On-chip NBTI and Gate-Oxide-Degradation Sensing and Dynamic Management in VLSI circuits

by

Prashant Singh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2011

Doctoral Committee:

Professor David Blaauw, Chair

Professor Dennis M. Sylvester

Associate Professor Jerome P. Lynch

Associate Professor Scott Mahlke

© Prashant Singh

2011

To Divine

and

My parents

Acknowledgements

First of all I can't thank my advisor Prof. David Blaauw enough for his guidance all through my graduate studies. He encouraged and helped me to go on even during the most difficult times. Thanks to Prof. Dennis Sylvester for co-advising me and for his invaluable suggestions all through these years. I would like to express my gratitude towards Prof. Scott Mahlke and Prof. Jerome Lynch for agreeing to be a part of my PhD defense committee.

There are several people who have directly or indirectly contributed to my research studies. Eric, thanks for helping me develop a strong footing in the reliability research. I would like to express gratitude towards the senior research students at Michigan, Sanjay, Visvesh, Carlos, who helped me at some point in time during the past five years. There were several people who helped me when I was going through very rough times. Zhiyoong, Matt, Scott and Mike, I can't thank you all enough for your help and patience.

I need to mention my friends at Michigan who made sure that I never had a dull moment here. Kaviraj, Manav, Shantanu, Ravikishore, Vivek, Rach, Animesh, Ayan, Ritesh – Thanks for the great times you all gave me here in Ann Arbor.

I am indebted to my parents for their support all this time while I was away from home. And lastly, no words can express my gratitude to the Lord, who looked after me as His son while I experienced one of the most important period of my life.

Table of Contents

Dedication ii

Acknowledgements iii

List of figures vii

List of Tables xii

Chapter

1. Introduction 1
   - 1.1. Reliability Management Techniques 2
   - 1.2. Contributions of this Thesis 6
2. Compact Degradation Sensor to Monitor Negative Bias Temperature Instability 7
   - 2.1. NBTI Degradation Mechanism 8
   - 2.2. Previous Measurement Techniques 10
   - 2.3. Proposed NBTI Sensor 10
     - 2.3.1. Circuit Design Principle 10
     - 2.3.2. Oscillator Modeling 13
     - 2.3.3. Calibration and Measurement Strategy 15
     - 2.3.4. Experimental Results 16
   - 2.4. Summary 26
3. A Unified 45nm NBTI and Oxide Degradation Sensor 28
   - 3.1. Gate-oxide Degradation Mechanism 28
   - 3.2. Previous Gate-oxide Degradation Measurement Techniques 30
   - 3.3. Proposed Unified Sensor Design and Operation 31
   - 3.4. Silicon Measurements from Sensors 34
   - 3.5. Summary 43

4. Dynamic NBTI Management (DNM) Using the 45nm Unified NBTI and Oxide Degradation Sensor 44
4.1. Concept 44
4.2. Number of Sensors 46
4.3. Reliability Model 48
4.4. Algorithm 49
4.5. Experimental Results 50
4.6. Conclusion 54
5. Early Detection of Oxide Breakdown Through *In situ* Degradation Sensing 55
5.1. Previous *In situ* OBD Detection Techniques 56
5.2. Proposed *In situ* OBD Detection Technique 56
5.3. Test Chip and Silicon Measurements 61
5.4. More Silicon Measurements 73
5.5. Summary 77
6. *In situ* Bias Temperature Instability (BTI) Sensing Technique and Dynamic BTI Management Implementation 78
6.1. Introduction 78
6.2. Concept 79
6.2.1. Threshold Voltage Shift Computation 82
6.2.2. Simulation Results 83
6.2.3. Overheads 85
6.3.Test-chip Implementation 86
6.4. Silicon Results 89
6.5. Dynamic BTI Management 93
6.6.Conclusion 96
7. Summary and the Road Ahead 97
7.1. Summary 97
7.2. Future Work 99
7.2.1. Statistical Modeling of NBTI Degradation 99
7.2.2. *In situ* Gate-oxide Degradation Sensing 99
7.2.3. Gate-oxide Degradation Recovery 99

7.2.4. *In situ* BTI Sensing 99
7.2.5. Mapping of Accelerated Stress Conditions to Nominal Conditions 100
7.2.6. Dynamic Gate-Oxide Reliability Management 100
7.2.7. Dynamic Reliability Management Implementation 101
8. Bibliography 102

List of Figures

Figure

1.1 Lifetime distribution of chips under different process, voltage and temperature (PVT) conditions. 3

1.2 Conceptual demonstration of DRM. 4

1.3 Block level representation of a DRM controller. 4

2.1 NBTI degradation mechanism. 9

2.2 Circuit schematic and layout for the proposed NBTI sensor and the block diagram for a bank of sensors on the test chip. 12

2.3 (Top) Discrete experimentally measured frequency and frequency given by calibrated model vs. emulated ΔV_{th} at several temperatures. 16

2.4 NBTI Recovery measured by making four quick measurements. 17

2.5 Comparison of sensor output with internal and external bias. 18

2.6 Verification of NBTI measurements. 19

2.7 (Left) Computation of ΔV_{th} using individual calibration (IC), die calibration (DC) and lot calibration (LC) for 46 sensors on a typical die. 20

2.8 Measurement from the NBTI sensor. 22

2.9 (Left) Correlation between amount of ΔV_{th} post-stress and the amount of recovery. 22

2.10 (Left) Intra-die and global variation for 15 dies. 23

2.11 (Left) Distribution of NBTI-induced Vth shift across DUTs from 11 dies. 25

2.12 Distribution of amount of recovery for DUTs recovering under different gate source/drain electric field conditions. 25

2.13 Chip microphotograph. 26

3.1 (Left) Soft-breakdown. 29

3.2 Oxide degradation mechanism. 30

3.3 (Top) Sensor circuit, (Left Bottom) Sensor bank Architecture, (Center Bottom) Chip Layout, (Right Bottom) Die shot. 32

3.4 Timing diagram of all the control signals and corresponding sensor modes of operation. 33

3.5 Gate-oxide stress and measurement results from the sensor. 35

3.6 After a step increase, the gate-leakage becomes noisy and unpredictable. 36

3.7 Early soft breakdown detection gives sufficient time for DRM. 37

3.8 No correlation observed between initial gate leakage and time to breakdown. 38

3.9 Intra-die distribution of initial gate-oxide sensor frequency of 3 dies. 38

3.10 NBTI stress and recovery measurements from the sensor (On:Off = 5:1). 39

3.11 NBTI measured under different stress conditions of voltage and temperature. 40

3.12 Degradation rate increases as the temperature is increased. 41

3.13 No noticeable change in degradation rate is observed when the temperature is decreased. 41

3.14 A weak positive correlation is observed between initial V_{th} and NBTI degradation. 42

3.15 Short sleep intervals result in 10-15% reduction in ΔV_{th} and performance improvement. 43

4.1 DRM illustrations. 45

4.2 Sensor-subset reasonably covers the ΔV_{th} distribution of 256 devices. 47

4.3 Chip layout showing the sensor-subset and the 'core' devices. 48

4.4 Sampling window of 200 samples gives an accurate model fit. 49

4.5 ΔV_{th} distribution after extrapolating the degradation of 32 sensors (at certain time) to T_{LT}. 50

4.6 SRM under worst conditions (Temp = 100C, Vstress = 2.2V). 51

4.7 Lowering the temperature increases NBTI margin. 52

4.8 One of the 32 sensors' readings, with model fit and supply voltage scaling in a DNM implementation. 53

4.9 ΔV_{th} distribution at T_{LT} with DNM at 55C. The slack was converted into performance by DNM. 53

5.1 (Top) Increasing gate-oxide degradation can be modeled as an increase in the value of K and decrease in the value of P. 58

5.2 The non linear nature of the Ig-Vgs characteristics of the gate-oxide becomes more linear with degradation. 59

5.3 The *in situ* monitoring technique is implemented by dividing a circuit into clusters using MTCMOS headers. 60

5.4 The nature of Vrail vs. Vbias curve changes with degradation. 60

5.5 DVA drops sharply with degradation which flags the onset of breakdown. 61

5.6 OBD detection technique implemented on GUTs. 62

5.7 OBD detection technique implemented on XOR Parity trees of different sizes. 63

5.8 (Top) Silicon measurement of Vrail vs Vbias curve for a stressed INVERTER at different points of degradation. 64

5.9 (Top) Silicon measurement of Vrail vs Vbias curves of a 64 gate XOR parity tree at 25C and 125C at three different points in degradation showing immunity of DVA measure to environmental conditions. 65

5.10 Delay and DVA measurements for an XOR parity tree with 64 gates. 66

5.11 Measurements show that the time to detection of onset of degradation increases with cluster sizes larger than 512 gates. 67

5.12 Measurements show a large variation in the time to onset of gate-oxide degradation. 67

5.13 Sensing circuit implemented on 16 bit, 8-tap FIR filter. 141 blocks were stressed and monitored. 68

5.14 (Top) The performance of the FIR degrades as clusters are detected with onset of gate-oxide degradation. 70

5.15 Virtual Rail vs Vbias measurements for the first cluster to be detected with failure, at different points in time. 71

5.16 Spatial map of the clusters indicating their time of failure detection. 72

5.17 Normalized current of clusters with stressed time. 72

5.18 (Left) Chip 1 microphotograph (Right) Chip 2 microphotograph. 73

5.19 Gate-oxide degradation recovery observed. 74

5.20 Simultaneous effect of gate leakage increase and threshold voltage shift at lower stress voltage. 75

5.21 Simultaneous effect of gate leakage increase and threshold voltage shift at lower higher stress voltage. 76

5.22 Stress interruption results in lower leakage current. 77

6.1 Inverter chain with leaking devices. 79

6.2 Header based methodology is used to detect BTI. 81

6.3 Shift in the VR vs Vbias curve is the measure of threshold voltage shift. 82

6.4 The error in measured threshold voltage shift decreases as the header and cluster widths are matched. 84

6.5 A system with BTI measurement technique implemented. 86

6.6 Test-chip with BTI implementation on individual gates. 87

6.7 Test-chip with BTI implementation on XOR parity trees. 88

6.8 Test-chip with BTI implementation on FIR Filter. 89

6.9 NBTI measurement on an inverter. 90

6.10 PBTI measurement on an inverter. 91

6.11 BTI measurement with periodic stress and recovery periods. 92

6.12 Threshold voltage shift and leakage measurement for FIR filter cluster. 92

6.13 Threshold voltage shift distribution among FIR filter clusters. 93

6.14 BTI at 125C and 75C leaves margins at the end of life-time. 94

6.15 Power law extrapolation used to predict BTI degradation till the end of life-time. 95

6.16 DBM trades excess BTI budget with performance by boosting the supply voltage at 75C and 125C. 96

List of Tables

Table

2.1 (Right) Mean and standard deviation of ΔV_{th} for 15 dies. 23

4.1 Comparison of $\Delta V_{th}(\mu+3\sigma)$ of different sensor-subsets 47

Chapter 1

Introduction

Silicon industry has come a long way since the invention of integrated circuits in 1958. It has contributed immensely to the improvement of life standards by advancing the field of research, computing, education, entertainment etc. This has been made possible by Moore's Law which allows for more and more transistors to be packed in a given area [1.1]. This in turn enables high level of integration on a single die. But Moore's Law has been sustained only by the technological advancement in lithography which has allowed reduction in the critical dimensions of the transistor or *technology scaling*. Technology scaling improves the performance of the transistor but at the same time compromises its reliability. Since scaling mainly targets high performance applications, to further boost the performance the supply voltage is not reduced in the same proportion as the critical dimensions are. As a result electric field increases across the transistor gate-oxide, channel, and interconnects which exacerbates transistor degradation mechanism such as gate-oxide wear-out, Hot Carrier Injection (HCI), Bias Temperature Instability (BTI) and Electromigration [1.2]. In older process nodes (above 100nm) the rate of degradation processes was low enough that it did not raise concern over end-of-lifetime failures. But in advanced process nodes (sub 100nm) the aforesaid degradation mechanisms threaten the fulfillment of reliability specifications of the chips. This makes it indispensable to

explore existing reliability management techniques to find out the most suitable of those for advanced process nodes.

1.1 Reliability Management Techniques

Designers have traditionally used static reliability management to meet reliability specifications. This method limits the supply voltage to a fixed maximum value for all fabricated chips such that a chip lifetime of the required level is attained. The lifetime of a chip, however,

is a statistical variable due to the innate randomness in the degradation process. Moreover, process variation, fluctuations in environmental conditions such as voltage and temperature, and state dependence of the oxide degradation also add to the randomness in lifetime [1.3]. Hence the voltage limit is set so that the *weakest* chip meets the lifetime requirements under worst-case conditions. But In reality these conditions will not be experienced by all chips at all times, making such assumptions very conservative. Hence, many chips will fail much later than the desired lifetime. Fig. 1.1 depicts the aforesaid situation with gate-oxide failure under consideration. A percolation-based oxide degradation model was used in this simulation [1.4]. It shows the lifetime distribution of an ensemble of chips under different process, voltage, and temperature conditions. As the voltage and temperature are lowered and the oxide thickness reduced, the mean and spread in the lifetime of the chips decrease [1.5].

If a lifetime of ten years and a yield of 99.9% are desired, then not more than 0.1% of the chips should fail at the end of ten years. As shown in Fig. 1.1 this does not hold true under the conservative worst process and operating condition assumptions. Hence, the voltage would have to be scaled down until the above mentioned criterion is

met. This results in reduced performance of the chips and a lifetime much greater than the specification for many chips, or a *reliability slack,* owing to pessimistic assumptions. This slack can potentially be traded with performance enhancement by increasing the supply voltage.

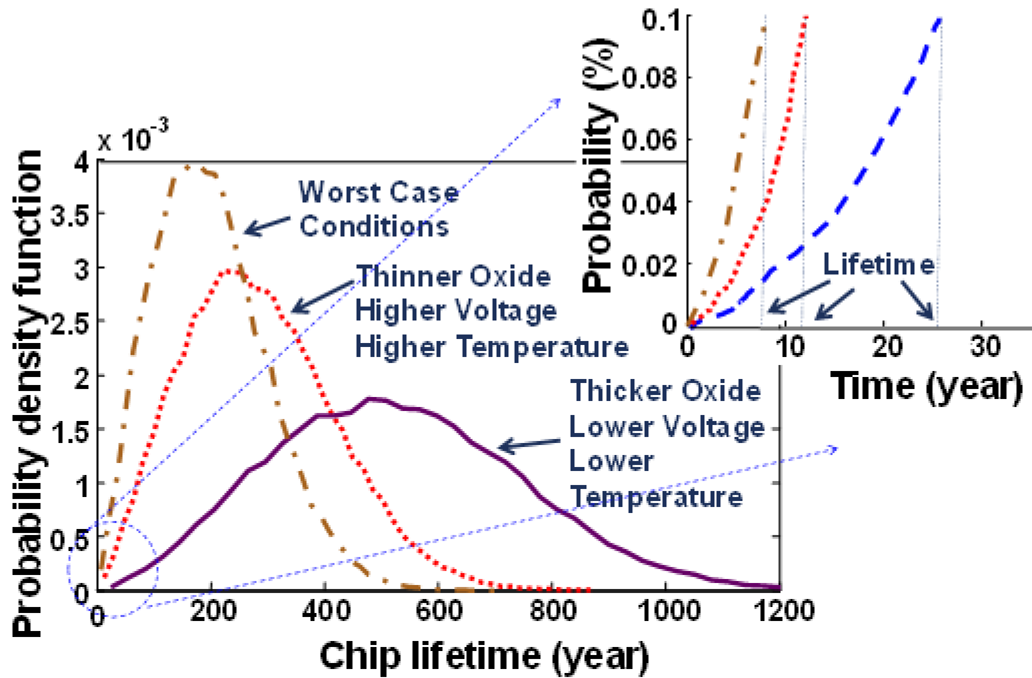


Fig 1.1 Lifetime distribution of chips under different process, voltage and temperature (PVT) conditions. The desired lifetime of 10 years is not met under worst PVT conditions [1.5].

However, this requires that the system is self-aware of its reliability. Such a system is then capable of dynamically adjusting the operating conditions of the chip (in particular, supply voltage and the maximum temperature limit) to achieve peak performance benefits while just meeting the lifetime specification. This approach has been referred to as Dynamic Reliability Management (DRM) [1.6, 1.7, 1.8]. Fig. 1.2 demonstrates the concept behind DRM. The figure shows the lifetime budget curve computed under worst case conditions, with the maximum supply voltage limit of 1.2V. This budget is completely used up under worst conditions. But under typical operating

conditions, the supply voltage is either less than the maximum limit due to less workload or is equal to the maximum allowed supply voltage at the time of maximum workload. Hence at the end of life-time lot of lifetime budget remains unutilized. But a system provided with DRM can exceed the maximum supply voltage limit at the time of peak workload. And hence the lifetime budget is used up to get performance enhancement. Fig. 1.3 shows a block level representation of a DRM controller.

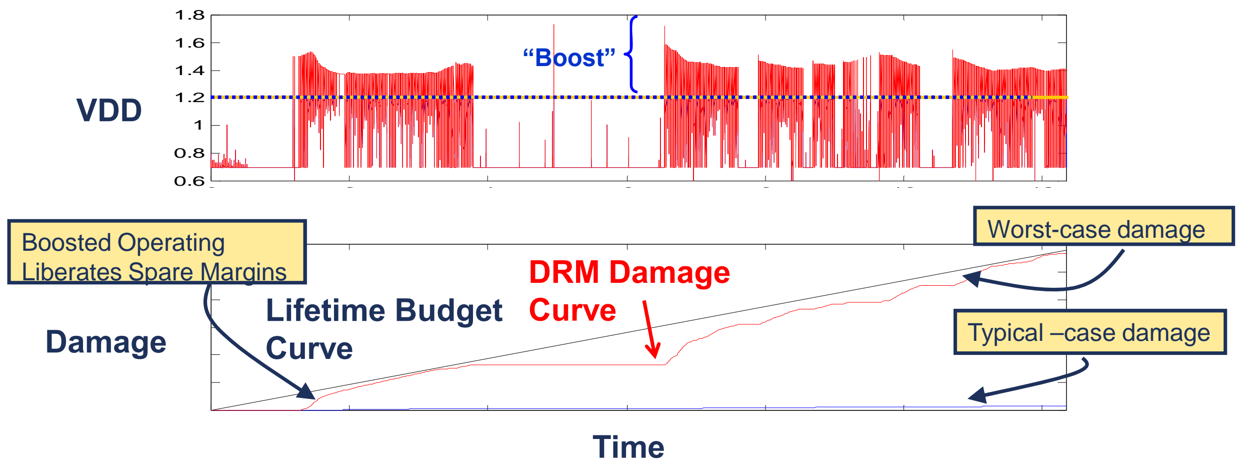


Fig. 1.2 Conceptual demonstration of DRM[].

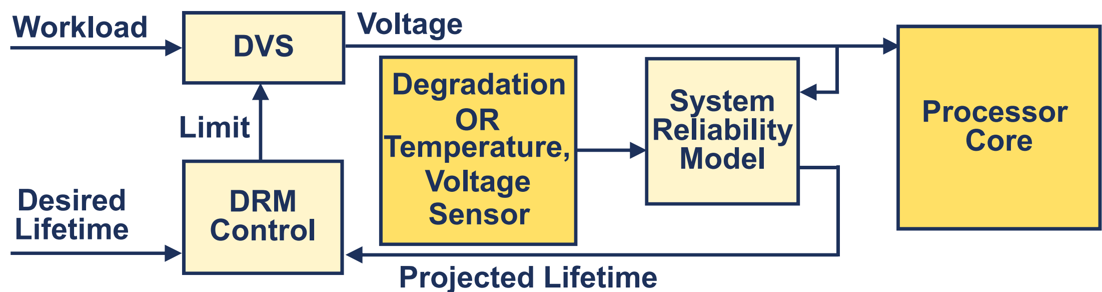


Fig. 1.3 Block level representation of a DRM controller

DRM can be implemented in three ways: model based, degradation sensor based, and *in situ* monitoring based.

In the degradation model-based approach, on-chip voltage and temperature sensors are employed to sense the operating conditions of the chip. The data from these sensors is

fed to a degradation model which is used to compute the expected reliability state of the chip [1.6]. DRM systems based on this approach must account for inaccuracies in the sensors and the degradation models themselves by adding margins that make this approach more conservative. In addition, this approach does not address any innate sources of variation in lifetime such as those due to process variation, state-dependence and the inherent randomness of oxide breakdown, and hence has to be used with considerable margins.

The degradation sensor-based approach obviates the need for voltage and temperature sensors as well as degradation models. Instead, it relies on special sensors that directly monitor the degradation of replicated transistor oxides. Degradation sensors are distributed across the core in large numbers so that the sensors experience the same environmental conditions as the devices in the actual circuit. Degradation data from these sensors is then used to estimate the degradation of the actual devices in the chip [1.3]. Since such degradation sensors experience the same process conditions as the actual devices and also do not incur model inaccuracies, they can operate with tighter margins. However, degradation sensors do not account for lifetime variations due to innate randomness of degradation mechanisms, process mismatch between the replicated oxides that are monitored and the functional oxides of devices in the chip, and state dependence of stress. Thus, considerable margins remain.

The *in situ* monitoring-based DRM scheme employs a direct approach to monitor degradation. In this methodology the actual devices used in the circuit are measured directly to determine their degradation. This approach addresses all sources of lifetime variation and hence provides the most accurate observation of degradation, resulting in

almost complete margin elimination. The key challenge is to achieve this with minimal invasiveness and overhead.

1.2 Contributions of this Thesis

So far DRM has been discussed in literature but not implemented or demonstrated in hardware. This work focuses on realizing DRM in hardware and demonstrating its effectiveness. We focus on developing sensor and *in situ* sensing based approaches to DRM. For this work we focused only on BTI and gate-oxide wear-out degradation mechanism. Other degradation mechanisms such as electromigration and HCI need to be addressed in future work.

Degradation sensors have been proposed in past to characterize the degradation mechanisms. These sensors have large area and power overhead and hence are not very well suited for degradation sensor based DRM. Keeping these factors in mind, in this work we propose two versions of NBTI and oxide degradation sensors with low area and power overhead numbers (Chapter 2 and Chapter 3). These sensors are based on standard cell design which reduces the effort in its integration with the core. Furthermore we implement Dynamic NBTI Management (DNM) using the NBTI component of the degradation sensors and quantify its benefits over traditional Static Reliability Management (Chapter 4). We then discuss our work on *in situ* degradation sensing techniques. In Chapter 5 we propose a method to do *in situ* gate-oxide wear-out detection which is capable of sensing the onset of gate-oxide wear-out of the actual core circuitry. We then propose an *in situ* BTI sensing technique and use it to perform Dynamic BTI Management (Chapter 6 and Chapter 7). Finally we conclude with the unanswered questions and the future work in this direction (Chapter 8).

Chapter 2
Compact Degradation Sensor to Monitor Negative Bias Temperature Instability

Semiconductor reliability is a growing issue as device-critical dimensions shrink and transistor integration continues to roughly double every 24 months. Aggressive oxide thickness scaling has led to large vertical electric fields in MOSFET devices in which Negative Bias Temperature Instability (NBTI) is a critical issue. Due to NBTI the threshold voltage of PMOS devices increase with time because of which the core logic gets slower and the SRAM bit-cells might become instable. Since NBTI is highly sensitive to operating conditions it is extremely hard to characterize a design with respect to NBTI during design time. This supports the case for on-chip structures to be used for real-time estimation of NBTI degradation in devices and circuits. Since the degradation is a statistical process, hundreds or even thousands of sensors are required to estimate bounds on overall chip performance degradation [1.3]. Hence it is essential that the sensors are small with low power consumption.

This work introduces a new compact structure to quantify the change in performance of devices undergoing NBTI [2.1]. The small size of the sensors makes them amenable to use in a standard cell design with minimal area and power overhead. Compact sensors can be implemented in large numbers to collect high-volume data on device degradation. For instance, based on results from the test chip in this chapter we

observe the effect of initial threshold voltage on the V_{th} shift due to NBTI and the correlation between amount of V_{th} shift and the amount of recovery, among other effects.

2.1 NBTI Degradation Mechanism

NBTI results in an increased absolute threshold voltage of p-channel MOSFETs, and hence a degradation in drain current and performance. Although NBTI is not a new phenomenon, it has recently become a major reliability issue due to high gate electric fields resulting from scaling, high operating temperatures due to large power consumption on-chip, and the addition of nitrogen to thermally grown SiO_2 (since hydrogen diffusion is enhanced in nitride-oxides) [2.2]. Most research on NBTI attributes the threshold shift to the creation of interface traps and oxide charge by a negative gate bias at elevated temperatures (Fig. 2.1). The mechanism involves breaking of Si-H bonds at the Si/SiO_2 interface by a combination of electric field, temperature, and holes. It results in dangling bonds or interface traps at that interface and positive oxide charge that may be due to H^+ [2.2]. The threshold voltage shift due to this mechanism is permanent and cannot be recovered upon removal of stress. The second mechanism involves hole-trapping due to electric field in the gate-oxide. Upon removal of the negative stress there is an immediate partial recovery from the threshold shift that occurred during stress. It is due to hole-detrapping in the gate-oxide [2.3].

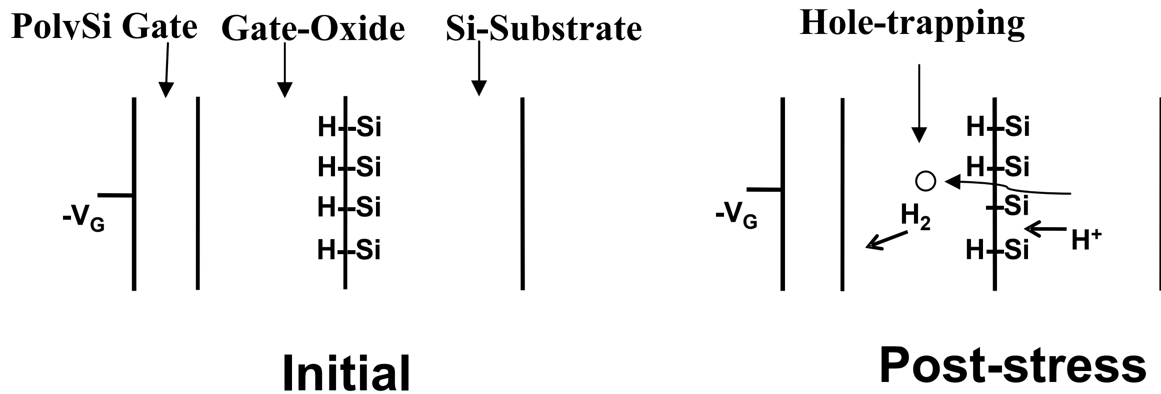


Fig. 2.1 NBTI degradation mechanism. Interface traps are created by breaking Si-H bonds due to negative gate voltage stress. Positively charged interface traps contribute to threshold voltage shift. Hole-trapping also contributes to vth shift.

NBTI poses a serious threat to chip reliability since significant ΔV_{th} can lead to marginal circuit operation (e.g., timing failures in digital logic or bias point drift in analog circuitry) [2.4, 2.5]. Additionally, NBTI degradation in PMOS devices has been shown to cause a reduction in the static noise margin of SRAM cells, leading to read-induced cell stability issues [2.6].

Upon removal of the negative stress, there is an immediate partial recovery from the threshold voltage shift that occurred during the stress [2.7]. This recovery arises due to the partial passivation of the interface traps when the hydrogen diffuses back to SiO_2-Si interface, as well as the removal of the holes trapped in the oxide defects when the negative stress is removed. NBTI recovery adds to the complexity of measuring NBTI effects. If the measurement interrupts the stress state of the device, the measurement time needs to be sufficiently small (on the order of ms) to avoid masking the actual V_{th} shift by inadvertent recovery. This poses the biggest challenge to any experiment seeking to characterize NBTI. The recovery effect also makes it difficult to assess the lifetime of a device undergoing NBTI since the behavior of a device under stress changes once it has been subjected to recovery.

2.2 Previous Measurement Techniques

The aim of most previous work in NBTI measurement has been in the direction of characterizing NBTI. In the past, researchers have used invasive probing methods that require direct access to the device-under-test (DUT) to monitor currents. One large class of work [2.8-2.14] employs a direct current probing approach. Ring oscillator based structures have been proposed in references [2.15, 2.16, 2.17]. All these structures consist of a pair of ring oscillators, one of which experiences accelerated stress. The structure proposed in references [2.15, 2.17] measures the beat frequency (i.e., difference of the two oscillator frequencies) which is attributed to the V_{th} shift due to NBTI. These structures produce a digital output which makes it easier to collect and process it. The structure proposed in reference [2.16] requires a controllable external analog bias to map the change in beat frequency to V_{th} shift. In reference [2.18] this analog bias is generated on-chip using a delay-locked loop and the ring oscillators is replaced with a voltage controlled delay line. The analog output makes it harder to collect the data from these two approaches [2.16, 2.18]. All of these four proposed structures [2.15-2.18] require a large number of delay stages to get high sensitivity to V_{th} change and hence are costly in terms of area and not ideal for use in large numbers (hundreds or thousands) as on-chip sensors.

2.3 Proposed NBTI Sensor

2.3.1. Circuit Design Principles

Fig. 2.2 shows the full schematic of the proposed NBTI sensor, its layout, and the placement scheme for sensors on the test chip. The NBTI measurement technique relies on a PMOS device (P1) to starve the current supplied to a 15-stage NAND gate ring oscillator. There are three modes of operation for this sensor: stress mode, measurement

mode, and recovery mode. In the stress mode, P1 is stressed with negative bias by grounding the input. The change in oscillation frequency during the lifetime of the sensor quantifies the change in V_{th} of P1. In the measurement mode, P1 is biased in subthreshold to exponentially sensitize the oscillation frequency to ΔV_{th}. Experimental results show that biasing P1 in subthreshold leads to a 53% change in oscillator frequency (f_{osc}) for 10% ΔV_{th}. During recovery mode, the gate of P1 is tied to VDD to allow NBTI recovery.

Subthreshold current is highly sensitive to temperature; therefore, we implemented a control PMOS header P0 to correct for temperature variation. Any change in the P0-starved oscillation frequency gives a measure of this variation. A mathematical model, explained in Section IV.B, maps ΔV_{th} and temperature to oscillation frequency.

