On-Chip Circuits for Characterizing Transistor Aging Mechanisms in Advanced CMOS Technologies

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

John P. Keane

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Chris H. Kim, Advisor

April 2010

Acknowledgements

First, thank you to Professor Chris Kim for making my graduate education an enjoyable and rewarding experience. Very few students find an advisor that incessantly pushes them to do good work, but can still be considered a friend. I look forward to representing our research group as I begin my career, and to watching it continue to grow stronger.

Next, I would like to thank the members of my final defense committee: Professors Kiarash Bazargan, Gerald Sobelman, and Antonia Zhai. I appreciate you taking time out of your busy schedules to critique my work. I am also grateful to Professor Pen-Chung Yew for serving on my preliminary oral defense committee.

I owe a great deal of gratitude to IBM for supporting two years of my graduate education through the IBM PhD Fellowship program, as well as inviting me to complete several internships in Rochester, MN and at the Austin Research Lab in Texas. In particular, I would like to thank each of the members of the Exploratory VLSI Design Group at ARL, especially my mentors Fadi Gebara and Jeremy Schaub. Thanks to all of you for making me feel at home in Austin.

I would like to thank all of the members of Professor Kim's VLSI Research Lab at the University of Minnesota for making it more fun to spend countless hours in front of computer screens or measurement equipment. In particular, thanks are due to Jie Gu and

Tony Kim who were there when I started in the lab, and became great friends during the nerve-wracking run-ups to tapeouts and paper deadlines.

I am grateful for the support I continue to receive from my entire family, of whom only a few can be named in this space unfortunately. First, thank you to my godfather, Brother Gabriel Fagan for helping me make my undergraduate education at Notre Dame possible, encouraging me to continue on for a PhD, and being a constant mentor on life in general. Next, I would like to thank my parents for everything they have done for my six siblings and me. Thank you for pushing us to always achieve great things, even from a young age when we might not have been very agreeable. We tend to take it for granted that you have raised seven people who have done so much already, but it is important to stop and acknowledge what you have accomplished through all of us. Thank you for your sacrifices and unending support. All of us will strive live up to the high standards you have set through your example.

Finally, thank you to my wife Sarah. Your patience with me is astounding, and a great lesson that I will continue to learn from for many years. You have made the final portion of my life as a graduate student infinitely more balanced and fun. Thank you for everything.

Abstract

The parametric shifts or circuit failures caused by Hot Carrier Injection (HCI), Bias Temperature Instability (BTI), and Time Dependent Dielectric Breakdown (TDDB) in CMOS transistors have become more severe with shrinking device sizes and voltage margins. These mechanisms must be studied in order to develop accurate reliability models which are used to design robust circuits. Another option for addressing aging effects is to use on-chip reliability monitors that can trigger real-time adjustments to compensate for lost performance or device failures. The need for efficient technology characterization and aging compensation is exacerbated by the rapid introduction of process improvements, such as high-k/metal gate stacks and stressed silicon.

Much of the device aging data gathered for process characterization is obtained through individual probing experiments. However, probing stations are expensive, and they have other drawbacks such as limited timing resolution. In order to resolve these issues, several on-chip systems have recently been proposed to measure device aging. In this thesis I will present five unique test chip designs that we have implemented for this purpose.

Performing reliability experiments with on-chip circuits provides us with several advantages, in addition to avoiding the use of expensive probing equipment. First, using on-chip logic to control the measurements enables much better timing resolution. This is critical when interrupting stress to record BTI measurements, as this mechanism is known to partially recover within microseconds or less. We will also see that a digital beat frequency detection system allows us to measure ring oscillator frequency shifts with resolution ranging down to a theoretical limit of less than 0.01% . That mix of speed and resolution is not possible with standard off-chip equipment. Next, standard logic can be used to control tests on several devices in parallel, resulting in a large experiment time speedup when monitoring statistical processes. Utilizing these benefits to obtain accurate CMOS aging information would allow manufacturers to avoid wasteful overdesign and frequency guardbanding based on pessimistic degradation projections, and hence more fully realize the benefits of CMOS scaling.

Table of Contents

List of Figures

Chapter 1

Introduction

The parametric shifts or circuit failures caused by Hot Carrier Injection (HCI), Bias Temperature Instability (BTI), and Time Dependent Dielectric Breakdown (TDDB) have become more severe with shrinking transistor sizes and voltage margins. We now have chips containing billions of transistors operating at breakneck speeds, with precariously small voltage margins between the supply levels and the threshold at which devices turn on. More switching activity means more heat density, which accelerates most aging mechanisms. Process improvements such as strained silicon and high-k/metal gate devices also introduce new degradation concerns, such as BTI in n-type devices. Finally, technology scaling has led to a massive increase in the number of operating conditions devices find themselves in, so there is a larger variation in their aging processes.

Semiconductor companies generally deal with this aging problem by playing it very safe. For example, they build generous guardbands into clock speeds in order to ensure that their products will continue to operate over their intended lifetimes. This means that clocks have to be slowed down to well under the limits for fresh circuits in order to account for the impending logic slow-down that comes with aging, among other variables. By doing so, manufacturers throw out a portion of the performance benefit that comes with scaling because of problems that could arise after long periods of use.

Device dimensions have now been pushed to the atomic scale, though, and we are approaching physical limitations where transistors no longer act as reliable switches. In addition, manufacturers are facing significant challenges in the fabrication process which could become too costly to surmount. In this environment where we cannot count on continued performance improvements from scaling alone, making conservative estimations about circuit aging will no longer do. Research, design, and process development groups are all now devoting significant resources to better understanding the aging mechanisms, and exploring strategies to reduce the cushion put into clock speeds or maximum operating voltages which prevent possible timing failures down the road.

