Sensor-Driven Reliability and Wearout Management

Prashant Singh and Cheng Zhuo  
University of Michigan

Eric Karl  
Intel Portland Technology Development

David Blaauw and Dennis Sylvester  
University of Michigan

Editor’s note:  
Gate oxide degradation is a key limiter to semiconductor reliability. Because of variations in gate oxide thickness, however, product reliability is often guaranteed by designing for the worst case. This article describes the use of oxide-degradation sensors and oxide-thickness sensors to reduce these design margins and enhance performance.

—Jim Tschanz, Intel

Semiconductor reliability is a growing concern in the face of shrinking device-critical dimensions and transistor integration that continues to roughly double every 24 months. Aggressive oxide thickness scaling has caused large vertical electric fields in MOSFET devices, a situation that makes oxide breakdown a crucial issue when supply voltage is not scaled as aggressively as transistor feature size. It therefore becomes increasingly difficult to ensure the reliability of ICs over their lifetime.

Gate oxide degradation leads to poor device characteristics resulting in delay and voltage swing degradation, and in increased gate leakage, both of which eventually lead to IC failure. The gate oxide degradation is nondeterministic and depends strongly on process variation, temperature, voltage, and device state. Traditionally, designers have handled these uncertainties by making worst-case assumptions for each variable. In practice, however, worst-case conditions rarely occur because some devices have thicker oxides than others, or they operate at lower temperatures. Also, because wearout is a cumulative process, periods of low stress lengthen the device’s overall lifetime. Therefore, assuming worst-case conditions is conservative, and it leads to overly large reliability margins that designers could otherwise have traded for performance by raising the operating voltage.

Dynamic reliability management (DRM)—which has been proposed as an alternative to the traditional, conservative approach to reliability—attempts to recover some of these unnecessary margins for chips operating in conditions that are better than worst case. DRM identifies chips that have excessive reliability margins due to operating conditions, and then raises the voltage on these chips to bring their reliability down to the specified constraint, thereby improving their performance.\(^1,2\) Current DRM methods use precalibrated degradation models that are fed temperature and operating voltage conditions obtained from sensors. These models compute the chip’s reliability, and the DRM system enhances or lowers the chip’s voltage and performance, based on the updated reliability budget. However, current DRM methods do not consider process variation, and therefore the models are calibrated assuming worst-process conditions. This limits the benefits that can be obtained with the approach. DRM also relies heavily on the calibration of reliability models with respect to operating conditions, which might raise questions of accuracy.

In this article, we propose two new approaches to improve existing DRM methodology. First, we propose reliability sensors that use small replicated circuits to directly measure device wearout on the chip.\(^3\) A direct degradation measurement by these sensors removes a layer of uncertainty introduced because of inaccurate calibration of the degradation models. Note that, despite using the degradation sensors, we still require the degradation models in order to make reliability projections for the chip’s remaining lifetime. Measurements from the degradation sensors...
can be used to improve the calibration of the models and correct their projections throughout the chip's lifetime. In the second approach, we propose the use of process variation sensors on a die. These sensors impart process information to the DRM system that is used to calibrate the degradation models and hence reduce pessimistic process assumptions. Together, these sensor-based approaches offer a means of managing IC wearout and improving the performance-reliability trade-off.

**Improving DRM: Two new approaches**

In our first approach, we directly measured the wearout of sensor devices that are exposed to the same conditions as the devices on the die. These sensors let us obtain a more direct measure of the wearout effect that includes the impact of process, process temperature, and voltage drops. One difficulty of monitoring wearout is that the monitoring is inherently statistical—one sensor cannot reliably predict the wearout of other devices, even if the sensor is exposed to exactly the same process and operating conditions. We therefore had to use a set of sensors to obtain a statistical reading of the wearout. From these sensors we made accurate predictions, with a specified confidence level, about the device's lifetime.

To determine how many sensors we needed, we explored the statistics of the degradation mechanism in the presence of process, voltage, and temperature (PVT) variations. If there is a very tight distribution of the time to failure (TTF), only a few sensors are required and the sensors' area is not a major concern. However, a large TTF variation across devices creates a strong case for using a large number of sensors.

