Unreliable Silicon: Circuit through System-Level Techniques for Mitigating the Adverse Effects of Process Variation, Device Degradation and Environmental Conditions

by

Eric Alexander Karl

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2008

Doctoral Committee:

Professor David Blaauw, Co-Chair
Associate Professor Dennis M. Sylvester, Co-Chair
Professor Trevor N. Mudge
Associate Professor Michael Flynn

© Eric A. Karl
2008

Dedicated to my mother and father,
for their unwavering support and patience
with me throughout my life.

Acknowledgements

I would like to acknowledge the invaluable contributions of my advisors Dennis Sylvester and David Blaauw during my time at the University of Michigan. Many times you refocused my work when the light seemed “dim”, inspired me and hardened my resolve to see my projects to a fitting conclusion. Thanks to Trevor Mudge and Michael Flynn for agreeing to serve on my thesis committee and for additional insight provided along the way. Special thanks to Todd Austin and the resilient group in GSRC; some of the quarterly meetings sparked new ideas in my work.

I’ve had the opportunity to work with some truly talented graduate students during the past six years, some of whom merit individual recognition. Prashant, thanks for your patience, hard work and your “resilience” on our many collaborative projects. Carlos, your attention to detail, willingness to help and overall demeanor make the research group a better place to work. Sanjay, Yushiang, Scott, Mike, Greg and those I haven’t mentioned, I’ve enjoyed the opportunities that I’ve had to work on our various tapeouts over the last few years. I have many fond memories from difficult times largely due to the teamwork and cooperation in our lab. Thanks to Ashish, Sarvesh, Kanak and Himanshu for the solid foundation and reputation that you helped to build for our research group. I will truly feel proud to be an alumni of this group.

A large number of mentors in industrial positions have positively influenced my work and my development through interactions on my interships and coops. In no particular order, thanks to Spencer Gold, Steve Kosonocky, Dan Knebel, George Gristede, Keith Bowman, Vivek De and Robert Schwabel. Each of you positively impacted me at various points in my development as a young engineer.

Table of Contents

Dedication ii

Acknowledgements iii

List of Figures vi

List of Tables ix

Chapter

1. Introduction 1
2. Timing Error Correction for On-Chip Memories 15
   - 2.1 Timing Error Detection and Correction 18
   - 2.2 Exploiting Address/Data Dependence 24
   - 2.3 180nm Simulation Results 26
   - 2.4 130nm Physical Implementation 32
   - 2.5 130nm Simulation Results 40
3. Reliability Modeling and Management 46
   - 3.1 Reliability Modeling 49
   - 3.2 System Level Modeling 63
   - 3.3 DRM System 67
   - 3.4 Results and Discussion 73
   - 3.5 Summary 82
4. Analysis of Real-time Reliability Monitoring 84
   - 4.1 Oxide Breakdown Simulation Methodology 87

4.2 OBD Monte Carlo Framework 90

4.3 Failure Distribution Analysis 94

4.4 Real-time Monitoring 98

4.5 Summary 102

5. A 130nm Oxide Degradation Sensor.. 104

5.1 Oxide Degradation Sensor Design 105

5.2 130nm Testchip Implementation.. 111

5.3 Measured Results and Discussion 114

5.4 Summary and Discussion... 117

6. A 45nm NBTI/Oxide Degradation Sensor 119

6.1 2nd Generation Oxide Degradation Sensor Design.................... 120

6.2 45nm Testchip Implementation.. 127

6.3 Simulated Results and Discussion 130

6.4 Summary 133

7. Conclusion 135

7.1 Future Work.. 138

Bibliography 140

List of Figures

Figure

1.1 2005 ITRS low power roadmap scaling of oxide thickness, threshold voltage and supply voltage. 3

1.2 Proposed ElastIC architecture, showing multi-processing elements in different operation modes. 10

2.1 Main and Shadow Sense Amplifiers latched and compared via XOR gate and multiplexed to an output bus.......... 19

2.2 Timing signal generation circuits and waveforms. 21

2.3 Proposed SRAM floorplan. 23

2.4 Error rate of memory as VDD reduces for gap00 memory trace. 27

2.5 Error rates for a set of benchmarks vs. supply voltage. 28

2.6 Instantaneous supply voltage and error rate during 10M cycle DVS simulation with varying workload. 31

2.7 Array Architecture and Output Path. 33

2.8 Sense amplifier and integrated 2:1 column multiplexer and layout of new local block showing area overhead of “RAZOR” or *shadow sense amplifiers*. 35

2.9 Output latch with added *shadow latch* and error correction and detection circuitry. 38

2.10 Distribution of Read Current for 130nm and 45nm technologies. 41

2.11 SRAM Read Path Timing and Read Current scaling with VDD for 130nm technology. 42

2.12 Minimum operating voltages for various combinations of threshold voltage variation for the implemented design of the time redundancy test chip. 44

3.1 Reliability Degradation over time. 47

3.2 Proposed piecewise NBTI calculation algorithm. 62

3.3 DRM System block diagram. 68

3.4 PID Controller gain values vs. peak performance improvement for example DRM system. 71

3.5 DRM Operation for workload C630 (16 seconds) and supply voltage assignment histogram. 76

3.6 DRM Operation over 10 year reliability simulation. 77

3.7 DRM PID controller step input response, representing sharp change in usage profile mid-life. 78

3.8 Peak performance and max supply voltage vs. workload activity considering NBTI. 80

3.9 Peak demand vs. peak performance improvement (with NBTI) 82

4.1 Percolation model placement of point defects in 3-dimensional oxide space (2D diagram shown). 88

4.2 Monte carlo simulation framework for generating oxide breakdown distribution. 91

4.3 Oxide failure distributions with different simulation components included. 95

4.4 Failure distribution for 1.9nm oxide with conditions varying from 1.15-1.25V and 60-110C. 96

4.5 25th to 75th percentile ranges for 100 die simulation of 1.9nm oxide first failures 97

4.6 95% confidence interval in TTF prediction as sensor count increases for 2 different dies. 101

5.1 Simple model of oxide degradation sensor measurement featuring series gate oxide resistance and thick oxide Schmitt trigger oscillator. 106

5.2 Oxide degradation sensor schematic. 107

5.3 Simulated waveforms demonstrating sensor operation. 110

5.4 Oxide degradation sensor standard cell compatible layout. 112

5.5 (Above) Complete 130nm multi-project testchip. (Below) Detailed view of oxide and NBTI degradation sensor test chip with block layout view inset.. 113

5.6 Oscillator frequency increase as the gate oxide is stressed at 2.5V and 130C for 3 sensors from a single die. 114

5.7 Final oxide degradation sensor frequency increase following 56 hours stress period at 2.5V and 130C. 115

5.8 Oxide degradation sensor frequency increase as a function of time. 116

5.9 Distribution of initial oxide degradation sensor oscillating frequencies. 117

6.1 45nm oxide degradation sensor circuit diagram. 122

6.2 Simulated operation of 2nd generation 45nm oxide degradation sensor. 126

6.3 45nm oxide degradation sensor cell layout. 127

6.4 45nm multi-project testchip including 2nd generation oxide and NBTI degradation sensors. 129

6.5 Oscillation frequency vs. oxide leakage scaling for extracted parasitic simulation. 131

6.6 45nm 2nd generation oxide degradation sensor temperature scaling. 132

6.7 Monte carlo simulation of 45nm oxide degradation sensor design. 133

List of Tables

Table

1.1 Constant Field Scaling Factors 1

2.1 Power Consumption at Various Operating Points 30

3.1 Simulation Technology Specification and Selected Model Parameters of Interest 74

4.1 Parameter Values for Simulation Framework... 92

4.2 Spatial Correlation between Blocks in the PV Model 98

4.3 Sensor Count and Prediction Error based on Ideal Sensors....................... 100

5.1 130nm Oxide Degradation Sensor Sizing 109

6.1 Subthreshold vs. Gate Leakage in 2 Process Technologies (Normalized).. 121

6.2 45nm Oxide Degradation Sensor Sizing 122

Chapter 1

Introduction

The continued, aggressive downscaling of dimensions in forthcoming CMOS technology generations [1] stands to increase the risk of significant reliability issues in integrated circuits. Conventional or constant electric field scaling law, in Table 1.1, dictates that supply voltages must be reduced to maintain a constant electric field as the critical dimensions of CMOS transistors and wires shrink. However, in recent years, the voltage scaling trend of constant electric field scaling [2] has been ignored to maintain the saturation current and other performance metrics [1].

TABLE 1.1
Constant Field Scaling Factors

Parameter	**Scaling Factor**
Length, Width, Oxide Thickness	$1/\alpha$
Doping Concentration	α
Supply Voltage	$1/\alpha$
Circuit Speed	α
Circuit Power	$1/\alpha^2$
Device Density	$1/\alpha^2$
Power Density	1

Over the past ten years, device designers have found that it impossible to scale supply voltage down to match the dimension scaling due to the opposing constraints

of standby power consumption and saturation current. The delay of digital circuits with short-channel MOSFETs is directly related to the velocity saturated drain current [3]. To continue scaling supply voltage and maintain a constant electric field, threshold voltage must be reduced to provide improvement in saturation current to offset the reduced supply voltage, according to the short channel velocity-saturation-limited current relationship [4] in Equation (1.1).

$$I_{dsat} = C_{ox} W \vartheta_{sat} (V_g - V_{th}) \tag{1.1}$$

$$V_{dsat} = \sqrt{2\vartheta_{sat} L (V_g - V_{th}) / m\mu_{eff}} \tag{1.2}$$

Notice that Equation (1.1) is independent of channel length and linear with respect to gate overdrive, V_g-V_{th}. The saturation voltage required to achieve velocity saturation, in Equation (1.2), is a function of channel length (L), but the magnitude of short channel velocity saturated current is not a function of channel length, in the first order. Successive technology generations scale width and length with other dimensions in Table 1.1 to increase circuit density and minimize area, to save fabrication cost. In Equation (1.1), equal reduction in transistor width and oxide thickness are offset by the changes in C_{ox} (increases with decreasing T_{ox}) and W. The reduction in oxide thickness traditionally must be accompanied by a reduction in supply voltage (V_g) to maintain reliability integrity of oxides, and control dynamic power consumption, as detailed in Table 1.1. Since the saturation velocities of hole and electrons are constant, any reduction in supply voltage (V_g) must be accompanied by a reduction in threshold voltage (V_{th}) to maintain the saturation current of the scaled MOSFET.

The dramatic increase in off-state leakage of MOSFETs as technology scales [5] prevents designers from actually scaling the threshold voltage to maintain the saturation current. Figure 1.1, from the 2005 ITRS update data clearly displays these trends; oxide thickness is constantly scaled, while supply voltage and particularly threshold voltage remain nearly constant across this 15 year span. The result of this compromise is that with each technology generation, the supply voltage is not reduced at the rate dictated by constant-field scaling theory and the electric field across thin-film oxides and junctions is steadily increasing. The drawback is that reliability issues and device degradation become increasingly probable for most breakdown mechanisms due to the elevated electric fields.

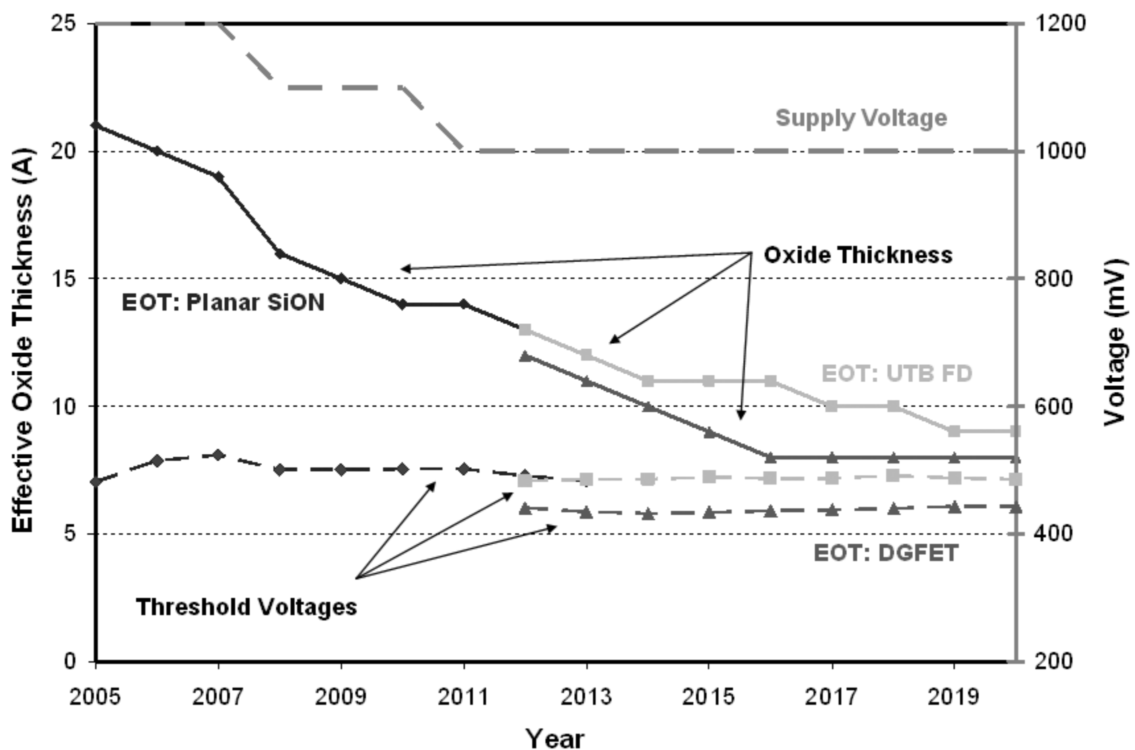


Figure 1.1 2005 ITRS low power roadmap scaling of oxide thickness, threshold voltage and supply voltage.

Oxide breakdown (OBD) [6], electromigration (EM) [7] and the negative bias temperature instability effect (NBTI) [8] are three reliability failure mechanisms for integrated circuits that are strongly impacted by electric fields in the gate oxide region. OBD and NBTI mechanisms degrade device structure through collisions between particles in the oxide lattice or the oxide-silicon interface region. Two widely accepted models for oxide breakdown modeling, the E_{ox} model [9] in (1.3) and the $1/E_{ox}$ model [10] in (1.4) clearly show exponential reduction in time to oxide breakdown with increasing oxide electric fields.

$$t_{BD} \propto e^{\frac{\Delta H_0}{k_B T} - \gamma E_{ox}} \tag{1.3}$$

$$t_{BD} \propto e^{G/E_{ox}} \tag{1.4}$$

The rate of threshold voltage shift in bias temperature instability also increases exponentially according to the NBTI stress model [11] in (1.5). The rate of interface trap generation and shift in threshold voltage (ΔV_{th}) is dependent upon gate voltage and oxide electric field, which are constant and increasing, respectively, under the current paradigm of transistor dimension scaling. Secondarily, increasing temperature also increases the rate of threshold shift in an exponential fashion (activation energy, E_a is typically negative).

$$\Delta V_{th} \propto A t_{ox} \sqrt{C_{ox}(V_{gs} - v_{th})}\left[1 - \alpha V_{ds}/(V_{gs} - V_{th})\right] e^{(E_{ox}/E_0)} e^{-\frac{E_a}{kT}} \tag{1.5}$$

Electromigration degradation occurs in metal wires with high current density when metal atoms shift along grain boundaries due to diffusion and collisions. Black's law [7] is the classic relationship between current density, temperature and

mean time to failure for wires, as show in (1.6). In (1.6), J is the current density, which typically increases, with high electric fields across MOSFET drains and gates, leading to a super-linear decrease in mean time to failure. Similar to the NBTI effect, electromigration lifetime decreases with increasing temperature, T.

$$MTF = AJ^{-n}\exp\left(\frac{E_a}{kT}\right) \tag{1.6}$$

In all of these mechanisms, higher electric fields and temperatures lead to vastly increased rates of degradation and higher incidence of failure. The continued non-scaling of supply voltage in process technologies is clearly exchanging lifetime reliability and wearout margin for maintaining saturation current expectations and off-state leakage budgets. The importance of reliability qualification techniques is now enhanced due to the diminishing margins through increasing electric fields in CMOS circuits.

Traditional stress-based reliability qualification techniques, such as the JEDEC JESD-47 Standard [12], qualify designs by stressing sample systems under pessimistic environmental conditions with a zero failure pass/fail criteria. While the traditional approach is an accepted method of ensuring reliability, the limits it places on supply voltage and temperature leave a significant and increasing reliability margin between circuit performance at worst-case conditions and at typical conditions. Widely varying environmental conditions and significant process variation increase the margin required to meet worst-case design time conditions. This unnecessarily limits the specifications of the majority of integrated circuits to the performance of a highly improbable set of process variation and environmental conditions.

In [13], Borkar identifies two critical sources of increasing process variation, random, discrete placement of dopant atoms in shrinking transistor channel regions and patterning shapes with feature sizes below the wavelength of the lithographic light source. Chen recently declared that line edge roughness and dopant fluctuation pose fundamental barriers to controlling device parameters [14]. Reliability mechanisms like NBTI are directly related to random dopant fluctuations, since the process variation shifts the initial threshold voltage of a device from its nominal value, and the NBTI effect adds a dynamic, time-varying shift to the threshold voltage. Combined the static and dynamic shift from process variation and reliability degradation severely limits the worst-case performance of a transistor. As discussed above, continued scaling increases the possible range of transistor threshold voltages by increasing the impact of dopant fluctuation in a smaller channel region (seen in Equation (1.7) [4]) and also increasing the electric field across gate oxides to exacerbate NBTI degradation.

$$\sigma V_{th} \propto \frac{q}{C_{ox}} \sqrt{\frac{N_a W_{dm}^0}{3LW}} \tag{1.7}$$

Using sub-wavelength light to pattern features results in significant variation in edge resolution and sharpness. This is primarily seen as line edge roughness that impacts wire width and the channel length of polysilicon gates. Variations in wire width and thickness complicate the projection of wire lifetime when considering electromigration. Variations in gate oxide area, as well as thickness, significantly alter the breakdown statistics for oxide breakdown modeling.

In addition to the impact of process variation, semiconductor products are used in an increasingly wide range of applications, with widely varying environmental

conditions. Increased range of ambient operating temperature, leads to a greater variation in on-chip temperature. From (1.5) and (1.6), temperature is an exponential factor on many reliability degradation mechanisms, including NBTI, electromigration, and thermal cycling. Temperature is another factor increasing the margin utilized by traditional worst-case reliability qualification techniques.

Hence, there are opportunities for alternative approaches to ensuring lifetime reliability under dynamic operating conditions. One alternative to stress-based qualification, knowledge-based risk assessment, is a sophisticated methodology when compared to simplistic corner-case stress testing. The knowledge based approach (a framework for knowledge-based qualification is defined by JEDEC JESD-34 [15]) requires careful characterization and analysis of individual failure modes to assess a reasonable system reliability risk factor given the reliability targets for the system. Despite the advantages of the knowledge-based risk assessment methodology, it still suffers from limited information on actual stress conditions on-chip and must use some type of a priori profiling that will introduce voltage or timing margins.

An alternative to placing the burden of reliability qualification entirely upon a priori characterization techniques is the recent proposal to utilize adaptive systems, capable of monitoring and self-diagnosis. Adaptive systems can self-regulate voltage, temperature or other critical parameters to dynamically maintain their specified reliability lifetime. The goal of the dynamic, adaptive system architectures and enabling circuits is to vastly reduce the margins on timing and voltage for typical silicon, while maintaining proper functionality for devices or structures with worst-

case stress conditions or variation. The contribution of this thesis is in adaptive circuit and architecture techniques for coping with highly unpredictable and unreliable silicon.

Early work on dynamic voltage scaling systems (DVS) targeted power reduction for digital circuits by dynamically reducing the power supply voltage when the target application or circuit did not require a full nominal voltage. Coincidentally, DVS systems are an extremely powerful way to improve reliability lifetime of integrated circuits due to the exponential dependence on electric field of many reliability mechanisms. DVS systems [16-18] initially targeted voltage reduction only where the application had periods of reduced computational demand. Frequency is typically reduced as the voltage was scaled, preserving voltage margin in the design despite the reduced performance required.

RAZOR [19, 20] was proposed as a DVS system that reduced voltage at a fixed frequency, aiming to achieve power reduction by exploiting data-dependent circuit delays and reduce the timing and reliability related voltage margin that is added through worst-case design methodologies. RAZOR utilized a novel timing error detection circuit built into flip-flop circuits that allowed detection of late arriving signal transitions due to process variation, voltage droop or wearout effects. Another novelty in the RAZOR work is the concept of scaling the voltage below the point of the first detected failure, leveraging the fact that first failures occur very infrequently. Recent work [21, 22] on the RAZOR concept has resulted in latches with integrated timing error detection circuitry that refines the approach to obtain up to 17%

reduction in supply voltage. From a reliability standpoint, 17% reduction in oxide electric field increases typical process oxide lifetime by several orders of magnitude.

The drowsy cache [23] was proposed by Flautner to reduce the standby leakage for inactive memory banks by using DVS to reduce their voltage. In Chapter 2, a technique similar to RAZOR is implemented in SRAM sense amplifiers to enable dynamic and standby power reduction for all banks of memory while improving the reliability lifetime of all memory circuits [24]. The technique utilizes a secondary sense amplifier to evaluate the bitline differential development through an extended evaluation phase in order to capture read timing related errors from decode logic, SRAM cell read current or sense amplifier threshold voltage mismatch. Simulated results show that this technique is successful in eliminating 200-300 mV of design voltage margin in 180nm process technology.

Beyond the initial DVS based work targeted at power reduction, Sylvester [25] proposed ElastIC, a broad adaptive architecture targeted at delivering reliable system functionality in an era of unpredictable and unreliable silicon devices. ElastIC is a framework for massively multi-processor systems based upon self-test and diagnosis, adaptive operation and stress relaxation techniques. As shown in Figure 1.2, the ElastIC architecture leverages research concepts in the device, circuit, architecture and testing fields.

The foundation of ElastIC is an array of processing elements with built-in self-test (BIST) and power, performance and reliability monitoring circuits. Each processing element (PE) is tunable through voltage, frequency or timing boundaries in tunable flip-flops. The diagnostic and adaptivity processing (DAP) unit is centrally

located to manage the individual testing of each PE and schedule the variety of adaptive features for each PE. Some PEs will have an elevated voltage for high performance, some have adjusted flip-flop timing signals to improve functional yield. Other elements are in a low power sleep state, or placed into a beneficial bias state to "heal" reversible reliability related degradation (such as NBTI or positive bias temperature instability effect (PBTI)). Highly redundant memory and processors combine to deliver high reliability despite underlying uncertainty in the silicon devices.

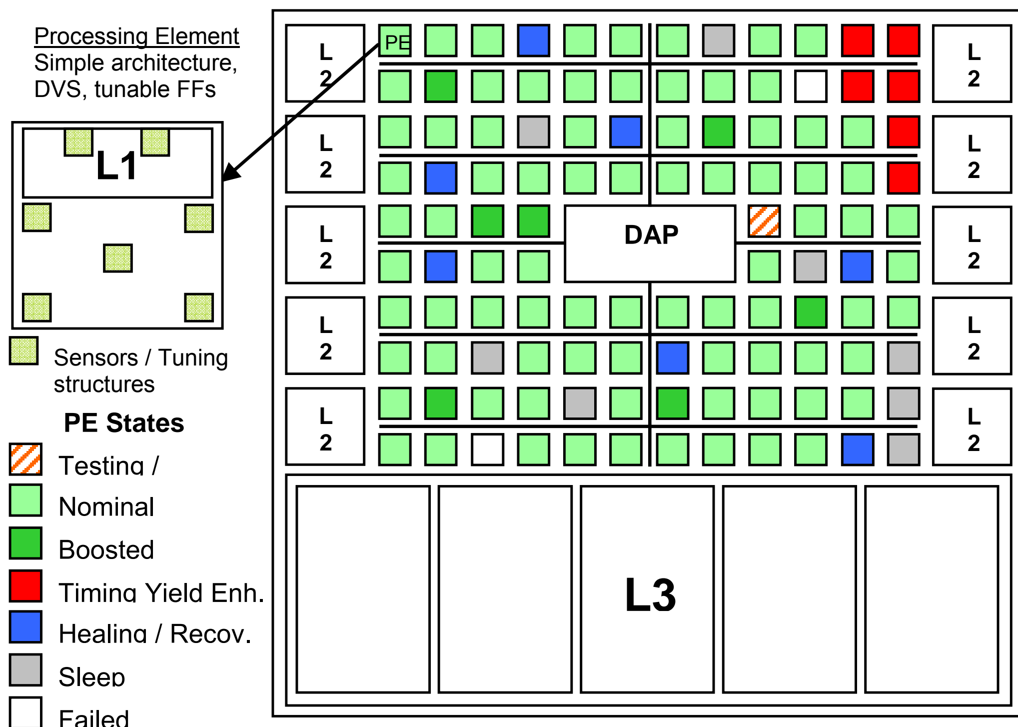


Figure 1.2 Proposed ElastIC architecture, showing multi-processing elements in different operation modes. Processing elements consist of a mini processor with dynamic voltage scaling and tunable flip-flop timing.

The precursor to ElastIC, was the initial proposal of dynamic reliability management systems (DRM) [26] by Srinivasan. DRM is a control architecture for integrated circuits and systems that attempts to dynamically ensure that the component meets its reliability lifetime specification by tuning voltage, restricting circuit activity and temperature in order to adjust the current rate of reliability degradation. The key concept introduced in DRM work was modeling reliability as a budget that can be consumed at different rates, based upon the stress and activity of the system. Srinivasan used a sum of failure rates modeling method while considering multiple reliability mechanisms and discussed the potential implementation of several control mechanisms.

Lu and co-authors [27] analyzed electromigration effects and suggested a dynamic thermal management (DTM) system. Lu’s proposed DTM system monitors temperature of junctions on-chip and throttles execution when necessary to keep system temperature within specified bounds. This system is considerably simpler than Srinivasan’s DRM system, yet also significantly reduces reliability uncertainty by restricting the range of operating temperatures dynamically.

McGowen and co-authors [28] describe an embedded feedback and control system implemented into a commercial microprocessor that simultaneously monitors and limits power consumption and junction temperature. Intel Foxton Technology (FT), as referred to in the original work, modulates voltage and frequency dynamically to maximize core clock frequency within specified power and temperature envelopes. The FT system is currently integrated on a shipping product, consuming 0.5% of the die area and 0.5% of the processor power budget.

In Chapter 3, a full multi-mechanism DRM system is described in detail using a PID controller to actuate voltage limits in a DVS system [29, 30]. System-level probabilistic modeling and a method for extrapolating simulation to full lifetime projection is detailed. With an implementation similar to Lu and McGowen, this work integrates a full DRM modeling system similar to the style of control needed by the DAP unit in the ElastIC architecture, relying upon on-chip measurement of voltage, temperature and reliability degradation to make accurate real-time reliability projection.

One drawback to the implementation of a full DRM system attempting to monitor reliability degradation in real-time, is the lack of suitable sensor circuits for directly observing the degradation on chip. In the work by McGowen, the power and temperature on chip are modeled directly, allowing the control system to accurately restrict power and temperature. Full DRM systems require similar on chip monitors for oxide degradation, NBTI and even electromigration to improve the feedback and control loop accuracy for improving system lifetime.

Kim proposes the Silicon Odometer [31] as an on-chip degradation monitor targeting generic delay degradation in digital circuits. The sensor consists of a pair of 105-stage ring oscillators, one of which is “stressed” and the other remains unstressed, that generate a beat frequency representing the difference in delay of the comparable oscillators. The results in this work attribute NBTI degradation for changes in oscillator frequency. The silicon odometer sensor consumes $265 \times 132 \mu m^2$ in a 130nm process, roughly equivalent to 4900 NAND3 gates in that technology.

Keane introduced a specific sensor for monitoring NBTI degradation [32] utilizing a delay-locked-loop (DLL) to lock to a stressed delay element, outputting a voltage representing the amount of NBTI degradation. In this case, the DLL locking time is potentially too great (20μs) to detect the maximal amount of threshold shift in the delay element, prior to significant recovery of the degradation. The design is implemented in 130nm technology and consumes 545 μm x 255μm, roughly equivalent to 19,300 NAND3 gates.

In Chapter 4, an analysis of the impact of voltage, temperature, process variation and state dependence on oxide breakdown is used to define constraints on the implementation and use of on-chip sensors directly monitoring breakdown mechanisms[33]. Analysis on the quantity of such sensors needed to improve projection of an inherently random breakdown mechanism is addressed. Final results from this work are used to develop 3 novel and effective on-chip reliability monitoring sensors discussed in Chapters 5 and 6.

In Chapter 5, the design of a novel oxide degradation sensor is presented [34]. This sensor is based upon measuring the change in gate leakage current over time as an indicator of oxide degradation. As oxide devices under test are stressed using local supply voltage, the gate leakage current increases with progressive soft breakdown events, mirroring degradation in actual circuits under similar voltage and temperature stress. Based upon the results in Chapter 4, the design is optimized for area and the layout is standard cell compatible. The implementation of the oxide degradation sensor in 130nm technology uses 7.20μm x 20.86μm, roughly equivalent to 21 NAND3 gates.

In Chapter 6, a 2nd generation design of a combined oxide degradation and NBTI degradation sensor in 45nm technology is presented. The oxide sensor principle of operation is identical to the design in chapter 5, utilizing gate leakage current change to measure progressive soft breakdown events. The implementation and specifically power consumption is improved significantly over the design in 130nm technology. The NBTI sensor is integrated in the 45nm cell based upon a successful standalone sensor design in 130nm [34]. The NBTI monitor detects the change in oscillation frequency of a ring oscillator current starved by a stressed PMOS device biased in subthreshold. The area for the combined sensor implementation 45nm technology is 8.5μm x 9.1μm, roughly equivalent to 54.8 NAND3 gates.

The sensor designs from Chapters 5 and 6 represent enabling circuits for the system architecture proposed in Chapter 3 and the ElastIC architecture. Collectively, the work presented in chapters 2-6 represent a series of circuit and architectural innovations to enable adaptive, reliable system operation on an unreliable silicon platform. The material in chapters 2-5 has been published in top peer-reviewed conferences and journals and was met with positive feedback and commentary.

Chapter 2

Timing Error Correction for On-Chip Memories

Increasing demand for large, fast on-chip memories has placed growing importance on designing high-speed memories with minimal power consumption while delivering high parametric yields in an era of probabilistic device performance. Traditionally, the least intrusive and safest method for dealing with such uncertainties in circuit performance is to add some timing margin or voltage margin to the design, to ensure operation within this margined specification. Beyond the nominal circuit delay spec or operating voltage, designers must add margin for process variation, high temperatures, supply voltage droop and noise and device degradation and wearout. Recent work on adaptive designs [21, 22] suggests that designers are margining their supply voltage by 17% over the required minimum for typical circuit operation within specifications. From a power perspective, this translates to roughly 30-35% power overhead in margin. Leakage power in these large on-chip memories is significant and designers struggle to minimize leakage during both standby and active modes of operation.

An adaptive system that can reduce the supply voltage to eliminate these margins can provide significant power and reliability benefits. Leakage power decreases roughly cubically with reduction in supply voltage [35]; therefore dynamic supply voltage scaling systems provide a powerful mechanism to control dynamic

and leakage power with reasonable complexity and area overhead (i.e., no need for multiple voltage supplies within the memory array). Additionally, voltage is a dominant (typically exponential) stress factor in most reliability and wearout mechanisms. Any marginal reduction in long-term voltage stress will yield significant reductions in degradation due to mechanisms such as electromigration, oxide breakdown or bias temperature instability effects.

This work presents an approach to dynamic voltage scaling (DVS) for SRAM-based memories that allow aggressive scaling of supply voltage to reduce active power and gate and subthreshold leakage power with a simple, non-invasive sensing scheme, while simultaneously improving the operational lifetime of the circuits by operating at a reduced supply voltage. Conventional DVS techniques are limited to a conservative critical voltage that includes overly pessimistic margins for worst-case process variations and temperature fluctuations [19]. In addition to eliminating voltage margins and process and environmental fluctuations, the proposed circuits also provide a measure of protection from uncertainty due to SRAM access device bitline leakage currents. Exponential growth in leakage current variability is detailed in [35] and threatens to dramatically reduce the number of SRAM access devices per bitline in aggressively scaled process technology.

The proposed approach dynamically converges to a minimum operating voltage through an embedded timing error detection and correction circuit. The SRAM voltage is adjusted in real-time by monitoring the rate of timing errors detected, even allowing operation at sub-critical voltages (voltages below the first timing error incidences) for tradeoffs of error rate vs. supply voltage scaling. A differential voltage

is developed on the bitlines in the SRAM array and a standard latch-type sense amplifier is triggered speculatively by an enable signal generated from a clock edge. After a delay, a second sense amplifier re-samples the bitline to confirm the value, relying upon a larger voltage differential to provide greater confidence in the measurement. If a timing error (an error in the circuit due to insufficient time to evaluate) is detected, the correct data is available one cycle later from the conservatively-clocked sense amplifier. This technique is particularly advantageous considering technology scaling, since (i) intra-die and ambient variations lead to greater safety margins, (ii) interconnect leads to increased delay variability between SRAM banks, and (iii) data-dependent bitline leakage variability reduces certainty in effective read currents. These factors combine to result in overly-pessimistic worst-case design [36].

2.1 Timing Error Detection and Correction

A basic requirement to ensure proper operation in a sequential system is to guarantee the proper values are stored and propagated through the intermediate storage elements. Commonly used sequential storage elements rely upon sampling input values at clock edges (clock boundaries). In the proposed single-cycle SRAM, there are two clock boundaries that require delay-error detection and correction circuitry: 1) at the clocked storage element near the input/output (I/O) interface and 2) at the sense amplifier. Existing RAZOR flip-flop circuit structures from [19] can detect and correct timing errors in the signal at the I/O interface clock boundary. The RAZOR storage mechanism consists of an additional delayed-clock latch that re-samples the final data to detect transient timing and voltage errors and returns the correct result with a one cycle penalty in the case of an error.

At the sense amplifier output on the read bitpath, two standard differential latch-type (DLT) sense amplifiers [37] are used to “double-sample” the bitline during a read operation in the SRAM. In Figure 2.1, the rising edge of the EN1 signal is generated from the falling edge of the clock to trigger the original sense amplifier. The output of this original sense amplifier is immediately stored in an unclocked S-R latch to guarantee stability of the static output bus during the precharge phase of the SRAM cycle. The rising edge of EN2 is delayed from EN1, while the bitlines in the bank continue to develop more significant differential voltage. As supply voltage is scaled down in the SRAM, the effective read current is decreased and the word-line pulse arrives later due to eroding performance in the decode logic. These effects

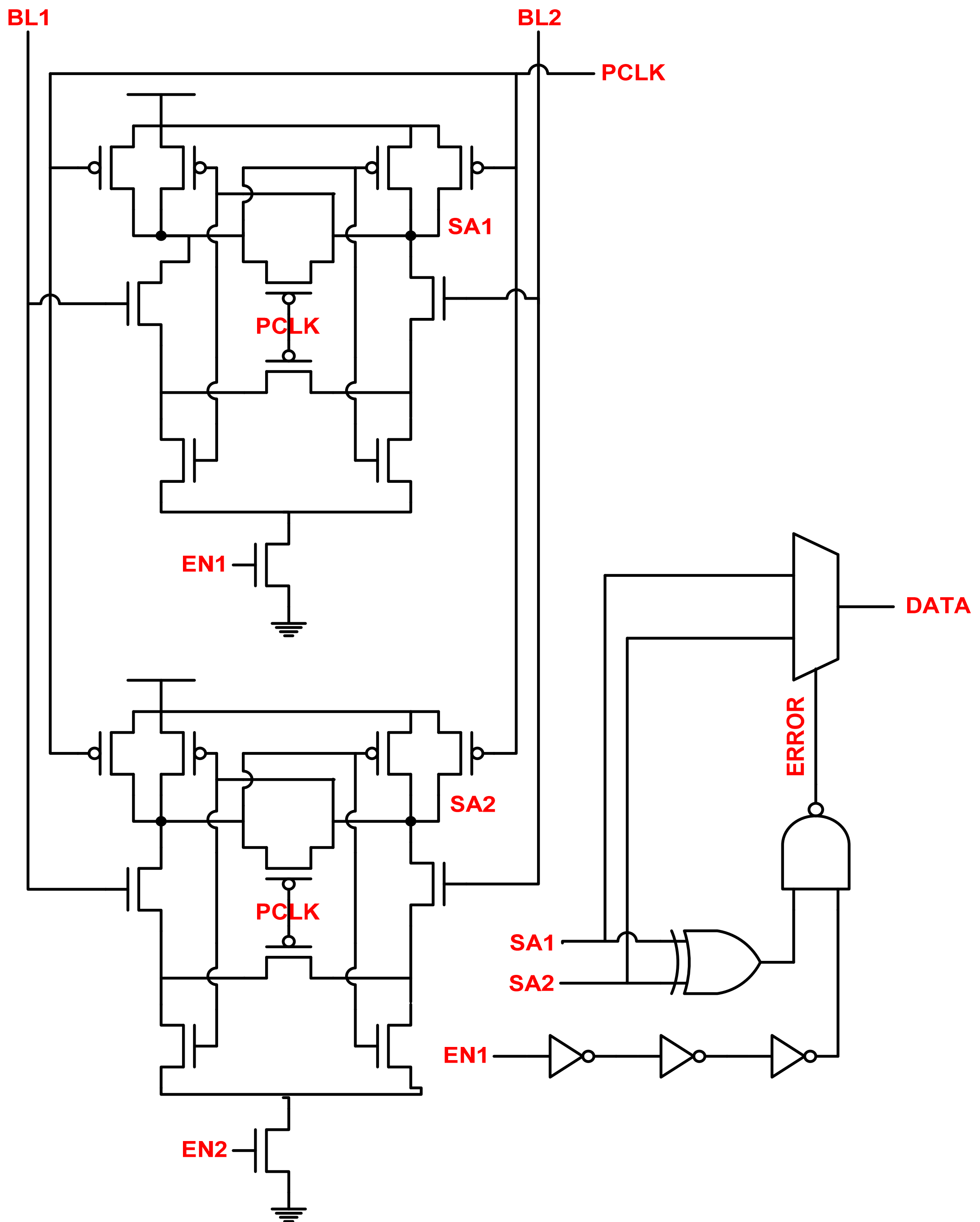


Figure 2.1 Main and Shadow Sense Amplifiers latched and compared via XOR gate and multiplexed to an output bus.

combine to result in a reduced differential voltage present on the bitlines given a fixed amount of time from the beginning of the read cycle. Re-sampling the bitline voltage with a delay from the original sense enable signal allows additional time for the bitlines to discharge and overcome process-variation induced offset voltages, data-dependent leakage currents, and activity-dependent internal voltages in the sense amplifier.