One critical aspect of that work involves developing accurate and efficient means to measure the effects of the different aging mechanisms, which is the objective of this thesis. In following chapters, we will presented five unique circuit designs that we have implemented in order to demonstrate the benefits of utilizing on-chip logic and a simple test interface to automate aging experiments. First we will go through a brief introduction to the transistor degradation mechanisms addressed by these test circuits, along with prior art in the field of on-chip aging sensors.

1.1 CMOS Transistor Aging Mechanisms

As shown in Fig. 1.1, CMOS devices suffer from HCI, BTI, and TDDB stress under standard digital operating conditions. HCI has become less prominent with the reduction of operating voltages, but remains a serious concern due to the large local electric fields in scaled devices [1]. Hot carriers (i.e., those with high kinetic energy) accelerated toward the drain by a lateral electric field across the channel lead to secondary carriers generated through impact ionization (Fig. 1.2(a)). Either the primary or secondary carriers can gain enough energy to be injected into the gate stack. This creates traps at the silicon substrate/gate dielectric interface, as well as dielectric bulk traps, and hence degrades device characteristics such as the threshold voltage (V_{th}). These "traps" are electrically active defects that capture carriers at energy levels within the bandgap.

NBTI (Negative Bias Temperature Instability) in PMOS transistors is often cited as the primary reliability concern in modern processes, especially after the introduction of nitrogen into gate stacks, which reduces boron penetration and gate leakage, but leads to worse NBTI degradation [2]. This mechanism is characterized by a positive shift in the absolute value of the PMOS V_{th}, which occurs when a device is biased in strong inversion, but with a small, or no, lateral electric field (i.e., $V_{DS} \approx 0$ V). The V_{th} shift is generally attributed to hole trapping in the dielectric bulk, and/or to the breaking of Si-H bonds at the gate dielectric interface by holes in the inversion layer, which generates positively charged interface traps (Fig. 1.2(c)) [2], [3]. When a stressed device is turned off, it immediately enters the "recovery" phase, where trapped holes are released, and/or the freed hydrogen species diffuse back towards the substrate/dielectric interface to anneal the broken Si-H bonds, thereby reducing the absolute value of the V_{th} (Fig. 1.2(d)). PBTI (Positive Bias Temperature Instability) in NMOS transistors was not critical in silicon dioxide dielectrics (such as those used in the test circuits presented in this thesis), but is now contributing to the aging of high-k gate stacks [4].

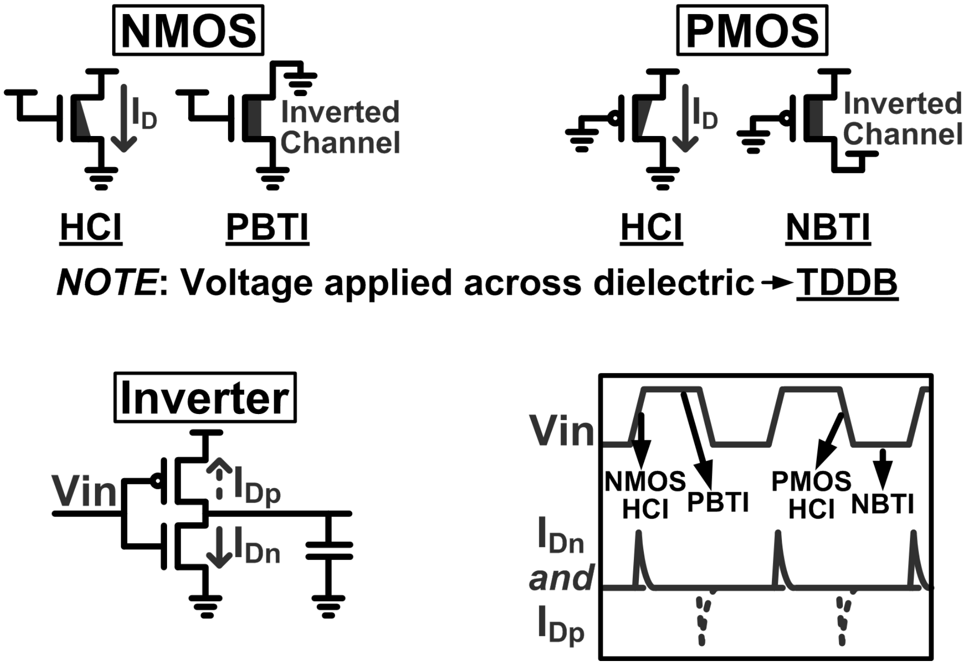


Fig. 1.1: HCI, BTI, and TDDB stress illustrated for NMOS and PMOS transistors, as well as for an inverter during standard operation.

Finally, any voltage drop across the gate stack can cause the creation of traps within the dielectric. These defects may eventually join together and form a conductive path through the stack in a process known as TDDB, or oxide breakdown (Fig. 1.2(b)) [5]. Breakdown has been a cause for increasing concern as gate dielectric thicknesses are scaled down to the one nanometer range, because a smaller critical density of traps is needed to form a conducting path through these thin layers, and stronger electric fields are formed across gate insulators when voltages are not reduced as aggressively as device dimensions. The scaling of the physical dimensions of gate stacks can now be slowed or reversed with the introduction of high-k dielectrics, but TDDB remains a critical aging mechanism in those materials, and is currently being studied by device physicists [4], [6].