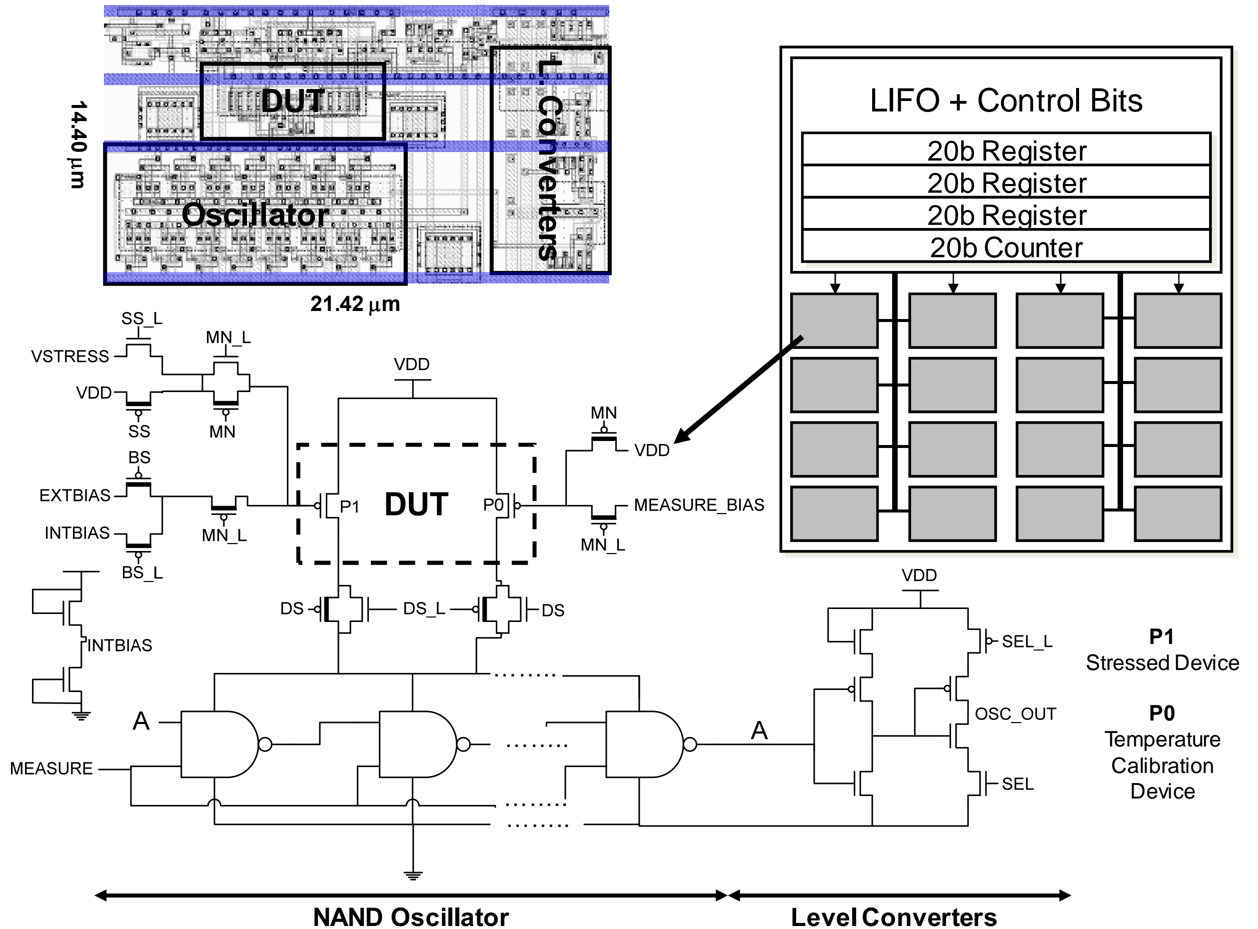


Fig. 2.2 Circuit schematic and layout for the proposed NBTI sensor and the block diagram for a bank of sensors on the test chip. The sensor operates in i) Measurement mode or ii) Stress/Recovery mode. In i) P1/P0 are biased by EXTBIAS or INTBIAS (~1V), while in ii) P1 is biased with VSTRESS (≤0) or VDD (recovery mode) and P0 is biased with VDD.

The estimated temperature and the P1-starved oscillation frequency are used to quantify the ΔV_{th} of P1 after stress. The additional circuitry consists of an internal bias generator (for measurement mode), four multiplexers (to switch between different modes of operation), and a level converter (at the oscillator output). Multiplexers selectively put 1.) P1 in stress or recovery mode (which allows the application of AC stress on P1), while P0 is always in the unstressed mode 2.) Enable switching from stress/recovery mode to measurement mode, 3.) Select between the internal bias generator and external

bias (from pads), and 4.) Select either P0 or P1 to starve the oscillator. Since the oscillator is strongly current starved, the oscillation amplitude is small. Thus a level converter is used to restore the amplitude of oscillations.

The sensors are arranged in an array and form a bank. The bank also includes one 20-bit counter and three subsequent serially connected 20-bit storage units. The counter and three storage units together allow four quick measurements when a stress cycle is interrupted to conduct measurements, and allow quantifying the fast recovery process. In total, a die contains 96 NBTI sensors.

2.3.2. Oscillator Modeling

The oscillation frequency of the NBTI sensor oscillator can be approximately formulated as

$$\frac{1}{f_{osc}} = \frac{KCV_{amp}}{I_{DUT}} \quad (1)$$

Where K is a fitting constant, C is the capacitance at each stage of the oscillator, V_{amp} is the amplitude of oscillations, and I_{DUT} is the average current through the DUT. Simulations show that the peak charging current, I_{charge}, for a stage of the ring oscillator is roughly equal to I_{DUT}. This helps to arrive at an expression for V_{amp}. Since P0 is strongly starved so that the ring oscillator operates in subthreshold, I_{DUT} and I_{charge} are governed by the subthreshold current equation:

$$I_{sub} = AT^2 e^{\frac{V_{GS} - V_{TH} - \gamma V_{BS} + \eta V_{DS}}{nT}} \left(1 - e^{-\frac{V_{DS}}{V_T}}\right) \quad (2)$$

Here A is a constant that is dependent on device parameters W, L, μ, and C_{ox}.

Using (2),

$$I_{DUT} = A_{DUT} T^2 e^{\frac{V_{DD} - V_{bias} - V_{thDUT} + \eta(V_{DD} - V_{amp})}{nT}} \quad (3)$$

Where V_{bias} is the bias voltage at the gate of the DUT.

Since V_{BS} is 0, the γV_{BS} term is dropped. Also, with $V_{DS} \gg V_T$ the last term can be dropped.

Similarly,

$$I_{charge} = A_{charge} T^2 e^{\frac{V_{amp} - V_{thROSC} + \eta V_{amp}}{nT}} \quad (4)$$

Where V_{thROSC} refers to the effective V_{th} of the devices in the ring oscillator.

Here also the same two terms are dropped by the same reasoning. Based on the previous argument,

$$I_{DUT} = I_{charge} \quad (5)$$

Using (3), (4), and (5), the expression for V_{amp} becomes

$$V_{amp}(V_{thDUT}, T) = A_1 T + A_2 V_{thDUT} + A_3 \quad (6)$$

where $A_1 = \frac{n}{1+2\eta} \ln\left(\frac{A_{DUT}}{A_{charge}}\right)$,

$$A_2 = -\frac{1}{1+2\eta}$$

$$A_3 = \frac{V_{DD}(1+\eta) - V_{bias} + V_{thROSC}}{1+2\eta}$$

Furthermore, V_{thROSC} can be modeled as

$$V_{thROSC}(T) = V_{tho} - T_{coff} \log\left[\frac{T}{T_i}\right] \quad (7)$$

where V_{tho} is the threshold voltage at $T = T_i$. The value for V_{tho} and the corresponding T_i, are taken from SPICE device models. After reducing the number of required fitting constants, (1), (4), (6), and (7) form a system of equations to give

$$f_{osc} = AT^2 \, e^{\frac{A_1 T + A_2 V_{thDUT} + A_3(1+\eta) - V_{tho}}{nT}} \frac{1}{A_1 T + A_2 V_{thDUT} + A_3}$$

$$= g(V_{thDUT}, T, \{A, \eta, n, A_1, A_2, A_3\}) \quad (8)$$

Here V_{tho} is taken from SPICE models, V_{thDUT} is the threshold voltage of the DUT, T is the temperature, and $A, \eta, n, A_1, A_2, A_3$ are fitting parameters.

2.3.3. Calibration and Measurement Strategy

The mathematical model developed in the last section must be calibrated, i.e., the fitting parameters need to be calculated. The calibration steps are shown in Fig. 2.3. The temperature and gate bias of the DUT are swept, and frequency is measured. The effect of sweeping gate bias is equivalent to varying threshold voltage. This data is used to curve-fit the model given by Equation 8. Results of the curve fitting at different temperatures are also shown in Fig. 2.3.

The values of the fitting parameters will differ across DUTs due to process variation. Using a particular fixed set of parameters for all DUTs leads to error in the computation of temperature and ΔV_{th}. This error varies with the calibration methodology employed. This work investigates three different methodologies and examines the resulting error in each case: 1) individual, 2) die, and 3) lot calibration. Using individual calibration, the fitting parameters are calculated separately for each DUT. This gives the minimum error in temperature and ΔV_{th} but at the cost of greatly increased test time and post-test computational requirements. Die calibration involves calibrating the model for only one DUT on each die and then using these derived parameters to model all other DUTs on the chip. With small intra-die variations, this method

delivers very reasonable accuracy with a significant reduction in test/calibration time. With lot calibration, just one DUT is examined for the entire wafer lot and fitting

parameters from that device are used for other DUTs within that lot. Experiments were performed to quantify the incurred error when different calibration methodologies are adopted; the results are discussed in the following sub-section.

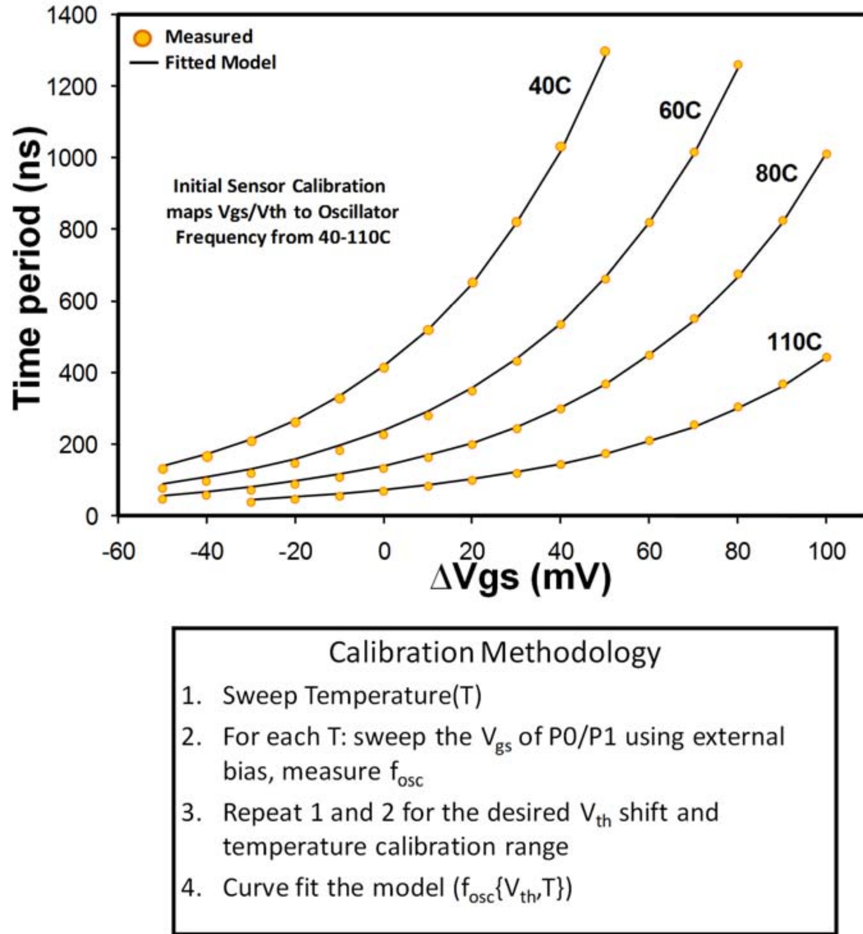


Fig. 2.3 (Top) Discrete experimentally measured frequency and frequency given by calibrated model vs. emulated ΔV_{th} at several temperatures. The model fits the experimental data to minimize the actual ΔV_{th} and $\Delta V_{th\text{-}predicted}$ by sensor. (Bottom) Overview of the calibration methodology.

2.3.4. Experimental Results

The test chip, fabricated in 1.2V/3.3V 130nm CMOS technology, contains 96 NBTI sensors arranged in six banks. Each bank contains 16 NBTI sensors, which are individually addressable for measurement using a 20-bit counter. The threshold voltage in

this technology has a nominal value of 355/-325mV for NMOS and PMOS devices, respectively.

The area of each NBTI sensor is 308 μm^2 with a stress mode power consumption of 4.5nW and a measurement power of 500nW. A measurement time of 100µs is achieved for results presented. The three serially connected 20-bit storage registers were used to perform four quick V_{th} measurements at the interval of 100µs, to see the effect of the measurement time on amount of recovery induced by measurements. The results shown in Fig. 2.4 indicate that the measurement period of 100µs induces minimal V_{th} recovery.

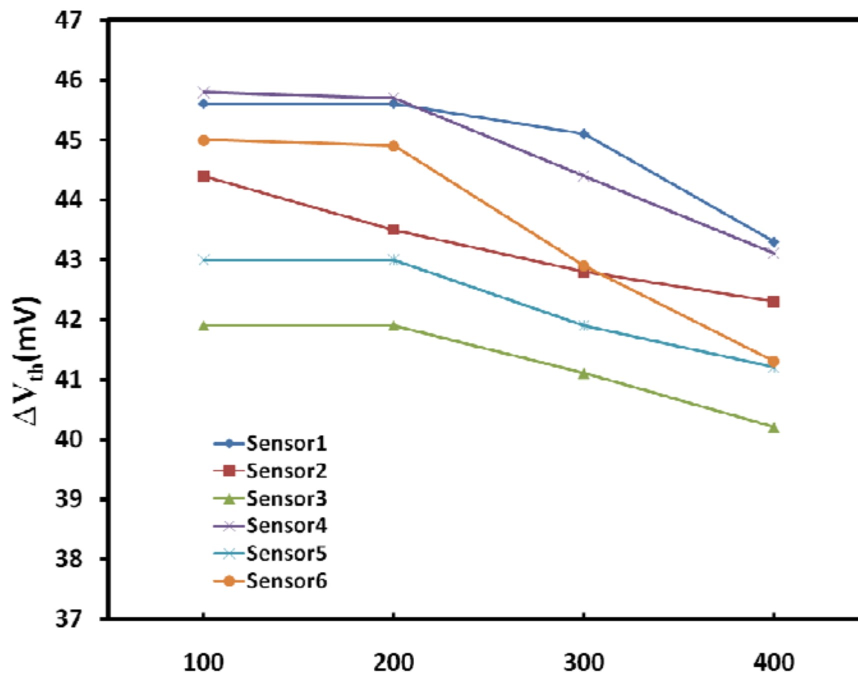


Fig. 2.4 NBTI Recovery measured by making four quick measurements.

Further, the measurement time can be vastly improved by using a high-speed clock for the counter and using the sensor output as an enable signal for the counter. This high speed clock can be as slow as 500MHz to give a high counter count. This clock can

be easily tapped off the system clock with miniscule area and power overhead over the system's area and power consumption. An internal bias generator is included in the design that incurs a maximum of 2.2% error in ΔV_{th} measurement relative to the external bias (Fig 2.5). The external bias is used for the results presented here.

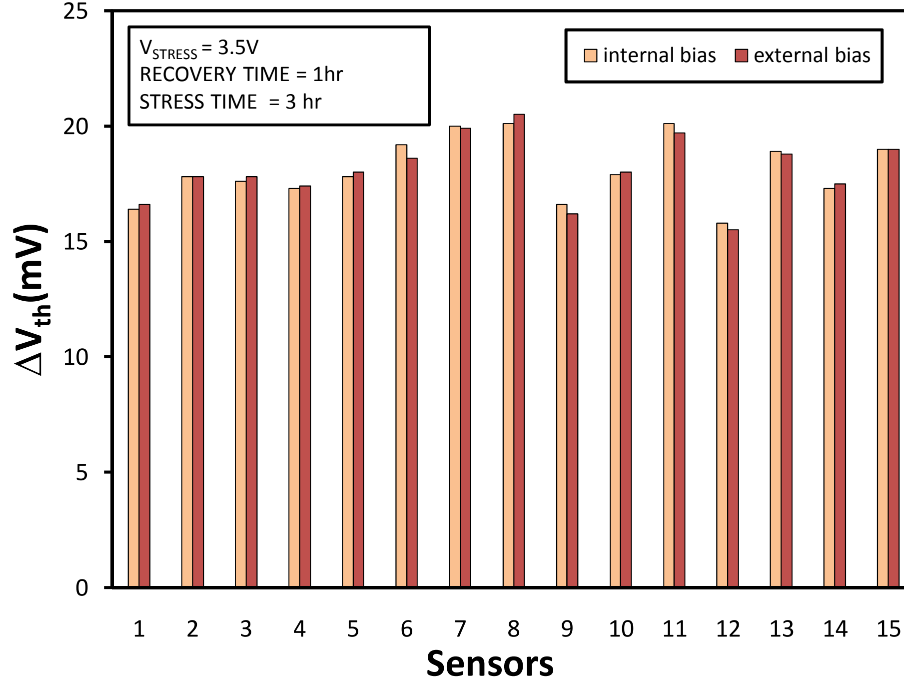


Fig. 2.5 Comparison of sensor output with internal and external bias. Internal bias incurs a maximum error of 2.2% over the external bias for the shown sample.

Fig. 2.6 shows results from an experiment to verify that the proposed method of NBTI ΔV_{th} estimation measures the actual shift in V_{th} of the device P1. In this experiment, the gate voltage of the DUT is swept, showing frequency vs. V_{gs} (Curve A). The DUT is then stressed and subsequently relaxed (allowing some recovery to occur). After the recovery saturates, the gate voltage of the DUT is again swept to obtain Curve B. The horizontal shift between the two curves gives the residual $\Delta V_{th\text{-}sweep}$. Then the sensor is used to find $\Delta V_{th\text{-}predicted}$ based on oscillation frequency. The sensor predicts ΔV_{th}

accurately with a 3σ error of 1.23 mV. Thirteen different points on Curves A and B are used to compute the mean and standard deviation for $\Delta V_{th\text{-}sweep}$.

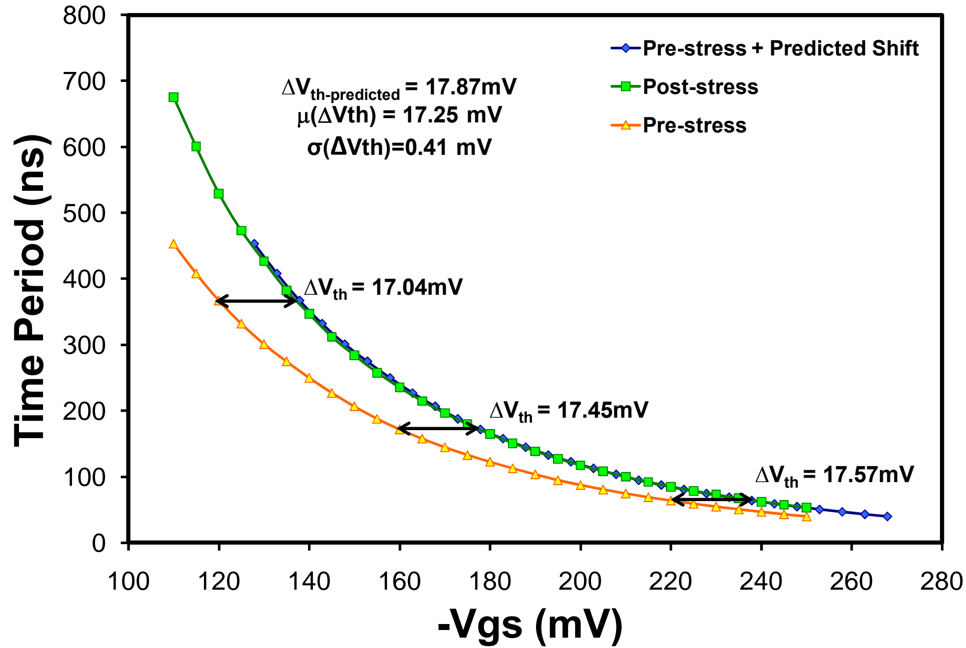


Fig. 2.6 Verification of NBTI measurements. V_{gs} for P1 is swept before stress and after stress under constant conditions of V_{dd} and T; the horizontal shift in the two profiles is equal to $\Delta V_{th\text{-}sweep}$. The pre-stress profile shifted horizontally by $\Delta V_{th\text{-}predicted}$ overlaps very well with the post-stress profile.

We also measured the error introduced by using die or lot calibration methodologies. Here the error is computed and normalized with respect to individual calibration. As shown in Fig. 2.7, $\mu(\Delta V_{th\text{-}error})$ = 1.95% and $\sigma(\Delta V_{th\text{-}error})$ = 1.28 % for die calibration and $\mu(\Delta V_{th\text{-}error})$ = 2.11% and $\sigma(\Delta V_{th\text{-}error})$ = 1.5 % for lot calibration. The confidence interval analysis shows that the margin of error in estimation of mean for die calibration is 0.31% and for lot calibration is 0.36% with confidence level of 90%.

These small errors imply that die and lot calibration provide a good trade-off between accuracy and calibration/testing effort. This has enormous implications on the usability of these sensors in actual chips. For a particular process-technology the sensor can be extensively calibrated and that calibration can be used for the sensors deployed on

chips in that process-technology. In that view the sensors would not need an external bias (primarily required for calibration process) which will reduce the routing overhead of this signal all over the chip. This study was extended to test for temperature correction (Fig. 2.7) and measurements were taken for a wide range of temperatures (10°C to 110°C). The estimates derived from the individual calibration methodology vary the least (σ/μ= 1.1 %) across the range, followed by die calibration (σ/μ= 2.1 %) and then lot calibration(σ/μ= 3.1 %).

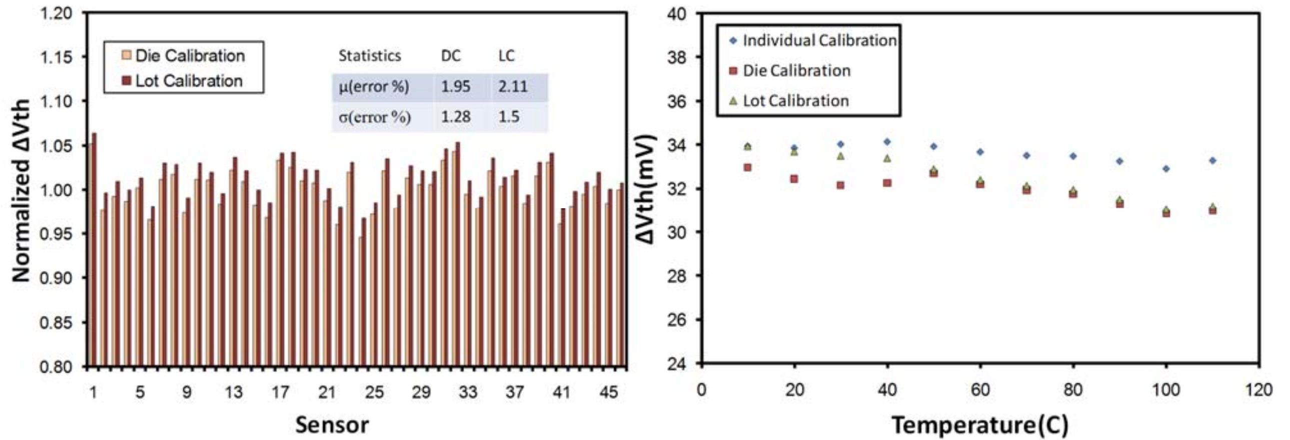


Fig. 2.7 (Left) Computation of ΔV_{th} using individual calibration (IC), die calibration (DC) and lot calibration (LC) for 46 sensors on a typical die. The maximum error incurred by lot calibration ~6%, which makes it quite reliable to use and simplifies the calibration procedure drastically. (Right) Temperature sensitivity of different calibration methods. IC varies the least across the temperature range, followed by DC and then LC.

Fig. 2.8 shows the characteristic saw-tooth curve generated by alternately stressing and then removing the stress with V_{stress} = 1.7V and T = 130°C. The jitter in oscillations of the sensor due to power supply and bias voltage noise results in noise in the ΔV_{th} estimate. Fig. 2.8 also shows the NBTI degradation for different stress conditions.

The statistical nature of NBTI degradation is one aspect that has been largely overlooked in the literature. To investigate the probabilistic nature of NBTI, we collected stress/recovery data for 705 sensors across 15 dies with V_{stress} = 3V and T = 120°C (Fig. 2.9). A measurement time of 100μs is used. We observed a strong positive correlation (correlation coefficient = 0.78) between the amount of ΔV_{th} during stress and the amount of recovery. Fig. 2.9 shows the distribution of ΔV_{th} post recovery. This shows that the devices suffering more NBTI degradation also recover more significantly if given sufficient recovery time. This finding can be used to dynamically mitigate the severity of NBTI by allocating sufficient recovery time for the blocks suffering higher threshold shifts. Fig. 2.10 shows intra-die distribution of ΔV_{th} across 15 dies after NBTI stress and the global distribution. The intra-die variation of ΔV_{th} across the dies is fairly consistent. The mean value of four dies (die 7, 9, 10, 15) is off from the cluster of mean ΔV_{th} cluster of other dies. The intra-die variation is much more pronounced as compared to inter-die variation. Due to this the global distribution looks very much similar to the intra-die distribution. Table 2.1 shows the mean and standard deviation of ΔV_{th} for each die and the global distribution. The statistical data shows that the impact of NBTI on circuit performance is more complex than previously considered. These findings also make the traditional reliability solution such as delay margining, more pessimistic and corroborate the case for dynamic reliability monitoring using NBTI sensors.

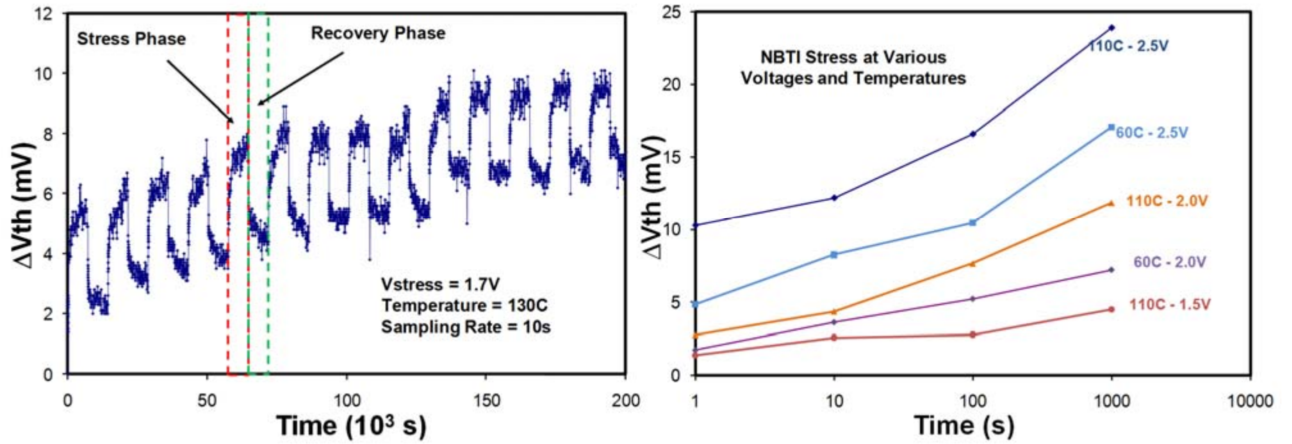


Fig. 2.8 Measurement from the NBTI sensor. (Left) Measurement for 15 cycles of periodic stress and recovery mode. (Right) NBTI measured from the sensor under various stress conditions

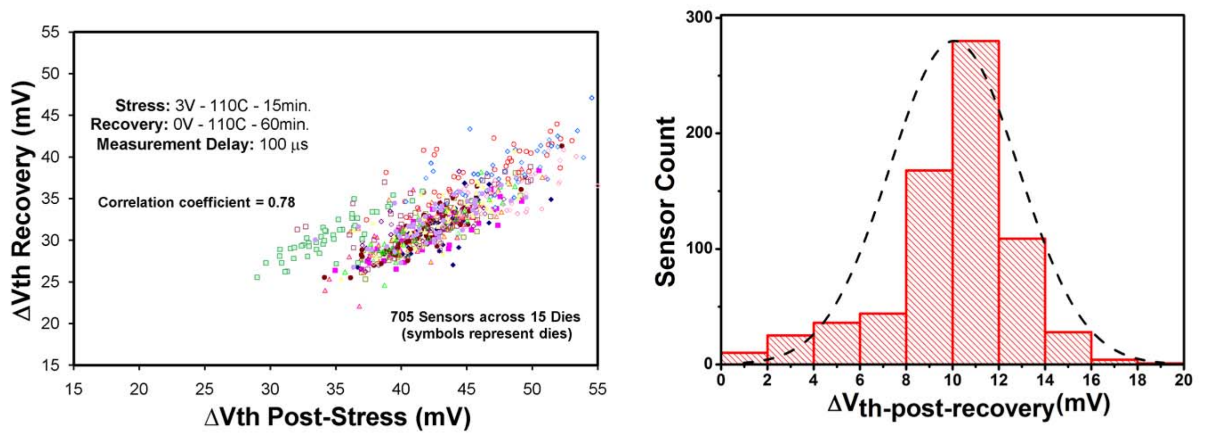
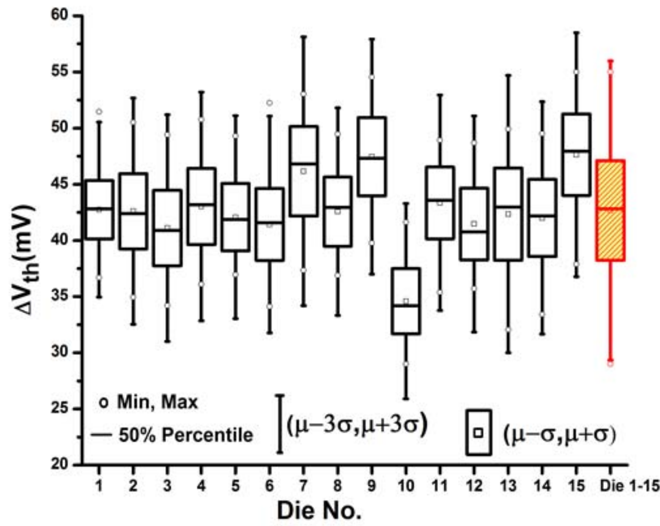


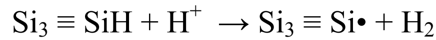
Fig. 2.9 (Left) Correlation between amount of ΔV_{th} post-stress and the amount of recovery.
(Right) Distribution of ΔV_{th} post recovery



Die no.	μ(Δ Vth) (mV)	σ(Δ Vth) (mV)
1	42.74787	2.597079
2	42.60326	3.357749
3	41.11042	3.364006
4	43.02913	3.393327
5	42.07383	3.010861
6	41.4275	3.215041
7	46.15894	3.987419
8	42.56688	3.081232
9	47.46043	3.487033
10	34.5934	2.89541
11	43.35128	3.199631
12	41.46396	3.20498
13	42.34354	4.112596
14	42.00958	3.445013
15	47.63091	3.620797
Die 1-15	42.67	4.44

Fig. 2.10 (Left) Intra-die and global variation for 15 dies.
Table 2.1 (Right) Mean and standard deviation of ΔV_{th} for 15 dies.