Enable signal pulses are generated from the falling edge of the clock using an inverter delay chain connected to a NAND gate as shown in Figure 2.2. A point on the delay chain is tapped to generate the falling edge of the precharge clock (PCLK) using a NAND gate with the unlatched BANK_EN signal. This ensures that the delayed pre-charge phase will begin as soon as the enable pulse (EN1) for the sense amp is de-asserted. EN1 is also used to clock the data bus mux enable to prevent XOR glitches from increasing latency. If the output of the error detection XOR is high when the enable signal is de-asserted, the NAND gate in Figure 2.1 will select the output of the second sense amp to be driven on the data bus. The rising edge of EN1 presets the multiplexor to select the output of the main sense amplifier to minimize delay impact.

When a speed path failure is detected at the sense amplifier, the correct value is muxed onto the static data bus. This result will not reach the I/O interface within the clock cycle, but the memory element at the I/O interface is capable of re-sampling the data bus and propagating the correct value via the latching mechanism proposed in [19]. If an error is detected at the latch column or at the sense amplifier block, the

SRAM can return a signal similar to a cache miss in a hierarchical memory system.

The corrected data is forwarded to the system at the end of the second cycle after

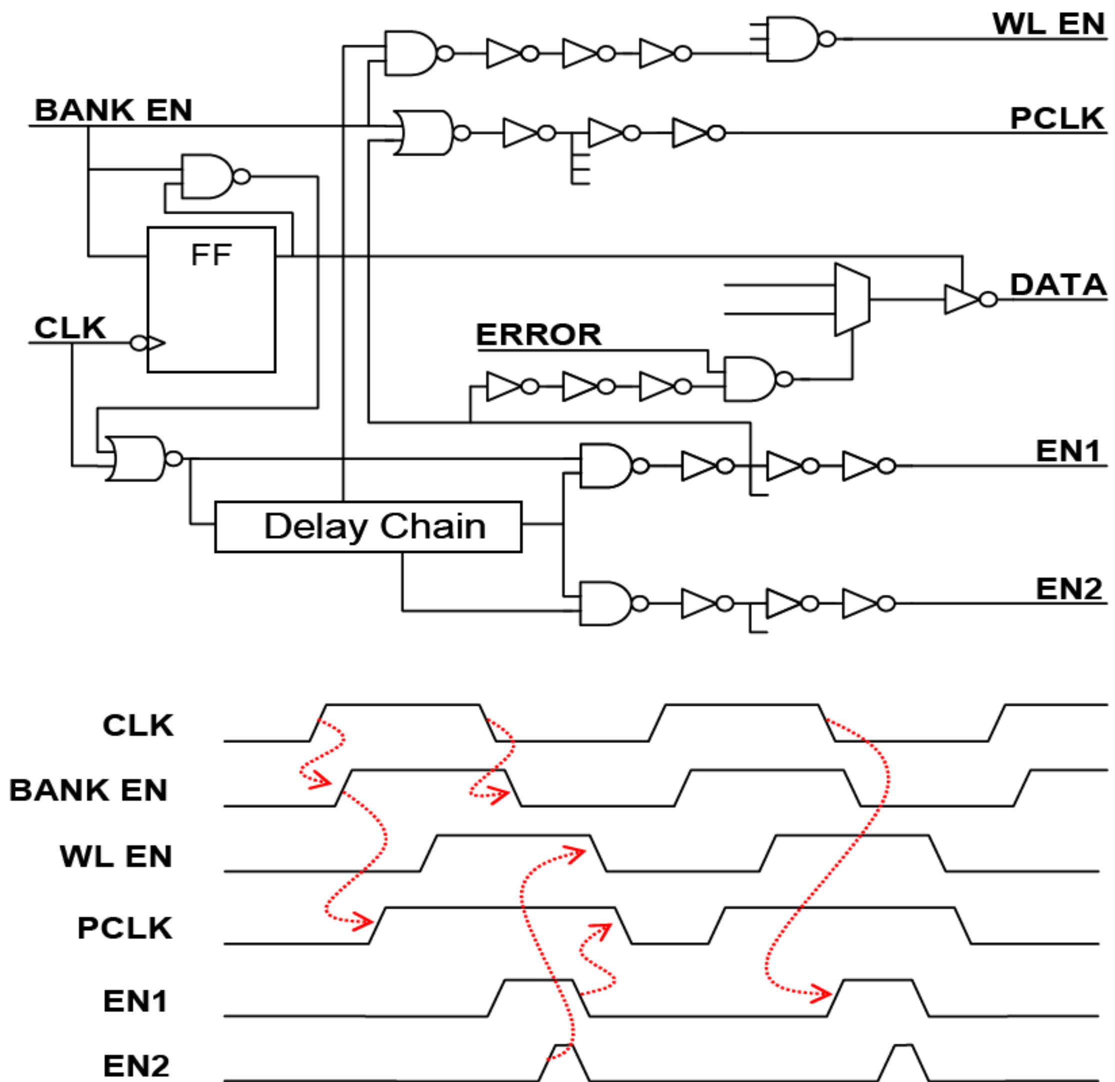


Figure 2.2 Timing signal generation circuits and waveforms. Two-phased timing generation features the precharge path derived from the rising edge of the system clock, and the read path derived from the falling edge of the clock.

the request rather than in one cycle. Many existing systems support hierarchical memory models and would require few changes to include an SRAM with timing speculation.

Delaying the pre-charge phase of the SRAM to accommodate the shadow sense amp requires a 2X increase in pre-charge device and driver sizes in a 64kB 180nm SRAM design. Total area overhead is projected to be less than 8% for an SRAM with similar block sizes and organization. The area overhead is localized near the sense amplifiers; therefore, the overhead is highly dependent upon the cells/sense amplifier. A greater cell to sense amplifier ratio reduces the fractional area overhead due to the proposed dual-sensing scheme. The overall structure of the proposed 32-bit SRAM is detailed below in Figure 2.3. The 64 kB SRAM is divided into 16 rectangular banks subdivided into four 1kB blocks. The I/O buses for the SRAM were routed to minimize wirelength in the routing channels between banks.

In order to explore address-dependent delay, a precharged dual-rail address bus is used to prevent glitching and false evaluate paths without requiring an additional clock boundary in the decode logic. Traditional designs have relied upon clocked decode networks or arrival pulse propagation alongside the address bus to prevent glitching and initiate the read/write sequence. Using the dual rail bus allows simple arrival pulse generation at each bank as the data arrives. Generating the pulse and propagating it along the address bus is another option, but simulations revealed less variability between enable pulse and data arrival when using local pulse generators with the dual-rail bus.

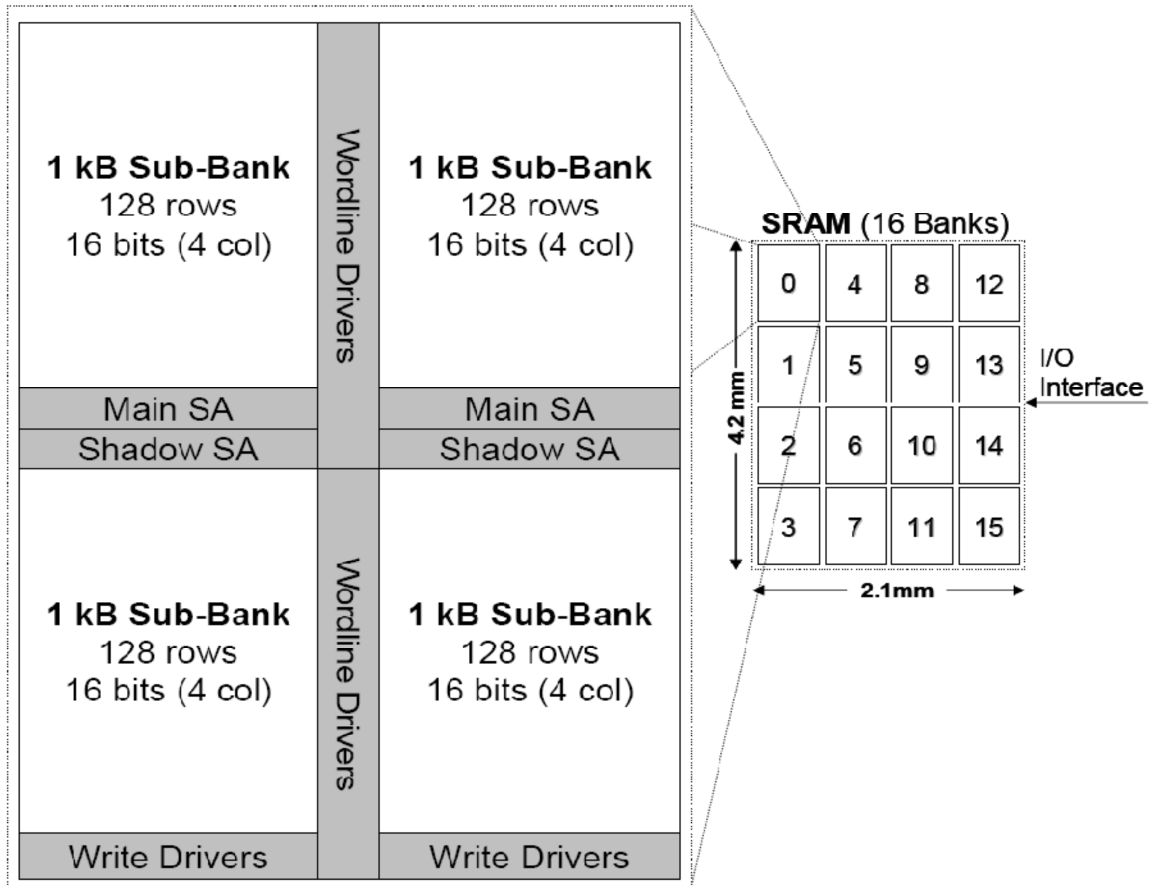


Figure 2.3 Proposed SRAM floorplan. Uneven buffering on input signals is used to allow banks closer to the I/O interface (9,10,13,14) to initiate read/write paths faster than further banks. Distant banks (0,1,2,3,4,7) have the most significant delay for read and write paths.

2.2 Exploiting Address/Data Dependence

Recent commercial SRAM designs consist of many banks spread over great distances on chip. Signal propagation and address decode delays typically dominate cycle times in large SRAM designs [38]. Typical design practice involves building a decode tree that equalizes the nominal delay from the boundary of the SRAM, to each bank. Although this hierarchy of decode provides roughly equivalent delays to each bank, banks in close proximity to the source of the access signals sacrifice performance in their decode network by being driven by the same signal nets with long wires and large drivers as banks that are millimeters away.

In an adaptive design that can detect and correct timing errors, limiting the performance of all portions of the address space to a pessimistic worst-case is neither necessary nor optimal. By using a combination of repeaters and long wire segments throughout the address bus, natural and substantial variation in the path delay between different banks in the SRAM can be exploited. This path delay variability causes portions of the address space in the SRAM to develop timing errors at widely varying supply voltages, allowing portions of the SRAM to continue functional operation at potentially lower voltages than other banks. Generating the arrival/enable pulses as the data arrives allows each bank to complete the read/write operation in the minimum cycle time rather than limiting quicker banks to a slower cycle. This benefits applications with high data access locality. When the SRAM is accessed, most accesses will target addresses from a few banks, allowing the SRAM to tune the supply voltage to the process and interconnect characteristics of the dominant banks in the active data set.

Beyond the benefits of exploiting address dependence in memory access, timing error correction allows a relaxation of the constraints placed by a pessimistic worst-case data state on a particular bitline. Bitline leakage depends upon the data stored in each cell connected to the bitlines. In addition to process-related variability, leakage currents dependent upon stored data influence the effective read currents of SRAM cells [39], particularly for long bitlines preferred for dense array designs. The bitline re-sampling technique allows designers to avoid margining for infrequent data storage patterns that lead to the worst-case read current. The shadow sense amplifier allows a speculative sensing phase with the detection/correction to maximize the cycle times of the SRAM in the presence of leakage-induced read current variability.

2.3 180nm Simulation Results

To evaluate the effectiveness and explore the trade-offs inherent to the shadow-sense amplifier, a 64 kB single-cycle SRAM was designed in 0.18um CMOS technology. Simulation results confirm that the SRAM operates at 250 MHz with worst-case device and interconnect models at 85ºC with a 10% margin on the 1.8V supply. Under typical process conditions at 25ºC the design approaches 400 MHz operation. The shadow sense amplifier is clocked with sufficient delay to reliably detect timing errors down to 1.3V in the worst-case corner at 250 MHz. Increasing the delay between sense amplifier enable signals decreases the minimum safe operating voltage, but increases the area penalty through a larger delay element and tighter constraints upon the pre-charge phase timing of the memory. Delay beyond 500ps between sense amplifier enable signals requires pre-charge logic differentiating read and write operations, allowing early precharge following a write.

Detailed variability simulations including parasitic models of the SRAM considering physical design were used to determine failure voltages for each SRAM bank. The variability model was adapted from industry-provided SPICE models and includes individual intra-die Gaussian distributions for W, L, and Vth and inter-die Gaussians for W, L, V_{th}, R_{dsw}, μ_0, and T_{ox}. Each bank of the SRAM was simulated with a fixed set of inter-die variation (-1σ was used) and intra-die variation for devices (-2σ to +2σ). Gaussian variations were generated for each relevant device, in each bank and combined to compose a “chip” for simulation purposes.

Figures 2.4 and 2.5 detail the timing error-rate vs. supply voltage with the SRAM used as a direct-mapped cache running memory traces obtained from 11 SPEC2000 benchmarks run on the SimpleScalar/Alpha v3.0 toolset [40]. The memory traces

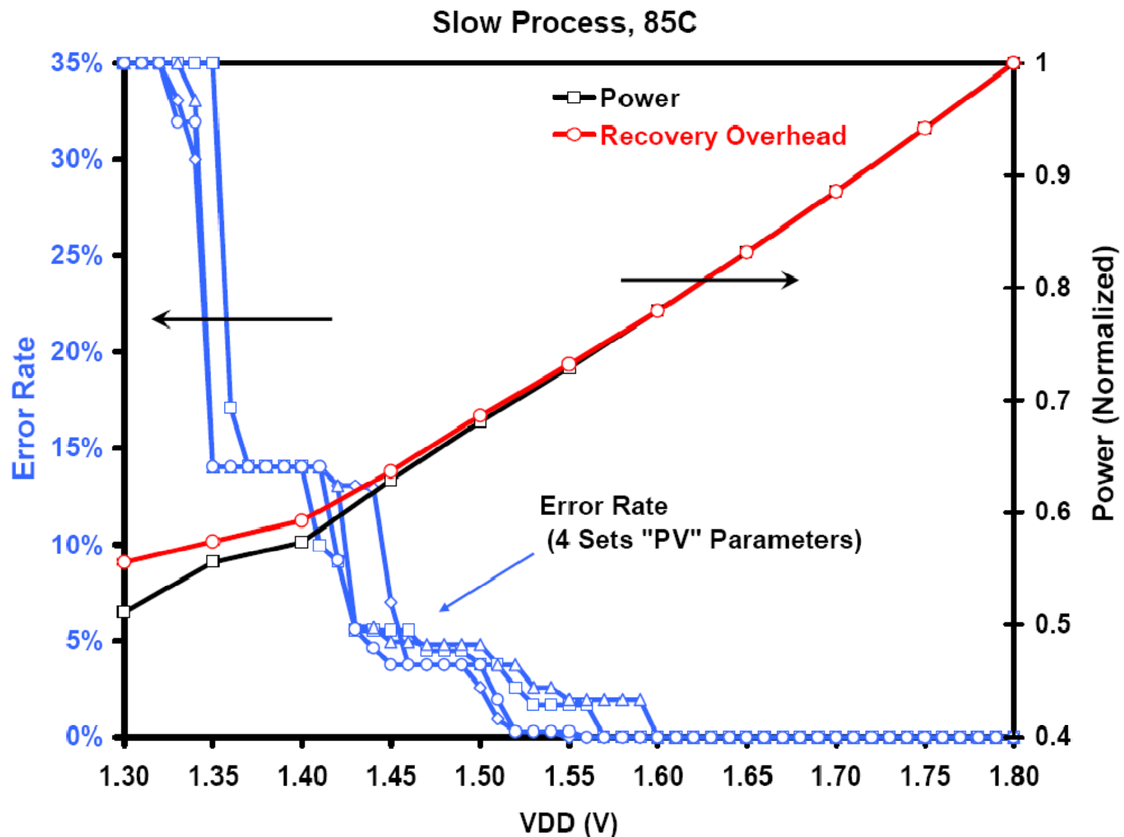


Figure 2.4 Error rate of memory as VDD reduces for gap00 memory trace. Four traces represent sets of intra-die variation with inter-die fixed at slow corner. Power both with and without recovery overhead is shown.

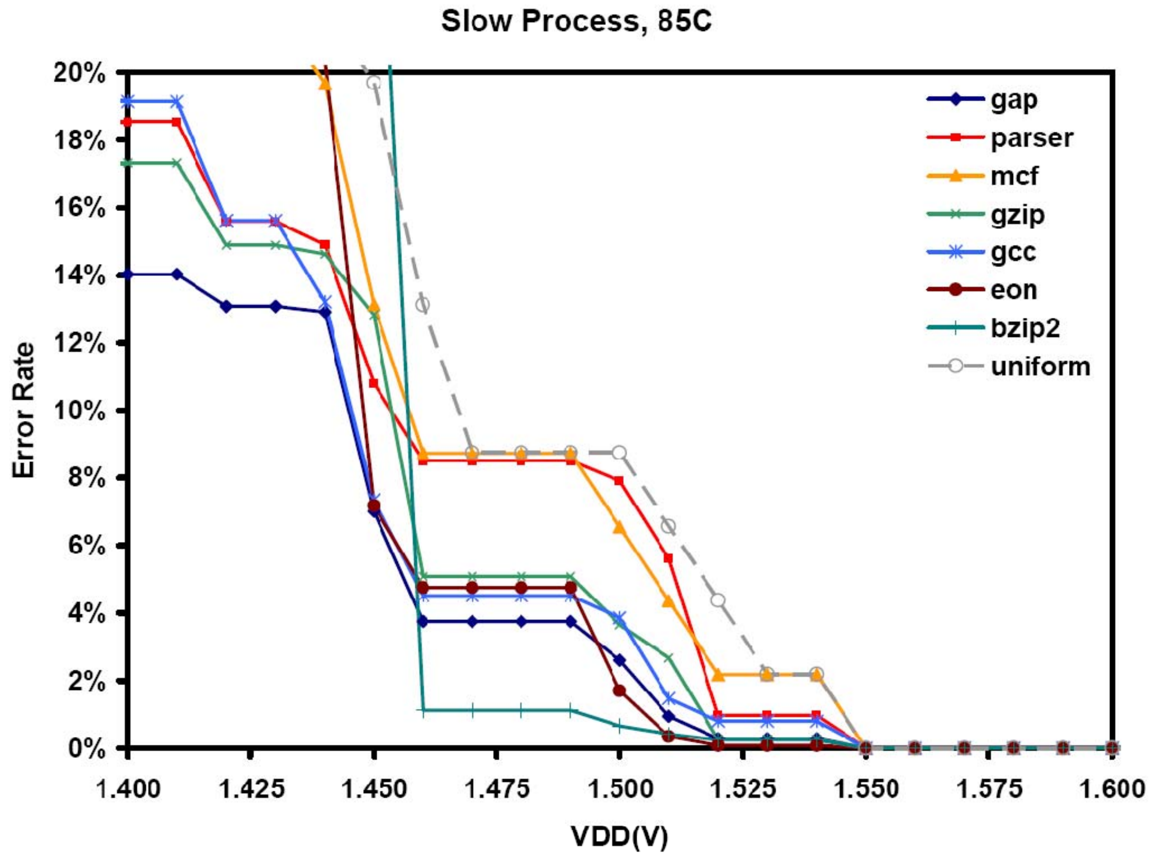


Figure 2.5 Error rates for a set of benchmarks vs. supply voltage. Fixed “slow corner” inter and intra-die variation. Error rate increases in discrete intervals since the variation in access time is dominated by the physical location of the bank accessed. In 180nm technology, very little within bank variation contributes to significant cycle time variation.

consisted of 10 million simulation cycles taken from an optimal point within the program execution to deliver typical memory access patterns using the Early SimPoint method [41]. Figure 2.4 displays the error rate on a trace of the gap00 benchmark for four different worst-case inter-die SRAMs defined by the Gaussian intra-die variability model described above. Figure 2.5 shows the impact of memory access patterns on error rate for a given chip. Applications with high data access locality (e.g., gzip) will exhibit dramatic fluctuations in error rate, while applications with lower data access locality (e.g., gap) will demonstrate frequent gradual increases in error rate as the voltage is lowered. Figure 2.6 is a trace of 100M cycles of SPEC benchmarks showing instantaneous voltage and error rates using a simple voltage control algorithm that updates every 10,000 cycles. The supply voltage ranges over 100mV during operation for a target error rate of 2%. Error rates surpass 2% in many instances due to the simple ± 10mV control algorithm.

In Table 2.1, the error detection/correction circuitry allows an SRAM design at the worst-case inter-die corner to operate at 1.55V at 85ºC with a zero error rate and 1.5V with 5% error-rate, as compared to the 1.8V supply for an SRAM without error detection and with a safety margin, saving 12% and 17% power, respectively, after considering the overhead of the additional circuitry and reducing the voltage stress on the entire circuit by 200-250 mV. Without an additional voltage safety margin, the SRAM operates at 1.62V in the worst-case corner. At operating temperatures less than 85ºC, the SRAM adjusted to 1.45V at 50C or 1.39V at 30ºC, saving 23% and 29% power, respectively, over a margined design operating at a fixed 1.8V. A typical part (i.e., not at the worst-case inter-die corner) can be operated at 1.3V

VDD, achieving up to 35% power savings over a conservatively margined SRAM and also significantly lowering reliability degradation due to NBTI/PBTI and oxide breakdown mechanisms. Static power, while not appreciable for this technology, is sensitive to supply voltage with both gate and subthreshold leakage benefiting from lower operating voltages and corresponding reduced ambient temperatures [35].

TABLE 2.1
Power Consumption at Various Operating Points

Operating Point	VDD	Total	Static
Single Sense Amp, WC, 85ºC	1.80V	76.5mW	16.30μW
Single Sense Amp, WC, 85ºC	1.62V	64.3mW	13.60μW
Dual Sense Amp, WC, 85ºC (Zero Errors)	1.55V	67.8mW	12.60μW
Dual Sense Amp, WC, 85ºC (5% Errors)	1.50V	63.8mW	11.90μW
Dual Sense Amp, WC, 50ºC (Zero Errors)	1.45V	58.9mW	5.56μW
Dual Sense Amp, WC, 30ºC (Zero Errors)	1.39V	54.0mW	3.92μW
Dual Sense Amp, TYP, 85ºC (Zero Errors)	1.30V	49.4mW	16.00μW

All power results for dual sense amplifiers include overhead for the required signal generation, circuitry and increased pre-charge device sizes.

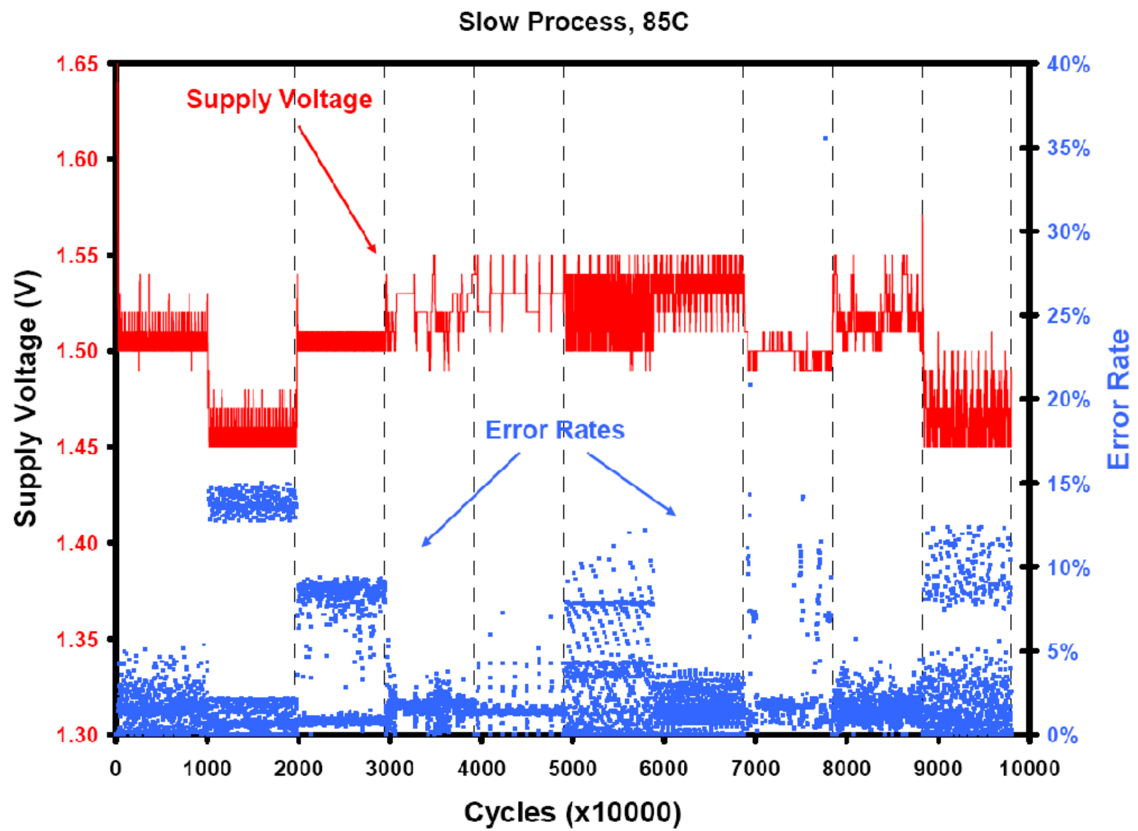
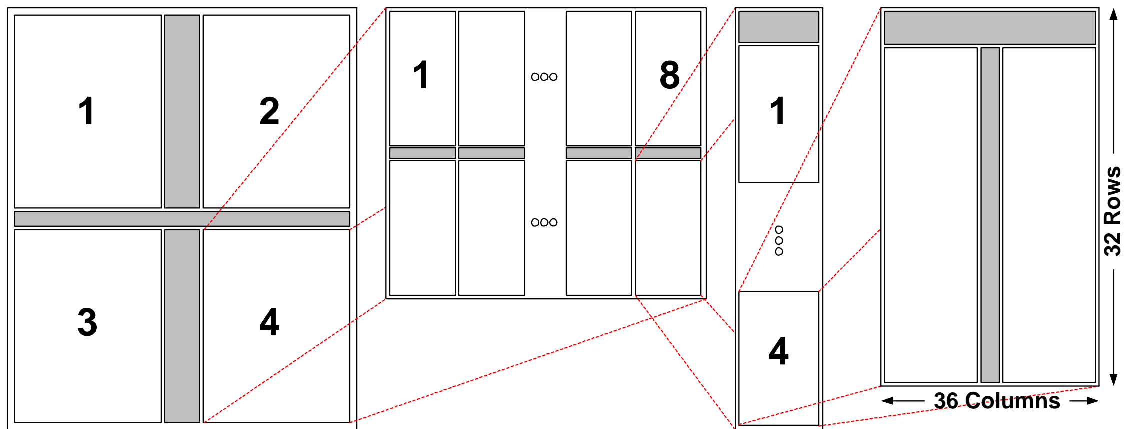


Figure 2.6 Instantaneous supply voltage and error rate during 10M cycle DVS simulation with varying workload. Uses error counter with ± 10mV V_{DD} update every 10000 cycles over 10 benchmarks.

2.4 130nm Physical Implementation

A 130nm testchip from an industrial foundry has been designed and fabricated, including a 36kB SRAM containing a prototype implementation of the proposed time redundancy technique. The implementation of the test chip is based upon a fully functional, industry-designed 32kB SRAM verified to operate across voltages from 0.9V to 1.5V. The original SRAM, detailed in [42], was designed using 4 levels of metal and it consumed around 1.7 mm^2 of silicon area. The 36kB test chip containing an implementation of the time redundancy circuitry is around 2.25 mm^2, a 32% increase in array area; although the test array contains 12.5% more 6T SRAM cells in the increase from 32kB to 36kB. The specific components driving the additional increases in area are discussed below during description of the hierarchy of the SRAM and selected implementation-specific design issues.

The smallest segment of the initial design, the local block, contains 32 rows of SRAM cells 16 bits wide, including adjacent timing generation circuitry and pitch-matched write drivers and sense amplifiers devoid of any column multiplexing. The relatively short bit-line (32b) and word-line (16b) limits the skew between the word-line pulse and the sense amplifier enable signal and greatly reduces bit-line delay variability due to access transistor leakage currents. Reducing the sub-array size and eliminating column multiplexing transistors, discussed in [42], are alternative techniques for reducing the delay variability present in the SRAM design and presents a similar drawback to the proposed technique, leading to area overhead and reduced array efficiency. Despite the already small sub-array size and lack of



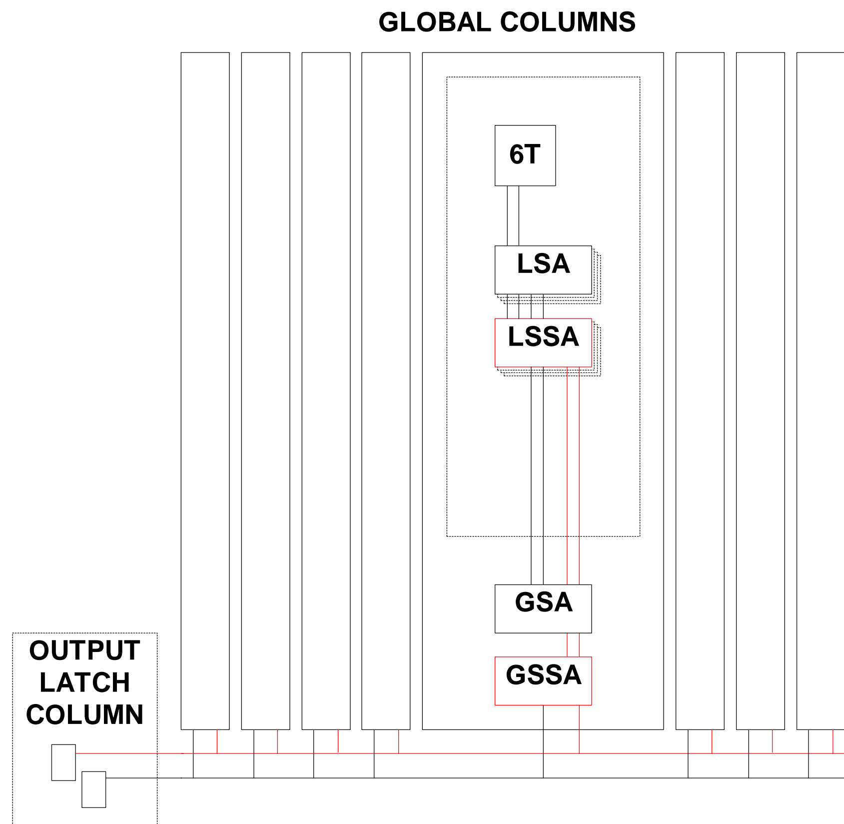
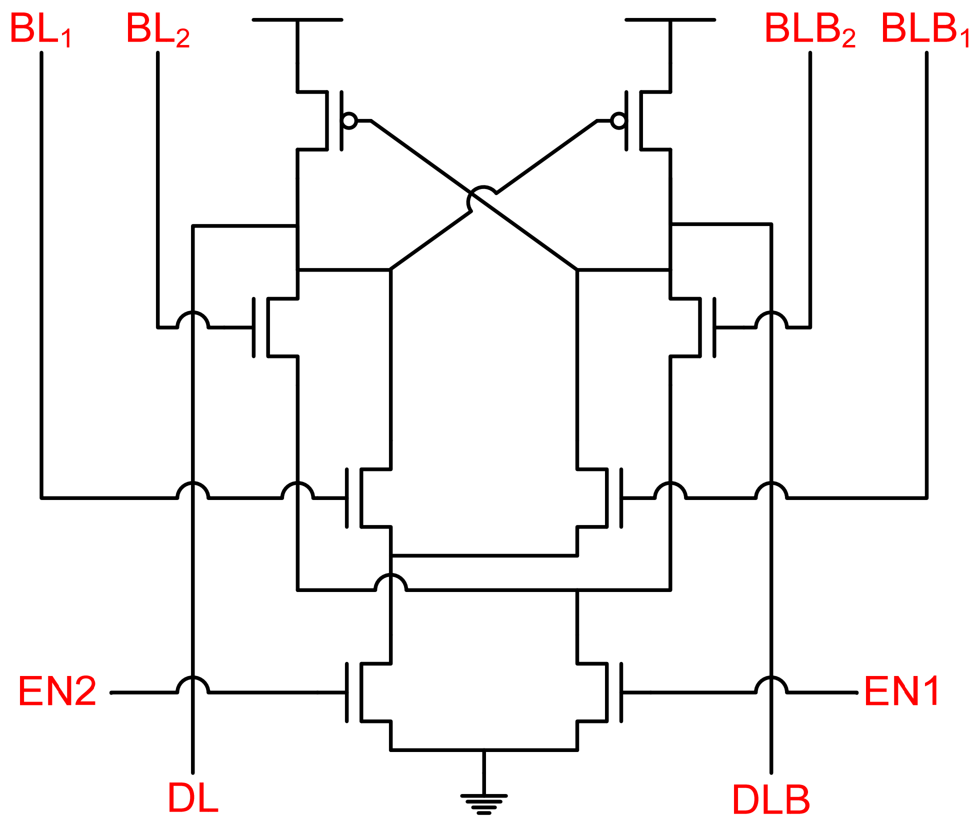


Figure 2.7 Array Architecture and Output Path. Each of 4 macros provides 18b of data. A macro is divided into 16 global columns and each global column into 4 sub-arrays. Sub-arrays contain 32 rows and 36 columns. Local Shadow SA connects to Global Shadow SA to separate output bus to shadow latch.

column multiplexing, the time redundancy circuits were added to the design in an effort to evaluate any potential improvements in delay variability related errors.

The original design includes 4 local blocks containing local sense amplifiers each connected to a global sense amplifier at the bottom of each global column. For the purposes of testing the output of both sense amplifiers before error detection via the scan chain, the output of the local shadow sense amplifier is connected to the input of a global shadow sense amplifier which drives a separate output bus connected to the shadow latch portion of the output latch bank. Figure 2.7 illustrates the configuration of the local and global sense amplifiers and the output latch bank. Since the design is limited to 4-levels of metal, this led to the first challenge associated with integrating the shadow sense amplifiers and routing their output through the congested metal tracks through the global column. No additional wiring tracks were available to support the shadow sense amplifiers, since M2 tracks contained local bit-lines and M4 contained global bit-lines and no column multiplexing was implemented.