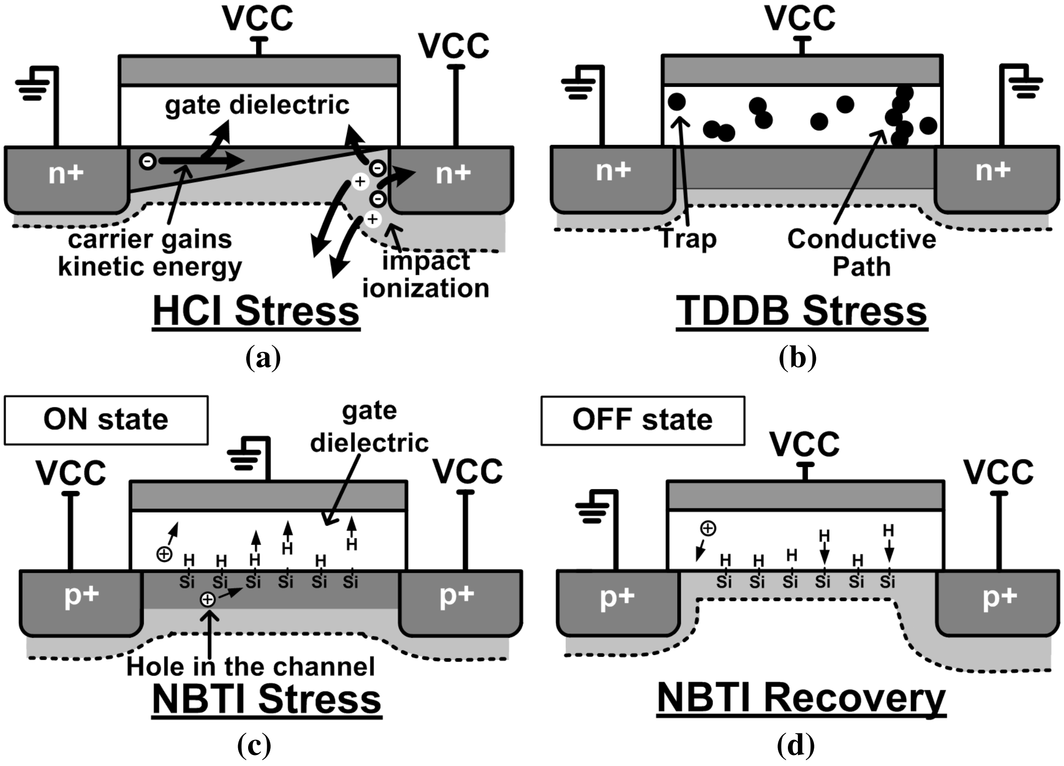


Fig. 1.2: Transistor cross sections illustrating (a) HCI, (b) TDDB, (c) NBTI stress, and (d) NBTI recovery.

These transistor degradation mechanisms lead to a host of problems in circuit performance over time. For example, as a CMOS system ages, certain logic paths that were not critical at design time may experience more significant stress, thereby becoming critical, and preventing proper timing closure. Aging can also cause degradation in the static noise margin of SRAM cells [7]-[9], and lowers the maximum operating frequency of aged circuits. In order for circuit designers to mitigate these effects without using costly over-design methods, such as liberally up-sizing stressed devices or using large guardbands in the system clock, accurate predictive models or real-time compensation schemes should be developed and incorporated into their suite of tools. These techniques must of course be solidly corroborated by reliable hardware data if they are to be effective.

Much of the device aging data gathered for process characterization is obtained through device probing experiments. The equipment used in those tests can be expensive (up to tens of millions of dollars for automated wafer probe stations), and testing each device individually leads to long experiment times. Several on-chip aging measurement systems have been proposed in recent years to address these issues and assist in the process of understanding and dealing with transistor aging.

Performing reliability experiments with on-chip circuits provides us with several advantages, in addition to avoiding the use of expensive probing equipment. First, using on-chip logic to control the measurements enables much better timing resolution. This is critical when interrupting stress to record NBTI measurements, as this mechanism is known to recover within microseconds or less [3], [10]. Moving on, we will see that a digital beat frequency detection system allows us to measure ring oscillator frequency shifts with resolution ranging down to a theoretical limit of < 0.01% [11]. That mix of sub-μs measurement times with high frequency shift resolution would be difficult, if not impossible to achieve with standard off-chip equipment. Finally, on-chip digital logic can be used to control tests on several devices in parallel, resulting in a large experiment time speedup when monitoring a statistical process like TDDB.

Although on-chip aging monitors are beneficial for these reasons, they do have drawbacks. For example, early technology characterization is often performed before many metallization layers are being fabricated. Therefore, process engineers will want to measure the characteristics of new transistors when only one or two metal layers are available, which would be difficult to do with anything but the most basic of on-chip circuits. In addition, extracting process parameters from on-chip tests generally involves some translations (e.g., frequency to threshold voltage) using approximations which lead to varying levels of error. In both of these cases, sensitive off-chip probing equipment may provide the optimal solution.

The benefits and drawbacks of previously proposed on-chip reliability monitors will be outlined in the following section, prior to introducing the contributions of this thesis.

