SAND2020-3531R

Advanced CMOS Reliability Update: Sub 20nm FinFET Assessment

Jeremy Walraven, Michael King, Robert Kaplar, Gad Haase and Thomas Grzybowski

March 2020

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Contents

LIST OF FIGURES

LIST OF TABLES

No table of figures entries found.

EXECUTIVE SUMMARY

Reliability testing and characterization of advanced CMOS technology (FinFETs) have demonstrated improvements to Bias Temperature Instability (BTI), Time Dependent Dielectric Breakdown (TDDB), and Hot Carrier Injection (HCI) over planar transistors. Improvements in these reliability issues are attributed to continued improvements to silicon fin and sub-fin dielectric processing. High-k (HK) dielectric materials with metal gate (MG) stacks have been optimized in the FinFET process to reduce the density of interface states at the silicon surface.

Heat dissipation or self-heating in FinFETs is becoming a major reliability concern. Planar and FinFET transistors dissipate heat through the substrate. FinFETs have less substrate contact for heat dissipation than planar transistors, impacting BTI, TDDB and HCI. In nMOS, PBTI charge defects and impacts threshold voltage (V_{th}), not mobility. Improvements in PBTI have been observed due to gate optimization and oxide thickness scaling. In pMOS, NBTI degradation is higher in (110) FinFETs vs. (100) FinFETs. A slight regression has been observed in pMOS, but overall FinFET BTI is improved in comparison to planar technologies.

In FinFETs, HCI degradation is critically dependent upon fin geometry. Sidewall flatness and channel doping have considerable impact, making FinFETs very sensitive to processing parameters. Narrow fin widths increase HCI resistance, but (110) FinFETs are more susceptible to HCI degradation compared to (100) FinFETs due to higher interface trap density formation. nMOS FinFETs are more susceptible to HCI due to their sensitivity to gate voltage.

Total ionizing dose (TID) testing of 14-nm FinFET devices has shown minimal radiation-induced threshold voltage shift. Off-state leakage current of nFETs exhibit a strong gate bias dependence, indicating that electrostatic gate control of the sub-fin region and the corresponding parasitic conduction path are the largest concerns for radiation hardness. FinFETs with the best irradiation performance have high V_{th} while those with low V_{th} exhibit more change in the off-state leakage current. In 14-nm FinFET technology, modeling indicates that devices with high channel stop doping show the most robust response to TID. Single- and multi-cell error robustness has improved with the transition to and scaling of FinFET devices, leading to overall better soft-error rates in terrestrial environments. Circuit design and geometry significantly impact the probability of these errors occurring.

The implication of these transistor-level reliability mechanisms must be accommodated in the device level design and layout to provide robustness to foundry process and use environment variations. Aging and Reliability Aware design methodologies have been developed to support this design optimization but adds to the complexity of the design cycle.

1. INTRODUCTION

Advances in FinFET design and fabrication enable manufacturing of denser, more compact integrated circuits (ICs) with substantially reduced leakage while shortening the channel-lengths. The same stress-induced leakage and breakdown degradation mechanisms that affect planar transistors also impact FinFET devices. Reliability concerns such as Bias Temperature Instability (BTI), Time Dependent Dielectric Breakdown (TDDB), and Hot Carrier Injection (HCI) become very important with changes to transistor geometry and fin sidewall crystal orientation. Recent testing has shown that FinFETs respond differently to radiation (radiation effects such as total ionizing dose) when compared to planar transistors. These reliability and radiation effects issues become very important when changing transistor geometry and scaling FinFETs towards smaller feature sizes (22-nm, 16-nm, 14-nm, 10-nm, and smaller critical dimensions). The comparable 2019 state of the art transistor densities in current high-volume manufacturing silicon-based foundries is 7-nm (TSMC, Samsung) and 10-nm (Intel) [www.anandtech.com,fuse.wikichip.org]. Released products include supporting components for the cellphone and commercial microprocessor markets respectively. Extensive development in the foundry industry is driving to a 5-nm technology node in late 2020.

Any electrical reliability study attempts to determine the useful operating lifetime of an electrical element and to employ methods intended to verify the expected operating lifetime. The purpose of such a reliability study is to ensure that the lifetime of a system is longer than the intended application lifetime, and that the failure rate during an applications-normal operating lifetime is acceptably small. Applications of CMOS technology vary from commercial parts for consumer electronics to high-reliability applications such as medical devices and aerospace. The reliability concerns specific to each application vary greatly, making reliability analysis of CMOS technology particularly challenging. The semiconductor industry has continued to scale CMOS processes to smaller feature sizes and reduced operating voltages in pursuit of performance and density goals. Continuing decreases in device dimensions and increasing electric fields have led to an increase in the number of physical processes that can critically impact device performance and operating lifetime.

The primary goal of this study is to highlight the appropriate physics, stress, analysis, and modeling techniques required to determine the time to failure for the major reliability concerns in advanced CMOS technology nodes, while emphasizing their impact at the device and circuit levels. The range of technology nodes considered in this study will focus on CMOS reliability mechanisms for the 20-nm planar and next-generation FinFETs with feature sizes smaller than 14-nm. Particular emphasis will be placed on the trade-offs in performance and reliability at highly scaled technology nodes. Questions to ask manufacturers, as well as the context for those questions, are included in this updated document. In addition, this updated document focuses on fabrication and manufacturing scaling with Moore's Law, pertinent reliability issues such as HCI, BTI and TDDB, soft error studies, total ionizing dose effects, ESD and 3D packaging are also discussed. Back-End-Of-Line (BEOL) material presented in this document remains the same as in the previous version, but this should be considered in subsequent updates of this document considering the growing emphasis on system-in-package solutions for complex electrical components.

2. TRANSISTOR COMPARISON AND DEFINITIONS

Transistors fabricated using planar technologies include 32nm, 45nm, 65nm, 90nm and larger nodes. In planar transistors, a conducting channel is formed under a gate dielectric. The on/off state of the transistor is defined by the conductivity of the channel. In addition, planar FETs are optimized to (100) silicon surfaces stemming from fabrication on (100) grown silicon wafers. FinFET technology is fabricated in a 3-dimensional structure where 3 sides of a vertical fin comprise the conducting channel. Differences in the physical construction of planar and FinFET technology are shown in Figure 1. The channel orientation of the fin impacts the sidewall crystal plane. A fin fabricated on (100) silicon crystal planes will behave differently than a fin fabricated on a (110) silicon crystal plane. FinFETs fabricated on (110) silicon exhibit higher hole mobility and reduced electron mobility, improving the drive current ratio between pFETs and nFETs. In addition, (110) silicon oxidizes more quickly than does (100) silicon and exhibits lower leakage current. The Si/SiO_2 interface of (110) silicon has a higher trap density than does this interface with (100) silicon, which impacts charge build-up and possible transistor cross-talk.

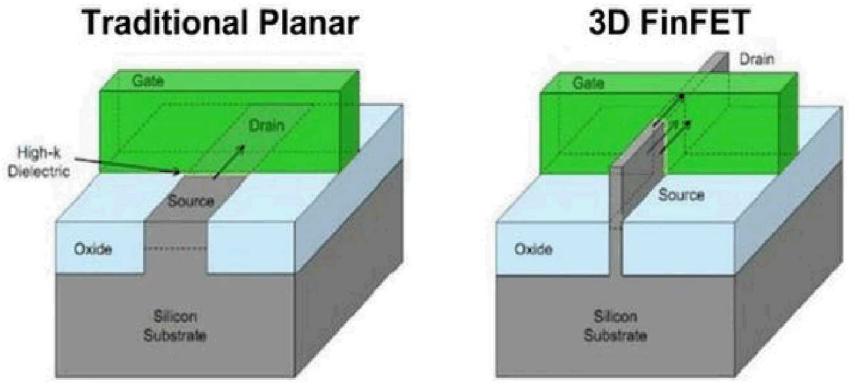


Figure 1. Schematic representations of a planar transistor (left) compared to a FinFET (right).

During the fabrication of planar and FinFET devices, materials such as high-k gate dielectrics, low-k dielectrics and metal gate electrodes are used. High-k dielectric materials are materials whose relative dielectric constant (k) is higher than that of silicon dioxide (SiO_2); such materials are often used to improve current leakage while achieving the same capacitance as SiO_2. Low-k dielectric materials are materials whose dielectric constant is lower than that of SiO_2 and are used to reduce parasitic capacitance, produce higher switching speeds and lower heat dissipation. These materials are used as a replacement for inter-layer dielectrics (ILDs), reducing charge build-up and cross-talk.

Questions to ask FinFET foundries/suppliers include: *What channel orientation is used on their FinFETs? If the vendor uses (110) orientation, is their process optimized to reduce trap densities? In addition, does the vendor use a high-k/metal gate?*

3. RELIABILITY CONCERNS FOR ADVANCED CMOS TECHNOLOGY

This section will highlight the major intrinsic reliability concerns for advanced CMOS technology. The topics considered here are by no means exhaustive but represent the primary concerns for reliability in advanced technology nodes. These topics include Bias Temperature Instability (BTI), dielectric integrity and breakdown (Time Dependent Dielectric Breakdown or TDDB), hot-carrier effects (Hot-Carrier Injection, or HCI), metallization degradation and electromigration, and stress voiding. Discussion of the recent application of 3D packaging methods and expansion of FinFET Self heating were also added due to required design/foundry optimization. Additional topics include Electrostatic Discharge (ESD), Radiation-induced Soft Errors (SER), Single Event Upsets (SEU), Multi-Cell Upsets (MCU), and Total Ionizing Dose (TID) effects.

3.1. Bias Temperature Instability (BTI)

BTI is a phenomenon that results in the generation of interface states at the Si-SiO_2 interface and the accumulation of positively charged traps. Particularly vulnerable are pMOSFETs, which are affected heavily by Negative Bias Temperature Instability (NBTI). NBTI has been shown to adversely impact key device operating characteristics such as threshold voltage (V_{th}), drain current (I_d) drive in the linear and saturation regions, and transconductance (g_m). Consequently, NBTI has become the most critical MOSFET reliability concern for circuit design in sub-45nm generation technology nodes. A primary reason for this reliability concern is self-heating of FinFET transistors during operation. While planar transistors dissipate heat through the substrate silicon, FinFETs have much less thermal contact area and consequently dissipate less heat through the substrate. For devices requiring high current, circuits are designed that contain multiple parallel fins. An example of multiple parallel fins is shown in Figure 2 where interior fins are unable to dissipate heat as effectively as those around the perimeter of the array. In these designs, self-heating of the interior fins becomes a critical issue, particularly for high-current devices [1]–[3]. This self-heating may impact the BTI of interior fins, causing irreversible damage. Modification to the doping profiles in the FinFET channel can reduce the electric field, reducing the positive BTI but at the risk of increasing the negative BTI.

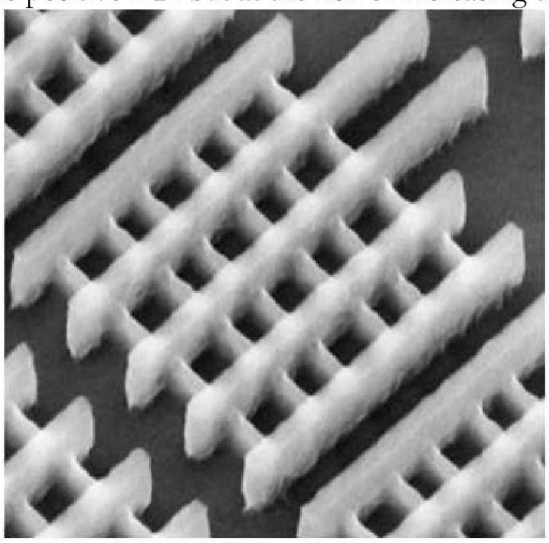


Figure 2. 30 fin FinFET cell.

The increase in pMOS sensitivity in highly-scaled devices is the result of increased electric fields at the Si-SiO_2 interface, nitridation of the gate dielectric material, and reduced power supply, which all

increase the NBTI damage to the device and impact circuit level parameters. The difficulty in evaluating NBTI in the laboratory involves the sensitivity of NBTI to subtle differences in CMOS processes, requiring careful attention to differences in the gate dielectric stack, front-end-of-line (FEOL) and back-end-of-line (BEOL) processes. The ideal stress conditions would involve relatively low electric fields, between 4-7 MV/cm, at elevated temperature to minimize concurrent degradation mechanisms contributing to parametric shifts in the device. All these complications lead to difficulty in defining a general framework for exploring the parametric degradation of pMOSFETs due to NBTI.

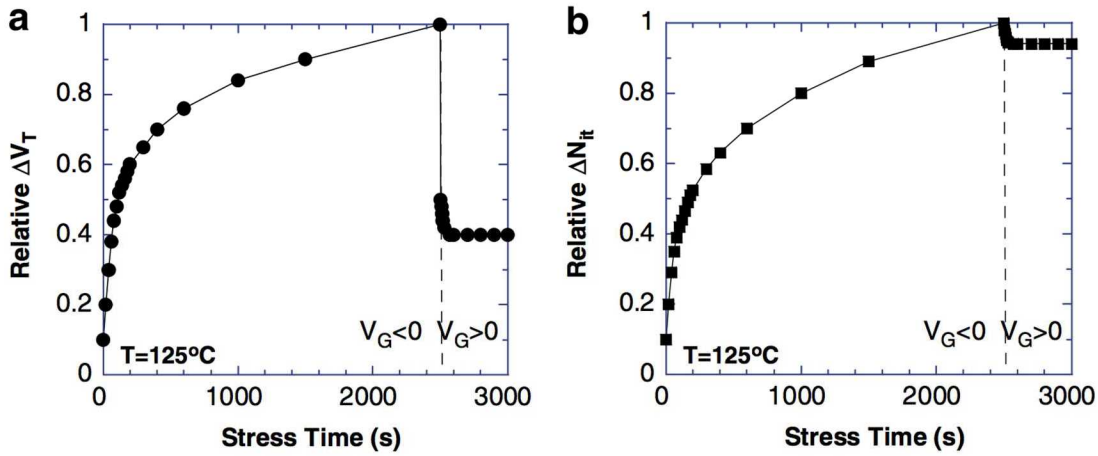


Figure 3. Relative shifts for a) ΔV_{th} and b) ΔN_{it} versus stress time for negative and positive gate voltages. [4]

The response of pMOSFETs under elevated temperature and voltage stress is shown in Figure 3, after [4]. An increase in relative threshold voltage and interface trap population can be seen during stress. After the removal of stress, the device exhibits a recovery behavior; however, the recovery is incomplete and does not restore the device to the original operating conditions. This leads to permanent device degradation whose magnitude increases with increasing stress time. Concurrently, the increase in interface trap population leads to a decrease in surface carrier mobility, decreased transconductance, and a decrease in drain current I_d.

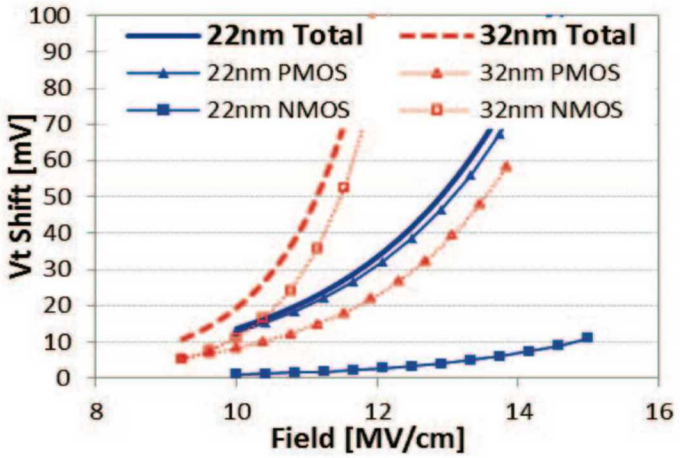


Figure 4. BTI comparison between 22nm FinFET and 32nm planar transistors. NMOS is significantly improved due to gate optimization, oxide scaling, and work function tuning. [5]

Figure 4 shows a comparison in the BTI response of an nMOSFET and a pMOSFET in a planar 32nm CMOS process and a tri-gate technology at the 22nm technology node. The magnitude of the threshold voltage shift shown in Figure 4 is significant in consideration of the baseline threshold voltages of these technologies, which is on the order of 0.35 – 0.5 V. The magnitude of shift for pMOSFET devices undergoing NBTI is greater than that for nMOSFET devices undergoing PBTI/NBTI for equivalent vertical field conditions. Changing the device geometry from a planar transistor to a tri-gate FinFET results in improvement of the BTI response of both nMOS and pMOS devices. Figure 4 also suggests that lowering the operating voltage may reduce device degradation due to BTI and is one way to extend expected device operating lifetime. In general, it has been observed that NBTI degradation is higher in high-k dielectrics in (110) fabricated FinFETs compared to (100) fabricated FinFETs. Positive BTI in nMOS structures charges defects in the high-k dielectric material. This impacts the threshold voltage (V_{th}), not the electron or hole mobility. nMOS devices show significantly improved PBTI due to gate optimization and oxide scaling while in pMOS devices, NBTI shows regression; but an overall BTI improvement is observed in FinFETs. This trend is expected to continue at the 14nm node and below. When discussing these reliability issues with a manufacturer it is important to ask: *Does the vendor monitor NBTI effects on the (110) sidewall FinFETs? In addition, is their fabrication process optimized to reduce NBTI induced failures?*

As the operating voltage of logic transistors dropped under 1.2 V, Channel Hot Carrier degradation (CHC or HCI), which caused threshold voltage (V_{th}) degradation in nMOS transistors of older transistor technology nodes (and still does for some high voltage transistors) became less important. Instead, the increased leakage through the gate dielectric (I_g) gave rise to V_{th} shifts in pMOS transistors and was shown to be enhanced by gate voltage (negative for pMOS) and temperature stress. It was also soon demonstrated that the V_{th} shift can partly recover with an exponential time dependence when the gate voltage is removed; thus, the degradation was found to depend on the frequency of operation. Digital circuits would experience less overall degradation than was originally suspected as a result. Many models concentrated on the leakage current removal of hydrogen atoms (that passivate Si dangling bonds) at the Si/SiO_2 interface, and how hydrogen can diffuse back leading to a partial V_{th} recovery [6]. However, more recently there has been more and more evidence that bulk oxide defects have a major contribution [7]. Since the newest technology nodes (typically < 32 nm) started using Hf-silicates on top of a thin SiO_2 layer as high-k gate dielectrics, and added metal gates, many more avenues to form charge traps appeared, giving rise to positive bias instability (PBTI) in nMOS transistors as well [8].

