FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 31, N° 3, September 2018, pp. 343-366
https://doi.org/10.2298/FUEE1803343S

A COMPARATIVE STUDY OF RELIABILITY FOR FINFET

Saleh Shaheen, Gady Golan, Moshe Azoulay, Joseph Bernstein

Faculty of Engineering, Dept. of Electrical Engineering, Ariel University, Ariel, Israel

Abstract. *The continuous downscaling of CMOS technologies over the last few decades resulted in higher Integrated Circuit (IC) density and performance. The emergence of FinFET technology has brought with it the same reliability issues as standard CMOS with the addition of a new prominent degradation mechanism. The same mechanisms still exist as for previous CMOS devices, including Bias Temperature Instability (BTI), Hot Carrier Degradation (HCD), Electro-migration (EM), and Body Effects. A new and equally important reliability issue for FinFET is the Self -heating, which is a crucial complication since thermal time-constant is generally much longer than the transistor switching times. FinFET technology is the newest technological paradigm that has emerged in the past decade, as downscaling reached beyond 20 nm, which happens also to be the estimated mean free path of electrons at room temperature in silicon. As such, the reliability physics of FinFET was modified in order to fit the newly developed transistor technology. This paper highlights the roles and impacts of these various effects and aging mechanisms on FinFET transistors compared to planar transistors on the basic approach of the physics of failure mechanisms to fit to a comprehensive aging model.*

Key words: *FinFET, Reliability, CMOS FinFET, BTI, HCD, electromigration, aging*

1. INTRODUCTION

1.1. CMOS FinFET Transistors

A Fin Field-effect transistor (FinFET) is a Tri-gate transistor built on a substrate where the gate is placed on three sides of the channel or wrapped around the channel, forming a double gate structure. These devices have been given the generic name "FinFETs" because the source/drain region forms fins on the silicon surface. FinFET devices exhibit significantly faster switching times and higher current density than the mainstream in the planner transistors technology. The term FinFET (fin Field Effect Transistor) was coined in 2001 by the University of California, Berkeley [1]. FinFET Devices are 3D transistors, where the current flows through a thin fin wrapped by a metal gate. In this structure, channel inversion is created in the three walls of the fin, which increases the gate control over the channel and reduces the short channel effects [2]. FinFETs present lower threshold voltage variations than

Received February 8, 2018
Corresponding author: Gady Golan
Faculty of Engineering, Dept. of Electrical Engineering, Ariel University, Ariel 40700, Israel
(E-mail: gadygolan@gmail.com)

planar transistors due to its undoped or slightly doped channel that reduces the impact of random doping fluctuations (RDF). On the other hand, thinner fins also increase the series resistance of the source/drain region due to its small transversal area, and limits the driving current of the FinFET. This reduction in current affects the reliability which will be discussed in the following sections.

1.2. Reliability Issues in Deep Sub-Micron Technologies

As previously mentioned, the semiconductor industry has witnessed remarkable growth and achievements in IC manufacturing through significant scaling in transistor dimensions. Such scaling has not only made the IC more compact and dense, but also enhanced its performance without an increase in its power consumption as long as the chip area was kept constant. Although the vision of Gordon Moore in 1965 that the complexities of an IC will be approximately doubled every two years seemed to be a dream at that time, it came true over four decades. This resulted in the production of more complex circuits [3]. In 2018, there are more than billion transistors in one single processor die! Thus, this high integration density has to be accompanied by tough efforts to increase the IC reliability, since the failure of several transistors in a circuit can lead to a complete failure of the whole system. Despite the fact that there were some claims in the semiconductor industry of hitting a "red brick" wall at100 nm technology node in 1998 [4], leading edge research and development is currently working towards developing of transistors even smaller than 10nm technology node and beyond [5].

As with the continuous downscaling of device dimensions, variations in transistor parameters are increasing drastically and lead to unexpected reliability issues [6, 7]. These issues are essentially classified to "Time-zero" variability issues [8] such as Line-Edge Roughness (LER), Random Dopant Fluctuations (RDFs), Metal Gate Granularity and Body Thickness Variation, that causes intra-die variations during manufacturing process, and "Time-dependent" variability issues that are considered to be a major source for performance degradation of scaled devices over their lifetime, such as Negative Bias Temperature Instability (NBTI) [9], Hot Carrier Injection (HCI) [10], and Time-Dependent Dielectric Breakdown (TDDB) [11] Electro-migration, Self-heating and Body Effects, these degradation mechanisms are caused by the formation of charged traps within the gate oxide layer due to the high electric field and temperature that lead to a change in the device parameters (e.g. threshold voltage, carrier mobility, drain current) over time , depending on the operating conditions and the workload over lifetime. Therefore, these issues degrade the reliability of the scaled devices and eventually may lead to an IC failure, when the variations reach a certain limit.

An example to demonstrate the impact of variability on the scaled devices, are evident in downscaling of technologies. Variation can reach up to 50% of Vth in advanced technology [12], which strongly affects SRAM functionality and pose a major challenge for the SRAM design. Reliability of digital integrated circuits has become one of the critical challenges at Deep Sub-Micron (DSM) semiconductor technologies. Researchers, nowadays, are studying these reliability issues at various levels such as design, process, transistor, and circuit. They also argue that these time-dependent mechanisms can be best described in terms of an ensemble of individual defects and their time, voltage, and temperature dependent properties that can be modeled and inserted into a circuit simulation, and thus, enabling reliability awareness at design [13], as will be discussed in the following paragraphs. In fact, the simulation and analysis of aging effects at higher design levels are basically difficult, since the

degradation rate depends on operating conditions and workloads over the lifetime. These factors are often unknown during the design of a circuit, since the change of workloads applied to a circuit will lead to various amounts of performance degradation, and thus, impose dramatic challenges to the design of digital integrated circuits. In order to improve design predictability and support robust design it is necessary to develop appropriate techniques that are efficiently able to predict the aging effects in existing and future technologies. In this paper, our objective is to provide a detailed and accurate study for predicting aging effects in FinFET compared to planar transistors, due to few of the above mentioned time-dependent phenomena, like NBTI mechanism, Hot Carrier, Self Heat, Body Effects and finally Electro-migration. The rest of this paper is organized as follows:

Paragraph 2 gives the background and physical concepts behind NBTI phenomenon and Hot Carrier, Self Heat, and finally Electro-migration. The mechanisms are explained and the ways of modeling this degradation mechanism in the transistor and gate levels are explained. A new Multiple Temperature Operational life test (MTOL) model [14] is discussed as well. Paragraph 3 presents a comprehensive analysis of the main differences between FinFET transistor and planner transistor from the reliability perspective and presents detailed FinFET Transistors Reliability and Aging Analysis. Paragraph 4 presents a Comparative Discussion, putting it all together. Finally, paragraph 5 summarizes the work that has been done and propose ideas for future improved reliability for FinFET devices.

2. Physical Mechanisms

2.1. Multiple-Temperature Operational Life and FIT

Reliability device simulators have become an integral part of the VLSI design process. These simulators successfully model the most significant physical failure mechanisms, such as Negative Bias Temperature Instability (NBTI), Electro-migration (EM) and Hot Carrier Injection (HCI) in modern electronic devices. 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, the chip level reliability prediction methods are mostly statistical. Chip level reliability prediction tools, today, 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 [14-17]. Chip and packaged system reliability is still measured by a Failure unIT, also defined as the Failure-In-Time (FIT). The FIT is a measure for the constant rate function (Poisson model) failure rate, λ. This model is time-independent, and the failure rate in FIT is 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, based on operation within the specified conditions of voltage, frequency, heat dissipation and more. 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 inverse FIT rate for every component. A FIT is defined in terms of an acceleration factor, AF, seen in equation 1 below:

$$FIT = \frac{\#failures}{\#tested*hours*AF} * 10^9 \quad (1)$$

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, AF, has to be provided by the manufacturer, since only they know the failure mechanisms that are being accelerated in the final High Temperature Operating Life (HTOL) test. This factor is generally based on a company proprietary variant of the MIL-HDBK-217 approach for accelerated life testing. The real task of reliability modeling, therefore, is to choose an appropriate value for AF based on the physics of the dominant failure mechanisms that would occur in the field for the device. The key innovation of the Multiple-Temperature Operational Life(MTOL) method is its success in separating different failure mechanisms in devices in such a way that actual reliability prediction scan be made for any user defined operating conditions.

This methodology is opposed to the common approach for assessing device reliability today, using High Temperature Operating Life (HTOL) testing, which is based on the assumption that just one dominant failure mechanism is acting on the device. However, it is known that multiple failure mechanisms act on the device simultaneously. The new approach, MTOL, deals with this issue, this method predicts the reliability of electronic components by combining the Failure in Time (FIT) of multiple failure mechanisms. Degradation curves are generated for the components exposed to the accelerated testing at several different temperatures and core stress voltage. The recent published data [18] clearly reveals that different failure mechanisms act on the components at different regimes of operation causing different mechanisms to dominate, depending on the stress and the particular technology. A linear matrix solution, allows the failure rate of each separate mechanism to be combined linearly to calculate the actual reliability as measured in FIT of the system based on the physics of degradation at specific operating conditions. In this paper, we present the most significant physical failure mechanisms in modern electronic devices, such as Negative Bias Temperature Instability (NBTI), Electro-migration (EM) and Hot Carrier Injection (HCI) but we will not present the MTOL analysis, due to the fact that we present the theoretical aspects of physical reliability, yet we will present the MTOL analysis on FinFET transistors that may be implemented in the near future.

