National Aeronautics and Space Administration

Microelectronics Reliability: Physics-of-Failure Based Modeling and Lifetime Evaluation

Mark White
Jet Propulsion Laboratory
Pasadena, California

Joseph B. Bernstein
University of Maryland
College Park, Maryland

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

JPL Publication 08-5 2/08

National Aeronautics and Space Administration

Microelectronics Reliability: Physics-of-Failure Based Modeling and Lifetime Evaluation

NASA Electronic Parts and Packaging (NEPP) Program
Office of Safety and Mission Assurance

Mark White
Jet Propulsion Laboratory
Pasadena, California

Joseph B. Bernstein
University of Maryland
College Park, Maryland

NASA WBS: 939904.01.11.10
JPL Project Number: 102197
Task Number: 1.18.5

Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, CA 91109

http://nepp.nasa.gov

This research was primarily carried out at the University of Maryland under the direction of Professor Joseph B. Bernstein and was sponsored in part by the National Aeronautics and Space Administration Electronic Parts and Packaging (NEPP) Program, the Aerospace Vehicle Systems Institute (AVSI) Consortium—specifically, AVSI Project #17: Methods to Account for Accelerated Semiconductor Wearout—and the Office of Naval Research.

PREFACE

The solid-state electronics industry faces relentless pressure to improve performance, increase functionality, decrease costs, and reduce design and development time. As a result, device feature sizes are now in the nanometer scale range and design life cycles have decreased to fewer than five years.

Until recently, semiconductor device lifetimes could be measured in decades, which was essentially infinite with respect to their required service lives. It was, therefore, not critical to quantify the device lifetimes exactly, or even to understand them completely. For avionics, medical, military, and even telecommunications applications, it was reasonable to assume that all devices would have constant and relatively low failure rates throughout the life of the system; this assumption was built into the design, as well as reliability and safety analysis processes.

Technological pressures on the electronics industry to reduce transitor size and decrease cost while increasing transitor count per chip, however, runs counter to the needs of most high-reliability applications where long life with exceptional reliability is critical. As design rules have become tighter, power consumption has increased and voltage margins have become almost non-existent for the designed performance level. In achieving the desired performance levels, the lifetime of most commercial parts is the ultimate casualty. Most large systems are built with the assumption that electronic components will last for decades without failure. However, counter to this assumption, device reliability physics is becoming so well understood that manufacturing foundries are designing microcircuits for a three- to seven-year useful life, as that is what most of the industry seeks. The military, aerospace, medical, and especially the telecommunications industries cannot afford to depend on custom parts for their most sophisticated circuit designs.

Hence, we have developed this guideline document as an approach for system designers and device reliability engineers to develop a better understanding of device failures as a result of wearout, and to provide a better understanding of how current reliability models are applied in practice. We describe the best possible approaches to modeling reliability concerns in some of the

more advanced microelectronic technologies, and provide in-depth descriptions on how to implement into reliability equivalent circuits for spacecraft, planets, instrument, C-matrix, events (SPICE) simulation. Within the inherent limitations of high-power, high-speed, commercial Complementary Metal Oxide Semiconductor (CMOS) devices, suggestions are developed on how to model the incipient failure rate, how to trade circuit performance with reliability, and how to obtain a predictable end-of-life or component-level system repair rate through realistic time-dependent reliability prediction.

The development of this handbook for evaluating and simulating microelectronic systems reliability has been an ongoing project of the Microelectronics Reliability Engineering program at the University of Maryland, College Park, for more than six years. The program has been funded by the Aerospace Vehicle Systems Institute (AVSI) Consortium and the NASA Electronic Parts and Packaging (NEPP) Program Scaled CMOS Reliability Task, as well as the Office of Naval Research. Several doctoral dissertations have resulted from this work and major contributions were carried out by a number of individuals, including Jöerg Walters, Xiaohu Zhang, Xiaojun Li, Bing Huang, Jin Qin, Mark White, Moshe Gurfinkel, Shahrzad Salami, Qinguo Fan, Zvi Gur, Michael Talmor, and Yoram Shapira.

ACRONYMS

ADC	Analog-to-Digital Converter
AHI	Anode Hole Injection
AHR	Anode Hydrogen Release
ALT	Accelerated Life Testing
AST	Accelerated Stress Tests
ATPG	Automatic Test Pattern Generation
AVSI	Aerospace Vehicle Systems Institute
BERT	Berkeley Reliability Tools
BIR	Built-In-Reliability
BTI	Biased Temperature Instability
CAD	Computer Aided Design
CADMP-2	Computer-Aided Design of Microelectronic Packages
CALCE	Computer-Aided Life-Cycle Engineering
CDF	Cumulative Distribution Function
CFR	Constant Failure Rate
CHC	Channel Hot Carrier
CHE	Channel Hot Electron
CMOS	Complementary Metal Oxide Semiconductor
COTS	Commercial-off-the-Shelf
DAC	Digital-to-Analog Converter
DAHC	Drain Avalanche Hot Carrier
DFR	Design-For-Reliability
DNL	Differential Nonlinearity
EM	Electromigration
EOS	Electrical Overstress
ETM	Effective Temperature Models

FaRBS	Failure Rate Based SPICE
FPGA	Field Programmable Gate Array
FIT	Failure in Time
FN	Fowler-Nordheim
GCA	Gradual Channel Approximation
GIDL	Gate-Induced Drain Leakage
GOS	Gate Oxide Short
HCD	Hot Carrier Degradation
HCI	Hot Carrier Injection
HISREM	Hot Carrier Induced Series Resistance Enhancement Model
HTOL	High Temperature Operating Life
ICs	Integrated Circuits
INL	Integral Nonlinearity
ITRS	International Technology Roadmap for Semiconductor
KCL	Kirchhoff's Current Law
LDD	Lightly Doped Drain
LEM	Lucky Electron Model
LNA	Low Noise Amplifier
LSB	Least Significant Bit
MACRO	Maryland Circuit-Reliability Oriented
MIL-HDBK	Military Handbook
MOS	Metal Oxide Semiconductor
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MSM	Matrix Stressing Method
MTBF	Mean Time between Failures
MTTF	Mean-Time-To-Failure
NBTI	Negative Bias Temperature Instability
NEPP	NASA Electronic Parts and Packaging Program

NMOS	N-Channel Metal Oxide Semiconductor
NMOSFET	N-Channel Metal Oxide Semiconductor Field Effect Transistor
PBTI	Positive Bias Temperature Instability
PMOS	P-Channel Metal Oxide Semiconductor
PMOSFET	P-Channel Metal Oxide Semiconductor Field Effect Transistor
PoF	Physics-of-Failure
RAC	Reliability Analysis Center
RAMP	Reliability Aware Micro-Processor
RF	Radio Frequency
RT	Room Temperature
SFDR	Spurious-Free-Dynamic-Range
SGHE	Secondary Generated Hot Electron
SHA	Sample-and-Hold Amplifier
SHE	Substrate-Hotelectron
S/H	Sample-and-Hold
SNDR	Signal-to-Noise-Plus-Distortion
SNM	Static Noise Margin
SPICE	Spacecraft, Planets, Instrument, C-matrix, Events
SRAM	Static Random Access Memory
SoC	System-on-Chip
SOFR	Sum-of-Failure-Rates
TBD	Time-to-Breakdown
TCAD	Technology Computer Aided Design
TDDB	Time-Dependent Dielectric Breakdown
UIUC	University of Illinois at Urbana-Champaign
VHDL	Very High Density Logic
VTC	Voltage Transfer Characteristics
VLSI	Very Large Scale Integration

CONTENTS

EXECUTIVE SUMMARY

This handbook presents a physics-of-failure approach to microelectronics reliability modeling and assessment. Knowledge of the root cause and physical behavior of key failure mechanisms in microelectronic devices has improved dramatically over recent years and has led to the development of more sophisticated reliability modeling tools and techniques. Some of these tools are summarized here.

Chapter 1 provides an overview of traditional reliability prediction approaches, i.e., MIL-HDBK-217 compared with some of the more recent reliability modeling and prediction approaches, including Reliability Aware Micro-Processor (RAMP) Model, Failure Rate Based SPICE (FaRBS) reliability simulation, and Maryland Circuit-Reliability Oriented (MaCRO) simulation. Chapter 2 describes the intrinsic wearout mechanisms of the electron device, including physics processes, mechanisms and models of electromigration (EM), hot carrier degradation (HCD), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI). In Chapter 3, the modules and processes of FaRBS reliability simulation, model parameter extraction, and derating of voltage and temperature for reliability are described. Sensitivity analysis and spacecraft, planets, instrument, C-matrix, events (SPICE) simulation of the wearout models are also discussed. To account for the effect of wearout mechanisms on circuit functionality and reliability, the device-level accelerated lifetime models are extended to microelectronic circuit-level applications and an analog-to-digital converter reliability simulation using the FaRBS application is provided. Lifetime and failure equivalent circuit models for HCI, TDDB, and NBTI are presented in Chapter 4, Microelectronic Circuit Reliability Analysis and MaCRO. This chapter includes an illustrative case study for the purpose of demonstrating how to apply MaCRO models and algorithms to circuit reliability simulation, analysis, and improvement. The most common circuit structures used in reliability simulations are the ring oscillator, the differential amplifier, and the Static Random Access Memory (SRAM). The SRAM is selected as a case study vehicle to show the applicability of MaCRO models and algorithms in circuit reliability simulation and analysis. Chapter 5, in conclusion, describes the microelectronic system aspect of reliability, including impact to the system of individual failure mechanism lifetime models, voltage and temperature acceleration, and qualification based on failure mechanism and application. A failure-

mechanism-based qualification methodology using specifically designed stress conditions over traditional approaches (i.e., one voltage and one temperature) can lead to improved reliability predictions for targeted applications and optimized burn-in, screening, and qualification test plans.

1 INTRODUCTION

1.1 Organization

Microelectronics integration density is limited by the reliability of the manufactured product at a desired circuit density. Design rules, operating voltage, and maximum switching speeds are chosen to ensure functional operation over the intended lifetime of the product. To determine the ultimate performance for a given set of design constraints, reliability must be modeled for its specific operating condition.

Reliability modeling for the purpose of lifetime prediction is, therefore, the ultimate task of a failure physics evaluation. Unfortunately, existing industrial approaches to reliability evaluation fall short of predicting failure rates or wearout lifetime of semiconductor products. This is mainly attributed to the lack of a unified approach for predicting device failure rates, and the fact that all commercial reliability evaluation methods rely on the acceleration of a single, dominant failure mechanism.

Over the last several decades, knowledge of the root cause and physical behavior of the critical failure mechanisms in microelectronic devices has grown significantly. Confidence in historical reliability models has led to more aggressive design rules that have been successfully applied to the latest Very Large Scale Integration (VLSI) technology. One result of improved reliability modeling has been accelerated performance; that is, performance beyond the expectation of Moore's Law. A consequence of more aggressive design rules has been a reduction in the significance of a single-failure mechanism. Hence, in modern devices, there is no single-failure mode that is more likely to occur than any other within a range of specified operating conditions. This is practically guaranteed by the integration of modern simulation tools in the design process. The consequence of more advanced reliability modeling tools is a new awareness that device failures result from a combination of several competing failure mechanisms.

1.2 Reliability Prediction from a Historical Perspective

Reliability modeling and prediction is a relatively new discipline. Only since World War II has reliability become a subject of study due to the relatively complex electronic equipment used during the war and the high failure rates observed.

Since then, there have been two different approaches for reliability modeling corresponding to different time periods. Until the 1980s, the exponential, or constant failure rate (CFR), model [1] had been the only model used for describing the useful life of electronic components. It was common to the six reliability prediction procedures that were reviewed by Bowles [2] and was the foundation of the military handbook for reliability prediction of electronic equipments (known as the Military-Handbook-217 [MIL-HDBK-217] [3] series) that became the de facto industry standard for reliability prediction. Although the CFR model was used without physical justification, it is not difficult to reconstruct the rationale for the use of the CFR model, which mathematically describes the failure distribution of systems wherein the failures are due to completely random or chance events. Throughout that period, electronic equipment complexity began to increase significantly. Similarly, the earlier devices were fragile and had several intrinsic failure mechanisms that combined to result in a constant failure rate.

During the 1980s and early 1990s, with the introduction of integrated circuits (ICs), more and more evidence was gathered suggesting that the CFR model was no longer applicable. Phenomena such as infant mortality and device wearout dominated failures; these failures could not be described using the CFR model. In 1991, two research groups, IIT Research Institute/Honeywell SSED and the Westinghouse/University of Maryland teams, both recommended that, on the basis of their research and findings, the CFR model should not be categorically applied [4] to further updates of MIL-HDBK-217. They further recommended that the exponential distribution should not be applied to every type of component and system without due awareness.

The end of the CFR as a sole model for reliability modeling was officially set with the publication of the “Perry Memo.” Responding to increasing criticism of CFR, Secretary of Defense William Perry issued a memorandum in 1994 that effectively eliminated the use of most defense

standards, including the MIL-HDBK-217 series. Many defense standards were cancelled at that time and, in their place, the Department of Defense (DoD) encouraged the use of industry standards, such as the ISO 9000 series for quality assurance.

Since then, the physics-of-failure approach has dominated reliability modeling. In this approach, the root cause of an individual failure mechanism is studied and corrected to achieve some determined lifetime. Since wearout mechanisms are better understood, the goal of reliability engineers has been to design dominant mechanisms out of the useful life of the components by applying strict rules for every design feature. The theoretical result of this approach is, of course, that the expected wearout failures are unlikely to occur during the normal service life of microelectronic devices. Nonetheless, failures do occur in the field and reliability prediction has had to accommodate this new theoretical approach to the virtual elimination of any one failure mechanism limiting the useful life of an electronic device.

1.2.1 Traditional Approach

MIL-HDBK-217

The first brick of all traditional (empirical) reliabililty improvement methodologies was laid with MIL-HDBK-217. It was published in 1965 to achieve the following goals:

- To organize the reliability-data collected from the field.
- To find the basis for better designs.
- To give the “quantitative reliability requirements.”
- To estimate the reliability before full-scale production [5].

MIL-HDBK-217 soon became a standard; it was subsequently updated several times to keep pace with technology advancement as well as the changes in prediction procedures. Meanwhile, other organizations started to develop their own prediction models suitable for their own industries.

In the 1990s, attempts were focused on finding an electronic system reliability assessment methodology, including causes of failures, that could be used in the design and manufacturing of electronic systems. To cover the vast range of electronic devices, the notion of a "similar-system" was invented. The term "similar-system" refers to a system that uses similar technology and is built for similar application, or performs a similar function. The next step was to find whether the "similar-system" was used for existing field data. The data from a predecessor system could be used to generate the prediction of a new "similar-system" to the extent that the new generation was evolutionary (not revolutionary). The key process was the translation of the almost-old data to the new similar-system by considering the differences reflected in complexity and temperature, as well as the environmental and learning factors [5].

The last version of MIL-HDBK-217 (MIL-HDBK-217F) covers a wide range of major electronic component categories used in modern military systems, from microcircuits and discrete semiconductors to passive components such as resistors and capacitors [6]; for each of these areas, the handbook presents a straightforward equation for calculating the failure rate in failures per million hours. According to its claim, the goal of the handbook is to "establish and maintain consistent and uniform methods for estimating the inherent reliability of the mature designs of military equipment and systems" [3].

It is possible to classify the concepts behind the traditional MIL-HDBK-217F prediction procedures as:

1. The constant-failure-rate: The constant-failure-rate reliability model is used by most of the empirical-electronic reliability prediction approaches. The failure rate of the system containing different components is the summation of its components, which means that all system components are in series.
2. π factors: Almost all of the traditional prediction methods have a base failure rate modified by several π factors. Microcircuits, gate/logic arrays, and microprocessors incorporate stress models as a combination of package and parts. Examples of π factors include π_{CF} (Configuration Factor), π_E (Environmental Factor), and π_Q (Quality Factor). These

multiplication factors are included in the total failure rate calculation, Equation (1.1); are defined in MIL-HDBK-217F; and are based on different configuration levels, environmental stress levels, and quality levels for the part.

3. Two basic methods for performing reliability prediction based on the data observation include the parts count and the parts stress analysis. The parts count reliability prediction method is used for the early design phases, when not enough data is available but the numbers of component parts are known. The information for parts count method includes generic part types, part quantity, part quality levels (when known or can be assumed), and environmental factors. The general expression for item failure rate with this method is:

$$\lambda_S = \sum_{i=1}^{n} N_i(\lambda_g \pi_Q)_i \tag{1.1}$$

where λ_S is the total failure rate, λ_g is the failure rate of the i^{th} generic part, π_Q is the quality factor of the i^{th} part, N_i is the quantity of the i^{th} generic part, and n is the number of the generic part categories. If the parts operating in the equipment are operating in more than one environment, the above equation is applied to each portion of the equipment in a distinct environment. The overall equipment failure rate is obtained by summing the failure rates for each environment.

The part stress model is based on the effect of mechanical, electrical, and environmental stress and duty cycles, such as temperature, humidity, vibration, etc., on the part failure rate. The part failure rate varies with applied stress and the strength-stress interaction determines the part failure rate [7]. This method is used when most of the design is complete and the detailed part stress information is available. It is applicable during later design phases as well. Since more information is available at this stage, the result is more accurate than the part count method. An example of the microelectronic circuit part stress is:

$$\lambda_p = (C_1 \pi_T + C_2 \pi_E)\pi_Q \pi_L \tag{1.2}$$

where λ_p is the part failure rate and C_1, C_2 are the complexity of the die base failure rate (such as the number of gates) and the complexity of the package type (such as pin count), respectively; π_T is the temperature acceleration factor for the related failure mechanism; π_E is the environmental factor; and π_L is the learning factor, which considers the maturity of the device manufacturing line [6].

Component quality affects the part failure rate and is based on the component quality level, which is determined by the tests and screening in the manufacturing process. Since there are different technologies in this regard, there are several types of quality levels.

The environmental π factor defines the sensitivity of environmental stress on the device. Different prediction methods have their own list of environmental factors suitable for their device conditions. For instance, the environmental π factors defined in MIL-HDBK-217F cover almost all of the environmental stresses suitable for military electronic devices (the exception is ionizing radiation).

The learning factor shows the maturity of the device and suggests that the first versions are less reliable than subsequent generations. For instance, the learning π factor in the military handbook tries to take into consideration the effect of the number of years that the product has been in production. Therefore, the appropriate acceleration models are applied to the failure rates.

Table 1.1 gives the temperature acceleration factor used in some of the traditional prediction procedures [2].

Table 1.1. *Different procedural temperature acceleration factors.*

Procedural Method	**Temperature Acceleration Factor**
MIL-HDBK-217F	$\pi_T = 0.1 \exp[-A(1/T_j - 1/298)]$
HRD4	$\pi_T = 1$, for $T_j \leq 70℃$; $2.6 \cdot 10^4 \exp[-3500/T_j] + 1.8 \cdot 10^{13} \exp[-11600/T_j]$, for $T_j > 70℃$
NTT	$\pi_T = \exp[3480(1/339 - 1/T_j)] + \exp[8120(1/356 - 1/T_j)]$
CNET	$\pi_T = A_1 \exp[-3500/T_j] + A2 \exp[11600/T_j]$
Siemens	$\pi_T = A \exp[E_{a1} \cdot 11605(1/T_{j1} - 1/T_{j2})] + (1 - A) \exp[E_{a2} \cdot 11605(1/T_{j1} - 1/T_{j2})]$

The common acceleration models are:

Arrhenius Law of Temperature

$$AF_T = \exp[\frac{E_a}{k}(\frac{1}{T_1} - \frac{1}{T_2})] \tag{1.3}$$

where E_a is the activation energy, k is the Boltzmann's constant, and T_1 and T_2 are temperatures in Kelvin.

Kemeny Model for Voltage Acceleration

$$\lambda = \exp[C_0 - \frac{E_a}{kT_j}] \exp[C_1(\frac{V_{cb}}{V_{cbmax}})] \tag{1.4}$$

where V_{cb} is the collector-base voltage, V_{cbmax} is the maximum collector-base voltage before breakdown, and C_0 and C_1 are material-related constants.

Peck's Law for Temperature Humidity

$$AF = (\frac{M_{use}}{M_{test}})^{-n} \exp[\frac{E_a}{k}(\frac{1}{T_{use}} - \frac{1}{T_{test}})] \quad (1.5)$$

where M_{use} is the moisture level in service, M_{test} is the moisture level in test, and n is a material constant.

Coffin-Mason Based Law for Fatigue

$$AF = \frac{N_{use}}{N_{test}} = [\frac{\Delta T_{test}}{\Delta T_{use}}]^n \quad (1.6)$$

Examples of traditional reliability prediction approaches include the Telcordia, CNET, RDF, SAE, BT-HRD-5, Siemens, NTT, PRISM, and FIDES procedures. Table 1.2 provides common applications and procedural methods for those traditional approaches [8]:

Table 1.2. *Procedural methods and applications.*

Procedural Method	**Application**
MIL-HDBK-217	Military
Telcordia SR-332	Telecom
CNET	Ground Military
RDF-93 and 2000	Civil Equipment
SAE Reliability Prediction Method	Automotive
BT-HRD-5	Telecom
Siemens SN29500	Siemens products
NTT Procedure	Telecom
PRISM	Commercial and Military
FIDES	Aeronautical and Military
	Aeronautical and Military

Telcordia

The Telcordia (also called Bellcore) methodology from May 2001 was developed by Bell Communication Research (Telcordia Technologies Inc.) and focuses on equipment for the telecommunications industry. The main concepts in MIL-HDBK-217 and Telcordia SR-332 are similar, but Telcordia's SR-332 includes the ability to incorporate burn-in, field, and laboratory test data using a Bayesian analysis.

The basis of the Telcordia model for devices is referred to as the black box technique. This parts count method defines a black box steady-state failure rate, λ_{BB}, for different device types as:

$$\lambda_{BB} = \lambda_G \cdot \pi_Q \cdot \pi_S \cdot \pi_T \tag{1.7}$$

where λ_G is the generic steady-state failure rate for the particular device, π_Q is the quality factor, π_S is the electrical stress factor, and π_T is the temperature factor.

Parts count steady-state failure rate for units, λ_{PC}, is defined by:

$$\lambda_{PC} = \pi_E \sum_{i=1}^{n} N_i \lambda_{SSi} \tag{1.8}$$

where λ_{SSi} is the steady-state device failure rate of device i, π_E is the unit's environment factor, N_i is the quantity of device type i, and n is the number of device types in the unit.

The system failure rate, λ_{SYS}, is the sum of all failure rates of the units contained in a system:

$$\lambda_{SYS} = \sum_{j=1}^{M} \lambda_{PCj} \tag{1.9}$$

where λ_{PCi} is the failure rate of unit j and M is the number of units in system.

PRISM

PRISM was developed by the Reliability Analysis Center (RAC) under contract with the U.S. Air Force in the 1990s. The latest version of the method, which is available in a software version, was released in July 2001. RAC Rates is the name of the PRISM mathematical model for component failure rates; the component models are based on data derived from several sources. PRISM applies Bayesian methods with empirical data to obtain a system-level prediction. This methodology considers the failures of components as well as those related to the system. However, the component models are the heart of the analysis. The methodology provides different models for capacitors, diodes, integrated circuits, resistors, thyristors, transistors, and software. The total component failure rate is composed of:

1. Operating conditions.
2. Non-operating conditions.
3. Temperature cycling.
4. Solder joint.
5. Electrical overstress (EOS).

For components without a defined RAC Rates model, PRISM provides “Non-electronic Parts Reliability and Electronic Parts Reliability Data” books for reference. A multitude of part types can be found in these references with failure rates for various environments.

The general PRISM failure rate of a system, λ_{SYS}, is:

$$\lambda_{SYS} = (PG)\sum_{i=1}^{N}(\lambda_P)_i + \lambda_{SW} \qquad (1.10)$$

where PG is the process grade, λ_P is the RAC Rate failure rate of the i^{th} component, and λ_{SW} is the RAC Rate failure rate of software.

Unlike other handbook constant failure rate models, RAC Rates models do not have a separate factor for part quality level. Quality level is implicitly accounted for by a method known as process grading. Process grading addresses factors such as design, manufacturing, part procurement, and system management, which are intended to capture the extent to which measures have been taken to minimize the occurrence of system failures.

FIDES

The FIDES [9, 10] prediction method attempts to predict the constant failure rate experienced in the useful life portion of the classic bathtub curve. This approach models intrinsic failures together with extrinsic failures resulting from equipment specification, design, production, and integration, as well as selection of the procurement route. The methodology takes into account failures resulting from development and manufacturing and the over-stresses linked to the application, such as electrical, mechanical, and thermal. At the highest level, the FIDES method is comprised of three basic factors:

$$\lambda = \lambda_{Phy} \cdot \pi_{Partmanufacturing} \cdot \pi_{Process} \qquad (1.11)$$

where λ_{Phy} is the physical contribution, $\pi_{Partmanufacturing}$ is a factor representing quality and manufacturing technical control, and $\pi_{Process}$ covers all processes from specification to field operation and maintenance. λ_{Phy} is expressed as:

$$\lambda_{Phy} = \left[\sum_{physicalcontributions} (\lambda_0 \cdot \pi_{acceleration}) \right] \cdot \pi_{induced} \qquad (1.12)$$

where λ_0 is the basic failure rate that depends on the technological characteristics, $\pi_{acceleration}$ is an environmental acceleration factor vs. use conditions, and $\pi_{induced}$ is the overstress factor.

Limitations

MIL-HDBK-217, as the origin for almost all traditional reliability approaches, has limitations. MIL-HDBK-217 has not been updated since 1995, and most ICs have not been updated since 1991. Therefore, more recent technologies are not included or defined. Table 1.3 shows a comparison and some of the limitations of MIL-HDBK-217, compared to the physics of failure approach [6].

Despite a variety of empirical prediction models, the majority of engineers still use MIL-HDBK-217. A Crane survey shows that almost 80 percent of the respondents use the Military handbook, while PRISM and Telcordia are second and third. Inconsistency among different traditional prediction methods is the main problem facing engineers.

Table 1.3. *A Comparison between MIL-HDBK-217 and Physics–of-Failure Approach.*

Issue	Mil-Hdbk-217	Physics of Failure
Model Development	Models can't provide accurate design or manufacturing guidance since they were developed from assumed constant failure-rate data, not root-cause, time-to-failure data. A proponent representative's quote is germane: "Therefore, because of the fragmented nature of the data and the fact that it is often necessary to interpolate or extrapolate from available data when developing new models, no statistical confidence intervals should be associated with the overall model results" [12].	Models based on science/engineering first principles. Models can support deterministic or probabilistic applications.
Device Design Modeling	The Mil-Hdbk-217 assumption of perfect designs is not substantiated due to lack of root-cause analysis of field failures. Mil-Hdbk-217 models do not identify wearout issues.	Models for root-cause failure mechanisms allow explicit of the impact that design, manufacturing, and operation have on reliability.
Device Defect Modeling	Models can't be used to 1) consider explicitly the impact of manufacturing variation on reliability, or 2) determine what constitutes a *defect*, or how to screen/inspect *defects*.	Failure mechanism models can be used to 1) relate manufacturing variation to reliability, and 2) determine what constitutes a *defect* and how to screen/inspect.
Device Screening	Mil-Hdbk-217 promotes and encourages screening without recognition of potential failure mechanisms.	Provides a scientific basis for determining the effectiveness of particular screens or inspections.
Device Coverage	Doesn't cover new devices for approximately the first 5–8 years. Some devices, such as connectors, weren't updated for more than 20 years. Developing and maintaining current design reliability models for devices is an impossible task.	Generally applicable—applies to both existing and new devices—since failure mechanisms are modeled, not devices. Thirty years of reliability physics research has produced and continues to produce peer-reviewed models for the key failure mechanisms applicable to electronic equipment. Automated computer tools exist for printed wiring boards and microelectronic devices.
Use of Arrhenius Model	Indicates to designers that steady-rate temperature is the primary stress designers can reduce to improve reliability. Mil-Hdbk-217 models will not accept explicit temperature change inputs. Mil-Hdbk-217 lumps different acceleration models from various failure mechanisms together, which is unsound.	The Arrhenius model is used to model the relationships between steady-state temperature and mean time-to-failure for each failure mechanism, as applicable. In addition, stresses due to temperature change, temperature rate of change, and spatial temperature gradients are considered, as applicable.
Operating Temperature	Explicitly considers only steady-state temperature. Effect of steady-state temperature is inaccurate because it is not based on root-cause, time-to-failure data.	The appropriate temperature dependence of each failure mechanism is explicitly considered. Reliability is frequently more sensitive to temperature cycling, provided adequate margins are given against temperature extremes [13].
Operational Temperature Cycling	Does not support explicit consideration of the impact of temperature cycling on reliability. No way of superposing the effects of temperature cycling and vibration.	Explicitly considers all stresses, including steady-state temperature, temperature change, temperature rate of change, and spatial temperature gradients, as applicable to each root-cause failure mechanism.
Input Data Required	Does not model critical failure contributors, such as materials architectures, and realistic operating stresses. Minimal data in, minimal data out.	Information on materials, architectures, and operating stresses—the things that contribute to failures. This information is accessible from the design and manufacturing processes of leading electronics companies.
Output Data	A proponent representative's quote offers some illumination: "Mil-Hdbk-217 is not intended to predict field reliability and, in general, does not do a very good job in an absolute sense" [12].	Provides insight to designers on the impact of materials, architectures, loading, and associated variation. Predicts the time-to-failure and failure sites for key failure mechanisms in a device or assembly. These failure times and sites can be ranked. This approach supports either deterministic or probabilistic treatment.
DoD/Industry Acceptance	Mandated by government, 30-year record of discontent. Not part of the US Air Force Avionics Integrity Program (AVIP). No longer supported by senior US Army leaders.	Represents the best practices of industry.
Coordination	Models have never been submitted to appropriate engineering societies and technical journals for formal peer review. Future tri-service coordination at issue.	Models for root-cause failure mechanisms undergo continuous peer review by leading experts. New software and documentation currently being coordinated with leading electronics companies worldwide, US Army, and AVIP to start.
Relative Cost of Analysis	Cost is high compared with value added. Can misguide efforts to design reliable electronic equipment.	Intent is to focus on root-cause failure mechanisms and sites, which is central to good design and manufacturing. Acquisition flexible, so costs are flexible. The approach can result in reduced life-cycle costs due to higher initial and final reliabilities, reduced probability of failing tests, reductions in *hidden factory*, and reduced support costs.

1.2.2 Physics-of-Failure Approach

Attempts, which began during the 1970s, to include physics-of-failure into military handbooks were not very successful. Although the need for a physics-of-failure methodology was realized in the 1970s, a physics-of-failure-like model for small-scale CMOS technology was not introduced until 1989. Even so, this approach, as an independent methodology, only started to attract attention during the 1990s in the form of recommendations to update the military handbook. The recommendations addressed the weaknesses of traditional approaches: (1) the misleading use of constant physics-of-failure, (2) the use of the Arrhenius temperature model, (3) the modeling of wearout mechanisms, and (4) modeling mechanisms such as brittle die fracture.

The physics-of-failure methodology can be summarized as follows:

- Identify potential failure mechanisms, e.g., chemical, electrical, physical, mechanical, structural, or thermal processes leading to failure, and the failure sites on each device.
- Expose the product to highly accelerated stresses to find the dominant root-cause of failure.
- Identify the dominant failure mechanism as the weakest link.
- Model the dominant mechanism (what and why the failure takes place).
- Combine the data gathered from acceleration tests and statistical distributions, e.g., Weibull distribution, Lognormal distribution.
- Develop an equation for the dominant failure mechanism at the site and its mean time-to-failure (MTTF).

Physics-of-failure modeling and simulation tools are the key elements in this approach. There are two computer-based modeling and simulation tools: Computer-Aided Design of Microelectronic Packages (CADMP-2) and Computer-Aided Life-Cycle Engineering (CALCE). The CADMP-2 assesses the reliability of electronics at the package level; CALCE assesses the reliability of electronics at the printed wiring board level. Together, these two models provide a framework to support a physics-of-failure approach for reliability in electronic systems design.

The CADMP-2 is a set of integrated software programs that can be used to design and assess the reliability of integrated circuit, hybrid and multi-chip module packages. Figure 1.1 shows the input and output of this software.

The CALCE software provides an environment for incorporating various tools associated with reliability, supportability, producibility, and costing tasks into the design of electronic systems in the earliest stages of the design process. Figure 1.2 shows the inputs and outputs of this software [24]. The main advantage of the physics-of-failure methodology is that contributing failure causes are based on scientific knowledge; that knowledge provides the scientific basis for reliability prediction, incorporating relevant information on materials, architectures, and operating stresses. Moreover, since accelerated stress tests are one of the main methods for finding the degradation model parameters, the test results could help provide the necessary test criteria for the product as well.

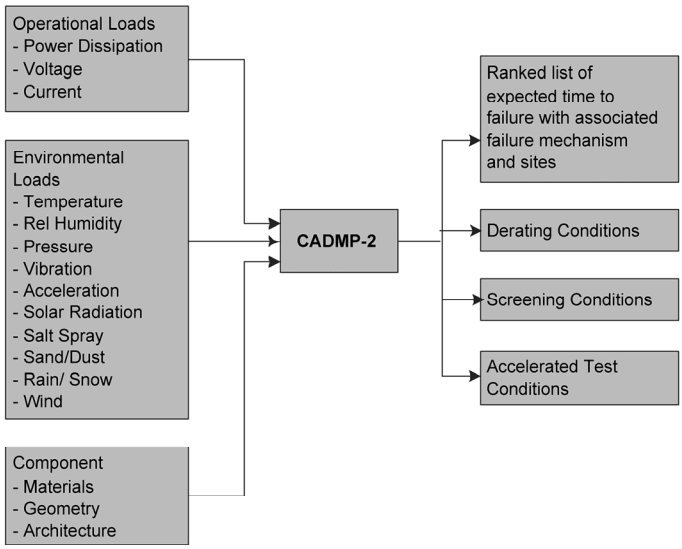

Fig. 1.1. *CADMP-2 inputs and outputs.*

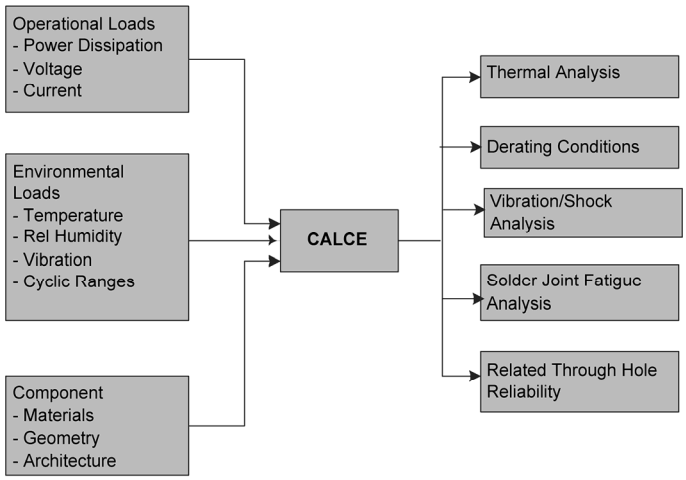

Fig. 1.2. *CALCE inputs and outputs.*

The objectives of the physics-of-failure methodology are "to develop a physics-of-failure handbook for reliability assurance containing a methodology for assessing system reliability on the basis of environmental and operating stresses, the materials used, and the packaging selected," and "development of mixture models which consider both early and premature wearout failures caused by the displacement of the mean and variability due to manufacturing, assembly, handling, and misapplication" [4]. However, there are some serious challenges in implementing this approach, Ref. Table 1.3.

1.2.3 Recent Approach: RAMP

In 2003, IBM published the Reliability Aware Micro-Processor (RAMP) model for processor reliability. RAMP models chip mean-time-to-failure as a "function of the failure rates of individual structures on the chip due to different failure mechanisms, and can be used to evaluate the reliability implications of different applications, architectural features and processor designs" [25]. In addition, it is claimed that the above-mentioned model is a self-standing module that could be attached to simulators to make power and temperature predictions [13].

In IBM's report, processor errors are classified into two main categories: soft and hard. Hard errors are subclassified into intrinsic and extrinsic failures. RAMP only models intrinsic processor failures because long-term processor reliability is dominated by wearout or intrinsic failures. However, IBM believes that RAMP can be extended to model soft errors.

Although main wearout intrinsic failure mechanisms affecting processors are those related to electromigration, stress migration, time-dependent dielectric breakdown, temperature cycling and hot carrier injection, negative-bias temperature inversion and corrosion, RAMP only considers the first four; however, it could cover the others as well. It should be noted that RAMP uses the Arrhenius model to show the dependence of processor failures on temperature; due to the direct processor reliability relation to the operating temperature, it is expected that many reliability problems are the result of elevated processor temperature.

RAMP uses MTTF to measure reliability. To calculate MTTF, RAMP assumes all failure mechanisms have constant failure rates. This assumption is clearly inaccurate; however, it allows RAMP to combine different failure mechanisms and provides a unified MTTF. Assuming a constant failure rate, the MTTF is calculated as the inverse of the failure rate. The system reliability model used in RAMP is the sum-of-failure-rates model. RAMP treats each structure on a chip as a separate component that can fail in different ways corresponding to various failure mechanisms. The "competing risk model" determines the dominant component failure mechanism, and the "series model" estimates the system failure rate (based on the failure rate of each component).

To calculate the failure rate of a component using the competing risk model, RAMP makes the following assumptions:

- Each failure mechanism proceeds independently of every other, at least until a failure occurs.
- The component fails when the first of all competing failure mechanisms reaches a failure state.
- Each of the failure mechanisms has a known life distribution model.

If there are k failure mechanisms and the failure rate of the component due to the i^{th} failure mechanism is $\lambda_i(t)$, then the failure rate of the component due to all failure mechanisms, $\lambda_C(t)$ is given by:

$$\lambda_C(t) = \sum_{i=1}^{k} \lambda_i(t) \tag{1.13}$$

In the case where a component has a constant failure rate λ_C, the MTTF for that component is given by:

$$MTTF_C = \frac{1}{\lambda_C} = \frac{1}{\sum_{i=1}^{k} \lambda_i} \tag{1.14}$$

where λ_i is the failure rate due to each failure mechanism.

The "series model" is also applied to estimate the systems reliability based on the components. By applying the same set of assumptions used for the "competing risk model," i.e., a system consisting of j components fails when the first component fails, the MTTF of the system is given by:

$$MTTF_{SYS} = \frac{1}{\sum_{i=1}^{j} MTTF_i} = \frac{1}{\sum_{i=1}^{j} \lambda_i} = \frac{1}{\sum_{i=1}^{j} \sum_{l=1}^{k} \lambda_{il}} \tag{1.15}$$

where λ_i is the failure rate of the i^{th} component and λ_{il} is the failure rate of the i^{th} component due to the l^{th} failure mechanism.

1.3 Reliability Modeling and Prediction Today

Reliability device simulators have become an integral part of the design process. These simulators successfully model the most significant physical failure mechanisms in modern

electronic devices, such as time-dependent dielectric breakdown (TDDB), negative bias temperature instability (NBTI), electromigration (EM), and hot carrier injection (HCI). These mechanisms are modeled throughout the circuit design process so that the system will operate for a minimum expected useful life.

Modern chips are composed of tens or hundreds of millions of transistors. Hence, chip-level reliability prediction methods are mostly statistical. Today, chip-level reliability prediction tools model the failure probability of the chips at the end of life, when the known wearout mechanisms are expected to dominate. However, modern prediction tools do not predict the random, post burn-in failure rate that can be seen in the field.

Chip and packaged system reliability is measured as rate of failure in time (FIT). The FIT is defined as one failure per billion part hours. The semiconductor industry provides an expected FIT rate for every product that is sold based on operation within the specified conditions of voltage, frequency, heat dissipation, etc. Hence, a system reliability model is a prediction of the expected mean time between failures (MTBF) for an entire system as the reciprocal of the sum of the FIT rates for every component.

The failure rate of a component can be defined in terms of an acceleration factor, *AF*, as:

$$\lambda = \frac{Number\,of\,failures}{Number\,of\,tested \times hours \times AF} \times 10^9 FIT \qquad (1.16)$$

where “Number of failures” and “Number of tested” are the number of actual failures that occurred as a fraction of the total number of units subjected to an accelerated test. The acceleration factor, AF, is generally established by the manufacturer for a given technology and product, as they know the impact of different failure mechanisms on their designs accelerated in the High Temperature Operating Life (HTOL); this information is generally based on a company proprietary variant of the MIL-HDBK-217 approach for accelerated life testing. The true task of reliability modeling,

therefore, is to choose an appropriate value for AF based on the physics of the dominant failure mechanisms that would occur in the field for the device.

The HTOL qualification test is usually performed as the final qualification step of a semiconductor manufacturing process. The test consists of stressing some number of parts, usually about 100, for an extended time, usually 1000 hours, at an accelerated voltage and temperature. Two features shed doubt on the accuracy of this procedure. One feature is lack of sufficient statistical data due to too few parts; the second is that manufacturers stress their parts under relatively low stress levels to guarantee zero failures during qualification testing. Unfortunately, with zero failures, little statistical data is acquired.

The accepted approach for measuring FIT would, in theory, be reasonably correct if there is only a single dominant failure mechanism that is excited equally by either voltage or temperature. For example, EM is known to follow Black's equation (described later) and is accelerated by increased stress current in a wire or by increased temperature of the device. If, however, multiple failure mechanisms are responsible for device failures, each failure mechanism should be modeled as an individual "element" in the system and the component survival is modeled as the survival probability of all the "elements" as a function of time.

If multiple failure mechanisms, instead of a single mechanism, are assumed to be time-independent and independent of each other, FIT (constant failure rate approximation) rates can be a reasonable measure of realistic field failures. Under the assumption of multiple failure mechanisms, each will be accelerated differently, depending on the physics that are responsible for each mechanism. If, however, an HTOL test is performed at an arbitrary voltage and temperature for acceleration based only on a single failure mechanism, then only that mechanism will be accelerated. If multiple failure mechanisms with different sensitivity functions to acceleration conditions exist, a choice of only one HTOL test point is biasing the results.

1.3.1 Competing Mechanisms Theory

Whereas failure-rate qualification has not improved over the years, the semiconductor industry's understanding of the reliability physics of semiconductor devices has advanced enormously. Every known intrinsic wearout failure mechanism is well understood and the processes are so tightly controlled that electronic components are designed to perform with reasonable life and with no single dominant failure mechanism. Standard HTOL tests generally reveal multiple intrinsic failure mechanisms during testing, which would suggest also that no single failure mechanism would dominate the FIT rate in the field. Therefore, in order to derive a more accurate model for FIT, a preferable approximation would be that all failures are equally likely and that the resulting overall failure distribution will resemble a constant failure rate process that is consistent with the military handbook, FIT rate approach.

The acceleration of a single failure mechanism is a highly non-linear function of temperature and/or voltage. The temperature acceleration factor (AF_T) and voltage acceleration factor (AF_V) can be calculated separately; this is the subject of most studies concerning reliability physics. The total acceleration factor of the different stress combinations will be the product of the acceleration factors of temperature and voltage.

This acceleration factor model is widely used as the industry standard for device qualification; however, it only approximates a single dielectric breakdown type of failure mechanism and does not correctly predict the acceleration of other mechanisms.

To be even approximately accurate, however, electronic devices should be considered to have several failure modes degrading simultaneously. Each mechanism "competes" with the others to cause an eventual failure. When more than one mechanism exists in a system, the relative acceleration of each one must be defined and averaged at the applied condition. Every potential failure mechanism should be identified and its unique *AF* should be calculated for each mechanism at a given temperature and voltage so that the FIT rate can be approximated for each mechanism separately, where each mechanism leads to an expected failure unit per mechanism, FIT_i. Unfortunately, again, individual failure mechanisms are not uniformly accelerated by a standard

HTOL test, and the manufacturer is forced to model a single acceleration factor that cannot be combined with the known physics-of-failure models.

The history of reliability prediction of microelectronic devices can be categorized into two distinct phases:

1. The first phase relied on traditional methods, or empirical models. These traditional methods are based on the data gathered either from laboratory tests or fielded applications; the statistical curve fitting of the component failure data provides the required mathematical model. Depending on the specific sources and environment used to collect the data, the models provide predictions for the relevant area. Due to data diversities, the predictions are different as well. Almost all of the procedures are based on the data gathered from the field with extrapolation from devices that are similar to each other. These procedures model past experience and data to estimate the reliability of similar or modified products and to deal with the early defects and random events.

2. The second phase, physics-of-failure, is an approach that uses the "knowledge of the root-cause failure mechanism" [14]. This approach tries to bring the prediction to increased scientific accuracy. Physics-of-failure focuses on device end-of-life failure mechanisms. Unlike traditional methodologies, this approach studies the impact of different parameters on single-device wearout mode.

Combining these two methodologies can develop into a powerful framework for predicting microelectronic device reliability.

1.3.2 FaRBS

FaRBS (Failure-Rate-Based SPICE [spacecraft, planet, instrument, C-matrix, events]) [15] is a circuit-level simulation method that is based on the physics-of-failure and sum-of-failure-rates (SOFR) models. It combines the modules of SPICE (simulation program with integrated circuit emphasis), semiconductor wearout models, integrated circuit system reliability models, accelerated

factor models, and the SOFR reliability model. The FaRBS simulation methodology is presented in Chapter 3.

1.3.3 MaCRO

MaCRO (Maryland Circuit-Reliability Oriented) [16] is an integrated circuits emphasis (SPICE) simulation method that was developed based on the rate-of-failure concept and failure-equivalent circuit-modeling techniques. MaCRO consists of a series of accelerated lifetime models and failure-equivalent circuit models for common silicon intrinsic wearout mechanisms, including HCI, TDDB, and NBTI. The MaCRO simulation is a first-order approach that does not fully characterize the micro-cosmic interactions among the wearout mechanisms. This assumption simplifies the device-wearout modeling process and makes the MaCRO compatible with the standard simulation tools. MaCRO has promised a way for system designers and device reliability engineers to better prepare for the reliability challenges that will be present in future-generation technologies. The overall simulation flow of MaCRO is straightforward; the SPICE routine is only called for a very limited number of times to simulate the impact of the device wearout on circuit functionality. The MaCRO simulation methodology is presented in Chapter 4.

1.4 Summary

As has been shown above, the latest generations of electronic-system-reliability prediction (RAMP, FaRBS, and MaCRO) have formed a framework for electronics reliability research. The paradigm began with the so-called traditional reliability approaches. Operational and non-operational data are gathered and formulated in the first steps of scientific theorization. As was expected, the chief disadvantage of these methodologies is the lack of scientific reasoning in device/system failures. Although traditional methodology is based on the empirical data, which is the foundation of any scientific method, it does not describe the physics behind the failures. However, it nevertheless provides a tool to deal with electronic device reliability prediction. Despite the fact that traditional approaches may not provide an accurate prediction, subsequent modified versions could fulfill the primary requirements. The history of traditional reliability prediction is evidence of this claim.

One could say that the traditional approach based on empirical data is the first stage of reasoning in the absence of scientific explanation in any paradigm; the fact that it only models the gathered data without providing an explanation shows the inability of scientists to model the physics-of-failure satisfactorily. The history of science is filled with accounts of using mathematical models/tools for physical phenomena before fully understanding the physics behind them.

Physics-of-failure is an approach that tries to reveal and model the root cause processes of device failures. This branch of reliability combines knowledge about the device with the statistical aspects of failure occurences. The fact that physics-of-failure is not widely used by engineers shows that it was not successful in achieving its goals. It seems that the key element of this lack of success is the complexity of modeling the MTTF of devices based on the underlying root causes. Moreover, the physics of device failures has not yet been clearly formulated. Scientists are still working on formulating the reasons behind each failure. Therefore, applying complex statistical tools to vague scientific principles adds several parameters to the equations, leading to a higher level of complexity. In contrast, a scientific model should give a simple explanation for the instances and then generalize the model. Until now, the physics-of-failure approach was not able to make accurate predictions or replace traditional approaches.

The electronic system (circuit/processor) reliability approach is a method built upon the advantages of both traditional and physics-of-failure methodologies; this approach combines the device physics-of-failure mechanisms with the constant failure rate model and applies them to the electronic system, which provides both a physical explanation for the electronic system failures, and a simplified statistical tool for reliability prediction. However, this approach can still:

- Use traditional prediction tools in specific field studies to obtain an approximate numerosity.
- Update the previous models based on statistical methods (like the Bayesian approach) and try to calculate the uncertainty growth of the electronic systems.
- Unify electronic-device failure mechanisms.
- Try to apply the new scientific models to electronic systems.

2 ELECTRON DEVICE PHYSICS OF FAILURE

The major wearout mechanisms of semiconductor-based micro-electronic devices are electromigration (EM) gate-oxide breakdown, also known as time-dependent dielectric breakdown (TDDB), and hot carrier (HC) effects. The latter are usually divided into hot carrier injection (HCI) and negative bias temperature instability (NBTI). These mechanisms are briefly reviewed.

2.1 Electromigration

2.1.1 Introduction

Interconnects that are embedded in interlayer dielectric material are the wire connections to supply electrical signals to these devices. Aluminum (*Al*) has been used as the major on-chip interconnect material. It has evolved from a single layer of *Al* to multiple levels of sandwiched *Ti/Al-Cu/TiN* metal layers. In recent technology development, *Cu* and new dielectric materials have been adapted to gain better resistance-capacitance delay and reliability resistance. Due to continuing transistor scaling, interconnects are now a significant limiter and are as important as transistors in determining an integrated circuit's (IC) density, performance, and reliability. Aggressive interconnect scaling has resulted in increasing current densities and associated thermal effects, which can cause reliability problems.

EM, the dominating failure mode of interconnects, is characterized by the migration of metal atoms in a conductor through which large direct-current densities pass [17]. Although EM has been intensely studied for more than 40 years, many aspects of EM are still not well understood. This lack of understanding is caused by two related issues: the existence of many factors that influence EM and the inability to isolate the effect of these factors experimentally. These factors include grain structure, grain texture, interface structure, stresses, film composition, physics of void nucleation and growth, thermal and current density dependencies, etc. [18]. According to experimental research, current density and temperature are among the most important factors.

Black [19] developed an empirical model relating the median time (t_{50}) of a metal line to the temperature (T) and current density (J); the model has the form

$$t_{50} = \frac{A}{J^n} e^{\frac{E_a}{kT}} \tag{2.1}$$

where A is a material and process-dependent constant and E_a is the activation energy for the diffusion processes that dominate the temperature range of interest. The importance of current density and temperature is shown in this equation.

As expected, the scaling of interconnects will increase current densities and temperature, thereby greatly reducing the median time. The reliability of the IC will decrease simultaneously. To better understand the interconnect-scaling effect, physical models and statistical models must be carefully developed. Section 2 focuses on the physical process, statistical models, and acceleration factors.

2.1.2 Basic Physics Process of EM

A significant amount of research has been done on the physics of EM; a detailed review can be found in [20]. Figure 2.1 summarizes the EM failure process. As IC technology increases device density, the interconnects that carry signals are consequently reduced in size, specifically, in height and cross section. This leads to extremely high current densities, on the order of at least 106 A/cm$^{2.}$ At these current densities, momentum transfer between electrons and metal atoms becomes important. The transfer, which is called the electron-wind force, results in a mass transport along the direction of electron movement. Once the metal atoms are activated by the electron wind, they are subject to the electric fields that drive the current. Since the metal atoms are positively ionized, the electric field moves them against the electron wind once they have been activated. The interplay of these two phenomena determines the direction of net mass transfer. This mass transfer manifests itself in the movement of vacancies and interstitials. The vacancies coalesce into voids or microcracks, and interstitials become hillocks. The voids, in turn, decrease the cross-sectional area of the circuit metallization and increase the local resistance and current density at that point in the

metallization. Both the increase in local current density and in temperature increase EM effects. This positive feedback cycle can eventually lead to thermal runaway and catastrophic failure.

