

Reliability in CMOS IC Design: Physical Failure Mechanisms and their Modeling

There are a number of physical failure mechanisms that can affect the reliability of a CMOS ASIC. Some of the common mechanisms can be mitigated by adhering to foundry design rules (Electromigration, Time Dependent Dielectric Breakdown (TDDB), and Hot Carrier Damage). Certain fabrication steps can cause stress that may lead to latent damage that may later reduce the useful lifetime of an ASIC. Contamination with Mobile Ions (most commonly Sodium) will render transistor characteristics unstable and encourage early TDDB. Process Induced Oxide Charging, caused by injection of charge into gate oxides during certain ion etching processes, will reduce TDDB lifetime and cause some transistor degradation similar to Hot Carrier Damage. Metal Stress Migration, which is caused by large thermal coefficient of expansion difference between metal interconnect and inter-level dielectrics (oxides), can lead to voiding of metal lines similar to damage caused by Electromigration. Foundries will expend great effort to control the above fabrication related failure mechanisms by designing the fabrication process to minimize the unnecessary stresses applied to wafers and to maintain an absolutely clean fabrication process to eliminate contamination (by Sodium and other materials). This document will provide an outline of reliability limiting physical mechanisms that are directly influenced by the ASIC designer and also provide some general design for reliability guidelines to ASIC designers.

Foundry Reliability Targets

CMOS foundries are strongly motivated to mitigate any physical mechanism that may cause an ASIC to fail due to some known physical failure mechanism within a specific reliability target. That target is typically in the neighborhood of 20 to 40 years. The choice of 20 to 40 years is not because they expect the CMOS device to be placed into consumer service for use that length of time. The target time to fail is chosen because the physical wear-out mechanisms are governed by a stochastic processes with a random distribution of failure times with some probability distribution. The choice of fail time target is selected such that a consumer product nominal service life (typically less than ten years) will fall somewhere far out in the tails of the failure mechanism probability distribution function (usually at least six sigma). The foundry assumes that the ASIC will be used under operating conditions (temperature, voltage, current density, etc.) that do not exceed the limits defined by the foundry.

In order to achieve these goals, a foundry must understand the physics of each failure mechanism and identify any wafer processing steps that may detrimentally influence each mechanism. Once the wafer processing is optimized for maximum lifetime set by each physical mechanism, the foundry develops design rules intended to prevent ASIC designers from over-stressing the devices and cause the expected lifetime to fall below foundry targets. These design rules will be embodied in the form of maximum operating voltage, transistor channel length constraints for service under certain bias conditions, maximum current per unit line width in metal interconnect, maximum current per contact or via and certain constraints upon interconnect layout of very wide metal lines. Failure to comply with the reliability design rules may lead to unpredictably shorter ASIC lifetime.

Common CMOS Failure Mechanisms

This section will outline the most common physical failure mechanisms and describe in general terms how a foundry is able to control them with processing design and with design rules. It is not the intent of this description to provide people with a comprehensive treatment of the physical failure mechanisms, but it is intended that this document will answer some common questions that designers raise about the reliability design rules. One common general misconception is that the physics of failure is thoroughly understood and that mathematical models exist that permit designers to tune ASIC reliability to their specific produce lifetime. Most mathematical models for physical failure mechanisms are based upon observations of accelerated reliability tests performed at elevated temperature and at higher than normal voltage or current density. These accelerated tests can produce some unrealistic stresses upon the materials resulting in failure statistics that do not represent stresses experienced at nominal operating conditions. There is great temptation to extrapolate accelerated test results, using these mathematical models developed from accelerated test data, back to "at-use" conditions to predict ASIC reliability in typical applications. Unfortunately, this involves extrapolating over many orders of magnitude of acceleration factors to normal operating conditions using these simple mathematical models that may not be valid under lower acceleration factors. Sometimes this is also attempted using data sample sizes that are too small to define the tails of the probability distribution function. In many cases, even the probability distribution function is in debate within the community of reliability physicists.

Even though commercial tools exist to help designers manage design reliability targets, the accuracy of fail time models is nowhere close to the accuracy of transistor models used in SPICE circuit simulations. The statistical properties of failure mechanisms can lead to fail times that differ from model predicted median fail times by orders of magnitude. The magnitude of errors between physical ASIC device lifetime and the model-predicted lifetime do not lend themselves for fine tuning ASIC lifetime.

Time Dependent Dielectric Breakdown (TDDB)

TDDB is wear-out of the insulating properties of silicon dioxide in the CMOS gate leading to the formation of a conducting path through the oxide to the substrate. With a conducting path between the gate and the substrate, it is no longer possible to control current flow between the drain and source by means of the gate electric field. TDDB lifetime is strongly affected by the number of defects in the gate oxide produced during wafer fabrication. Therefore, foundries strive to produce an ultra-clean oxide in their process to maximize the TDDB lifetime. Even if a foundry could produce a perfectly defect free oxide, TDDB would remain a concern for ASIC designers.

