CMOS scaling impacts on reliability, What do we understand?

Seyab, Nor Zaidi Haron, and Said Hamdioui.
Delft University of Technology, Computer Engineering Laboratory,
Mekelweg 4, 2628 CD Delft, The Netherlands. E-mail: {seyab,nor-zaidi,said}@ce.et.tudelft.nl

Abstract—The physical miniaturization have increased CMOS fault sensitivity to the extent that many reliability constraints pose threat to the device normal operation. This paper addresses the impact of physical miniaturization on CMOS reliability constraints. The focus will be mainly on five technological aspects namely Negative Bias Temperature Instability (NBTI), Time Dependent Dielectric Breakdown (TDDB), electromigration induced failure, hot carriers degradation and single particle induced soft errors. First the concept of each of the aspect is explained, then the impact of scaling on it is analyzed. Thereafter various mitigation techniques, to alleviate the issues related to each aspect are presented.

I. INTRODUCTION

In semiconductor perspective reliability is the ability of a device to conform to its electrical and visual/mechanical specifications over a specified period of time under specified conditions at a specified confidence level [1]. Since the beginning reliability has been remained an important part of semiconductor industry. For the last six decades device reliability have improved with each scaled generation of technology. Manufacturers of devices with critical applications of semiconductor industry. For the last six decades reliability has been remained an important part of semiconductor industry. For the last six decades device reliability have improved with each scaled generation of technology. Manufacturers of devices with critical applications like military, automotive and medical mainly contributed to initiate and develop semiconductor reliability field.

In order to maintain the growing operating speed and integration density over a given silicon area, the CMOS scaling should go into sub-10nm for future semiconductor devices. The aggressive miniaturization have fronted reliability different aspects: (a) wafer process technology, (b) packaging (c) handling (d) mounting and (e) environmental hazards both at the research and production level [2].

This paper mainly focuses on major CMOS reliability constraints related to wafer process technology as they are critical to CMOS operation. The major CMOS technological constraints are classified into:

- **Negative Bias Temperature Instability (NBTI):** The threshold voltage of PMOS may shift when the gate terminal is subjected to negative potential at elevated temperature.
- **Time Dependent Dielectric Breakdown (TDDB):** Long time operation under high oxide field causes Time Dependent Dielectric Breakdown in MOS IC’s, especially in those with large oxide areas such as DRAM [3].
- **Electromigration Induced Failure:** The increasing current density in scaled interconnect and contacts initiates electromigration to produce open, short circuits or leakage problems.
- **Hot carrier Degradation:** The energetic hot electrons changes current/voltage characteristics of CMOS by trapping in the oxide layer near drain or entering the interface.
- **Single event soft error:** High energy radiations and particles changes state of CMOS based storage devices by creating electron hole pair in storage and sensor parts of the circuit to produce soft errors.

This paper discusses the above technological reliability issues in detail, together with the impact of physical miniaturization on them. The paper further provides different measures to pacify them. The rest of the paper is organized as follow: Section II addresses the negative bias temperature bias instability, including the impact of scaling on it, and the major measures taken to prevent it. Section III to Section VI in a similar way deal with ‘time dependent dielectric breakdown’, ‘electromigration induced failure’, ‘hot carrier degradation’ and ‘single event soft error’ respectively. Finally, Section VII summarizes the paper.

II. NEGATIVE TEMPERATURE INSTABILITY (NBTI)

NBTI refers to the shift in the PMOS threshold voltage when the gate terminal is subjected to negative potential at elevated temperature for a prolonged period of time. The degradation is attributed to breaking of Si-H bonds at the Si/SiO₂ interface and the resultant diffusion of hydrogen species into the gate oxide. Unlike other degradation processes the NBTI is a self annealing process and the broken bonds recover when the applied potential is removed.