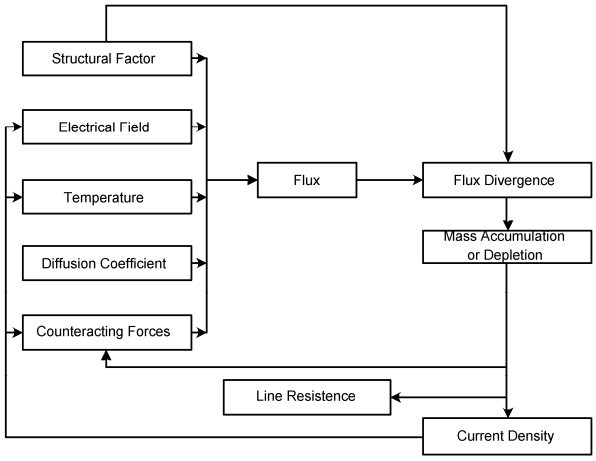

Fig. 2.1. *Summary of EM failure process [21].*

According to the quantum theory of electron transport in a metal, the Bloch waves representing the moving electrons can transfer energy to the lattice only by interacting with defects or through phonon–electron interactions. Thus, the microstructure of the interconnect will play a significant role in the EM process. Grain boundaries become important as potential defect sites and as possible transport channels. In the electron-defect interaction process, an electron is scattered by an ion in a defect site. The electron's momentum is reversed, causing an average change in momentum in the transport direction equal to 2 *mv* where *m* is the electron mass and *v* is the mean velocity of the electron in the direction of the current flow. The friction force on the ion is:

$$F_f = \frac{2m\overline{v}}{\tau_{col}} \tag{2.2}$$

where τ_{col} is a time constant representing the time between collisions. From elementary transport theory, the electron current density J_e is:

$$J_e = ne\overline{v} \tag{2.3}$$

where n is the density of electrons available for transport. From Equation (2.3) we can get the v, and the friction force can be expressed as:

$$F_f = \frac{2mJ_e}{ne\tau_{col}} \tag{2.4}$$

Assuming the ion transport is simply proportional to the force applied:

$$v_i = \mu F_f \tag{2.5}$$

and

$$J_i^f = eN\mu F_f \tag{2.6}$$

where v_i is the mean ion velocity in the direction of transport, μ is the ion mobility, J^f_i is the ion current density due to electron momentum transfer, and N is the density of ions available for transport. According to the Nernst-Einstein equation, the ion diffusion coefficient, D, and μ are related as:

$$\frac{D}{\mu} = \frac{kT}{e} \tag{2.7}$$

where k is the Boltzmann constant.

Combining Equations (2.5) through (2.7) yields:

$$J_i^f = \frac{e^2DN}{kT}F_f \tag{2.8}$$

From Equation (2.4), we see that F_f is proportional to J_e:

$$F_f = C_1 J_e \tag{2.9}$$

where C_1 is the proportionality constant derived from Equation (2.4). Thus, we get:

$$J_i^f = \frac{e^2 DNC_1}{kT} J_e \tag{2.10}$$

From Ohm's law, this equation becomes:

$$J_i^f = \frac{e^2 DNC_1}{kT} \rho_e E \tag{2.11}$$

where ρ_e is the electron resistivity and E is the electric field.

The electric field will also induce an ion current, J^E_i, that is counter to the friction current. Using the basic transport relation and the Nernst-Einstein relationship to describe the field-induced ion current J^E_i:

$$J_i^E = \frac{e^2 NDE}{kT} \tag{2.12}$$

The total ion current, J_i, is given by:

$$J_i = J_i^f - J_i^E = (eN)(C_1 \rho_e - 1)(\frac{eD}{kT})E \tag{2.13}$$

To simplify this equation, we use Z' to denote ($C_1\rho_e - 1$), then:

$$J_i = J_i^f - J_i^E = (eN)Z'(\frac{eD}{kT})E \tag{2.14}$$

The simple interpretation of this equation is that the ion current is equal to the effective charge on the ion, multiplied by the density of ions available for transport, the ion mobility, and the electric field.

There are other physical effects that might give rise to net ion currents and to the ion current divergence necessary for void formation. The temperature gradients occurring in the interconnect will also create the ion flux divergences responsible for open-metal device failures. The reason is that the ion diffusion coefficients will become position dependent. Mobilities will be greater in the hotter region and less in the cooler region. The ion will move from the hotter regions to the cooler regions to form a hillock.

The stress in the conducting strip also affects the EM. Just as an electron field causes ion drift, a gradient of stress σ acts as a generalized force to induce ion motion. Ions preferentially migrate from compressively (σ more negative) stressed regions and accumulate at locations stressed in tension (σ more positive), while vacancies diffuse the other way. The resulting stress gradient causes a backflow of matter; this effect plays a significant role in short conductors. Thermodynamics argues that the free energy per atom in a stress field depends on the stress and is equal to $\varphi = -\sigma\Omega$, where Ω is the atomic volume. The force per atom in a stress field is the negative gradient of the free energy:

$$F = \omega\frac{d\sigma}{dx} \tag{2.15}$$

Furthermore, compressive and tensile stresses might increase or decrease ion migration activation energy, changing the diffusion coefficient D.

A more complete equation that accounts for these effects is:

$$J_i = eZ' N(\frac{eD}{kT})E - eD\nabla N - \frac{eDN\omega\nabla N}{\beta N_0 kT} \qquad (2.16)$$

where ω is the atomic volume, β is the film compressibility, and N_0 is the atom density of the film. The first term accounts for friction flow, the second accounts for the concentration gradient, and the third accounts for film stress.

Material Structure Inhomogeneities

Thus far, the powered interconnect is assumed to have a homogeneous structure. The electrically induced diffusion of metal atoms alone is not sufficient to cause EM. The growth of voids and hillocks requires the saturation of vacancies and the supersaturation of interstitials [21], which requires not only a diffusion of metal atoms but also a divergence in the diffusion flux.

Microscopic examination of most deposited thin films indicates a pronounced cellular structure to the film generally referred to as the "grain structure" of the film. These cells arise as a result of the processes of nucleation and growth that form the film. The grain structure depends on the deposition conditions and has a profound effect on EM damage. For example, powered single-crystal *Al* strips have been shown to exhibit virtually "infinite" life.

Grain density is determined by surface conditions and film growth parameters, such as substrate temperature, and the rate of arrival of metal atoms to the growth surface. Grain boundaries represent interfaces with associated free energy of surface formation. During growth and subsequent annealing cycles, some grains might grow and others disappear in order to minimize the free energy. The grain boundaries represent relatively low-resistance ion conducting channels. At standard IC operating temperatures, bulk ion migration processes are slow and the grain boundaries carry the bulk of the ion current [22]. The grain structures and boundaries enable

considerable refinement of material models of EM. Specifically, there are three properties that have immediate impact on reliability models. They are:

- The orientation of the boundary with respect to the electric field.
- The angles of the grain boundaries with respect to each other.
- Changes in the number of the grains per unit area–grain density.

Each of these properties can give rise to the ion divergences necessary to create voids in metal strips. The effects of these properties will be discussed below.

With increasingly shrinking interconnect stripe widths, broad area or blanket metallization leave greater numbers of grain boundary "triple points," and grains line up in a bamboo structure after patterning. Figure 2.2 shows the confluence of three grain boundaries at a triple point. If the boundary to the left is parallel to the applied field, the angle θ_1 equals 0, and the apparent ion mobility is highest along that boundary. Migration along the two adjacent boundaries is the result of a projected field component and is lower. Under this condition, it is apparent that fewer ions leave the triple point than enter it, and a mass accumulation is favored; otherwise, voids form.

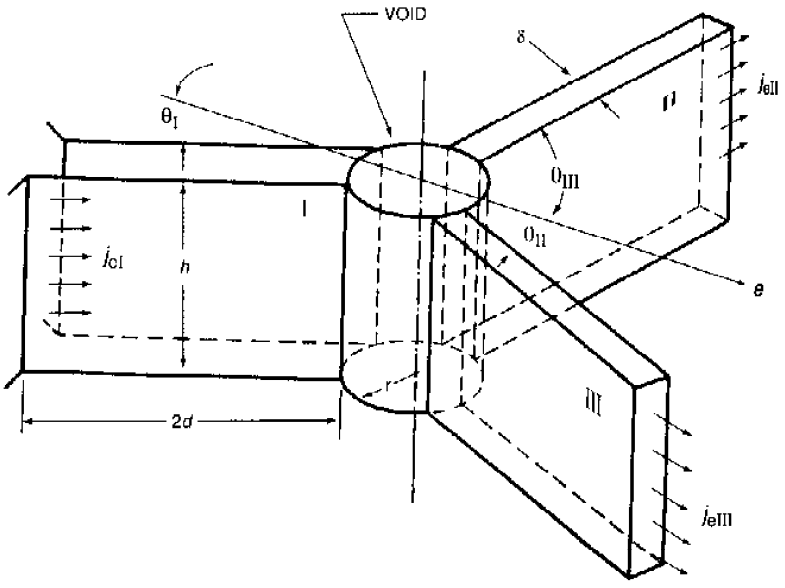

Fig. 2.2. *Grain boundary "triple point" structure [23].*

The grains themselves can be viewed as composed of linear arrays of dislocations characterized by some Burger vector, b. The most common assumption relating grain boundary diffusion to tilt is that the ion mobility is simply proportional to the number of dislocations in the boundary. Thus:

$$\mu_i = \frac{Ab}{2d} \tag{2.17}$$

where A is a proportionality constant, and d is the separation distance of dislocation.

If the grain size changes along the strip, the density of ion conduits into and out of a region must also change. A densely grained region will channel ions out more effectively than a sparsely grained region. This creates the ion current divergence necessary to form a void. Similarly, ion pile-up can occur in regions wherein the grain sizes increase in the direction of electron flow. Another factor that affects the ion conduction is the texture of the oriented crystallite in the metal films.

To summarize, the important factors of EM include current density, electrical field, temperature, grain structure, and boundary stress. To model EM accurately, all of these factors need to be considered. Consideration of all of the factors causes increased complexity of the model, making this approach nearly impossible. In this situation, statistical models based on empirical data will help us understand the EM mechanisms and realize reliability design goals.

2.1.3 Statistical Models of EM

Lognormal Distribution

Knowledge of the correct failure distribution is critical in predicting IC reliability. Traditionally, lognormal failure distribution has been used to characterize EM failures. Assuming that the time-to-failure, t, is a random variable, the lognormal probability density function $f(t)$ is

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} exp[-\frac{1}{2}(\frac{ln(t) - ln(t_{50})}{\sigma})^2] \tag{2.18}$$

where t_{50} is the median time-to-failure and σ is the lognormal standard deviation. The value of t_{50} can be estimated by using Black's equation. The lognormal standard deviation, σ, which is related to the ratio of the linewidth to the grain size [24] and current density [25], ranges from 0.28 to 1.4 [26].

Various arguments have been offered to justify the use of the lognormal distribution.

Normally distributed activation energy—Schwarz [27] used temperature-ramp resistance analysis to determine the distribution of activation energies for EM damages in *Al* and *Al*-4%*Cu* thin film interconnects. He found that the activation energies for the pure aluminum conductors are well represented by a normal distribution. For *Al*-4%*Cu* interconnects, there are three lognormal distributed subpopulation of activation energies.

Normally distributed conductor temperatures—Lloyd [28] demonstrated that a normal temperature distribution, given the variation is small compared to the mean temperature, can produce a lognormal failure distribution in EM lifetime experiments. Bobbio [29] also justified the applicability of the lognormal distribution based on the temperature dispersion of conductors during life testing.

Lognormal distributed grain sizes—Based on the grain boundary diameter distribution that was microscopically measured, Attardo et al. [30] used Monte Carlo simulation to get the failure times. Their simulation results determined that the statistical failure rate distribution best fits a lognormal curve for wide lines.

Experimental results—Although many EM experiments have been conducted, their sample sizes were too small to demonstrate strong support of the lognormal distribution. Towner [31]

performed EM lifetests on a variety of aluminum-alloy films using sample sizes ranging from 35 to 120. His results showed that the lognormal distribution rather than the logarithmic extreme distribution is a better fit when the grain size is smaller than the linewidth. Where the grain size exceeds the linewidth, either distribution can be used to represent the data. Gall [32] did an experiment utilizing large interconnect arrays in conjunction with a Wheatstone Bridge. Over a temperature range from 155 to 200°C, a total of more than 75,000 interconnects were tested. The results indicated that the EM failure mechanism in this experiment followed perfect lognormal behavior down to the four sigma level.

Although lognormal distribution is widely used in EM tests, it cannot be used as an element failure distribution that can be applied with the "weakest link" model. If $F_1(s)$ is the cumulative distribution function (CDF) of strength of a single link, then, statistically, the CDF of a chain of N independent (in strength) links is:

$$F_N(s) = 1 - [1 - F_1(s)]^N \tag{2.19}$$

In general, the form of the distribution F_N depends on the number of links (failure elements) in the chain (series); therefore, "scaling up" the model for longer chains affects the choice of modeling distribution. The lognormal distribution does not scale in Equation (2.19) and, therefore, cannot be the failure distribution for elements in series. This means lognormal distributions can approximate true failure times only in a finite percentile interval.

A Generalized Black Model has been proposed to characterize EM failures [33]:

$$t_{50} = A \times J^{-n} \times T^{-m} \times \exp(\frac{E_a}{\kappa T}) \tag{2.20}$$

where A is a process and material-related constant, J is the average current density, κ is the Boltzmann's constant, T is temperature in Kelvins, and E_a is an experimentally determined activation energy. The various values of n and m are determined by the particular failure physics

and the conductor's geometry. If $n = 2$, $m = 0$, we have the original Black model. With respect to failure physics, for all nucleation-dominated failures, $n = 2$, and $m = 0$; if a failure is growth-dominated, $n = 1$, and $m = 0$. With respect to the conductor's geometry, it has been observed that for wide lines (defined as those where the average grain size is smaller than the line width), $n = 2$; whereas for narrow lines, $n = 1$.

For engineering applications, there is not a significant difference between which n and m values are used; however, calculations show the combination of $n = 2$ and $m = 2$ would produce very good lifetime predictions and that the extrapolated activation energies would be reasonably accurate [33]. The side effect of using the Shatzke and Lloyd Model is the nonlinearity when the activation energy is extrapolated, making parameter extraction difficult.

Weibull Distribution

The Weibull probability density function can be expressed as:

$$f(t) = (\frac{\beta}{\alpha^{\beta}})t^{\beta-1}exp[-(\frac{t}{\alpha})^{\beta}] \tag{2.21}$$

where α is the characteristic lifetime and β is the shape parameter. Generally, the lognormal distribution fits the electromigration experimental data better. However, lifetest data of aluminum conductors showed that lognormal and Weibull distribution fit well equally at large failure rates (0.1 to 1%); at lower failure rates, the projected failure rates differ by several orders of magnitude [30]. There are different views concerning the usage of the Weibull distribution. For example, Gall et al. [32] used simulation results to roll out the Weibull distribution in the analysis of their experiment data. Pennetta [34] simulated EM damage in metallic interconnects by biased percolation of a random resistor network in the presence of degradation and recovery processes. Lognormal distribution and Weibull distribution both fit the simulation result well in this case.

Bimodal Lognormal Distribution

A single lognormal distribution is observed if only one physical mechanism dominates the failure process. It is possible that two different failure modes act in parallel within a sample or even a single specimen. This failure distribution might appear nonlinearly in the lognormal probability plotting paper. In this situation, a bimodal lognormal distribution might help us model the data. Suppose there are two different EM failure mechanisms. Each individual mechanism is described by a lognormal distribution $CDF_A(t)$, (respectively, $CDF_B(t)$) over time t, with median time-to-failure t_{50A} (respectively, t_{50B}) and standard deviation σ_A (respectively, σ_B). In addition, the failure mechanisms might have different activation energies, E_A and E_B, and current density exponents n_A and n_B. There are two different models of an overall bimodal failure distribution [35]:

1. Superposition Model

Consider a sample in which the failure scenario is influenced by the presence or absence of a particular physical property in the test device. Its presence forces a specimen to fail due to mechanism A; its absence exclusively due to B. The property appears with a probability $P(A)$, the property is absent with a probability $P(B)=1-P(A)$, and the overall CDF of all the specimens is:

$$CDF(t) = P(A) \cdot CDF_A(t) + (1 - P(A)) \cdot CDF_B(t) \tag{2.22}$$

The resulting CDF appears s-shaped in the probability-plot.

2. Weak-Link Model

In this scenario, different failure mechanisms can cause the interconnect failure in a serial fashion. If the failure mechanisms act statistically independently, the overall CDF is given by:

$$CDF(t) = 1 - (1 - CDF_A(t)) \cdot (1 - CDF_B(t)) \tag{2.23}$$

This CDF is “hook-shaped” in the probability plot.

The bimodal lognormal distribution is often seen in copper via EM tests. Lai [36] described two EM failure mechanisms: via related and metal-stripe-related. Ogawa [37] reported two distinct failure modes in dual-damascene Cu/oxide interconnects. One model described void formation within the dual-damascene via; the other reflected voiding that occurs in the dual-damascene trench. These models formed a bimodal lognormal distribution.

2.2 Hot Carrier Degradation

2.2.1 Introduction

Hot carrier degradation (HCD) has been studied for more than 30 years as an important failure mechanism that must be mitigated in the design of aggressively scaled VLSI devices. Extensive work has focused on the physical mechanisms, life estimation, and technology improvement over the last three decades. A physical understanding of HCD and the respective models are briefly introduced.

2.2.2 Hot Carriers

For semiconductors in thermal equilibrium, electrons and holes continually absorb and emit acoustical phonons (low-frequency lattice vibrations), resulting in an average energy gain of zero. Such electrons have kinetic energies (E) that are normally slightly higher than that of the conduction band edge (E_C) by an amount kT_r (T_r is room temperature). Similarly, for holes, E is slightly less than the valence band edge (E_V) by kT_r. In the case of low electrical fields, the carrier velocity is field-independent and kT_r is only 0.025 eV; small compared to the carrier kinetic energy corresponding to E_C and E_V. However, if the electrical field is very high (for example, 100 kV/cm), the carriers gain more energy than they lose by scattering. Such accelerated electrons have energies of $E_C + kT_e$, where T_e is an effective temperature such that $kT_e > kT_r$. With effective temperatures ($\sim E_C / kT$) of tens of thousands of degrees Kelvin, these electrons are at the very top of the Fermi distribution and are known as *hot electrons*.

During the operation of the metal oxide semiconductor field effect transistor (MOSFET), if the gate voltage is comparable to or lower than V_{DS}, the inversion layer is much stronger on the source side than the drain side and the voltage drop due to channel current is concentrated on the drain side (if $V_D > V_S$). The field near this side can be so high that carriers can gain enough energy between two scattering events to become hot carriers. The majority of these hot carriers simply continue toward the drain, but a small number of them gain enough energy to generate electrons and holes by impact ionization. In the n-channel MOSFET (NMOSFET), the vast majority of the generated holes are collected by the substrate and give rise to the substrate current (I_{Sub}) and the generated electrons enhance the drain current (I_D). Photon emission might also occur during hot carrier generation in the drain.

Some of the hot carriers with enough energy (approximately 3.2 eV for electrons and 4.7 eV for holes) [38] can surmount the energy barrier at the $Si{-}SiO_2$ interface and be injected into the oxide, with a small gate current (I_G). Some energetic injected carriers might break some $Si{-}H$ or similar weak bonds in the oxide or at the $Si{-}SiO_2$ interface. If the hot carrier injection lasts long enough, the trapped charge or generated defects will permanently modify the electric field at the $Si{-}SiO_2$ interface and, hence, the electrical characteristics of the MOSFET. Figure 2.3 schematically shows the process of HCD.

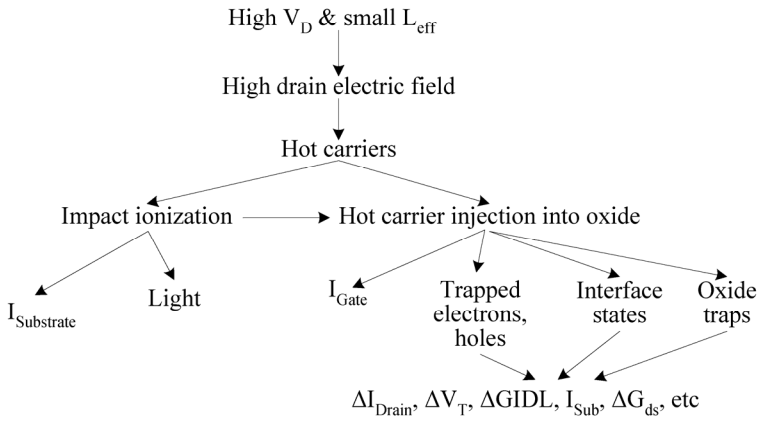

Fig. 2.3. *Hot carrier generation and degradation in MOSFETs.*

2.2.3 Hot Carrier Injection Mechanisms

According to Takeda [39], there are three main types of hot carrier injection modes:

1. Channel hot electron (CHE) injection.
2. Drain avalanche hot carrier (DAHC) injection.
3. Secondary generated hot electron (SGHE) injection.

CHE injection is due to the escape of "lucky" electrons from the channel, causing a significant degradation of the oxide and the $Si{-}SiO_2$ interface, especially at low temperature (77 K) [40]. On the other hand, DAHC injection results in both electron and hole gate currents due to impact ionization, giving rise to the most severe degradation around room temperature. SGHE injection is due to minority carriers from secondary impact ionization or, more likely, bremsstrahlung radiation, and becomes a problem in ultra-small metal oxide semiconductor (MOS) devices. Fowler–Nordheim tunneling and direct tunneling might also cause hot carrier injection. For deep sub-micrometer devices, it is important to attempt to account for the effects resulting from combinations of some if not all of these injection processes.

Channel Hot Electron (CHE) Injection

The CHE injection process is shown in Figure 2.4. In this figure, CHE injection occurs when the gate voltage (V_G) is comparable to the drain voltage (V_D) (an NMOSFET). The gate current (I_G) rises as V_G initially increases, peaks when V_G is roughly equal to the drain-source potential V_D, and drops thereafter. There are two reasons that cause the I_G to increase. First, the inversion charge in the channel increases, so that more electrons are present for injection into the oxide. Second, the stronger influence of the vertical electric field in the oxide prevents electrons in the oxide from detrapping and drifting back into the channel. It was reported [39] that if an n-channel MOSFET is operating at $V_G = V_D$ the conditions would be optimum for CHE injection of "lucky electrons." Such electrons gain sufficient energy to surmount the $Si{-}SiO_2$ barrier without suffering an energy-losing collision in the channel. In many cases, this gate current is responsible for device degradation as a result of carrier trapping. No gate current can be measured for $V_G < V_D$, since

CHE injection is retarded. However, if V_D is large enough, a reduction of V_G intensifies the electric field at the drain to the point where avalanche multiplication due to impact ionization might substantially increase the supply of both hot electrons and hot holes.

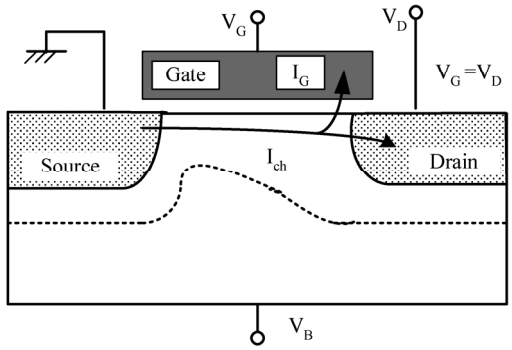

Fig. 2.4. *Channel-hot-electron injection. Occurs when (V_G) of NMOSFET is comparable to (V_D).*

Drain Avalanche Hot Carrier (DAHC) Injection

The DAHC injection process generally occurs when V_D exceeds V_G. It is schematically shown in Figure 2.5. This mechanism first depends on an impact-ionization avalanche to create carriers. These secondary electrons then become hot and cause degradation. In the case of high substrate-bias voltages, additional secondary hot electrons generated from deeper S_i substrate regions can also be injected into the oxide. These secondary electrons produce less damage than the primary hot electrons. Analyzing DAHC behavior is difficult because hot holes and hot electrons are injected simultaneously into the oxide and across the drain junction just below the substrate surface.

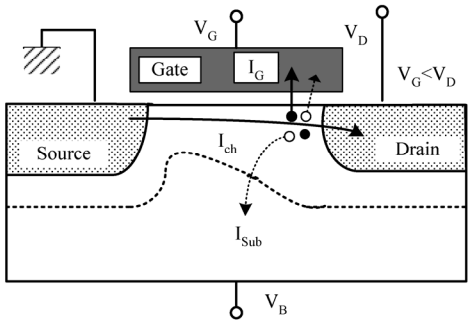

Fig. 2.5. *Drain avalanche hot-carrier injection. Occurs when* $V_D > V_G$.

Secondary Generated Hot Electron (SGHE) Injection

Secondary impact ionization by hot holes and photo-induced generation processes have been reported as secondary minority carrier generation mechanisms. This injection process is shown in Figure 2.6. Takeda [39] experimentally demonstrated that photo-induced generation is the main physical mechanism. The temperature dependence of I_{Sub} and that of electron diffusion current, I_D, were compared to each other for a device with t_{ox} = 7 nm and L_{eff} = 2.0 μm. The experiment results imply that a photo-induced generation process, believed to be bremsstrahlung radiation, rather than secondary impact ionization, is more likely to be the origin of the SGHE.

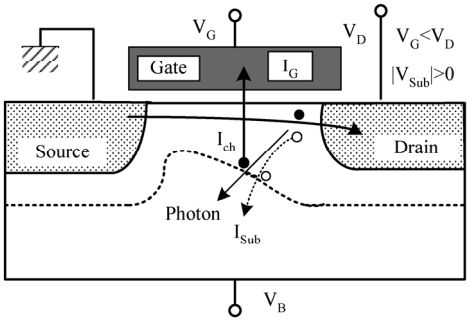

Fig. 2.6. *Secondarily generated hot electron injection*

2.2.4 HCD Models

Lucky Electron Model

The lucky electron approach of modeling the hot electron distribution was originated by Shockley. Verwey [41] applied it in the study of substrate hot electron injection, which was subsequently refined and verified by Ning [42]. Hu [43] modified the substrate lucky electron injection model and applied it to CHE in MOSFETs.

The basic assumption of the lucky electron model hinges on a supply of hot electrons that are "lucky." For channel electrons to reach the gate oxide, two conditions have to be met. The first requires that electrons gain sufficient kinetic energy from the channel field to become "hot." Second, the electron momentum must be redirected perpendicularly, so that hot electrons can enter the oxide. The probability that a channel electron will travel a distance d or more without suffering any collision is equal to $\exp[-d/\lambda]$, where λ is the mean free path between scattering events. Consider an electron of charge q traveling a distance λ in the channel electric field ε_c: the probability that it will reach energy φ without suffering a collision is given by $\exp[-\Phi/(\lambda q\varepsilon_c)]$, as $d = \Phi/(q\varepsilon_c)$. The probability of a hot electron's redirection to the $Si-SiO_2$ interface without suffering any collision is essentially a function of oxide field ε_{ox} [44]. The measurable consequences of these processes are substrate (I_{Sub}) and gate currents (I_G) whose magnitudes depend on sufficient electron energies for impact ionization (Φ_i) and for surmounting the $Si-SiO_2$ energy barrier (Φ_b), respectively. Hu [45] presented a general model for the hot electron effects. The equations are shown below:

$$I_{Sub} = C_1 I_D \exp[-\frac{\phi_i}{\lambda q\varepsilon_c}], \tag{2.24}$$

$$I_G = C_2 I_D \exp[-\frac{\phi_b}{\lambda q\varepsilon_c}], \tag{2.25}$$

$$\tau = C_3 \frac{W}{I_D} \exp[\frac{\phi_{it}}{\lambda q \varepsilon_c}] \tag{2.26}$$

where C_1, C_2, and C_3 are constants; τ is the device lifetime; W is the channel width; Φ_{it} is interface trap creation energy; and I_D is the drain current flow that supplies some of the eventually lucky electrons. From Equation 2.24 and 2.25, eliminating $\lambda q \varepsilon_c$ yields the correlation between I_G and I_{Sub}:

$$\frac{I_G}{I_D} = C_2 [\frac{I_{Sub}}{C_1 I_D}]^m, \tag{2.27}$$

where $m = \Phi_b/\Phi_i$. This equation, which applies to the case where V_G exceeds V_D, has been verified in n-channel transistors. It is found that m is approximately 3, which is roughly consistent with values of $\Phi_i \approx 1.3$ eV and $\Phi_b \approx 3.5$ eV.

From 2.24, 2.25, and 2.26, the device lifetime τ is

$$\frac{\tau I_D}{W} \propto [\frac{I_{Sub}}{I_D}]^{-\phi_{it}/\phi_i}. \tag{2.28}$$

The model above assumes static, or DC, voltages and currents. As device operation involves AC time-dependent wave forms, dynamic degradation needs to be considered. By integrating or time averaging over substrate and drain currents, the τ [46] is

$$\tau^{-1} = \frac{B}{T_c} \int_0^{T_c} I_D (\frac{I_{Sub}}{I_D})^m dt, \tag{2.29}$$

where T_c is the full cycle time.

The lucky electron model (LEM) has two major limitations [47]: a) it relates HCD to a local field (E_L), thus neglecting the space and time lag of carriers in reaching local equilibrium with the field; and b) since the potential energy is the only source of energy available to the carriers, the maximum attainable energy is limited to qV_{TOT} where V_{TOT} is the total voltage drop experienced by the carriers. Therefore, the LEM predicts no HCD at voltages smaller than the threshold energy.

There are higher order models that attempt to overcome the above-noted limitations: a) "non-local" LEMs; and b) effective temperature models (ETM). Non-local LEMs replace local field with "non-local" quantities, such as the potential drop along the current flowlines or a suitable electric field. ETMs assume quasi-equilibrium Maxwellian distributions whose effective temperature (T_e) is a function of the local field. In their simplest forms, LEM and ETM predict the relationships between I_{Sub} and I_G:

$$I_{Sub} \propto I_D \exp(-\phi_i/E^*), \tag{2.30}$$

$$I_G \propto B(E_{OX}) I_D \exp(-\phi_b/E^*), \tag{2.31}$$

$$I_G/I_D \propto (I_{Sub}/I_D)^{\phi_b/\phi_i} \tag{2.32}$$

where $B(E_{OX})$ models the collecting efficiency of the gate, $E^* = q\lambda E_L$ for LEM, and $E^* = k_B T_e$ for ETM.

Empirical Models

Power law model

This model was proposed by Takeda [48] based on the following assumptions:

1. Avalanche hot carrier injection due to impact ionization at the drain, rather than channel hot electron injection composed of "lucky electrons," imposes the severest constraints on device design.
2. Device degradation (V_{th} shift and G_m change) resulting from drain avalanche hot carrier injection has a strong correlation to impact ionization induced substrate current.

The V_{th} shift, ΔV_{th}, or G_m degradation, $\Delta G_m/G_{m0}$, can be empirically expressed as

$$\triangle V_{th}(or \triangle G_m/G_{m0}) = At^n \tag{2.33}$$

This expression is particularly valid for short stress times, while for long stress times, ΔV_{th} and/or $\Delta G_m/G_{m0}$ begins to saturate. The slope n or ΔV_{th}, in a log-log plot is strongly dependent on V_G, but has little dependence on V_D. This suggests that n changes according to the hot carrier injection mechanism. The magnitude of degradation, A, is strongly dependent on V_D and has little dependence on V_G. In particular,

$$A \propto \exp(-\alpha/V_D) \tag{2.34}$$

Therefore, the lifetime τ can be expressed as

$$\tau \propto \exp(b/V_D) \tag{2.35}$$

where $b = \alpha/n$.

Takeda [39] and Hu [45] both reported $\tau \propto I^{m}_{Sub}$, while m ranging between 3.2–3.4 given by Takeda and 2.9 by Hu.

Other empirical models

Several other empirical models have been proposed, including the logarithmic law, the mixed law, the federative law, the saturation law, and the body effect law.

1. Logarithmic law.

   For degradation under long stress time, the logarithmic law [49] is described as

$$\Delta = B\ln(t) - C \tag{2.36}$$

   where B and C are technology parameters.

2. Mixed law.

In the study of sub-0.25 μm bulk S_i MOSFETs, Szelag [50] proposed the mixed law, which combines a power and logarithmic time dependence. The mixed law is written as

$$\Delta G_{m_{max}}(t) = A[log(t)]^n \tag{2.37}$$

where $G_{m\text{-}max}$ is the maximum transconductance.

3. Federative law.

Marchand [51] suggested to apply the federative law, which was first found in the analysis of stress-induced leakage current. The relationship is expressed as

$$J = J_0 \exp(-Bt^n). \tag{2.38}$$

It can be further expressed as

$$t\frac{d\ln(J)}{dt} = nBt^{-n}. \tag{2.39}$$

Equation 2.39 is a straight line in log-log scale where the slope gives the exponent n and quickly informs about the saturating (negative slope) and non-saturating (positive slope) nature of the law.

4. Saturation law.

Chan [52] proposed the saturation model in the study of oxide-spacer lightly doped drain (LDD) NMOSFETs HCD. In this model, two factors are considered: the increase in the series resistance in the LDD region and the reduction of the carrier mobility in the channel and subdiffusion region. By combining both factors, the linear-current degradation for LDD NMOSFETs is expressed as

$$\frac{\Delta I_D}{I_D}(t) = f(R_D(t)) + [\frac{I_D}{WH}(\frac{I_{Sub}}{I_D})^m t]^n \tag{2.40}$$

where W is the device width; n is the degradation rate coefficient; m and H are technology-related parameters; and $f(R_D(t))$ represents the current reduction due to the increase in the drain series resistance. The lifetime correlation can be expressed as

$$\frac{\tau I_D}{W} = H[\frac{\Delta I_D}{I_D}|_{LifetimeCriterion} - f(R_D(\infty))]^{\frac{1}{n}}(\frac{I_{Sub}}{I_D})^{-m} \tag{2.41}$$

where $f(R_D(\infty))$ is the maximum amount of current reduction due to the series resistance increase.

5. Body effect law.

Koike [53] considered the body effect caused by the secondary hot electrons and developed a drain avalanche HCD model for NMOSFET. This model assumes that the recombination of hot holes and hot electrons injected into the gate oxide causes the interface trap generation around maximum I_{Sub} condition and at small $|V_B|$. Both primary hot electron ($e1$) and secondary hot electron ($e2$) contribute to the hot electron current density I_{he}. The lifetime τ is assumed inversely proportional to a product of hot hole current density I_{hh}/W and hot electron current density I_{he}/W, where W is the channel width. τ can be modeled as

$$\tau^{-1} = \tau_{e1}^{-1} + \tau_{e2}^{-1}. \tag{2.42}$$

$$\tau^{-1} \propto (I_{hh}/W)(I_{he1}/W). \tag{2.43}$$

$$\tau^{-1} \propto (I_{hh}/W)(I_{he2}/W). \tag{2.44}$$

On the basis of the lucky electron concept, τ_{e1} is modeled as

$$\tau_{e1}(\frac{I_D}{W})^2 = (\Delta D_f)^{1/n} H_{e1}(\frac{I_{Sub}}{I_D})^{-m_{e1}}, \tag{2.45}$$

where ΔD_f is the degradation criteria for lifetime and H_{e1}, m_{e1}, and n are the first impact ionization parameters. τ_{e2} is modeled as

$$\tau_{e2}(\frac{I_D}{W})^2 = (\Delta D_f)^{1/n} H_{e2}(\frac{I_{Sub}}{I_D})^{-m_{e2}} \exp(\frac{a_{e2}}{|V_B|}), \tag{2.46}$$

where H_{e2}, m_{e2}, and a_{e2} are the second impact ionization parameters.

Statistical Models

Literature research did not reveal any statistical model that has been widely accepted in hot carrier analysis. Engineers tend to use lognormal distribution to analyze HCI failure data.

2.2.5 Acceleration Factors

Hot carrier effects are enhanced at low temperature. The main reason is an increase in electron mean free path and impact ionization rate at low temperature. As shown in [54], substrate current at 77 K is five times greater than that at room temperature (RT), and CHE gate current is approximately 1.5 orders of magnitude greater than that at RT. At low temperature, the electron trapping efficiency increases and the effect of fixed charges becomes large [38]. This accelerates the degradation of G_m at low temperature. The degradation of V_{th} and G_m at low temperatures is more severely accelerated for CHE-induced effects than for DAHC. Hu [45] showed the temperature coefficient of CHE gate and substrate current to be negative.

The temperature acceleration factor is expressed as

$$AF = \exp[E_\alpha(\frac{1}{T_1} - \frac{1}{T_2})] \tag{2.47}$$

where T_1 and T_2 are operating temperatures and E_a is the activation energy, with a value around −0.1 eV ~ −0.2 eV [55].

2.3 Time-Dependent Dielectric Breakdown

2.3.1 Introduction

TDDB is a wearout phenomenon of SiO_2, the thin insulating layer between the control “gate” and the conducting “channel” of the transistor. SiO_2 has a very high bandgap (approximately 9 eV) and excellent scaling and process integration capabilities, which makes it the key factor in the

success of MOS-technology. Dielectric layers as thin as 1.5 nm can be obtained in fully functioning MOSFETs with gate lengths of only 40 nm [56]. Although SiO_2 has many extraordinary properties, it is not perfect and suffers degradation caused by stress factors, such as a high oxide field. Oxide degradation has been the subject of numerous studies that were published over the past four decades. Even today, a complete understanding of TDDB has not yet been reached. Basic models, such as E model and 1/E model, have been proposed and are still debated in the reliability community. The statistical nature of TDDB is well described by the Weibull distribution, since TDDB appears to be a "weakest link" type of failure mechanism. Percolation theory has been successfully applied to the statistical description of TDDB. As oxide continues to scale down, new findings will help researchers gain a better understanding of this complicated process.

2.3.2 Physics of Breakdown

The exact physical mechanism of TDDB is still an open question. The general belief is that a driving force such as the applied voltage or the resulting tunneling electrons create defects in the volume of the oxide film. The defects accumulate with time and eventually reach a critical density, triggering a sudden loss of dielectric properties. A surge of current produces a large localized rise in temperature, leading to permanent structural damage in the silicon oxide film.

Charge in Silicon Dioxide and at the Silicon-Oxide Interface

Silicon dioxide is far from perfect, as oxide rings have the tendency to create shallow oxide vacancies. Particularly, the interface at the $Si\text{-}SiO_2$ interface is prone to dangling bonds that require H passivation. Both dangling bonds and vacancies are electron trap sites that result in threshold shifts and a degradation of Gm. There are charges inside the oxide and near the silicon-oxide interface. These charges can be mobile ionic charges, electrons, or holes trapped in the oxide layer. They can also be fabrication-process-induced fixed oxide charges near the silicon-oxide interface, and charges trapped at the surface states at the silicon-oxide interface. Electrons and holes can make transitions between the crystalline states near the silicon-oxide interface to the surface states. These charges will definitely affect the electrical characteristics of devices and are

important factors in TDDB. Figure 2.7 shows the names and locations of charges inside silicon dioxide and at the silicon-oxide interface.

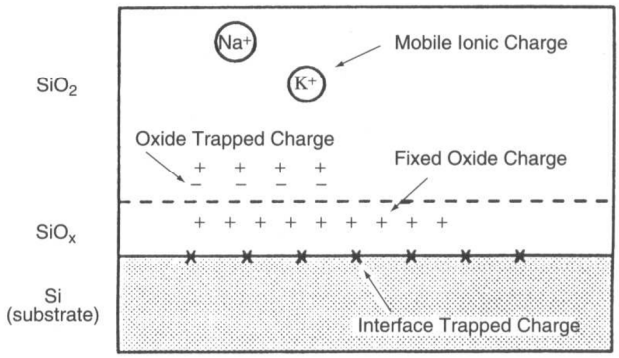

Fig. 2.7. *Location and identification of charges in Si−SiO_2 and at the oxide-silicon surface [17].*

1. Interfacial oxide charge: This charge is located within 0.2 nm of the *SiO_2−Si* surface. The interfacial oxide charge arises from oxide vacancies, metal impurities, dangling bonds, and broken bonds due to charge injection. These interfacial states are amphoteric: that is, they are acceptor-like in the upper half of the Si band gap and donor-like in the lower half of the band gap.
2. Fixed oxide charge: Fixed oxide charge is a positive charge located some 3 to 5 nm from the *Si–SiO_2* interface. Due to the nature of modern electronics, bulk properties of modern oxides are harder to define. Fixed and trapped oxide charges are generally likely to occur at oxygen vacancy sites. The most common defects are the E' and E' delta sites, and are primarily due to excess silicon species introduced during oxidation and postoxidation heat treatment. They are fixed and largely uninfluenced by the normal operating voltages of the MOS transistor.
3. Oxide trapped charge: This charge is also likely to occur at oxygen vacancy sites. The sources of this charge include the oxide growth process, fabrication of device [57], and high-energy electrons. A fabrication-introduced charge can be removed through low-temperature annealing.
4. Mobile Na+ and K+ ionic charge: These charges have been virtually eliminated as a source of reliability problems.

It is the generation of oxide charge states under high electric fields that ultimately leads to dielectric breakdown. There are processes such as Fowler–Nordheim tunneling, direct tunneling, and trap-assisted tunneling that contribute to the overall creation and persistence of oxide charges.

Tunneling Current

Fowler–Nordheim Tunneling

Fowler–Nordheim tunneling is a quantum mechanical tunneling process where the electrons penetrate through the oxide barrier into the conduction band of the oxide under the assistance of a high electric field. The complete theory of Fowler–Nordheim tunneling is complicated and not discussed fully here. Figure 2.8 illustrates electron tunneling from the silicon surface inversion layer to the SiO_2 conduction band.

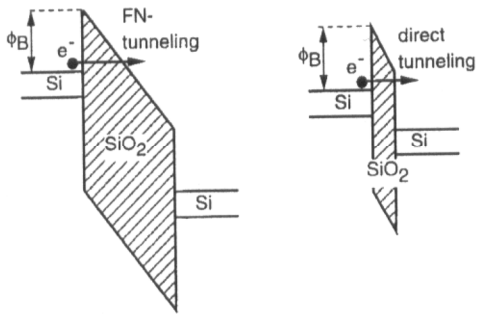

Fig. 2.8. *Schematic illustration of the Fowler-Nordheim tunneling current and direct tunneling current.*

The Fowler–Nordheim tunneling current density is given by

$$J_{FN} = \frac{q^3 \varepsilon_{ox}^2}{16\pi^2 \hbar \phi_{ox}} \exp\left(-\frac{4\sqrt{2m^\star}\phi_{ox}^{\frac{3}{2}}}{3\hbar q \varepsilon_{ox}}\right) \tag{2.48}$$

where ε_{ox} is the electric field in the oxide, Φ_{ox} is the silicon-silicon dioxide interface potential barrier for electrons, m^* is the electron effective mass, and h is Planck's constant. Equation 2.48 shows that Fowler–Nordheim tunneling current is characterized by a straight line in a plot of log (J/ε^2) versus $1/\varepsilon_{ox}$. Fowler–Nordheim tunneling current is dependent on the oxide field, thus the voltage applied to the gate oxide. It can occur in any gate oxide, provided the voltage is sufficient for electrons to tunnel through the barrier.

Direct Tunneling

Direct tunneling is the dominant current conduction mechanism through sub-3-nm oxide layers. The tunneling probability is given by the well depth and breadth. The field modulates the well depth. There is no simple dependence of the tunneling current density on voltage or electric field, without a closed analytic form of expression. The direct tunneling current can be very high for thin oxide layers in modern oxides, as shown by Lo [58].

Trap-Assisted Tunneling and Other Effects

Tunneling depends on well depth and breadth. Traps effectively subdivide the well into shorter sections (i.e., breadth) while the field modulates the trap electric level (depth). Trap-assisted tunneling is dependent on the density of the traps and the electric field.

Other factors can influence the tunneling current, the gate-drain and the gate-source overlap regions. Another factor engineers must take into account is valence band tunneling. Valence band tunneling becomes more important with very thin oxides. Another mechanism that has been observed is electron hopping. Electron hopping is caused by the jump of thermally excited electrons between isolated states. Field emission, or the tunneling of trapped electrons to the conduction band, is an additional factor affecting the tunneling current. Finally, Poole-Frenkel emission, or the tunneling of trapped electrons into the conduction band due to the barrier lowering, can affect the overall gate current.

Trap-Generation Mechanisms

Trap generation is the key factor determining the oxide degradation and breakdown. Several models have been proposed and discussed. These models include the "anode hole injection" (AHI) model, the "thermochemical" model or "E-model," and the "anode hydrogen release" (AHR) model.

AHI Model

The AHI model (1/E model) was initially proposed by Schuegraf and Hu [59]. This model suggests that breakdown is caused by holes that are injected from the anode. Electrons injected from the cathode into the oxide undergo impact ionization events that generate holes in the process. These holes become trapped in the oxide near the cathode, distorting the band diagram, and increasing the field nearby, as shown in Figure 2.9. Electron tunneling is enhanced in the high field according to Fowler–Nordheim tunneling Equation 2.48, thus resulting in greater current injection. Another mechanism occurs at the anode side of the oxide as the electron drops down to the Fermi level and leads to the release of its energy of at least 3.1 eV to the lattice at the SiO_2 interface [60]. This energy is sufficient to break the $Si{-}O$ bond. The breaking of bonds proceeds from anode to cathode and forms a convenient conductive path for discharge that causes dielectric breakdown. In both cases, the injected oxide charge is accumulated inside the oxide until a critical hole charge density is reached for dielectric breakdown.

The reciprocal field expression of time-to-breakdown (TBD) based on the AHI model takes the form *G(T)*

$$T_{BD} = \tau_0(T) \exp[\frac{G(T)}{\varepsilon_{OX}}] \tag{2.49}$$

where ε_{ox} is the electric field across the dielectric in MV/cm. Constants $\tau_0(T)$ and $G(T)$ are temperature-dependent and given by $\tau_0(T)$=5.4 $\times 10^{-7}$exp[−0.28 eV/kT](sec), and $G(T)$ = 120+ 5.8/kT MV/cm, where k is Boltz-mann's constant and T is the absolute temperature.

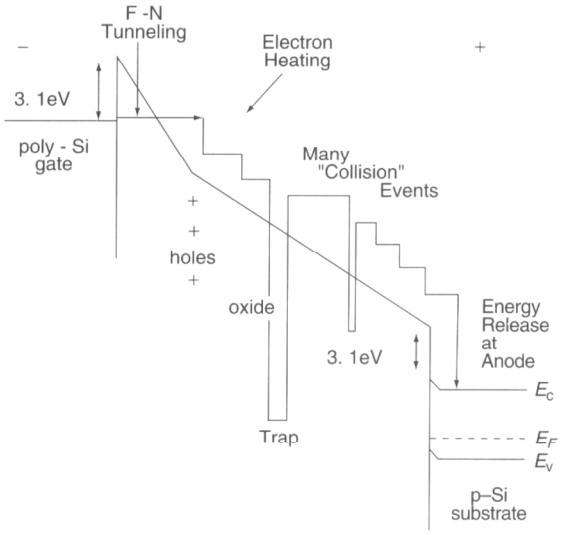

Fig. 2.9. *Band diagram models of AHI Model.*

Support of the AHI model comes from a theoretical treatment of AHI by surface plasmon excitations, and experimental data showing the expected dependence on the anode material [61]. This model was criticized for its inability to account for the substrate currents measured at low voltages. According to the plasmon model [62], the gate voltage threshold for positive charge generation by hole trapping due to AHI is 7 to 8 V. The AHI model has been further criticized because there are other origins of substrate current at low voltage besides tunneling holes; these include generation-recombination processes in the substrate and photoexcitation due to photons generated by hot electrons in the gate [61]. However, some recent experimental evidence and modeling have demonstrated the effectiveness of the AHI model in ultrathin oxides at a low gate voltage [63].

Thermochemical Model

The thermochemical model (E model) is another widely cited dielectric breakdown model. McPherson [64] reviewed the development of this model and proposed a physical explanation. This model proposes that defect generation is a field-driven process, and the current flowing through the oxide plays at most a secondary role. The interaction of the applied electric field with the dipole moments associated with oxygen vacancies (weak $Si-Si$ bonds) in SiO_2 lowers the activation energy required for thermal bond breakage and accelerates the dielectric degradation process. Eventual charge trapping at the broken bond sites and their wave function overlap and lead to a conduction subband formation. Consequently, severe Joule heating occurs at the stage of oxide breakdown. McPherson [64] also showed that allowing for a distribution of energies of the weak bonds could account for a wide range of observations of the temperature and field dependence of dielectric breakdown times. The E model suggests T_{BD} is given by

$$T_{BD} = A_0 \exp[-\gamma \varepsilon_{OX}] \exp[\frac{E_a}{kT}], \tag{2.50}$$

where:

A_0 : arbitrary scale factor; dependent upon materials and process details

γ : field acceleration parameter; temperature dependent, $\gamma(T) = a/kT$ where a is the effective dipole moment for the molecule

ε_{OX} : externally applied electric field across the dielectric.

The E model has attained widespread acceptance on the basis of experimentally verified exponential dependence of TBD on the electric field [61]. However, this alone is not enough to prove the validity of this model. It was observed that for very thin oxides, the breakdown times are no longer a function of only the field, but also strongly decrease with thickness at the same oxide field. The decreasing breakdown times are consistent with the increasing direct-tunneling leakage currents in the ultrathin oxides. An AHI-like mechanism was proposed, suggesting that the strong

increase in current leads to an increase in-hole injection, and that these holes are trapped at oxygen vacancies, further reducing the activation energy for bond rupture. Substrate-hot-electron (SHE) injection experiments showed that TBD is inversely related to the current density, demonstrating that breakdown is dominated by the effect of the energetic electrons and not the field in the oxide [65].

AHR Model

There is evidence supporting the AHR model that involves the release of atomic hydrogen from the anode by energetic tunneling electrons [66]. The released hydrogen diffuses through the oxide and can generate electron traps. Experiments have shown that exposure of bare SiO_2 films to atomic hydrogen radicals, even without any electric field, will produce electrically active defects essentially identical to those produced by electrical stress or radiation [61]. DiMaria [66] showed that the desorption rate of hydrogen from silicon surfaces is similar to the voltage dependence of the trap generation process. Based on data taken at IBM, he determined that hydrogen release requires electrons with energy levels of at least 5 eV in the anode and 2 eV in the oxide.

The primary argument against the hydrogen release process for oxide breakdown is the apparent lack of any isotope effect for the breakdown process compared to the large effect observed for hydrogen/deuterium desorption and channel hot electron induced interface degradation. The observation of TBD does not appear to improve if an isotope of hydrogen is used to passivate the silicon–oxide interface [67].

Degraeve and coauthors [68] gave an outline of these three models on neutral electron trap generation, which is shown in Figure 2.10. Figure 2.10(a) shows an overview of the AHI model. The AHR model is included in “Other Mechanism” of Figure 2.10(b), and the thermochemical model is shown in Figure 2.10(c).

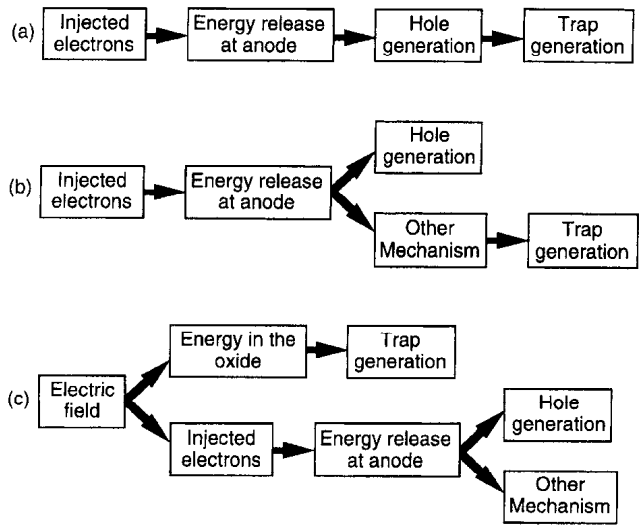

Fig. 2.10. *Outline of neutral electron trap generation.*

In addition to the basic models described above, there are other factors that will affect the dielectric breakdown. Researchers have observed a strong dependence of breakdown on anode and cathode type. This dependence is explained by the difference in current density between an n-type cathode and a p-type cathode. This shifts the trap creation threshold relative to the silicon anode Fermi level. Furthermore, hot carrier effects can shorten the time to breakdown data. Hot electrons can increase the trap generation rate, causing accelerated breakdown.