2.2. The Physical Mechanism Behind BTI

Bias Temperature Instability (BTI) is a time-dependent degradation mechanism, it has been known since 1966 [7] and a model for understanding its effects was first proposed in 1977. BTI has emerged as a key reliability concern due to its increasingly negative impact on performance of modern electronic devices. BTI effects worsen as a transistor ages, and lead to severe shifts of important transistor parameters. Therefore, understanding the impacts of BTI degradation is of primary importance for current and near future CMOS technologies. The continuous MOSFET miniaturization trends (i.e., aggressive oxide thickness scaling) resulted in higher oxide fields and temperature [19]. Consequently, more charge traps are able to tunnel through the gate oxide. These traps capture some of the charged carriers which are responsible for the current flow between source and drain. Therefore, it results in the formation of a narrower transistor channel due to this charge loss. This means that less current can flow through the device and consequently, the device performance will degrade. These effects show up themselves at the circuit level by increasing circuit delays and in turn circuit timing errors. In order to maintain the drain current to its pre-degradation state, a higher voltage bias needs to be applied on the gate.

Therefore, a higher voltage will be needed before the transistor begins to conduct. This means that the threshold voltage (Vth) increases significantly over time. The threshold voltage shift (ΔVth) is accelerated by elevated temperature or supplied voltage and it is a direct measure of the device degradation and is widely used in the literature to evaluate BTI impacts [20, 21]. BTI mechanism, as illustrated below in figure 1, occurs in two phases, firstly, the transistor is in stress phase when the voltage Vgs is applied to the gate over a period of time. During this phase, charged traps are generated at the gate oxide layer and the transistor threshold voltage increases (degrades). Secondly, when the stress voltage is removed, the transistor is in recovery phase. During the recovery phase, trapped charges are released and the threshold voltage partially recovers to the level that was prior to the stress. The transistor enters into stress and recovery phases alternately, when the input is dynamic.

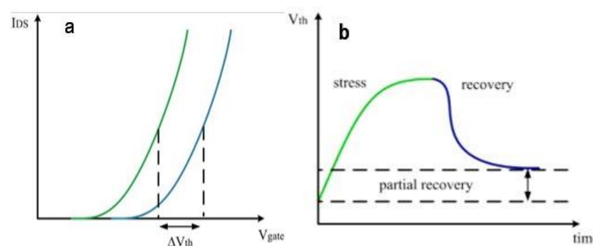

Fig. 1 (a) Threshold voltage shift due to BTI aging,
(b) Two phases of BTI. The transistor does not fully recover

Thus the amount of BTI degradation depends on the stress history that is reflected at the duty factor and calculated by the device stress and relaxation periods. Devices in arithmetic units and memory circuits tend to present an un-balanced duty factor, while devices in clock circuitry are an example of duty cycle factor of 50%. The impact of BTI is observed at both NMOS and PMOS transistors; both are susceptible to Positive Bias Temperature Instability (PBTI) and Negative Bias Temperature Instability (NBTI) respectively. Hard *et al.* [22] have analyzed the impact of BTI in four scenarios: 1. NMOS under negative gate bias, 2. NMOS under positive bias, 3. PMOS under negative bias and 4. PMOS under positive bias. The study has clearly shown that PMOS devices are more susceptible to BTI, regardless of negative or positive bias. It also proves that PMOS under negative bias is the case that presents the largest threshold voltage shift. It is unfortunate, since in digital circuits PMOS devices are negatively biased. This is the reason why BTI is often referred to as NBTI which is attributed to threshold voltage shifts in PMOS devices only. Therefore, most of published data [23, 24] are focused only to study the impact of NBTI on the circuit reliability. In the last decade, accurately modeling of BTI has become a major concern for industry. Several approaches have been proposed to understand the origin of this phenomenon and predict its impacts, while there is a good agreement on the fact that BTI is caused due to the generation of traps in oxide layer when bias is applied at the gate. In addition, the BTI degradation impact can be directly measured as a shift in threshold voltage. The acceleration factor (AF) of the NBTI is presented below in Equation 2 [18]:

$$A_{\text{f-NBTI}} = \exp\big(\beta_{\text{NBTI}}(V_{\text{stress}} - V_{\text{use}})\big)\exp\left[\frac{E_a}{K}\left(\frac{1}{T_{\text{use}}} + \frac{1}{T_{\text{stress}}}\right)\right] \tag{2}$$

where:
T_{use} – Use Temperature in Kelvin
T_{stress} – Life test stress temperature in Kelvin
K = $8.63 * 10^{-5}$ [ev/k] – Boltzmann's constant
E_a – Activation energy
V_{use} – Use voltage
V_{stress} – Life test stress Voltage
β_{NBTI} – Acceleration factor for NBTI

The next paragraph gives an overview of the developed models for NBTI mechanism at the transistor-level and the reasons behind the choice of the used model to carry out the research. Moreover, BTI impacts at the gate-level are also presented.

2.3. The Physical Mechanism Behind HCD

A physical understanding of HCD and the respective models are briefly introduced. For semiconductors in thermal equilibrium, electrons and holes continually absorb and emit acoustical phonons (low-frequency lattice vibrations), resulting in an average energy gain of zero. Such electrons have kinetic energy (E) that are normally slightly higher than that of the conduction band edge (EC) by an amount KTr (Tr is room temperature). Similarly, for holes, E is slightly less than the valence band edge (EV) by KTr. In the case of low electrical fields, the carrier's velocity is field-independent and KTr is only 0.025 eV ; which is small compared to the carriers kinetic energy corresponding to EC and EV. However, if the electrical field is very high (for example, 100 KV/cm), the carriers gain more energy than they lose by scattering. Such accelerated electrons have energies of EC + KTe, where Te is an effective temperature such that KTe >KTr. With effective temperatures (~EC /KT) of tens of thousands of degrees Kelvin, these electrons are at the very top of the Fermi distribution and are known as hot electrons.

While a MOSFET is in active operation, if the gate voltage is comparable to or lower than Vds, the inversion layer is much stronger on the source side than the drain side and the voltage drops due to the channel current is concentrated on the drain side (if Vd>Vs). The field near this side can be so high that carriers can gain enough energy between two scattering events to become hot carriers. The majority of these hot carriers simply continue toward the drain, but a small number of them gain enough energy to generate electrons and holes by impact ionization. In the n-channel MOSFET (N-MOSFET), the vast majority of the generated holes are collected by the substrate and give rise to the substrate current (Isub) and the generated electrons enhance the drain current (Id). Photon emission might also occur during the generation of the hot carrier in the drain. Some of the hot carriers with enough energy have been calculated (approximately 3.2 eV for electrons and 4.7 eV for holes [25]) and can surmount the energy barrier at the Si-SiO2 interface and be injected into the oxide, with a small gate current (Ig). Some energetic injected carriers might break some S-H or similar weak bonds in the oxide or at the Si-SiO2 interface. If the hot carriers injection last long enough, the trapped charge or generated defects will permanently modify the electric field at the Si-SiO2 interface, and hence, the electrical characteristics of the MOSFET.

Hot Carrier Injection Mechanisms

Hot Carrier Injection Mechanisms According to Takeda [26], are three main types of hot carrier injection modes:

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

CHE injection is due to the escape of “lucky” electrons from the channel, causing a significant degradation of the oxide and the Si-SiO2 interface, especially at low temperature (77 K) [27]. On the other hand, DAHC injection results in both electrons and holes gate currents due to impact ionization, giving rise to the most severe degradation around room temperature. SGHE injection is due to minority carriers from secondary impact ionization or, more likely, bremsstrahlung radiation, and becomes a problem in ultra-small metal oxide semiconductor (MOS) devices. Fowler–Nordheim tunneling and direct tunneling might also cause hot carrier injection. For deep sub-micrometer devices, it is important to attempt for the effects resulting from combinations in some or all of these injection processes.

Channel Hot Electron (CHE) Injection:

The CHE injection occurs when the gate voltage (Vg) is comparable to the drain voltage (Vd) (in N-MOSFET). The gate current (Ig) rises as Vg initially increases, reaching to a maximum peak when Vg is roughly equal to the drain-source potential Vd, and drops thereafter. There are two reasons that cause the Ig to increase. First, the inversion charges in the channel increases, so that more electrons are present for injection into the oxide. Second, the stronger influence of the vertical electric field in the oxide prevents electrons in the oxide from detrapping and drifting back into the channel. It has been reported [28] that if an n-channel MOSFET is operating at Vg = Vd the conditions would be optimum for CHE injection of “lucky electrons.” Such electrons gain sufficient energy to surmount the Si-SiO2 barrier without suffering an energy losing collision in the channel. In many cases, this gate current is responsible for device degradation as a result of carrier trapping. No gate current can be measured for Vg<Vd, since CHE injection is retarded. However, if Vd is large enough, a reduction of Vg intensifies the electric field at the drain to the point where avalanche multiplication due to impact ionization might substantially increase the supply of both hot electrons and hot holes.