The test chip also allowed the investigation of DUTs with two different (high and low) threshold voltages to examine differences in their NBTI characteristics. Fig. 2.11 shows that high-Vth DUTs exhibit larger NBTI degradation overall. To our knowledge, this is the first time such an observation has been made. A possible explanation could be based on the NBTI model proposed by Tsetseris *et al.* [2.19]. According to this model the following chemical reaction occurs at the Si-SiO_2 interface under stress conditions:



$$Si_3 \equiv SiH + H^+ \rightarrow Si_3 \equiv Si\bullet + H_2$$

Here $Si_3 \equiv SiH$ is a hydrogen-terminated interface trap and $Si_3 \equiv Si\bullet$ an interface trap with the dot representing a dangling bond. The hydrogen is believed to originate from the phosphorus-hydrogen bonds in the n-type Si substrate. The P-H bonds dissociate and the hydrogen attracts a hole as it moves to the SiO_2/Si interface, becoming H+, then reacts with the H from the SiH bond to form H_2. The net result is a positively charged Si dangling bond (or trapping center) that contributes to the Vth shift. Since the phosphorus

doping is larger for high Vth devices, there is a higher concentration of H+ available for the reaction, resulting in more traps and a consequently greater Vth shift. Another possible explanation could be that due to higher doping there are more Si atoms in the lattice which are not at the Si lattice centers. These dangling Si bonds are passivated by hydrogen. Since there are more number of SiH bonds which could be broken during NBTI stress, it would lead to higher Vth shift.

We also investigated the impact of electric field on recovery by designing DUTs to recover in three different conditions of gate bias: 1) zero gate-source/drain bias, 2) positive gate-source/drain bias, 3) and positive gate-drain bias with zero gate-source bias. Fig. 2.12 shows a significant difference in the mean ΔV_{th} of zero and positive bias conditions, indicating that the recovery rate is enhanced in the presence of positive electric field. As explained in section II, positive holes trapped in oxide defects also contribute to the threshold shift during negative stress. These charged species are neutralized when the negative stress is removed, which contributes to threshold voltage recovery. When a positive bias is applied at the gate, more of these trapped holes are neutralized, resulting in larger recovery.

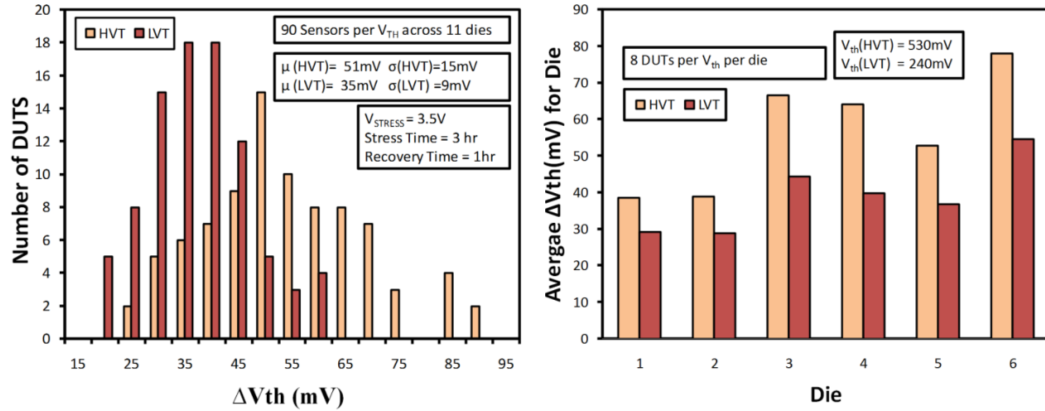


Fig. 2.11 (Left) Distribution of NBTI-induced Vth shift across DUTs from 11 dies. HVT devices show higher degradation. (Right) The mean Vth shift of HVT devices is larger across dies. The respective V_{th} of devices are 530mV (HVT) and 240mV (LVT).

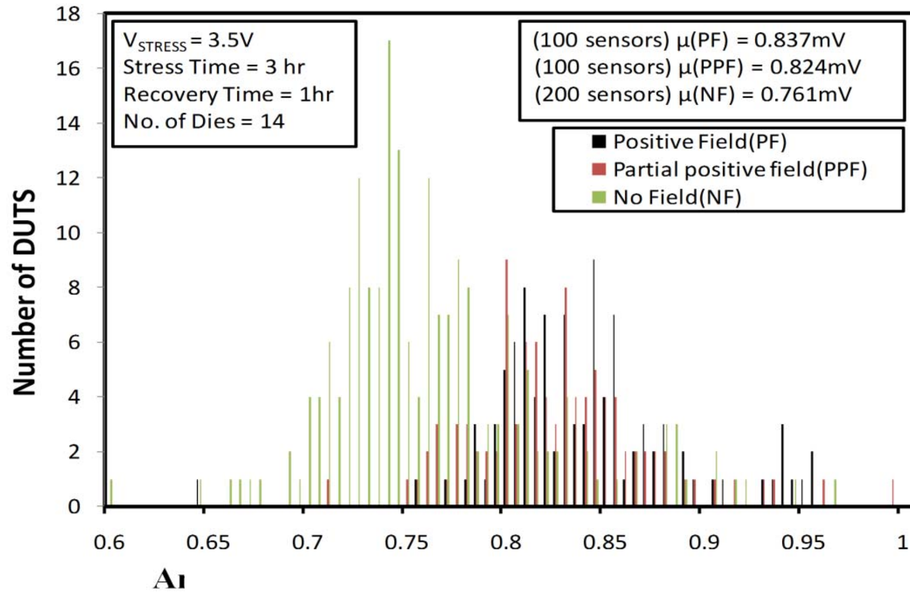


Fig. 2.12 Distribution of amount of recovery for DUTs recovering under different gate source/drain electric field conditions. Application of a positive field increases the amount of recovery in a given time.

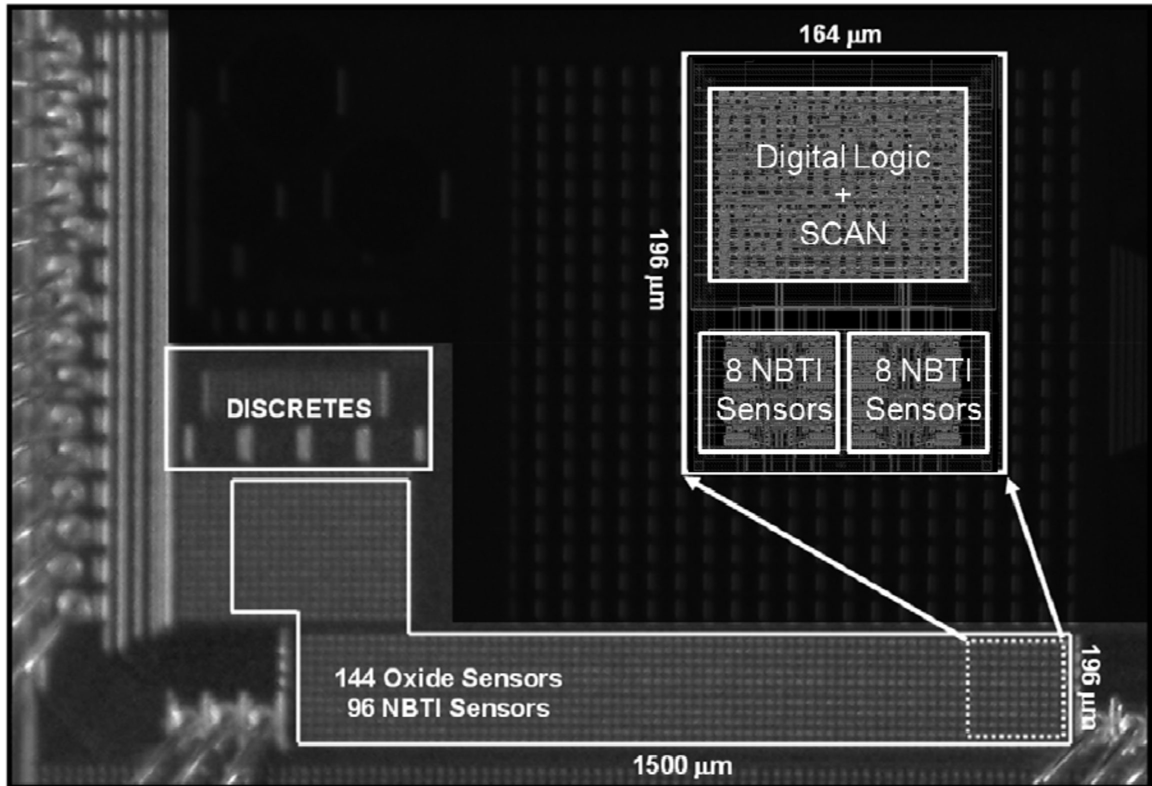


Fig. 2.13 Chip microphotograph

2.4 Summary

This work presents compact NBTI degradation sensor with digital outputs. Due to their small size and simple frequency outputs they are amenable for use in standard cell-based designs. The sensors enable high-volume data collection and the monitoring of chip reliability throughout system lifetime. The degradation data supplied by these sensors also aids in understanding and modeling the complex degradation mechanisms. For example, results from a test chip in 130nm CMOS provide new insight concerning the statistical nature of NBTI and gate oxide degradation, the impact of electric field on NBTI recovery, and the relationship between initial V_{th} and NBTI-induced V_{th} shift.

The NBTI sensor (308 μm^2) is based on a subthreshold ring oscillator concept and is 110X smaller than previous work. We propose a simple calibration method to process

the sensor output data. We observed a maximum error of 2.2% for the NBTI sensor under process/voltage/temperature variations, yielding ΔV_{th} measurements with 3σ accuracy of 1.23mV from 40-110°C.

Chapter 3
A Unified 45nm NBTI and Oxide Degradation Sensor

Process engineers have continuously pushed gate-oxide thickness limits to gain performance. In the advanced process nodes the gate-oxide thickness is less than 15Å, which is approximately 3-4 atomic layers. Since the supply voltage is not scaling proportionately due to performance needs the electric field across the gate-oxide is increasing. In Chapter 2 we discussed how high electric fields exacerbate degradation mechanism such as NBTI. In this chapter in our discussion we will encompass the gate-oxide wear out mechanism as well. Gate-oxide wear-out can be critical than NBTI in the sense that gate-oxide failure of even one transistor can lead to failure of a chip.

In this chapter we propose a unified NBTI and gate-oxide wear-out sensor designed in a 45nm process node. The integration of NBTI and oxide degradation sensing enables efficient reliability monitoring with reduced sensor-deployment effort and overhead. We will discuss the mechanism of gate-oxide wear-out, the previous works on techniques to measure it, and finally we give details on the proposed unified NBTI and oxide degradation sensor and experimental results.

3.1 Gate-oxide Degradation Mechanism

When a voltage is applied across the gate-oxide of a transistor, current flows through it due to quantum-mechanical tunneling. The magnitude of current is determined by the

energy barrier offered by the gate-oxide dielectric [3.1]. As carriers flow through the oxide, they are trapped in the free energy-state traps and consequently alter the electric field inside the oxide. This leads to the formation of more energy-traps. When a chain of defects connects the two sides of gate-oxide there is a sudden increase in the gate-oxide current. This condition is called *soft-breakdown* (Fig. 3.1). As more defects paths are formed and they reach a critical density, a step increase in current is observed which leads to catastrophic break-down of the gate-oxide and as a result the oxide no longer behaves as an insulator (Fig. 3.1). Fig. 3.2 illustrates this phenomenon using energy band diagrams.

It is very critical to sense the gate-oxide wear-out early in its onset so that sufficient time is available to manage its reliability and avoid the catastrophic breakdown of the gate-oxide.

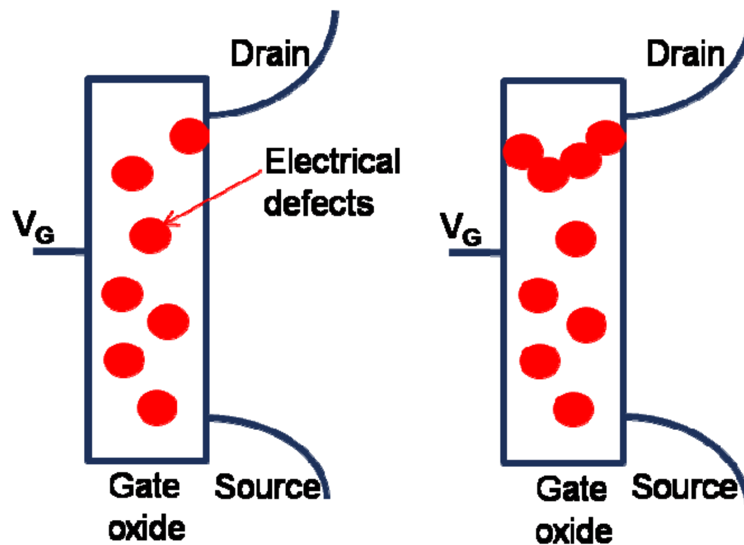


Fig. 3.1 (Left) Soft-breakdown. (Right) Hard-breakdown

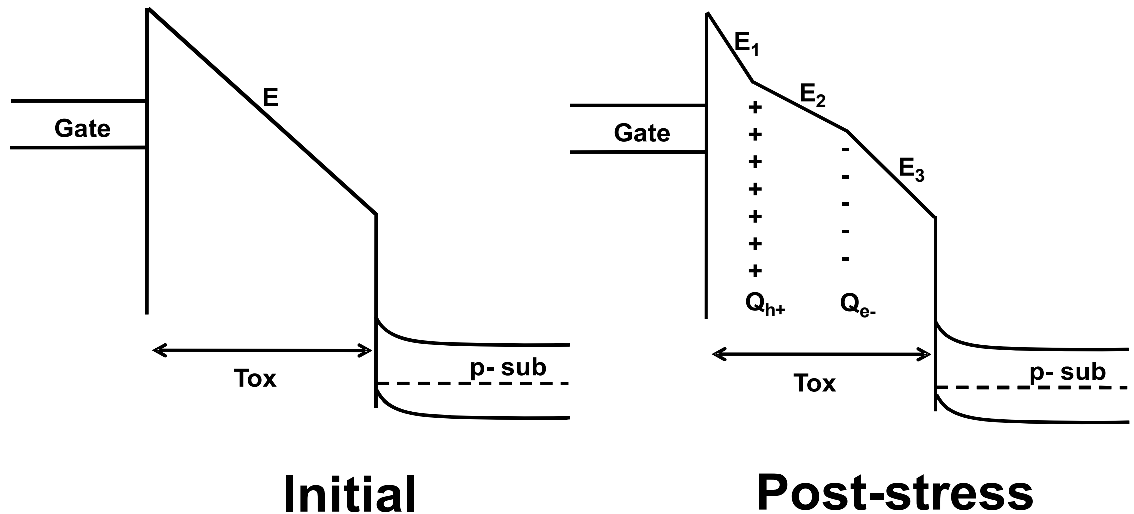


Fig. 3.2 Oxide degradation mechanism. Electrons flowing through the oxide generate defects which increase the local electric field in the oxide, causing more electrons to tunnel through that region and generating new defects. This positive feedback eventually causes the defects to reach a critical density, leading to destructive breakdown [3.1].

3.2 Previous Gate-oxide Degradation Measurement Techniques

As in the case of NBTI, all previous oxide breakdown measurement techniques have been invasive, requiring direct access to DUTs and peripheral circuitry. Uraoka [3.2] evaluates gate oxide reliability using a luminescence method. The set-up requires an optical microscope, photon counting camera, and image processor. References [3.3] and [3.4] propose statistical methods to monitor the yield of gate oxide layers in a manufacturing production line. This technique can be useful in statistically binning the oxide reliability of ICs at manufacturing time but cannot dynamically monitor chip reliability throughout its lifetime.

Gate oxide reliability for high voltage analog power transistors is addressed in [3.5]. A power amplifier (PA) is designed with an oxide reliability monitor for the output stage. The monitoring circuit uses the elements already present in the PA, such as resistors and a pre-driver stage, to measure the conductance of output stage transistors.

Extending this approach to digital circuits would result in large silicon area and power overheads due to analog components. Also, the output of the monitor is an analog voltage, increasing testing costs.

To characterize gate-oxide breakdown reference [3.6] proposed array based test structures in which the gate-oxide of the test devices are stressed. This approach is useful to characterize hard-breakdowns but has limitations measuring the initial gate-oxide wear-out early in its life time due to spurious leakage currents which overwhelm the gate-oxide current. For DRM enabled systems it is important to capture the early on-set of gate-oxide degradation so that the system can take appropriate measures soon enough to increase the remaining life-time of the chip. Finally, in [2.1] oxide degradation sensor was proposed which was smaller in area but consumed high power. In the proposed work we improve upon the oxide-degradation sensor design in [2.1] and reduce its power approximately by 10^5. Also the design of the sensor is less complicated than the one proposed in [2.1].

3.3 Proposed Unified Sensor Design and Operation

The proposed sensor consists of 2 DUTs (D1, D2), which are stressed and then measured for gate oxide and NBTI degradation, respectively. The other components of the circuit include: 1) muxes, to switch the sensor between stress and measurement modes, 2) a ring oscillator, which is shared between oxide and NBTI sensing circuitry, 3) a Schmitt trigger, to improve the slew of the signal originating from node *N2*, 4) level converters, to output NBTI measurements (Fig. 3.3).

To monitor the gate-leakage of a device it is critical to isolate it from other dominant sources of leakage such as sub-threshold leakage. This made the design in [2.1]

complicated as it had to use oxide-stack divider to isolate the critical node from any source and drain connections. In 45nm the gate leakage is comparable to sub-threshold leakage and we took advantage of this fact to simplify the sensor design by using a transistor stack (S) to stress the node N2. This also obviates the amplifier used in [2.1], which results in power reduction of the proposed sensor design.

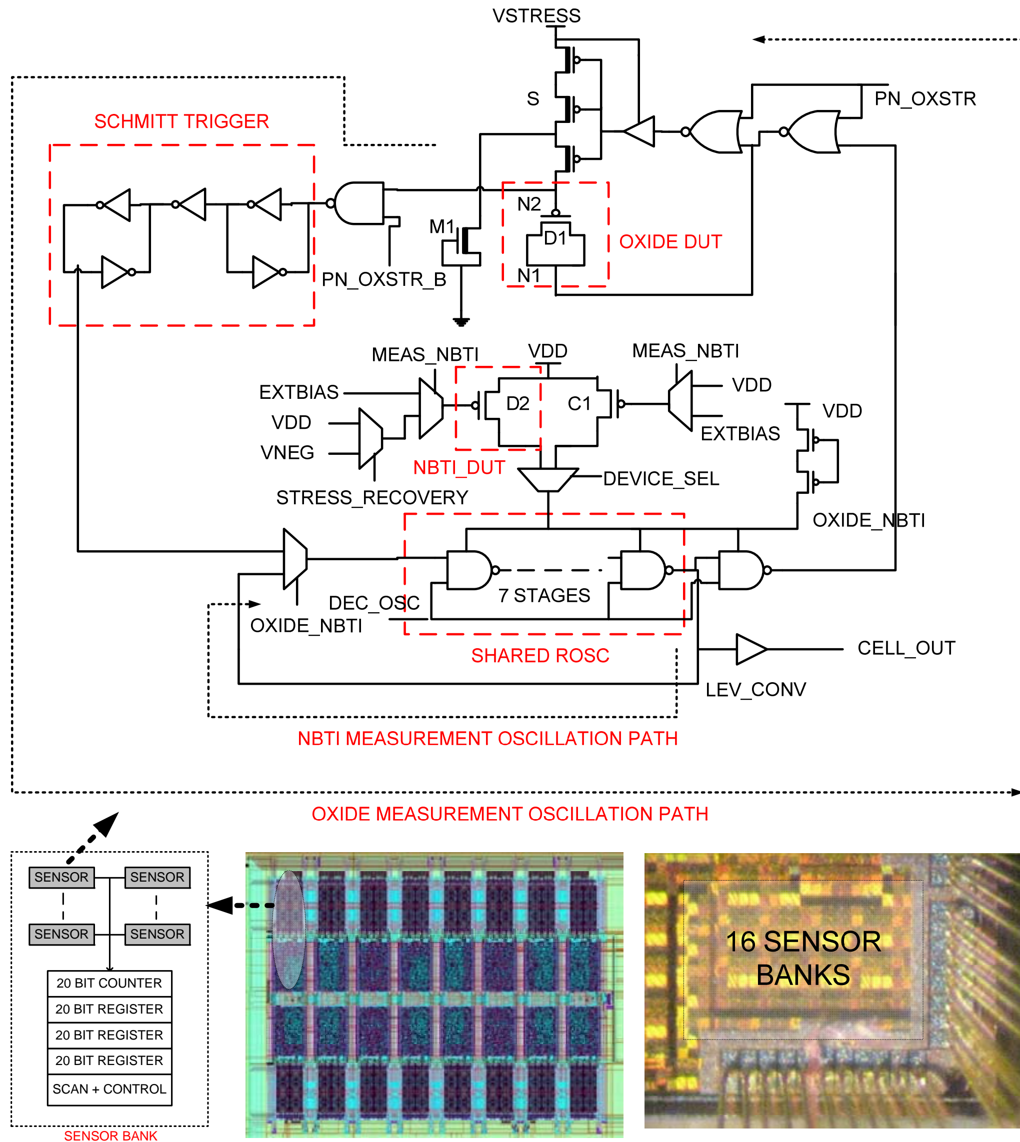


Fig. 3.3 (Top) Sensor circuit, (Left Bottom) Sensor bank Architecture, (Center Bottom) Chip Layout, (Right Bottom) Die shot.

The modes of operation and respective timing diagram of all the control signals and critical nodes are shown in Fig. 3.4. D1's gate oxide is stressed by charging *N2* to *VSTRESS* through the transistor stack S while *N1* is held at ground. During gate-oxide leakage measurement S is cut-off, allowing *N2* to be discharged through the gate leakage of D1. To prevent subthreshold leakage from affecting the discharge time, a cut-off device M1 is added to sink the stack subthreshold current. N2 drives a cross-coupled inverter-based Schmitt trigger to improve the slew of the signal originating from *N2*. The Schmitt trigger drives the ring oscillator, which is shared with the NBTI sensing circuitry. As D1's oxide degrades, its gate leakage increases and N2 is discharged faster, increasing the sensor frequency.

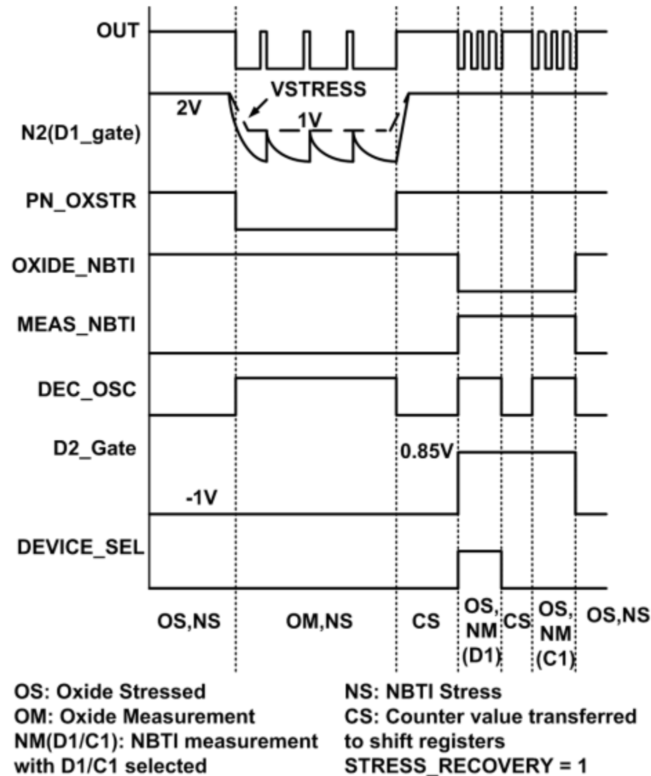


Fig. 3.4 Timing diagram of all the control signals and corresponding sensor modes of operation.

For NBTI sensing, D2 is stressed by muxing in a negative voltage at its gate. In the measurement mode D2 is biased in subthreshold so that any change in its V_{th} impacts the starved ring oscillator frequency exponentially. Since subthreshold circuits are extremely sensitive to temperature changes, a control header C1 is added to correct for any temperature change. A temperature calibration scheme as described in [2.1] is used.

To record the frequencies a 20 bit counter, along with three 20 bit parallel registers are used. The counter and the storage registers collectively enable four consecutive fast frequency measurements. This feature is particularly useful to capture the phenomenon of NBTI recovery when the stress is interrupted to make measurements [2.9].

3.4 Silicon Measurements from Sensors

The sensor is implemented in a 45nm CMOS process. The test chip consisted of 16 banks, each bank containing 16 sensors, 80-bit storage which includes 20-bit counter, control and scan logic (Fig. 3.3). The area of the sensor is 77.3μm^2 (6 Flip-flops). Sensor stress mode power is 8.6nW (>100,000X lower than [2.1]) while measurement mode power is 84.7nW. Hence the power overhead of laying out thousands of sensors would only be a few hundreds of µW at maximum, which is a small fraction of power relative to a microprocessor core. The area and power overhead of the storage and other logic can be amortized by inculcating more sensors in a bank.

Fig. 3.5 shows the sensor output during gate oxide stress measurement results. The sensor captures the initial gradual increase in the gate-oxide leakage as the oxide wears out. This data would be used by the DRM system to detect onset of gate-oxide wear-out or *soft breakdown*, and raise an alert.

Soft breakdown was defined as a 10% increase in gate leakage. In nominal conditions soft breakdown will occur over years. This is followed by a phase where the leakage remains relatively constant with occasional fluctuations. The fluctuation in gate-leakage occurs because defects are continuously being injected and neutralized during the process of degradation [3.7]. This noisy behavior of the gate-leakage is also indicative of the worn-out oxide as it allows more defect formation. At this point the DRM system would take measures to reduce the rate of wear-out. Finally there is a *hard breakdown* of the oxide marking the end of life of transistor, as a result of which there is a step increase of more than 10X in gate leakage and sensor frequency. The DRM system would ensure that hard breakdown does not occur in the core circuitry, by introducing some pessimism in its analysis while adjusting the supply voltage and temperature limit on the core.

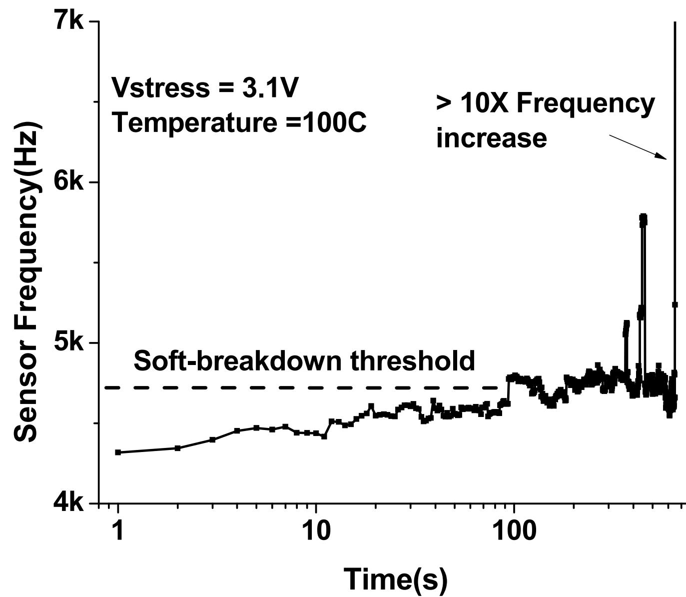


Fig. 3.5 Gate-oxide stress and measurement results from the sensor.

Fig. 3.6 shows the noisy and unpredictable behavior of the gate leakage once it reaches hard breakdown. Fig. 3.7 shows the distribution of time available to the DRM

system after soft-breakdown detection (SBD), normalized to total life time. In this experiment 256 sensors were stressed out of which only 70 had hard breakdown at the end of the experiment. For 95% of the oxides soft breakdown occurred at less than half of their life time. This means that significant time is available to DRM system for reliability management. In remaining 5% of oxides, it can occur as late as up to the last 10% of their lifetime. Even though this number appears small, at nominal conditions it could translate to years which is still a significant time for reliability management.