In more recent ultrathin oxide studies, Wu [69] proposes a non-Arrhenius temperature relationship and a soft-breakdown of the dielectric. Wu proposes the the following model to describe interrelationship and impact of voltage and temperature on breakdown:

$$T_{BD} = T_{BD0}(V)\exp[\frac{a(V)}{T} + \frac{b(V)}{T^2}] \tag{2.51}$$

where T_{BD0} and coefficients a and b are voltage dependent.

2.3.3 Oxide Breakdown Models

Projection of dielectric lifetime of a product from data collected by stressing test structures under accelerated test conditions requires correct models and assumptions. The voltage, temperature acceleration, and area scaling must be performed using the proper model.

Weibull Distribution

The statistics of gate oxide breakdown are usually described using the Weibull distribution

$$F(t) = 1 - \exp[-(t/\alpha)^{\beta}], \tag{2.52}$$

which is an extreme-value distribution in *ln(x)* and is a "weakest link" type of problem. Here *F* is the cumulative failure probability, *t* can be either time or charge, α is the scale parameter (63.2 percentile), and β is the shape parameter. The "weakest link" model was formulated by Sune et al. [70] and described oxide breakdown and defect generation via a Poisson process. In this model, a capacitor is divided into a large number of small cells. It is assumed that during oxide stressing, neutral electron traps are generated at random positions on the capacitor area. The number of traps in each cell is counted, and at the moment that the number of traps in one cell reaches a critical value, breakdown will occur. Dumin [71] incorporated this model to describe failure distributions in thin oxides.

Gate oxide failure is a weakest-link type of problem because the whole chip fails if any one device fails, and a device fails if any small portion of the gate area of the device breaks down. Statistically, if the probability of any one unit failing is p, then the probability of any one of N-independent units failing is

$$F = 1 - (1 - p)^{N}. \tag{2.53}$$

Therefore,

$$ln[-ln(1-F)] = lnN + ln[-ln(1-p)]. \tag{2.54}$$

From Equation (2.52) and Equation (2.54), an extremely useful property of Weibull distribution is derived: if the oxide area is increased by a factor *(A′ /A)* then the curve shifts vertically by *ln*(*A′ /A*) and the scale parameter α decreases to *α′*, according to

$$\frac{\alpha'}{\alpha} = (\frac{A}{A'})^{1/\beta}. \tag{2.55}$$

This is helpful in relating breakdown tests on individual small area capacitors to the reliability of an integrated circuit containing many millions of gates.

The Weibull slope β is an important parameter for reliability projections. A key advance was the realization that β is a function of oxide thickness t_{OX}, becoming smaller as t_{OX} decreases [72, 73, 74]. The smaller β for thinner oxide is explained because the conductive path in the thinnest oxides consists of only a few traps and, therefore, has a larger statistical spread.

Log-normal distribution has also been used to analyze accelerated test data of dielectric breakdown. Although it may fit failure data over a limited sample set, it has been demonstrated that the Weibull distribution more accurately fits large samples of TDDB failures [75]. An important disadvantage of the lognormal distribution is that it does not predict the observed area dependence of TBD for ultrathin gate oxides.

Percolation Theory

The percolation theory was applied to modeling the intrinsic breakdown distribution by Degraeve et al. [76]. Stathis [72] used a computer simulation to demonstrate the thickness dependence of the number of defects at breakdown using percolation theory.

Figure 2.11 shows the percolation model for oxide breakdown. It is assumed that electron traps are generated inside the oxide at random positions in space. Around these traps, a sphere is defined with a fixed radius *r*, which is the only parameter of this model (see Figure 2.11(a)). If the spheres of two neighboring traps overlap, conduction between these traps becomes possible. The two interfaces are modeled as an infinite set of traps in Figure 2.11(b). This mechanism of trap generation continues until a conducting path is created from one interface to the other and breakdown occurs (see Figure 2.11(c)).

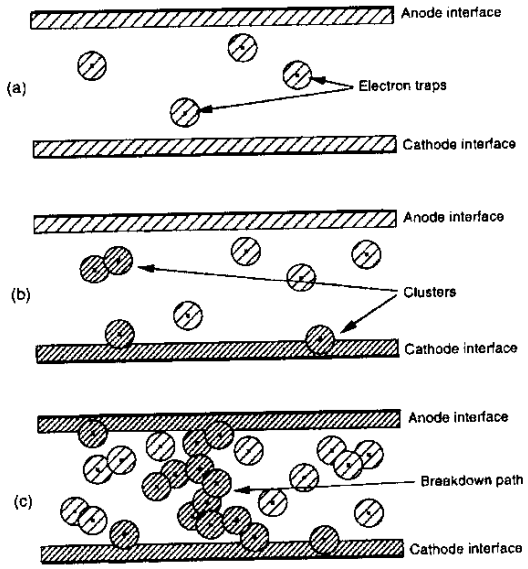

Fig. 2.11. *The percolation model for oxide breakdown.*

The percolation model for breakdown is able to explain quantitatively two important experimental observations: (i) as the oxide thickness decreases, the density of oxide traps needed to trigger

breakdown decreases; and (ii) as the oxide thickness decreases, the Weibull slope of breakdown distribution decreases [68].

2.3.4 Acceleration Factors

For TDDB, the calculation of acceleration factors depends on the model that has been chosen.

AF of 1/E Model

$$AF_{1/E}(T_0, \varepsilon_0; T_1, \varepsilon_1) = \frac{\tau_0(T_0)}{\tau_0(T_1)} \exp[\frac{G(T_0)}{\varepsilon_0} - \frac{G(T_1)}{\varepsilon_1}] \tag{2.56}$$

where T_0 and ε_0 are temperature and electric field of environment 0, T_1 and ε_1 are temperature and electric field of environment 1.

AF of E Model

$$AF_E(T_0, \varepsilon_0; T_1, \varepsilon_1) = \exp[-\gamma(\varepsilon_0 - \varepsilon_1)] \exp[\frac{E_a}{k}(\frac{1}{T_0} - \frac{1}{T_1})] \tag{2.57}$$

where T_0 and ε_0 are temperature and electric field of environment 0, T_1 and ε_1 are temperature and electric field of environment 1.

2.4 Negative Bias Temperature Instability

2.4.1 Introduction

NBTI occurs to p-channel MOS (PMOS) devices under negative gate voltages at elevated temperatures. Bias temperature stress under constant voltage (DC) causes the generation of interface traps (N_{IT}) between the gate oxide and silicon substrate, which translate to device threshold voltage (V_t) shift and loss of drive current (I_{on}). The NBTI effect is more severe for PMOSFETs than NMOSFETs due to the presence of holes in the PMOS inversion layer that are

known to interact with the oxide states. The degradation of device performance is a significant reliability concern for today's ultrathin gate oxides where there are indications that NBTI worsens exponentially with thinning gate oxide. NBTI has been studied and modeled since the 1960s [77]. Deal [78] named it "Drift VI" and discussed the origin in the study of oxide surface charges. Goetzberger et al. [79] investigated surface state change under combined bias and temperature stress through experiments utilizing MOS structures formed by a variety of oxidizing, annealing, and metalizing procedures. They found that an interface trap density D_{it} peak in the lower half of the band gap and p-type substrates gave higher D_{it} than n-type substrates. The higher the initial $D_{it,}$ the higher the final stress-induced $D_{it.}$ Jeppson et al. [80] were the first to propose a physical model to explain the surface trap growth of MOS devices subjected to negative bias stress. The surface trap growth was described as diffusion controlled at low fields and tunneling limited at high fields. The power law relationship ($t^{1/4)}$ was also proposed for the first time. Study of NBTI has been very active in recent years since the interface trap density induced by NBTI increases with decreasing oxide thickness, which means NBTI is more severe with ultrathin oxide devices. New developments of NBTI modeling and surface trap analysis have been reported in recent years. At the same time, effects of various process parameters on NBTI have been studied to minimize the NBTI. Schroder et al. [81] did an extensive review of those models and the effects of manufacturing process parameters. In this report, the failure mechanism, models, and related parameters of NBTI will be briefly discussed.

2.4.2 NBTI Failure Mechanisms

Silicon dioxide, the critical component of silicon devices, serves as insulation and passivation layers and is never completely electrically neutral. Mobile ionic charges, oxide trapped electrons or holes, fabrication-process-induced fixed charges, and interface trapped charges are four main categories of charges inside oxide and at the silicon–oxide interface. The electrical characteristics of a silicon device are very sensitive to the density and properties of those charges. As is already known, the threshold voltage of a p-channel MOSFET is given by:

$$V_{TH} = V_{FB} - 2\phi_B - |Q_B|/C_{ox} \tag{2.58}$$

where $\Phi_B = (kT/q)\ ln(N_D/n_i)$, $|Q_B| = (4_{q\varepsilon Si}\ \Phi B\ N_D)^{1/2}$, and C_{ox} is the oxide capacitance per unit area. The flat band voltage is given by:

$$V_{FB} = \phi_{MS} - \frac{Q_f}{C_{ox}} - \frac{Q_{it}(\phi_s)}{C_{ox}} \tag{2.59}$$

where Q_f is the fixed charge and Q_{it} is the interface trapped charge. From Equation (2.58) and Equation (2.59) it can be found that the only parameters to change the threshold voltage are Q_f and Q_{it}. Most of the NBTI failure mechanism research has been focused on the generation of Q_{it}.

Interface Trap Generation: Reaction-Diffusion Model

Jeppson and Svensson [80] were the first to propose the reaction-diffusion model to explain the generation of interface states at low fields. In their model, it is assumed that the silicon interface contains a large number of defects. Those defects are electrically inactive and can be activated through chemical reaction like this:

$$(surface\ defect) \rightleftharpoons (surface\ trap) + (surface\ charge)^{+} + X_{interface} + e^{-} \tag{2.60}$$

$$X_{interface} \rightleftharpoons^{diffusion} X_{bulk} \tag{2.61}$$

where $X_{interface}$ is a diffusing species that is formed at the interface in the reaction. Based on the infrared measurements report that showed large numbers of $Si{-}H$ groups existed in bulk silicon and probably also at the interface, Jeppson et al. proposed this reaction:

$$Si_3 \equiv SiH + O_3 \equiv SiOSi \equiv O_3 \rightleftharpoons Si_3 \equiv Si\cdot + O_3 \equiv Si^{+} + O_3 \equiv SiOH + e^{-} \tag{2.62}$$

where $Si_3 \equiv SiH$ is the surface defect, $Si_3 \equiv Si$ is the surface trap, $O_3 \equiv Si+$ is the oxide charge, and $O_3 \equiv SiOH$ is the diffusing X. When the defect is activated, the H of SiH bond is released by some

dissociation mechanisms and reacts with the SiO_2 lattice to form an *OH* group bonded to an oxide atom, leaving a trivalent *Si* atom in the oxide to form a fixed charge and one trivalent *Si* atom at the *Si* surface to form an interface trap. This chemical reaction is schematically shown in Figure 2.12. The number of interface traps relationship, $N_{it} \sim t^{1/4}$ was observed and mathematically proven by assuming the process is diffusion limited rather than reaction-rate limited.

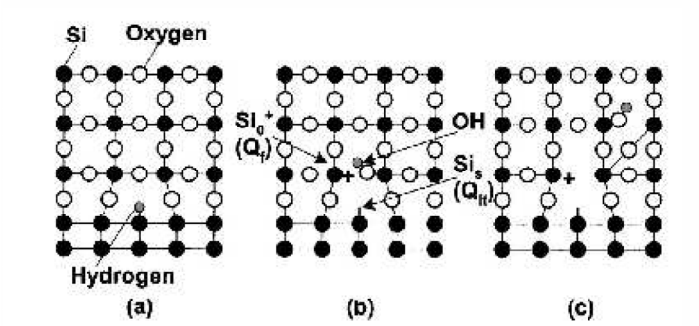

Fig. 2.12. *Schematic two-dimensional representation of the Si−SiO_2 interface, showing (a) the* ≡ S_i *defect, (b) how this defect may be electrically activated during NBTI to form an interface trap, a fixed oxide charge, and a hydroxyl group, and (c) the* OH *diffuses through the oxide. Adapted from [80].*

Various mechanisms have been proposed for the dissociation process. Ogawa et al. [82] listed three of those:

1. High-electric field dissociation

   $Si_3 \equiv SiH - \rightarrow Si_3 \equiv Si + H_i$, where H_i, the neutral species *X*, is an interstitial hydrogen atom.

2. Interstitial atomic hydrogen attack $Si_3 \equiv SiH + H_i - \rightarrow Si_3 \equiv Si + H_2$, molecular hydrogen H_2 is the species *X*.

3. Dissociation involves holes

   $Si_3 \equiv SiH + H+ \rightarrow Si_3 \equiv Si + H+$.

The actual diffusing species X have not yet been identified. Possibilities include interstitial atomic hydrogen (H_i) [82], molecular hydrogen (H_2), hydroxyl (*OH*) group, and proton (*H*+) [83].

Rashkeev et al. [83] did first-principles calculations to show proton is the only stable charge state of *H* at the $Si{-}SiO_2$ interface. The protons can react directly with *SiH* to form H_2 and leave behind positively charged dangling bonds.

Fixed Charge Generation

The fixed charge Q_f is a positive charge in the oxide and near the $Si{-}SiO_2$ interface. It cannot be charged or discharged by varying the silicon surface potential, and is primarily due to excess silicon species introduced during oxidation and during postoxidation heat treatment [77]. Ogawa et al. [82] determined fixed oxide charge densities from capacitance-voltage measurements and interface trap densities from conductance measurements of MOS capacitors under low field stress range from −1.6 to −5.0 MV/cm. The formulated expressions for the number of fixed and interface traps, N_f and N_{it}, are:

$$\Delta N_f(E_{ox}, T, t) = BE_{ox}^{1.5} t^{0.14} exp(-0.15/kT) \tag{2.63}$$

$$\Delta N_{it}(E_{ox}, T, t, t_{ox}) = CE_{ox}^{1.5} t^{0.25} exp(-0.2/kT)/t_{ox} \tag{2.64}$$

where *B* and *C* are two constants that are independent of E_{ox}, *T*, and T_{ox}. The thickness of oxide in their experiments ranged between 4.2 to 30 nm; together with an early report [77] that stated no thickness dependence of fixed charges for 40 through 100 nm oxides, showed that ΔN_f is independent of oxide thickness across a wide range. ΔN_{it} is inversely proportional to oxide thickness as is shown in Equation (2.64), which means NBTI is worse for thinner oxides.

Saturation and Recovery

Recent results indicate NBTI shifts tend to saturate over time [81, 84]. One possible reason is the reaction limitation mechanism; the generation of *Si*+ decreases as the number of available *SiH* bonds reduces with time. Threshold voltage shifts as a result of NBTI also tend to recover over

time after annealing [80]. This means that NBTI might exhibit different characteristics, depending on whether the device is operating at AC or DC. For example, Abadeer et al. [85] reported a threefold increase in the magnitude of threshold voltage shift under DC operation, compared to the shift under AC operation. In another NBTI experiment, Rangan [86] showed that recovery is independent of stress voltage, time, and temperature (under 25°C), but can reach 100% at 25°C. The mechanism of recovery is still under investigation and there is much work in this area. One explanation is that the diffusion species, X, moves back to the $Si-SiO_2$ interface under the influence of positive gate voltage and passivates the Si-dangling bond [87].

2.4.3 NBTI Models

The time dependence of the threshold voltage shift (ΔV_{TH}) is found to follow a power-law model

$$\Delta V_{TH}(t) = At^n \tag{2.65}$$

where A is a constant that depends on oxide thickness, field, and temperature. The theoretical value of the time dependence parameter n is 0.25 according to the solution of diffusion equations [80]. Reported value of n is in the range from 0.2 to 0.3. According to Chakravarthi [84], the values of n vary around 0.165, 0.25, and 0.5 depending on the reaction process and the type of diffusion species. The temperature dependence of NBTI follows the Arrhenius law with activation energies ranging from 0.18 to 0.84 eV [88, 89].

Improved models have been proposed after the simple power-law model. Considering the temperature and gate voltage, ΔV_{TH} can be expressed as

$$\Delta V_{TH}(t) = A\exp(\beta V_G)\exp(-E_\alpha/kT)t^{0.25} \tag{2.66}$$

where A and β are constants and V_G is the applied gate voltage. Liu et al. [90] proposed a new model considering the reversible reaction and diffusion processes,

$$\Delta V_{TH}(t) = B_1[1 - \exp(-t/\tau_1)] + B_2[1 - \exp(-t/\tau_2)] \tag{2.67}$$

where B_i represents the total number of trivalent silicon bonds in reaction (i) and τ_i is time constant for reaction (i), which considers both the forward and the reverse reaction.

3 FAILURE RATE BASED SPICE (FARBS) RELIABILITY SIMULATION

3.1 Introduction

The FaRBS (Failure Rate Based SPICE [spacecraft, planets, instrument, C-matrix, events]) reliability simulation tool provides microelectronic chip designers and device reliability engineers a method to quantify product reliability, as well as to make accurate performance and reliability tradeoffs in the product-development stage. Chip designers and device reliability engineers utilize SPICE circuit simulators to determine device real-time operating parameters, then employ state-of-the-art semiconductor wearout failure models to estimate a device's lifetime or failure-in-time (FIT) value. These wearout failure models are based on the same principles as those for accelerated lifetime tests; therefore, their model parameters can be derived from stress experiments. By assuming that all failures will be random and scalable and that devices have no dominating failure mode by design, system designers can apply the sum-of-the-failure-rate model and the improved acceleration factor and FIT model to calculate the overall mean-time-to-failure (MTTF)/FIT values of the systems. Finally, based on the above FaRBS simulation results, more accurate performance and reliability trade-offs can be made to guide the development of long-life or high-reliability electronic systems. Chip designers and device reliability engineers do not have the transistor-level SPICE netlist, they can use SPICE macro models of common circuit blocks (Op Amp, static random access memory [SRAM] blocks, field programmable gate array [FPGA], analog-to-digital converter / digital-to-analog coverter [ADC/DAC], and other digital logic blocks) from reliability libraries supplied by FaRBS to do similar reliability simulation. SPICE macro models can be simulated and improved with FaRBS to include value-added reliability parameters.

3.2 Modules and the Process of FaRBS

The modules and process of FaRBS are shown in Figure 3.1. The function of each constituent module is summarized below.

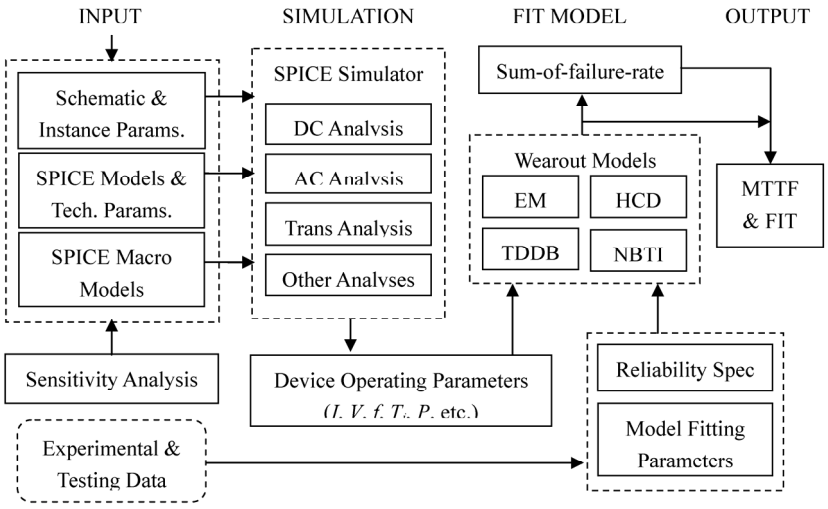

Fig. 3.1. *Modules and data flows in FaRBS.*

3.2.1 Sensitivity Analysis

It is a significant amount of work for large-scale electronic-system designers to perform full-scale SPICE simulation due to the ever-increasing complexity of the circuits. Oftentimes, transistor-level schematics and SPICE netlists for commercial parts are considered proprietary, which also prohibits the full-scale transistor-level SPICE simulation. From a functional perspective, not every block contributes equally to system failure. In most mixed-signal applications, the probability of failure from some specific blocks is an order of magnitude higher than others. In this case, hardware failure sensitivity analysis can be employed to help identify the critical or failure-prone blocks in the system. System designers only need to carry out transistor-level SPICE simulation for these identified blocks. Other blocks can be characterized by behavior-level SPICE macro models, which are normally provided by device suppliers and easy to access. Hardware failure sensitivity analysis at the transistor-level schematic can be performed by selecting nodes in the circuit according to failure distribution and by injecting appropriate faults using fault-injection algorithms at these nodes. Some wearout failures leading to increased leakage current can be modeled as current sources at corresponding nodes. The theoretical basis for sensitivity analysis can be found in [91]. Basically, there are two environments suitable for hardware failure sensitivity analysis: one is SPICE simulation [91], the other is VHDL modeling

[92]. The fault injection models and algorithms for sensitivity analysis regarding those wearout failure mechanisms are under development.

3.2.2 SPICE Simulation

As CMOS scaling pushes device performance to technological limits and diminishes reliability margins, performance and reliability tradeoffs based on the worst-case DC stress analysis are no longer accurate and will result in more stringent design constraints and increased cost. In real operating conditions, devices experience time-variant voltage or frequency stresses. The effective AC stress time during which the transistor undergoes real stresses is only a fraction of worst-case DC stress time. If a lifetime factor is defined as the ratio of worst-case DC age to AC age, in a 100-MHz application, the lifetime factors of the n-channel metal oxide semiconductor field effect transistor (NMOSFET) and the p-channel MOSFET (PMOSFET) reported in [93] are 120 and 300, respectively. Therefore, to accurately model device lifetime, real-time operating parameters should be determined from SPICE.

In FaRBS, the process for SPICE simulation is fairly straightforward. After the sensitivity analysis, the transistor-level schematics or SPICE netlists of reliability critical blocks and the behavior-level SPICE macro models are fed into the SPICE simulator. MOSFET models and technology files are then selected according to the system specifications. The DC analysis of SPICE determines the DC operation point of the circuit. This is automatically performed prior to a transient analysis to determine the transient initial conditions and prior to an AC small-signal analysis to determine linearized, small-signal models for nonlinear devices. The AC small-signal analysis of SPICE computes the AC output variables as a function of frequency. The transient analysis of SPICE computes the transient output variables as a function of time over a user-specified time interval [94]. The outputs of these SPICE analyses are device real-time operating parameters, such as voltage, current, and frequency.

3.2.3 Wearout Models

Wearout models for deep submicron CMOS devices, including electromigration (EM), hot carrier injection (HCI), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI), have been discussed. The model equations are recapitulated here.

1. EM

$$MTTF_{EM} = A_{EM}(J \times T)^{-2} \exp(\frac{E_{aEM}}{kT}) \tag{3.1}$$

where A_{EM} is an empirically determined constant, E_{aEM} is the activation energy of EM, and J is the current density flowing the interconnects. J can be determined by SPICE simulation with,

$$J = \frac{CV_{dd}}{WH} \times f \times P_r \tag{3.2}$$

where C is the parasitic capacitance, W and H are the width and thickness of a metal line, f is the clock frequency, and Pr is the probability that the line toggles in a clock cycle.

2. HCI

$$MTTF_{HCI} = A_{HCD} \exp(\frac{\theta}{V_{ds}}) \tag{3.3}$$

where A_{HCD} and θ are constants determined from lifetime testing and V_{ds} is the drain-to-source voltage. This simple equation is only valid for a small range of voltages near the maximum substrate current.

3. TDDB

$$MTTF_{TDDB} = A_{TDDB}A_G(\frac{1}{V_{gs}})^{(\alpha-\beta T)} \exp(\frac{X}{T} + \frac{Y}{T^2}) \tag{3.4}$$

where V_{gs} is the gate voltage; T is the temperature; α, β, X, and i are fitting parameters; and A_{TDDB} is an empirically determined constant. A_G is the total gate oxide surface area.

4. NBTI

$$MTTF_{NBTI} = A_{NBTI}(\frac{1}{V_{gs}})^{\gamma}\exp(\frac{E_{aNBTI}}{kT}) \tag{3.5}$$

where A_{NBTI} is process-related constant, γ is voltage acceleration factor, and E_{aNBTI} is activation energy.

3.2.4 System Reliability Model

The lifetime of each wearout failure mechanism for each interconnect and MOSFET in a circuit can be determined by Equations (3.1), (3.3), (3.4), and (3.5). To obtain the lifetime for the entire circuit, we need to combine the effects of these different wearout mechanisms across different structures. The system reliability can be estimated by the competing failure model, as follows:

$$MTTF_s = \frac{1}{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{MTTF_{ij}}} \tag{3.6}$$

where $MTTF_s$ is the lifetime of a circuit composed of n units and $MTTF_{ij}$ is the lifetime of each failure mechanism for each unit.

3.3 Parameter Extraction Model

How to accurately determine those model parameters in wearout models is an essential part of FaRBS simulation method. Parameter extraction always involves tedious processes; therefore, some commercial tools have been developed to facilitate this extraction work. Sometimes, there will be a distinct discrepancy in the extracted parameter values, depending on the extraction methodology employed. System designers might have to set up complex accelerated stress experiments to calibrate and model parameters per system specifications. One of the most efficient and accurate model-extraction tools on the market is BSIMProPlus from Cadence, which offers

active and passive device modeling solutions for various process technologies. BSIMProPlus supports lifetime parameter extraction for HCI and NBTI reliability modeling. The stressing and data collection for NBTI and HCI are automated through simultaneous control of various types of measurement hardware. BSIMProPlus provides real-time stress status; monitors device degradation during stress through periodic measurement of MOSFET threshold voltage, transconductance, gate, and bulk current characteristics; measures and saves I-V characteristics during stress; and makes data files available for viewing. It accommodates two lifetime extraction models (substrate-current and gate-current models) and enables links to circuit reliability and full-chip reliability simulators. The detailed description and application of BSIMProPlus can be found in [95].

3.4 Derating Voltage and Temperature for Reliability

Technology is mainly driven by a few fast-moving markets, such as wireless communication systems and entertainment electronics, wherein devices are customized and fabricated to explore their performance to the limits, sometimes by sacrificing reliability. As a result, most Commercial-off-the-Shelf (COTS) devices currently available in the market have high performance but short lifetimes, which might limit the lifetimes of the systems in long-life applications if such COTS devices are incorporated. The lifetime models developed in the previous chapters offer a way to address this problem: operating devices at reduced voltage, frequency and temperature levels than their original ratings, i.e., derating. In CMOS circuits, power dissipation is determined by frequency, voltage and temperature. Reduction in voltage will significantly reduce the power dissipation; similarly, reduction in frequency and temperature will also lead to a corresponding reduction in power dissipation. There is a positive relationship between the peak power dissipation of CMOS digital circuits and many wearout mechanisms; consequently, even though derating voltage, frequency, and temperature does degrade the performance of a device, it also reduces the physical stresses on the device, thereby increasing its expected useful life [96, 97].

Derating voltage and temperature for reliability improvement is one of the major applications of FaRBS. Based on lifetime models of failure mechanisms, if all model parameters are calibrated from testing, FaRBS can accurately predict device and circuit failure rate and characterize circuit

derating behaviors under different voltage and temperature stresses through SPICE simulation. Through the introduction of a unified derating factor, FaRBS simulation is capable of formulating practical derating design guidelines for improving product lifetime and reliability in long-life applications.

3.4.1 Circuit Design and Simulation

A 17-stage ring oscillator consisting of CMOS inverters and interconnecting capacitors is simulated in order to investigate voltage and temperature derating effects. A CMOS ring oscillator has been widely used as a test circuit for monitoring process variations and characterizing reliability behaviors because its oscillating frequency is sensitive to SPICE model parameters. Figure 3.2 shows the schematic diagram of the 17-stage ring oscillator. The BSIM3v3 model is used to characterize the MOSFETs Q_n and Q_p, and the model parameters are taken from a TSMC 0.18-μm CMOS process. The TSMC 0.18-μm CMOS process supports 1.8-V and 3.3-V applications. 1.8-V technology is widely used in general purpose and low-power design, while 3.3-V technology is used for high-quality mixed-signal or RF devices. In the following simulation, the rated value of power supply voltage V_{DD} is selected as 3.3 V to set a wider voltage derating range. To obtain symmetrical transfer characteristics, the device gate widths (W_n of Q_n and W_p of Q_p) are designed to follow the well-known relationship:

$$\frac{W_p}{W_n} = \frac{I_n}{I_p} = \frac{\mu_n}{\mu_p} \tag{3.7}$$

The extracted values of electron and hole mobilities are μ_n = 263.8 cm^2/V and μ_p = 118.3 cm^2/V therefore, the gate geometries of Q_n and Q_p are designed to be $L_n = L_p$ = 0.35 μm, W_n = 1.10 μm, and W_p = 2.60 μm. The overall simulation is divided into three steps to investigate voltage scaling effects, temperature scaling effects, and DC transfer characteristics, respectively.

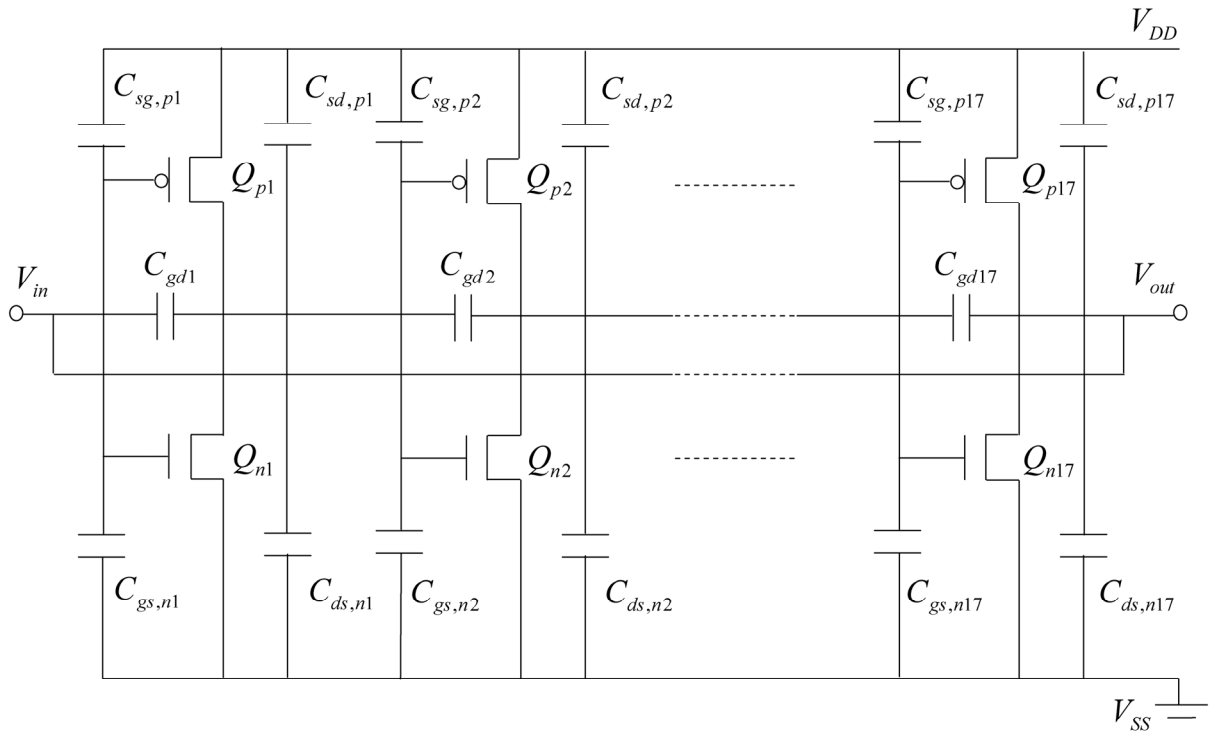

Fig. 3.2. *The schematic design of the ring oscillator, which consists of 17-stage CMOS inverters and interconnecting capacitors.*

3.4.2 Simulation Results and Analysis

Voltage Derating Analysis

The transient analysis is performed by sweeping the power supply voltage V_{DD} from 1.0 V to 4.0 V with incremental steps of 0.1 V in order to investigate voltage derating behaviors. The ambient temperature is set to 27°C. When V_{DD} is scaled, the oscillating frequency monotonically increases from 80.91 MHz to 418.5 MHz.

For a CMOS inverter, if the pull-down delay τ_n of an NMOSFET is defined as the time for the output voltage to decrease from V_{DD} to $V_{DD}/2$, then τ_n can be expressed as:

$$\tau_n = \frac{CV_{DD}}{2I_{Nsat}} = \frac{CV_{DD}}{\mu_{neff}C_{ox}(W/L)(V_{DD}-V_{tn})^2} \tag{3.8}$$

where V_{tn} is the threshold voltage, μ_{neff} is the electron effective mobility, W is the channel width, L is the channel length, and C is the output loading capacitance. When $V_{DD} >> V_{tn}$, τ_n is approximately proportional to the inverse of V_{DD}. The similar expression can be derived for the pull-up delay τ_p of PMOSFET. The transition delay τ of the CMOS inverter is the arithmetic average of τ_n and τ_p (i.e., $\tau = (\tau_n+\tau_p)/2$). Therefore, when V_{DD} is scaled down in proper range, the operating frequency of the ring oscillator will decrease proportionally. The power consumption of CMOS circuits mainly comes from switching periods in dynamic operation because their static power dissipations are negligible; the total average power consumption P_D can be estimated as:

$$P_D = \frac{1}{T} C_L V_{DD}^2 = C_L V_{DD}^2 f \tag{3.9}$$

where C_L is the total loading capacitance on the chip, and f is the frequency at which the circuit switches [98]. Figure 3.3 is the simulation result of frequency and power dissipation derating trends with respect to V_{DD}. Both Equation (3.9) and simulation results show that voltage derating significantly affects power dissipation. When voltage increases 4 times, the frequency increases approximately 5 times, whereas the power dissipation increases up to 100 times. Thus, the net result of the dependence of the power dissipation on the voltage is much stronger than a simple quadratic relationship.

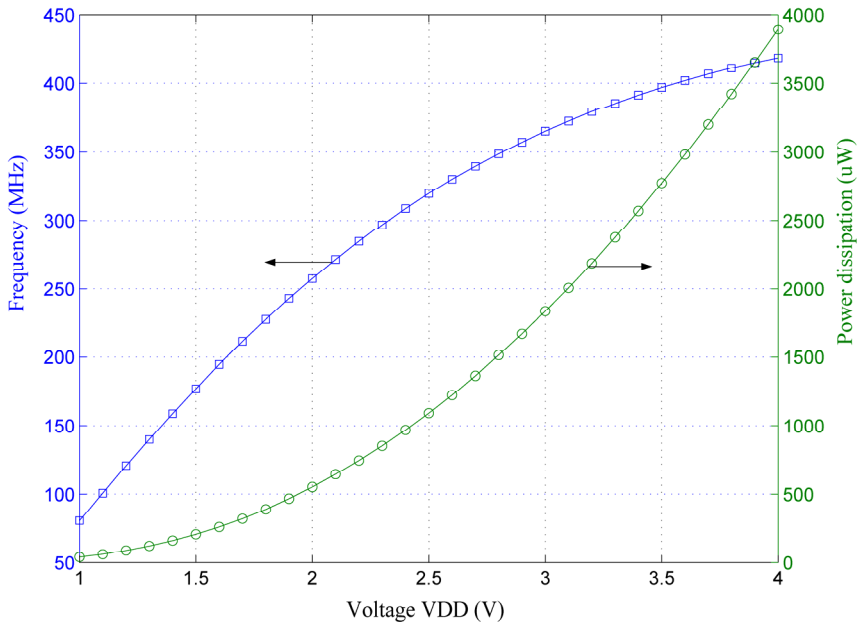

Fig. 3.3. *The derating relationship of frequency and power dissipation vs. V_{DD}. When voltage increases 4 times, the frequency increases approximately 5 times, whereas the power dissipation increases approximately 100 times.*

The above voltage derating analysis is based on a quasi-static assumption that the device response is quick enough compared with the switching speed of its terminal voltage. This assumption is valid only when the input signal's rise and fall times are much longer than the carrier transit time across the channel. For very short channel devices, the carrier transit time τ_{tran} is determined by the carrier saturation transit time $\tau_{sat} = L/\upsilon_{sat}$ or average transit time $\tau_{avg} = L^2/(\mu_{eff} V_{DD})$, whichever is larger. For the 0.18-µm NMOSFET SPICE parameters, $\tau_{sat} = 3.9$ ps. When V_{DD} is greater than 1.2 V, the electron average transit time across the channel τ_{avg} is smaller than τ_{sat}; therefore, during the whole range of voltage derating (from 4.0 V down to 1.0 V), the device response time (i.e., τ_{tran}) is determined by τ_{sat} and, as a result, keeps constant. The simulated minimum switching delay of terminal voltage for the CMOS inverter is much larger than τ_{tran}. This means the quasi-static assumption is held for the above simulation and the voltage derating behaviors in light of frequency and power dissipation, given by Figure 3.3, are valid.

The above voltage derating and timing response analysis formulate a guideline for setting proper lower-bounds of V_{DD} in some special applications. For high frequency applications where switching delay is comparable to τ_{tran}, and in some mixed-signal applications where long channel devices coexist with short channel devices, τ_{tran} will be greater than the switching delay of device terminal voltage if V_{DD} is derated below some critical value. Thus, the quasi-static assumption would no longer be valid. In these situations, a non-quasistatic model should be incorporated in the simulation to account for possible new voltage, derating behaviors.

Temperature Derating Analysis

Temperature is another controllable and reliability-sensitive design parameter because a number of important device parameters, such as mobility, threshold voltage, and saturation velocity, are temperature dependent. To determine the temperature derating behaviors of frequency and power dissipation, the temperature transient analysis of the same ring oscillator is performed by sweeping the temperature from 0°C to 150°C with steps of 10°C. V_{DD} is set to 3.3 V during the process.

Carrier mobility is a well-known temperature-dependent parameter. Phonon scattering, surface scattering, and impurity scattering are major scattering mechanisms governing the characteristics of carrier mobility; these mechanisms follow different temperature dependencies. At low temperature, impurity scattering dominates and the mobility increases with rising temperature; while at high temperature, phonon scattering starts to prevail and the mobility will decrease and follow the trend $\mu_{eff} \propto T^{-3/2}$. These competing temperature effects result in a non-monotonic dependence of the mobility on temperature and lead to the existence of a maximum carrier mobility value [99]. According to the discussion in the above section, for long-channel devices operated at very low V_{DD}, the carrier transit time τ_{tran} sets device switching speed and is determined by $\tau_{avg} = L^2/(\mu_{eff} V_{DD})$, in which carrier mobility μ_{eff} is the only temperature-dependent factor. Therefore, the derating relationship between temperature and frequency of long channel devices operated at low voltage is mainly governed by μ_{eff}. Due to the aforementioned non-monotonic dependence of the mobility on temperature, in the derating curve of frequency vs. temperature,

there should exist a maximum frequency value at which the mobility is maximal and has a relatively weak temperature sensitivity.

If the device channel length is very short and V_{DD} is high, device operating speed is determined by the interconnecting and parasitic capacitances (refer to Equation (3.8)). In the BSIM3v3 model, parasitic capacitances are temperature dependent but not in linear relation; therefore, device operation frequency will demonstrate nonlinear behavior when temperature is derated. The relations of frequency and power dissipation vs. temperature in this case are plotted in Figure 3.4, which shows a minimum frequency value at temperature 120°C and, therefore, demonstrates a different behavior from that of long-channel devices.

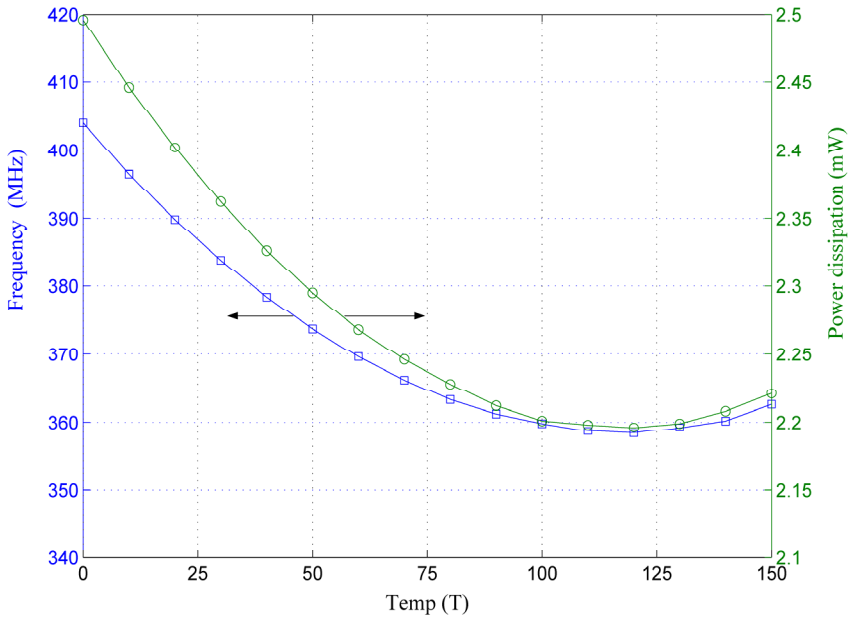

Fig. 3.4. *The derating curves of frequency and power dissipation vs. temperature. Operation frequency and power dissipation follow nonlinear trends when temperature is derated and simulation shows a minimum frequency value at approximately 120°C.*

Temperature derating behavior around these maximum or minimum frequency values has an interesting implication in the process of derating temperature for reliability. Simulation identifies relatively flat regions around these extreme value points, so temperature derating within these flat

regions will cause little variation in circuit speed and power consumption, which simplifies performance and reliability tradeoffs.

The above analysis formulates a guideline for effectively derating temperature for the sake of reliability improvement. The flat region of temperature derating curves must be properly identified. Otherwise, short channel devices might not obtain lifetime enhancement even though temperature is derated, and long channel devices might lose the potential to gain lifetime extension by scaling temperature without sacrificing performance.

Threshold voltage, saturation velocity and parasitic drain and source resistances are other important parameters that are sensitive to temperature. Threshold voltage V_t increases as temperature decreases due to the shifts of Fermi level and bandgap energy. Saturation velocity υ_{sat} is determined by the critical field and carrier effective mobility μ_{eff}, thereby also varying with temperature. Although μ_{eff} has complicated temperature dependency, υsat is actually a weak function of temperature and usually demonstrates a simple dependence on temperature: υsat decreases as temperature increases [100]. Parasitic drain/source series resistance R_{ds} consists of contact resistance, drain and source diffusion sheet resistance, and spreading resistance resulting from current crowding at the edge of the inversion layer. In the BSIM3v3 model, V_t, υ_{sat} and R_{ds} are all modeled with linear relations to temperature [101].

Derating temperature alone does not influence device performance as much as derating voltage; but reducing temperature and voltage together will produce an order of magnitude reliability improvement. This significant improvement results from the modification of device junction temperature $T_{j,}$ which is dependent on the power dissipation P_D, the ambient temperature T_a, and the thermal impedance θ_{ja}:

$$T_j = \theta_{ja} P_D + T_a \tag{3.10}$$

The dependence of T_j on V_{DD} is given by:

$$T_j = T_a - \frac{V_{DD}(V_{DD} - V_t)^2(T_a - T_j^0)}{V_{DD}^0(V_{DD}^0 - V_t)^2} \tag{3.11}$$

where V^0_{DD} and T^0_j denote normal operating values for voltage and junction temperature, V_{DD} and T_j represent derated values for voltage and junction temperature, V_t is threshold voltage, and T_a is the ambient temperature. Each of these parameters can be controlled in circuit design [102]. Temperature derating does provide an alternative to improve device reliability; however, the above temperature derating behaviors are only valid within the temperature range of −50°C to 150°C due to the SPICE model limitation. Beyond this range, complicated scattering mechanisms start to dominate and significantly change the temperature behaviors of low V_{DD} devices; consequently, an additional temperature derating model will be required to characterize any new temperature behaviors [101].

Voltage Transfer Analysis

Digital integrated circuits consist of various kinds of interconnected logic gates; the voltage signals are always contaminated by noise. To characterize the noise tolerance or immunity of a circuit to undesired external perturbations, designers normally need to explore and properly set the noise margin parameter that is the difference of equivalent voltage levels between output and input of consecutive gates. Noise magnitude must be within the noise margin to make logic gates work at the correct input and output voltage levels. There are two noise margin parameters: $NM_L = V_{IL} - V_{OL}$ for low signal levels, and $NM_H = V_{OH} - V_{IH}$ for high signal levels, where V_{IL} is input low voltage, V_{IH} is input high voltage, V_{OL} is output low voltage, and V_{OH} is output high voltage. These parameters characterize the DC input-output voltage behaviors and determine the circuit noise tolerance to external signal perturbations. Setting proper values for these noise margins is a basic design consideration for realizing intended functions and enabling correct voltage derating.

The simulation results for the two noise margin parameters NM_L and NM_H vs. V_{DD} are plotted in Figure 3.5, which shows that over the voltage derating range of 4.0 V to 1.2 V, NM_L and NM_H

approximately decrease linearly with V_{DD}. Therefore, derating does not change the ratios of noise margin to voltage.

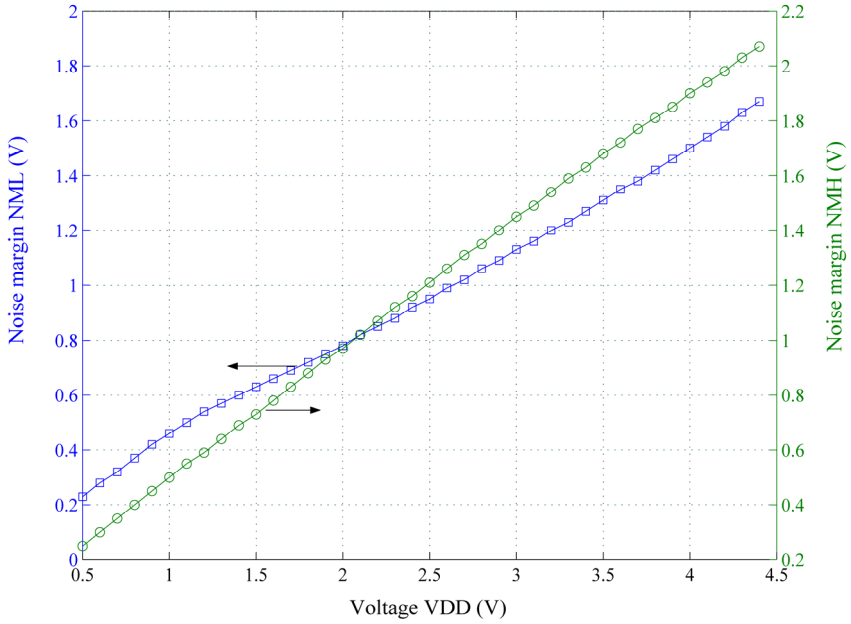

Fig. 3.5. *The simulation results for* NM_L *and* NM_H *vs.* V_{DD}. *Over the voltage derating range of 4.0 V to 1.2 V,* NM_L *and* NM_H *approximately decrease linearly with* V_{DD}.

Sufficient noise margin is very important for a circuit in severe environments where noise can corrupt the circuit signals. Figure 3.5 shows that when V_{DD} is very small, noise margins will decrease to very low levels. Therefore, in low-power applications where noise is ubiquitous, noise margin might impose lower limits on voltage derating. Nevertheless, the frequency can be decreased more than what is required only by voltage reduction to reduce the noise sensitivity, and a derated device can have greater noise tolerance than its full performance specification. Figure 3.6 is the plot of DC voltage transfer characteristics (VTC) under different input voltage dynamic range (from 0.5 V to 4.5 V). For an ideal CMOS inverter, the output dynamic range is from 0 to V_{DD}. When V_{DD} scales down, it is obvious that the width of the uncertain region (i.e., transition region) of VTC reduces proportionally. Reducing the width of the uncertain region is one of the most important design objectives for lowering power and boosting speed; however, there exists a

limit for excessively reducing the width of this transition region due to the MOSFET threshold voltage requirements.

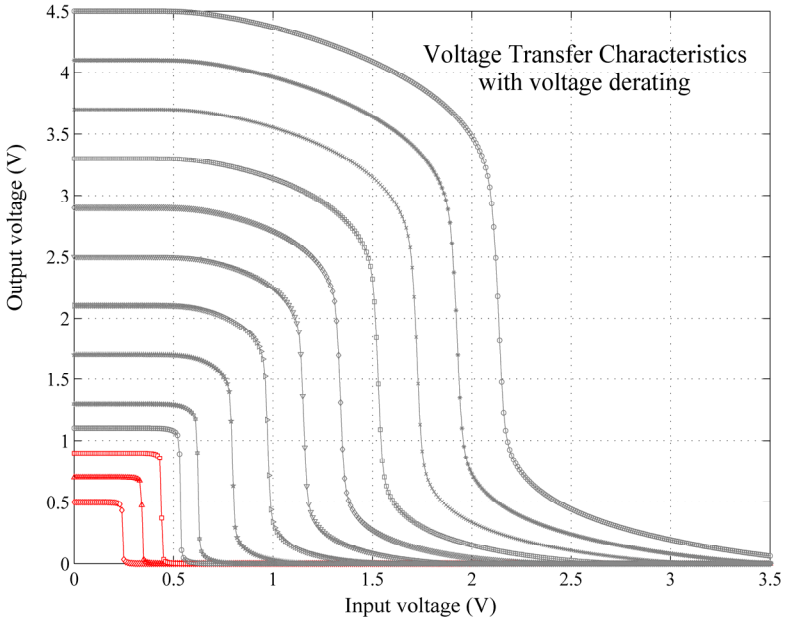

Fig. 3.6. *DC VTC curves at different power supply voltage (from 0.5 V to 4.5 V). When V_{DD} scales down, the transition region of VTC reduces proportionally. When V_{DD} is lower than 0.8 V, the transition region disappears, and the VTC exhibits a hysteresis behavior.*

Figure 3.6 shows that when V_{DD} is lower than 0.8 V, the transition region disappears and V_{IL} quickly approaches V_{IH}. In this case, the inverter will operate with a region wherein none of the transistors is conducting. This means the inverter can no longer function correctly. In theory, the lower limit for V_{DD} is bounded by the summation of NMOSFET and PMOSFET threshold voltages:

$$V_{DD}(min) = V_{tn} + |V_{tp}| \tag{3.12}$$

For the 0.18-μm technology being considered, V_{tn} = 0.368 V and V_{tp} = 0.435 V, their summation is approximately 0.8 V; the simulation result, therefore, conforms to the theory. When V_{DD} is lower than V_{DD} *(min)*, the VTC will contain a cut-off region. The output voltage within this region will maintain its previous state due to the charge preservation at the output node. Thus, the inverter VTC exhibits a hysteresis behavior at very low supply voltages [97].

Derating Model and Derating Factor

Designers may inappropriately draw a conclusion from the above simulation that the voltage and temperature derating margin is very large, when in fact, designers do not have that much flexibility because they must adhere to the published device specifications. When MOS devices go down to deep submicron dimensions, the nominated supply voltages are also lowered to subdue electric fields; e.g., in 0.13-μm technology, the supply voltage V_{DD} is as low as 1.2 V. According to the ITRS 2002 Update, for 90-nm technology, V_{DD} will be even lower than 1.0 V. However, threshold voltages have not been scaled down proportionately over technology generations, and some aforementioned mechanisms also impose lower limits on V_{DD}. Consequently, derating V_{DD} in valid ranges is not trivial. With technology advancement, an in-depth understanding of derating behaviors, accurate derating models, and practical design guidelines become more and more important for circuit designers.