Drain Avalanche Hot Carrier (DAHC) Injection:

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

Secondary Generated Hot Electron (SGHE) Injection:

Secondary impact ionization by hot holes and photo-induced generation processes have been reported as secondary minority carrier generation mechanisms. Takeda [26] experimentally demonstrated that photo-induced generation is the main physical mechanism. The temperature dependence of Isub and that of electron diffusion current, Id, were compared to each other for a device with Tox of 7 nm and Leff of 2.0μm. The experiment results imply that a photo-induced generation process, believed to be bremsstrahlung radiation, rather than secondary impact ionization, is more likely to be the origin of the SGHE.

The acceleration factor (AF) of the hot carrier injection is presented below [18] in equation 3.

$$A_{\text{f-HCI}} = \exp\left(\text{Ɣ}\left(\frac{1}{V_{use}} + \frac{1}{V_{stress}}\right)\right)\exp\left[\frac{E_a}{K}\left(\frac{1}{T_{use}} + \frac{1}{T_{stress}}\right)\right] \quad (3)$$

Acceleration factor of manufacturing- Ɣ

2.4. The Physical Mechanism Behind Electromigration

The overall EM challenges due to technology scaling come from the widening of the gap in current density limits for metal lines between the design needs and the technology capability. This issue was highlighted in the ITRS (International Technology Roadmap for Semiconductor) roadmap. The current limit needed by the circuit/chip design increases rapidly from technology node to node, while the metallization process struggles to maintain the constant current carrying capability for the metal lines without invoking major innovations. Interconnects that are embedded in interlayer dielectric material are the wire connections to supply electrical signals to these devices. Aluminum (Al) has been used as the major on-chip interconnects material. It has evolved from a single layer of Al to multiple levels of sandwiched Ti/Al-Cu/TiN metal layers. In the recent technology development, Cu and new dielectric materials have been adapted to gain better resistance-capacitance delay and reliability resistance. Due to continuing transistor scaling, interconnects are now a significant limiter and are as important as transistors in determining an integrated circuits (IC) density, performance, and reliability. Aggressive interconnect scaling has been resulted in increasing current densities and associated thermal effects, which can cause reliability problems. EM is the dominating failure mode of interconnects, it is characterized by the migration of metal atoms in a conductor through which large direct-current densities pass [29]. Although EM has been intensely studied for more than 40 years, many aspects of EM are still not well understood. This lack of understanding is caused by two related issues: the existence of many factors that influence EM and the inability to isolate the effect of these factors experimentally. These factors include grain structure, grain texture, interface structure, stresses, film composition, physics of void nucleation and growth, thermal and current density dependencies [29]. According to experimental research, current density and temperature are among the most important factors. Black [30] developed an empirical model relating the median time (t50%) of a metal line to the temperature (T) and current density (J); the model has the form as equation 4 below.

$$t_{50} = \frac{A}{J^n}\exp\left(\frac{E_a}{KT}\right) \quad (4)$$

Where: A is a material and process-dependent constant and E_a is the activation energy for the diffusion processes that dominate the temperature range of interest. The importance of current

density and temperature is shown in this equation**.** As expected, the scaling of interconnects will increase current densities and temperature, thereby greatly reducing the median time. The reliability of the IC will decrease simultaneously. To better understand the interconnect-scaling effect, physical models and statistical models must be carefully developed. A significant amount of research has been done on the physics of EM as IC technology increases device density, the interconnects that carry signals are consequently reduced in size, specifically, in height and cross section. This leads to extremely high current densities, on the order of at least 106 A/cm2 [28]. At these current densities, momentum transfer between electrons and metal atoms becomes important. The transfer, which is called the electron-wind force, results in a mass transport along the direction of electron movement. Once the metal atoms are activated by the electron wind, they are subject to the electric fields that drive the current. Since the metal atoms are positively ionized, the electric field moves them against the electron wind once they have been activated. The interplay of these two phenomena determines the direction of net mass transfer. This mass transfer manifests itself in the movement of vacancies and interstitials. The vacancies coalesce into voids or micro-cracks, and interstitials become hillocks. The voids, in turn, decrease the cross-sectional area of the circuit metallization and increase the local resistance and current density at that point in the metallization. Both the increase in local current density and in temperature increase EM effects. This positive feedback cycle can eventually lead to thermal runaway and catastrophic failure.

The acceleration factor (AF) of the EM is presented as equation 5 below [18]:

$$A_{\text{f-EM}} = \left(\frac{J_{stress}}{J_{use}}\right)^{n} \exp\left[\frac{E_a}{K}\left(\frac{1}{T_{use}} + \frac{1}{T_{stress}}\right)\right] \tag{5}$$

Where:

J_{stress} – Use current density

J_{use} – Life test stress current destiny.

2.5. The Physical Mechanism Behind Self-Heating

The basic method for calculating the total heat generation rate (power dissipated) within a lumped (semi) conducting element is to write it as the product of the current and voltage:

$$Q = I \times V \tag{6}$$

Here, the voltage drop is just that across the device, excluding its contacts. Hence, this formula must be applied with care in describing the power dissipated in a structure with relatively large contact resistance, e.g. a nanotube or molecular device. This expression will also tend to overestimate the total heat dissipated in a quasi-ballistic device, i.e. one that is only a few electrons mean free paths long. In such a device, electrons will gain energies comparable to qV but will generally not undergo enough inelastic scattering events to completely thermalize and provide this energy to the lattice (in the form of self-heating) by the time they exit. Hence, relatively hot electrons will escape through the contacts, and some portion of the $I \times V$ power will be deposited there instead [31]. In other words, the power dissipated inside the device is less than suggested above in formula (6), and the rest of power dissipation occurs in the contacts. In addition, the simple formula above gives just an estimate of the total power dissipated, not of the physical location of its peak (if any) or its make-up in terms of the emitted phonon frequencies. However, this formula is very well suited for quick, first order estimates. In the context of a semiconductor device simulator, heat generation due

to electric current flow is most often calculated with the classical drift-difusion approach. The main component of this heat generation expression is the product of the electric field E and current density J, as computed at every grid node within the simulation as Eq (7) below:

$$Q^m = J * E + (R - G)(E_g + 3k_B T) \quad (7)$$

Where: J = q*n*Ve, with n being the electron density and Ve the average electron velocity. Note the notation of Q^m (power density per unit volume, i.e. W/cm^3) versus Q in Eq. (6) (total power in Watts). The total power Q in this formulation can be recovered by integrating Eq. (7) over the device volume. The first term represents the Joule heating rate, which is usually positive (power generation) as electrons drift down the band structure slope under the influence of the electric field, and gradually lose energy through net phonon emission. It should be noted that Joule heat can also be negative (power consumption). When electrons diffuse against an energy barrier and the energy required to move up to the conduction band, the slope is extracted from the lattice through net phonon absorption [32]. The second term of the above equation is the heat generation rate due to non-radiative electrons and holes generation and recombination processes.

When an electron and a hole, both with an average energy of (3/2) KBT recombine, the excitation energy Eg +3KBT is given either directly to the lattice, or to another charge carrier (Auger transition). In the latter case, the excited particle eventually gives the energy to the lattice by phonon emission as well. Equation 7 may include other higher order terms, accounting for electron drift along a temperature gradient or across a discontinuity in the band structure, e.g. a hetero junction like in a semiconductor laser [33]. Unfortunately, this field-dependent method does not account for the microscopic non-locality of phonon emission near a strongly peaked electric field region, such as in the drain of the transistor. Although electrons gain most of their energy at the location of the electric field peak, they typically travel several mean free paths before releasing all of it to the lattice, in decrements of (at most) the optical phonon energy. In silicon transistors, for example, electrons can gain energies that are a significant fraction of an eV, while the optical phonon energy is only about 60 meV.