A novel method of 2:1 column multiplexing is used to create an additional M4 wiring track for the outputs of the shadow sense amplifier, while avoiding adding pass transistors traditionally used in column multiplexing schemes. As Figure 2.8 illustrates, the cross-coupled PMOS devices in the sense amplifier are shared between separate matched pairs and footers connected to adjacent bit-line pairs. Each matched pair is activated by separate enable signals routed horizontally in M1, guaranteed that only one of the two enable signals can be activated in a given cycle. The shadow sense amplifiers are column muxed in the same manner and now can



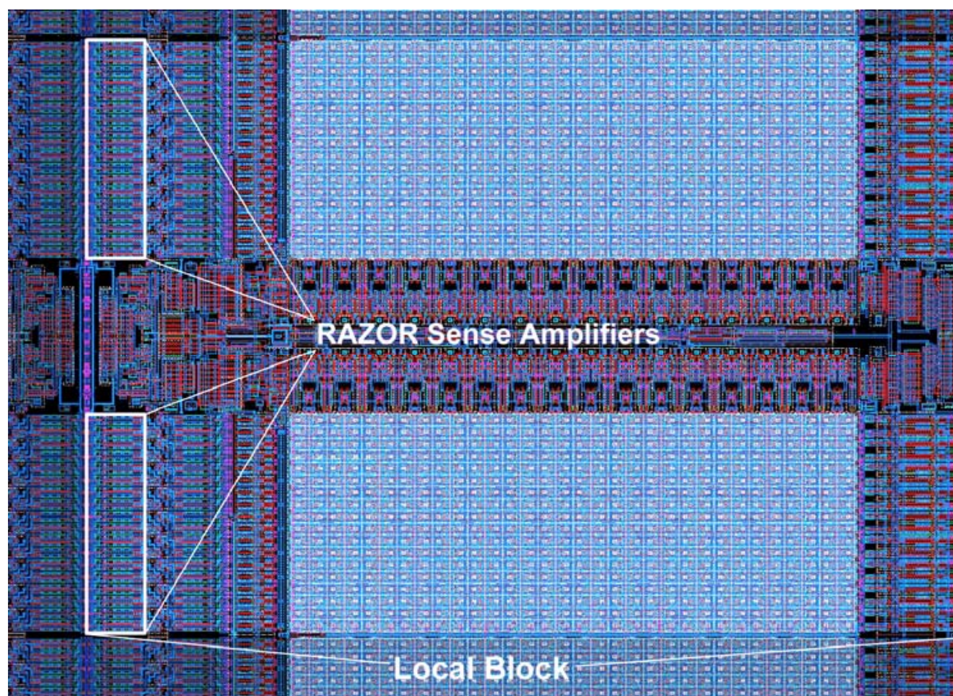


Figure 2.8 Sense amplifier and integrated 2:1 column multiplexer and layout of new local block showing area overhead of "RAZOR" or shadow sense amplifiers.

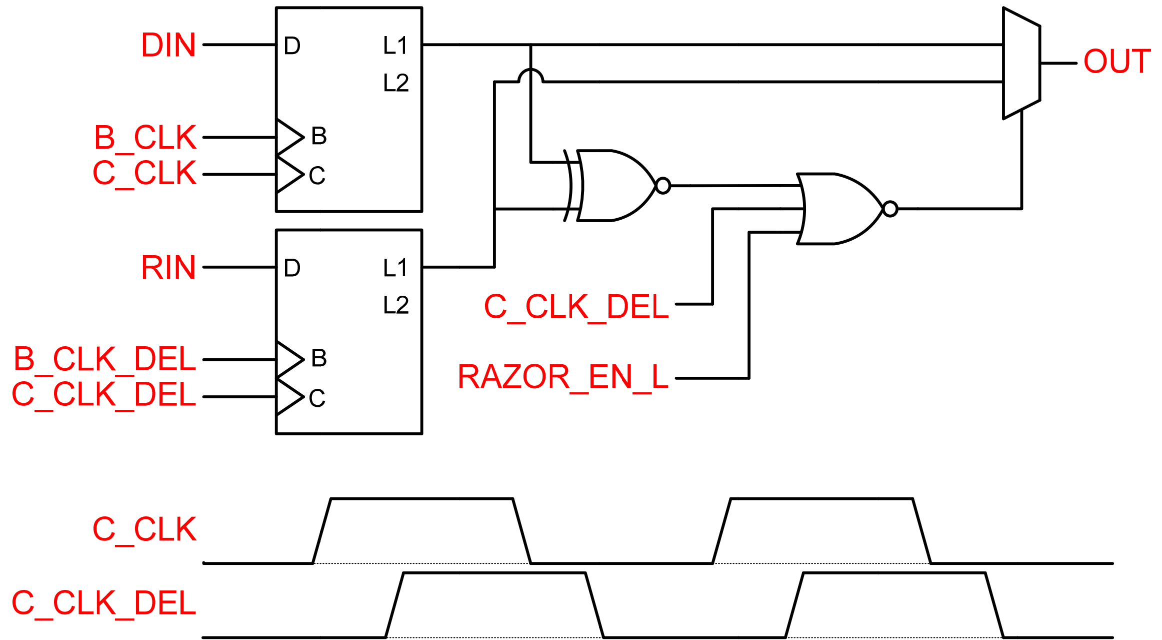
connect to a global bit-line in the M4 level adjacent to the output of the conventional sense amplifier pair. A similar design modification is necessary to allow the write drivers to multiplex two bit-line pairs. With this scheme, each local block would produce 9 bits of output rather than the 18 bits the SRAM required. Rather than adding to the width of the word-line to increase the output of each local block, the local block was copied, including word-line drivers and timing generation logic, and mirrored to produce a new larger local block seen in Figure 2.8. This technique allowed each new local block to generate 18 bits of output without requiring a re-design of the timing generation and word-line driver circuits that would invalidate the previous measured functionality of these circuits.

Combining two local blocks to form a new local block and incorporating the 2:1 column multiplexing requires some alterations to the address decode network in the SRAM. Formerly, 16 global columns created a half-quad in the SRAM and the four "G" address bits selected the global column to access. Reducing the 16 global columns to 8 global columns allows G<0:2> to be utilized to decode the global column enable signal, freeing G<3> to be used to control the column multiplexing scheme. Previously, 3-input NAND structures were used to generate static, partial global column enable signals from G<2:3> and B<0> address bits. The same structures are used with the new decode network by holding the input to the G<3> at "1" and generating the global column enable signal from only G<2> and B<0>. This simple alteration imposes minimal timing changes on this signal and the selection of signal G<3> to be removed from the global column select path ensures there is no

impact on the more critical pulsed G<0:1> path that controls both edges of the word-line and sense amplifier enable pulses.

Routing an additional output bus for the signals from the shadow sense amplifier path required increasing the width of the channel between half-quad blocks by around 10um. Gaps in the data input driver path can be seen where bit width was decreased from 18b to 9b per half global column as seen on the left and right edges of Figure 2.9, directly adjacent to the global block output drivers. The standard output bus is connected to the DIN input of the bank of output latches and the shadow output bus is connected to the RIN input of the output latches, corresponding to the shadow latches. Figure 2.9 illustrates the configuration of the output latch enabling input of two signals to produce one corrected output.

In Figure 2.9, the DIN signal is provided from the conventional sensing bit-path and RIN is the output signal provided at a later time from the shadow sense amplifier path. B_CLK and C_CLK are used to latch the signal from DIN, and the delayed version of each clock, B_CLK_DEL and C_CLK_DEL are used to latch the later-arriving RIN signal. When C_CLK is high the latch is transparent and data flows through and the rising edge of C_CLK_DEL clock presets the output of the multiplexer to the output of the conventional bit-path data latch. If the DIN and RIN signals differ after C_CLK_DEL falls, the multiplexer is switched to transmit the data from the late arriving signal. This mechanism allows the later arriving signal to be forwarded to the next stage in the pipeline following the SRAM. In this implementation of the SRAM, the correction circuitry in Figure 2.9 is designed to allow use of the standard library latches and gates without manual design efforts.



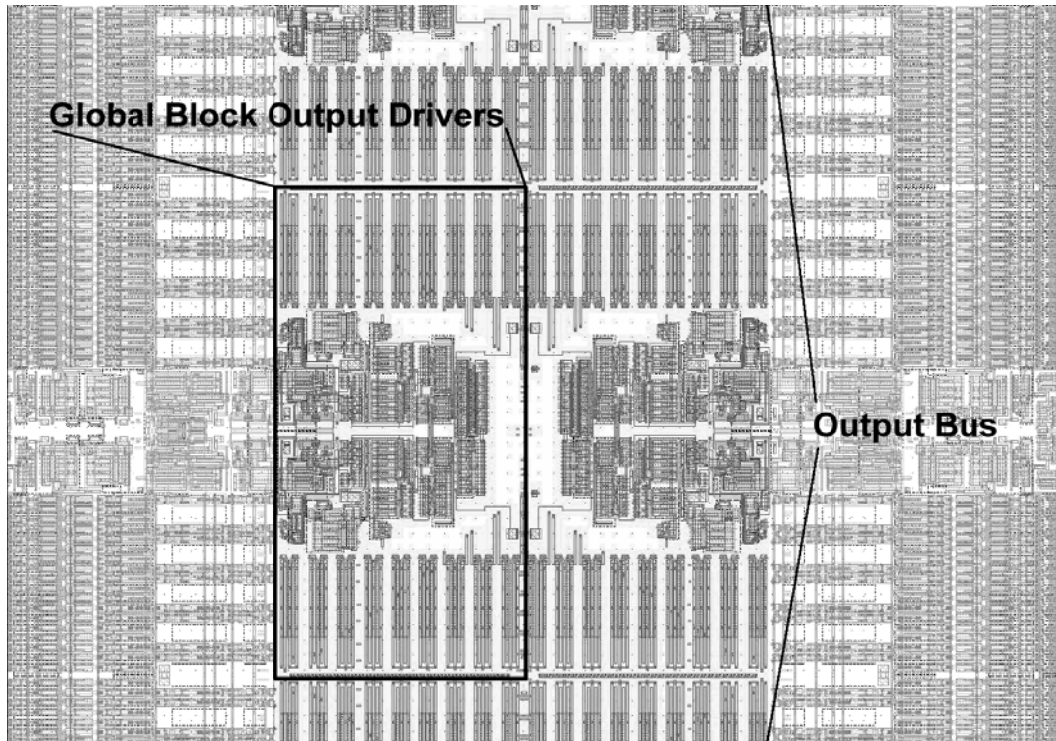


Figure 2.9 Output latch with added shadow latch and error correction and detection circuitry. Layout of the routing channel containing drivers for the additional output bus from the shadow sense path.

The multiplexer in the diagram, which is the only significant source of delay overhead, can be relocated to an internal feedback path in the conventional latch itself, to reduce the delay impact to only capacitive loading effects on internal latch nodes. In [19], a latch designed in this manner is presented and shown to cause a minimal impact on the D-Q delay path in the latch.

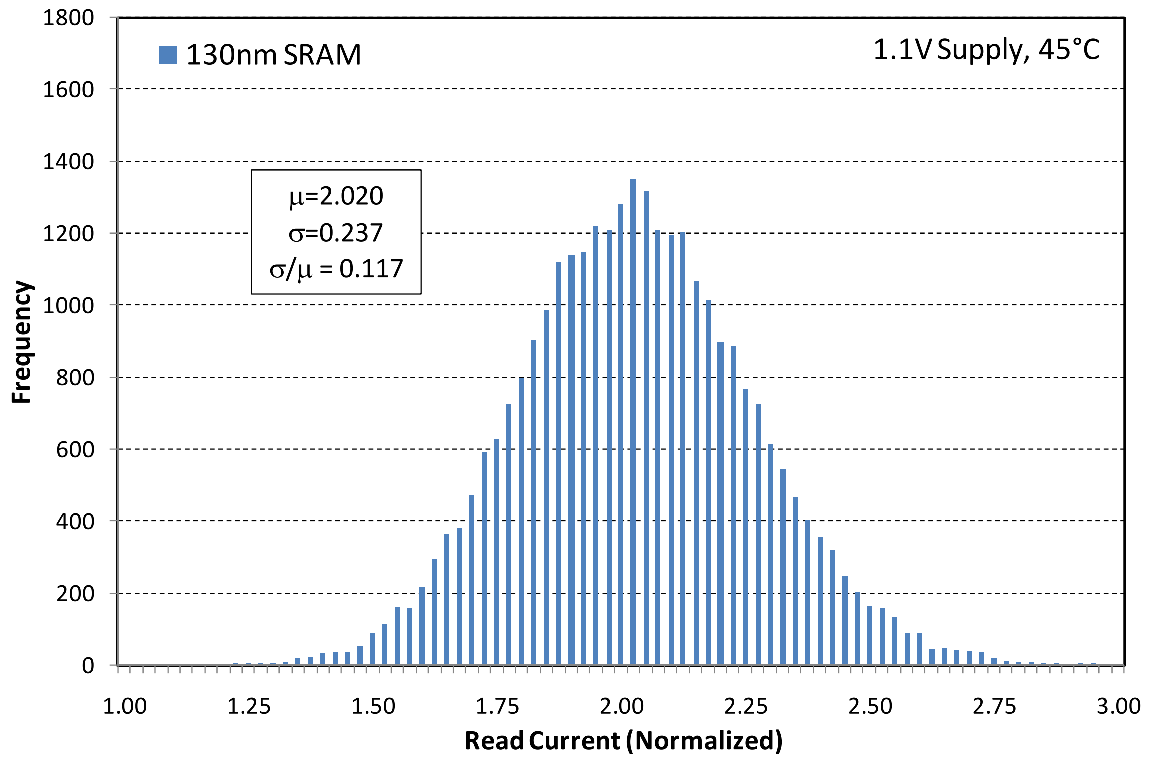
In this implementation of the proposed technique, altering the timing generation of the word-line and conventional sense amplifier enable signal is avoided to maintain the integrity of the reference SRAM design. The shadow sense amplifier enable signal is delayed from the initial sense amp enable signal with 5 inverters and a nand gate, which effectively creates a 90ps delay between enable signals at 1.5V supply and an 850ps delay between enable signals at 0.6V supply. This delay was set partially by area and time constraints in the layout and due to the realization that the goal of this test chip was to demonstrate the potential to capture delay errors using the technique rather than designing the entire system to exploit the traits of the time redundancy technique. Each successive enable signal in the output path of the shadow sense amplifier is delayed using an equal length and size delay chain to develop the delay required to accommodate the late-arriving shadow sense output signal.

Additional structures are added to count the number of errors detected at the output latch bank and to compress the output patterns from multiple cycles into a readily comparable signature (MISR).

2.5 130nm Simulation Results

This section presents some basic variability analysis of the SRAM cell in 130nm and 45nm process technologies, followed by a brief summary of findings from some simulations covering a fraction of the variability space in the local bank of the designed test SRAM.

Figure 2.10 shows the wide range of SRAM read currents possible at 1.1V supply, ranging from normalized values of 1.00 to 3.00, yielding variations of greater than 43% from the average read current value over a Monte Carlo sample of 30,000 cells. The complete 36kB SRAM contains nearly 300,000 SRAM cells making it likely to encounter read current variations greater than +/- 3σ from the mean. The Monte Carlo simulation was conducted using HSPICE BSIM4 and PSP models including variability modeling for 130nm and 45nm process technologies. SRAM-specific device models are used in the SPICE simulation to provide the best possible simulation-based prediction of actual read current variation. Designing the read bit-path to accommodate the outlying read currents from this distribution leads to a sacrifice in read delay that the proposed technique aims to improve by allowing more aggressive timing of the sense amplifier enable signal. When scaling supply voltage from 1.5V to 0.6V there is a slight upward trend in the ratio of standard deviation to mean read current, representing an increasing opportunity to achieve performance or reliability lifetime improvement via the time redundancy technique at lower supply voltages.



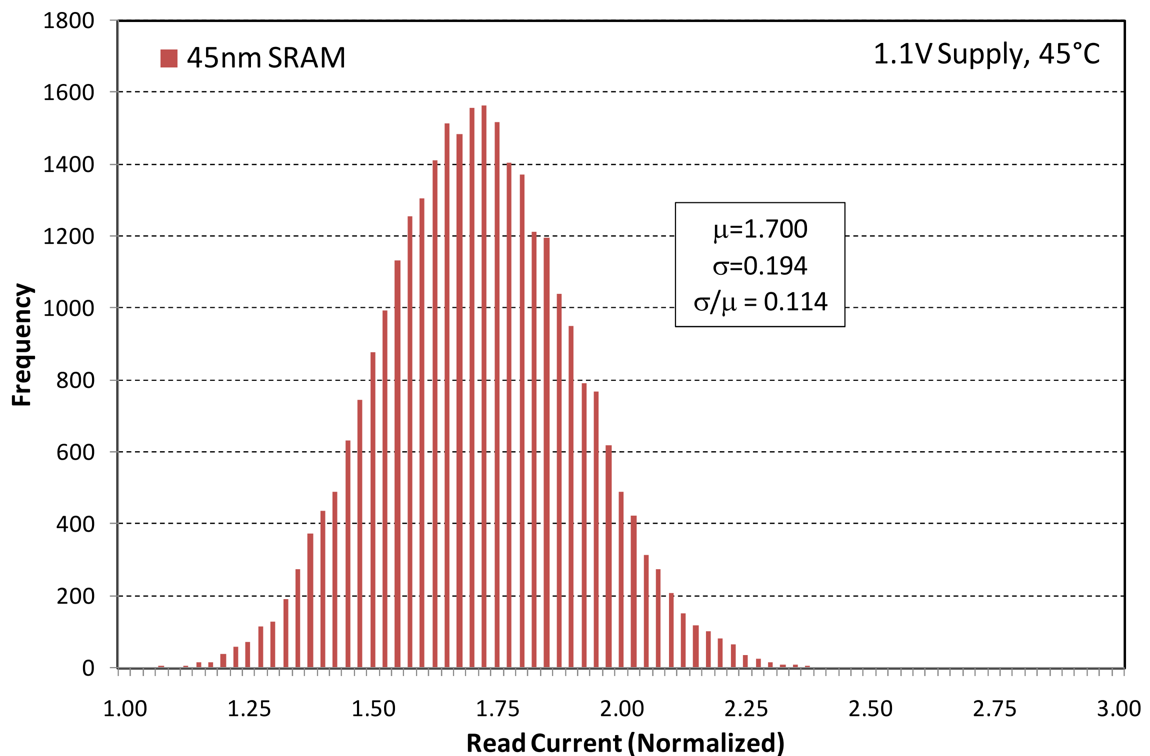


Figure 2.10 Distribution of Read Current for 130nm and 45nm technologies.

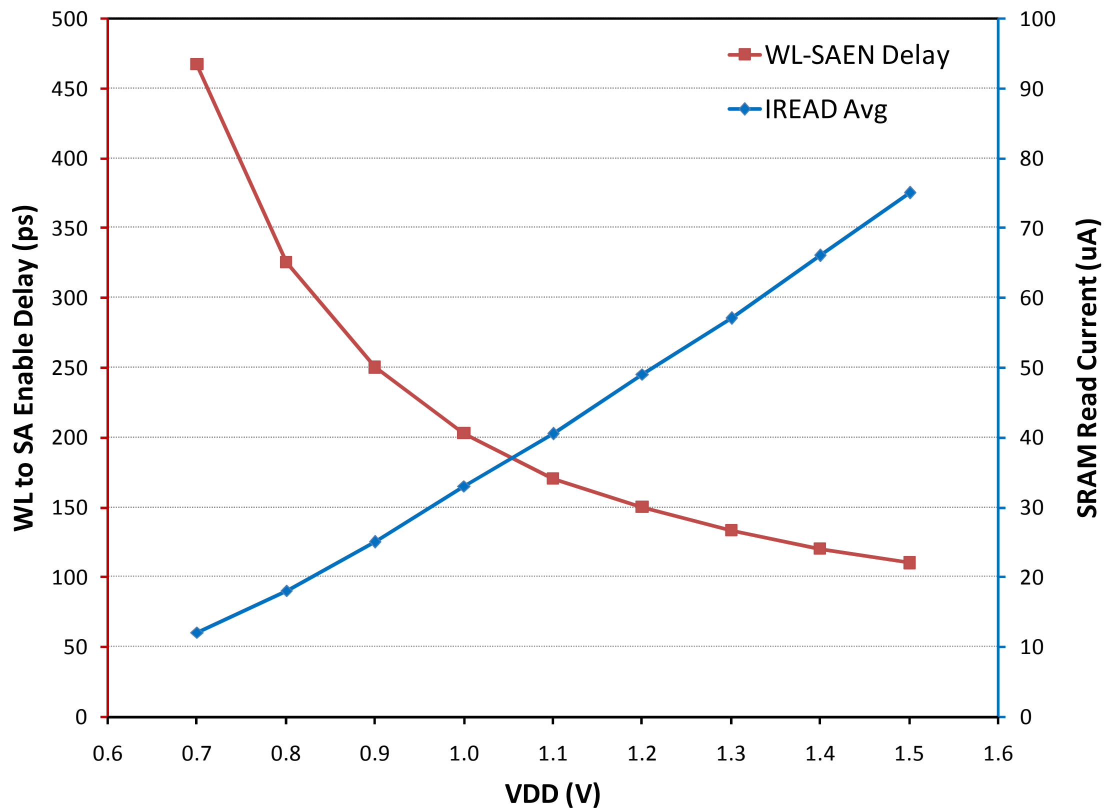


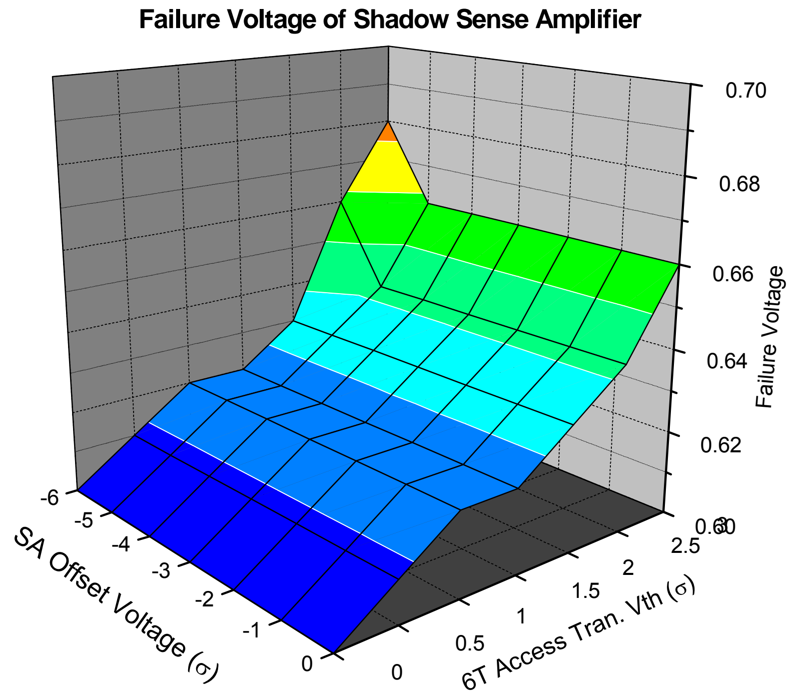
Figure 2.11 SRAM Read Path Timing and Read Current scaling with VDD for 130nm technology. As supply voltage is decreased, the word-line to sense amp enable delay increases rapidly.

Figure 2.11 depicts the scaling of read current and the delay between word-line assertion and the sense amp enable pulse as power supply is decreased. The average read current decreases from 75.2 μA at 1.5V to 11.9 μA at 0.6V – a decrease of ~6.3X. The sense amp enable delay increases from 109 ps to 468 ps – an increase of nearly ~4.3X. This slight discrepancy in scaling reveals that in order to provide robust operation at low voltages near 0.7V, the design includes unnecessary timing margin at the higher range of the voltage scale. Unfortunately, this situation indicates that reducing supply voltage will be a less effective way of stimulating delay errors during testing of the SRAM including the time redundancy circuits. Scaling the cell supply voltage separately from the peripheral supply voltage was explored as a technique to degrade the read current of the SRAM

intentionally in an effort to ease the measurement and evaluation of the proposed technique. As the difference between cell supply voltage and peripheral supply voltage increases, the internal “0” storage node voltage during a read increases proportionally and results in read stability errors in simulation before effecting sufficient reductions in read current to allow more effective delay error detection with the proposed circuits.

Transient simulations of the fabricated design of the test SRAM were conducted exploring the variability space including threshold voltage variation in access transistors, sense amplifier offset voltage variation and bit-line capacitance variation. Combinations of sense amplifier offset variation and access transistor threshold voltage variation showed the only positive results in which the shadow sense amplifier is noticeably more robust and scales to lower voltages than the conventional sense amplifier. Only with $+2.5\sigma$ to $+3\sigma$ increase in access transistor threshold mismatch and very unlikely values of -6σ sense amp offset variation did the shadow sense amplifier catch any delay errors. The failure voltages of each sense amplifier are plotted below in Figure 2.12, against sense amp offset voltage variation and access transistor variation.

Figure 2.12 summarizes the outlook for the proposed technique in the test SRAM. Comparing the minimum operating voltage of each memory, the failure voltages are very similar across the range of variation in access transistor and sense amplifier offset. Only in very extreme cases of variation in access transistor threshold mismatch and sense amplifier offset can the shadow sense amplifier provide any voltage margin prior to complete failure. Finding such a combination of



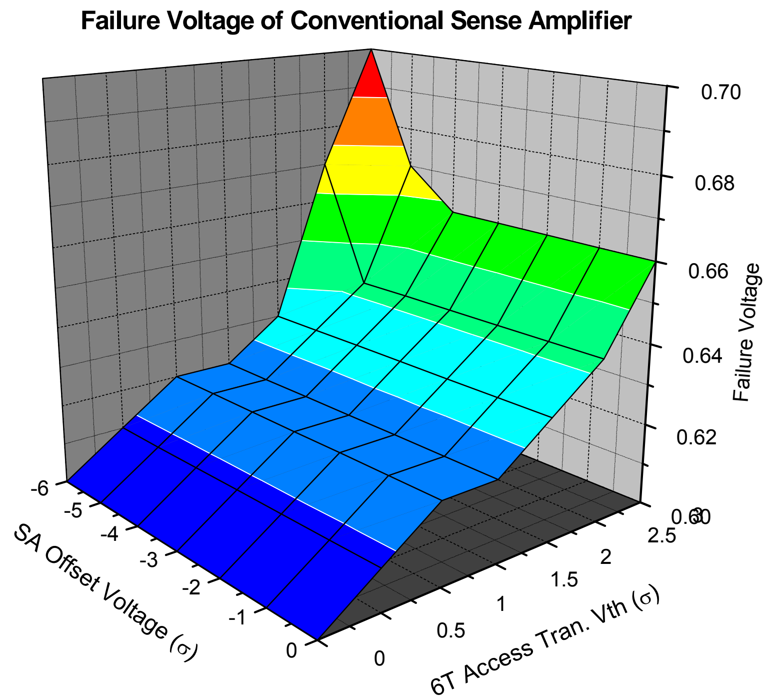


Figure 2.12 Minimum operating voltages for various combinations of threshold voltage variation for the implemented design of the time redundancy test chip.

pessimistic variations in hardware, if possible, is extremely difficult with limited testing time. Many failures in Figure 2.12 are the result of the inability to write to the SRAM cell with the high levels of variation in 6T access transistor threshold voltage.

The design of the local sense amplifier in the original SRAM contributes to the inability of the proposed redundancy technique to provide additional reliability. The sense amplifier is weakly cross-coupled by only PMOS devices and the footer device is roughly $1/6^{th}$ of the width of the devices in the matched NMOS pair. These factors cause the sensing time of the amplifier to be relatively slow and allow a great noise margin, allowing incremental discharge of the bit-line during the enable signal pulse to contribute to the final amplification. The weak cross-coupling in the sense amplifier prevents the amplifier from accelerating the sensing of the voltage differential when read current is greatly reduced, since the PMOS devices are nearly never out of subthreshold conduction in these situations. Previous work on this technique studied the behavior of a strongly cross-coupled sense amplifier with a larger footer device, which leads to a sensing operation that amplified the voltage differential on the bit-lines detected at the beginning of the enable pulse, rather than the slower sensing scheme used in this SRAM.

Chapter 3

Reliability Modeling and Management

Traditional stress-based reliability qualification techniques, such as the JEDEC JESD-47 Standard [12], qualify designs by stressing sample systems under pessimistic environmental conditions with a zero failure pass/fail criteria. While the traditional approach is an accepted method of ensuring reliability, the limits it places on supply voltage and temperature leave a significant and increasing reliability margin between circuit performance at worst-case conditions and at typical conditions. Widely varying environmental conditions linked to portable products combine with dynamic power reduction techniques to exacerbate the limitation of this conventional worst-case qualification methodology.

Hence, the need for alternative approaches to ensuring lifetime reliability under dynamic operating conditions is clear. Knowledge-based risk assessment is one alternative to simplistic corner-case stress testing that increases the complexity of qualification considerably. The knowledge based approach (a framework for knowledge-based qualification is defined by JEDEC JESD-34 [15]) requires careful characterization and analysis of individual failure modes to assess a reasonable system reliability risk factor given the reliability targets for the system. In this paper, we propose the use of so-called dynamic reliability management (DRM), where real-

time workloads and thermal information provides accurate inputs to real-time knowledge-based reliability models for projecting the degradation caused by various failure mechanisms. We then use the projected failure probability to control the maximum voltage assigned by a dynamic voltage scaling (DVS) algorithm.

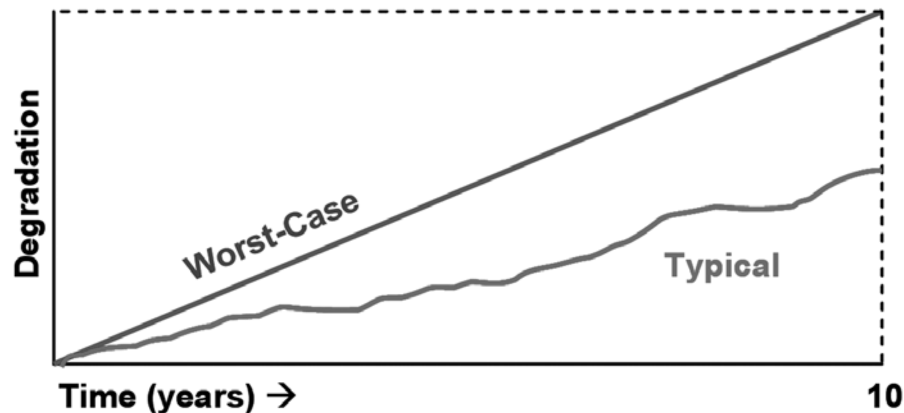


Figure 3.1 Reliability Degradation over time. Typical usage profile results in less degradation than a worst-case usage profile mandated by traditional qualification methods published by JEDEC.

The concept of DRM is conceptually motivated in Figure 3.1. The line labeled worst-case profile shows the accumulated damage due to a failure mechanism, such as oxide breakdown, over a 10 year time span under worst-case operating conditions (maximum operating frequency, voltage and ambient temperature). In traditional analysis, the maximum supply voltage is set such that at the accumulated damage at the 10 year mark results in a failure probability that meets the specified constraint (such as 63.2%, or $1\text{-}e^{-1}$). However, performance traces collected from an actual desktop processor show that the processor spends over 85% of its time in low power or sleep modes where the incurred damage rate is significantly lower. The accumulated damage from such typical usage is shown in Figure 3.1 with the line labeled typical profile and results in a much lower final damage at the 10 year mark.

Hence, the failure probability for this typical usage is well below that of the specification and the maximum allowed operating voltage was unnecessarily constrained resulting in a loss of potential performance.

Since the maximum supply voltage is currently set at design time or during post-fabrication testing, worst-case conditions must be assumed. It is becoming increasingly difficult to anticipate the actual usage of a part in order to construct reasonable worst-case corners for qualification. Under DRM, however, it is possible to dynamically monitor the operating voltage and temperature during part operation. With this operating condition history we propose the use of failure mechanism models to project the expected reliability and dynamically adjust the maximum supply voltage available to a DVS algorithm such that the required reliability constraint is met while the delivered processing performance during peak demand is maximized. DRM has the added potential benefit of theoretically allowing a user to select a desired lifetime or degradation envelope, post-fabrication. Furthermore, DRM provides designers and architects with the ability to control lifetime reliability independently for individual projects in a manner that is transparent to the manufacturing technology.

The concept of DRM was first introduced in [26] using a sum of failure rates method while considering multiple reliability mechanisms. Lu and co-authors [27] analyzed electromigration effects and suggested dynamic thermal management (DTM). However, both approaches focus on short time scales that are not indicative of realistic reliability requirements, and more critically do not propose an actual control system required to obtain performance gains.

In this chapter, we explore a DRM framework for digital logic using physics-based failure models for oxide breakdown, electromigration, thermal cycling and NBTI expressed as incremental damage mechanisms using a linear cumulative damage model referred to as Miner's rule [43]. The reversibility of NBTI damage is modeled and lifetime projection and the recovery effect on achievable system performance are explored. The performance impact of DRM in systems with DVS control techniques is analyzed with a focus on macro-level user collected processor usage profiles rather than traditional benchmark applications. The DRM system sets a maximum supply voltage based on the degradation characteristics modeled during operation, and exceeds the nominal supply voltage when possible while meeting the required reliability lifetime constraints. With the proposed implementation of a specific DRM control algorithm, this work demonstrates and quantifies the potential performance improvements of DRM utilizing dynamic voltage scaling.

This chapter is organized as follows: In 3.1, the adapted reliability models used to estimate failure rates under variable conditions are described. In section 3.2, the projection of failure rate at a desired lifetime using inputs from the reliability models is explored. In 3.3, The DRM control system that enables operation beyond nominal voltages is discussed and the simulation setup and results are presented. Finally, a summary of the simulation results and contributions of the work completes this chapter.

3.1 Reliability Modeling

In order to implement a real-time dynamic reliability management scheme, we require accurate models that can comprehend dynamic stress behavior with minimal

computational expense. High level compact models for oxide breakdown, electromigration and thermal cycling are addressed in the following sections. The models are adapted from state-of-the-art physics of failure work and applied to real-time DRM. In our proposed approach, we cast all reliability models such that they express wear-out in terms of an accumulated damage, or fraction of lifetime consumed. This approach allows simple projections of the failure rate at the desired lifetime and is key to the efficient computation of total failure probability which drives the proposed DRM control method. It also allows the use of degradation dependent models for each reliability mechanism, a capability that is lost when dealing directly with probabilities.

Oxide Breakdown

Oxide breakdown, or dielectric breakdown, is a degradation mechanism that results in a low-impedance path through an insulating or dielectric barrier. During normal operation, each electron passing a dielectric barrier has a small probability to enter a high-energy state to tunnel through the insulating layer. Defect paths in the dielectric barrier reduce the energy level required for conduction through the layer, and therefore increase the probability that electrons will travel through the layer.

Each tunneling charge has a small probability of creating a defect when passing through the oxide. This probability of defect generation is the wear-out mechanism for thin dielectric films. When a critical defect density is reached, there is a high probability that a low-impedance defect path exists in the oxide and a runaway current path through the insulating film will develop. The exact microstructure and nature of the defects is not well understood and less than 1% of defect paths

ultimately lead to an uncontrolled current path and oxide breakdown. The relationship between charge tunneling through the oxide and the defect density is expressed below in (3.1), where N_{BD} is the defect density, P_{DG} is the probability of defect generation, and I_{tunnel} is the tunneling current, V is the voltage across the oxide, and T is the temperature [6].

$$N_{BD} \approx \int_0^t P_{DG}(V,T)\, I_{tunnel}(V,T)dt \tag{3.1}$$

A simple simulation methodology for estimating the critical defect density required for a low-impedance defect path was originally developed by Degraeve [44] using a percolation concept. The percolation model places defects of a certain size into a 3-D oxide volume until a path of overlapping defects is created between the top and bottom planes. By running this simulation repeatedly for a given dielectric thickness, one can obtain a probability density function modeling the probability of a defect path related to the defect density. From this PDF, the approximate reliability of a thin-film dielectric is determined. The PDF generated by Monte Carlo simulation of the percolation model is fit to a Weibull distribution and used to calculate the probability of entering the onset of defect-induced oxide breakdown for an individual device.

The tunneling current through a gate oxide is calculated using BSIM4 model [45] equations, yet alternative methods could be employed. The BSIM4 model for gate oxide leakage is well-suited for this calculation due to readily available parameters for most processes and the significant validation efforts to ensure accuracy. The probability of defect generation is a technology-specific term with an increasing exponential trend with increasing supply voltage and an Arrenhius temperature

relationship. In this work, published defect generation relationships from an IBM technology node are used in the simulations [6]. This oxide breakdown model allows an incremental summation of defect density at variable supply voltage and temperature stress conditions. This closed-form, high-level oxide breakdown model is therefore ideal for a real-time DRM system considering dynamic stress conditions.

Electromigration

Electromigration (EM) is a failure mechanism caused by the movement of metal atoms through wires, creating voids (vacancies) and hillocks (deposits) that force open and short circuits in the surrounding wire networks. The transport phenomenon is primarily caused by electrical current, temperature gradients and diffusion processes in the conductors. Black's formula [7] is a well-known relationship between the mean time to failure of an interconnect and the current density, temperature and physical dimensions of the wire as shown in (3.2):

$$MTF = AJ^{-n}\exp\left(\frac{E_a}{kT}\right) \tag{3.2}$$

The term *A* is a constant related to the materials and the geometric structure of the wire and it generally increases with both width and thickness of the structure. *J* is the current density, E_a is activation energy for atom transport, *k* is the Boltzmann constant, and *T* is temperature. The value of *n* is also a constant that depends on the criterion for EM failure and the treatment of wire-self heating. Typical values of *n* lie in the range of 1.0 to 2.0 when wire self-heating is considered, although larger values may fit data more accurately when self-heating is not considered. When the criterion for failure is related to a critical void size, a value of *n* close to 1.0 is used,

whereas when considering a critical value of stress, a value of 2.0 is commonly used. The results in this paper are presented using a value of n=2.0.