In the second approach, we investigated process variation sensors to reduce the pessimism in chip reliability estimation. These sensors provided limited process information on the die and therefore allowed a more accurate, less pessimistic, die-specific prediction of the process conditions. This prediction creates more-accurate reliability and performance margins.

For example, several oxide thickness sensors can detect that a particular die has thicker oxides compared to other dies; therefore, the die can be operated more aggressively without compromising reliability specifications.

One difficulty with this approach is that the number of oxide thickness sensors is inherently limited. Thus, although they increase the visibility of a die's process conditions, sensors cannot predict the exact process condition for every device on a die, and the measured values must be combined with a statistical model of the oxide variation. However, we demonstrate that a few sensors are adequate to significantly improve the process variation prediction, leading to performance improvements of up to 26%.

To investigate these two approaches, we first studied the oxide breakdown process. We analyzed the statistical behavior of oxide degradation and studied the effects of different factors (such as process, state, temperature, and voltage) on it. On the basis of our statistical analysis, we investigated the number of sensors required to make accurate predictions about the chip's degradation state. Then we built a prototype of an oxide breakdown sensor for use with the test chip. Finally, in our second approach we present a methodology to reduce reliability margining using on-chip process variation sensors.

**Oxide breakdown modeling**

Oxide, or dielectric, breakdown is a degradation mechanism that results in a low-impedance path through an insulating or dielectric barrier. Device failures related to this low-impedance path are typically manifested as abnormally high gate leakage current, changes in circuit switching delay, or failure to switch (in severe cases of degradation). Researchers have developed several models to explain the mechanisms of oxide breakdown. A commonly used model is the anode hole injection model, according to which injected electrons generate holes at the anode that can tunnel back into the oxide and recombine with electrons to generate electron traps. Another model, known as an electron trap density model, suggests that high gate-oxide electric fields may induce sufficient energy to trigger oxide breakdown. Although researchers are still debating many details of the breakdown mechanisms, we can learn from either model that defect generation is a nondeterministic process.

To handle the breakdown statistics, a model was proposed by Suñé et al. to capture the scaling of breakdown distribution with oxide area; however, this model lacked predictive power about the dependence on oxide thickness. This model was then improved and replaced by the well-known percolation model, proposed by DeGraeve et al., which generates defects of tunneling charges to model the wearout for thin dielectric films. In this model,
when a critical defect density is reached inside the oxide volume, there is a high probability that a low-impedance defect path ultimately leads to uncontrolled current and oxide breakdown.

The percolation model places defects of a certain size into a 3D oxide volume until a path of overlapping defects is created between the top and bottom planes. Repetitive simulation at a specified dielectric thickness yields a probability density function of the defect density required to form a path between the top and bottom planes.

In applying the percolation model, we applied a Monte Carlo-based simulation methodology to generate a variety of oxide breakdown distributions with differing voltage, temperature, variation, and stress time inputs. Figure 1 outlines the hierarchical methodology we used to compute the oxide breakdown distributions. Additional details about the simulation methodology are available elsewhere.8

Failure distribution analysis

The simulation methodology we’ve just described lets us determine the oxide breakdown, or failure distributions, that we analyze further here. An important goal for the failure analysis is to explore the impact of various factors—namely process variation, variable operating condition, and the inherent randomness in degradation—on oxide breakdown. The nominal voltage and temperature in our simulation was 1.2 V and 350 K and the process node was 130 nm.

Figure 2 displays the 25th and 75th percentile values for the first failures of a simulation of 100 dies with 25,000 oxides on each die. The simulation was done with the different sets of the aforementioned factors varying at any one time: the inherent randomness (IR), IR with process variation (PV-IR), IR with state variation (State-IR), combined (PVState-IR), PVState-IR with temperature variation, and PVState-IR with voltage variation. As we added the process, state, voltage, and temperature statistics to the basic IR simulation, typically we found that the TTF was reduced. For example, the baseline IR simulation predicted a 25th-percentile TTF of 58.3 years; this became 37.1 years when we considered process variation, state, and temperature effects.