Among various NBTI degradation models presented in the literature, the Reaction Diffusion model (R-D model) illustrated in Fig.1 is consistent with experimental results. The model considers that negative potential applied at the gate produce interface traps IT at Si/SiO₂ interface. The electro-chemical reaction break Si-H bonds to produce hydrogen leaving behind an interface traps [4]. The released hydrogen species diffuse in the oxide layer toward gate electrode to increase threshold voltage. The R-D model is based on following two equations.

\[
\frac{dN_{IT}}{dt} = k_f(N_o - N_{IT}(t)) - k_r N_{IT}(t) N_{H^0}(t) \quad (1)
\]
A. Impact of scaling

The aggressive scaling of oxide thickness without a corresponding reduction in the power supply voltage have increased oxide electric field significantly [5]. For NBTI with H₂ diffusion the $N_{H^o}$, $k_r$ and $D_H$ in eq.1 are not affected by field variation as they are restricted to formation and diffusion of neutral hydrogen species. The field variation will only come through field dependent $k_f$ which can be written as [6]:

$$k_f = k_o \times T \times p \times \sigma \times e^{-(E_{Fo} - E_{ox})/k_BT}$$  \hspace{1cm} (3)

The forward reaction rate have three fold dependence on the electric field $E_{ox}$. Firstly the number of charges already present at the Si/SiO₂ are $p$, and can be written as:

$$Q(p \times q) = C_{ox} \times (E_{ox})$$  \hspace{1cm} (4)

so

$$p \propto E_{ox}$$  \hspace{1cm} (5)

Secondly the charges generated tunnel inside the Si-H bond to approximately 1.5-2Å depending on transmission co-efficient $T$. The transmission co-efficient depend exponentially on the electric field.

$$T \propto e^{\gamma TE_{ox}}$$  \hspace{1cm} (6)

Finally the barrier potential $E_{Fo}$ required to generate interface trap decrease by factor $aE_{ox}$ with applied electric field. And most importantly thinner poly-silicon gate have brought the Si/SiO₂ interface closer to the gate, so the hydrogen diffusion front reaches poly-interface within the stress phase and causes higher $V_T$ shift.

B. Mitigation

Nitrogen presence near the oxide/substrate interface enhances the NBTI degradation due to ease of hydrogen diffusion in nitrided oxide. The NBTI degradation decreases by adopting plasma nitridation instead of oxide nitrides [7]. In plasma nitridation the nitrogen diffuses from oxide/gate end toward oxide/substrate interface. The decreasing nitrogen concentration near the gate restrict hydrogen diffusion and improve NBTI immunity.

Secondly the time power exponent $n$ in threshold voltage shift $\Delta V_T = t^n$ have a range of values 0.25-0.30. The exponent variation is due to delay in measurement after removing stress and varying diffusion coefficient of hydrogen species in amorphous oxide. The hydrogen diffusion in amorphous oxide have a range of values due to variable hoping distances and timings. The interface trape generation rate in amorphous oxide is written as [5]:

$$N_{IT} = (k_f N_{o}/k_r)^{1/2}(D_H t)^{n-\frac{a}{2}}$$  \hspace{1cm} (7)

and the diffusion coefficient of hydrogen in amorphous oxide is given by:

$$D_H = D_{Ho} (\sigma t)^{n-\frac{a}{2}}$$  \hspace{1cm} (8)

where with $a = 0.1-0.2$, results in a lower power exponent range 0.225-0.20. The lowering of the power exponent suggest that NBTI performance can be improved by using more amorphous oxide with deeper traps level and longer release time [5].

III. TIME DEPENDENT DIELECTRIC BREAKDOWN(TDDB)

In TDBB the dielectric material isolating gate and substrate suffers form short circuit failure due to intense electric field applied across them [8]. TDBB is a two step linked process consisting wearout and thermal runaway. In the first step charge traps accumulate in bulk oxide and silicon/oxide interface, with the passage of time their density reaches to a critical value. The step is followed by sufficient local electric field and current that causes thermal runaway and melting of microscopic regions. Thus wear out is a global while runaway is a local phenomena.

