases that employ this kind of technique:

- Lockheed–Martin, Manassas, VA [31].
- Honeywell Solid State Electronics Center, Plymouth, MN [32].
- Sandia National Laboratories.
- Aeroflex UTCM [33].

RHBP techniques try to protect circuits from the physical point of view, and all of them have a preventive character. This means that they try to avoid the possibility that an impact from a particle could produce a failure in the system. In other words, the idea is to prevent the energy from being transferred by the particle after the impact with the semiconductor producing a change in the transistor state.

This can be achieved by reducing the physical capacity of the material to absorb charge, which would mean a decrease in the linear energy transfer (LET). This is a measure of the energy transferred to the material as an ionizing particle travels through it. However, as technology shrinks and transistor channels are reduced, the critical charge needed to produce a bitflip in a device also decreases, which makes electronic systems more sensitive in general. As a result of this, two situations are becoming more likely:

- The effects of radiation arise in lower altitudes. Before, only devices operating in space or aviation applications used to suffer this kind of effect. But nowadays, soft errors are being reported for applications at ground level.
- MBUs are becoming more frequent. A single event is more likely to produce more than one bitflip in adjacent cells. This makes the error detection process more complex.

As a complement to the radiation-hardening process, and motivated by the increasing sensitivity of devices, applications usually stop performing critical operations when the environment is especially hostile to the electronic components. In this way, forecast of the operation conditions is usually performed [34] to schedule time frames where devices go inactive or operate with a reduced functionality.

There are areas where especially intense radiation activity is well known. This happens, for example, with the South Atlantic Anomaly, where the density of charged particles is much higher than in other areas and so is the probability of impact. This is due to the so-called Van Allen belts, which, due to their nature, trap more particles in that particular area.

In the following, some of the most extended methods to build rad-hard circuits will be explained.

14.4.1 Process to Reduce the Charge Generation and Accumulation

One of the most used technologies in this category is known as silicon-on-insulator (SOI). It is based on circuit hardening against radiation by using substrates with an isolation capability higher than the traditional semiconductor wafers. This drastically reduces the amount of induced charge due to impacts [35-39]. They differ from conventional devices in that the silicon junction is above an electrical insulator, typically silicon oxide (SiO). Exceptionally, sapphire is also used in the SOS form [40].

One of the techniques to produce SOI devices is called separation by implantation of oxygen (SIMOX). It consists of a direct injection of purified oxygen particles in the silicon wafer at high temperature. Oxygen reacts with silicon and creates a thin layer of silicon oxide.

While chips produced through a standard process can resist between 5 and 10 krad, devices that have been isolated with this method can support several orders of magnitude more. According to some studies, the SOI technology can provide a tolerance of up to 1 Mrad.

14.4.2 Mitigation of SET Generation and Propagation

This group of techniques tries to block SETs that have been induced in devices so that they are not registered by storage elements, since in this case they would turn into SEUs. To achieve this, all these techniques have a common goal, which is to increment the critical charge of memory cells.

There are two possible alternatives to this:

- Resistive hardening is an efficient way to increase fault tolerance in RAM cells, with just a small impact in the circuit density. It is based on the use of cross-coupled gate resistors to minimize SET propagation [41].
- Design of more resilient memory cells, which are known as heavy-ion tolerant cells. The objective of this process is the implementation of structures that are immune to SEUs and at the same time with a minimal cost. Among others, examples of these types of cells are HIT [42], DICE [43], Whitaker [44,45], and Rockett [46].

14.5 TECHNIQUES TO PROVIDE FAULT TOLERANCE IN ELECTRONIC DEVICES

Apart from radiation hardening by process, which was explained in the previous section, another group of techniques try to mitigate the effect of soft errors by design. These are called radiation hardening by design (RHBD). This means that the structure of the circuit is such that, although a soft error finally happens, it does not have any effect on the behavior. Therefore, if it is masked and does not propagate to the outputs, the system will behave correctly.

This kind of technique is based on redundancy to provide a suitable level of protection [47,48]. Redundancy has proved to be an efficient approach to mask soft errors in circuits. However, it also usually implies a cost overhead that can be unfeasible in some applications.

Redundancy can be applied on three different levels: spatial, temporal, and information.

14.5.1 Spatial Redundancy

Spatial redundancy consists of replicating the module to protect an odd number of times. A majority voter is added that selects the dominant behavior. Generically, it can be addressed as N-modular redundancy (NMR). In practice, the number of redundant copies is usually three. This would reduce the area overhead while providing a good protection level. In this case, this technique is known as triple modular redundancy (TMR).

One of the advantages of TMR is that this technique is generic. This means that it can be easily applied to different modules and applications without effort. On the other hand, as mentioned before, power consumption and area can be an issue (over three times in area vs. the unprotected module is a typical value, considering the majority voter).

TMR is usually applied to the sequential elements (e.g., registers, flip-flops) of the circuits. When extra protection is needed and the combinational modules also need to be replicated (typically logical gates), then this technique is known as functional triple modular redundancy (FTMR) [4,47,49].

14.5.2 Temporal Redundancy

In this case, no extra hardware is required to achieve protection. Temporal redundancy implies repeating the operation to protect several times. Each time, the result is monitored, and at the end, the result that has occurred the most times is selected as valid. This produces a worsening in performance instead of an area overhead, as happened with the spatial redundancy. This technique is especially useful when most failures tend to be transient.

An alternative to this, which also uses principles associated with TMR, is the temporal-spatial redundancy. In this case, several delays are added to the clock signal that controls the different TMR copies. In this way, there are still three copies of each protected module (spatial redundancy), but each one operates at different time instants (temporal redundancy), which is useful to eliminate transient errors like SETs. The delay introduced in the clock signals has to be estimated as the duration of a typical SET. The problem with this technique is that it introduces delays in the critical path (thus reducing performance) and that implementing delays is not a trivial task.

14.5.3 Information Redundancy

This technique is usually applied to large storage elements, as memories. Since the information stored in these elements is usually large, no spatial redundancy is usually employed since the area overhead would have a very negative impact. Instead, some extra information is added to data to detect and correct errors, which usually implies much less area than the previous approach. One of the most popular approaches in this category is called error detection and correction (EDAC). There are many different types of EDAC codes, each of them with different detection and correction capabilities. They are based on codifying data words of k-bits using extra redundant bits, forming final words of n-bits. The redundant bits, n-k, are used to determine if there has been any bitflip in data and to eventually correct them. There is a limit to the number of errors that a particular code is able to detect and correct. For example, a simple implementation is called single-error correction, double-error detection (SEC-DED). This implementation is able to detect up to two errors in the same word, but it can correct only isolated errors. There are more complex codes that can handle more simultaneous errors, but at the expense of a higher number of redundant bits. One of the most popular implementations of this kind of redundancy is Hamming codes. Other alternative codes are Golay, Hadamard, and Reed-Solomon [50].

14.6 AD HOC PROTECTION TECHNIQUES FOR DIGITAL FILTERS

In the previous sections, several techniques to protect circuits against soft errors were presented, each of them with some advantages and drawbacks. However, a different alternative to achieve this protection is to use circuit-specific techniques designed ad hoc for each system. This approach has the potential of reducing the overhead associated with the technique, since the structure and functionality of the circuit are analyzed in detail to protect only the critical elements and to reuse functional elements. The main drawback of this technique is that it requires an in-depth knowledge of the circuit to be protected and therefore a larger engineering effort.

One area where a number of ad hoc protection techniques have been proposed is signal processing [51]. This is due to the wide use of signal processing circuits in a variety of applications and to the regular structure of most of these circuits that makes them suitable for ad hoc protection techniques.

For a given circuit, another factor that is key when designing ad hoc protection is to consider only the errors that cause an issue for the system or application. In many cases, failures may not have an impact on the circuit functionality or may be detected and corrected using elements of the circuit. For example, in a circuit that implements a communications receiver, a soft error can cause a bit error in the received sequence, but if soft errors occur rarely, the receiver may still meet the bit error rate target without further impacts. In this case, soft errors are not critical, and protection would not be an issue. However, if the error rate needs to be reduced, soft errors have to be handled accordingly. This example illustrates how the knowledge of both the circuit and the application requirements can help in optimizing the protection.

Continuing with the previous example, in some cases the receiver has a Viterbi Decoder [52] to reduce the bit error ratio (BER). When this is the case the error caused by a soft error at the input may be corrected as part of the decoding process. In this way an element of the circuit itself introduces protection for soft errors on other elements. This is also the case for other signal processing filters like adaptive filters where the adaptation is able to remove the errors caused by a soft error [53].

A number of techniques have been proposed to protect different signal processing circuits from the effects of soft errors. For example, in [54-56] techniques to protect fast Fourier transforms (FFTs) are presented, whereas in [57-60] the protection of convolutions is addressed. In both cases the structure and properties of the underlying signal processing algorithms are used to derive efficient protection techniques. Another approach presented in [61,62] is to use reduced precision replicas of the circuit to detect and correct errors. This strategy can be applied to finite impulse response (FIR) filters among other circuits. Other protection techniques, like the use of Hamming codes, have also been proposed for FIR filters [63].

To illustrate circuit-specific protection techniques for signal processing applications, moving average filters are studied in the following [64]. The goal is to show how the applications requirements in conjunction with the circuit structure can be used to design efficient ad hoc protection techniques.

Digital FIR filters are good candidates for ad hoc protection techniques, as they exhibit a regular structure and are frequently used in many applications [51]. A FIR filter performs the following operation:

$$y[n]=\sum_{i=0}^{N-1}x[n-i]\cdot h[i] \tag{14.1}$$

where $x[n]$ is the input signal, $y[n]$ the output, and $h[n]$ the impulse response of the filter whose nonzero values are all in the 0 to $N-1$ interval.

A number of structures have been proposed to implement FIR filters [51]. Two of the most common ones are illustrated in Figure 14.1. Although both implementations are functionally equivalent, they exhibit different behaviors in the presence of soft errors. For example, a soft error on the registers in the first structure will cause an error in the output that can last for $N-1$ clock cycles at most, whereas in the second one, a soft error on the registers will only corrupt $y[n]$ for one clock cycle. In fact, in the first case, the error in the delay element propagates to the output filtered by a portion of the impulse response depending on the position of the delay element that suffered the soft error. As can be seen in this example, just looking at the effects of soft errors on existing filter implementations can yield interesting results.

Moving average filters are a special type of FIR filter that show some interesting properties for implementation and are also used in many applications [65]. A moving average filter performs the following operation [51]:

$$y[n]=\frac{1}{N}\sum_{i=0}^{N-1}x[n-i] \tag{14.2}$$

which is a particular case of (14.1). Normally, N is a power of two, so that the division can be implemented with a shift operation. In this case, the filter needs only adders.

A more efficient implementation can be derived by rewriting (14.2) as follows:

$$y[n]=y[n-1]+\frac{1}{N}(x[n]-x[n-N]) \tag{14.3}$$

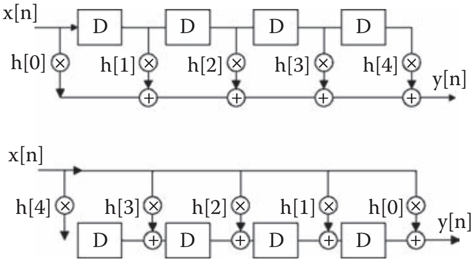
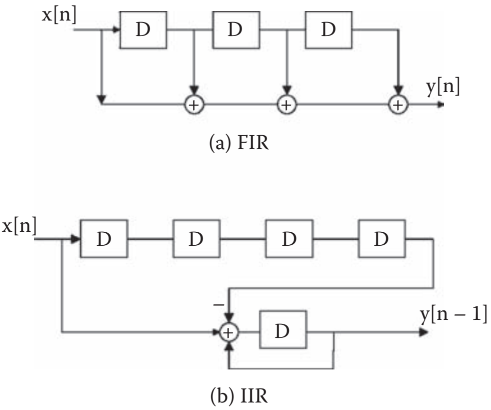


FIGURE 14.1 FIR filter implementations. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)



(a) FIR

(b) IIR

FIGURE 14.2 Examples of moving average filter implementations. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

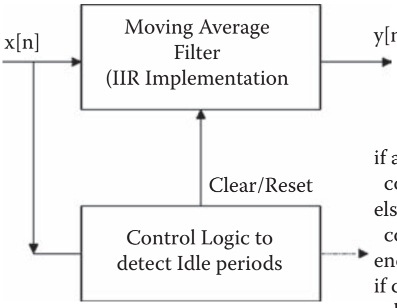
In this case, only two adders are needed irrespective of the value of *N*. In fact, this implements the FIR filter using a recursive structure normally used in infinite impulse response (IIR) filters. In the following we will refer to this implementation as IIR or recursive. A diagram for both implementations is shown in Figure 14.2.