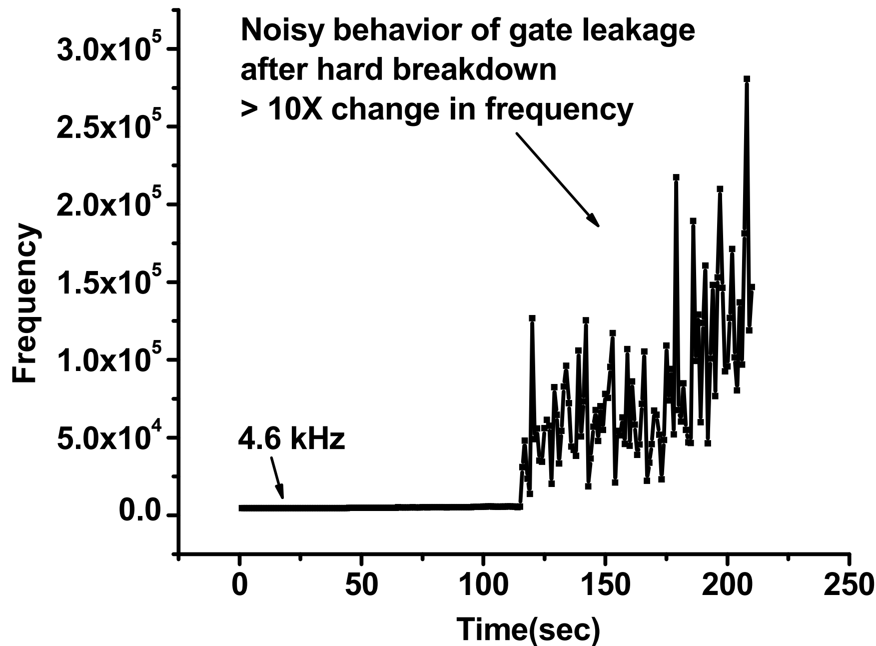


Fig. 3.6 After a step increase, the gate-leakage becomes noisy and unpredictable.

Fig. 3.8 shows little correlation between pre-stress oscillation frequency and time to hard failure. The pre-stress frequency is indicative of the initial gate leakage, or gate oxide thickness. An outlier with high frequency or low oxide thickness will fail early, however similar failure times are observed even for nominal oxide thickness sensors. This confirms the significant randomness inherent in the formation of oxide defects that lead

to failure [3.8] and supports the need for a larger number of low-power and compact reliability sensors on a chip to construct statistical bounds on expected lifetime.

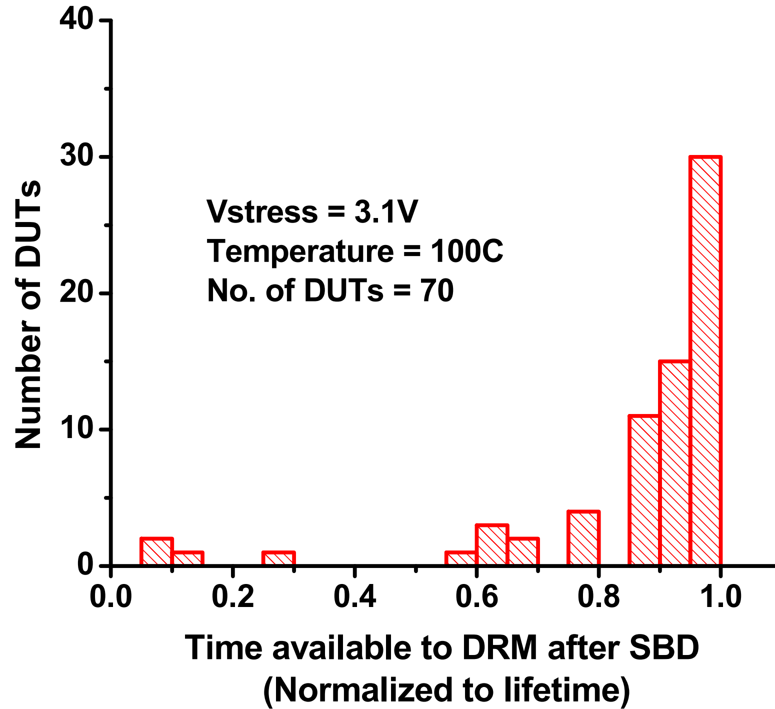


Fig. 3.7 Early soft breakdown detection gives sufficient time for DRM.

Fig. 3.9 shows the intra-die distribution of initial gate-oxide sensor frequency for 3 dies. Die 1 has a clear outlier with low gate-oxide thickness. Knowing that the gate-leakage is exponentially dependent on gate-oxide thickness, it can be seen in general the gate-oxide thickness is very well controlled in this process.

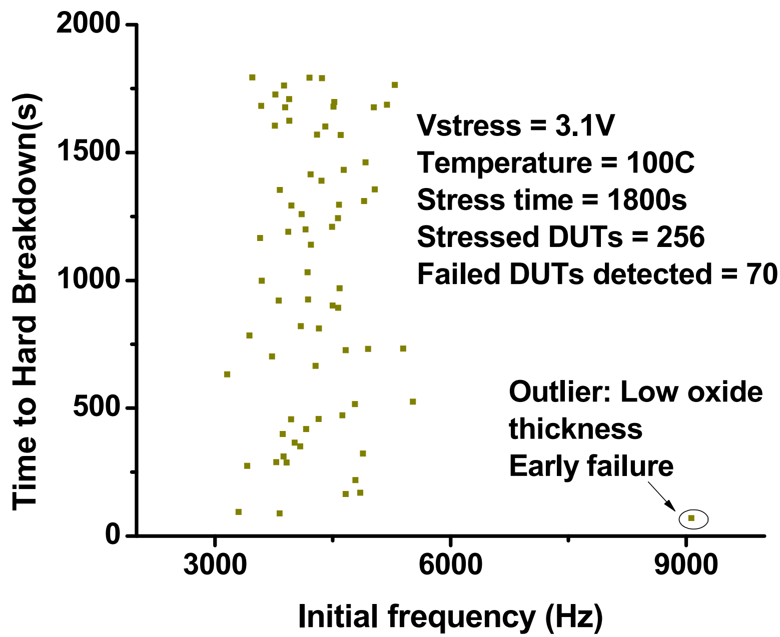


Fig. 3.8 No correlation observed between initial gate leakage and time to breakdown. This shows inherent random nature of oxide breakdown.

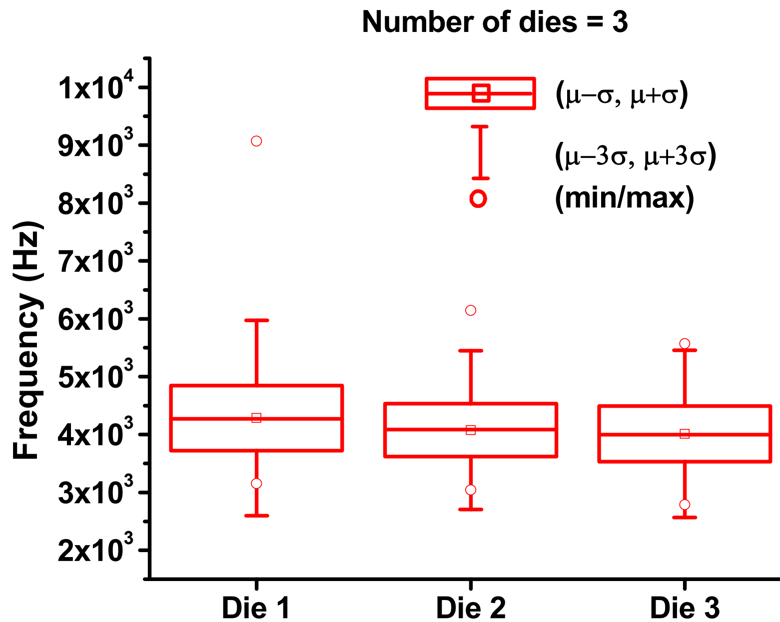


Fig. 3.9 Intra-die distribution of initial gate-oxide sensor frequency of 3 dies.

Fig. 3.10 shows typical saw-tooth curve for NBTI degradation of device D2. The measurement time in all NBTI measurements is 100μs, which is same as in [2.1]. Fig. 3.11 shows ΔV_{th} of D2 due to NBTI, determined using sensor frequency measurements under different accelerated conditions of temperature and voltage. Comparing these threshold shifts in 45nm with results in [2.1] (in 130nm) under similar stress conditions confirms that scaling has significantly increased NBTI effects.

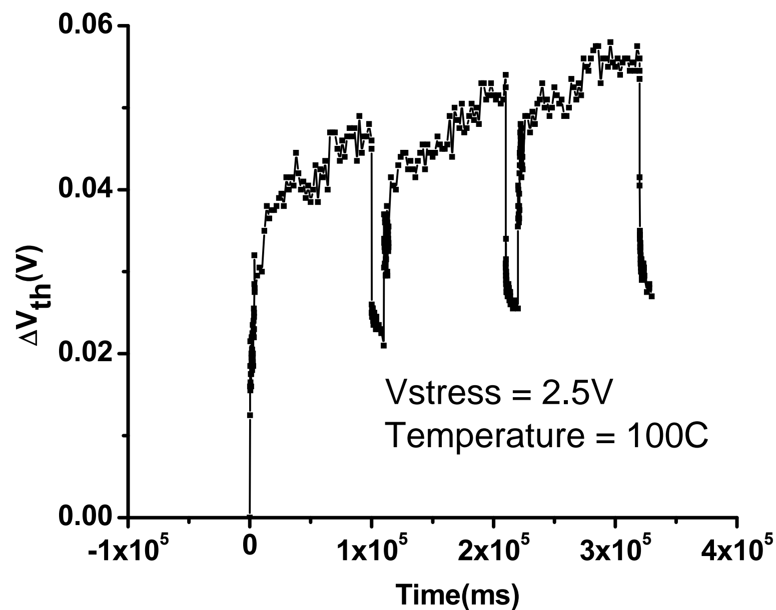


Fig. 3.10 NBTI stress and recovery measurements from the sensor (On:Off = 5:1).

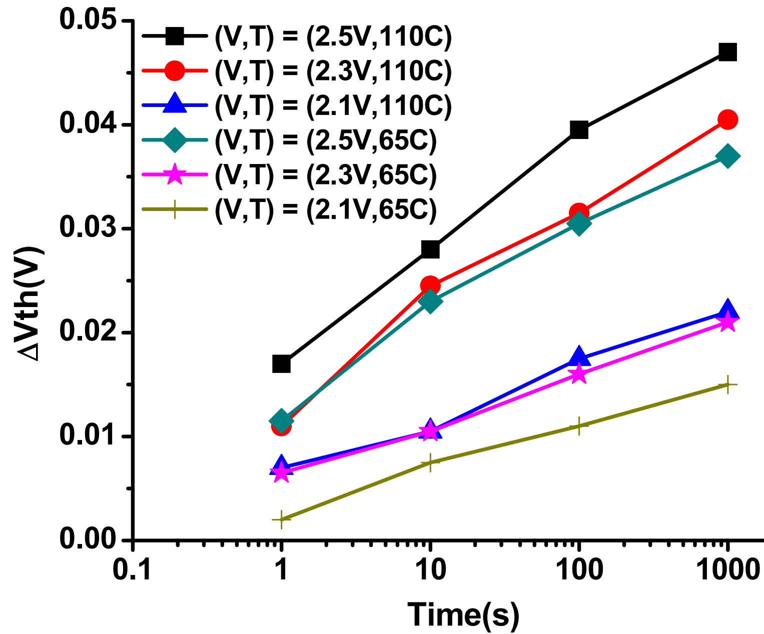


Fig. 3.11 NBTI measured under different stress conditions of voltage and temperature. As expected the degradation shows power law dependence on time.

We experimented to see the effect of dynamic temperature variation on NBTI. Fig. 3.12 shows the experiment in which the sensor was stressed at 55C for some time and then the temperature was increased to 100C. The result shows that the degradation rate increases as the temperature is increased. This happens because the number of SiH bonds broken is a strong function of temperature. So the major contribution to the change in V_{th} due to temperature mostly comes from the breaking of SiH bonds. The contribution of hole-trapping to V_{th} shift is weakly dependent on temperature and strongly dependent on the stress voltage [2.3]. It must be noted that the Si-H bonds cannot be passivated once broken. This is more evident from another experiment, in which we stressed the sensor at high voltage for some time and then reduced the temperature (Fig. 3.13). No change in degradation curve was observed as the SiH bonds cannot be passivated once they are broken and the hole-detrapping/detrapping is a weak function of temperature.

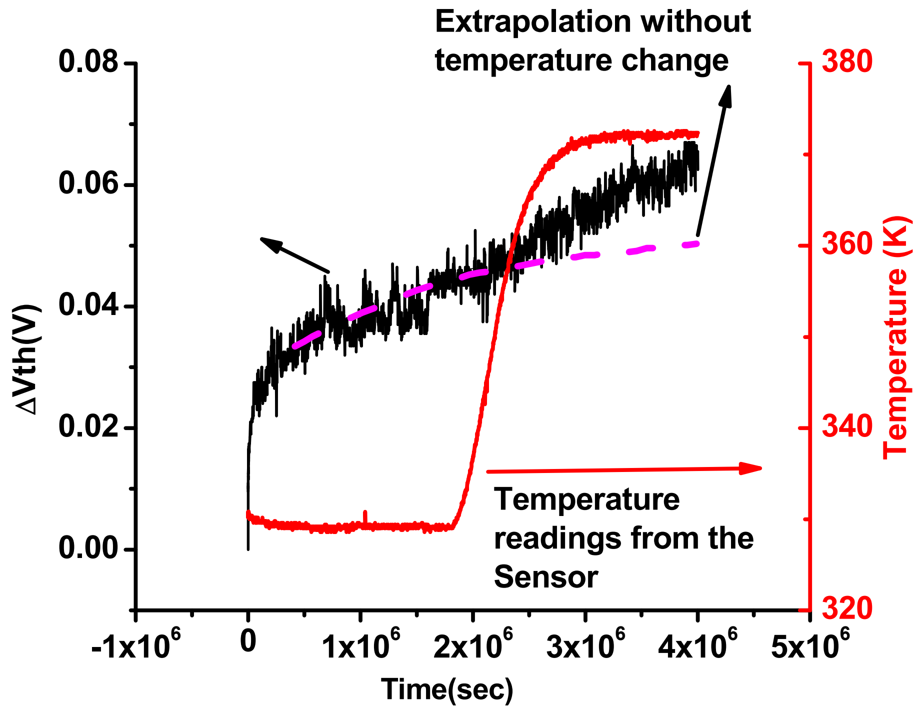


Fig. 3.12 Degradation rate increases as the temperature is increased.

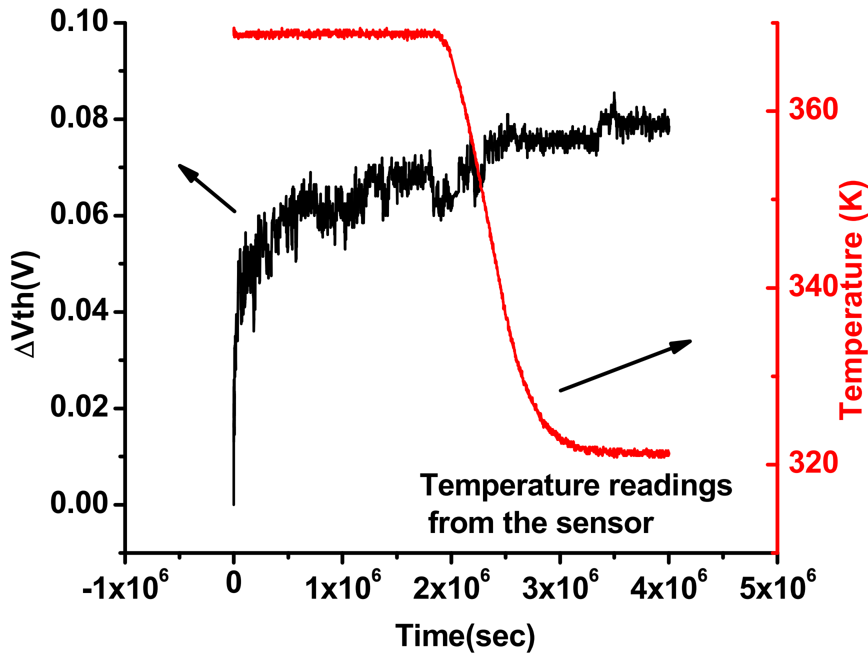


Fig. 3.13 No noticeable change in degradation rate is observed when the temperature is decreased.

The pre-NBTI V_{th} differs among devices due to process variation. Fig. 3.14 shows weak positive correlation between pre-NBTI relative V_{th} and ΔV_{th} post-stress due to NBTI. Higher V_{th} devices are more likely to have large V_{th} shifts compared to low V_{th} devices. This indicates that slow corner chips degrade at a higher rate than nominal or fast corner chips. Furthermore, slower chips may operate at higher voltages to meet performance, further accelerating their degradation.

Fig. 3.15 shows results for an implemented reliability scheme where a single active period (DC stress) is divided into ten active periods, each separated by a short sleep (recovery) time with an active: sleep ratio of 25:1. The recovery intervals reduce the degradation rate and hence can improve the overall performance and reliability of the chip.

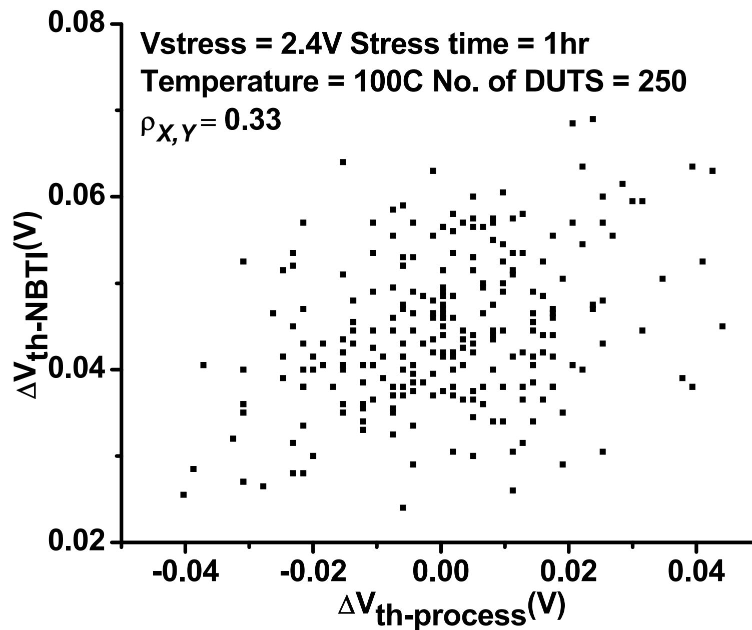


Fig. 3.14 A weak positive correlation is observed between initial V_{th} and NBTI degradation.

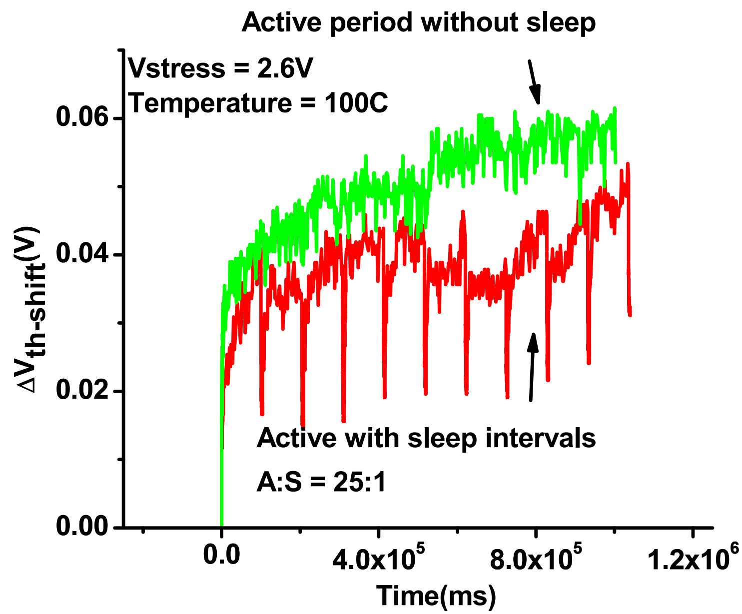


Fig. 3.15 Short sleep intervals result in 10-15% reduction in ΔV_{th} and performance improvement.

3.5 Summary

A low power unified oxide and NBTI degradation sensor is proposed. Cell power consumption is 10^5 lower than previously proposed sensors. The unified nature enables efficient reliability monitoring with reduced sensor deployment effort and overhead. The sensor sheds insight into key degradation concepts, enabling and potentially improving dynamic reliability management.

Chapter 4
Dynamic NBTI Management (DNM) Using the 45nm Unified NBTI and Oxide Degradation Sensor

4.1 Concept

As described previously due to pessimism adopted in *Static Reliability Management* excess reliability slack is present in the design. This slack can be reduced by *Dynamic Reliability Management*. To implement DRM a system must be able to estimate its reliability. This can be done in a few ways. One proposed approach involves monitoring the voltage and temperature (V, T) of the chip and then estimating its reliability using statistical models [1.6]. But these models become inaccurate and difficult to calibrate under time varying conditions of (V, T) and hence considerable margins remain.

Sensors that directly measure degradation improve on this by eliminating the need for models, thereby enabling greater accuracy of DRM. The degradation sensors lie alongside the core circuitry and the test devices embedded in the sensor are exposed to the same environmental conditions as the core. The measurement circuitry in the sensor then detects the degradation of the test device. Since the environmental conditions vary spatially across the chip, the sensors need to be sprinkled across the chip. Moreover, since the degradation is statistical in nature, hundreds or even thousands of sensors are

required in each locality to capture the statistics of the wear-out. This imposes severe constraints on the area and power of the sensors.

As shown in Fig. 4.1, $D(t_o)$ is the degradation statistics of the chip at time t_o. Using a reliability model *g* for a particular mechanism, $D(t_o)$ is extrapolated to t_{LT} to find $Dpred(t_{LT})$. If $Dpred(t_{LT})$ is less than the degradation budget, based on the workload requirements the supply voltage is boosted and/or operation at higher temperature is allowed. On the other hand if $Dpred(t_{LT})$ is higher than the degradation budget the supply voltage is lowered. Apart from designing sensors with low area and power overhead, another challenge in this approach is to come up with accurate reliability model *g*. Inaccuracy in *g* will allow more pessimism to creep in, and the benefits from DRM would be reduced.

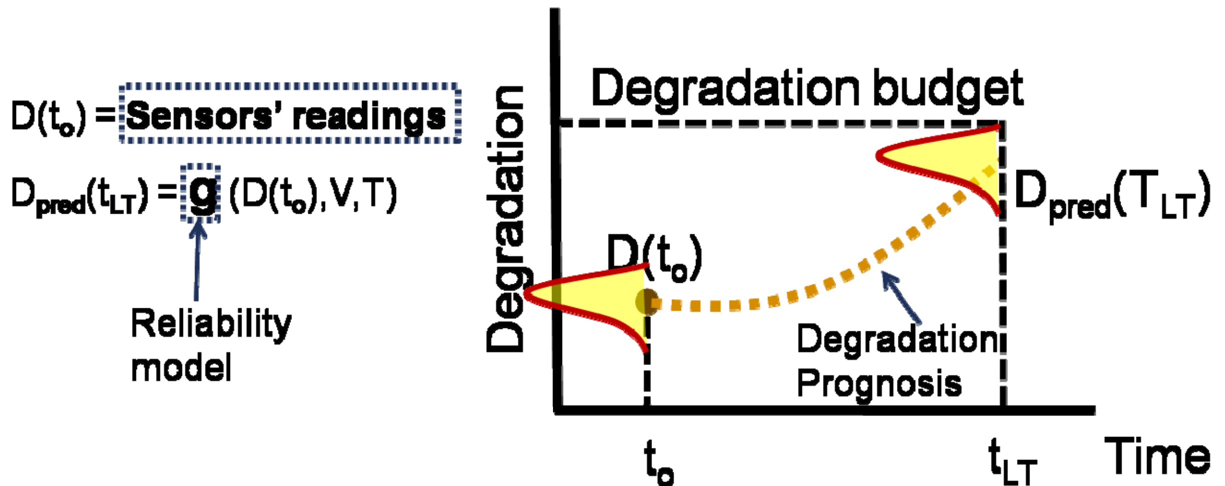


Fig. 4.1 DRM illustrations. $D(t_o)$ is the degradation distribution at time t_o. Reliability model *g* uses $D(t_o)$, voltage and temperature as input to extrapolate it to t_{LT}.

$D(t_o)$ is the degradation distribution at time t_o. Reliability model *g* uses $D(t_o)$, voltage and temperature as input to extrapolate it to t_{LT}.

Another approach to degradation sensing is to measure the degradation of the actual core devices rather than estimating it from sensors lying beside the core devices. Though this approach removes a layer of uncertainty and pessimism, the technique is more intrusive and complicates the design methodology.

4.2 Number of Sensors

We used the NBTI sensor component of the proposed sensor to perform Dynamic NBTI Management. We performed SRM using worst conditions of temperature and voltage. Since NBTI varies strongly with temperature, we found out the NBTI slack resulting from the operation at a lower temperature. We then use DNM to convert this slack to supply voltage boost and consequently to performance

Since there was no actual core implemented in our test-chip, we selected a subset of sensors from the 256 sensors to monitor NBTI while the rest of the sensors are treated as core devices. The challenge in selecting the subset was that it should be small so that the area overhead is small but at the same time it should be able to encompass the ΔV_{th} distribution of all the 256 sensors. We compared the '$\mu+3\sigma$' of sensor subsets of different sizes to the '$\mu+3\sigma$' of the 256 devices (Table 4.1). Based on that study we selected the size of the sensor subset to be 32 sensors. The normalized distribution of 32 sensor-subset and 256 devices is compared in Fig. 4.2. The subset reasonably covers the distribution of 256 devices. The location of the subset is shown in Fig. 4.3.

Sensor-subset size(n)	ΔV_{th} ($\mu+3\sigma$) (mV)
256	66
128	67.5
64	69.3
32	67.1
16	61.5

Table 4.1 Comparison of $\Delta V_{th}(\mu+3\sigma)$ of different sensor-subsets

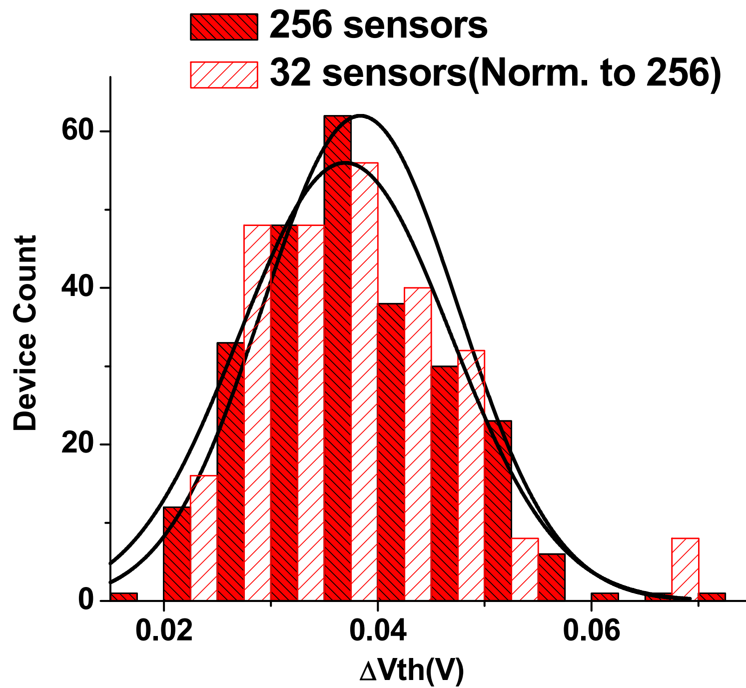


Fig. 4.2 Sensor-subset reasonably covers the ΔV_{th} distribution of 256 devices.

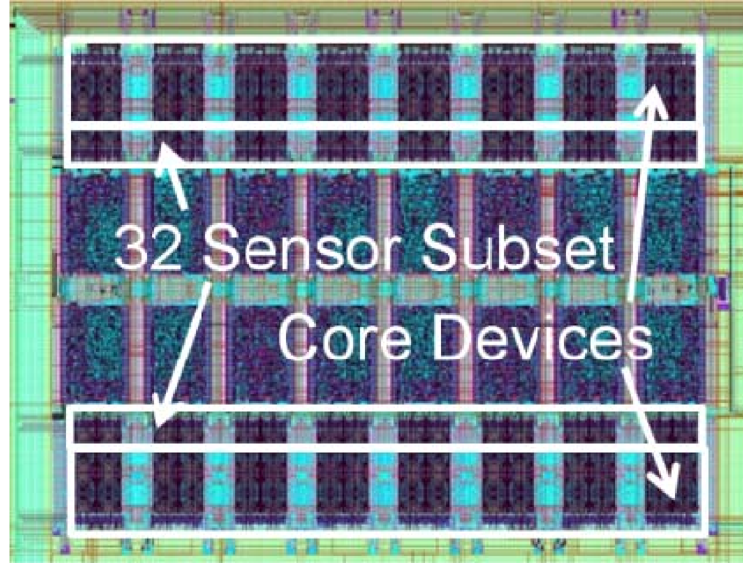


Fig. 4.3 Chip layout showing the sensor-subset and the 'core' devices.

4.3 Reliability Model

The next step is to come up with an accurate reliability model which can extrapolate the reliability history to the end of life-time. We used a power law model for this purpose which is represented by

$$\Delta V_{th}(t) = K\, t^{n} \qquad (1)$$

where K and n fitting coefficients. We used a generic power law model because to our knowledge no work has been published which models NBTI with dynamic stress voltage variation. The model is fit over the slow-saturation regime of $\Delta V_{th}(t)$ curve because that regime gives more information about the long-term prognosis of NBTI. Fig. 4.4 shows the effect of the size of fitting- window on the accuracy of predicted ΔV_{th} at the end of life-time ($\Delta V_{th\text{-}LT\text{-}PRED}$). As shown, a sampling window of 200 samples, as used in this paper, results in an accurate fit. The frequency of sample measurement is 10 seconds.

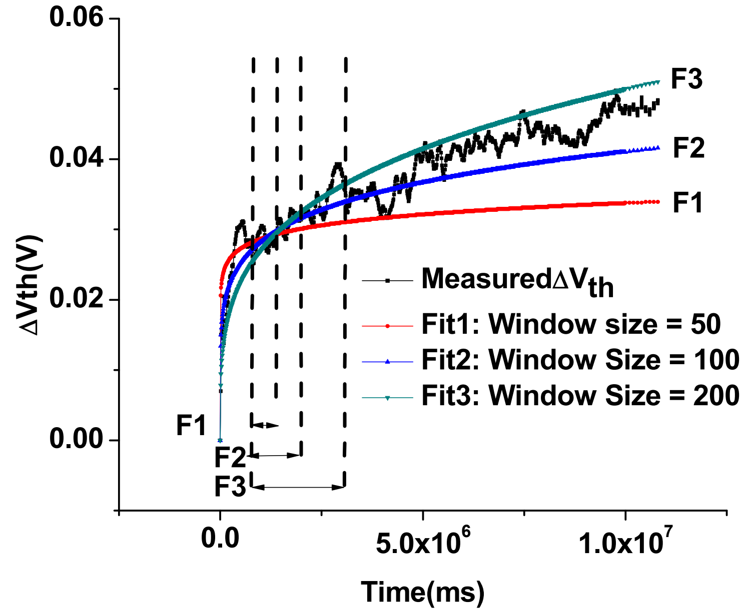


Fig. 4.4 Sampling window of 200 samples gives an accurate model fit.