Assuming an electron velocity of 10^7 cm/s (the saturation velocity in silicon) and an electron-phonon scattering time around 0.05–0.10 ps in the high-field region, the electron-phonon mean free path is then on the order of 5–10 nm. The full electron energy relaxation length is therefore even longer, on the order of several inelastic mean free paths. While such a discrepancy may be neglected on length scales of microns, or even tenths of a micron, it must be taken into account when simulating heat generation rates on length scales of 10 nm, as in a future generation transistor. The highly localized electric field in such devices leads to the formation of a nanometer-sized hot spot in the drain region, that is spatially displaced (by several mean free paths) from this drift diffusion prediction. In addition, the J·E formulation of the Joule heating also does not differentiate between electron energy exchange with the various phonon modes, and does not give any spectral information regarding the types of phonons emitted. The heating rate can also be computed with the more sophisticated hydrodynamic approach, as a function of the electron temperature and an average energy relaxation time [34] equation **8**:

$$Q^m = \frac{3}{2} k_B \frac{n(T_e - T_L)}{\tau_e - L} + (R - G)\left[E_g + \frac{3}{2} k_B (T_e + T_L)\right] \quad (8)$$

Where: the holes have been assumed in thermal equilibrium with the lattice (T_L), but the electrons are described by their own temperature (T_e), energy relaxation time (τ_e−L), and

number density (n). This is the situation in which electrons are the majority current carriers, but the holes and the hole temperature can be incorporated in a similar way. Unlike the J·E method, the hydrodynamic approach has been shown to be better suited for capturing non-local transport effects near highly peaked electric field regions. However, the hydrodynamic approach suffers from simplifications inherent to using a single (averaged) carrier temperature and relaxation time, since scattering rates are strongly energy dependent. Similar to the previously described methods, this average carrier temperature-based approach also does not differentiate among electron energy exchange with the various phonon modes, and does not give information regarding the emitted frequencies of phonons. Such spectral information is important because it is well-known that the emitted phonons travel at different velocities and have widely varying contributions to heat transport and to device self-heating [35, 36].

The mechanism through which lattice self-heating occurs is that of electron scattering with phonons, and therefore only a simulation approach which deliberately incorporates all such scattering events will capture the full microscopic, detailed picture of self-heating. In FinFET, self-heating and reliability issues are more crucial as compared to other planar transistors. Self-heating issues in conventional planar transistors are controlled by moving the generated heat away down into the bulk of the device. However, in FinFET, self-heating is a significant issue due to its complex geometry and lack of heat release path. Self-heating of FinFET causes many problems like increase in temperature of the metal interconnects resulting at the enhancement of the Electro-migration effect.

3. Analysis of FinFET Reliability and Aging

3.1. Background

As tri-gate transistor technologies continue to scale to smaller dimensions, a variety of aging mechanisms become important to be included in models to accurately predict end-of-life transistor performance. Traditional aging effects such as BTI and hot carriers continue to play a major role. However, modeling these mechanisms becomes more complicated with the addition of recovery, variation, and local self-heating. Further, second-order effects are starting to accumulate, such as recovery interactions, minority carrier gate injection, damage localization, and interactions between hot carrier and BTI. This work highlights the roles and impacts of these various effects and how they will need to be fitted into a comprehensive aging model. This paragraph presents a comprehensive analysis of the roles and impacts of various effects (second-order effects are starting to accumulate, such as recovery interactions, minority carrier gate injection, damage localization, and interactions between hot carrier and BTI) and how they are needed to fit into a comprehensive aging model. Many models exist to treat BTI and hot carrier (HC) degradation, which continue to play a significant role in the total aging, as well as BTI recovery. In scaled tri-gate devices, second-order mechanisms can play an even more significant role. Here we will describe the influence of these, including: Self-heating, tri-gate/HC interactions, BTI/HC interactions, and body bias effects [37].

3.2. Detailed Theoretical Consideration

In their simplest appearance, aging models are generally assumed to be a function of stress voltage, temperature and time as described below in Equation 9:

$$\%I_D = F(V_{G_Stress}, T_{stress}, time_{stress}) \tag{9}$$

The model is calibrated with a DC stress at different stress conditions. To apply the function to a time-varying waveform, a quasi-static approximation is made so that the waveform is discretized and the aging tabulated at each time step. The total aging for the waveform is then the sum of the contributions from each time step. In this manner, the net aging is independent of frequency and simply a function of the time spent at each voltage and temperature. Though simple, this approximation has been shown to work reasonably well for modeling BTI under repeating waveforms. With the recognition of recovery effects becoming important in BTI modeling [38], additional methods were needed to accurately capture the aging. While many methods have been proposed [39], the recovery can be included as an additional term in the quasi-static approach to generally capture the behavior of repeating waveforms [40], Equation 10:

$$\%I_D = F(V_{GS}, T, time)Rec(Stress, RecCond) \tag{10}$$

Equations (9) and (10) represent aging for the drive current which is ultimately the parameter of interest, especially for digital circuit operation. However, the observed trans-conductance degradation observed for NBTI in PMOS suggests that there are both mobility and threshold voltage components to the aging [41]. These components can be built into the aging compact model to provide more accurate recovery of the device characteristics [42]. However, this complicates the drive current dependence on aging since the percent drive loss is by definition a function of the threshold voltage, Vth, and mobility. For example, starting with a simple formula for the source-drain current in the linear regime, as Equation 11:

$$\%I_{Ds_Linear} = A\mu(V_{GS} - VT) \tag{11}$$

Where: μ is mobility, and A is a constant, the percent change in current after stress simplifies to as Equation 12:

$$\%I_{Ds_Linear} = \frac{(1-\%\mu)(V_{GS}-V_T-\Delta V_T)}{(V_{GS}-V_T)} - 1 \tag{12}$$

For the sake of notational simplicity, we will generally discuss aging in terms of simply %Ids degradation, though in reality, all functions for %Ids would be more accurately conveyed as degradation functions for each of the underlying device parameters, such as Vth and mobility shifts. Modeling hot carrier degradation complicates the formula in equation 10, since HC depends not only on the gate voltage at stress, but also the drain voltage. However, neglecting layout dependences such as drawn channel length, equation 10 can be simply extended to include a drain voltage dependence to capture hot carrier degradation. This drain voltage dependence can either be an explicit term in the formula, or implicitly embedded in a term that depends on substrate current, which was a traditional method for modeling hot carrier degradation [43]. Importance in circuit modeling is not only the typical behavior for device aging, but also the variation to predict aging for a worst-case device. The so-called exponential Poisson distribution [44] has been used to model PMOS BTI variation in a wide range of technologies [45]. In this approach, BTI is considered to be due to discrete traps, each

of which contributes a certain amount to Vth shift depending on their location in the channel. This distribution of Vth shifts follows an exponential distribution, and the number of traps follows a Poisson distribution. With this formulation, the aging at any point can be predicted. In terms of the general aging model formula, including the variation necessitates the inclusion of a term that depends on these aging variation characteristics. The overall aging formulation presented so far works reasonably well for modeling single devices under controlled stress conditions. However, real circuits, experience a variety of BTI and HC conditions as the transistor switches. As a result, the aging model needs to be generalized to provide a total degradation from all mechanisms. In the most simplistic approximation, the mechanisms can be treated as completely independent, so that the aging from each component can be effectively added (though the underlying math will be more complicated with any Vth and mobility dependencies) as Equation 13:

$$\%I_D = F_{BTI}(V_{G_Stress}, \mathrm{time}, \mathrm{Rec}) + F_{HC}(V_{G,D_Sress}, T, \mathrm{time}, \mathrm{Rec}) \quad (13)$$

However, as will be discussed in the upcoming sections, these simple approximations to treat repeating waveforms, modeling the underlying physics, and combining aging mechanisms, will be challenged when applied to highly scaled tri-gate devices.

3.3. The Tri-gate Influences on Aging

Trigate devices are known to influence aging due to the three dimensional architecture of the device. For example, the sidewall crystal orientation is generally chosen to be <110> to improve transistor performance [46], but can affect the interface quality. Additionally, top corners can enhance local electric fields [47] and the vertical sidewalls require special integration schemes to ensure appropriate gate [48] and junction formation. Perhaps most critically, the fin architecture limits heat dissipation from the channel, which can lead to increased local temperatures [49].

BTI in Tri-gate

Despite the crystal orientation and corners, multiple authors have shown that BTI in tri-gate is matched to planar devices [50, 51, and 52]. For example, PMOS NBTI is matched to planar for both degradation and recovery. This indicates that the same basic modeling formulation can be extended from planar devices and applied to tri-gate. With a potential dependence on fin height and width, the BTI variation could be expected to be influenced by the vertical tri-gate architecture. However, data from Intel's 22nm technology indicate that the variability characteristics are matched between planar and tri-gate [50]. For example, the exponential distribution characteristic parameter for the average Vth shift per trap, which indicates that the behavior in tri-gate devices is identical to that of planar.

Hot Carrier in Tri-gate

Hot carrier degradation in tri-gate devices is known to be complicated by dependences on tri-gate features such as fin width, field profiles, and junction formation [53,54]. Further, the tri-gate architecture itself can increase degradation due to increased likelihood that the gate will capture energetic channel carriers. As with planar devices, the damage is expected to be localized near the drain end of the channel, where fields are the highest. Tri-gate devices can enhance this localization due to the improved control over short channel effects,

which effectively means that the gate better controls the potential in the channel, leaving more of the drain-source voltage drop at the drain end of the channel. The effect of this localization is reflected in the degradation characteristics, where, the Ids degradation is measured at different applied biases after hot carrier stress. During this monitoring, a constant high gate bias is applied and the Ids degradation is measured at different source or drain biases. As the monitor drain voltage increases from low to high, the measured degradation decreases. This can be interpreted as a reduction of interaction between the channel electrons and gate oxide damage at the drain end of the channel as the depletion region increases. Conversely, as the source voltage is increased, the measured degradation increases, which indicates that a greater fraction of the channel is contributing to the measured degradation as the source-side depletion region increases. As a result, the degradation impact to a circuit will depend on the bias conditions during operation. Therefore, the aging formula needs to be extended to include a dependence on the playback bias:

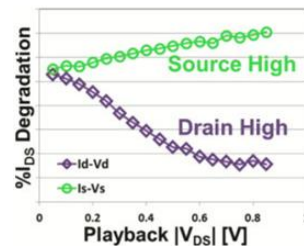

Fig. 2 The measured degradation after NMOS HC stress is observed to be a strong function of applied bias during playback. As the drain voltage increases, measured degradation decrease. Conversely, as the source voltage increases, the measured degradation increases. Both effects are due to damage localization at the drain end of the channel. The data come from Intel's 14nm tri-gate technology, and the applied gate voltage is constant for all biases.