A first look at the effect of soft errors on both structures shows that in the case of the more efficient IIR implementation, soft errors in the delay line or in the accumulator can cause errors in the output that will persist until the filter is reset. This was previously noted in [65] as a drawback of the IIR structure and used to advocate the use of an FIR structure for cascaded moving average filters. The previous discussion clearly shows how the presence of soft errors can influence the choice of the implementation structures for digital filters and suggests the interest of ad hoc protection techniques.

As mentioned before, to come up with optimal soft error protection techniques we need to take the requirements of the application in which the filter is used into account. To illustrate this point, three distinct scenarios will be presented, each of them with different protection requirements. By carefully assessing these requirements, efficient implementations will be generated that are more convenient than general (nonspecific) protection techniques.

14.6.1 First Scenario (Low Protection Requirements)

Let us consider a first scenario where the filter is used to detect Ethernet link pulses [66] in the presence of noise. In this application, idle periods in between pulses are present, and the detected pulses drive a state machine that ensures that occasional misdetection causes no problem to the application. In this situation, errors on the output of the filter can be tolerated as long as they are transient. In this case, the FIR-like implementation can be directly used, as a soft error creates only a transient error


x[n]
Moving Average Filter (IIR Implementation
y[n]
Clear/Reset
Control Logic to detect Idle periods

Logic Description

```
if abs(x[n]) < Threshold
  count_next = count_next + 1
else
  count_next = 0
end
if count == n + 1
  clear filter;
  count_next = 0;
end
if positive clockedge
  count = count_next;
end
```

FIGURE 14.3 Illustration of the first protection technique. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

at the output. However, to use the IIR-like or recursive implementation, a protection technique is needed to ensure that the effects of soft errors become transient.

The traditional approach to deal with soft errors would be to apply TMR to all storage elements. However, a better approach would be to detect that we are in an idle period and if the accumulator has a value that is greater than the maximum expected value for that case then reset the accumulator and the delay line registers. In this way, an extra counter can be added that will compute the number of consecutive idle cycles in the system. By idle, we mean cycles in which the system input is below a given threshold, which would indicate the presence of noise in the filter. On the other hand, inputs over the noise threshold would be considered as active data. Therefore, if the counter detects N consecutive idle cycles (N being the number of taps in the circuit), this would mean that anything inside the circuit (stored in the delay line) comes from acquired noise and therefore can be discarded (reset).

With this mechanism, the filter actually is reset whenever N idle cycles are detected. No matter when a soft error hits the system, it will only last until the next reset of the circuit, making its effect transient. Notice that the mechanism will work only if frequent N idle cycles can be guaranteed, but since the system knowledge assures that this application can expect this kind of input, the proposed technique will be good enough with a cost much lower than TMR. This is the core of the first technique, as illustrated in Figure 14.3.

Special care must be taken to ensure that soft errors in the added control logic do not cause errors; this can occur, for example, if a soft error in one of the bits of the counter triggers the reset of the filter during the detection of a pulse. With a careful selection of the threshold value for the counter we can ensure that this situation will never happen. If N is a power of two, then the threshold to reset can be set to $N + 1$ samples rather than to N. In this way, a single soft error in only one bit will never cause

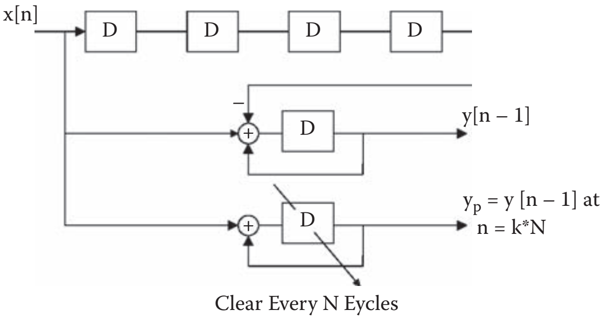


FIGURE 14.4 Illustration of the second technique. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

the mentioned reset, since the value of the counter would be zero during pulse detection, and to get to $N + 1$ (reset condition), two bits flipped to 1 would be needed.

14.6.2 Second Scenario (Average Protection Requirements)

As a second scenario, let us take an application that can tolerate occasional errors on the output of the filter (like in the first scenario) but that has no guarantee of idle periods that can be used to reset the accumulator and delay line registers.

In this situation, the computation of the output can be done in parallel by another structure added to the filter, for the sake of comparison. Obviously, if this added structure is a replica of the filter itself, we would be doubling the complexity of the system. To avoid this situation, this parallel structure will be implemented with a decimated filter (as illustrated in Figure 14.4), which has a structure simpler than a regular filter, with the drawback that it computes only the right output 1 of N cycles.

In this way, the output of both structures would be compared at each $y[n*N]$, which is 1 of N times. A comparison highlights several cases:

- Both structures have the same output. This means the system is error-free.
- The structures have different outputs. This means that a soft error has hit the original filter or the secondary (decimated) one. To determine where the error is, the decimated filter is reset, and after N cycles the comparison is made again. After this, the following may happen:
    - The error is gone, which means the soft error occurred in the decimated filter.
    - The error is still present, which means the soft error occurred in the main filter, which needs to be reset to eliminate it.

Regardless of which of the previous situations happens, the effects of the soft error will always be transient, which satisfies the application requirement, but in this case no idle periods are needed. Therefore, the protection level has been increased with a minimum complexity increment, thanks again to the use of the knowledge of the system.

14.6.3 Third Scenario (High Protection Requirements)

As a third and final scenario, let us assume that we want to provide protection such that soft errors do not cause any misbehavior in the output of the filter. One alternative to the use of TMR in all registers is to take advantage of the fact that the registers for the FIR implementation in the upper part of Figure 14.1 and for the IIR-like or recursive implementation are connected in such a way that their values do not suffer changes as they move across the delay line.

This can be used to compute a two-dimensional parity as follows:

- For each input value, compute a "vertical" parity bit, *Pv.*
- For each bit position on the input value, compute a "horizontal" parity bit, *Ph,* across all the bits that have that position on the registers of the delay line.

Pv is updated only at the input of the delay line, while *Ph* is updated every clock cycle with the bit entering the delay line and the one leaving it. These two sets, *Pv* and *Ph*, form the accumulated parity of the circuit, which is constantly being updated.

Dynamically, each clock cycle, both the horizontal and vertical parity are rechecked and compared with the accumulated values. Then, several situations can arise:

- The actual and accumulated values are the same. There is no problem with the system, and its behavior can be taken as correct.
- There is a discrepancy between a bit of the accumulated and actual *Ph* and between a bit of the accumulated and actual *Pv.* If both differences happen, that means a soft error has affected a register in the delay line. The bit affected by the soft error is the crossing point of the discrepant *Ph* and *Pv.* In this way, since it has been identified, it can be corrected instantly; therefore, the system behavior remains correct.
- There is a discrepancy between a bit of the accumulated and actual *Ph* or between a bit of the accumulated and actual *Pv.* If only one of the parity registers shows the discrepancy, it would mean a soft error has affected the discrepant parity register itself. It is important to remember that all the extra structure added for protection can also be affected by soft errors.

This is the outline of the third proposed technique, as illustrated in Figure 14.5.

Note that if a soft error strikes on Ph_{x} the error can be permanent and therefore needs to be corrected. One alternative is to use TMR on Ph_{x}. However, it seems that a better option is to generate a signal $Err_{\mathrm{v}} = Err_{\mathrm{v1}}$ or Err_{v2} or … or Err_{vn} indicating if there is an error on the vertical parity bits. Then if ($Err_{\mathrm{v}} = 0$) and ($Err_{\mathrm{hi}} = 1$) we know that there has been a soft error on Ph_{i}, and we can do $Ph_{\mathrm{i}} = \mathrm{not}(Ph_{\mathrm{i}})$. This is illustrated in Figure 14.5c.

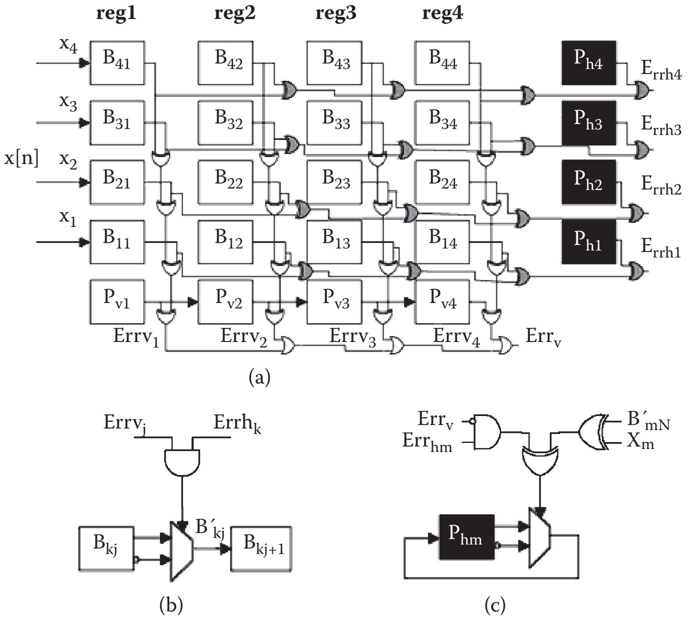


FIGURE 14.5 Illustration of the third technique. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

As already mentioned, at each clock cycle *Phs* and *Pvs* are recomputed and compared with the stored values. If there are no differences $B_{\mathrm{m,n}}$ would be updated at each clock cycle with $B_{\mathrm{m,n-1}}$. If both Pv_{m} and $Ph_{\mathrm{n-1}}$ show differences with the stored values, then $B_{\mathrm{m,n}}$ would be updated with $B_{\mathrm{m,n-1}}$ negated. This correction logic is illustrated in Figure 14.5b.

The overhead of this approach is $N + M$ flip-flops plus the logic to compute the parities and detect the errors. In principle, it would work for both the FIR and IIR implementations. There is, however, one potential problem with the proposed technique that can be explained by analyzing Figure 14.5b. Basically, if a soft error occurs in the middle of a clock cycle, it will not be corrected until the new value of B_{kj} propagates through the parity checking logic and sets Err_{hk} and Err_{vj} to 1 (so that the inverted output of the flip-flop is selected). So, potentially, if the soft error occurs near the end of the clock cycle the wrong value may be stored in the next element of the delay line before the correction can take place.

To avoid this problem, buffers could be used to delay the inputs to the multiplexer so that they have approximately the same delay as the signal that controls that multiplexer. However, this may be complex to do and would also complicate synthesis and place and route that would require manual intervention. This would make the design technology dependent and therefore harder to reuse.

A better alternative can be found by noting that for the IIR or recursive implementation, this problem is an issue only if it occurs in the last element (register) of

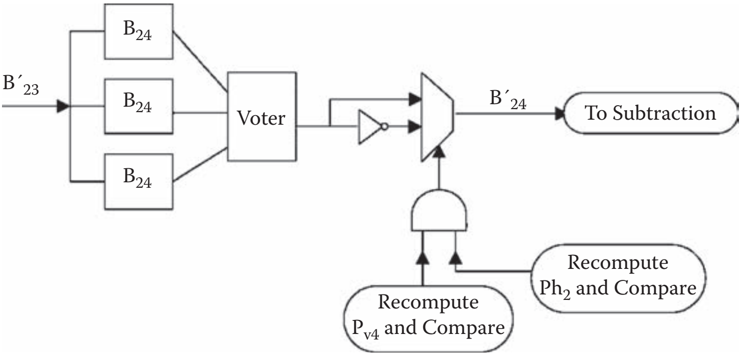


FIGURE 14.6 Proposed implementation for the last delay line register. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

the delay line. If it happens on previous elements, it would be corrected during the next clock cycle and would never cause a difference in the filter output (since only the last element of the delay line is used for subtraction). This observation could be used to solve the problem by adding TMR to this last register only, such that soft errors in this register would not propagate to the multiplexer. In this situation, for this last register, we need to correct only soft errors that occurred in the previous delay line register and that have been propagated to this because they occurred at the very end of the clock cycle (other soft errors will be eliminated by TMR). The overall strategy is shown in Figure 14.6. The additional cost in terms of area is quite low, as TMR is added to only one register.