4.4 Algorithm

The 32 sensor-subset is periodically monitored (every 10 sec) and ΔV_{th} is computed. Once NBTI reaches saturation regime, and the sample window is 200 samples wide, the NBTI history is fit to the power law model and extrapolated to T_{LT}. This gives another ΔV_{th} distribution of the estimated ($\Delta V_{th\text{-}LT\text{-}PRED}$) at the end of lifetime (Figure 4.5). The $\Delta V_{th\text{-}LT\text{-}PRED}$ is fitted to a Gaussian distribution and (μ+3σ) point is computed to predict the maximum ΔV_{th} ($\Delta V_{th:\mu+3\sigma}$) for all of the core devices (with 99.7% confidence). If $\Delta V_{th:\mu+3\sigma}$ > NBTI slack/budget, then the supply voltage (in our case an accelerated supply voltage), is decremented by 100mV, and vice-versa. After adjusting the voltage, we wait again till the transients are settled and we start monitoring the sensors again till a sample window of 200 samples is reached. We repeat this process till time T_{LT}. In Fig. 4.5 since $\Delta V_{th:\mu+3\sigma}$ is less than NBTI slack/budget ($\Delta V_{th\text{-}LT\text{-}MAX}$), DNM increments the stress voltage by 100mV. After this, the DNM algorithm waits for at least 100 samples to

compute the new fit, so that $\Delta V_{th}(t)$ curve has stabilized. The algorithm waits to get a good fit to the data from all the sensors before it reevaluates $\Delta V_{th:\mu+3\sigma}$.

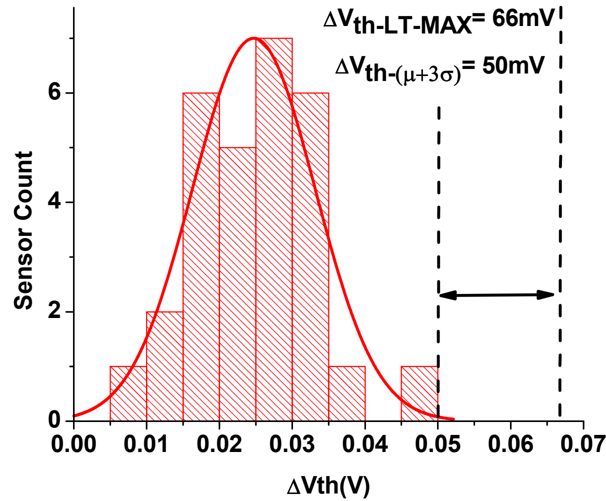


Fig. 4.5 ΔV_{th} distribution after extrapolating the degradation of 32 sensors (at certain time) to T_{LT}. The $\mu + 3\sigma$ point is compared with the budget. In this case since it has not exceeded the budget, supply voltage is increased by 100mV.

4.5 Experiment Results

We used SRM methodology to set the NBTI budget ($\Delta V_{th\text{-}LT\text{-}MAX}$). To do that 256 sensors from a chip are stressed under worst conditions of temperature (100C) and voltage (2.2V) for 3hrs, to get a ΔV_{th} distribution (Fig. 4.6). Under accelerated conditions lifetime (T_{LT}) was 'assumed' to be 3hrs. ($\mu + 3\sigma$) point of the distribution is computed to set $\Delta V_{th\text{-}LT\text{-}MAX}$ to 66mV.

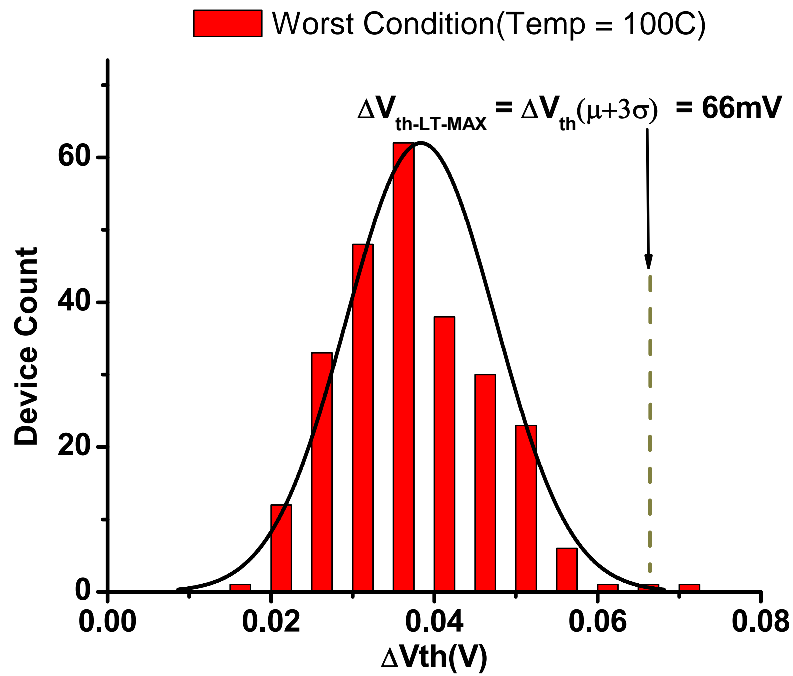


Fig. 4.6 SRM under worst conditions (Temp = 100C, Vstress = 2.2V).

To find the NBTI slack we repeat the same experiment for another chip but at a more typical temperature of 55C (Fig. 4.7). As can be seen in the figure the $\Delta V_{th\text{-}margin}$ = 22.5mV.

For a chip operating at typical temperature (55C), our implementation of DNM aims to consume this slack by operating at higher voltage and consequently boosting the performance.

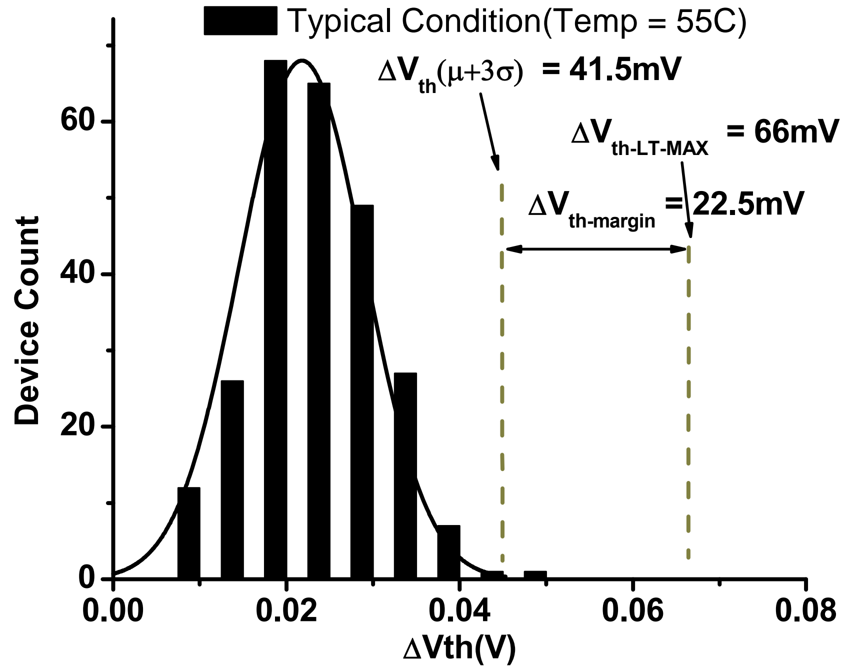


Fig. 4.7 Lowering the temperature increases NBTI margin.

Fig. 4.8 shows the readings from one of the sensors whose stress voltage is controlled by DNM. It also shows the corresponding model fit and the stress voltage scaling governed by DNM. As can be seen, the supply voltage is adjusted twice in time T_{LT}. The data processing and the DRM algorithm are implemented in MATLAB. Fig. 4.9 shows the measured ΔV_{th} distribution for all the 256 sensors after DNM at 55C. $\Delta V_{th}(\mu+3\sigma)$ for the distribution is 58mV which implies 14.5 mV of excess slack was consumed (when compared to No-DNM at 55C), whereas an average boost of 90mV in the accelerated supply voltage was obtained.

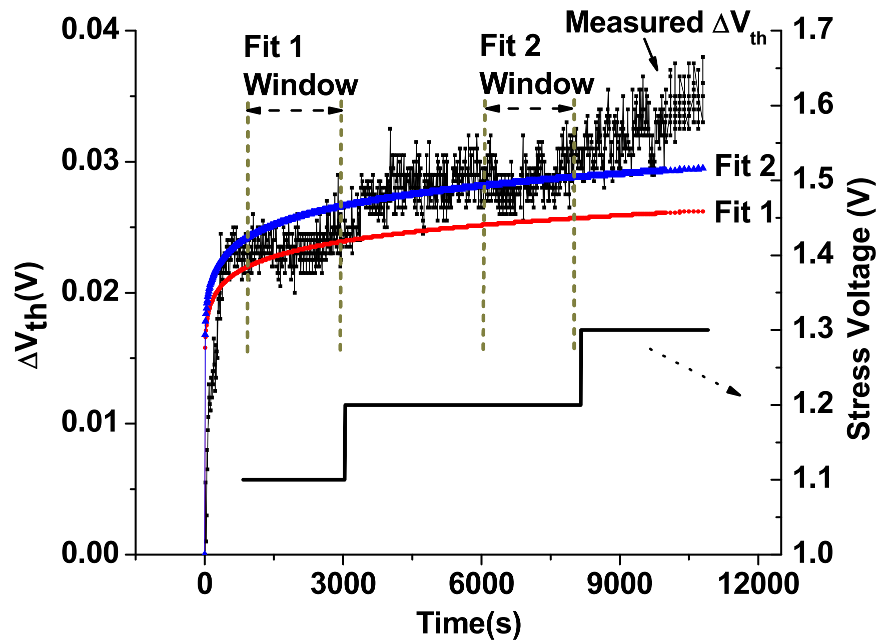


Fig. 4.8 One of the 32 sensors' readings, with model fit and supply voltage scaling in a DNM implementation.

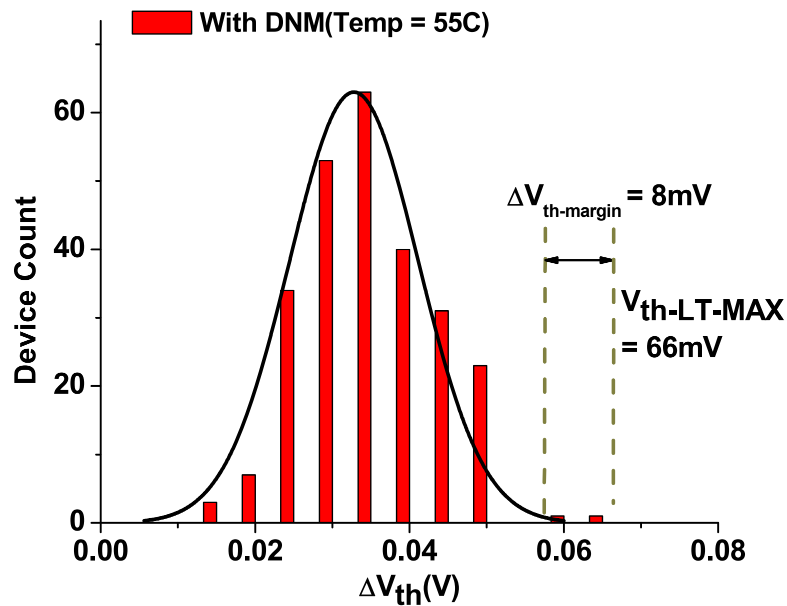


Fig. 4.9 ΔV_{th} distribution at T_{LT} with DNM at 55C. The slack was converted into performance by DNM.

4.6 Conclusion

DRM aims to trade the unused reliability margins present in a chip (due to static reliability margining) with performance. We proposed a unified NBTI and oxide wear-out sensor in 45nm process node which enables efficient DRM. The low area and power consumption (> 10^5 times lower then a previous sensor) of the sensor enables their use in large numbers. To our knowledge, a sensor based DNM was tested in silicon for the first time. For the typical case, the proposed DNM allows for an average boost of 90mV in the accelerated supply voltage while reducing the excess NBTI margin of 22.5mV to 8mV where the total budget for NBTI was 66mV.

Chapter 5
Early Detection of Oxide Breakdown Through *In situ* Degradation Sensing

So far we have talked about degradation sensors which are used to collect the statistics of degradation, and then the statistics are used to estimate the degradation of the core devices. But due to dependence of degradation on the process mismatch, the circuit state dependence and the inherent statistical nature of the degradation, the sensor based approach does not give an accurate estimate of the degradation of the core devices. Due to lack of accuracy pessimism is adopted in the degradation estimation. The only way to remove this layer of uncertainty or pessimism is to measure the degradation of the actual core devices of interest. The benefits of this approach would be that since it eliminates almost all of the pessimism in the estimation of degradation, and hence DRM controller can be more aggressive with supply voltage scaling and setting temperature limits. The down side of this approach is that it may make the design methodology of the core more intrusive.

In the context of gate-oxide break down (OBD) *in situ* detection becomes more important because gate-oxide failure of even one transistor may cause a chip to fail. The statistics from the degradation sensors are useful but it may not be able to capture the outlier devices in the core. Moreover to ensure that corrective measures are taken well before a device fails completely it is critical that degradation monitoring detects the *onset*

of breakdown (or "soft breakdown"), when the oxide becomes leaky yet the device continues to function correctly. This can be attained only through *in situ* degradation monitoring.

5.1 Previous *In situ* OBD Detection Techniques

To perform *in situ* OBD there have a couple of approaches based on measurement of change in delay. For *in situ* OBD detection, path delay monitors were proposed [5.1, 5.2]. However, the delay change of a single gate that is close to failure may be obscured by other longer paths. Also it is hard to isolate delay change due to degradation in the presence of varying operating conditions. Furthermore, we experimentally show that at the onset of soft breakdown a gate delay may increase by only 2 – 4%, and in some cases can actually improve, complicating this approach. Recently, OBD detection for a power amplifier by measuring oxide resistance using resistors and a pre-driver stage was shown [3.5]. This technique was targeted for very specific transistors in the circuit and cannot be extended to a large circuit block without significant area and power overhead.

5.2 Proposed *In situ* OBD Detection Technique

Gate leakage is the most direct measure of gate oxide degradation. However, it is difficult to measure the gate current in the presence of background currents such as subthreshold leakage, band-to-band tunneling, and gate-induced drain leakage. In addition, voltage and temperature strongly impact the total measured current, further complicating the measurement of changes in gate current. In our proposed approach the key in detecting oxide degradation is sensing the change in the Ig-Vgs characteristics of a degraded device. In [5.3] the leakage of a degraded gate oxide is modeled as follows:

$$I_g = K\,(V_{gs})^P, \qquad (1)$$

where I_g is gate current and V_{gs} is gate-source voltage.

Fig. 5.1 illustrates how the values of K and P vary with degradation in time and consequently how the gate leakage increases. The values of K and P are based on measurements reported in [5.3].

As the gate oxide degrades, defects, or trap sites, are formed in the oxide. The defects are at lower energy levels than the barrier height introduced by the insulator and hence the defects alter the exponential dependence of the gate current on voltage across the gate oxide. Due to this phenomenon the Ig-Vgs characteristics of the device become more linear as a device degrades, which is illustrated by progressively straighter lines in Fig. 5.2. It is this key behavior that we

use to detect the degradation of a device. We monitor this behavior for a cluster of gates to reduce monitoring overhead.

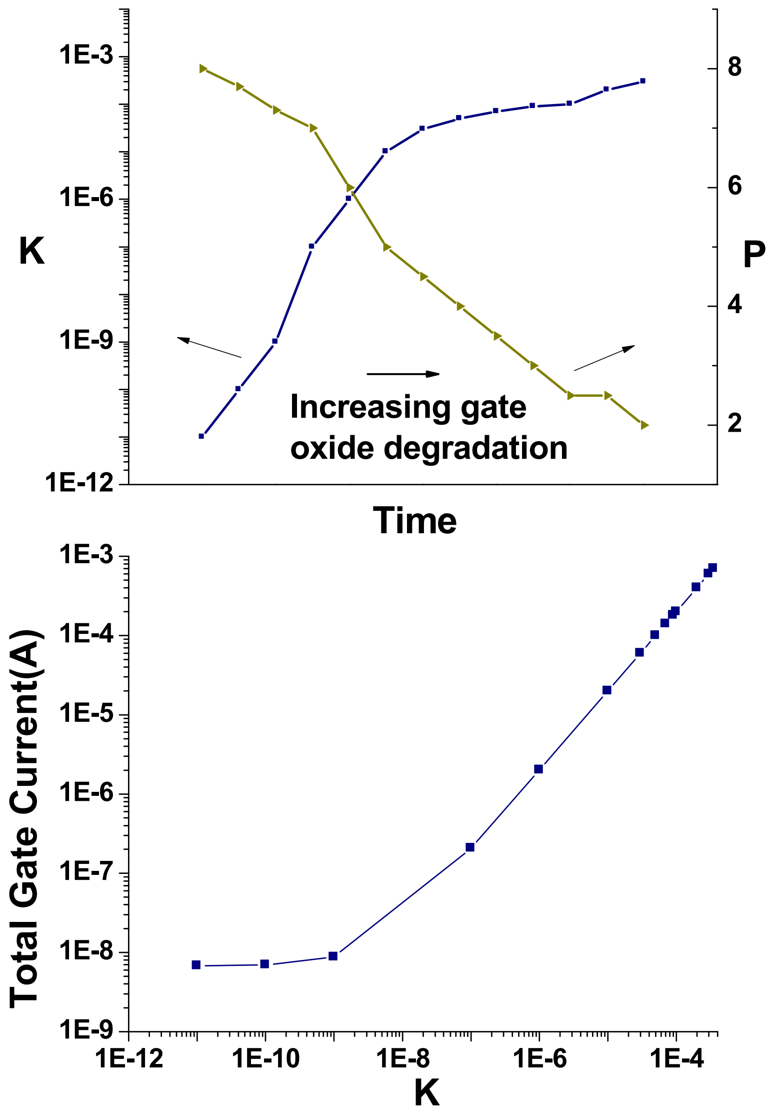


Fig. 5.1 (Top) Increasing gate-oxide degradation can be modeled as an increase in the value of K and decrease in the value of P.
(Bottom) The gate current increases by orders of magnitude due to gate-oxide degradation.

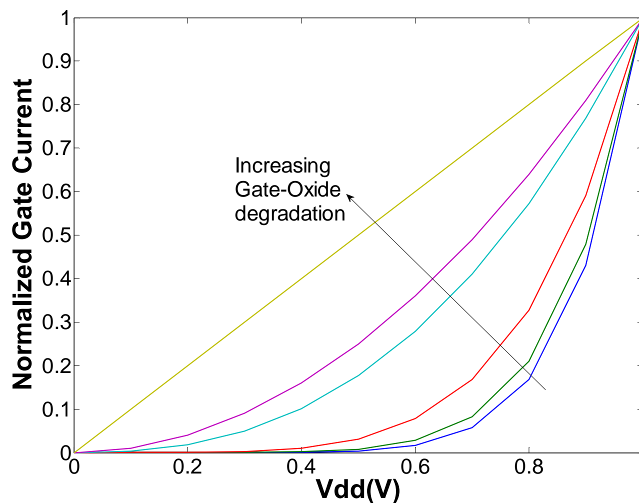


Fig. 5.2 The non linear nature of the Ig-Vgs characteristics of the gate-oxide becomes more linear with degradation.

As introduced in [5.4], the proposed *in situ* monitoring implementation leverages the prevalence of MTCMOS-based designs with PMOS header switches, a common technique to reduce standby power [5.5] with relatively low overhead. Fig. 5.3 shows a circuit block partitioned into clusters, each connected to the power supply through a standard high-Vt MTCMOS switch and a weak PMOS device (WP) with a controllable gate voltage, Vbias. The design is periodically taken offline and tested for oxide degradation by sweeping the gate voltage of WP from 0 to VDD using an on-chip DAC, while the virtual rail voltage is recorded using an

on-chip ADC. The resulting Vbias vs. Vrail (V/V) curve is then analyzed to detect the onset of oxide breakdown (OBD). Initially, both oxide leakage and sub-threshold leakage are strongly non-linear with supply voltage, resulting in a characteristic "hockey stick" curve. However, as the device degrades the V/V curve flattens (Fig. 5.4). Based on this key behavior shift, we define a figure of merit called the degradation voltage angle (DVA) that measures the angle of a straight line fitted to the V/V curve over the 90 –

10% Vrail interval. As a gate degrades, the oxide displays more linear resistive behavior and a sharp drop in DVA is observed (Fig. 5.5).

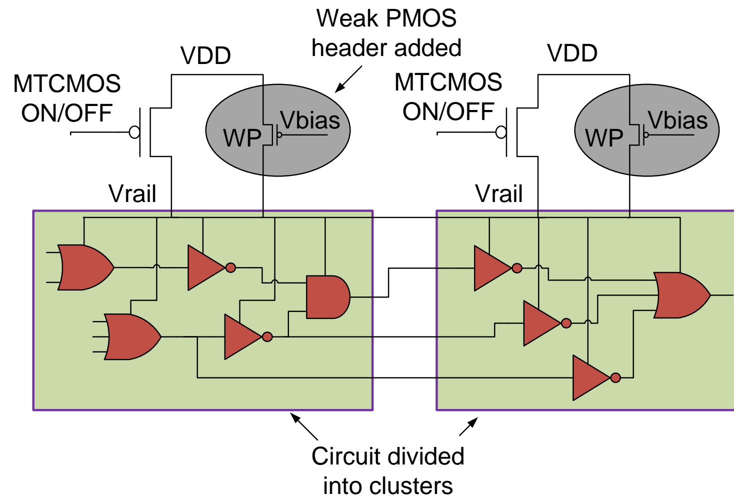


Fig. 5.3 The *in situ* monitoring technique is implemented by dividing a circuit into clusters using MTCMOS headers. Weak PMOS headers are added to monitor the conductance of the clusters.

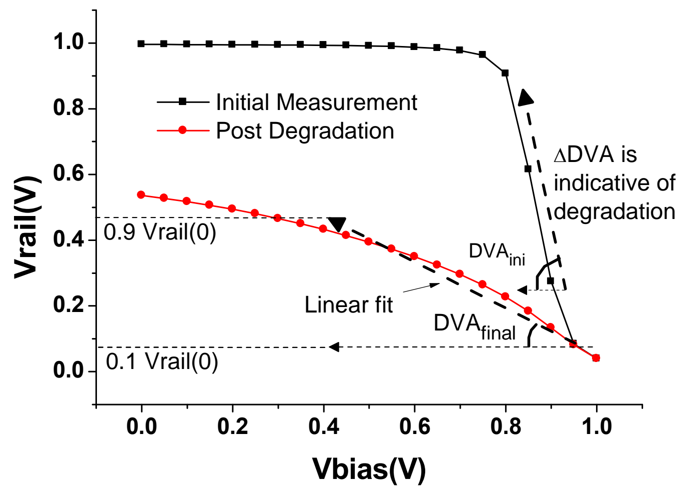


Fig. 5.4 The nature of Vrail vs. Vbias curve changes with degradation. DVA is defined to quantify this change in behavior.

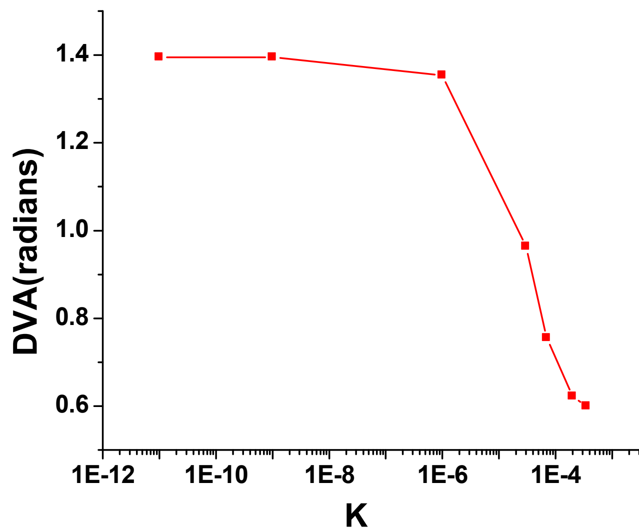


Fig. 5.5 DVA drops sharply with degradation which flags the onset of breakdown.

5.3 Test Chip and Silicon Measurements

The technique was implemented in two test chips fabricated in 65nm CMOS. The first chip applies the technique to individual gates and parity trees for a detailed study of the OBD effect while the second chip implements a FIR filter to demonstrate applicability to larger circuit blocks. The technique can be applied to larger designs in which case, the overhead would be further amortized.

The first chip consists of an array of 12 INV, 2 NAND, and 2 NOR gates (referred to as GUTs) as well as 5 parity circuits ranging from 64 – 1024 gates (Fig. 5.6 and Fig. 5.7). Three modes of operation are supported: stress mode, oscillation mode (to measure performance degradation), and OBD measurement mode. To accelerate stress for testing purposes a PMOS header switch (TP) is implemented using a thick oxide transistor to transfer the stress voltage (VddST) to the virtual rail (STR_OP = 0, Fig. 5.6). A thick oxide NMOS device TN protects the weak PMOS (WP) from degradation during

accelerated testing (note that during regular non-accelerated operation all terminals of WP are at Vdd which prevents its degradation). During oscillation mode, VddST is brought to 1.0V and TP remains on while the circuit is placed in a feedback loop. To isolate the GUT delay, the measurement is repeated with the GUTs bypassed to allow the non-GUT portion of the delay to be subtracted. During OBD measurement TP is super-cutoff while TN is ON and Vbias is swept to record measurements.

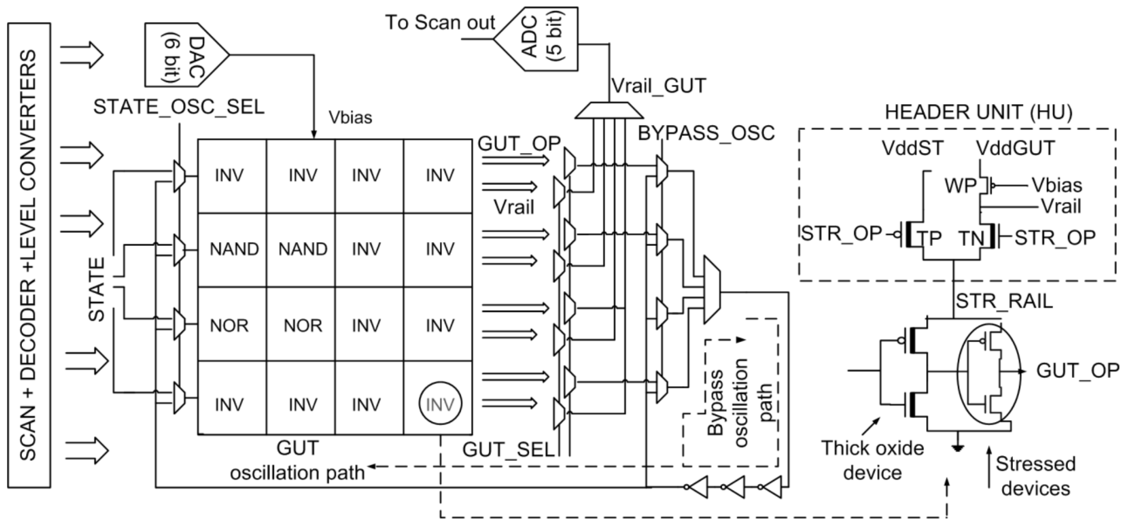


Fig. 5.6 OBD detection technique implemented on GUTs.

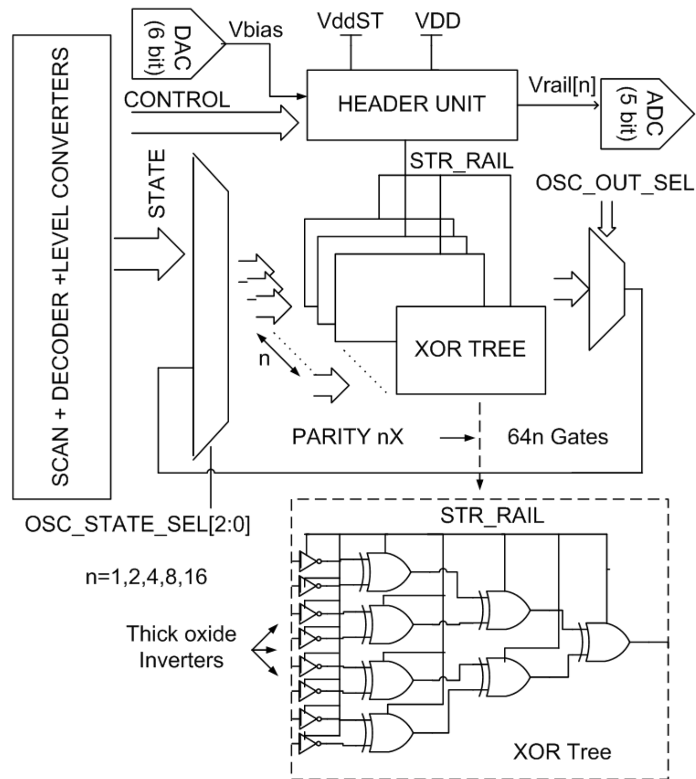


Fig. 5.7 OBD detection technique implemented on XOR Parity trees of different sizes.

Fig. 5.8 shows the measured V/V, DVA, and gate delay curves for an inverter. We define 15% drop in DVA as the detection point of OBD onset. In this case the proposed technique detects the onset of gate-oxide degradation with as little as a 3% increase in the delay of an individual gate. This illustrates that delay monitors such as the one proposed in [5.1, 5.2, 5.6] will not be able to detect degradation until the delay is severely affected. Impact of stress on delay shows non-monotone behavior, at times resulting in faster gate delays. This is expected and is caused by the suppression of voltage swing under certain failure modes [5.3]. The corresponding degradation is captured by the DVA. Eventually the gate delay increases to 20X and then fails completely.