A further complication of hot carrier degradation is the tendency for "minority carriers" to be injected into the gate, causing an increase in drive current rather than degradation. For example, a PMOS device in a stress condition of high drain voltage but low gate voltage will have fields favoring the injection of electrons into the gate. This can cause a |VT| reduction and drive current increase. A recent example of this behavior in tri-gate devices is shown in Fig. 2 where a PMOS device was stressed with drain-high and gate-low, and the drive current is seen to increase after stress [54]. As a consequence of this mechanism, a general aging model will need to include a term for this minority carrier injection as indicated in equation 14. Here, the mechanisms are assumed to be independent and therefore additive, though the actual physical picture is likely far more complicated.

$$\%I_D = F_{majority}\,(\text{Stress, Rec, Playback}) + F_{minority}(\text{Stress, Rec, Playback}) \quad (14)$$

The discrete trap behavior from hot carrier degradation is expected to have a similar influence on variation as BTI. As such, the variation can be modeled with the same variation formulation as BTI [55]. However, as discussed in Ref. [53], hot carrier degradation depends on many features of trigate device, so the variation magnitude may be affected as well. Recent evaluations suggest that this may be the case, though additional research is needed in this area [54].

Self-Heat in Tri-gate

One of the most important aging-related mechanisms in the tri-gate device is the channel temperature during operation. As discussed in Ref. [47], the channel temperature may be increased in tri-gate devices due to the reduced heat conduction to the substrate, as shown below in Fig.3 the channel temperature is known to depend on the size and layout of the transistors. As a result, the total channel temperature must be calculated as the combination of the ambient temperature plus a self-heat function of both power dissipation and transistor layout.

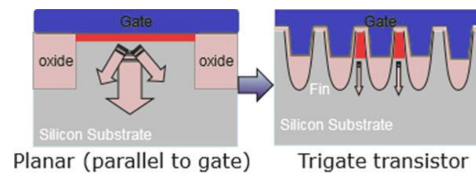

Fig. 3 Tri-gate devices have a constricted path for heat conduction from the channel to the substrate. As a result, the local temperature in the channel can increase above planar devices. [49]

A critical complication from self-heating is that the thermal time-constant is generally much longer than transistor switching times in modern technologies [54]. As a result, the channel will heat up during a transistor switching event (the only time a transistor in digital operation dissipates significant power) and then this generated self-heat will persist long after the switching event itself. As a result, the temperature at any given time will depend on the activity history of the device, and the channel temperature needs to be expressed as Equation 15 below:

$$T = f\,(\text{power, layout, activity}) \tag{15}$$

This is of critical importance since it violates the quasi-static approximation on which the entire aging model is based. For a repeating waveform that switches with some regularity, an average temperature can be obtained for the entire waveform and a net aging for the waveform can be calculated. However, if the waveform is irregular then the quasi-static approximation breaks down, and an explicit evaluation of the aging at each point in the waveform will need to be evaluated. A further modeling challenge from the channel self-heat effect is from the empirical calibration standpoint. During standard DC hot carrier stress on a device, the channel temperature can increase far beyond conditions found during normal switching operations due to the total power dissipated. As a result, evaluation of the temperature dependence for the hot carrier model must take into account both the ambient and self-heat temperature when extracting the temperature dependence. Furthermore, a particular concern is that the temperature dependence deviates from expected behavior at very high channel temperatures, which can lead to overly optimistic predictions of aging behavior at regular use conditions [55].

Body Effects in Tri-gate

One of the advantages of tri-gate devices is the excellent gate control over the channel which renders the device less susceptible to body-bias influences. For example, figures 4 and

5 show the NBTI at PMOS [50] and HC at NMOS [56] sensitivity to body bias, which only affects the aging at very high body voltages. While the behavior observed is well beyond normal operating voltages, there are some stacked configurations where devices are placed on high voltage supplies and therefore the body bias effect may need to be considered. In these cases, the aging model formulation will need to be extended to include a dependence on the body bias during stress. To accurately include this body bias effect, the underlying physics needs to be comprehended. The body bias can influence the aging either through modulating existing mechanisms by altering channel fields or by introducing a completely new mechanism such as substrate hot carrier injection. Depending on the underlying mechanism, the body bias will need to be added to the model either as a completely separate additive component or as a modulation of the existing terms. It is beyond the scope of this work to detail the exact physics involved with these effects in tri-gate devices, and for the sake of notational simplicity in the equations shown here, the body bias is assumed to simply modulate existing mechanisms.

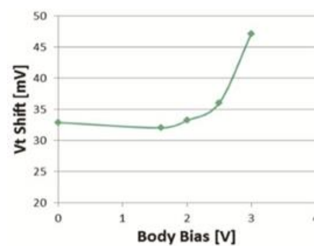

Fig. 4 PMOS BTI in tri-gate devices is generally insensitive to body-bias until biases are applied far outside normal operating voltages. [48]

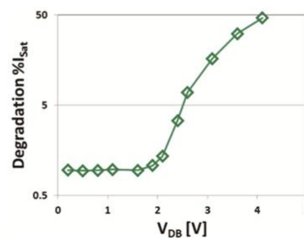

Fig. 5 NMOS HC in tri-gate devices is relatively immune to body-bias effects, but at a high voltage threshold, the degradation can increase dramatically. [55]

A feature unique to tri-gate is the possibility of traps forming in the sub-fin oxide, especially after high body-bias stress. Depending on polarity, these traps can form either a parasitic conduction path or pinch off sub-fin leakage. For simplicity, these effects can be included in the aging model as terms for either the body bias aging influence or the minority carrier injection term. However, the physics is likely more complicated and a more precise model will need to have explicit terms and interactions included.

Electromigration in Tri-gate

The overall EM challenges due to technology scaling come from the widening of the gap in current density limits for metal lines between the design needs and the technology capability. The current limit needed by the circuit/chip design increases rapidly from technology node to node, while the metallization process struggles to maintain the constant current carrying capability for the metal lines without invoking major innovations. The major driving forces for technology scaling include enhancing the chip performance and reducing the cost per device. Technology scaling includes physical scaling, material scaling, electrical scaling and new integration schemes. Physical scaling refers to dimensional shrink [57]. One obvious benefit of the physical scaling is to allow smaller devices and denser designs (more devices per chip area). This is essential to packing more functions, and more importantly to have more chips per wafer. The technology scaling has a direct undesirable consequence which is the increase of the overall process cost. To counter this cost increase for processing, packing in more functions per chip and producing more chips per wafer allows the cost per function and the cost per chip to keep decreasing, though the cost per wafer may increase from technology node to node.

Materials scaling refers to using materials which are more efficient or preferable for performance enhancement. One example from front end of line (FEOL) is to replace SiO_2 based material with materials having higher dielectric constant (K), such as HfO_2 based material for gate dielectric. In back end of line (BEOL), the latency from the interconnect RC effect has become a major contributor to the overall performance degradation. To lower the conductor resistance, the Al based metallization has been replaced with Cu based (higher electric conductivity) metallization since the 180nm technology node. Materials with lower K values have been introduced as inter- and intra- metal dielectrics (ILD) since the 130nm technology node to alleviate the interconnect capacitance effect. Electrical scaling refers to the operating voltage (V_{DD}) and power reduction.

Lowering the power at chip level has become a major desire for advanced applications. Lower V_{DD} can directly result in lower power, lower electric field, and lower current, which has major benefits for time dependent dielectric breakdown (TDDB) and EM reliability. However, due to device leakage concerns, and driving for faster speed (performance), V_{DD} has not been scaled as fast as the dimensional scaling. Integration scaling refers to the new integration schemes or innovations to enhance performance and pack more devices. This includes two very different integration aspects: (1) the interconnect fabrication/processing integration innovation driven by the scaling; and (2) chip or system level integration. The trend for the chip and system level integration scaling is growing from 2-D to 3-D schemes. The examples include adopting FinFets in FEOL[58], and chip stacking through silicon vias (TSV) [59] for BEOL and packaging. Most of these scaling aspects are not in favor of the technology reliability, they bring various new reliability challenges. However, these scaling aspects are not in favor of the technology reliability, they bring various new reliability challenges. From the circuit/chip design side, to keep the performance scaling following the so called Moore’s law, on one hand, the circuits and chips need to be smaller in size, or aggressively shrinking in dimensions both horizontally and vertically. On the other hand, the operating voltage scales at a much slower rate. The current density to flowing through a metal line may be computed as follows equation 16 as below:

$$j = \frac{CV_{DD}}{WH} fp \tag{16}$$

Where: C stands for capacitance, W and H are the metal line width and height; V_{DD} is the supply voltage to devices, "f "is the clock frequency and "p" is the device switching factor. Generally, W and H scale by a factor of 0.7, or the electric current conducting cross sectional area (W * H) reduces by about 50% for each technology node. For Cu interconnects, from 180nm node to 10nm node, the Cu cross sectional area for a minimum width metal line has reduced from 0.03 lm^2 to 0.0015 lm^2, a 95% reduction. On the other hand, the operating voltage (V_{DD}) is only scaled down from 1.8 V to 0.9 V, merely 50% reduction. The net effect is 10 times increase of (V_{DD}/ W H) ratio.

