Available online at www.sciencedirect.com
ScienceDirect
Microelectronics Reliability 46 (2006) 1957–1979
MICROELECTRONICS RELIABILITY
www.elsevier.com/locate/microrel

Introductory Invited Paper

Electronic circuit reliability modeling

Joseph B. Bernstein *, Moshe Gurfinkel, Xiaojun Li, Jörg Walters, Yoram Shapira, Michael Talmor

Reliability Engineering Center, Mechanical Engineering Department, University of Maryland at College Park, MD 20742, USA

Received 6 December 2005; received in revised form 12 December 2005
Available online 8 February 2006

Abstract

The intrinsic failure mechanisms and reliability models of state-of-the-art MOSFETs are reviewed. The simulation tools and failure equivalent circuits are described. The review includes historical background as well as a new approach for accurately predicting circuit reliability and failure rate from the system point of view.

1. Introduction

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 insure functional operation over the intended lifetime of the product. Thus, in order to determine the ultimate performance for a given set of design constraints, the 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, all the industrial approaches to reliability evaluation fall short of predicting failure rates or wearout lifetime of semiconductor products. This is attributed mainly to two reasons: 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, our knowledge about the root cause and physical behavior of the critical failure mechanisms in microelectronic devices has grown significantly. Confidence in the reliability models have led to more aggressive design rules that have been successfully applied to the latest VLSI technology. One result of improved reliability modeling has been accelerated performance, beyond the expectation of Moore's Law. A consequence of more aggressive design rules has been a reduction in the weight of a single failure mechanism. Hence in modern devices, there is no single failure mode that is more likely to occur than any other as guaranteed by the integration of modern failure physics modeling and advanced simulation tools in the design process.

The consequence of more advanced reliability modeling tools is a new phenomenon of device failures resulting from a combination of several competing failure mechanism. Hence, a new approach is required for reliability modeling and prediction. This paper reviews the existing modeling and prediction methods and presents an approach for accurate system reliability modeling in the competing mechanisms era. In addition, a new simulation tool that based on this new approach is presented.

1.1. Reliability prediction from historical perspective

In order to understand the current trends in reliability modeling and prediction, we have to look at it from historical point of view. Reliability modeling and prediction is a relatively new discipline. Only since World War II has reliability become a subject of study. This came about because of the relatively complex electronic equipment

* Corresponding author.
E-mail address: joey@umd.edu (J.B. Bernstein).

0026-2714/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.microrel.2005.12.004

used during the war and the rather high failure rates observed.

Since then, we can indicate two different approaches for reliability modeling corresponding to different time periods. Until the 1980s, the exponential, or the constant failure rate (CFR) model [1] had been the only model used for describing the useful life of electronic components. It was common to six reliability prediction procedures, which were reviewed by Bowles [2] and was the foundation of the military handbook for reliability prediction of electronic equipments, known as MIL-HDBK-217 [3] series. It 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 in which the failures are due to completely random or chance events. Indeed, throughout that period, electronic equipment complexity has begun to increase significantly. Similarly, the earlier devices were fragile and had several intrinsic failure mechanisms, which combined together 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 the field failures and they 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 suggested that the CFR model should not be used [4], based on their research to provide guidelines to update the MIL-HDBK-217. They suggested that the exponential distribution was unacceptable and should not be blindly applied to every type of component and system.

The end of the CFR as a sole model for reliability modeling was officially set with the publication of the "Perry Memo" [5]. Responding to increasing criticism of CFR, in 1994, Secretary of Defense William Perry issued a memorandum 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 DOD encouraged the use of industry standards, such as ISO 9000 series for quality assurance.

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

1.2. 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 would be seen in the field.

Chip and packaged system reliability is still measured by a Failure unIT (FIT). The FIT is a rate, defined as the number of expected device failures per billion part hours. A FIT is assigned for each component multiplied by the number of devices in a system for an approximation of the expected system reliability. The semiconductor industry provides an expected FIT 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 sum of the FIT rates for every component.

A FIT is defined in terms of an acceleration factor, A_F, as

$$\mathrm{FIT} = \frac{\#\text{failures}}{\#\text{tested} * \text{hours} * A_F} \times 10^9, \tag{1}$$

where #failures and #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, A_F, must be supplied by the manufacturers since only they know the failure mechanisms that are being accelerated in the high temperature operating life (HTOL) and it 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 A_F based on the physics of the dominant device failure mechanisms that would occur in the field.

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 h, at an accelerated voltage and temperature. Two features shed doubt on the accuracy of this procedure.

One feature is the lack of sufficient statistical data and the second is that companies generally present zero failures results of their qualification tests and hence stress their parts under relatively low stress levels to guarantee zero failures during qualification testing.

Unfortunately, with zero failures, no statistical data are acquired. Another feature is their calculation of the acceleration factor A_F. If the qualification test results in zero failures, which allows the assumption (with only 60% confidence!) that no more than 1/2 a failure occurred during the accelerated test. This would result, based on the example parameters, in a reported FIT $= 5000/AF$, which can be almost any value from less than 1 FIT to more than 500 FIT, depending on the conditions and model used for the voltage and temperature acceleration.

The accepted approach for measuring FIT would, in theory, be reasonably correct if there was only a single dominant failure mechanism that was excited equally by either voltage or temperature. For example, electromigration is known to follow Black's equation (described later) and is accelerated by increased current stress in a conductor or by increasing the device temperature. 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) should be a reasonable approximation for realistic field failures. Under the assumption of multiple failure mechanisms, each will be accelerated differently depending on the physics that is 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. In that instance, which is generally true for most devices, the reported FIT (especially one based on zero failures) will be meaningless with respect to other failure mechanisms.

1.3. Competing mechanism theory

1.3.1. Multiple failure mechanism model

Whereas the failure rate qualification has not improved over the years, the semiconductor industry understanding of reliability physics of semiconductor devices has advanced enormously. Every known failure mechanism is so 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 failure mechanisms during testing, which suggests also that no single failure mechanism dominates the FIT rate in the field. Therefore, in order to make a more accurate model for FIT, a preferable approximation is that all failures are *equally likely* and the resulting overall failure distribution resembles *constant failure rate process* that is consistent with the mil-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 and are the subject of most studies of reliability physics. The total acceleration factor of the different stress combinations are the product of the acceleration factors of temperature and voltage:

$$AF = \frac{\lambda(T_2, V_2)}{\lambda(T_1, V_1)} = AF_T \cdot AF_V = \exp\left(\frac{E_a}{k}\left(\frac{1}{T_1} - \frac{1}{T_2}\right)\right)\exp(\gamma_1(V_2 - V_1)). \quad (2)$$

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, 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, then the relative acceleration of each one must be defined and averaged under the applied condition. Every potential failure mechanism should be identified and its unique AF should then be calculated for each mechanism at given temperature and voltage so the FIT rate can be approximated for each mechanism separately. Then, the final FIT is the sum of the failure rates per mechanism, as described by

$$\mathrm{FIT}_{\mathrm{total}} = \mathrm{FIT}_1 + \mathrm{FIT}_2 + \cdots + \mathrm{FIT}_i, \quad (3)$$

where each mechanism leads to an expected failure unit per mechanism, FIT_i. Unfortunately, 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.

1.3.2. Acceleration factor

The qualification of device reliability, as reported by a FIT rate, must be based on an acceleration factor, which represents the failure model for the tested device. If we assume that there is no failure analysis (FA) of the devices after the HTOL test, or that the manufacturer does not report FA results to the customer, then a model should be made for the acceleration factor, AF, based on a combination of competing mechanisms. This will be explained by way of example. Suppose there are two identifiable, constant rate competing failure modes (assume an exponential distribution). One failure mode is accelerated only by temperature. We denote its failure rate as $\lambda_1(T)$. The other failure mode is only accelerated by voltage, and the corresponding failure rate is denoted as $\lambda_2(V)$. By perform-

ing the acceleration tests for temperature and voltage separately, we can get the failure rates of both failure modes at their corresponding stress conditions. Then we can calculate the acceleration factor of the mechanisms. If for the first failure mode we have $\lambda_1(T_1)$, $\lambda_1(T_2)$, and for the second failure mode, we have $\lambda_2(V_1)$, $\lambda_2(V_2)$, then the temperature acceleration factor is

$$AF_{\mathrm{T}} = \frac{\lambda_1(T_2)}{\lambda_1(T_1)}, \quad T_1 < T_2 \tag{4}$$

and the voltage acceleration factor is

$$AF_{\mathrm{V}} = \frac{\lambda_2(V_2)}{\lambda_2(V_1)}, \quad V_1 < V_2. \tag{5}$$

The system acceleration factor between the stress conditions of (T_1, V_1) and (T_2, V_2) is

$$AF = \frac{\lambda_1(T_2, V_2) + \lambda_2(T_2, V_2)}{\lambda_1(T_1, V_1) + \lambda_2(T_1, V_1)} = \frac{\lambda_1(T_2) + \lambda_2(V_2)}{\lambda_1(T_1) + \lambda_2(V_1)}. \tag{6}$$

The above equation can be transformed to the following two expressions:

$$AF = \frac{\lambda_1(T_2) + \lambda_2(V_2)}{\dfrac{\lambda_1(T_2)}{AF_{\mathrm{T}}} + \dfrac{\lambda_2(V_2)}{AF_{\mathrm{V}}}} \tag{7}$$

or

$$AF = \frac{\lambda_1(T_1)AF_{\mathrm{T}} + \lambda_2(V_1)AF_{\mathrm{V}}}{\lambda_1(T_1) + \lambda_2(V_1)}. \tag{8}$$

These two equations can be simplified based on different assumptions.

When $\lambda_1(T_1) = \lambda_2(V_1)$ (i.e. equal probability under normal operating conditions):

$$AF = \frac{AF_{\mathrm{T}} + AF_{\mathrm{V}}}{2}. \tag{9}$$

Therefore, unless the temperature and voltage is carefully chosen so that AF_{T} and AF_{V} are very close, within a factor of about 2, then one acceleration factor will overwhelm the failures at the accelerated conditions. Similarly, when $\lambda_1(T_2) = \lambda_2(V_2)$ (i.e. equal probability during accelerated test condition) then AF will take this form:

$$AF = \frac{2}{\dfrac{1}{AF_{\mathrm{T}}} + \dfrac{1}{AF_{\mathrm{V}}}} \tag{10}$$

and the acceleration factor applied to normal operating conditions will be dominated by the individual factor with the greatest acceleration. In either situation, the accelerated test does not accurately reflect the correct proportion of acceleration factors based on the understood physics of failure mechanisms.

This discussion can be generalized to incorporate situations with more than two failure modes. Suppose a device has n independent failure mechanisms, and $\lambda_{\mathrm{LT}_{\mathrm{FM}i}}$ represents the ith failure mode at accelerated condition, $\lambda_{\mathrm{use}_{\mathrm{FM}i}}$ represents the ith failure mode at normal condition, then A_{F} can be expressed. If the device is designed, such that the failure modes have equal frequency of occurrence during normal operating *conditions*:

$$AF = \frac{\lambda_{\mathrm{use}_{\mathrm{FM1}}} \cdot AF_1 + \lambda_{\mathrm{use}_{\mathrm{FM2}}} \cdot AF_2 + \cdots + \lambda_{\mathrm{use}_{\mathrm{FM}n}} \cdot AF_n}{\lambda_{\mathrm{use}_{\mathrm{FM1}}} + \lambda_{\mathrm{use}_{\mathrm{FM2}}} + \cdots + \lambda_{\mathrm{use}_{\mathrm{FM}n}}} = \frac{\sum_{i=1}^{n} AF_i}{n}. \tag{11}$$

If the device is designed, such that the failure modes have equal frequency of occurrence during the *test conditions*:

$$AF = \frac{\lambda_{\mathrm{LT}_{\mathrm{FM1}}} + \lambda_{\mathrm{LT}_{\mathrm{FM2}}} + \cdots + \lambda_{\mathrm{LT}_{\mathrm{FM}n}}}{\lambda_{\mathrm{LT}_{\mathrm{FM1}}} \cdot AF_1^{-1} + \lambda_{\mathrm{LT}_{\mathrm{FM2}}} \cdot AF_2^{-1} + \cdots + \lambda_{\mathrm{LT}_{\mathrm{FM}n}} \cdot AF_n^{-1}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{AF_i}}. \tag{12}$$

From these relations, it is clear that only if the acceleration factors for each mode are almost equal, i.e., $AF_1 \approx AF_2$, the total acceleration factor will be $AF = AF_1 = AF_2$, and certainly not the product of the two (as is currently the model used by industry). If, however, the acceleration of one failure mode is much greater than the second, the standard FIT calculation (Eq. (2)) could be incorrect by many orders of magnitude.

Due to the exponential nature of the acceleration factor as a function of V or T, if only a single parameter is changed, then it is not likely for more than one mechanism to be accelerated significantly compared to the others for any given V and T. In the next section, at least four mechanisms should be considered. Also, the various voltage and temperature dependencies must be considered in order to make a reasonable reliability model for electron devices. The assumption of equal failure probability under normal operating conditions is the most conservative and probably the most accurate. In fact, the exact proportions will not alter the result significantly since the proportional factor is only linearly related to the final acceleration factor, as compared to the exponential and power-law factors of the related physics models.

2. MOS failure mechanisms models

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

2.1. Electromigration

Electrons passing through a conductor transfer some of their momentum to its atoms. At sufficiently high electron current densities (greater than 10^5 A/cm^2 [6]), atoms may shift towards the anode side. The material depletion at the cathode side causes circuit damage due to decreased

electrical conductance and eventual formation of open circuit conditions. This is caused by voids and micro-cracks, which may increase the conductor resistance as the cross-sectional area is reduced. Increased resistance alone may result in device failure, yet, the resulting increase in local current density and temperature may lead to thermal runaway and catastrophic failure [7], such as an open circuit failure. Alternatively, short circuit conditions may develop due to excess material buildup at the anode. Hillocks form where there is excess material, breaking the oxide layer, allowing the conductor to come in contact with other device features. Other types of damage include whiskers, thinning, localized heating, and cracking of the passivation and inter-level dielectrics [8].