By analyzing failure distributions using the simulation methodology, we answered several crucial questions regarding wearout’s random behavior. First, the
oxide breakdown effect has an innate randomness, and outlying failures are typical. Second, voltage and temperature have a dominant effect on predicted failure time, and the effect of state dependence and process variation alter the shape of the distribution from a pure Weibull function to nearly lognormal. These observations guided us in exploring real-time reliability monitoring and predictions.

Real-time monitoring

As we’ve mentioned, DRM can be improved with the use of in situ degradation sensors that can enable real-time degradation monitoring. A degradation sensor consists of a dummy oxide device that is exposed to similar voltage, temperature, and systematic process variation as the core devices in its vicinity. But because of inherent randomness in the degradation process, the dummy device’s degradation (which would be the sensor output) differs from the core-device degradation. Consequently, we need sufficient dummy devices or sensors in order to obtain a distribution, and therefore a range, with certain confidence, on the degradation of core devices.

Based on the near-lognormal distribution shapes we have discussed, to fit a lognormal distribution to samples of simulated distributions we used two methods: least-squares and maximum likelihood estimation. The least-squares method—when applied to the earliest 30% of the simplest set (to ensure a good fit in the early failure range)—gives the most accurate fit. In subsequent analysis of our real-time monitoring approach, we used this least-squares fitting method.

The high degree of innate randomness in oxide breakdown failures necessitated that we use multiple sensors to obtain any information when we directly measured oxide degradation. In our work, we simulated a single die to analyze the effects of the number of sensor samples needed to effectively predict the die’s TTF. The die’s failure times were sampled with different sensor counts, ranging from 35 to 5,000 sensors on the die. Table 1 lists the prediction error of TTF (using the modified least-squares method).

The actual failure time of the die was 4.841 years. The sensor prediction approached a 10% average error with 1,000 sensors.

More quantitative analysis in relation to the required number of degradation sensors can be found elsewhere. Implementing 1,000 or more sensors to directly monitor oxide degradation places

<table>
<thead>
<tr>
<th>No. of sensors</th>
<th>Absolute mean error (years)</th>
<th>Min. error (years)</th>
<th>Max. error (years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>3.7148</td>
<td>-3.5437</td>
<td>18.0896</td>
</tr>
<tr>
<td>50</td>
<td>2.5499</td>
<td>-3.3444</td>
<td>18.7818</td>
</tr>
<tr>
<td>100</td>
<td>1.4777</td>
<td>-3.8320</td>
<td>6.8266</td>
</tr>
<tr>
<td>250</td>
<td>0.9150</td>
<td>-2.3437</td>
<td>3.6080</td>
</tr>
<tr>
<td>500</td>
<td>0.7514</td>
<td>-1.8663</td>
<td>3.1098</td>
</tr>
<tr>
<td>1,000</td>
<td>0.4415</td>
<td>-1.1737</td>
<td>2.3487</td>
</tr>
<tr>
<td>1,500</td>
<td>0.3537</td>
<td>-1.2094</td>
<td>1.0934</td>
</tr>
<tr>
<td>2,000</td>
<td>0.3339</td>
<td>-0.9953</td>
<td>1.3249</td>
</tr>
<tr>
<td>5,000</td>
<td>0.1929</td>
<td>-0.5549</td>
<td>0.6913</td>
</tr>
</tbody>
</table>

Figure 2. 25th to 75th percentile ranges for a 100-die simulation of 1.9-nm oxide first failures. The bars indicate different sets of factors—innate randomness (IR), process variation (PV), state, temperature, voltage—that we varied in the simulation to demonstrate their impact on failure. The nominal voltage and temperature conditions are set to 1.2 V and 350 ºK respectively.

Table 1. Error in the time-to-failure (TTF) estimate with a varying number of gate-oxide-degradation sensors on a die. The nominal oxide thickness of the gate oxide was 1.8 nm. The die TTF was 4.841 years.
strict constraints on the sensor area to realize a feasible system. The sensor size must be on the order of a standard cell macroblock to permit placement and routing for such a large number of blocks. In an era of billion-transistor systems, 1,000 to 5,000 compact sensors can be easily spared to control the difficult problem of a priori reliability qualification.