The idea of derating for reliability originated with the principles of accelerated stress tests (ASTs), wherein devices are over-stressed to precipitate failures within a reasonably short time span, and then their reliability parameters are extrapolated back to normal operation conditions. Derating can be treated as a reverse application of AST; furthermore, designers need simple rules to derate devices for desired lifetime improvement without affecting performance. A new factor, derating factor D_f, as a counterpart to the acceleration factor in AST, is introduced. This factor is defined as the ratio of the MTTF of a device operating at derated conditions ($MTTF_d$) to its MTTF at rated operation conditions ($MTTF_0$):

$$D_f = \frac{MTTF_d}{MTTF_0} \tag{3.13}$$

D_f can represent the total effect of various wearout mechanisms at the circuit level. From the lifetime models presented in the previous chapters, it is easy to obtain the expressions of D_f for the four wearout mechanisms: HCI, TDDB, NBTI, and EM, respectively.

$$D_{f_{HCI}} = (\frac{I_{sub}^0}{I_{sub}})^n \exp[\frac{E_{aHCI}}{\kappa}(\frac{1}{T_j} - \frac{1}{T_j^0})] \tag{3.14}$$

$$D_{f_{TDDB}} = \frac{(V_{gs})^{a+bT_j}}{(V_{gs}^0)^{a+bT_j^0}} \exp[c(\frac{1}{T_j} - \frac{1}{T_j^0}) + d(\frac{1}{(T_j)^2} - \frac{1}{(T_j^0)^2})] \tag{3.15}$$

$$D_{f_{NBTI}} = (\frac{V_{gs}^0}{V_{gs}})^{\frac{1}{\beta}} [\frac{(1+2\exp(-\frac{E_1}{\kappa T_j^0}))^{-1} + (1+2\exp(-\frac{E_2^0}{\kappa T_j^0}))^{-1}}{(1+2\exp(-\frac{E_1}{\kappa T_j}))^{-1} + (1+2\exp(-\frac{E_2}{\kappa T_j}))^{-1}}]^{\frac{1}{\beta}} \tag{3.16}$$

$$D_{f_{EM}} = (\frac{J^0}{J})^n (\frac{T_j^0}{T_j})^m \exp[\frac{E_{aEM}}{\kappa}(\frac{1}{T_j} - \frac{1}{T_j^0})] \tag{3.17}$$

where I^0_{sub}, V^0_{gs}, J^0, E^0_2 and T^0_j denote rated operating values for nominal use conditions, while I_{sub}, V_{gs}, J, E_2, and T_j represent expected derated values. The above four individual derating factors are related to the total derating factor D_f with a function f_d:

$$D_f = f_d(D_{f_{HCI}}, D_{f_{TDDB}}, D_{f_{NBTI}}, D_{f_{EM}}) \tag{3.18}$$

The most important part of a derating model is to determine the function f_d. Derivation of the explicit expression for f_d is complicated and requires detailed information of circuit architecture and stress conditions. However, for a simple analysis, designers can assume that D_{fHCI}, D_{fTDDB},

D_{fNBTI}, and D_{fEM} are independent of each other; therefore, within small derating scales, f_d can be approximated with a linear relation:

$$f_d = C_{HCI}D_{f_{HCI}} + C_{TDDB}D_{f_{TDDB}} + C_{NBTI}D_{f_{NBTI}} + C_{EM}D_{f_{EM}} \tag{3.19}$$

where C_{HCI}, C_{TDDB}, C_{NBTI}, and C_{EM} are constants and their values can be determined from experimentation or simulation. When the derated condition is the same as the rated condition, there is no derating and the total derating factor D_f equals unity:

$$D_f = f_d(1,1,1,1) = 1 \tag{3.20}$$

Equation (3.20) indicates that the summation *of* C_{HCI}, C_{TDDB}, C_{NBTI}, and C_{EM} always equals unity for any derating process:

$$C_{HCI} + C_{TDDB} + C_{NBTI} + C_{EM} = 1 \tag{3.21}$$

From Equations (3.14)–(3.17), designers can determine D_{fHCI}, D_{fTDDB}, D_{fNBTI}, and D_{fEM} under any derated voltage and temperature conditions. If C_{HCI}, C_{TDDB}, C_{NBTI}, and C_{EM} are calibrated from simulation or testing work, the overall circuit derating factor D_f can be predicted from Equations (3.18) and (3.19).

Derating Factor and Simulation

For the purpose of obtaining knowledge on the derating factor and understanding its influence on circuit reliability improvement, the dependence of D_f on V_{DD} is simulated and a derating graph is generated within a reasonable scale.

Currently, there is no universally accepted lifetime distribution model for all device wearout mechanisms; however, failure information extracted from an avionics maintenance database has

been researched, and statistical analysis has been performed to obtain information relating to how circuits fail [103]. Overwhelming evidence points to an exponentially distributed failure pattern for aerospace circuits [102], prompting an assumption that circuit lifetime distribution is approximately exponential, no matter what the lifetime distribution is for each of the device wearout mechanisms.

A simple method to calculate C_{HCI}, C_{TDDB}, C_{NBTI}, and C_{EM} starts from an assumption that each wearout mechanism contributes equally to the total derating effect. This is a plausible assumption; otherwise, if any wearout mechanism is more significant than others, designers and manufacturers will develop techniques to attenuate its effect. A good example is the development of the lightly doped drain (LDD) structure for suppressing HCI effect. Upon this assumption, C_{HCI}, C_{TDDB}, C_{NBTI}, and C_{EM} will conform to the following ratios:

$$C_{HCI} : C_{TDDB} : C_{NBTI} : C_{EM} = \frac{1}{D_{f_{HCI}}} : \frac{1}{D_{f_{TDDB}}} : \frac{1}{D_{f_{NBTI}}} : \frac{1}{D_{f_{EM}}} \tag{3.22}$$

Combining Equations (3.21) and (3.22), we can easily find the expressions for C_{HCI}, C_{TDDB}, C_{NBTI}, and C_{EM} as follows:

$$C_{HCI} = \frac{\frac{1}{D_{f_{HCI}}}}{\frac{1}{D_{f_{HCI}}} + \frac{1}{D_{f_{TDDB}}} + \frac{1}{D_{f_{NBTI}}} + \frac{1}{D_{f_{EM}}}} \tag{3.23}$$

$$C_{TDDB} = \frac{\frac{1}{D_{f_{TDDB}}}}{\frac{1}{D_{f_{HCI}}} + \frac{1}{D_{f_{TDDB}}} + \frac{1}{D_{f_{NBTI}}} + \frac{1}{D_{f_{EM}}}} \tag{3.24}$$

$$C_{NBTI} = \frac{\frac{1}{D_{f_{NBTI}}}}{\frac{1}{D_{f_{HCI}}} + \frac{1}{D_{f_{TDDB}}} + \frac{1}{D_{f_{NBTI}}} + \frac{1}{D_{f_{EM}}}} \tag{3.25}$$

$$C_{EM} = \frac{\frac{1}{D_{f_{EM}}}}{\frac{1}{D_{f_{HCI}}} + \frac{1}{D_{f_{TDDB}}} + \frac{1}{D_{f_{NBTI}}} + \frac{1}{D_{f_{EM}}}} \tag{3.26}$$

Substituting Equations (3.23)–(3.26) into (3.18) and (3.19), we conclude with a very simple derating factor model:

$$D_f = \frac{4}{\frac{1}{D_{f_{HCI}}} + \frac{1}{D_{f_{TDDB}}} + \frac{1}{D_{f_{NBTI}}} + \frac{1}{D_{f_{EM}}}} \tag{3.27}$$

The voltage derating trends governed by this simple D_f model are simulated with typical model parameters from the 0.18-μm technology. V_{DD} is derated within the range [100%–80%] of its rated value (i.e., V^0_{DD} = 3.3 V). Figure 3.7 is the plotting of the relation between D_f and V_{DD}/ V^0_{DD}, which shows that within the derating range, the dependency of D_f on V_{DD}, after being normalized to V^0_{DD}, is an exponential relation. Figure 3.7 also indicates that the variations of individual derating factors are different by up to three orders of magnitude. Another derating factor for the lower-rated voltage V^0_{DD} = 1.8 V is also plotted in Figure 3.7. These two derating factors at different rated voltages almost follow the same trend, which reveals that no matter what the rated voltage is, if voltage is derated to the same ratio, the reliability gain is nearly the same. This is a very important derating guideline. The above derating analysis has been verified by the experimental work in [104].

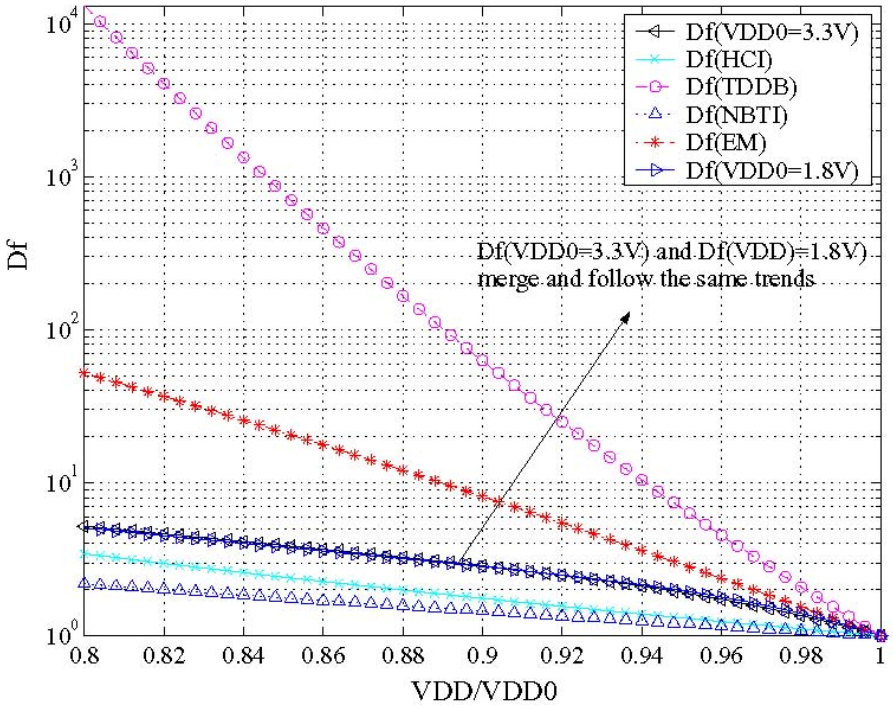

Fig. 3.7. *Trends of D_f vs. V_{DD}/V^0_{DD} with typical model parameters from the 0.18–μm technology. V_{DD} is derated within the range [100%–80%] of its rated value V^0_{DD} =3.3 V. The trend of D_f when V^0_{DD} = 1.8 V is also plotted for comparison.*

3.5 FaRBS Application: An Analog-to-Digital Converter Reliability Simulation

3.5.1 Introduction

An Analog-to-Digital Converter (ADC) is a data-acquisition device that interconnects and converts real-world signals into digital codes that will be processed by digital units with high speed and low cost. Due to its critical role in systems, many conversion techniques and architectural styles have been proposed to implement analog-to-digital conversion functionality. ADCs can be broadly classified into high-speed and high-resolution structures. Flash, folding and interpolating, multistep, and pipelined architectures belong to the high-speed category. Successive approximation, delta-sigma, and integrating architectures implement high-resolution conversions. These two categories trade off speed and accuracy [91], as is shown in Figure 3.8.

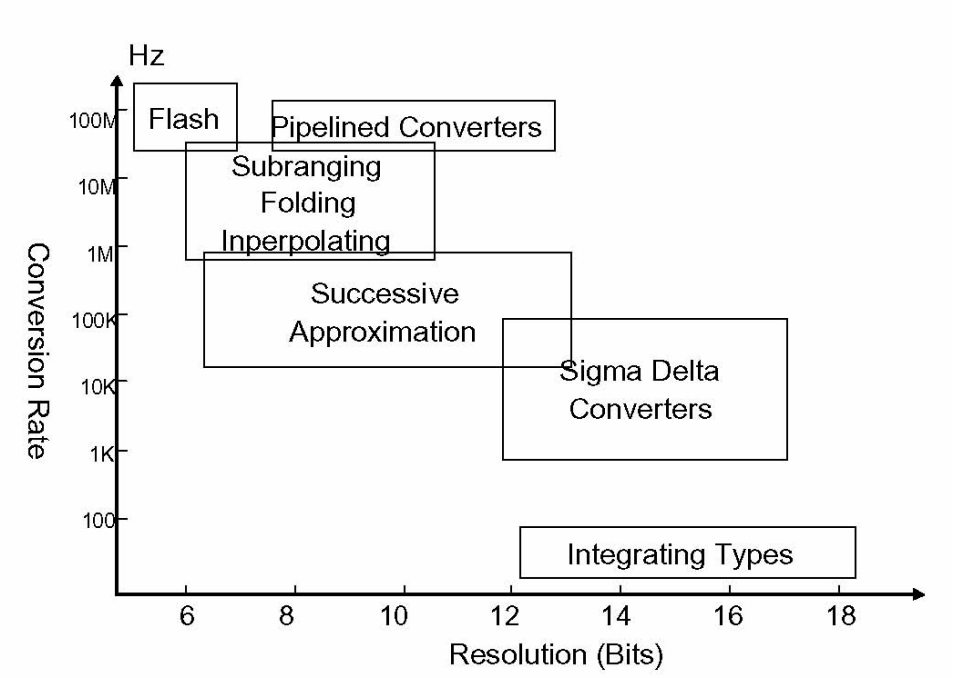

Fig. 3.8. *ADC speed and resolution tradeoff.*

Compared to digital circuitry, little work has been done on reliability modeling and simulation of analog and mixed-signal counterparts, while the reliability constraints of the technology dictate that the analog circuit operates at the same or even worse stress levels. Significant reliability concerns encountered by analog devices (ADC in particular) have been discussed in [105], where TDDB and HCD are identified as two primary failure mechanisms experienced by switched-capacitor implementation of analog circuits in the process of voltage scaling for reliability. In this work, for simplicity, we only estimate lifetimes of ADCs due to HCD. Full analysis including EM, TDDB, and NBTI will be performed in the work of SRAM reliability simulation with FaRBS.

3.5.2 ADC Circuits

A flash ADC, also known as a fully parallel architecture, is conceptually the simplest and fastest architecture and is easy to understand. An n-bit flash ADC consists of a ladder of 2^n-1 comparators and a set of 2^n-1 equally distributed voltage reference values normally generated from a poly resistance string. Each of the comparators samples the analog input signal and compares the signal to its reference value. Each comparator then generates a digital output (low voltage or high

voltage) indicating whether the input signal is larger or smaller than the reference assigned to that comparator. The output pattern of these 2^n-1 comparators is often referred to as a thermometer code. This name is derived from the fact that if the comparator outputs are listed in a column and ordered according to the reference values associated with the comparator that produced them, the ones would all be at the bottom, and the zeroes all at the top. The level of the boundary between ones and zeroes would indicate the value of the signal, much as the level of mercury in a mercury thermometer indicates the temperature. Finally, a digital encoder unit converts the thermometer code produced by the comparators to a binary code and outputs this code to the digital signal processing units that follow. The two primary drawbacks to the flash ADC are the large hardware requirement and sensitivity to comparator offsets.

A well-designed, 6-bit, 1.3-G sample/s, flash, high-speed ADC based on 0.35-µm CMOS process has been developed by Michael Choi et al. at UCLA [106]. In this work, the purpose is not to investigate design techniques for ADC to achieve high speed and high performance, but rather to illustrate how FaRBS will be applied to devices for reliability modeling and simulation. For this purpose, a typical and simple circuit structure is considered. Therefore, based on the similar circuit techniques for the 6-bit, 1.3-G sample/s, flash, high-speed ADC, a simplified 3-bit flash ADC was designed as a candidate example.

Figure 3.9 is the circuit block diagram of the 3-bit flash ADC. The clock circuit synchronizes the operations of all other circuit blocks. In most designs, much effort will be put on the clock circuit because it is vital for other blocks. As a result, problems associated with clock (e.g., skew, jitter, feedthrough, etc.) and its hardware failures are normally kept to a minimum. In this design, for simplicity, a square-wave voltage source (not transistor-level implementation) is used to generate the clock signal; the reliability impact is not considered in the whole system. The sample-and-hold (S/H) circuit tracks and samples the differential input signals, then outputs the sampled values to the preamplifier circuits. Each of the seven preamplifiers magnifies the difference between its input differential signal and its threshold voltage (i.e., reference voltage for comparison) that is generated from the resistor ladders. Then, each comparator of the two-stage comparator array will amplify the output of its preamplifier to a rail-to-rail signal. These rail-to-rail signals will be processed by the 7-to-3 digital encoding circuit and converted to the desired digital

code. The encoding circuit is implemented by a NAND-NAND structure. The function of resistive averaging networks between preamplifiers and the first comparator array, and between the first and the second comparator arrays, is to lower the impact of offsets.

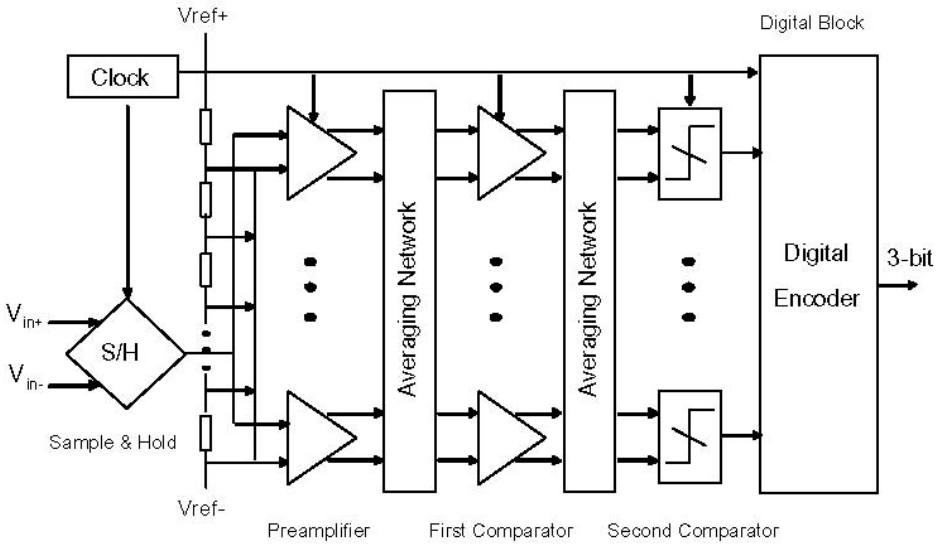

Fig. 3.9. *Block diagram of 3-bit ADC.*

Suppose the peak-to-peak full-scale range of the input signals is 1.6 V, and the input common voltage is 0.5 V; in this example each of the input differential signals will vary from 0.1 V to 0.9 V. The flash ADC will then convert these analog input signals to 3-bit digital codes.

1. Sample-and-hold circuit

The S/H circuit samples the differential input signals with two 0.1-pF capacitors for a certain amount of time during the high clock level. The circuit for S/H is very simple and is plotted in Figure 3.10. When the clock goes low, the input signals are disconnected from the S/H circuit; and the voltages saved on the capacitors will be amplified by the p-channel metal oxide semiconductor (PMOS) source followers. The n-well for PMOS is connected to its source to reduce the nonlinear body effect. In the following stages, the differential input signal ranging from -0.8 V to 0.8 V will

be compared with the linear increased reference voltages generated from resistance ladders. These reference voltages will be applied on the gates of NMOS transistors to bias them in the saturation region; therefore, reference voltages cannot start from values lower than threshold voltages of these NMOS transistors. This means the sampled input voltages must be raised to a higher range. PMOS source followers realize this level-shifting function. The common mode voltage from PMOS source followers will be raised to 1.64 V. Based on the above information, the range of reference voltages for comparators is set from 1.24 V to 1.64 V.

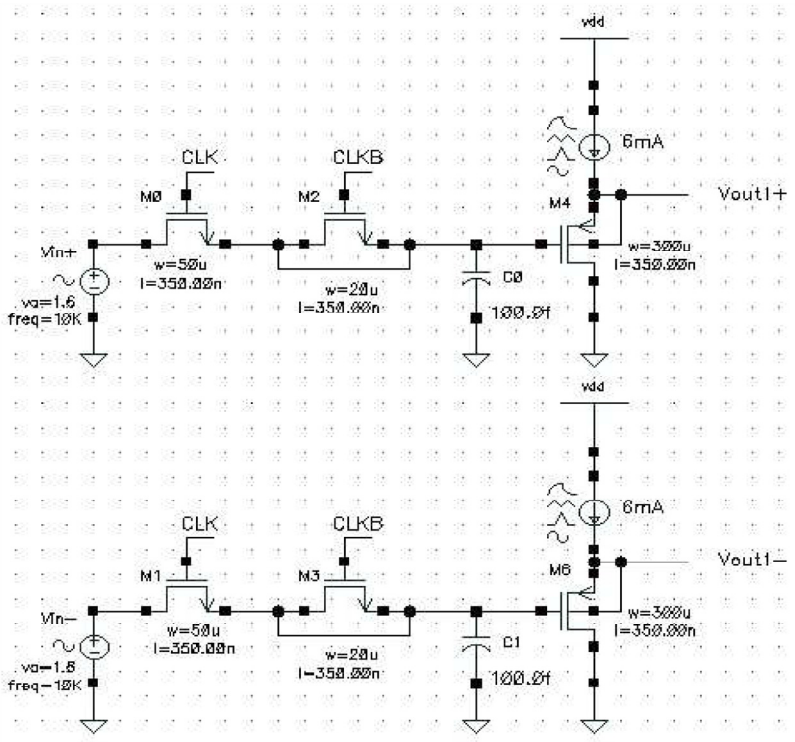

Fig. 3.10. *Sample-and-hold circuit.*

The dummy switches that are driven by the reverse of the clock signal and located between the inputs and the sampling capacitors will reduce the effect of charge injection and clock feedthrough.

2. Preamplifier circuit

The Preamplifier circuit is shown in Figure 3.11. The preamplifier circuit is designed with two cross-connected pairs of active-load amplifiers in a differential input/differential output structure. It magnifies the difference between input signals and voltage references and provides high gain to overcome the comparator offsets that follow. During the positive level of clock, the NMOS M323 bridging between the differential outputs will reset and eliminate the residual voltages of previous cycles.

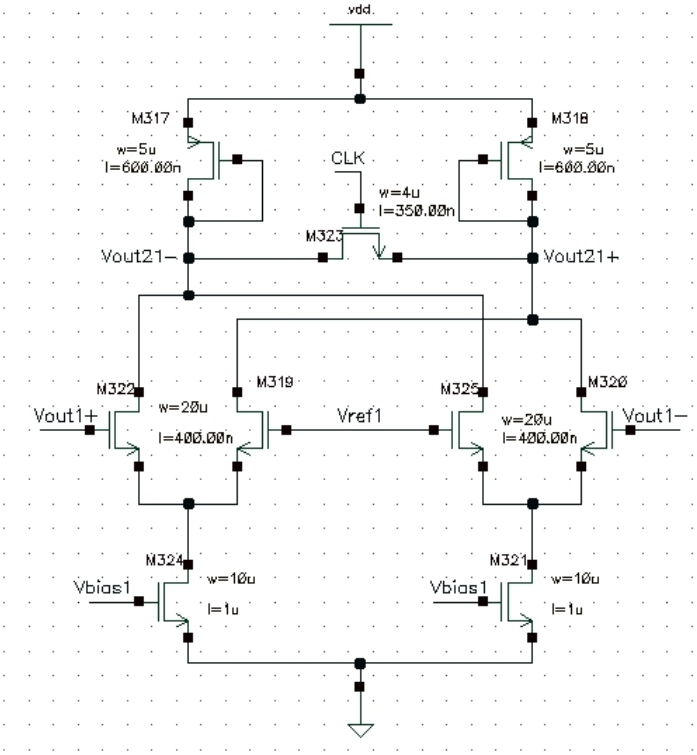

Fig. 3.11. *Preamplifier circuit.*

3. First-stage comparator circuit

The first-stage comparator consists of a differential pair and a latch pair that are synchronized by the negative level and positive level of clock, respectively, and therefore work in an alternate way. When the clock is high, the latch pair holds the output results of the previous cycle; when the clock goes low, the differential pair is activated and its differential output follows its differential input. At the same time, the switch between the differential outputs turns on and erases the memory of the previous decision. The first-stage comparator circuit is plotted in Figure 3.12.

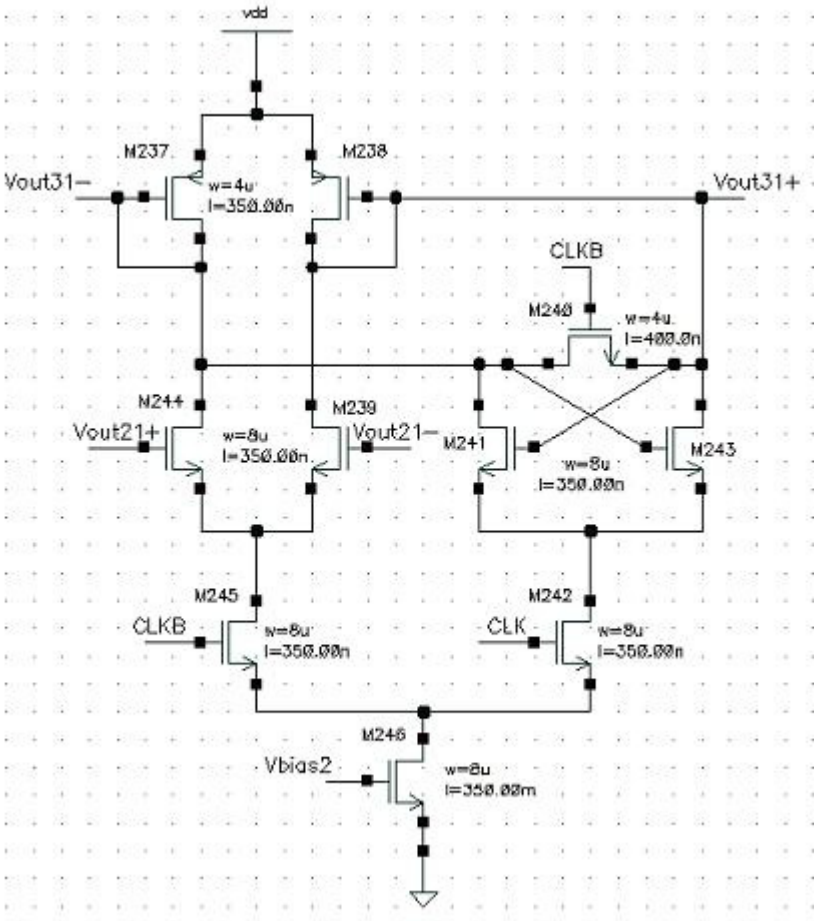

Fig. 3.12. *First-stage comparator circuit.*

4. Second-stage comparator circuit

The second-stage comparator will generate rail-to-rail outputs to the digital encoding circuit that follows. The second-stage comparator circuit is plotted in Figure 3.13. When the clock is high, the output is reset through two parallel discharge paths for fast overdrive recovery and two cascade amplifiers with symmetrical structure will sample and amplify the input signals. When the clock goes low, the cross-coupled latch pair speeds up the regeneration process and quickly drives the output voltage to a rail-to-rail range.

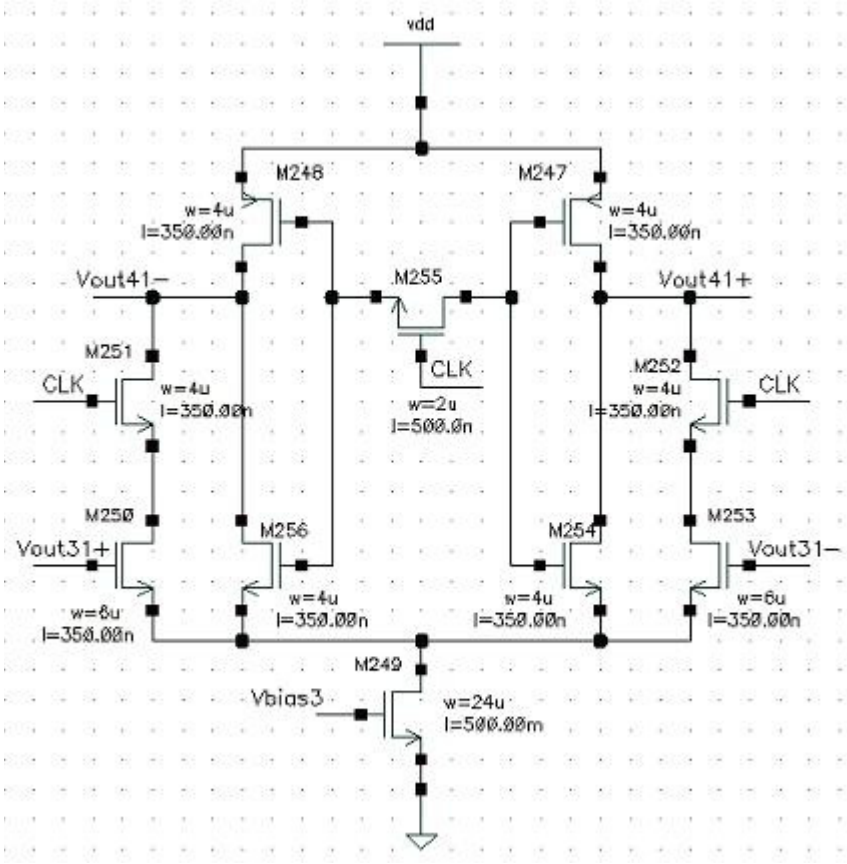

Fig. 3.13. *Second-stage comparator circuit.*

5. Digital encoder circuit

Digital encoder converts the thermometer codes from the second-stage comparators to the desired binary codes. The circuit for encoding the least significant bit (LSB) A1 is shown in Figure 3.14. For simplicity, the digital correction circuit was not designed-in and the possible bubble errors or glitch errors caused by comparator metastability in the thermometer codes were not considered. The design of a digital encoder whose function is converting thermometer codes to standard three bits or less binary codes is straightforward. The inputs of the digital encoder are seven signals (as well as their reverse signals) generated from the second-stage comparator array; the outputs of the encoder are 3-bit binary codes. From the truth-table, the logic function for each output digit can be expressed by an OR function whose four product terms are combinations of the seven input signals or their reverse signals. Then, the encoding function can be realized with 7-input NAND to 4-input NAND structures. The circuits for the other two digits are the same as Figure 9 except that the input gate signal of each transistor in the array is different according to its logic functions. However, this implementation structure is only good for encoding three bits or less because, with the linear increase of output bits, the long series-connected pull-down NMOS chains will introduce non-linear body effect, degrade circuit transmission speed, and exponentially increase the areas of the pull-up PMOS. For encoding a higher number of bits, other design style and circuit structures should be employed to overcome these problems. As discussed before, digital circuits are normally driven by rail-to-rail signals and by output rail-to-rail signals. As a result, device parameter degradation has a relatively (compared with analog circuits) minor effect on their functionality. For mixed-signal circuits, such as ADC, the reliability bottleneck is obviously set by analog blocks in this work; therefore, reliability simulation and analysis on digital encoder circuits are not performed.

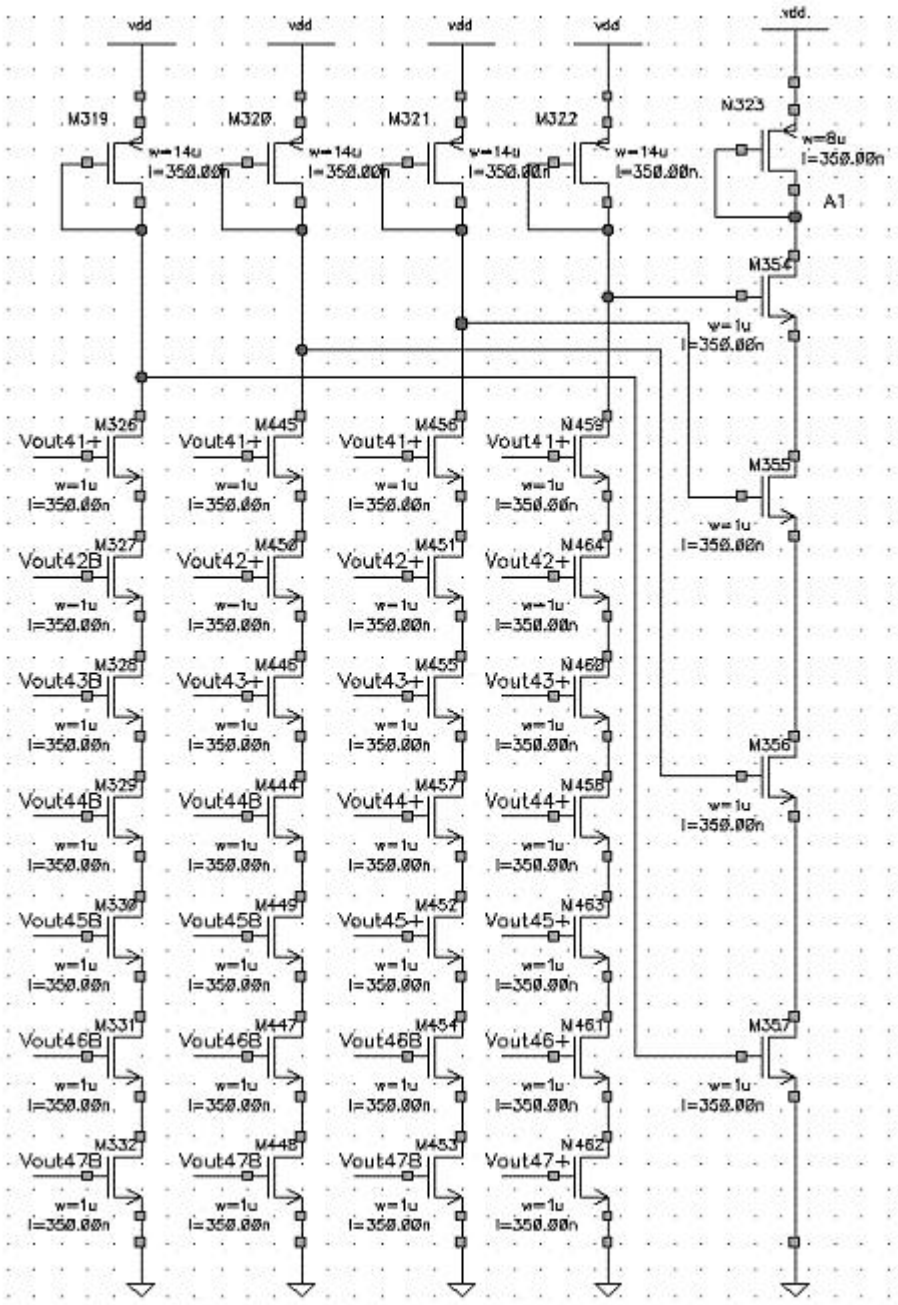

Fig. 3.14. *Digital encoder circuit for bit A1.*

6. Averaging network

Averaging technique is unavoidable in flash ADC design because the random offsets of the comparator chain in the flash ADC structure are significant. In this design, the average resistor network is nothing more than a series of resistors that sum the output of all amplifiers and balance them to an appropriate operation point. The reliability of the averaging network is relatively insignificant in this work and, therefore, is not discussed here.

3.5.3 FaRBS Analysis of ADC Reliability

SPICE simulation can evaluate the performance of the ADC circuit. Some important parameters of the ADC (such as integral nonlinearity (INL), differential nonlinearity (DNL), signal-to-noise-plus-distortion ratio (SNDR), and spurious-free-dynamic-range (SFDR)) can be determined from SPICE simulation. In FaRBS, the purpose of SPICE simulation is to obtain the real-time operating parameters of each transistor; functional verification and performance analysis are second-tier issues. Therefore, we will not simulate or discuss the performance parameters of ADCs.

The ADC is a kind of special circuit that has regular repetitive structures; therefore, simulation of each transistor is not necessary. Most of the transistors work under the same operating conditions and similar stress levels. However, as mentioned before, since a practical analog or mixed-signal circuit does not have repetitive subcircuits and normally contains huge numbers of transistors, it is not trivial work to simulate the operating parameters of each transistor and calculate its contributions to the overall FIT value. The best way to overcome this constraint is sensitivity analysis.

1. Sensitivity analysis

Sensitivity analysis for wearout failures such as HCD and TDDB is still under development. For this work, we directly apply the results of [91] to identify the reliability critical subcircuits. Circuits in space systems are constantly exposed to α-particle interference and are prone to suffer transient failures. An α-particle transient model has been proposed in [91], and a simple injection

model was used in SPICE to analyze the system functional errors induced by α-particle in ADCs. The sensitivity analysis results are shown in Figure 3.15. The horizontal axis represents the constituent subcircuits of the ADC; the vertical axis shows the normalized maximum relative error (i.e., sensitivity value) of each subcircuit. It is obvious that an S/H amplifier subcircuit (SHA) is the most sensitive block for α-particle transient failures. Even though the failure behaviors of ADCs due to α-particle might be different from those due to wearout failures, we still can draw the conclusion that the SHA subcircuit is more functionally important than others in the whole ADC. Averaging network mainly consists of resistors that are passive devices; parameter drifts due to wearout or degradation generally have more severe effects on the functionality of analog subcircuits than digital counterparts. Therefore, for ADC reliability simulation, we will focus on transistors in the SHA and neglect averaging network and digital encoder subcircuits. We still have to consider transistors in preamplifiers and the two-stage comparators because there are large numbers of them, even though they each make a relatively small contribution to the overall FIT value.

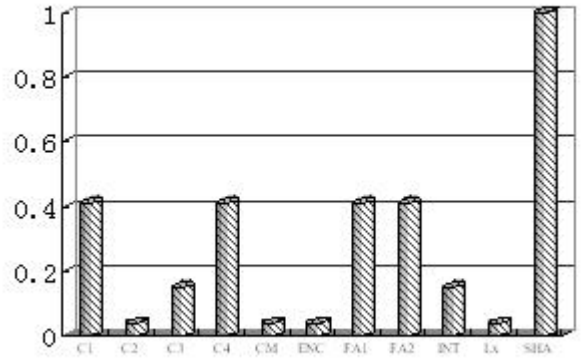

Fig. 3.15. *Sensitivity of ADC's subcircuits [2].*

2. Model parameters

After we identify which transistors should be included in the reliability simulation, the SPICE simulation process in order to obtain their operating parameters is straightforward. The next step is subbing these parameters into the wearout models that were introduced before. There is a gap between the parameters from the SPICE simulation and the parameters in the wearout models

because the wearout model parameters are calibrated under DC stress conditions, while the simulated parameters are for real-time conditions. Hot-carrier reliability design rules for translating device degradation to CMOS digital circuit degradation have been proposed in [93], where the relation of device hot carrier lifetime between DC and AC stresses has been established. In our ADC circuit, there is only one clock signal to synchronize all subcircuits. If gate-to-source voltage is the only stress factor, if the temperature effect is negligible, and if we do not consider the self-healing effect, then most transistors are stressed exactly half of the time. For this case, we simply double the MTTF result calculated from wearout models with the well-defined clock signal voltage level. However, in reality, these preconditions are usually not well established. For HCD, Equation (3.3) shows that drain-to-source voltage, not gate-to-source voltage, is the stress factor and temperature effect is small. However, drain-to-source voltage continuously changes with time; therefore, we cannot estimate lifetime by simple inspection. For this case, SPICE simulation is necessary. The average value of the simulated drain-to-source voltage of a transistor can be used in the HCD wearout model.

The fitting and empirical parameters of the wearout models have to be extrapolated first before the calculation of MTTF or FIT with simulated transistor operating parameters. The development of a systematic algorithm for model parameter extrapolation is considered future work. For the current stage, we use data in literature [107] and fit them by inspection. The fitting process is straightforward but time consuming; therefore, for simplicity, only the final results are given here. For HCD, its lifetime is determined by Equation (3.3), and the model parameters are A_{HCD} = $4x10^{-5}$s and $\theta = 90$ v.

3. SPICE simulation

The circuit is simulated with SPICE in Cadence. A TSMC 0.35-µm technology with a 3.3-V power supply voltage is used. The full-swing scale of the input differential signal is −0.8 V to 0.8 V. The frequency is 100 MHz. The sample frequency (clock signal) is 800 MHz. The sampled output of SHA circuit with respect to the differential input is shown in Figure 3.16. The outputs of SHA are compared with reference voltages by the preamplifiers, and the difference is amplified.

The differential input and output of the preamplifier (the lowest one in the preamp array) are plotted in Figure 3.17.

The differences (generated from preamplifiers) between the sampled voltages and related reference voltages are first amplified to a larger range by first-stage comparators (shown in Figure 3.18) and then driven to rail-to-rail voltage signals by second-stage comparators (shown in Figure 3.19). These rail-to-rail signals are applied to the digital encoder circuit to generate the required 3-bit binary codes.

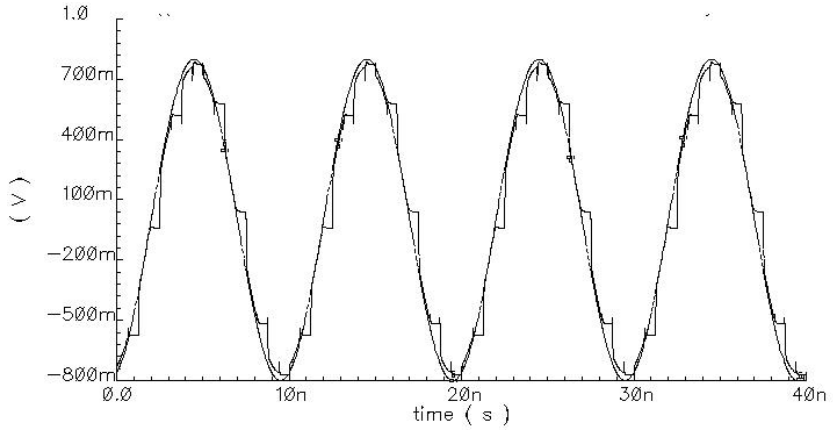

Fig. 3.16. *Differential output and input of SHA.*

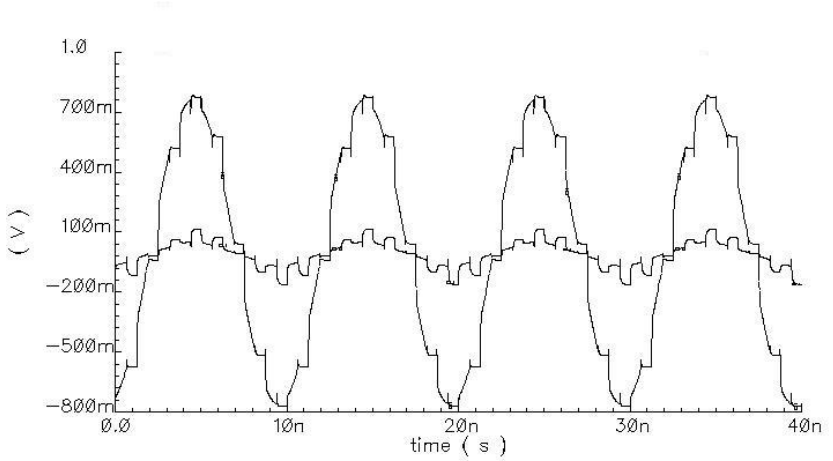

Fig. 3.17. *Differential output and input of a preamplifier.*

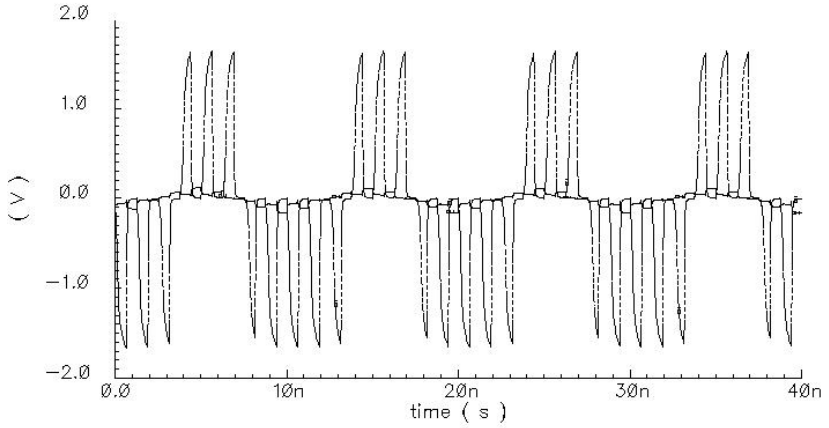

Fig. 3.18. *Differential output and input of a first-stage comparator.*

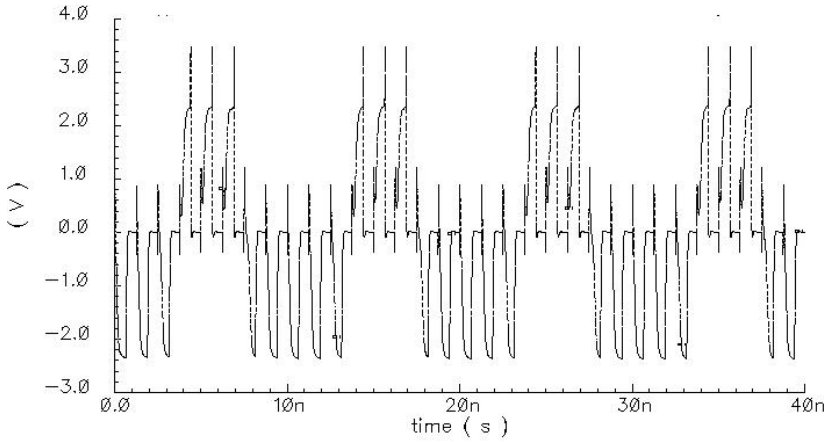

Fig. 3.19. *Differential output and input of a second-stage comparator.*

4. Lifetime calculation

From SPICE simulation, it is easy to determine the time-related operating parameters of each transistor and the passive elements in a circuit. Cadence Analog Design Environment provides a

waveform calculator tool with embedded functions to calibrate the average value of a signal. For HCD analysis, therefore, we can easily calculate the average value of drain-to-source voltage of each transistor with which we are concerned. Based on the previous discussion, we will only select the most reliability-critical transistors that make a significant contribution to system functionality. These transistors are identified in Table 3.1. Their average drain-to-source voltages given by SPICE simulation are also listed in the same table.

Table 3.1. *Critical transistors and their average V_{ds} (V).*

Subcircuit	**Critical Transistors with Average V_{ds} (V).**
SHA	M0 (34m), M1 (35m), M4 (1.64), M6 (1.64)
Preamp	M324 (428m), M321 (418m), M322 (127m), M319 (114m), M325 (137m), M320 (124m), M317 (2.75), M318 (2.76)
1st Comparator	M246 (288m), M245 (144m), M244 (1.8), M239 (1.7), M237 (1.34), M238 (1.49)
2nd Comparator	M249 (182m), M250 (668m), M253 (540m), M251 (2.02), M252 (1.95), M248 (622m), M247 (811m)

The preamplifier and the two-stage comparator circuits are repetitive structures. We assume a similar transistor in these similar subcircuits is approximately working under similar voltage stress levels; therefore, each transistor identified in the above table (except those in SHA) will be multiplied by seven to calculate its contribution to the overall FIT value. In simulation, signals are not perfect and might have glitches that will complicate the estimation of their average value. For now, we ignore this effect. From the HCD lifetime model, higher drain-to-source voltages will lead to a shorter lifetime. The above table shows that stresses on some transistors are much higher than others, so their failures will dominate the lifetime of the whole ADC circuit. Subbing V_{ds} values in the above table and HCD model parameters set into Equation (3.3), we can estimate the FIT value of the ADC. Based on these data, the FIT (due to HCD) for the ADC is approximately 2.1, which corresponds to a lifetime of 15 years. A comprehensive analysis, including all wearout models, can be performed if the accurate model parameters are available. The above calculation indicates that accurate wearout model parameters and SPICE simulation results are the most important factors for the accuracy of the final FIT value.

4 MICROELECTRONIC CIRCUIT RELIABILITY ANALYSIS AND MACRO

4.1 Introduction

In this chapter, lifetime and failure equivalent circuit models for common silicon intrinsic wearout mechanisms including hot-carrier injection (HCI), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI) are presented. Additionally, this chapter includes an illustrative case study for the purpose of demonstrating how to apply Maryland Circuit-Reliability Oriented (MaCRO) spacecraft, planets, instrument, C-matrix, events (SPICE) [16] models and algorithms to improve circuit reliability simulation and analysis. MaCRO has an integrated-circuits emphasis that was developed based on the rate-of-failure concept and failure-equivalent circuit-modeling techniques; it consists of a series of accelerated lifetime models and failure-equivalent circuit models for intrinsic wearout mechanisms.

The most common circuit structures used in reliability simulations are the ring oscillator, the differential amplifier, and the SRAM. The SRAM is selected as a case study vehicle to show the applicability of MaCRO models and algorithms in microelectronic circuit reliability simulation and analysis.

4.2 Hot Carrier Injection

Most HCI lifetime models are based on the "lucky electron" model, upon which the hot carrier stress on an n-channel metal oxide semiconductor (NMOSFET), in terms of generated interface traps ΔN_{it}, can be related to the electric field E_m at the drain, the drain-to-source current I_{ds}, and the stress time t in a simple power-law relation [107]:

$$\Delta N_{it} = C_1[\frac{I_{ds}}{W}\exp(-\frac{\Phi_{it,e}}{q\lambda_e E_m})t]^n \tag{4.1}$$

where W is the channel width, $\Phi_{it,e}$ is the critical energy for electrons to create an interface trap ($\Phi_{it,e}$ =3.7 eV [45]), λ_e is the hot-electron mean-free path (λ_e = 6.7 nm [108]), and C_1 is a process constant. The dynamics of interface trap generation is similar to the rate of thermal oxide growth. At the initial stage, the interface trap generation rate is reaction limited, therefore, $N_{it}(t) \propto t$ and n = 1. At the later stage, the generation is diffusion limited, then $N_{it}(t) \propto t^{1/2}$ and n = 0.5. The overall process is the compromised result of these two competing processes and, as a result, the parameter n falls within the range between 0.5 and 1 [45]. In *FaRBS*, the default value of n is set to 0.65.

The most important parameter in Equation (4.1) is the electric field E_m, which cannot be determined accurately by simple calculation. A semi-quantitative analytical E_m model has been given in [45]:

$$E_m = \frac{V_{ds} - V_{dsat}}{\sqrt{3 t_{ox} x_j}} \tag{4.2}$$

where t_{ox} is the gate oxide thickness and x_j is the drain junction depth. $\sqrt{3 t_{ox} x_j}$ is the characteristic length that models the effective thickness of the channel "pinchoff" region whose typical values are within $\sqrt{100}$ nm to $\sqrt{300}$ nm. The factor 3 in $\sqrt{3 t_{ox} x_j}$ is derived from the ratio of $\varepsilon_{Si}/\varepsilon_{SiO2}$ [109]. In *FaRBS*, the default value of $\sqrt{3 t_{ox} x_j}$ is 10 nm.

In Equation (4.2), V_{dsat} is the potential at the channel "pinchoff" point. There are many models for V_{dsat}, among which the simplest one is $V_{dsat} = V_{gs} - V_t$, where V_{gs} is gate-to-source voltage and V_t is the threshold voltage. For short channel devices, V_{dsat} is channel length (L) dependent, and the relation is often modeled as [45]:

$$V_{dsat} = \frac{(V_{gs} - V_t) L E_{cr}}{V_{gs} - V_t + L E_{cr}} \tag{4.3}$$

where E_{cr} is the critical field for velocity saturation and its value is approximately 5×10^4 V/cm.

In the above discussion, the only unknown parameter in Equation (4.1) is the coefficient C_1, which is a process-determined constant. For each technology, this constant only needs to be characterized once. The typical values of C_1 are within 1.9 ~ 2 according to [108].

In addition to the interface trap generation model given by Equation (4.1), the other two important models for hot carrier effects are the substrate current (I_{sub}) model and the gate current (I_{gate}) model:

$$I_{sub} = C_2 I_{ds} \exp(-\frac{\Phi_i}{q\lambda_e E_m}) \tag{4.4}$$

and

$$I_{gate} = C_3 I_{ds} \exp(-\frac{\Phi_b}{q\lambda_e E_m}) \tag{4.5}$$

where Φ_i is the minimum energy (in electron-volts) for a hot electron to create an impact ionization (Φ_i = 1.3 *eV*) and Φ_b is the barrier energy (also in electron-volte) at the *Si–SiO*$_2$ interface. The formula for Φ_b is given by Equation (3.9) in [108]. The constants C_2 and C_3 are given in [45] as C_2 = 2 and $C_3 = 2 \times 10^{-3}$.