This diffusive process, known as electromigration, is still a major reliability concern despite vast scientific research as well as electrical and materials engineering efforts. Electromigration can occur in any metal when high current densities are present. In particular, the areas of greatest concern are the thin-film metallic interconnects between device features, contacts and vias [8].

2.1.1. EM physics

At high current densities, the force exerted by electrons scattering off the positively charged metal ions becomes stronger than the electrostatic pull force toward the cathode. Thus, the diffusion of the ions is biased in the direction of the electron flow, leading to electromigration. Its effects are expected to be characteristic of the material, such that the activation energy for electromigration is dependent on the material type, the size and orientation of the grains, stress, temperature and even the length of the conductor. Even low concentration doping may have great impact on the EM features. As an example, the EM activation energy of bulk Al is 1.4 eV, while adding small amounts (0.3–5%) of Cu reduces this activation energy by about 0.5–0.8 eV [8].

Grain size and pattern also have substantial impact on the effective EM activation energy of the metal. For instance, the activation energy ranges between 1 and 2 eV for thin films with large grain sizes. For very fine grained samples, the activation energy may be as low as 0.4–0.6 eV. Thus, mass transport-induced damage is more severe at grain boundaries and is greatest where three or more grains meet. When small dimension conductors are used, columnar growth of the metal lowers the grain boundary density and increases the electromigration lifetime.

Stress gradients also affect electromigration since they can induce atomic motion within the metal. Atoms migrate from regions of compressive stress to regions of tensile stress. When a conductor is shorter than a critical length, L_c, known as the 'Blech Length', the stress-induced flow of atoms counters the EM driving force and EM is eliminated [9].

Temperature gradients, caused by high current Joule heating also affect electromigration. While these gradients may only span a temperature change of some tens of degrees, the temperature change over a few microns results in large gradients [9]. Since EM is a thermally activated process, the temperature gradients produce flux divergences such as those found at contacts or other device features.

Increasingly, low resistivity Cu interconnects have been made use in ICs since Cu has a lower atomic diffusivity than Al. However, the surface self-diffusion in copper appears to be faster than grain-boundary self-diffusion. Thus, Cu does not provide the desired solution and the reliability of Cu interconnects may be improved by suppressing the interface and surface diffusion [10].

2.1.2. Lifetime prediction

Modeling electromigration median time to failure (MTTF) from the first principles of the failure mechanism is difficult. While there are many competing models attempting to predict time-to-failure from first principles, there is no universally accepted model.

Currently, the favored method to predict time to failure is an approximate statistical one given by the *Black's* equation, which describes the MTTF by

$$\mathrm{MTTF} = A(j_e)^{-n}\exp(E_a/kT), \tag{13}$$

where j_e is the current density and E_a is the EM activation energy. Failure times are described by the lognormal distribution [11]. The symbol A is a constant, which depends on a number of factors, including grain size, line structure and geometry, test conditions, current density, thermal history, etc. Black determined the value of n to equal 2. However, n is highly dependent on residual stress and current density [8] and its value is highly controversial.

A range of values for the EM activation energy, E_a, of aluminum (Al) and aluminum alloys is also reported. The typical value is $E_a = 0.6 \pm 0.1$ eV. The activation energy can vary due to mechanical stresses caused by thermal expansion. Introduction of 0.5% Cu in Al interconnects may result in $n = 2.63$ and an activation energy of $E_a = 0.95$ eV. For multi-level Damascene Cu interconnects, an activation energy of $E_a = 0.94 \pm 0.11$ eV at a 95% confidence interval (CI) and a value of the current density exponent of $n = 2.03 \pm 0.21$ (95% CI) were found [12].

2.1.3. Lifetime distribution model

Traditionally, the EM lifetime has been modeled by the lognormal distribution. Most test data appear to fit the lognormal distribution, but these data are typically for the failure time of a single conductor [13]. Through the testing of over 75,000 Al(Cu) connectors, Gall et al. [13] showed that the electromigration failure mechanism did follow the lognormal distribution. This is valid for the TTF of the first link with the assumption that the first link failure will result in device failure. The limitation is that a lognormal distribution is not scalable. A device with different numbers of links fails with a different lognormal distribu-

tion. Thus, a measured failure distribution is valid only for the device on which it is measured. Gall et al. also showed that the Weibull (and thus the exponential) distribution is not a valid model for electromigration.

Even though the lognormal distribution is the best fit for predicting the failure of an individual device due to EM, the exponential model is still applicable for modeling EM failure in a system of many devices where the reliability is determined by the first failure of the system.

2.1.4. Lifetime sensitivity

The sensitivity of the electromigration lifetime can be observed by plotting the lifetime as a function of the input parameters. For EM, the most significant input parameters corresponding to lifetime are the temperature (T) and current density (j_e). The lifetime may be normalized using an acceleration factor.

Substituting *Black's* equation and assuming an exponential failure distribution into

$$A_f = \lambda_{rated}/\lambda \tag{14}$$

provides the acceleration factor for EM

$$A_{f,EM} = (j_e/j_{e,rated})^{-n} \exp[(E_{a,EM}/kT)(1/T - 1/T_{rated})]. \tag{15}$$

Obviously, T has a much greater impact on A_f than j_e.

As device features continue to shrink and interconnect current densities grow, EM will remain a concern. New technologies may reduce the EM impact of increasing densities but new performance requirements emerge that require increased interconnect reliability under conditions of decreased metallization inherent reliability [9]. Thus, EM will remain a design and wearout issue in future semiconductor designs.

2.2. Time dependent dielectric breakdown

Time dependent dielectric breakdown (TDDB), also known as oxide breakdown, is a source of significant reliability concern. When a sufficiently high electric field is applied across the dielectric gate of a transistor, continued degradation of the material results in the formation of conductive paths, which may short the anode and cathode [14]. This process will be accelerated as the thickness of the gate oxide decreases with continued device down-scaling.

The TDDB process takes place in two stages [15]. In the first stage, the oxide is damaged by the localized hole and bulk electron trapping within it and at its interfaces. The second stage is reached when the increasing density of traps within the oxide form a percolation (conduction) path through the oxide. This short circuit between the substrate and gate electrode results in oxide failure. This process has been successfully modeled using Monte Carlo simulations.

The formation of a percolation path may result in one of two types of failure. Once a conduction path forms, current flows through the path causing a sudden energy burst, which may cause runaway thermal heating. The result may be a *soft breakdown* if the device continues to function. Local melting of the oxide will destroy the gate and is thus denoted as *hard breakdown*. It has been speculated that soft breakdown does not even significantly affect transistor operation, although it may still lead to the failure of short channel devices. While the change in both threshold voltage and leakage from soft breakdown is small and initially does not affect device operation, the effects are cumulative. Multiple soft breakdowns may result in an increase in leakage current to unacceptable levels [16].

2.2.1. TDDB physics

Trap generation is the key factor determining oxide degradation and breakdown. Three general models are discussed in the literature for trap generation. These models are the "anode hole injection" (AHI) model, the "thermo-chemical" model, and the "anode hydrogen release" (AHR) model.

The AHI model ($1/E$ model) was proposed by Schuegraf and Hu [17] and studied by many researchers. In this model, electrons injected from the gate metal cathode into the oxide undergo impact ionization events, which generate holes in the process. Some of these holes tunnel back into the cathode and create electron traps in the oxide. The physics of the trap creation process is still speculative.

The thermo-chemical model (E model) is another widely cited dielectric breakdown model. McPherson and Mogul [18] 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 a secondary role at most. The interaction of the applied electric field with the dipole moments associated with oxygen vacancies leads to a conduction sub-band formation and to severe Joule heating at the stage of oxide breakdown.

In the AHR model, the energy release of the incoming electrons at the anode may activate hydrogen release at the anode, besides creating holes. The released hydrogen diffuses through the oxide and can generate electron traps. There have been contradicting opinions on the exact field acceleration law of time-to-breakdown – t_{BD}. According to the AHI model ($1/E$ model) the field dependence of the t_{BD} takes the form:

$$t_{BD}(t) = \tau_0 \exp\left(\frac{G}{E_{OX}}\right), \tag{16}$$

where E_{OX} is the electric field across the dielectric and τ_0 and G are constants.

According to the thermo-chemical model (E model) the field dependence of the t_{BD} is of the form:

$$t_{BD}(t) = t_0 \exp(-\gamma E_{OX}), \tag{17}$$

where t_0 and γ are constants.

There is still no consensus on the correct acceleration law and the discussion of the E and $1/E$ models continues.

The debate about E vs. $1/E$ models is most applicable for thick oxides. For ultra-thin oxides evidence shows that gate voltage is the primary driver of the breakdown process [19]. Additionally, there is evidence that the temperature dependence of ultra-thin oxides is non-Arrhenius, but rather the temperature acceleration factor is larger at higher temperatures. To account for these observations, Wu et al. [19] have proposed a relationship in the form of

$$\mathrm{MTTF} = T_{\mathrm{BD0}}(V)\exp(a(V)/T + b(V)/T^2), \tag{18}$$

where $T_{\mathrm{BD0}}(V)$, a and b are voltage dependent factors. The second order term, b/T^2, is included in order to account for any non-Arrhenius temperature effects.

2.2.2. Lifetime distribution model

The time-to-breakdown (t_{BD}) is a statistically distributed parameter. At high fields, a very wide distribution of breakdown times is found. It is commonly assumed to be distributed according to the Weibull statistics, which is typical for 'weakest link processes'. The cumulative distribution function can then be described as

$$F(t) = 1 - \exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right], \tag{19}$$

where β is the shape factor of the distribution, often called the Weibull slope, and η is a scale factor. Eq. (19), can be rearranged such that

$$\ln[-\ln(1 - F(t))] = \beta\ln(t) - \beta\ln(\eta), \tag{20}$$

which implies that a plot of $\ln[-\ln(1 - F)]$ as a function of the logarithm of t yields a straight line with a slope of β.

Lognormal 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 numbers of TDDB failures [20]. An important disadvantage of lognormal distribution is that it does not predict the observed area dependence of t_{BD} for ultra-thin gate oxides.

2.2.3. The breakdown event

The breakdown event itself is usually described using a "weakest link" model. 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.

The first 'weakest link' model was formulated by Sune et al. [21] in the early 1990s 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. Once the number of traps in a cell reaches a critical value, breakdown will occur.

The disadvantage of Sune's model is its two-dimensional nature. A new three-dimensional model, based on the percolation concept has been suggested in Ref. [22] and has been thoroughly elaborated in Refs. [23,24]. The model assumes 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 the model. If the spheres of two neighboring traps overlap, conduction between these traps becomes possible by definition.

This mechanism of trap generation continues until a conducting path is created from one interface to the other, which defines the breakdown condition. The percolation model for oxide breakdown is able to quantitatively explain two important experimental observations: (i) as the oxide thickness decreases, the density of oxide traps needed to trigger breakdown decreases [23–25], and (ii) as the oxide thickness decreases, the Weibull slope of the breakdown distribution decreases and approaches unity, i.e. a larger spread of the t_{BD}-values is observed [23,24,26,27].

Oxide thickness will continue to be scaled in future devices because of the need to improve and optimize circuit performance. The effect of TDDB in the case of ultrathin oxides is still arguable. The contrasting viewpoints indicate the need for better understanding of TDDB as device features shrink.

2.3. Hot carrier injection

Hot carriers in the semiconductor device are the cause of a distinct wearout mechanism, the hot carrier injection (HCI). Hot carriers are produced when the source–drain current flowing through the channel attains high energy beyond the lattice temperature. Some of these hot carriers gain sufficient energy to be injected into the gate oxide, resulting in charge trap and interface state generation. The latter may lead to shifts in the performance characteristics of the device, e.g., the threshold voltage, transconductance, or saturation current, and eventually to its degradation.

The rate of hot carrier injection is directly related to the channel length, oxide thickness and the operating voltage of the device. Since the latter are minimized for optimal performance, the scaling has not kept pace with the reduction in channel length. Current densities have been increased with a corresponding increase in device susceptibility to hot carrier effects.

2.3.1. HCI physics

Hot carriers are generated during the operation of semiconductor devices as it switch states. As carriers travel through the channel from source to drain, the lateral electric field near the drain junction causes carriers to become hot [28]. A small part of these hot carriers gain sufficient energy—higher than the Si–SiO_2 energy barrier of about 3.7 eV—to be injected into the gate oxide. In nMOS (negative-channel metal-oxide semiconductor) devices, hot electrons are generated while hot holes are produced in pMOS

(positive-channel metal-oxide semiconductor) devices. Injection of either carrier results in three primary types of damage: trapping of electrons or holes in pre-existing traps, generation of new traps, and the generation of interface traps [29]. These traps may be classified by location [30] while their effects vary.

Interface traps are located at or near the $Si–SiO_2$ interface and directly affect transconductance, leakage current and noise level. Oxide traps are located further away from the interface and affect the long term MOSFET stability, specifically the threshold voltage. Effects of defect generation include threshold voltage shifts, transconductance degradation and drain current reduction [28]. NBTI seems to have similar degradation patterns, except for pMOS, so both will be treated similarly here.