Prototype: Oxide-degradation sensor

The feasibility of realizing a system with real-time reliability monitoring depends on the realization of an oxide-degradation measurement technique or an oxide-degradation sensor that can be easily included in the system. Many oxide-breakdown measurement techniques developed by other researchers have been invasive, requiring direct access to the devices under test (DUTs) and peripheral circuitry resulting in large area overhead. Uraoka et al., for example, evaluated gate oxide reliability using a luminescence method. The setup required an optical microscope, photon counting camera, and image processor. More recently, Keane et al. proposed an array-based test structure to statistically characterize the gate oxide breakdown. As an alternative to these techniques, we designed a compact oxide breakdown sensor— which does not require direct access to DUTs—for runtime wear-out monitoring.

Figure 3 shows our prototype sensor’s design. It consists of a pair of parallel MOSFETs connected in series with a small Schmitt trigger–based ring oscillator and a differential pair to drive the gate oxide with a high voltage when stressed. The oscillation frequency was set by the gate leakage through the PMOS devices and the size of the capacitor used at N2, the node driven by gate leakage. During measurement, the stress signal was driven low to short both drain-source-bulk nodes of the PMOS devices to the oscillator’s output and to deactivate the output of the differential amplifier connected to N3. The oscillation signal was driven high to enable the ring oscillator to toggle, limited by the

Figure 3. Circuit schematic for oxide degradation–sensing oscillator: gate leakage through PMOS devices between N1/N3 and N2 controlled the sensor oscillating frequency. High voltage was applied at N3 and ground to N1 during the stress phase of operation (VDD12: nominal supply = 1.2 V; Vstress: stress voltage; VDD33: supply for thick oxide devices = 3.3 V) (a). Oxide-degradation sensor: initial oscillation frequency distribution across dies (b).
gate leakage current through the parallel PMOS devices. To stress the oxide devices, node N2 was driven through a series-resistive gate leakage divider formed by M0 and M1, while N3 was driven to a stress voltage \(V_{\text{stress}}\) (in operation this may be local \(V_{\text{DD}}\)) and N1 was tied to ground. In this case, we held N2 to a nominal voltage that was half of the applied voltage at N3; this remained fairly constant until breakdown occurred, since an increase in leakage in one PMOS device in series led to an increased voltage drop in the other, eventually evening the imbalance in degradation.

A scheme of two gate oxides in a stack was essential to isolate node N2 from spurious source and drain diffusion currents. During our measurements, node N2 was driven solely by the gate leakage of M0 and M1. In the 130-nm technology in which we implemented this circuit, subthreshold current was much larger than gate leakage, eliminating the possibility of any source- or drain-connected transistors at node N2. In newer technologies, it might be possible to reduce subthreshold leakage current below gate leakage current levels by using appropriate circuit techniques. This would obviate the use of two stacked MOS devices. This sensor can be converted into an oxide-thickness sensor by disabling the stress mode, as its output is proportional to the oxide thickness.

To test the proposed prototype sensor, we fabricated a test chip consisting of 144 oxide-breakdown sensors in a 130-nm technology node. In this process, the nominal gate oxide thickness for standard devices is 2.2 nm. The oxide sensor area of our test chip was 150 \(\mu\text{m}^2\), approximately 3.3 times the size of a D flip-flop.

The sensor's small size meant that we could use a large number of them, facilitating much more statistical degradation data. Based on 860 sensor measurements across nine sample dies, the statistical distribution of the initial gate leakage values indicated a \(3\sigma/\mu\) value of 57% within a die and 98% across the nine dies (see Figure 3b).

To accelerate oxide-degradation testing, we used ambient temperatures ranging from 130°C to 175°C and \(V_{\text{stress}}\) voltages of 5–6 V that were sent to the sensor to achieve a stress voltage of 2.5–3 V across each PMOS device. The sensor was stressed for 56 hours at 2.5 V and 130°C across each oxide. Results showed a sharp initial rise in oscillator frequency, followed by a steady increase, ultimately exhibiting 19% average degradation with a range of observed results from 5% to 40% (see Figure 4). More details on the sensor operation are available elsewhere.