Ideally a model for DRM should be expressed as a wear-out mechanism, with a quantifiable stress or damage term that is summed over time. Black's formula is instead expressed as a lifetime estimate based upon a single current density and temperature. We therefore use Miner's rule [43] (linear cumulative damage) to estimate the EM lifetime of a conductor by adding the percentage of lifetime consumed during each period of varying stress.

$$\sigma_{life} = \sum_{t} \frac{MTF_{ref}}{MTF(J,T)} \Delta t \tag{3.3}$$

Equation (3.3) summarizes the adaptation to Black's formula that allows variable stress conditions to be expressed as a percentage of lifetime, σ_{life}. MTF_{ref} is a reference value that would be characterized at worst-case conditions for the design, and $MTF(J,T)$ is an MTF calculation that is performed with varying current density and temperature averaged over a time window, Δt.

A Weibull distribution is again used to convert the percentage of lifetime figure (σ_{life}) to a probability of failure. Due to the scarcity of published distributions of failure relating to electromigration, the specific parameters are not available and a Weibull curve similar to the oxide breakdown curve is used. This Weibull distribution would need to be characterized for the specific process and geometric structures in the interconnect stack to provide sufficient accuracy for DRM. A self-consistent temperature is calculated for wires in each layer considering the thermal effects of wire resistance and the current density at a given supply voltage [46]. Equation (3.4)

relates the resistance of the wire to the temperature (T_{wire}), thickness of the wire (t), and resistivity of the material (ρ_0 *at* T_0):

$$R_{wire} = \rho_0/t[1 + \alpha(T_{wire} - T_0)] \quad (3.4)$$

$$T_{wire} = T_{sub} + R_{thermal}P_{wire} \quad (3.5)$$

Equation (3.5) demonstrates that the temperature of the wire is a function of the power, which is a function of the resistance R_{wire}. The thermal resistance, $R_{thermal}$ depends upon the layer of the interconnect stack and increases for upper levels of metal which are typically dominated by power and ground wires. The EM modeling in this work is limited to unidirectional currents (power/ground network) due to the greatly reduced experimental observation of failures in wires with bi-directional current [47-49]. However, the analysis could be extended to include bi-directional current carrying interconnects as well.

Thermal Cycling

Thermal cycling related failures are a growing concern in microelectronic devices as continued scaling has led to rising power densities and temperatures. Systems with power saving techniques (such as DVS or sleep modes) exacerbate the incidence of thermal cycling by modulating the power consumption, and therefore temperature, at a much greater frequency than in a conventional system. Thermal cycling is a mechanical stress mechanism that is manifested in many locations on an integrated circuit including solder connections and thin-film interfaces. As the temperature of the component materials on a chip changes, the

components will expand and contract at differing rates, since most materials will have different thermal coefficients of expansion. The intermolecular bonds in materials will actually change length as they store increased amounts of energy, leading to a change in volume for a component. These changes in volume over time can eventually create adhesion problems between layers, or potentially create shorts or opens in extreme cases. Reducing the number of thermal cycles a system undergoes will decrease the rate at which such mechanical stress abnormalities are observed.

Blish [50] related the number of cycles of thermal fatigue of various materials on a silicon die to the thermal swing via the well-known Coffin-Manson Equation [51]:

$$N_{cyc} \approx (\Delta t)^{-m} \tag{3.6}$$

$$N_{cyc} = \sum_{t} \frac{{\Delta T_{ref}}^{-m}}{\Delta T^{-m}} \tag{3.7}$$

The number of cycles, N_{cyc}, before breakdown is related to the thermal swing ΔT and a coefficient depending upon the materials involved. In this work, we consider thin-film cracking damage with a Coffin-Manson exponent of 8.4 [50] and again use Miner's rule to express (3.6) as a percentage of lifetime. Equation (3.7) formalizes the use of Miner's rule with the Coffin-Manson equation.

In (3.7), ΔT_{ref} is a reference thermal swing that the system is designed to withstand, and ΔT represents a measured thermal swing that may differ from the reference. Temperature traces are monitored in real-time to detect thermal swings. Equation (3.7) is used to sum the damage caused by thermal swings to express their contribution in terms of equivalent cycles of a larger swing ΔT_{ref}. Due to the lack of

availability of information (related to the difficulty in isolating this effect) regarding the probabilistic behavior of thermal cycling related wearout, some assumptions are made regarding the distribution of failures. A Weibull distribution centered at 10,000 cycles of max thermal stress is used to approximate the effect of thermal swings on system reliability. As empirical data becomes available, this approximation should be improved and validated.

Negative Bias Temperature Instability

The Negative Bias Temperature Instability effect (NBTI) leads to shifts in parameter values (V_{TH} and I_{DSAT}) in PMOS devices after extended periods of stress at negative voltages across the gate to channel region. The effect is caused by the dissociation of hydrogen atoms from Si-H bonds that are present near the interface of the dielectric oxide layer and the doped silicon channel region. NBTI is primarily observed at elevated temperatures when the device is biased in the inversion regime, when interaction with holes weakens the Si-H bonds. The dissociation of a hydrogen atom leaves a dangling Si+ bond that serves as an interface trap for free electrons in the surrounding area. The creation of interface states near the oxide-silicon surface leads to a reduction in the effective saturation current (I_{DSAT}) of PMOS devices and is often modeled as an increase in the threshold voltage (V_{TH}) of the MOSFET device.

A unique characteristic of the NBTI effect, is the recovery phenomenon that occurs when the electric field and temperature are relaxed. The dissociated hydrogen atoms return to the oxide interface and anneal many of the interface states that were created during the period of stress, leading to a partial recovery in the

saturation current of the PMOS device. Although the exact mechanisms governing the recovery effect are under debate in various physics journals, data collected from several test chips [52] indicates a strong link between the temperature of the device during stress and the extent to which recovery is possible. The model used in the proposed DRM system attempts to capture the dynamics and qualitative nuance of the stress phase and the temperature-limited recovery phase.

The Reaction-Diffusion model (R-D model) [8] is the standard chemistry model for the reaction that governs interface state (N_{it}) creation and annealing. The R-D model is detailed below in (3.8) and (3.9):

$$\frac{dN_{it}}{dt} = k_f[N_0 - N_{it}] - k_r N_{it} {N_H}^0 \tag{3.8}$$

$$dN_{it} = D - \frac{dN_H}{dx}|_{x=0} + \frac{\delta dN_H}{2dt} \tag{3.9}$$

Equation (3.8) represents the reaction-rate component of the R-D model where the forward reaction constant, k_f, depends on the initial Si-H bond density N_o and the current density of interface states, N_{it}. The reverse reaction is governed by k_r, the reverse reaction rate, N_{it} and the concentration of hydrogen at the interface, ${N_H}^o$. The diffusion equation models the outflow of hydrogen from the interface and the inflow of hydrogen across the oxide interface of dimension, δ. The reaction rate equation dominates initial creation of interface traps, leading to a rapid increase in NBTI damage when stress is applied. After the initial period of damage, the diffusion component of the reaction becomes the limiting factor and subsequent generation of interface traps slows considerably.

The R-D model matches measured circuit degradation well, yet is unsuitable for use in a DRM system due to the computational requirements of solving each iteration numerically. The following model, adapted from Vattikonda [11] uses a piecewise function that is a numerical solution of the R-D model for a hypothetical stress phase and recovery phase. A piecewise function is an excellent candidate for use in a traditional integrated circuit with a static power supply, or even with a sleep mode integrated circuit with well defined stress and recovery phases. In a DRM system using DVS to actuate the reliability mechanisms, it is difficult to define stress and recovery phases.

$$N_{it} = \sqrt{K_v^2 t_s^{2n} + N_{it0}^2} \tag{3.10}$$

$$K_v = A t_{ox} \sqrt{C_{ox}\left(V_{gs} - v_{th}\right)}\left[1 - \alpha V_{ds}/\left(V_{gs} - V_{th}\right)\right] e^{(E_{ox}/E_0)} e^{-\frac{E_a}{kT}} \tag{3.11}$$

Equation (3.10) and (3.11) model the accumulation of interface traps based upon the voltage and temperature stress and the period of time the circuit has been stressed. K_v is the stress factor, t_s is the time under stress, n is a technology dependent factor that is usually around 0.25-0.3 and N_{it0} is the initial concentration of interface traps when the stress period was initiated. The stress factor, K_v, is a strong function of V_{gs} and the related E_{ox}, and follows an arrenhius relationship with temperature. If V_{ds} does not equal zero, the NBTI stress is reduced at the drain or source end and results in a lower incidence of hydrogen dissociation. The t^{2n} term in the stress phase equation captures the initial rapid increase in interface traps and the transition to a diffusion limited reaction where trap generation slows.

$$N_{it} = (N_{it0} - \delta)\left[1 - \sqrt{n(t_r - t_0)/t_r}\right] \tag{3.12}$$

Equation (3.12) is the model used by Vattikonda for NBTI recovery when the electric field across the oxide is removed and hydrogen atoms have a probability of annealing the interface traps contributing to NBTI degradation. N_{it0} is the initial interface trap concentration at time t_0, n is a technology dependent variable around 0.35, and t_r is the recovery time elapsed since t_0. This piecewise function that alternates between stress and recovery has some significant drawbacks that required resolution for use in the proposed DRM system.

The stress equation cannot handle varying voltage and temperature due to the reliance on a fixed time component t_s in the formulation. For example, if voltage increases slightly, the term K_v will see an increase and t_s will remain the same, leading to a discontinuity in the calculated interface trap concentration N_{it} that is not present in measured data in the literature. To utilize this stress equation in a DVS system with frequently changing voltage and temperature stress, each time K_v changes, the time t_s must be recalculated. The recalculated t_s is an "effective time" given the previous stress, to maintain continuity in the trap generation curve and correctly model the future trend in trap generation.

$$t_{s,eff} = \left({N_{it}}^2/{K_{v,new}}^2\right)^{-2n} \tag{3.13}$$

Equation (3.13) is the simple method used to calculate the "effective" time at stress $K_{v,new}$ that is required to reach the current interface trap density. This value of t_s is used with the NBTI stress equation until another change in the stress value of K_v is encountered or the model transitions to the recovery function.

Consider the following scenario: NBTI stress begins and the initial N_{it} curve follows the t^{2n} trend, a brief period of stress relaxation occurs, then the stress is reapplied. In this circumstance, there should be a minor period of recovery, but the overall interface trap concentration should follow a trend line similar to that of an uninterrupted period of stress, as the simulation continues. If the N_{it0} term is used in (3.10) following the recovery period and t_s is reset to 0, the final trend line for interface trap generation will greatly exceed that of an uninterrupted period of stress. Intuitively and chemically, there is no evidence that this behavior should occur. Using the "effective" time calculation from (3.13) and ignoring the N_{it0} term in (3.10) will prevent this abnormality that comes up quite frequently in a DVS system modeled by these equations, as seen in (3.14). In (3.14), time t_0 is the time that the stress factor $K_{v,new}$ was recalculated and N_{it0} was the accumulated interface trap density at that time.

$$N_{it} = \sqrt{K_{v,new}\left[\left(\frac{{N_{it0}}^2}{{K_{v,new}}^2}\right)^{-2n} + (t - t_0)\right]^{2n}} \qquad (3.14)$$

Another simple modification to the recovery model is to account for the limited annealing following stress at high temperature. There are no known models for this damage "lock-in" effect of high temperature, but adding a generic function will allow more sophisticated models of this effect to be added as they are verified and come into use. A simple model of the recovery is shown in (3.15).

$$N_{it} = N_{it0}\left(1 - \varphi\left(Vdd, T, N_{it,max}\right)\sqrt{\frac{n(t_r - t_0)}{t_r}}\right) \qquad (3.15)$$

$$Let\ \varphi\left(Vdd, T, N_{it,max}\right) \approx 0.4$$

The φ function is estimated to be around 0.4 for circuits with stress temperatures in the 85-100ºC range from published data in [53]. This effectively represents a limitation to a recovery of 60% of the interface traps that were created during the stress period preceding.

The final challenge in implementing the piecewise function modeling NBTI degradation and annealing in a DVS system, is defining what a recovery phase is and when is the circuit in a stress phase. Particularly important is the transition from recovery to stress phases following a slight reduction in stress. When the system is binary, either at maximum voltage or in sleep mode, the definition is simple. However, the task of defining the NBTI degradation when the system experiences a 300 mV reduction in supply voltage or a 40ºC reduction in temperature is addressed in Figure 3.2. When the NBTI model is in the recovery state, and the stress is unchanged, both an updated interface traps value using the recovery phase model and the stress phase model are calculated using the current conditions. This models the annealing effect of recovery and the simultaneous creation of new traps, allowing a shift to the stress phase of the model when the $N_{it,stress}$ value exceeds the $N_{it,recov}$ value. These modifications to Vattikonda's NBTI model allow it to be used effectively within a DVS-DRM framework.

One final step to projecting NBTI damage over a varying workload is needed, since the computation overhead for simulating year-long traces of behavior is prohibitive.

```
Calculate K_v
If (K_v >> K_v,prev){
    CASE 1: Increasing Stress
    Calculate T_eff, stress
    Calculate N_it, N_it,perm
}
Else If (K_v << K_v,prev) {
    CASE 3: Initiate Recovery
    Save T_current
    Calculate N_it (N_it,perm influences recovery)
    Calculate T_eff,stress
}
Else {
    CASE 2: Continuation
    If (Recovery){  // Case 2.3
            Calculate N_it,recov (using recovery model)
            Calculate N_it,stress (using T_eff,stress)
            If( N_it,stress >= N_it,recov ) { Enter Damage Mode }
            Else{ N_it = N_it,recov }
    }
    Else{ // Case 2.1
            Calculate N_it, N_it,perm
    }
}
```

Figure 3.2 Proposed piecewise NBTI calculation algorithm.

The previous reliability mechanism models, especially oxide breakdown and thermal cycling are particularly amenable to projection, since the defect density and thermal cycles can be directly summed to create a lifetime estimate. This allows shorter application traces to be characterized with the detailed models and a lifetime trace to be constructed by superposing the results from the short traces to construct a longer trace. With the N_{it} value in the NBTI model, the generation is non-linear and a simple summation does not suffice for projecting shorter traces to a lifetime value. The approximation used in this work is translating the N_{it} value from a trace

to an “effective” time, in (3.13), of stress considering nominal voltage and temperature, and summing the damage in the time domain to create a reasonable estimate of lifetime degradation due to the NBTI mechanism.

3.2 System Level Modeling

This section presents an efficient approach to calculating system-level probability of failure that can be tailored to the desired level of detail. In this work, a single failure due to any reliability mechanism for any structure on the chip is a sufficient condition to declare that the chip has failed. In reality, individual dielectric breakdown events or electromigration voiding effects may not induce total system failure and certain components in a design (i.e., memory) may have built-in redundancy. However, this assumption will not significantly alter the conclusions reached.

The models outlined in section 3.1 calculate actual probabilities of failure for individual oxides and wires, considering the historical stress pattern of voltage and temperature in a dynamic system. Since the primary parameters that cause correlation between these failure mechanisms (voltage, temperature) are directly used in the calculation of accumulated stress over the simulated time span, the correlation between these failure mechanisms is naturally considered. This allows system-level probability calculations using probabilistic independence and greatly simplifies the mathematical formulation, as now described.

In order to derive the total projected system failure probability at the end of lifetime, t_{life}, we perform two tasks: 1) Based on the existing stress history and accumulated damage for a particular failure mechanism and device at the current

time t_1 , we project the probability of failure for that failure mechanism and device at t_{life}. 2) We combine the failure probabilities for all considered failure mechanisms and devices. We discuss each step below.

For each mechanism discussed in section 3.1, some concept of the degradation over time is maintained, typically expressed as a damage value. For oxide breakdown, the relevant metric is an estimate of defect density in a typical oxide layer, and NBTI damage is tabulated as the density of interface traps near the oxide-silicon interface, N_{it}. Thermal cycling damage is counted in cycles normalized to the maximum expected thermal swing, and EM is defined as the projected MTF due to average current and stress conditions for lack of a suitable damage variable.

In the system, the damage (D_1) at time t_1 is extrapolated to the damage (D_{life}) at time t_{life} based on history information about the rate of damage up to time t_1 using the following simple linear extrapolation:

$$D_{life} = D_1 \frac{t_{life}}{t_1} \tag{3.16}$$

Equation (3.16) accounts for environmental conditions and workload history intrinsically, providing a lifetime projection that is tailored to the exact stress conditions historically experienced on the chip. The model implicitly assumes that the future is similar to the past. However, we show in section 3.4 that, even under use profiles that display significant shifts over time, the proposed DRM algorithm provides stable control. Given the projected accumulated damage at t_{life}, the probability of failure of an individual structure is then calculated using a cumulative distribution function for the relevant reliability mechanism characterized for the given process technology.

Equation (3.16) is useful for scaling the damage done in the near-linear damage model equations, such as oxide breakdown, thermal cycling and the basic electromigration model described previously, but it is inaccurate for a non-linear mechanism like NBTI. To extrapolate the NBTI damage, the relationship between time and damage is implemented from the actual NBTI model itself to provide decent projections of actual NBTI damage.

$$D_{life,NBTI} = D_1 \left(\frac{t_{life}}{t_1}\right)^{0.25} \quad (3.17)$$

Equation (3.17) is very similar to (3.16), where the damage, D_1, at time t_1 is scaled to a lifetime damage prediction at t_{life} using the $(t_{life}/t_1)^{0.25}$ relationship from the NBTI model presented in section 3.1.

Individual device reliability projections are used to compute a chip-level reliability projection across all devices and all failure mechanisms using the follow expression:

$$\begin{aligned}
(1 - P_{ox}) &= \prod^{b}\prod^{d}(1 - P'_{ox-b}) \\
(1 - P_{EM}) &= \prod^{b}\prod^{l}\prod^{n}(1 - P'_{EM-b-l}) \\
(1 - P_{cyc}) &= \prod^{b}(1 - P'_{cyc-b}) \\
(1 - P_{NBTI}) &= min(1 - P'_{NBTI-b})
\end{aligned} \quad (3.18)$$

P_{ox} is the probability of oxide failure, P_{EM} is the probability of an electromigration failure, and P_{cyc} is the probability of a thermal cycling failure. Oxide breakdown failure probability is calculated based on the number of devices per functional unit (d,b) with a specific failure rate for a device from each individual functional unit (P'_{ox-}

$_b$). Electromigration failure rate is projected from the individual failure rate (P'_{EM-b-l}) across the number of wires (n), in each layer (l) and in each block (b). Thermal cycling failure is calculated as a component from each functional unit, since separate blocks undergo vastly different temperature traces. In future work, thermal cycling will consider the temperature gradients between functional units for this projection. NBTI failure probability is calculated differently from the other mechanisms, since the nature of an NBTI failure is much different (circuit timing failure vs. fundamental device/material failure). NBTI reduction in saturation current of PMOS devices is tracked at the block level and the minimum probability of correct operation for any block is used to represent the NBTI failure contribution, P_{NBTI}. The total chip failure rate is estimated by using the contributions of each failure mechanism in (3.18).

$$P_{failure} = 1 - \left((1 - P_{NBTI})(1 - P_{ox})(1 - P_{EM})(1 - P_{cyc})\right) \qquad (3.19)$$

Equation (3.19) is the combination of the failure rates due to each individual mechanism contributing to overall failure, $P_{failure}$. The simplicity of the chip failure rate calculation allows it to be used directly to drive a DRM control algorithm, which is described in the following section.

3.3 DRM System

Dynamic reliability management is implemented in this work using dynamic voltage scaling, which selects clock frequency and supply voltage pairs based upon workload demand and reliability model feedback. The scope of the DRM in this work includes digital logic blocks degrading from electromigration, NBTI, oxide breakdown and thermal cycling. The framework can be extended to include other wear-out mechanisms, such as channel hot carrier effects, or to consider the impact on analog, I/O circuits, or even packaging degradation with a similar modeling approach. For this work, DVS is an ideal control scheme for managing reliability concerns, since oxide breakdown and electromigration are both strongly voltage dependent and reductions in supply voltage greatly reduce the effect of these wear-out mechanisms. Thermal cycling is intuitively exacerbated by the increased variation in power consumption in a DVS system, yet thermal swings can be indirectly limited by capping the absolute maximum supply voltage, which limits the maximum temperature. If necessary, it is possible to further address reliability degradation due to temperature cycling by limiting the rate of voltage change in the DVS algorithm. In our analysis, however, this was found to be unnecessary.

Figure 3.3 details the organization of the DRM system that is implemented for maximizing the peak performance of a microprocessor system. Processor utilization traces are used to generate voltage/frequency traces for the DVS microprocessor. The selection of a voltage/frequency pair is converted to a block-based power consumption value that is derived from Wattch [54] application traces.

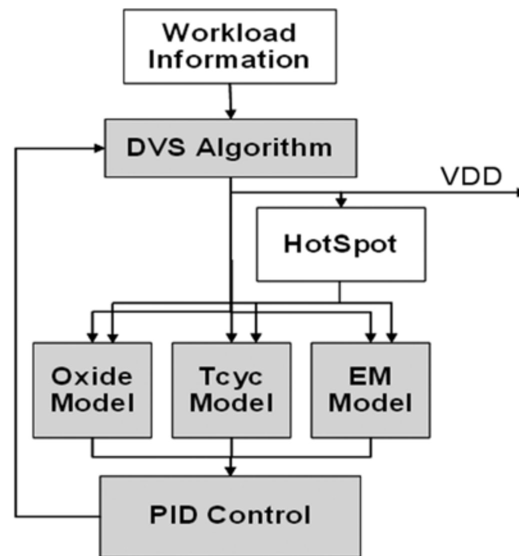


Figure 3.3 DRM System block diagram. Original system design includes Oxide, EM and thermal cycling wearout models.

Thermal information for each block is calculated using HotSpot [55] in the simulation flow, but can be replaced by a thermal sensor in an actual silicon implementation. The combination of voltage and thermal information is supplied to the four reliability mechanism models described in section 3.1 and the output of each model is combined to generate the probability of chip failure at the desired lifetime as described in section 3.2. Chip failure probability information is then used in a Proportional-Integral-Derivative (PID) control algorithm to set a maximum allowable voltage used in the DVS voltage assignment step.

A PID-based control algorithm [56, 57] is proposed as the key mechanism to provide maximum improvement in peak circuit performance when necessary, without affecting steady-state performance or comprising reliability. Equation (3.20) describes the behavior of a PID control system, where *e(t)* is an error signal, and *v(t)* is the output being controlled:

$$v(t) = v(t - t_0) + P\left[e(t) + R\int e(t)dt + D\frac{\delta e(t)}{\delta t}\right] \quad (3.20)$$

P is the proportional gain, *R* is the reset or integral gain, and *D* is the derivative gain. In general, proportional gain controls the response time of the controller, integral gain corrects for offset, and the derivative gain limits overshoot in the error term. In the proposed DRM system, *e(t)* is the probability of system failure projected to the lifetime, t_{life}, and *v(t)* is the maximum voltage available to the DVS algorithm.

The DRM system described is a discrete, non-linear, time-varying control system. Most of the reliability models are inherently non-linear with stress input and any models with a time-dependence or recovery mechanism (NBTI) are also time-varying, preventing a straightforward expression of the transfer function of the system. Therefore, it is unfeasible to present a general proof of system stability under all conditions. The results section presents some evidence of system stability under dramatic workload shifts using the impulse response of the system. An alternative to a closed-form proof of system stability could be achieved by fitting a mathematical model of the system response for a given implementation to collected data. Given this model, and the general PID equation, a proof of stability should be possible. Since the system is discrete in reality, (3.21) reflects the modifications made to the theoretical model of PID control in (3.20).

$$v(t) = v(t-1) + P\left[e(t) + R\sum_{-\infty}^{t} e(k) + D\big(e(t) - e(t-1)\big)\right] \quad (3.21)$$

Tuning of the control algorithm is dependent upon the desired response time and the length of time to correct offset issues. Overshoot, or selecting a maximum voltage that is too large, leads to a number of negative side-effects in a DVS system.

To compensate for the excessive amount of wear-out damage, the algorithm will reduce the clock frequency below nominal which could limit performance in subsequent time periods. In severe cases of poorly set proportional gain, oscillation between high and low voltages is observed in a classic case of an unstable feedback loop. The integral gain plays a very small and inconsequential role in this system, and the control system could be reduced to a P-D system without much impact on the selected voltages. Setting a relatively high derivative gain (on the order of proportional gain) delivers near-optimal control performance in the proposed DRM system by allowing a decent response time to processing demand requests, yet minimizing overshoot and undershoot when correcting the voltage setting.

In order to evaluate the effectiveness of the PID control system in achieving gains in processor frequency by using the available “reliability slack”, the figure of merit, peak performance improvement (PPI) is defined. PPI is a measure of the relative improvement in processor frequency possible when the system is operating at its peak demand. Essentially, it is a measure of how far the frequency can be “overclocked” during peak usage to deliver critical results. PPI is not a measure of overall system speed-up since it does not include any information regarding the proportion of total calculation time the processor may spend at the elevated voltage/frequency pairing, however, applications in parallel processing systems can be limited by the peak performance of a critical thread. Many parallel algorithms require global or semi-global reductions or calculations that often function in a sequential fashion that benefit greatly from the potential increased single processor performance offered by the proposed system. Synchronization in parallel

applications is another bottleneck that can potentially benefit from elevated peak performance to minimize the contention for data. Any system that involves user input and interaction can also potentially benefit from the peak performance gains available from the DRM-DVS concept. Although the PPI figure is not a measure of total system performance, it is certainly a metric of interest in many applications.

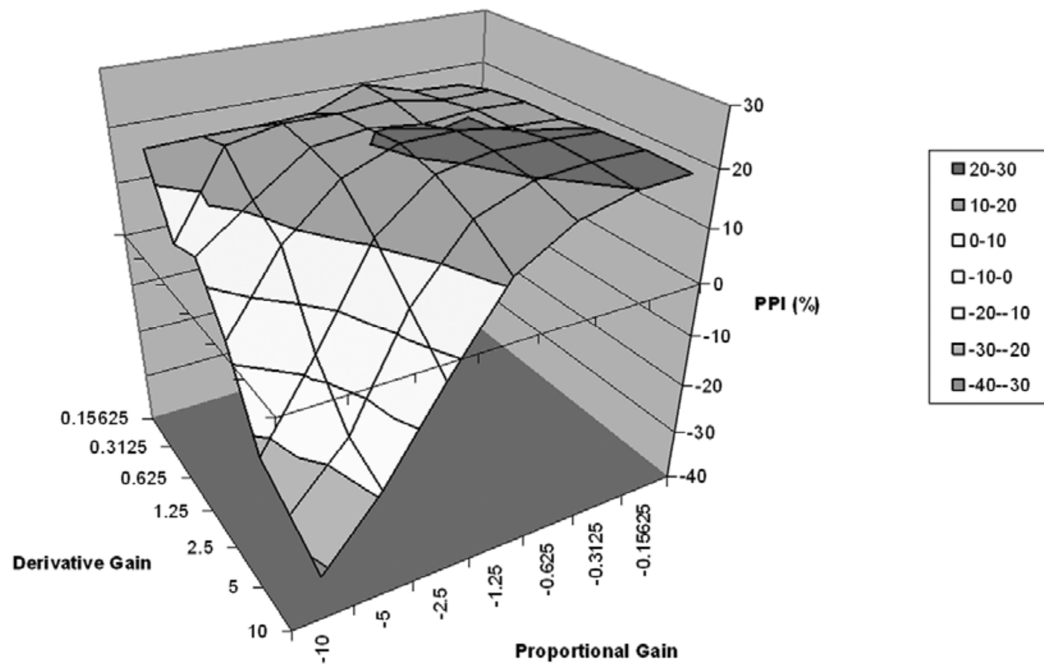


Figure 3.4 PID Controller gain values vs. peak performance improvement for example DRM system.

Figure 3.4 shows the improvement in peak performance (PPI) using the PID control system with varying values of proportional gain and derivative gain. Tuning the control system is somewhat dependent upon the variation in processor demand seen in the workload traces. In this plot, the average PPI is plotted for each set of values over a compilation of varying workload traces. An explanation of the workload trace collection is included in the section 3.4.

An actual implementation of the DRM system could take several forms, from purely hardware driven to purely software controlled. Several factors point in the direction of a predominantly software-oriented approach with limited hardware support. 1) Calculations to update the projected lifetime of the system and update the maximum assignable DVS voltage are needed infrequently, as the timescale of degradation is significantly greater than the timescale of computation. 2) Updating the models for a given system after development will be significantly easier for a system designed in software. Hardware components may consist of distributed sensors and a communication network, or temporary memory-mapped storage registers to maintain information on voltage and temperature history or the output of the sensors. Given infrequent updates, the overhead of the DRM system should be minimal in terms of performance when implemented in software, particularly when run on systems with significant "sleep" time. Area overhead should also be minimal for a software controlled system, with reasonable amounts of sensors. Assuming flexibility in placement of the sensors (placement in available whitespace), and routing, no more than 1-2% area overhead should be incurred. The remaining factor is the overhead of the DVS system, which has been implemented in an existing industrial processor [17] and shown to have no significant area overhead.

3.4 Results and Discussion

Workload data from several desktop computers was collected over several months to provide realistic processor utilization information with a wide-range of system behavior. A processor layout similar to the Alpha 21264 is used with process parameters based on 130nm industrial models, which are summarized in Table 3.1. The hypothetical processor is divided into 15 sub-blocks representing individual functional units on the chip (i.e. ALU, memory, decode unit). Initial power estimates are generated by Wattch and used with the processor utilization data collected to generate workload-based voltage, frequency, and power traces. The PID controller assigns voltage-frequency pairs based upon the requested performance and the reliability state of the system. HotSpot 2.0 is used to calculate temperatures for each functional unit in the design using power numbers adjusted according to the selected supply voltage. The PID controller updates the maximum voltage every 50μs in the short-time limit simulations presented and every hour in ten-year lifetime simulations.

The technology specification utilized for simulation is an aggressive 130nm technology based on values from several industrial models, predictive technology models, the SIA roadmap and available figures from the literature on the relevant reliability mechanisms. The relative impact of each mechanism is a strong factor in the simulation results. Given the values used in this study, oxide breakdown was the dominant mechanism in the results presented for the work with oxide breakdown, electromigration in power wires and thermal cycling. The electromigration model had a moderate impact on results, with thermal cycling

showing a minimal impact (when considering the on-chip component, thin film cracking). When considering the NBTI effect, it had a similar magnitude to the oxide breakdown mechanism, using aggressive figures for degradation. The NBTI effect in this work scaled up to 15-17% reduction in circuit speed, which is somewhat higher than figures from more recent published models, which cite 8-9% expected degradation [58].

TABLE 3.1
Simulation Technology Specification
and Selected Model Parameters of Interest

Symbol	Quantity	Value
L_{drawn}	channel length	130nm
V_{th0}	device threshold voltage	250mV
V_{DDnom}	nominal supply voltage	1.2V
T_{ox}	oxide thickness	1.8 nm
W_L	wire width (local)	140 nm
T_L	wire thickness (local)	350 nm
W_G	wire width (global)	450 nm
T_G	wire thickness (global)	1200 nm
ρ_0	wire resistivity (em model)	1.68×10^{-8} ohm-m
T_0	wire resistivity reference temperature	293.15 K
K_{ox}	wire thermal conductivity	0.25 W/K-m
n_{EM}	technology constant (em model)	2.00
A_{TFC}	thermal cycling constant	5.65×10^{21}
α_{TFC}	thermal cycling constant	-0.33
m_{TFC}	thermal cycling constant	8.4
n_{NBTI}	NBTI constant (stress/recovery)	0.25 / 0.35
A_{NBTI}	NBTI constant	1.80×10^{7}

Results (without NBTI model)

A figure of merit to quantify the performance gains available with DRM that will be used throughout the results discussion is "peak performance improvement" (PPI). This figure is a measure of the improvement in attainable frequency (%) during periods of peak CPU demand. This is a convenient measure of how well DRM provides additional performance when it is needed most by the user or application. The traces in Figure 3.5 are of a DRM simulation over a time span of only 16 seconds, allowing a detailed look at the interplay between the voltage assignment, workload, temperature and the projected failure rate at the 10 year lifetime. The horizontal line across the supply voltage trace is the nominal supply voltage, 1.2V, and the line across the failure rate curve is the target failure rate of 63.2% at 10 years. The plot clearly shows the increase in projected failure rate during periods of high supply voltage and temperature across this high activity profile. Longer simulations result in a much smoother failure rate projection curve, as the slope of the damage projection becomes more stable over time.

The histogram in Figure 3.5 shows the frequency of voltage assignments in the conventional voltage range and the boosted DRM voltage range. The distribution is for the Alpha 21264 system running the workload that was plotted in detail in Figure 3.5. The plot is bimodal since all tasks that require peak performance are executed at the maximum supply voltage allowed by the PID controller. Although there is no voltage limit upon the system, Figure 3.6 shows no data points beyond 1.7V and the majority of the boost voltage usage occurs below 1.5V. In an actual implementation of a DRM system, an upper bound on voltage could therefore be

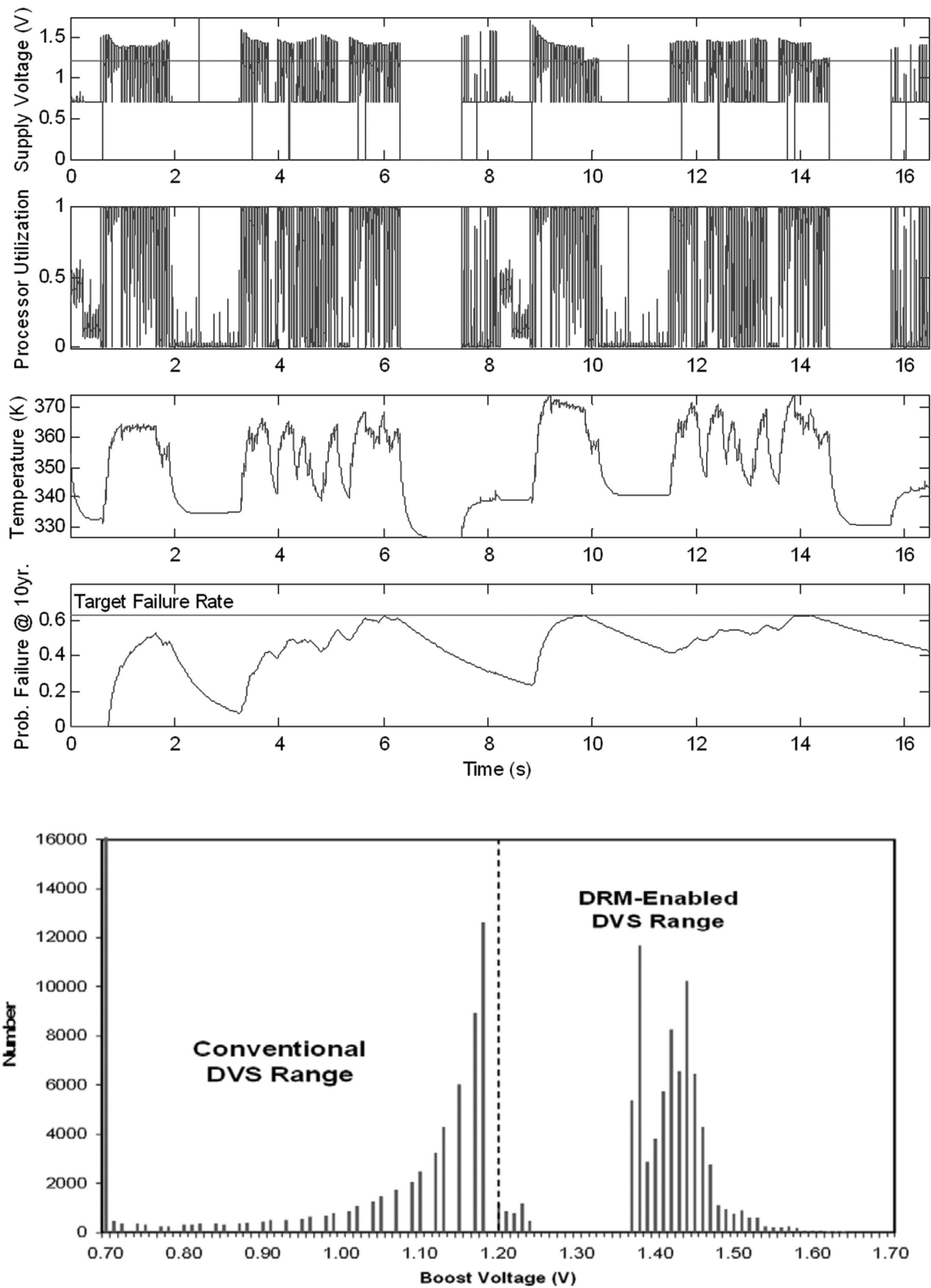


Figure 3.5 DRM Operation for workload C630 (16 seconds) and supply voltage assignment histogram.