Hu et al. [31] proposed the 'lucky' electron model for hot carrier effects. This is a probabilistic model proposing that a carrier must first gain enough kinetic energy to become 'hot', and then the carrier momentum must become redirected perpendicularly so the carrier can enter the oxide. The current across the gate is denoted by i_{gate} and during normal operation its value is negligible. Degradation due to hot carriers is proportional to i_{gate}, making the latter a good monitor of the former. If the rate of change of the HCI-induces damage, designated by Δ, is proportional to i_{gate}, then

$$d\Delta/dt \sim i_{gate} = A(\Delta)/W \cdot i_{drain} \cdot (i_{sub}/i_{drain})^m, \tag{21}$$

where W is the width of the MOSFET. By letting $B = A(\Delta)/W$, and knowing that MTTF depends on the reciprocal of $d\Delta/dt$, the failure rate is found from

$$\lambda = B \cdot i_{drain} \cdot (i_{sub}/i_{drain})^m. \tag{22}$$

This equation assumes static (dc) voltages and currents. To account for dynamic degradation λ has to be integrated over a full cycle time.

Temperature plays an interesting, though small role in hot carrier injection, since the activation energy is negative, implying that HCI diminishes with increasing temperature. At low temperatures, the substrate current increases because the drain current increases. According to Acovic et al. [32], the effects of oxide degradation are stronger at low temperatures because the electrons, having lower thermal energy, are more confined within the negatively charged degraded zone. Another possibility is that freeze-out of impurities in the drain at low temperatures makes n-MOSFETs more sensitive to electrons trapped in the drain region, increasing degradation. Degradation decreases at high temperatures because the drain current and the mean free path decrease.

2.3.2. Lifetime prediction

The lucky electron model does not fully predict hot carrier injection lifetime. Since there is no direct method of measuring device lifetime, the Arrhenius relationship remains a favored lifetime prediction tool. The following models are from JEP-122A [14]. It contains two models.

The *N-Channel* model is for nMOS devices. In these devices the substrate current is an indicator of hot carriers. The MTTF equation is

$$\mathrm{MTTF} = B(i_{sub}) - N\exp(E_a/kT), \tag{23}$$

where B is a scale factor, which is a function of doping profiles, sidewall spacing, dimensions, etc., i_{sub} is the substrate current, N ranges from 2 to 4, and E_a is the activation energy in the range of -0.1 eV to -0.2 eV.

In pMOS devices, hot holes do not show up as substrate current. However, the gate current can serve as an indicator of hot carriers. Thus the *P-Channel* model is

$$\mathrm{MTTF} = B(i_{gate})^{-M}\exp(E_a/kT), \tag{24}$$

where B and E_a are the same as before while i_{gate} is the peak gate current during stressing and M ranges from 2 to 4. However, the Arrhenius term is not necessarily appropriate for these mechanisms.

2.3.3. Lifetime distribution model

There is little discussion in literature about a proper statistical lifetime distribution model for hot carrier injection. A logical hypothesis for the lifetime distribution would be the exponential one. This is a good assumption because as a device becomes more complex, with millions of gates, it may be considered as a system.

The failure probability of each individual gate is not most likely an exponential distribution. However, the cumulative effect of early failures and process variability, ensuring each gate has a different failure rate, widens the spread of the device failures. The end result is that intrinsic hot carrier injection becomes statistically more random as the failures occur at a constant rate.

2.3.4. Lifetime sensitivity

As for EM, hot carrier injection lifetime is sensitive to changes in the input parameters. The acceleration factor for hot carrier injection is

$$A_{F,HCD} = \exp(B(1/V_{dd} - 1/V_{dd,max})). \tag{25}$$

HCI continues to be a reliability concern as device feature sizes shrink. HCI is a function of internal electric fields in the device and as such is affected by channel length, oxide thickness and device operating voltage. Shorter channel lengths decrease reliability but the oxide thickness and the voltage may also be reduced to help alleviate the reduction in reliability. Another way of improving hot carrier reliability may be by shifting the position of the maximum drain so it is deeper in the channel [32]. This would result in hot carriers being generated further away from the gate and $Si–SiO_2$ interface, reducing the likelihood of injection into the gate. Another method is to reduce the substrate current by using a lightly doped drain (LDD) where part of the voltage drop is across a lightly doped drain extension not covered by the gate. Annealing the oxides in NH_3, N_2O or NO or growing them directly in N_2O or NO improves

their resistance to interface state generation by the hot carriers.

2.4. Negative bias temperature instability

NBTI differs from hot carrier injection in that NBTI causes a shift in the device threshold voltage. The mechanism for NBTI damage are holes trapped within the interface between the SiO_2 gate insulator and the Si substrate. NBTI damage is most prevalent in p-MOSFET devices where holes are thermally activated and gain sufficient energy to disassociate the interface/oxide defects near the lightly doped drain (LDD) regions. This happens at the LDD regions because of the higher hole concentrations near the gate edge.

NBTI mainly occurs in p-channel MOS devices stressed with negative gate voltages at elevated temperatures. It appears to be negligible for positive gate voltage and for either positive or negative gate voltages in n-MOSFETs. NBTI manifests itself as decrease in absolute drain current I_{Dsat} and transconductance g_m while the absolute "off" current I_{off} and threshold voltage V_{th} increase.

The typical stress conditions of NBTI are temperatures of 100–250 °C and oxide electric fields below 6 MV/cm. These stress conditions are typical during burn-in. The threshold voltage V_{th} and flat-band voltage V_{FB} of a MOSFET are given by

$$V_{th} = V_{FB} - 2\phi_F - \frac{|Q_B|}{C_{ox}}, \tag{26}$$

$$V_{FB} = \phi_{MS} - \frac{Q_f}{C_{ox}} - \frac{Q_{it}(\phi_s)}{C_{ox}}. \tag{27}$$

The fixed oxide charge Q_f and the interface-trapped charge density Q_{it} are the two factors determining the threshold voltage shift. Positive increases of these two parameters lead to a negative threshold voltage shift:

$$\Delta V_{th} = -\frac{\Delta Q_{it}(\phi_s)}{C_{ox}} - \frac{\Delta Q_f}{C_{ox}}. \tag{28}$$

During the NBTI degradation, the threshold voltage shifts to more negative direction, affecting either the interface traps or the fixed oxide charges.

The simplest form of the on-state driving current I_{Dsat} and transconductance g_m of a MOSFET is given by

$$I_{Dsat} = \frac{W}{2L}\mu_{eff}C_{ox}(V_{gs} - V_{th})^2, \tag{29}$$

$$g_m = \frac{W}{L}\mu_{eff}C_{ox}(V_{gs} - V_{th}). \tag{30}$$

These equations show that the parameters leading to I_{Dsat} and g_m degradation are the threshold voltage and the mobility μ_{eff}. The mobility degradation is mostly induced by interface trap generation, leading to additional surface-related scattering.

2.4.1. NBTI physics

Since first observed by Deal et al. [33] in 1967, NBTI has been intensively investigated and many models for its physical mechanisms have been proposed. The three most prominent models in the literature feature holes injection into the oxide, electron tunneling and electrochemical reactions.

The hole-trapping model is based on avalanche hole injection measurements of unstressed MOS capacitors and NBTI tests [34–36]. This model proposes that the negative midgap voltage shift (which is believed to be a monitor of the positive oxide charge with no contribution from interface states), is due to population of intrinsic hole traps. All the positive charge generated by preceding negative bias stresses can be removed by the positive bias stress. However, the exact mechanism for hole injection into the oxide is still unknown.

The thermally assisted electron tunneling model was established by Breed [37,38]. According to this model, the neutral or positive centers, which are the charge traps, are located near the oxide interface. Under negative bias stress, the centers are excited. The electrons tunnel from the excited states into empty states of the Si conduction band. This is a thermally assisted tunneling process.

Several authors proposed the electrochemical reaction model or the reaction-diffusion model, which has recently been accepted by many researchers [39–42]. This model explains the NBTI effect in terms of electrochemical reactions.

NBTI became evident with the advent of 0.13 μm processes as devices required much thinner gate oxides and introduced nitrides in the SiO_2 to prevent boron penetration into the gate. Another source of concern is plasma-induced damage during interconnect deposition resulting in driving hydrogen atoms into the Si–SiO_2 interface.

2.4.2. Lifetime prediction

Generally, a threshold voltage shift ΔV_{th} caused by NBTI can be expressed as

$$\Delta V_{th} = A f_1(t) f_2(V_g) \exp\left(\frac{-E_a}{kT}\right). \tag{31}$$

Here f_1 and f_2 are functions account for the time dependence and gate voltage dependence. Based on the physical mechanisms and experimental data, several models for the time dependence have been suggested.

2.4.2.1. Logarithmic time dependence

$$\Delta V_{th} = A\log(t). \tag{32}$$

This model was established on the ideal of charge trapping, wherein carriers tunnel into existing traps [33]. According to this model, the NBTI is field accelerated, and there is little or no temperature activation. The saturation behavior is due to the finite trap density. There is significant deviation at long time when using this model. However, it is frequently observed in recent high-k gate dielectrics experiments.

2.4.2.2. Exponential time dependence

$$\Delta V_{\text{th}} = A \exp\left(\frac{t}{\tau}\right) \tag{33}$$

or

$$\Delta V_{\text{th}} = A \exp\left(\frac{t}{\tau_1}\right) + B \exp\left(\frac{t}{\tau_2}\right). \tag{34}$$

A single exponential time dependence model was established at first based on the first order reaction, which was limited by the hole concentration [39]. Different from the logarithmic time dependence, this model suggested the temperature activation from the reactions. Later, a two-exponential model was further suggested [43,44].

2.4.2.3. Power law time dependence.

$$\Delta V_{\text{th}} = At^n. \tag{35}$$

The power law model was based on the reaction diffusion mechanism. According to this model, the hydrogen profile determines the time dependency [40,41]. The temperature dependence arises from the reaction and diffusion processes and the saturation behavior comes from the diffusion barrier or the available Si–H bonds. Compared with the other two models, this model has the most observed features and is widely accepted.

3. Review of existing reliability simulation tools

IC reliability simulation is not a new concept and a number of reliability models and simulation methodologies such as BERT [45] and ARET [46] have been developed during the past decade. Most state-of-the-art reliability simulation methods try to emulate the degradation process of aged devices in a repetitive scheme. They are based on the physical failure mechanisms and contain the major wearout models for EM, HCI, NBTI and TDDB. A set of parameters for each of these failure mechanisms are identified and the algorithms of extracting these parameters for a given technology are developed by accelerated tests on test structures. A circuit simulator, such as SPICE, is employed to calculate the electrical parameters of fresh and degraded devices to predict their degradation or failure from these parameters.

This reliability simulation method can help designers understand how the devices degrade over time, identify the reliability bottlenecks within the circuits and make design tradeoffs between performance and reliability in the product design stage. It can also help manufacturers build their circuits such that no known wearout mechanism will dominate over the life of an operating device and assure adequate reliability for the product. In what follows, two commercial state-of-the-art reliability simulation methods are reviewed, then a set of failure equivalent circuits for the most important intrinsic silicon wearout mechanisms including HCI, TDDB and NBTI are reviewed. Finally, in the next section, a new failure rate-based SPICE reliability simulation methodology is introduced to investigate product reliability in different ways.

3.1. Degradation-based reliability simulation tools

Hot carrier reliability simulation models and methods have been implemented and widely used in the semiconductor industry for many years. To some extent, the accuracy of hot carrier reliability simulation represents the robustness and efficiency of the entire reliability simulator, therefore, for the purpose of simplicity, HCI simulation is employed as the vehicle to deliver the basic concepts and flows realized in some commercial degradation-based reliability simulation methods.

3.1.1. Hot carrier reliability simulation in Virtuoso UltraSim

Virtuoso UltraSim is the Cadence FastSPICE circuit simulator capable of predicting and validating timing, power and reliability of mixed-signal, complex digital and System-on-Chip (SoC) designs in advanced technology of 0.13 μm and below. It has a set of specialized reliability models (AgeMos) for HCD and NBTI simulation [47]. In the simulation, an Age parameter is calculated for each nMOS device with the following formula:

$$\text{Age}(\tau) = \int_{t=0}^{t=\tau} \left(\frac{I_{\text{sub}}}{I_{\text{ds}}}\right)^m \frac{I_{\text{ds}}}{WH}\, \mathrm{d}t, \tag{36}$$

where W refers to the width of the transistor; m and H are technology dependent parameters and determined from experiments; I_{sub} is the substrate current; I_{ds} is the drain current; τ is the time for stress. For pMOS devices, the gate current I_{gate} is used instead of I_{sub} to determine the Age parameter. The degree of MOS device degradation has been experimentally found to be a function of this *Age* parameter for wide ranges of channel length and stress conditions and the relationship has a plausible theoretical basis [48].

The simulation starts with device parameter extraction and modeling. From the SPICE model parameters of fresh devices, some other device parameters are added to accurately model I_{sub}. Saturation current I_{dsat}, threshold voltage, V_{th}, or the maximum transconductance, g_{max}, can be used as a degradation monitoring parameter. I_{dsat} is a good degradation monitor for digital circuits, while V_{th} is suitable for analog applications. Normally, the stress time resulting in 10% decrease of one of these degradation monitoring parameters is arbitrarily set as the device lifetime. The final step is *AgeMos* extraction. Based on the *Age* parameter calculated after the fresh simulation, the *AgeMos* applies the degradation models, which can be input to most SPICE-like simulators, for the aged circuit simulation. Reliability simulation with Virtuoso UltraSim is an iterative process, in which several iterations are often needed in order to get accurate modeling. The simulation can calculate and output the degradation results to predict the lifetime of each MOS instance [49]. The overall simulation flow is depicted in Fig. 1.