**Figure 4. Silicon measurements from oxide-degradation sensor. Relative frequency shift after 56-hour stress period (a). Frequency change during oxide stress time (b).**

**Process variation sensor-driven DRM**

We've just explained about introducing in situ degradation sensors to provide an accurate degradation state of the chip to the DRM system. The DRM system then uses the degradation models to project the chip's reliability for the remaining lifetime of the chip. Here, we will discuss our second approach, which uses process variation sensors (oxide thickness variation sensors) to calibrate the degradation models rather than using a worst-case process for that purpose. This lets us make a less pessimistic reliability projection, using limited process variation sensor readings to statistically estimate chip oxide thickness,
and hence allows for more performance improvement. As we mentioned earlier, such a process variation sensor can be made by modifying the prototype oxide-degradation sensor. We conducted this study in a 65-nm process node with a nominal oxide thickness of 1.67 nm.

As discussed earlier, the variation in oxide thickness creates a wide distribution in a chip’s TTF. To determine chip lifetime distribution, we conducted a Monte Carlo simulation of 50,000 chips distributed across different wafers, with each chip consisting of 0.5 million devices (see Figure 5). The chips on the same wafer are assumed to have similar gate oxide thickness distribution. The chip lifetime distribution (curve with the square) obtained by the simulation (using $3\sigma/\mu = 4\%$) had the lognormal shape with a long tail, the 99.9% reliability confidence point of which is 25.5 years (the 99.9% confidence point is defined as the time in which first 0.1% of the total chips fail). However, the lifetime distribution of the chips on a particular wafer is much tighter. Figure 5 shows the lifetime distribution for chips on two different wafers—one with thinner gate oxides (for the curve with the triangle, the 99.9% reliability confidence point is 11.6 years) and the other with thicker gate oxides (for the curve with the circle, the 99.9% reliability confidence point is 38.7 years). If the oxide thickness variation can be estimated using limited oxide thickness measurements (using process variation sensors), then the reliability margin available in the thick-oxide chips can be traded for higher performance.

Estimating the variation using limited measurements in a mathematically rigorous manner is a non-trivial problem. First, we must account for the limited number of measurements, whereas the number of transistors might exceed millions or even billions per chip. This imbalance between measurement and prediction complexity makes it difficult to fully utilize the measurement information. Second, the conventional interpolation and smoothing techniques can’t be employed because, unlike a thermal profile, oxide thickness across the chip shows non-continuous characteristics.

To address those challenges, we used a postfabrication measurement-driven statistical methodology. We assumed that the process variation sensor supplied the oxide thickness for a set of devices on a chip. Our methodology leveraged the spatial correlation between sensor readings and then constructed a conditional die-specific oxide thickness distribution. Once we achieved estimation of the oxide thickness and the corresponding variance, we provided this information to the degradation model in the DRM system. For our analysis, the degradation model assumed a Weibull relationship between device reliability and oxide thicknesses. Using the process information, the model deduced the chip reliability distribution, which could then be projected to the TTF distribution given a certain reliability requirement. The TTF distribution was tighter than what worst-process variation assumption gave, and hence enabled the DRM system to boost system performance by selecting a higher supply voltage limit while the lower bound of the TTF distribution still met the design requirement.

To assess the gains from this technique, we applied this technique to a DRM system without dynamic voltage scaling. Based on the computed TTF distribution, the DRM system determines the supply voltage of the chip. With the results obtained from this technique, we performed a one-time postsilicon supply voltage tuning. Table 2 summarizes the results from this simulation experiment on 10,000 chips with 0.5 million devices on each chip using 25 and 100 process variation sensors. It was assumed that the reading from the process variation sensor was 100% accurate. The table...
shows that we can achieve a performance improvement of 15% on average and 26% at maximum for 10,000 chips, using only 25 process variation sensors, to reach the reliability target. The average runtime to execute this scheme is only approximately 0.4 seconds per chip.