By adding this protection, we ensure that soft errors will not cause any error on the filter output; therefore, the protection requirements stated for this scenario are met.

Note that this can be done only for the IIR structure. However, for the FIR one, we can have errors in the filter output caused by soft errors that occur in any delay line register (all are connected through adders to the filter output) at a time, such that the wrong value propagates to the output before it can be corrected. The amount of time during which a soft error can cause this error is the difference between the delay of the correction logic and the data input to the multiplexer. If this is small compared with the clock cycle, then most of the soft errors will not cause errors in the filter output, and this technique may still be useful. If that is not the case, buffers could be used as discussed before.

A similar problem to the one just discussed can occur when correcting Ph_k if a soft error has changed B_{kj}, and Err_{hk} becomes 1 before Err_v does. In this case, Ph_k would be inverted, and if that happens at the end of a clock cycle the inverted Ph_k could be stored. However, in this situation, the error in Ph_k will not cause errors on the filter output and will be corrected in the next clock cycle. Finally, the two previous problems could also theoretically occur simultaneously. In that case, Ph_k would be inverted, and B_{kj} would not be corrected, creating an error in practice. Nevertheless, this is unlikely, since Err_{hk} should be 1 for the first problem to occur, and in that case Err_{vj} needs to be 0 for the second problem to occur. In most cases, the delay to generate Err_{hk} would be similar to or larger than the one to generate Err_{vj} (assuming that we have more taps in the filter than bits in each register).

The proposed technique cannot correct multiple errors if they occur simultaneously or very close in time in all cases. For example, if a soft error hits one *Pv* and then, later, another one hits one of the bits of the register that contains the erroneous *Pv*, the vertical parity checking would give the correct result, and the erroneous bit will not be corrected.

A final observation is that the proposed scheme for the FIR implementation of the moving average filter, with the limitations previously outlined, can in fact be used for any FIR filter implemented with that structure.

14.6.4 Evaluation of the Protection Techniques

After a protection technique has been proposed, the next step is to evaluate its effectiveness for actual implementations. The proposed techniques have been implemented in VHDL, and then the circuits have been synthesized for a commercial application-specific integrated circuit (ASIC) library, as discussed later. This will also allow assessing the efficiency of the proposed techniques in terms of circuit complexity and comparing it with the traditional approaches. Then, a simulation environment will be used to insert soft errors in the circuit and evaluate their effects.

A block diagram of the environment is shown in Figure 14.7. This environment uses Matlab to generate the reference signals for the input and output of the filter

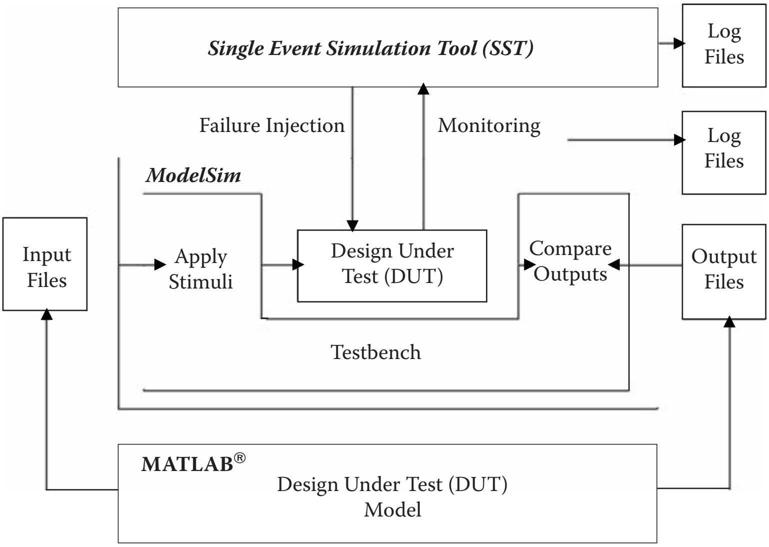


FIGURE 14.7 Simulation environment block diagram. (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

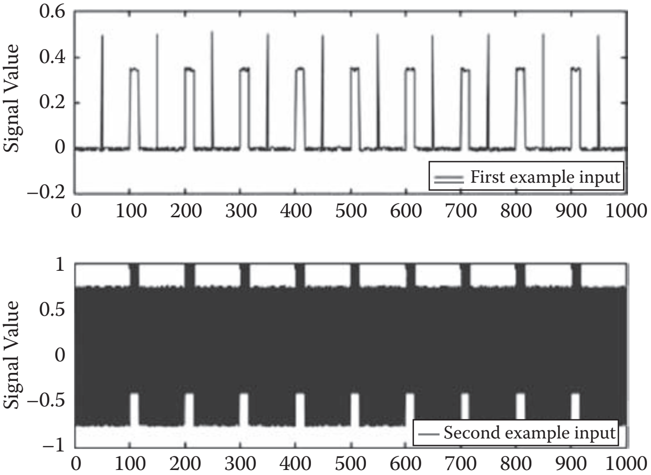


FIGURE 14.8 Input signals for examples 1 (top) and 2 (bottom). (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

that are stored in files. These files are then used to drive the circuit implementation using a commercial VHDL simulator (Modelsim), whose output is compared with the reference output to detect any discrepancies. Soft errors are introduced using the SST developed at the European Space Agency [16], which has been updated so that it works with the latest releases of Modelsim and extended in functionality [67]. This environment is flexible in the way that different signals are easily generated in Matlab to exercise the filter under different conditions. The SST is also convenient to define when and where to insert soft errors in a design.

To better illustrate these features, the first protection technique described in the previous section has been evaluated, assuming an eight-bit input signal, $x[n]$, with range –1,1 and N set to 16. The threshold to reset the accumulator is 3 LSBs. We start by generating a sequence (see top of Figure 14.8) of pulses (instants multiple of 100) plus impulsive noise (instants 50, 150, 250, etc.) in Matlab, and then a soft error is introduced in the accumulator on cycle 20 to check the discrepancies at the output. The results (Figure 14.9) show that there is no permanent error and also demonstrates how the circuit recovers from such errors. We have also used the SST to insert soft errors only during the reception of pulses and only in the protection logic to check that pulse detection is not affected by soft errors in this case.

For the second example, a signal that contains the pulses plus a high-frequency noise has been generated (see bottom of Figure 14.8), and then the SST has been used to insert a soft error in cycle 10. In this case, technique 1 would not work, as the incoming signal continuously exceeds the threshold. As can be seen in Figure 14.10, technique 2 removes the error in approximately 2*N samples. Besides, the SST has been employed to check that soft errors in the decimated filter used for correction do not cause errors in the output of the filter.

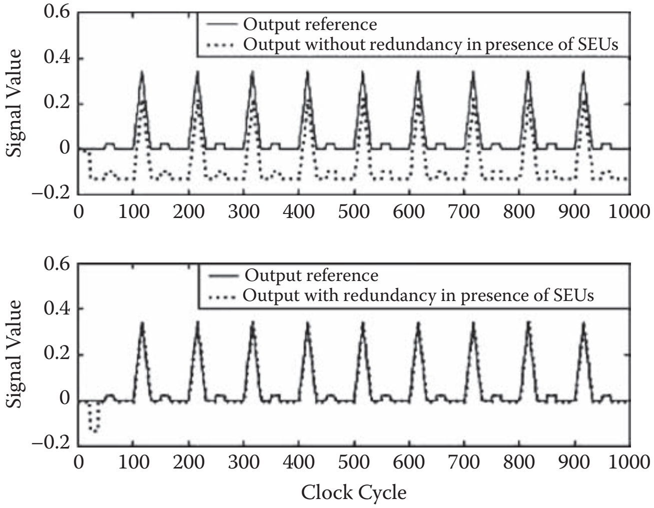


FIGURE 14.9 Output of an unprotected IIR (top) and of an IIR protected with technique 1 (bottom). (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

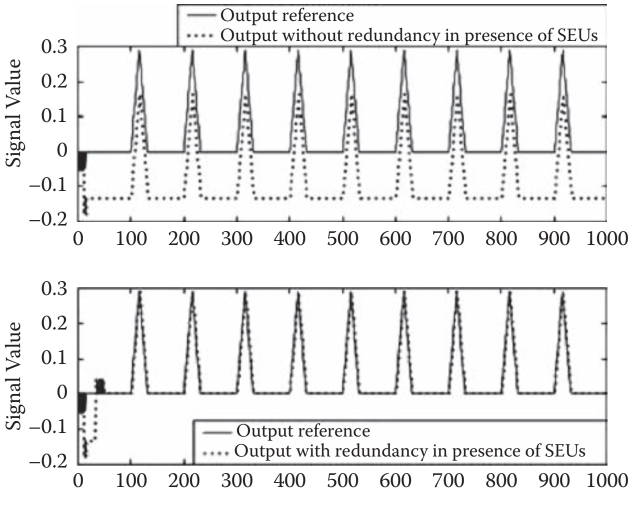


FIGURE 14.10 Output of an unprotected IIR (top) and of an IIR protected with technique 2 (bottom). (Reprinted with permission from Reyes, P., Reviriego, P., Maestro, J.A., Ruano, O., "New protection techniques against SEUs for moving average filters in a radiation environment." *IEEE Transactions on Nuclear Science,* Volume: 54 Issue: 4, 957–964.)

For the third example, some soft errors have been introduced in the registers of the delay line to check that they are corrected. Then, other soft errors have been applied to the parity bits, in order to verify that they do not trigger erroneous corrections. Both VHDL simulations (introducing delays in the parity checking and delay back-annotated simulations on the synthesized netlist) have been run to check that the soft error propagation described in the previous section causes errors in the filter output when TMR is not present in the last delay register element. Reciprocally, it has been proven that this does not happen once TMR is added. With the final implementation we have verified that the output of the filter does not show any discrepancy with the reference output when inserting random soft errors that are at least *N* clock cycles apart. We have also checked that soft errors closer than *N* clock cycles apart can create errors on the filter output, as described before.

Although we have focused on analyzing the effects of soft errors affecting some particular bits or registers, the fault insertion campaigns applied to all three examples are far more extensive, including in all cases thousands of random soft errors.

14.6.5 Comparison with TMR

After checking the effectiveness of the protection techniques, the cost in terms of circuit complexity has to be analyzed and compared with general protection techniques such as TMR. To do it, we have focused on the number of equivalent gates, as an estimate of circuit complexity. The results (see Table 14.2 and Table 14.3) have been generated for a TSMC 0.25 um library and assuming a 50 Mhz clock and an

TABLE 14.2
Number of Gates for the FIR Implementations

	***N* = 8**	***N* = 16**	***N* = 32**
FIR	713	1,591	3,788
FIR with TMR	1,686	3,538	7,500

TABLE 14.3
Number of Gates for the IIR-Like Implementations

	***N* = 8**	***N* = 16**	***N* = 32**
IIR	517	866	1,550
IIR with Tech. 1	690	1,141	2,024
IIR with Tech. 2	835	1,300	2,186
IIR with Tech. 3	1,452	2,332	4,029
IIR with TMR	1,633	2,962	5,591

eight-bit data path. For the FIR case, the structure shown at the top of Figure 14.1 has been used, as it is the one that results in a lower gate count.

The first thing to note is that the IIR implementation is more efficient than the FIR one, as expected, and that for large values of *N*, the difference can be more than double. This explains why the IIR implementation is normally used in the absence of soft errors. It should also be remarked that although not shown in the tables, the IIR or recursive implementation is also faster; therefore, for the same technology, the filter can operate at higher clock frequencies.