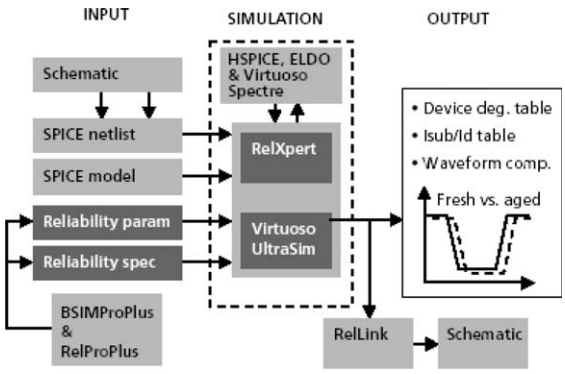

Fig. 1. Hot carrier reliability simulation flowchart in Virtuoso UltraSim [45].

The fundamental models and methodology of reliability simulation was first proposed and implemented in BERT (Berkeley Reliability Tools), introduced by Chenming Hu in 1992 [48]. This approach to reliability equivalent circuits was commercially realized in Virtuoso UltraSim and similarly in ELDO, the simulator developed by Mentor Graphics. The main advantages of the BERT simulation methodologies are accuracy and SPICE modeling technology compatibility. However, it imposes a burden on designers to correctly extract the device fresh and degraded parameters and leads to non-physical trends, which prevents its popularity in reliability design process. These tools are very important for the IC designers, but once a circuit is produced, no flexibility remains to alter the reliability if a new application or set of operating parameters is applied.

3.1.2. Hot carrier reliability simulation in Eldo

Eldo and UltraSim both deliver all the capability and accuracy of SPICE-level simulation for complex analog circuits and SoC designs. However there are some subtle differences. In Eldo, the substrate current I_{sub} is not selected as the primary reliability parameter as in UltraSim. In general, the drain current I_d, threshold voltage V_t or transconductance g_m is often used as a degradation monitoring parameter, and the stress time resulting in 10% decrease of one of these monitoring parameters is arbitrarily set to the device lifetime. Hot carrier reliability simulation in Eldo adopts I_d as the degradation monitoring parameter and characterizes it with a compact ΔI_d model, which directly models the difference of drain currents between fresh and aged devices.

There exist two competing mechanisms, which lead to the obvious hot carrier induced drain current variations between fresh and degraded devices: the deviation of I_d from its linear dependence on V_{ds} due to velocity saturation effects and the decreasing of $\Delta I_d/I_d$ due to the reduction of charged interface states [51]. In Eldo, the ΔI_d is modeled with Eq. (41) to (40), which unify the subthreshold, linear and saturation regions with a simple relation for both forward and reverse operation modes [50]:

$$\frac{\Delta I_d}{I_d} = \left(\frac{B_6(1-e^{-B_1 V_{gs}})+B_2}{1+B_5(V_{gs}-B_3 V_{th})}\right)\left(\frac{N_{it}L_{it}}{L_{eff}}\right) \times \left(\frac{1}{1+\alpha(V_{ds}-V_{low})+\beta V_{ds}}\right), \quad (37)$$

$$V_{low} = A_3 V_{dsat}, \quad (38)$$

$$\alpha = \frac{A_1}{1+A_4(V_{gs}-V_{th})^{A_2}}, \quad (39)$$

$$\beta = A_5 V_{gs} + A_6, \quad (40)$$

where N_{it} is the interface trap density, L_{it} is the extension of the damage within the channel, L_{eff} is the effective channel length, V_{gs} is the gate-to-source voltage, V_t is the threshold voltage, V_{ds} is the drain-to-source voltage, V_{dsat} is the drain saturation voltage, A_1 to A_6 and B_1 to B_6 are model fitting parameters.

The same *Age* parameter defined by Eq. (36) is incorporated to model the "age" of each transistor. The HCI aging process is simulated in an iterative way as depicted schematically in Fig. 2.

The period, T_{age}, at which the circuit performance is to be tested is divided into smaller time intervals, T_1. The *Age* table is calculated at the end of each time interval and a new simulation with Eldo is carried forward. This process is repeated until T_{age} is reached. This iterative scheme can account for the gradual change of bias conditions as a result of device wearout.

The ΔI_d modeling approach provides the possibility to have a relatively simpler parameter extraction process. It is suitable to model bi-directional stress and asymmetrical drain current behavior. However, because this approach also adopts both *Age* parameter and small-step iterative algorithm in the degradation simulation process, it inherits the same limitations of the BERT-like tools as discussed before.

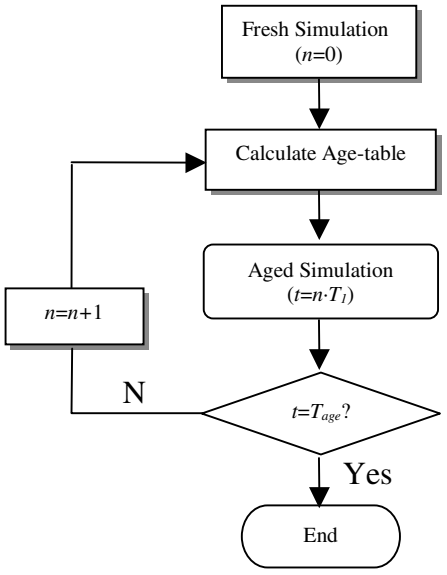

Fig. 2. HCI reliability simulation in Eldo [50]. A large number of SPICE simulation iterations have to be carried out to obtain accuracy.

3.2. Failure mechanism equivalent circuits

In order to account for the effect of the failure mechanisms on circuit functionality and reliability, the device-level lifetime models have to be extended to circuit-level applications. The bridge connecting the gap between device wearout degree and circuit performance drift is no doubt the circuit models. The underlying concept of the 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 circuit operation environment. The values of these additional lumped elements are determined by device wearout parameters (such as ΔN_{it} which 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 magnitudes reflect the device wearout degree. The larger the magnitude of these values, the more severe is the damage to circuit functionality. As a result, circuit designers can quickly analyze circuit reliability behavior at any given time with these circuit models.

3.2.1. HCI

Several HCI circuit models have been developed in the past years and some of them have been built into commercial reliability simulation tools. In this section, some of these circuit models are briefly reviewed, followed by the introduction of the HCI circuit model and its implementation.

BERT has been the most successful circuit reliability simulation tool. BERT directly models n-MOSFET 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 famous *Age* parameter introduced in the previous section. In BERT, ΔI_d is implemented as an asymmetrical voltage controlled current source in parallel with the original n-MOSFET. The p-MOSFET HCI effect is modeled with the concept of channel shortening and drain resistance increase [45]. The BERT ΔI_d model is shown in Fig. 3. Here one can see asymmetry in the forward and reverse I–V characteristics, allowing the 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 [52]. The main contribution of BERT ΔI_d model is the ability to characterize bi-directional hot carrier stress effects, however it requires extraction of six process parameters from device testing, which is a non-trivial work.

Experiments have proven that HCI-induced interface traps in n-MOSFET are localized above the channel near the drain junction. More specifically, these interface traps are localized within 100 nm from the drain [53]. Based on this observation, Leblebici et al. at UIUC [54,55] developed a two-transistor HCI circuit model, which consisted of an HCI damaged parasitic transistor with fixed channel length

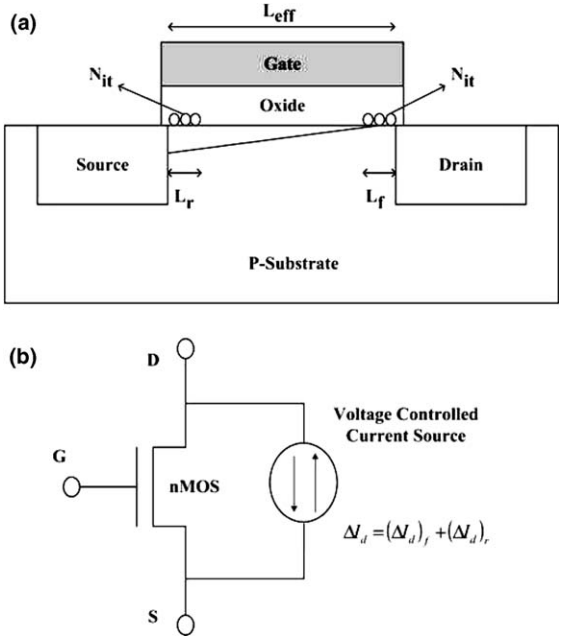

Fig. 3. BERT n-MOSFET HCI circuit model. (a) Bidirectional interface trap generation near both drain and source. L_f and L_r represent forward and reverse hot carrier damaged regions. (b) HCI drain current ΔI_d circuit model [51].

L_2 ($L_2 \approx 0.1$ μm) in series connection with the original transistor whose channel length was shrunk to $L - L_2$. The primary assumption for this model is that all generated interface traps are occupied with electrons, which equals to considering only negative fixed charge. The model is illustrated in Fig. 4.

From Fig. 4a, the interface trapped charge Q_{it} due to HCI can be readily derived as

when $(0 \leqslant x < L_1)$:

$$Q_{it}(x) = 0 \tag{41}$$

or when $(L_1 \leqslant x < L)$:

$$Q_{it}(x) = \frac{Q_M}{L_2}(x - L_1), \tag{42}$$

where Q_M denotes the largest interface charge, $L_1 = L - L_2$, 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: that the triangular charge density distribution is over simplified, and that an accurate Q_M value is difficult to extrapolate.

The simplest HCI circuit model has been the hot carrier induced series resistance enhancement model (HISREM), also named ΔR_d model, which is proposed by Hwang et al. at Oregon State University [58]. 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 n-MOSFET can

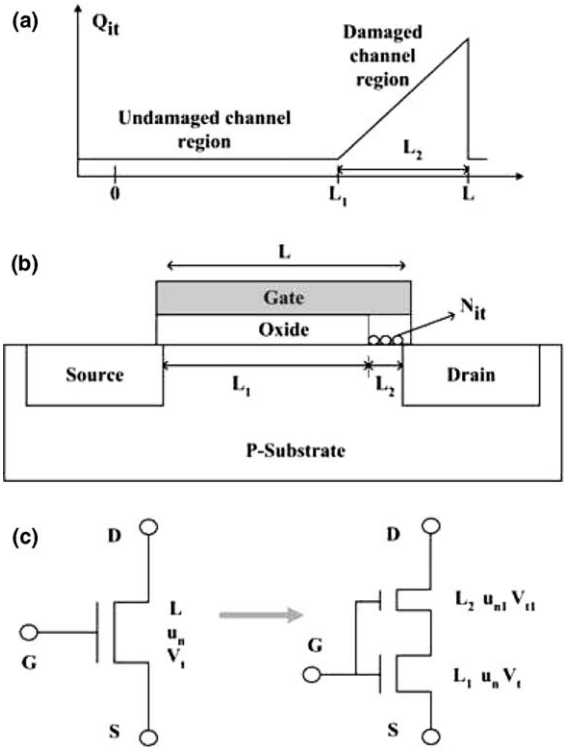

Fig. 4. UIUC n-MOSFET HCI two-transistor series model. (a) Triangular oxide charge distribution profile used in model derivation. (b) Cross-sectional view of n-MOSFET with hot carrier damage, L_2 is the damaged channel region. (c) Two-transistor series circuit model. The parasitic transistor has different channel mobility and threshold voltage with the channel length L_2 set to 0.1 μm [54,56,57].

reflect the process of hot carrier induced interface trap generation and therefore accounts 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 n-MOSFET. ΔR_d is a function of the applied voltages and the hot carrier induced interface trapped charge ΔN_{it}.

The behavior of the damaged n-MOSFET is emulated by the original undamaged device operated with a reduced drain-to-source voltage, which is controlled by this additional drain resistor ΔR_d. Because ΔN_{it} is a time dependent parameter, ΔR_d model is able to predict drain current degradation at any given time. HISREM is also capable of modeling self-limiting effects of hot carrier damage because the increase in series drain resistance of an n-MOSFET suppresses hot carrier stress. The most advantageous feature of the HISREM model is that only one parameter, ΔN_{it}, needs to be extrapolated from device testing work. Consequently, HISREM can be easily used by circuit designers to perform an expeditious reliability analysis.

The HCI circuit model is based on the above ΔR_d model with some improvements. The major improvement is that ΔR_d value is considered to be determined by both interface trapped charge ΔN_{it} and oxide trapped charge ΔN_{ox}. The contribution of ΔN_{ox} to device wearout is often neglected, but recent experimental work recognizes that they can

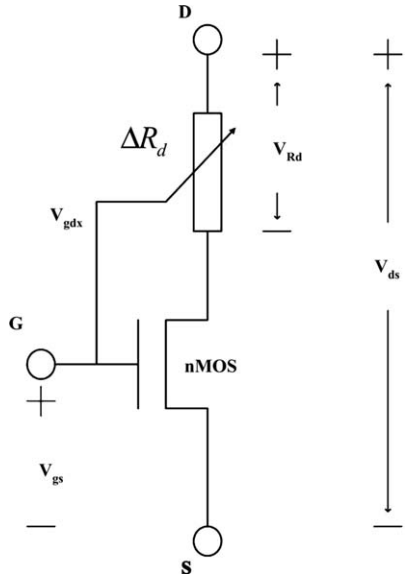

Fig. 5. HCI circuit model. In the model: $V_{gdx} = V_{gs} - V_t - V_{ds}$ and $V_{R_d} = I_{ds}\,\Delta R_d$. V_t is threshold voltage and I_{ds} is the current from node D to S.

account for some of the observed enhanced degradation effects in n-MOSFETs which could not be explained solely by ΔN_{it} generation.

The HCI circuit model is illustrated in Fig. 5. The derivation of ΔR_d is carried out assuming that (1) all interface traps are acceptor-like and occupied by electrons, and (2) the channel mobility degradation, μ, is caused by both ΔN_{it} and ΔN_{ox}. The assumption (1) means the net charge in interface traps is a fixed negative charge for n-MOSFET in strong inversion operation. Assumption (2) leads to:

$$\mu = \frac{\mu_0}{(1 + \alpha \cdot \Delta N)}, \tag{43}$$

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 $\alpha \approx 2.4 \times 10^{-12}\ cm^2$ [58].