By defining the device hot carrier lifetime tf as the time to reach a fixed amount of interface trap density, we can combine Equation (4.1) and Equation (4.4) into a very useful lifetime equation:

$$\frac{t_f I_{ds}}{W} = C_4 [\frac{I_{sub}}{I_{ds}}]^{-\Phi_{it,e}/\Phi_i} \tag{4.6}$$

Equation (4.6) is used in many hot carrier reliability simulation tools derived from Berkeley Reliability Tools (BERT) [110]. From this equation, a very simple accelerated lifetime model for HCI can be obtained:

$$t_f = C_5 \exp(\frac{\theta}{V_{ds}}) \tag{4.7}$$

where C_5 and θ are technology-related constants whose values are determined from accelerated lifetime tests and V_{ds} is the drain-to-source voltage. The power of Equation (4.7) is that it relates a device's HCI lifetime to only one operating parameter, which can be directly calibrated from SPICE simulation. The main problem with this simple relation is that it is only valid for a small range of gate voltages near the maximum substrate current [107].

To account for more realistic hot carrier stressing profiles in the circuit environment, a more general lifetime model is incorporated in FaRBS that relies on the substrate current model. I_{sub} has been identified as the best hot carrier reliability monitor for NMOSFETs. According to [111], the device parameter degradation due to HCI can be modeled as:

$$\Delta P = C_6 \left(\frac{I_{sub}}{W}\right)^{\alpha} t^{\beta} \tag{4.8}$$

where I_{sub}/W is the normalized substrate current and α, β, and C_6 are technology-related constants.

Temperature acceleration is often treated as a minor effect in most HCI models; however, to consider possible large temperature excursions, FaRBS includes a temperature acceleration effect based on the HCI lifetime model given in [55]. The combination of temperature effect and Equation (4.8) produces a more comprehensive HCI lifetime model:

$$t_f = A_{HCI} \left(\frac{I_{sub}}{W}\right)^{-n} \exp\left(\frac{E_{aHCI}}{\kappa T}\right) \tag{4.9}$$

where E_{aHCI} is the apparent activation energy (E_{aHCI} is within $-0.1\ eV$ to $-0.2\ eV$), W is the device gate width, κ is Boltzmann's constant ($\kappa = 8.62 \times 10^{-5}\ eV/K$), T is temperature in Kelvin, n is a technology-dependent constant, and A_{HCI} is the model prefactor. In FaRBS, the default values for n and E_{aHCI} are $n = 1.5$ and $E_{aHCI} = -0.15\ eV$, respectively.

There are two ways to determine I_{sub}: one is from Equation (4.4), the other is from BSIM3 model equations, as follows:

$$I_{sub} = \frac{\alpha_0 + \alpha_1 L_{eff}}{L_{eff}} V_{ds}^{'} \exp(\frac{-\beta_0}{V_{ds}^{'}}) \frac{I_{ds0}(1 + V_{ds}^{'}/V_A)}{1 + R_{ds} I_{ds0}/I_{dseff}} \tag{4.10}$$

$$V_{ds}^{'} = V_{ds} - V_{dseff} \tag{4.11}$$

The meaning of the above model parameters is given in BSIM3 Model User Manual [112]. This BSIM3 I_{sub} model is quite similar to the I_{sub} model proposed in iProbe-d [113]; therefore, the iProbe-d I_{sub} model is an alternative if some SPICE simulator does not support BSIM3 I_{sub} calculation.

The degradation of P-channel MOSFETs (PMOSFETs) under hot carrier stress is becoming one of the important contributors to circuit reliability. The hot-carrier-induced PMOSFET degradation effect on circuit performance is different from that of NMOSFET in that it might lead to reverse shifts (compared to NMOSFET) in device and circuit parameters due to significant negative charge trapping in oxide rather than excessive interface trap generation. The circuit performance degradation can be characterized more accurately if PMOSFET HCI effect is also considered. Even though the wearout dynamics and device parameter degradation trends of PMOSFETs are different from those of NMOSFETs, with minor modifications, the above NMOSFET's accelerated lifetime can be applied to PMOSFETs.

$$t_f = A_{HCI,p} (\frac{I_{gate}}{W})^{-m} \exp(\frac{E_{aHCI,p}}{\kappa T}) \tag{4.12}$$

where $E_{aHCI,p}$ is the apparent activation energy ($E_{aHCI,p}$ is within $-0.1\ eV \sim -0.2\ eV$) and W is the device gate width. m and $A_{HCI,p}$ are technology-related constants whose default values in FaRBS are $m = 12.5$ and $E_{aHCI,p} = -0.15\ eV$, respectively. The I_{gate} is given by Equation (4.5).

In developing FaRBS HCI accelerated lifetime model, we assume a quasi-static approximation that averages device dynamic operation parameters (e.g., I_{ds}, V_{ds}, V_{gs}) in terms of simulation time; therefore, I_{sub} and I_{gate} in Equations (4.9) and (4.12) are average values calculated from Equation (4.4) and Equation (4.5), respectively.

4.2.1 Failure-Equivalent Circuit Model

To account for the effect of device hot carrier damage on circuit functionality and reliability, the device-level accelerated lifetime models have to be extended to circuit-level applications. The bridge connecting the gap between device wearout degree and circuit performance drift is the failure-equivalent circuit models. The underlying concept for failure-equivalent circuit models is modeling degradation of device parameters with some additional lumped circuit elements (resistors, transistors or dependent current sources, etc.) to capture the behavior of a damaged MOSFET in the circuit operation environment. The values of these additional lumped elements are determined by device wearout parameters (such as ΔN_{it}) that are time dependent and by device terminal voltage and current waveforms; therefore, at any time, t values of these lumped elements can be predicted accurately and their magnitude will reflect the device wearout degree. The larger the magnitude of these values, the more severe the damage to circuit functionality. As a result, circuit designers can quickly analyze circuit reliability behaviors at any given time with these failure equivalent circuit models.

Several HCI failure equivalent circuit models have been developed in the past years; some of them have been built into commercial reliability simulation tools. Almost all failure equivalent circuit models are based on the SPICE simulation platform, which is a de facto tool in circuit design. In this section, we first briefly review some of the failure equivalent circuit models, then introduce the equivalent circuit models adopted in FaRBS.

BERT has thus far been the most successful circuit reliability simulation tool. BERT directly models NMOSFET hot carrier damage in drain current degradation. The drain current degradation, ΔI_d, results from channel mobility degradation, which again results from HCI-induced interface

traps ΔN_{it}. ΔN_{it} is modeled in terms of the *Age* parameter. In BERT, ΔI_d is implemented as an asymmetrical-voltage-controlled current source in parallel with the original NMOSFET. The PMOSFET HCI effect is modeled with the concept of channel shortening and drain resistance increase [110]. The BERT ΔI_d model shown in Figure 4.1 captures the asymmetrical forward and reverse $I-V$ characteristics and allows simulation of devices undergoing bi-directional stresses (such as devices in a transmission gate).

The detailed ΔI_d model equations and parameters are defined in [115]. The main contribution of the BERT ΔI_d model is the ability to characterize bi-directional hot carrier stress effects; however, this model requires the extraction of six process parameters from device testing experiments, which are not easy to implement.

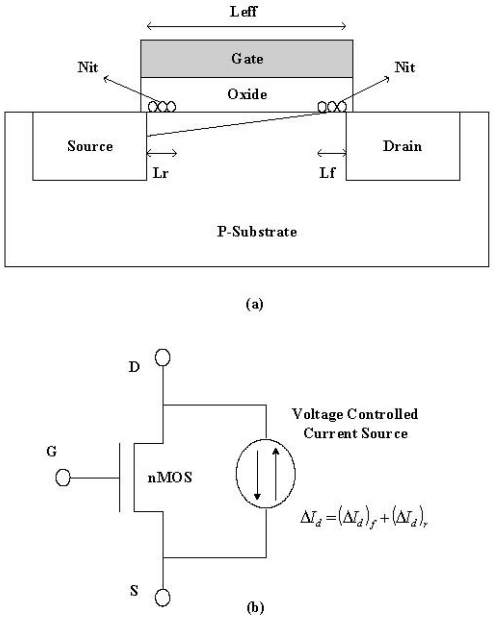

Fig. 4.1. *BERT NMOSFET HCI failure equivalent circuit model. (a) Bi-derectional interface trap generation near both drain and source. L_f and L_r represent forward and reverse hot carrier damaged regions. (b) HCI drain current ΔI_d failure equivalent circuit model [114].*

Experiments have proven that HCI-induced interface traps in NMOSFET are localized above the channel near the drain junction. More specifically, these interface traps are localized in the vicinity of 100 nm from the drain [45]. Based on this observation, Leblebici et al. at University of Illinois at Urbana-Champaign (UIUC) [108][116] developed a two-transistor HCI failure equivalent circuit model that consists of an HCI damaged parasitic transistor with fixed channel length L_2 ($L_2 \approx 0.1$ μm) in series connection with the original transistor whose channel length is shrunk to $L-L_2$. The primary assumption for this model is that all generated interface traps are occupied with electrons, which amounts to considering only a negative fixed charge. The model is illustrated in Figure 4.2.

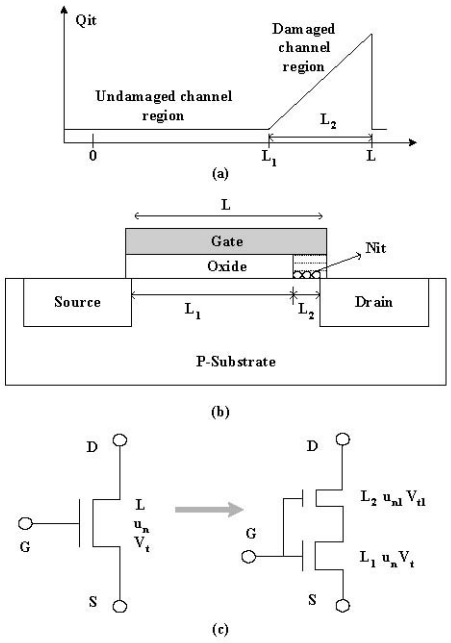

Fig. 4.2. *UIUC NMOSFET HCI two-transistor series model. (a) Triangular oxide charge distribution profile used in model derivation. (b) Cross section view of NMOSFET with hot carrier damage (L_1 is undamaged channel region and L_2 is damaged channel region). (c) Two-transistor series failure equivalent circuit model. The parasitic transistor has different channel mobility and threshold voltage with the channel length L_2 set to 0.1 μm [108][113][117].*

From Figure 4.2 (a), the interface trapped charge Q_{it} due to HCI can be readily derived as:

$$Q_{it}(x) = 0 \tag{4.13}$$

when ($0 \leq x \leq L_1$)

$$Q_{it}(x) = \frac{Q_M}{L_2}(x - L_1) \tag{4.14}$$

when ($L_1 \leq x \leq L$)

where Q_M denotes the largest interface charge and $L_1 = L - L_2$, and L_2 represents the length of the damaged channel region. This two-transistor model characterizes the amount of hot carrier damage with only two parameters Q_M and L_2; therefore, the model parameter extraction work is greatly reduced. The drawbacks of this model are related to two aspects: the triangular charge density distribution is oversimplified, and it is not easy to extrapolate the Q_M value.

Until now, the simplest HCI failure equivalent circuit model has been the Hot Carrier Induced Series Resistance Enhancement Model (HISREM), also named the ΔR_d model, which was proposed by Hwang et al. at Oregon State University [118]. Based on the fact that the increase of HCI-induced series drain resistance is due to the injection of hot carriers close to the drain edge, a series resistance ΔR_d added to the drain of the NMOSFET can reflect the process of hot-carrier-induced interface-trap generation and, therefore, account for the channel mobility reduction and threshold-voltage drifts. HISREM consists of a voltage-dependent drain resistor ΔR_d connected in series with the original NMOSFET. ΔR_d is a function of the applied voltages and the hot-carrier-induced interface trapped charge ΔN_{it}. The behavior of the damaged NMOSFET is emulated by the original undamaged device operated with a reduced drain-to-source voltage that is controlled by this additional drain resistor ΔR_d. Because ΔN_{it} is a time-dependent parameter, the ΔR_d model is able to predict drain current degradation at any given time. HISREM is capable of modeling self-limiting effects of hot carrier damage because the increase in series drain resistance of an NMOSFET suppresses hot carrier stress. The most advantageous feature of the HISREM model is that only one parameter, ΔN_{it}, needs to be extrapolated from experiments. Consequently, the HISREM model can be easily used by circuit designers to perform a rapidly derived reliability analysis.

The HCI failure-equivalent circuit model in FaRBS is based on the above ΔR_d model with some improvements. The major improvement is that the ΔR_d value is considered to be determined by both interface-trapped charge ΔN_{it} and oxide-trapped charge ΔN_{ox}. Although the contribution of ΔN_{ox} to device wearout is often neglected, recent experimental work recognizes that it can account for some of the observed degradation effects in NMOSFETs that could not be explained solely by ΔN_{it} generation.

The FaRBS HCI failure equivalent circuit model is illustrated in Figure 4.3. The derivation of ΔR_d is carried out under the assumptions that (1) all interface traps are acceptor-like and occupied by electrons and (2) channel mobility degradation, μ, is caused by both ΔN_{it} and ΔN_{ox}. The assumption (1) means that the net charge in interface traps is a fixed negative charge for NMOSFET in a strong inversion operation. The assumption (2) leads to the following equation:

$$\mu = \frac{\mu_0}{1 + \alpha \Delta N} \tag{4.15}$$

where $\Delta N = \Delta N_{it} + \Delta N_{ox}$ (in unit cm^{-2}), μ_0 is the original channel mobility, α is a process-dependent constant, and $\alpha \approx 2.4 \times 10^{-12} cm^2$ [118].

The charge in the conducting channel, $Q_{ch}(y)$, is modeled as:

$$Q_{ch}(y) = -C_{ox}(V_{gs} - V_t - \frac{q\Delta N}{C_{ox}} - V_{ch}(y)) \tag{4.16}$$

where C_{ox} is the gate oxide capacitance per unit area, $V_{ch}(y)$ is the potential along the channel, and y is the horizontal axis pointing to the drain and along the channel. All other parameters in Equation (4.16) assume their normal meaning.

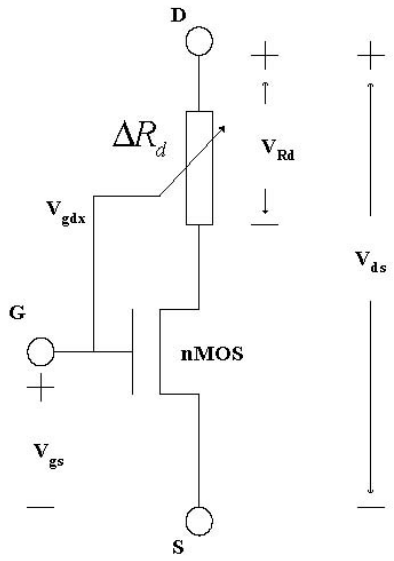

Fig. 4.3. *HCI failure equivalent circuit model in FaRBS. In the model* $V_{gdx} = V_{gs} - V_t - V_{ds}$ *and* $V_{Rd} = I_{ds}\Delta R_d$. V_t *is the threshold voltage and* I_{ds} *is the current from node D to S.*

Applying gradual channel approximation (GCA) and combining Equation (4.15), the drain current, I_{ds}, with inclusion of the hot carrier effect, is obtained as:

$$I_{ds} = \frac{\mu_0}{1+\alpha\Delta N} C_{ox} \frac{W}{L} (V_{gs} - V_t - \frac{q\Delta N}{C_{ox}} - \frac{V_{ds}}{2}) V_{ds} \tag{4.17}$$

Consider the failure equivalent circuit in Figure 4.3, where n-channel metal oxide semiconductor (NMOS) is the undamaged device with mobility μ_0 and threshold voltage V_t and is in series connection with ΔR_d, the current from node *D* to *S* can be obtained as:

$$I_{ds} = \mu_0 C_{ox} \frac{W}{L} (V_{gs} - V_t - \frac{V_{ds} - V_{Rd}}{2})(V_{ds} - V_{Rd}) \tag{4.18}$$

where V_{Rd} is voltage drop across ΔR_d. Combining Equation (4.17) and Equation (4.18) and then solving for V_{Rd} yields:

$$V_{Rd} = -V_{gdx} + \sqrt{V_{gdx}^2 + 2V_{ds}\Delta N[\frac{\alpha(V_{gdx} + \frac{V_{ds}}{2})}{1 + \alpha\Delta N} + \frac{q}{C_{ox}}]} \tag{4.19}$$

where $V_{gdx} = V_{gs} - V_t - V_{ds}$ for linear region and $V_{gdx} = 0$ for saturation region.

According to Equation (4.17), when $\Delta N = 0$ at $t = 0$, the undamaged drain current flows through NMOS, which is defined as I_{ds0}:

$$I_{ds0} = \mu_0 C_{ox} \frac{W}{L}(V_{gs} - V_t - \frac{V_{ds}}{2})V_{ds} \tag{4.20}$$

If ΔN is small, from Equation (4.17) and Equation (4.20), a simple relation between fresh and degraded drain-to-source current occurs:

$$I_{ds} = \frac{I_{ds0}}{1 + \alpha\Delta N} \tag{4.21}$$

Based on the above deduction, we finally obtain a function of ΔR_d that is determined by ΔN and terminal voltages and currents:

$$\Delta R_d = \frac{1 + \alpha\Delta N}{I_{ds0}} V_{Rd} \tag{4.22}$$

where I_{ds0} is given by Equation (4.20) and V_{Rd} is given by Equation (4.19). In a quasi-static operation, ΔN is a time-dependent parameter; therefore, ΔR_d is also time dependent. At any time, t, if ΔN is known, ΔR_d will be solely determined. The models for ΔN_{it} and ΔN_{ox} have been well documented in literature [45][108]. ΔN_{it} can be obtained from Equation (4.1) if technology-related constant C_1 is extrapolated from device testing. The models and model parameters for ΔN_{ox} are given in [108]. For convenience, they are recapitulated as follows.

The modeling of ΔN_{ox} starts from a simple injection current model, I_{ei}, which describes a one-dimensional process of electron injection into oxide based on quasi-elastic scattering assumption.

$$I_{ei} = \frac{1}{2}\frac{I_{ds}}{WL}\frac{t_{ox}}{\lambda_r}R^2 P_i(E_{ox})exp(-\frac{1}{R}) \tag{4.23}$$

where L is channel length, W is channel width, E_m is given by Equation (4.2), λ_r is re-direction mean-free path (λ = 61.6 *nm*), and t_{ox} is oxide thickness. $R = \lambda E_m/\varphi_b$, where λ is the scattering mean-free path of the hot electron (λ = 9.2 *nm*), and φ_b denotes the silicon and oxide energy barrier ($\varphi_b \approx 3.2$ *eV* for NMOSFET).

The most important term in Equation (4.23) is $P_i(E_{ox})$, which denotes the probability that a hot electron can enter the gate oxide by surmounting the surface potential barrier. An empirical expression for $P_i(E_{ox})$ is given as:

$$P_i(E_{ox}) = \frac{\alpha E_{ox}}{1 + E_{ox}/\beta} \times \frac{1}{1 + \frac{\gamma}{L}exp(-E_{ox}t_{ox}/1.5)} + \eta \tag{4.24}$$

where $E_{ox} = (V_{gs} - V_{ds})/t_{ox}$. Other model fitting parameters are given in [108]. Equation (4.24) is for the case $E_{ox} \geq 0$; if $E_{ox} < 0$, it is simplified to $P_i(E_{ox}) = \eta$.

Based on Equations (4.23) and (4.24), and for simulation purposes, a two-term kinetic equation is given in Equation (4.25) to model the relationship between oxide trapped charge density N_{ox} and electron injection current:

$$N_{ox} = N_1(1 - e^{-\sigma_1 I_{ei} t}) - N_2(1 - e^{-\sigma_2 I_{ei} t}) \tag{4.25}$$

A set of typical model fitting parameters for Equation (4.25) has been given in [108].

The above new ΔR_d model inherits all the merits of the HISREM model and is physically more comprehensive in characterizing hot carrier damages. The drawback of this improved ΔR_d model is the inclusion of one more parameter, ΔN_{ox}, which complicates parameter extraction work.

Currently, FaRBS does not provide PMOSFET HCI failure-equivalent circuit model because HCI physical effects on PMOSFETs are weaker than those of NMOSFETs. With further scaling of CMOS devices, PMOSFET might suffer from more pronounced HCI damage than ever before. In future work, FaRBS will include the PMOSFET HCI failure-equivalent circuit model based on the channel shortening theory and SPICE FaRBS models proposed in [119].

4.3 Time-Dependent Dielectric Breakdown

The TDDB defect-generation mechanism and device-wearout dynamics have been extensively investigated in the past. Many distinct, even controversial and contradicting, models have been proposed in literature. After many years of development, three successful models—the thermochemical model, the anode hole injection (AHI) model, and the voltage driven model—are singled out and have gained broad application.

The thermochemical model, also known as E model, assumes a direct correlation in existence between the electric field and the oxide degradation. The weak chemical bonds (*Si–Si* bonds) in SiO_2 associated with oxygen vacancies experience heavy strains due to the high electric field applied across the oxide. Some bonds might obtain enough thermal energy to break off and create defects or traps which, when accumulated to a large amount, will lead to oxide breakdown. According to the thermochemical model, if the logarithm of time-to-failure, t_f, is plotted against applied electric field E, a straight line will be observed; therefore, lifetime can be modeled as:

$$t_f = B_1 \exp(-\gamma E_{ox}) \exp(E_a/\kappa T) \tag{4.26}$$

where E_{ox} is an externally applied electric field across the dielectric in unit *MV/cm*, γ is field acceleration factor (with typical value of 1.1 decade per *MV/cm* [63], [120]), E_a is the thermal

activation energy ($E_a = 0.6 \propto 0.9\ eV$ [55]), and B_1 is a technology constant. The E model has been proved to provide a good fit to data from long-term, low-field TDDB stresses.

The AHI model assumes that gate oxide breakdown is triggered by the trapping of holes at localized regions in oxide, which either enhances the cathode field or leads to oxide electron trap generation, and increases the local current density, further facilitating local hole trapping and trap generation in a positive loop, and eventually leading to sudden breakdown of oxide [121]. The lifetime t_f function in an earlier version of the AHI model, derived a reciprocal electric field dependence ($1/E$) from the functional form of Fowler-Nordheim (FN) electron tunneling current, which is the driving force for oxide defect generation and impact ionization coeffcient in SiO_2. In this case, t_f can be approximated as:

$$t_f = B_2 \exp(\beta/E_{ox}) \exp(E_a/\kappa T) \tag{4.27}$$

where β is the electric field acceleration factor (with a typical value of 350 MV/cm) and B_2 is a process-dependent prefactor (the default value is $1 \times 10^{-11} s$) [55]. The $1/E$ model has been proven to provide a good fit to data from long-term, high-field TDDB stresses. It is important to note that the AHI model does not predict a strict $1/E$ dependence [121] and that there exists a model that predicts a much stronger $1/E$ effect ($t_f \propto \exp(\beta/E_{ox})(1/E^2_{ox})$ [122]).

Each of the two models (E and $1/E$) can only fit data in a limited range of the electric field, which might lead to significant errors in lifetime extrapolation if we exclusively use only one of them in reliability analysis. Researchers have proposed parallel competing models, i.e., combined models in terms of E and $1/E$ models trying to account for TDDB data in a larger electric field range [121] [123].

The applicability of E and $1/E$ models is mainly valid for oxides thicker than 5 nm where non-ballistic electron injection due to FN tunneling is dominant. When gate oxide thickness is smaller than 5 nm, e.g., ultrathin oxide, gate oxide lifetime dramatically shortens with the increase in direct tunneling current. In this situation, the validity of electric-field-driven models becomes problematic

because the injected electrons will travel ballistically through oxide without entering the oxide-conduction band, and the electron energy at the anode is controlled by the applied gate voltage [107]. This new phenomenon of electron injection in ultrathin oxides prompts the generation of voltage-driven breakdown models. The dependence of lifetime t_f on gate voltage V_{gs} is given by [124] [125]:

$$t_f = B_3 \exp(-\theta V_{gs}) \exp(E_a/\kappa T) \tag{4.28}$$

where θ is the voltage acceleration factor and B_3 is technology constant. The typical values of θ and activation energy E_a are given in [124][126].

All the TDDB lifetime models presented so far are based on exponential law for field or voltage acceleration, and Arrhenius law for temperature acceleration. Recent work shows that these two acceleration laws might not be accurate as gate oxide thickness scales below 5 nm; the extrapolation of ultrathin oxide lifetime with these exponential relations might produce erroneous or even absurd results. According to experimental data, the exponential law for time-to-breakdown voltage dependence cannot hold over a wide range of gate voltage; otherwise, the extrapolation of lifetime down to normal use conditions will predict that (1) the lifetime of smaller-area structures would be shorter than that of larger-area structures and that (2) the lifetime of thinner oxide devices would ultimately exceed that of thicker oxide devices, both of which are a contradiction to oxide degradation physics [127]. Therefore, new TDDB acceleration laws for voltage and temperature must be explored, commensurate with CMOS technology development.

Voltage and temperature dependencies and their interrelationship in oxide breakdown are critical factors for understanding ultrathin oxide reliability. More recent experimental data show that oxide time-to-breakdown evolution with temperature does not precisely follow an Arrhenius law: the activation energy increases with temperature. This behavior might be explained either by the non-thermochemical origin for the breakdown mechanism, or by a competing model involving two distinct mechanisms with different activation energies [125]. Wu et al. at IBM [128][69] proved with convincing data that the voltage dependence of time-to-breakdown follows a power law behavior rather than an exponential law, as had been commonly assumed. The ultrathin oxide

power law dependence of lifetime on gate voltage is consistent with the experimental facts that voltage exponential law acceleration factor θ (shown in Equation (4.28) and defined as $\theta = -\partial\, lnt_f /\partial V_{gs}$) is (1) temperature dependent at a fixed gate voltage, and (2) voltage dependent at a fixed temperature. Due to these new oxide time-to-breakdown voltage and temperature dependencies and the complicated interaction between voltage and temperature, TDDB lifetime modeling becomes more difficult than ever before. In another respect, however, the power law voltage dependence and non-Arrhenius temperature acceleration provide possible relief in circuit reliability margin. This margin continues to diminish with thinner oxide thicknesses [129].

The ultrathin oxide accelerated lifetime model in FaRBS is similar to the model proposed by Wu et al. at IBM [127][69] with some improvements, including the addition of oxide Poisson area scaling statistics and the cumulative failure percentile law. The original Wu model (power law voltage acceleration and non-Arrhenius temperature acceleration) has been implemented in the Reliability Aware Micro-Processor (RAMP) model jointly developed by UIUC and IBM for long-term processor-reliability prediction [13][130].

On the basis of extensive experimental investigation, ultrathin oxide lifetime dependence on voltage (power law acceleration) can be accurately captured by two simple empirical formulae [127][69]:

$$\frac{V_{gs}}{t_f}\frac{\partial t_f}{\partial V_{gs}} = n(T) \tag{4.29}$$

and

$$\frac{d}{dT}\left(\frac{1}{t_f}\frac{\partial t_f}{\partial V_{gs}}\right)\Big|_{t_f(\%)} = 0 \tag{4.30}$$

where $n(T)$ denotes the temperature-dependent voltage acceleration factor, T is the absolute temperature in Kelvin, and $t_f(\%)$ is the mean lifetime for a fixed cumulative percentile of failure (for example 63%).

Equation (4.29) reflects the power law dependence of time-to-breakdown on voltage: if $t_f = t_0 V^{n(T)}$, then $\partial t_f/\partial V = n(T) t_0 V^{n(T)-1} = n(T) t_f/V$, so $(V/t_f)(\partial t_f/\partial V) = n(T)$. Equation (4.29) also reflects the experimental fact that θ ($\theta = -\partial \ln t_f/\partial V$) is a voltage-dependent voltage acceleration factor: $\theta = -n(T)/V$. Equation (4.29) shows that voltage power law acceleration factor $n(T)$ is temperature dependent; for simplicity, we assume a linear relation $n(T) = a + bT$ (note: $n(T)$ should be always less than 0). This leads to the first part of the TDDB accelerated lifetime equation in FaRBS:

$$t_f \propto V_{gs}^{a+bT} \tag{4.31}$$

Equation (4.30) reflects the experimental fact that at a fixed accumulative failure percentile lifetime, the voltage exponential law acceleration factor θ is temperature independent. In reliability tests, we normally stress a large number of samples to a high cumulative percentile of failure (e.g., $F = 63\%$) and calculate lifetime at this percentile (e.g., $t_f(63\%)$); we then extrapolate lifetime to a low cumulative percentile of failure (e.g., $F = 0.01\%$) at normal use conditions. To take into account the effect of different cumulative failure percentiles being selected in use conditions for different devices from the same technology and tested at the same high cumulative-failure percentile, we need to incorporate Weibull statistics of oxide breakdown in accelerated lifetime model development.

According to the Weibull distribution, the cumulative failure probability $F(t)$ is:

$$F(t) = 1 - \exp[-(t/\alpha)^{\beta}] \tag{4.32}$$

where α is the characteristic life (i.e., lifetime at 63%) and β is the slope parameter that represents trends of failure rate. Weibull distribution is an extreme-value distribution in $ln(t)$ and can model weakest-link types of failure mechanisms. TDDB is a weakest-link mechanism because the first breakdown of any small portion in the gate oxide of a device will lead to the failure of the device and of the whole circuit [61].

Equation (4.32) can be rearranged and modified to:

$$t_f = \alpha[\ln \frac{1}{1-F}]^{\frac{1}{\beta}} \tag{4.33}$$

At normal use conditions, lifetime is often defined as the time to a very small cumulative percentile of failure (e.g., F = 0.01%); therefore, applying the logarithmic approximation law in Equation (4.33), we obtain the second part of the TDDB accelerated lifetime equation in FaRBS:

$$t_f \propto F^{\frac{1}{\beta}} \tag{4.34}$$

Another effect we have to consider in the TDDB accelerated lifetime model is that the gate oxide areas of sampled devices in accelerated tests are normally significantly different from those of devices in circuits. Experimental observations prove that the lifetime of TDDB is a function of the total gate oxide surface area due to the weakest-link character of oxide breakdown [127]. This gate oxide area effect has been modeled in [127][129][61], which is the third part of TDDB accelerated lifetime equation in FaRBS:

$$t_f \propto (\frac{1}{WL})^{\frac{1}{\beta}} \tag{4.35}$$

where W is the channel width and L is the channel length.

Finally, for the temperature acceleration effect, a non-Arrhenius model has been proposed in [127][69], which is the fourth part of TDDB accelerated lifetime equation in FaRBS:

$$t_f \propto \exp(\frac{c}{T} + \frac{d}{T^2}) \tag{4.36}$$

where c and d are voltage-dependent constants. In Equation (4.36), the second term d/T^2 empirically inserts non-Arrhenius temperature effects in the lifetime model.

Combining Equations (4.31), (4.34), (4.35), and (4.36), we can obtain a complete TDDB accelerated lifetime model for ultrathin oxide:

$$t_f = A_{TDDB}(\frac{1}{A})^{\frac{1}{\beta}} F^{\frac{1}{\beta}} V_{gs}^{a+bT} \exp(\frac{c}{T} + \frac{d}{T^2}) \tag{4.37}$$

where $A = W \times L$ is the device gate oxide area, β is Weibull slope parameter, F is cumulative failure percentile at use conditions (assuming the same cumulative failure percentile at test conditions), V_{gs} is gate-to-source voltage, T is temperature, a, b, c, d, and A_{TDDB} are model fitting parameters determined from experimental work. A set of typical values of these parameters is: β = 1.64, F = 0.01%, a = −78, b = 0.081, c = 8.81×10^3, and d = −7.75×10^5 [13][130].

It is important to note that Equation (4.37) is best applicable to cases when the gate oxide thickness is thinner than 5 nm (corresponds to 0.25-μm technology and beyond). If the gate oxide thickness is larger than 5 nm, to simplify parameter extrapolation work, Equation (4.28) should be used instead, with the default value of θ as 32. If the gate oxide thickness is much larger than 10 nm, the E or 1/E model (Equation (4.26) and (4.27), respectively) should be used depending on the magnitude of power supply voltage.

4.3.1 Failure Equivalent Circuit Model

It is challenging work to develop an effective equivalent circuit model for gate oxide breakdown because device post-breakdown behaviors are extremely complicated, often even perplexing. Device I−V characteristics after gate oxide breakdown rely on many parameters, including breakdown location, transistor type, voltage polarity, device operation mode (accumulation or inversion), oxide area, and even poly-gate doping type. Nevertheless, a literature review reveals an interesting phenomenon: TDDB failure equivalent circuit modeling is a very active area and more than a dozen circuit models have been developed by various research institutes and industrial labs. All of this work attempts to develop quantitative methodologies for predicting the response of circuits to device gate oxide breakdown events [131]. In this section, we

first review some of the most successful TDDB failure models, we then present the TDDB failure model adopted in FaRBS.

Starting with the observation that CMOS inverters' transfer curves under gate oxide stresses can be fitted by a combination of a threshold voltage shift (caused by charge trapping prior to breakdown) and a gate-drain leakage current model, which follows the form of a power-law relation as $I = KV^{p}_{gd}$, Rodriguez et al. at IBM [132, 133, 134] developed a simple TDDB damaged equivalent circuit model. The model consists of two voltage-dependent current sources bridging gate-to-drain and gate-to-source, respectively, which allow the oxide breakdown leakage current in a transistor to be simulated in a circuit. This power-law leakage current model is illustrated in Figure 4.4. The effects of gate oxide breakdown on the stability of SRAM cells and ring oscillators have been analyzed with this power-law leakage current model. Results show that for SRAM cells, oxide breakdown at different locations (drain, p-source, and n-source) leads to different trends in noise margin degradation, while for ring oscillators, oxide breakdown changes the loading of neighboring inverter stages and degrades the voltage-transfer characteristics [132].

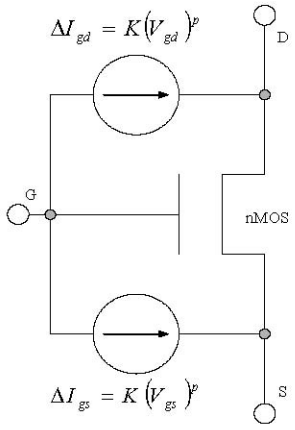

Fig. 4.4. *Power-law leakage current model. The exponent p varies from 5 to 2 as the degredation level increases. K reflects the "size" of the breakdown spot.*

Rodriguez et al. [134] noted that a linear ohmic oxide breakdown resistance is not sufficient to model the experimental data. The ohmic model provides good results only for hard breakdown; however, the power-law leakage current model predicts progressive oxide breakdown behaviors more effectively prior to the final hard breakdown.

In a MOSFET, the oxide breakdown changes isolations of the device's internal structures by forming an abnormal conduction path. This effect can be modeled with parasitic ohmic or rectifying device elements, depending on the relative doping of the internal structures being shorted. Based on the facts that oxide post-breakdown behavior depends on breakdown location (gate-to-substrate, gate-to-drain, and gate-to-source), transistor type (NMOSFET and PMOSFET) and poly-gate doping type (n^+ poly-gate and p^+ poly-gate), Segura et al. [135][136] developed a complete set of gate oxide short (GOS) electrical models (altogether 12 different GOS models) to account for all combinations of these location and doping effects. Among these models, the most important one is the model for gate-to-substrate breakdown of NMOSFET with n^+ poly-gate. For this kind of device, the gate-to-substrate breakdown path between n^+ poly-gate and n type inversion channel can be modeled as a gate-to-substrate resistance R_{GOS}. The formation of this resistance-like breakdown path splits the whole channel into two parts, which is physically equivalent to two transistors connected in series. This model is illustrated in Figure 4.5. For other combinations of location and doping effects, the models can be readily deduced with a similar principle. For example, when the breakdown path appeared between the gate and the drain (or the source) terminals of the NMOSFET, a $n^{++} - n^+$ barrier (n^+ poly-gate to n^+ drain/source diffusion) forms. In this case, the breakdown is modeled with a resistance between gate-to-drain/source.

With these GOS electrical models, Segura et al. [135] explored testing considerations at the circuit level to sensitize GOS under various logic fault situations (stuck-at, stuck-open, and stuck-on faults) and concluded that GOS does not behave as a bridge in normal cases and stuck-at based Automatic Test Pattern Generation might not detect GOS depending on the gate topology.

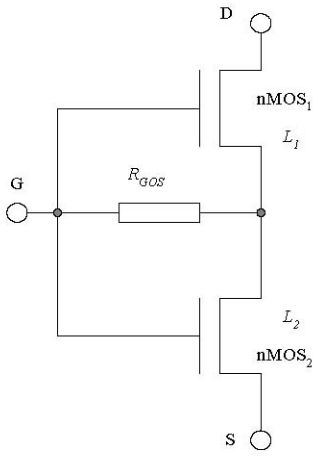

Fig. 4.5. *TDDB GOS model for gate-to-substrate breakdown of NMOSFET with n^+-poly gate. The channel lengths of $NMOS_1$ and $NMOS_2$ follow the relation: $L_1 + L_2 = L$ where L is the undamaged NMOSFET channel length. The parameter R_{GOS} is related to the size and location of the breakdown path. A value of R_{GOS} as low as 3KΩ was used in the simulation in [136].*

Gate oxide breakdown equivalent circuit models for analog circuits and RF circuits are also developed in an attempt to expand model applicability and explore the oxide breakdown effect beyond digital circuits. For typical analog circuits, oxide breakdown changes parameters of transistors in differential pairs in an asynchronous way and, therefore, leads to mismatches that prompt the offset generation and compromise circuit functionality [137]. As for RF circuits, they are very sensitive to device parameter drift; therefore, oxide breakdown is expected to have a more severe impact on their functionality and performance [138].

Yang et al. [138] [139] developed an equivalent RF circuit model for gate oxide breakdown and investigated the effect of TDDB on a Low Noise Amplifier (LNA) circuit. This RF equivalent model is shown in Figure 4.6, which consists of the original NMOSFET, the terminal series resistances (R_G, R_D, R_S), the substrate parasitic resistances (R_{DB}, R_{SB}, R_{DSB}), gate overlap parasitic capacitances (C_{GDO}, C_{GSO}), the junction capacitances (C_{jDB}, C_{jSB}), and the two inter-terminal resistances (R_{GD}, R_{GS}). R_G and the "H" type substrate *RC* network are included for more accurate

RF modeling. The two resistances R_{GD} and R_{GS} vary in opposite directions, representing different breakdown locations along the channel from source to drain. If one of them is significantly smaller than the other, breakdown is gate-to-source or gate-to-drain depending on which resistance is dominant.

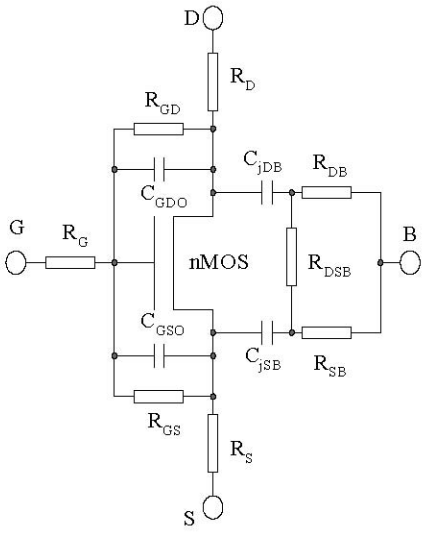

Fig. 4.6. *TDDB RF equivalent ciruit model. Model parameters for simulation in [138] are set as:* $R_G = 85.4\Omega$, $R_D = R_S = 12.14\Omega$, $R_{GD} = 6.88K\Omega$, $R_{GS} = 23K\Omega$, $C_{GDO} = C_{GSO} = 15.3fF$, $C_{jDB} = C_{jSB} = 7fF$, $R_{DSB} = 80K\Omega$, $R_{DB} = R_{SB} = 49.37\Omega$..

Based on this TDDB RF circuit model, the performance degradation of 0.16-μm NMOSFET devices and a 1.8-GHz LNA circuit is analyzed [138]. For the device *S*-parameters, the inclusion of R_{GD} and R_{GS} changes device input impedance *S*11; provides an additional connection between gate and drain, and, therefore, degrades reverse transmission coefficient *S*12; changes the output impedance *S*22 at the drain; and decreases transconductance g_m, which is equivalent to forward transmission coefficient *S*21. For the LNA circuit, oxide breakdown has a significant impact on its performance: most *S*-parameters drift dramatically and fail to meet usual performance requirements; input impedance matching is disturbed due to increased gate leakage current, and

noise figure obviously deteriorates with the breakdown path formation across the gate oxide, which adds another noise source to the transistor.

Until now, the most frequently discussed TDDB failure equivalent circuit model is the one proposed by Kaczer et al. at IMEC [140, 141, 142, 143, 144, 145]. In this model, the breakdown path is assumed to be formed by n-type silicon, and a microscopic structure of the device is explored to investigate the exact configuration and connection of device internal parts after gate oxide breakdown. For an NMOSFET (n^+ poly-gate/p substrate/n^+ drain and source diffusion) with an oxide breakdown path formed between gate and substrate, if the gate voltage is negative ($V_G < 0$), the device is in accumulation state and no inversion layer is developed below the $Si - Si_2$ interface. The contact region of the breakdown path (n-type) and the substrate (p-type) is a forward biased pn junction. Electrons emit from n^+ poly-gate, flow through n-type breakdown path, diffuse along the substrate, and are collected by the source and the drain junctions. This mechanism is exactly that of a bipolar transistor with an emitter at the breakdown path, a base at the substrate, and a collector at the source and the drain. Therefore, NMOSFET with oxide breakdown and operated at negative gate voltage can be modeled with a gate resistor, two bipolar transistors, and the original NMOSFET [140][142]. Because NMOSFETs rarely operate in a negative gate voltage bias situation, this complicated two-bipolar-transistor model for ($V_G < 0$) is not of primary interest.

When gate voltage is positive enough such that NMOSFET is in a strong inversion state, an n-type conduction channel will form under the gate oxide connecting the source and the drain. Note that, under these conditions, the contact region of the breakdown path (n-type) and the channel (n-type) is an ohmic connection. The positive gate voltage forces an electric field to penetrate through the breakdown path and deplete the contact region of the breakdown path and substrate. This contact region serves as an electron sink and, therefore, can be treated as an additional drain in the middle of the channel. Based on this microscopic example, an equivalent electrical circuit for NMOSFET with a hard gate oxide breakdown and operated in positive gate voltage can be constructed. This is illustrated in Figure 4.7. Apart from the original NMOSFET (NMOS), the model contains a constant resistance (R_G) corresponding to the breakdown path, two adjacent parasitic NMOSFETs (M_S and M_D, characterized by level-1 SPICE models), and two resistors (R_S

and R_D), characterizing the resistance in the source and the drain extensions, respectively. The effects of the breakdown location are represented by varying the gate lengths of M_S and M_D. Gate-to-substrate breakdowns in the vicinity of the drain or the source are represented by logarithically varying extension resistances R_S or R_D [140]. For gate-to-source (or gate-to-drain) breakdowns, the model can be simplified to a circuit containing only R_G, R_S (or R_D), and the original NMOS transistor.

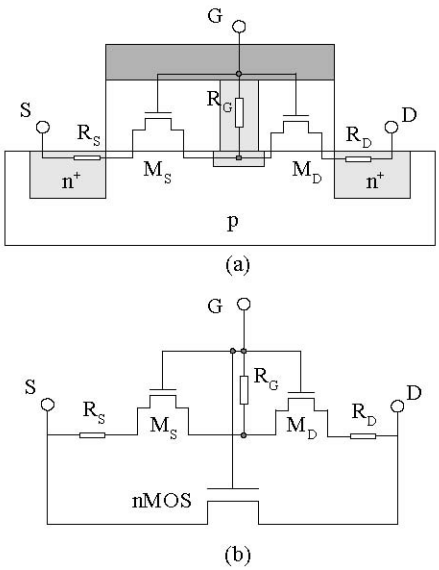

Fig. 4.7. *TDDB equivalent circuit model for NMOSFET with hard gate oxide breakdown and operated in positive gate voltage. (a) Cross section view of breakdown structure. (b) Equivalent circuit model. Model parameters for simulation in [140] are set as R_G = 1KΩ, L_{MS} + L_{MS} = 0.09 μm, W_{MS} = W_{MS} = 0.25 μm, R_D and R_S vary from 2.5KΩ (at source and drain) to 12.5KΩ (at the middle of the channel).*

This model has been used in a CMOS ring oscillator oxide breakdown analysis [141]. The simulation shows that gate-to-substrate breakdowns have a minor effect on circuit operation; however, breakdowns at the very edges of the gate significantly damage the circuit performance. This observation reveals that progressive breakdown (i.e., soft breakdown) occurs mainly in the transistor channel, while the hardest, circuit-killing breakdowns occur above the source and the drain extension regions [143]. This conclusion can be explained with the help of the Kaczer model:

in the extension regions where contact resistances are low, the power dissipation during the breakdown is very high and leads to accelerated wearout of the breakdown path. This corresponds to hard breakdown behaviors: that is, if a breakdown happens in the transistor channel region, where resistance (i.e., channel resistance) of the discharge path is higher, a soft breakdown will be triggered.

Even though a lot of work has been done to maturate this model, careful evaluations in [138] and our critical examinations have identified several limitations of this Kaczer model: (1) the level-1 M_S and M_D models are obsolete; (2) the model only applies to linear operation situation (that is, if a breakdown path forms above the saturation region where the channel has "pinched-off," the inclusion of the two parasitic transistors (M_S and M_D) is not valid); (3) M_S and M_D bring two more diffusion regions, which do not physically exist; (4) a simulator cannot handle the breakdown position from zero to the whole channel length; (5) it is problematic to preserve the original NMOSFET in the model if M_S and M_D are already included because they have represented all device internal structures after oxide breakdown; and (6) the prime assumption that the breakdown path is *n*-type silicon is arbitrary and not physically justified. The last two points are most important, and they prompted us to develop a physically justifiable circuit model for gate oxide breakdown.

Besides those that have been briefly reviewed above, there are many other successful models worth mentioning [146, 147, 148, 149, 150, 151]. A PMOSFET gate-to-channel short model is proposed in [146] and is used to investigate its effect on logic gate failures. A pair of breakdown models for NMOSFET and PMOSFET (only gate-to-diffusion breakdown) is proposed and used to transform the effect of oxide breakdown into a delay fault or a logic fault [147]. Yeoh et al. [148][149] conducted a thorough investigation of oxide breakdown modes and developed a set of complex models by combining resistors, diodes, and transistors in different ways to model device internal connections after the oxide breakdown path formed at different locations. Based on the work of a linear non-split MOS model and a non-linear two-dimensional channel split MOS model [150], a non-linear non-split MOS oxide breakdown model is developed in [151] in an attempt to

enable circuit simulation of the gate-to-channel effect on minimum length transistors. Even though none of these models is superb, the development of fundamental concepts, physical principles, and modeling techniques in these models is the foundation for constructing any advanced oxide breakdown circuit models. A vivid example is the improved TDDB failure equivalent circuit model adopted in FaRBS, which is presented as follows.

From a semiconductor materials point-of-view, we should not assume the breakdown path as *n*-type silicon diffusion because this is not physically substantiated and the oxide breakdown path is actually defect-assisted electron conduction rather than a reliable physical connection. Therefore, we could not use only resistance to model gate-to-substrate and gate-to-diffusion breakdowns. The correct modeling method should be based on the channel potential re-distribution concept. The oxide breakdown path disturbs device channel surface potential in the vicinity below the breakdown path, where GCA is broken; therefore, a new three-dimensional channel potential model must be developed for this purpose. According to [136], if we define a three-dimensional coordinate system in terms of the gate oxide surface with x along the channel length L direction from source to drain, y perpendicular to the gate oxide, and z along the channel width W direction, the contact point of the breakdown path to channel surface can be defined as: $x = L_1$, $y = 0$ and $z = W_1$ (refer to Figure 10 in [136]). The drain current I_D of a defect-free MOSFET can be obtained from:

$$I_D = \frac{W}{L}[f(\Psi(x = L)) - f(\Psi(x = 0))] \tag{4.38}$$

where $\Psi(x)$ is the channel surface potential at x and f is a function of channel mobility, oxide capacitance, threshold voltage, and device terminal voltages.

If the breakdown defect located at ($x = L_1$, $y = 0$, and $z = W_1$) is considered, the two-dimensional channel can be divided in two regions and, similar to Equation (4.38), the drain and source currents of the damaged MOSFET can be written as [136]:

$$I_D = \frac{W}{L - L_1}[f(\Psi(x = L)) - f(\Psi(x = L_1))] \tag{4.39}$$

and

$$I_D = \frac{W}{L_1}[f(\Psi(x = L_1)) - f(\Psi(x = 0))] \tag{4.40}$$

where $\Psi(x = L_1)$ is the surface potential under the breakdown path. Equations (4.39) and (4.40) show that an NMOSFET with gate oxide breakdown is equivalent to the series connection of two devices with gate geometries of (W, L_1) and (W, $L - L_1$).

No matter what the breakdown path is made of, its electrical effect is that it provides a conduction path to inject electrons from the gate into the channel; therefore, we can use a voltage-dependent current source I_{OX} connecting between the gate and the channel to model this effect. Based on the above discussion, a new TDDB failure equivalent circuit model is obtained and illustrated in Figure 4.8. It seems this model requires two model parameters (L_1 and V_i, which is voltage at the connection point of M_1 and M_2); however, with some practical simplifications, V_i can be reduced to a function dependent on L_1. Therefore, there is only one independent model parameter left and requiring characterization, which facilitates the application of this model.

Suppose the original drain-to-source current of a fresh NMOSFET is I_{DS0}, and neglecting the effect of R_D, R_S, and short-channel effect (to simplify equation derivation), we can write I_{DS0} as:

$$I_{DS0} = \mu_n C_{ox} \frac{W}{L}[(V_{GS} - V_t)V_{DS} - \frac{1}{2}V_{DS}^2] \tag{4.41}$$

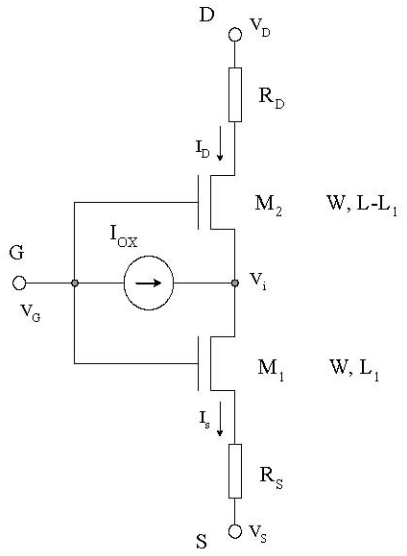

Fig. 4.8. *FaRBS TDDB equivalent circuit model NMOSFET with hard gate oxide breakdown. $I_{OX} = I_S - I_D$ is a voltage-dependent current source representing breakdown path current injection effect. R_D and R_S characterize the resistance in the source and the drain extensions, respectively. L_1 represents breakdown location in terms of the source edge.*

Applying Kirchhoff's Current Law (KCL) to Figure 4.8 (for simplicity, neglect R_D and R_S), we derive the following equations:

$$I_{OX} = I_S - I_D \tag{4.42}$$

$$I_D = \mu_n C_{ox} \frac{W}{L_2}[(V_G^{'} - V_i)(V_D - V_i) - \frac{1}{2}(V_D - V_i)^2] \tag{4.43}$$

$$I_S = \mu_n C_{ox} \frac{W}{L_1}[(V_{GS} - V_t)(V_i - V_S) - \frac{1}{2}(V_i - V_S)^2] \tag{4.44}$$

where $L_2 = L - L_1$ is the channel length of M_2, $V_G = V_G - V_{t2}$ and V_{t2} is the threshold voltage V_t plus body effect induced enhancement with source-to-substrate bias of $V_{sb} = V_i$. V_i represents the channel potential at the breakdown location.