The specific operative mechanism and the equations that define TDBB are still matter of controversy [9], however main TDBB models include the E-model, $1/E$ model and current dependent power law model. The E-model based on the thermo-chemical mechanism states that the TDBB follow field driven thermo-chemical breakage of Si-Si bond in SiO₂ [10]. The decomposition and hence time to fail have exponential dependence on the applied oxide field as follow [11]:

$$MTTF = A exp(-\gamma E) exp(\frac{E_a}{kT})$$  \hspace{1cm} (9)
where $A$ is a constant, $\gamma$ is field acceleration parameter, $E$ is oxide field, $E_a$ is activation energy, $k$ is Boltzman constant and $T$ is absolute temperature. The activation energy is linearly dependent on oxide field and the field acceleration parameter decrease with increasing temperature.

In daily reliability work the linear E-model is used and is fine upto oxide field of 4.8MV/cm, however the model ignore role of tunneling current and is un-adequate for ultra thin oxides [12].

The 1/E model states that the TDDB is a current driven mechanism caused by Fowler-Nordheim tunneling current and anode hole injection in SiO$_2$ layer [10] [8]. The model can be expressed as:

$$MTTF = A exp\left(\frac{G}{E}\right) exp\left(\frac{E_a}{kT}\right)$$

where $G$ is a constant. The model behaves well at higher stress values however ignores the thermal/diffusion that takes place in all material over time even in the absence of electric field.

Although voltage is the driving force for current but the leakage current also depend on oxide thickness, trap density, and defect generation rate. In order to include effects of all these parameters, a newly proposed model cover overall response of the system in the form of mean value of average leakage current $I_{avg}$. The model states that TDDB of thin gate oxide with thickness range of 1.7nm-6.8nm decreases inversely as power law rather than exponential of the mean value of the average leakage current during the constant voltage stress and selected temperature range [13] as below:

$$MTTF = A \frac{1}{I_{avg}^n} exp\left(\frac{E_a}{kT}\right)$$

where $A$ is a constant and the value of exponent $n$ increases from 1 to about 10 when oxide thickness reduces from 6.8nm–1.7nm. The power law prediction of TDDB reduction with scalings is in accordance with the percolation theory [15].

A. Impact of scalings

Historically the TDDB have received little reliability attention due to dielectric thickness and lower operating field. However MOS scaling increased the electric field $E$ across the gate oxide films that reduces the TDDB activation energy $E_a$ by following relation [11].

$$E_a = \Delta H_o - a \times E$$

where $\Delta H_o$ is the enthalpy of activation and $a$ is the effective dipole movement. Additionally under high electric field, frequent charges mobility increases oxide temperature that results in lower field acceleration parameter $\gamma$. The strong dependence of activation energy and acceleration parameter on oxide field reduces the TDDB with oxide scaling as shown in Fig 2.

Additionally the TDDB dependence on oxide scaling comes from defects in oxide layer and tunneling currents flowing through it. The TDDB relation to them is given by [14].

$$TDDF = \frac{N_{BD}}{kJ}$$

where $N_{BD}$ is the number of defects needed for dielectric to breakdown, $k$ is a constant and $J$ is the tunneling current through oxide layer. Percolation model states the breakdown only occur if a connecting path is formed across the gate oxide. The formation of such a path is a function of defect density and oxide thickness. For a given defect density the path formation is more likely for thinner oxides thus reducing TDDB [15].

Direct tunneling current starts for ultra thin oxide below 4nm and increase by one order of magnitude for every 0.2-0.3nm reduction in oxide thickness. Additionally at very low oxide thickness the electronic wave function of silicon conduction band may become large enough to reduce the tunneling barrier.

B. Mitigation

Introduction of high $k$ dielectric minimize the unacceptable leakage current in the oxide layer. Silicon nitride have a dielectric strength is about 10 MV/cm and is alternative for silicon dioxide layer [16]. The Silicon nitride at SiO$_{2}$/poly interface minimize defect generation by restricting boron penetration from highly doped gate material and have excellent thermal compatibility with Si substrate.