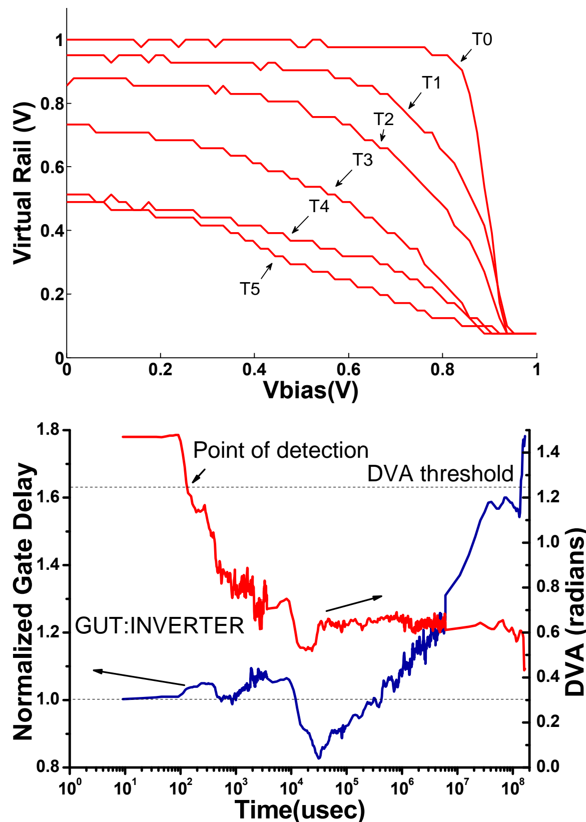


Fig. 5.8 (Top) Silicon measurement of Vrail vs Vbias curve for a stressed INVERTER at different points of degradation.
(Bottom) Silicon measurement of DVA and Delay of a stressed INVERTER.

Since the subthreshold current is a strong function of temperature, the leakage of the non-failing gates can overwhelm the change in the gate leakage of the failing gate. To determine the sensitivity of the V/V curve to temperature, Fig. 5.9 shows the V/V, DVA, and path delay curves for an XOR tree consisting of 64 gates at 25C and 125C. There is a 10X change in subthreshold leakage for this 100C change in temperature, which changes the V/V curves significantly. However, the DVA metric changes by only 7% (less than the picked failure detection threshold of 15%), showing that DVA provides a robust measure of gate oxide degradation across temperature. The excellent robustness of DVA

metric eliminates the need to calibrate or compensate for temperature, which would increase the complexity of the approach.

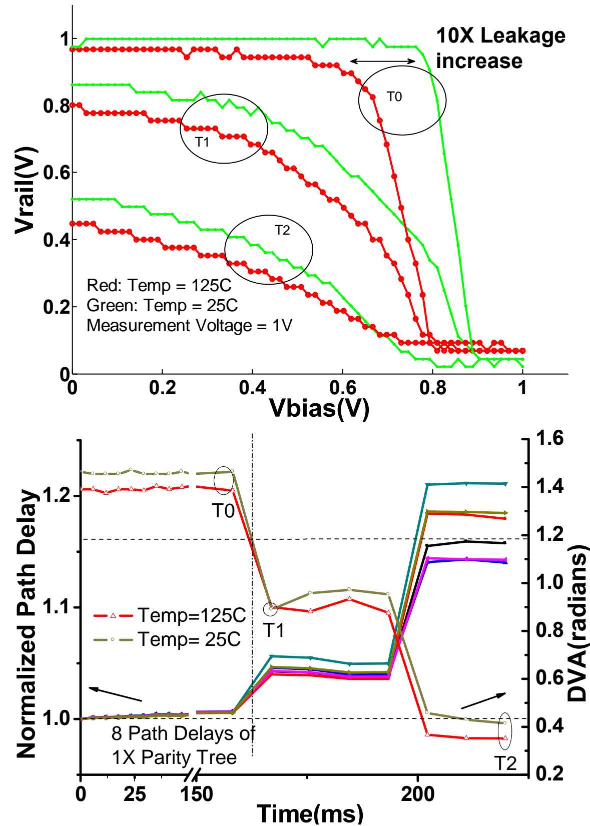


Fig. 5.9 (Top) Silicon measurement of Vrail vs Vbias curves of a 64 gate XOR parity tree at 25C and 125C at three different points in degradation showing immunity of DVA measure to environmental conditions.
(Bottom) Measured DVA (at 25C and 125C) and Delay of a stressed 64 gate XOR parity tree.

We conducted stress experiments on the XOR parity trees as well. Fig. 5.10 shows the DVA measurements and the delay of the eight paths of a 64 gate XOR parity tree. The measured DVA decreases as the gate-oxides of the devices in the XOR parity tree degrade. Eventually one of the 8 paths of the circuit fails while the DVA decreases sharply at that time.

Fig. 5.11 shows the effectiveness of this technique as the cluster size varies. As the cluster size increases past 512 gates the failure detection is delayed since the leakage increase of the failing gate(s) is masked by the background leakage of the non-failing gates. It must be noted here that in a parity tree all the paths are critical. Hence degradation in any transistor would have a performance penalty. In an actual microprocessor the number of critical paths is a very small fraction of the total number of paths. So, the performance penalty at the time of wear-out detection would be much less.

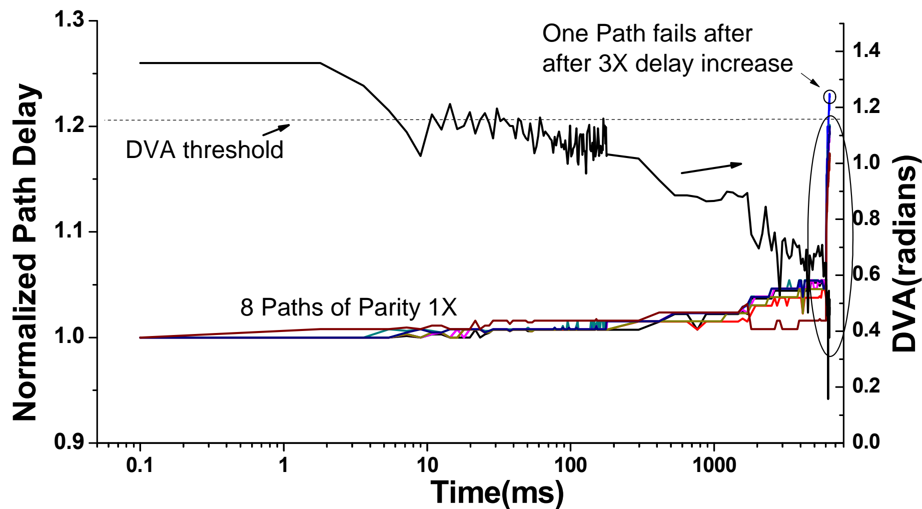


Fig. 5.10 Delay and DVA measurements for an XOR parity tree with 64 gates.

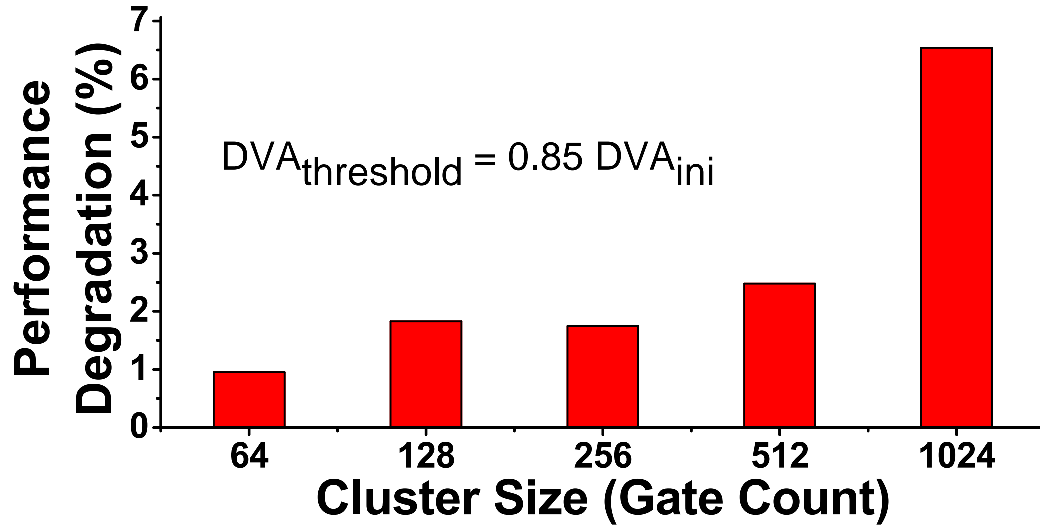


Fig. 5.11 Measurements show that the time to detection of onset of degradation increases with cluster sizes larger than 512 gates.

Fig. 5.12 shows a large variation in time to onset of OBD for 63 inverters, illustrating the statistical nature of oxide degradation.

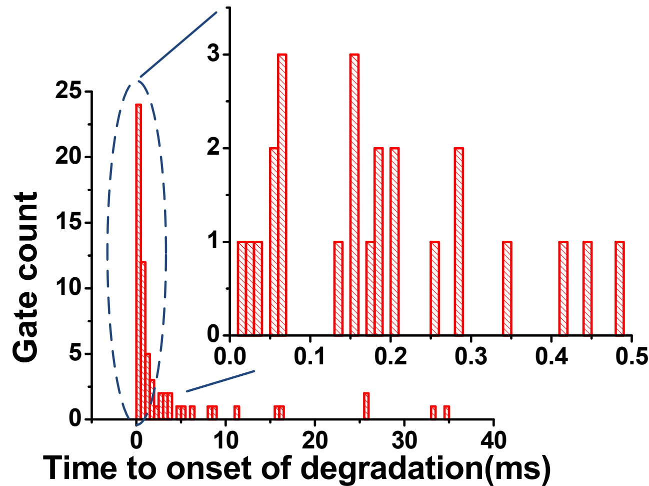


Fig. 5.12 Measurements show a large variation in the time to onset of gate-oxide degradation.

The second test chip applies the approach to a 16-bit, 8-tap FIR filter consisting of 7K gates (Fig. 5.13). The FIR is divided into 360 blocks of ~20 gates placed into 36 rows

and 10 columns. To monitor each of the 360 virtual rails (VRs), a low leakage 360x1 two-stage mux is used. Since the VRs are driven by small leakage currents it is extremely important to isolate the selected VR. To this end, a unity gain buffer mirrors the voltage seen on the selected VR onto the other VRs. The thick oxide devices in the HU result in a large overhead in this case, but are only required for accelerated testing. In a regular design with only thin oxide devices the design area overhead is 17% compared to a design without MTCMOS and 5% compared to a standard MTCMOS design. The overhead can be reduced by increasing the block size, which would reduce the number of HUs and VRs.

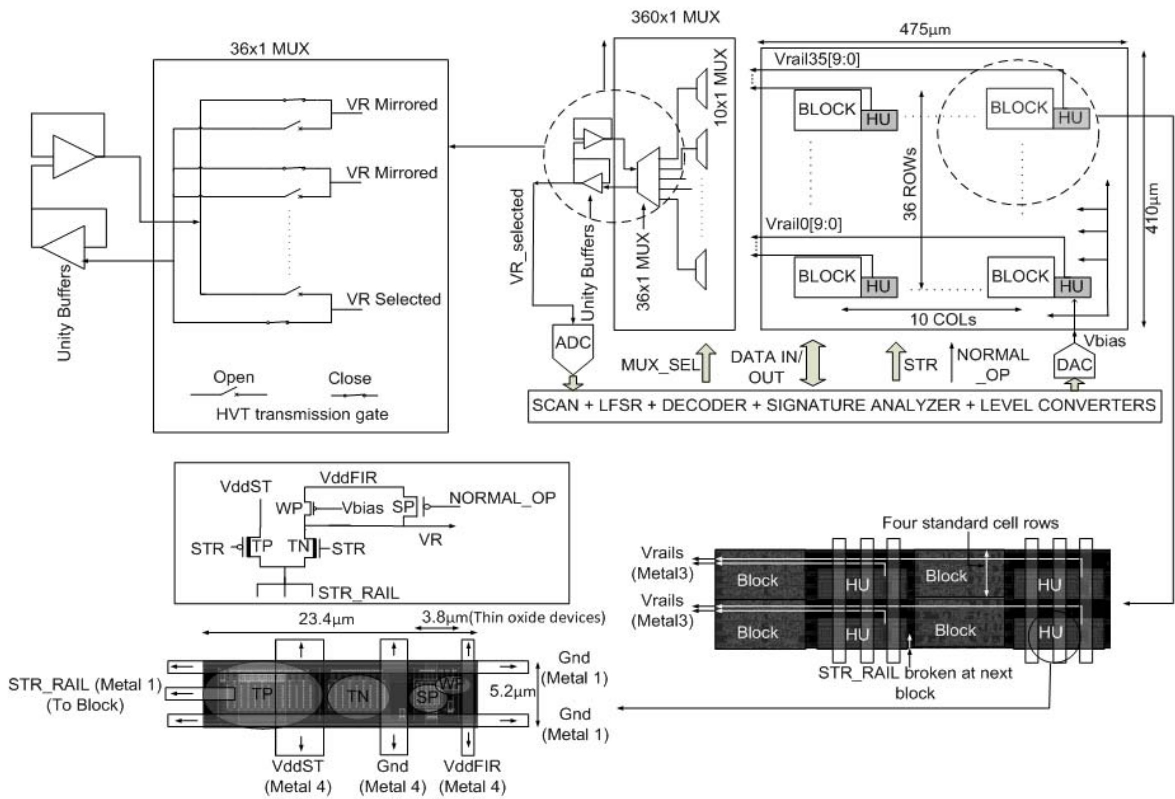


Fig. 5.13 Sensing circuit implemented on 16 bit, 8-tap FIR filter. 141 blocks were stressed and monitored.

We conducted stress and DVA measurements on the FIR filter. In this experiment 141 clusters of the FIR filter were stressed and periodically measured for performance and DVA.

Fig. 5.14 shows the accumulation of block failures over time and their spatial distribution. The first detected block failure corresponds to a performance degradation of the FIR by 0.5%.

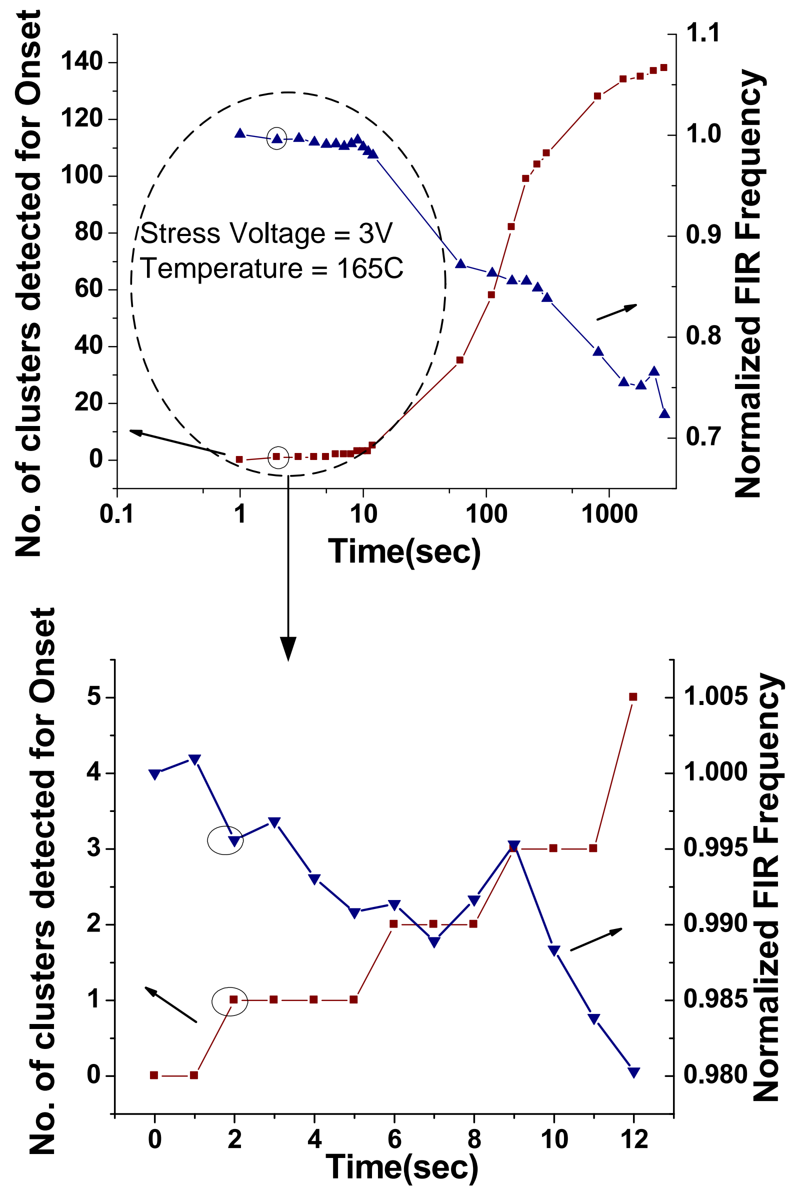


Fig. 5.14 (Top) The performance of the FIR degrades as clusters are detected with onset of gate-oxide degradation.
(Bottom) The performance degradation is 0.5% when the first cluster is flagged for onset of degradation.

Fig. 5.15 shows the VR vs Vbias curves for the first cluster to fail (i.e. the first cluster to cross the DVA threshold), at different points in time.

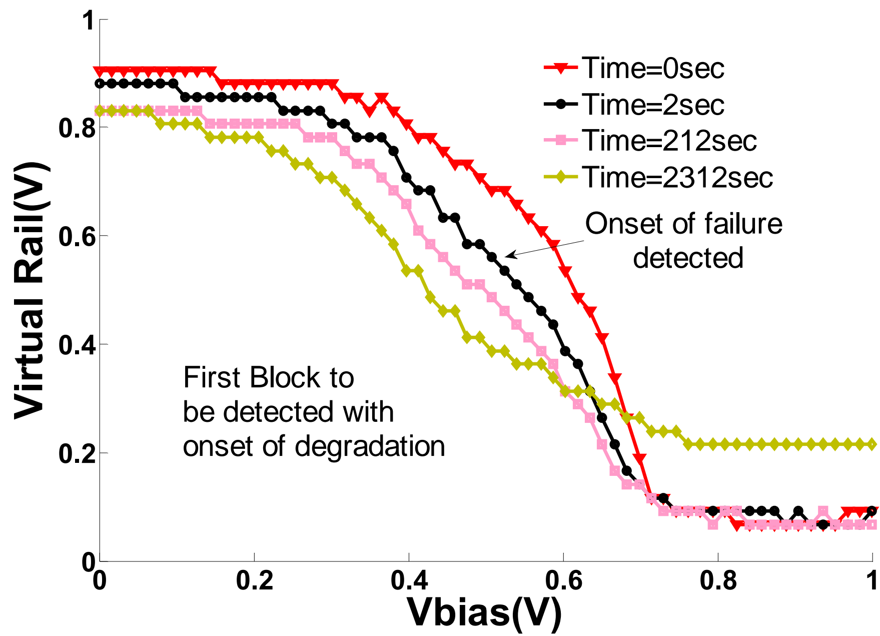


Fig. 5.15 Virtual Rail vs Vbias measurements for the first cluster to be detected with failure, at different points in time.

We monitored the failing clusters and constructed a spatial map of the clusters, based on their physical location, indicating the times at which they were detected for failure. Fig. 5.16 shows this spatial map. We did not observe any spatial correlation in the times of failure detection.

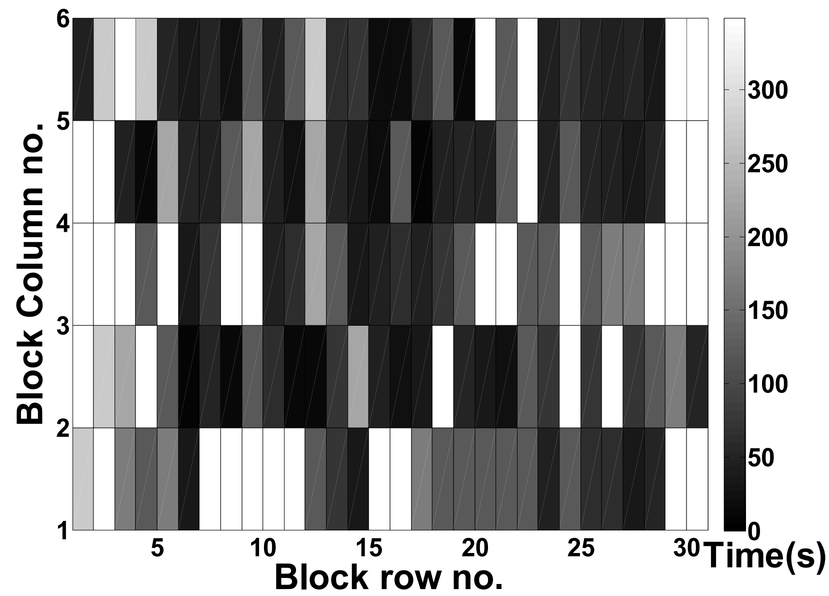


Fig. 5.16 Spatial map of the clusters indicating their time of failure detection.

Fig. 5.17 shows the normalized gate-leakage for 141 stressed clusters with time.

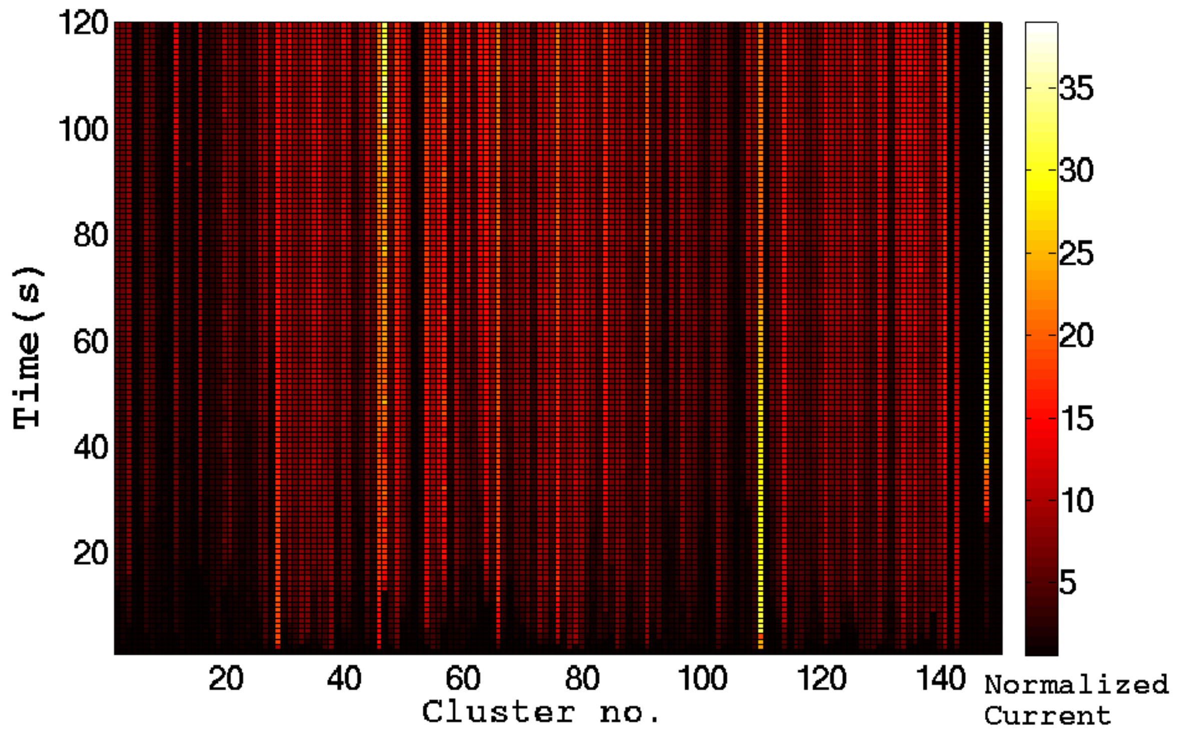


Fig. 5.17 Normalized current of clusters with stressed time.

Fig 5.18 shows the die photo of the two test-chips.

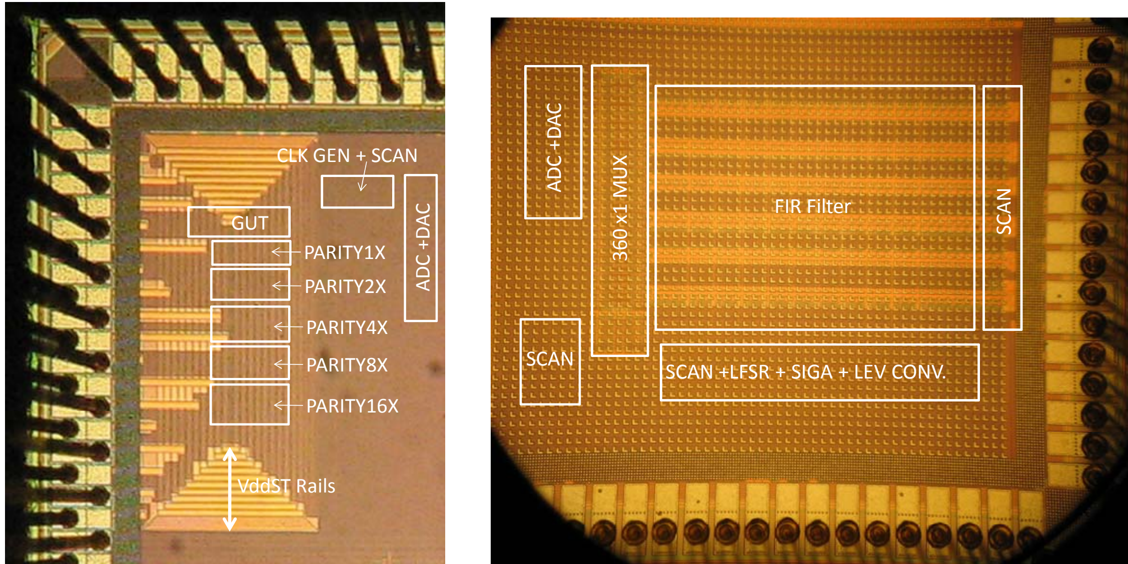


Fig. 5.18 (Left) Chip 1 microphotograph (Right) Chip 2 microphotograph

5.4 More Silicon Measurements

During the stress experiments on individual gates we observed the phenomenon of gate-oxide recovery. Fig. 5.19 shows the silicon results of the stress experiment on a NOR gate. In the circled region, the delay of the gate increases and the DVA decreases. But the delay recovers after sometime and so does the DVA. This is explained by the neutralization of the electron and hole traps formed in the gate-oxide [5.7].

We performed some stress experiments on the FIR filter at lower stress voltages (2V-2.5V) and monitored the leakage current in the FIR filter. Fig. 5.20 shows the results when the stress voltage is 2V. Initially the total leakage increases due to gate-oxide wear-out which causes the gate-leakage component to increase. However, as more gate-oxide

wear-out occurs the threshold voltage of the degrading devices increase and consequently their sub-threshold leakage decreases and the total leakage decreases.

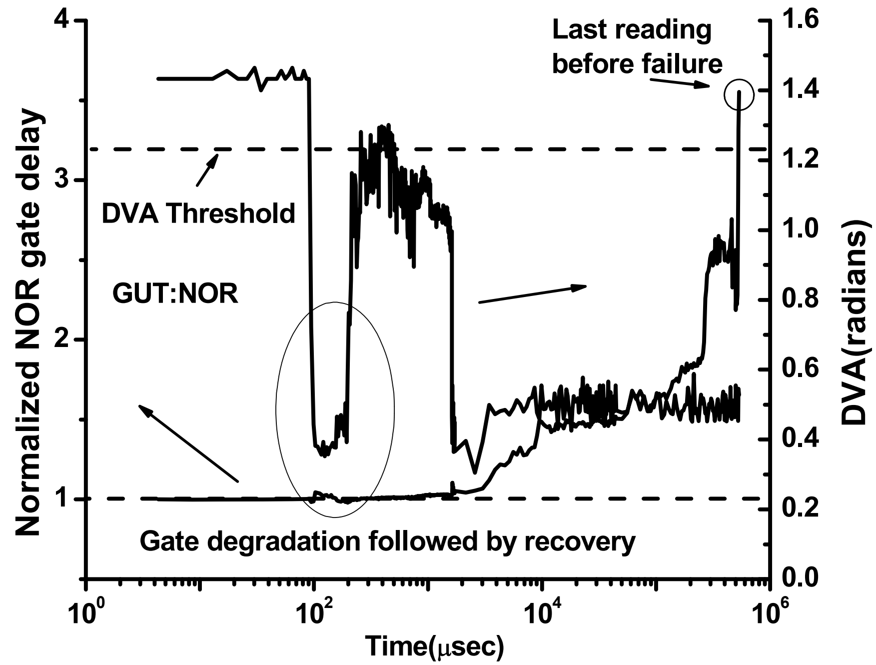


Fig. 5.19 Gate-oxide degradation recovery observed.

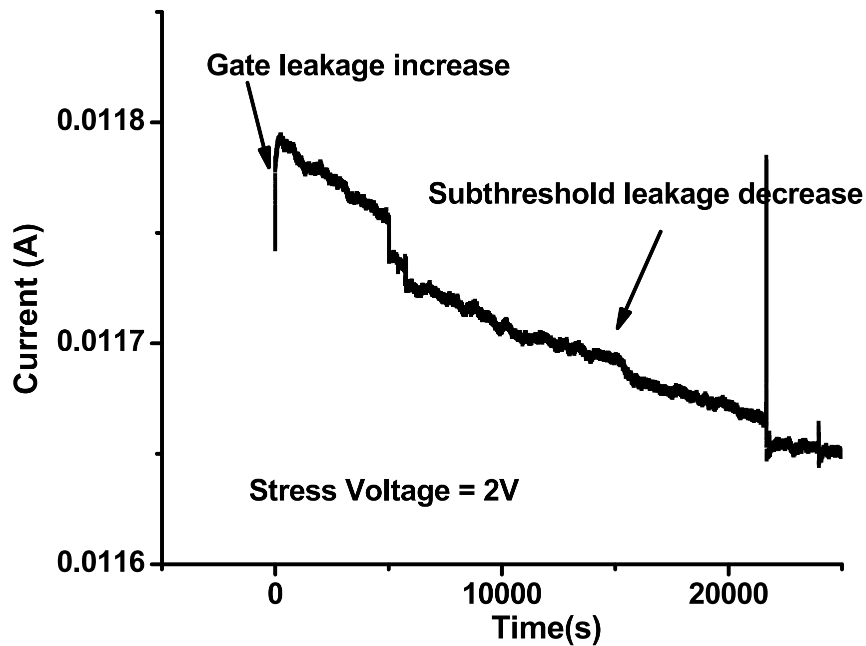


Fig. 5.20 Simultaneous effect of gate leakage increase and threshold voltage shift at lower stress voltage.