The main effects of gate oxide breakdown on device characteristics are abrupt gate current and substrate current generation; as a result, gate voltage cannot control and sustain channel current as strong as before, which leads to degradation of drain current. Therefore, a good assumption in Figure 4.8 is that the source current I_S maintains its value as before, whereas injection of I_{OX} degrades I_D current at the drain. This means $I_S = I_{DS0}$. So from Equation (4.41) and Equation (4.44), we can solve for V_i:

$$V_i = V_{G_{on}} - \sqrt{V_{G_{on}}^{\prime 2} - (V_S^2 + 2V_{ov}V_S + \frac{2I_{DS0}L_1}{\mu_n C_{ox} W})} \tag{4.45}$$

where $V_{Gon} = V_G - V_t$ and $V_{ov} = V_{Gon} - V_s$ is the gate overdrive voltage. If V_S is tied to ground, Equation (4.45) is reduced to:

$$V_i = (V_G - V_t) - \sqrt{(V_G - V_t)^2 - \frac{2I_{DS0}L_1}{\mu_n C_{ox} W}} \tag{4.46}$$

Equation (4.46), or Equation (4.45) if $V_S \neq 0$, shows that V_i is solely determined by L_1. Therefore, the number of model parameters is reduced from two to only one. If the breakdown location parameter L_1 is characterized from experimental work, from Equations (4.41) ~ (4.46), the voltage-dependent current source I_{OX} can be obtained.

The above NMOSFET TDDB failure equivalent circuit model can be easily extended to PMOSFET by properly changing current flowing directions in Figure 4.8 and voltage/current signs in model equations.

4.4 Negative Bias Temperature Instability

As process technology develops, gate oxides are becoming much thinner in the deep submicron regimes and experience an increased oxide electric field, which is one of the major incentives for

NBTI effects. Nitrogen is commonly introduced in PMOSFET's oxide to prevent boron diffusion, increase dielectric constant, suppress gate leakage current, and improve hot carrier immunity. However, the inclusion of nitrogen in processes exacerbates NBTI effects [81] [152].

In device physics, NBTI becomes a more important reliability concern as device feature sizes continue to shrink. Interface traps and oxide traps generated from the dissociation of interface $Si{-}H$ bonds increase carrier surface-related scattering and disturb the local electric field in oxide, leading to channel mobility degradation and threshold voltage shift. The electrical effects of NBTI influence on PMOSFETs manifest in decreasing saturated drain current (I_{dsat}) and transconductance (g_m), increasing threshold voltage (V_t), and temporarily decreasing off-state current [81] [152].

During circuit operation, NBTI is different from HCI in that HCI stresses devices only during the dynamic switching periods when current flows through the device, whereas NBTI stresses devices even when they are in static state operation [153][154]. The different stress time windows of HCI and NBTI in an inverter VTC plot and an input-output waveform plot are illustrated in Figure 4.9, which shows that the PMOSFET suffers from NBTI stress when the inverter input voltage is low and output voltage is high. In contrast, the PMOSFET only experiences HCI stress during the inverter output pulling-up period when C_o is charging up, while the NMOSFET suffers from HCI stress during the opposite dynamic stage when the inverter output is discharged to a low voltage level [155]. The fact that NBTI has a much larger stress time window, which even extends to device steady state operation periods, leads to the obvious result that duty cycle has a much more severe effect on the NBTI mechanism. This complicates circuit NBTI behaviors and compels us to address NBTI effects in the circuit design stages.

The most obvious NBTI-induced device degradation phenomenon is the threshold voltage shift $\Delta V_t(t)$; therefore, in developing NBTI lifetime models, $\Delta V_t(t)$ is unanimously used as an NBTI degradation monitor to characterize device wearout degree and, accordingly, time to a fixed $\Delta V_t(t)$ value (e.g., 50 mV or 100 mV) is often defined as the NBTI lifetime. Due to the electrochemical reaction-diffusion processes in NBTI, the time dependence of $\Delta V_t(t)$ is both mathematically

derived and experimentally observed to follow a fractional power law relation $\Delta V_t(t) \propto t^n$, where the exponent n ranges from 0.15 ~ 0.3 with a typical value of 0.25 [153][84]. The fractional value of n gives rise to a saturation behavior at a long time t, which conforms to experimental observations.

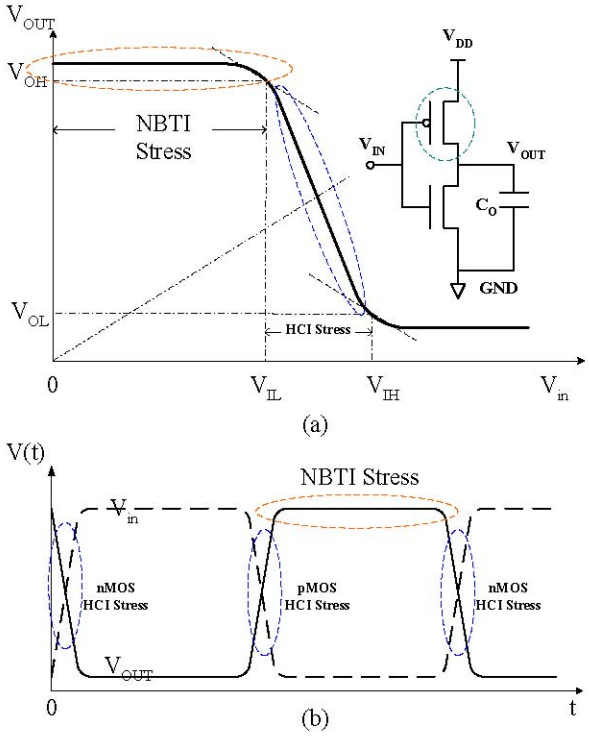

Fig. 4.9. *The different stress time windows of HCI and NBTI for an inverter in (a) voltage transfer curve (VTC) plot and (b) input-output waveform plot. HCI stresses devices only during the dynamic switching periods when both gate voltage and drain voltage must be high enough and there is current flowing through the device. NBTI stresses PMOS devices mainly during the period when they are in one of the two static operation states when gat voltage is negative with respect to both drain and source voltages.*

The voltage dependence of $\Delta V_t(t)$ is phenomenologically modeled with an exponential law $\Delta V_t(t) \propto \exp(\beta V_G)$ [156, 157, 158]. The temperature dependence of $\Delta V_t(t)$ is empirically modeled

with the well-known Arrhenius law $\Delta V_t(t) \propto \exp(-E_a/\kappa T)$ [157][158]. In terms of all the above relations, $\Delta V_t(t)$ is often given as [159][84]:

$$\Delta V_t(t) \propto \exp(\beta V_G)\exp(-E_a/\kappa T)t^n \tag{4.47}$$

NBTI lifetime t_f corresponds to the time to a fixed $\Delta V_t(t)$ value; therefore, by rearranging Equation (4.47), we obtain a frequently used NBTI accelerated lifetime model:

$$t_f = A_0 \exp(-\beta^{'} V_G)\exp(E_a^{'}/\kappa T) \tag{4.48}$$

where $\beta = \beta/n$, $E_a = E_a/n$, and A_0 is a process constant.

In deriving Equation (4.48), the assumption of exponential law for voltage dependence is not justified and does not fit recent experimental data very well. Therefore, we have to develop a more suitable acceleration law for NBTI voltage dependency. In our previous work, we developed a power law voltage acceleration model in [15]. A phenomenological DC model suggests that shifts in threshold voltage result from the increase in positive fixed charge $\Delta N_f(t)$ and the generation of donor type interface traps in the lower half of the silicon bandgap $\Delta N_{it}(t)$ [160]:

$$\triangle V_t(t) \propto \frac{q}{C_{ox}}(\triangle N_f(t) + \triangle N_{it}(t)) \tag{4.49}$$

where C_{ox} is the oxide capacitance. For ultrathin oxide, $\Delta N_f(t)$ and $\Delta N_{it}(t)$ are determined by temperature T, oxide electric field E_{ox}, oxide thickness t_{ox}, and stress time t:

$$\triangle N_{it}(t) \propto E_{ox}^m t^{n_1} \frac{1}{t_{ox}} \exp(-\frac{E_{a_1}}{\kappa T}) \tag{4.50}$$

and

$$\triangle N_f(t) \propto E_{ox}^m t^{n_2} \exp(-\frac{E_{a_2}}{\kappa T}) \tag{4.51}$$

where n_1 = 0.25, E_{a1} = 0.2 *eV* for $\Delta N_{it}(t)$, and n_2 =0.14, E_{a2} = 0.15 *eV* for $\Delta N_f(t)$, respectively, and m = 1.5 for both cases [81]. Equation (4.50) shows thickness-dependence of $\Delta N_{it}(t)$ on t_{ox}, while Equation (4.51) means $\Delta N_f(t)$ is thickness-independent. These dependencies prompt us to make an assumption that for smaller t_{ox}, $\Delta N_{it}(t)$ will dominate over $\Delta N_f(t)$ in Equation (4.49) (this assumption is supported in [161]). Substituting Equation (4.50) into Equation (4.49) and neglecting $\Delta N_f(t)$, C_{ox} and t_{ox} will cancel each other (because $C_{ox} = \varepsilon_{ox}/t_{ox}$) in Equation (4.49). If we replace the oxide electric field E_{ox} with the gate bias voltage V_{gs} (for p+ poly-Si gate PMOSFETs, $E_{ox} = (V_{gs} - 0.2\ V)/t_{ox} \approx V_{gs}/t_{ox}$ according to Equation (22) in [81]), then we get a new NBTI accelerated lifetime model:

$$t_f = A_1 (\frac{1}{V_{gs}})^{\gamma} \exp(\frac{E_a}{\kappa T}) \tag{4.52}$$

where E_a is activation energy, A_1 is a process-related constant, and γ is the voltage acceleration factor. This voltage power law relation is also reported in [162]. In literature, the typical value of E_a is reported as 0.9 ~1.2 *eV*, and the γ value is approximately 6–8 [162][163].

Quick development of NBTI testing and analyzing techniques have discovered some new phenomena of NBTI effects, including dynamic recovery effect [156][157][84] and a $\Delta V_t(t)$ saturation effect [157]. These new phenomena demand new physics-based lifetime models to account for and predict NBTI impact on circuit performance and functionality. Based on the model proposed by Zafar [164] [165], we developed a new NBTI accelerated lifetime model that is based on degradation physics and statistics mechanics. This new model provides a new statistical explanation for the $\Delta V_t(t)$ saturation effect and a physical explanation for dynamic recovery effect in the same framework. Based on the same Zafar model, we also developed a new NBTI failure equivalent circuit model that is the first electrical model in this area for modeling NBTI effect on circuit functionality.

According to [165], by applying statistical mechanics to calculating the decrease in interfacial *Si–H* density as a function of stress conditions, we can mathematically derive a new time dependence of $\Delta V_t(t)$ as:

$$\Delta V_t(t) = \Delta V_{max}[1 - e^{-(\frac{t}{\tau})^{\beta}}] \tag{4.53}$$

where ΔV_{max}, τ, and β are three model parameters. The parameter ΔV_{max} is the maximum $\Delta V_t(t)$ shift that would occur when all the interfacial *Si–H* bonds have been de-passivated. The parameter τ is the time when $\Delta V_t(t)$ increases to 63.2% of ΔV_{max} and, therefore, is a measure of the NBTI degradation rate. The parameter β ($0 < \beta < 1$) is a measure of dispersion in hydrogen diffusion; this value decreases from 1 to 0 as dispersion increases. β is independent of stress oxide field E_{ox} [165].

τ and ΔV_{max} have been derived in [165] as:

$$\tau = B_1 E_{ox}^{-\frac{1}{\beta}} \tag{4.54}$$

and

$$\Delta V_{max} = B_2[\frac{1}{1 + 2\exp(-\frac{E_1}{\kappa T})} + \frac{1}{1 + 2\exp(-\frac{E_2}{\kappa T})}] \tag{4.55}$$

where B_1 and B_2 are model prefactors. E_1 and E_2 are material and oxide electric field-dependent parameters. Their values are given as:

$$E_1 = E_{it} - E_g + E_F \tag{4.56}$$

and

$$E_2 = E_{fx} - E_F + \gamma E_{ox}^{\frac{2}{3}} \tag{4.57}$$

where E_{it} and E_{fx} are trap energy level at the oxide/*Si* interface and trap energy in the oxide, respectively; E_F is Fermi energy with respect to valence band edge in bulk *Si*; E_{ox} is applied electric field across the oxide; γ is a constant; and $\gamma E^{2/3}_{ox}$ represents the decrease in the electronic energy

due to band bending in the substrate. A set of typical values for these parameters is given in [165]: $E_{it} = 0.24\ eV$, $E_{fx} = -0.16\ eV$, $E_g = 1.12\ eV$, $E_F = 0.98\ eV$, $\gamma = 6.64\times10^{-7}$. Based on these values, we obtain $E_1 = 0.10\ eV$ and $E_2 = 0.14\ eV$ (if we assume $V_{ox} = 1V$ and $t_{ox} = 10\ nm$). E_1 is a process-determined parameter, while E_2 is a circuit-operation-dependent parameter due to the fact that V_{ox} is a function of V_{gs}.

If we define $F(t) = \Delta V_t(t)/\Delta V_{max}$, then we can rewrite Equation (4.53) in the form:

$$F(t) = 1 - e^{-(\frac{t}{\tau})^{\beta}} \tag{4.58}$$

Equation (4.58) is exactly the same as the Weibull function. If we define $f(t) = \partial F(t)/\partial t$, then $f(t)$ represents the rate-of-change in $\Delta V_t(t)$ (normalized to ΔV_{max}). Based on the above transformations, we can explain NBTI time-dependent degradation behaviors (power law at initial period followed by a gradual saturation effect) with Weibull statistics. When $(t/\tau)^{\beta}$ is very small (corresponds to initial NBTI stress), with the mathematical approximation $1 - e^{-x} \approx 1 - x$, Equation (4.58) can be simplified to:

$$F(t) = 1 - e^{-(\frac{t}{\tau})^{\beta}} \approx [1 - (1 - (\frac{t}{\tau})^{\beta})] = (\frac{t}{\tau})^{\beta} \propto t^{\beta} \tag{4.59}$$

Equation (4.59) shows that, at the initial state, NBTI-induced $\Delta V_t(t)$ follows a power law time dependency.

From Weibull statistics, we know that if the slope parameter β is smaller than 1, the probability density function $f(t)$ will decrease with time t. In Equation (4.58), β is always smaller than 1 (i.e., $0 < \beta < 1$), which means that the rate-of-change in $\Delta V_t(t)$ (normalized to ΔV_{max}) will decrease with time. Therefore, after a very-long-time t, $\Delta V_t(t)$ will gradually saturate. The above Weibull equivalent explanations justify the validity of Equation (4.53) from a statistical point-of-view. From Equation (4.53), we can derive a new NBTI accelerated lifetime model that sufficiently explains NBTI dynamic recovery effects.

Rearranging Equation (4.53) and solving for time *t*, we obtain:

$$t = \tau[\ln \frac{1}{1 - \frac{\Delta V_t(t)}{\Delta V_{max}}}]^{\frac{1}{\beta}} \tag{4.60}$$

Substituting Equation (4.54) into Equation (4.60), we obtain:

$$t = B_1 E_{ox}^{-\frac{1}{\beta}}[\ln \frac{1}{1 - \frac{\Delta V_t(t)}{\Delta V_{max}}}]^{\frac{1}{\beta}} \tag{4.61}$$

The relations between E_{ox} and applied gate voltage V_{gs} is given as (according to Equation (21) in [81]):

$$E_{ox} = \frac{V_{gs} - V_{FB} - \phi_s}{t_{ox}} \approx \frac{V_{gs} - 0.2V}{t_{ox}} \tag{4.62}$$

where V_{FB} is flat-band voltage and Φ_s surface potential. Equation (4.62) can be written in a general form as:

$$E_{ox} \propto V_{gs} - \alpha \tag{4.63}$$

where α is a technology-related potential constant with typical value of 0.2 V for PMOSFETs with *p*+ poly-gate.

Equation (4.61) can be rewritten to Equation (4.64) by subbing with Equation (4.63):

$$t = B_1 (V_{gs} - \alpha)^{-\frac{1}{\beta}}[\ln \frac{1}{1 - \frac{\Delta V_t(t)}{\Delta V_{max}}}]^{\frac{1}{\beta}} \tag{4.64}$$

According to the mathematical approximation that if x is very small, then $\ln[1/(1-x)] \approx x$ (e.g., if $x = 0.1$, $\ln[1/(1-x)] = 0.1054$ and the relative error is only 5.4%), we can further simplify Equation (4.64). Because most device service times at normal use conditions are much shorter than devices' end-of-life lifetimes, we can intuitively assume $\Delta V_t(t)/\Delta V_{max}$ to be a very small quantity (the $1/\beta$ exponent of it tends to further shrink the difference between $\ln[1/(1-x)]$ and x). Therefore, Equation (4.64) is reduced to:

$$t = B_1(V_{gs} - \alpha)^{-\frac{1}{\beta}}\left[\frac{\Delta V_t(t)}{\Delta V_{max}}\right]^{\frac{1}{\beta}} \tag{4.65}$$

Substituting Equation (4.55) into Equation (4.65) and neglecting the effect of α on V_{gs} shift (if V_{gs} is much larger than 0.2 V), we obtain a new physics and statistics-based NBTI accelerated lifetime model:

$$t_f = A_{NBTI}V_{gs}^{-\frac{1}{\beta}}\left[\frac{1}{1+2\exp(-\frac{E_1}{\kappa T})} + \frac{1}{1+2\exp(-\frac{E_2}{\kappa T})}\right]^{-\frac{1}{\beta}} \tag{4.66}$$

where the typical value of β is 0.3 [165] and E_1 and E_2 are given by Equation (4.56) and Equation (4.57), respectively.

Equation (4.66) is the mathematical transformation of Equation (4.53); therefore, it inherits all the merits of Equation (4.53). This means our new NBTI accelerated lifetime model inherently accounts for NBTI $\Delta V_t(t)$ power law and saturation behaviors having been discussed before. Another main feature of this new model is its accountability for NBTI dynamic recovery and AC effects. Traditional NBTI analysis neglects these significant new effects observed from the latest experimental work that lead to relaxed NBTI degradation [156][87]. If these effects are not considered, an overly pessimistic NBTI lifetime will be extrapolated. In dynamic digital circuit operations and analog circuit AC operations, NBTI effects can be treated as a two-step stress process: a high-stress period and a low-stress recovery period. According to our new NBTI accelerated lifetime model, E_2 is voltage dependent (Equation (4.57)); therefore, E_2 will be larger at high-stress periods and smaller at low stress periods. According to Equation (4.66), a higher E_2

leads to a shorter t_f, and a lower E_2 leads to a longer t_f. The final t_f for the whole process is a result of these two processes. Therefore, Equation (4.66) provides a prediction method for NBTI dynamic recovery and AC effects. The above discussion concludes that the new NBTI accelerated lifetime model outperforms other peer models (Equation (4.48), Equation (4.52), and the model in [166]) in that it accounts for nearly all known NBTI effects in a unified framework for reliability analysis.

4.4.1 Failure Equivalent Circuit Model

To date, the majority of the work of NBTI investigation has been concentrated on discrete transistor parameter drift, rather than on circuit performance degradation [154][167]. Recently, the interest of NBTI literature has been gradually elevated to characterize impacts of NBTI on digital circuit reliability [153][154][167, 168, 85, 169] and on analog and RF circuit reliability [170, 171, 172]. Reddy et al. [154][167] developed an NBTI circuit degradation model to investigate the first-order impact of NBTI-induced PMOSFET degradation on ring oscillator and SRAM circuit performances. This model establishes a simple relationship between inverter propagation delay and device threshold voltage shift, thereby enabling circuit frequency degradation simulation due to NBTI-induced device parameter drift. Compared to HCI reliability, it is more difficult to identify NBTI critical subcircuits because of the obvious absence of an effective NBTI equivalent electrical model. Contradictorily, NBTI degrades device parameters even when they are in static state; therefore, NBTI critical subcircuits must be identified as early as possible in the design cycle [168]. For example, DC biased circuits are very important for circuit operation (especially for analog and mixed-signal circuits), but they are prone to NBTI degradation. If the most NBTI-sensitive subcircuits in biasing networks were not identified and properly designed, the overall circuit could not be NBTI-robust.

It is very important to be able to simulate the impact of NBTI at the circuit level using SPICE simulation [168]. In most of the work on NBTI circuit SPICE simulation, simulation was performed in such a way that degraded circuit behaviors were simulated with SPICE transistor model parameter *Vt* (threshold voltage) being arbitrarily perturbed and shifted by a fixed value [154]. This kind of simulation method cannot physically relate circuit performance degradation to

the device NBTI wearout process in dynamic operation situations because the parameter *t* (NBTI stress time) is not set in. The most effective way to build up this kind of relation is through an NBTI failure equivalent circuit model. However, to our best knowledge, there is no electrical model of this kind in the literature. Based on the previously introduced Weibull law time dependence of $\Delta Vt(t)$ model (Equation 4.53), we have developed a new NBTI failure equivalent circuit model that is the first electrical model relating the time-dependent NBTI physical degradation parameter $\Delta Vt(t)$ to lumped electrical model elements, thereby enabling effective and quick NBTI circuit reliability simulation.

As mentioned earlier, the most severe NBTI effect is PMOSFET threshold voltage increase $\Delta Vt(t)$, which is equivalent to PMOSFET absolute gate-to-source voltage decrease. Therefore, if we split the PMOSFET gate connection and add a gate resistance R_G between the original gate biasing point *G* (voltage at this point is preserved as before by biasing circuit) and the PMOSFET gate immediate terminal *G'*, and construct a gate leakage current flowing mechanism (voltage controlled current sources between gate-to-drain and gate-to-source), then the gate leakage current will flow through this gate resistance R_G and increase the PMOSFET effective gate voltage at point *G'*. Because PMOSFET source is held at the highest potential, the inclusion of R_G and gate-leakage current leads to the decrease of PMOSFET absolute gate-to-source voltage, thereby imitating the NBTI threshold voltage degradation. Based on this concept, the NBTI failure equivalent circuit model is constructed as shown in Figure 4.10.

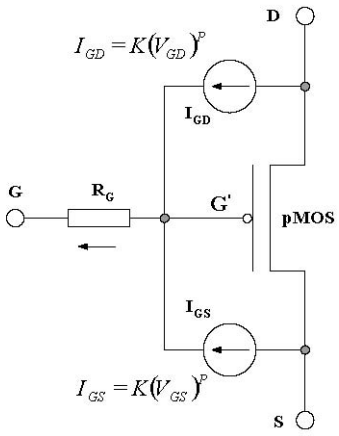

Fig. 4.10. *FaRBS NBTI failure equivalent circuit model. NBTI-induced PMOSFET threshold voltage increase is modeled as absolute gate-to-source voltage decrease. Gate tunneling current flowing through the gate resistance R_G leads to the increase of voltage at point G'. This corresponds to the decrease of PMOSFET absolute gate-to-source voltage and therefore mimics the threshold voltage degradation effect. Gate tunneling current is modeled with two voltage controlled current sources which follow the form of a power law relation as: $I = KV^P$.*

In this model, R_G is a voltage-dependent resistance because gate leakage currents are voltage dependent. R_G is also a time-dependent resistance because voltage drop across R_G at any specific time t is equal to threshold voltage shift $\Delta Vt(t)$, which is time dependent. According to [134], gate leakage current due to oxide breakdown conduction can be modeled as gate-to-diffusion leakage with a power law dependence of the formula $I = KV^p$ (where K and p are fitting parameters). We adopt the same power law voltage dependency in modeling gate leakage currents in Figure 4.10. As a result, for the gate-to drain leakage current, $I_{GD} = K(V_{GD})^p$, and for the gate-to-source leakage current, $I_{GS} = K(V_{GS})^p$. In FaRBS, the default value of p is set to 5, and the default value of K is 3×10^{-6} [134].

In Figure 4.10, the voltage drop across R_G is:

$$V_{R_G}(t) = V_{G'} - V_G = \Delta V_G(t) = (I_{GD} + I_{GS})R_G \tag{4.67}$$

Threshold voltage degradation $\Delta Vt(t)$ due to NBTI is already given by Equation (4.53); therefore, from the relation $\Delta V_G(t) = \Delta Vt(t)$, we can obtain an analytical solution for R_G:

$$R_G = \frac{\Delta V_{max}}{KV_{GD}^p + KV_{GS}^p}[1 - e^{-(\frac{t}{\tau})^\beta}] \tag{4.68}$$

The typical values and extraction methods for the model parameters ΔV_{max}, K, p, τ, and β have been given and discussed during the process of deriving Equation (4.68).

One of the more important points regarding Figure 4.10, is that this model more accurately incorporates both NBTI and possible oxide breakdown effects than a simple model that only inserts a voltage source between G and G'.

For an NMOS positive bias temperature instability (PBTI) failure equivalent circuit model, we adopt a similar model as that of PMOS NBTI in Figure 4.10, except that all current flowing directions are reversed and the model fitting parameters of the threshold voltage model ΔVt (Equation 4.53) are determined from NMOS PBTI stress testing. For the two current sources (I_{GD} and I_{GS}) in the NMOS PBTI circuit model, we adopt a better gate leakage model proposed by Lee et al. [173] as follows:

$$I_{GS} = \frac{1}{2}AL\exp(\alpha V_{GS} - \beta t_{ox}^{-\gamma}) \tag{4.69}$$

and

$$I_{GD} = \frac{1}{2}AL\exp(\alpha V_{GD} - \beta t_{ox}^{-\gamma}) \tag{4.70}$$

where I_{GS} and I_{GD} are in unit μA, L is effective channel length in nanometer, t_{ox} is oxide thickness in nanometer, A = 127.04, α = 5.61, β = 10.6, and γ = 2.5. These typical values for NMOSFETs were obtained by fitting industrial data and found to be good for technologies across many generations to 0.13 μm. They are also found to maintain good stability in SPICE simulation [173].

4.5 MaCRO Application: An SRAM Reliability Simulation and Analysis

4.5.1 Introduction

The lifetime models and circuit models for HCI/TDDB/NBTI failure mechanisms as well as the overall reliability simulation algorithms in MaCRO have been presented in the previous chapters. This chapter is an illustrative case study for the purpose of demonstrating how to apply MaCRO models and algorithms to circuit reliability simulation, analysis, and improvement.

The most common circuit structures used in exemplary reliability simulations are the ring oscillator, differential amplifier, and SRAM. Compared with the other two circuits, SRAM includes many typical subcircuits such as the cross-connected 6-T memory cell, precharge, peripheral control logic and a sense amplifier. The magnitude of the MOSFET's wearout mechanisms and their effects on circuit performance and functionality depend on the types of circuits involved [174]. Moreover, for a typical SoC circuit, SRAM occupies more than 40% of the chip area [175]. The ever-increasing integration of SRAM in embedded SoC indicates that the reliability of modern VLSI systems depends on the reliability of on-chip memories [176]. Therefore, SRAM is selected in this case study as a vehicle to show the applicability of MaCRO models and algorithms in circuit reliability simulation and analysis.

4.5.2 SRAM Circuit Design and Simulation

To simplify the circuit structure, reduce reliability simulation complexity, and magnify the effects of each failure mechanism on circuit operation, only a one-bit SRAM cell and its operation control functions are implemented. The address decoder and complex timing control subcircuits are intentionally omitted. The SRAM circuit chosen for this consideration includes one 6-T cell,

precharge, read/write control, and sense amplifier. The SRAM structural block diagram is shown in Figure 4.11. The detailed structure and function of each block are introduced in this section. The overall circuit is implemented with a commercial 0.25-μm technology with a gate oxide thickness of 5.7 nm and power supply voltage of 2.5 V.

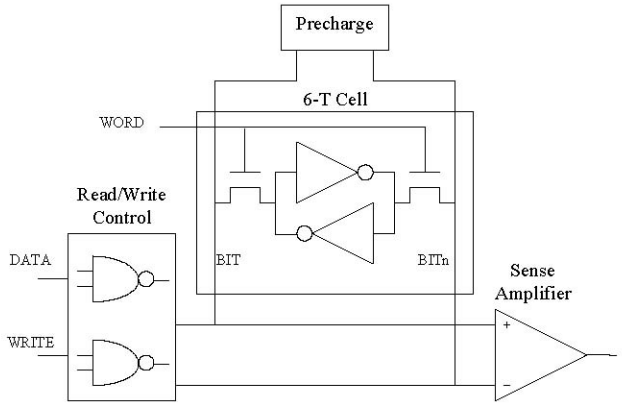

Fig. 4.11. *The one-bit SRAM structural block diagram. The circuit consists of one bit 6-T cell, read/write control logic, and output sense amplifier.*

The most important functional block in Figure 4.11 is the one-bit 6-T SRAM cell, which consists of a pair of cross-connected inverters and two NMOSFET pass transistors. The schematic of the SRAM cell is shown in Figure 4.12. Transistors M1–M4 form a regenerative structure for storing a single bit "1" or "0" at the node "Store," depending on the differential voltages of BIT/BITn lines during write cycles. The WORD line controls pass transistors M5 and M6 and enables charging/discharging paths between the nodes Store/Storen and BIT/BITn lines during read/write cycles. The cell transfer ratio of pass transistor to pull-down NMOSFET widths (i.e., width ratio of M5 to M1, and M6 to M2) is designed to 1. The proper value of this ratio is important for cell stability during read operations [175]. The two transmission gates (consisting of M41–M44) provide bidirectional paths and connect BIT/BITn lines to the write control circuit during write operation, and to the sense amplifier during the read operation.

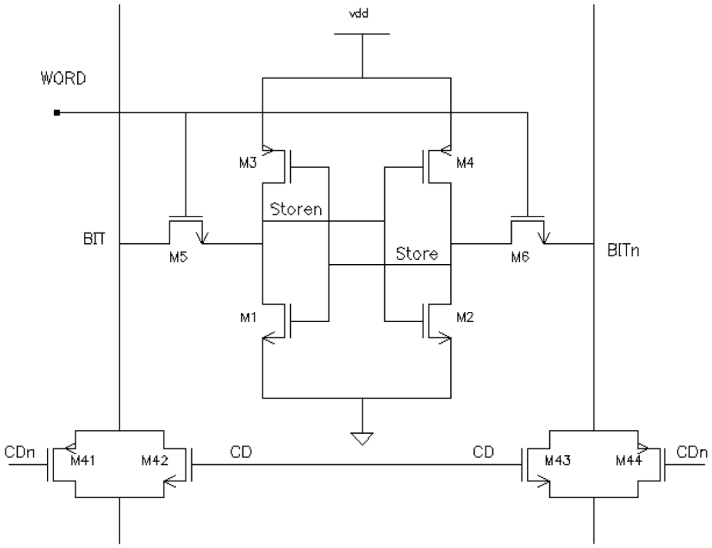

Fig. 4.12. *Schematic of on bit 6-T SRAM cell. Store/Storen represent cell state. WORD line enables the two pass transistors M5 and M6 during memory read and write cycles.*

The function of the precharge circuit is to pre-charge BIT and BITn lines to the same level before each read and write operation. The schematic of the precharge circuit is shown in Figure 4.13. When PRE signal is high, M21–M25 turn on, equalizing and charging up BIT/BITn lines to the same voltage level $V_{DD} - 2Vt$. Because NMOSFET threshold voltage Vt = 0.65 V, the pre-charge voltage level is approximately set to the middle of V_{DD}, which avoids full rail-to-rail signal transitions in subsequent read/write operation, thereby improving circuit operation speed. The high-speed transition of PRE on M21 - M25 might introduce charge injection effects on BIT/BITn lines.

These transient charges will increase voltage overshooting and reduce cell stability. For high-speed, high-volume SRAM circuits wherein node capacitances on BIT/BITn lines are very large and the swings of BIT/BITn signals are very small, transient charge injection has a more deleterious effect. The inclusion of transistors M26–M29 is for suppressing these transient charge effects and smoothing BIT/BITn signals during switching. Simulation shows for this simplified

SRAM circuit, which exhibits large BIT/BITn swings (because of small node capacitances associated with the one bit cell), failures of these transistors have minor effects on circuit functionality; therefore, M26–M29 are neglected in the following MaCRO reliability analysis.

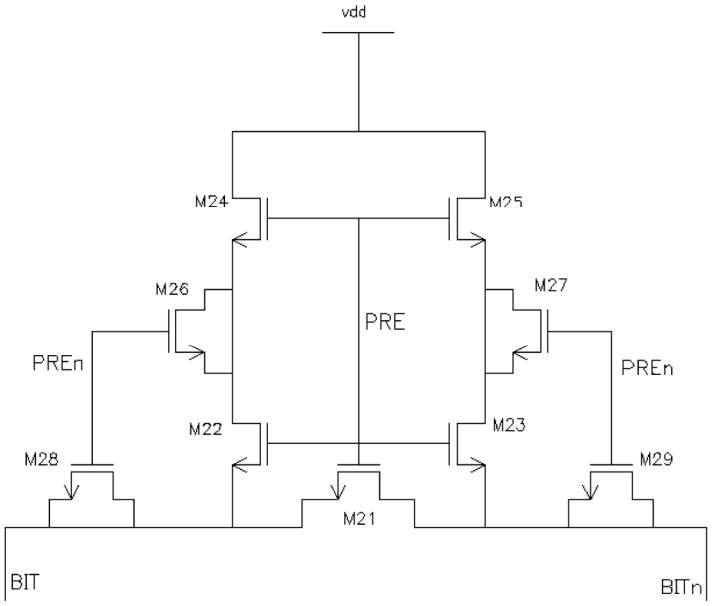

Fig. 4.13. *Schematic of the precharge circuit. BIT/BITn lines are pre-charged to the same voltage level before each read and write operation. M26–M29 are included for reducing transient charge injection effects.*

The write control logic circuit is straightforward (see Figure 4.14). WRITE signal controls the operation of the sandwiched NMOSFET and PMOSFET in the two stacked inverters, thereby gate-keeping the connection between DATA line and the SRAM cell.

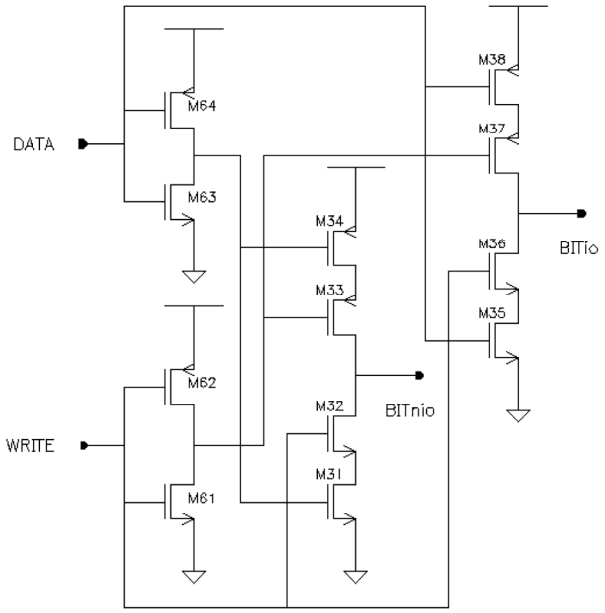

Fig. 4.14. *Schematic of the write control circuit. WRITE signal controls the connections of DATA line and BITio/BITnio lines. BITio/BITnio lines are connected to BIT/BITn by the two transmission gates (M41–M44).*

A latch-type sense amplifier (rather than a current mirror amplifier) is selected due to the small node capacitances and large voltage swings of BIT/BITn lines. A READ signal applies to M55 and M60 and controls the read operation. If the READ signal is high, the latch amplifier, consisting of M51–M55, quickly pulls BIT/BITn apart in reverse directions to the full digital levels. M56–M59 form the output buffer and help to generate smooth rail-to-rail output signals. The overall schematic of the sense amplifier circuit is illustrated in Figure 4.15.

The function of the SRAM is simulated in SPICE to perform a set of sequential “write 0, read 0, write 1, read 1” operations. The duration of each operation cycle is 2 ns, and the circuit is simulated for 8 ns, with an operation speed of 500 MHz. The timing of input signals is given in Figure 4.16.

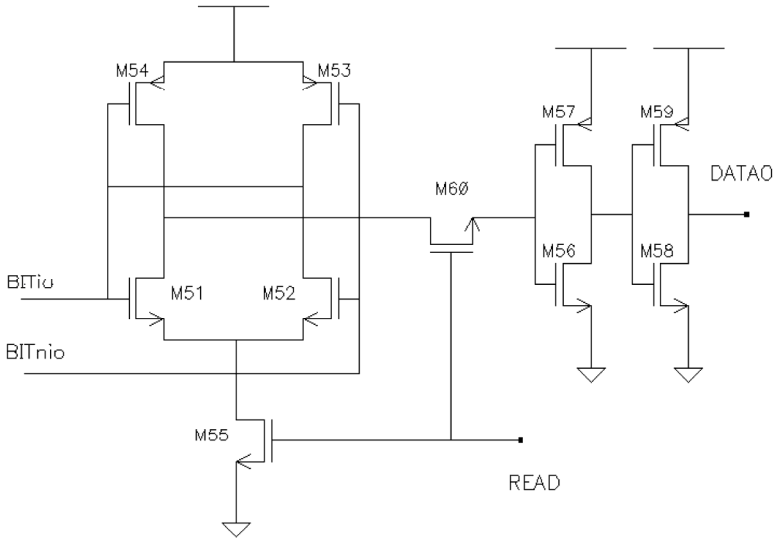

Fig. 4.15. *Schematic of the sense amplifier. READ signal controls the operation of the latch amplifier and the connection between BIT/BITn and the output. The latch amplifier magnifies BIT/BITn line swings to full digital levels.*

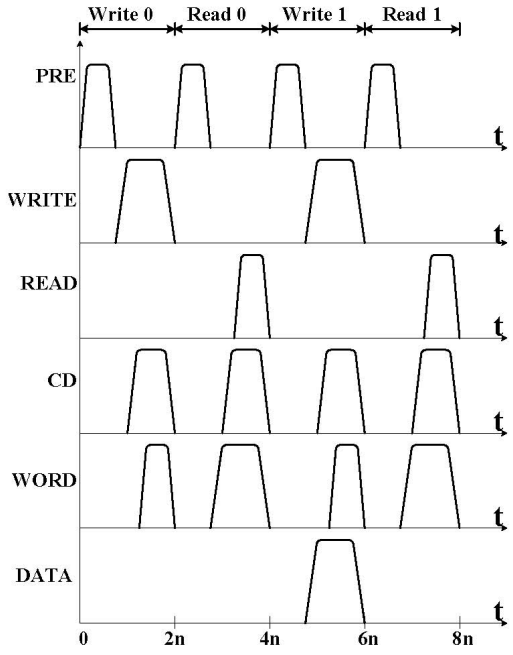

Fig. 4.16. *SRAM SPICE simulation stimuli. PRE exerts before each read/write operation. CD signal enables the transmission gates M41–M42 and WORD signal enables the pass transistors M5–M6 during each read/write operation. The "0" or "1" is available on DATA line during each write operation.*

The SPICE simulation results are shown in Figure 4.17, which (a) demonstrates precharging states and swings of BIT/BITn signals during read/write operations, (b) indicates SRAM cell state stored at Store/Storen nodes, (c) shows results of the two write operations, and (d) shows results of the two read operations. These simulation waveforms illustrate the SRAM operation process: within 1–2ns, "0" on the DATA line is written into the SRAM cell; within 3–4ns, "0" state stored in the SRAM cell is read out to the output data line DATAO; within 5–6 ns, "1" on the DATA line is written into the SRAM cell; and within 7–8 ns, "1" state stored in the SRAM cell is read out to DATAO. These timing relations will be compared later with MaCRO reliability simulation results after the SRAM experiences HCI/TDDB/NBTI stresses.

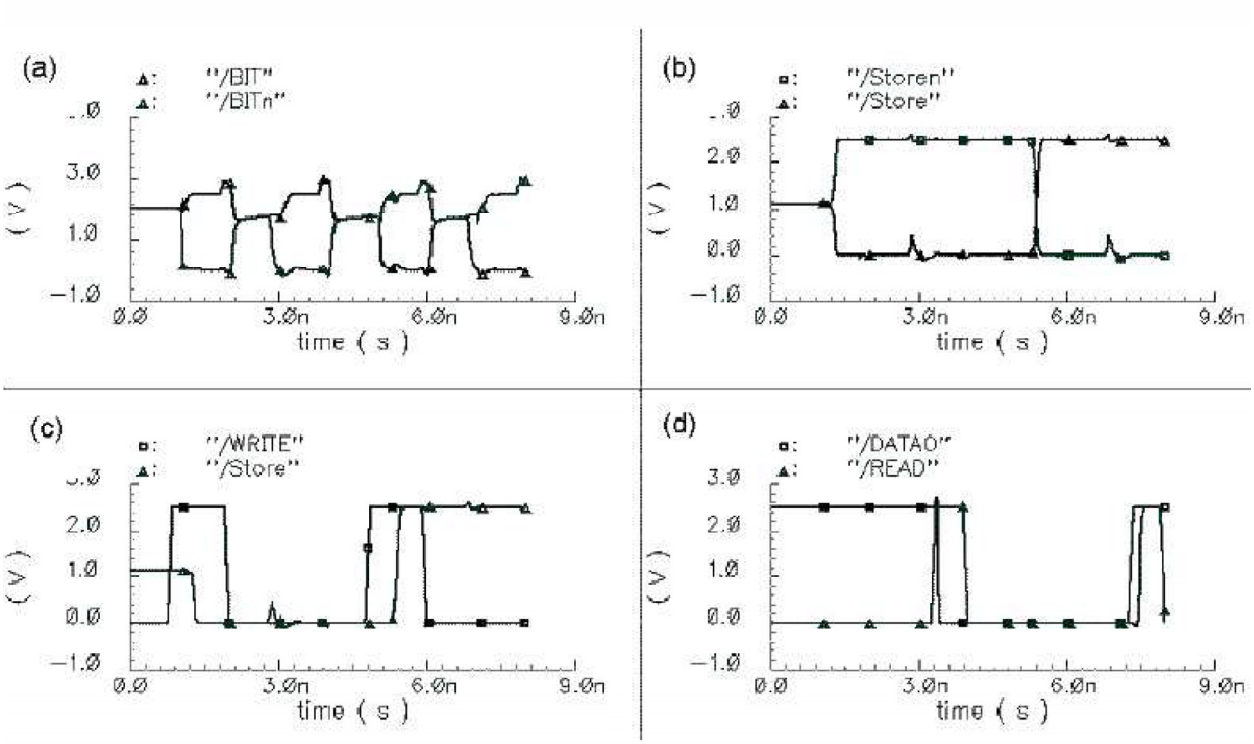

Fig. 4.17. *SRAM SPICE simulation results. (a) shows waveforms of BIT/BITn signals, (b) shows SRAM cell state signals Store/Storen, (c) is write operation result, and (d) is read operation result.*

4.5.3 Preview of SRAM Failure Behaviors

For the sake of facilitating the understanding of MaCRO reliability simulation results, a brief overview of SRAM reliability behaviors and failure effects presented in the literature is given in this section.

The main effects of HCI on device electrical characteristics are threshold-voltage drift and transconductance (gm) degradation. Pass transistors in an SRAM cell receive more severe damages because of bidirectional HCI stresses. This is proved by MaCRO simulation that follows. The gm degradation of these pass transistors gradually reduces the driving capability and cell transfer ratio [174] and increases access time after long-term operation [111, 177]. The physical origin of this enhanced HCI damage on pass transistors is explained in [178]. The sense amplifier also suffers from significant HCI stress, which results in increased input offset voltage [179] and decreased drain output resistance and small-signal voltage gain [180].

TDDB has the most deleterious effects on SRAM cell stability. There are only four topologically distinct oxide breakdown locations in the SRAM cell shown in Figure 4.12: Store-to-Storen, Store-to-VDD, Store-to-gnd, and gate-to-diffusion of pass transistors. Any other possible oxide breakdown location is equivalent to one of these categories [175]. Store-to-Storen breakdown and gate-to-diffusion breakdown of pass transistors reduce BIT/BITn differential voltage and output swing, whereas breakdowns at Store-to-VDD and Store-to-gnd increase leakage current at the opposite transistors and degrade cell stability and Static Noise Margin (SNM) [132]. The leakage currents of 20–50 μA at the NMOSFET source can result in a 50% reduction in SNM [133, 181]. Most SRAM cells become unstable without sufficient SNM [182].

A thorough investigation of different gate oxide breakdown effects on SRAM subcircuits is presented in [183, 135].

NBTI leads to device mismatches in the SRAM cell and input offset voltages in the sense amplifier. The SNM degradation due to NBTI increases as VDD decreases [154]. Experimental work of an operational amplifier to end-of-life degradation indicates little change in output characteristics, suggesting that PMOSFET NBTI-induced device mismatch is not the fundamental reason for circuit failures [170]. This conclusion is also supported by MaCRO reliability simulation results that follow.

4.5.4 Device Lifetime Calculation

The lifetime model for each failure mechanism (HCI/TDDB/NBTI) is introduced in previous chapters. These lifetime equations are recapitulated here for convenience:

$$t_f(HCI) = A_{HCI}(\frac{I_{sub}}{W})^{-n}\exp(\frac{E_{aHCI}}{\kappa T}) \tag{4.71}$$

$$t_f(TDDB) = A_{TDDB}(\frac{1}{A})^{\frac{1}{\beta}}F^{\frac{1}{\beta}}V_{gs}^{a+bT}\exp(\frac{c}{T}+\frac{d}{T^2}) \tag{4.72}$$

$$t_f(NBTI) = A_{NBTI}V_{gs}^{-\frac{1}{\beta}}[\frac{1}{1+2\exp(-\frac{E_1}{\kappa T})}+\frac{1}{1+2\exp(-\frac{E_2}{\kappa T})}]^{-\frac{1}{\beta}} \tag{4.73}$$

On the basis of SPICE simulation results, if all the model parameters are determined from device-testing work, designers can calculate device lifetime for each failure mechanism at any use condition. However, from the perspective of circuit functionality, the absolute value of device lifetime is not of primary interest. The main purpose of lifetime calculation is to identify the weakest and most damaged devices; we can conclude, therefore, that only relative lifetime (i.e., normalized lifetime) can be calculated for each device by lumping all common model parameters into a single factor. Based on this concept, Equations (4.71)–(4.73) can be rewritten in the following simplified forms:

$$t_f(HCI) = \tau_1(\frac{I_{sub}}{W})^{-n} \tag{4.74}$$

$$t_f(TDDB) = \tau_2(\frac{1}{W})^{\frac{1}{\beta}} V_{gs}^{a+bT} \tag{4.75}$$

$$t_f(NBTI) = \tau_3 V_{gs}^{-\frac{1}{\beta}} [E_1^{'} + \frac{1}{1+2\exp(-E_2/\kappa T)}]^{-\frac{1}{\beta}} \tag{4.76}$$

where τ_1–τ_3 are the lumped factors and defined as benchmarks for normalized lifetimes and W is the channel width, $E_1^{'}$ is a process-dependent constant. In deriving Equations (4.74)–(4.76), device junction temperature and the ambient temperature are not differentiated. The temperature effects of various failure mechanisms are discussed in [184]. The method to model device junction temperature with respect to device power dissipation and ambient temperature is given in [176].

In the normalized lifetime calculation process, it is unnecessary to characterize τ_1–τ_3 factors because they are common to all devices in the same circuit. This reduces the number of model parameters and simplifies parameter testing and extraction work. In Equations (4.74)–(4.76), I_{sub}, V_{gs}, and E_2 can be predicted from SPICE simulation. After obtaining the reduced set of model parameters necessary to Equations (4.74)–(4.76), designers can easily calculate device normalized lifetimes for each failure mechanism. The lifetime results are shown in Figure 4.18, wherein the horizontal axis denotes the transistor's index (e.g., "1" represents "M1") and the vertical axis denotes lifetime value normalized to τ_1–τ_3, respectively (e.g., for HCI: t_f (M1) = 4.2893τ_1). Compared with other devices, M33, M34, M37, M41, and M43 have very large NBTI lifetimes. To show details of other devices' relatively smaller lifetime values, normalized lifetime values of these transistors are not drawn to scale in (c) of Figure 4.18.

The trends that follow can be easily observed from inspecting Figure 4.18: for HCI effect, pass transistors generally experience more damage due to bidirectional stresses and more frequent switching operations, shown by M5, M6, M21, M42, M44; NMOSFETs in inverters suffer from less HCI stress, shown by M35, M56, M63; in stacked inverters, NMOSFETs on the top receive more HCI damage, shown by comparisons of M31 to M32 and M35 to M36, respectively; and the

sense amplifier is sensitive to HCI because distinct HCI damages on M51 and M52 lead to increased device mismatches and input offset voltages. For TDDB effect, PMOSFET is easier to suffer from TDDB due to its relatively larger channel area, and area scaling has a significant effect on device lifetimes, shown by M62, M64, whose channel widths are very large (12 μm). For NBTI effect, PMOSFETs in latch structure receive more imbalanced NBTI damage, which also leads to increased device mismatches and input offset voltages, shown by M3, M4 and M53, M54.

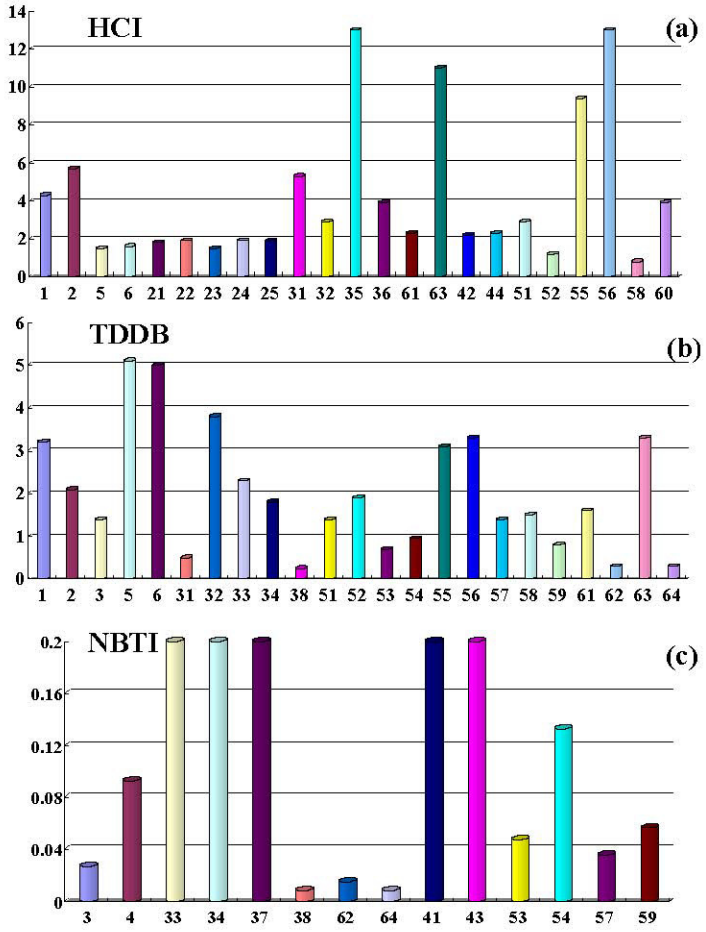

Fig. 4.18. *Device lifetime calculation results for the three failure mechanisms: (a) HCI, (b) TDDB, and (c) NBTI. The horizontal axis denotes device's index, and vertical axis denotes lifetime value normalized to τ_1–τ_3, respectively.*

It is easy to identify the most damaged transistors for each failure mechanism from Figure 4.18. For HCI, M5, M6, M52, and M58 are the most damaged transistors. M58 has the shortest lifetime; however, it can be excluded after a careful analysis. In the initial schematic, the two stages of inverters after the sense amplifier were designed with the sizing ratio of 1. If scaling up the channel widths of M58 and M59 and increasing the sizing ratio to 3, the lifetime of M58 increases from $0.84\tau_1$ to $7.79\tau_1$. The reason for this significant improvement is the reduction in inverter transition delay after proper sizing of the inverter chain, as shown in Figure 4.19. Proper inverter sizing improves both transition speed and device lifetime with the penalties of larger chip area and loading to neighboring gates. Therefore, circuit designers need to perform detailed lifetime calculation and functional simulation to make appropriate tradeoffs.