Zirconium and Hafnium silicate (ZrSi$_x$O$_y$ and HfSi$_x$O$_y$) with 3-8% of Hf or 2-5% Zr are suitable for replacing SiO$_2$ in future ultra scaled gate oxide. The silicates have higher dielectric constant, lower leakage current and are stable in direct contact with the Si [17]. The modification is achieved by adding Zr and Hf to the SiO$_2$ and then following the rest of the CMOS processing procedure.

IV. ELECTROMIGRATION INDUCED FAILURE

The electromigration is characterized by metal ion drift in an interconnect with high current density. The high energy electrons force metal ions to move in their own direction mainly due to momentum exchange [18]. The material flux...
cause voids formation at their previous positions and protrusions at the collection points. As a result short and open circuit faults are produced the interconnects as shown on the fig 3.

For CMOS devices the material migration is more obvious at the silicon interconnect junction. The effective drift velocity of material flux derived in Blench Model is given as [19]:

$$v_d = \frac{J}{n} = \frac{D_{eff}}{kT} \left[ e j \rho Z_{eff} - \Omega \frac{\Delta \sigma}{\Delta x} \right]$$  \hspace{1cm} (14)

where $v_d$ is the drift velocity, $J$ is atomic flux, $n$ is the atomic density, $D_{eff}$ is the effective diffusion coefficient, $k$ is the Boltzman’s constant, $T$ is the absolute temperature, $Z_{eff}$ is the effective charge number, $e$ is the absolute value of electron charge, $\rho$ is the metallic resistivity, $\sigma$ is the atomic volume and $\frac{\Delta \sigma}{\Delta x}$ is the electromigration induced stress gradient along the length of the interconnect.

**A. Impact of scaling**

The increasing chip density and growing performance have increased the current density in scaled interconnects. Eq.14 shows that atomic flux $J$ in narrow interconnect increase with increasing current density $j$ at constant temperature. Additionally the line width dependence of the material migration comes from diffusion constant $D_{eff}$. The $D_{eff}$ is a sum of diffusion in bulk, grain boundary and line width at the interface [19].

$$D_{eff} = n_b D_b + \left( \frac{\delta_{gb}}{d} \right) \left( 1 - \frac{d}{w} \right) D_{gb} + \delta_i \left( \frac{2}{w} + \frac{1}{h} \right)$$  \hspace{1cm} (15)

where $n_b$ is the fraction of atoms in the bulk, $\delta_{gb}$ is the effective thickness of the grain boundaries, $d$ is the grain size and $\delta_i$ is the effective conductor width at the interface. The smaller grain sizes in scaled lines results in higher diffusions and current density. The higher diffusion rate increases the drift velocity that causes a decrease in the MTTF as shown in the Fig. 4.

For the worst case of non-scaled voltage the decreasing line width result in increasing current density exponentially, and from Black’s equation the lifetime dependency on current density $j$ is written as [20]:

$$MTTF = A j^n \exp \left( \frac{E_a}{kT} \right)$$  \hspace{1cm} (16)

where $E_a$ is activation energy, $j$ is current density, $E_a$ is the activation energy and $T$ is temperature. For constant temperature with current density below $25\text{mA}/(\mu\text{m}^2)$ value of $n$ is 1, and may reach to 1.8 for current density $25\text{mA}/(\mu\text{m}^2) \leq j \leq 140\text{mA}/(\mu\text{m}^2)$ thus reducing the mean time to fail due to electromigration.

**B. Mitigation**

The electomigration is due to push of electrons in high current density interconnects. By keeping the current density below the maximum allowed density number of electrons colliding metal ions decreas thus the problem of electromigration is alleviated.

Increasing copper concentration in Al(Cu) lines can alleviate the material migration issue. For interconnect with smaller linewidth the activation energy of Cu is nearly equal (about 1eV) to Al(Cu) but the absolute lifetime of Cu lines is about $50 \times$ that of Al(Cu) interconnects. The MTTF improvement is due to the higher resistivity to electromigration and higher melting point [21]. Furthermore good selection and deposition of the passivation over the metal interconnect reduces the electromigration damage by limiting extrusion and suppressing surface diffusion.