TDDB occurs at all gate voltage bias conditions. The goal of the foundry is to trade off gate oxide thickness with operating voltage specifications to achieve both speed and lifetime targets for the technology. The lifetime of a particular gate oxide thickness is determined by the total amount of charge that flows through the gate oxide by tunneling current. The electron tunneling current is modeled by the Fowler-Nordheim equation for oxide current density, J :

$$J = AE^2 e^{\left(\frac{B}{E}\right)},$$

where, A and B are constants related to effective mass and barrier height and E is the electric field. It is clear from this relationship that operating a CMOS device at voltages greater than foundry specification results in an exponential increase in the amount of oxide current. Once electrons have breached the oxide potential barrier they are accelerated through the oxide by the electric field which is determined by the applied voltage and the oxide thickness. Charge accelerated in the gate oxide achieves greatest energy at the oxide-silicon interface presuming there have been no collisions in transit. At the end of its travel through the oxide, it deposits its energy at the oxide-silicon interface.

The oxide-silicon interface has some special properties produced by the fact that there is a large thermal coefficient of expansion (TCE) difference between silicon dioxide and silicon (silicon: 2ppm/°C vs. Silicon Dioxide: 0.3ppm/°C). This large difference in TCE leads to strained chemical bonds that can be broken by the accelerated charge. Once the chemical bonds have been broken, the sites become locations where charge can become trapped. This trapped charge will have an influence upon the channel carrier mobility in transistors and reduce their gain. This trapped charge also has the effect of increasing the electric field locally and therefore increasing the local tunneling current. The process has positive feedback that leads to rapid charge build-up until the tunneling current is large enough to literally burn a hole through the gate oxide.

Modeling TDDDB lifetime has been a challenging task. The specific physical process details leading to failure are somewhat complex, and as a result, simple models tend to be inaccurate if the gate oxide is substantially different from the thickness used for collecting the data used to develop the model. A great deal of material has been published about TDDDB modeling with various relationships of electric field to lifetime. Debate has raged for some time over whether the lifetime is related to E (electric field), or 1/E, or just applied gate Voltage. Currently it appears that simple models will only be valid over certain ranges of gate oxide thickness. Moderately thick oxides seem to have a lifetime related to 1/E (at high electric field) or E (at low electric field) while for very thin oxides (thinner than 5nm) the lifetime appears to be related to applied Voltage.

All of the following models are based upon data fitting from experimental data under different amounts of electric field stress with different oxide thickness. Tests are performed at very high electric field (greater than under normal use conditions) and at elevated temperature to reduce testing time to a few weeks instead of tens of years. These models include the temperature dependence.

For thick oxides at relatively low electric field (moderately greater than at-use conditions) TDDDB lifetime has been shown to obey [1]:

$$t_{BD1} = t_1 \exp(-\mathcal{E}_{ox}), \text{ where } t_1 = 6.3E14 \text{ sec. and } (=2.66\text{cm/MV}$$

$$\text{Temperature effects: } t_1 = t_{10} \exp\left(\frac{\Delta H_0}{k_B T}\right) \text{ and } \mathcal{E} = b + \frac{c}{T}$$

Similarly, for thick oxides under high field stress (much greater than at-use conditions):

$$t_{BD2} = t_2 \exp\left(\frac{G}{E_{ox}}\right), \text{ where } t_2=1.0E-11 \text{ sec. and } G=350 \text{ MV/cm}$$

$$\text{Temperature effects: } t_2 = t_{20} \exp\left[-\frac{E_b}{k_B}\left(\frac{1}{T} - \frac{1}{300}\right)\right] \text{ and } G = G_0 \left[1 + \frac{d}{k_B}\left(\frac{1}{T} - \frac{1}{300}\right)\right]$$

Both of the above equations contain constants that are extracted from data and will vary depending upon specific oxide properties and defect content. Foundries will characterize their process using experimental accelerated TDDB lifetime data to extract model parameters. From these parameters they are able to tune the fabrication process to produce gate oxide that has at least 20 to 40 year lifetime at the specified operating voltage.

When oxides are thinner than about 5nm the lifetime modeling begins to show a voltage dependence as discussed in a paper by P. E. Nicollian, et. al. [2]. Even with the insights from this paper the prediction of TDDB lifetime is not an exact science.

As a result of the imprecise modeling of TDDB lifetime over all, it is not possible to accurately engineer a device lifetime using the existing TDDB models. Also, it is not possible to use model parameters from one fabrication line to extrapolate TDDB lifetime on another fabrication line. It is much safer for ASIC designers to stay within the foundry operating voltage specifications.