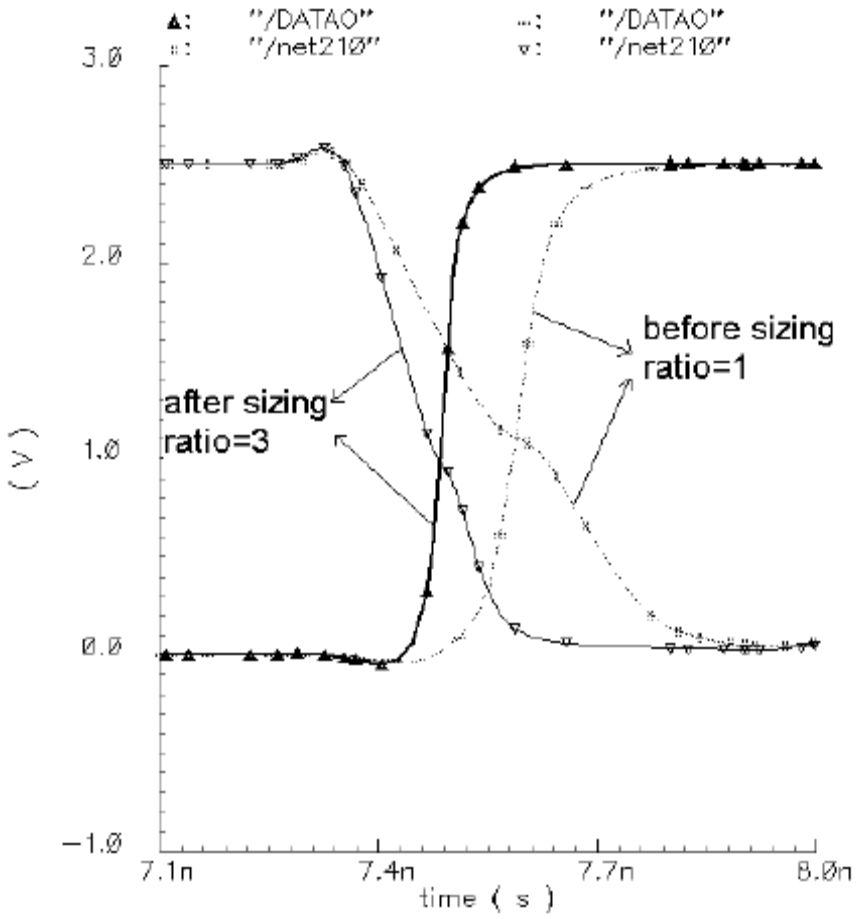

Fig. 4.19. *Comparison of transition delay of M58 before and after inverter sizing. Proper sizing significantly reduces dynamic switching delay, thereby suppressing HCI effect. Wn = 0.6 um before sizing and Wn = 1.8 um after sizing.*

For TDDB, M3, M31, M38, M53, M 62, and M64 are the most damaged transistors. M62 and M64 are PMOSFETs in the write-control logic subcircuit; their channel widths are designed to be very large to quickly generate inverse signals of WRITE and DATA. Their widths can be properly scaled down to improve lifetime; therefore, it is unnecessary to include them in the weakest device list. M3 is included because it is within the SRAM cell and its oxide breakdown has a significant effect on SRAM operation. All transistors in precharge circuit (M21–M29) are not selected because, during all operation periods, their gate-to-source/drain voltages are negligible.

For NBTI, M3, M38, and M53 are selected as the most damaged transistors. Although lifetimes of M62, M64, M57, and M59 are comparable to those of being selected, on the basis of the reason given above, they are not included in the weakest-device list.

In summary, the devices identified to be the most damaged for each failure mechanism are: HCI—M5, M6, and M52; TDDB—M3, M31, M38, and M53; and NBTI—M3, M38, and M53. These transistors will be substituted with corresponding circuit models in the following SPICE simulation.

4.5.5 SPICE Reliability Simulation with Circuit Models

The model equations and methods to determine circuit model parameters have been presented in previous chapters. Most of these model parameters are time dependent; therefore, SPICE simulation with these circuit models has to be performed several times to pinpoint the time at which the circuit function fails. The most effective way to find this failure time is by a three-step progressive process: first, only consider HCI failure electrical models and find the circuit HCI lifetime T_a; then, add on TDDB electrical models and simulation circuit operation at times shorter than T_a and find the corresponding circuit HCI+TDDB failure lifetime T_b ($T_b \leq T_a$); finally, include all failure electrical models and find the circuit failure lifetime T_c ($T_c \leq T_b$) at which the circuit cannot maintain correct operations. In this step-by-step process, circuit failure behaviors and response due to each failure mechanism can be efficiently characterized. The detailed algorithm of

this process is given in Chapter 2. The SRAM reliability simulation and analysis are performed according to this three-step process.

HCI

There is only one parameter in the HCI circuit model: ΔR_d, which characterizes drain current reduction due to mobility degradation resulting from HCI-induced interface charge and oxide charge. ΔR_d values of M5, M6, and M52 at different stress times are plotted in Figure 4.20. These HCI-induced series parasitic resistances are not in simple logarithmic relation to stress time t because the horizontal axis is not drawn in linear scale. M5 and M6 receive bidirectional HCI stresses; consequently, each of them has two resistances (ΔR_{d1} and ΔR_{d2}) associated with drain and source, respectively.

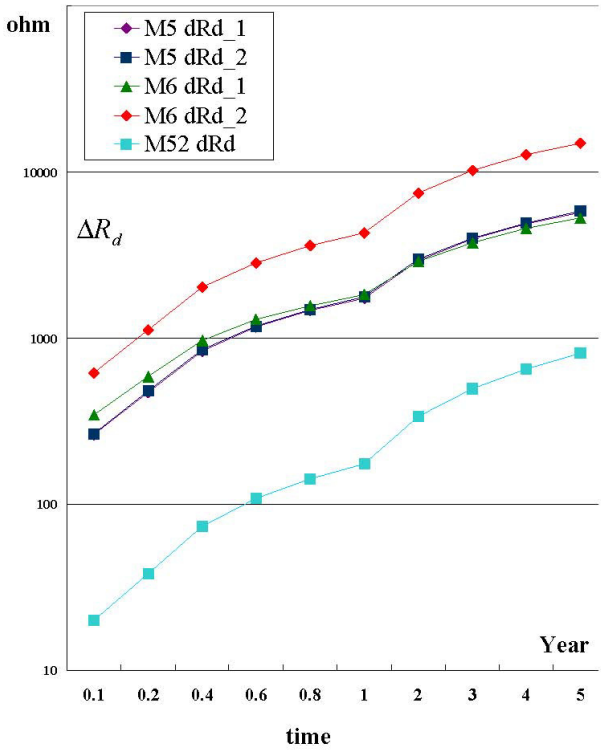

Fig. 4.20. *ΔRd values of M5, M6, M52 at different stress times. The unit of horizontal axis is time in years; the vertical axis is in logarithmic scale and in unit Ohm.*

The SRAM circuit with these HCI-induced ΔR_d elements is simulated at different stress times to check its functionality. Figure 4.21 shows the waveforms of the SRAM cell state (i.e., Store signal) and output state (i.e., DATAO signal) after different stress times. It indicates that the SRAM circuit operates correctly until 0.8 year, and fails at 1 year, at which time the Store signal does not switch as expected during the "write 1" cycle. The gradual degradation of the Store signal is clearly shown in Figure 4.21. The corruption of the Store signal occurring more rapidly than that of DATAO implies that malfunction of this SRAM circuit mainly results from HCI damage of M5 and M6, rather than M52; this verifies other researchers' work on the relation between the pass transistor's HCI degradation and SRAM cell stability.

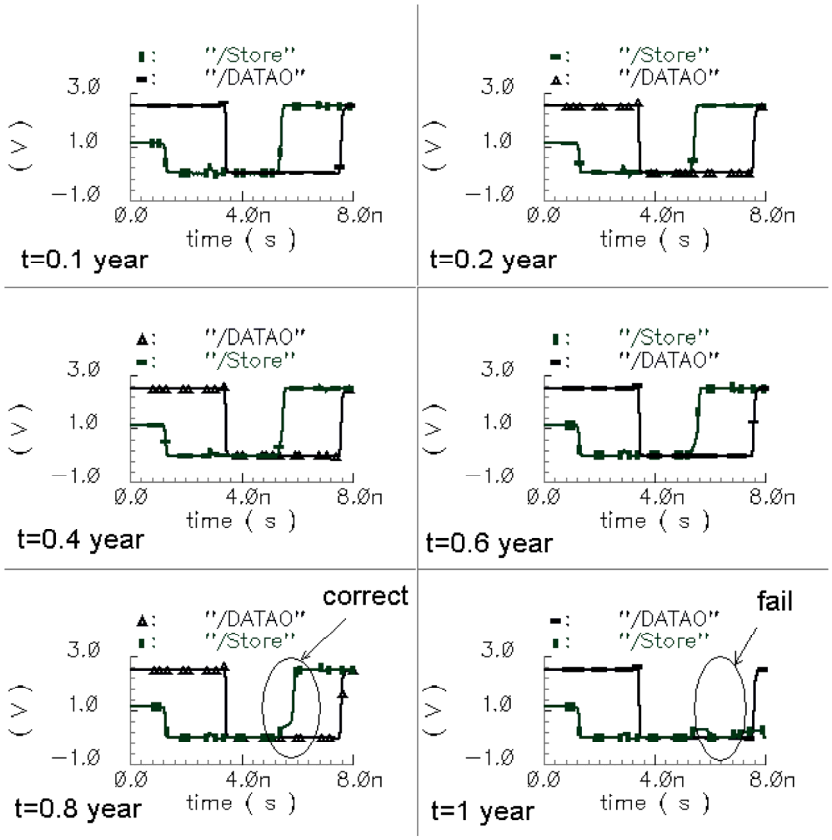

Fig. 4.21. *The simulated waveforms of the SRAM cell Store signal and output DATAO signal after different stress times. At t = 1 year, Store signal does not jump to high as expected during the "write 1" cycle indicating failure of SRAM cell.*

A closer look at BIT/BITn and Store/Storen waveforms before and after SRAM cell failure reveals more reliability information. Figure 4.22 compares and shows how these signals corrupt. It is clearly shown that the addition and increase of HCI-induced series resistances in M5 and M6 degrade BIT/BITn signals and reduce cell transfer ratio. As a result, the high BIT line signal at "write 1" cycle cannot be effectively written into the SRAM cell. Store/Storen signals cannot switch when a reverse value is being written to the SRAM cell.

From the above SPICE simulation with HCI circuit models, the HCI lifetime of the SRAM circuit is predicted to be 0.9 year. If considering the effect of duty cycle and assuming that the average access frequency of the SRAM is one full operation per 1 μs at normal use condition, the predicted 0.9 year corresponds to a circuit HCI lifetime of 112.5 years.

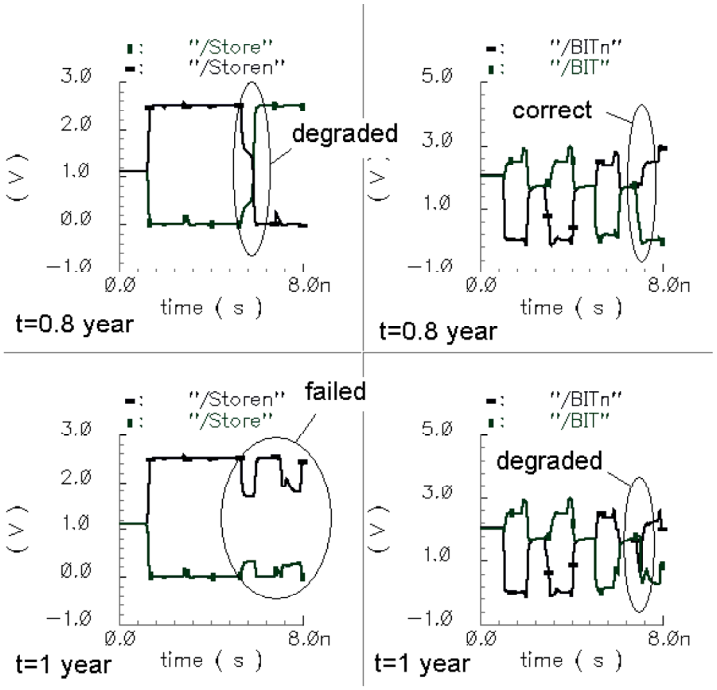

Fig. 4.22. *The waveforms of SRAM Store/Storen signals and BIT/BITn signals before and after circuit failure. Store/Storen signals do not flip due to the degredation in BIT/BITn signals when a reverse value is being written to the SRAM cell.*

HCI+TDDB

The second step in the SRAM circuit reliability simulation is the inclusion of both TDDB and HCI circuit models. Only gate-to-channel breakdown is considered, and the breakdown location is intentionally set to the middle point of the channel. As a result, only one parameter (I_{ox}) needs to be characterized for each identified TDDB damaged transistor. The values of I_{ox} have been calculated as: I_{ox}(M3) = −50.719 μA, I_{ox}(M31) = 25.642 μA, I_{ox}(M38) = −18.07 μA, and I_{ox}(M53) = −101.05 μA.

The SPICE simulation that results when taking into account both HCI and TDDB effects are illustrated in Figure 4.23. The SRAM circuit survives until 0.4 year but fails to function at 0.6 year.

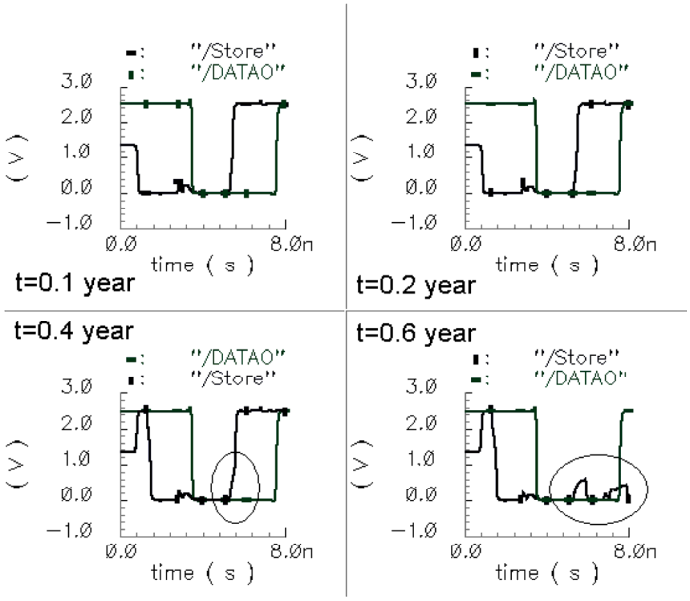

Fig. 4.23. *The simulated waveforms of the SRAM cell Store signal and output DATAO signal at different HCI+TDDB stress times. At t = 0.6 year, Store signal does not jump to high during the "write 1" cycle indicating failure of the SRAM cell.*

The addition of TDDB failure electrical models significantly reduces circuit lifetime. Figure 4.24 shows the interaction between the HCI effect and the TDDB effect, wherein the BIT/BITn

and Store/Storen waveforms before and after circuit failure (at 0.4 year and 0.6 year, respectively) are plotted. At 0.6 year, the corruption of Store/Storen signals and the degradation of BIT/BITn signals during the final "write 1, read 1" cycles are very similar to those at 1 year in Figure 4.22, wherein only the HCI effect is considered. Moreover, if the TDDB effect on M3 is disabled at 0.6 year, the circuit function restores and the waveforms without TDDB at 0.6 year are quite similar to the waveforms with TDDB at 0.4 year. These similarities imply that gate-to-channel breakdown of TDDB accelerates SRAM cell instability; it does not, however, introduce new failure behavior at the circuit level. This result is not observed by other researchers because most work on SRAM cell instability analysis does not combine HCI and TDDB effects together and because that simulation work includes worse-case gate-to-diffusion breakdown mode rather than the more frequent and less severe gate-to-channel breakdown mode of TDDB.

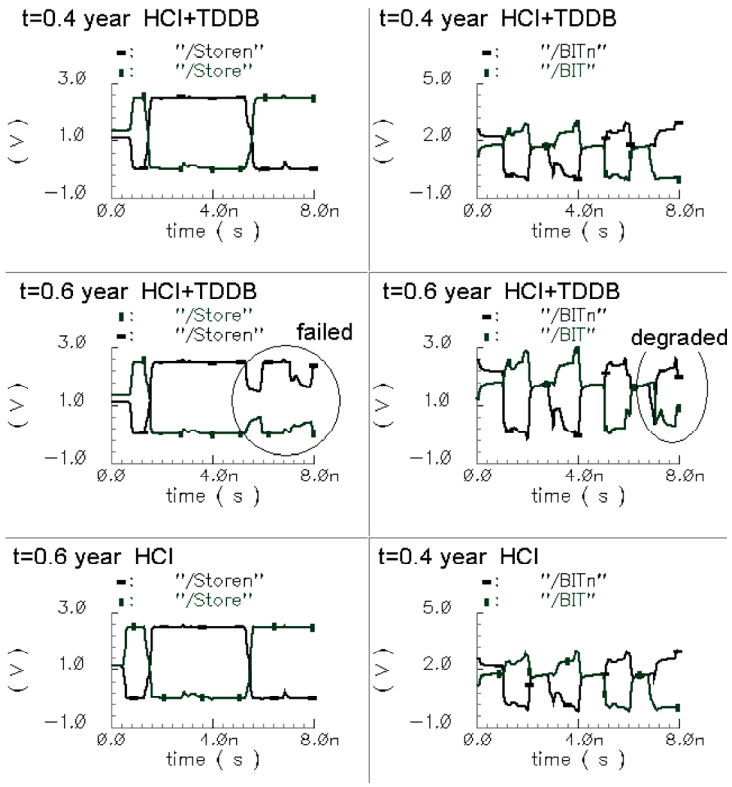

Fig. 4.24. *The waveforms of the SRAM Store/Store signals and BIT/BITn signals before and after circuit failure.*

Besides the TDDB effects of M3 on circuit operation, it is also necessary to investigate TDDB effects of M31, M38, and M53 on circuit performance. Simulation proves breakdowns of M31 and M38 (both belong to inverters in the write control subcircuit) have minor effects on SRAM operation, but breakdown of M53 has a significant effect. Figure 4.25 shows the TDDB effect of M53 on sense amplifier input signals. The breakdown in M53 provides an additional current path between sense amplifier input and V_{DD} and tends to pull up this input signal. The erratic jumps in the amplifier input signal shown in Figure 4.25 reduces amplifier output stability.

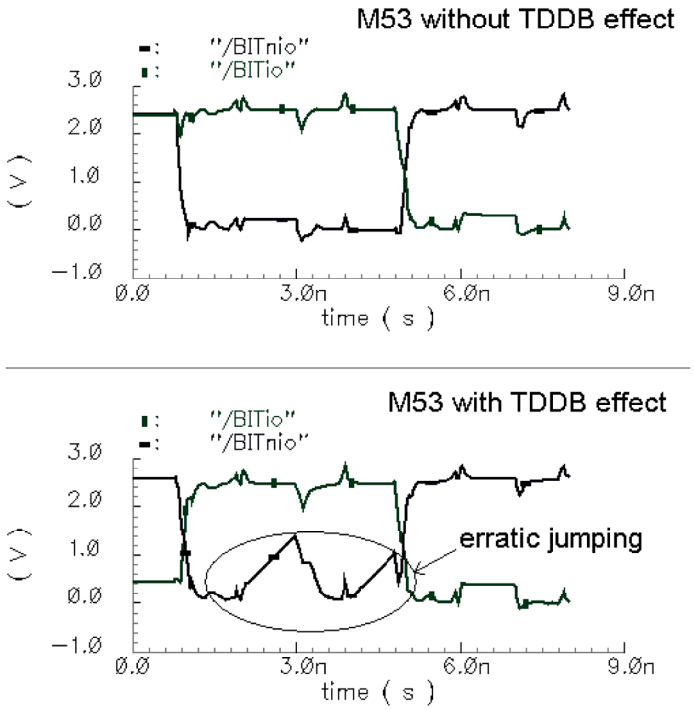

Fig. 4.25. *The TDDB effect of M53 on sense amplifier output stability. The breakdown in M53 provides additional current path between BITnio and VDD and tends to pull up BITnio when it is at low level in "read 0" and "write 1" cycles.*

HCI+TDDB+NBTI

The last step is the inclusion of NBTI circuit models. M3, M38, and M53, being identified for suffering the most NBTI damage, also receive the most TDDB damage; therefore, designers need

to properly combine the NBTI and TDDB electrical models together for these PMOSFETs. If they simply add all NBTI failure circuit model elements into the TDDB model, the oxide breakdown effect will be overestimated, which results in suppressing or overshooting of SRAM cell state signals (i.e., Store/Storen) and unexpected jumps of sense amplifier input signals. These negative phenomena are observed in simulation results. The correct TDDB+NBTI failure electrical model for a PMOSFET is illustrated in Figure 4.26.

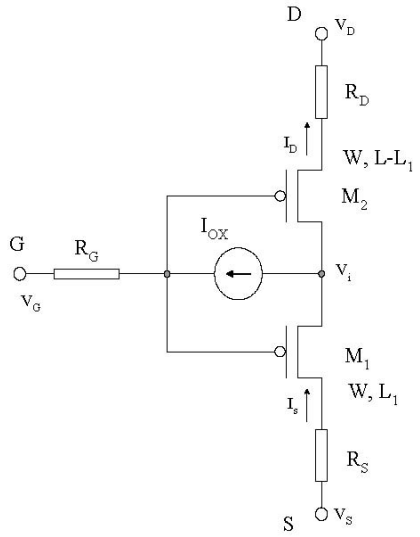

Fig. 4.26. *The TDDB+NBTI circuit model for a PMOSFET. R_G and I_{OX} account for threshold voltage degradation due to NBTI. I_{OX} and the two split PMOSFETs represent TDDB damage. R_D and R_S characterize the resistances in drain and source extensions. They are excluded in this SRAM case study in order to simplify simulation work.*

With the previous HCI+TDDB simulation results, it is only required to calculate R_G for each of M3, M38, and M53 at time 0.4 year. Their values are: R_G(M3) = 6.6KΩ, R_G(M38) = 965.4Ω, and R_G(M53) = 3.3KΩ. Simulation indicates that NBTI has relatively weak effects on SRAM cell stability and functionality. Its most obvious influence observed from simulation is that NBTI degrades SRAM cell transition speed. This effect is shown in Figure 4.27, where the switching of Store/Storen signals slows down when the NBTI model is set in. Simulation also shows NBTI has minor effects on functionality of the latch type sense amplifier. The degradation in input signals is very small.

SPICE DC voltage transfer function simulation along the path from BITn line to Storen line encompasses all of the three failure mechanisms (HCI of M6, TDDB and NBTI of M3); therefore, degradation in VTC for BITn-to-Storen at different combinations of these failure mechanisms can reflect their individual influence on SRAM cell stability. These VTC curves, plotted in Figure 4.28, indicate that HCI and TDDB have reverse effects on VTC drift, while NBTI has no observable effect.

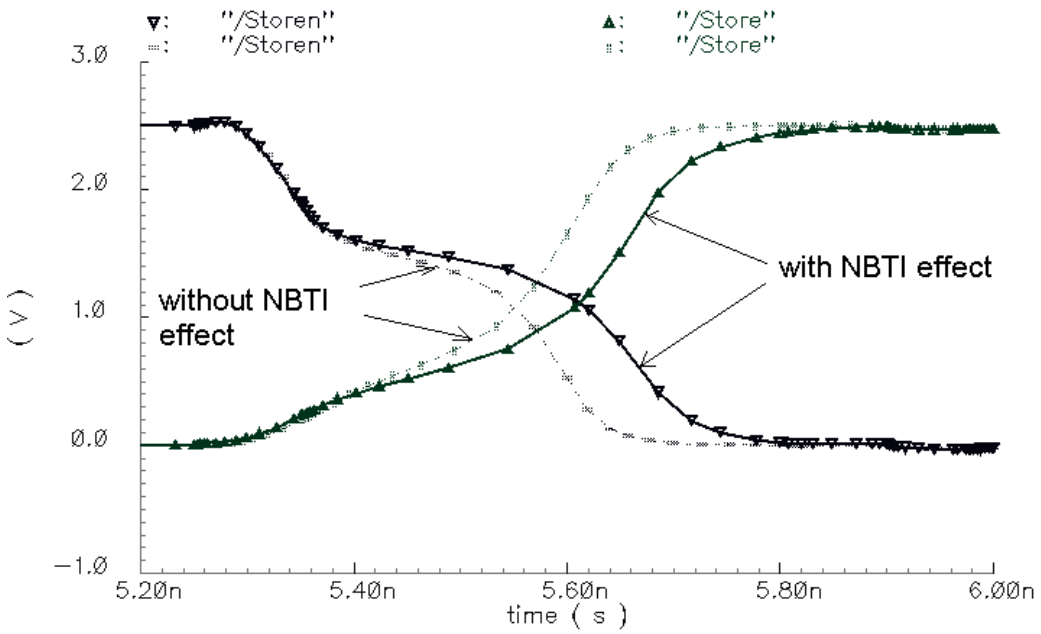

Fig. 4.27. *The NBTI effects on SRAM cell transition speed. The switching speed of SRAM cell Store/Storen signals degrades when NBTI damage on M3 is considered.*

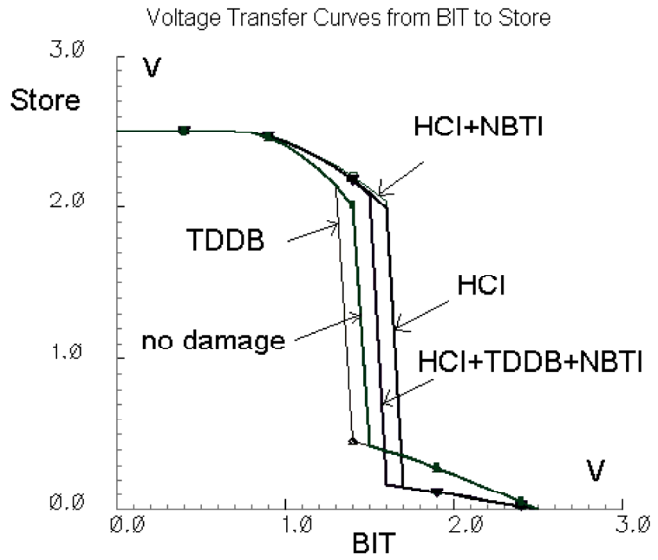

Fig. 4.28. *Voltage transfer curves of BITn-to-Storen for different combinations of failure mechanisms. From left to right, the curves represent effects of TDDB, no damage, HCI+TDDB+NBTI, HCI, and HCI+NBTI, respectively. NBTI has negligible effect on SRAM cell stability.*

SNM is the most important factor in SRAM circuit reliability analysis. On the basis of the SPICE DC transfer analysis, SNM butterfly plots for various combinations of the failure mechanisms are generated in Figure 4.29. The size of the two maximized embedded squares in the butterfly plots represents the magnitude of SNM. In Figure 4.29, (a) represents failure free operation, (b) shows SNM degradation due to TDDB effect, (c) shows the combined effect of TDDB+NBTI on SNM, and (d) is the combination of plots (a) to (c) for the sake of easy comparison. These curves are obtained by setting failure circuit model parameters at stress time 0.4 year. It is indicated from these butterfly plots that SRAM cell noise margin shrinks due to TDDB and NBTI stresses and TDDB has the dominant effect. The gate-to-channel breakdown of M3 leads to symmetrical shrinkage of the two embedded squares, which is distinct and in contrast to the case of gate-to-diffusion breakdowns presented in [133, 181], where asymmetrical scales of the sizes of the two embedded squares resulted from p-source breakdown. It is expected that the gate-to-diffusion breakdown model of TDDB would accelerate SNM degradation. At 0.4 year, even

though SNM is significantly reduced, the two transfer curves still cross and form two stable states; therefore, SRAM cell function is maintained.

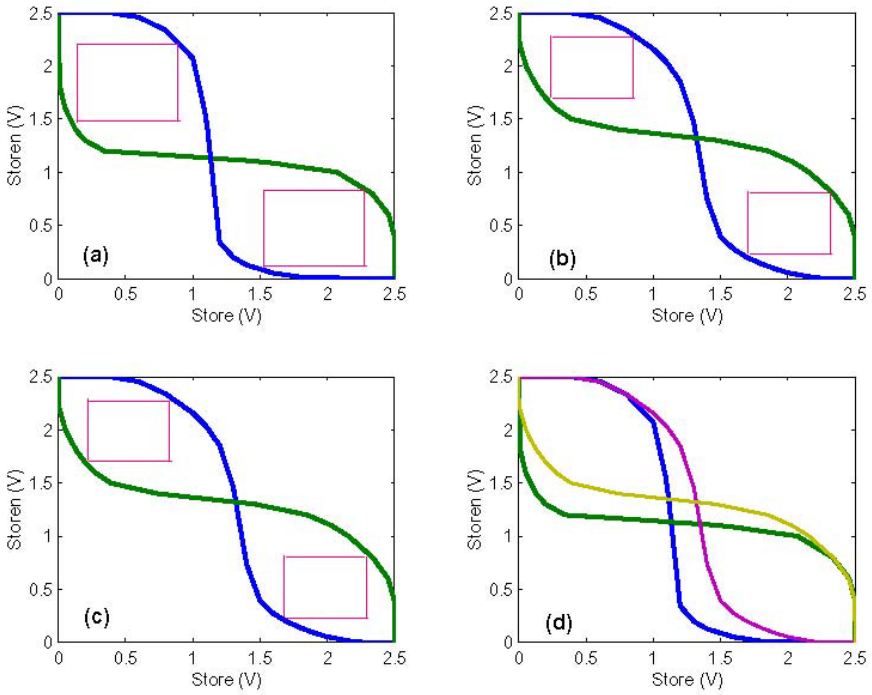

Fig. 4.29. *Butterfly plots for various failure mechanisms. (a) denotes the no-damage operation, (b) shows SNM degredation due to TDDB, (c) shows the combined effect of TDDB+NBTI, and (d) is the combination of the previous three plots. The difference in (b) and (c) is very small indicating that NBTI is not a dominant effect.*

The SRAM circuit survives to 0.4 year but fails at 0.6 year. If the same duty cycle and usage profile are assumed as before, the HCI+TDDB+NBTI lifetime of this SRAM circuit under normal use condition is approximately 62.5 years.

4.5.6 Reliability Design Techniques

After exploring circuit degradation effects and reliability behaviors with MaCRO models, designers need to make design iterations to improve circuit lifetime if the initial design falls short of reliability specifications. Traditionally, this is arduous work due to the lack of a systematic and convenient reliability analysis method to help pinpoint reliability weaknesses and characterize

circuit degradation in performance and functionality. With MaCRO models and simulation algorithms, designers can perform a quick reliability analysis and gain knowledge on circuit failure behaviors. Equipped with this reliability knowledge, they can develop their own expertise on reliability improvement through proper design iterations.

In the literature, there are some reliability design techniques available for suppressing different failure mechanisms. Reliability design techniques for HCI, including transistor sizing, gate topology, and input signal scheduling are presented in [108]. Some design improvement concepts for TDDB are introduced in [185]. A design technique to reduce gate-to-source voltage during static state operation and improve NBTI reliability is introduced in [161]. Even though some progress has been achieved from individual work, design techniques for TDDB and NBTI continue to be investigated.

4.5.7 Summary

In this section, a simple SRAM circuit is designed and simulated with MaCRO models and algorithms to illustrate how to apply this method to circuit reliability simulation and analysis. Simulation shows that HCI and TDDB have significant effects on SRAM cell stability and voltage transfer characteristics, while NBTI mainly degrades cell transition speed when the cell state flips. This case study of SRAM reliability simulation work demonstrates that with MaCRO lifetime models and circuit models, microelectronic circuit designers and device reliability engineers can develop an in-depth understanding of circuit failure behaviors and the damage effects of HCI/TDDB/NBTI on circuit operation. With this knowledge, they can better estimate circuit lifetime, make appropriate performance and reliability tradeoffs, and formulate practical design guidelines to improve microelectronic circuit reliability.

5 MICROELECTRONIC SYSTEM RELIABILITY

5.1 Introduction

Voltage and temperature are two important stresses in semiconductor device reliability analysis, especially in accelerated testing. Until now, most of the research study has been focused on voltage and temperature acceleration effects toward single-failure mechanisms. At the system level, due to the complexity of VLSI circuit and dynamic operating conditions, extrapolation of system voltage and temperature acceleration factors from individual failure mechanisms remains a formidable challenge. Although various models have been proposed to describe the voltage acceleration effect for a single failure mechanism, because of the unique physical process underlying each failure mechanism, e.g., electromigration (EM), hot carrier injection (HCI), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI), no universal voltage acceleration model has been established. For temperature acceleration, the Arrhenius model has been widely applied in reliability practice. However, each failure mechanism might have its own unique activation energy. These factors complicate system acceleration modeling. As a matter of fact, $AF_S = AF_T\, AF_V$ is the prevailing system acceleration model wherein AF_S is the system acceleration factor, AF_T is the temperature acceleration factor, and AF_V is the voltage acceleration factor. Generally, AF_T is modeled through the Arrhenius relationship:

$$AF_T = \exp[\frac{E_a}{k}(\frac{1}{T_0} - \frac{1}{T_A})] \tag{5.1}$$

where E_a is the activation energy, k is Boltzmann constant, T_A is accelerated temperature, and T_0 is nominal operating temperature. AF_V is modeled by the exponential law:

$$AF_V = \exp\left[\gamma(V_A - V_0)\right] \tag{5.2}$$

where γ is the voltage acceleration coefficient, V_A is the accelerated voltage, and V_0 is the nominal operating voltage. This multiplication model gains popularity because it is easy to apply reliability projections without building a lifetime model that fits a range of temperatures and voltages. However, companies usually neglect the multiple-failure-mechanisms' effect at the system level and simply assume E_a and γ are stress-independent. They only provide one E_a for AF_T and one γ for AF_V at each technology generation. Without solid proof, this kind of practice only provides a rough reliability estimation, which is not enough to fully exploit the tradeoffs between performance and reliability. Simulation shows that E_a and γ depend on stress voltage and temperature when multiple intrinsic failure mechanisms are involved.

5.2 Individual Failure Mechanism Lifetime Models

Relentless scaling for better performance keeps generating new reliability challenges in every aspect of process technology. EM, the main reliability concern of interconnects, needs to be handled carefully because of the dual threats posed by decreasing feature size and increasing temperature. To meet performance and reliability requirements, copper interconnects have gradually taken the place of aluminum-alloy metallization, due to its low resistivity and high resistance towards electromigration. Copper interconnects have different EM characteristics compared to aluminum. They are interface-dominated [186] and have larger activation energies [187]. TDDB has always received a lot of attention because device scaling keeps driving the oxide thickness down; however, the supply voltage scaling does not keep up with the pace. The direct impact of this non-ideal voltage scaling is an increase in gate leakage and tunneling current, which decreases the oxide lifetime. An empirical observation is that as gate oxide thickness reduces by ΔT_{ox} (in nm), by scaling, the leakage current will increase by 10 ΔT_{ox} [188] and TDDB lifetime will reduce by the same 0.22 factor. Oxide-breakdown-related failures are often reported in device burn-in tests of deep submicron technologies [189, 190].

Device scaling also increases susceptibility to another failure mechanism, NBTI, which occurs primarily in p-channel metal oxide semiconductors (PMOSFETs) with a negative gate voltage bias. The interface-trap density generated by NBTI has an inverse proportionality to oxide thickness (T_{ox}), which means NBTI becomes more severe for ultrathin oxides [82], while the

NBTI-generated fixed charge has no thickness dependence. Like NBTI for a p-channel metal oxide semiconductor (PMOS), HCI induces interface states and causes degradation of n-channel MOSFETs (NMOSFETs). Although well contained by channel engineering, it still shows up in real applications [191].

To model system reliability, all of the intrinsic failure mechanisms should be considered since any one of them might cause system failure. Various lifetime models have been proposed for each failure mechsnism. As the goal is to show the unique characteristics of system lifetime and voltage and temperature acceleration, we will adapt the generally accepted models here.

Failure rate model and acceleration factors for EM, HCI, TDDB, and NBTI are listed below.

1. EM

From the well known Black's equation [192] and Arrhenius model, failure rate of EM can be expressed as:

$$\lambda_{EM} \propto (J)^n \cdot \exp[\frac{-E_{aEM}}{kT}] \tag{5.3}$$

where J is the current density in the interconnect, k is Boltzmann's constant, T is absolute temperature in Kelvin, E_{aEM} is the activation energy, and n is a constant. Both E_{aEM} and n depend on the interconnect metal. Recently, copper/low-K dielectric material has been rapidly replacing aluminum alloy/SiO_2-based interconnect. For copper, n has been reported to have values between 1 and 2 [186] and E_{aEM} varies between 0.7 eV and 1.1 eV [130].

In Equation (5.3), current density, J, can be replaced with a voltage function [193]:

.

$$J = \frac{C \cdot V_D}{W \cdot H} \cdot f \cdot p \tag{5.4}$$

where C, W, and H are the capacitance, width, and thickness of the interconnect, respectively. f is the frequency and p is the toggling probability; therefore, λ_{EM} is also a function of voltage:

$$\lambda_{EM} \propto (V_D)^n \cdot \exp[\frac{-E_{aEM}}{kT}] \tag{5.5}$$

The EM acceleration factor is:

$$AF_{EM}^{V_O,T_O;V_A,T_A} = (\frac{V_A}{V_O})^n \cdot \exp[\frac{E_{aEM} \cdot (T_A - T_O)}{k \cdot T_A \cdot T_O}] \tag{5.6}$$

.

2. HCI

Based on the empirical HCI voltage lifetime model proposed by Takeda [48] and the Arrhenius relationship, HCI failure rate λ_{HCI} can be modeled as:

$$\lambda_{HCI} \propto \exp[\frac{-\gamma_{HCI}}{V_D}] \cdot \exp[\frac{-E_{aHCI}}{kT}] \tag{5.7}$$

where γHCI is a technology-related constant and EaHCI is the activation energy, which varies between −0.1 eV to −0.2 eV [55]. The negative activation energy means HCI becomes worse at low temperature. The HCI acceleration factor is:

$$AF_{HCI}^{V_O,T_O;V_A,T_A} = \exp[\gamma_{HCI} \frac{(V_A - V_O)}{V_A \cdot V_O}] \cdot \exp[\frac{E_{aHCI} \cdot (T_A - T_O)}{k \cdot T_A \cdot T_O}] \tag{5.8}$$

.

3. TDDB

The exponential law for TDDB failure-rate voltage dependence has been widely used in gate oxide reliability characterization and extrapolation. Combining with the Arrhenius relationship for temperature dependence, the TDDB failure rate is:

$$\lambda_{TDDB} \propto \exp[\gamma_{TDDB} \cdot V_G] \cdot \exp[\frac{-E_{aTDDB}}{kT}] \tag{5.9}$$

where γ_{TDDB} is a device-related constant and E_{aTDDB} is the activation energy. E_{aTDDB} normally falls in the range of 0.6 eV to 0.9 eV [55]. The TDDB acceleration factor is:

$$AF_{TDDB}^{V_O,T_O;V_A,T_A} = \exp[\gamma_{TDDB} \cdot (V_A - V_O)] \cdot \exp[\frac{E_{aTDDB} \cdot (T_A - T_O)}{k \cdot T_A \cdot T_O}] \tag{5.10}$$

4. NBTI

Like TDDB, NBTI voltage dependence can also be modeled by the exponential law [84]. Considering the temperature dependence together, the NBTI failure rate is:

$$\lambda_{NBTI} \propto \exp[\gamma_{NBTI} \cdot V_G] \cdot \exp[\frac{-E_{aNBTI}}{kT}] \tag{5.11}$$

where γ_{NBTI} is a constant, and E_{aNBTI} is the activation energy, which has been reported to vary from 0.1 eV to 0.84 eV [89, 194]. The NBTI acceleration factor is:

$$AF_{NBTI}^{V_O,T_O;V_A,T_A} = \exp[\gamma_{NBTI} \cdot (V_A - V_O)] \cdot \exp[\frac{E_{aNBTI} \cdot (T_A - T_O)}{k \cdot T_A \cdot T_O}] \tag{5.12}$$

5.3 Microelectronic System Voltage and Temperature Acceleration

In a simplified example, assuming there is no interaction among failure mechanisms, the system's failure rate can be obtained by the sum-of-failure-rates since all failure mechanisms may contribute to microelectronic failures.

$$\lambda_S = \lambda_{EM} + \lambda_{HCI} + \lambda_{TDDB} + \lambda_{NBTI} \tag{5.13}$$

The microelectronic system acceleration factor can be expressed as:

$$AF_S = \frac{\lambda_S^{V_A,T_A}}{\lambda_S^{V_O,T_O}} = \frac{\lambda_{EM}^{V_A,T_A} + \lambda_{HCI}^{V_A,T_A} + \lambda_{TDDB}^{V_A,T_A} + \lambda_{NBTI}^{V_A,T_A}}{\lambda_{EM}^{V_O,T_O} + \lambda_{HCI}^{V_O,T_O} + \lambda_{TDDB}^{V_O,T_O} + \lambda_{NBTI}^{V_O,T_O}} \tag{5.14}$$

Given the models of individual failure mechanisms, the system acceleration factor (5.14) can be further expressed as:

$$AF_S = P_E^{V_O,T_O} \cdot AF_{EM} + P_H^{V_O,T_O} \cdot AF_{HCI} + P_T^{V_O,T_O} \cdot AF_{TDDB} + P_N^{V_O,T_O} \cdot AF_{NBTI} \qquad (5.15)$$

where $P_E^{VO,TO}$, $P_H^{VO,TO}$, $P_T^{VO,TO}$, and $P_N^{VO,TO}$ are failure percentages of EM, HCI, TDDB, and NBTI at stress conditions (V_O, T_O), respectively. The advantage of using these failure percentages here is to simplify the derivation process without the need to find the absolute failure rate for each failure mechanism.

Due to proprietary issues, manufacturer microelectronic device lifetime data is rarely reported in the literature. To reveal the characteristics of temperature and voltage acceleration at the system (component) level, we can perform lifetime simulation by using the models given above. In this example, the component is assumed to be made with 0.13-μm technology with an oxide thickness of 3.2 nm. Nominal operating conditions are V_O = 1.3 V and T_O = 50°C. HCI, TDDB, and NBTI are assumed to contribute equally to system failures at nominal conditions. All of the acceleration parameters are extracted from published results related to 0.13 μm technology (HCI [195], TDDB [129], and NBTI [81]) and are listed in Table 5.1 along with the simulation parameters. We assume V_O = 1.3 V and T_O = 75°C.

Table 5.1. *Simulation parameters for EM, HCI, TDDB and NBTI*

	Voltage acceleration parameter	Activation energy (eV)	Failure percentage
EM	2	1.2	25%
HCI	16	-0.2	25%
TDDB	12	0.7	25%
NBTI	6	0.4	25%

5.3.1 Non-Arrhenius Temperature Acceleration

Activation energy is designated as $E_{aSYS}^{Vi,Ti}$, which is estimated from accelerated tests at (V_i,T_i) and (V_i,T_A). If the Arrhenius relationship still holds at the V_i,T_i component level, $E_{aSYS}^{Vi,Ti}$ should be the same for all T_i and V_i. The system temperature acceleration factor AF_S^T can be calculated as:

$$AF_S^T = P_E^{V_i,T_i} \cdot AF_{EM}^T + P_H^{V_i,T_i} \cdot AF_{HCI}^T + P_T^{V_i,T_i} \cdot AF_{TDDB}^T + P_N^{V_i,T_i} \cdot AF_{NBTI}^T \quad (5.16)$$

where $P_E^{Vi,Ti}$, $P_H^{Vi,Ti}$, $P_T^{Vi,Ti}$, and $P_N^{Vi,Ti}$ are the percentages of EM, HCI, TDDB, and NBTI failure at (V_i,T_i), respectively.

Using the parameters given in Table 5.1 and setting T_A=125°C, E_{aSYS} estimation at various T_i was simulated at three voltages: 1.17 V, 1.30 V, and 1.43 V and is shown in Figure 5.1. The simulation result clearly shows that E_{aSYS} is not a constant. It depends on the stress voltage V_i and the stress temperature T_i. At a given V_i, $E_{aSYS}^{Vi,Ti}$ is an increasing function of T_i. The reason for this is that failure mechanisms with larger activation energies will increase their failure percentage at high temperature at a given stress voltage. As an illustration, if $|T_A - T_i|$ is considerably small, then system activation energy can be approximated by:

$$E_{aSYS}^{V_i,T_i} = P_{EM}^{V_i,T_i} \cdot E_{aEM} + P_{HCI}^{V_i,T_i} \cdot E_{aHCI} + P_{TDDB}^{V_i,T_i} \cdot E_{aTDDB} + P_{NBTI}^{V_i,T_i} \cdot E_{aNBTI} \quad (5.17)$$

From Equation (5.17), we find that at a given E_{aEM}, E_{aHCI}, $E_{aTDDB,}$ and E_{aNBTI}, $E_{aSYS}^{Vi,Ti}$ depends on $P_E^{Vi,Ti}$, $P_H^{Vi,Ti}$, $P_T^{Vi,Ti}$, and $P_N^{Vi,Ti}$. The failure mechanism with the largest activation energy will be accelerated the most as temperature increases, and its failure percentage increases accordingly.

As E_{aSYS} is generally estimated from high-temperature acceleration testing, using that activation energy tends to suggest an overly optimistic projection at low temperature. For example, if the acceleration tests were done at (1.43 V, 125°C) and (1.43 V, 115°C), the estimated E_{aSYS} is 1.0 eV. Using this activation energy to extrapolate the system failure rate at (1.43 V, 50°C) will get an optimistic estimation that is ***1/14*** of the real rate because the 'true' E_{aSYS} is 0.60 eV.

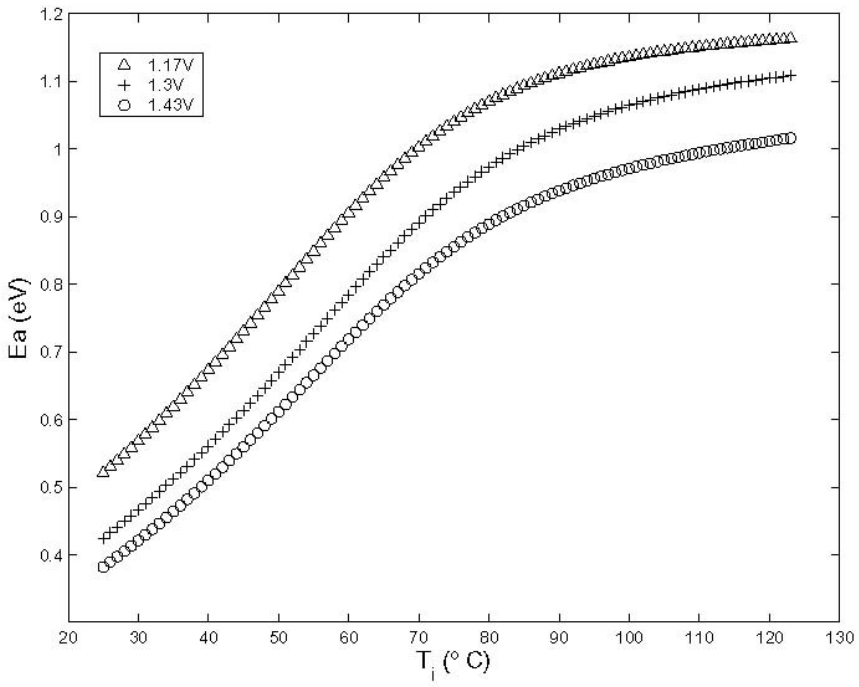

Fig. 5.1. *System activation energies estimated from simulated failure rate at (V_i, T_i) and (V_i, T_A). V_i =1.17 V, 1.30 V and 1.43 V. At a given V_i, T_A = 125°C and T_i varies from 25°C to 124°C.*

5.3.2 Stress-Dependent Voltage Acceleration Factor

To show the characteristic impact of voltage acceleration, we assume AF_S^V follows the exponential law.

$$AF_S^V = \exp[\gamma_{SYS}^{V_i,T_i} \cdot (V_A - V_i)] \tag{5.18}$$

where $\gamma_{SYS}^{Vi,Ti}$ is the voltage acceleration parameter. AF_S^V is shown below.

$$AF_S^V = P_E^{V_i,T_i} \cdot AF_{EM}^V + P_H^{V_i,T_i} \cdot AF_{HCI}^V + P_T^{V_i,T_i} \cdot AF_{TDDB}^V + P_N^{V_i,T_i} \cdot AF_{NBTI}^V \tag{5.19}$$

where $P_E^{Vi,Ti}$, $P_H^{Vi,Ti}$, $P_T^{Vi,Ti}$, and $P_N^{Vi,Ti}$ have the same meaning as presented in Equation (5.16). A simulation was performed with the parameters given in Table 5.1; the estimated γ_{SYS} is shown in Figure 5.2.

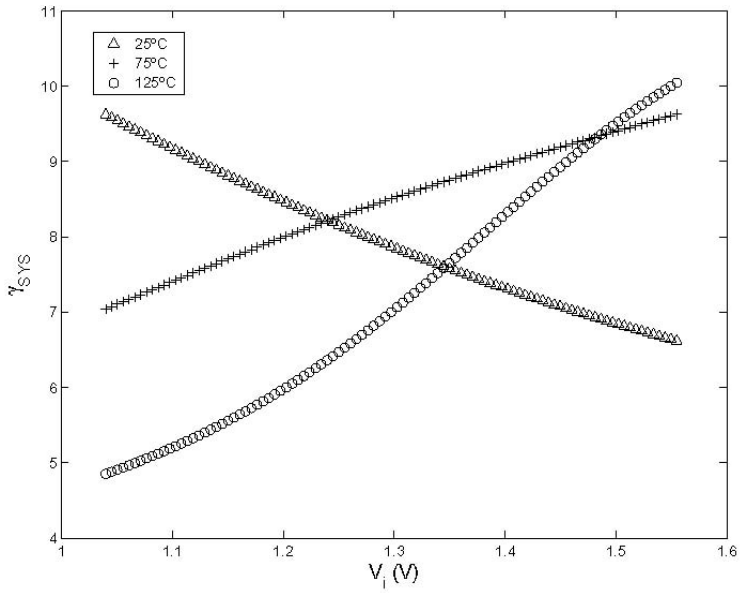

Fig. 5.2. *Estimated γ_{SYS} from failure rates at accelerated conditions (V_i, T_i and V_A, T_i). T_i = 25°C, 75°C, and 125°C. For each T_i, V_A = 1.56 V, V_i varies from 1.04 V to 1.55 V.*

The results show that γ_{SYS} varies according to V_i and T_i. As an approximation, if the difference between V_A and V_i is reasonably small, γ_{SYS} can be approximated by:

$$b_{SYS}^{V_i,T_i} = P_E^{V_i,T_i} \cdot \frac{n}{V_i} + P_H^{V_i,T_i} \cdot \frac{\gamma_{HCI}}{V_i^2} + P_T^{V_i,T_i} \cdot \gamma_{TDDB} + P_N^{V_i,T_i} \cdot \gamma_{NBTI} \tag{5.20}$$

Like $E_{aSYS}^{Vi,Ti}$, $\gamma_{SYS}^{Vi,Ti}$ also depends on the failure percentages and the voltage acceleration parameters. As shown in Figure 5.2, at 125°C, γ_{aSYS} is larger at higher stress voltages because TDDB, together with NBTI, dominate and the higher voltage accelerates them more than EM and HCI.

Using γ_{SYS} estimated at (125°C, 1.55 V) to extrapolate a component failure rate at low voltage will give an overly optimistic estimation. At 125°C and V_i = 1.55 V, γ_{SYS} is estimated to be 10.0, while we will get 7.0 if V_i = 1.30 V. In this case there is an approximate ***5X*** difference in failure rate extrapolation.

5.3.3 Combined Voltage and Temperature Acceleration Factor

Considering the voltage and temperature acceleration effect together, system acceleration is further complicated by the interplay between voltage and temperature acceleration, as shown above. Since there is no universal E_{aSYS} and γ_{SYS} of multiple failure mechanisms, using AF_T with one activation energy and AF_V with one voltage acceleration parameter for reliability extrapolation is not appropriate. Taking the simulation above as an example, we find out that failure rate estimation using the multiplication model gives an overly optimistic result. The real system failure rate at (50°C, 1.30 V) is ***20X*** that of the estimated failure rate using the multiplication model with E_{aSYS} and γ_{SYS} from high-temperature, high-voltage acceleration testing at (125°C, 1.55 V).