Scaling for performance also drives higher and higher clock frequency "f", and switching factor "p". All these factors point to that higher and higher "j" is needed for circuit and chip design. In addition to the physical scaling (dimensional shrink), the material scaling impact on EM reliability can also be significant. There are two aspects of this impact, direct impact from the new material properties and the indirect impact from the process integration changes driven by accommodating the new material properties .Replacing Cu with Al gave a significant boost to the interconnect EM capability [60], due to Cu’s higher melting point (1083 LC vs 660 LC of Al) and higher EM activation energy (0.9 eV for Cu vs 0.8 eV for Al). However, the subsequent aggressive dimensional shrink from technology scaling has led to rapid EM performance degradation for Cu interconnects. This is because the decrease of the critical void volume to cause EM failure and the increase of the Cu drift velocity. EM failure time may be expressed as a function of the critical void volume and Cu drift velocity [61, 62] as equation 17:

$$t_{fail} = \frac{V_{critical}}{sv_d} \tag{17}$$

Another important factor to increase the interconnect current carrying capability is to take advantage of the short length benefit. From equation 18 below:

$$v_d = \frac{D_{0,eff}}{KT} e^{-\frac{\Delta H}{KT}} \left(Z * e\rho j - \Omega \frac{\Delta\sigma}{\Delta L} \right) \tag{18}$$

As the EM process proceeds, a higher and higher backflow stress gradient ($\Omega\Delta\sigma/\Delta L$) will be built up at the anode to slow down the Cu drift rate, V_d. When this backflow stress gradient becomes sufficiently high, it can completely balance the driving force (Z *epj), and makes the net Cu mass flow to zero. At this steady state, equation 19can be written as:

$$(jL)_c = \frac{\Omega\Delta\sigma}{Z * ep} \tag{19}$$

$(jL)_c$ is called the threshold jL product [63]. In theory, if the jL in the interconnect is below the threshold product $(jL)_c$, the interconnect should not suffer EM damage. Even when the jL product is higher than $(jL)_c$ to some extent, the EM damage can occur but with longer time to fail, based on the following modified Black equation 20 [64].

$$\mathrm{MTTF} = \frac{A}{(j - j_c)} \exp\left(\frac{\Delta H}{KT}\right) \tag{20}$$

$$j_{max} = j_{stress} \left(\frac{t_{50stress}}{t_{lifetime}} \right)^{\frac{1}{n}} e^{\left[\frac{Z\sigma}{n} + \frac{\Delta H}{nk} \left(\frac{1}{T_{use}} + \frac{1}{T_{stress}} \right) \right]}$$

Where: MTTF is the median time to fail (i.e. t_{50}), A is geometry and material related constant, j_c is the threshold current density. All the mentioned equations serve as the basis for the fact that shorter lines can have higher maximum allowed current densities. Taking advantage of

this feature, i.e. making short interconnections, has been proven to be powerful in circuit design to solve some of the EM challenges. Another advantage of utilizing short length benefit for EM is the low temperature sensitivity. If some devices are known to have high power, high frequency and high activity factors, the local interconnect temperature has a potential to be much higher than the nominal junction temperature due to joule heating. As shown in equation 20, the maximum allowed current density decreases exponentially with the interconnect temperature for the regular EM process. A severe j_{max} de-rating may be needed for such circuits to account for the local temperature rise. Using wider metal lines, which only increase the current linearly with line width, may not provide sufficient relief to compensate for the j_{max} de-rating. Under such circumstances, taking advantage of the short length benefit becomes essential to overcome the local high temperature issues. Since $(jL)_c$ has very low sensitivity to temperature [65], as long as the jL is sufficiently below $(jL)_c$, the EM reliability will not decrease much with the local joule heating.

Breaking the long interconnect into short segments to take advantage of the short length benefit may not always be feasible due to spacing limitations and resistance sensitivity. To allow the backflow stress to build, physical barriers at both cathode and anode of the interconnect is required. There are alternative ways to establish some pseudo-barriers to enable local backflow stress build up, and form some short length effects [66, 67]. One example is to have blocking islands on top of a metal line or using multiple levels of metal lines with period of vias or bar vias connecting them [68].Those blocks on the Cu surface may not create a complete physical barrier for Cu diffusion, since they have a direct metal to metal interface in the blocking islands, the Cu diffusion along those local interface areas will be much slower, and some degree of backflow stress gradient will be built up to create partial short length effects. One of the major challenges for short length EM benefit applications is the line length definition [66]. Actual circuits are often more complicated than the simple metal line segment with vias at each end. They often have fingers, branches/wings, passive reservoirs, passing vias, dropdowns and width transitions. In such cases, the line length and $(jL)_c$ to be used for the short length benefit calculations become very challenging, not only to the EDA tools, but also for the design engineers. Appropriate engineering judgments are often sought to solve these issues. Due to the variability control and the liner thinning, lower $(jL)_c$ values are expected for the future technologies. Furthermore, the distribution complexity should be closely watched as well when applying short length benefits for circuit designs.

4. Comparative Discussion

The discussion so far was focused on the details of each independent mechanism and its influence on aging. In real circuits, devices will experience a variety of biases, temperatures, activity, and aging conditions, as illustrated in Fig.6 below. In this type of generalized waveform, interactions between the aging mechanisms become important. One of the more challenging modeling complications from these interactions are the recovery effects. For example, BTI recovery is more sensitive to the applied bias. Therefore, as a circuit switches to a low voltage state either in analog operation or from dynamic voltage power management schemes, the transistors will undergo recovery differently.

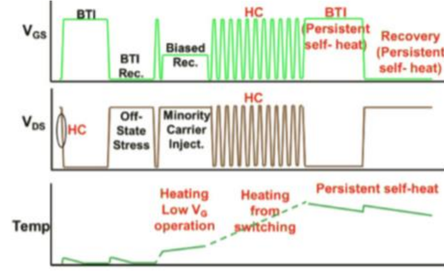

Fig. 6 Devices in circuits may see a wide variety of biases with associated aging mechanisms. The persistent self-heat effect makes the aging sensitive to history, modulating the aging and recovery behavior in regions of the waveform after sections of high self-heat generation.

Additionally, recovery is known to be sensitive to temperature, with more recovery occurring at higher temperatures [68]. Modeling this effect is further complicated by the persistent self-heat behavior, so that within a general waveform the temperature during a recovery phase will depend on the immediate history of the waveform. One final challenge for recovery modeling is the interaction with minority carrier injection during hot carrier stress. These minority carriers can passivate BTI traps enhancing recovery. For example, in PMOS NBTI a hole trap may be passivated by an injected electron during a subsequent hot carrier stress phase. Comprehending these recovery interactions in a model requires that the recovery term will be updated to include a dependence on bias, temperature, activity, and minority carrier injection. Putting all of these effects into a single equation depends on the physics involved, but a simplistic additive form is summarized below in 21:

$$\begin{aligned}
\%I_D = {} & F_{BTI}\begin{pmatrix} V_{G,B_Stress}, \\ T(pwr, activity, layot), \\ Rec\left(V_{G,D,B}, F_{minority}, activity, T\right), \\ V_{D,G,B_playback}, \sigma \end{pmatrix} \\
& + F_{HC_{majority}}\begin{pmatrix} V_{D,G,B_{Stress}}, \\ T(pwr, activity, layot), \\ Rec\left(V_{G,D,B}, F_{minority}, activity, T\right), \\ V_{D,G,B_playback}, \sigma \end{pmatrix} \\
& + F_{HC_minority}\begin{pmatrix} V_{D,G,B_Stress}, \\ T(pwr, activity, layot), \\ Rec\left(V_{G,D,B}, activity, T\right), \\ V_{D,G,B_playback}, \sigma \end{pmatrix}
\end{aligned} \tag{21}$$

In this summarizing equation, a key feature is still missing, which is any interaction between BTI and hot carriers. This is an area that still requires significant research, but the basic issue hinges on whether the traps associated with BTI and HC are independent or have some interaction. For simplicity, the traps are often assumed to be separate and independent based on their location. However, recent work has shown that there is a strong interaction between the mechanisms in some cases, so that the independent approximation may lead to overly conservative predictions of total aging [69]. This interaction poses further challenges for aging since there may be sequence dependence to the total aging, as indicated in Fig. 7

below [69]. As a result, the net aging after a general waveform will not necessarily match a quasistatic integration of the waveform, but rather a more detailed model is needed for the recovery of the degradation at each point in time.