The dynamics of NBTI and PBTI effects on both permanent and recoverable V_{th} shifts can be measured quite well for more recent technology nodes in either quasi-DC stress/sense experiments on different single-transistor test structures, or at high frequency in various ring oscillator configurations. Because the V_{th} shifts can be fit with relatively simple mathematical expressions as a function of time and gate voltage, for each transistor type in a particular technology and manufacturing process, they are frequently included in the associated circuit simulator software models and process design kits. A circuit designer can easily simulate the effect of the time-dependent V_{th} shifts in all transistors, determine the impact on the overall circuit performance, and make adjustments to correct for both short- and long-term parametric and performance degradation [9, 10].

3.2. Dielectric Integrity

Degradation of dielectric integrity refers to a reduction in the electrical insulating qualities of gate dielectric materials due to a combination of thermal and electrical stress. The dielectric material serves to provide capacitive control of carriers in the channel of MOSFETs while providing a barrier to

carrier transport. The ideal gate dielectric material would have a large relative permittivity and while simultaneously allowing zero leakage current, serving as an extremely high-resistance path in the device. While no insulator is perfect, the Si-SiO_2 interface has proven robust for technology scaling and exhibits a low defect density.

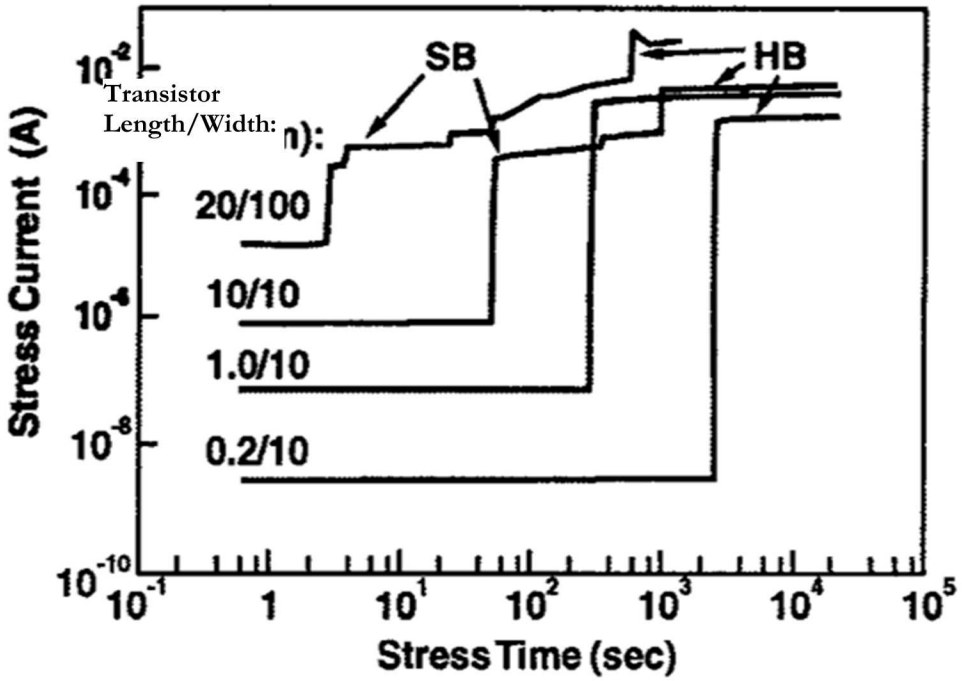


Figure 5. Observed soft and hard breakdown in stress current vs. time. [11]

The classic concern for dielectric integrity is dielectric breakdown, where the normally electrically insulating gate becomes a low-resistance, conducting leakage path. An example of time dependent dielectric breakdown is shown in Figure 5. In this experiment, a capacitor was held at constant gate voltage and elevated temperature conditions, and the current through the dielectric was measured as a function of time. The capacitors had equivalent thicknesses, and the areas of the devices were varied according to the inscriptions in Figure 5. Larger-area capacitors show larger total leakage current and tend to exhibit breakdown characteristics sooner than do smaller-area devices, since the likelihood of generating a sufficient number of defects in close enough proximity to form a conductive path increases with gate area. However, not all breakdown events are destructive. Soft-breakdown events may not necessarily result in a non-functional gate capacitor and the gate can remain operational, although with a significant switching time penalty between the on- and off-states. This type of breakdown is increasingly common in advanced technology nodes. Hard-breakdown remains a significant reliability concern as well and results in a permanent low-resistance, conductive path in the gate dielectric material. This type of breakdown is not recoverable.

It is widely agreed that oxide breakdown results from the generation of defects in the dielectric during electrical stress. There has been a significant amount of literature generated over the last 40 years regarding the generation of defects in SiO_2 and the formation of conductive percolation paths that lead to increased leakage current and eventually dielectric breakdown [11, 14]. When an oxide or dielectric undergoes electrical stress, defects are continuously generated in the bulk and at the interface. The rate at which defects are generated depends on the electric field and temperature. The generation of defects leads to an increasing degradation of carrier mobility, carrier lifetime, and reduction in

switching frequency that cause timing penalties at the circuit level, and increases low-field leakage current through the gate.

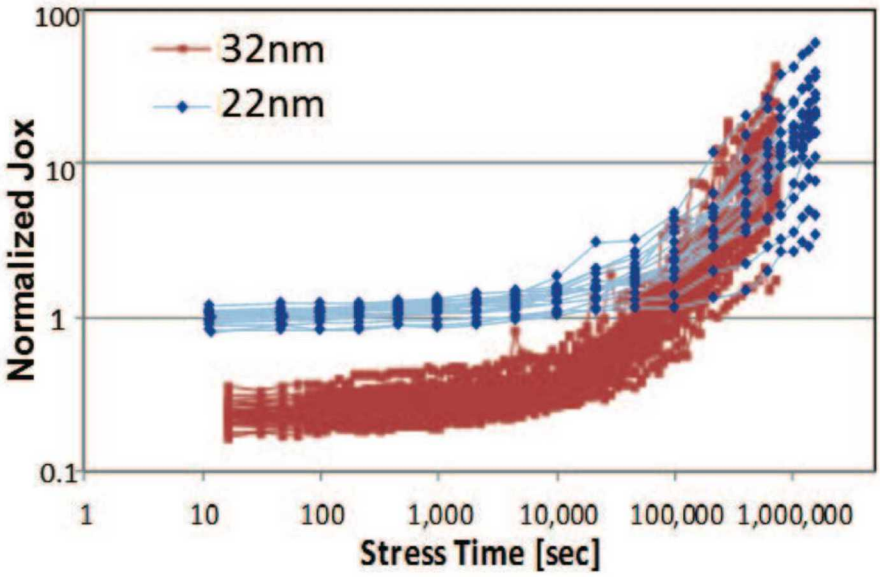


Figure 6. NMOS gate leakage increase with stress is less in 22nm FinFETs than in 32nm planar technology. EOL leakage is matched between the technologies. [5]

Stress-induced degradation is not always a catastrophic destructive effect. Figure 6 shows the normalized leakage through the gate of nMOSFETs in the 22nm and 32nm nodes as a function of increasing stress time, a phenomenon known as Stress-Induced Leakage Current (SILC). SILC is associated with electrons tunneling into and out of traps (and is thus also called trap-assisted tunneling) in oxides thinner than 5 nm. For short stress times (less than 10^4 seconds) the leakage current density in 32nm devices is smaller than that of in the 22nm nMOS devices. Whereas the 22nm devices maintain a fairly constant current density with stress time up to 10^4 seconds, the 32nm devices show increasing gate leakage with increasing stress time, indicating a time-evolving defect structure in the gate dielectric at a relatively short operating life. Near their designed end of life (greater than 10^5 seconds), the gate current density of the 22nm and 32nm nMOSFETs is comparable, indicating that both processes undergo degradation with increased operational stress. The relative change of nMOS devices in the 22nm node is less than that for 32nm devices, indicating a more stable and optimized process.

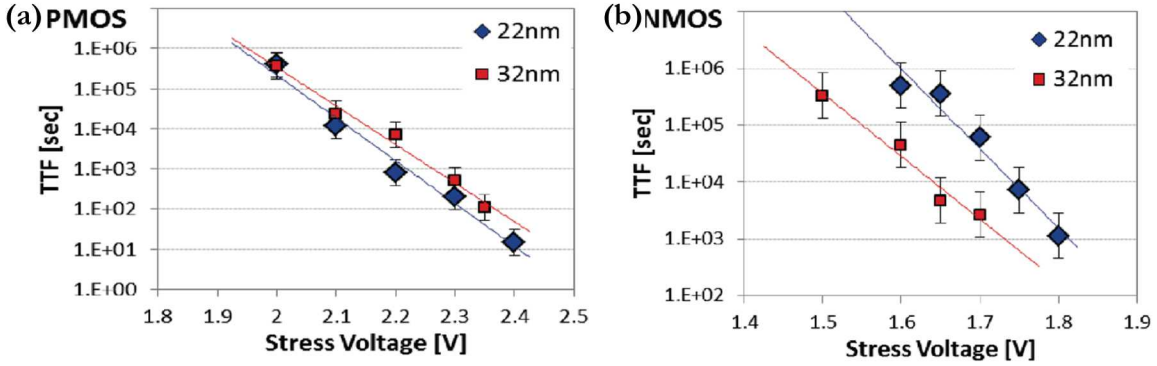


Figure 7. pMOS TDDB (a) has a slightly higher voltage acceleration factor due to oxide scaling, resulting in matched behavior to 32nm TDDB at the respective operating voltages. nMOS TDDB (b)

in the 22nm technology is improved over 32nm TDDB and indicates the intrinsic robustness of the tri-gate architecture. [5]

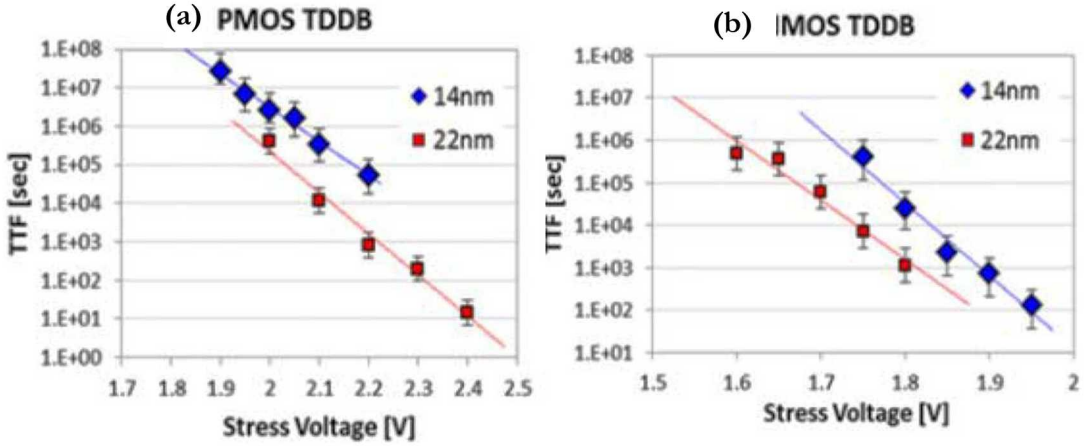


Figure 8. pMOS TDDB (a) has a higher mean TTF observed in 14nm FinFETs compared to 22nm technology at respective operating voltages. nMOS TDDB (b) is improved in 14nm technology compared to 22nm technology. Improvements in FinFET transistors scale with shrinking technology feature sizes.

It is well-established that the generation rate of defects within the gate dielectric is dependent on the applied electric field, or equivalently the gate voltage [15]. This can be seen in Figure 7, which shows mean time to failure (TTF) as a function of stress voltage for gate dielectric breakdown. Figure 7 shows a decrease in TTF with increasing stress voltage, suggesting again that reducing the nominal operating voltage of a design yields an increase in the meantime to destructive failure of nMOS devices. Similarly, Figure 7 shows that pMOSFET devices exhibit an increase in TTF with decreasing gate voltage. Comparing the MTTF for nMOS and pMOS devices fabricated in the 22nm and 32nm nodes in Figure 7 indicates that scaling improves the MTTF of nMOS devices by more than an order of magnitude for equivalent voltage stress. However, pMOS devices in the 22nm node show a slight reduction in the TTF compared to the 32nm node. Additional data comparing 22nm to 14nm FinFETs shows technology scaling improving TTF with decreasing gate voltage for pMOS. These results are shown in Figure 8.

For older technologies (or for high-voltage IO transistors in newer technologies) with SiO_2 (or nitrided oxides) gate dielectric of thickness > 3 nm, it has traditionally been assumed that there would not be any observable degradation in the transistor or the circuit functionality until a catastrophic breakdown occurred abruptly at end of life. However, it is possible that the leakage across the gate dielectric (aka stress-induced leakage, SILC) will increase gradually over time and eventually give rise to increased power consumption, contributing to self-heating of the affected devices.

As CMOS technologies kept scaling the gate oxide thickness to lower operating voltages and increased operating frequency, most chip manufacturers hit equivalent gate oxide thickness of ~2nm beginning with either the 130- or 90nm technology node. The introduction of very thin gate oxides led to a condition where a single electron trap between the gate electrode and device channel is sufficient to open a trap-assisted tunneling path through the oxide, which gives rise to increased leakage current that can lead to catastrophic failure through heating-induced run-off. However, due to the low voltages used (< 1.2 V), this degradation may also recover, resulting in an increase in leakage that is not necessarily catastrophic in nature. Indeed, gate oxide of ~2 nm thickness showed increases in leakage current under mild stress, which were named "Soft" Breakdowns (SBDs) [16] and it was

soon demonstrated that a digital circuit could still function, albeit with a timing penalty for affected circuits, [17] after SBDs until a hard breakdown occurred, which shorted the gate completely.

Ultra-thin gate oxides (< 1.5-2 nm) under stress show "progressive breakdown" behavior with many small SBD events, occasional recoveries, but an overall increasing SILC before a catastrophic HBD event. Yet at typical operating voltages, the leakage degradation has been acceptable, and seemed to still be able to provide the needed lifetime (although potentially with increased power consumption) when it was discovered that the oxide degradation rate could be fit to a power-law model [18].

However, SILC oxide degradation also leads to a loss of transistor transconductance which reduces necessary drive currents to turn-on/turn-off subsequent logic gates. This represents a slowing of the function block's operating speed with transistor aging.

3.3. Hot-Carrier Injection (HCI)

Carriers transporting through a device can acquire large kinetic energies in regions of high electric field. When the kinetic energy of a carrier, electron or hole, is greater than the average energy of a carrier in the lattice it is called a "hot" carrier. These hot carriers can gain sufficient energy to be injected into the gate oxide or can cause damage to the semiconductor / gate oxide interface. The generation of defects at the interface results in fluctuations in the electrical characteristics and operating points of MOSFETs. Hot-carrier injection (HCI) is typically a concern when the design of a device allows the presence of "high" electric fields during operation.

Channel hot carriers are associated with electrons or holes accelerated through the depletion region when a MOSFET is operated in the saturation region with sufficiently high drain voltage to achieve saturation velocity, and with carrier multiplication resulting from impact ionization. While many channel hot carriers contribute to the gate current, a small fraction of them cause damage when injected into the gate oxide or near the drain-side gate edge. For nMOSFET devices at low V_{gs} (less than V_{th}), hot holes are primarily injected, resulting in the generation of interface traps and the build-up of positive charge in the oxide. At medium V_{gs} (from V_{th} to V_{ds}), channel hot carriers generate additional interface states. At high V_{gs} (greater than V_{ds}), damage results from hot electron injection and contributes to high gate current conditions that lead to the injection and build-up of negative charge in the gate oxide and generation of interface states. All of these effects combined contribute to dynamic and unstable device operating conditions that are highly dependent on the stress and temperature conditions as well as on the generation mechanism of hot carriers in the device.

At the device level, damage from channel hot carriers contributes to varying V_{th}, decreased carrier lifetime, increased substrate current, degraded g_m and subthreshold slope, and reduced drive current at nominal operating voltages in nMOS devices.

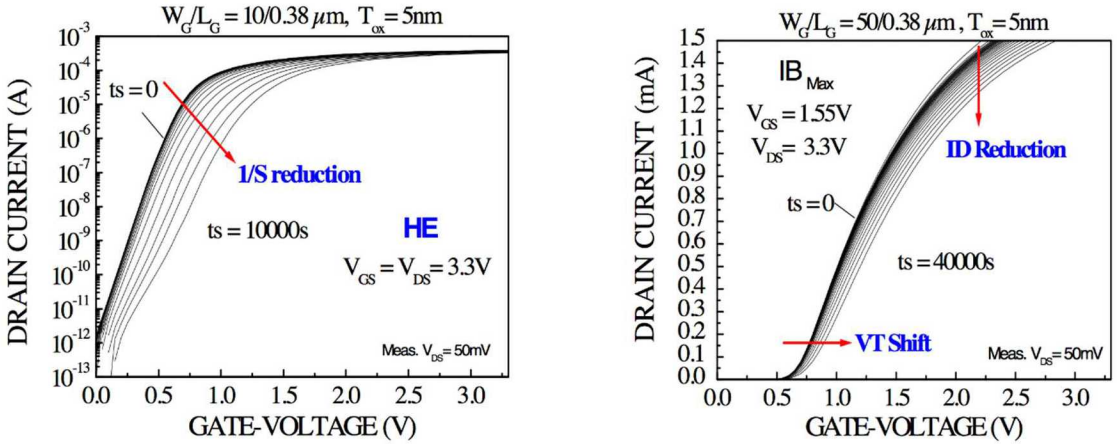


Figure 9. Channel hot electron damage to an nMOSFET device results in decreased g_m, reduced I_d and increased V_{th} with increasing stress (log scale on left, linear scale on right).

The predominant device level effects can be observed in Figure 9, where hot electron stress is shown to decrease device performance in the subthreshold region, reduce drain current in the saturation region, and increase the threshold voltage of nMOS devices with increasing stress time. At the circuit level, these effects would lead to switching speed and timing penalties, reducing device operating frequency and posing significant issues for digital and combinational logic. Further, these effects will contribute to imbalances in operational and differential amplifiers as well as many other common analog design elements.