1.2 Overview of Selected On-Chip Reliability Monitors

Denais *et al.* proposed an "on-the-fly" BTI measurement technique in order to avoid the recovery inherent in most measurement setups [12]. In this method, the stress voltage is kept quasi-constant, and the linear drain current ($I_{D,lin}$) of the device under test (DUT) is periodically measured to monitor device degradation. In [13], the on-the-fly technique was extended to characterize the recovery after stress conditions are removed. However, the on-the-fly method relies on a translation of $\Delta I_{D,lin}$ into ΔV_{th} (i.e., the threshold voltage shift) which requires some approximations, and the authors of [10] state that this method underestimates the total degradation due to a slow initial measurement which causes unrecorded degradation as well. Additionally, the time required for each measurement is typically in the range of milliseconds, and it is difficult to get an accurate reading of $\Delta I_{D,lin}$ at the stress voltage level, all of which could make on-the-fly results less reliable [14]. Next, Shen et al. used a 100 ns I-V sweep technique to monitor NBTI degradation, and demonstrated the fundamental differences in NBTI that are observed with ultra-high speed measurements [10]. This technique will experience drawbacks associated with high speed off-chip device probing, though, such as losses and cross-talk.

Kim *et al.* presented the first version of the "Silicon Odometer," which is a digital reliability monitor for high resolution frequency shift measurements [11]. This technique measures the beat frequency between two ROSCs, where one is stressed and the other is unstressed to maintain a fresh reference. They achieved 50X higher delay sensing resolution than prior schemes in the early stages of degradation. This concept is utilized and expanded upon in the present work.

Karl *et al.* proposed two separate compact circuits for monitoring NBTI and TDDB, with the goal of facilitating real-time characterization [15]. First they measured the frequency shift of a ROSC with a PMOS header that is placed under NBTI stress, and then biased in subthreshold during measurements for high ΔV_{th} sensitivity. Their work relies on a complex mathematical model to map temperature and V_{th} variations to the measured ROSC frequencies after extensive calibration. Next, the TDDB aging results were provided in the form of a frequency shift of a Schmitt trigger oscillator which is modified by the increasing gate leakage through a pair of stressed PMOS transistors.

Singh et al. recently introduced an in situ monitor for providing an early indication of the onset of TDDB [16]. In this design, circuits are periodically taken offline, and a PMOS header switch's gate bias is swept while recording the virtual supply rail voltage between the header and the circuit. This V_{bias} vs. V_{rail} characteristic is strongly non-linear in the fresh circuit, but becomes more linear as current paths are formed through the circuits' gate stacks. The authors state that the differences in this curve can be used as highly sensitive indications of early TDDB degradation. Utilizing it as an in-situ sensor is unlikely, though, due to the requirement of voltage generators, ADCs, the small number of gates that can be used per header, and more.

Finally, Saneyoshi *et al.* presented a fast method to detect NBTI degradation in delay lines by monitoring the number of stages an input edge could travel through before and after stress using edge capture logic [17]. It should be noted that the author of this thesis presented a roughly identical design to Saneyoshi's "hybrid" approach in July of 2007 and filed for patent protection with the U.S. Patent Office on July 19th, 2008 [18].

1.3 Summary of Thesis Contributions

The remainder of this work will explore the benefits of five test chip designs that we have implemented to accurately monitor CMOS transistor aging mechanisms. These designs build upon and refine the ideas behind many of the sensors described in the previous subsection. The first two circuits are based on the Silicon Odometer beat frequency detection system [11], and provide methods to individually monitor multiple aging mechanisms simultaneously, or to record statistical aging data. The Odometer framework allows us to perform those measurements in ROSCs under test with timing and frequency resolution that is unmatched by any traditional measurement system.

The third design contains a delay-locked loop, in which the increase in the PMOS threshold voltage due to NBTI stress is translated into a control voltage shift in the DLL for an average sensing gain of 10X. The final two circuits are array-based systems that facilitate the fast and efficient measurement of statistical time-to-gate breakdown data. These TDDB test circuits stress many devices in parallel, which speeds up experiments by a factor proportional to the number of DUTs when compared with standard probing tests. This is a significant benefit when characterizing TDDB, where up to thousands of test samples are needed to correctly define the breakdown behavior.

Chapter 2

An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI, and TDDB

2.1 Introduction to the All-In-One Silicon Odometer

As detailed in Chapter 1, the three main degradation mechanisms impacting modern CMOS transistors are BTI, HCI, and TDDB. Although some of the underlying physical explanations for these mechanisms are similar, each of them has different sensitivities to operating conditions and process changes, and can be more critical in certain circuit topologies. Therefore, they should each be examined separately. Although many methods have been proposed recently to monitor CMOS aging, none has been presented to isolate the effects of these three major reliability mechanisms in a single test structure.

In this work, we accomplish that task with a pair of ring oscillators (ROSCs) which are representative of standard circuits [19]. We use a "backdrive" concept in which one ROSC drives the transitions in both structures during stress, such that the driving oscillator ages due to both BTI and HCI, while the other suffers from only BTI. The latter ROSC is gated off from the supplies during stress so that no current is driven through the channels of its transistors, and therefore the carriers cannot become “hot.” In addition, long term or high voltage experiments facilitate TDDB measurements. It is now well known that BTI degradation recovers on a sub-µs timescale after the removal of stress conditions [3]. Therefore, we use a beat frequency detection method to take sub-µs measurements and avoid unwanted device recovery during stress interruptions. Sub-ps frequency measurement resolution is achieved for finely-tuned HCI and BTI readings, and experiments are automated through a simple digital interface. This design allows us to test the frequency, temperature, and voltage dependencies of the stress mechanisms. In addition, we can monitor both sustained stress and recovery characteristics, and can observe the effects of increased load capacitance on the frequency shift induced by aging.