In the presence of soft errors, the unprotected IIR implementation will suffer permanent errors, while the FIR one would show only temporary ones. In this case, what needs to be compared is the unprotected FIR versus the IIR protected with techniques 1 and 2 that mitigate the effects of soft errors so that they become temporary. It can be seen that for large values of *N*, the IIR with techniques 1 and 2 is still significantly smaller than the FIR. For $N = 16$, the reduction is 28% and 18%, respectively, while for $N = 32$, the reduction is over 40% in both cases. This is so because for these techniques, the amount of redundancy introduced is almost independent of the value of *N*. The IIR protected with these techniques still has the speed advantage versus the FIR that has been mentioned before.

If we now focus on implementations that ensure that isolated soft errors do not cause errors on the filter output, we can compare the IIR protected with technique 3 versus the IIR protected with TMR (note that the FIR with TMR is not considered, as it would be equivalent in this case to the IIR with TMR but with a larger number of gates). Technique 3 results in a reduction of 10% to 30% in the number of gates. These results show that technique 3 may be a good choice in this case.

In summary, the proposed techniques 1 and 2 enable the use of the IIR structure when transient errors are acceptable at the filter output, while technique 3 provides effective protection when all the errors are to be avoided at the filter output. In both cases, significant savings in terms of area are obtained when compared with traditional techniques.

From the discussion of the moving average filter protection techniques it can be seen that techniques that exploit both application and system knowledge to provide a more intelligent protection can result in a lower circuit complexity compared to TMR. More generally, the same broad idea of using system knowledge to derive specific or ad hoc protection techniques can also be applied to other types of filters like IIR filters, adaptive filters, filter banks, and FFTs.

This approach requires a larger engineering effort but can provide a more effective implementation in terms of circuit area and power. Therefore, depending on the objectives of a given design, it may be an interesting option compared with general techniques like TMR.

14.7 CONCLUSIONS

In this chapter the issue of soft errors on electronic devices has been presented. The different effects of radiation on electronic components have been described along with the main techniques used to mitigate them. The different mechanisms to emulate or simulate the effects of soft errors on electronic circuits have also been covered.

In the final part of the chapter ad hoc protection techniques for digital filters have been presented and evaluated. The results show that the use of specific protection techniques that exploit the circuit structure or the application requirements can result in an effective protection at a lower implementation cost than generic techniques like TMR.

REFERENCES

1. J. F. Ziegler and W. A. Lanford, "The effect of sea level cosmic rays on electronic devices," *J. Appl. Phys.,* vol. 52, pp. 4305–4318, 1981.
2. C. A. Gossett, B. W. Hughlock, M. Katoozi, G. S. LaRue and S. A. Wendler, "Single event phenomena in atmospheric neutron environments," *IEEE Transactions on Nuclear Science,* vol. 40, pp. 1845–1856, 1993.
3. R. D. Schrimpf and D. M. Fleetwood, "Radiation effects and soft errors in integrated circuits and electronic devices," World Scientific Publishing, 2004.
4. P. E. Dodd and L. L. Massengill, "Basic mechanisms and modeling of single-event upset in digital microelectronics," *IEEE Transactions on Nuclear Science,* vol. 50, no. 3, June 2003.
5. M. Nicolaidis, "Design for soft error mitigation," *IEEE Transactions on Device and Materials Reliability,* vol. 5, no. 3, September 2005.
6. T. C. May and M. H. Wood, "A new physical mechanism for soft errors in dynamic memories," in *Proc. 16th Annual Int. Reliability Physics Symp.,* pp. 33–40, 1978.
7. E. Normand, "Single event upset at ground level," *IEEE Transactions on Nuclear Science,* vol. 43, pp. 2742–2750, 1996.
8. D. G. Mavi and P. H. Eaton, "Soft error rate mitigation techniques for modern microcircuits," in *Proc. 40th Annual Reliability Physics Symp.,* pp. 216–225, 2002.
9. D. Radaelli, H. Puchner, S. Wong and S. Daniel, "Investigation of multi-bit upsets in a 150 nm technology SRAM device," *IEEE Transactions on Nuclear Science,* vol. 52, no. 6, December 2005.
10. D. Tipton, J. A. Pellish, R. A. Reed, R. D. Schrimpf, R. A. Weller, M. H. Mendenhall, et al., "Multiple-bit upset in 130 nm CMOS technology," *IEEE Transactions on Nuclear Science,* vol. 53, no. 6, Part 1, pp. 3259–3264, December 2006.
11. A. M. Chugg, M. J. Moutrie and R. Jones, "Broadening of the variance of the number of upsets in a read-cycle by MBUs," *IEEE Transactions on Nuclear Science,* vol. 51, no. 6, Part 2, pp. 3701–3707, December 2004.
12. A. M. Chugg, M. J. Moutrie, A. J. Burnell and R. Jones, "A statistical technique to measure the proportion of MBU's in SEE testing," *IEEE Transactions on Nuclear Science,* vol. 53, no. 6, Part 1, pp. 3139–3144, December 2006.
13. E. Jenn, J. Arlat, M. Rimen, J. Ohlsson and J. Karlsson, "Fault injection into VHDL models: the MEFISTO Tool," *Proc. FTCS-24,* pp. 66–75, 1994.
14. L. Berrojo, F. Corno, L. Entrena, I. González, C. López, M. Sonza, et al., *New techniques for speeding-up fault injection campaigns,* Paris/Francia, 2002.
15. K. K. Goswami, R. K. Iyer, Fellow, IEEE, L. Young, Member, IEEE, "DEPEND: A simulation-based environment for system level dependability analysis," *IEEE Transactions on Computers,* vol. 46, no. 1, January 1997.
16. D. Gonzalez-Gutierrez, "Single event upset simulation tool functional description," ESA Report TEC-EDM/DCC-SST2, July 2004.
17. V. Sieh, O. Tschäche and F. Balbach, "VHDL-based fault injection with VERIFY," Internal Report No. 5/96.
18. V. Sieh, O. Tschäche and F. Balbach, "VERIFY: evaluation of reliability using VHDL-models with embedded fault descriptions," *Twenty-Seventh Annual International Symposium on,* pp. 32–36, June 24–27, 1997.

19. J. Arlat et al., "Fault injection for dependability validation: a methodology and some applications," *IEEE Transactions on Software Engineering,* vol. 16, no. 2, February 1990.
20. J. V. Carreira, D. Costa and J. G. Silva, "Spectrum fault injection spot-checks computer system dependability," *IEEE Spectrum*, vol. 36, no. 8, pp. 50–55, August 1999.
21. R. Velazco, S. Rezgui and R. Ecoffet, "Predicting error rate for microprocessor-based digital architectures by C.E.U. injection," *IEEE Trans. on Nuclear Science,* vol. 47, no. 6, pp. 2405–2411, December 2000.
22. A. Benso, M. Rebaudengo, and M. Sonza, "FlexFi: a flexible Fault Injection environment for microprocessor-based systems," in *SAFECOMP 1999: 18th International Conference on Computer Safety*, *Reliability and Security* (Lecture Notes in Computer Science), ed. A. Pasquini, Springer Verlag, pp. 323–335.
23. G. A. Kanawati, N. A. Kanawati and J. A. Abraham, "FERRARI: A flexible software-based fault and error injection system," *IEEE Trans. on Computers,* vol. 44, no. 2, pp. 248–260, February 1995.
24. J. Carreira, H. Madeira, and J. Silva, "Xception: software fault injection and monitoring in processor functional units," *DCCA-5,* Conference on Dependable Computing for Critical Applications, Urbana-Champaign, IL, pp. 135–149, September 1995.
25. M. A. Aguirre, J. Noel, V. Baena, F. Muñoz, A. lbañez, A. Fernández, et al., "Ft-Unshades: a new system for SEU injection, analysis and diagnostics over post synthesis netlist," 2005 MAPLD International Conference, Washington, DC.
26. L. T. Young, R. Iyer and K. K. Goswami, "A hybrid monitor assisted fault injection experiment," *Proc. DCCA-3*, pp. 163–174, 1993.
27. P. Civera, L. Macchiarulo, M. Rebaudengo, M. Sonza Reorda and M. Violante, "Exploiting circuit emulation for fast hardness evaluation," *IEEE Transactions on Nuclear Science,* vol. 48, no. 6, pp. 2210–2216, December 2001.
28. H. J. Barnaby, "Will radiation hardening by design work?", *Nuclear & Plasma Sciences Society News,* no. 1, March 2005.
29. K. Norvag, "An introduction to fault tolerant systems," Department of Computer and Information Science, IDI Technical Report 6/99, Revised July 2000.
30. K. A. LaBel, "Space radiation effects on electronics: simple concepts and new challenges," Available at: http://radhome.gsfc.nasa.gov/radhome/papers/MRS04_LaBel.pdf
31. Lockheed–Martin, Manassas, VA, Available at: http://www.lockheedmartin.com
32. Honeywell Solid State Electronics Center, Plymouth, MN, Available at: http://www.ssec.honeywell.com/about/
33. Aeroflex UTCM, Available at: http://www.utcm.com
34. C. Dier, "Radiation effects on spacecraft & aircraft," *ESA SP-477,* pp. 505–512, 2002.
35. G. E. Davis, L. R. Hite, T. G. W. Blake, C.-E. Chen, H. W. Lam and R. DeMoyer, "Transient radiation effects in SOI memories," *IEEE Trans. Nucl. Sci.* vol. 32, no. 6, pp. 4432–4437, December 1985.
36. J. P. Colinge, "*Silicon-on-insulator technology: materials to VLSI,*" 2nd ed., Boston: Kluwer Academic, 1997.
37. R. E. Mikawa and M. R. Ackermann, "Transient radiation effects in SO1 static RAM cells," *IEEE Transactions on Nuclear Science*, vol. NS-34, no.6, pp. 1648–1703, December 1987.
38. J. L. Leray, E. Dupont-Nivet, J. F. Pere, Y. M. Coic, M. Raffaelli, A. J. Auberton, et al., "CMOS/SOI hardening at 100 Mrad(SiO)," *IEEE Trans. Nucl. Sci.,* vol. 37, pp. 2013–2019, December 1990.
39. S. T. Liu, W. C. Jenkins and H. L. Hughes, "Total dose radiation hard 0.35 pm S0I CMOS technology," *IEEE Trans. Nucl. Sci.,* vol. 45, no. 6, pp. 2442–2449, December 1998.
40. F.-L. Hsueh and L. S. Napoli, "CMOS/SOS high soft-error threshold memory cell," *IEEE Transactions on Nuclear Science,* vol. NS-32, no.6, pp. 4155–4158, December 1985.