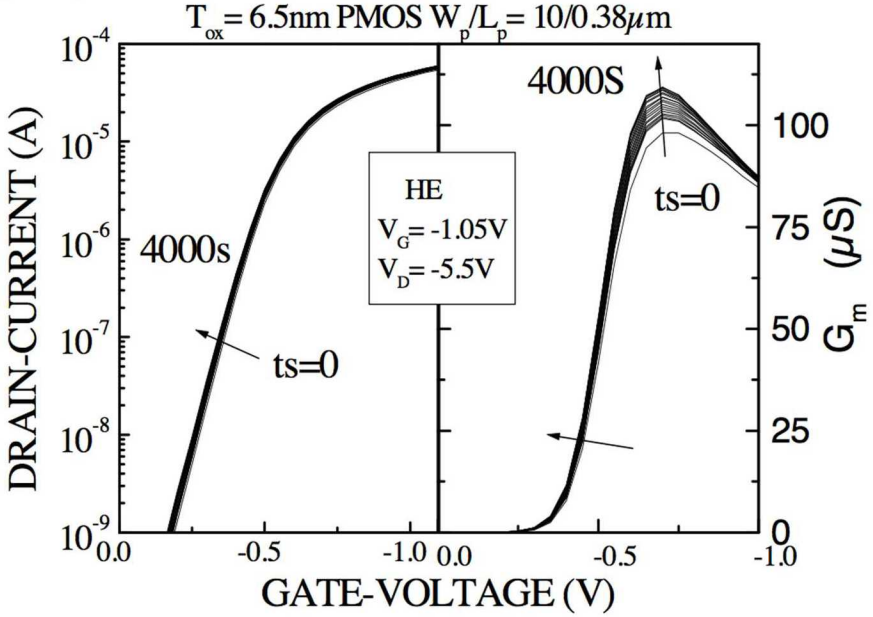


Figure 10. Channel hot-carrier-induced damage in pMOS devices exhibits a reduction in threshold voltage and a change in transconductance, while increasing saturation drain current. [5]

Similar hot-carrier effects can be seen in pMOS devices, as shown in Figure 10. The change in sub-threshold characteristics and threshold voltage shown in Figure 10 is less dramatic than it is for nMOS devices. The reduction in threshold voltage with increasing stress for pMOS devices leads to an earlier turn-on and higher drive currents in saturation. These types of parametric shifts lead to further imbalances in SRAM cells and analog designs such as current mirrors and differential

amplifiers, where precise matching of device geometry and drive characteristics influences data retention in the case of SRAM, and linear amplifier response in the latter designs.

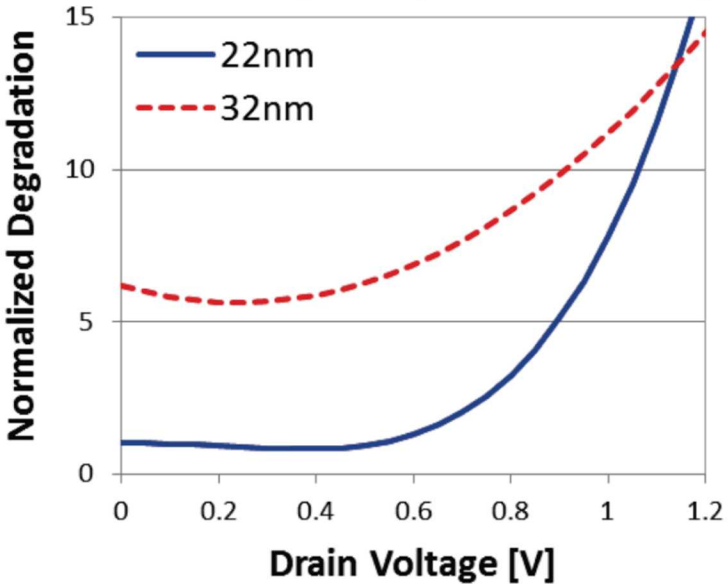


Figure 11. Normalized degradation of 32nm and 22nm nMOSFETs as a function of drain voltage. Results show that the reduction in channel length at low drain bias improves hot carrier-induced stress in the 22nm node compared to the 32nm node. [5]

An important aspect of hot-carrier-induced damage in nMOS devices is the continued scaling of feature sizes, such that the degradation effects increase with decreasing gate length. Figure 11 shows that overall the hot-carrier degradation of nMOS devices in the 22nm FinFET node appears to be better than in the 32nm planar node for a large range of the standard operating voltages of devices fabricated in these nodes. It is important to note that at the onset of hot-carrier-induced damage in the 22nm node (V_{ds} = 0.5 V), the rate of degradation in the 22nm devices increases more rapidly than for the 32nm node. This lends further incentive to providing a reduced operating voltage whenever operating frequency and drive characteristics are a secondary concern as design parameters.

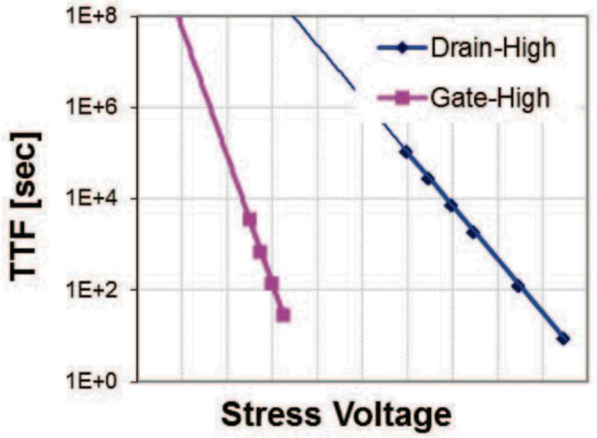


Figure 12. Comparison of functional device failure between hot-carrier stress and gate-dielectric breakdown. NMOS TDDB data for 22nm NFETs stressed with drain high compared to gate-high stress. Drain-high results in significantly better TTF than gate high stress, indicating robust junction and gate-edge quality. [5]

While we have explored several important mechanisms that persist primarily at the device level, with serious circuit-level implications for performance and operating frequency, it is important to evaluate the limiting factor for the majority of designs. Figure 12 shows the time-to-failure for gate-high and drain-high stress, corresponding to degradation of the device due to dielectric breakdown and hot-carrier and bias-temperature instability, respectively. Data show that the drain-high stress condition has a significantly longer TTF for all relevant operating voltage conditions than does the gate-high stress, suggesting that FinFETs can exhibit high-quality junction and gate-edge features. This result suggests that at the device level, the primary concern remains degradation of the gate dielectric material, and mitigation strategies intended to increase the operating lifetime of circuit designs utilizing modern FinFETs should focus on reducing the gate voltage of devices in order to achieve the desired reliable device operating lifetime.

Recent studies on 20nm FinFET technology have shown HCI degradation at increased temperature impacts pFET devices more severely than it does nFETs. These results are anomalous and do not follow traditional HCI temperature behavior based on impact-ionization mechanisms in FinFETs. At elevated temperatures (125°C), pFETs have shown up to 3x more susceptibility to HCI degradation than their nFET counterparts. In addition, the pFET HCI dependence is strongly modulated by the number of fins in the device, indicating the pFET HCI mechanism is different than that in nFETs. These results may be indicative of a BTI issue intertwined with HCI, but nonetheless demonstrate the importance of self-heating on FinFET reliability [22]. These results are shown in Figure 13.

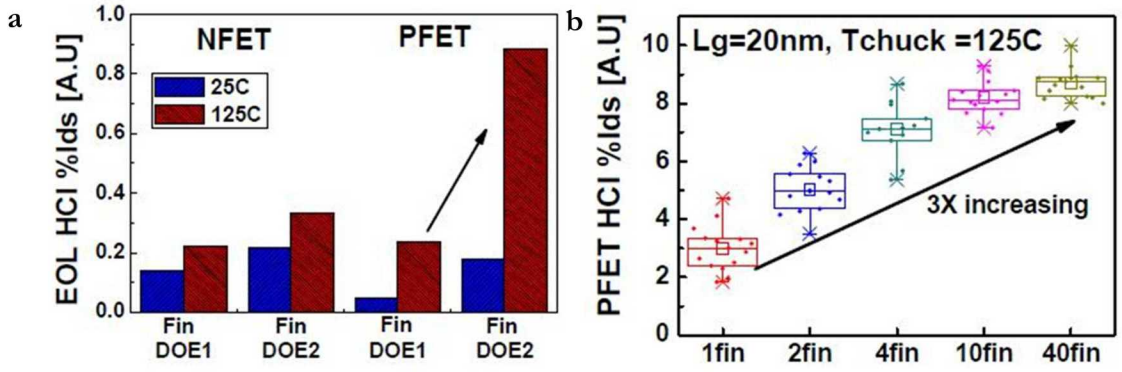


Figure 13. Effects of elevated temperature on a) nFETs and pFETs show that pFETs are more susceptible to degradation at elevated temperature, b) pFET HCI degradation dependence on the number of fins, where no dependence is observed for nFETs. [22]

Overall, HCI degrades FinFET devices when state-changes or peak current events occur, making HCI degradation very dependent upon fin geometry. nFET devices suffer from HCI degradation more severely than do pFETs due to a) holes having smaller impact ionization rates, and b) holes having a higher Si-SiO_2 barrier than electrons. Narrow fin width increases HCI resistance; however, as the fins become taller, FinFETs may become more susceptible to HCI. FinFETs manufactured using (110) silicon are more susceptible to HCI degradation compared to FinFETs manufactured using (100) silicon. This is due to the higher interface trap density formed on (110) oriented silicon. HCI is typically worse for nFET devices (due to dependence on V_g), while pFET structures suffer degradation in V_{th} with increasing stress. When visiting a manufacturer, it is important to ask the vendor: *How do they measure HCI-induced degradation, particularly for FinFETs with frequent state changes? Is the dependence of HCI on gate voltage established prior to gathering their reliability data?*

3.4. Back end of line / Electromigration / Stress Voiding

The Back-End-Of-Line (BEOL) is the portion of a technology process that provides chip-level communication, power, and clock signals to the various transistors, capacitors, and resistors on a die. The BEOL consists of multiple interconnecting layers of metal in a complex routing scheme that provides electrical contact with individual devices, isolation in the form of dielectrics, and bonding/probing sites for external input/output to a chip.

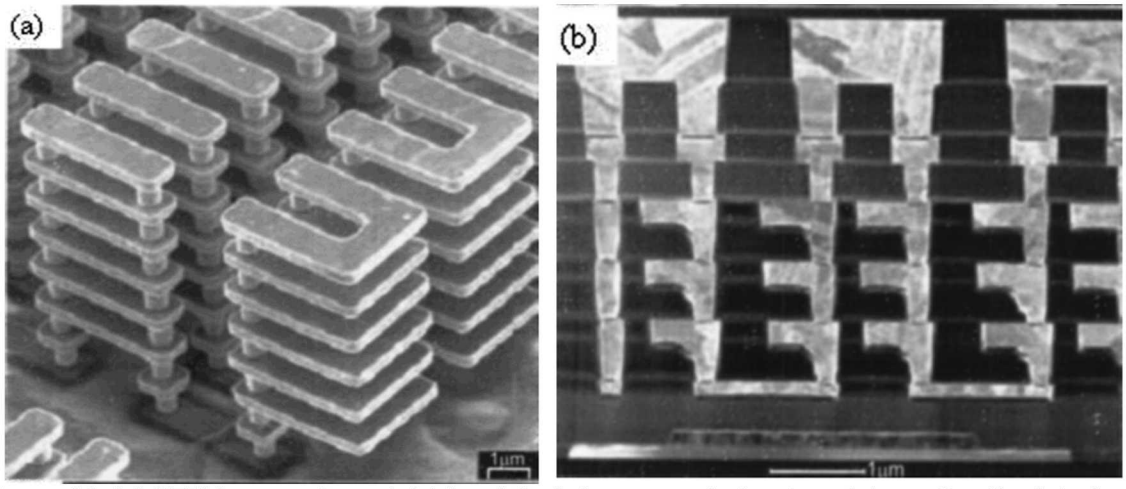


Figure 14. (a) SEM image of an eight-level Cu interconnect structure taken after the interlevel dielectric was etched away. (b) Cross-sectional TEM image of a six-layered dual damascene Cu interconnect structure on a Si surface. The diameter of the narrowest Cu via is 0.25 um. [12]23

Figure 14 (a) and (b) show an example BEOL from a process that features an 8-level Cu metal stack. More modern processes possess up to 10 or 11 metal layers. Figure 14 (b) shows the BEOL cross section. Routing between metal layers requires the use of either Cu or W plugs, or vias, that provide connection from one layer to another. Higher metal layers are the widest and thickest, and have the largest spacing between adjacent routing lines. Higher metal lines have a lower resistance and RC time constant penalty, and are used for power and clock routing at the chip level. Lower metal lines are thinner, narrower, and shorter, with smaller spacing between wires. These lower metal lines are used for local device and circuit connections.

The BEOL provides the backbone for device connections and communications at the device, circuit, and chip levels. This makes the fabrication and operation of reliable metal routing and power lines a fundamental piece of technology design and reliability. Modern BEOL materials consist of a wide variety of metals and alloys that increase the complexity of integrating the BEOL process while simultaneously ensuring formation of high-quality thin films. Properties of thin-film materials can be very different from those of bulk materials, further exacerbating reliability testing of highly-scaled BEOL materials and requiring extensive investigation of the thin-film material properties and degradation mechanisms of each new technology node.

Failure in the BEOL is typically the result of open connections, increased resistance, unintended short circuits, and leakage current. Open connections result from line discontinuities, whereas shorts cause signal or power loss through unintentional connections to other lines. Increased resistance is more difficult to detect and can result in timing failures at the circuit and system levels, failure in memory operations, and erratic device switching behavior. Leakage-related failures can cause signal loss below acceptable margins and can reduce the functionality of circuit and memory elements.

Detecting leakage and resistance-related failures is extremely difficult, as these mechanisms result in intermittent failures and random anomalous device behavior.

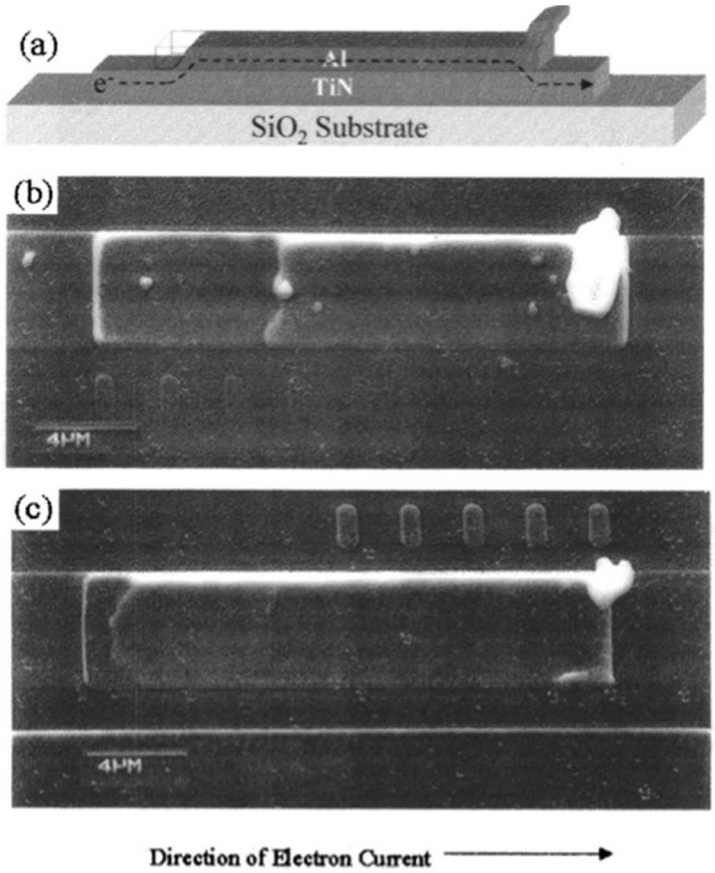


Figure 15. (a) Schematic diagram of the Blech structure consisting of a short Al strip on a TiN base line for electromigration tests. A void at the cathode and a hillock at the anode are depicted. (b) SEM image of the top view of a Cu strip tested for 99 h at 350°C with current density of 5×10^5 A/cm^2. (c) SEM image of the top view of a Cu (2 wt% Sn) strip tested at the same condition. [23]

There are currently two main concerns for metallization of the BEOL, electromigration and stress voiding. The two are similar phenomenon insomuch as stress voiding is the release of shear stress resulting in vacancies, while electromigration processes can cause vacancies in metal lines through heating and collisions. The four best conducting materials are aluminum, copper, silver, and gold due to their low intrinsic resistance and high thermal capacity. Normally in physics and engineering, metal wires are treated as electrical conductors with resistance and heat properties specific to the material. The conductor is generally considered to remain unaffected by the current flowing through it. This is a reasonable assumption for relatively low currents. However, when considering electrical failure in semiconductor technology it is important to note that the size of metal lines is much smaller than what is found in a house or power grid. The current densities in BEOL metal lines can therefore be very large, resulting in Joule heating. This aspect of metal line integrity is carefully monitored in technology processes. However, metal line failure can result from an applied current causing the migration of metal in Al and Cu lines [24,25].

Prolonged high current densities forced through the thin-film metal lines of a modern BEOL process can cause a significant migration of metal in electrical wires, leading to voids and/or extrusions in the wires and eventually to open- or short-circuit behavior. Figure 15 shows a schematic of a Blech structure. In this example an electrical current flows from left to right through an Al wire. The

electrical current causes collisions between transporting electrons and Al atoms, resulting in momentum transfer and displacement of Al atoms. Displacement of metal atoms results in vacancies that can accumulate and gives rise to large voids, as seen by the empty box on the left of the Al line in Figure 15. With sufficient current stress through metal lines, voids can contribute to open-circuit effects or increase in resistance. The open-circuit connection issues are easily seen in the time evolution of electromigration and voiding shown in Figure 16, which shows an increase in Cu vacancies for high current density conditions at elevated temperature with increasing stress time. Concurrently, the migration of Cu from the left of the metal line leads to extrusions at the opposite end of the metal line. These extrusions can penetrate into the low-k dielectric material and cause shorts between adjacent metal lines.