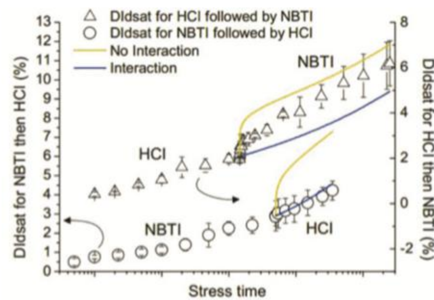

Fig. 7 Aging after BTI and HC is observed to depend on the sequence of stresses, indicating both an interaction between the mechanisms as well as a history dependence. [69]

The solutions to the EM challenges due to technology scaling have relied on various innovations from all aspects, including process development, circuit design and chip/system integrations. Though these innovations have been proven successfully, they have been projected more and more difficulties for the future technologies. From process development point of view, any innovative schemes to enhance the EM performance will have to overcome the challenges of line/via electrical resistance, Cu grain growth and variability. The rapid resistance increase of the interconnect has become one of the bottle necks and diminished the performance gain from technology scaling. For the advanced Cu interconnect, historically, all process integration schemes to boost EM reliability came with a certain degree of sacrifice of the electrical resistance. To slow down the resistance increase trend with scaling, new integration measures are needed not only to minimize the Cu resistivity deterioration, but also to maximize the Cu volume fraction in the trenches and vias. While the former metal technologies faces the challenges of the fundamental physics (size effect from electron diffraction), the later has significant potential implications with reliability and manufacturability. Certain liner thickness in the trenches and via has been proven to be critical for good Cu fill and slow Cu mass flow along the sidewalls. A new liner deposition process is needed to overcome these challenges. Though technological solutions can and will be developed to meet these challenges discussed above, the real potential barrier ahead of the technology scaling could be the economics, i.e. whether these technology solutions can still provide viable economic benefits. Rather than developing costly technology solutions to cover ''universal'' applications, an approach to tailor the technology for specifically targeted applications and reliability may have to be adopted. Therefore, a co-optimization of process development along with circuit and chip design becomes essential .Circuit and chip designers will need to actively participate in the technology definition and process window evaluations. On one hand, the circuit and chip design teams need to understand the process capability and take advantage of the process strength and avoid the process weakness. On the other hand, the process development team knows what the critical needs are from circuit and chip designs and optimizes the process windows around those critical constructs of circuit and chip designs.

5. Conclusions and Future Work

Conclusions

In scaled tri-gate devices, BTI, recovery, hot carrier and EM continue to play a significant role in aging. Modeling these effects can be accomplished with a similar modeling framework as methods established for planar devices. However, in addition to these primary aging

mechanisms additional phenomenon need to be included in any aging model general enough to capture the arbitrary waveforms found in real circuit use. Mechanisms such as self-heat, minority carrier injection, body bias effects, BTI/HC interactions, and recovery dependence on bias, temperature and activity all need to be included in the model. In the most limiting cases, these effects will violate the fundamental quasistatic approximation for many simple aging models, which will require restructuring the basic model formulation. Technology scaling results in severe EM challenges to the advanced interconnect. To balance the EM reliability with performance, cost and cycle time, joint efforts are needed from all aspects, including process development, manufacturing, circuit designs and chip integration. While innovative process integration schemes are essential to enhance the interconnect current carrying capabilities, the robust circuit design and chip level budgeting are also important to make a product with high EM reliability and optimized performance. NBTI and hot carrier and EM and aging are the most critical challenges for the future of semiconductor industry. They all affects the overall reliability of nano-scaled circuits and potentially causes system failures. Being able to predict all the mentioned mechanisms degradations crucial for the development and long-term success of novel transistor structures such as the FinFETs.

Yet to be explored are:

1. FinFETs evaluation of the performance degradation on levels of abstraction such as processor-level or system-level. For that, new methodologies are required for predicting aging behavior such as Dynamic Reliability Management (DRM) techniques for new FinFETs devices (7nm and below).
2. Studying of mitigation techniques for NBTI aging effects for FinFETs. Mitigation designates any method of limiting or controlling BTI and its impacts on innovative electronic devices. Currently, a number of research projects are conducted to develop fully automated schemes that are able to eliminate BTI effects with little or no trade-offs in terms of performance, power consumption or costs.
3. Extending the reliability-aware digital flow to cover Statistical Static Timing Analysis (SSTA) to improve the accuracy of aging predictions. In reality, the distributions of logic gates in a circuit are correlated depending on their statistical properties. To obtain the information of path timing degradation, statistical timing analysis techniques to handle this correlation should be incorporated.

Acknowledgement: *Sponsored by US Dept. of Defense (ONR and AFOSR).*

REFERENCES

[1] FinFET Modeling for IC Simulation and Design: Using the BSIM-CMG Standard 114–117, April 19, 2001.
[2] S.-H. Oh, D. Monroe, J.M. Hergenrother, “Analytic description of short-channel effects in fully-depleted double-gate and cylindrical, surrounding-gate MOSFETs,”*IEEE Electron Device Letters,* vol. 21, no. 9, 445-447, 2000.
[3] G. E. Moore, “Cramming more components onto integrated circuits,” *Proc. of Electronics*, vol. 38, 114–117, April 19, 1965.
[4] S. Thompson, P. Packan, and M. Bohr, “MOS scaling: transistor challenges for the 21st century,” *Intel Technology Journal*, vol. 2, pp. 1–19, 1998.
[5] K. J. Kuhn, “CMOS scaling for the 22nm node and beyond: Device physics andtechnology,” In Proceedings of the International Symposium on VLSI Technology, Apr. 2011, pp. 1–2.
[6] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J.Pearson, and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime andbeyond," *IBM Journal of Research and Development*, vol. 50, pp. 433–449, Jul-Sep 2006.