2.2 All-In-One Odometer Circuit Techniques

A block diagram of our proposed reliability monitor for separating the effects of HCI and BTI is shown in Fig. 2.1. This circuit contains four ROSCs in total: two stressed, and two unstressed to maintain fresh reference points. Each of the stressed oscillators is paired with its identical, fresh reference during measurements, and its frequency degradation is monitored with the Silicon Odometer beat frequency detection circuit [11].

2.2.1 Illustration of the Backdrive Concept

Fig. 2.2 presents the pair of stressed ROSCs in both (a) stress and (b) measurement modes. (Note that all body terminals are connected to their respective supply levels.) During stress, the BTI_ROSC stages are gated off from the power supplies, while the

DRIVE_ROSC maintains a standard inverter configuration with the supply set at VSTRESS. Both ROSC loops are opened, and the input of the DRIVE_ROSC is driven by a stress clock generated by an on-chip voltage controlled oscillator (VCO) whose output is level-shifted up to VSTRESS. The switches between these two ROSCS are closed so the DRIVE_ROSC can drive the internal node transitions for both structures.

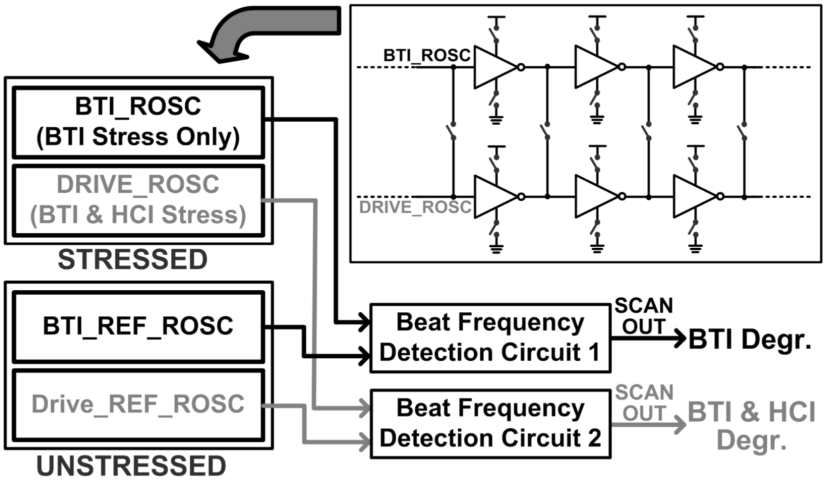


Fig. 2.1: High-level diagram of the All-In-One Silicon Odometer.

Simulated voltage and current waveforms are shown in Fig. 2.2(c). The internal nodes of the BTI_ROSC switch between the supply level (VSTRESS) and 0 V, as would be the case in standard operation. However, the peak drain current though the "on" devices in this structure is only 3-5% of that in the DRIVE_ROSC, since their sources are gated off from the supplies. Note that the sources of these "on" devices in the stressed BTI_ROSC are held at their respective supply levels due to the backdriving action of the DRIVE_ROSC. Therefore, the BTI_ROSC will age due only to BTI stress, while the DRIVE_ROSC suffers both BTI and HCI. We can extract the contribution of HCI to the latter ROSC's frequency degradation with the equation $HCI_{DEG} = DRIVE_{DEG} - BTI_{DEG}$, where DEG stands for degradation. During measurement periods, both ROSCs are connected to the digital logic power supply (VCC) and the switches between them are opened, so they each operate independently in a standard closed-loop configuration.