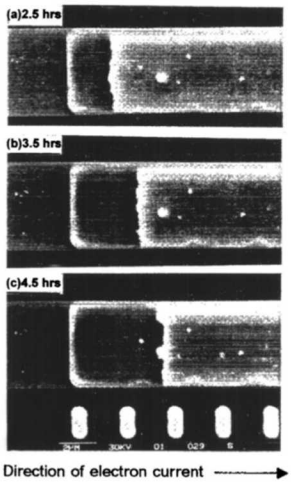
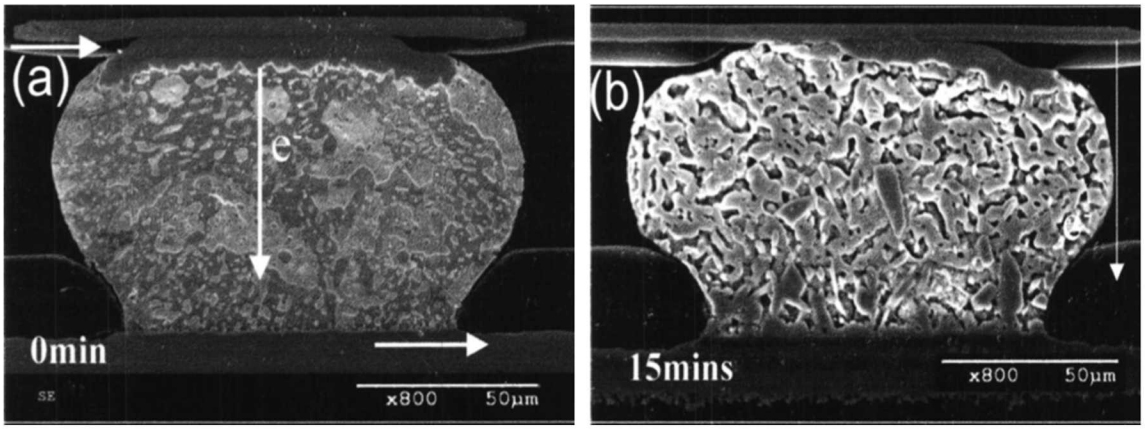


Figure 16. (a) SEM image of mass depletion at the cathode of a Cu strip at 2.5 h at 400°C with current density of 2.1×10^6 A/cm^2, (b) at 3.5 h, and (c) at 4.5 h. The drift velocity is about 2 um/h as measured from the three images. [23]

Electromigration and stress voiding can also occur at external electrical connections to a chip, as shown in Figure 17. High current density stress leads to pitting and voids in solder ball connections and eventually causes an open in Cu metal lines and dissolution of the solder ball connection, leaving a high-resistance contact resulting in loss of electrical connection to devices connected to the voided metal line. An effective mitigation strategy for minimizing damage from stress-induced voiding and electromigration is to limit the current density per solder ball and subsequently flowing through wires. This allows one to appropriately size the metal lines to reduce electrical stress below acceptable levels.



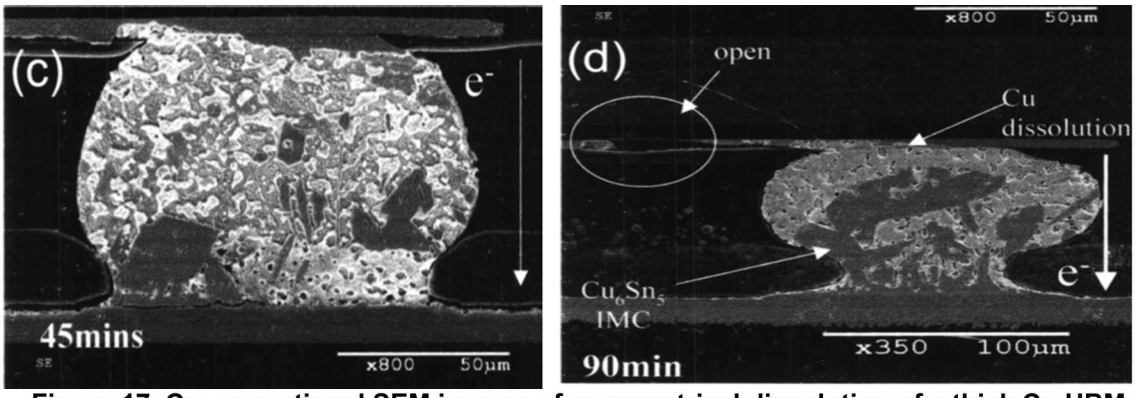


Figure 17. Cross-sectional SEM images of asymmetrical dissolution of a thick Cu UBM due to current crowding by electromigration at 125°C with an applied current of 1.2 A: (a) 0 min, (b) 15 min, (c) 45 min, and (d) 90 min. The electron current entered the bump from the upper-left corner. As the solder replaced the Cu at the upper-left corner, more and more intermetallic Cu_6Sn_5 formed inside the solder bump near the anode. [23]

In summary, the resulting highly localized die level electromigration related failure modes can be broken into 3 areas: 1) interconnect open circuits due to gross line void formation, 2) increased interconnection RC time constants due to increased ohmic resistance and 3) formation of high leakage or shorting between parallel interconnection paths due to metal line extrusion. Failure Mode 1 typically leads to a localized intermittent followed by hard device open circuit failure. Failure Mode 2 results in non-gross increased parametric electrical interconnection resistance which directly impacts the RC time constant and increasing interconnection signal propagation delays. Failure Mode 3 results in intermittent electrical arcing and then high current shorting related feature failures.

In many cases, the highly localized and intermittent nature of these failure modes can appear as electrical "glitches" and system level impacts could be modeled as such. The circuit location of these electrical reliability glitches is typically associated with aggressively designed areas of interconnect layout.

3.5. Middle of Line Concerns

Reliability concerns for dielectric integrity and stress voiding have manifested as an emerging concern at the Middle-Of-Line (MOL) beginning with the 32nm technology node. The MOL refers to issues between the diffusion and active silicon layers (the Front-End-Of-Line, FEOL) and Metal 1 (M1, the beginning of the BEOL). Aggressive scaling of device feature sizes has required shrinking the pitch between gate and diffusion contacts. When discussing dielectric integrity in Sec. 3.2 it was noted that the probability of breakdown occurring shows a strong dependence on oxide/dielectric thickness. The same principle holds true for the MOL, where decreasing distance between the gate and contact or silicide provides an additional leakage path that only became significant in devices fabricated in the 32nm node. These leakage paths can be seen in schematic diagrams shown in Fig. 18, Fig. 19, and Fig. 20.

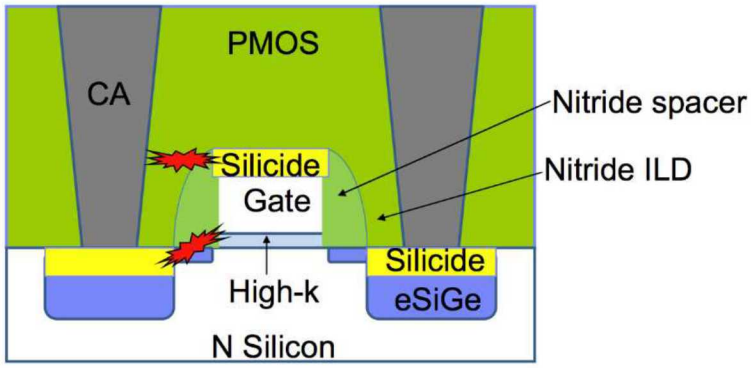


Figure 18. Middle of line reliability concerns show leakage and breakdown paths along gate to diffusion and gate to contact leakage.

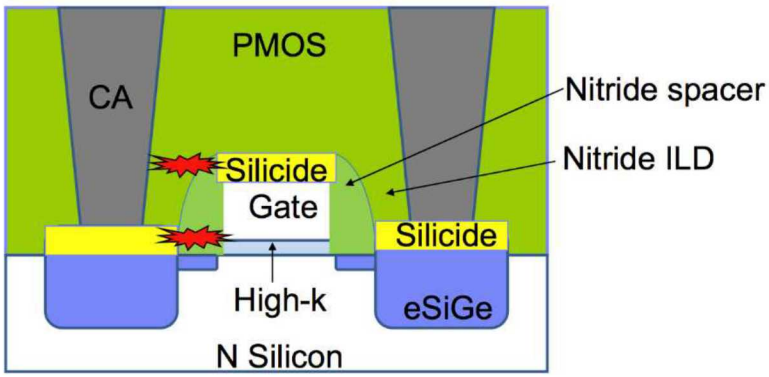


Figure 19. Dominant leakage paths for MOL reliability in a gate-first, raised-drain geometry. Elevated silicide and small spacing between M1-DIFF and gate silicide lead to contact-diffusion leakage path/breakdown paths from the gate to the contact.

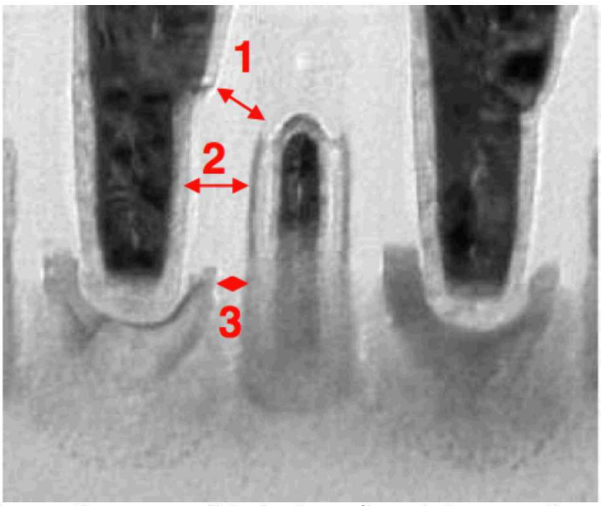


Figure 20. TEM of FinFET shows three possible leakage/breakdown paths due to bulging contacts, small contact-gate spacing, and gate-diffusion spacing.

Increased leakage or breakdown can shift the threshold voltage and impede nominal device operation. Other factors influencing MOL reliability include metal salt contamination within the nitride spacer and lining, contact defect bulging, and vanishing spacer causing encroachment of the diffusion silicide towards the gate. These concerns can largely be addressed by optimizing the process, including increasing gate-contact pitch, to reduce lateral gate-to-contact fields.

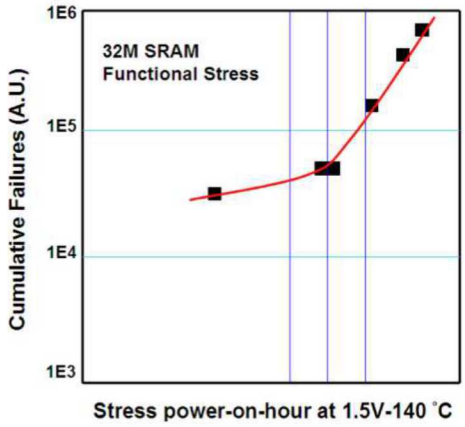


Figure 21. Poly-diffusion V_{bd} with different contact numbers transformed based on fatal area scaling from the un-optimized process. [26]

Figure 21 shows the impact of MOL reliability on failure rates for on-state stress at 140°C in a 32 Mbit SRAM fabricated in the 32nm node that exhibits a gradual slope due to extrinsic defects (contact bulging, vanishing spacer, metallic contamination of spacer/liner) and a higher failure rate due to intrinsic reliability (leakage and breakdown). Each of these concerns needs to be optimized for a reliable process.

Intrinsic reliability can be addressed by providing sufficient space between contacts and the gate stack, which in turn takes care of extrinsic issues related to bulging contacts. Further optimization of contact formation yields improved results for failure statistics as shown in Figure 22. Further, "cleaning" of the wafer yields an improvement in breakdown statistics by removing metal salts contamination of the nitride spacer and liner near the gate stack, Figure 23. By optimizing contact spacing, the desired reliability can be achieved in the MOL for a technology node. Figure 24 shows the intrinsic TDDB reliability of a 28nm process.

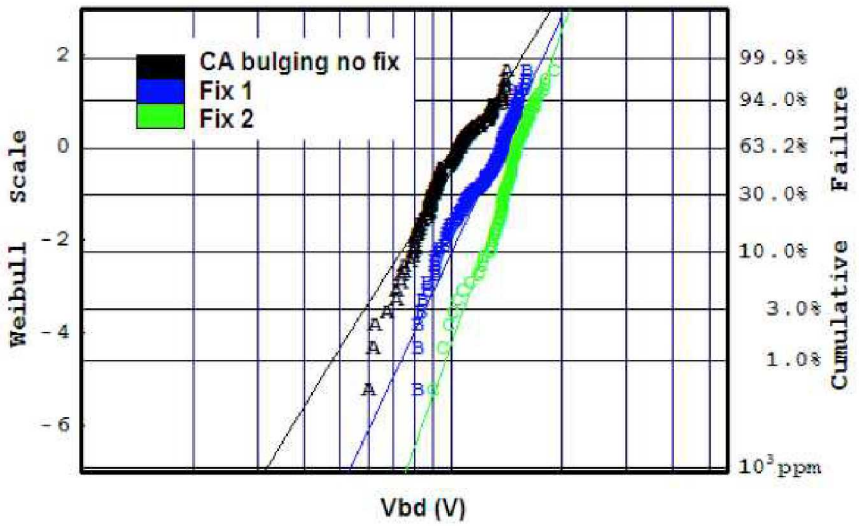


Figure 22. PC-to-CA Vbd difference among CA bulging splits. [18].

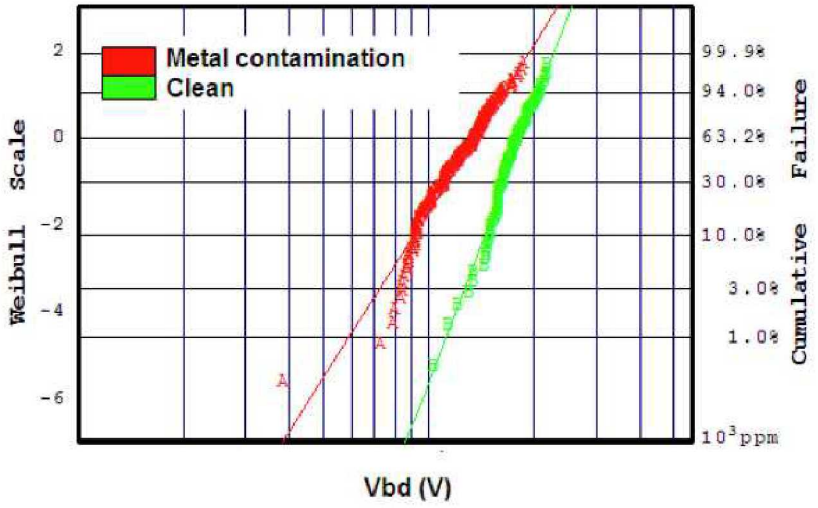


Figure 23. Comparison of failure statistics for a wafer with metal contamination in the nitride spacer and lining of the gate stack and a clean wafer. [26]

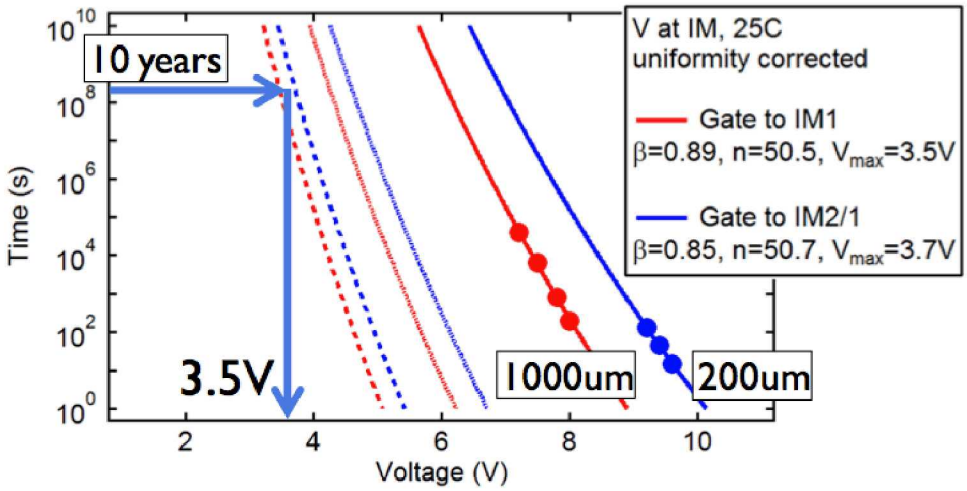


Figure 24. The TDDB lifetime extrapolation of the gate-to-IM1 and gate-to-IM2/1 dielectric after correcting for die-to-die variability. Scaling to a total contact length of 100m and to 0.01% failure results in a 10 year Vmax of 3.5V. This is the intrinsic TDDB lifetime at 25°C, clearly meeting the VDD spec. [27]

Addressing the intrinsic and extrinsic defect reliability concerns effectively improves product lifetime in the MOL beyond requirements for a 10-year product at 3.5 V in the 28nm technology node. Further process optimization is required for continued scaling of advanced CMOS nodes, but many of the inherent reliability concerns overlap many technology nodes with small feature sizes. The most difficult problems remaining to be addressed revolve around decreasing cell size and shrinking gate-contact pitch with each further technology node.

4. PACKAGE RELIABILITY AND PACKAGE-DIE INTERACTION

Although the focus of this report is on Silicon based transistor technology nodes and circuits, the accelerated drive to add product functionality and performance with a reduced package footprint has driven the semiconductor industry to move to highly complex multichip package solutions. The interaction of these advance package solutions with silicon die performance is significant and must be accommodated in the final product design and manufacturing to meet customer expectations. In this section, we will highlight some of the critical package/die technologies and interactions to provide guidance for their mitigation.

4.1. Package Trends

Historically, die packaging involved wire bonding interconnect between the die and the package lead frame. At that time, the bonding point between gold wires and IC aluminum pads suffered Al/Au inter-metallic phase reliability issues (known as the "purple plague") which caused bond weakening and high-resistance contacts, but these problems have largely been solved. Wire bonding is still being used on some occasions, such as to connect MEMS and other multiple silicon die technologies, on to multichip packages.