41. H. T. Weaver et al., “An SEU tolerant memory cell derived from fundamental studies of SEU mechanisms in SRAM,” *IEEE Transactions on Nuclear Science,* vol. NS-34, no. 6, pp. 1281–1286, December 1987.
42. D. Bessot and R. Velazco, “Design of SEU-hardened CMOS memory cells: the HIT cell,” *IEEE RADECS,* 1993.
43. M. J. Myjak, D. R. Blum and J. G. Delgado-Frias, “Enhanced fault-tolerant CMOS memory elements circuits and systems,” MWSCAS, *2004 47th Midwest Symposium on* vol. 1, pp. 453–456, July 2004.
44. M. N. Liu and S. Whitaker, “Low power SEU immune CMOS memory circuits,” *IEEE Transactions on Nuclear Science,* vol. 39, no. 6, pp. 1679–1684, December 1992.
45. S. Whitaker, J. Canaris and K. Liu, “SEU hardened memory cells for a CCSDS Reed Solomon encoder,” *IEEE Transactions on Nuclear Science,* vo1. 38, no. 6, pp. 1471–1477, December 1991.
46. L. R. Rockett, “An SEU-hardened CMOS data latch design,” *IEEE Transactions on Nuclear Science,* vo1. 35, no. 6, pp. 1682–1687, December 1988.
47. C. Carmichael, J. Fabula, R. Padovani and R. Reis, “A fault injection analysis of Virtex FPGA TMR design methodology Lima,” 6th European Conference on Radiation and Its Effects on Components and Systems, RADECS 2001.
48. W. Heidergott, “SEU tolerant devices circuits and processor design,” DAC 2005, June 13–17, 2005, Anaheim, CA.
49. P. Dodd, M. Shaneyfelt, J. Felix and J. Schwank, “Production and propagation of single-event transient in high-speed digital logic ICs,” *IEEE Transactions on Nuclear Science,* vol. 51, no. 6, pp. 3278–3284, December 2004.
50. P. Sweeny, “*Error control coding, from theory to practice*,” John Wiley and Sons, 2002.
51. A. V. Oppenheim and R. W. Schafer, *Discrete time signal processing*, Prentice Hall, 1999.
52. B. Sklar, “*Digital communications: fundamentals and applications*,” Prentice Hall, 1988.
53. P. Reviriego, J. A. Maestro, and O. Ruano, “Efficient protection techniques against SEUs for adaptive filters: an echo canceller case study,” *IEEE Transactions on Nuclear Science,* vol. 55, no. 3, pp. 1700–1707, June 2008.
54. A. L. N. Reddy and P. Banarjee, “Algorithm-based fault detection for signal processing applications,” *IEEE Transactions on Computers,* vol. 39, no. 10, pp. 1304–1308, October 1990.
55. S. Wang and N. K. Jha, “Algorithm-based fault tolerance for FFT networks,” *IEEE Transactions on Computers,* vol. 43, no. 7, pp. 849–854, July 1994.
56. S. Lu, J. Shih and S. Huang, “Design-for-testability and fault-tolerant techniques for FFT processors,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 13, no. 6, pp. 732–741, June 2005.
57. G. R. Redinbo, “System level reliability in convolution computations,” *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 37, no. 8, pp.1241–1252, August 1989.
58. P. E. Beckmann and B. R. Musicus, “Fast fault-tolerant digital convolution using a polynomial residue number system,” *IEEE Transactions on Signal Processing,* vol. 41, no. 7, pp. 2300–2313, July 1993.
59. S. Sundaram and C. N. Hadjicostis, “Fault-tolerant convolution via Chinese remainder codes constructed from non-coprime moduli,” *IEEE Transactions on Signal Processing,* vol. 56, no. 9, pp. 4244–4254, September 2008.
60. A. B. O´Donnell and C. J. Bleakley, “Area efficient fault tolerant convolution using RRNS with NTTS and WSCA,” *Electronics Letters*, vol. 44, no. 10, pp. 648–649, May 8, 2008.

61. B. Shim, S. R. Sridhara and N. R. Shanbhag, “Reliable low-power digital signal processing via reduced precision redundancy,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 12, no. 5, pp. 497–510, May 2004.
62. B. Shim, N. R. Shanbhag and S. Lee, “Energy-efficient soft error-tolerant digital signal processing,” *Signals, Systems and Computers Conference Record of the Thirty-Seventh Asilomar,* pp. 1493–1497, November 2003.
63. R. Hentschke, F. Marques, F. Lima, L. Carro, A. Susin, and R. Reis, “Analyzing area and performance penalty of protecting different digital modules with Hamming code and triple modular redundancy,” *15th Symp. on Integrated Circuits and Systems Design,* pp. 95–100, 2002.
64. P. Reyes, P. Reviriego, J. A. Maestro and O. Ruano, “New protection techniques against SEUs for moving average filters in a radiation environment,” *IEEE Transactions on Nuclear Science,* vol. 54, no. 4, pp. 957–964, August 2007.
65. H. Sato, “Moving average filter,” U.S. Patent 6,304,133, October 16, 2001.
66. IEEE 802.3i Ethernet over Unshielded Twisted Pairs Standard (10BaseT), 1990.
67. O. Ruano, J. A. Maestro, P. Reyes and P. Reviriego, “A simulation platform for the study of soft errors on signal processing circuits through software fault injection”, *Proc. of IEEE International Symposium on Industrial Electronics (ISIE’07),* Vigo, Spain, pp. 3316–3321, June 2007.

15 Fault-Injection Techniques for Dependability Analysis: An Overview

Massimo Violante

CONTENTS

15.1 INTRODUCTION

Designers of electronic systems are relying on simulation and emulation tools for their work: the common design practice entails the building of models of the system being designed that are simulated to evaluate whether the design solutions meet the user requirements. Nowadays a top-down design flow is typically used: starting from a very abstract model of the system, which captures the behavior (i.e., the algorithm) the system has to implement, details are added by refining iteratively the model, eventually obtaining a cycle-accurate description of the structure of the system, which can be emulated using an appropriate hardware platform, to verify whether the design actually meets all the user requirements and which can be used for manufacturing of the system. When a reasonable confidence about the correctness of the model is reached, the manufacturing can start. Thanks to this approach, shorter time to market and lower costs can be achieved, as design bugs and inconsistency with user requirements can be captured when production is not yet started.

The electronic design automation (EDA) industry is offering designers a wide range of products that support the top-down design flow. Hardware description languages (HDL) are available (e.g., VHDL, Verilog [1], and SystemC [2]) that can be used for describing systems, and a number of HDL simulators are available for studying the dynamic behavior of the obtained models. Both the description languages and the simulation tools support different abstraction levels and domains of representations so that designers can start reasoning on the system behavior during the early design stages and following the top-down flow can end up with a structural description of the system that is represented as a netlist connecting logic gates.

Besides simulation tools, a number of EDA solutions are available for supporting designers in the refinement of the model. Tools are available for analyzing of the partitioning between hardware and software implementations of the system functions, and synthesis tools are available to implement in automatic fashion the transitions from higher, less detailed, abstraction levels/representation domains to lower, mode detailed, abstraction levels/representation domains.

The EDA tools are mostly focused on the analysis and the synthesis of the functionalities the system must implement to fulfill the user requirements. If we limit our analysis to simulation tools, we can see that they provide efficient ways for evaluating how the model reacts to input stimuli representing the typical workload the system has to process, and they provide advanced debug features to simplify the identification of design bugs or part of the model that does not fulfill adequately the user requirements. Moreover, they can provide timing and power information to simplify the tuning of parameters like speed and power consumption.

Nowadays simulation tools are essential ingredients of any successful design flow, and they are widely adopted in industries because they provide the type of support designers need: the capability of producing useful information and of dealing with ever growing designs by exploiting clever simulation algorithms and possibly taking advantage of dedicated hardware emulators. To continue to provide useful support to designers of electronic systems, we expect that in the coming years simulation tools will start providing features related to dependability evaluation.

The concept of dependability, which can be seen as the property of an electronic system of being able to deliver service that can justifiably be trusted [3], is well known to developers of electronic systems employed in mission- or safety-critical applications, where failures may lead to loss of money or human lives. An essential part of the design flow of this kind of system is indeed the dependability evaluation process, which must provide evidence of the capability of the system to fulfill the dependability requirements. Since a few years ago, *dependability* and *dependability evaluation* were concepts bound mostly to very specific application domains, like space, military, medical, and transportation. With the advent of deep submicron (DSM) manufacturing technologies, things are changing, and dependability is also becoming an important concept for developer and commodity applications. Moreover, faulty scenarios that were bound to very specific application domains (e.g., space, nuclear science, and medical) are becoming more and more important as DSM technology becomes widely used.

DSM technology is offering multi-GHz operational frequency, very low voltages, and very high integration capabilities. Although these results have positive effects on the applications built on top of DSM technology, the scaling of feature sizes and operation voltages are reducing significantly both noise margins and the amount of charge needed to store information in memory elements. Consequently, low-energy radioactive particles that are present at ground level can easily introduce perturbations to DSM-based electronic systems, which may eventually show unexpected errors [4]. Moreover, as pointed out in [5], systems manufactured using DSM technologies are more likely to be affected by failures than those manufactured using older technologies.

The higher sensitivity of DSM technologies to perturbations induced by the environment and the higher probability for DSM-based systems of being affected by failures during the normal operational life, are posing new challenges to designers. When developing DSM-based systems, designers must be aware of a nonnegligible failure probability, must employ suitable countermeasures to deal with failures, and must rely on tools suitable to evaluate the impact of failures on their designs and to validate the capabilities of the adopted countermeasures. As a result, developers of DSM-based systems, even those aiming at commodity application, will benefit from the know-how established in those application domains targeting mission- or safety-critical applications by adopting the same dependability evaluation techniques that are likely to be integrated in the most common tool designer exploits: the simulation tool.

Among the different techniques for performing dependability evaluation, the one we believe is most suitable for being introduced in simulation tools is fault injection [6]. The concept of fault injection consists of inoculating a fault in a system and observing how the fault propagates within it, eventually reaching the system outputs. This concept perfectly fits with the purpose of simulation tools that offer the capability of studying the dynamic behavior of a system model. Fault injection can be provided as an additional feature of simulation tools, which will allow designers studying the dynamic behavior of system models when affected by faults, along with the more traditional and more established features oriented to design debug, performance, and power evaluations.

In this chapter we will first outline the general architecture of a fault-injection system and then will describe fault-injection techniques that are suitable for being integrated within simulation tools. Moreover, as the complexity of systems is ever growing, we will present additional fault-injection techniques that can be employed when using emulation hardware often exploited to speed up the simulation of very complex models. Finally, as more and more systems include embedded software running on embedded processors, specific fault-injection techniques developed for attacking software-based systems will be presented.

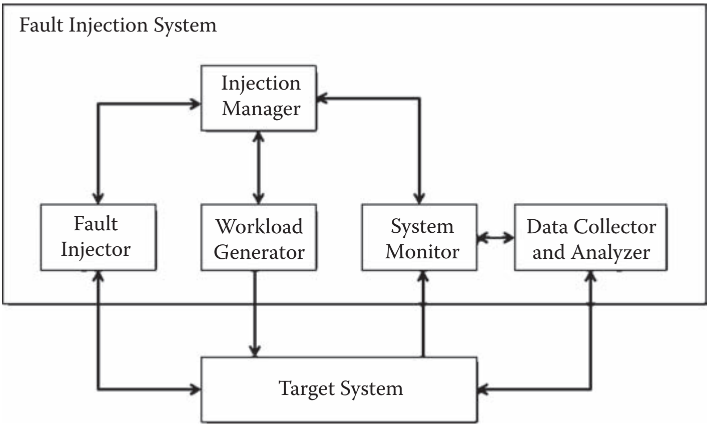


FIGURE 15.1 Overview of a fault-injection system.

15.2 OVERVIEW OF A FAULT-INJECTION SYSTEM

The basic components of a fault-injection system are depicted in Figure 15.1. The target system is the object of the investigation, where faults will be inoculated and whose behavior will be monitored to study the impact of the injected faults. As this discussion is general, we do not refer to a specific type of target system; for the sake of this section the target system can be a behavioral model of an electronic system, a structural model, a hardware-emulated model, or even a prototype implementing the functionalities of the model.

With the term *fault* we refer to the malfunctions the target system may be subject to during its lifetime, which may let the target system behavior deviate with respect to the intended one. The malfunctions that affect a system may be as follows:

- *Permanent*, in which case the malfunction always affects the system.
- *Transient*, in which case the malfunction affects the system when a certain set of conditions is met. Otherwise, the system functions normally.

During injection, fault models are used to capture such malfunctions, which are described using the same modeling approach used for the target system. Among them, the most widely used is the *single stuck-at*, which models a permanent malfunction that may affect a circuit by tying permanently to one or to zero one net of the structural model of the target system.

In more recent years a new fault model, the *bitflip* fault model, is beginning to be used in the industry to represent the malfunctions provoked by the charge deposition induced by a radioactive particle in a semiconductor device [7]. The bitflip fault model consists in the random mutation of the content of one memory element of a design, which changes its content from 0 to 1 (or vice versa). Bitflips are random in both time and space: they can affect a system during its entire lifetime and can strike any of the memory elements the system embeds. It is worthwhile to remark here that the bitflip fault model refers to the modification of the stored information that is corrupted and not to the storage cell that preserves its correct functionality.