The drain current drain flowing through an undamaged nMOS (when $\Delta N = 0$ at $t = 0$) is defined as I_{ds0}:

$$I_{ds0} = \mu_0 C_{ox} \frac{W}{L}\left(V_{gs} - V_t - \frac{V_{ds}}{2}\right)V_{ds}. \tag{44}$$

When ΔN is small, the relation between fresh and degraded drain-to-source current is

$$I_{ds} = \frac{I_{ds0}}{1 + \alpha \cdot \Delta N}. \tag{45}$$

This leads to an expression of ΔR_d which is determined by ΔN and the terminal voltages and currents

$$\Delta R_d = \frac{V_{R_d}}{I_{R_d}} = \frac{1 + \alpha \cdot \Delta N}{I_{ds0}} V_{R_d}, \tag{46}$$

where I_{ds0} is given by Eq. (43) and V_{R_d} is calculated using [59]:

$$V_{R_d} = -V_{gdx} + \sqrt{V_{gdx}^2 + 2V_{ds}\Delta N\left[\frac{\alpha\left(V_{gdx}+\frac{V_{ds}}{2}\right)}{1+\alpha\cdot\Delta N}+\frac{q}{C_{ox}}\right]}, \quad (47)$$

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

In quasi-static operation, ΔN is a time-dependent parameter, therefore, ΔR_d is also time-dependent. At any given time t, if ΔN is known, ΔR_d will be uniquely determined. The models for ΔN and ΔN_{ox} have been well documented in literature [53,54]. ΔN_{it} can be calculated using

$$\Delta N_{it} = C_1\left[\frac{I_{ds}}{W}\exp\left(-\frac{\Phi_{it,e}}{q\lambda_e E_m}\right)t\right]^n, \quad (48)$$

where W is the channel width, $\Phi_{it,e}$ is the critical energy for electrons to create an interface trap, λ_e is the hot-electron mean free path and C_1 is a process constant.

The models and model parameters for ΔN_{ox} are given in [55] (on pp. 59–66). For convenience, they are recapitulated as

$$\Delta N_{ox} = N_1[1-\exp(-\sigma_1 I_{ei}t)] - N_2[1-\exp(-\sigma_2 I_{ei}t)]. \quad (49)$$

A set of typical model fitting parameters, for Eq. (49), have been given in [55] (p. 65). The above new ΔR_d model inherits all the merits of HISREM model and it 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.

3.2.2. NBTI

Simulating the impact of NBTI at the circuit level using SPICE is very important [60]. Most of the work on NBTI SPICE simulation is performed such that the degraded circuit behavior is simulated using the transistor parameter V_t, which is shifted by a fixed value [61]. This kind of simulation method cannot physically relate circuit performance degradation to NBTI wearout under dynamic operation conditions since it does not include the NBTI stress time as a parameter. The most effective way to develop such a relation is by developing a NBTI circuit model. However, to the best of our knowledge, no such electrical model exists in literature. Stretched exponential time dependence describes the threshold voltage degradation as

$$\Delta V_t(t) = \Delta V_{max}\left[1-\exp\left(-\frac{t}{\tau}\right)^{\beta}\right]. \quad (50)$$

Thus, a new NBTI circuit model is proposed, which is an electrical model relating the time dependent NBTI physical degradation parameter ΔV_t to lumped electrical model elements, thereby enabling effective and quick NBTI circuit reliability simulation.

The most severe NBTI effect is p-MOSFET threshold voltage increase—ΔV_t, which is equivalent to a decrease in the p-MOSFET absolute gate-to-source voltage. This is equivalent to adding a voltage source at the gate. However, we chose to split the p-MOSFET gate connection and add a gate resistance R_G between the original gate biasing point G and the p-MOSFET gate terminal G'. This allows inclusion of a gate leakage current flowing mechanism (voltage controlled current sources between gate and drain and between gate and source), so that the NBTI model can be included with the TDDB model without a conflict developing by having a voltage source and current source at the same node. In fact, there should be no difference between using a voltage source or a current source to represent ΔV_t, whereby the gate leakage current flows through the gate resistance R_G and increases the p-MOSFET effective gate voltage at point G'. This gate current would depend on the gate voltage because of R. The current may be somewhat un-physical so a substrate bias source may be substituted instead, but many possibilities have yet to be explored in this area.

Since the p-MOSFET source is held fixed at its highest potential, the inclusion of R_G and gate leakage current leads to a decrease of the p-MOSFET absolute gate-to-source voltage, thereby imitating the NBTI threshold voltage degradation. Based on this concept, the NBTI circuit model is constructed and shown in Fig. 6. The advantage of this configuration will become clear when compared to the model for TDDB, presented next.

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 ΔV_t which is time dependent. According to [62], the gate leakage current due to oxide breakdown can be

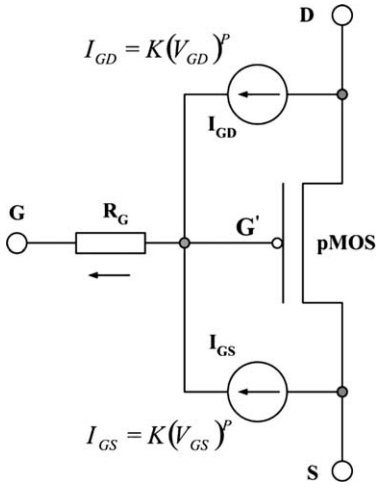

Fig. 6. NBTI circuit model. NBTI-induced p-MOSFET threshold voltage increase is modeled as absolute gate-to-source voltage decrease. Gate tunneling current flowing through the gate resistance R_G leads to the increase of voltage at point G'. This corresponds to the decrease of p-MOSFET absolute gate-to-source 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: $I = KV^p$.

modeled as a gate-to-diffusion leakage current, with a power law dependence of the formula $I = KV^p$ (where K and p are fitting parameters). The same power law voltage dependency, shown in Fig. 6, is adopted in modeling gate leakage currents. As a result, for the gate-to-drain leakage current, $I_{GD} = K(V_{GD})^p$, and for the gate-to-source leakage current, $I_{GS} = K(V_{GS})^p$. The default value of p is set to 5, and the default value of K is 3×10^{-6} [62].

In Fig. 6, the voltage drop across R_G is

$$V_{R_G}(t) = V_{G'} - V_G = \Delta V_G(t) = (I_{GD} + I_{GS})R_G. \quad (51)$$

Threshold voltage degradation ΔV_t due to NBTI is already given by Eq. (50). Therefore, from the relation $\Delta V_G(t) = \Delta V_t(t)$, we obtain an analytical solution for R_G:

$$R_G = \frac{\Delta V_{max}}{KV_{GD}^p + KV_{GS}^p}\left[1 - e^{-\left(\frac{t}{\tau}\right)^\beta}\right]. \quad (52)$$

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 Eq. (52).

One of the most important points shown in Fig. 6 is that this new model is much better than a simple model which only inserts a voltage source between G and G' representing threshold voltage shift in that it inherently incorporates both NBTI and possible oxide breakdown effects.

For nMOS positive bias temperature instability (PBTI) circuit model, a similar structure to that for pMOS NBTI shown in Fig. 6 can be constructed, except that all current flowing directions are reversed and the model fitting parameters of the threshold voltage model ΔV_t (Eq. (50)) are determined from nMOS PBTI stress testing. For the two current sources I_{GD} and I_{GS} in nMOS PBTI circuit model, a better gate leakage model, proposed by Lee et al. [63] is adopted:

$$I_{GS} = \frac{1}{2}AL\exp(\alpha V_{GS} - \beta t_{ox}^{-\gamma}) \quad (53)$$

and

$$I_{GD} = \frac{1}{2}AL\exp(\alpha V_{GD} - \beta t_{ox}^{-\gamma}), \quad (54)$$

where I_{GS} and I_{GD} are in μA, L is effective channel length in nanometer, t_{ox} is oxide thickness in nanometer, $A = 127.04$, $\alpha = 5.61$, $\beta = 10.6$ and $\gamma = 2.5$. These typical values for n-MOSFETs were obtained by fitting industrial data and found to be good for technologies across many generations up to 0.13 μm. These new leakage models are able to maintain good stability in SPICE simulation [63].

3.2.3. TDDB

It is onerous work to develop an effective circuit model for gate oxide breakdown because device post-breakdown behavior is extremely complicated, and has been perplexing the reliability physics community for decades. Device I–V characteristics after gate oxide breakdown relies on many parameters including, breakdown location, transistor type, voltage polarity, device operation mode (accumulation or inversion), oxide area and even poly-gate doping type. Interestingly, literature searches reveal that TDDB failure modeling is a very active area and more than a dozen circuit models have been developed by various research institutes and industrial labs. All this work attempts to develop quantitative methodologies for predicting the response of circuits to gate oxide breakdown events [64]. In this section, some of the most successful TDDB failure circuit models are reviewed. The first equivalent circuit, developed by Rodriguez et al., models oxide soft-breakdown as a progressive wearout phenomenon, whereas the others model oxide breakdown as a catastrophic singular event.

Starting from the observation that a CMOS inverter's transfer curve 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-to-drain leakage current model, which follows the form of a power-law relation as $I = KV_{gd}^p$. Rodriguez et al. at IBM [65,66,62] developed a simple TDDB circuit model, which consists of a voltage-dependent current source bridging either gate-to-drain or gate-to-source, depending on the breakdown location, allowing the oxide breakdown leakage current in a transistor to be simulated in a circuit. This power-law leakage current model is illustrated in Fig. 7.

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 VTC [65].

Rodriguez et al. [62] note that a linear Ohmic oxide breakdown resistance is not sufficient to model the experimental data. The Ohmic model only provides good results

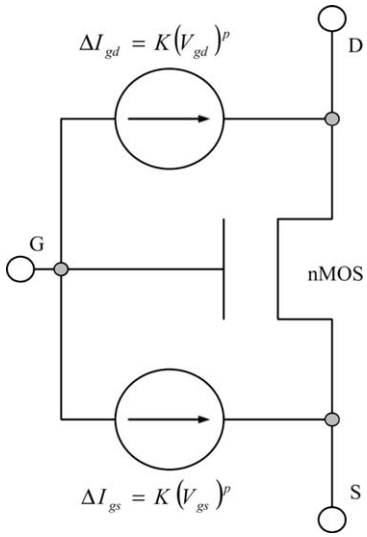

Fig. 7. Power-law leakage current model. The exponent p varies from 5 to 2 as the degradation level increases. K reflects the "size" of the breakdown spot.

for hard breakdown but the power-law leakage current model predicts progressive oxide breakdown behavior prior to the final hard breakdown much better. This model makes the assumption that hard breakdown is part of a continuum of progressive breakdowns and allows their inclusion in a degradation equivalent circuit.

In a MOSFET, the oxide breakdown changes isolation of the device's internal structures by forming an abnormal conduction path and 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 fact that oxide post-breakdown behavior depends on breakdown location (gate-to-channel, gate-to-drain and gate-to-source), transistor type (n-MOSFET and p-MOSFET) and poly-gate doping type (n^+ poly-gate and p^+ poly-gate), Segura et al. [67,68] 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-channel breakdown of n-MOSFET with n^+ poly-gate. For this device family, the gate-to-channel breakdown path between the n^+ poly-gate and n type inversion channel can be modeled as a gate-to-channel resistance R_{GOS}. The formation of this resistance-like breakdown path splits the channel into two parts, which are physically equivalent to two transistors connected in series. This model is illustrated in Fig. 8.

For other combinations of location and doping effects, the models can be readily deduced with the similar principle. For example, when the breakdown path appears between the gate and the drain (or the source) terminals of the n-MOSFET, an n^{++}–n^+ barrier (i.e., n^+ poly-gate to n^+ drain/source diffusion) will form. In this case, the breakdown is modeled with a resistance between gate-to-drain/source.

With these GOS electrical models, Segura et al. [67] 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 (ATPG) may not detect GOS depending on the gate topology.

Gate oxide breakdown equivalent circuit models for analog circuits and RF circuits are also developed in an attempt to expand model applicability and explore 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, which accelerates the offset generation and compromises circuit functionality [69]. As for RF circuits, they are very sensitive to device parameter drift. Therefore, oxide breakdown is expected to have more severe impact on their functionality and performance [70].

Yang et al. [70,71] developed an RF failure 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 Fig. 9 which consists of the original n-MOSFET, terminal series resistances (R_G, R_D, R_S), 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 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

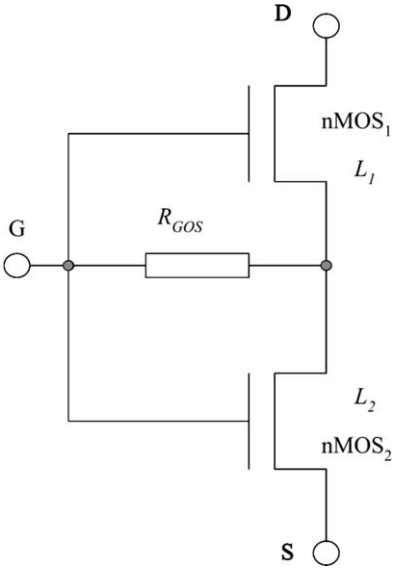

Fig. 8. TDDB GOS model for gate-to-channel breakdown of n-MOSFET with n^+-poly gate. The channel lengths of nMOS1 and nMOS2 follow the relation: $L_1 + L_2 = L$ where L is the undamaged n-MOSFET channel length. The parameter R_{GOS} is related to the size and location of the breakdown path. A value of R_{GOS} as low as 3 KΩ was used in the simulation in [68].