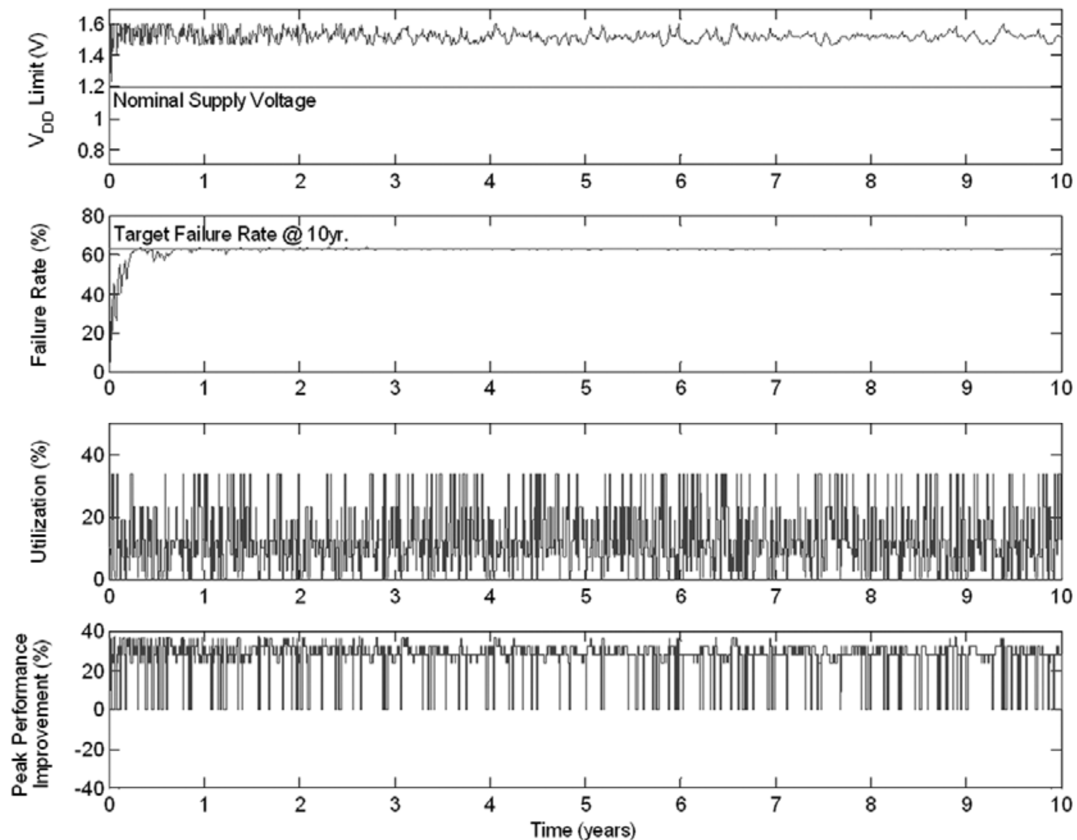


Figure 3.6 DRM Operation over 10 year reliability simulation. Slices of short-time scale operation used to construct 10 year lifetime trace.

placed at 1.5V to accommodate power distribution or voltage regulator limitations.

Figure 3.6 displays the 10-year performance of the DRM control algorithm over a randomized selection of 10 representative 1-hour workloads collected from an actual desktop machine. The randomization process selects a workload and then selects a random duration for that workload (ranging from 1 hour to 2 weeks) to be repeated using the 1 hour trace to fill that duration. The V_{DD} limit graph in each section of the plot represents the upper limit placed upon the DVS algorithm by the DRM mechanism, not the actual voltage during the entire trace. The PID controller does an excellent job of maintaining the target error rate over the long lifetime simulation.

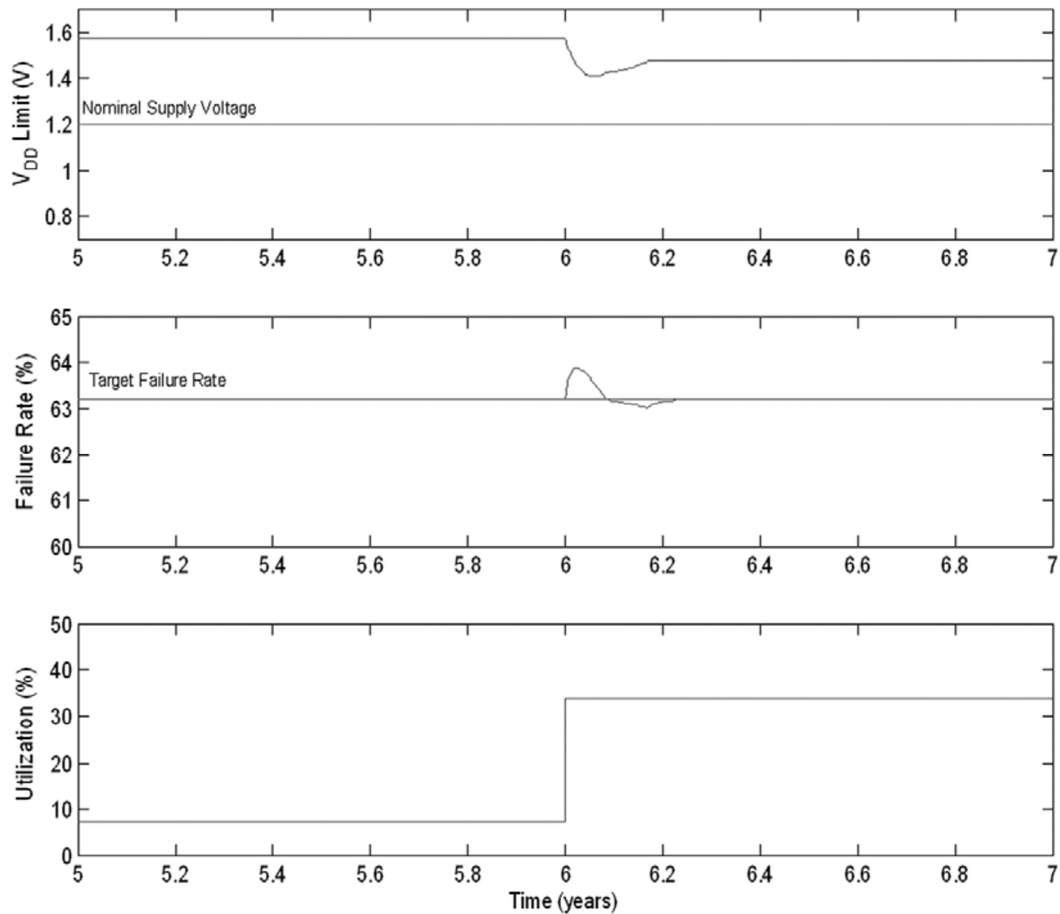


Figure 3.7 DRM PID controller step input response, representing sharp change in usage profile mid-life.

Although the nature of the calculation of error rate prevents any straightforward analysis on the stability of the control algorithm, it is possible to provide evidence of stability with extreme inputs to the system.

Figure 3.7 represents the response of the control algorithm to a sudden change in workload at the 6 year period, representing a pessimistic scenario for the proposed system. Tuning the controller involves a tradeoff between response time and stability. In Figure 3.7, there is a very small undershoot on the voltage limit, demonstrating the ability of the control algorithm to respond to sudden changes in

the workload. The response time of the system can be improved at the expense of the magnitude of the undershoot which could result in unnecessary performance throttling where the V_{DD} limit drop below the nominal supply voltage.

Peak performance gains over the 10-year workloads ranged from 20-35% compared to a nominal DVS controlled system. Specifically, the profile in Figure 3.6 shows a 26.7% peak performance improvement. In cases where the system is operating below the maximum operating temperature, significant overall performance improvements are possible. In Figure 3.6, with a design rated to a maximum on-die temperature of 125ºC, an ambient temperature of 60ºC allows 12.5% overall performance improvement before considering the workload.

Results (with NBTI model)

In Figure 3.8, the peak performance improvement is plotted vs. the workload activity factor. The workload profile for this plot is constructed with oscillations between 100% and 0% utilization at a period of 5 min. with a variable duty cycle that equals the activity factor. For extremely inactive systems, the voltage may be boosted dramatically above the nominal voltage, delivering a maximum of 34% peak performance improvement during periods of peak CPU demand. The sharp roll off in performance gains as activity factor is increased is related to the higher temperatures that are reached as the chip spends longer periods of time above nominal supply voltage. As the activity factor of the workload approaches 1.0, the performance gains are reduced to 12.5%. At this point, the benefits of DRM are derived not from periods of low voltage stress in the workload, but from a lower

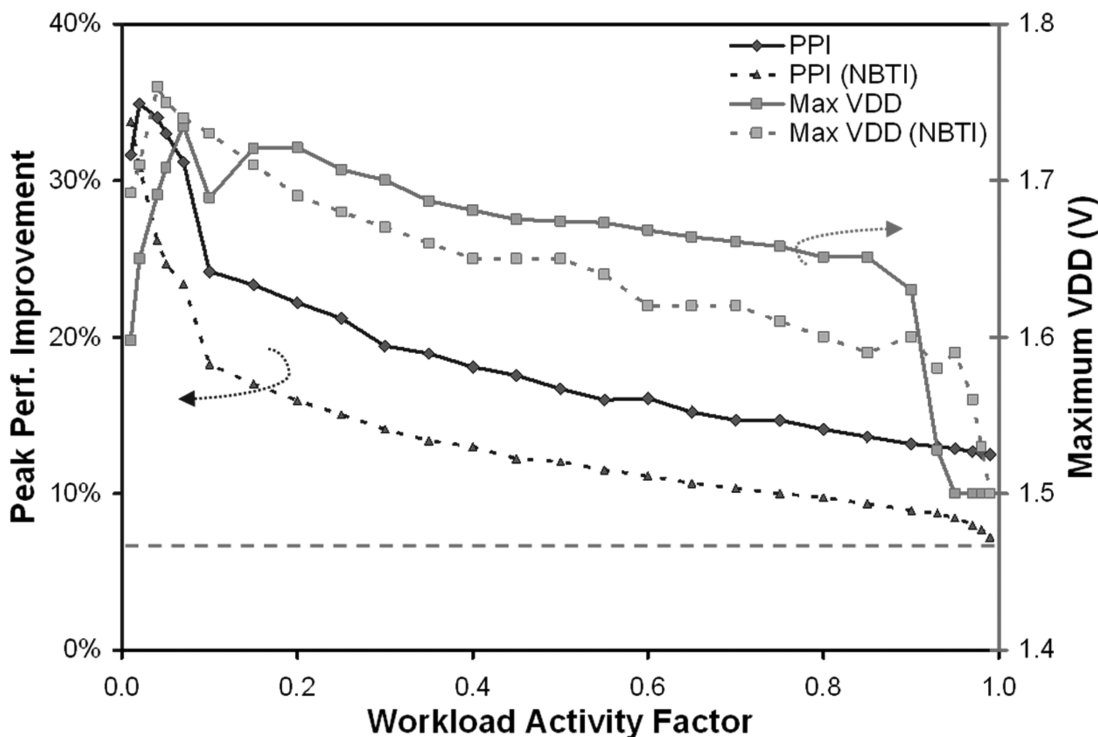


Figure 3.8 Peak performance and max supply voltage vs. workload activity considering NBTI

operating temperature than specified in the worst-case reliability corner.

The simulated chip is designed to operate at nominal voltage (1.2V) and a maximum on-die temperature of 125°C for 10 years. In this simulation, with an ambient temperature of 60°C, a 12.5% performance improvement is possible even with an extremely pessimistic workload. This plot demonstrates 15-20% peak performance improvement for workload profiles below a 100% maximum performance. Collected usage profiles from actual desktop machines indicate workload activity factors between 0.10-0.15 are typical for user-driven systems for which peak performance improvement is approximately 25%. The overall performance improvement when considering NBTI is also in the range of 0.0-2.6% depending upon the workload, and gains up to 7.5% were observed when operating significantly below the maximum on-die operating temperature.

Providing boosted supply voltages above nominal voltage may require a greater number of pads devoted to the power and ground network to handle the additional current associated with a higher clock frequency/voltage pairing. While the absolute cost, in terms of area or packaging, is difficult to quantify in a general sense, it is helpful to consider the effects on peak power consumption and lifetime energy consumption for systems with DRM implementations. Energy consumption over the lifetime was found to track the performance gains closely, and is relatively unaffected by a change in the maximum allowable voltage. Approximately 20.1% additional energy consumption is required to obtain peak performance gains of 20.68% over the lifetime of the chip under the workload of Figure 3.5. Peak power increases were found to be somewhat larger but naturally tend to be short in duration due to the feedback from the PID controller.

In Figure 3.9, the number of cycles with peak demand requested by the workload is linked to the attainable PPI on the left axis. Similar to the workload activity plot in Figure 3.8, there is a strong dependence on peak demand and the PPI, however this data is collected from the workload traces from actual desktop profiles rather than the synthetic benchmark used to generate Figure 3.8. The traces plotted here are 250,000 cycles in length, so data points on the left are near 0% workload activity and the right is near 100% workload activity. The maximum assigned voltage at each point is roughly constant at 1.8 V, yet this value is above the typical value seen in the DVS trace. The high assignment of 1.8 V is an artifact of the PID controller error function beginning the trace uninitialized leading to an exaggerated voltage assignment during the first period of peak demand.

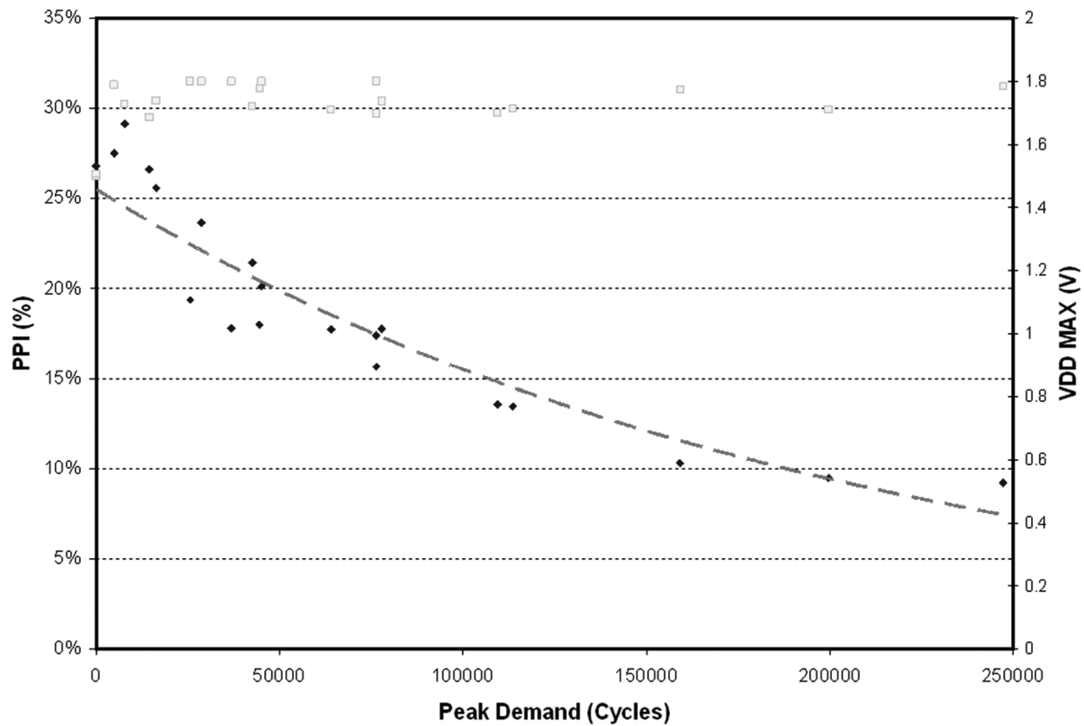


Figure 3.9 Peak demand vs. peak performance improvement (with NBTI)

An exponential curve is fitted to the simulated data in Figure 3.9, showing a rough exponential trend between 10-30% PPI depending upon the number of samples of peak demand witnessed in the simulated workloads. With proper characterization, the PPI attainable for a given system may be predicted with decent accuracy a priori, allowing some potential high-level optimizations in a multi-processing environment.

3.5 Summary

A framework for implementing dynamic reliability management is presented, including a rigorous model for failure rate prediction under four common failure mechanisms and a PID-based control system that balances increased throughput in

peak-demand periods with the remaining reliability lifetime. Workload and processor utilization information collected over months for typical users is used to quantify achievable gains in peak processor performance. On-chip real-time reliability monitoring allows supply voltages to be boosted beyond nominal values set during worst-case profiling and qualification, enabling maximum responsiveness during periods of critical computational demand. Despite minimal overall performance gains typically in the range of 0-2.6%, we observe typical peak performance gains of 20-35% over a variety of real-world workloads and lifetime usage profiles, without exceeding the specified lifetime budget. Considering the impact of NBTI reduces the achievable gains by 8-10% in simulation, yet still allows a peak performance gain of 15-25% over the typical range of workload activity of 0.05-0.20. The design of low-overhead on-chip sensors (temperature, tunneling current, etc.) is a key requirement in order to enable realistic silicon implementations and provide some validation of the modeling-based work.

Chapter 4

Analysis of Real-time Reliability Monitoring

Before a reasonable approach to controlling reliability breakdowns and degradation can be proposed, there are important questions to answer about the nature of the problem. Understanding the statistical nature of breakdown mechanisms and the influence of exogenous factors on their rate of wear-out or degradation will motivate the design constraints for any on-chip monitoring structures and define the potential and limitations for a real-time monitoring strategy.

Is it typical to see outlying device failures or will failures occur steadily after the first failure? If parametric outliers fail far earlier than other devices, it may be beneficial to design a monitoring scheme to detect actual individual device failures from large circuit blocks using in-situ testing of actual circuits. This scenario may present a challenge for predicting the relative importance of the outlying device failure. It may represent an impending widespread problem across the circuit block or a prolonged period of normal operation may follow. If the reliability and degradation-related failures occur regularly with increasing quantity, sampling a subset of devices may provide sufficient distribution information to assess circuit lifetime. This scenario may favor an array of sensors or “canary circuits” that are

monitored to provide the system or user with information regarding the “health” of circuits. The key to sampling the subset is tracking a progressive development of symptoms that lead to a failure, rather than the failure itself. Detecting a first failure in a subset of devices is unlikely to occur before there are widespread failures in the larger actual circuit. By tracking the progressive development of breakdown symptoms, reliability lifetime projection can be enhanced by knowledge of the state of degradation. An example of this idea is tracking the threshold shift in MOSFETs due to the bias temperature instability effect or measuring local impedance of power grid nets to detect electromigration-related thinning of metallization.

How will temperature, voltage, process variation and even circuit activity impact the probability of failure or degradation? Simulating the impact of temperature, supply voltage and process variation helps to quantify the sensitivity of each reliability mechanism to each factor. Understanding the magnitude of each effect on reliability degradation will help design appropriate control strategies to maximize circuit lifetime in a system utilizing adaptivity to manage the reliability lifetime challenge. If temperature is identified as the most significant stress factor, a system with thermal throttling could be an ideal solution to minimizing degradation. Likewise, in a voltage-stress-constrained system, dynamic voltage scaling could provide a near optimal method for stress relaxation. Additionally, understanding the impact of temperature, voltage and variation on the symptoms that preclude reliability failure events will allow a sensor design that is insensitive to the dominant factors when measuring the “health” of devices. This is essential for utilizing sensor

networks in the field, since temperature and supply voltage fluctuations cannot be assumed constant, as in an ideal testing situation.

Is there significant spatial correlation between failure events? When designing a sensor network for monitoring device degradation, significant levels of spatial correlation would indicate the need to disperse sensors evenly across a design to ensure that the sample space is representative of the actual circuit. In the absence of strong spatial correlation, a sensor array may be able to provide adequate measurement from a concentrated cluster of sensors, enabling potential area savings and minimizing routing and communication overhead.

This chapter attempts to address the aforementioned issues and determine the most effective methods for reducing the probability of failure or degradation through voltage, temperature or activity reduction. The conclusions from the analysis are used to explore the limitations of a real-time monitoring approach to understanding and controlling reliability issue. Assuming ideal sensors, confidence bounds for the error of real-time projection methods are presented as a function of sensor count.

First, the modeling platform used to simulate the failure distributions for the oxide breakdown mechanism is presented. Oxide breakdown is an ideal mechanism for this study since there are a variety of mature modeling approaches and it is a significant limiting factor on the scaling of device dimensions in future technologies. Results from simulation on the distribution of failure for the oxide breakdown mechanism and related conclusions regarding the factors that influence the mechanism follow the initial modeling discussion. Finally, an analysis of the results

details the implications on sensor-based monitoring techniques and defines the value a sensor array may have for real-time management or test characterization.

4.1 Oxide Breakdown Simulation Methodology

Oxide breakdown, or dielectric breakdown, is a degradation mechanism that results in a low-impedance path through an insulating or dielectric barrier. Failures related to this low-impedance path are typically manifest as abnormally high off-state leakage current, changes in circuit switching delay or even failure to switch in severe cases of degradation. Oxide reliability has been a primary concern during the development of new IC technology and several competing models that exhibit strong correlation with measured results have been proposed in the literature.

The two most well-known models for oxide breakdown, the E-model and the 1/E-model, were developed to relate the electric field to the time to breakdown (T_{BD}). In the E-model, the logarithm of T_{BD} is proportional to the electric field, E. Conversely, in the 1/E-model, the logarithm of time to breakdown is shown to be proportional to the reciprocal electric field, 1/E. Despite the seemingly conflicting forms of the E and 1/E model, the proponents of each model have presented compelling measured data to support their work. For the purpose of this work, the drawback of these models is that they primarily predict time to failure based solely upon electric field strength and they provide no information regarding the actual distribution shape or spread given a set of environmental and process conditions.

The percolation model, proposed by DeGraeve [44], treats oxide degradation as a series of traps or defects generated in the oxide layer. During operation, each

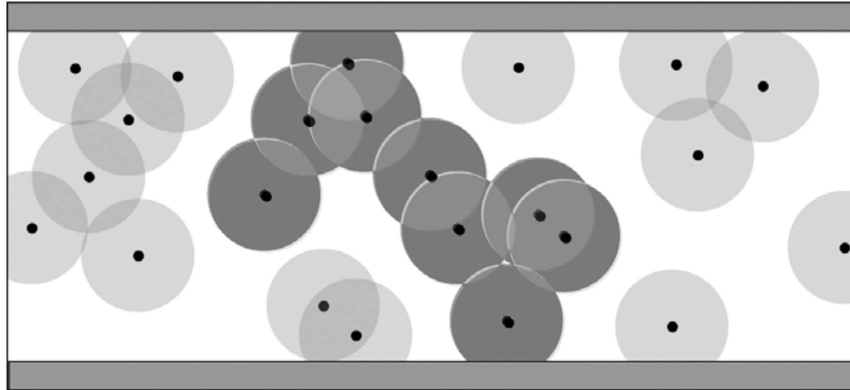


Figure 4.1 Percolation model placement of point defects in 3-dimensional oxide space (2D diagram shown). The darker point defects represent a complete chain of defects between the top and bottom plate of the oxide, forming a low impedance breakdown path.

electron passing a dielectric barrier has a small probability to enter a high-energy state to tunnel through the insulating layer and collide with particles in the lattice, possibly creating oxide traps or defects. Chaining paths of these oxide defects in the dielectric barrier reduce the energy level required for conduction through the layer, and therefore increase the probability that electrons will travel through the layer. Although, the E, 1/E and percolation models have all been experimentally validated with measured results, only the percolation model provides the capability of exploring a potential distribution of oxide failure through simulation rather than a single predicted number. The percolation model allows a thorough analysis of the factors affecting oxide breakdown and is used in the subsequent sections of this chapter.

The probability of defect generation of a tunneling charge is the wear-out mechanism for thin dielectric films in the percolation model. When a critical defect density inside the oxide volume is reached, there is a high probability that a low-impedance defect path exists in the oxide and a runaway current path through the insulating film will develop. The exact microstructure and nature of the defects is not

well understood and a low percentage of defect paths are assumed to ultimately lead to an uncontrolled current path and oxide breakdown. The relationship between charge tunneling through the oxide and the defect density is expressed below in (4.1), where N_{BD} is the defect density, P_{DG} is the probability of defect generation, and I_{tunnel} is the tunneling current, V is the voltage across the oxide, and T is the temperature [4].

$$N_{BD} \approx \int_0^t P_{DG}(V,T)\, I_{tunnel}(V,T)dt \tag{4.1}$$

A simple simulation methodology for estimating the critical defect density required for a low-impedance defect path was originally developed by Degraeve using a multi-dimensional placement simulation. The percolation model places defects of a certain size into a 3-D oxide volume until a path of overlapping defects is created between the top and bottom planes, as illustrated in Figure 4.1. By running this simulation repeatedly for a given dielectric thickness, one can obtain a probability density function for a chain of defects as a function of the defect density. Using the defect density, the probability of defect generation and the injected charge or tunneling current are used to calculate the time to formation of the defect chain, according to the relationship in (4.1).

The tunneling current through a gate oxide is calculated using BSIM4 model equations, yet alternative methods could be employed. The BSIM4 model for gate oxide leakage [45] is well-suited for this calculation due to readily available parameters for most processes and the accurate modeling of temperature and process variation related parameters. The BSIM4 model divides gate current into several components, representing gate-drain, gate-source and gate-body currents.

A complete description of the BSIM4 model is available in the model guide available on the web [59].

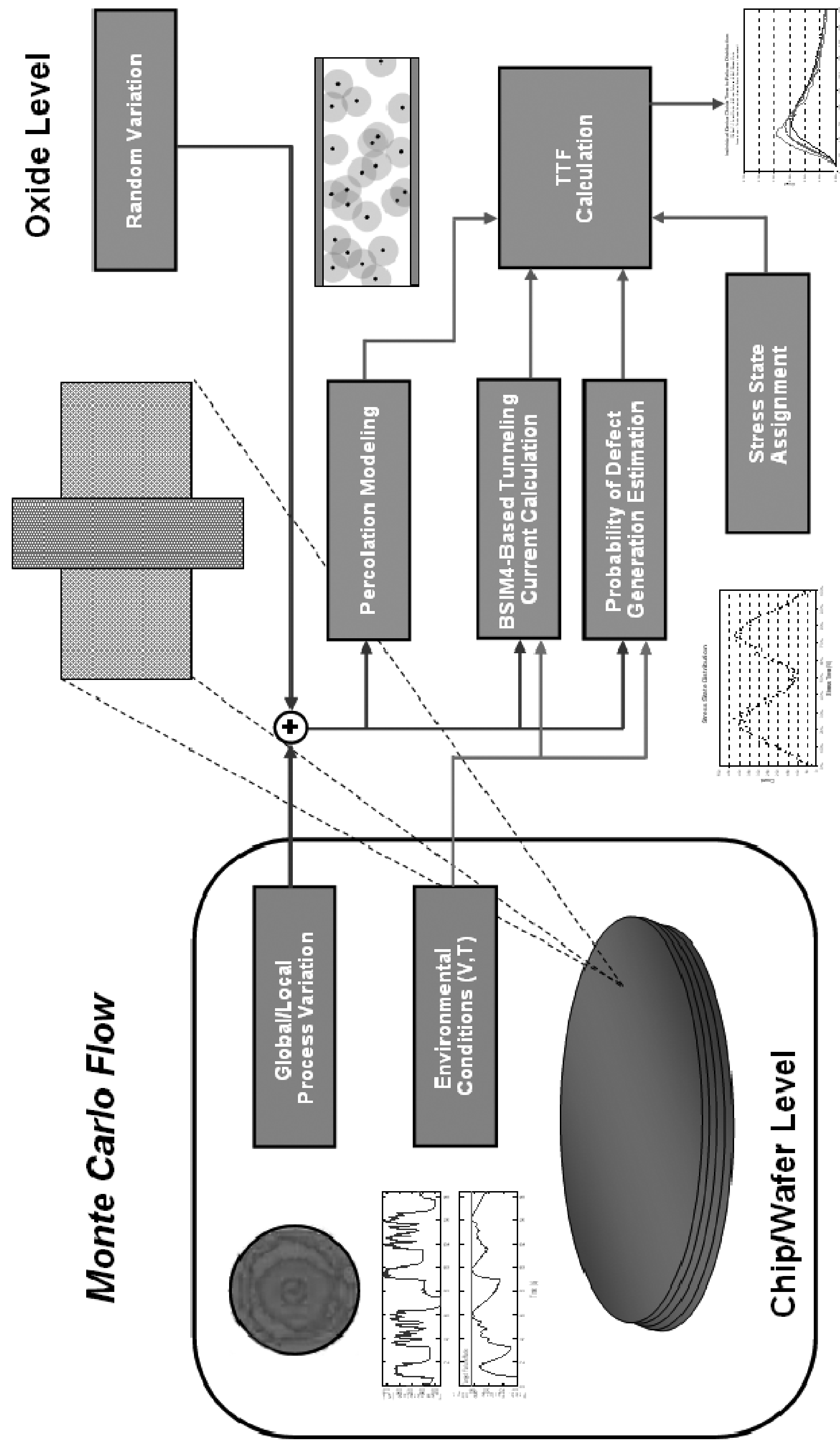
The probability of defect generation is a technology-specific term with an increasing exponential trend with increasing supply voltage and a non-arrenhius temperature relationship. In this work, published defect generation relationships from an IBM technology node are used in the simulations [6, 60]. The empirically collected data is fit to an equation relating the voltage and temperature [61] of the oxide layer to the probability of defect generation. In (4.2), V_{DD} is the stress voltage applied to the oxide and *T* is the temperature of the oxide.

$$P_{DG} = 4.9\ x\ 10^{-18} exp(9.5946\ V_{DD}) exp(3.85(1000/T)^2 - 29.679(1000/T) + 45.749) \quad (4.2)$$

This oxide breakdown model allows an incremental summation of defect density at variable supply voltage and temperature stress conditions. This closed-form, high-level oxide breakdown model is therefore ideal for a real-time DRM system considering dynamic stress conditions.

4.2 OBD Monte Carlo Framework

Using the oxide breakdown model detailed in the previous section, a monte carlo simulation methodology is utilized to generate a variety of oxide breakdown distributions with differing voltage, temperature, variation and stress time inputs. The diagram depicting the basic method used to generate the oxide breakdown distributions is outlined in Figure 4.2.


Monte Carlo Flow
Oxide Level
Random Variation
Percolation Modeling
BSIM4-Based Tunneling Current Calculation
Probability of Defect Generation Estimation
TTF Calculation
Stress State Assignment
Global/Local Process Variation
Environmental Conditions (V,T)
Chip/Wafer Level

The simulation method begins by generating a mapping of spatial variation in oxide length, width and thickness using a multi-level variation model. The variation model, is a four level model, with a global value, a 2nd level split into four regions with a 3rd tier split into 16 regions. The 4th level represents random variations but it is calculated on an individual oxide level after the die level variation model is created. Direct summation is used on the variation from the four-level model to deliver a final parameter variation for each device in the simulation (in this case, oxide). The values are taken from a normal distribution with parameter appropriate levels of variation listed in Table 4.1. For the simulations included in this study, the voltage and temperature values are statically set, although a simulation can be constructed that has a time-varying voltage or temperature to analyze the impact of voltage noise or activity-based temperature traces.

TABLE 4.1
Parameter Values for Simulation Framework

Symbol	Quantity	Value
L_{drawn}	channel length	130 nm
V_{th0}	device threshold voltage	250 mV
VDD_{nom}	nominal supply voltage	1.2 V
T_{ox}	oxide thickness	1.7-1.9 nm
$\sigma_{1\text{-}TOX}$	global variation sigma for T_{ox}	0.036 nm
$\sigma_{2\text{-}TOX}$	2nd tier variation sigma for T_{ox}	0.009 nm
$\sigma_{3\text{-}TOX}$	3rd tier variation sigma for T_{ox}	0.009 nm
$\sigma_{R\text{-}TOX}$	random variation sigma for T_{ox}	0.036 nm
$\sigma_{1\text{-}W/L}$	global variation sigma for W/L	4.0 nm / 2.6 nm
$\sigma_{2\text{-}W/L}$	2nd tier variation sigma for W/L	1.0 nm / 0.65
$\sigma_{3\text{-}W/L}$	3rd tier variation sigma for W/L	1.0 nm / 0.65
$\sigma_{R\text{-}W/L}$	random variation sigma for W/L	4.0 nm / 2.6 nm

At the individual oxide level, random variation is generated and combined with values from the multi-level chip model of variation to initiate the percolation simulation. The variation of length, width and oxide thickness is used to initialize the 3-dimensional oxide container for the percolation placement simulation. The small variations may influence the density required to create a continuous chain between the top and bottom plates of the oxide. The percolation simulation is completed and the defect density needed to complete the chain is saved and used to calculate the approximate time-to-failure as described by the oxide breakdown model section.

The variation information and the voltage/temperature data is used by the BSIM4 model and the empirically fit probability of defect generation relationships to calculate the TTF of the oxide being simulated. The stress state, or the fraction of real time that the oxide is under max stress, is sampled from a simple bimodal distribution with peaks around 20% and 80%. This value directly modifies the "rate of injected charge" in a linear fashion, adding a realistic factor to spread the distribution of oxide breakdown that would be typical in a system not containing all oxides stressed continuously.

In the final step, the probability of defect generation, gate current and stress state are combined using the relationship in (4.1) to calculate the TTF for a single oxide in the simulated chip. The simulation can be used to characterize a sufficient number of oxides from one "chip-level" variation map and voltage/temperature profile. The entire process loops at the chip-level to develop distributions for different sets of process variation under the same parameters and voltage or temperature traces.

4.3 Failure Distribution Analysis

The simulation framework was used to generate a variety of oxide breakdown distributions at different oxide thicknesses, temperatures and voltages. An important goal for the analysis is to explore the impact of various inputs to the reliability model for oxide breakdown. To initially present the impact of process variation, state dependence and the inherent randomness of oxide breakdown, a simulation with fixed oxide dimensions, voltage and temperature is explored. The voltage is fixed to 1.2V, temperature at 350K and the oxide is 400nm x 130nm.

Graphically, in Figure 4.3, the lifetime distribution for a chip composed of 25,000 oxides is presented and the change in the distribution can be seen as the factors are added to the analysis one by one. Initially, the percolation model is used alone to observe the innate randomness of the process, and this results in a distribution that clearly matches a Weibull, which is used in most oxide breakdown projection methodologies. However, as the effect of process variation is added, the curve PV-Percolation in Figure 4.3 shows a distribution that has shifted to appear lognormal. The variation in the oxide thickness and the oxide size causes an exponential effect on the tunneling current and injected charge, which introduces this lognormal shape. The effect of state dependence, the fraction of lifetime that the oxide spends at high stress (when not at stress, degradation is assumed to be 0), spreads the distribution, pushing the peak of early failures to an earlier predicted year and smoothing the long tail on the right side. The trends in distribution shape shown in Figure 4.3 determine the appropriate fitting functions for real-time lifetime projection.

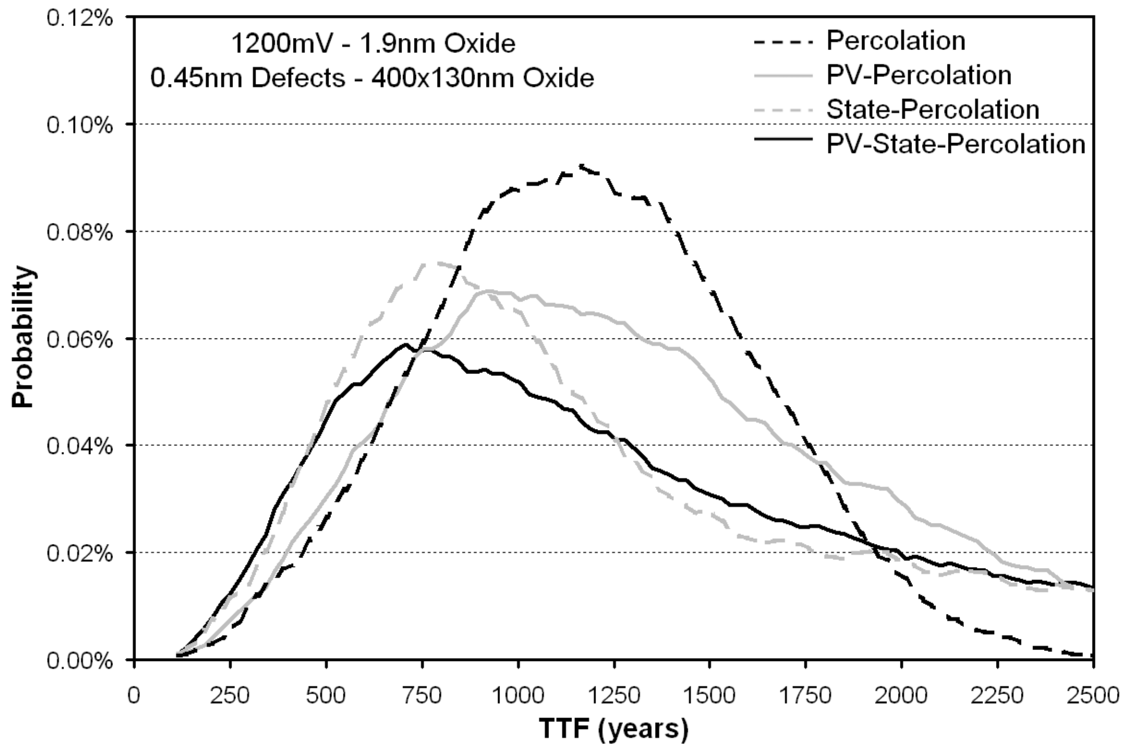


Figure 4.3 Oxide failure distributions with different simulation components included. Percolation represents the inherent randomness in the placement of defects, PV represents the impact of variation on the W,L and T of the oxide, state dependence represents the impact of the fraction of time a device is in a stressed state.