Most of the multi-pin packages of current-generation, fast microprocessors are built in a flip-chip configuration. Instead of the traditional gold or Al wiring that was previously used to connect the aluminum pads at the top metal layer of the die to the package substrate lead-frame, the IC is mounted upside-down in the package and small solder balls are used to connect the IC's aluminum pads directly with the package substrate, Figure 25. An underfill of thermally conductive, electrically insulating polymer is added between the solder balls. Defects in the solder balls may lead to metal voids that can grow by electromigration, similar to the stress and high-current-density concerns of the BEOL reliability discussed previously.

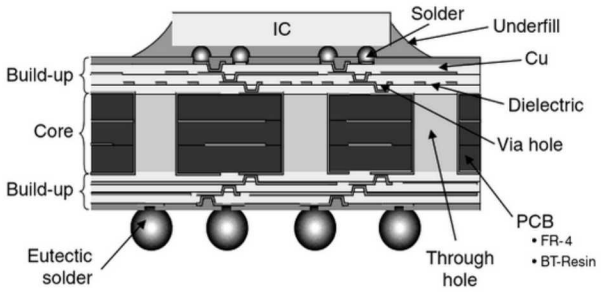


Figure 25. Traditional Flip-chip Ball Grid Array Package. Ref [28]

Note that for some new technology node ICs, such as parts that are used in cell phones where price is an issue, flip-chip packaging may be implemented with a conductive composite (silver-particle-filled) polymer that serves to make electrical contact between the balls and the package substrate while providing insulation between the balls. This type of packaging is less reliable, especially at high temperatures. Extensive reliability testing and customer use condition modelling has established that the many of these advanced technologies can meet final commercial and consumer product reliability expectations.

4.2. 2&3D Die Stacking Packaging Options

The accelerated drive to add product functionality with a reduced package footprint has driven the semiconductor industry to move from single highly complex SOC chips to 2D and 3D stacked multichip package technologies. See Figure 26. Some benefits to using 3D stacked package technologies include; smaller device footprint, enhanced power management, ability to handle large amounts of data, create new and development of new, interesting classes of products. In some cases, these advanced packaging concepts represent the only solution path for certain applications. All 3D packaged devices need to be considered and modelled as a highly integrated physical and electrical system from a quality, performance and reliability perspective.

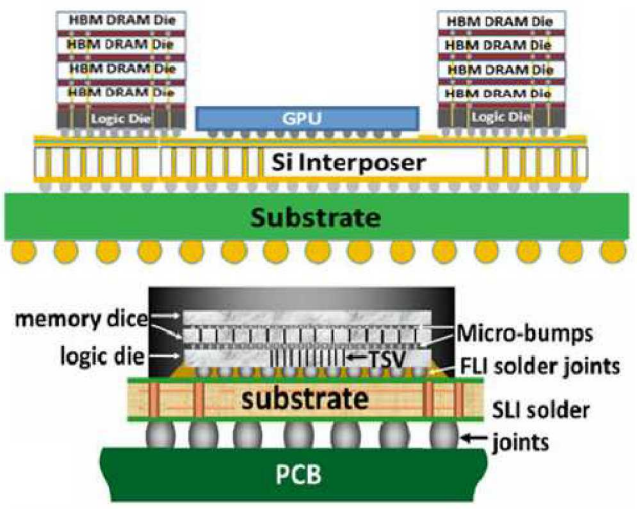


Figure 26. Schematic view of several 3D packaging concepts allowing integration of multiple die technologies to optimize device functionality, performance and yield. (Adapted from Ref 29)

Some of the advantages of these 3D package methods are:

1. Improved integrated device performance due to reduced die separation and line RC delays.
2. Integration of multiple die technologies like FLASH, DRAM memory, MEMS or optical devices into a single packaged device.
3. Each component die could be pretested and matched at a wafer level prior to integration into the final package increasing the yield of the final packaged product.

Some disadvantages of these packaging solutions include:

1. Increase in packaged device complexity which increases the cost of final packaging and electrical test process solutions.
2. Increased utilization of maturing State-of-the-Art die thinning, silicon to silicon and silicon to package interconnect technologies. These have been driven by low cost commercial products and meet the reliability requirements for those markets. These include utilization of TSVs, Through Silicon Vias, silicon to silicon die bump technologies, as well as wafer to wafer physical bonding techniques.
3. Reduction in ability to physically, thermally and electrically isolate die to provide a good failure analysis capability.

4. Increased performance sensitivity to final application, testing and thermal environments. Higher internal die count has led to increased package/application- specific power dissipation which can lead to strong internal temperature and mechanical stress gradients affecting device reliability and performance[10,31,33].

Die stacking methods have evolved from a single layer of multiple adjacent wire bonded die and discrete components onto an over molded BGA PWB substrate (2D) to the current vertical 3D stacking of flip-chip die directly on top of one another utilizing through die via to provide intra-die communication [29]. This is exhibited in the FPGA example below.

The consumer cell phone market has also aggressively driven a total package thickness reduction to enable physically lighter weight and thinner phones. This has driven a substantial move to die thinning techniques allowing the stacking of a greater number of die while maintaining a standard total package height. Many memory products now use this 3D package technology to increase the amount of memory per package. The capability to manage high local data flow rates has been an issue due to bandwidth limitations traditionally seen using individual packaged parts like microprocessor and packaged memory. Packaged I/O bandwidth has not kept pace with Moore's law but 3D heterogeneous integration & packaging improves I/O bandwidth enabling more rapid data processing.

Some advanced commercial FPGA devices like the Xilinx Virtex-7, Figure 27 incorporate silicon interposer-based 3D Stacked Silicon Interconnect, SSI, to increase yields and improve performance. Latest generation cellphone chip sets like the Apple A12 Bionic and Qualcomm Snapdragon incorporate the latest foundry 7nm generation silicon process and advanced 3D packaging integrating the processor, memory and support chips into a single package. This allows for the addition of advanced product features, increased clock rates and reduced power consumption.

The complexity and benefits associated with these stacked die packages have increased but are also subject to increased physical and electrical reliability considerations. The increased number of buried die interconnects, physically larger material interfaces, thermal expansion mismatches and multiple die technologies increase the impact of most environmental stresses. Total die stack and package layout must be carefully designed to ensure that thermal and moisture related bending stresses are minimized during standard use conditions and/or reliability testing to avoid internal interface delamination which can easily break internal intra-die interconnect or die resulting in device failure.

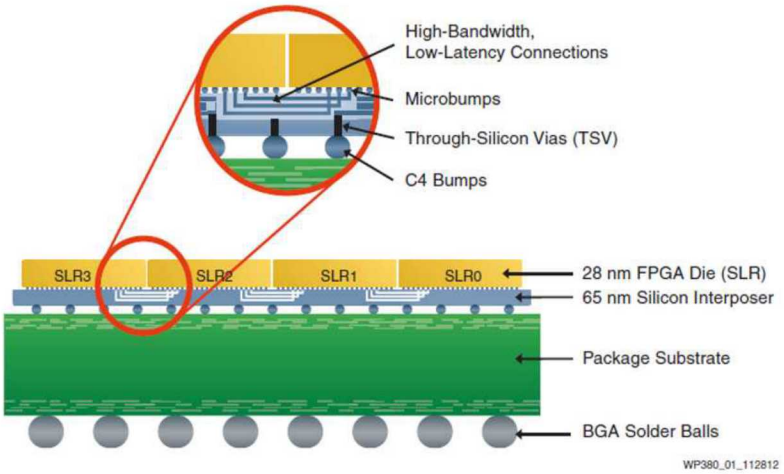


Figure 27. Schematic Xilinx Virtex®-7 2000T FPGA Enabled by Stacked Silicon Interconnect, SSI, Technology. From Reference [32]

4.1. Packaging Process Reliability Considerations

Although work is continuing on 3D package technology development and reliability, there are several noteworthy findings when considering utilizing these advanced packaging technologies. These plastic packages are not hermetic and contain several thermal expansion mismatched material systems. This is an important consideration due to the very high temperature cycles associated with their solder attachment, assembly and repair, to the application printed circuit board. Highlighted previously, several die foundry manufacturing process and design optimizations leads to significant transistor parametric interactions that need to be accommodated for long term operational performance. These will be discussed below.

4.1.1. Package Assembly Reliability

If the package is of the PBGA (Plastic Ball-Grid Array) type, the flip-chip frame connects directly to a solder-ball grid array that can later be soldered directly to a board by applying flux and increasing the temperature above the solder melting point for a few minutes.

The various steps in building the package, and then soldering it to the board, often involve polymer curing or solder re-flow at elevated temperatures. When the package and die cool, they contract differently, and stress may build up in the die to a point where it may "potato-chip". Sometimes, every step in the packaging process or soldering may increase or relieve the accumulation of stress. Since new microprocessors have many interconnect levels, and the metal lines and vias are embedded in very soft (low modulus, low toughness) low-k dielectric, delamination of the interconnect layers over time, as well as moisture ingress, are reliability risks [33]. Further, the surface stress at the silicon can affect the nominal operating frequency of transistors substantially [34], as pMOS and nMOS transistors respond to stress at perpendicular directions due to changes in hole and electron mobilities, respectively.

Finally, plastic packages are not effective moisture barriers. If a micro-crack occurs on the protective overcoat in the die, moisture ingression will eventually occur, leading to corrosion and early electrical breakdown of dielectrics in the BEOL and gate stack. Further, if the plastic part is left in ambient air above the manufacturer's specified temperature and then soldered at high temperature, "pop-corning" or delamination of the package component may occur due to flash boiling of the moisture that resides in the package. It is therefore important to bake the parts before soldering them to the board if they were not sealed in dry gas.

When engaging with a die packaging houses for a new die level IC product, questions addressing package robustness against your specific downstream manufacturing process and use conditions should be asked. *These include; Is the proposed IC package currently in high volume production at their site indicating that it is a mature process? Does the package meet all IPC/JEDEC packaging reliability requirements for comparably sized die? Does the packaging house have experience with special customer requirements such as package moisture level 1 capability and high temperature and/or humidity operation? Where is the actual packaging manufacturing performed and sources for all package piece parts such as lead frames, printed wiring board interposers, etc. (domestic or international supplier sites)?*

4.1.2. Silicon Die Thinning Reliability Impacts

The goal of moving to 3D packaging is to reduce the overall function block device package footprint and thickness. Due to the needed vertical stacking of individual die and need for Thru Silicon

Via interconnection, these dice must be thinned from their original ~750 μm wafer thickness to thicknesses far less than 150 μm.

Published work has established that reduced die thickness does impact the stress state within the active CMOS channel which directly impacts the performance of p and n MOS transistors differently[35]. Figure 28 provides a demonstration of die thickness impacts to the transistor critical Idsat saturation current. These pMOS and nMOS transistor impacts can lead to CMOS cell balance issues on the final device. This is particularly important to meet SRAM memory cell balance performance requirements. At a higher level, this will also impact transistor related logic switching times due to required drive currents needed to manage subsequent RC timing parasitics. These die thickness related CMOS balance issues must be modelled and mitigated through both circuit design and/or foundry process optimization to ensure proper performance of the final product customer application.

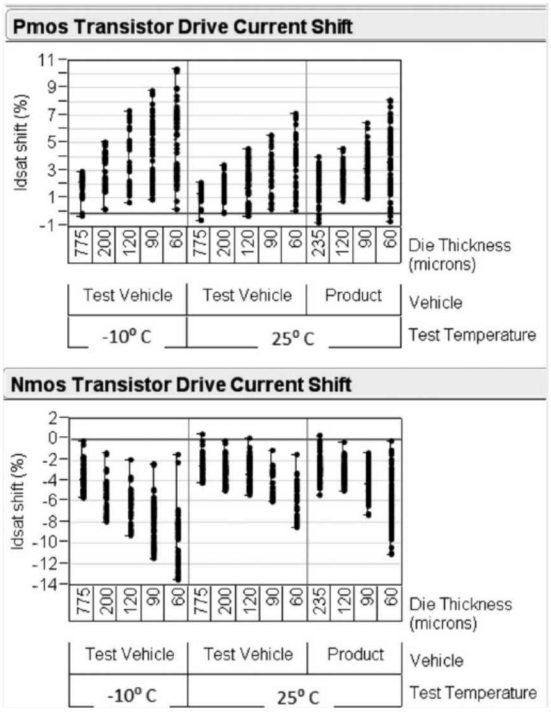


Figure 28. Transistor saturation drive current (Idsat) shifts by test temperature, vehicle, and die thickness (as a proxy for mechanical stress). Idsat shifts are calculated from the observed shifts in P- and N-dominated ring oscillators. [35]

Although the individual transistor parametric performance is affected, work is continuing to understand and quantify the transistor level reliability impacts of these non-mitigated CMOS cell balance effects. Shifts of the CMOS cell Idsat will directly influence timing delays in associated logic structures. The impact of these timing related failure modes within logic applications will be discussed further in Section 8 of this report.

Typically, the use of these advanced 3D packaging technologies require extensive modelling of the die/package parametric performance interactions. Most die foundries are familiar with advanced package requirements. *When engaging with a foundry for a new integrated 3D packaged die, ask for their recommended layout design considerations, available foundry process flow optimizations, captive or external suppliers for wafer thinning to ensure successful packaged product integration?*

4.1.3. Through Silicon Via, TSV, Reliability Impacts

TSV have similar reliability issues associated as the previously discussed with the BEOL via structures. The TSV larger dimensions and aspect ratios through the thinned silicon wafers lead to higher stress intensification factors in the active area. Significant additions to the silicon foundry manufacturing process are required to accommodate their integration leading to additional reliability and quality risks. A good description of these TSV manufacturing process and reliability challenges are described in Ref [36]. The TSV manufacturing foundry process is maturing and incorporates 3 possible foundry flows, Via- First, Via-Last and Via-in-the-middle as described in Figure 29. Beyond the added TSV layout considerations and manufacturing process risks, high speed electrical connections must now be reliably made to die above and/or below this die as demonstrated in Figures 26 and 27.

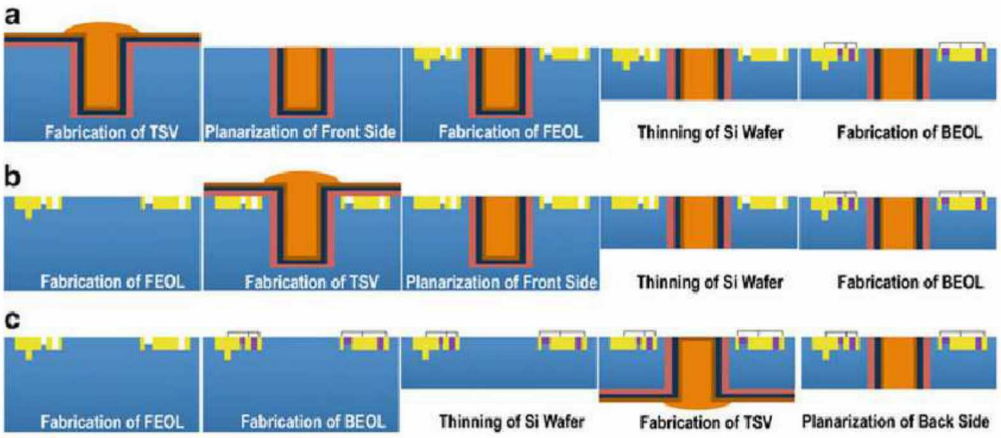


Figure 29. Flow-process sequence for fabricating TSV assembly with respect to the fabrication of FEOL and BEOL structures: (a) via-first, (b) Via-middle and (c) via-last. It should be noted that some of the sequences may be interchanged depending on the process optimization; for example, BEOL structures may be formed before thinning the Si wafer in the via-first process sequence, etc. From Ref [36].

A number of reliability studies have been completed that have established the reliability of TSV designs utilized in many application usage environments and TSV are now used extensively in many commercial 3D packaged devices[37, 38, 29, 66].

Several key design and reliability take-aways from TSV integration are [36]:

1. Significant active area TSV "keep away zones" are required due to the local mechanical stress induced transistor parametric shift by the large embedded copper TSV structure.
2. Copper diffuses quickly through silicon generating deep energy recombination centers resulting in poor transistor performance. This is mitigated through the use of a high-quality barrier layer along the total length of the TSV which is very similar to BEOL via structures.
3. Depending on the TSV process utilized, processing of the wafer's front and backside may be required necessitating appropriate required backside wafer alignment capabilities.

Most die foundries have a TSV structure capability. Since there are several possible TSV foundry process flows, when engaging with a foundry on a new integrated 3D packaged die incorporating TSV structures, ask for their recommended

layout design considerations/ rules and available optimized TSV foundry process flow to ensure successful final packaged product integration?

5. INTEGRATED DIE DESIGN AND PROCESS RELIABILITY ISSUES

5.1. Electrical Latchup

Electrical latchup is a type of short circuit that occurs in an integrated circuit, or more specifically, a low-resistance path between the power and reference supply rails of an IC. This triggers a parasitic structure that can disrupt the nominal operation of a device and can lead to the destruction of a device due to an overcurrent condition. The correction procedure for latchup requires a power-line cycle that allows the condition to self-quench.

The parasitic structure that leads to latchup is inherent in standard CMOS processes where n- and p-wells exist on a common substrate and is called a thyristor or Silicon Controlled Rectifier (SCR) where a *pnpn* structure exists. This configuration acts as a coupled *PNP* and *NPN* bipolar transistor pair. During a latchup event, one of the transistors is conducting, and the other also begins conducting. The pair of transistors creates a positive feedback condition where they keep each other in the saturation region, leading to a large current density that can cause damage to the device at the material level. Latchup need not necessarily occur between the power rails; it can occur anywhere the SCR structure is present, for example in SRAM structures and IO buffers.

Latchup conditions are detected by a voltage transient on the input or output of an IC that exceeds the power rail potential by more than the diode voltage drop (~0.7 V at 297 K). Latchup can also lead to an electrostatic discharge event, which will be discussed in the following section. Additionally, latchup can occur due to ionizing radiation, which makes this a significant issue in electronics for space and high-altitude applications, as well as terrestrial applications.