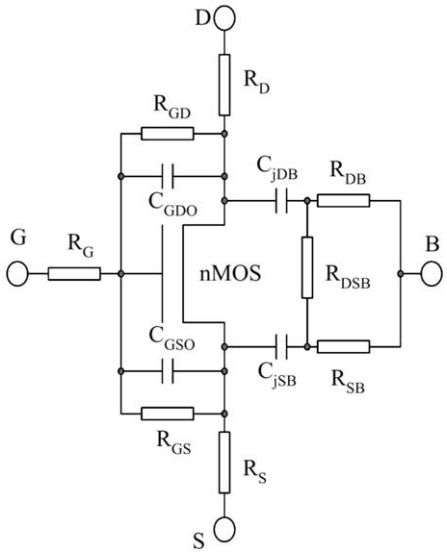

Fig. 9. TDDB RF equivalent circuit model. Model parameters for simulation in [70] are set as: $R_G = 85.4\ \Omega$, $R_D = R_S = 12.14\ \Omega$, $R_{GD} = 6.88$ KΩ, $R_{GS} = 23$ KΩ, $C_{GDO} = C_{GSO} = 15.3$ fF, $C_{jDB} = C_{jSB} = 7$ fF, $R_{DSB} = 80$ KΩ, $R_{DB} = R_{SB} = 49.37\ \Omega$.

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 occurs at gate-to-source or gate-to-drain depending on which resistance is dominant.

Based on this TDDB RF circuit model, the performance degradation of 0.16 μm n-MOSFET devices and a 1.8 GHz LNA circuit was analyzed [70]. For the device *S*-parameters, the inclusion of R_{GD} and R_{GS} changes the device input impedance *S11*, provides an additional connection between gate and drain and therefore degrades the reverse transmission coefficient *S12*, changes the output impedance *S22* at the drain, and also decreases transconductance g_m, which is equivalent to the forward transmission coefficient *S21*. Oxide breakdown has significant impact on the LNA circuit performance: most *S*-parameters drift dramatically and fail to meet the usual performance requirements, input impedance matching is disturbed due to increased gate leakage currents, and the noise figure obviously deteriorates with breakdown paths forming across the gate oxide, which adds another noise source to the transistor.

The most frequently discussed TDDB circuit model has been the one proposed by Kaczer et al. at IMEC [72–77]. In this model, the breakdown path is assumed to form 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 n-MOSFET (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 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) becomes a forward biased p–n junction. Electrons are emitted from the n^+ poly-gate, flow through the 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, an n-MOSFET with oxide breakdown operated at negative gate voltage can be modeled with a gate resistor, two bipolar transistors and the original n-MOSFET [72,74]. Because n-MOSFETs rarely operate at a negative gate voltage situation, this complicated two-bipolar-transistor model for ($V_G < 0$) is not of primary interest.

When the gate voltage is positive enough such that the n-MOSFET is in strong inversion, an n-type conduction channel will form under the gate oxide connecting the source and the drain. Then, the contact region of the breakdown path (n-type) and the channel (n-type) is an Ohmic contact. The positive gate voltage forces the electric field to penetrate the breakdown path and deplete the contact region of the breakdown path and substrate. This contact region serves as an electron sink and can therefore be treated as an additional drain in the middle of the channel.

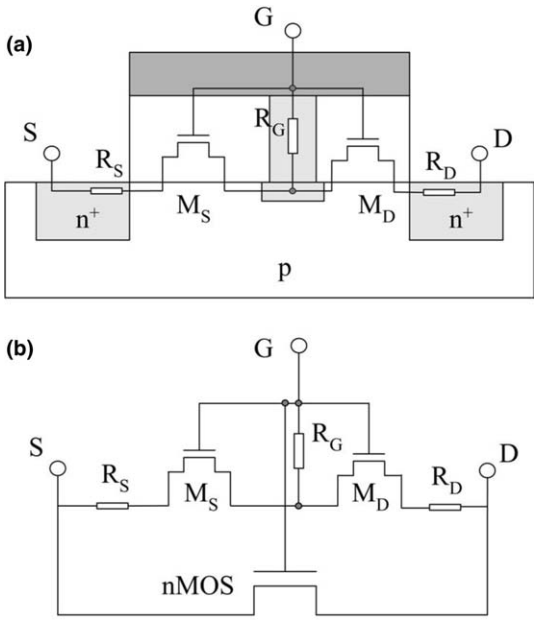

Fig. 10. TDDB circuit model for n-MOSFET with hard gate oxide breakdown and operated in positive gate voltage. (a) Cross-sectional view of breakdown structure. (b) Equivalent circuit model. Model parameters for simulation in [72] are set as: $R_G = 1\ K\Omega$, $L_{M_S} + L_{M_D} = 0.09\ \mu m$, $W_{M_S} = W_{M_D} = 0.25\ \mu m$, R_D and R_S vary from 2.5 KΩ (at source and drain) to 12.5 KΩ (at the middle of the channel).

Based on this microscopic picture, an equivalent electrical circuit for n-MOSFET with hard gate oxide breakdown and operated at positive gate voltage has been constructed and shown in Fig. 10.

Apart from the original n-MOSFET (nMOS), the model contains a constant resistance (R_G) corresponding to breakdown path, two adjacent parasitic n-MOSFETs (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 effect of breakdown location is represented by varying the gate lengths of M_S and M_D. Gate-to-channel breakdowns in the vicinity of the drain or the source are represented by logarithmically varying extension resistances R_S or R_D [72]. 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.

This model has been used in a CMOS ring oscillator oxide breakdown analysis [73]. The simulation shows that gate-to-channel breakdowns have minor effects on the circuit operation but 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 [75]. This conclusion can be explained with the help of the hard breakdown model of Kaczer: 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 behavior. If breakdown happens in the transistor channel region, where the resistance (i.e., channel resistance) of the discharge path is higher, soft breakdown will be triggered.

Even though a lot of work has been done to improve the Kaczer model, careful evaluation in [70] and our critical examination have identified several limitations of this model: (1) The level-1 M_S and M_D models are obsolete; (2) The model only applies to linear operation situation. If breakdown path forms above the saturation region where channel has "pinched-off", the inclusion of the two parasitic transistors, i.e., M_S and M_D, is not valid. (3) M_S and M_D bring two more drain diffusion regions, which do not physically exist. (4) Simulators cannot handle the breakdown position from zero to the entire channel length. (5) It is problematic to preserve the original n-MOSFET in the model if M_S and M_D are included because they already represent all device internal structures after oxide breakdown. Specifically, the entire conducting channel has been physically characterized by M_S and M_D. Therefore, it is erroneous to keep the original n-MOSFET in the post-breakdown TDDB circuit model. (6) The assumption that the breakdown path is n-type silicon is arbitrary and physically unjustified. The last two points are the most important ones and they call for developing a physically justifiable circuit model for gate oxide breakdown. A new equivalent circuit model, which overcomes these last two limitations of the previous models, is presented.

Besides the brief review above, there are many other successful models worth mentioning [78–83]. A p-MOSFET gate-to-channel breakdown model is proposed in [78] and used to investigate its effect on logic gate failures. A pair of breakdown models for n-MOSFET and p-MOSFET (only gate-to-diffusion breakdowns) is proposed and used to transform the effect of oxide breakdown into a delay fault or a logic fault [79]. Yeoh et al. [80,81] 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 the device internal connections after oxide breakdown path formed at different locations. Based on the work of linear non-split MOS model and non-linear two-dimensional channel split MOS model [82], a non-linear non-split MOS oxide breakdown model is developed in [83] in an attempt to enable circuit simulation of gate-to-channel effect on minimum length transistors. Even though these models do not accurately model all aspects of breakdown, the development of fundamental concepts, physical principles and modeling techniques in these models is good infrastructure for constructing any advanced oxide breakdown circuit models. Following this conclusion, a new TDDB circuit model is developed below.

From the semiconductor materials point of view, it is improper to assume the breakdown path as n-type silicon diffusion because this is not physically substantiated. The oxide breakdown path is actually a defect-assisted electron conduction rather than a reliable physical connection. Therefore, the resistance cannot be solely used to model gate-to-channel and gate-to-diffusion breakdowns. The correct modeling method should be based on the channel potential re-distribution concept. The oxide breakdown disturbs the channel surface potential around the breakdown path, where the gradual channel approximation (GCA) fails, so a new three-dimensional channel potential model has to be developed. According to [68], in a 3-D coordinate system with x along the channel length (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 the channel surface can be defined as: $x = L_1$, $y = 0$ and $z = W_1$ (see 10 in [68]). The drain current I_D of a defect-free MOSFET can be obtained from

$$I_D = \frac{W}{L}[f(\Psi(x = L)) - f(\Psi(x = 0))], \tag{55}$$

where $\Psi(x)$ is the channel surface potential at x, f is a function of channel mobility, oxide capacitance, threshold voltage and the 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 into two regions, and similar to Eq. (55), the drain and source currents of the damaged MOSFET can be written as [68]:

$$I_D = \frac{W}{L - L_1}[f(\Psi(x = L)) - f(\Psi(x = L_1))] \tag{56}$$

and

$$I_D = \frac{W}{L_1}[f(\Psi(x = L_1)) - f(\Psi(x = 0))], \tag{57}$$

where $\Psi(x = L_1)$ is the surface potential under the breakdown path. Eqs. (56) and (57) show that an n-MOSFET 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 to provide a conduction path to inject electrons from the channel into the gate. Therefore, a voltage dependent current source I_{OX} connecting between gate and channel can be used to model this effect. Based on the above discussion, a new TDDB circuit model is obtained and illustrated in Fig. 11.

It seems that this model requires two model parameters (L_1 and V_i, which is the voltage at the connection point of M_1 and M_2), but 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 requiring characterization, which facilitates the application of this model.

Suppose the original drain-to-source current of a fresh n-MOSFET is I_{ds0}, and neglect the effect of R_D, R_S and the short-channel effect (in order to simplify equation derivation), we can write I_{ds0} as

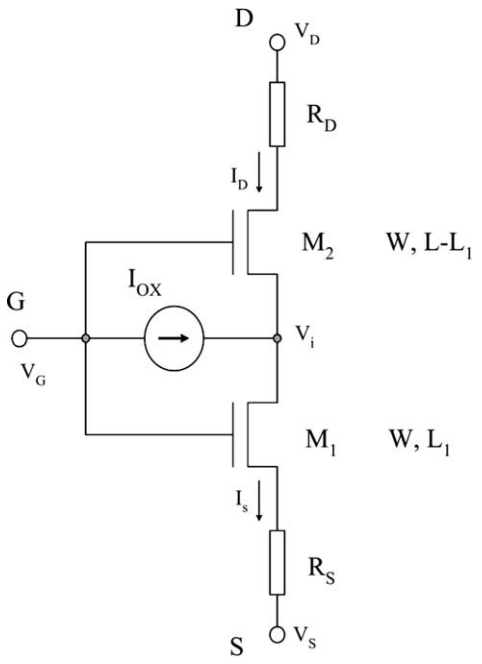

Fig. 11. TDDB circuit model for n-MOSFET with hard gate oxide breakdown. $I_{OX} = I_S - I_D$ is a voltage dependent current source representing breakdown path current injection effect. R_D and R_S characterize the resistance in the source and the drain extensions, respectively. L_1 represents breakdown location away from the source edge.

$$I_{ds0} = \mu_n C_{ox} \frac{W}{L}\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right] \tag{58}$$

Applying Kirchhoff's Current Law (KCL) to Fig. 11 (for simplicity, neglecting R_D and R_S), we get the following equations:

$$I_{OX} = I_S - I_D, \tag{59}$$

$$I_D = \mu_n C_{ox} \frac{W}{L_2}\left[(V'_G - V_i)(V_D - V_i) - \frac{1}{2}(V_D - V_i)^2\right], \tag{60}$$

$$I_D = \mu_n C_{ox} \frac{W}{L_1}\left[(V_{GS} - V_t)(V_i - V_S) - \frac{1}{2}(V_i - V_S)^2\right], \tag{61}$$

where $L_2 = L - L_1$ is the channel length of M_2, $V'_G = V_G - V_{t2}$, V_{t2} is the original threshold voltage V_t plus body bias ($V_{sb} = V_i$) induced enhancement effect. 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. The gate voltage cannot control and sustain channel current as well as before, which leads to degradation of the drain current. Therefore, a good assumption in Fig. 11 is that the source current I_S maintains its value as before, whereas the injection of I_{OX} degrades I_D. This means $I_S = I_{ds0}$. From Eqs. (58) and (61), we solve for V_i:

$$V_i = V_{Gon} - \sqrt{V_{Gon}'^2 - \left(V_S^2 + V_{ov}V_S + \frac{2I_{ds0}L_1}{\mu_n C_{ox} W}\right)}, \tag{62}$$

where $V_{Gon} = V_G - V_t$, $V_{ov} = V_{Gon} - V_S$ is the gate overdrive voltage. If V_S is tied to ground, Eq. (62) is reduced to

$$V_i = (V_G - V_t) - \sqrt{(V_G - V_t)^2 - \frac{2I_{ds0}L_1}{\mu_n C_{ox} W}}. \tag{63}$$

Eq. (63) (or Eq. (62) 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, according to Eqs. (58)–(63), the voltage dependent current source I_{OX} can be obtained.

The above n-MOSFET TDDB circuit model can be easily extended to p-MOSFET by properly changing the current flow directions in Fig. 11 and voltage/current signs in the model equations.

4. Failure-rate-based SPICE (FaRBS) simulator

4.1. Assumptions of the FaRBS model

(1) *Constant failure rate assumption*. Eventhough the four failure mechanisms we are investigating (EM, HCI, NBTI, and TDDB) do not all follow exponential distributions, we have justified that for a complex electronic system with multiple failure mechanisms, an exponential distribution can be used to approximate the overall failure rate. We have also proven that in the rate-based reliability analysis method, the distribution for each failure mechanism is not very important for determining the failure rate. All trends in hazard rate of each failure mechanism will be averaged out to yield a constant level of failure rate.