Fig. 5.21 shows the leakage results at a higher stress voltage (2.1V). Similar trend is observed but this time the gate-leakage increase rate is higher. This is evident from the fact that after sometime the gate-leakage starts overwhelming the decrease in sub-threshold leakage and the total leakage starts increasing.

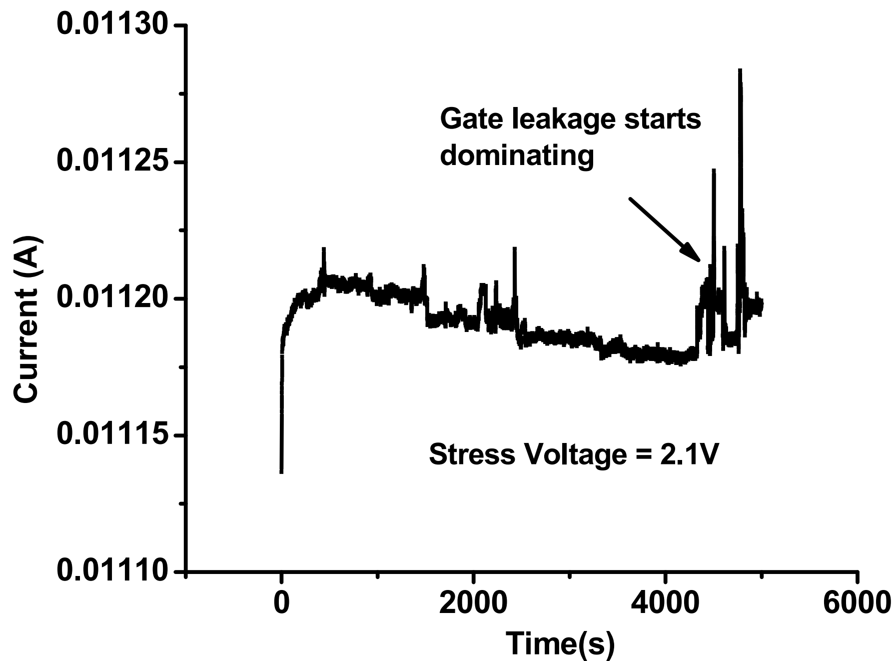


Fig. 5.21 Simultaneous effect of gate leakage increase and threshold voltage shift at lower higher stress voltage.

We repeated this experiment at 2V but interrupted the stress after every 2000sec for 20min (Fig. 5.22). During this interruption the supply voltage was reduced to 1V. After 20 min stress is resumed by increasing supply voltage to 2V. When the stress is resumed after interruption the leakage is considerably reduced from the pre-interruption period and slowly it starts increasing again. We cannot account for this behavior after stress resumption. If the decrease in leakage is attributed to decrease in subthreshold leakage then the Vth of the devices must increase further during the interruption i.e. when the supply voltage is low. No physical reasoning seems apt for this behavior.

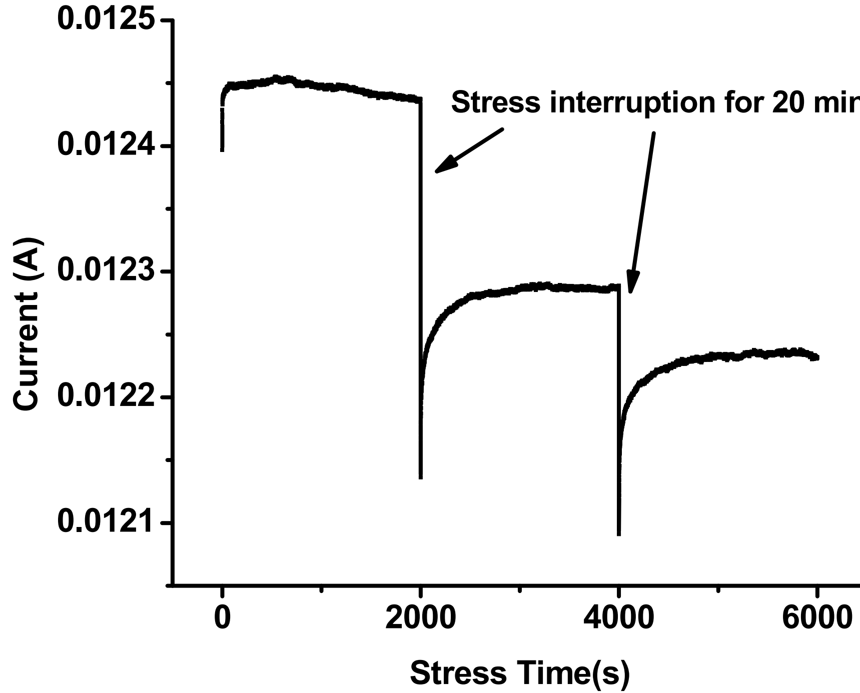


Fig. 5.22 Stress interruption results in lower leakage current.

5.5 Summary

We proposed a method for *in situ* monitoring of the gate oxide degradation of the *actual* devices of the core. The detection is based on the change in gate-oxide's I-V characteristics from non-linear to linear. To quantify this change we defined a metric called Degradation Voltage Angle (DVA). The DVA measurement on a FIR filter shows that it captures the gate-oxide wear out very early in the life-time when the performance degradation is as low as 0.5%. Hence this technique provides sufficient time for a DRM controller to manage chip reliability.

Chapter 6
In situ Bias Temperature Instability (BTI) Sensing Technique and Dynamic BTI Management Implementation

6.1 Introduction

In the previous chapters we have talked about the impact of NBTI degradation in advanced process nodes and why it is becoming a serious concern for designers and process engineers. While NBTI affects the threshold voltage of a PMOS device, there is an analogous degradation mechanism in NMOS devices, Positive Bias Temperature Instability (PBTI) which increases the threshold voltage of an NMOS device [6.1]. As with NBTI, the phenomenon of recovery accompanies PBTI as well. Even though NBTI emerged as a major concern with the advent of sub-100nm process nodes, PBTI is also becoming a serious concern in advanced process nodes, especially with the inclusion of high K metal gate dielectric [6.2]. Hence in the context of Dynamic Reliability Management (DRM), it is important to detect PBTI along with NBTI. Traditionally both degradation mechanisms are collectively termed as Bias Temperature Instability (BTI).

We have already talked about how *in situ* detection takes into account all the sources of lifetime variation and hence gives a more accurate measurement when compared to

degradation sensors. In this chapter we will propose an *in situ* technique to measure BTI and then demonstrate Dynamic BTI Management (DMB) using the *in situ* technique.

6.2 Concept

Due to BTI degradation the threshold voltage of the devices increase which results in performance degradation, but at the same time the leakage of the devices also decreases. The change in leakage can be measured to detect an "average change" in the threshold voltage of the underlying devices. Fig. 6.1 shows an inverter chain with leaking transistors.

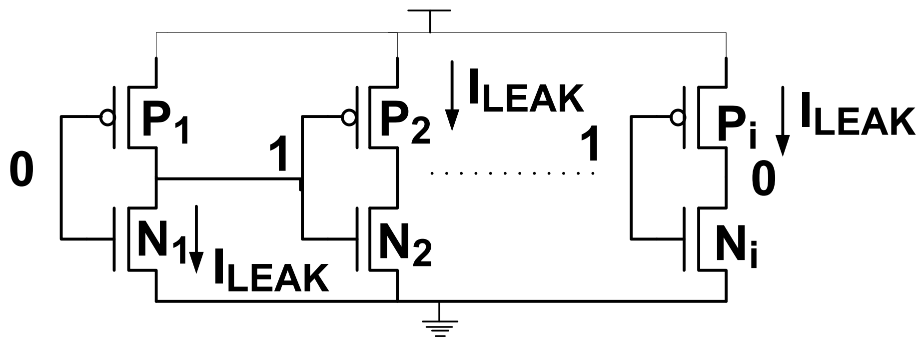


Fig. 6.1 Inverter chain with leaking devices.

It is only the OFF devices that contribute to leakage and not the ON devices even though they are undergoing BTI and their threshold voltage would be increasing. This implies that the leakage measurement will reflect on the change in threshold voltage of only the OFF devices. So the input vector to the circuit determines the devices which are being measured for BTI. If we consider a simple circuit consisting of identical gates, and all the devices undergo same threshold voltage shift, the average threshold voltage shift measured will be equal to the threshold voltage shift of an individual device irrespective of the input vector. But as we know in practice every device will degrade by a different amount in which case the average threshold voltage shift measured will vary with the

input state. Moreover a mixture of different gates in the circuit will also make the analysis more involving, especially with in stacked devices. A study of this technique with these variations taken into account for different kinds of circuit topologies would be a part of future work.

The phenomenon of recovery adds to the challenges in using the leakage current based idea. As the threshold voltage of the OFF devices recovers the leakage will increase as the threshold voltage decreases. Hence the leakage measurements need to be fast. If the circuit was already in a particular state for long time in the normal circuit operational mode, it is best to change the state of the circuit when switching to the measurement mode. Again this can be also very tricky because ideally we would want all the circuit nodes to change their state in the measurement state, which may not be possible. So it is best to make measurements with many different input vectors to span all the devices in a gate.

It is impractical to have an on-chip circuitry to measure leakage current. Instead we use a header based methodology to directly measure the change in threshold voltage of the underlying circuits. As shown in Fig. 6.2, two headers separate the inverter chain from the supply rail. One header is the conventional MTCMOS header (P_{MTCMOS}), which is ON during the normal operational mode of the circuit. The other header is the BTI Measurement header ($P_{MEAS\text{-}BTI}$) which turns ON only during BTI measurement. Since measurement time is very small as compared to the lifetime of the chip (could be as low as $1/100^{th}$), the BTI degradation in $P_{MEAS\text{-}BTI}$ would be insignificant. The gate voltage (Vbias) of $P_{MEAS\text{-}BTI}$ is controllable. Also the virtual rail (VR) of the circuit is monitored. Fig. 6.3 shows how the VR behaves when Vbias is swept from 0 to Vdd. When $P_{MEAS\text{-}BTI}$

is strongly ON it holds VR at Vdd. When Vbias is near Vdd, $P_{MEAS\text{-}BTI}$ is weakly ON or is in sub-threshold and VR falls below Vdd and eventually falls to 0.

With time, the threshold voltage of the underlying devices increases and the consequently their leakage decreases. As a result the VR voltage is higher for the same Vbias and consequently VR vs Vbias curve starts moving towards right. The amount of shift in this curve towards right is used to get a quantitative measure of the average threshold voltage shift in the devices due to BTI.

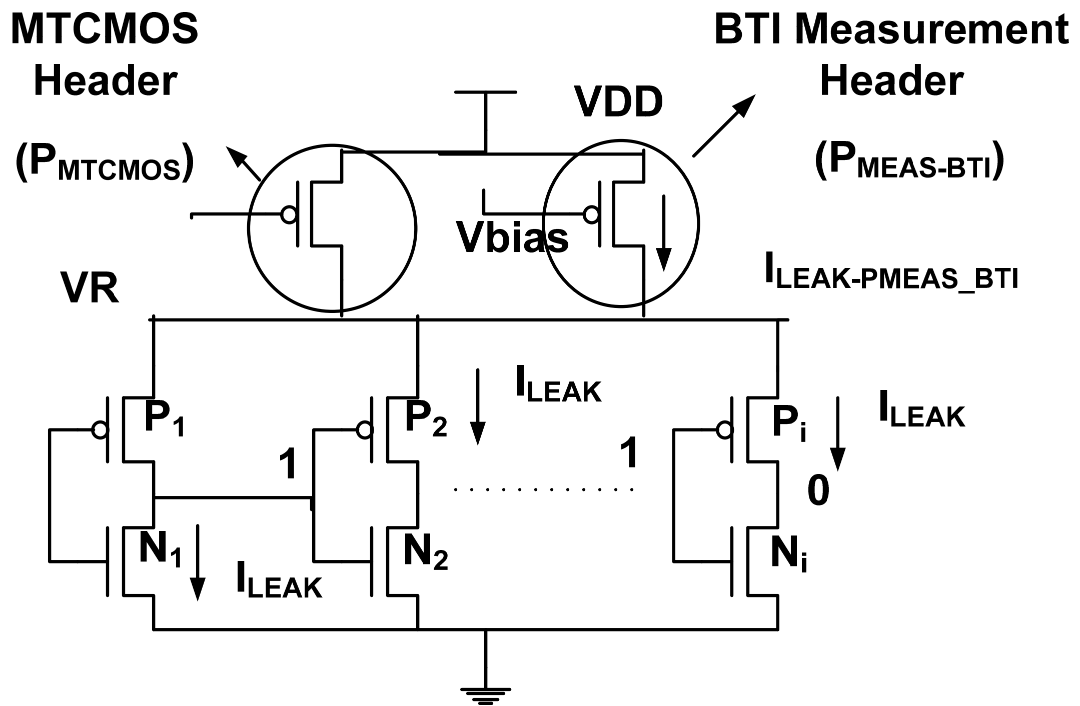


Fig. 6.2 Header based methodology is used to detect BTI.

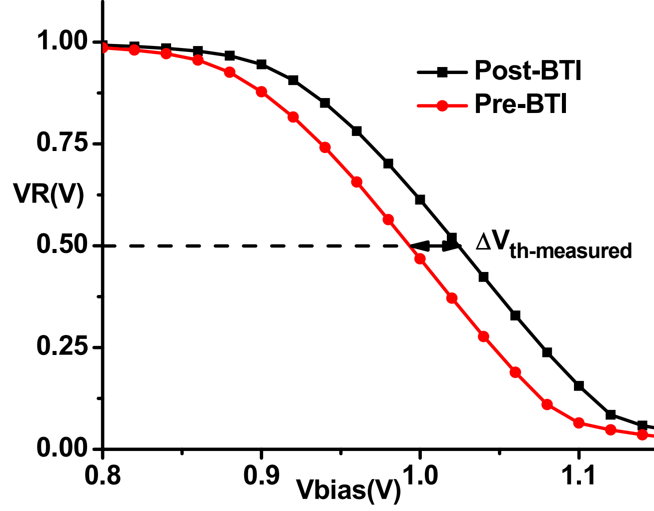


Fig. 6.3 Shift in the VR vs Vbias curve is the measure of threshold voltage shift.

6.2.1 Threshold voltage shift computation

Sub-threshold current in a transistor can be written as

$$I_{sub} = A\exp(Vgs - Vth + \eta Vds/nV_T)(1 - \exp(-Vds/V_T))$$

(6.1)

where A is dependent on technology constants, W and L of the device, and temperature. η is the DIBL coefficient. For OFF transistors Vgs = 0. If Vds > $4V_T$ the last term can be ignored. Simplifying (6.1)

$$I_{sub} = A\exp(-Vth + \eta Vds/nV_T)$$

(6.2)

For a block shown in Fig. 6.2 the currents through $P_{MEAS\text{-}BTI}$ and the cluster of underlying devices when VR = VDD/2 will be given by

$$I_{PMEAS} = A_{PMEAS} \exp(Vdd - Vbias_1 - Vth_{PMEAS} + 0.5\,\eta Vdd / nV_T)$$

(6.3)

The MTCMOS must be in super cutoff so that it does not contribute to the leakage currents.

Moreover for this equation to be valid, $P_{MEAS\text{-}BTI}$ must be sized such that it is sub-threshold when VR= Vdd/2. When VR= Vdd/2, the leakage current through the cluster would be

$$I_{CLUSTER} = \sum A_i \exp(-Vth_i + 0.5\,\eta Vdd / nV_T)$$

(6.4)

I_{PMEAS} and $I_{CLUSTER}$ can be equated to have

$$A_{PMEAS} \exp(Vdd - Vbias_1 - Vth_{PMEAS} + 0.5\,\eta Vdd / nV_T) = \sum A_i \exp(-Vth_i + 0.5\,\eta Vdd / nV_T)$$

(6.5)

After BTI the Vth of the cluster changes to Vth + ΔVth and the following equation can be written

$$A_{PMEAS} \exp(Vdd - Vbias_2 - Vth_{PMEAS} + 0.5\,\eta Vdd / nV_T) = \sum A_i \exp(-Vth_i - \Delta Vth + 0.5\,\eta Vdd / nV_T)$$

(6.6)

Using (6.5) and (6.6)

$$\Delta Vth = Vbias_2 - Vbias_1$$

(6.7)

And hence we define the horizontal shift in the VR vs Vbias curve at VR=Vdd/2 as a measure of ΔV_{th}.

6.2.2 Simulation Results

We verified Eq. 6.7 in simulation by applying this methodology on an inverter chain (Fig. 6.2).

In the experiment, the threshold voltage of all the devices in the inverter chain (Pi's, Ni's) was increased by 30mV and the VR vs Vbias curve was plotted. The measured shift was equal to 30.7mV. Since $P_{MEAS\text{-}BTI}$ and P_i/N_i are biased at different voltages their sub-threshold slopes differ, introducing error in Eq. 6.7. If $P_{MEAS\text{-}BTI}$ is sized to match the effective $(W/L)_{CLUSTER}$, its bias point moves closer to P_i/N_i, leading to minimal $\Delta V_{th\text{-}cluster}$ error. This is verified in Fig. 6.4. Alternatively ΔV_{th} can be measured at VR < Vdd/2, so that $P_{MEAS\text{-}BTI}$ is deeper in cut-off.

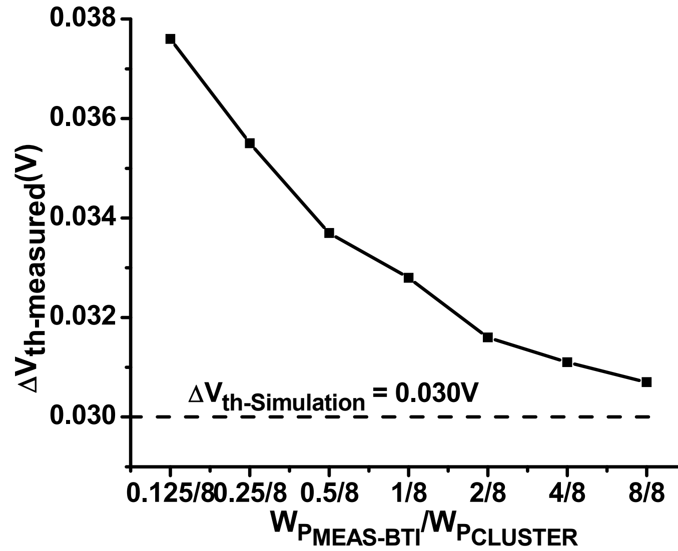


Fig 6.4 The error in measured threshold voltage shift decreases as the header and cluster widths are matched

Due to the statistical nature of BTI, ΔV_{th} differs across transistors *i* in a cluster, in which case $\Delta V_{th\text{-}cluster}$ will be a weighted average of their $\Delta V_{th\text{-}i}$. The MTCMOS header P_{MTCMOS} will also undero BTI. To determine its ΔV_{th} V/V measurements with $P_{MEAS\text{-}BTI}$

must be followed by a V/V measurement with P_{MTCMOS} as the header ($V/V_{PMTCMOS}$). With BTI degradation in P_{MTCMOS}, the $V/V_{PMTCMOS}$ curve will shift to the left when compared to measurements using P_{MEAS}. By comparing this shift and $\Delta V_{th\text{-}cluster}$ (computed from measurements with P_{MEAS}), the $\Delta V_{th\text{-}PMTCMOS}$ can be computed. In our experiments we did not stress P_{MTCMOS}.

6.2.3 Overheads

It is important to address the overheads associated with this method. If a design already has power gating implemented (MTCMOS), then the overheads are considerably reduced. In that case, most of the design-flow overhead, such as header insertion would already have been accounted for. The insertion of $P_{MEAS\text{-}BTI}$, routing of V_{bias} (for controllability), and virtual rails (for monitoring) would incur area overhead. These overheads would be dependent on the number of clusters into which the circuit is being divided into. If the granularity is high (i.e more number of clust ers), the measured Vth will be closer to the actual degradation of the devices. On the other hand if the granularity is low, the statistics will be more averaged out but the overhead will be low. Moreover, on-chip ADC and DAC would need to be implemented to generate Vbias and monitor VR. The rate of BTI recovery would determine the speed requirements for the ADC, DAC. Based on the experiments on the NBTI sensor (proposed in Chapter 2), 10-15% recovery occurs in 100 µs. So a sub MHz implementation of the data converters would suffice to make accurate measurements. A resolution of 20mV is enough for Vbias generation and VR measurements. If the supply voltage is 1V then a resolution of 20mV translates to 6 bit data-converters. Fig. 6.5 shows the implementation of this technique on a system.

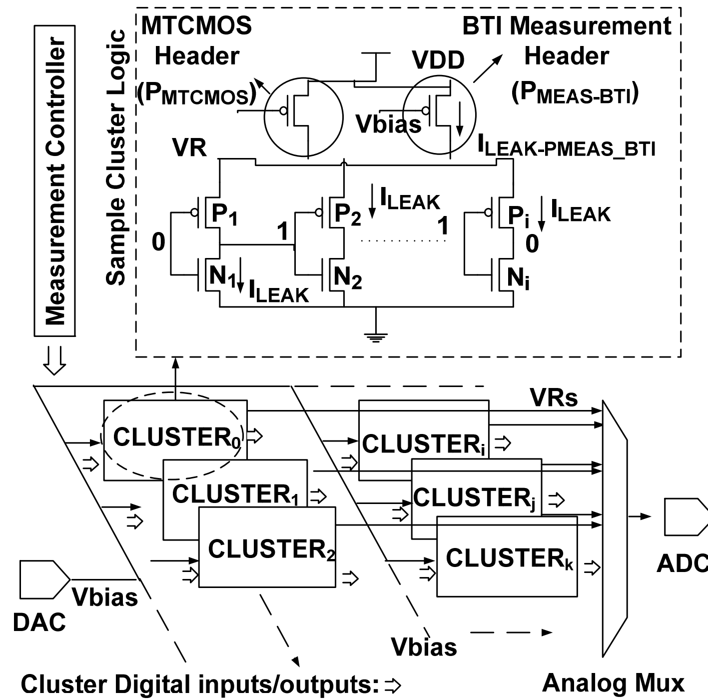


Fig 6.5 A system with BTI measurement technique implemented.

6.3 Test-chip Implementation

The circuit implementation of this technique is very similar to the *in situ* gate-oxide wear-out detection technique. We designed two test-chips to test *in situ* gate-oxide wear-out detection and later we use the same designs to test *in situ* detection of BTI. It must be kept in mind that these chips were not optimally designed for BTI detection.

This technique is tested on two test chips in 65nm technology node. On the first test chip

the technique was implemented on standalone inverters, NAND and NOR gates (Fig. 6.6) . Apart from standalone gates, we implemented this technique on XOR parity tree blocks with different gate-counts namely 32, 64, 128, 256, 512, 1024 gates (Fig. 6.7). Vbias is generated on-chip using a 6 bit DAC and VR is monitored using a 5 bit DAC.

Scan-flops are used to control the DAC and to store the ADC results. Since only one point in VR vs Vbias curve can be measured at a time, to generate the curve with the full Vbias sweep multiple scans are required. Hence considerable recovery (more than 40%) would have occurred by the time full VR vs Vbias curve is generated and threshold-shift is measured. This problem can be easily solved by having an on-chip state machine to control the Vbias and on-chip memory elements to store the ADC values.

The second test chip applies the proposed technique to a 16-bit, 8-tap FIR-filter. Its architecture is shown in Fig. 6.8. The FIR-filter is divided into 360 clusters, with each cluster consisting of ~20 gates. Each cluster has its own header unit (HU) with $(W/L)_{PMEAS\text{-}BTI}$= 10/0.06. $P_{MEAS\text{-}BTI}$ is used for normal FIR-filter operation as well as V/V measurements. For in-field implementations where thick-oxide transistors would not be required, the area overhead of inserting HUs at this granularity would be 17% when compared to a design without MTCMOS and 5% when compared to a standard MTCMOS design.

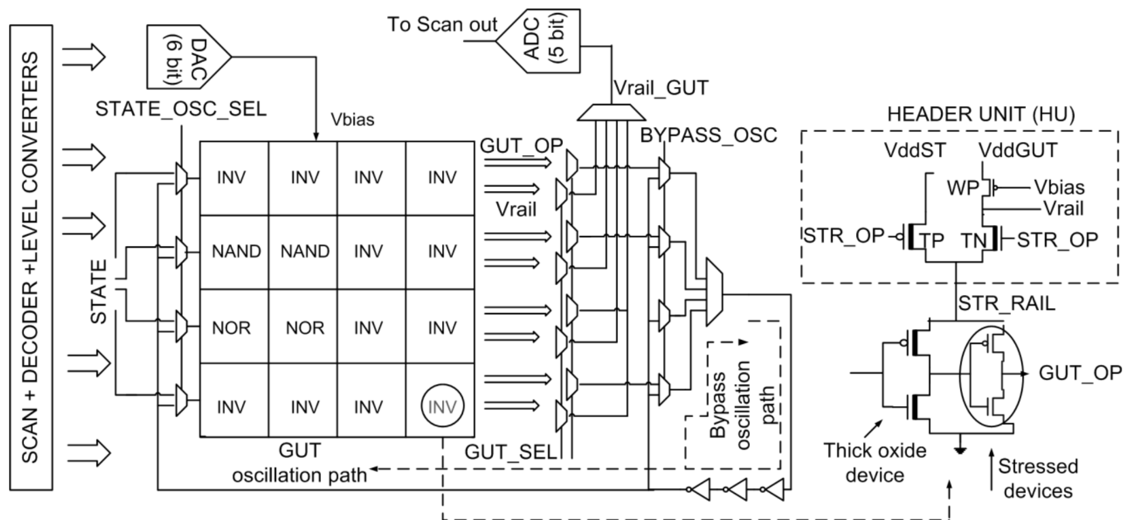


Fig. 6.6 Test-chip with BTI implementation on individual gates.

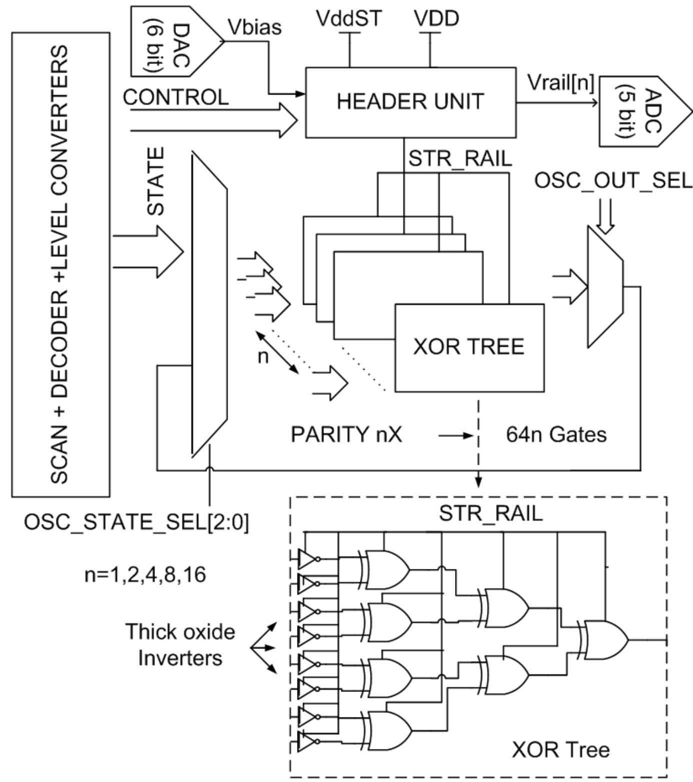


Fig. 6.7 Test-chip with BTI implementation on XOR parity trees.

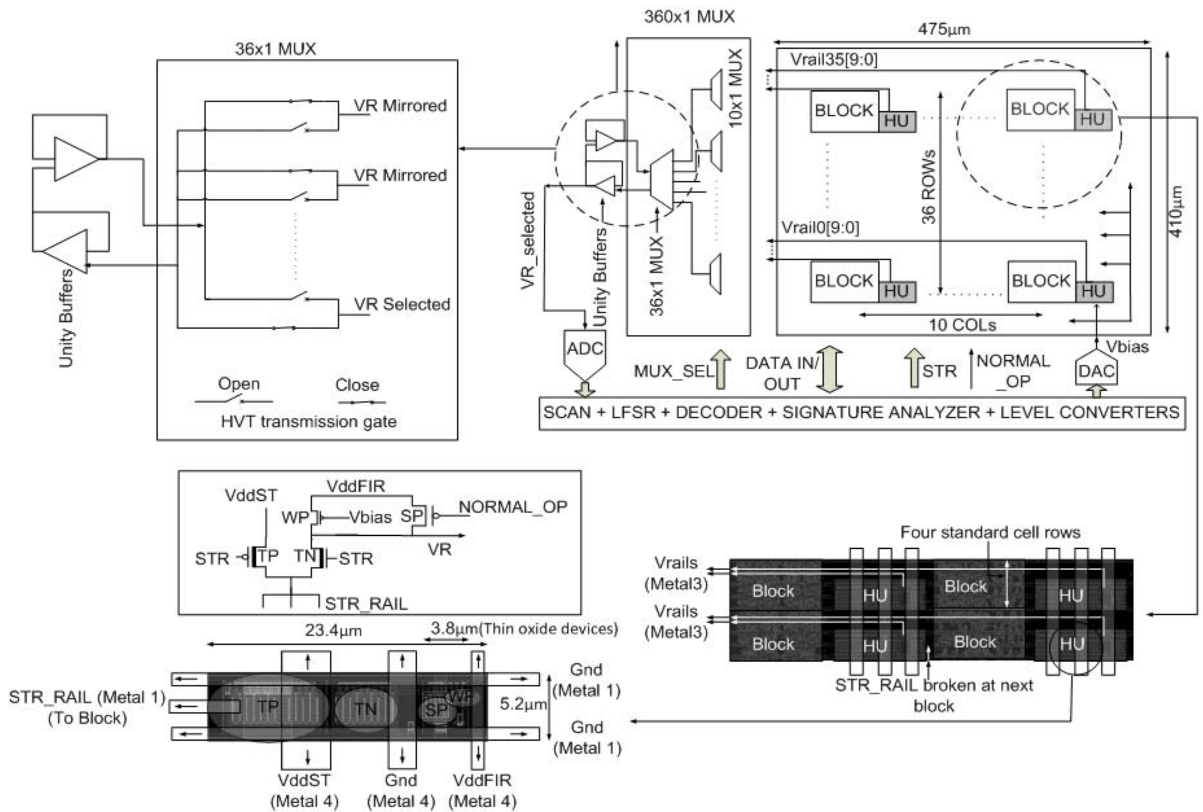


Fig. 6.8 Test-chip with BTI implementation on FIR Filter.