The main components of a typical fault-injection system are as follows:

1. *Injection manager*: supervises the injection campaign. Given the list of faults that have to be inoculated in the system (i.e., the *fault list*), it runs an injection experiment for each of them. Each experiment encompasses the following steps:
   a. The first fault, *f*, to be injected is selected from the fault list.
   b. The target system, the workload generator, the data collector, and analyzer are set to the reset state.
   c. The fault injector is programmed to inoculate fault *f* and the system monitor is programmed accordingly. For example, let's suppose fault *f* has to be injected in the target system at time t_f, where t_0 is the time when the first input stimuli is applied. The fault injector is instructed to stop the system at time t_f, to inoculate the fault (according to the type of the target system), and to resume the target system after fault injection to let the fault propagate. Moreover, the system monitor is programmed to trigger data collection from time t_f until the end of the workload.
   d. The workload generator is activated. The input stimuli are applied to the target system; in accordance with step (c) fault *f* is injected in the target system, and data are collected from t_f onward.
   e. Upon the completion of the workload, the fault effect is classified, and the whole procedure is repeated from step (a).
2. *Fault injector*: inoculates a fault in the target system as it activated by the by the workload generator. As detailed in the following sections, a number of techniques can be used to achieve fault inoculation, depending on the type of the target system.
3. *Workload generator*: generates the input stimuli to activate the target system during the fault-injection experiment. The input stimuli can be synthetic workload generated ad hoc, or real inputs taken from the application where the system will be deployed.
4. *System monitor*: observes the target system and, when necessary, triggers the collection of data from the target system.
5. *Data collector and analyzer*: when triggered by the monitor, collects from the target system data that are useful to classify the impact of the inoculated fault. For example, it can collect the outputs produced by the target system as well as status information. The collected data are then processed by the data analyzer, which produces the classification of the fault effect. This task is normally performed by comparing the data collected on the faulty system with those produced by the fault-free system when activated by the same workload used during the injection experiment.

A number of different implementations are possible for the fault-injection system, which can be grouped in different categories as a function of the type of the target system. We can have the following:

1. *Simulation-based fault injection*: This category collects all the methods that have been developed to inoculate faults in a model of the target system whose dynamic behavior is evaluated through the use of simulation tools. The category can be further divided into subcategories by considering the abstraction level at which the target system is modeled:
   a. *System-level simulation*: where the system is described in terms of complex components like processors executing software, memories, input/output (I/O) peripherals, possibly connected through a network infrastructure. This abstraction level is suitable for modeling complex systems, such as for a cluster of computers that build a server farm.
   b. *Register-transfer-level simulation*: where the system is described in terms of components like registers, arithmetic and logic units, and caches. This abstraction level is suitable when the target system is a component in a large infrastructure, such as with one of the network card of a computer in a server farm.
   c. *Gate-level simulation*: where the system is described in terms of logic gates and simple memory elements. This abstraction level is suitable when the target system is a component in a larger infrastructure, such as with a protocol controller chip inserted in a network interface card.
2. *Emulation-based fault injection*: where the system is first described as in the register-transfer-level case, and it is then emulated using dedicated hardware, such as field-programmable gate array (FPGA)-equipped boards. This subcategory can be seen as an evolution of register-transfer-level simulation, where hardware emulation is exploited to boost the simulation performance. Indeed, when very complex workloads and very large fault lists have to be considered, the time spent for each fault-injection experiment can be prohibitive, and means to reduce it are needed.
3. *Software-based fault injection*: where the system is a physical model (i.e., a prototype), composed of processors, memories, and I/O peripherals, possibly connected through a network infrastructure. Fault injection is implemented by means of specially crafted software that is added to the software the target system executes to implement the desired functionality. This subcategory is intended for performing dependability evaluation when the prototype of the target system is available and can be applied only to processor-based systems.

The following sections discuss in further detail the different categories of fault-injection systems.

15.3 SIMULATION-BASED FAULT INJECTION

Simulation-based fault injection refers to all the methods that have been developed to inoculate faults in a model of the target system whose dynamic behavior is evaluated through the use of simulation tools. Referring to Figure 15.1, we have the following:

- The target system is an executable model written in a description language (e.g., VHDL, Verilog, C/C++). Depending on the phase of the design flow when fault injection is used, the model can be at the system, RT, or gate level. Lower levels (transistor or device level) are possible but are not considered here.
- The workload generator is normally a testbench [8] for the target system model, and it is applied to the target system by exploiting a simulation tool. Commercial off-the-shelf tools can be used for this purpose, which are the same ones used during the normal design practice.
- Injection manager, fault injector, system monitor, and data collector and analyzer are ad hoc software. They can be stand-alone software modules that run on top of commercial off-the-shelf simulation tools, or they can be integrated within the simulation tool in case it provides dependability-oriented features. In particular, fault injection can be implemented as follows.
    - The simulation tool is enriched with algorithms that allow not only the evaluation of the faulty-free target system, as normally happens in VHDL or Verilog simulators, but also their faulty counterparts. This solution is very popular as far as certain fault models are considered (e.g., commercial tools exist that support the evaluation of permanent faults like the stuck-at or the delay one [9]). Conversely, there is limited support for fault models that represent radiation-induced faults; therefore, designers have to rely on prototypical tools either built in-house or provided by universities.
    - The model of the target system is enriched with special data types or with special components that are in charge of supporting fault injection. This approach is quite popular since it offers a simple solution to implement fault injection that requires limited implementation efforts, and several tools are available related to adopting it [10-13].
    - Both the simulation tool and the target system are left unchanged, while fault injection is performed by means of simulation commands. Nowadays, it is quite common to find, within the instruction set of simulators, commands for forcing desired values within the model [14]. By exploiting this feature it is possible to support a wide range of fault models.

15.3.1 An Example of Fault Injection Using System-Level Simulation

This section describes the Fault Injection Using Virtual Platforms (FI-VP) tool developed at Politecnico di Torino for performing the injection of bitflips in target systems described at the system abstraction level. The purpose of the tool is to perform dependability analysis at that step of the design flow where the system architecture has been established, potentially exploiting the application software the system will run, but a prototype of the system is not available yet. The system is modeled as a structure of interconnected components that can be described according to different styles. They can be instruction set simulators for processor cores; behavioral/structural models of standard components like memories, network cards, and processor bridges; and user-defined behavioral/structural models of custom hardware. Virtutech's Simics [15] is used for performing system simulation.

Figure 15.2 depicts how the generic structure of a fault injection system has been customized while implementing the FI-VP tool. A custom program coded in C language implements the fault injector and the system monitor, while workload generator and target system are modeled and executed using Virtutech's Simics. The data collector and analyzer is a user-provided C function called by the FI-VP core at the end of each fault-injection experiment to implement fault effect classification.

For each injection experiment the fault injector generates a command script file that tells Simics at which time the fault has to be injected and how to perform fault injection (the commands Simics offers run/stop the model and to alter the content of any simulation object are used for this purpose). The outputs of the target system are collected (through Simics commands) and are stored in a trace log file, which is processed at the end of the simulation by the data collector and analyzer.

We adopted such architecture as any target system has its own peculiarity, and it is up to the user to define which outputs and which status information have to be collected and used for fault effect classification. When preparing the fault-injection campaign, the user is thus asked to provide a C function, which is linked to the FI-VP core, to process the trace log generated by each fault injection. For implementing data analysis, before starting the fault-injection campaign, the fault-free system is executed once, collecting the reference trace log to be used during fault effect classification. The main benefit of the architecture lies in its generality. Any model can undergo

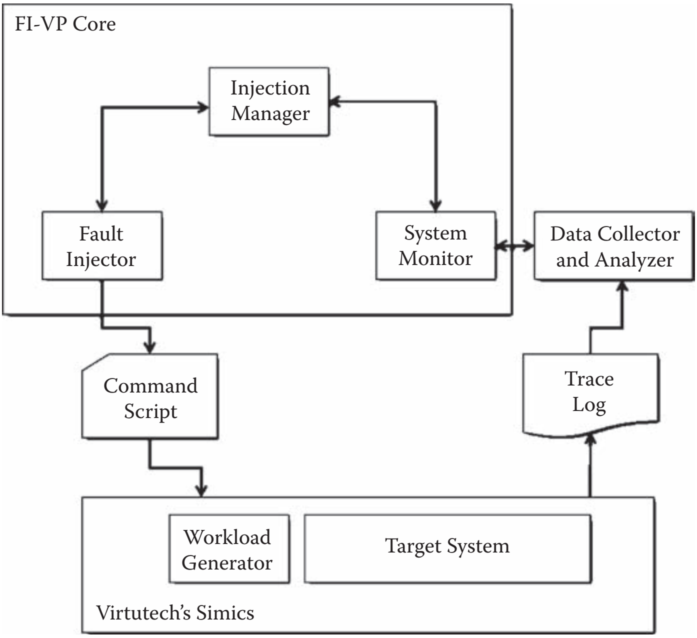


FIGURE 15.2 Architecture of the FI-VP tool.

fault injection, reusing most of the FI-VP architecture as is, with the exception of the data collector and analyzer module, which has be provided by the user.

The FI-VP was used to perform fault injection in a target system composed of one PowerPC440GP and 128 Mbytes of RAM. The target system runs a Linux 2.6 operating system and an application that performs 10,000 times the product of two 10 × 10 matrices. When injecting 10,000 faults in the processor program counter, we recorded the following fault effects.

- 32.94% of injected faults do not affect the system behavior. The workload produces the expected results.
- 26.18% of the injected faults corrupt the matrix multiplication application, which produces results different from the expected one.
- 10.20% of the injected faults lead to a CPU trap, indicating that the error triggered an error detection mechanism the PowerPC embeds.
- 30.59% of the injected faults corrupt lead to operating system crash.

15.3.2 An Example of Fault Injection Using Register-Transfer-Level Simulation

This section describes the Fault Injection Using VHDL (FI-VHDL) tool developed at Politecnico di Torino for performing injection of bitflips in target systems described at the register-transfer abstraction level. The purpose of the tool is to support designers in developing dependable components to be implemented as application-specific integrated circuits (ASICs) of FPGAs. The system is modeled as a structure of interconnected modules described using the VHDL language either at the behavioral, data flow, or structural representation domain. Target system simulation is implemented using the Mentor Graphics Modelsim tool [16]. The architecture of FI-VHDL is the same as FI-VP, while Simics is replaced with Modelsim.

In the case of register-transfer-level simulations, injection campaigns may require huge amounts of time (many hours or days) for their execution, depending on the complexity of the model of the target system, the efficiency of the VHDL simulator adopted, of the workstation used for running the experiments, as well as the number of faults that have to be injected. To overcome this limitation, we presented in [14] a technique aimed at minimizing the time spent for running fault injection. The technique encompasses three steps:

1. *Golden run execution*: the target system is simulated without injecting any fault and a trace log file is produced, gathering information on the target system behavior and on the state of the simulator.
2. *Static fault analysis*: given an initial list of faults that must be injected, by exploiting the information gathered during the golden run execution we identify those faults whose effects on the target system can be determined a priori and remove them from the fault list. Since the injection of each fault encompasses the simulation of the target system, by reducing the number of faults that we need to inject we are able to reduce the time needed for the whole experiment.

3. *Dynamic fault analysis*: during the injection of each fault, the state of the target system is periodically compared with the golden run at the correspondent time instant. The simulation is stopped as soon as the effect of the fault on the target system becomes known (e.g., the fault triggered some detection mechanisms, disappeared from the target system, or manifested itself as a failure). Although the operations needed for comparing the state of the target system with that of the golden run come at a nonnegligible cost, the benefits they produce on the time for running the whole experiment are significant. In general, a fault is likely to manifest itself (or to disappear) a few instants after its injection. As a result, by monitoring the evolution of the fault for a few simulation cycles after its injection, we may be able to stop the simulation execution in advance with respect of the completion of the workload. We can thus save a significant amount of time. Similarly, in case the fault is still latent until a few simulation cycles after its injection, it is likely to remain latent, or manifest itself, until the completion of the workload. In this case, the state of the target system and that of the golden run are no longer compared, thus saving execution time until the end of the injection experiment.

15.3.3 Final Remarks on Simulation-Based Fault Injection

Simulation-based fault injection is a powerful technique that we expect to appear in the next years as an additional feature of simulation tools, as we anticipate that dependability evaluation will become a primary design issue. In particular, we expect that simulation-based fault injection will be used to debug and validate the fault detection and correction embedded in future designs target DSM technology.