[1] C. Hu, et. al., “A Unified Gate Oxide Reliability Model”, IRPS, 1999

[2] P. E. Nicollian, et. al., “Experimental Evidence for Voltage Driven Breakdown Models in Ultrathin Gate Oxide”, IRPS 2000.

Hot Carrier Damage

Hot carrier damage causes the transistor transconductance to slowly degrade and eventually may cause the transistor threshold to change near the drain edge of the channel such that it can not form a channel in the drain region. This mechanism can be more damaging to digital circuits because it will cause parts of the digital circuit to have longer delay than originally intended, leading to logic race conditions. With minimum channel length transistors, hot carrier damage occurs even when operating voltages are with foundry specification. Its preferential impact upon short channel length transistors makes it is less often a problem in analog circuits because analog circuits are rarely designed using minimum channel length transistors.

Carriers in a channel traveling from source to drain experience an electric field magnitude that varies depending upon the transistor bias conditions. As the transistor channel is forming at low drain-source voltage the electric field is uniformly divided over the length of the channel by the presence of the inverted channel. When the drain-source voltage is increased the channel begins to pinch off. A point is reached when the electric field is largely confined to the region between the pinched off channel edge and the drain diffusion. This causes the magnitude of the electric field to reach high values and cause the channel carriers to accelerate through the pinched off region reaching a high velocity that is greater than thermally limited diffusion drift velocity, hence the term “hot carriers”. If a channel hot carrier collides with a crystal atom near the drain region, it may produce an electron-hole pair in an impact ionization event.

Carriers generated by impact ionization are subjected to the same large electric field that had lead to their formation. In an N-channel transistor, for example, the electrons are accelerated to the drain while the hole is driven into the substrate by the gate electric field. This hole current is a measurable quantity that serves as an indicator of the amount of impact ionization that is occurring at a specific set of bias conditions. Energy imparted to the carriers formed in an impact ionization event can cause electrons (and holes) to be scattered toward the gate oxide interface with the substrate. As with TDDB described above, these scattered carriers can generate interface states that will degrade channel carrier mobility and reduce transconductance of the transistor. Scattered electrons with the highest energy may be injected into the gate oxide producing a trapped charge that will also degrade transistor performance. Further more, the trapped charge also enhances gate oxide electric field which exacerbates TDDB problems.

Hot carrier damage rate is highest in a transistor with channel length at minimum design rule length and when drain-source voltage is maximum permitted voltage while the gate-source voltage is around of half of the drain-source voltage. Worst case bias condition can readily be observed by measuring the substrate current as a function of bias conditions and noting when substrate current is maximum. This is precisely the circumstances that occur in digital circuits in the transition between a logical “1” and a logical “0”. Analog circuits tend to be designed with longer than minimum channel lengths so hot carrier damage rates are much lower at all bias conditions.

Mathematical modeling of hot carrier damage rates is very difficult. Most common models are not sufficiently accurate in predicting damage rate at all bias conditions to be useful in a detailed design tradeoff.

Accumulation of interface states (or change in the number of interface states) at the gate oxide-semiconductor interface can be estimated with:

$$\Delta N_{it} \propto \left[N_s \left(\frac{1}{WH} \right) \int_0^{T_s} I_D^{1-m}(t) I_{SUB}^m(t) dt \right]^n,$$

where N_s is the total number of transitions (switches)
and n, m and H are extracted from data.

When hot carrier damage accumulates sufficiently to effect drain saturation current, then:

$$\frac{\Delta I_d}{I_{d0}} \propto (Age)^n, \text{ where } Age = \frac{I_d}{WH} \left(\frac{I_{SUB}}{I_d} \right)^m t$$

The above equations are from recently published papers [3][4] on hot carrier damage.

The change in drain current relative to its value when the transistor is “fresh” is experimentally obtained to extract the constants n, m and H as a function of channel width W using accelerated hot carrier test results. Again, the extracted constants will apply for only the process from which the parameters were extracted.

Generally, foundries will provide design rules governing minimum channel length transistors in various applications in which the nominal power supply voltage is used and when bias conditions favor hot carrier damage. Any attempt to extrapolate these design rules to be useful with higher power supply voltage is likely to lead to large errors in predicted damage rates. Operating CMOS circuits at voltage greater than foundry specification voltage may lead to unexpectedly early circuit failure due to hot carrier damage.

[3] W. Jiand, et. al., “Assessing Circuit-Level Hot Carrier Reliability”, IRPS 1998.

[4] P. Chen, et. al., “A unified Compact Scalable) I_d Model for Hot Carrier Reliability Simulation”, IRPS 1999.