[7] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deepsubmicron silicon semiconductor manufacturing," *Journal of Applied Physics*, vol. 94, pp. 1–18, Jul. 2003.
[8] X. Wang, B. Cheng, A. R. Brown, C. Millar, J. B. Kuang, S. Nassif, and A. Asenov, "Statisticalvariability and reliability in nanoscale finfets," in Proceedings of the IEEE Int. Electron Devices Meeting (IEDM), 1–4, 2011.
[9] B. Kaczer, T. Grasser, P. J. Roussel, J. Franco, R. Degraeve, L. Ragnarsson, E. Simoen, G. Groeseneken, and H. Reisinger, "Origin of NBTI variability in deeply scaled pFETs," in Proc. of the IEEE IRPS, 2010, pp. 26–32.
[10] P. Woerlee, P. Damink, M. van Dort, C. Juffermans et al., "The impact of scaling on hot-carrier degradation and supply voltage of deep-submicron NMOS transistors," in Proceedings of the IEEE Int. ElectronDevices Meeting (IEDM), 1991, pp. 537–540.
[11] Y. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon, "Prediction of logicproduct failure due to thin-gate oxide breakdown," in Proceedings of the IEEE IRPS, 2006, pp. 18–28.
[12] The International Technology Roadmap for Semiconductors (ITRS), 2009. http://public.itrs.net
[13] V. Huard, F. Cacho, Y. Mamy Randriamihaja, and A. Bravaix, "From defects creation to circuitreliability—A bottom-up approach," *Microelectron. Eng.*, vol. 88, no. 7, pp. 1396–1407, Jul.2011.
[14] V. Huard V, F. Cacho, Y. Mamy Randriamihaja, A. Bravaix, "From defects creation to circuit reliability – a bottom-up approach," *Microelectro. Eng.*, vol. 88, pp. 1396-1407, 2011.
[15] JEDEC Publication. Failure mechanisms and models for semiconductor devices, JEP-122G, October 2011.
[16] Joseph B Bernstein, Gurfinkel Moshe, Li Xiaojun, Walters Jörg, Shapira Yoram, Talmor Michael, "Electronic circuit reliability modeling," *Microelectron Reliab*, vol. 46, pp. 1957–1979, 2006.
[17] RF Drenick "Mathematical Aspects of the reliability problem," *J Soc Ind Appl Math*, vol. 8, pp. 125–149, 1960.
[18] "Reliability prediction with MTOL" by Joseph B. Bernstein, Alain Bensoussan, Emmanuel Bender
[19] Y. Miura and Y. Matukura, "Investigation of silicon-silicon dioxide interface using MOSstructure," *Japanese Journal of Applied Physics*, vol. 5, pp. 180, 1966
[20] S. Khan, S. Hamdioui, H. Kükner, P. Raghavan, and F. Catthoor, "BTI impact on logical gates innano-scale cmos technology," in Proceedings of the IEEE 15thInternational Symposium on Design and Diagnosticsof Electronic Circuits Systems (DDECS), 2012, pp. 348–353.
[21] H. Kükner, P. Weckx, P. Raghavan, B. Kaczer, F. Catthoor, L. Van Der Perre, R. Lauwereins,and G. Groeseneken, "Impact of duty factor, stress stimuli, and gate drive strength on gate delaydegradation with an atomistic trap-based BTI model," in Proceedings of the 15thEuromicro Conf. on DSD, 2012, pp. 1–7.
[22] V. Huard, M. Dennis and C. Parthasarathy, "NBTI degradation: From physical mechanisms to modelling," *Microelectronics Reliability*, vol. 46, pp. 1–23, 2006.
[23] T. Grasser, B. Kaczer, W. Goes, T. Aichinger, P. Hehenberger, and M. Nelhiebel, "A two-stage model for negative bias temperature instability," in Proceedings of the IEEE IRPS, 2009, pp. 33–44, 2009.
[24] W. Wang, S. Yang, S. Bhardwaj, S. Vrudhula, T. Liu, and Y. Cao, "The impact of NBTI effect on combinational circuit: Modeling, simulation, and analysis," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems, vol. 18, no. 2, pp. 173–183, 2010.*
[25] Acovic, G. L. Rosa, and Y.-C. Sun, "A review of hot carrier deration mechanisms in MOSFETs," *Microelectronics Reliability*, vol. 36, pp. 845–869, 1996.
[26] E. Takeda, C. Y. Yang, and A. Miura-Hamada, Hot-Carrier Effects in MOS Devices, ch. 2, pp. 49–58. Academic Press, 1995.
[27] M. Song, K. P. MacWilliams, and J. C. S. Woo, "Comparison of NMOS and PMOS hot carrier effects from 300 to 77 k," *IEEE Transactions on Electron Devices*, vol. 44, pp. 268–276, 1997.
[28] M. Ohring, Reliability and Failure of Electronic Materials and Devices, ch. 5, p. 259. Academic Press, 1998.
[29] D. G. Pierce and P. G. Brusius, "Electromigration: A review," *Microelectron Reliability*, vol. 37, pp. 1053–1072, 1997.
[30] J. R. Black, "Mass transport of aluminum by moment exchange with conducting electrons," in Proceedings of the 6th Annual International Reliability Physics Symposium, pp. 148–159, 1967.
[31] R. Lake and S. Datta, "Energy balance and heat exchange in mesoscopic systems," *Phys. Rev. B*, vol. 46, no. 8, pp. 4757–4763, 1992
[32] U. Lindefelt, "Heat generation in semiconductor devices," *J. Appl. Phys.*, vol. 75, no. 2, pp. 942–957, 1994.
[33] J. Lai and A. Majumdar, "Concurrent thermal and electrical modeling of submicrometer silicon devices," *J. Appl. Phys.*, vol. 79, no. 9, pp. 7353–7361, 1996.
[34] M. Artaki and P. J. Price, "Hot phonon effects in silicon field-effect transistors," *J. Appl. Phys.*, vol. 65, no. 3, pp. 1317–1320, 1989.
[35] P. Lugli and S. M. Goodnick, "Nonequilibrium longitudinal-optical phonon effects in GaAs-AlGaAs quantum wells," *Phys. Rev. Lett.*, vol. 59, no. 6, pp. 716–719, 1987.
[36] S. Ramey et al, "Frequency and recovery effects in High-k BTI Degradation," IRPS 2009. pp. 1023-1027.
[37] S. Ramey, Y. Lu, I. Meric, S. Mudanai, S. Novak, C. Prasad, J. Hicks. "Aging model challenges in deeply scaled tri-gate technologies", In Proceedings of the IEEE International Reliability Workshop (IIRW2015), 2015, pp. 56-62.

[38] T. Grasser et al, "The Universality Of NBTI Relaxation and its Implications For Modeling And Characterization," IRPS 2007, pp. 268-280.
[39] S. Pae, et al, "Reliability Characterization of 32nm High-K and Metal Gate Logic Transistor Technology," IRPS 2010. pp. 3D2.1-3D2.6
[40] A. Krishnan, et al., "NBTI Impact on Transistor & Circuit: Models, Mechanisms, & Scaling Effects," In Proceedings of the IEDM 2003. pp. 14.5.1-14.5.4
[41] C. Hu, "Lucky-electron model of channel hot electron emission," In Proc. of the IEDM, 1979. pp. 22-25.
[42] B. Kaczer, et al., "Origin of NBTI Variability in Deeply Scaled pFETs," In Proceedings of the IRPS 2010, pp. 2A3.1-2A3.7.
[43] C. Prasad, et al., "Bias temperature instability variation on SiON/Poly, HK/MG and trigate architectures," In Proceedings of the IRPS 2014, pp. 6A.5.1-6A.5.7.
[44] P. Packan, et al, "High Performance Hi-K + Metal Gate Strain Enhanced Transistors on (110) Silicon," In Proceedings of the IEDM 2008. pp.1-4.
[45] G. Groeseneken, et al, "Reliability issues in MUGFET Nanodevices," In Proc of IRPS 2008. pp.52-60.
[46] J. Kim, et al, "Effects of Gate Process on NBTI Characteristics of TiN Gate FinFET," In Proceedings of the IRPS 2012. pp.GD6.1-GD6.4.
[47] C. Prasad, et al., "Self-heat reliability considerations on Intel's 22nm Tri-Gate technology," In Proceedings of the IRPS 2013. Pp.5D.1.1-5D.1.5.
[48] S. Ramey et al, "Intrinsic Transistor Reliability Improvements from 22nm Tri-Gate Technology," In Proceedings of the IRPS 2013. p.4C.5.1.-4C.5.5.
[49] K.T. Lee, et al , "Technology scaling on High-K & Metal-Gate FinFET BTI reliability," In Proceedings of the IRPS 2013. pp. 2D.1.1 - 2D.1.4.
[50] C.C. Wu, et al, "High performance 22/20nm FinFET CMOS devices with advanced high-K/metal gate scheme," In Proceedings of the IEDM 2010. pp. 27.1.127.1.4.
[51] S. Ramey, et al., "Transistor reliability variation correlation to threshold voltage," In Proceedings of the IRPS 2015. pp. 3B2.1-3B2.6.
[52] M. Cho, et al, "Off-state stress degradation mechanism on advanced pMOSFETs," In Proceedings of the ICICDT 2015. pp.1-4.
[53] B. Kaczer, et al., "Origins and Implications of Increased Channel Hot Carrier Variability in nFinFETs," In Proceedings of the IRPS 2015. pp. 3B.5.1 - 3B.5.6.
[54] C. Xu, et al, "Analytical Thermal Model for Self-Heating in Advanced FinFET Devices With Implications for Design and Reliability," IEEE
[55] International Technology Roadmap for Seminconductors. <www.itrs.net>.
[56] D. Hisamoto, "Multi-gate FETs," In Proceedings of the IEEE int electron dev meet (IEDM). Short course; 2003.
[57] Baozhen Li, Cathryn Christiansen, Dinesh Badami, Chih-Chao Yang. "Electromigration challenges for advanced on-chip Cu interconnects", *Microelectronics Reliability*, vol. 54, no. 4, pp. 712-724, 2014.
[58] D. Edelstein D et al, "Full copper wiring in a Sub-0.25 pm CMOS ULSI technology," Technical Digest. In: IEEE int electr dev meeting, 1997. p. 773–6.
[59] IA Blech, "Electromigration in thin aluminum films on titanium nitride," *J Appl Phys,* vol. 47, pp.1203–1208, 1976.
[60] Li B et al, "Threshold electromigration failure time and its statistics for Cu interconnects," *J Appl Phys*, vol. 100, pp. 114516, 2006.
[61] C-K Hu et al, "Impact of Cu microstructure on electromigration reliability," In Proceedings of the IEEE intern interconnect tech. conf (IITC); 2007 [Section 6.1].
[62] JJ. Clement, "Electromigration modeling for integrated circuit interconnect reliability analysis," *Trans Dev Mater Rel*, vol. 1, pp. 33–42, 2001.
[63] C. Christiansen, B. Li, J. Gill, "Blech effect and lifetime projection for Cu/low-K interconnects," In Proceedings IEEE intern interconnect tech. conf. (IITC), 2008. p. 114–6.
[64] B. Li B et al, "Short line electromigration characteristics and their applications for circuit design," In Proceedings of the IEEE int rel phys symp (IRPS), 2013, 3F2.
[65] C-K Hu et al, "Electromigration challenges for nanoscale Cu wiring," In Proceedings of the AIP Conf 2009, 1143:3–11.
[66] S. Ramey, et al., "BTI Recovery in 22nm tri-gate technology," In Proceedings of the IRPS 2014, pp. XT2.1-XT2-6.
[67] F. Cacho et al, "HCI/BTI coupled model: The path for accurate and predictive reliability simulations," In Proceedings of the IRPS 2014. pp.5D4.1-5D4.5.
[68] M. Song, K. P. MacWilliams, and J. C. S. Woo, "EM reliability" *IEEE Transactions on Electron Devices*, vol. 44, pp. 268–276, 1997.