5.4 Qualification Based on Failure Mechanism

It is a matter of great complexity to build a system lifetime model to fit all temperatures and voltages if there are multiple failure mechanisms at work. The conventional extrapolation method using one E_{aSYS} and γ_{SYS} tends to give an overly optimistic estimation. For systems with strict reliability requirements (such as aerospace avionics), more accurate reliability projections are necessary for system design and qualification. Using acceleration parameters obtained at high-temperature, high-voltage acceleration testing cannot be justified because stress conditions tend to accelerate failure mechanisms with a high positive activation energy and a larger voltage acceleration parameter, such as TDDB, while EM and HCI failures are more common in field applications. To improve the accuracy of reliability qualification, all failure mechanisms should be considered in the qualification approach.

For reliability qualification considering multiple failure mechanisms, acceleration tests should be designed to accelerate the target failure mechanism with specific stress conditions. This is workable because each failure mechanism has its unique activation energy and voltage acceleration parameter. Among these failure mechanisms, only HCI has a negative activation energy; others are positive. This means lowering the stress temperature will accelerate HCI while decelerating the other three failure mechanisms. HCI also has a comparably large γ. In an application at low temperature and at a reasonable high voltage, HCI failures will dominate. For EM, since copper

interconnect has a large activation energy and a small γ (≤ 2), an acceleration test should be designed with high temperature and low voltage. The traditional acceleration test with high temperature and voltage can be applied to accelerate TDDB and NBTI since both have a large voltage acceleration parameter and activation energy. A simulation of failure percentages of each failure mechanism at various accelerated conditions is shown in Figure 5.3.

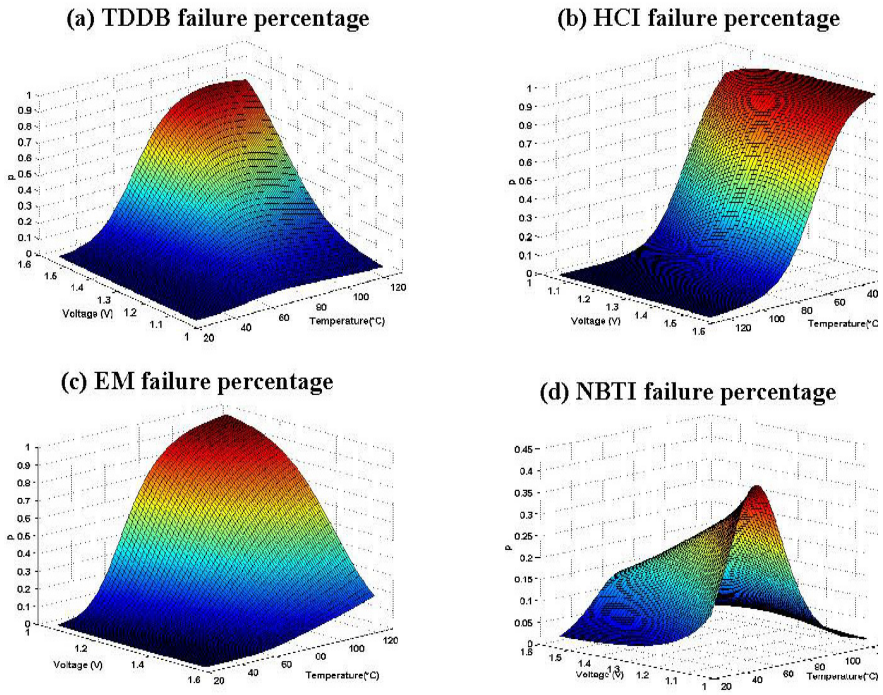

Fig. 5.3. *Failure percentages of EM, HCI, TDDB and NBTI at different accelerated conditions.*

5.5 Summary

For semiconductor devices, reliability modeling at the system or component level is complicated by the involvement of multiple failure mechanisms that have the same stress factors: voltage and temperature. The Arrhenius relationship with one activation energy for all temperatures has shown to not be valid at the component level if these failure mechanisms do not have the same activation energy. The same result can be observed when modeling voltage dependence. Using the exponential law with only one constant coefficient is a good option for an individual failure mechanism, but not for the component. A failure-mechanism-based qualification methodology using specifically designed stress conditions over traditional approaches (i.e., one

voltage and one temperature) can lead to improved reliability predictions for targeted applications and optimized burn-in, screening, and qualification test plans.

REFERENCES

1. J. A. Mclinn, "Constant failure rate—a paradigm in transition?," *Quality and Reliability Engineering International*, vol. 6, pp. 237–241, 1990.
2. J. B. Bowles, "A survey of reliability-prediction procedures for microelectronic devices," *IEEE Transactions on Reliability*, vol. 41, pp. 2–12, March 1992.
3. *MIL-HDBK-217F Reliability Prediction of Electronic Equipment.* Department of Defense, 1991.
4. M. G. Pecht and F. R. Nash, "Predicting the reliability of electronic equipment," *Proceedings of the IEEE*, vol. 82, pp. 992–1004, July 1994.
5. W. Denson, "The history of reliability prediction," *IEEE Transactions on Reliability*, vol. 47, no. 3, pp. SP321–SP328, 1998.
6. S. F. Morris and J. F. Reilly, "MIL-HDBK-217—a favorite target," in *Annual Reliability and Maintainability*, pp. 503–509, IEEE, Jan 1993.
7. *Electronic Reliability Design Handbook*. Department of Defense, 1998.
8. B. Foucher, J. Boullié, B. Meslet, and D. Das, "A review of reliability prediction methods for electronic devices," *Microelectronics Reliability*, vol. 42, pp. 1155–1162, Aug 2002.
9. P. Charpenel, F. Favenel, R. Digout, M. Giraudeau, M. Glade, J. Guerveno, N. Guillet, A. Lauriac, S. Male, D. Manteigas, R. Meister, E. Moreau, D. Perie, F. Relmy-Madinska, and P. Retailleau, "The right way to assess electronic system reliability: FIDES," *Microelectronics Reliability*, vol. 43, pp. 1041–1404, 2003.
10. J. J. Marin and R. W. Pollard, "Experience report on the FIDES reliability prediction method," in *Annual Reliability and Maintainability Symposium*, pp. 8–13, IEEE, 2005.
11. M. Pecht and W.-C. Kang, "A critique of MIL-HDBK-217E reliability prediction methods," *IEEE Transactions on Reliability*, vol. 37, pp. 453–457, Dec 1988.
12. "IEEE guide for selecting and using reliability predictions based on IEEE 1413." IEEE Standards Coordinating Committee 37, 2002.
13. J. Srinivasan, S. V. Adve, P. Bose, J. A. Rivers, and C.-K. Hu, "RAMP: A model for reliability aware microprocessor design," IBM Research Report, 2003.
14. M. J. Cushing, D. E. Mortin, T. J. Stadterman, and A. Malhotra, "Comparison of electronics-reliability assessment approaches," *IEEE Transactions on Reliability*, vol. 42, no. 4, pp. 542–546, 1993.
15. X. Li, B. Huang, J. Qin, X. Zhang, M. Talmor, Z. Gur, and J. B. Bernstein, "Deep submicron CMOS integrated circuit reliability simulation with SPICE," in *IEEE Proceedings of the Sixth International Symposium on Quality Electronic Design*, pp. 382–389, IEEE, 2005.
16. X. Li, J. Qin, B. Huang, X. Zhang, and J. B. Bernstein, "SRAM circuit-failure modeling and reliability simulation with SPICE," *IEE*E *Transactions on Device and Materials Reliability*, vol. 2, no. 2, pp. 235–246, 2006.
17. M. Ohring, *Reliability and Failure of Electronic Materials and Devices*, ch. 5, p. 259. Academic Press, 1998.

18. D. G. Pierce and P. G. Brusius, “Electromigration: A review,” *Microelectron Reliability*, vol. 37, pp. 1053–1072, 1997.
19. J. R. Black, “Mass transport of aluminum by moment exchange with conducting electrons,” in *6th Annual International Reliability Physics Symposium*, pp. 148–159, 1967.
20. A. Scorzoni, B. Neri, C. Caprile, and F. Fantini, “Electromigration in thin-film interconnection lines: Models, methods and results,” *Materials Science Reports*, vol. 7, pp. 143–220, 1991.
21. D. Young and A. Christou, “Failure mechanism models for electromigration,” *IEEE Transactions on Reliability*, vol. 43, pp. 186–192, June 1994.
22. A. Christou, *Electromigration & Electronic Device Degradation*, ch. 10, pp. 324–331. John Wiley Sons, 1994.
23. M. Ohring, *Reliability and Failure of Electronic Materials and Devices*, ch. 5, p. 270. Academic Press, 1998.
24. J. Cho and C. V. Thompson, “Grain size dependence of electromigrationinduced failures in narrow interconnects,” *Applied Physics Letter*, vol. 54, pp. 2577–2579, 1989.
25. A. S. Oates, “Electromigration failure distribution of contacts and vias as a function of stress conditions in submicron IC metallizations,” in *34th Annual Proceedings of Reliability Physics Symposium*, (Dallas, TX), pp. 164–171, 1996.
26. K. Hinode, T. Furusawa, and Y. Homma, “Dependence of electromigration lifetime on the square of current density,” in *IEEE International Reliability Physics Symposium 1993*, pp. 317–326, 1993.
27. J. A. Schwarz, “Distributions of activation energies for electromigration damage in thin-film aluminum interconnects,” *Journal of Applied Physics*, vol. 61, pp. 798–800, 1987.
28. J. R. Lloyd, “On the log-normal distribution of electromigration lifetimes,” *Journal of Applied Physics*, vol. 50, pp. 5062–5064, 1979.
29. A. Bobbio and O. Saracco, “On the spread of time-to-failure measurements in thin metallic films,” *Thin Solid Films*, vol. 17, no. 1, pp. S13–S16, 1973.
30. M. J. Attardo, R. Rutledge, and R. C. Jack, “Statistical metallurgical model for electromigration failure in aluminum thin-film conductors,” *Journal of Applied Physics*, vol. 42, pp. 4343–4349, 1971.
31. J. M. Towner, “Are electromigration failures lognormal distributed?,” in *IEEE International Reliability Physics Symposium*, pp. 100–105, 1990.
32. M. Gall, C. Capasso, D. Jawarani, R. Hernandez, and H. Kawasaki, “Statistical analysis of early failures in electromigration,” *Applied Physics*, vol. 90, no. 2, pp. 732–740, 2001.
33. J. R. Lloyd, “Reliability modeling for electromigration failure,” *Quality and Reliability Engineering International*, vol. 10, pp. 303–308, 1994.
34. C. Pennetta, L. Reggiani, G. Trefan, F. Fantini, A. Scorzoni, and I. D. Munari, “A percolative approach to electromigration in metallic lines,” *Journal of Physics D: Applied Physics*, vol. 3, pp. 1421–1429, 2001.
35. A. H. Fischer, A. Abel, M. Lepper, A. E. Zitzelsberger, and A. V. Glasow, “Experimental data and statistical models for bimodal EM failures,” in *IEEE 38th Annual International Reliability Physics Symposium*, (San Jose), pp. 359–364, 2000.
36. J. B. Lai, J. L. Yang, Y. P. Wang, S. H. Chang, R. L. Hwang, Y. S. Huang, and C. S. Hou, “A study of bimodal distributions of time-to-failure of copper via electromigration,” in *International Symposium on VLSI Technology, Systems, and Applications*, pp. 271–274, 2001.

37. E. T. Ogawa, K. D. Lee, H. Matsuhashi, et al., "Statistics of electromigration early failures in Cu/Oxide dual-damascene interconnects," in *39th Annual International Reliability Physics Symposium*, (Orlando, Florida), pp. 341–349, 2001.
38. A. Acovic, G. L. Rosa, and Y.-C. Sun, "A review of hot carrier deration mechanisms in MOSFETs," *Microelectronics Reliability*, vol. 36, pp. 845–869, 1996.
39. E. Takeda, C. Y. Yang, and A. Miura-Hamada, *Hot-Carrier Effects in MOS Devices*, ch. 2, pp. 49–58. Academic Press, 1995.
40. M. Song, K. P. MacWilliams, and J. C. S. Woo, "Comparison of NMOS and PMOS hot carrier effects from 300 to 77 k," *IEEE Transactions on Electron Devices*, vol. 44, pp. 268–276, 1997.
41. J. F. Verwey, R. P. Kramer, and B. J. de Maagt, "Mean free path of hot electrons at the surface of boron-doped silicon," *Journal of Applied Physics*, vol. 46, pp. 2612–2619, 1975.
42. T. H. Ning, C. M. Osbum, and H. N. Yu, "Emission probability of hot electrons from silicon into silicon dioxide," *Journal of Applied Physics*, vol. 48, pp. 286–293, 1977.
43. C. Hu, "A lucky-electron model of hot-electron emission," in *International Electron Devices Meeting Technical Digest*, pp. 22–25, 1979.
44. S. Tam, P.-K. Ko, and C. Hu, "Lucky-Electron model of channel hot-electron injection in MOSFET's," *IEEE Transactions on Electron Devices*, vol. ED-31, pp. 1116–1125, 1984.
45. C. Hu, S. C. Tam, F.-C. Hsu, P.-K. Ko, T.-Y. Chan, and K. W. Terrill, "Hotelectron-induced MOSFET degradation-model, monitor, and improvement," *IEEE Journal of Solid-State Circuits*, vol. SC-20, pp. 295–305, 1985.
46. W. Weber, "Dynamic stress experiments for understanding hot-carrier degradation phenomena," *IEEE Transactions on Electron Devices*, vol. 35, pp. 1476–1486, 1988.
47. E. Sangiorgi, "Historical perspective and recent developments of hot-carrier generation modeling for device analysis," in *International Conference on Simulation of Semiconductor Processes and Devices*, pp. 5–8, 1997.
48. E. Takeda and N. Suzuki, "An empirical model for device degradation due to hot-carrier injection," *IEEE Electron Device Letters*, vol. EDL-4, pp. 111–113, 1983.
49. S.-H. Renn, J.-L. Pelloie, and F. Balestra, "Hot-carrier effects and reliable lifetime prediction in deep submicron n-and p-channel SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. 45, pp. 2335–2342, 1998.
50. B. Szelag, S. Kubicek, K. D. Meyer, and F. Balestra, "Time-dependent degradation law for reliable lifetime prediction in sub-0.25 μm bulk silicon NMOSFETs," *Electronics Letters*, vol. 5, pp. 1385–1386, 1999.
51. B. Marchand, G. Ghibaudo, F. Balestra, and G. Guegan, "A new hot carrier degradation law for MOSFET lifetime prediction," *Microelectronics Reliability*, vol. 38, pp. 1103–1107, 1998.
52. V.-H. Chan and J. E. Chung, "Two-stage hot carrier degradation and its impact on submicronmeter LDD NMOSFET lifetime prediction," *IEEE Transactions on Electron Devices*, vol. 42, pp. 957–962, May 1995.
53. N. Koike and H. Yonezawa, "A modeling methodology and body effect analysis for hot-carrier reliability simulation of logic circuits*," IEICE Transactions on Electronics*, vol. E85-C, pp. 1356–1365, 2002.
54. E. Takeda, C. Y. Yang, and A. Miura-Hamada, *Hot-Carrier Effects in MOS Devices*, ch. 5, pp. 124–125. Academic Press, 1995.

55. JEDEC, *Failure Mechanisms and Models for Semiconductor Devices*. JEDEC Solid State Technology Association, 2003.
56. H. Iwai, H. S. Momose, M. Saito, M. Ono, and Y. Katsumata, "The future of ultra-small-geometry MOSFETs beyond 0.1 micron*," Microelectronic Engineering*, vol. 28, pp. 147–154, 1995.
57. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, ch. 2, pp. 85–85. Cambridge University Press, 1998.
58. S.-H. Lo, D. A. Buchanan, and Y. Taur, "Modeling and characterization of quantization, polysilicon depletion and direct tunneling effects in MOSFETs with ultrathin oxides," *IBM Journal of Research and Development*, vol. 43, pp. 327–337, 1999.
59. K. F. Schuegraf and C. Hu, "Hole injection SiO_2 breakdown model for very low voltage lifetime extrapolation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 761–767, 1994.
60. J. F. Verweij and J. H. Klootwijk, "Dielectric breakdown i: A review of oxide breakdown," *Microelectronics*, vol. 27, pp. 611–622, 1996.
61. J. H. Stathis, "Physical and predictive models of ultrathin oxide reliability in CMOS devices and circuits," *IEEE Transactions on Device and Materials Reliability*, vol. 1, pp. 43–59, 2001.
62. M. V. Fischetti, "Model for the generation of positive charge at the $Si–SiO_2$ interface based on hot-hole injection from the anode," *Physical Review B*, vol. 31, pp. 2099–2113, 1985.
63. J. S. Suehle, "Ultrathin gate oxide reliability: Physical models, statistics, and characterization," *IEEE Transactions on Electron Devices*, vol. 49, pp. 958–971, 2002.
64. J. W. McPherson and H. C. Mogul, "Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO_2 thin films," *Journal of Applied Physics*, vol. 84, pp. 1513–1523, 1998.
65. E. M. Vogel, J. S. Suehle, M. D. Edelstein, B. Wang, Y. Chen, and J. B. Bernstein, "Reliability of ultrathin silicon dioxide under combined substrate hot-electron and constant voltage tunneling," *IEEE Transactions on Electron Devices*, vol. 47, pp. 1183–1191, 2000.
66. D. J. DiMaria and J. W. Stasiak, "Trap creation in silicon dioxide produced by hot electrons," *Journal of Applied Physics*, vol. 65, pp. 2342–2356, 1988.
67. J. Wu, E. Rosenbaum, B. MacDonald, E. Li, J. Tao, B. Tracy, and P. Fang, "Anode hole injection versus hydrogen release: The mechanism for gate oxide breakdown," in *International Reliability Physics Symposium*, (San Jose, CA), pp. 27–32, 2000.
68. R. Degraeve, B. Kaczer, and G. Groesenken, "Degradation and breakdown in thin oxide layers: Mechanisms, models and reliability prediction," *Microelectronics Reliability*, vol. 39, pp. 1445–1460, 1999.
69. E. Wu, J. Sune, W. Lai, E. Nowak, J. McKenna, A. Vayshenker, and D. Harmon, "Interplay of voltage and temperature acceleration of oxide breakdown for ultrathin gate oxides," *Solid-State Electronics*, vol. 46, pp. 1787–1798, 2002.
70. J. Sune, I. Placencia, N. Barniol, E. Farres, F. Martin, and X. Aymerich, "On the breakdown statistics of very thin SiO_2 films," *Thin Solid Films*, vol. 185, pp. 347–362, 1990.
71. D. J. Dumin, S. K. Mopuri, S. Vanchinathan, R. S. Scott, R. Subramoniam, and T. G. Lewis, "High field related thin oxide wearout and breakdown," *IEEE Transactions on Electron Devices*, vol. 42, pp. 760–772, 1995.
72. J. H. Stathis, "Percolation models for gate oxide breakdown," *Journal of Applied Physics*, vol. 86, pp. 5757–5766, 1999.

73. E. Y. Wu and R. P. Vollertsen, "On the Weibull shape factor of intrinsic breakdown of dielectric films and its accurate experimental determination— Part I: Theory, methodology, experimental techniques," *IEEE Transactions on Electron Devices*, vol. 49, pp. 2131–2140, 2002.
74. E. Y. Wu, J. Sune, and W. Lai, "On the Weibull shape factor of intrinsic breakdown of dielectric films and its accurate experimental determination—Part II: Experimental results and the effects of stress conditions," *IEEE Transactions on Electron Devices*, vol. 49, pp. 2141–2150, 2002.
75. E. Y. Wu, W. W. Abadeer, L.-K. Han, S.-H. Lo, and G. R. Hueckel, "Challenges for accurate reliability projections in the ultrathin oxide regime," in *International Reliability Physics Symposium*, pp. 57–65, 1999.
76. R. Degraeve, G. Groeseneken, R. Bellens, J. L. Ogier, M. Depas, P. J. Roussel, and H. E. Maes, "New insights in the relation between electron trap generation and the statistical properties of oxide breakdown," *IEEE Transactions on Electron Devices*, vol. 45, pp. 904–911, 1998.
77. B. E. Deal, M. Sklar, A. S. Grove, and E. H. Snow, "Characteristics of the surface-state charge of thermally oxidized silicon," *Journal of the Electrochemical Society*, vol. 114, pp. 266–274, 1967.
78. B. E. Deal, "The current understanding of charges in the thermally oxidized silicon structure," *Journal of the Electrochemical Society*, vol. 121, pp. 198c–205c, 1974.
79. A. Goetzberger, A. D. Lopez, and R. J. Strain, "On the formation of surface states during stress aging of thermal Si–SiO_2 interfaces," *Journal of the Electrochemical Society*, vol. 120, pp. 90–96, 1973.
80. K. O. Jeppson and C. M. Svensson, "Negative bias stress of MOS devices at high electric fields and degradation of MNOS devices," *Journal of Applied Physics*, vol. 48, pp. 2004–2014, 1977.
81. D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron semiconductor manufacturing," *Journal of Applied Physics*, vol. 94, pp. 1–18, 2003.
82. S. Ogawa and N. Shiono, "Generalized diffusion-reaction model for the low-field charge-buildup instability at the Si–SiO_2 interface," *Physical Review B*, vol. 51, no. 7, pp. 4218–4230, 1995.
83. S. N. Rashkeev, D. M. Fleetwood, R. D. Schrimpf, and S. T. Pantelides, "Proton-induced defect generation at the Si–SiO_2 interface," *IEEE Transactions on Nuclear Science*, vol. 48, pp. 2086–2092, 2001.
84. S. Chakravarthi, A. T. Krishnan, V. Reddy, C. F. Machala, and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," in *42nd International Reliability Physics Symposium*, pp. 273–282, IEEE, 2004.
85. W. Abadeer and W. Ellis, "Behavior of NBTI under AC dynamic circuit conditions," in *International Reliability Physics Symposium*, pp. 17–22, 2003.
86. S. Rangan, N. Mielke, and E. C. C. Yeh, "Universal recovery behavior of negative bias temperature instability," *in IEEE International Electron Devices Meeting*, pp. 341–344, 2003.
87. G. Chen, M. F. Li, C. H. Ang, J. Z. Zheng, and D. L. Kwong, "Dynamic NBTI of PMOS transistors and its impact on MOSFET scaling," *IEEE Electron Device Letters*, vol. 23, pp. 734–736, 2002.

88. G. Haller, M. Knoll, D. Braunig, F. Wulf, and W. R. Fahrner, "Biastemperature stress on metal-oxide-semiconductor structures as compared to ionizing irradiation and tunnel injection," *Journal of Applied Physics*, vol. 56, p. 184, 1984.
89. P. Chaparala, J. Shibley, and P. Lim, "Threshold voltage drift in PMOSFETS due to NBTI and HCI," in *Integrated Reliability Workshop*, pp. 95–97, IEEE, 2000.
90. C.-H. Liu, M. T. Lee, C.-Y. Lin, J. Chen, Y. T. Loh, F.-T. Liou, K. Schruefer, A. A. Katsetos, Z. Yang, N. Rovedo, T. B. Hook, C. Wann, and T.-C. Chen, "Mechanism of threshold voltage shift caused by negative bias temperature instability in deep submicron PMOSFETs," *Japanese Journal of Applied Physics*, vol. 41, pp. 2423–2425, 2002.
91. M. Singh and I. Koren, "Fault-sensitivity analysis and reliability enhancement of analog-to-digital converters," *IEEE Transactions on VLSI Systems*, vol. 11, pp. 839–852, 2003.
92. D. Gil, J. C. Baraza, and J. V. Busquets, "Fault injection into VHDL models: Analysis of the error syndrome of a microcomputer system," in *Euromicro Conference*, 1998.
93. K. N. Quader, P. Fang, and J. T. Yue, "Hot carrier reliability design rules for translating device degradation to CMOS digital circuit degradation," *IEEE Transactions on Electron Devices*, vol. 41, pp. 681–691, 1994.
94. U. of California, "SPICE version 2g.6 user's guide," 1994.
95. C. D. Systems, "BSIMProPlus datasheet," 2003.
96. J. D. Walter and J. B. Bernstein, "Semiconductor device lifetime enhancement by performance reduction," tech. rep., University of Maryland, ENRE, 2003.
97. S. M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design*. McGraw-Hill College, 2002.
98. J. M. Rabaey, A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*. Pearson Education, Inc., 2nd ed., 2003.
99. R. S. Muller, T. I. Kamins, and M. S. Chan, *Device Electronics for Integrated Circuits*. John Wiley & Sons, 3rd ed., 2002.
100. W. Liu, *MOSFET Models for SPICE Simulation, Including BSIM3v3 and BSIM4*. John Wiley & Sons, 2001.
101. Y. H. Cheng and C. M. Hu, *MOSFET Modeling & User's Guide*. Kluwer Academic Publishers, 1999.
102. J. D. Walter, *Methods to Account for Accelerated Semiconductor Wearout in Longlife Aerospace Applications*. PhD thesis, University of Maryland, College Park, 2003.
103. M. J. Lakey, "Statistical analysis of field data for aircraft warranties," in *Proceedings of Annual Reliability Maintainability Symposium*, pp. 340–344, 1991.
104. Y. Chen, D. Nguyen, S. Guertin, J. Bernstein, M. White, R. Menke, and S. Kayali, "A reliability evaluation methodology for memory chips for space applications when sample size is small," in *IRW*, pp. 91–94, IEEE, 2003.
105. A. M. Abo, *Design for Reliability of Low-Voltage, Switched-Capacitor Circuits*. PhDthesis, University of California, Berkeley, 1999.
106. M. Choi and A. A. Abidi, "A 6-b 1.3-Gsample/S A/D converter in 0.35 μm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 36, pp. 1847–1858, 2001.
107. G. Groeseneken, R. Degraeve, T. Nigram, G. V. D. Bosch, and H. E. Maes, "Hot carrier degradation and time-dependent dielectric breakdown in oxides," *Microelectronic Engineering*, vol. 49, pp. 27–40, 1999.

108. Y. Leblebici and S. M. Kang, *Hot Carrier Reliability of MOS VLSI Circuits*, p. 61. Kluwer Academic Publisher, 1993.
109. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, p. 157. Cambridge University Press, 1998.
110. C. C. Li, K. N. Quader, E. R. Minami, C. Hu, and P. K. Ko, "A new bidirectionalPMOSFET hot carrier degradation model for circuit reliability simulation," in *IEEE International Electron Devices Meeting*, vol. 1992, pp. 547–550, 1992.
111. W. J. Hsu, B. J. Sheu, S. M. Gowda, and C. G. Hwang, "Advanced integrated-circuit reliability simulation including dynamic stress effects," *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 247–257, 1992.
112. W. Liu, X. Jin, J. Chen, M. Jeng, Z. Liu, Y. Cheng, K. Chen, M. Chan, K. Hui, J. Huang,R. Tu, P. K. Ko, and C. Hu, *BSIM3V3. 2. 2 MOSFET Model User's Manual*. University of California, Berkeley, 1999.
113. P.-C. Li, G. I. Stamoulis, and I. N. Hajj, "iProbe-d: A hot carrier and oxide reliability simulator," in *International Reliability Physics Symposium*, pp. 274–279, 1994.
114. K. N. Quader, C. C. Li, R. Tu, E. Rosenbaum, P. K. Ko, and C. Hu, "A bidirectional NMOSFET current reduction model for simulation of hot-carrierinduced circuit degradation," *IEEE Transactions on Electron Devices*, vol. 40, pp. 2245–2254, 1993.
115. K. N. Quader, C. Li, R. Tu, E. Rosenbaum, P. Ko, and C. Hu, "A new approach for simulation of circuit degradation due to hot-electron damage in NMOSFETs," in *IEEE International Electron Devices Meeting*, pp. 337–340, 1991.
116. Y. Leblebici and S.-M. Kang, "Modeling of NMOS transistors for simulation of hot-carrier induced device and circuit degradation*," IEEE Transactions on Computer-Aided Design*, vol. 11, pp. 235–246, 1992.
117. Y. Leblebici and S. M. Kang, "Modeling and simulation of hot-carrier-induced device degradation in MOS circuits," *IEEE Journal of Solid-State Circuits*, vol. 28, pp. 585–595, 1993.
118. N. Hwang and L. Forbes, "Hot-carrier induced series resistance enhancement model (HISREM) of NMOSFET's for circuit simulation and reliability projections," *Microelectronic Reliability*, vol. 35, pp. 225–239, 1995.
119. C. C. Li, K. N. Quader, E. R. Minami, C. Hu, and P. K. Ko, "A new bidirectional PMOSFET hot carrier degradation model for circuit reliability simulation," in *International Electron Devices Meeting*, pp. 547–550, 1992.
120. Y. Chen, J. S. Suehle, C. C. Shen, J. Bernstein, C. Messick, and P. Chaparala, "The correlation of highly accelerated q_{bd} tests to TDDB life tests for ultrathin gate oxides," in *IEEE Annual International Reliability Physics Symposium*, pp. 87–91, 1998.
121. Y. C. Yeo, Q. Lu, and C. Hu, "MOSFET gate oxide reliability: Anode hole injection model and its applicaitons," *International Journal of High Speed Electronics and Systems*, vol. 11, pp. 849–886, 2001.
122. N. Shiono and M. Itsumi, "A lifetime projection method using series model and acceleration factors for TDDB failures of thin gate oxides," in *1993 International Reliability Physics Symposium*, (Atlanta, GA), pp. 1–6, IEEE, 1993.
123. C. Hu and Q. Lu, "A unified gate oxide reliability model," *in IEEE Annual International Reliability Physics Symposium*, pp. 47–51, 1999.

124. F. Monsieur, E. Vincent, D. Roy, S. Bruyere, G. Pananakakis, and G. Ghibaudo, "Time to breakdown and voltage to breakdown modeling for ultrathin oxides (Tox<32A)," in *2001 Integrated Reliability Workshop*, pp. 20–25, 2001.
125. G. Ribes, S. Bruyere, F. Monsieur, D. Roy, and V. Huard, "New sights into the change of voltage acceleration and temperature activation of oxide breakdown," *Microelectronics Reliability*, vol. 43, pp. 1211–1214, 2003.
126. F. Monsieur, E. Vincent, D. Roy, S. Bruyere, J. Vildeuil, G. Pananakakis, and G. Ghibaudo, "A thorough investigation of progressive breakdown in ultrathin oxides. physical understanding and application for industrial reliability assessment," in *International Reliability Physics Symposium*, pp. 45–54, 2002.
127. E. Y. Wu, E. J. Nowak, A. Vayshenker, W. L. Lai, and D. L. Harmon, "CMOS scaling beyond the 100-nm node with silicon-dioxide-based gate dielectrics," *IBM Journal of Research and Development*, vol. 46, 2002.
128. E. Wu, J. Sune, and W. Lai, "Interplay of voltage and temperature acceleration of oxide breakdown for ultrathin oxides," *Microelectronic Engineering*, vol. 59, pp. 25–31, 2001.
129. E. Wu and J. Sune, "Power-law voltage acceleration: A key element for ultrathin gate oxide reliability," Microelectronics Reliability, 2005.
130. J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "The case for lifetime reliability-aware microprocessor," in *IEEE Proceedings of 31st Annual International Symposium on Computer Architecture*, 2004.
131. J. H. Stathis, "Physical models of ultra thin oxide reliability in CMOS devices and implications for circuit reliability," in *International Workshop on Gate Insulators*, pp. 26–29, 2001.
132. J. H. Stathis, R. Rodriguez, and B. P. Linder, "Circuit implications of gate oxide breakdown," *Microelectronics Reliability*, vol. 43, pp. 1193–1197, 2003.
133. R. Rodriguez, J. H. Stathis, and B. P. Linder, "The impact of gate-oxide breakdown on SRAM stability," *IEEE Electron Device Letters*, vol. 23, pp. 559–561, 2002.
134. R. Rodriguez, J. H. Stathis, and B. P. Linder, "A model for gate-oxide breakdown in CMOS inverters," *IEEE Electron Device Letters*, vol. 24, pp. 114–116, 2003.
135. J. Segura, C. D. Benito, A. Rubio, and C. F. Hawkins, "A detailed analysis of GOS defects in MOS transistors: Testing implications at circuit level," in *IEEE International Test Conference*, pp. 544–551, 1995.
136. J. Segura, C. D. Benito, A. Rubio, and C. F. Hawkins, "A detailed analysis and electrical modeling of gate oxide shorts in MOS transistors," *Journal of Electronic Testing: Theory and Applications*, vol. 8, pp. 229–239, 1996.
137. A. Avellan and W. H. Krautschneider, "Impact of soft and hard breakdown on analog and digital circuits," *IEEE Transactions on Device and Materials Reliability*, vol. 4, pp. 676–680, 2004.
138. H. Yang, J. S. Yuan, Y. Liu, and E. Xiao, "Effect of gate oxide breakdown on RF performance," *IEEE Transactions on Device and Materials Reliability*, vol. 3, pp. 93–97, 2003.
139. H. Yang, J. S. Yuan, and E. Xiao, "Effect of gate oxide breakdown on RF device and circuit performance," in *IEEE Annual International Reliability Physics Symposium*, pp. 1–4, 2003.

140. B. Kaczer, R. Degraeve, M. Rasras, A. D. Keersgieter, K. V. D. Mieroop, and G. Groeseneken, "Analysis and modeling of a digital CMOS circuit operation and reliability after gate oxide breakdown: A case study," *Microelectronics Reliability*, vol. 42, pp. 555–564, 2002.
141. B. Kaczer, R. Degraeve, M. Rasras, K. V. de Mieroop, P. J. Roussel, and G. Groeseneken, "Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability," *IEEE Transactions on Electron Devices*, vol. 49, pp. 500–506, 2002.
142. B. Kaczer, R. Degraeve, A. D. Keersgieter, K. V. D. Mieroop, V. Simons, and G. Groeseneken, "Consistent model for short-channel NMOSFET after hard gate oxide breakdown," *IEEE Transactions on Electron Devices*, vol. 49, pp. 507–513, 2002.
143. R. Degraeve, B. Kaczer, A. D. Keersgieter, and G. Groseneken, "Relation between breakdown mode and location in short-channel NMOSFETs and its impact on reliability specifications," *IEEE Transactions on Device and Materials Reliability*, vol. 1, pp. 163–169, 2001.
144. B. Kaczer, F. Crupi, R. Degraeve, P. Roussel, C. Ciofi, and G. Groeseneken, "Observation of hot-carrier-induced nFET gate-oxide breakdown in dynamically stressed CMOS circuits," in *International Electron Devices Meeting*, pp. 171–174, 2002.
145. B. Kaczer, R. Degraeve, A. D. Keersgieter, K. V. D. Mieroop, T. Bearda, and G. Groeseneken, "Consistent model for short-channel NMOSFET post-hard-breakdown characteristics," in *Symposium on VLSI Technology Digest of Technical Papers*, pp. 121–122, 2001.
146. M. Shaheen and S. Mourad, "Gate to channel shorts in PMOS devices: Effects on logic gate failures," in *Proceedings of IEEE International Symposium on Circuits and Systems*, pp. 211–214, IEEE, 1998.
147. X. Lu, Z. Li, W. Qiu, H. Walker, and W. Shi, "A circuit level fault model for resistive shorts of MOS gate oxide," in *5th International Workshop on Microprocessor Test and Verification*, 2004.
148. T.-S. Yeoh, N. R. Kamat, R. S. Nair, and S.-J. Hu, "Gate oxide breakdown model in MOS transistors," in *Proceedings of IEEE International Reliability Physics*, pp. 149–155, IEEE, 1995.
149. T.-S. Yeoh and S.-J. Hu, "Influence of MOS transistor gate oxide breakdown on circuit performance," in *Proceedings of the IEEE International Conference on Semiconductor Electronics*, pp. 59–63, 1998.
150. M. Renovell, J. M. Galliere, F. Azais, and Y. Bertrand, "MDelay testing of MOS transistor with gate oxide short," in *Proceedings of the Twelfth Asian Symposium*, pp. 168–173, 2003.
151. M. Renovell, J. M. Galliere, F. Azais, and Y. Bertrand, "Modeling gate oxide short defects in CMOS minimum transistors," in *Proceedings of the Seventh IEEE European Test Workshop*, pp. 15–20, 2002.
152. V. Huard, M. Denais, and C. Parthasarathy, "NBTI degradation: From physical mechanisms to modeling," *Microelectronics Reliability*, 2005.
153. V. Reddy, J. Carulli, A. Krishnan, W. Bosch, and B. Burgess, "Impact of negative bias temperature instability on production parameter drift," in *IEEE ITC International Test Conference*, pp. 148–155, 2004.

154. V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, "Impact of negative bias temperature instability on digital circuit reliability," *Microelectronics Reliability*, vol. 45, pp. 31–38, 2005.
155. A. S. Goda and G. Kapila, "Design for degradation CAD tools for managing transistor degradation mechanisms," in *IEEE Proceedings of the Sixth International Symposium on Quality Electronic Design*, pp. 416–420, 2005.
156. M. Ershov, S. Saxena, S. Minehane, P. Clifton, M. Redford, R. Lindley, H. Karbasi, S. Graves, and S. Winters, "Degradation dynamics, recovery, and characterization of negative bias temperature instability," *Microelectronics Reliability*, vol. 45, pp. 99–105, 2005.
157. H. Aono, E. Murakami, K. Okuyama, A. Nishida, M. Minami, Y. Ooji, and K. Kubota, "Modeling of NBTI saturation effect and its impact on electronic field dependence of the lifetime," *Microelectronics Reliability*, 2005.
158. A. Narr and A. Lill, "Lifetime prediction for PMOS and NMOS devices biased on adegradation model for gate-bias-stress," *Microelectronics Reliability*, vol. 37, pp. 1433–1436, 1997.
159. Y. H. Lee, S. Jacobs, S. Stadler, N. Mielke, and R. Nachman, "The impact of PMOS bias-temperature degradation on logic circuit reliability for performance," *Microelectronics Reliability*, vol. 45, pp. 107–114, 2005.
160. G. L. Rosa, "NBTI challenges in PMOSFETs of advanced CMOS technologies," 2003. IEEE International Reliability Physics Symposium Tutorial Notes.
161. R. Thewes, R. Brederlow, C. Schlunder, P. Wieczorek, B. Ankele, A. Hesener, L. Holz, S. Keseek, and W. Weber, "MOS transistor reliability under analog operation," *Microelectronics Reliability*, vol. 40, pp. 1545–1554, 2000.
162. Z. Chen, K. Hess, J. Lee, J. W. Lyding, E. Rosenbaum, I. Kizilyalli, S. Chetlur, and R.Huang, "On the mechanism for interface trap generation in MOS transistors due to channel hot carrier stressing," *IEEE Electron Device Letters*, vol. 21, pp. 24–26, 2000.
163. N. T. Do, T. Q. Vu, G. Warren, G. P. Li, and C. S. Tsai, "Negative bias instability in silicon-on-sapphire n-channel MOSFETs," in *Proceedings 1998 IEEE International SOI Conference*, pp. 85–86, 1998.
164. S. Zafar, B. H. Lee, J. Stathis, A. Callegari, and T. Ning, "A model for negative bias temperature instability (NBTI) in oxide and high k pFETs," in *2004 Symposium on VLSI Technology Digest of Technical Papers*, pp. 208–209, IEEE, 2004.
165. S. Zafar, "Statistical mechanics based model for negative bias temperature instability indced degradation," *Journal of Applied Physics*, vol. 97, pp. 1–9, 2005.
166. J. Srinivasan, S. V. Adve, P. Bose, and J. Rivers, "Exploiting structural duplication for lifetime reliability enhancement," in *Proceedings of the 32nd International Symposium on Computer Architecture*, 2005.
167. V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost, and S. Krishnan, "Impact of negative bias temperature instability on digital circuit reliability," in *International Reliability Physics Symposium*, pp. 248–254, IEEE, 2002.
168. H. Punchner and L. Hinh, "NBTI reliability analysis for a 90nm CMOS technology," in *Proceedings of the 34th European Solid-State Device Research Conference*, pp. 257–260, 2004.

169. A. T. Krishnan, V. Reddy, S. Chakravarthi, J. Rodriguez, S. John, and S. Krishnan, "NBTI impact on transistor and circuit: Models, mechanisms and scaling effects," in *International Electron Devices Meeting*, pp. 349–352, 2003.
170. M. Agostinelli, S. Lau, S. Pae, and P. Marzolf, "PMOS NBTI-induced circuit mismatch in advanced technologies," in *IEEE International Reliability Physics Symposium*, pp. 171–175, 2004.
171. C. Schlunder, R. Brederlow, B. Ankele, W. Gustin, K. Goser, and R. Thewes, "Effects of inhomogeneous negative bias temperature stress on p-channel MOSFETs of analog and RF circuits," *Microelectronics Reliability*, vol. 45, pp. 39–46, 2005.
172. C. Schlunder, R. Brederlow, B. Ankele, A. Lill, K. Goser, and R. Thewes, "On the degradation of PMOSFETs in analog and RF circuits under inhomogeneous negative bias temperature stress," in *IEEE Proceedings of International Reliability Physics Symposium*, pp. 5–10, 2003.
173. D. Lee, D. Blaauw, and D. Sylvester, "Gate leakage current analysis and reduction for VLSI circuits," *IEEE Transactions on VLSI systems*, vol. 12, pp. 155–166, 2004.
174. M. Fukuma, H. Furuta, and M. Takeda, "Memory LSI reliability," in *Proceedings of the IEEE*, vol. 81, pp. 768–775, 1993.
175. R. Rodriguez, R. V. Joshi, J. H. Stathis, and C. T. Chuang, "Oxide breakdown model and its impact on SRAM cell functionality," in *International Conference on Simulation of Semiconductor Processes and Devices*, pp. 283–286, 2003.
176. X. Li, J. D. Walter, and J. B. Bernstein, "Simulating and improving microelectronic device reliability by scaling voltage and temperature," in *International Symposium on Quality of Electronic Design*, pp. 496–502, March 2005.
177. B.-K. Liew and A. R. Alvarez, "Circuit reliability of hot electron induced degradation in high speed CMOS SRAM," in *Custom Integrated Circuits Conference*, pp. 30.2.1–30.2.4, IEEE, 1993.
178. D. Goguenheim, A. Bravaix, D. Vuillaume, M. Varrot, N. Revil, and P. Mortini, "Hot-carrier reliability in NMOSFETs used as pass-transistors," *Microelectronics and Reliability*, vol. 38, no. 4, pp. 539–544, 1998.
179. S. Z. Mohamedi, V. H. Chan, J. T. Park, F. Nouri, B. W. Scharf, and J. E. Chung, "Hot-electron-induced input offset voltage degradation in CMOS differential amplifiers," in *IEEE 30th International Annual Proceedings of Reliability Physics Symposium*, pp. 76–80, 1992.
180. J. E. Chung, K. N. Quader, C. G. Sodini, P.-K. Ko, and C. Hu, "The effects of hot-electron degradation on analog MOSFET performance," in *International Electron Device Meeting*, pp. 553–556, 1990.
181. R. Rodriguez, J. H. Stathis, B. P. Linder, S. Kowalczyk, C. T. Chuang, R. V. Joshi, G. Northrop, K. Bernstein, A. J. Bhavnagarwala, and S. Lombardo, "The impact of gate oxide breakdown on SRAM stability," *IEEE Electron Device Letters*, vol. 23, no. 9, p. 9, 2002.
182. S. Ikeda, Y. Yoshida, K. Ishibashi, and Y. Mitsui, "Failure analysis of 6T SRAM on low-voltage and high-frequency operation," *IEEE Transactions on Electron Devices*, vol. 50, no. 5, pp. 1270–1276, 2003.
183. J. Segura and A. Rubio, "A detailed analysis of CMOS SRAM's with gate oxide short defects," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 10, pp. 1543–1550, 1997.

184. C. Yao, J. Tzou, R. Cheung, and H. Chan, "Temperature dependence of CMOS device reliability," in *Annual Proceedings—Reliability Physics Symposium*, pp. 175–182, 1986.
185. B. Kaczer, R. Degraeve, E. Augendre, M. Jurczak, and G. Groeseneken, "Experimental verification of SRAM cell functionality after hard and soft gate oxide breakdowns," in *The 33rd Conference on European Solid-State Device Research*, pp. 75–78, 2003.
186. C.-K. Hu and R. Rosenberg, "Scaling effect on electromigration in on-chip Cu wiring," in *IEEE International Interconnect Technology Conference*, pp. 267–269, 1999.
187. C.-K. Hu, L. Gignac, and R. Rosenberg, "Electromigration of Cu/low dielectric constant interconnects," *Microelectronics Reliability*, vol. 46, pp. 213–231, 2006.
188. J. H. Stathis, "Reliability limits for the gate insulator in CMOS technology," *IBM Journal of Research and Development*, vol. 46, no. 2/3, pp. 265–286, 2002.
189. T. J. Anderson and J. M. Carulli Jr., "Modeling and monitoring of product DPPM with multiple fail modes," in *International Reliability Physics Symposium*, pp. 545–551, 2006.
190. Y.-H. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon, "Prediction of logic product failure due to thin-gate oxide breakdown," in *International Reliability Physics Symposium*, pp. 18–28, 2006.
191. W. Bornstein, R. Dunn, and T. Spielberg, "Field degradation of memory components due to hot carriers," in *International Reliability Physics Symposium*, pp. 294–298, 2006.
192. J. R. Black, "Mass transport of aluminum by momentum exchange with conducting electron," in *Proceedings of the Sixth Annual Reliability Physics Symposium*, pp. 148–159, 1967.
193. A. Dasgupta and R. Karri, "Electromigration reliability enhancement via bus activity distribution," in *33rd Design Automation Conference*, (Las Vegas, Nevada), 1996.
194. S. Mahapatra, P. B. Kumar, and M. A. Alam, "Investigation and modeling of interface and bulk trap generation during negative bias temperature instability of PMOSFETs," *IEEE Transactions on Electron Devices*, vol. 51, no. 9, pp. 1371–1379, 2004.
195. J.-C. Lin, S.-Y. Chen, H.-W. Chen, Z.-W. Jhou, H.-C. Lin, S. Chou, J. Ko, T.-F. Lei, and H.-S. Haung, "Investigation of dc hot-carrier degradation at elevated temperatures for n-channel metal-oxide-semiconductor field-effecttransistor of 0.13μm technology," *Japanese Journal of Applied Physics, Part 1: Regular Papers and Short Notes and Review Papers*, vol. 45, pp. 3144–3146, Apr 2006.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
25-02-2008	JPL Publication	N/A

4. TITLE AND SUBTITLE
Microelectronics Reliability: Physics-of-Failure Based Modeling and Lifetime Evaluation

5a. CONTRACT NUMBER
NAS7-03001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Mark White, Joseph B. Bernstein

5d. PROJECT NUMBER
102197

5e. TASK NUMBER
1.18.5

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91009

8. PERFORMING ORGANIZATION REPORT NUMBER
JPL Publication 08-5

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSORING/MONITOR'S ACRONYM(S)
NASA NEPP

11. SPONSORING/MONITORING REPORT NUMBER
0

12. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified—Unlimited

Subject Category Engineering-33

Availability: NASA CASI (301) 621-0390 Distribution: Nonstandard

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This handbook presents a physics-of-failure approach to microelectronics reliability modeling and assessment. Knowledge of the root cause and physical behavior of key failure mechanisms in microelectronic devices has improved dramatically over recent years and has led to the development of more sophisticated reliability modeling tools and techniques. Some of these tools are summarized here. Chapter 1 provides an overview of traditional reliability prediction approaches, i.e., MIL-HDBK-217 compared with some of the more recent reliability modeling and prediction approaches, including Reliability Aware MicroProcessor (RAMP) Model, Failure Rate Based SPICE (FaRBS) reliability simulation, and Maryland Circuit-Reliability Oriented (MaCRO) simulation. Chapter 2 describes the intrinsic wearout mechanisms of the electron device, including physics processes, mechanisms and models of electromigration (EM), hot carrier degradation (HCD), time-dependent dielectric breakdown (TDDB), and negative bias temperature instability (NBTI). In Chapter 3, the modules and processes of FaRBS reliability simulation, model parameter extraction, and derating of voltage and temperature for reliability are described. Sensitivity analysis and SPICE simulation of the wearout models are also discussed. To account for the effect of wearout mechanisms on circuit functionality and reliability, the device-level accelerated lifetime models are extended to microelectronic circuit-level applications and an analog-to-digital converter reliability simulation using the FaRBS application is provided. Lifetime and failure equivalent circuit models for Hot Carrier Injection (HCI), TDDB, and NBTI are presented in Chapter 4, Microelectronic Circuit Reliability Analysis and MaCRO. This chapter includes an illustrative case study for the purpose of demonstrating how to apply MaCRO models and algorithms to circuit reliability simulation, analysis, and improvement. The most common circuit structures used in reliability simulations are the ring oscillator, the differential amplifier, and the SRAM. The SRAM is selected as a case study vehicle to show the applicability of MaCRO models and algorithms in circuit reliability simulation and analysis. Chapter 5, in conclusion, describes the microelectronic system aspect of reliability, including impact to the system of individual failure

mechanism lifetime models, voltage and temperature acceleration, and qualification based on failure mechanism and application. A failure-mechanism-based qualification methodology using specifically designed stress conditions over traditional approaches (i.e., one voltage and one temperature) can lead to improved reliability predictions for targeted applications and optimized burn-in, screening, and qualification test plans.

15. SUBJECT TERMS
Microelectronics Reliability, Physics-of-Failure, Modeling, Lifetime Evaluation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			STI Help Desk at help@sti.nasa.gov
U	U	U	UU	214	19b. TELEPHONE NUMBER (Include area code) (301) 621-0390

JPL 2659 R 10 / 03 W

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18

NASA Supplementary Instructions To Complete SF 298 (Rev. 8-98 version)

NASA uses this inter-governmental form that does not allow customization. Look for special notes (NOTE) if NASA's procedures differ slightly from other agencies.

Block 1 NOTE: NASA uses month and year (February 2003) on the covers and title pages of its documents. However, this OMB form is coded for block 1 to accept data in the following format: day, month, and year (ex.: day (23), month (02), year (2003) or 23-02-2003, which means February 23, 2003. For this block, use the actual date of publication (on the cover and title page) and add 01 for the day. Example is March 2003 on the cover and title page, and 01-03-03 for block 1.

Block 2: Technical Paper, Technical Memorandum, etc.

Block 3: Optional for NASA

Block 4: Insert title and subtitle (if applicable)

Block 5a: Complete if have the information

b: Complete if have the information

c: Optional for NASA

d: Optional for NASA; if have a cooperative agreement number, insert it here

e: Optional for NASA

f: Required. Use funding number (WU, RTOP, or UPN)

Block 6: Complete (ex.: Smith, John J. and Brown, William R.)

Block 7: NASA Center (ex.: NASA Langley Research Center)
City, State, Zip code (ex.: Hampton, Virginia 23681-2199)
You can also enter contractor's or grantee's organization name here, below your NASA center, if they are the performing organization for your center

Block 8: Center tracking number (ex.: L-17689)

Block 9: National Aeronautics and Space Administration
Washington, DC 20546-0001

Block 10: NASA

Block 11: ex.: NASA/TM-2003-123456

Block 12: ex.:
Unclassified – Unlimited
Subject Category http://www.sti.nasa.gov/subjcat.pdf

	Availability: NASA CASI (301) 621-0390 Distribution: (Standard or Nonstandard) If restricted/limited, also put restriction/limitation on cover and title page
Block 13:	(ex.: Smith and Brown, Langley Research Center. An electronic version can be found at http:// _____ , etc.)
Block 14:	Self-explanatory
Block 15:	Use terms from the NASA Thesaurus http://www.sti.nasa.gov/thesfrm1.htm, Subject Division and Categories Fact Sheet http://www.sti.nasa.gov/subjcat.pdf, or Machine-Aided Indexing tool http://www.sti.nasa.gov/nasaonly/webmai/
Block 16a,b,c:	Complete all three
Block 17:	UU (unclassified/unlimited) or SAR (same as report)
Block 18:	Self-explanatory
Block 19a:	STI Help Desk at email: help@sti.nasa.gov
Block 19b:	STI Help Desk at: (301) 621-0390