There are several strategies to mitigation and prevention of latchup in advanced technology nodes. The first is to introduce trench isolation surrounding both nMOSFETs and pMOSFETs. This acts to break the parasitic SCR structure. This approach, while effective, leads to an increased area for standard designs that may be undesirable, depending on the application. Another strategy is to fabricate an IC on lightly-doped epitaxial layers grown on heavily-doped substrates. The heavily-doped substrate acts as a blocking layer where excess charge can quickly recombine, preventing the positive-feedback condition where charge injection from either of the parasitic bipolar devices can reinforce the operating condition of the other. The introduction of latchup prevention circuitry is another method that detects voltage spikes on the power rails and acts to power-down the IC for a predetermined amount of time to allow the latchup condition to self-quench, before powering up the IC again in the correct order. Finally, the use of silicon-on-insulator technology is an effective strategy for latchup mitigation as the fundamental device geometry consists of a buried oxide (BOX) below the active silicon region, and no electrical path exists between adjacent nMOS and pMOS transistors.

5.2. Electrostatic Discharge

Electrostatic discharge (ESD) is the sudden flow of electrical current between two charged objects through contact, short circuit, or dielectric breakdown. Static charge can accumulate through triboelectric charging (the act of electrically charging an object through friction) or electrostatic induction. An ESD event occurs when two oppositely charged objects are brought close together or the dielectric between them breaks down (similar to breakdown of a gate dielectric). This may involve, but not necessarily require, a visible spark between the two objects.

The sudden discharge of an electrical current can cause damage to sensitive electronics if the appropriate precautions are not taken. Manufacturers have gone to great lengths to reduce the potential for ESD by establishing electrostatic-discharge-free environments to prevent charging of materials, clothing, and individuals. This includes, but is not limited to, removal of readily-charging

materials from the environment, grounding human workers, using special conductive clothing and equipment connected to ground, using special floor tiling, and controlling humidity.

The inputs of digital and analog ICs are vulnerable to ESD and are typically protected with input protection diode structures that are intended to reduce the magnitude of voltage transients on the input of any logic or device gate structure. Normally, these diodes are configured in the reverse operating mode where only a small amount of leakage is present. When the input of a gate terminal experiences an input transient due to ESD, the forward diode characteristics clamp the voltage signal by providing a short-circuit current path either to ground or power, provided the diode forward current characteristics are sufficient to bleed off the charge on the input line. By reducing the magnitude of these voltage transients, the vulnerable gate dielectric is somewhat protected from large fields resulting from ESD on the device input terminal.

Compared to planar transistors, the breakdown voltage for FinFET circuits is much lower, causing these circuits to be more susceptible to ESD damage and making ESD a reliability concern. Increased current density and self-heating make FinFET ESD protection circuitry less efficient than their planar counterparts. Typical failure modes for FinFET ESD events include metal burn-out and interconnect failure due to current crowding (high current density). Nearly 55% of recorded ESD failures are interconnect related. Proper ESD circuit design includes simulation and modeling to identify areas of high current density (Shanmugvel, 2013). This enables design engineers to accurately place ESD diodes and clamps in their circuits. Since locations and current densities change with scaling feature sizes, ESD circuitry is often redesigned with each technology node. When engaging a manufacturer, questions addressing analysis, ESD and functional testing include: *What ESD design changes were implemented to mitigate higher current densities? What are the root causes of failure from ESD-based events, and how do these track with FinFET designs/shrinking technology nodes? What are the defect levels, and what defects account for the majority of yield loss? What structures and how much functionality are tested? What test/repair algorithms are used to address memories prone to dynamic faults? What are the manufacturing redundancies, and do they have hard/soft repair support? What is the manufacturer's repair area reserve?*

6. HIGH RELIABILITY ELECTRONICS VIRTUAL CENTER (HIREV)

The High Reliability Electronics Virtual (HiREV) center is a collection of laboratory efforts that engage in evaluating microelectronics and advanced devices for high reliability applications, including strategic applications. Their mission is to ensure timely delivery of independent high-fidelity lifetime estimates for electronic device technologies, as well as their corresponding underlying physics and chemistry of degradation and failure, to enable their qualification for critical applications. The group is a large collaboration between the Air Force Institute of Technology, Arizona State University, Georgia Institute of Technology, Iowa State University, Naval Postgraduate School, Purdue University, SUNY Albany, Vanderbilt University, AFOSR, ONR, and others.

By evaluating the physics of degradation and reliability of emerging technologies, HiREV approaches technology qualification proactively by addressing critical concerns of past, current, and next-generation technology. This includes, but is not limited to, device degradation physics, lifetime statistics, and determining the radiation response of these technologies to enable quick insertion of technology for high-reliability applications.

HiREV has previously analyzed the reliability and radiation concerns of digital 150, 90, and 65nm CMOS, Xilinx V-4 FPGAs (90nm), HNY HX5000 (150nm), and NG SONOS EEPROM, as well as analog mixed-signal evaluation of SERDES from HNY/BAE (150nm). Current efforts are focused on digital CMOS 90, 65, and 45nm technology node feature sizes, Xilinx V-5 FPGAs (65nm), and MRAM from HNY and Aeroflex, as well as analog/mixed signal for CMOS 90, 65, 45nm nodes and SiGe BiCMOS (130, 90nm), and finally SiC- and GaN-based power electronics. Their future efforts include evaluating the 32, 28, and 22nm CMOS technology nodes, Xilinx V-7 FPGAs (28nm), BAE/Achronix (150nm), carbon nanotube transistors, and Freescale 45nm SOI. Other future efforts include 32nm CMOS for mixed signal and analog applications, InP HBTs, and continued efforts for SiC- and GaN-based power electronics.

7. RADIATION EFFECTS FOR ADVANCED CMOS TECHNOLOGY

Radiation interacts with matter through physical processes that can be described either as ionizing or nonionizing. Ionizing radiation effects can further be divided into destructive and transient (non-destructive) effects. Here we will consider the effects of total ionizing dose (accumulation of charge) and transient collection of charge. We will also consider non-ionizing lattice damage (displacement damage or knock-ons) on the response of devices and circuits.

7.1. Total Ionizing Dose (TID)

The interaction of energetic charged particles, x-rays, and γ-rays with electronic materials results in the generation of *e-h* pairs in semiconductors and oxides. This is known as ionization; the result of ionization is that valence band electrons are excited to the conduction band, where they are highly mobile under the presence of an electric field. Ionization causes the material (semiconductor or oxide) to conduct more current than it would typically. When a valence band electron is excited into the conduction band it leaves behind a hole that is much less mobile than an electron in the conduction band. The production and subsequent trapping of holes in oxides has been reported to lead to significant parametric degradation of MOS and bipolar devices.

In MOS devices, holes are trapped in the gate oxide and in field oxides or sidewall oxides. MOS gate oxides typically trap holes, resulting in a semi-permanent charge sheet that will alter applied electric fields, alter bonds at the Si-SiO_2 interface, and change the density and energy distribution of defects at the oxide-semiconductor interface. At the device level, hole trapping in the bulk of an oxide results in negative threshold voltage shifts in n- and p-MOSFET devices, causing the nMOSFET to "turn-on" more easily, while the pMOSFET becomes harder to switch from the *off*- to the *on*-state. There is typically a change in the subthreshold behavior of MOSFETs due to the accumulation of interface traps. Simultaneously, the buildup of interface traps can cause a positive or negative threshold voltage shift in nMOSFET devices and pMOSFET devices, depending on their occupancy and the position of the Fermi level at the interface. The rates at which oxides trap holes and populate interface traps are not equal, and depend on the oxide quality, bias condition of the device, surface passivation techniques, and thickness of the material. Consequently, the balance of threshold voltage and subthreshold performance of MOSFETs irradiated with x-rays, γ-rays, and light ions (electrons and protons) changes with increasing dose and may undergo non-monotonic behavior (rebound effects) where the magnitude of the threshold voltage shift may change with increasing time.

At the circuit level, these effects are best correlated with hot-carrier and bias-temperature instability effects since they cause parametric shifts in critical operating points and lead to timing degradation of circuits. Total ionizing dose effects are typically most detrimental in SRAM cells and synchronous circuits where timing penalties cause significant system performance degradation. These effects are typically considered dynamic in that some of the effects can be annealed away with increased temperature and time.

7.1.1. Comments on TID in Planar and FinFET Geometries

Recent progression of CMOS technology downscaling beyond 32 nm is proceeding to FinFET technology for high performance or Ultra-Thin-Body and BOX (UTBB) planar technology for low-power applications. Channel length scaling of partially-depleted (PD) CMOS/SOI has resulted in technology progressively less sensitive to TID effects. Commercial 32 nm PD-SOI nMOSFET transistors are much less sensitive to TID than are those from earlier commercial 150 nm PD-SOI technology. These trends are shown in Figure 30. As shown in Figure 30, preliminary results indicate that scaling from 32 nm to 14 nm results in an increase in TID sensitivity or an inflection in the TID

vulnerability trend. Bulk and SOI 14nm FinFETs and UTBB SOI show this reversal in the TID response. For example, the TID-induced increase in off-state current for 14 nm SOI FinFET technology is nearly the same as 90 nm PD CMOS/SOI.

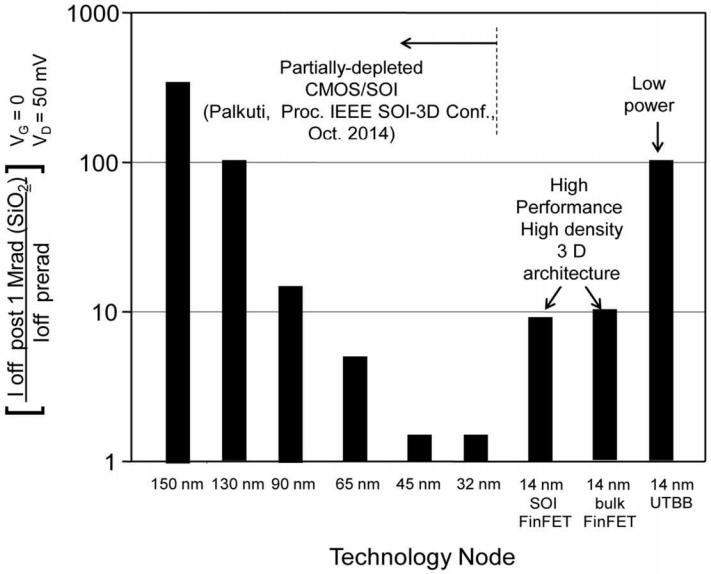


Figure 30. Relative change in leakage current for off-state irradiation of PD-SOI in technology nodes from 150nm through 32nm. The recent adoption of 3-D transistor geometries has resulted in an increased impact of TID on the latest technology.

Experimental results and modeling have shown FD-SOI technologies to be extremely soft to TID [40, 41]. In fact, their use as radiation dosimeters has long been recommended [42]. The sale of IBM's foundry business has meant the abandonment of 14nm SOI with no commercial alternatives. Therefore, we continue the discussion focusing on the bulk 14nm FinFET variants.

7.1.2. TID Effects in a Bulk 14nm FinFET Technology

Figure 31 shows I_{ds}-V_{gs} characteristics representative of the low- (upper) and high- (lower) V_{th} transistors irradiated under *on*-state bias conditions. The transistors were irradiated in incremental steps to a total dose of 1 Mrad(SiO_2). Several features present in the *I-V* curves warrant discussion. The radiation-induced shift in $I_{ds,off}$ and V_{th} for three irradiation bias conditions is shown in Figure 31 for several irradiation conditions.

First, the most dramatic impact of TID on device performance is clearly the modulation of *off*-state leakage current, $I_{ds,off}$ with increasing dose. The change in *off*-state leakage is not simply a static increase in drain current, but has a strong dependence on the applied gate voltage during irradiation. This is consistent with previous reports and modeling efforts [43–50] that attribute TID-induced leakage currents to charge accumulation in the STI near the sub-fin region of the device that contains the CSD or Super-Steep Retrograde Well (SSRW) [51].

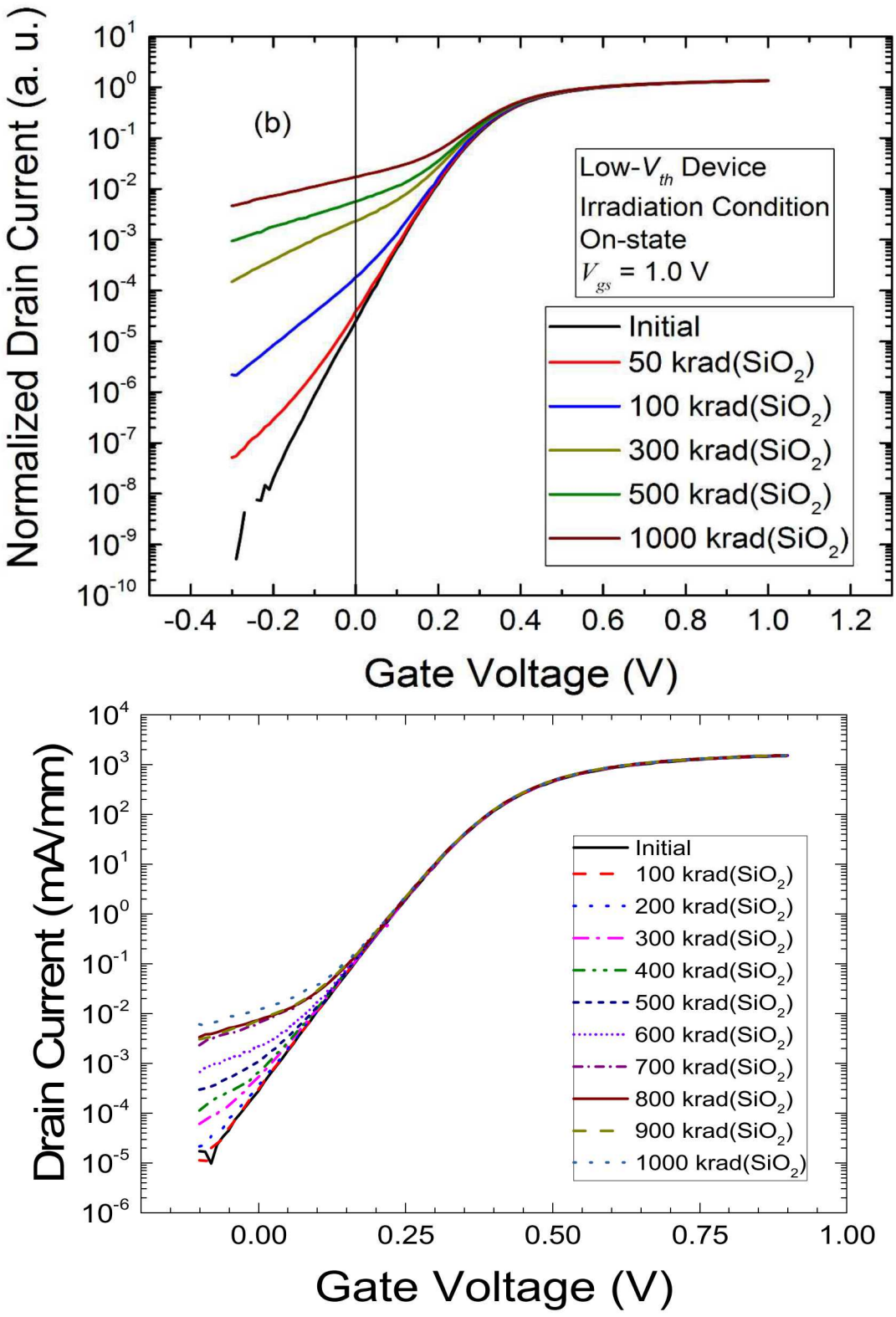


Figure 31. The difference in low- (upper) and high- (lower) Vth transistor response to TID when irradiated in the on-state. Large increases in leakage current are observed for the lower threshold voltage device variant.

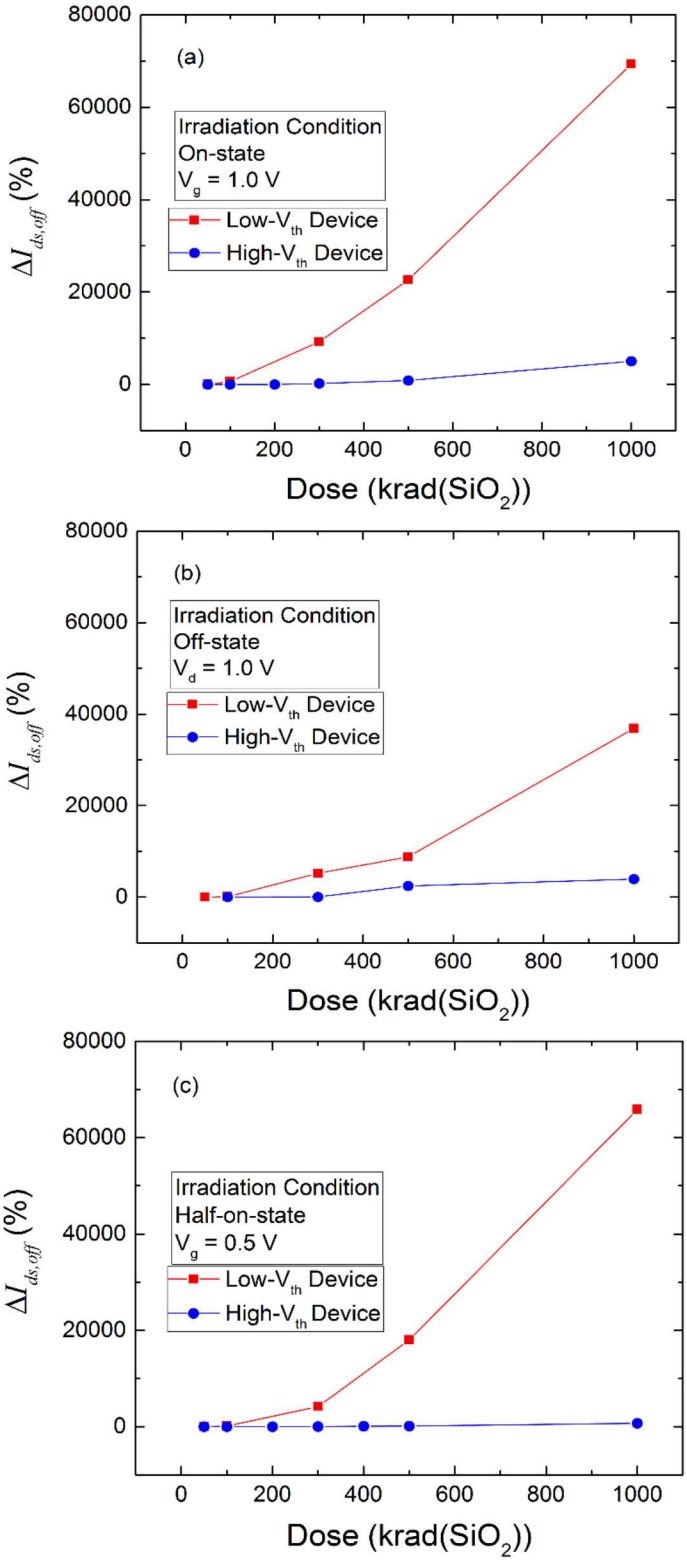


Figure 32. Percent change in Ids,off as a function of dose for the low-Vth and high-Vth 14nm FinFET devices in the off- (a), on- (b), and half-on-state (c) during irradiation shows the worst-case bias condition for leakage current response is the on-state condition. The higher threshold voltage device shows a less severe response for all bias conditions.