6.4 Silicon Results

For silicon results we stressed the circuits under accelerated conditions of voltage and temperature. The stress voltages were around 2V while the maximum temperature was chosen to be 125C. The stress voltage was kept below 2.5V so that the gate-oxide wear-out does not dominate the BTI degradation.

To stress the circuits the supply voltage is raised to stress voltage, and during measurement the supply voltage is brought back to nominal value (1V). To test the concept we stressed the PMOS of an inverter by applying a zero at its gate in the stress

mode and during measurement we applied a 'one' at its gate. The VR vs Vbias curve was generated before and after the stress and the horizontal shift at VR = Vdd/2 was measured. The measured threshold shift was equal to 25mV (Fig. 6.9). To investigate the severity of PBTI in this technology we stressed the NMOS by putting a 'one' at its gate during the stress mode, and putting a zero in the measurement mode. The measured Vth was insignificant (Fig. 6.10).

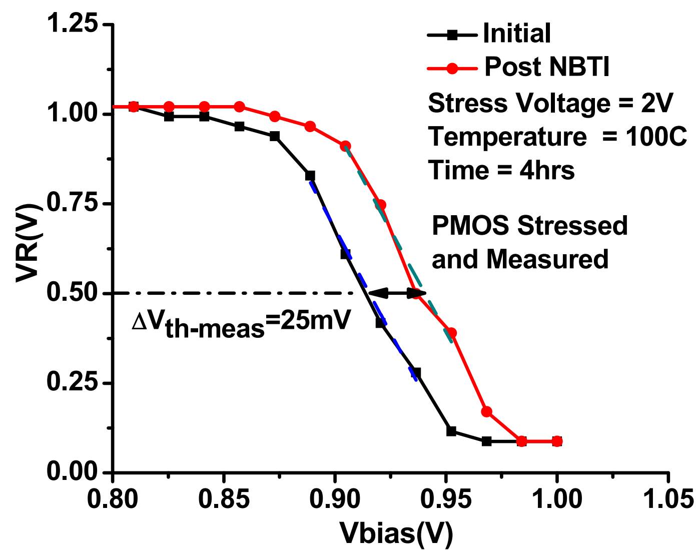


Fig. 6.9 NBTI measurement on an inverter

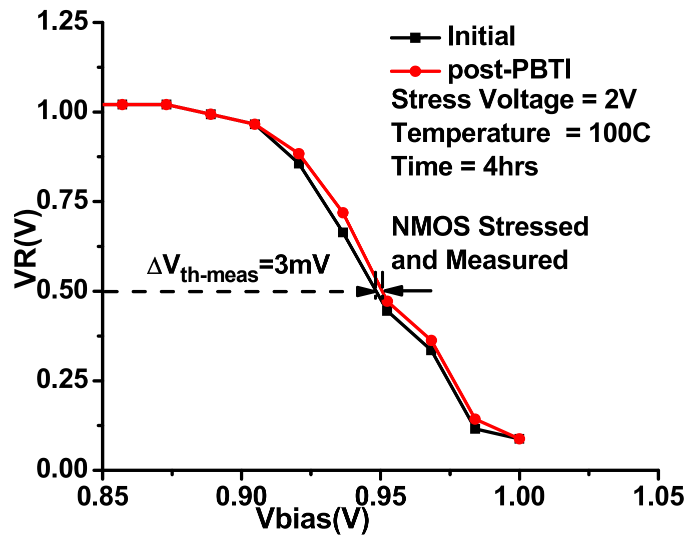


Fig. 6.10 PBTI measurement on an inverter.

To characteristic saw-tooth behavior of BTI is shown in Fig. 6.11. A periodic waveform with 50% duty cycle was applied to the inverter input during stress-mode. The stress mode was interrupted with measurement intervals. As the V_{th} increases and recovers, gate delay also increases and recovers. The recovery behavior confirms that what we are measuring is indeed BTI.

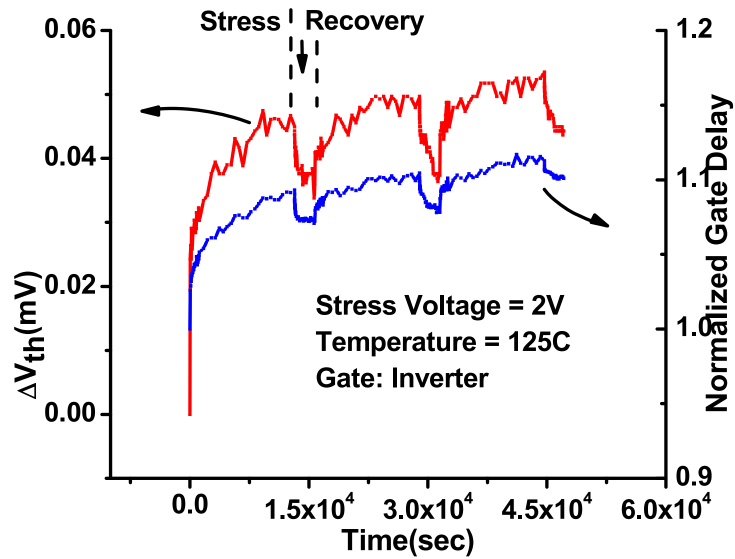


Fig. 6.11 BTI measurement with periodic stress and recovery periods.

Next we conducted experiments on the FIR filter for BTI measurements. One of the clusters of the FIR filter was stressed and measured for BTI. Fig. 6.12 shows that the Vth of the cluster increases with time. The leakage of the cluster was also monitored and as it can be seen the leakage reduces as the threshold voltage increases.

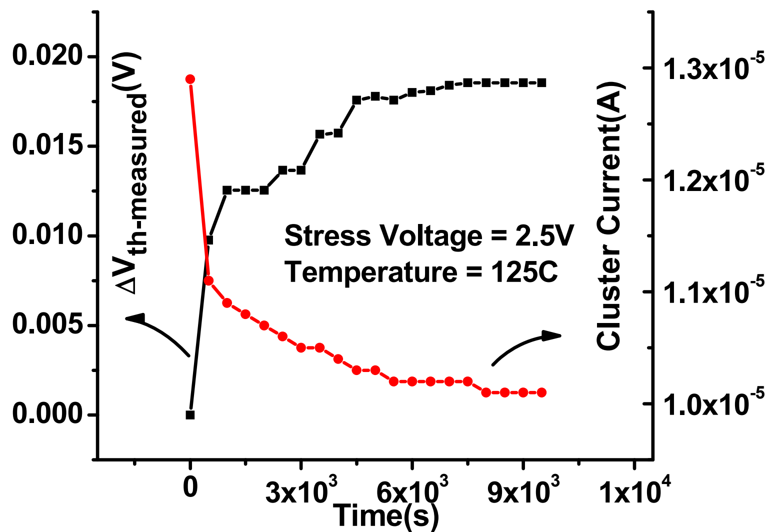


Fig. 6.12 Threshold voltage shift and leakage measurement for FIR filter cluster.

Next we stressed 141 clusters of the FIR filter and measured the threshold voltage shift of all the clusters. Fig. 6.13 shows the plot of the distribution of Vth among 141 clusters. There is a large spread in the distribution of the measured Vth shifts.

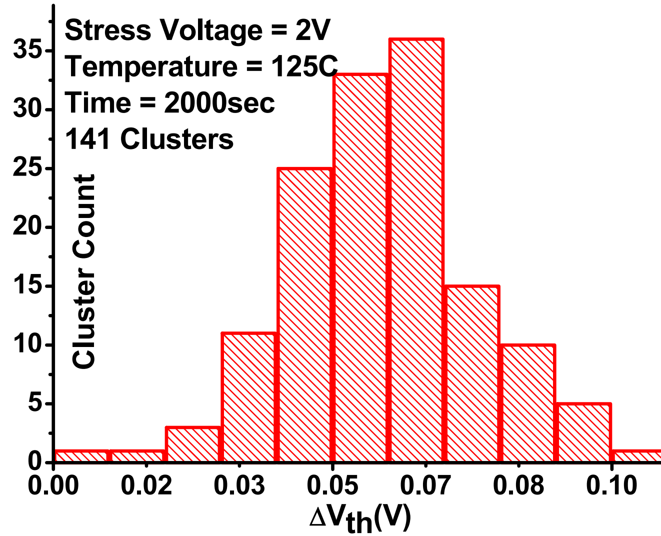


Fig. 6.13 Threshold voltage shift distribution among FIR filter clusters.

6.5 Dynamic BTI Management

We used the *in situ* BTI detection technique to perform Dynamic BTI Management. We demonstrate DBM on a 32-gate XOR parity block to trade excess BTI budget with performance. Under accelerated conditions of temperature and voltage we assumed a target lifetime (T_{LT}) of 15.6 hrs and a nominal voltage of 2V. To set the BTI budget we stressed the circuit under the worst conditions of temperature (125C). The resultant Vth shift at the end of T_{LT} was 29mV. Based on this we set a BTI budget of 6mV i.e. $\Delta V_{th\text{-}LT}$ = 35mV (Fig. 6.14). To compute the slack resulting due to temperature change we repeat this experiment at 75C, which gives a Vth shift of 12mV and hence a slack of 23mV.

Now the DBM is implemented at 75C to convert the slack into supply voltage boost. The DBM is also implemented at 125C as well to make sure that the degradation does not exceed the budget. To implement DBM, the Vth measurements are made every 256 seconds and reliability projections are made every 50 readings. Power law model $\Delta Vth = A + K\,t^n$ is used to extrapolate the degradation to T_{LT} (Fig. 6.15). A, K and n are fitting constants. Power law is typically an accepted model for BTI. The constant K has been added to account for the change in degradation curve due to dynamic variation in supply voltage. After extrapolation the predicted shift ($\Delta V_{th\text{-}Pred\text{-}TLT}$) is compared with $\Delta V_{th\text{-}TLT}$. If $\Delta V_{th\text{-}Pred\text{-}TLT} < \Delta V_{th\text{-}TLT}$ then the supply voltage is increased by 100mV, otherwise it is reduced by 100mV. After the supply voltage is modulated the DBM algorithm waits for next 50 readings to evaluate the degradation and adjust the supply voltage. This process is repeated till time T_{LT}.

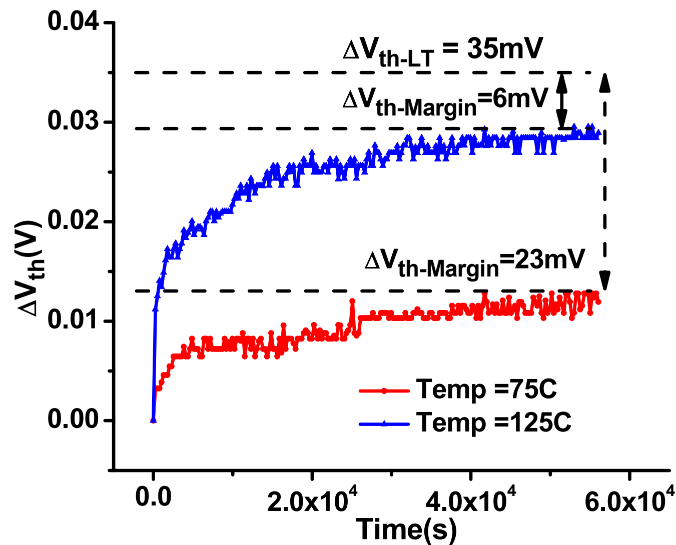


Fig. 6.14 BTI at 125C and 75C leaves margins at the end of life-time.

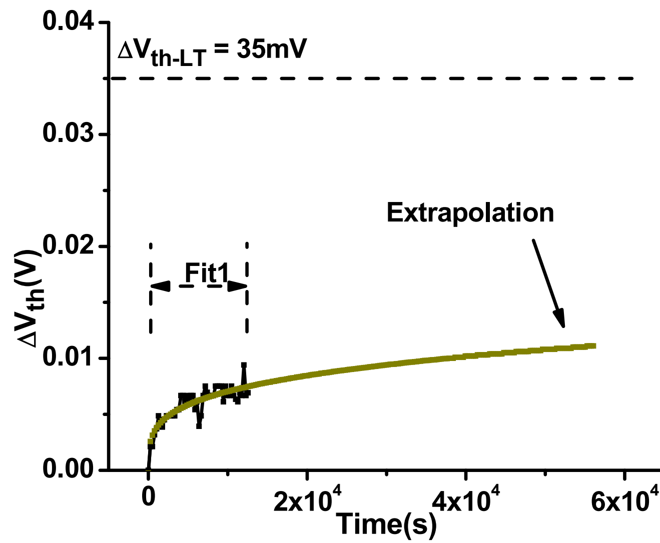


Fig. 6.15 Power law extrapolation used to predict BTI degradation till the end of life-time.

DBM at 75C results in an average supply (accelerated) boost of 8.6% while trading 9mV of BTI budget. At 125C it allows an average supply boost of 7.7% while consuming the entire excess BTI budget and making sure that ΔV_{th} at T_{LT} does not overshoot the reliability specification (Fig. 6.16).

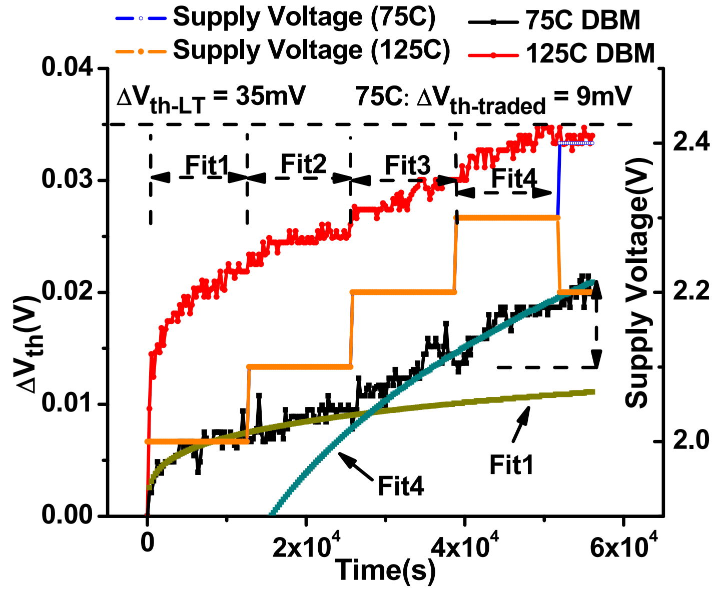


Fig. 6.16 DBM trades excess BTI budget with performance by boosting the supply voltage at 75C and 125C.

6.6 Conclusion

In this chapter *in situ* BTI detection technique was proposed for the first time. The technique uses the change in the sub-threshold leakage of devices to detect the change in their sub-threshold voltage. Since measuring on-chip current is impractical, a header based methodology is used. The technique has low overhead when implemented on a power gated design with MTCMOS headers. The technique was implemented on two test chips in 65nm process node. The BTI measurement results from the test-chip verify the technique. The BTI technique is used to implement Dynamic BTI Management which trades excess reliability slack at lower temperature by allowing operation at higher voltage.

Chapter 7
Summary and the Road Ahead

Static Reliability Management is traditionally employed by industry to ensure that the chips meet the reliability specifications. The process engineers assume worst process, voltage, temperature and state (PVTS) conditions to do reliability analysis for a particular technology node. But this method results in a considerable reliability slack for a large ensemble of chips. Dynamic Reliability Management aims to trade this slack with performance. In this scheme, a system monitors its reliability during the infield operation and it uses the reliability history to estimate the degradation at the end of lifetime specification. Based on the projections it increases or decreases the operating voltage or temperature limits and hence manages the reliability while maximizing the performance.

The focus of this work has been around the concept of Dynamic Reliability Management. We worked on developing degradation sensors and *in situ* degradation sensing techniques and then use them to perform Dynamic Reliability Management. The degradation mechanisms which we focused on were Bias Temperature Instability and Gate-oxide wear out.

7.1 Summary

We proposed an NBTI sensor based on sub-threshold ring oscillator. The sub-threshold operation gives it exponential sensitivity to threshold voltage shifts. The sensor was tested on a test-chip in 130nm technology node. The sensor was laid out in a

standard cell fashion, was small in area (6 D Flip flops) and had low power consumption. This allowed its deployment in large numbers with minimal area and power over-head. The digital output assists in convenient data collection and temperature compensation allows for its implementation in real systems operating in field.

Next we proposed a unified NBTI and Gate-oxide wear out sensor. The test-chip was designed in 45nm technology node. While the NBTI component of the sensor was simply ported from the 130nm sensor design, the gate-oxide wear-out sensor improved on the previous works by reducing the power consumption by more than 10^5 times. Using the NBTI component of the sensor, dynamic NBTI management was performed for the first time. We demonstrated that dynamic NBTI management trades the excess reliability slack with performance while meeting the lifetime target.

From sensor based degradation sensing we moved to *in situ* sensing techniques. *In situ* techniques give a more accurate estimate of the degradation as they account for all the factors affecting reliability and measure the degradation of the actual devices of the core. We proposed an *in situ* gate-oxide wear-out sensing technique which allows for early detection of the gate-oxide wear-out. The early detection allows significant time to manage the reliability of the circuits. The technique was implemented and tested on two test-chips in 65nm.

Lastly we proposed an *in situ* BTI sensing technique. The design methodology for *in situ* BTI sensing technique is very similar to *in situ* gate-oxide wear out sensing technique. The technique was tested on two test-chips in 65nm.

In the following section we will discuss the unanswered questions which have resulted from this work and possible future work in this field.

8.2 Future Work

7.2.1 Statistical Modeling of NBTI Degradation

Our results show that there is a fair amount of variation in NBTI. The literature in the scientific community does not talk a lot about the statistical nature of NBTI. It would be interesting to study the physical sources of the variation and model it. Any systematic and random components would be useful in assessing

7.2.2 In situ Gate-oxide Degradation Sensing

The proposed work for *in situ* gate-oxide degradation sensing defines a metric called Degradation Voltage Angle (DVA). When the DVA crosses a certain threshold we define that as an onset of gate-oxide wear out. It would be useful to map this metric to gate leakage increase because that would give a better estimate of gate-oxide degradation. Moreover, DVA does not change much during the later stages of degradation. This is because DVA is defined very simply as the slope of the measurement curves in a certain region. More information could be extracted from the curves to compute degradation.

7.2.3 Gate-oxide Degradation Recovery

We observed gate-oxide degradation recovery in our experiments. There have been explanations by the scientific community giving the physical phenomenon happening during recovery. This gate-oxide recovery can potentially be used to increase the life of gate-oxide and provide a knob for gate-oxide reliability management.

7.2.4 In situ BTI Sensing

This work presents only a precursor to the BTI sensing technique. More detailed analysis must be done while considering different classes of gates and input vectors. Also

the circuit technique implementation in this work was not optimal for BTI sensing allowing large recovery during measurement.

7.2.5 Mapping of Accelerated Stress Conditions to Nominal Conditions

All the laboratory tests in this work were done under accelerated conditions of voltage and temperature so that significant measurable degradation could be attained in hours or at maximum a few days. It is a question of utmost value that how does life-time under accelerated conditions map to nominal conditions. Unfortunately it is not possible to wait for years to verify the accelerated test models. Nevertheless the answer to this question would show the extent of threat to the reliability of systems.

7.2.6 Dynamic Gate-oxide Reliability Management

In this work even though we proposed and implemented sensing techniques to measure gate-oxide wear out, we did not perform Dynamic Gate-Oxide Reliability Management. The reason for that was the unpredictable nature of the gate-oxide degradation. It is not possible predict gate-leakage years in advance by just observing the history of the gate-leakage current. The gate-current can abruptly increase many folds without showing much signs. Better understanding of the physical sources of gate-oxide current increase and statistical study of the phenomenon may lead to better extrapolation and prediction of gate-oxide degradation.

Another challenge in Dynamic Gate-Oxide Reliability Management is the severity of gate-oxide failure. Unlike BTI degradation, which would not lead to catastrophic failure of the chip if a few devices exceed the BTI budget, even a single gate-oxide failure may render chip useless. Due to this more pessimism needs to be incorporated in the future prediction of gate-oxide failure.

7.2.7 Dynamic Reliability Management Implementation

It can be a question of debate whether DRM algorithm must be implemented on the silicon or if a major portion of it must be implemented in software. The hardware implementation would be area and power intensive the algorithm would require extensive modeling and would be mathematically intensive. But it would take care of issues such as revealing silicon data to the operating system.

This was the first time that Dynamic Reliability Management was implemented and demonstrated, and potential benefits from the technique were illustrated. Its adoption by industry would depend on the fact that how severe it considers the threat of reliability so that its implementation is worth the investment of resources.

Bibliography

[1.1] Gordon E. Moore," Cramming more components onto integrated circuits," Electronics, Volume 8, April19, 1965.

[1.2] International Technology Roadmap from Semiconductors 2009 Edition (Design)

[1.3] E. Karl, D. Blaauw, and D. Sylvester, "Analysis of System-Level Reliability Factors and Implications on Real-Time Monitoring Methods for Oxide Breakdown Device Failures," in Proceedings of IEEE International Symposium on Quality Electronic Design (ISQED),2008, pp. 391-395.

[1.4] R. Degraeve, et. al, "A consistent model for intrinsic breakdown in ultra-thin oxides," International Electron Devices Meeting, Dec. 1995, pp. 863-866.

[1.5] Cheng Zhuo, David Blaauw, and Dennis Sylvester, "Post-Fabrication Measurement-Driven Oxide Breakdown Reliability Prediction and Management," in Proceedings of IEEE/ACM International Conference on Computer-Aided Design, San Jose, November 2009, pp.441-448.

[1.6] E. Karl, D. Blaauw, D. Sylvester, T. Mudge," Multi-Mechanism Reliability Modeling and Management in Dynamic Systems," in IEEE Transactions On VLSI Systems, April 2008, pp. 476-487.

[1.7] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "The case for lifetime reliability-aware microprocessors," in Proceedings of 31st Annual International Symposium on Computer Architecture, 2004, pp. 276–287.

[1.8] Z. Lu, W. Huang, M. R. Stan, K. Skadron, and J. Lach, "Interconnect lifetime prediction under dynamic stress for reliability-aware design," in Proceedings of IEEE/ACM International Conference Computer-Aided Design, 2004, pp. 327–334.

[2.1] E. Karl et al., "Compact *in situ* Sensors for Monitoring NBTI and Oxide Degradation," IEEE International Solid-State Circuits Conference, pp. 410-623, 2008**.**

[2.2] D. K. Schroder, **"**Negative Bias Temperature Instability: What do we understand?," Microelectronics Reliability, Volume 47, Issue 6, pp. 841-852, June 2007.

[2.3] Huard *et al*, "New characterization and modeling approach for NBTI degradation from transistor to product level," *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, vol., no., pp.797-800, 10-12 Dec. 2007.

[2.4] B. Paul et al., "Impact of NBTI on the Temporal Performance Degradation of Digital Circuits", IEEE Electron Device Letters, pp. 560-562, August 2005.

[2.5] W. Wang et al., "The Impact of NBTI on the Performance of Combinational and Sequential Circuits", ACM/IEEE Design Automation Conference, pp. 364-369, 2007.

[2.6] S. Kumar et al., "Impact of NBTI on SRAM Read Stability and Design for Reliability", IEEE International Symposium on Quality Electronics Design, pp. 210-218, 2006.

[2.7] R. Vattikonda et al., "Modeling and Minimization of PMOS NBTI Effect for Robust Nanometer Design," ACM/IEEE Design Automation Conference, pp. 1047-1052, 2006.

[2.8] G. Chen et al., "Dynamic NBTI of PMOS Transistors and its Impact on Device Lifetime", IEEE Electron Device Letters, pp. 734-736, December 2002.

[2.9] S. Rangan et al., “Universal recovery behavior of negative bias temperature instability”, IEEE International Electron Devices Meeting, pp. 341–344, 2003.

[2.10] V. Huard et al., “NBTI degradation: from physical mechanisms to modeling,” Microelectronics Reliability, Volume 46, Issue 6, pp. 1-23, 2006.

[2.11] M. Denais et al., “On-the-fly Characterization of NBTI in Ultra-Thin Gate Oxide PMOSFET’s,” IEEE International Electron Devices Meeting, pp. 109-112, 2004.

[2.12] S. Aota et al., “A New Method for Precise Evaluation of Dynamic Recovery of Negative Bias Temperature Instability,” IEEE International Conference on Microelectronic Test Structures, pp. 197-199, 2005.

[2.13] C. Shen et al., “Characterization and Physical Origin of Fast Vth Transient in NBTI of pMOSFETs with SiON Dielectrics,” IEEE International Electron Devices Meeting, pp. 1-4, 2006.

[2.14] R. Fernández et al., “AC NBTI Studied in the 1Hz – 2GHz Range on Dedicated On-Chip Circuits,” IEEE International Electron Devices Meeting, pp. 337-340, 2006.

[2.15] J. Keane et al.,” An all-in-one silicon Odometer for separately monitoring HCI, BTI, and TDDB,” IEEE Symposium on VLSI Circuits, pp. 108-109, 2009.

[2.16] M. B. Ketchen et al., “Ring Oscillator Based Test Structure for NBTI Analysis,” IEEE International Conference on Microelectronic Test Structures, pp. 42-47, 2007.

[2.17] T-H. Kim et al., “Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits”, IEEE Journal of Solid-State Circuits, vol. 43, Issue 4, pp. 874–880, 2008.

[2.18] J. Keane et al., “An on-chip NBTI sensor for measuring PMOS threshold voltage degradation”, ACM/IEEE International Symposium on Low Power Electronics and Design, pp. 189-194, 2007.

[2.19] L. Tsetseris et al, “Physical mechanisms of negative-bias temperature instability,” Applied Physics Letters, 86(14), art. no. 142103, pp. 1-3, 2005.

[3.1] Chen, I.C.; Holland, S.; Hut, C.; , "A quantitative physical model for time-dependent breakdown in SiO2," *Reliability Physics Symposium, 1985. 23rd Annual* , vol., no., pp.24-31, 25-29 March 1985.

[3.2] Y. Uraoka, N. Tsutsu, Y. Nakata, and S. Akiyama, “Evaluation technology of VLSI reliability using hot carrier luminescence,” IEEE Trans. on Semiconductor Manufacturing, vol. 4, no. 3, pp. 183-192, 1991.

[3.3] S. R. Nariani and C. T. Gabriel, “A simple wafer-level measurement for predicting oxide reliability,” IEEE Electron Device Letters, vol. 16, pp. 242–244, 1995.

[3.4] M. Wang et al.,” Statistical method of monitoring gate oxide layer yield,” United States Patent 6289291, 2001.

[3.5] M. Acar et al.,”Digital Detection of Oxide Breakdown and Life-Time Extension in Submicron CMOS Technology,” IEEE International Solid-State Circuits Conference, pp. 530-633, 2008.

[3.6] J. Keane et al.,” An array-based test circuit for fully automated gate dielectric breakdown characterization,” IEEE Custom Integrated Circuits Conference, pp. 121-124, 2008.

[3.7] Reiner, J.C.; , "Pseudo-progressive breakdown of ultra-thin nitrided gate oxide," *Integrated Reliability Workshop Final Report, 2004 IEEE International* , vol., no., pp. 151- 153, 18-21 Oct. 2004.

[3.8] B. P. Linder, J. H. Stathis, "Statistics of progressive breakdown in ultra-thin oxides, Microelectronic Engineering", Volume 72, Issues 1-4, Proceedings of the 13th Biennial Conference on Insulating Films on Semiconductors, April 2004, Pages 24-28.

[5.1] Y. Kim, Y. Kameda, H. Kim, M. Mizuno and S. Mitra, "Low-Cost Gate -Oxide Early-life Failure Detection in Robust Systems," *Symposium VLSI Circuits,* Honolulu, Hawaii, June 2010.

[5.2] A. Drake, R. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, and V. Pokala, "A distributed critical-path timing monitor for a 65 nm high-performance microprocessor," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2007, pp. 398–399.

[5.3] R. Rodriguez, J. H. Stathis, and B. P. Linder, "Modeling and experimental verification of the effect of gate oxide breakdown on CMOS inverters," in Proceedings of IEEE International Reliability Physics Symposium, Dallas, TX, 2003, pp. 11–16.

[5.4] P. Singh et al.," Early Detection of Oxide Breakdown Through *In Situ* Degradation Sensing", in IEEE International Solid State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2010, pp. 190–191.

[5.5] G. Gerosa et al., "A sub 1 W to 2 W low power IA processor for mobile internet devices and ultra-mobile PCs in 45 nm high-k metal gate CMOS," in IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers, Feb. 2008, pp. 256–257.

[5.6] J. A. Blome, S. Feng, S. Gupta, and S. Mahlke, “Online timing analysis for wearout detection,” in 2nd Workshop on Architectural Reliability(WAR-2), Dec 2006.

[5.7] Cheung, K.P.; , "Ultrathin gate-oxide breakdown-reversibility at low voltage," *Device and Materials Reliability, IEEE Transactions on*, vol.6, no.1, pp. 67- 74, March 2006.

[6.1] Zafar, S *et al*, "A Comparative Study of NBTI and PBTI (Charge Trapping) in SiO2/HfO2 Stacks with FUSI, TiN, Re Gates," *VLSI Technology, 2006. Digest of Technical Papers. 2006 Symposium on* , vol., no., pp.23-25.

[6.2] Seok Joo Doh *et al* , "Improvement of NBTI and electrical characteristics by ozone pre-treatment and PBTI issues in HfAlO(N) high-k gate dielectrics," *Electron Devices Meeting, 2003. IEDM '03 Technical Digest. IEEE International* , vol., no., pp. 38.7.1-38.7.4, 8-10 Dec. 2003.