Electromigration

Electromigration is the diffusion of metal atoms along the conductor in the direction of electron flow. This directional diffusion process occurs because the momentum transfer between the electrons and the metal atoms increases the probability that an aluminum atom will move in the direction of the electron flow. Since the mass of the electron is very much smaller than aluminum atoms the transfer of momentum is only enough to have a statistical effect upon the diffusion of aluminum. This diffusion process will preferentially fill metal ion vacancies found in crystal defects, leaving a vacancy in the location from which the metal atom came.

Aluminum conductors in IC's are comprised of a large number of small crystal domains with random crystal axis orientation. Most of the vacancy sites lie in the interfaces between these crystal domains. As a result, the preferential diffusion paths are along crystal boundaries where the majority of the vacancies exist. High current density in the aluminum wires will cause the metal ion diffusion to flow along the crystal boundaries in the general direction of electron flow.

The flow of metal ions in the direction of electron flow in itself does not cause damage to the aluminum wires. Damage only occurs when there is a metal ion flux divergence that causes metal ions to be taken from one part of a wire and deposit them in some other location without replenishing the metal ions from a reservoir. The flux divergence causes vacancies to coalesce in one location to produce a void. This void volume grows until there is no aluminum in a short length of the wire, causing current flow is interrupted leading to circuit failure.

The accepted model for electromigration median time to fail (time required for 50% of the test samples to fail) is the classic Black's Model:

$$t_{50} = A \left(\frac{1}{J} \right)^n \exp \left(\frac{E_a}{kT} \right),$$

where A is material dependent, n is current density exponent, E_a is the thermal activation energy, J is current density, k is Boltzmann's constant and T is temperature in K.

This model is a surprisingly accurate prediction of electromigration lifetime for aluminum and its dilute alloys with copper and silicon. It assumes uniform temperature along the length of the line. The line must also be long enough to contain enough vacancies to cause a void large enough to open the line. There must not be any large mechanical stress gradients, temperature gradients or

current density gradients. Also, the operating temperature must not be above maximum temperature specified for the process. However, there are some lingering questions about the accuracy of the model prediction of t_1 (time to 1% failure) or $t_{0.1}$ (time to 0.1% failure) which are reliability numbers that are far more interesting to actual applications. Usually the tails of the time to fail distribution are corrupted by other statistical populations such as defect related early fail mechanisms.

Unfortunately, current submicrometer CMOS technologies fail to meet many of the above conditions. To complicate matters further, metal conductors are not made of a uniform aluminum alloy. They are constructed of a composite layered structure with a refractory metal layer on top and at the bottom of the aluminum alloy core material. For composite metal interconnect, the failure model must be changed from open circuit failure to one in which the line resistance increases by some percentage (10% or 20% in many applications) or in which the test line resistance changes by a specific resistance value. In this modified failure model, the Black's equation can still serve as a useful extrapolation tool.

When a significant thermal gradient or a significant current density gradient exists then the Black equation is not useful for extrapolation of median lifetime. If these conditions are present either in the test structures used for extraction of equation coefficients or in the application circuit then there is no model available that will provide a useful prediction of median lifetime. Foundries deal with these situations by de-rating electromigration design rules to mitigate the reduced reliability performance that these environmental conditions cause. At the present no electromigration model has been developed that will account for the above gradient conditions.

From the perspective of the IC designer, it would be very useful to have Black's equation model parameters in circumstances where metal interconnect area is a limiting factor in the ASIC layout. One could trade-off the foundry conservative lifetime performance using the foundry electromigration rules against a smaller expected electromigration lifetime in exchange for reduced metal area. However, foundries generally do not provide any information about Black's model parameters because of the many uncertainties involved in real applications. After all, the model parameters are extracted using test structures of a specific geometry and do not in any way represent the essentially infinite ways in which the metal lines are structured in an ASIC. Therefore, electromigration model parameters are regarded as proprietary information related to the development and maintenance of a high quality metal interconnect fabrication process. By strictly adhering to foundry electromigration design rules an IC designer is assured of maximum metal interconnect reliability from the foundry.

General Comments

Foundries usually extract model parameters for the above failure mechanisms to serve as process control parameters to maintain the reliability performance they need for their products. The parameters are considered to part of the proprietary process information that is not shared with ASIC designers. Even if these reliability parameters were provided to designers, they are not useful for predicting arbitrary circuit performance because they are valid only for the test structures and accelerated test conditions that were used to extract them. Certainly these reliability model parameters are useless for comparing one foundry with another foundry.

ASIC designers are strongly advised to use the foundry reliability design rules in laying out ASIC devices to maximize chip reliability. Designers are strongly advised to operate ASIC devices inside the operating voltage and current density limits defined by the foundry. Any ASIC device operated at voltages and/or current densities in excess of these limits may be exposed to unknown reliability hazards and will be solely at the risk of the ASIC designer.