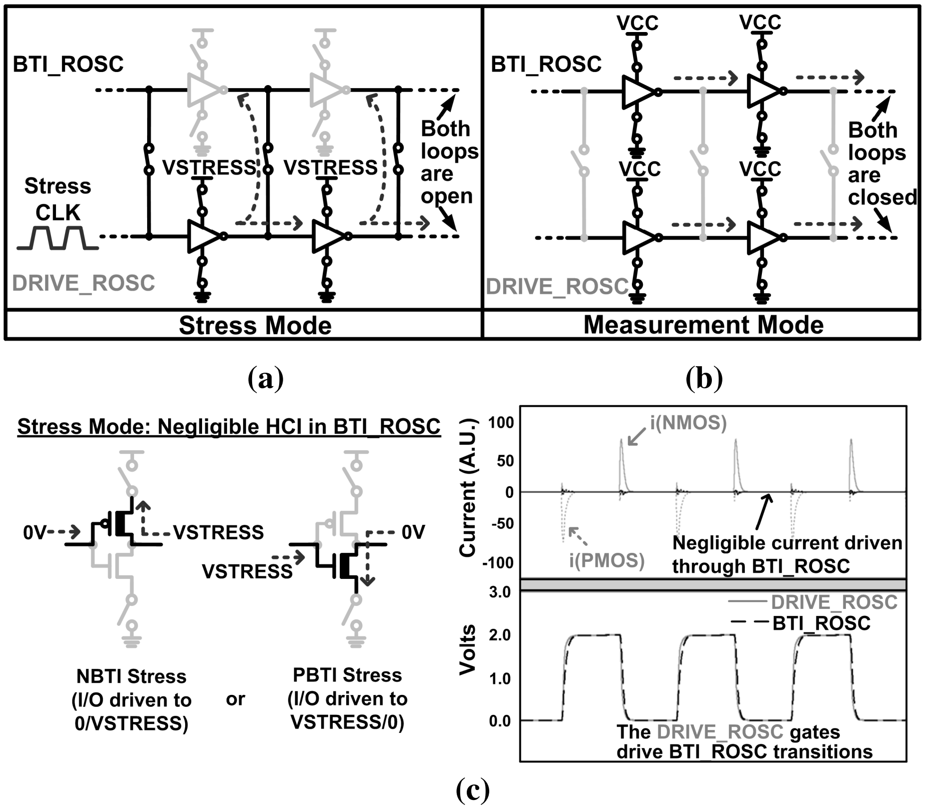
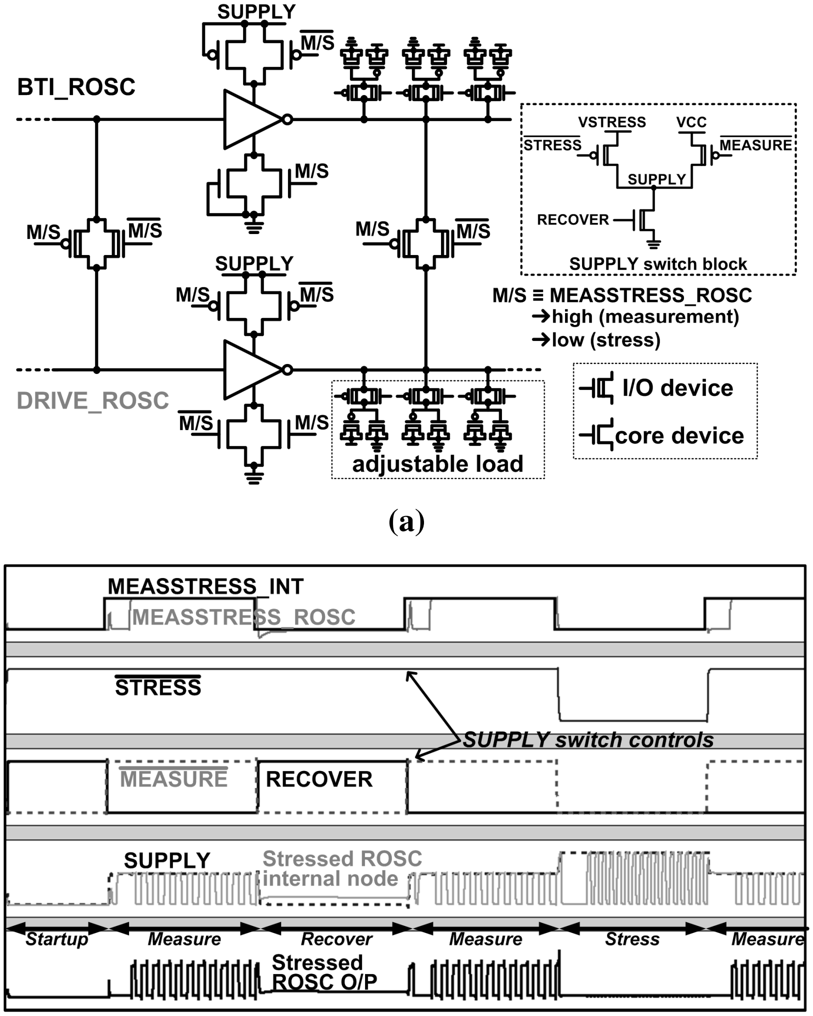


Fig. 2.2: ROSC configuration during (a) stress and (b) measurement modes. (c) The BTI_ROSC transistors suffer the same amount of BTI as the DRIVE_ROSC transistors during stress, but with negligible HCI degradation, since very little current is driven through the channels of the devices under test the former structure.

2.2.2 ROSC Design Details for Backdrive

A detailed schematic of one stage of the paired ROSCs is shown in Fig. 2.3(a). The thick oxide I/O devices should not age appreciably during stress experiments aimed at the thin oxide core transistors. All core devices are either stressed devices under test

(DUTs), or have no voltage drops across any pair of terminals during stress, so they will not age. The header and footer transistors in each inverter pin the source nodes of those gates to the supply levels when closed. The M/S signal here is used to start and end measurement periods. This signal is timed and driven by the on-chip finite state machine (FSM) after the external MEASSTRESS_EXT signal is asserted.



(a)

(b)

Fig. 2.3: (a) Schematic of one stage of the paired ROSCs. (b) Simulation waveforms from a stressed ROSC during measurement, stress, and recovery periods. Note that any initial lone pulses seen at the stressed ROSC output are rejected by the beat frequency detection logic.

Both ROSCs contain three levels of adjustable fanout which allows us to test the effects of additional load capacitance on aging. Extracted simulations show that turning on each additional stage of fanout increases the transition times by an average of roughly 22%. It is expected that these changes will adjust the balance between HCI and NBTI stress during normal voltage switching operation, as longer input and output transition times result in an increasing number of hot carriers [20], [21].

Fig. 2.3(b) contains waveforms from a stressed ROSC during measurement, stress, and recovery periods. After MEASSTRESS_INT is driven high by the FSM, there is a short delay before MEASSTRESS_ROSC goes high, which then causes the tapped output from the stressed ROSC to be connected to the input of the Odometer measurement system. This delay allows the SUPPLY node to settle after being switched to the standard operating supply of VCC, and having the ROSC loop closed. An external control signal is set high any time we wish to enter a recovery mode between measurements, but will not take effect until the end of the subsequent measurement period.