Figure 4.4 is a plot of the failure distribution of a 1.9nm oxide with 400x130nm dimensions considering percolation, process variation, stress state and varying environmental conditions. The effects of environmental conditions on the failure distribution of the oxides is enormous in this plot, particularly so for the wide range of temperatures (60-110ºC). The voltage range is selected to be on the order of a static offset in a regulator or a consistent small amount of noise in a power supply network. Even small voltage discretions lead to large changes in potential oxide lifetime due to exponential relationships in voltage and temperature. One of the greatest advantages of real-time monitoring of these effects is the ability to capture the impact of the environmental conditions on each die, which can be difficult

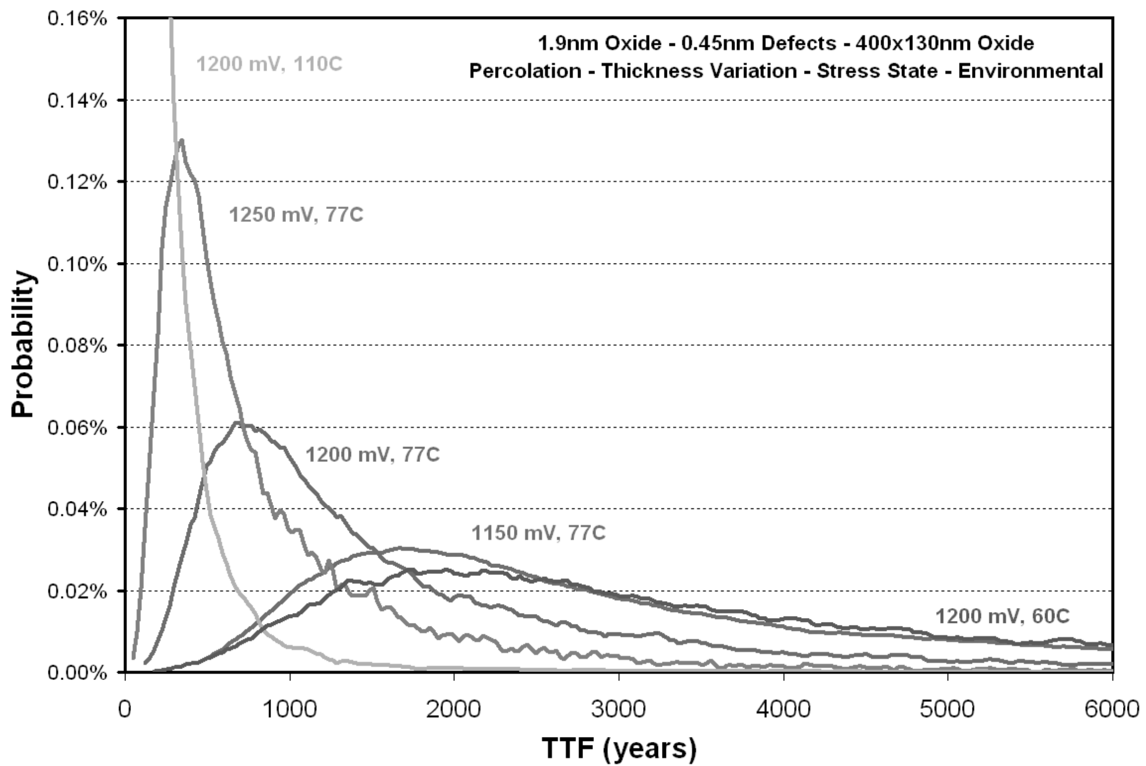


Figure 4.4 Failure distribution for 1.9nm oxide with conditions varying from 1.15-1.25V and 60-110C.

to estimate for general-purpose components.

Figure 4.5 displays the 25th and 75th percentile values for the first failures of a simulation of 100 dies with 25,000 oxides on each die. A simulation size of 25,000 oxides provides reasonable simulation time and exhibits comparable results to a 250,000 oxide simulation. Starting with the basic percolation simulation at the top, the added effects typically reduce the TTF, for example the baseline percolation simulation minimum observed TTF of 12.17 years, becomes 1.01 years when considering process variation, state, and temperature effects. The effects of spatial correlation within the multi-level process variation model are analyzed in Table 4.2. From the simulations of 100 dies consisting of 25,000 oxides, there are no signs of

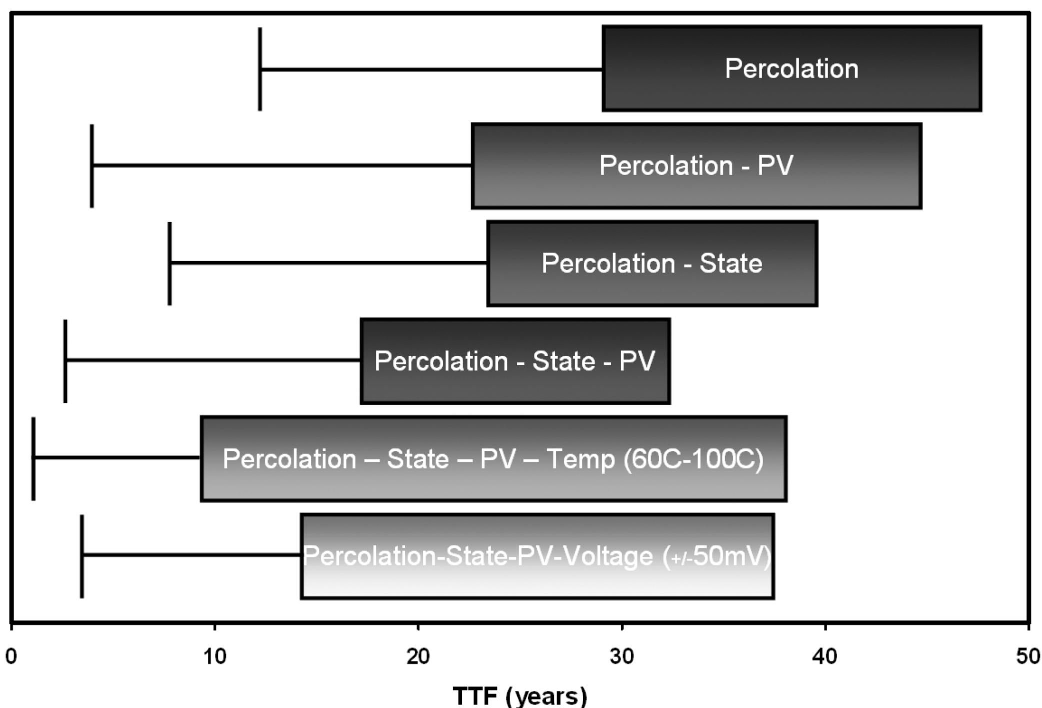


Figure 4.5 25th to 75th percentile ranges for 100 die simulation of 1.9nm oxide first failures. The bar to the left of each range indicates the minimum TTF outlier. Each bar includes a different set of simulation components to demonstrate effects of process variation, state or temperature on failure.

direct spatial correlation between the first observed failure and the second observed failure. However, when considering the first 5 or 10 failures, the data supports a moderate level of spatial correlation in large (25% die area) blocks. Monitoring circuits may be able to detect large areas that may be more susceptible to oxide failure, but to pinpoint a region for a second failure following an observed first failure is not likely.

The analysis of failure distributions using the simulation methodology answers many of the crucial questions at the beginning of this chapter. There is an innate randomness to the oxide breakdown effect, and outlying failures are typical. Voltage and temperature have a dominant effect on predicted failure time and effects like state dependence and process variation alter the shape of the distribution from a

pure Weibull shape to nearly lognormal. These observations guide the exploration of the use of real-time monitoring and the design of sensors to achieve real-time reliability monitoring.

TABLE 4.2

Spatial Correlation between Blocks in the PV Model

Description	Expected	Simulated
Second failure in 3rd level block of	6.25%	3.20%
Second failure in 2nd level block of	25.00%	22.30%
5 of first 5 failures share 2nd level	0.39%	3.20%
4 of first 5 failures share 2nd level	1.56%	9.80%
3 of first 5 failures share 2nd level	6.25%	29.50%

The 2nd level blocks are ¼ of the die area in the process variation model, and the 3rd level blocks are 1/16th of the die area in the model. Details of the process variation model are discussed in the OBD Monte Carlo Framework section.

4.4 Real-time Monitoring

Recent research into dynamic systems and reliability management has proposed the use of in-situ sensors to improve the inputs to model-based algorithms. The results of the previous simulation show that much can be gained from a reliability standpoint if you have some awareness of environmental conditions and the process variation of the die. Process variation and environmental conditions can be measure in real-time with known circuit techniques, yet even with knowledge of these values, a layer of modeling is needed to extrapolate the impact on reliability mechanisms. If a sensor could be designed to isolate and directly measure the degradation due to a particular mechanism, this layer of modeling uncertainty could be eliminated. Based upon the prior analysis, if an ideal sensor to detect the TTF for

an oxide device under test could be designed, this section aims to explore the bounds on accuracy and the requirements to effectively implement and utilize a real-time monitoring system.

Based upon the near-lognormal distribution shapes when including process variation and state dependence, least-squares, modified least-squares and maximum likelihood estimators for lognormal distributions were used to fit samples of simulated distributions to obtain a prediction on lifetime reliability of a set of oxides. The most accurate fitting method for prediction was least-squares fitting with the fit range censored to the earliest 30% of the sample set to ensure a good fit in the early failure range. The censoring method is more effective in fitting early failures that form the basis of any reliability failure or lifetime projection standard. Subsequent results on the quality of real-time monitoring approaches assumes the modified least-squares fitting method with censoring of mid-to-late life oxides.

From the high degree of innate randomness in oxide breakdown failures, it is clear that a multitude of sensors will be needed to obtain statistically significant information when directly measuring the oxide degradation. To analyze the effects of the number of sensor samples needed to effectively predict the TTF for a die, a simulation of a single die was performed and the failure times for the die are sampled (assuming the samples are the available sensors) with different sensor counts, ranging from 35 sensors to 5000 sensors on the die. The error from the predicted value of TTF using the modified LS method is used as a figure of merit and the findings are listed in Table 4.3. For the die in question, the actual failure time is 4.841 years (1.8nm oxide). From this table, the sensor prediction approaches 10%

average error around 1000 sensors. The clear point is that if direct measurement is intended, many sensors will be required and they will need to be very compact and accurate to provide a reasonable oxide lifetime projection.

TABLE 4.3
Sensor Count and Prediction Error based on Ideal Sensors
TTF of Die = 4.841 years

Sensors	Abs. Mean Error	Min Error	Max Error
35	3.7148	-3.5437	18.0896
50	2.5499	-3.3444	18.7818
100	1.4777	-3.8320	6.8266
250	0.9150	-2.3437	3.6080
500	0.7514	-1.8663	3.1098
1000	0.4415	-1.1737	2.3487
1500	0.3537	-1.2094	1.0934
2000	0.3339	-0.9953	1.3249
2500	0.3199	-0.8333	0.7647
5000	0.1929	-0.5549	0.6913

Figure 4.6 displays the 95% confidence bounds for 2 different dies using the modified LS method as sensor count increases. The worst case PVT line represents a typical corner estimate of the TTF for a chip from this process considering 85C, high voltage and pessimistic process variation. With low amounts of sensors, the 95% confidence bound can be worse than the corner estimate, yet with 500-1000 sensors, even a 95% confidence bound will result in a much better prediction than the corner model labeled worst-case PVT. A carefully designed real-time monitoring system can realize excellent improvements in TTF awareness (and the dynamic control schemes possible with this knowledge) over traditional corner-based reliability qualification.

Facing implementation of 500 or more sensors to directly monitor reliability mechanisms places strict constraints to realize a feasible system. The sensor design needs to be on the order of a standard cell macro block to ease placement and routing for such a large number of blocks. The accuracy requirement is high, since under the ideal assumption in this first look, large numbers of sensors are needed to reach an accurate prediction. In an era of billion transistor systems, 500-5000 sensors are a feasible budget to control the difficult problem of a priori reliability qualification.

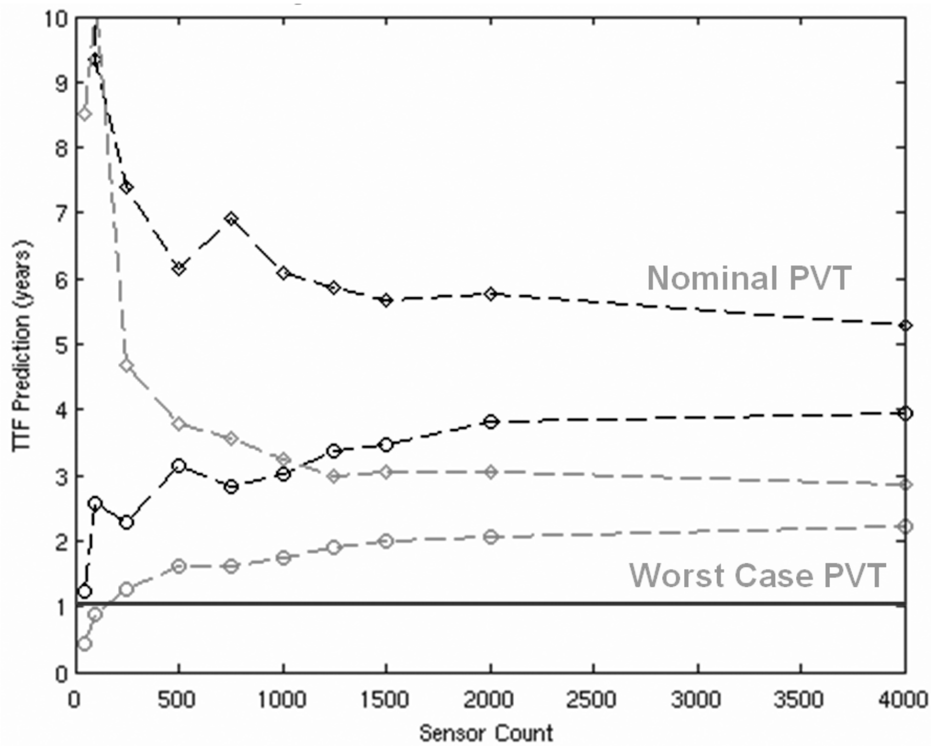


Figure 4.6 95% confidence interval in TTF prediction as sensor count increases for 2 different dies.

4.5 Summary

This work presents a new method for oxide lifetime simulation that delivers lifetime distributions using a computationally reasonable approach. The percolation model for oxide breakdown provides computationally efficient means for generating oxide lifetime distributions. The integration of percolation simulation, SPICE tunneling current modeling, the 4-level process variation model and state dependence distributions yields a monte carlo analysis flow ideal for analyzing the impact of temperature, voltage, variation and state on oxide lifetime.

The conclusions from the analysis conducted on oxide breakdown lifetime are clear. First, assuming an ideal sensor is available; 1000-2000 sensors throughout a chip are needed to provide lifetime predictions with less than 10% average error. The necessity of 1000-2000 sensors demands that the sensor is compact and energy efficient, minimizing overhead. The dominance of the inherent randomness, random process variation and supply voltage in determining the effective lifetime of the oxide indicates that the sensor array may not need to be dispersed widely around the chip, allowing a potential compact clustered design to minimize routing congestion and communication overhead. Although the simulation revealed only slight spatial correlation of failure events, this result is strongly dependent upon the variation model used for the study, and, as such, the conclusion needs to be revisited for each process technology.

A secondary goal of the study was to identify the optimal way to mitigate the progression of oxide breakdown. Supply voltage reduction is the most effective method for minimizing the degradation of oxide layers. In the event that it is not

practical to modulate the supply voltage of a given design, temperature throttling is also an effective method to minimize the likelihood of oxide breakdown events. State dependence of individual node voltages returns only linear benefits to lifetime, which is likely not worth the overhead required to use state control as a method to maximize oxide lifetime.

The following chapters attempt to draw from the insight gained from this monte carlo analysis of oxide breakdown factors to design effective monitoring structures under the following assumptions: 1) the sensor must be small and low-power, on the order of a series of NAND gates, 2) the sensor needs to be compatible with standard cell design methodologies, and 3) sensor accuracy should not be compromised by supply voltage noise or temperature fluctuation.

Chapter 5

A 130nm Oxide Degradation Sensor

Proposed dynamic control systems that modify the voltage, sleep state, and workload of components of large systems [17, 28-30, 62] further complicate *a priori* reliability qualification and call for on-chip structures for real-time estimation of various reliability mechanisms [31, 32, 63]. In Chapter 3, a dynamic reliability management system was presented that adaptively managed the supply voltage limit within a DVS system to maximize available performance within a reliability degradation budget. The prior work relied upon tracking the stress inputs to the system, such as voltage stress or temperature in order to model the impact upon the integrated devices. In this chapter, the design of a sensor that can track the reliability degradation actually occurring in devices on a chip is presented and discussed. These sensors enable the elimination of several layers of uncertain modeling required by the prior work by directly measuring the symptoms of each reliability mechanism that precede functional or parametric failure. An oxide degradation monitoring sensor design is presented that tracks gate oxide leakage current (a symptom of oxide degradation that increases as the oxide is damaged) in an ultra-compact, standard cell compatible circuit providing a digital output.

5.1 Oxide Degradation Sensor Design

Oxide degradation progresses to hard breakdown via increasing current through the insulating oxide layer, until it can no longer be considered an insulator. This undesirable current leads to failures in digital circuits by increasing or decreasing switching delay prior to creating a "stuck-at" fault or short, when the tunneling becomes comparable to the transistor I_{on} of connected devices in proximity. In analog circuits, the spurious current will slowly shift bias points which typically prevent the circuit from operating at nominal specifications. The premise of the proposed oxide degradation sensor is to measure the gate oxide leakage current through a typical MOSFET device over time, attempting to identify excursions from the initially measured quantity of gate leakage current.

The oxide-degradation sensor, in its simplest form, is composed of a miniature ring oscillator using the extremely high resistance of a gate oxide leakage to limit the charging time of a critical node driven solely by gate leakage current, as shown in Figure 5.1. The charging time of this critical node is sufficiently high to ignore the delay of the remainder of the ring oscillator gates, allowing a direct link between the frequency of this oscillator and the resistance of the gate oxide under test. The challenging aspect of this circuit is stressing the oxide under test via the critical node that is driven by gate leakage while making a measurement. In many technologies, any device connected via source or drain to the critical node will obscure the gate leakage current via subthreshold conduction. Additionally, to isolate the gate leakage of the devices under test, all other gate oxide connected to this node must use a thick oxide to limit gate leakage.

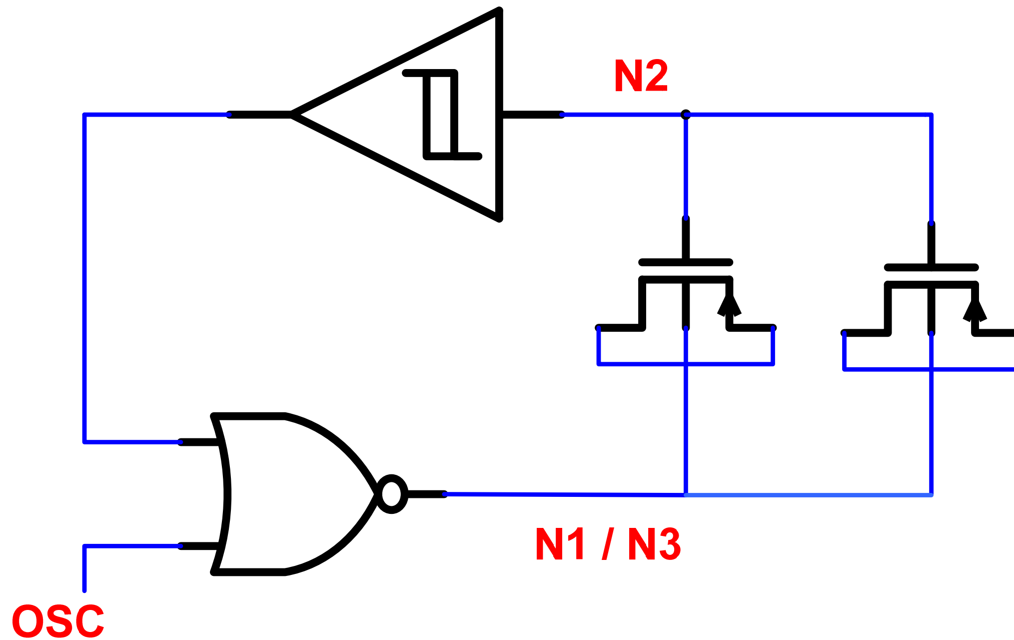


Figure 5.1 Simple model of oxide degradation sensor measurement featuring series gate oxide resistance and thick oxide Schmitt trigger oscillator.

In detail, the oxide degradation sensor consists of a pair of parallel PMOS devices connected in series with a Schmitt-trigger ring oscillator and a differential pair to drive a stress voltage to the gate oxide, in Figure 5.2. The frequency of oscillation is set by the gate leakage through the PMOS devices and size of the capacitor used at the node driven by gate leakage, N2. For measurement, the STRESS signal is driven low to short both drain-source-bulk nodes of the PMOS devices to the output of the oscillator and to deactivate the output of the differential amplifier circuit connected to N3. The OSC signal is driven high to enable oscillation, which is limited in frequency by a combination of the gate-leakage current charging N2 through the parallel PMOS devices and the coupling capacitance between N1/N3 and the critical node, N2. Since node N2 is driven only by gate leakage (ideally), careful sizing and layout is required to minimize the impact of coupling noise from

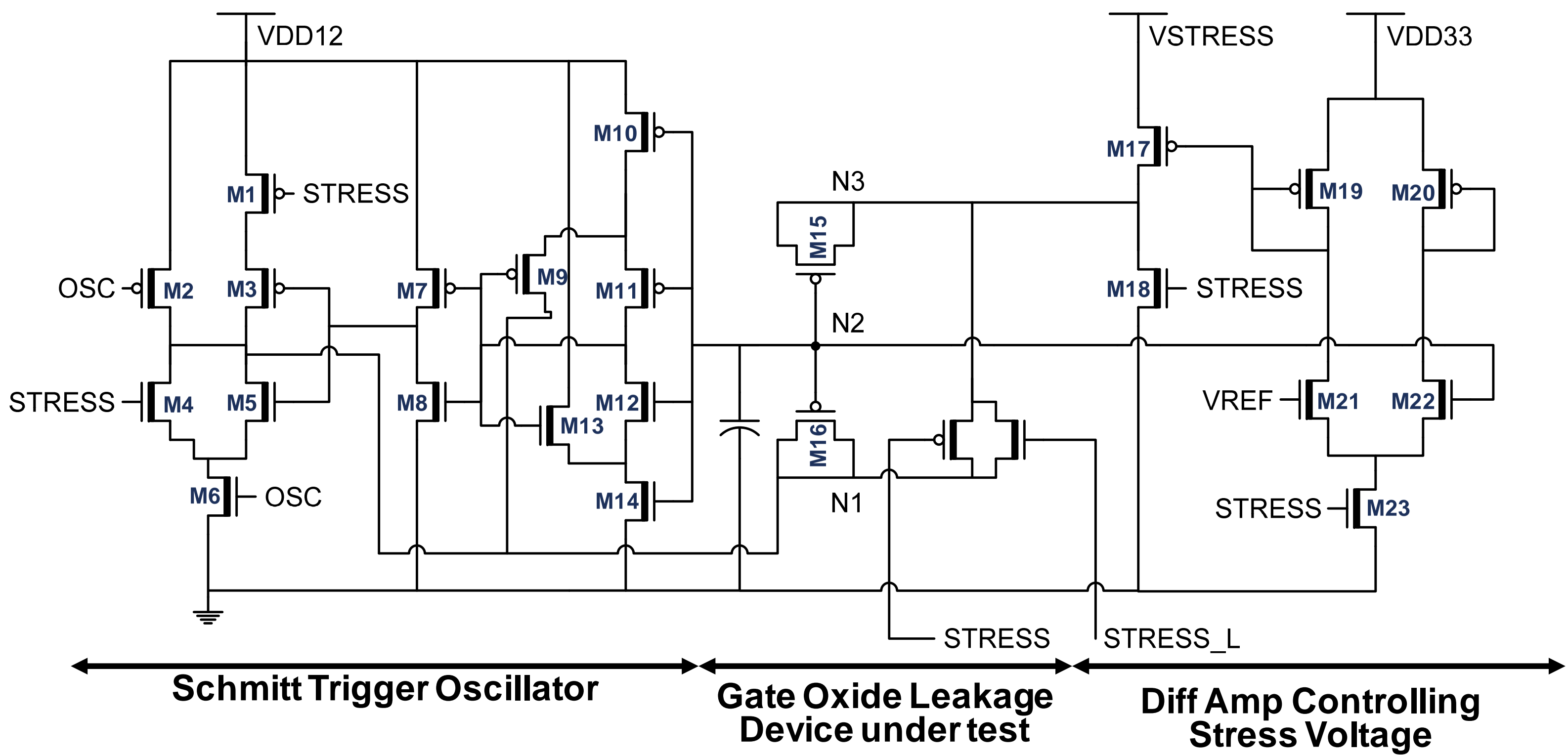


Figure 5.2 Oxide degradation sensor schematic, Schmitt trigger oscillator is limited in frequency by gate leakage charging critical node N2. A differential amplifier controls the stress device M17 which applies a high voltage stress across the gate oxide resistive divider.

the oscillator while minimizing the increase in cell area. In this design, the coupling noise to node N2 is mitigated by using a large capacitance to minimize the effect of the charge transfer from gate-to-drain and gate-to-source overlap capacitance. Node N2 oscillates with a peak-to-peak voltage of 430mV, so the capacitance is sized to limit coupling noise on this node to 10% of the peak-to-peak voltage swing, in this case 43 mV under nominal conditions.

To stress the oxide devices, node N2 is driven through a series resistive gate-leakage divider where N3 is driven to a voltage by a reference VSTRESS (in operation this reference may be local V_{DD}) and N1 is held to ground. In this case, N2 is held to a nominal voltage that is half of the applied voltage at N3 and will remain fairly constant until breakdown occurs, since an increase in leakage in one PMOS device in series will lead to an increased voltage drop in the other, eventually evening the imbalance in degradation. Although this balancing behavior may be considered undesirable, it is a tradeoff taken to isolate the gate leakage current through the device under test during measurement without attaching a device source of drain to the critical node, N2. Transistor M17 uses supply VSTRESS, while the differential amplifier circuit uses VDD33. In practice, VSTRESS cannot be increased beyond VDD33, or the stress voltage cannot be deactivated effectively. This portion of the design severely limits the maximum voltage used during accelerated stress testing and prevents the collection of data from advanced stages of oxide degradation. Table 5.1 below details the transistor sizing used in the fabricated implementation of the 130nm oxide degradation sensor.

TABLE 5.1
130nm Oxide Degradation Sensor Sizing

Transistor	Width	Length	Device	Oxide
M1	1.0μm	420nm	PMOS	3.3V
M2	1.0μm	420nm	PMOS	3.3V
M3	1.0μm	420nm	PMOS	3.3V
M4	680nm	420nm	NMOS	3.3V
M5	680nm	420nm	NMOS	3.3V
M6	680nm	420nm	NMOS	3.3V
M7	1.3μm	420nm	PMOS	3.3V
M8	800nm	420nm	NMOS	3.3V
M9	2.0μm	420nm	PMOS	3.3V
M10	1.0μm	420nm	PMOS	3.3V
M11	1.0μm	420nm	PMOS	3.3V
M12	800nm	420nm	NMOS	3.3V
M13	1.6 μm	420nm	NMOS	3.3V
M14	800nm	420nm	NMOS	3.3V
M15	1.0μm	120nm	PMOS	1.2V
M16	1.0μm	120nm	PMOS	1.2V
M17	3.2μm	400nm	PMOS	3.3V
M18	800nm	500nm	NMOS	3.3V
M19	800nm	400nm	PMOS	3.3V
M20	800nm	400nm	PMOS	3.3V
M21	6.0μm	400nm	NMOS	3.3V
M22	6.0μm	400nm	NMOS	3.3V
M23	500nm	1.0μm	NMOS	3.3V

The simulated operation of the oxide degradation sensor is show in Figure 5.3 for both oxide stress and measurement. From the beginning of the simulation until 150.0s on the x axis, the circuit is performing a measurement of the gate oxide leakage device under test. The STRESS signal is set low and the OSC signal is set high, so nodes N1 and N3 are shorted together via the pass-transistor mux. Node N2 oscillates via gate leakage at 95.76 mHz in this simulation. Sub-hertz oscillation frequencies simultaneously provide a benefit and a limitation. Measurement is very slow compared to picosecond FO4 delays in this 130nm technology; however

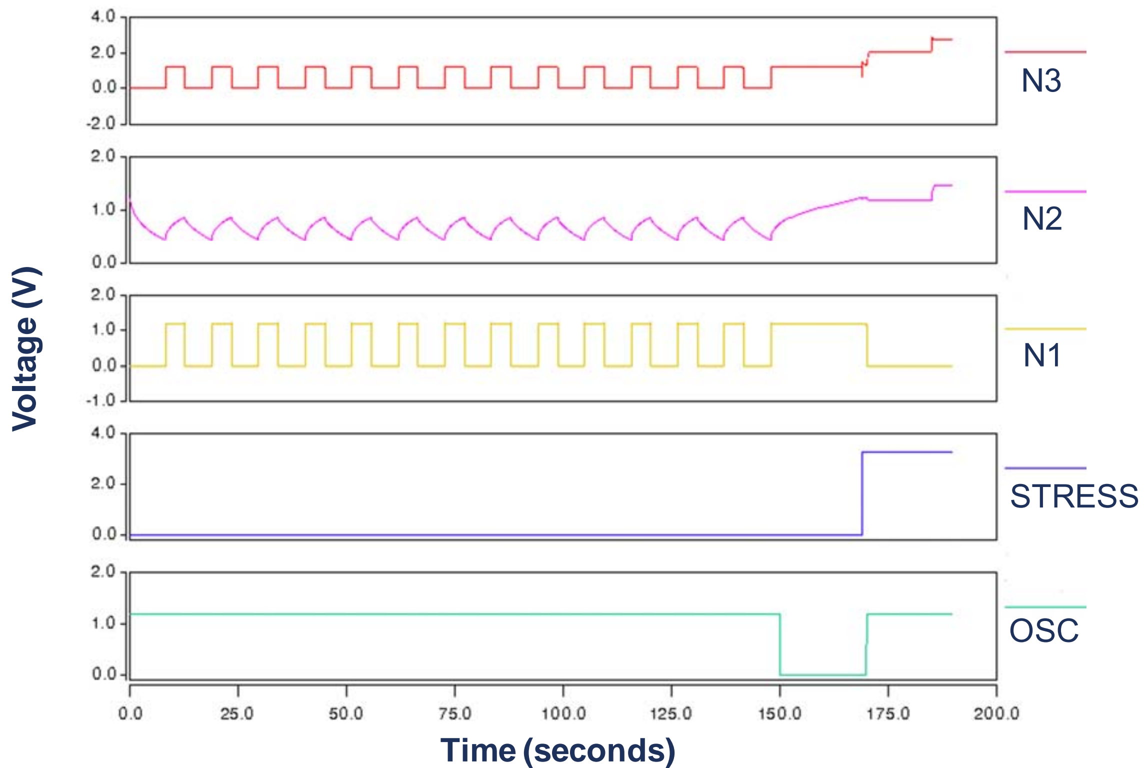


Figure 5.3 Simulated waveforms demonstrating sensor operation. Measurement period from 0-150s and stress begins at 150-180s on the time scaled axis.

the long period of oscillation time ensures that supply noise is averaged throughout the course of the measurement, preventing a worst-case droop or spike from severely impacting measurement.

In Figure 5.3, the oxide stress phase begins at 150 seconds, when the OSC signal is set low. This pulse lasts for ~25s and allows the internal node N2 to reach nearly VDD. It is essential to allow this precharging phase prior to stress to ensure that device M15 in Figure 5.2 does not see a voltage drop from VSTRESS to ground when the differential amp is initially activated. The precharged node caps the maximum initial voltage drop across device M15 to be VSTRESS-VDD12. In operation, VSTRESS should not be much more than twice the value of VDD12, therefore the initial voltage drop upon entering stress phase is limited to avoid

causing a premature breakdown event in the M15 oxide, causing the oxide degradation sensor to fail to track the degradation of typical devices on the chip.

Following the initial precharge phase before the high voltage stress is applied, the STRESS signal is set high and the OSC signal is set high. This activates M18 and M23 in the differential amplifier, allowing the device M17 to apply voltage stress from VSTRESS to devices M15 and M16. Node N1 is driven low by the series NMOS stack of M4 and M6. Devices under test M15 and M16 operate as a gate leakage resistive divider, which the internal node N2 rising to the reference voltage supplied to the differential amplifier, VREF. This fixes the voltage drop across M16 to be VREF, although it does not simultaneously control the drop across the M15 oxide. This is a limitation of this design, where it is possible to undergo uneven stressing of M15 and M16, potentially skewing the measurement result and indicating a premature oxide leakage increase. Under nominal conditions, identical devices M15 and M16 result in an equal voltage drops and equivalent oxide stress, and this is accepted as a satisfactory tradeoff to minimize the area of the sensor.

5.2 130nm Testchip Implementation

A combined test-chip containing 144 oxide-degradation sensors and 96 NBTI sensors was implemented in 130nm technology and fabricated through MOSIS [34]. The nominal gate-oxide thickness for standard devices is 2.2nm, while the nominal threshold voltage is 355mV for NMOS devices and -325mV for PMOS devices. Nominal supply voltage of 1.2V is used for all thin oxide devices and thick oxide devices in the core and pads utilized a 3.3V supply.

The oscillation frequency of the sensor described in the previous section is the digital output collected on the testchip. The testchip used 11 20-bit counters, connected to shared buses containing between 16-32 sensors. The 20-bit counter accumulated the pulses output from the oxide degradation selected by an address stored in control registers connected to a 1098 bit scan chain. A four entry LIFO queue is attached to the output of the counter, controlled by an external clock allowing 4 rapid measurements prior to a scan operation to extract the data.

Figure 5.4 shows the standard cell compatible layout fitting within 2 tracks of standard cell layout despite the predominantly thick oxide device cell design. Cell area is 150.19 μm^2 (7.20μm x 20.86μm), of which 30% is consumed by a large gate capacitance to minimize coupling noise on the oscillating node. Figure 5.5 shows the entire testchip design and detailed image showing the positioning of each block of degradation sensors on the chip. The block layout inset shows the 20-bit counter and 4 20-bit storage registers and 32 oxide sensors in 196μm. x 164μm.

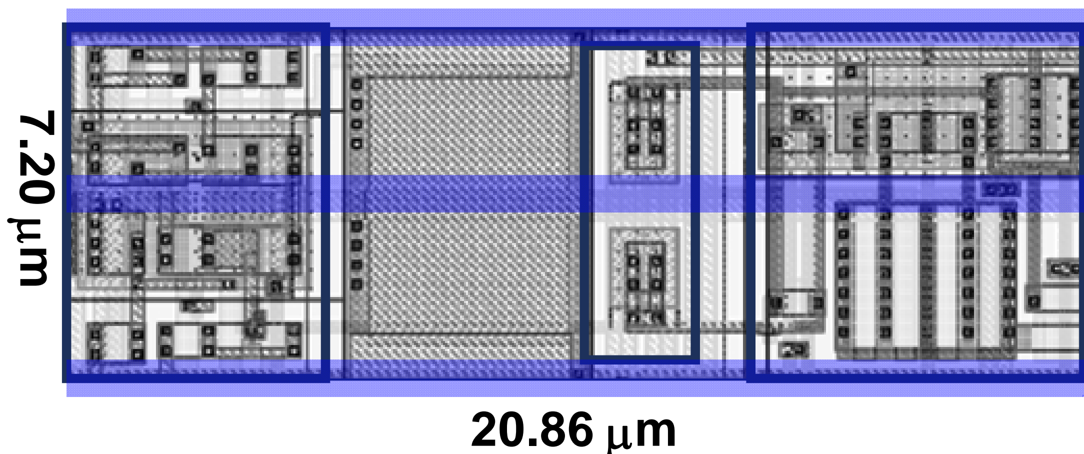
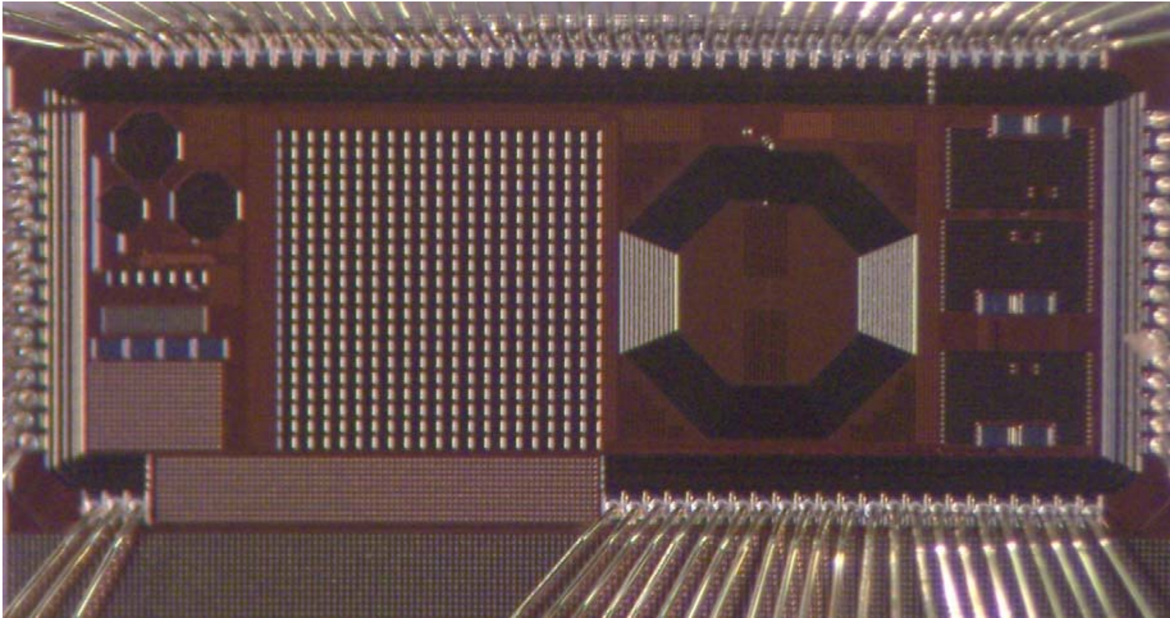


Figure 5.4 Oxide degradation sensor standard cell compatible layout.