As simulation is a slow process and the complexity of designs is ever growing, we expect that simulation-based fault injection will be mainly used for injecting a small set of carefully selected faults for outlining possible design bugs. The adoption of a top-down design flow, based on exploiting system-level simulation for analyzing complex infrastructures, and demanding to lower level simulations the analysis of detailed models of infrastructure components, can alleviate the simulation speed problem.

However, when very complex workloads and huge amounts of faults have to be injected we will reach the limit of the capability of simulation-based fault injection. Simulation time will be unfeasible, even in the case of very abstract system models, and alternative solutions to speed up injection campaigns will be needed. The next section presents the concept of emulation-based fault injection, along with an example that is an effective solution to the simulation speed problem.

15.4 EMULATION-BASED FAULT INJECTION

As the complexity of target systems is ever growing, and so is the execution time needed for running simulation-based fault-injection campaigns, several researchers proposed to boost performance by running the target system in hardware instead of using simulation tools. For this purpose FPGAs are used to implement a prototype of

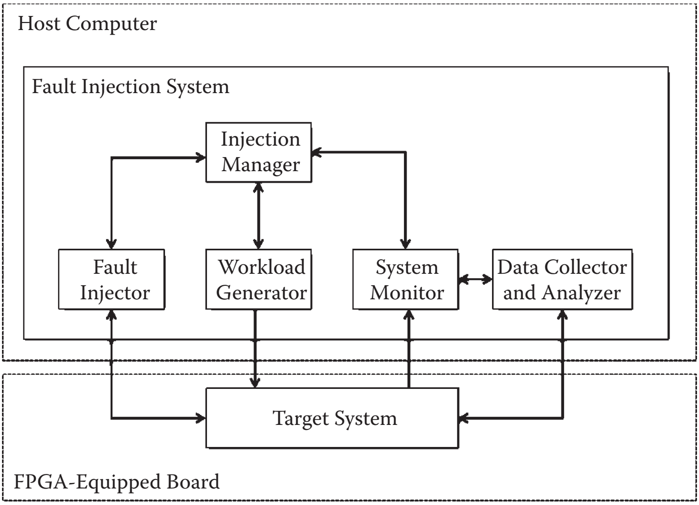


FIGURE 15.3 Conceptual architecture of an emulation-based fault-injection system.

the target system so that workload evaluation is done using actual hardware [17-19]. The conceptual architecture of an emulation-based fault-injection system is outlined in Figure 15.3.

A host computer runs the software modules implementing the fault-injection system, while an FPGA-equipped board emulates the target system. The idea is potentially very effective as the speed ratio between the simulation of a model of the target system and hardware emulation of the same model can be several orders of magnitude.

According to the conceptual architecture of Figure 15.3, input stimuli, the collected data, and the operations needed to implement fault injection have to travel from the host computer to the FPGA-equipped board (and vice versa). As a result, the communication link between these two components becomes the bottleneck, which may limit the actual performance improvement.

As far as the operations needed for fault injection are considered, two options have been proposed so far:

- Insert into the target system features to support the inoculation of the fault model of concern. For example, the approach presented in [17] replaces every memory element of the device with a special cell that offers fault-injection capabilities. Thanks to this approach every FPGA model can be used for target system emulation, but modifications to it have to be inserted. As a result, this approach is applicable only when the model of the target system is accessible and modifiable, while it cannot be exploited when the model includes encoded intellectual property (IP) cores. Moreover, issues can be raised on the intrusiveness of the method: as the model of the target system that undergoes fault-injection

is different from that of the system that will be deployed in the application, efforts have to be spent in validating the representativeness of the attained observations.
- Adopt the features of the FPGA device used to emulate the target system. Approaches like [18,19] exploit the reconfiguration capabilities of Xilinx FPGA for inoculating different types of faults. The main benefit of this approach lies in the capability of performing fault injection without altering the structure of the target system; therefore, it is possible to inject faults also in models that embed encoded IP cores, and the representativeness of the achieved result is less questionable, as no modifications to the model of the target system are required.

As far as the communication link bottleneck is considered, two solutions have been analyzed:

- Adopt a high-speed communication bus, like USB 2.0 or PCI, to maximize the transfer rate between the host computer and the FPGA-equipped board. This solution is preferable when commercial off-the-shelf boards are used for implementing the FPGA-equipped board. The architecture remains the same as in Figure 15.3, and the FPGA is used for emulating the target system only.
- Move the workload generator, system monitor, and data collector and analyzer to the FPGA-equipped board by exploiting a custom-designed board. Thanks to this approach, which is typically more expensive than the previous one as it may require the design of custom equipment, it is possible to optimize the communication link used to deliver input stimuli and to collect output data from the target system. To improve further the performance, the fault injector is normally moved to the FPGA-equipped board, thus leaving to the host computer only the supervising task implemented by the injection manager.

15.4.1 An Example of Emulation-Based Fault Injection

As an example of emulation-based fault injection we present FT-UNSHADES [19]. The tool was developed at the School of Engineering of the University of Seville with the support of the European Space Agency to perform the analysis of the effects of single bitflips in designs intended for space applications.

The tool supports the emulation of target systems modeled as IP cores; hence limited knowledge of the internal structure of the target system is needed for running injection campaigns. Moreover, no modifications are needed to the target system to support fault inoculation; thus, FT-UNSHADES can be credited with providing accurate analysis on the faulty behavior of the actual system when deployed in the field.

FT-UNSHADES implements the conceptual architecture of emulation-based fault injection, as depicted in Figure 15.4.

An ad hoc FPGA-equipped board is exploited where an FPGA device (Xilinx Virtex 8000 device) implements three functionalities:

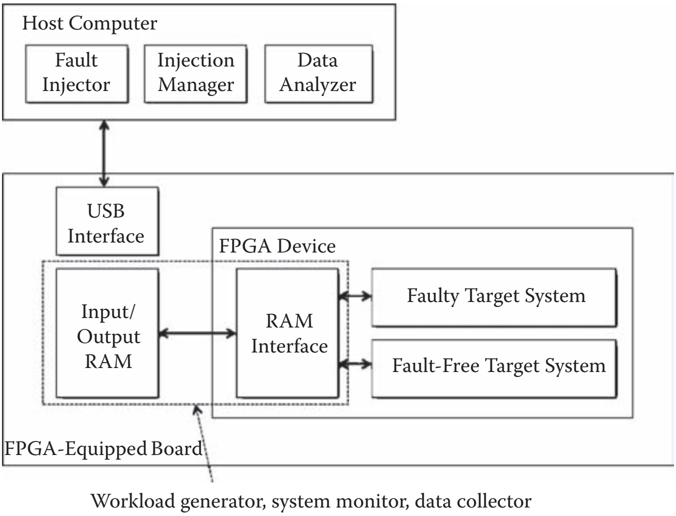


FIGURE 15.4 Architecture of FT-UNSHADES.

1. A RAM interface circuit to access the I/O RAM chips located outside the FPGA. These chips store the workload to be used during injections and the responses of the injection experiment.
2. An instance of the target system that will be subject to bitflip fault inoculation, the so-called faulty target system.
3. An instance of the target system that will not be affected by faults, the so-called fault-free target system.

The board also features I/O chips for workload and output data storage, and a USB chip to communicate with the host computer.

By moving the workload generator close to the device used for target system emulation, the designers of FT-UNSAHDES minimized the volume of data that has to be exchanged with the host computer during testing. As a result, the injection speed no longer depends on the throughput of the interface between the host computer and the FPGA-equipped board, but only on the much faster on-board bus between the I/O RAM and the FPGA device.

Moreover, the adoption of two instances of the target system simplifies the system monitor and data collection functions. The same input stimuli are processed by the faulty and fault-free target systems at the same time; therefore, the system monitor can be implemented by comparing the output signals produced by the two instances, looking for mismatches. Data collection is triggered only in case a mismatch is detected, and upon such an event the injection experiment is stopped. As a result, the amount of output data to be sent to the host computer for each experiment is minimized: it is a status indicating either no mismatches or that a mismatch occurred.

The host computer implements the fault injector, injection manager, and data analyzer functions. Among them, the most interesting is the fault injector, which performs the inoculation of single bitflips in the faulty target system.

Fault inoculation is obtained by exploiting the partial reconfiguration feature of the Xilinx FPGA device used for emulating the target system. Using such a feature, the FT-UNSHADES software operates according to the following algorithms:

1. Activate workload generator until the injection time T_i is reached.
2. Read from the FPGA device the value *v* of the flip-flop *ff* to be affected by bitflip.
3. If $v = 0$, perform a partial reconfiguration of the FPGA to assert the set control signal of *ff* and proceed to 5.
4. If $v=1$, perform a partial reconfiguration of the FPGA to assert the reset control signal of *ff*.
5. Reactivate the workload generator until the workload is completed, or a mismatch indication is received.

Thanks to this approach, and exploiting the one-hot behavior of the flip-flip reset/set control signals, one partial configuration operation is needed for each fault in the fault list. The amount of data that have to be exchanged between the host computer and the FPGA-equipped board thus account for two configuration frames (few hundreds of 32-bit words). As a result, thanks to the speed of the USB interface, and the limited amount of data that have to be exchanged, injection speed is not bounded by the host computer/FPGA-equipped board interface.

Thanks to its custom design, FT-UNSHADES can reach notable fault injection speed. As an example, the injection of faults in a design with 796 flip-flops activated by a workload encompassing 200,000 clock cycles takes 0.13 seconds on the average for each fault.

15.4.2 Final Remarks on Emulation-Based Fault Injection

Emulation-based fault injection is an effective solution to the problem of assessing the dependability of complex designs when very large fault lists and extensive sets of input stimuli have to be considered. To achieve such a goal, the following conditions have to be met:

1. A suitable emulation-based fault injection system must be available that is capable of hosting the system under test and able to provide efficient communications mechanisms to achieve high throughputs. To optimize performance, ad hoc hardware is likely to be needed.
2. For emulation purposes the model of the system under test should be suitable for being implemented using FPGA devices. This requirement imposes some limitations on the style the designers have to use to code the model and is likely to be enforced only when many of the design decisions have been taken. As a result, emulation-based fault injection is likely to be used only in the late phase of the design flow, when the detailed model of the system is available.

15.5 SOFTWARE-BASED FAULT INJECTION

Virtually any electronic system today embeds processor cores running application software, and the systems used in critical applications follow the same trend. During the early phases of the design flow, models of the processor cores can be used for running simulation- or emulation-based fault injection. Later in the design flow, when a system prototype is available, different approaches are in general required. Indeed, when the prototype is built using ad hoc manufactured ASICs and discrete processor chips, different techniques from those based on the features of simulator tools and FPGA devices are needed. In the following we will focus on processor chips only, letting the reader refer to other sources (e.g., [20]) for an overview of techniques that can be used for inoculated faults in ASICs.

The concept of software-based fault injection consists of enriching the application software the processor runs with a routine—the injection routine—whose purpose is to implement fault inoculation. The injection routine is never used during the normal operations of the system, as it does nothing useful for the system user. The injection routine is activated only when the fault under investigation has to be inoculated, and its execution is stopped as soon as the injection is performed.

In the past years several techniques have been proposed to put in practice software-implemented fault injection. All of them use similar injection routines that, when activated, can modify any of the user-accessible resources. As fault inoculation takes place by means of software, it can affect only those parts of the system that can be reached using the processor instruction set and that are hence visible to the programmer. As an example, in case we are interested in inoculating bitflips in a processor, software-implemented fault injection allows reaching all the registers of the instruction set architecture (e.g., general and special purpose registers, control, and status registers), as well as all the memory addresses the processor is able to reach. Conversely, those registers that are embedded in the processor but that are invisible to the programmer (e.g., the boundary registers in the processor pipeline) cannot be attacked by injection, as instructions are not available to alter their contents.