(2) *Equal contribution assumption*. This assumption can be changed if failure analysis results show the actual distribution of field failures. We assume here that the device is designed properly with no dominant failure mechanism. As a result of improved knowledge of device failure mechanisms, electronic components are designed at the edge of 'reasonable' life under tightly controlled specifications. Therefore, if any failure mechanism is more significant than the others, specific design and manufacturing techniques will be developed to address this dominant failure. This assumption is the extension of the previous one. When no single failure mechanism dominates, all are equally likely and the resulting failure distributions resemble constant rate processes.

(3) *Proportional acceleration assumption*. Constant failure rate-based reliability for electronic components allows the manufacturer to test parts under accelerated conditions on condition that all failure mechanisms are accelerated in approximately the same proportion. The resulting failure rate could then be extrapolated to operating conditions considering temperature, frequency and applied voltage. If this assumption is wrong, the accelerated stress test will lose its ground.

(4) *Linear superposition assumption*. System failure rate is linearly modeled as a sum of individual failure mechanisms. We do not consider interactions between different failure mechanisms. For example, HCI will precipitate the occurring of TDDB, but the interrelation is very complex and normally insignificant and negligible.

(5) *Technology dependence assumption*. All model parameters for each failure mechanism are only technology dependent. For different technologies, model parameters may vary, but for the same technology, they are assumed to be constant. This assumption allows manufacturers to design, accelerate and measure simple but typical sample structures to determine each model parameter. Then, designers can use these extracted model parameters to estimate failure rate of a device manufactured by the same technology.

$$\mathrm{MTTF}_{\mathrm{EM}} = A_{\mathrm{EM}}(J \times T)^{-2} \exp\left(\frac{E_{\mathrm{aEM}}}{kT}\right), \tag{64}$$

$$\mathrm{MTTF}_{\mathrm{HCD}} = A_{\mathrm{HCD}} \exp\left(\frac{\theta}{V_{\mathrm{ds}}}\right), \tag{65}$$

$$\mathrm{MTTF}_{\mathrm{TDDB}} = A_{\mathrm{TDDB}} A_{\mathrm{G}} \left(\frac{1}{V_{\mathrm{gs}}}\right)^{(\alpha-\beta T)} \exp\left(\frac{X}{T} + \frac{Y}{T^2}\right), \tag{66}$$

$$\mathrm{MTTF}_{\mathrm{NBTI}} = A_{\mathrm{NBTI}} \left(\frac{1}{V_{\mathrm{gs}}}\right)^{\gamma} \exp\left(\frac{E_{\mathrm{aNBTI}}}{kT}\right). \tag{67}$$

The lifetime of each wearout failure mechanism for each interconnect and MOSFET in a circuit can be determined by Eqs. (64)–(67). To obtain the lifetime for the entire circuit, we need to combine the effects of these different wearout mechanisms across different structures. This requires information of the time-dependent lifetime distribution for each wearout mechanism. In engineering applications, the FIT value is normally used to qualify product reliability, which represents the number of failures per 10^9 device-hours of accelerating test. Since most FIT calculation methods only apply to systems with constant failure rate for each failure mechanism, care must be used for systems having failure modes with time-varying characteristics [84].

With further developments in deep submicron technologies, IC's become increasingly complex, so both the physical dimensions and logic functions to their limits. Every unit is prone to fail in a shorter time, and if it does fail, the system may be greatly impaired or may even fail altogether. We can therefore approximate a complex integrated circuit by a competing failure system, i.e., a series failure system. Another practical approximation is that each failure mechanism has an exponential lifetime distribution. In this way, the failure rate of each failure mechanism is treated as a constant. This assumption is normally inaccurate for wearout failure mechanisms because their failure rates have decreasing values at the beginning of the component lifetime and go up as the component ages. However, since our intention is to characterize product reliability with MTTF or FIT, we will deal with the average value or mean value of the failure rate. Therefore, the trends of decreasing and then increasing failure rates will be averaged out to an approximately constant process. With the above two assumptions, we can apply the standard sum-of-failure-rates (SOFR) model widely used in industry to determine a system's failure rate from its individual failure mechanism [85].

From the SOFR model, the lifetime MTTF_s of a circuit composed of n units can be related to the lifetime MTTF_{ij} of each unit due to each of its m individual failure mechanisms:

$$\mathrm{MTTF}_s = \frac{1}{\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{\mathrm{MTTF}_{ij}}}. \tag{68}$$

The FIT is interchangeable with MTTF according to its definition for constant failure rate system:

$$\mathrm{FIT}_s = \frac{10^9}{\mathrm{MTTF}_s}. \tag{69}$$

Based on the four wearout failure models proposed in Section 2 and the SOFR model, a new failure rate-based SPICE reliability simulation methodology is proposed, whose basic philosophy and assumptions are the same as those of an acceleration test. This methodology emphasizes the effects of the device or circuit operating parameters (such as current, voltage, frequency, temperature and power) on their failure rates, rather than on their degradation behavior. MTTF and FIT are the primary reliability parameters to be investigated, and no Age parameter is required to model the aging process.

A SPICE circuit simulator is used to determine the operating parameters for each interconnect and MOSFET in a circuit. Other model constants, fitting parameters and activation energies in the wearout models will be extracted from experimental and testing data. Then, the MTTF and FIT of the circuit can be readily calculated with Eqs. (64)–(69). It is obvious that the failure rate-based methodology does not simulate the degradation of devices in a repetitive scheme as other methods introduced in Section 3.1. Therefore, the parameter extraction work for degraded devices is greatly simplified. Another advantage is that this simulation method elevates the reliability simulation from transistor level to circuit level and can be easily used to estimate lifetimes for various device families.

The flow chart of the entire simulation process is depicted in Fig. 12.

The above reliability simulation methodology provides a way for reliability projections and scaling behavior predictions of CMOS devices. By employing the latest MOS SPICE models (such as BSIM3v3 or BSIM4) and fitting parameters as well as the projected process parameters for deep submicron technologies, the reliability trend of each wearout failure mechanism may be simulated according to its lifetime model.

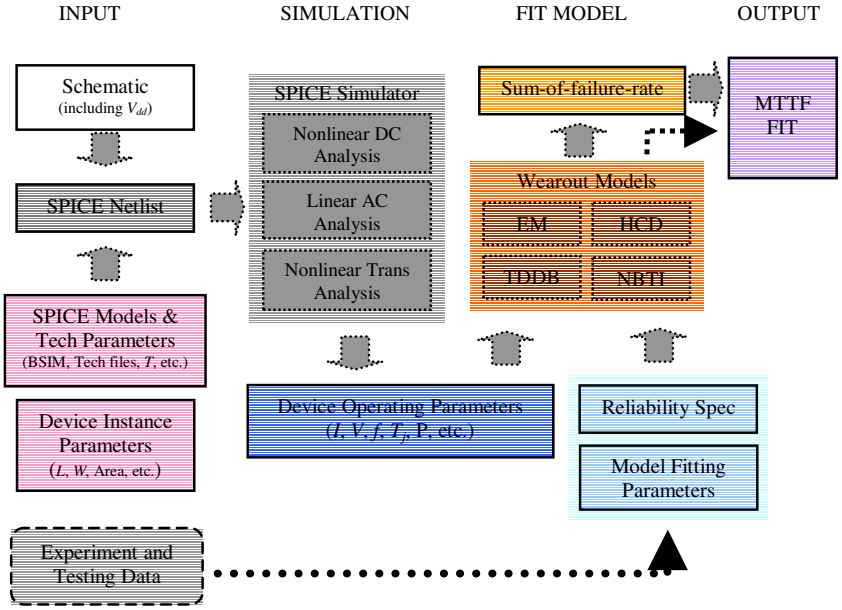

Fig. 12. Flow chart of the failure rate-based simulation process.

This simulation method can also be used to investigate the derating characteristics of semiconductors and provide ways to validate derating guidelines. In order to quantify the lifetime improvement for a device working under derated conditions, a term derating factor, D_f, is defined as the ratio of measured MTTF of a semiconductor device under manufacturer-rated operating conditions to the MTTF of an identical device operating under derated conditions [86]:

$$D_f = \frac{\mathrm{MTTF}_{\mathrm{derated}}}{\mathrm{MTTF}_{\mathrm{rated}}} \tag{70}$$

According to the four wearout failure models, the derating factors for EM, HCD, TDDB and NBTI can be given, respectively, as:

$$D_{f\mathrm{EM}} = \left(\frac{J_0 T_0}{JT}\right)^2 \exp\left(\frac{E_{\mathrm{aEM}}}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right), \tag{71}$$

$$D_{f\mathrm{HCD}} = \exp\left(\theta\left(\frac{1}{V_{\mathrm{ds}}} - \frac{1}{V_{\mathrm{ds0}}}\right)\right), \tag{72}$$

$$D_{f\mathrm{TDDB}} = \frac{(V_{\mathrm{gs0}})^{\alpha-\beta T_0}}{(V_{\mathrm{gs}})^{\alpha-\beta T}} \exp\left(X\left(\frac{1}{T} - \frac{1}{T_0}\right) + Y\left(\frac{1}{T^2} - \frac{1}{T_0^2}\right)\right), \tag{73}$$

$$D_{f\mathrm{NBTI}} = \left(\frac{V_{\mathrm{gs0}}}{V_{\mathrm{gs}}}\right)^{\gamma} \exp\left(\frac{E_{\mathrm{aNBTI}}}{k}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right). \tag{74}$$

Eqs. (71)–(74) provide the basis to estimate the lifetime enhancement when derating temperature, frequency, current and voltage. With reduced voltage and frequency, resulting power dissipation may be accurately calculated based on the SPICE simulation. Since these are all in proportional, the ratios may be applied to a full circuit and the scenario of the operational characteristics at lower voltage/speed operation is obtained. Derating operating parameters may reduce device performance, but the reduction of electrical or thermal stresses within the device will also reduce the degradation rate, thereby improving device reliability and lifetime.

It is worthwhile to emphasize that the MTTF or FIT value alone does not portray a complete picture of the circuit reliability. The time distribution of device lifetime due to each failure mode is very important. With this information, the failure rate-based SPICE reliability simulation method may provide insight into and ways to investigate the accurate ratios of the voltage and temperature acceleration factors for any individual failure mechanism. The simulation results can then be compared to the tested voltage/temperature acceleration data from the manufacturers to calculate the expected FIT values for parts operating in the field.

The main purpose of utilizing a failure-rate-based reliability analysis over standard time-to-fail calculations is that it eliminates the demand for analyzing each detail of every node in each circuit. Our approach assumes that the appropriate precautions were taken by using the BERT methodology when the circuit was designed, so that there is no vulnerable node or circuit that will lead to consistent single-mode field failures. Those calculations were critical for the chip designer, but are not so important for the user. FaRBS takes over when multiple mechanisms may lead to failure of a system including many circuits, where many approximations over many failure modes over the whole chip will result in an average that customers will be able to accept as a probability of failure over time. There is still much work to be done from this point when FaRBS is implemented for specific circuit classes (e.g., SRAM), and these results are compared with industrial evaluation.

5. Conclusion

In this paper, two state-of-the-art degradation-based reliability simulation methodologies are reviewed and a new failure rate-based SPICE reliability simulation methodology is proposed to address some limitations inherent in the former methods. Both types of simulation are based on the same wearout failure physics but addressing reliability from different perspectives. Therefore, they are both necessary for product designers, users and manufacturers. The chip designers need degradation-based approach to make sure that there are no design-vulnerable parts inherent in the chip. The failure-rate based approach is for the users or system designers to qualify product reliability by assuming that all failures will be random and scaleable and circuits do not have any one failure mode dominating by design, otherwise, the device designers or part manufacturers will develop methods to address those more significant failure mechanisms. The simplicity and power of this reliability simulation method make it an additional tool for designers and users to estimate product reliability and allow system designers to de-rate products for longer life applications.

References

[1] Mclinn JA. Constant failure rate—a paradigm in transition. Qual Reliab Eng Int 1990;6:237–41.

[2] Bowles JB. A survey of reliability-prediction procedures for microelectronic devices. IEEE Trans Reliab 1992;41:2–12.

[3] MIL-HDBK-217F reliability prediction of electronic equipment. Washington, DC 20301: Department of Defense, December 1991.

[4] Pecht MG, Nash FR. Predicting the reliability of electronic equipment. Proc IEEE 1994;82:992–1004.

[5] Perry WJ. Specifications and standards—a new way of doing business. US Department of Defense Policy Memorandum, 1994.

[6] Jensen F. Electronic component reliability. New York: John Wiley & Sons; 1995.

[7] Young D, Christou A. Failure mechanism models for electromigration. IEEE Trans Reliab 1994;43:186–92.

[8] Ohring M. Reliability and failure of electronic materials and devices. San Diego: Academic Press; 1998.

[9] Electromigration for designers, an introduction for the non-specialist. Available from: <www.cadence.com/whitepapers/withpdf/4095ElectromigrationWP.pdf>, 1999.

[10] Hau-Riege CS, Thompson CV. Electromigration in Cu interconnects with very different grain structures. Appl Phys Lett 2001;78:3451–543.

[11] Fischer AH, Abel A, Lepper M, Zitzelsbergr A, Von Glasow A. Modeling bimodal electromigration failure distributions. Microelectron Reliab 2001;41:445–53.

[12] Yokogawa S, Okada N, Kakuhara Y, Takizawa H. Electromigration performance of multi-level damascene copper interconnects. Microelectron Reliab 2001;41:1409–16.

[13] Gall M, Capasso C, Jawarani D, Hernandez R, Kawasaki H. Statistical analysis of early failures in electromigration. J Appl Phys 2001;90:732–40.