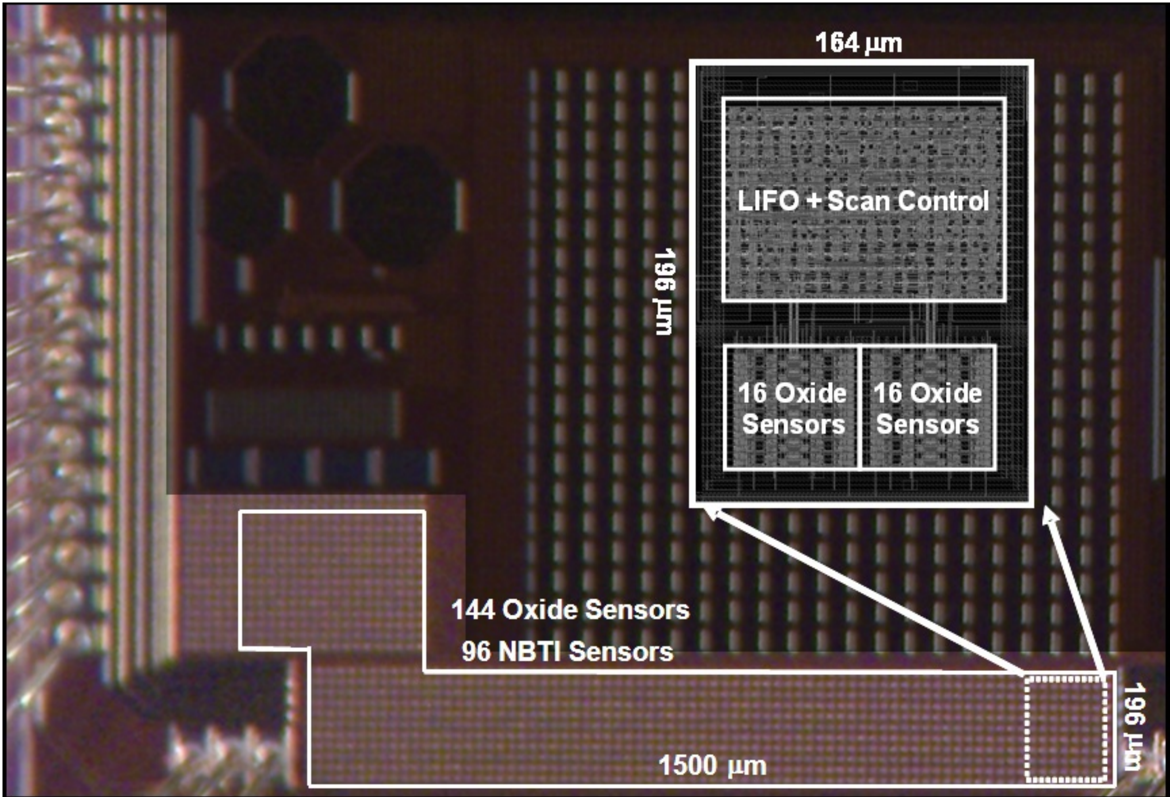


Figure 5.5 (Above) Complete 130nm multi-project testchip. (Below) Detailed view of oxide and NBTI degradation sensor test chip with block layout view inset.

5.3 Measured Results and Discussion

To accelerate testing of oxide degradation, ambient temperatures from 130 to 175°C and V_{STRESS} voltages to the sensor between 5 and 6V are used to achieve a stress voltage of 2.5 to 3V across each PMOS device. A simulation completed using extracted parasitic shows power consumption of the proposed sensor is 14.03mW for measurement and 469.5 mW for high voltage stress testing. Figure 5.6 shows the change over time in oscillator frequency as sensors from a single die are stressed at 2.5V across each oxide and a temperature of 130C. Consistent results from each of the 3 plotted sensors show ~22-27% increase in oscillator frequency as the oxide undergoes stress and gate leakage increases.

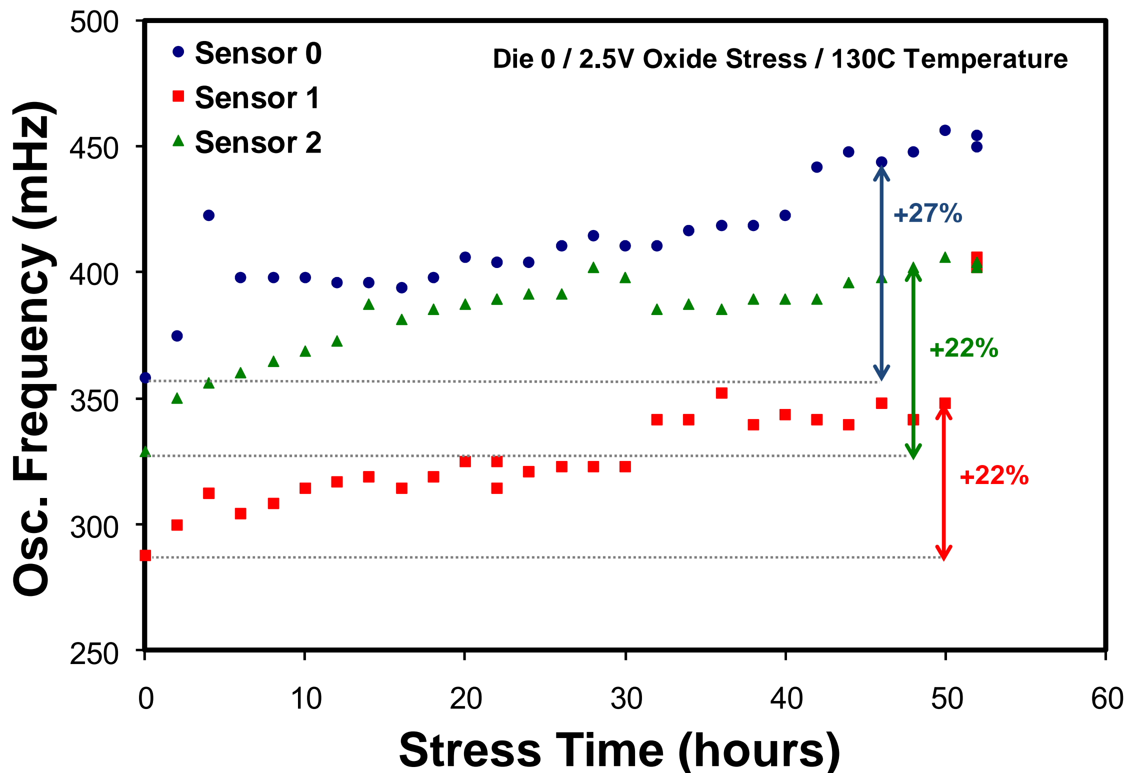


Figure 5.6 Oscillator frequency increase as the gate oxide is stressed at 2.5V and 130C for 3 sensors from a single die.

Figure 5.7 shows results from a series of 16 sensors from the same cluster from 2 different dies. Oscillator frequency increase from 56 hours of 2.5V oxide stress at 130C is variable within a group of sensors located in close proximity, highlighting the inherent randomness of oxide degradation and breakdown mechanisms. On average, this series of sensors exhibits 19% increase in oscillator frequency at the conclusion of the 56 hour test, with results ranging from 0-40% change.

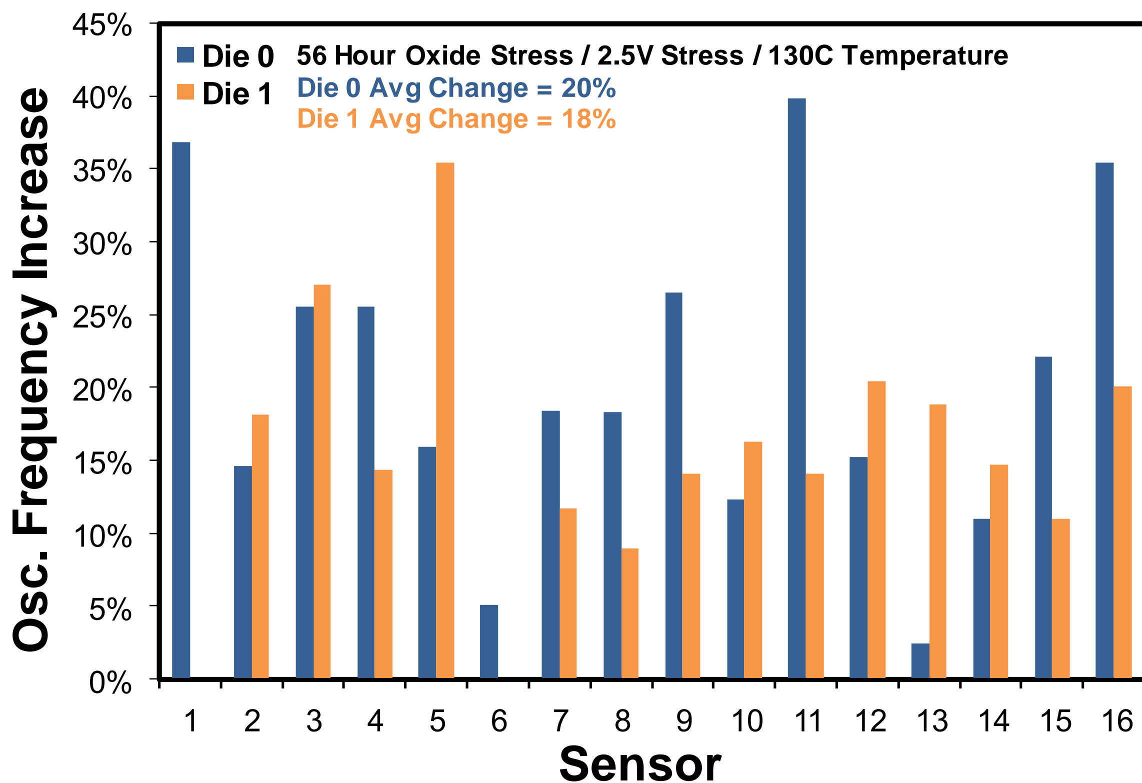


Figure 5.7 Final oxide degradation sensor frequency increase following 56 hours stress period at 2.5V and 130C.

Figure 5.8 shows the percentage oxide degradation sensor frequency increase over time as 2.5V and 130C stress is applied to an oxide. This plot again highlights the randomness in the effect of oxide stress on gate leakage current. For this die, after 50 hours of stress, 14-35% increase in the output frequency is observed. Note that some oxides are suffering from a sharp increase in frequency output in the first

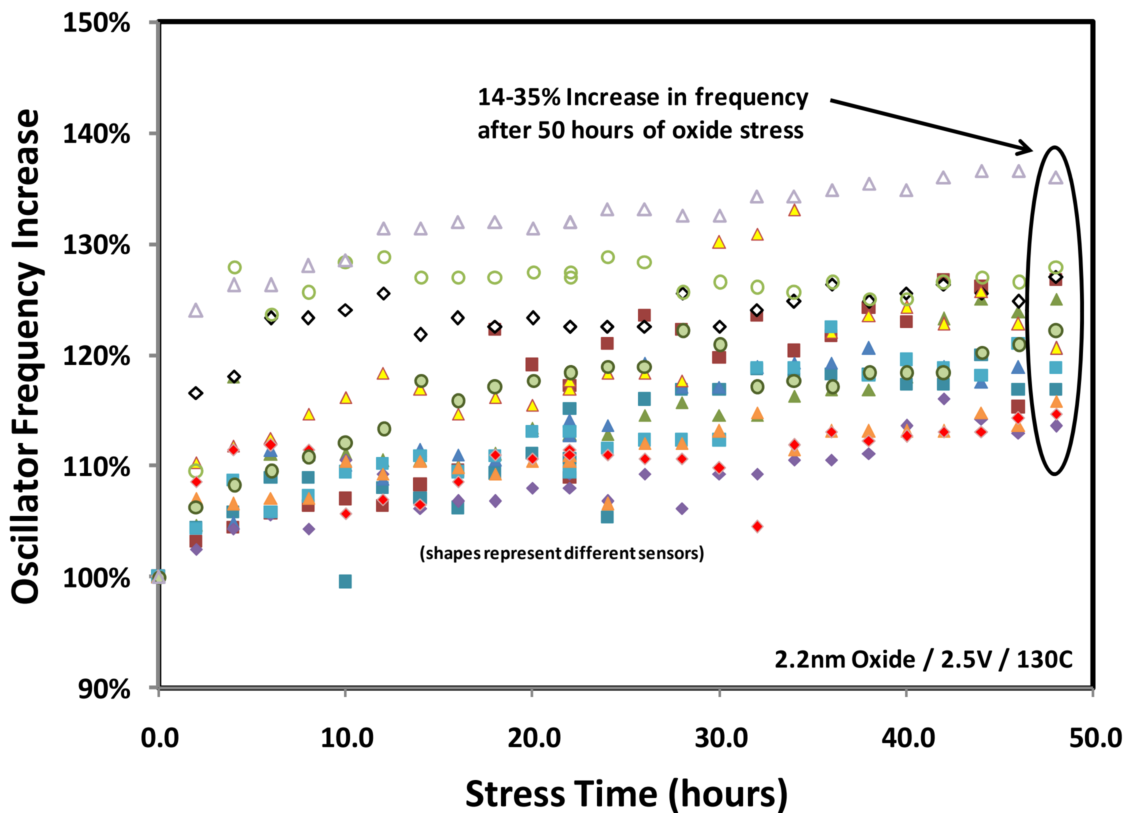


Figure 5.8 Oxide degradation sensor frequency increase as a function of time.

few hours that stress is applied. This is an interesting phenomenon that merits further investigation in future work.

An alternative use of this oxide degradation sensor is a simple method to measure gate leakage variability across a large number of devices. The initial oscillation frequency of the oscillator is an ideal way to measure the spread in gate leakage current over an array of sensors. Figure 5.9 fits a rough distribution to a series of oxide sensor measurements from 4 different dies. One die has an average oscillator frequency of 40mHz with another at 80mHz. Any system including an array of similar sensors can also report the initial spread of gate leakage current variability on die.

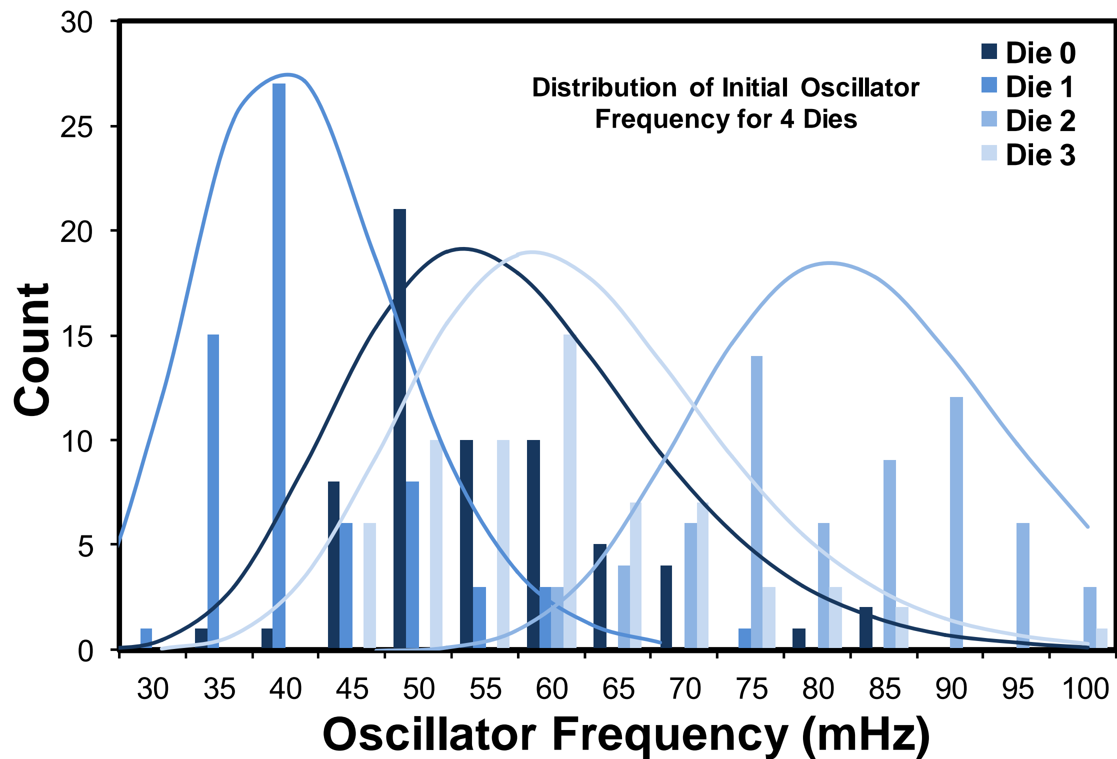


Figure 5.9 Distribution of initial oxide degradation sensor oscillating frequencies.

In the next section, the results and design of the oxide degradation sensor are summarized. A discussion of potential issues and limitations of this design is presented.

5.4 Summary and Discussion

The design and implementation of a novel oxide degradation sensor for use in real-time monitoring or characterization of degradation was presented in chapter 5. With accelerated stress testing, the oxide degradation sensor consistently shows increases in output frequency of ~19% after 50 hours of 2.5V stress at 130C. Due to the delicate nature of any charge on the critical node of the circuit drive only by gate leakage, it was not possible to directly probe the device under test to confirm the

exact source of the frequency shift in oscillator operation. In the following chapter, a discussion of the limitations of the initial oxide degradation sensor design and the design changes proposed to alleviate them is presented in detail.

Chapter 6

A 45nm NBTI/Oxide Degradation Sensor

In Chapter 5, the design and implementation of a novel oxide degradation sensor was detailed. Despite promising initial measurements, some drawbacks were discovered during testing and during the design of a 45nm second generation combined oxide and degradation sensor. The high power consumption during oxide stress, the potential for uneven voltage stress across the parallel MOSFETs under test, the extremely slow sub-Hz oscillation output and high voltage supply required to stress the gate leakage resistive divider were all identified as areas for improvement in the 2nd generation design of the oxide degradation sensor. Additionally, the integration of a successful NBTI degradation sensor into a combined NBTI/oxide degradation sensor was an additional goal of the updated design.

The opportunity to redesign the sensor in an emerging 45nm technology is also an excellent opportunity to gather valuable data on the magnitude of the reliability wearout problem in a scaled technology. Particularly, highlighting any increases or decreases in oxide degradation or NBTI threshold voltage shift is a goal of the new design. Additionally, noting any difference in the initial distribution of threshold voltages or oxide leakage is another benefit that can now be compared to the data obtained from the 130nm testchip.

6.1 2nd Generation Oxide Degradation Sensor Design

In an attempt to address the high voltage supply required, the high power consumption during stress and the potential for uneven stress across parallel leaking oxides, the entire stress delivery circuit from Figure 5.2, including the differential amplifier and high voltage supply were removed in favor of a simpler solution. The key observation is detailed in Table 6.1, showing the normalized gate leakage and subthreshold leakage for each transistor types in the 130nm process used for the original oxide degradation sensor and the leakage in the new 45nm technology. In the 130nm technology, the standard and low-power NMOS devices have gate leakage currents between 32.1-44.3, yet the high voltage thick oxide transistors required to deliver accelerated testing stress voltages have subthreshold leakage values of 3678.0/5880.8, more than 2 orders of magnitude greater than the gate leakage current under test. This observation led to the conclusion that it was not possible to design an oxide degradation sensor with a gate leakage driven node that had any transistor diffusion connection (source/drain connection). This design constraint severely limits the methods for stressing the oxides under test and forces the use of the gate leakage resistive divider and high voltage stress voltage supply that boosts power consumption and results in potentially uneven stressing oxide stress.

The initial investigation into the 45nm technology revealed that the gap between standard oxide thickness gate leakage (383.4-1500.0) and thick oxide subthreshold leakage (2901.5-5440.4) is not nearly as significant as in the 130nm technology used for the original oxide degradation sensor. When using an NMOS device for the

gate oxide under test and a series of thick oxide PMOS devices, the ratio of subthreshold leakage to gate leakage current falls to 1.934 (2901.5 to 1500.0). This promising ratio allows the exploration of designs that use carefully sized transistors with a source or drain connection to the critical node (see node N2 in Figure 5.2) driven by gate leakage.

TABLE 6.1

Subthreshold vs. Gate Leakage in 2 Process Technologies (Normalized)

Device	130nm Igate	130nm Isub	45nm Igate	45nm Isub
Standard NMOS	32.1	63212.4	1603.6	797927.5
Standard PMOS	1.5	50000.0	430.1	898963.7
Low Power NMOS	44.3	370.4	1500.0	74093.3
Low Power PMOS	1.0	435.2	383.4	98704.6
Thick NMOS	0.0	5880.8	8.5	5440.4
Thick PMOS	0.0	3678.8	1.5	2901.5

Figure 6.1 shows the modified circuit design for the 2nd generation of the oxide degradation sensor. To deliver voltage stress to the device under test (M1), 3 series stacked thick oxide PMOS devices with minimum width and greater than design-rule minimum length are used to drive critical node N2 to V_{STRESS} while node N1 is held to ground. This solution simultaneously solves three problems with the original design, reducing the power consumption by eliminating the need for the analog bias currents in the differential amplifier, creating a uniform and regular voltage stress across the single device under test (M1) and requiring the high voltage supply, V_{STRESS}, only for accelerated stress testing. During actual operation, V_{STRESS} can be attached to the local power supply grid, alleviating the need for any external voltage or bias for the 2nd generation oxide degradation sensor. In this case, the level-converter driving M7, M8 and M9 can be removed to minimize the cell area.

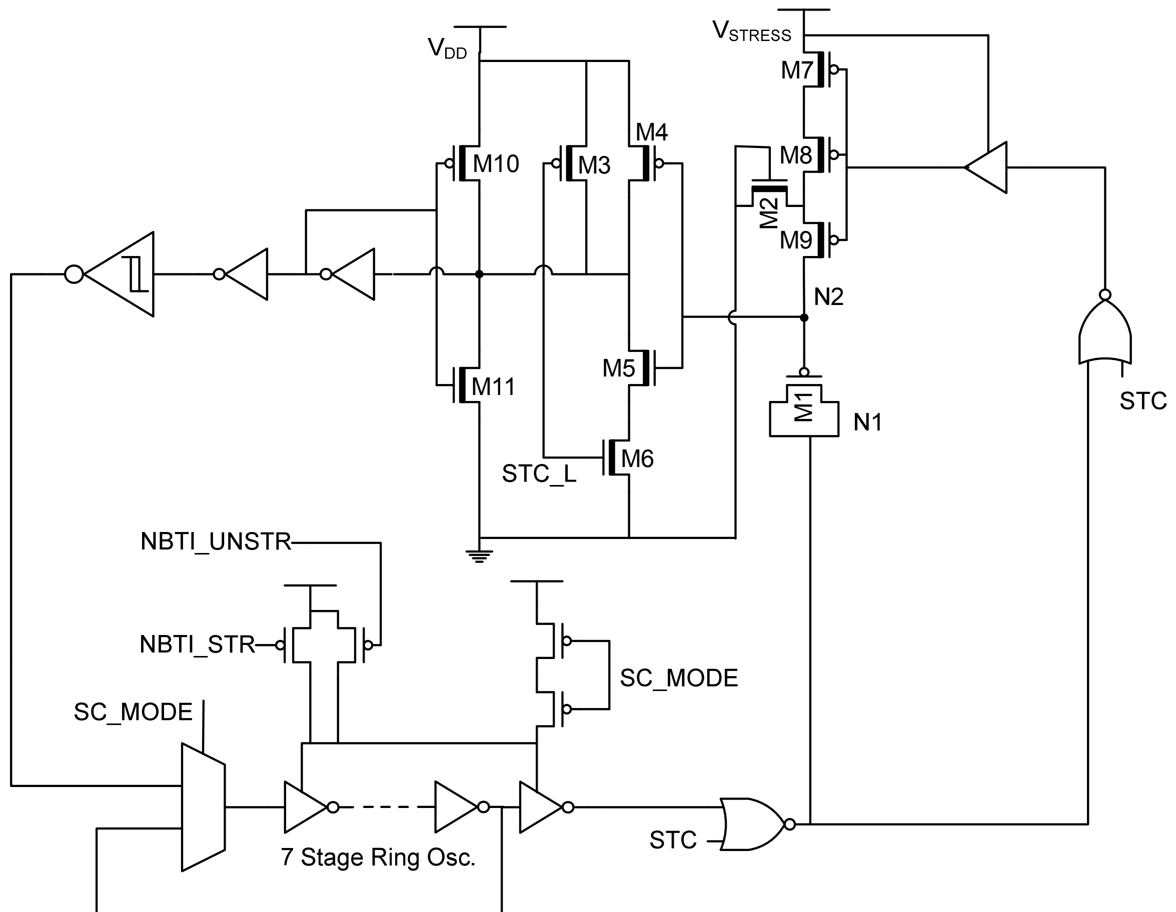


Figure 6.1 45nm oxide degradation sensor circuit diagram.

TABLE 6.2
45nm Oxide Degradation Sensor Sizing

Transistor	Width	Length	Device	Oxide
M1	0.32 µm	2.00 µm	NMOS	1.1V
M2	0.32 µm	3.00 µm	NMOS	1.8V
M3	0.40 µm	0.20 µm	PMOS	1.8V
M4	1.50 µm	0.15 µm	PMOS	1.8V
M5	0.50 µm	0.15 µm	PMOS	1.8V
M6	0.50 µm	0.15 µm	PMOS	1.8V
M7	0.32 µm	1.26 µm	PMOS	1.8V
M8	0.32 µm	1.26 µm	PMOS	1.8V
M9	0.32 µm	1.26 µm	PMOS	1.8V
M10	0.40 µm	0.35 µm	PMOS	1.8V
M11	0.40 µm	0.35 µm	NMOS	1.8V

An additional feature of the design is the thick oxide NAND Schmitt trigger circuit attached to the critical measurement node N2. In addition to the actual gate leakage through the device M1, the oscillation frequency is highly sensitive to Schmitt trigger switching points. During stress operation, the output of the NAND gate is forced high to minimize oxide stress across the thick oxide devices during accelerated test, and to also minimize the voltage stress across the PMOS devices. This is done to avoid excessive shift in threshold voltage due to NBTI during measurement and stress, which would have a significant impact upon the lower switch point of the Schmitt trigger. In the event of NBTI in device M3, the threshold voltage of M3 would increase, weakening the Ion current of this device, increasing the minimum input voltage required at the input of device M5 to initiate a transition on the output of the NAND gate. The extra step to reduce the NBTI stress on device M3 to prevent a time-dependent critical switching point on the Schmitt trigger NAND gate will allow reliable measurement of gate leakage increase when taking measurements over time.

During the design of the 45nm sensor, the 45nm PSP simulation model revealed a significant difference in gate leakage current between a device which had formed an inversion channel and a device that has no inversion channel. A MOSFET with an inversion channel exhibits several orders of magnitude greater gate leakage using the PSP simulation model in this 45nm process technology. This difference was not present in the BSIM4 based model used for the 130nm design. This has significant implications on the design of the measurement circuitry, since current flow

differs significantly depending upon the direction through the oxide device under test, M1.

The measurement functionality of the sensor in Figure 6.1 relies upon the gate leakage of a single transistor M1, in a single direction. When node N2 is high, the device under test leaks the charge to node N1 which is held to ground by the output of a NOR gate. However, unlike the sensor design in Figure 5.2, the charging of node N2 is performed by the triple stack of PMOS devices, M7-M9, controlled by an additional NOR gate connected to N1 and STC, the stress control voltage. The result of this design, is that only discharging N2 is performed by gate leakage, when an inversion channel is created in the oxide layer. This ensures that if a breakdown region forms at any point in the transistor gate, near the overlap regions or in the center of the channel, the gate leakage current in this state will include the breakdown path leakage. In the reverse direction, when the gate is at a lower potential than the drain and source, the channel is not formed, and breakdown paths to the channel may not impact the leakage current in this case, resulting in a poor measurement operation with a hard upper limit on the oscillation frequency change despite the damage caused to the oxide. The upper limit is defined by a breakdown region away from the overlap regions, such that only the discharge of node N2 is affected and the rising transition charging N2 in reverse would limit the oscillation frequency to a 2X increase as the falling phase time shrinks to insignificance. By discharging node N2 through device M1 with an inverted channel, the new sensor design avoids any limitation on the increase in oscillation frequency.

The design of the charging and discharging paths of node N2 in the sensor design shown in Figure 6.1 requires several tradeoffs. Devices M7-M9 are sized (detailed in Table 6.2) to minimize the subthreshold leakage current to node N2 during the discharge phase when gate leakage to node N1 is removing charge. This required very long channel devices and NMOS device M2 to reduce the undesired leakage current from preventing proper oscillation in monte carlo simulations considering all sources of process variation. Long channel, thick oxide devices consume significant area, so every effort to minimize the M7-M9 was made. The leaking device M1 is minimum width to reduce the coupling capacitance between the weakly driven node N2 and N1. This reduction in device width of M1 slows the oscillator down and places a constraint upon minimizing the leakage through the M7-M9 PMOS stack to ensure that the oscillation frequency properly tracks gate leakage change.

Included in the combined sensor is an NBTI sensor [34] attached to a 7 stage ring oscillator in the path of the oxide degradation oscillator. The premise of this circuitry is current-starving the 7 stage ring oscillator using a differential pair of PMOS devices and measuring the frequency of the ring oscillator. One of the devices is stressed and the other remains unstressed, allowing differential measurements of oscillator frequency with the PMOS header devices biased in subthreshold mode to detect any change in the threshold voltage of the stressed device. Subthreshold bias is used to maximize the sensitivity of the oscillator to threshold shifts. The PMOS header circuitry and oscillator is simplified in this diagram, yet the oxide degradation sensor uses double stacked PMOS devices to

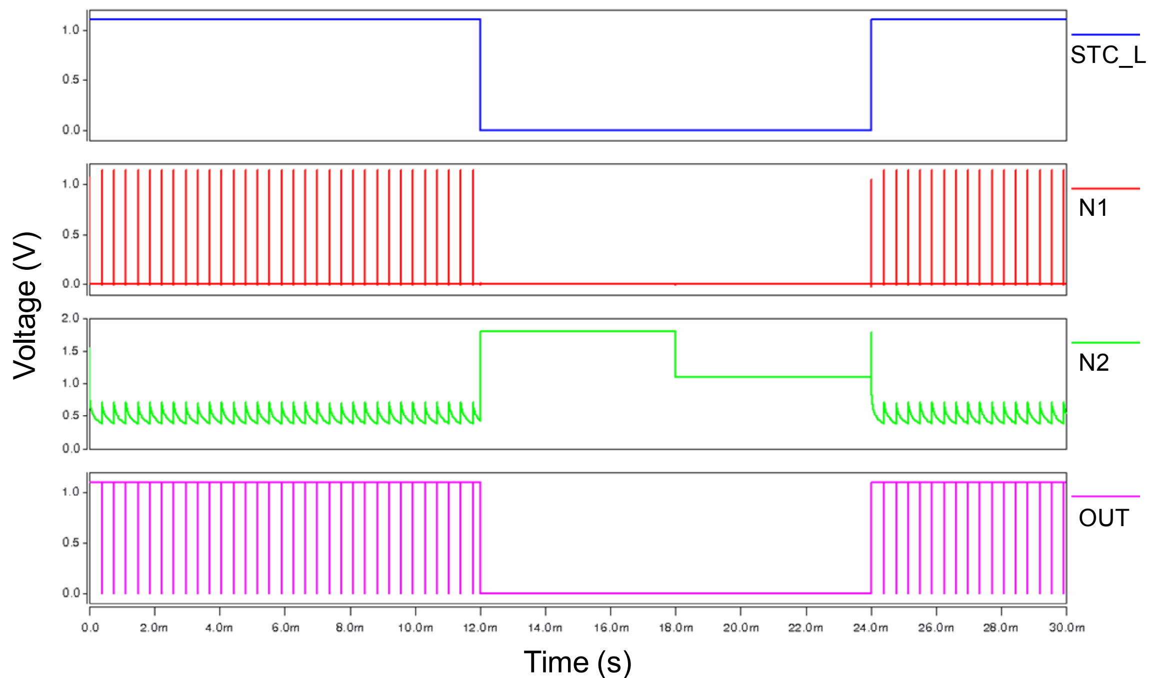


Figure 6.2 Simulated operation of 2nd generation 45nm oxide degradation sensor. Measurement operation from 0-12ms, then stress phase with 2 voltages from 12-24ms.

provide current to the oscillator during oxide degradation measurements. Simulation results of the NBTI portion of this sensor show proper operation of the NBTI circuitry despite sharing the ring oscillator circuitry in both portions of the design.

Figure 6.2 shows simulated waveforms of the modified sensor design. From time 0 to 12ms, the sensor is in measurement mode and the prominent discharge delay at node N2 sets the oscillation frequency of the sensor. The ultimate output of the sensor has changed from an even duty cycle clock-like signal to a short 10ns pulse when the discharge completes. At 12ms the STC_L signal is set low and the STC signal is set high to activate the high voltage stress phase. For 6ms, a voltage of 1.8V is driven to node N2, and then the voltage is reduced to 1.1V. At 24ms, the simulation returns to measurement operation.

6.2 45nm Testchip Implementation

A multi-project 45nm, 7-metal testchip was designed including 256 combined NBTI and oxide degradation sensors and is currently being fabricated through CMP. The nominal supply voltage is 1.1V and the thick oxide supply voltage for pads and voltage stress circuitry is 1.8V. The testing methodology is identical to the testing

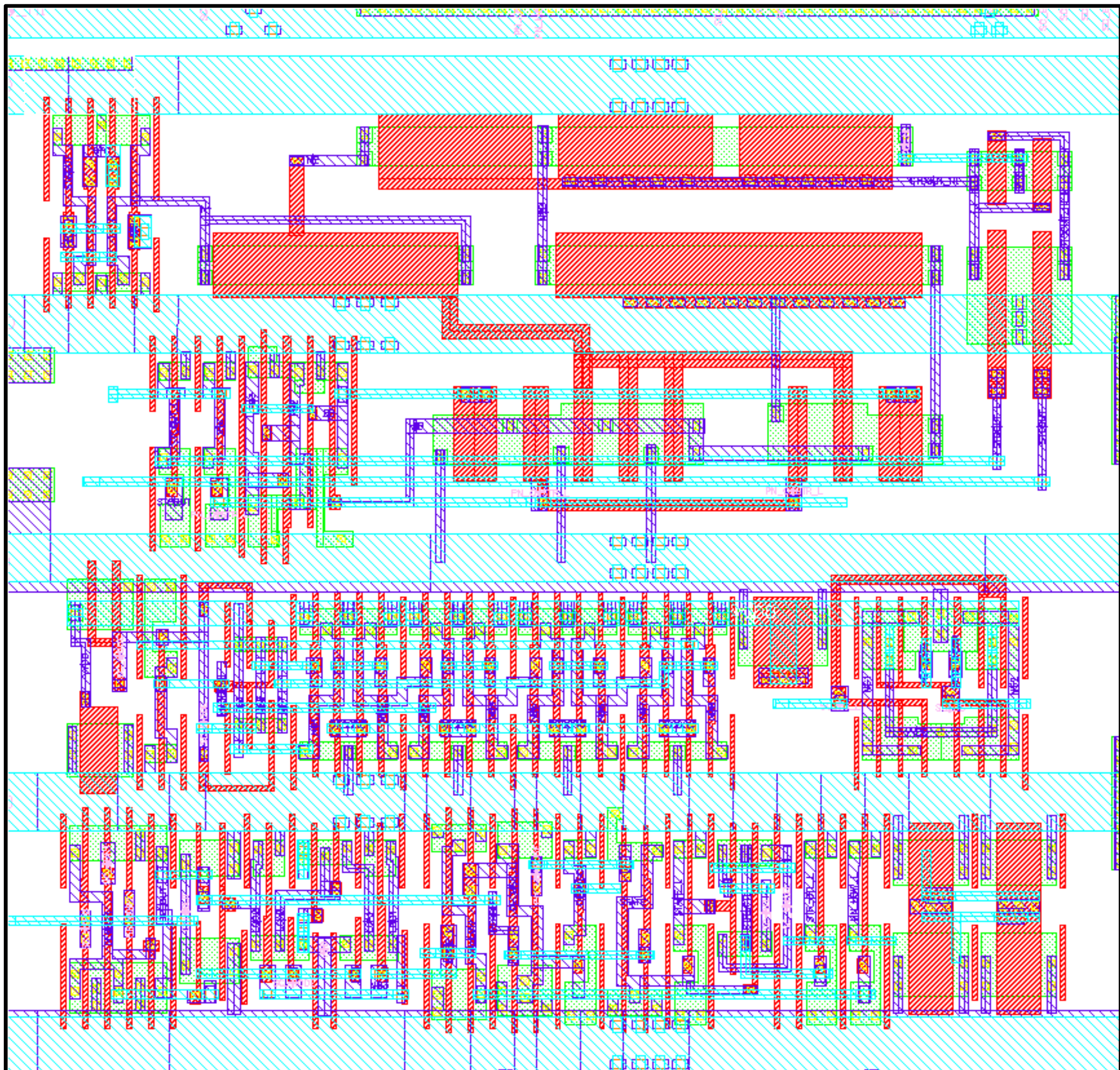


Figure 6.3 45nm oxide degradation sensor cell layout. Cell area includes NBTI sensor and oxide sensor in 77.35 μm^2 (8.5μm x 9.1μm).

setup used in the 130nm testchip presented in Chapter 5. A 20 bit counter is used to accumulate the oscillator frequency output from a shared output bus with 16 oxide degradation sensors attached using tri-state output drivers activated by a 4 bit address. Four 20-bit storage registers are connected in a LIFO queue configuration (last-in, first-out). Data is collected from the storage registers via a 1,440 bit scan chain that spans 16 blocks, allowing concurrent measurement of 16 different sensors.