As is well known, the magnitude of change in *off*-state leakage current is dependent on the bias condition during irradiation. The most dramatic shift in off-state leakage current occurs for the nMOSFET irradiated in the *on*-state with the gate high and all other pins grounded. This is clearly illustrated in Figure 32 where the percent change in off-state leakage current as a function of total dose for each of the irradiation bias conditions for the low-V_{th} and high-V_{th} devices is shown as a red or a blue trace, respectively. The data indicate that the on-state is the worst-case bias condition for nMOSFETs in a FinFET geometry with short channel length. This is different than the results from Chatterjee in [52] that concluded the worst-case was the off-state condition. Differences in geometry and manufacturing processes, and therefore electric field profiles, give rise to differences in charge yield and trapping behavior between our work and devices studied previously. Second, the change in V_{th} is observed to be minimal with a corresponding trend in the $I_{ds,on}$ parameter (not shown). The typical change in threshold voltage is on the order of 30-50 mV at a total dose of 1 Mrad(SiO_2) with a corresponding 3-5% change in total drive current at V_{gs} = 1 V in the *on*-state with V_{ds} = 50 mV. Lastly, the $I_{ds,on}/I_{ds,off}$ current ratio decreases from ~10^6 pre-irradiation to closer to ~10^2-10^3 at 1 Mrad(SiO_2).

The change in $I_{ds,on}/I_{ds,off}$ ratio with radiation preferentially impacts nMOSFET devices, with implications for timing of digital circuits due to the resulting drain current drive mismatch between *n*- and *p*-channel devices. Additionally, overall static and dynamic current dissipation of the device will increase significantly, impacting logic states in digital and memory applications [53]. For the low-V_{th} device, this can be very slightly mitigated in part by employing a reduced supply voltage condition, as seen in Figure 32(c). Intuitively, a lower operating voltage should result in lower electric fields that will reduce charge yield in the STI and impact charge trapping in the STI dielectric material. A reduction in operating voltage has the obvious advantage of reducing power dissipation and enabling more energy-efficient applications for near-threshold computing [54] or dynamic voltage and frequency scaling [55], of particular use when power and not speed is a primary design consideration.

7.1.3. TID Response of Low-V_{th} vs High-V_{th} Transistors

Figure 32 also shows the radiation-induced change in *off*-state leakage for low-V_{th} as compared to high-V_{th} devices irradiated in the *on*- (31 (a)), *off*- (31 (b)) and *half-on* (31 (c)) bias conditions. The high-V_{th} device shows a significantly smaller response in $\Delta I_{ds,off}$ post-irradiation than does the low-V_{th} device for all bias conditions evaluated in this work. This is presumably due to higher fin doping and threshold adjust implant of the high-V_{th} device and differences in the CSD and SSRW device isolation required to achieve a higher V_{th} device as compared to a low-V_{th} device. Conceptually, higher body doping would require more charge trapping in the STI dielectric than what is required for a lower body doping to deplete and invert the STI side-wall transistor. Consistent with the low-V_{th} results observed in the previous section, we see that the *on*-state irradiation condition induces the worst case parasitic leakage current response for high-V_{th} devices (*on*-state condition $\Delta I_{ds,off} \approx$ 5000%, *off*-state condition $\Delta I_{ds,off}$ = 3900%). Lower supply voltage conditions represent a marginal improvement in the post-irradiation TID response of these devices for both the high- and low-V_{th} devices, with $\Delta I_{ds,off}$ of 710% for the high-V_{th} device. Intuitively, a lower operating voltage should result in smaller electric fields and impact the charge yield and trapping behavior for the gate and STI dielectrics [56]. The reduction in trapped charge, in turn, has less of an impact on devices with higher threshold-voltage-adjust implant, requiring a higher total dose to deplete and eventually invert the sidewall region.

7.1.3.1. Single Event Effects (SEEs)

Single-Event Effects (SEEs) are caused by a single, energetic particle and can be divided into non-destructive effects, also called transient effects, and destructive effects that lead to permanent,

potentially catastrophic, damage of devices and circuits. We will discuss a few types of SEEs that are important for various types of circuits. In the space environment, spacecraft designers have traditionally been concerned with SEEs resulting from cosmic rays and high-energy protons and electrons trapped in the radiation belts surrounding Earth and emitted from the Sun. For cosmic rays, SEEs are typically caused by heavy ions, specifically ions with the mass of a proton or greater. At the terrestrial level, the primary concerns for SEEs are from atmospheric neutrons, muons, and alpha particles from packaging contamination. There are additional environments found at the terrestrial level that can lead to a concern for SEEs at the terrestrial level, such as nuclear reactor environments, high energy physics facilities (accelerators), and weapons environments.

7.1.3.2. Transient Effects

Single-Event Upsets (SEUs) are a type of non-destructive soft error and normally appear as transient pulses in logic or support circuitry, or in the change in state of a memory cell. When charge collection due to a single ionizing particle occurs in logic, the resulting circuit response is referred to as a Single-Event Transient (SET) and may not necessarily lead to an SEU in memory. SEUs may result from the propagation of an SET through logic that becomes latched in memory, or the direct interaction of a charged particle with the sensitive region of an SRAM cell or latch. The sensitive regions of an SRAM are considered to be the reverse-biased *pn* junctions of the cross-coupled inverters. The charge collection mechanisms include: charge drift in the drain region when an ion crosses the drain, and charge diffusion via *pn* junctions to the drain region. As devices have continued to scale below the 65nm technology node feature size, this estimation has become increasingly complicated, but to first order is a reasonable approximation. A calculation has been developed estimating the magnitude of soft errors based on the charge collection mechanisms stated above.

$$\boldsymbol{S = FAs/exp(-\frac{Qcoll}{Qcrit})}$$

Where S is the number of soft errors, F is the particle flux incident on the silicon surface, A_s is the sensitive area for charge collection, Q_{coll} is the amount of charge collected and Q_{crit} is the critical amount of charge needed to flip the bit. The exponential function represents the probability of an error occurring. This probability is dependent on the energetic particle involved. In general, the downward trend of S with shrinking FinFET feature sizes occurs because the A_s decreases faster than the decrease in Q_{crit}. [57]

During the transition from planar to FinFET transistors, a reduction in sensitive charge collection volume occurred. The initial charge collection volume decreases from planar to FinFET architecture showed a reduction in Soft Error failure Rates (SER). As FinFET feature sizes continued to shrink, further decreases in SER were observed. As shown in Figure 33, the transition from planar to FinFET technology and shrinking of FinFET feature size shows a reduction in SER as the charge collection area shrinks. These findings have been observed and published by FinFET manufacturers Samsung and TSMC. When addressing SER with FinFET manufacturers, it is important to ask the following questions: *What environments does the manufacturer consider when determining SERs? What memory types is the manufacturer testing/qualifying? Does the manufacturer have evidence of soft error trends improving? Does SER reduction scale with FinFET technology nodes (feature sizes)?*

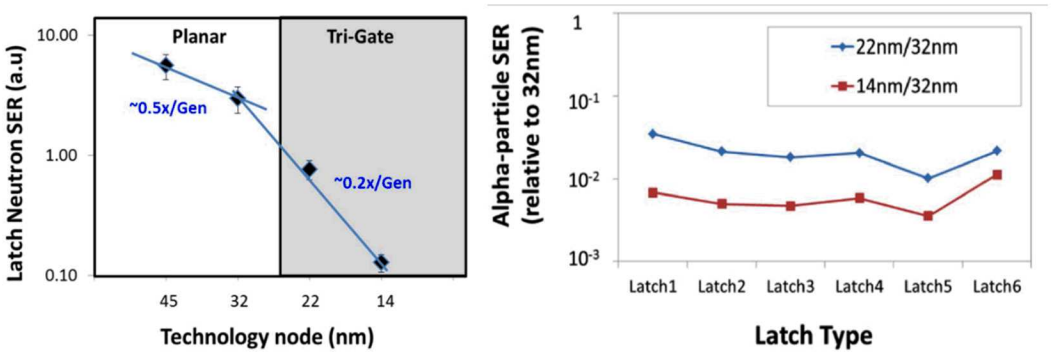


Figure 33. Reduction in SER with shrinking technology nodes in a) planar and FinFET devices, and b) 14nm and 22nm FinFET compared to 32nm planar transistor SER [58].

In addition to single-event susceptibility, Multi-Cell Upset susceptibility (MCU) can also occur. This phenomenon can be caused by two different mechanisms: 1) bipolar amplification of charge collection formed in the source, bulk or drain and 2) charge sharing and diffusion of charge to adjacent cells. The charge sharing and diffusion is the dominant mechanism [57], [59], [60]. Charge sharing indicates FinFET cell geometry significantly impacts the probability of MCUs occurring. Results for single-bit and multi-cell upsets in planar and FinFET technologies manufactured by TSMC are shown in Figure 34. a) alpha, thermal neutron and high-energy neutron single bit upsets for 40, 28 and 20nm planar transistors compared to 16nm FinFET, and b) High energy neutron induced multi-cell upsets across same technology nodes.

.

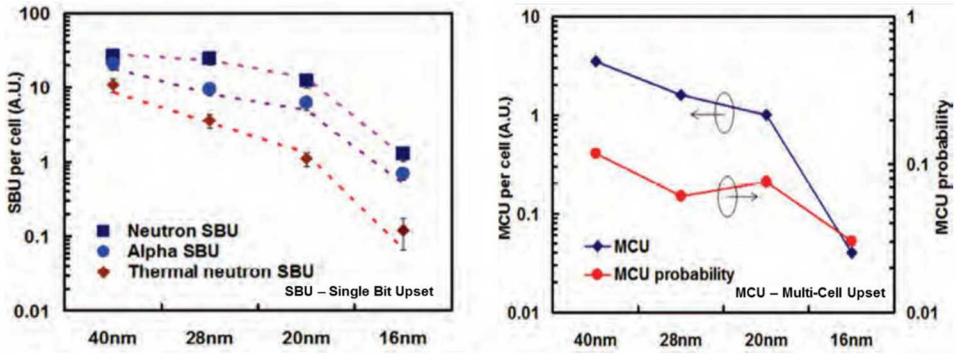


Figure 34. a) alpha, thermal neutron and high-energy neutron single bit upsets for 40, 28 and 20nm planar transistors compared to 16nm FinFET, and b) High energy neutron induced multi-cell upsets across same technology nodes.

Typically, manufacturers with well-controlled processes will account for these errors by including 5–10% margin in their design. Higher redundancy percentages would indicate a potential lack of control over certain process variables. When discussing SEU/MCU issues with FinFET manufacturers, a pertinent question is: *How much redundancy does the manufacturer have in their memory array to compensate for cell failure loss?*

Protons, usually trapped in the earth's radiation belts from solar flares or coronal mass ejections may cause SEEs in very sensitive devices, specifically in sub-65nm technology node feature sizes. Protons may more typically cause a nuclear reaction near a sensitive device area, and thus create an indirect ionization effect potentially causing an SEE.

7.1.3.3. Destructive Effects

In contrast to the soft errors discussed above, hard errors can potentially have destructive, catastrophic effects at the material, device, and circuit levels. Single-Event Latchup (SEL) is the

analogue to electrical latchup, but in the radiation environment. In addition to electrical latchup, when struck by a heavy ion with sufficient Linear Energy Transfer (LET) the *n-p-n-p* structure will enter the latchup state. SEL is a positive feedback loop, just like electrical latchup, and will persist until power to the circuit is cycled. The high-current condition induced by SEL can lead to self-heating and melting of the material system and is a primary concern for high-reliability applications in the space radiation, weapons, and nuclear reactor environments.

A second potentially destructive mechanism primarily impacts power semiconductor devices and is known as Single-Event Burnout (SEB). SEB occurs due to a high current state in a power MOSFET, leading to significant device heating and destruction of the material system. SEB occurs when the MOSFET is in the *off*-state, with a high voltage on the drain and the gate potential below threshold, and a heavy ion generates a large amount of charge in the drift region of the device that is collected due to the high applied field between source and drain. There are few ways to prevent this type of effect except use of lateral devices in place of vertical devices. Other types of SEB exist, including Single-Event Gate Rupture (SEGR), which is the formation of a conducting path through the MOSFET gate oxide. This causes a high current to flow through the thick gate stack, leading to a permanent degradation of device functionality.

7.2. Displacement Damage

An incident particle transporting through solid-state materials will displace atoms from their equilibrium positions through Coulomb scattering, nuclear elastic collisions, and nuclear inelastic collisions. For this type of interaction, the displacement of a single atom results in a point defect or vacancy-interstitial pair, where the vacancy is the empty lattice location and the interstitial is the relaxation of the displaced atom into a non-equilibrium position. The result is a defect in the material system. Defect generation is not limited to single atom displacements; in fact, large clusters of defects can be generated by neutron events and for charged particles near their end of range.

In semiconductor materials, the generation of defects can lead to increased resistance in the region in which damage was created, introduction of defect levels within the bandgap of the material, and the associated generation-recombination centers. At the device level, the introduction of defects can lead to increased resistivity of substrates and reduction in the electrostatic potential of the well/substrate. In depletion regions, the introduction of generation-recombination centers leads to an increased generation current, disturbing the *off*-state characteristics of Charge-Coupled Devices (CCDs), pixel sensors, or *pn* junctions. The presence of defects in the depletion region also disturbs the free carrier population, which in turn impacts the capacitance response of the *pn* junction, and device switching time decreases.

Several categories of devices have been shown to exhibit sensitivity to single particle displacement damage, from power rectifiers (as described in the previous paragraph), stuck bits in SDRAMs, change of the breakdown voltage for avalanche diodes, and reduction of the noise floor and dark pixels in CCDs and pixel sensors.

7.3. Commercial off-the-shelf electronics and radiation effects

As technology has driven more and more to reduce parameters such as power, weight, volume, and cost, while requiring increased functionality, emerging commercial technologies have been pressed into service in radiation environments. These technologies include high-speed and low-power CMOS and fiber optics. The types of integrated circuits (ICs) that utilize these technologies range from complex microprocessors to dense SRAMs. These technologies, without qualification and understanding of the response of mission-critical devices and circuits to ionizing and non-ionizing radiation, can lead to catastrophic mission failure.

8. IMPACT OF TRANSISTOR RELIABILITY DEGRADATION ON CIRCUIT/SYSTEM PERFORMANCE AND MITIGATION THROUGH DESIGN

In previous sections we discussed some of the key semiconductor device parametric reliability degradation mechanisms. In this section, we will discuss higher level device integrated reliability impacts associated with accumulated transistor parametric degradation and will not highlight random environmental events like radiation. We will also discuss required efforts to mitigation their impact through circuit design optimization utilizing reliability modeling.

At a systems level, there has been significant work to define, assess and model the discussed cumulative transistor reliability aging affects and provided methodologies for their mitigation. Due to the numerous possible system level failure modes, in this section, we will focus primarily on "soft" logic failure modes associated with signal propagation timing related failure mechanisms. These are important since they can occur long before the "hard" physical failure of the active semiconductor device defined by many transistor reliability models. These soft errors can be very difficult to detect and mitigate. These aging related logic timing failures are statistical in nature and they typically express themselves initially as infrequent erratic logic failure events. The frequency of these erratic events increase as more transistor/interconnect level timing degradation accumulates. If the circuit and system design is optimized, these erratic failures will only occur beyond device End-of-Life as our designed in timing margins go to zero.

Most of the parametric degradation mechanisms discussed in Section 3, BTI, HCI, gate oxide breakdown and EM interconnect resistance lead to significant non-recoverable changes in the transistor and interconnection operating performance. Specifically, this transistor aging leads to CMOS circuit level timing changes and ultimately to system logic related failures. Efforts have been made to statistically model these known reliabilities related timing shifts to allow for the prediction of circuit failures and optimization of circuit margins through layout and design improvements[61,64,30,65].

The switching speed of a CMOS cell transistor is measured by its ability to drive the required current to change the state of the next logic CMOS gate. This is directly related to its interconnect output and input RC impedances as well as CMOS n/p cell balance. This has the net effect of generating switching related timing delays in the signal propagation to the next logic gate.

As detailed in Section 3, the amount of transistor parametric reliability degradation is based upon the local transistor operating conditions such as signal frequency, V_g, V_{ds}, voltage duty cycle, temperature and stress times. Figure 35 provides a schematic view of a digital signal highlighting the portion of each signal cycle that can generate BTI and HCI related transistor V_{th} and I_d current degradation. Gate dielectric degradation occurs when V_{gs} is maximized resulting in increased defect related gate leakage to flow (SILC). Although the transistor can still switch, the added SILC breakdown current increases the required input switching loads resulting in effective propagation delays. Hence, by controlling the local signal duty cycles and voltages, the level of accumulated delay related degradation can be optimized.

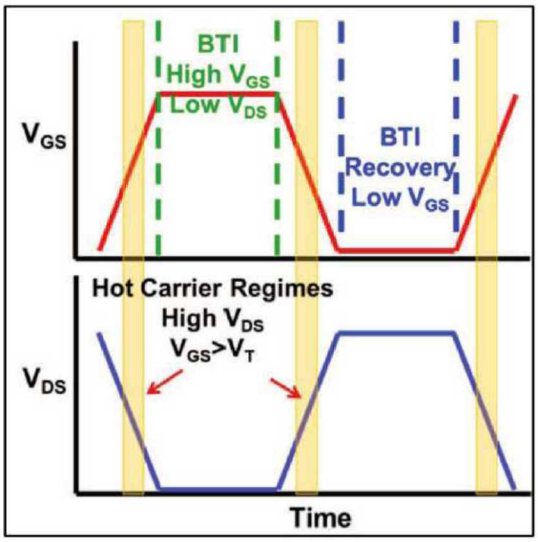


Figure 35.An example of a digital switching waveform on an inverter (only NMOS traces are shown, PMOS traces would be complementary). Sections where BTI, recovery, and hot carrier (HCI) degradation occur are called out. HCI only occurs during switching events in digital logic. From Prasad Ref[30].