The material flux depend on the length of the line that allow electromigration to occur. Any line shorter than *Blench length* will not suffer from electromigration. The improvement in electromigration resistivity is due to reversed mechanical stress buildup $\frac{\Delta \sigma}{\Delta x}$ in eq.14, the stress cause a reverse migration process to compensate the material flux [18]. To avoid the electromigration in line the product of wire current and wire length must be smaller than process...
technology dependent threshold value \((J)_{th}\).

V. HOT CARRIERS DEGRADATION

Hot carriers are energetic electrons and holes in the channel and pinch off region of a transistor that effect transistor on state and off state currents. Initially the carriers gain enough kinetic energy in excess of thermal energy to enter substrate region. If they continue to gain more energy (3.2-3.8 eV) they are injected into the oxide layer. The substrate current produce impact ionization and finally CMOS latchup while they are injected into the oxide layer to the formation of oxide states and trapped oxide charges.

The well known hot carrier degradation model is Lucky electron model. According to the model the Lucky electron gain sufficient energy from the field to become hot and their momentum is directed toward oxide. The electrons moves away from channel that results in a reduced MOSFET on state current and higher off state current [9].

A. Impact of scaling

The hot carrier degradation strongly depends on the channel length at constant supply voltage [9].

\[
MTTF = A \times (L_{eff})
\]

where A is a constant and \(L_{eff}\) is the channel length. The channel length dependency can be derived from modification in Lucky electron model. According to Lucky electron model for a given values of pinch off potential \(V_{dsat}\), drain potential \(V_d - V_{dsat}\) and gate thickness the impact ionization rate \(\alpha\) is given by [23]:

\[
\alpha = \frac{I_{sub}}{I_d} = \exp[-(1/(V_d - V_{dsat}))]
\]

the model argue that \(\alpha\) is independent of gate length at constant \(-1/(V_d - V_{dsat})\). However it is found that \(\alpha\) increases with scaling the gate length to sub-micron level because the carriers gain energy from drain field as well as channel electric field i-e electric field between source edge and pinch off region [23]. The drain field is independent of channel length but channel electric field \(E_{ch}\) increases with decrease in channel effective length \(L_{eff}\) by following the relation [23].

\[
E_{ch} = \frac{(V_{dsat} - V_s)}{L_{eff}}
\]

The channel electric field increases the electron velocity in the channel. The increase in carrier velocity enhances temperature distribution in the channel and the pinch-off regions that increases the impact ionization rate both in NMOS and PMOS by following equation [23].

\[
\alpha = \exp(-\pi/kT).
\]

Additionally the drain field increase with scaling the gate oxide thickness by following eq [25].

\[
E_m = \frac{(V_d - V_{dsat})}{l} = \frac{(V_d - V_{dsat})}{(0.22T_{ox}^{1/3}x_j^{1/2})}
\]

where \(x_j\) is the drain pn junction depth and \(T_{ox}\) is the oxide thickness. The hot carrier reliability of scaled CMOS degrades even at supply voltage of \(V_d = 1.0 V\) due to higher drain field[23][24]. Fig. 5 show the impact ionization rate dependence on oxide thickness and gate length at peak substrate current and constant \(V_d - V_{dsat}\).

\[
\alpha(T_{ox}) = \frac{1}{1 + T_{ox}^3}
\]

\[
\alpha(L) = \frac{1}{1 + L^3}
\]

The lateral field along the channel \(E_{ch}\) decreases by introduction of a Lightly Doped Drain LDD region which make use of a lightly doped region (n−) between the channel and the drain n+ region [26]. However voltage drop in the LDD lateral resistance degrade current reliability. In order to minimize the side effects modified LDD structures such as Inverse-T gate LDD ITLDD, Large Tilt Angle Implant Drain LATID and Buried LDD can be efficiently used.