The techniques developed so far differ in the method used for triggering the injection routine. The following methods have been proposed:

1. In case the processor runs an operating system, the services the operating system provides are used to trigger the injection routine. As an example, FERRARI [21] exploits the ptrace() system call of the Unix/Linux operating system to break the execution of the application whose behavior has to be studied in the presence of faults. Other approaches implement injection routine triggering by means of ad hoc developed device drivers [22].
2. In case the processor does not run an operating system, or in case modifications to the application software are not possible, the activation of the injection routine can be implemented using an interrupt request. According to this technique, the injection routine is set to be the handler of an interrupt not used by the system so that injection takes place when the corresponding interrupt is triggered. The Xception tool [23] exploits processor self-generated interrupt (e.g., software traps) to perform injection when one of the following events is

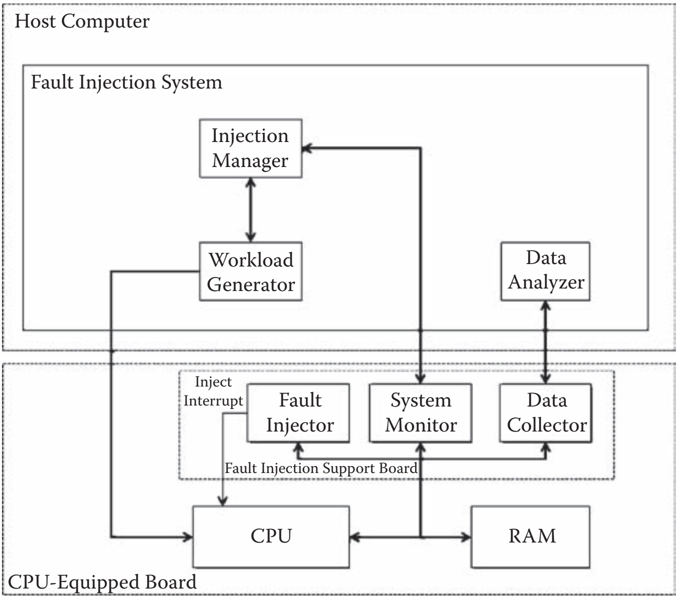


FIGURE 15.5 Architecture of FISB.

detected: instruction fetch from a specific address, operand load/store at a specific address, a specific time is reached, a combination of the aforementioned events. In [24] the authors suggest activating the injection routine through the processor self-generate debug trap; moreover, they suggest exploiting the Background Debug Mode (BDM) Motorola processors have to avoid performance reduction. Conversely, the FISB [25] exploits a programmable board that monitors the processor bus and triggers the interrupt request to which the injection routine is attached. The board can be instructed to trigger injection when one of the following events is detected: a certain number of instructions fetch is reached, or a specific address is accessed.

15.5.1 An Example of Software-Implemented Fault Injection

In this section we present the software-implemented fault-injection system introduced in [25], which makes use of the custom-developed fault-injection support board (FISB) to perform the injection of bitflip in a processor-based system.

The purpose of the system shown in Figure 15.5 is to allow fault injection in a processor-based system, minimizing the time overhead. The CPU-equipped board that implements the system under investigation is enriched with an ad hoc board (i.e., FISB) whose purpose is to implement system monitor, fault injector, and data collector functionalities, while the other components of the fault injection system are implemented in software and run by a host computer.

FISB is composed of an FPGA and 256 k-words, each 7 bytes wide. It is connected to the processor bus to monitor the processor behavior during application software execution, and it is seen by the processor as a memory mapped device. During its operation, FISB collects the following information:

1. The number of instructions the processor fetched from the memory until the application started.
2. The address and the value read/written from/to the memory by the application.

The fault injection manager can program FISB to trigger the inject interrupt as soon as a desired number of instructions has been executed so that the associated injection routine is activated and fault injection can take place.

A typical fault injection run encompasses the following operations:

1. The fault injection manager programs FISB with the desired injection time T_i, and the desired fault injection mask M, which tells which register/memory address has to be corrupted during injection and which bit of the register/memory address has to be altered by a bitflip.
2. The fault injection manager activates the workload generator so that the application software run by the CPU-equipped board is executed.
3. At injection time T_i, after the execution of the desired number of instructions, FISB issues the inject interrupt request so that the injection routine is executed.
4. The injection routine accesses FISB to read the injection mask M and inoculates the bitflip accordingly.
5. The execution of the application software continues, and the system monitor module that FISB implements observes the behavior of the system under investigation, while the data collector gathers useful data to define the fault effect.
6. At application software completion, the collected data are analyzed, and the fault effect is classified.

The main advantage of using FISB lies in the possibility of running the application software at nominal speed, without any performance degradation. Indeed, FISB works in parallel with the system under investigation, and it interacts with it only when it is time for the injection routine to be activated (via the inject interrupt). As a result, real-time systems can be analyzed as the fault injection system has minimal impact on the system timing.

15.5.2 Final Remarks on Software-Implemented Fault Injection

Software-implemented fault injection is an effective technique to inoculate faults in processors running application software, and several tools are available for supporting such a technique. However, the following limitations have to be noted:

1. When compared with simulation- and emulation-based techniques, software-implemented fault injection results are less versatile in terms of fault

models and fault-injection location. Being based on a software code running on a physical device, certain faults cannot be inoculated (e.g., delay faults). Moreover, only user-accessible resources can be the target of injection, while other hidden resources cannot be attacked. The latter can become an issue when very complex processors are exploited, where many resources remain hidden. Indeed, if we consider modern high-speed processors that can find their way in critical applications demanding very high computing resources, we can see that they include a lot of resources for implementing advanced features like speculative execution and deep pipelines, which are not accessible through the instruction set. Therefore, software-implemented fault injection can hardly be used for assessing the impact of faults in such resources.

2. Software-implemented fault injection requires the modification of the system software to insert the injection routine, which has to be triggered when needed. The size of the injection routine is normally negligible, and thus it does not introduce a significant memory occupation overhead. Conversely, the technique used to trigger injection may impact heavily on the system performance, in particular when special operational modes of the processor are used (e.g., the debug mode), or very computational-intensive features of the operating system are used (e.g., ptrace). In case an interrupt line is used, it must be available.

15.6 CONCLUSIONS

Dependability evaluation is a problem that can find solutions in fault injection, which may have many different implementations. None of them can be considered as the ultimate solution to the dependability evaluation problem, as each of them has its own benefits and its own limitations. However, all of them can be put to work together so that designers can positively exploit their benefits. In an ideal design flow, fault injection should be used extensively since the initial design phases, from the system conception down to its prototypical implementation:

1. Simulation-based fault injection should be used to assist designers in debugging the error detection and correction mechanisms the system under investigation embeds. In this way bugs will be identified as soon as possible, and important parameters such as system performance in the presence of faults can be evaluated at a time in the design flow where modifications have a low impact on the time to market, as they entail low development costs.
2. Emulation-based fault injection should be used when the system model has been consolidated, to perform exhaustive validation of the error detection and correction mechanisms the system embeds. Such a validation is mandatory to avoid producing incorrect systems that will require very expensive design respins. As the number of faults is likely to be several orders of magnitude higher than those considered for design debug, emulation-based fault injection is likely to be the only possibility to keep the injection time

acceptable. In case the system is processor based, software-implemented fault injection should be used as well to complete the validation of the system under investigation before committing it to production.

ACKNOWLEDGMENT

The author wishes to thank the following persons: M. A. Aguirre, M. Alderighi, N. Battezzati, F. Casini, M. Murciano, M. Rebaudengo, M. Sonza Reorda, and L. Sterpone. Many thanks to the students who have worked under my tutorship in the past years: it would be too long to list them all.

REFERENCES

1. N. M. Botros, *HDL Programming Fundamentals: VHDL and Verilog,* Charles River Media, 2005.
2. T. Grotker, S. Liao, G. Martin, and S. Swan, *System Design with System C,* Springer, 2002.
3. A. Avizienis, J. C. Laprie, B. Randell, and C. Landwehr, “Basic Concepts and Taxonomy of Dependable and Secure Computing,” *IEEE Trans. on Dependable and Secure Computing,* Vol. 1, No. 1, 2004, pp. 11–33.
4. J. L. Leray, “Effects of Atmospheric Neutrons on Devices, at Sea Level and in Avionics Embedded Systems,” *Microelectronics Reliability,* Vol. 47, 2007, pp. 1827–1835.
5. M. White and Y. Chen, “Scaled CMOS Technology Reliability Users Guide,” *JPL Publication* 08-14 3/08, 2008.
6. M. C. Hsueh, T. K. Tsai, and R. K. Iyer, “Fault Injection Techniques and Tools,” *IEEE Computer,* April 1997, pp. 75–82.
7. P. Dodd and L. Massengill, “Basic Mechanisms and Modeling of Single-Event Upset in Digital Microelectronics,” *IEEE Transactions on Nuclear Science,* Vol. 50, No. 3, 2003, pp. 583–602.
8. J. Bergeron, *Writing Testbenches: Functional Verification of HDL Models,* Springer, 2003.
9. TetraMAX, Available at: http://www.synopsys.com
10. E. Jenn, J. Arlat, M. Rimen, J. Ohlsson, and J. Karlsson, “Fault Injection into VHDL Models: the MEFISTO Tool,” *Proc. FTCS-24*, 1994, pp. 66–75.
11. T. A. Delong, B. W. Johnson, and J. A. Profeta III, “A Fault Injection Technique for VHDL Behavioral-Level Models,” *IEEE Design & Test of Computers,* Winter 1996, pp. 24–33.
12. D. Gil, R. Martinez, J. V. Busquets, J. C. Baraza, and P. J. Gil, “Fault Injection into VHDL Models: Experimental Validation of a Fault Tolerant Microcomputer System,” *Dependable Computing EDCC-3,* September 1999, pp. 191–208.
13. J. Boué, P. Pétillon, and Y. Crouzet, “MEFISTO-L: A VHDL-Based Fault Injection Tool for the Experimental Assessment of Fault Tolerance,” *Proc. FTCS´98,* 1998.
14. B. Parrotta, M. Rebaudengo, M. Sonza Reorda, and M. Violante, “New Techniques for Accelerating Fault Injection in VHDL Descriptions,” *IEEE International On-Line Test Workshop,* 2000, pp. 61–66.
15. http://www.virtutech.com
16. http://www.mentor.com
17. P. Civera, L. Macchiarulo, M. Rebaudengo, M. Sonza Reorda, and M. Violante, “Exploiting Circuit Emulation for Fast Hardness Evaluation,” *IEEE Transactions on Nuclear Science,* Vol. 48, No. 6, 2001, pp. 2210–2216.

18. L. Antoni, R. Leveugle, and B. Fehér, "Using Run-Time Reconfiguration for Fault Injection in Hardware Prototypes," *Proc. IEEE Intl. Symp. on Defect and Fault Tolerance in VLSI Systems,* 2000, pp. 405–413.
19. H. Guzman-Miranda, J. N. Tombs, and M. A. Aguirre, "FT-UNSHADES-uP: A Platform for the Analysis and Optimal Hardening of Embedded Systems in Radiation Environments,*" IEEE Intl. Symposium on Industrial Electronics,* 2008, pp. 2276–2281.
20. A. Benso and P. Prinetto, *Fault Injection Techniques and Tools for Embedded System Reliability Evaluation,* Kluwer Academic Publishers, 2003.
21. G. A. Kanawati, N. A. Kanawati, and J. A. Abraham, "FERRARI: A Flexible Software-Based Fault and Error Injection System," *IEEE Transactions on Computers*, Vol. 44, No. 2, February 1995, pp. 248–260.
22. G. Cabodi, M. Murciano, and M. Violante, "Boosting Software Fault Injection for Dependability Analysis of Real-Time Embedded Applications," *ACM Transactions on Embedded Computing Systems,* 2009.
23. J. Carreira, H. Madeira, and J. G. Silva, "Xception: Software Fault Injection and Monitoring in Processor Functional Units," *Proc. Fifth IEEE Working Conf. Dependable Computing for Critical Applications,* 1995, pp. 135–149.
24. M. Rebaudengo and M. Sonza Reorda, "Evaluating the Fault Tolerance Capabilities of Embedded Systems via BDM," *IEEE VLSI Test Symposium,* 1999, pp. 452–457.
25. A. Benso, P. L. Civera, M. Rebaudengo, and M. Sonza Reorda, "A Low-Cost Programmable Board for Speeding-Up Fault Injection in Microprocessor-Based Systems," *Annual Reliability and Maintainability Symposium,* 1999, pp. 171–177.