Device level computational and logic signal integrity is affected by the local summation of all input transistor level timing characteristics and supporting clock references. In many cases, these can be modelled using Spice like tools, and guard banded to ensure that the system logic will support the required functional reliability lifetime. This becomes progressively more difficult as logic blocks and physical die sizes grow. Larger and more complex logic die require routing paths incorporating large numbers, millions, of CMOS gates and multiple interconnect metal lines which contribute to a significant and varied signal delay which must be matched to the appropriate reference clocks when they are logically combined in subsequent function blocks. These timing delays across the device can become very large and vary significantly due to local transistor aging and operating temperature. Spice and Aging-aware based statistical circuit reliability analysis methods have been developed to help model and mitigate the reliability impact through circuit and layout design[30, 39, 62, 63].

A key take-away from this discussion should be considered for security related logic blocks such as authentication and encryption engines. As the circuit/system ages, the logic operating timing margins decrease due to the localized accumulated transistor level reliability degradation described above. This timing degradation will eventually drive the logic circuit from a consistent "Passing" operational state to a marginal/erratic and then to a fully stuck-at state with further environmental aging. As the system's operating lifetime approaches End-of-Life, the marginal/erratic operational state may also become more susceptible to normal application temperature and stress environments. If a security related logic block becomes erratic, then the loss of key security features and information is possible. This has been an important driver for the generation of aging-aware design methodologies described above to ensure that local logic blocks will not fail prior to the expected application reliability lifetime.

Questions regarding the susceptibility of an IC design to reliability aging related degradation should be addressed to the appropriate circuit design team members who can utilize the current advanced logic modelling techniques to assess and mitigate these risks. Most foundries will recommend tools that are consistent with their available parametric design data.

9. CONCLUSION

This work highlights many of the important reliability considerations that need to be comprehended in current CMOS die and packaging technologies since they now strongly interact. Although Moore's Law is slowing for silicon technology node implementations, it continues to drive the integration of advanced product and systems features. The modelling and mitigation of these reliability concerns are required to meet product application specific reliability needs. As discussed in this document, the ability to meet all reliability expectations can be accomplished through the application of both product design and manufacturing optimization.

REFERENCES

[1] C. Xu, S. K. Kolluri, K. Endo, and K. Banerjee, "Analytical Thermal Model for Self-Heating in Advanced FinFET Devices With Implications for Design and Reliability," IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 32, no. 7, pp. 1045–1058, Jul. 2013.

[2] S. E. Liu et al., "Self-heating effect in FinFETs and its impact on devices reliability characterization," in 2014 IEEE International Reliability Physics Symposium, 2014, p. 4A.4.1-4A.4.4.

[3] C. Prasad et al., "Self-heat reliability considerations on Intel's 22nm Tri-Gate technology," in Reliability Physics Symposium (IRPS), 2013 IEEE International, 2013, p. 5D.1.1-5D.1.5.

[4] D. K. Schroder, "Negative bias temperature instability: What do we understand?," Microelectron. Reliab., vol. 47, no. 6, pp. 841–852, Jun. 2007.

[5] S. Ramey et al., "Intrinsic transistor reliability improvements from 22nm tri-gate technology," in Reliability Physics Symposium (IRPS), 2013 IEEE International, 2013, p. 4C.5.1-4C.5.5.

[6] D. K. Schroder and J. A. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," J. Appl. Phys., vol. 94, no. 1, pp. 1–18, Jul. 2003.

[7] T. Grasser, H. Reisinger, P.-J. Wagner, F. Schanovsky, W. Goes, and B. Kaczer, "The time dependent defect spectroscopy (TDDS) for the characterization of the bias temperature instability," in Reliability Physics Symposium (IRPS), 2010 IEEE International, 2010, pp. 16–25.

[8] K. Zhao, J. H. Stathis, A. Kerber, and E. Cartier, "PBTI relaxation dynamics after AC vs. DC stress in high-k/metal gate stacks," in Reliability Physics Symposium (IRPS), 2010 IEEE International, 2010, pp. 50–54.

[9] M. Agarwal, B. C. Paul, M. Zhang, and S. Mitra, "Circuit Failure Prediction and Its Application to Transistor Aging," in 25th IEEE VLSI Test Symposium, 2007, 2007, pp. 277–286.

[10] M. Yabuuchi and K. Kobayashi, "Circuit characteristic analysis considering NBTI and PBTI-induced delay degradation," in Future of Electron Devices, Kansai (IMFEDK), 2012 IEEE International Meeting for, 2012, pp. 1–2.

[11] E. Wu, E. Nowak, J. Aitken, W. Abadeer, L.-K. Han, and S. Lo, "Structural dependence of dielectric breakdown in ultra-thin gate oxides and its relationship to soft breakdown modes and device failure," in Electron Devices Meeting, 1998. IEDM '98. Technical Digest., International, 1998, pp. 187–190.

[12] J. H. Stathis, "Percolation models for gate oxide breakdown," J. Appl. Phys., vol. 86, no. 10, pp. 5757–5766, Nov. 1999.

[13] J. W. McPherson and H. C. Mogul, "Underlying physics of the thermochemical E model in describing low-field time-dependent dielectric breakdown in SiO2 thin films," J. Appl. Phys., vol. 84, no. 3, pp. 1513–1523, Aug. 1998.

[14] J. W. McPherson, R. B. Khamankar, and A. Shanware, "Complementary model for intrinsic time-dependent dielectric breakdown in SiO2 dielectrics," J. Appl. Phys., vol. 88, no. 9, pp. 5351–5359, Nov. 2000.

[15] E. Y. Wu and J. Suñé, "Power-law voltage acceleration: A key element for ultra-thin gate oxide reliability," Microelectron. Reliab., vol. 45, no. 12, pp. 1809–1834, Dec. 2005.

[16] E. Miranda and J. Sune, "Analytic modeling of leakage current through multiple breakdown paths in SiO2 films," in Reliability Physics Symposium, 2001. Proceedings. 39th Annual. 2001 IEEE International, 2001, pp. 367–379.

[17] B. Kaczer, R. Degraeve, M. Rasras, K. Van de Mieroop, P. J. Roussel, and G. Groeseneken, "Impact of MOSFET gate oxide breakdown on digital circuit operation and reliability," IEEE Trans. Electron Devices, vol. 49, no. 3, pp. 500–506, Mar. 2002.

[18] R. Degraeve et al., "Statistical model for stress-induced leakage current and pre-breakdown current jumps in ultra-thin oxide layers," in Electron Devices Meeting, 2001. IEDM '01. Technical Digest. International, 2001, p. 6.2.1-6.2.4.

[19] I. A. Blech, "Electromigration in thin aluminum films on titanium nitride," J. Appl. Phys., vol. 47, no. 4, pp. 1203–1208, Apr. 1976.

[20] I. A. Blech and C. Herring, "Stress generation by electromigration," Appl. Phys. Lett., vol. 29, no. 3, pp. 131–133, Aug. 1976.

[21] H. W. Yao, P. Justison, and J. Poppe, "Stress-induced-voiding risk factor and stress migration model for Cu interconnect reliability," in Reliability Physics Symposium (IRPS), 2013 IEEE International, 2013, p. 2C.5.1-2C.5.8.

[22] M. Jin et al., "Hot carrier reliability characterization in consideration of self-heating in FinFET technology," 2016, p. 2A–2–1–2A–2–5.

[23] K. N. Tu, "Recent advances on electromigration in very-large-scale-integration of interconnects," J. Appl. Phys., vol. 94, no. 9, pp. 5451–5473, Nov. 2003.

[24] J. R. Black, "Electromigration failure modes in aluminum metallization for semiconductor devices," Proc. IEEE, vol. 57, no. 9, pp. 1587–1594, Sep. 1969.

[25] J. R. Black, "Electromigration - A brief survey and some recent results," IEEE Trans. Electron Devices, vol. 16, no. 4, pp. 338–347, Apr. 1969.

[26] F. Chen et al., "Investigation of emerging middle-of-line poly gate-to-diffusion contact reliability issues," in Reliability Physics Symposium (IRPS), 2012 IEEE International, 2012, p. 6A.4.1-6A.4.9.

[27] T. Kauerauf et al., "Reliability of MOL local interconnects," in Reliability Physics Symposium (IRPS), 2013 IEEE International, 2013, p. 2F.5.1-2F.5.5.

[28] Nishi, Y, & Doering, R (eds) 2007, Handbook of Semiconductor Manufacturing Technology, p 32-7, CRC Press LLC, Baton Rouge.

[29] Y. Li, Deepak Goyal Editors (2017), "3D Microelectronic Packaging from Fundamentals to Applications", Springer, pp 2.

[30] Prasad, C., S. Ramey, and L. Jiang. "Self-heating in advanced CMOS technologies." In 2017 IEEE International Reliability Physics Symposium (IRPS), pp. 6A-4. IEEE, 2017.

[31] Oprins, H., V. O. Cherman, B. Vandevelde, G. Van der Plas, P. Marchal, and E. Beyne. "Numerical and experimental characterization of the thermal behavior of a packaged DRAM-on-logic stack." In 2012 IEEE 62nd Electronic Components and Technology Conference, pp. 1081-1088. IEEE, 2012.

[32] Kirk Saban, "Xilinx Stacked Silicon Interconnect Technology Delivers Breakthrough FPGA Capacity, Bandwidth, and Power Efficiency", Xilinx White Paper, WP380 (v1.2) December 11, 2012.

[33] T. A. Tran, D. Dorinski, S. Gonzales, P. Vu, and K. Y. Yow, "Packaging challenges in low-k silicon with thermally enhanced ball grid array (TE-PBGA)," in Electronics Packaging Technology Conference (EPTC), 2010 12th, 2010, pp. 179–184.

[34] G. S. Leatherman, J. Xu, J. Hicks, B. Kilic, and D. Pantuso, "Die-package stress interaction impact on transistor performance," in Reliability Physics Symposium (IRPS), 2012 IEEE International, 2012, p. 2E.4.1-2E.4.6.

[35] G. S. Leatherman, J. Hicks, B. Kilic, D. Pantuso and G. Xu, "Transistor Performance Impact Due to Die–Package Mechanical Stress," in IEEE Transactions on Device and Materials Reliability, vol. 13, no. 2, pp. 350-356, June 2013.

[36] Gambino, Jeffrey P., Shawn A. Adderly, and John U. Knickerbocker. "An overview of through-silicon-via technology and manufacturing challenges." Microelectronic Engineering 135 (2015): 73-106.

[37] Tu, K. N. "Reliability challenges in 3D IC packaging technology." Microelectronics Reliability 51, no. 3 (2011): 517-523.

[38] Premachandran, C. S., Salvatore Cimino, Sean Ogdan, Zhuo-Jie George Wu, Daniel Smith, Sukesh Kannan, Linjun Cao et al. "Comprehensive 3D TSV Reliability Study on 14nm FINFET technology with thinned wafers." In 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), pp. 37-40. IEEE, 2018.

[39] Agarwal, Mridul, Bipul C. Paul, Ming Zhang, and Subhasish Mitra. "Circuit failure prediction and its application to transistor aging." In 25th IEEE VLSI Test Symposium (VTS'07), pp. 277-286. IEEE, 2007.

[40] P. Paillet et al., "Total ionizing dose effects on deca-nanometer fully depleted SOI devices," IEEE Trans. Nucl. Sci., vol. 52, no. 6, pp. 2345–2352, Dec. 2005.

[41] I. S. Esqueda et al., "Modeling the Radiation Response of Fully-Depleted SOI n-Channel MOSFETs," IEEE Trans. Nucl. Sci., vol. 56, no. 4, pp. 2247–2250, Aug. 2009.

[42] J. B. Yau et al., "FDSOI radiation dosimeters," in 2011 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA), 2011, pp. 1–2.

[43] M. R. Shaneyfelt, P. E. Dodd, B. L. Draper, and R. S. Flores, "Challenges in hardening technologies using shallow-trench isolation," IEEE Trans. Nucl. Sci., vol. 45, no. 6, pp. 2584–2592, Dec. 1998.

[44] F. Faccio and G. Cervelli, "Radiation-induced edge effects in deep submicron CMOS transistors," IEEE Trans. Nucl. Sci., vol. 52, no. 6, pp. 2413–2420, Dec. 2005.

[45] I. Chatterjee et al., "Bias Dependence of Total-Dose Effects in Bulk FinFETs," IEEE Trans. Nucl. Sci., vol. 60, no. 6, pp. 4476–4482, Dec. 2013.

[46] H. Hughes et al., "Total Ionizing Dose Radiation Effects on 14 nm FinFET and SOI UTBB Technologies," 2015 IEEE Radiat. Eff. Data Workshop REDW, pp. 1–6, Boston, MA 2015.

[47] F. Faccio et al., "Total ionizing dose effects in shallow trench isolation oxides," Microelectron. Reliab., vol. 48, no. 7, pp. 1000–1007, Jul. 2008.

[48] M. Turowski, A. Raman, and R. D. Schrimpf, "Nonuniform total-dose-induced charge distribution in shallow-trench isolation oxides," IEEE Trans. Nucl. Sci., vol. 51, no. 6, pp. 3166–3171, Dec. 2004.

[49] I. S. Esqueda, H. J. Barnaby, K. E. Holbert, F. El-Mamouni, and R. D. Schrimpf, "Modeling of Ionizing Radiation-Induced Degradation in Multiple Gate Field Effect Transistors," IEEE Trans. Nucl. Sci., vol. 58, no. 2, pp. 499–505, Apr. 2011.

[50] A. H. Johnston, "Super Recovery of Total Dose Damage in MOS Devices," IEEE Trans. Nucl. Sci., vol. 31, no. 6, pp. 1427–1433, Dec. 1984.

[51] N. Damrongplasit et al., "Comparative Study of Uniform Versus Supersteep Retrograde MOSFET Channel Doping and Implications for 6-T SRAM Yield," IEEE Trans. Electron Devices, vol. 60, no. 5, pp. 1790–1793, May 2013.

[52] I. Chatterjee et al., "Impact of Technology Scaling on SRAM Soft Error Rates," IEEE Trans. Nucl. Sci., vol. 61, no. 6, pp. 3512–3518, Dec. 2014.

[53] M. R. Shaneyfelt et al., "Wafer-level radiation testing for hardness assurance," IEEE Trans. Nucl. Sci., vol. 38, no. 6, pp. 1598–1605, Dec. 1991.

[54] R. G. Dreslinski, M. Wieckowski, D. Blaauw, D. Sylvester, and T. Mudge, "Near-Threshold Computing: Reclaiming Moore's Law Through Energy Efficient Integrated Circuits," Proc. IEEE, vol. 98, no. 2, pp. 253–266, Feb. 2010.

[55] G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott, "Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling," in Eighth International Symposium on High-Performance Computer Architecture, 2002. Proceedings, 2002, pp. 29–40.

[56] M. R. Shaneyfelt, D. M. Fleetwood, J. R. Schwank, and K. L. Hughes, "Charge yield for cobalt-60 and 10-keV X-ray irradiations of MOS devices," IEEE Trans. Nucl. Sci., vol. 38, no. 6, pp. 1187–1194, Dec. 1991.

[57] N. Tam et al., "Multi-cell soft errors at the 16-nm FinFET technology node," in 2015 IEEE International Reliability Physics Symposium, 2015, p. 4B.3.1-4B.3.5.

[58] N. Seifert et al., "Soft Error Rate Improvements in 14-nm Technology Featuring Second-Generation 3D Tri-Gate Transistors," IEEE Trans. Nucl. Sci., vol. 62, no. 6, pp. 2570–2577, Dec. 2015.

[59] Y. P. Fang and A. S. Oates, "Neutron-Induced Charge Collection Simulation of Bulk FinFET SRAMs Compared With Conventional Planar SRAMs," IEEE Trans. Device Mater. Reliab., vol. 11, no. 4, pp. 551–554, Dec. 2011.

[60] I. Chatterjee et al., "Impact of Technology Scaling on SRAM Soft Error Rates," IEEE Trans. Nucl. Sci., vol. 61, no. 6, pp. 3512–3518, Dec. 2014.

[61] Weir, Bonnie, Vani Prasad, Shahriar Moinian, SangJune Park, Joseph Blasko, Jason Brown, and Jayanthi Pallinti. "Utilizing a Thorough Understanding of Critical Aging and Failure Mechanisms in finFET Technologies to Enable Reliable High Performance Circuits." In 2019 IEEE International Reliability Physics Symposium (IRPS), pp. 1-5. IEEE, 2019.

[62] Agarwal, Mridul, Bipul C. Paul, Ming Zhang, and Subhasish Mitra. "Circuit failure prediction and its application to transistor aging." In 25th IEEE VLSI Test Symposium (VTS'07), pp. 277-286. IEEE, 2007.

[63] Khoshavi, Navid, Rizwan A. Ashraf, Ronald F. DeMara, Saman Kiamehr, Fabian Oboril, and Mehdi B. Tahoori. "Contemporary CMOS aging mitigation techniques: Survey, taxonomy, and methods." Integration 59 (2017): 10-22.

[64] Reis, Ricardo, Cao, Yu, and Wirth, Gilson, eds. 2014. Circuit Design for Reliability. New York, NY: Springer. Accessed February 28, 2020. ProQuest Ebook Central.

[65] Maricau, Elie, and Gielen, Georges. 2013. Analog Ic Reliability in Nanometer CMOS. New York, NY: Springer.

[66] Premachandran, C. S., Salvatore Cimino, Sean Ogdan, Zhuo-Jie George Wu, Daniel Smith, Sukesh Kannan, Linjun Cao et al. "Comprehensive 3D TSV Reliability Study on 14nm FINFET technology with thinned wafers." In 2018 IEEE 2nd Electron Devices Technology and Manufacturing Conference (EDTM), pp. 37-40. IEEE, 2018.