2.2.3 Silicon Odometer Background and Theory

The Silicon Odometer measures frequency changes in the stressed ROSCs with the concept illustrated in Fig. 2.4 (further details in [11]). During the short measurement periods, a phase comparator uses a fresh reference ROSC to sample the output of an identical stressed ROSC. The output signal of this phase comparator exhibits the beat frequency: $f_{PC} = f_{ref} - f_{stress}$. A counter is used to measure the beat frequency by counting the number of reference ROSC periods during one period of the phase comparator output signal (see Fig. 2.9(a), to be covered later). This count is recorded after each stress period to calculate the shift down in the stressed ROSC frequency.

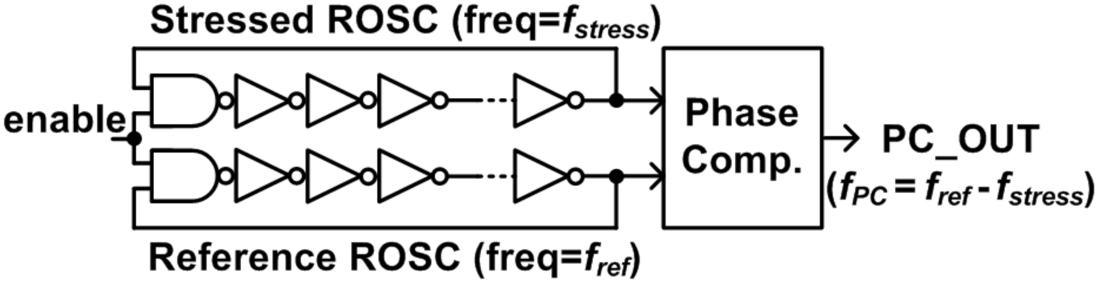

Stressed ROSC (freq=f_{stress})
enable
Phase Comp.
PC_OUT
($f_{PC} = f_{ref} - f_{stress}$)
Reference ROSC (freq=f_{ref})

Fig. 2.4: Beat frequency detection between a stressed and an unstressed ROSC. This system achieves sub-ps frequency shift resolution for the stressed ROSC, with sub-µs measurement times [11]. (More system details covered later in Fig. 2.9(a)).

The details of the beat frequency calculation can be found in the previous publication [11], but are summarized here for convenience. If the initial frequency of the reference ROSC is called f_{ref}, that of the fresh ROSC to be stressed is f_{stress}, and the initial Odometer output count is N_1, then assuming f_{ref} is higher than f_{stress}, we have:

$$1/f_{ref} \cdot N_1 = 1/f_{stress} \cdot (N_1 - 1) \tag{1}$$

The $(N_1 - 1)$ term arises from the fact that the stressed ROSC with the lower frequency, f_{stress}, will take one less period to cycle back to the same point in the reference ROSC period while both are oscillating. After a stress period ends, f_{ref} will remain unchanged, but f_{stress} will be decreased due to aging, and we call the new frequency f_{stress}' (later we will show that these calculations result in very small errors even if f_{ref} is modified by temporal variations along with f_{stress}). We also have a new output count (N_2), so the resulting equation is:

$$\frac{1}{f_{ref}} \cdot N_2 = \frac{1}{f_{stress}'} \cdot (N_2 - 1) \tag{2}$$

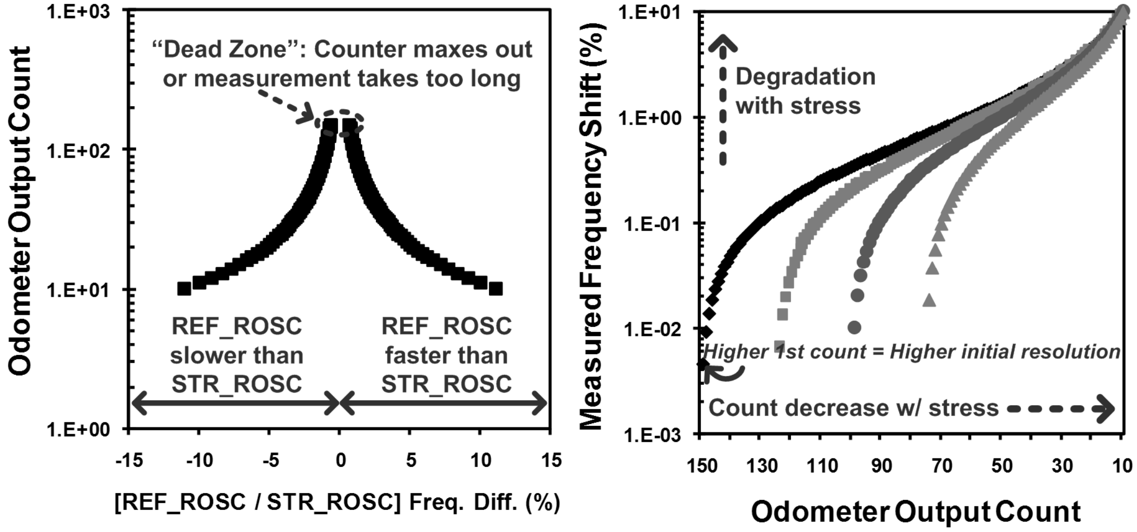
Using these two equations, we can calculate the frequency shift during stress as follows:

$$\frac{f_{stress}'}{f_{stress}} - 1 = \frac{N_1 \cdot (N_2 - 1)}{N_2 \cdot (N_1 - 1)} - 1 = \frac{(N_2 - N_1)}{N_2 \cdot (N_1 - 1)} \tag{3}$$

Those simple calculations show that if f_{ref} is only slightly higher than f_{stress}, the output count is high. For example, the count is 100 for a 1% difference. This slight difference can be ensured with trimming capacitors and calibration. The subsequent small decreases in f_{stress} due to aging cause a large change in this count. For instance, a 2% difference between the ROSC frequencies gives a count of 50, so a 1% shift to that point is translated into a decreased count of 50. Therefore, with high frequency ROSCs, the beat frequency detection system achieves sub-ps frequency shift measurement resolution.