[14] JEP122-A. Failure mechanisms and models for semiconductor devices. JEDEC Publication, JEDEC Solid State Technology Association, 2002.

[15] Quddus MT, DeMassa TA, Sanchez JJ. Unified model for Q(BD) prediction for thin gate oxide MOS devices with constant voltage and current stress. Microelectron Eng 2000;51–52:357–72.

[16] Alam M, Weir B, Silverman P. A future of function or failure. IEEE Circuits Dev 2002;18:42–8.

[17] Schuegraf KF, Hu C. Hole injection SiO_2 breakdown model for very low voltage lifetime extrapolation. IEEE Trans Electron Dev 1994;41:761–7.

[18] McPherson JW, Mogul HC. Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO_2 thin films. J Appl Phys 1998;84:1513–23.

[19] Wu E, Suné J, Nowak E, McKenna J, Harmon D. Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate oxides. Solid-State Electron 2002;46:1787–98.

[20] Wu EY, Abadeer WW, Han LK, Lo SH, Hueckel GR. Challenges for accurate reliability projections in the ultrathin oxide regime. In: IEEE international reliability physics symposium proceedings, 1999. p. 57–65.

[21] Sune J, Placencia I, Barniol N, Farres E, Martin F, Aymerich X. On the breakdown statistics of very thin SiO_2 films. Thin Solid Films 1990;185:347–62.

[22] Massoud HZ, Deaton R. Percolation model for the extreme-value statistics of dielectric breakdown in rapid-thermal oxides. In: ECS spring meeting, 1994. p. 287–8.

[23] Degraeve R, Groeseneken G, Bellens R, Depas M, Maes HE. A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides. IEDM Tech Dig 1995:863–6.

[24] Degraeve R, Groeseneken G, Bellens R, Ogier JL, Depas M, Roussel Ph, et al. New insights in the relation between electron trap generation and the statistical properties of oxide breakdown. IEEE Trans Electron Dev 1998;45:904–11.

[25] DiMaria DJ, Stathis JH. Explanation for the oxide thickness dependence of breakdown characteristics of metal-oxide semiconductor structures. Appl Phys Lett 1997;70:2708–10.

[26] Stathis JH. Quantitative model of the thickness dependence of breakdown in ultra-thin oxides. Microelectron Eng 1997;36:325–8.

[27] Paulzen GM. Q_{bd} dependencies of ultrathin gate oxides on large area capacitors. Microelectron Eng 1997;36:321–4.

[28] Mahapatara S, Parikh CD, Rao VR, Viswanathan CR, Vasi J. Device scaling effects on hot-carrier induced interface and oxide-trapped charge distributions in MOSFET's. IEEE Trans Electron Dev 2000;47:789–96.

[29] Groeseneken G, Degraeve R, Nigam T, Den Bosch GV, Maes H. Hot carrier degradation and time-dependent dielectric breakdown in oxides. Microelectron Eng 1999;49:27–40.

[30] Maes H, Groeseneken G, Degraeve R, Blauwe JD, Den Bosch GV. Assessment of oxide reliability and hot carrier degradation in CMOS technology. Microelectron Eng 1998;40:147–66.

[31] Hu C, Tam SC, Hsu FC, Ko PK, Chan TY, Terrill KW. Hot-carrier-induced MOSFET degradation—model, monitor, and improvement. IEEE Trans Electron Dev 1985;32:375–84.

[32] Acovic A, Rosa GL, Sun YC. A review of hot-carrier degradation mechanisms in MOSFETs. Microelectron Reliab 1996;36:845–69.

[33] Deal BE, Sklar M, Grove AS, Snow EH. Characteristics of the surface-state charge (Q_{ss}) of thermally oxidized silicon. J Electrochem Soc 1967;114:266.

[34] Hofstein SR. Stabilization of MOS devices. Solid-State Electron 1967;10:657–70.

[35] Lu D, Ruggles GA, Wortman JJ. Effects of processing conditions on negative bias temperature instability in metal-oxide-semiconductor structures. Appl Phys Lett 1988;52:1344–6.

[36] Haywood SK, DeKeersmaecker RF. Hole trapping and interface state generation during bias-temperature stress of SiO_2 layers. Appl Phys Lett 1985;47:381–3.

[37] Breed DJ. Non-ionic room temperature instabilities in MOS devices. Solid-State Electron 1974;17:1229.

[38] Breed DJ. A new model for the negative voltage instability in MOS devices. Appl Phys Lett 1975;26:116–8.

[39] Blat CE, Nicollian EH, Poindexter EH. Mechanism of negative-bias-temperature instability. J Appl Phys 1991;69:1712–20.

[40] Jeppson KO, Svensson CM. Negative bias stress of MOS devices at high electric fields and degradation of NMOS devices. J Appl Phys 1977;48:2004–14.

[41] Ogawa S, Shimaya M, Shiono N. Interface-trap generation at ultrathin SiO_2 (4-6NM)–Si interfaces during negative-bias temperature aging. J Appl Phys 1995;77:1137–48.
[42] Gerardi GH, Poindexter EH, Caplan PJ, Harmatz M, Buchwald WR, Johnson NM. Generation of PB centers by high electric-field—thermochemical effects. J Electrochem Soc 1989;136:2609–14.
[43] Krishnan MS, Kol'dyaev V, Morifuji E, Miyamoto K, Brozek T, Li XL. Series resistance degradation due to NBTI in PMOSFET. Microelectron Reliab 2002;42:1433–8.
[44] Zafar S, Lee YH, Stathis J. Evaluation of NBTI in HfO2 gate-dielectric stacks with tungsten gates. IEEE Electron Dev Lett 2004; 25:153–5.
[45] Tu RH, Rosenbaum E, Chan WY, Li CC, Minami E, Quader K, et al. Berkeley reliability Tools-BERT. IEEE Trans Computer-aided Des Integrated Circuits Syst 1993;12:1524–34.
[46] Xuan X, Chatterjee A, Singh AD, Kim NP, Chisa MT. IC reliability simulator ARET and its application in design-for-reliability. In: Proceedings of the 12th Asian test symposium, 2003. p. 19–22.
[47] Cadence Virtuoso UltraSim Full Chip Simulator Datasheet. Cadence Design Systems. Available from: http://www.cadence.com/.
[48] Hu C. IC reliability simulation. IEEE J Solid-State Circuits 1992;27: 241–6.
[49] Cadence White paper. Reliability Simulation in Integrated Circuit Design. Available from: http://www.cadence.com/.
[50] Karam M, Fikry W, Ragai H. Implementation of hot-carrier reliability simulation in Eldo. Mentor Graphics Deep Submicron Technical Publication, 2000.
[51] Quader KN, Li CC, Tu R, Rosenbaum E, Ko PK, Hu C. A bidirectional NMOSFET current reduction model for simulation of hot-carrier-induced circuit degradation. IEEE Trans Electron Dev 1993;40:2245–54.
[52] Quader KN, Li CC, Tu R, Rosenbaum E, Ko PK, Hu C. A new approach for simulation of circuit degradation due to hot-electron damage in NMOSFETs. IEDM Technical Digest 1991:337–40.
[53] Hu C, Tam SC, Hsu FC, Ko PK, Chan TY, Terrill KY. Hot-electron-induced MOSFET degradation—model, monitor, and improvement. IEEE J Solid-State Circuits 1985;20:295–305.
[54] Leblebici Y, Kang SM. Hot-carrier reliability of MOS VLSI circuits. Norwell: Kluwer Academic Publishers; 1993.
[55] Leblebici Y, Kang SM. Modeling of nMOSFET transistors for simulation of hot-carrier-induced device and circuit degradation. IEEE Trans Computer-Aided Des 1992;11:235–46.
[56] Li PC, Stamoulis GI, Hajj IN. iProbe-d: a hot-carrier and oxide reliability simulator. In: IEEE proceedings of international reliability physics symposium, IRPS, 1994. p. 274–9.
[57] Leblebici Y, Kang SM. Modeling and simulation of hot-carrier-induced device degradation in MOS circuits. IEEE J Solid-State Circuits 1993;28:585–95.
[58] Hwang N, Forbes L. Hot-carrier induced series resistance enhancement model (HISREM) of nMOSTFET's for circuit simulation and reliability projections. Microelectron Reliab 1995;35:225–39.
[59] Li XJ, Bernstein JB. Private communication.
[60] Puchner H, Hinh L. NBTI reliability analysis for a 90 nm CMOS technology. In: Proceeding of the 34th European solid-state device research conference. ESSDERC'04, 2004. p. 257–60.
[61] Reddy V, Krishnan AT, Marshall A, Rodriguez J, Natarajan S, et al. Impact of negative bias temperature instability on digital circuit reliability. Microelectron Reliab 2005;45:31–8.
[62] Rodriguez R, Stathis JH, Linder BP. A model for gate-oxide breakdown in CMOS inverters. IEEE Electron Dev Lett 2003;24: 114–6.
[63] Lee D, Blaauw D, Sylvester D. Gate leakage current analysis and reduction for VLSI circuits. IEEE Trans Very Large Scale Integration (VLSI) Syst 2004;12:155–66.
[64] Stathis JH. Physical models of ultra thin oxide reliability in CMOS devices and implications for circuit reliability. IWGI 2001:26–9.
[65] Stathis JH, Rodriguez R, Linder BP. Circuit implications of gate oxide breakdown. Microelectron Reliab 2003;43:1193–7.
[66] Rodriguez R, Stathis JH, Linder BP, Kowalczyk S, Chuang CT, Joshi RV, et al. The impact of gate-oxide breakdown on SRAM stability. IEEE Electron Dev Lett 2002;23:559–61.
[67] Segura J, Benito CD, Rubio A, Hawkins CF. A detailed analysis of GOS defects in MOS transistors: testing implications at circuit level. IEEE International Test Conference 1995:544–51.
[68] Segura J, Benito CD, Rubio A, Hawkins CF. A detailed analysis and electrical modeling of gate oxide shorts in MOS transistors. J Electron Testing: Theory Applicat 1996;8:229–39.
[69] Avellan A, Krautschneider WH. Impact of soft and hard breakdown on analog and digital circuits. IEEE Trans Dev Mater Reliab 2004; 4:676–80.
[70] Yang H, Yuan JS, Liu Y, Xiao E. Effect of gate oxide breakdown on RF performance. IEEE Trans Dev Mater Reliab 2003;3:93–7.
[71] Yang H, Yuan JS, Xiao E. Effect of gate oxide breakdown on RF device and circuit performance. In: IEEE annual international reliability physics symposium, IRPS'03, 2003. p. 1–4.
[72] Kaczer B, Degraeve R, Rasras M, De Keersgieter A, Van De Mieroop K, Groeseneken G. Analysis and modeling of a digital CMOS circuit operation and reliability after gate oxide breakdown: a case study. Microelectron Reliab 2002;42:555–64.
[73] Kaczer B, Degraeve R, Rasras M, Van De Mieroop K, Roussel PJ, Groeseneken G. Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability. IEEE Trans Electron Dev 2002; 49:500–6.
[74] Kaczer B, Degraeve R, De Keersgieter A, Van De Mieroop K, Simons V, Groeseneken G. Consistent model for short-channel nMOSFET after hard gate oxide breakdown. IEEE Trans Electron Dev 2002;49:507–13.
[75] Degraeve R, Kaczer B, De Keersgieter A, Groeseneken G. Relation between breakdown mode and location in short-channel nMOSFETs and its impact on reliability specifications. IEEE Trans Dev Mater Reliab 2001;1:163–9.
[76] Kaczer B, Crupi F, Degraeve R, Roussel PH, Ciofi C, Groeseneken G. Observation of hot-carrier-induced nFET gate-oxide breakdown in dynamically stressed CMOS circuits. IEEE IEDM 2002: 171–4.
[77] Kaczer B, Degraeve R, De Keersgieter A, Van De Mieroop K, Bearda T, Groeseneken G. Consistent model for short-channel nMOSFET post-hard-breakdown characteristics. In: Symposium on VLSI technology, 2001. p. 121–2.
[78] Shaheen M, Mourad S. Gate to channel shorts in pMOS devices: effects on logic gate failures. In: Proceedings of IEEE international symposium on circuits and systems, ISCAS'98, vol. 6, 1998. p. 211–4.
[79] Lu X, Li Z, Qiu W, Walker H, Shi W. A circuit level fault model for resistive shorts of MOS gate oxide. In: 5th international workshop on microprocessor test and verification, 2004.
[80] Yeoh TS, Kamat NR, Nair RS, Hu SJ. Gate oxide breakdown model in MOS transistors. Proc IEEE Int Reliab Phys 1995:149–55.
[81] Yeoh TS, Hu SJ. Influence of MOS transistor gate oxide breakdown on circuit performance. In: Proceedings IEEE international conference on semiconductor electronics, 1998. p. 59–63.
[82] Renovell M, Galliere JM, Azais F, Bertrand Y. Delay testing of MOS transistor with gate oxide short. In: Proceedings of the twelfth Asian symposium ATS'03, 2003. p. 168–73.
[83] Renovell M, Galliere JM, Azais F, Bertrand Y. Modeling gate oxide short defects in CMOS minimum transistors. In: Proceedings of the 7th IEEE European test workshop, 2002. p. 15–20.
[84] JESD85. Methods for calculating failure rates in units of FITs. JEDEC Publication, JEDEC Solid State Technology Association, 2001.
[85] Srinivasan J, Adce SV, Bose P, et al. The case for microarchitectural awareness of lifetime reliability. In: Proceedings of the 31st annual international symposium on Computer Arch, 2004.
[86] Walter JD. Methods to account for accelerated semiconductor device wearout in longlife aerospace applications. PhD dissertation, University of Maryland, College Park, 2003.