**Oxide Breakdown is a Major** degradation mechanism that continues to threaten IC reliability in sub-32-nm technology nodes. We studied the statistical nature of oxide breakdown, which is a random process. PVT variations further increase this randomness. Due to this statistical nature of oxide degradation, designers traditionally have been forced to assume worst conditions to solve this problem. However, this approach prevents designers from using the full potential offered by cutting-edge process nodes while exceeding the reliability specification for most of the ICs.

The goal with DRM is to manage reliability by sensing the operating conditions and balancing those against performance gain; however, DRM makes pessimistic assumptions for process variation and suffers from inaccuracy owing to inaccurate degradation models. The two approaches we've proposed reduce this pessimism: first, by directly measuring the oxide degradation using in situ oxide-degradation sensors; second, by using process variation sensors to give process information to the DRM system. With the first approach, we've demonstrated a prototype of an oxide breakdown sensor. The compact size of the sensor (3.3 times the size of a minimum-size D flip-flop) enabled real-time monitoring of reliability with low area overhead. In our second approach, applying process variation sensor data, the DRM uses a more accurately calibrated degradation model, which permits a boost in supply voltage and thereby results in a performance improvement as high as 26%. Designers can easily combine these two approaches to effectively control the difficult problem of a priori reliability qualification while utilizing the full potential offered by cutting-edge process nodes in the sub-32-nm design era.

### References

---

**Table 2. Optimization results from 10,000 chips for a design with 0.5 million devices (nominal thickness is 1.67 nm for 65-nm LP devices) using 25 and 100 measurements (the nominal voltage is 1 V and temperature is 70°C). The results are compared with a guardband approach that sets supply voltage at 0.858 V to meet reliability target. The table summarizes the minimum, maximum, mean, and standard deviation of optimized voltage and performance by constructing the histogram of 10,000 optimized chips.**

<table>
<thead>
<tr>
<th>No. of chips</th>
<th>No. of measurements</th>
<th>Optimized supply voltage (V)</th>
<th>Performance improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>10,000</td>
<td>25</td>
<td>0.9</td>
<td>1.05</td>
</tr>
<tr>
<td>10,000</td>
<td>100</td>
<td>0.92</td>
<td>1.05</td>
</tr>
</tbody>
</table>


Prashant Singh is a doctoral candidate in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include high-speed interconnects, circuit reliability, and error-tolerant designs. Prashant has a BS in electrical engineering from the Indian Institute of Technology Kanpur, India.

Cheng Zhuo is a doctoral candidate in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include circuit analysis and optimization, with an emphasis on design for yield and reliability. Cheng has an MS in information science and electronic engineering from Zhejiang University, Hangzhou, China. He is a student member of the IEEE.

Eric Karl is a design engineer in Intel’s Portland Technology Development organization in Hillsboro, Oregon. His research interests include emerging memory technologies, low power circuits and design for reliability, variation-tolerance and manufacturability. He has a PhD in electrical engineering from the University of Michigan, Ann Arbor. He is a member of the IEEE.

David Blaauw is a professor in electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include VLSI design and CAD, with an emphasis on circuit design and optimization for high-performance, low-power applications. Blaauw has a PhD in computer science from the University of Illinois, Urbana Champaign. He is a member of the IEEE.

Dennis Sylvester is an associate professor of electrical engineering and computer science at the University of Michigan, Ann Arbor. His research interests include low-power circuit design and design automation, design for manufacturability, and on-chip interconnect modeling. Sylvester has a PhD in electrical engineering from the University of California, Berkeley. He is a senior member of the IEEE.

Readers may contact Cheng Zhuo about this article at the University of Michigan, 2260 Hayward St., CSE 4844, Ann Arbor, Michigan 48109-2121; czhuo@umich.edu.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.
Get the building blocks you need.

Take your career to the next level in software development, systems design, and engineering with:

- Article collections from the IEEE Computer Society
- Materials from Harvard Business School Publishing
- Computer discounts
- Online courses and certifications

Our experts. Your future.

www.computer.org/buildyourcareer