Figure 6.3 shows the implementation of the cell in the configuration that would be used in a typical system, without the high voltage multiplexing circuitry for accelerated testing. The cell layout is compatible with the standard cell pitches in the 45nm technology and has power and ground stripes alternating through the design in metal 1 for easy integration with a standard cell flow. The thick oxide devices for the oxide degradation sensor are clustered in the upper right corner of Figure 6.3, with the NBTI header devices in the lower right corner, and the oscillator and control circuitry spread throughout the middle and lower left portions of the design.

On the 130nm testchip, the NBTI sensor used 308 μm^2 and the oxide degradation sensor was 150 μm^2. This version of combined NBTI/oxide degradation sensor is 8.5 μm x 9.1 μm, consuming 77.35 μm^2. Design rule limitations when using the thick oxide OD18 devices in close proximity to standard devices and the requirements of dummy polysilicon material at the edge of each minimum channel length device prevented any further area savings through combining the design.

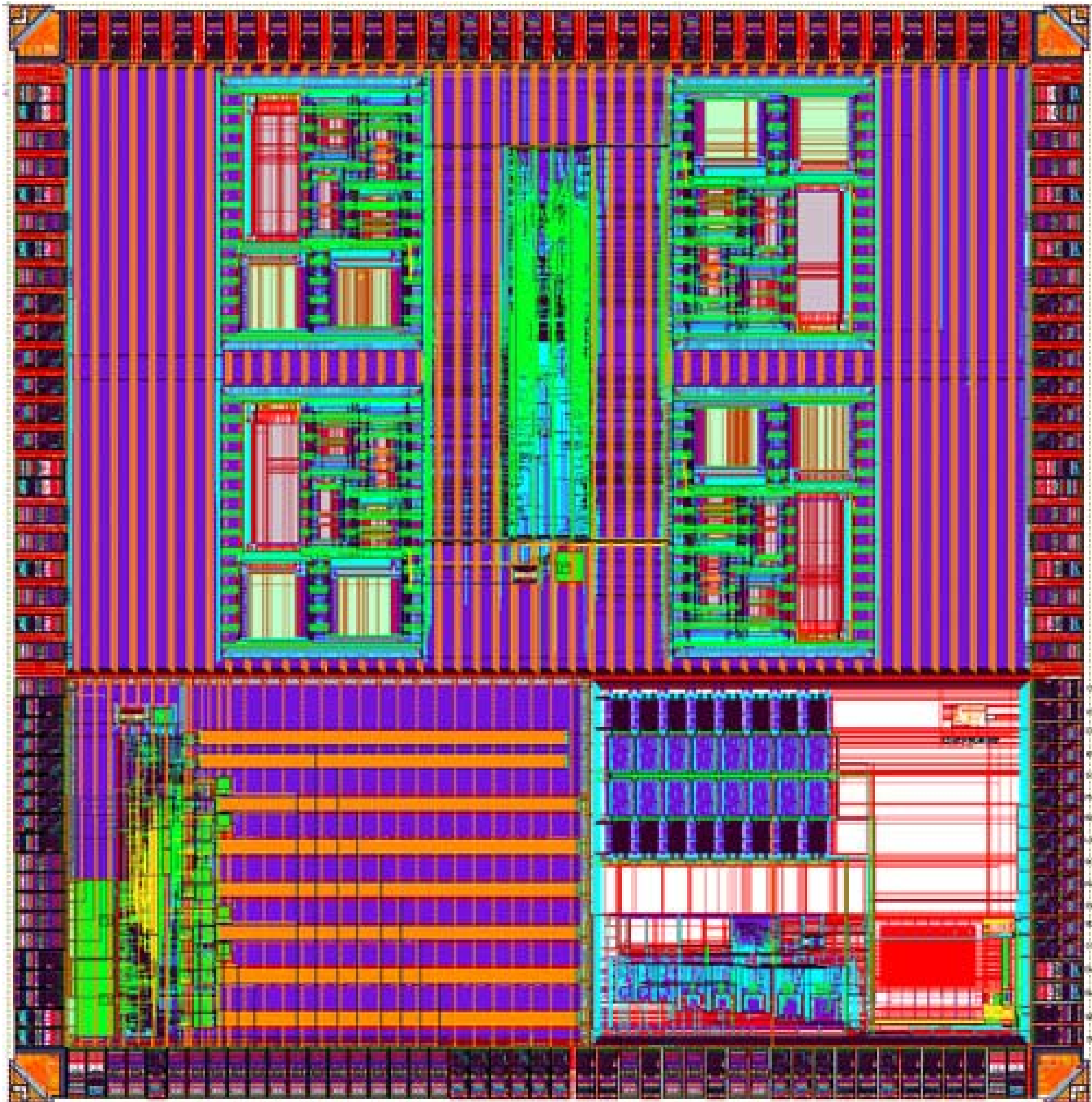


Figure 6.4 45nm multi-project testchip including 2nd generation oxide and NBTI degradation sensors.

Figure 6.4 shows the 5mm^2 multi-project testchip containing 256 oxide and NBTI degradation monitoring sensors and the related control and testing circuitry. The 16 bank array of sensors is 469.76 μm x 342.40 μm, or 160,845.8 μm^2. The NBTI and oxide sensor block uses only 5 pads beyond the typical power, ground, and scan control signals shared amongst the entire testchip.

6.3 Simulated Results and Discussion

Simulated results with extracted parasitics are presented in this section due to the unavailability of the silicon measurements while waiting for fabrication to be completed. The simulated power consumption of the combined sensor is 25.2 μW during measurement of oxide degradation and 4.2 nW during 1.1V oxide stress. This is a 111,785X reduction in oxide stressing power consumption compared to the 130nm design in Chapter 5, with results of 469.5 μW for oxide stress using the differential amplifier. The stress mode power consumption is a critical figure, since the majority of time will be spent in stress mode under normal operation. The modest increase from 14.03μW to 25.2μW during measurement mode is due to the addition of NBTI circuitry, and the larger 7 stage ring oscillator. This increase is offset by the higher frequency of the oscillator, from 100 mHz to 2.5kHz, allowing measurements to be made more rapidly for a given level of accuracy, allowing actual savings in energy consumption.

The most critical aspect of the new design is to analyze the ability of the modified sensor to track gate leakage increase to an oscillator frequency increase. Figure 6.5 shows promising simulation results relating gate leakage increase in device M1 from Figure 6.1 to an output frequency increase. Up to a 100X increase in gate leakage through the device corresponds to a linear 100X increase in oscillation frequency, as shown by the tight fit to the ideal 1:1 scaling line included on the plot. Results are shown for 85ºC operation, but have been confirmed across

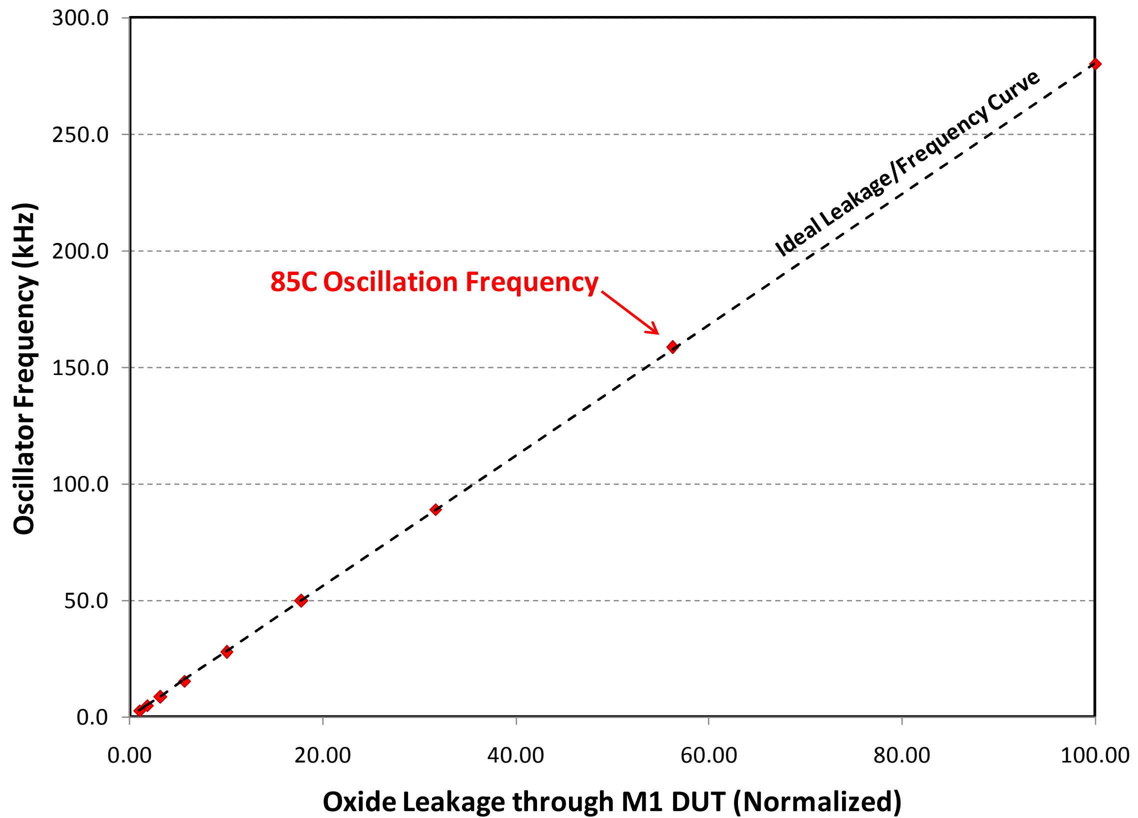


Figure 6.5 Oscillation frequency vs. oxide leakage scaling for extracted parasitic simulation. Dashed line represents the ideal 1:1 scaling between leakage and frequency.

25º-125ºC operation, and considering process variation in a set of monte carlo simulations.

Figure 6.6 shows the scaling of oxide sensor frequency output as temperature is scaled. At higher temperatures, the oscillator frequency increases from 1.52kHz at 25C to 3.86 kHz at 125C. The peak-to-peak voltage swing at node N2 in Figure 6.1 is also plotted in Figure 6.6, showing the primary cause of the increase in oscillator frequency. As the temperature increases the switching points of the NAND Schmitt trigger gate connected to the critical node converge to result in less time spent discharging node N2. This result indicates that it will be critical to utilize an unstressed oxide sensor to act as a cheap temperature sensor as described in [34],

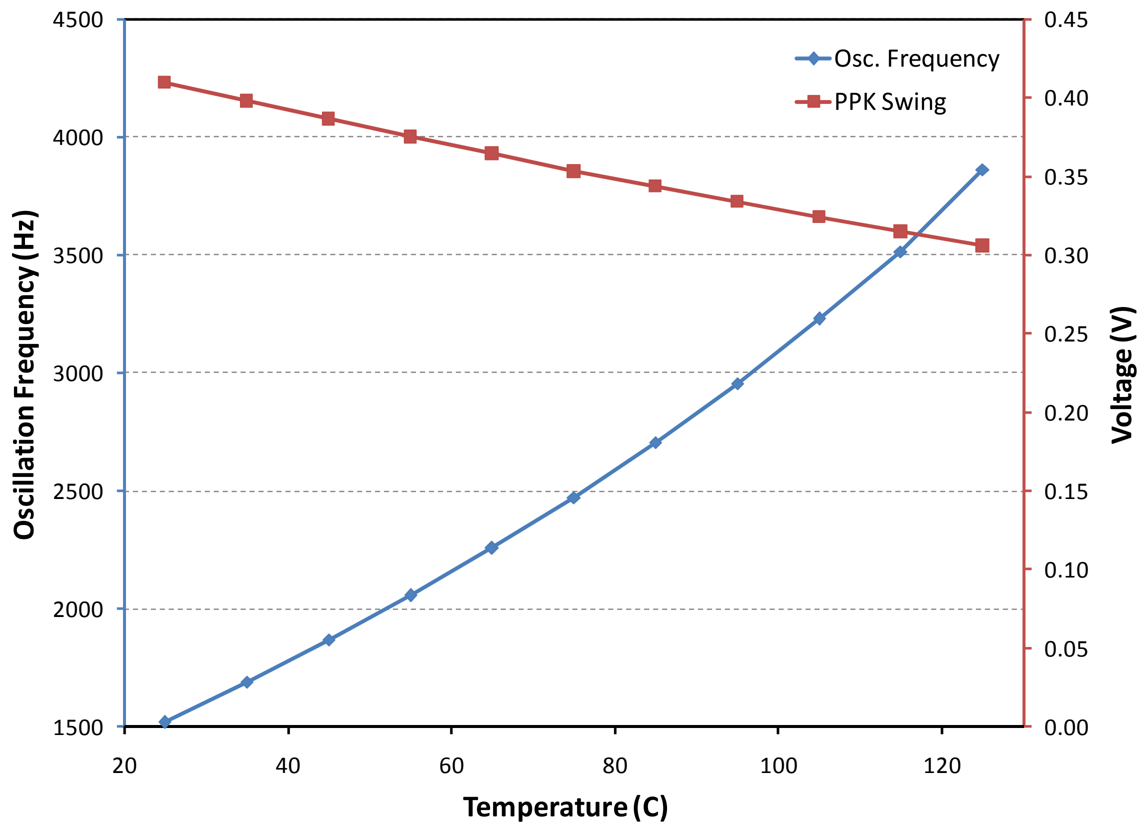


Figure 6.6 45nm 2nd generation oxide degradation sensor temperature scaling. Peak to peak swing on node N2 is plotted against oscillator frequency.

to allow a curve-fit model to eliminate the temperature dependence. This allows a series of measurements to be made accurately at different temperatures.

Figure 6.7 shows the results of a monte carlo simulation of the measurement phase of the oxide degradation sensor. The lower voltage switching point is used as the X-axis, since this voltage is a critical factor influencing the delay of the gate leakage discharging node N2. When this voltage is reached, the PMOS triple stack is activated to charge N2 and begin another discharge cycle. Any variation in this switch point will affect the output frequency. As seen in Figure 6.7, there is a slight positive correlation between the Schmitt trigger lower switch point and the oscillation frequency. Since the oxide degradation sensor is primarily used to measure the

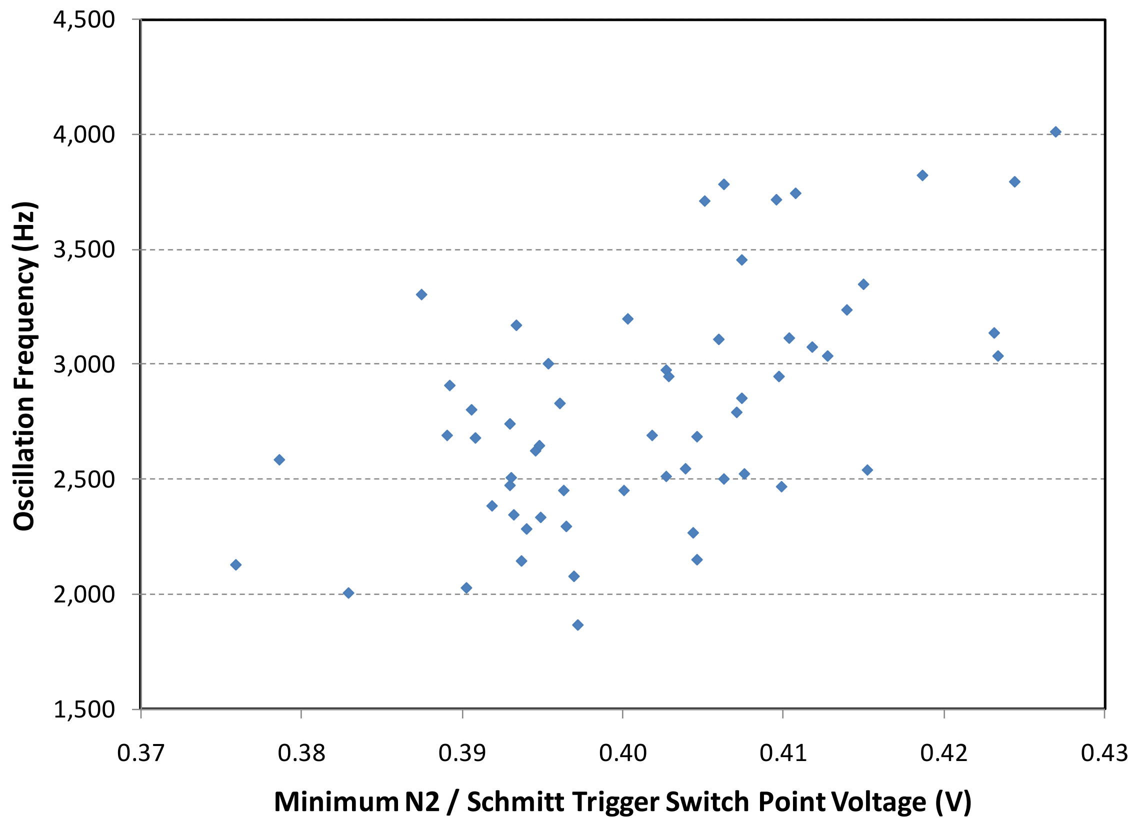


Figure 6.7 Monte carlo simulation of 45nm oxide degradation sensor design. Scatter plot of Schmitt trigger lower switch point vs. oscillation frequency.

change in time of a given sensor, this is not a critical concern, as long as the switch point does not change over time. This figure also reveals that the oxide degradation sensor is not an optimal way to measure gate oxide leakage in absolute terms, due to the potential for process variation affecting the switching point of the Schmitt trigger.

6.4 Summary

The 2nd generation design of the oxide degradation sensor improves upon the original oxide sensor design by eliminating the need for a high voltage supply and removing the potential for uneven oxide stress. The stress mode power

consumption was reduced by over 111,000X while combining an NBTI and oxide sensor into a single cell. The new combined sensor design consumes the area equivalent to 54.86 NAND3 cells (77.35 mm2), while the 130nm combined sensor area is 63.61 NAND3 cells (458.64 mm2), a reduction of 13.7% in process-adjusted area. The 2nd generation oxide sensor exhibits ideal 1:1 gate leakage to frequency behavior in extracted simulations and uses a new method of charging the critical gate leakage node that may remove a measurement limitation discovered in simulations using PSP models. The redesigned oxide degradation sensor is significantly improved in the key metrics of power consumption, area and accuracy.

Chapter 7

Conclusion

The non-scaling of supply voltage in integrated circuits due to the desire to maintain saturation current without increasing off-state leakage has lead to increasing electric fields across oxides and junctions with each successive generation of process technology. Increased electric fields increase the incidence of wearout and degradation from reliability mechanisms such as oxide breakdown, negative bias temperature instability and electromigration. As technology scaling continues without reducing supply voltage significantly, reliability qualification, modeling and mitigation techniques have increased importance as reliability-related timing and voltage margins are traded for increased performance.

Increasing process variation and wide ranges of environmental stress factors experienced by mobile electronics lead to conditions that are statistical outliers, but add significant margin to timing or voltage through worst-case design methodologies. Adaptive and dynamic designs that monitor conditions in real-time or attempt to maximize performance or minimize power in real-time can offset the negative impact of excessive margining.

In Chapter 2, a novel method for detecting read path timing errors in on-chip SRAM was presented with 180nm simulations and a 130nm implementation. By lowering the supply voltage until the actual memory under test experiences timing related failures, voltage margin is eliminated. This simultaneously reduces dynamic and leakage power and improves reliability lifetime of the entire memory through the diminished electric field and lower temperature in the memory circuits. Simulation results indicate the potential to reduce supply voltage from 1.8V to 1.55V with no errors for typical performance designs, eliminating the penalty of worst-case design methodology and saving up to 12% dynamic power and 23% leakage power.

A multi-mechanism modeling and management system was proposed in Chapter 3, modeling oxide breakdown, NBTI, thermal cycling and electromigration in real-time. The failure mechanism based modeling system calculated a probabilistic system lifetime based upon historical and current reliability stress factors, such as voltage and temperature. A PID controller modifies the limits of a dynamic voltage scaling system to maximize performance within a lifetime reliability budget. This system allows user controlled lifetime, ensures preventative DVS limits for worst-case dies and provides improvement in the peak performance possible by relaxing upper limits on voltage assignment when dies are significantly below the worst-case degradation trajectory.

In Chapter 4, an analysis of the factors influencing oxide breakdown and the utility of sensor arrays in projecting inherently probabilistic degradation mechanisms. The analysis confirmed that voltage reduction is the method of choice for reducing the incidence of oxide breakdown, and that inherent randomness in defect

generation and placement requires large quantities of ideal sensors to make an accurate projection of oxide lifetime. Minor spatial correlation was discovered between successive breakdown events in monte carlo simulation, indicating some value to a dispersed sensor network on-chip. Sample distribution fitting and projections indicate that 1000 or more sensors on chip will be needed to deliver a reasonable lifetime projection considering inherent randomness.

Chapter 5 presents the design of a novel, compact sensor for oxide degradation, aimed at detecting progressive soft breakdown events over time. The oxide sensor is standard cell compatible and requires the area of ~21 NAND3 gates in the 130nm technology in which it was implemented. Measured results from a 130nm testchip show 14-35% increase in gate oxide leakage following 50 hours of accelerated stress testing at 2.5V and 130ºC.

Chapter 6 details a significant redesign of the oxide degradation sensor from Chapter 5 and the inclusion of NBTI monitoring circuitry in a unified sensor cell. The unified sensor is implemented in 45 nm technology and requires the area of ~54.8 NAND3 cells (77.35 μm^2). The stress mode power consumption of the unified sensor was reduced to 4.2 nW, a 111,000X improvement over the previous design. Simulation results with extracted parasitics show an ideal 1:1 correspondence between gate leakage increase and sensor output frequency.

The techniques presented throughout chapters 2-6 form the framework for a truly dynamic, self-diagnosing architecture, like ElastIC. Dynamic reliability modeling in real-time using accurate measurement sensors as proposed in Chapter 5 and 6 presents an opportunity for user or application control over the reliability-

performance tradeoff. In many cases, where electronic products will be replaced frequently, DRM systems may allow the user to set system performance targets unattainable under the restrictive reliability qualification regime of today's electronic devices. If a user knows his system will be replaced every 18 months, the user can consume the system's reliability "budget" at a higher rate, obtaining vastly increased performance, as suggested by peak performance improvement modeling in Chapter 3.

In many applications demanding high-reliability in medical, transportation, or military systems, setting extremely conservative system lifetime targets using DRM would allow the use of a general purpose integrated circuit, that may have required a reliability oriented ASIC in the past. In the end, many consumers prefer reliability and resiliency over all, and DRM at the minimum allows graceful aging of semiconductor devices. As degradation occurs on-chip, voltage and frequency pairings can gradually adjust to maintain operation at slightly lower levels of performance. This avoids the worst scenario of all, a completely dead integrated circuit that provides no functionality.

7.1 Future Work

Significant challenges remain to be solved in order to realize adaptive, failure-mechanism-aware system architectures, such as ElastIC [25]. Additional validation of the newly developed sensors in Chapter 5 and 6 with actual failing circuits is necessary to prove their utility and convince system designers and architects that a truly self-aware, self-diagnosing system can become reality.

The control algorithm for DRM presented in Chapter 3 needs to grow in scope to actuate more than voltage limits in a DVS system, but also simultaneously handle voltage assignments and scheduling for multiprocessing workloads, assigning enforced healing periods for degraded circuits, forced sleep state, activity reduction limiting and other methods of minimizing reliability degradation. The modeling work needs to be updated to include input from the various reliability sensors developed in this work and other research groups currently.

Statistically aware reliability mechanism modeling needs to take place for mechanisms of interest. Much of the analytical work in this thesis is conducted on oxide breakdown models due to the maturity of the literature about this phenomenon and the availability of a simulation model that considers the probabilistic nature of the degradation and ultimate failure. Statistical data needs to be collected and modeled for NBTI/PBTI, electromigration and hot carrier effects. Probabilistic modeling of these effects will also significantly improve the reliability qualification method concerning these degradation mechanisms.

Bibliography

[1] *ITRS 2006 Update Report*. [cited 2008 2/24]; Available from: http://www.itrs.net/Links/2006Update/2006UpdateFinal.htm.

[2] Taur, Y., "CMOS Scaling and Issues in Sub-0.25um Systems", in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, Editor. 2001, IEEE Press: Piscataway, NJ. p. 27-28.

[3] Rabaey, J., A. Chandrakasan, and B. Nikolic, *Digital Integrated Circuits: A Design Perspective*. 2nd ed. 2003, Upper Saddle River, NJ: Pearson Education, Inc.

[4] Taur, Y. and T.H. Ning, *Fundamentals of Modern VLSI Devices*. 1st ed. 1998, Cambridge, UK: Cambridge University Press.

[5] Borkar, S., "Design challenges of technology scaling"*. Micro, IEEE*, 1999. **19**(4): p. 23-29.

[6] Stathis, J.H. "Physical and predictive models of ultra thin oxide reliability in CMOS devices and circuits", in *Reliability Physics Symposium, 2001. Proceedings. 39th Annual. 2001 IEEE International*, 2001, p. 132-149.

[7] Black, J.R., "Electromigration failure modes for aluminum metallization in semiconductor devices"*. Proceedings of the IEEE*, 1969. **57**(9): p. 1587-1593.

[8] Alam, M.A. and S. Mahapatra, "A comprehensive model of PMOS NBTI degradation"*. Microelectronics Reliability*, 2005. **45**(1): p. 71-81.

[9] Chen, I.C., et al., "Substrate hole current and oxide breakdown"*. Applied Physics Letters*, 1986. **49**(11): p. 669-671.

[10] McPherson, J.W. and R.B. Khamankar, "Molecular model for intrinsic time-dependent dielectric breakdown in SiO2 dielectrics and the reliability implications for hyper-thin gate oxide"*. Semiconductor Science and Technology*, 2000. **15**(5): p. 462-470.

[11] Vattikonda, R., W. Wenping, and C. Yu. "Modeling and minimization of PMOS NBTI effect for robust nanometer design", in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, p. 1047-1052.

[12] *Stress-Test-Driven Qualification of Integrated Circuits: JESD47E*, J.S.S.T. Assocation, Editor. 2007: JEDEC Standard.

[13] Borkar, S., "Designing reliable systems from unreliable components: the challenges of transistor variability and degradation". *Micro, IEEE*, 2005. **25**(6): p. 10-16.

[14] Chen, T.-C. "Where CMOS is going: trendy hype vs. real technology", in *Solid-State Circuits, 2006 IEEE International Conference Digest of Technical Papers*, 2006, p. 1-18.

[15] *Failure-Mechanism-Driven Reliability Qualification of Silicon Devices*, J.S.S.T. Assocation, Editor. 1993: JEDEC Standard.

[16] Martin, S.M., et al. "Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads", in *Computer Aided Design, 2002. ICCAD 2002. IEEE/ACM International Conference on*, 2002, p. 721-725.

[17] Nowka, K.J., et al., "A 32-bit PowerPC system-on-a-chip with support for dynamic voltage scaling and dynamic frequency scaling". *Solid-State Circuits, IEEE Journal of*, 2002. **37**(11): p. 1441-1447.

[18] Burd, T., et al. "A dynamic voltage scaled microprocessor system", in *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*, 2000, p. 294-295, 466.

[19] Ernst, D., et al. "Razor: a low-power pipeline based on circuit-level timing speculation", in *Microarchitecture, 2003. MICRO-36. Proceedings. 36th Annual IEEE/ACM International Symposium on*, 2003, p. 7-18.

[20] Austin, T., et al., "Making typical silicon matter with Razor". *Computer*, 2004. **37**(3): p. 57-65.

[21] Blaauw, D., et al. "Razor II: In Situ Error Detection and Correction for PVT and SER Tolerance", in *International Solid State Circuits Conference*, 2008, p. 400-401.

[22] Bowman, K., et al. "Energy-Efficient and Metastability-Immune Timing Error Detection and Instruction-Replay-Based Recovery Circuits for Dynamic-Variation Tolerance", in *IEEE International Solid State Circuits Conference*, 2008, p. 402-403.

[23] Flautner, K., et al. "Drowsy caches: simple techniques for reducing leakage power", in *Computer Architecture, 2002. Proceedings. 29th Annual International Symposium on*, 2002, p. 148-157.

[24] Karl, E., D. Sylvester, and D. Blaauw. "Timing error correction techniques for voltage-scalable on-chip memories", in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, 2005, p. 3563-3566 Vol. 4.

[25] Sylvester, D., D. Blaauw, and E. Karl, "ElastIC: An Adaptive Self-Healing Architecture for Unpredictable Silicon"*. Design & Test of Computers, IEEE*, 2006. **23**(6): p. 484-490.

[26] Srinivasan, J., et al. "The case for lifetime reliability-aware microprocessors", in *Computer Architecture, 2004. Proceedings. 31st Annual International Symposium on*, 2004, p. 276-287.

[27] Lu, Z., et al., "Improved thermal management with reliability banking"*. Micro, IEEE*, 2005. **25**(6): p. 40-49.

[28] McGowen, R., et al., "Power and temperature control on a 90-nm Itanium family processor"*. Solid-State Circuits, IEEE Journal of*, 2006. **41**(1): p. 229-237.

[29] Karl, E., et al. "Reliability modeling and management in dynamic microprocessor-based systems", in *Design Automation Conference, 2006 43rd ACM/IEEE*, 2006, p. 1057-1060.

[30] Karl, E., et al., "Multi-Mechanism Reliability Modeling and Management in Dynamic Systems"*. IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2008 (to appear).

[31] Kim, T.-H., R. Persaud, and C.H. Kim. "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits", in *VLSI Circuits, 2007 IEEE Symposium on*, 2007, p. 122-123.

[32] Keane, J., T.-H. Kim, and C.-H. Kim. "An On-Chip NBTI Sensor for Measuring PMOS Threshold Voltage Degradation", in *International Symposium on Low-Power Electronic Design*, 2007, p. 189-194.

[33] Karl, E., D. Blaauw, and D. Sylvester. "Analysis of System-Level Reliability Factors and Implications on Real-time Monitoring Methods for Oxide Breakdown Device Failures", in *International Symposium on Quality Electronic Design*, 2008 (to appear), p.

[34] Karl, E., et al. "Compact In-Situ Sensors for Monitoring Negative-Bias-Temperature-Instability Effect and Oxide Degradation", in *International Solid State Circuits Conference*, 2008, p. 410-411.

[35] Krishnamurthy, R.K., et al. "High-performance, low-power, and leakage-tolerance challenges for sub-70nm microprocessor circuits", in *Solid-State Circuits Conference, 2002. ESSCIRC 2002. Proceedings of the 28th European*, 2002, p. 315-321.

[36] Bhavnagarwala, A.J., T. Xinghai, and J.D. Meindl, "The impact of intrinsic device fluctuations on CMOS SRAM cell stability". *Solid-State Circuits, IEEE Journal of*, 2001. **36**(4): p. 658-665.

[37] Chrisanthopoulos, A., et al., "Comparative study of different current mode sense amplifiers in submicron CMOS technology". *Circuits, Devices and Systems, IEE Proceedings -*, 2002. **149**(3): p. 154-158.

[38] Bhavnagarwala, A.J., S. Kosonocky, and J.D. Meindl. "Interconnect-centric array architectures for minimum SRAM access time", in *Computer Design, 2001. ICCD 2001. Proceedings. 2001 International Conference on*, 2001, p. 400-405.

[39] Agawa, K., et al., "A bitline leakage compensation scheme for low-voltage SRAMs". *Solid-State Circuits, IEEE Journal of*, 2001. **36**(5): p. 726-734.

[40] Austin, T., E. Larson, and D. Ernst, "SimpleScalar: an infrastructure for computer system modeling". *Computer*, 2002. **35**(2): p. 59-67.

[41] Sherwood, T., et al., *Automatically characterizing large scale program behavior*, in *Proceedings of the 10th international conference on Architectural support for programming languages and operating systems*. 2002, ACM: San Jose, California.

[42] Bhavnagarwala, A.J., et al. "A pico-joule class, 1 GHz, 32 KByte/spl times/64 b DSP SRAM with self reverse bias", in *VLSI Circuits, 2003. Digest of Technical Papers. 2003 Symposium on*, 2003, p. 251-252.

[43] Miner, M.A., "Cummulative damage in fatigue". *Journal of Applied Mechanics*, 1945. **67**: p. A159-A164.

[44] Degraeve, R., et al. "A consistent model for the thickness dependence of intrinsic breakdown in ultra-thin oxides", in *Electron Devices Meeting, 1995., International*, 1995, p. 863-866.

[45] Cao, K.M., et al. "BSIM4 gate leakage model including source-drain partition", in *Electron Devices Meeting, 2000. IEDM Technical Digest. International*, 2000, p. 815-818.

[46] Casu, M.R., et al., "An electromigration and thermal model of power wires for a priori high-level reliability prediction"*. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 2004. **12**(4): p. 349-358.

[47] Maiz, J.A. "Characterization of electromigration under bidirectional (BC) and pulsed unidirectional (PDC) currents", in *Reliability Physics Symposium, 1989. 27th Annual Proceedings., International*, 1989, p. 220-228.

[48] Tao, J., N.W. Cheung, and H. Chenming, "An electromigration failure model for interconnects under pulsed and bidirectional current stressing"*. Electron Devices, IEEE Transactions on*, 1994. **41**(4): p. 539-545.

[49] Tao, J., N.W. Cheung, and H. Chenming, "Modeling electromigration lifetime under bidirectional current stress"*. Electron Device Letters, IEEE*, 1995. **16**(11): p. 476-478.

[50] Blish, R.C., III. "Temperature cycling and thermal shock failure rate modeling", in *Reliability Physics Symposium, 1997. 35th Annual Proceedings., IEEE International*, 1997, p. 110-117.

[51] Nelson, W., *Accelerated Testing: Statistical Models, Test Plans and Data Analyses*. 1990, New York: Wiley.

[52] Ershov, M., et al. "Transient effects and characterization methodology of negative bias temperature instability in pMOS transistors", in *Reliability Physics Symposium Proceedings, 2003. 41st Annual. 2003 IEEE International*, 2003, p. 606-607.

[53] Rangan, S., N. Mielke, and E.C.C. Yeh. "Universal recovery behavior of negative bias temperature instability [PMOSFETs]", in *Electron Devices Meeting, 2003. IEDM '03 Technical Digest. IEEE International*, 2003, p. 14.3.1-14.3.4.

[54] Brooks, D., V. Tiwari, and M. Martonosi. "Wattch: a framework for architectural-level power analysis and optimizations", in *Computer Architecture, 2000. Proceedings of the 27th International Symposium on*, 2000, p. 83-94.

[55] Skadron, K., et al. "HotSpot: Techniques for Modeling Thermal Effects at the Processor-Architecture Level", in *Proceedings of the 2002 International Workshop on THERMal Investigations of ICs and Systems (THERMINIC)*, 2002, p. 169-172.

[56] Shigemasa, T., M. Yukitomo, and R. Kuwata. "A model-driven PID control system and its case studies", in *Control Applications, 2002. Proceedings of the 2002 International Conference on*, 2002, p. 571-576 vol.1.

[57] Kiam Heong, A., G. Chong, and L. Yun, "PID control system analysis, design, and technology"*. Control Systems Technology, IEEE Transactions on*, 2005. **13**(4): p. 559-576.

[58] Kumar, S.V., C.H. Kim, and S.S. Sapatnekar. "An Analytical Model for Negative Bias Temperature Instability", in *Computer-Aided Design, 2006. ICCAD '06. IEEE/ACM International Conference on*, 2006, p. 493-496.

[59] Hu, C., et al. *BSIM4 Homepage*. 2008 [cited 2008 2/24/2008]; Available from: http://www-device.eecs.berkeley.edu/~bsim3/bsim4.html.

[60] Stathis, J.H. "Gate Oxide Reliability for Nano-Scale CMOS", in *Microelectronics, 2006 25th International Conference on*, 2006, p. 78-83.

[61] Wu, E.Y., D.L. Harmon, and H. Liang-Kai, "Interrelationship of voltage and temperature dependence of oxide breakdown for ultrathin oxides"*. Electron Device Letters, IEEE*, 2000. **21**(7): p. 362-364.

[62] Borkar, S., P. Dubey, and K. Kahn. *Platform 2015: Intel Processor and Platform Evolution for the Next Decade*. 2005 [cited 2007 Nov. 17th]; Available from: http://www.intel.com/technology/magazine/computing/Platform-2015-0305.pdf.

[63] Denais, M., et al. "On-the-fly characterization of NBTI in ultra-thin gate oxide PMOSFET's", in *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International*, 2004, p. 109-112.