The hot carrier reliability significantly improves by lowering the power supply voltage in accordance with \(V_{dsat}\). For halving the channel length the \(V_{dsat}\) may only to be reduced by less than 0.5V in future MOSFETs so \(V_{cc}\) reduction of 0.5V will guarantee hot carriers reliability.

VI. SINGLE EVENT SOFT ERROR

Radiation and particle of broad energy spectrum ionizes the semiconductor materials to cause a non-destructive change in state of CMOS based devices. The naturally occurring alpha particle imping on CMOS to generate electron hole pairs in several picoseconds as shown in Fig. 6. The charges generated in the depletion region or near the depletion region are separated by electric fields [27].

The charges in depletion region drift to opposite polarity nodes in first few picoseconds, thus effecting the stored charges. On the other hand charges generated outside the depletion region diffuse slowly toward the collection nodes.
The Alpha particles generate electron hole pair, to produce soft error. The drift and slow diffusion of charges induce bipolar current spike, with a bulk of charges collection at nodes. If the charge collected overcomes the critical $Q_{ crt }$ required, the state of CMOS based storage device change causing soft error [30]. The soft errors are random, nonrecurring, single bit and temporary i.e. no physical defect is associated with the failure [27]. The error can reduce the threshold voltage as a result MOS can turn on without any external voltage [9].

A. Impact of scaling

As technologies move toward the smaller feature size the critical charge $Q_{ crt }$ representing state of CMOS based devices and the capability to maintain their state against the spurious signals decreases. The critical charge $Q_{ crt }$ relation to device dimensions $l$ is [9].

$$Q_{ crt } = 0.023l^2$$  \hspace{1cm} (22)

For modern ultra scaled CMOS devices the critical charge has decreased to the extent that the naturally occurring alpha particle traveling in straight lines have longer range (2-3 $\mu m$) than current device dimensions. The the soft error rate dependence on the critical charge is shown in the fig. 7.

The CMOS supply scaling reduces the switching voltage amplitude that directly decreases the critical charge required [28]. The state of CMOS based storage device operating at lower supply voltage can be easily altered by radiation and particles with LET as shown in the fig 8. Additionally for a given heavy ion Linear Energy Transfer $LET$ the soft error immunity of CMOS decreases as the microcircuit frequency increases.

B. Mitigation

The only way to improve the soft error immunity is to reduce the sources of radiation or restrict their access to the sensitive chip area. The alpha particle emission can be reduced by using high purity material and shielding the chip with a thick polymeride layer prior to packaging [29]. The soft error induced by high energy neutron can be reduced by using insulator that do not contain boron.

The soft error resistance can be improved by reducing the charge collecting efficiency or increasing the critical charge. The charge collection decreases by using well isolation with silicon on insulator (SOI) substrate material. The well isolation limits the funneling effect and the likelihood of parasitic bipolar collection path [30].

VII. SUMMARY AND FUTURE PROSPECTS OF CMOS SCALING

In this paper we have discussed that aggressive miniaturization of CMOS will cause higher sensitivity to defects. The ultra scaling of gate oxide will increase field intensity that will enhance NBTI and TDDB issues. Plasma nitridation and use of amorphous oxide decreases the NBTI and the introduction of Zr or Hf silicate minimize the TDDB. The use of Cu in the Cu(Al) and optimizing gate length mitigates the open and short faults due to electromigration in the shrinking lines. The use of LDD and modified LDD structures reduce the hot carriers impact ionization rate and oxide trape generaton in scaled CMOS. Finally highly pure material, annealing and proper shielding minimize the soft errors. Table I summarizes the issues and the solutions. The above scaling constraints and their solutions reveal that inspite of aggressive downsizing the CMOS reliability constraints are manageable even MOSFET gate length is be scaled to the atomic limits [31]. The advent of new technology and improvement in the architecture design and process development will enhance CMOS performance even at gate length of 10-20nm. The current silicon CMOS is the most successful ‘nano device’ and no realistic replacement
of the silicon devices could be thought even if the scaling reaches to the downsize limits.

REFERENCES