The Odometer output count relationship with the difference between the reference ROSC (REF_ROSC) and stressed ROSC (STR_ROSC) frequencies is illustrated in Fig. 2.5(a). This figure shows that the Odometer operates correctly with a reference ROSC frequency that is either slower or faster than the stressed ROSC. In the former case, the output count will increase with stress, while it decreases in the latter. A slower reference frequency is accounted for in equations (1) through (3) by changing the *($N_{\#}$ – 1)* terms to *($N_{\#}$ + 1)*, because the faster stressed ROSC in this case goes through *one more* period than the slow reference during the beat frequency measurement, rather than *one less*.

Additionally, it is possible for the reference frequency to transition from being slower than to faster than the stressed ROSC, but this involves moving through a "dead zone" where the output count will either equal the counter max value, or if the counter is large enough, the measurement time will become excessively long as the difference between the two ROSC frequencies becomes extremely small.



(a) (b)

Fig. 2.5: (a) Silicon Odometer output count vs. the frequency difference between the reference and stressed ROSCs. When the two frequencies become extremely close, a high output count is observed, which requires a larger counter and longer measurement time. (b) Output count vs. frequency shift during a stress experiment. Curves are shown for varied initial counts, where a higher count corresponds to a smaller frequency difference between the two ROSCs, as was shown in part (a) of this figure.

We chose to start our experiments with a reference frequency that is slightly faster than the stressed ROSC frequency, so that we obtained a monotonic decrease in the output counts with stress. This allowed us to maximize the frequency measurement resolution in the early phases of stress, and to avoid the dead zone. Fig. 2.5(b) shows measurement result characteristics with monotonic count decreases, and four different initial counts. Note again that a smaller difference between the two ROSC periods leads to a higher initial count, and therefore a higher initial frequency resolution, while lengthening the measurement time. We achieved maximal starting counts of ~125 in our hardware measurements, which corresponds to initial frequency shift measurements ranging down to 0.0065%. The resolution decreases with time, but we are primarily concerned with the small initial degradation steps that can be obtained with stress that is closer to real operating conditions. It has been shown that stress at excessively high voltages, for example, can lead to unrealistic degradation characteristics that are not useful for predicting device lifetimes under standard operating conditions [5], [22].

The plot in Fig. 2.6 shows the theoretical maximum frequency measurement resolution for three measurement setups during a fixed time. In the "1 ROSC T-Counter" system (where T stands for period), a single ROSC's degradation is recorded with a single period counter during an externally controlled measurement time. The "2 ROSC T-Counter" measures the degradation in one stressed ROSC by counting the number of periods it cycles through while a set number of periods in a fresh reference ROSC are counted (see Fig. 2.8). Since the resolution of these period counters is simply the measurement time divided by the ROSC period, while that of the Odometer can be derived from equation (3), we see that the Odometer reaches a maximum resolution of 0.01% within only 0.3 μs in the ideal cause with a single measurement recorded, while the other systems require 100X more time. A large improvement is still seen when three counts are recorded during each Odometer measurement period for averaging, or to eliminate unpredictable initial counts (see Section 2.2.4). The longer measurement times in the standard period counter systems would result in unacceptable unwanted BTI recovery.

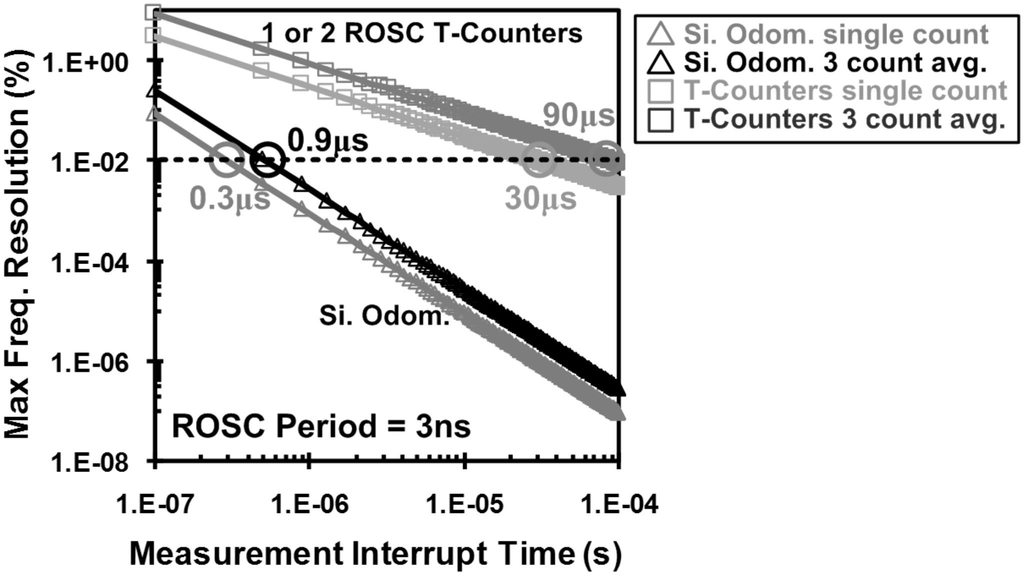
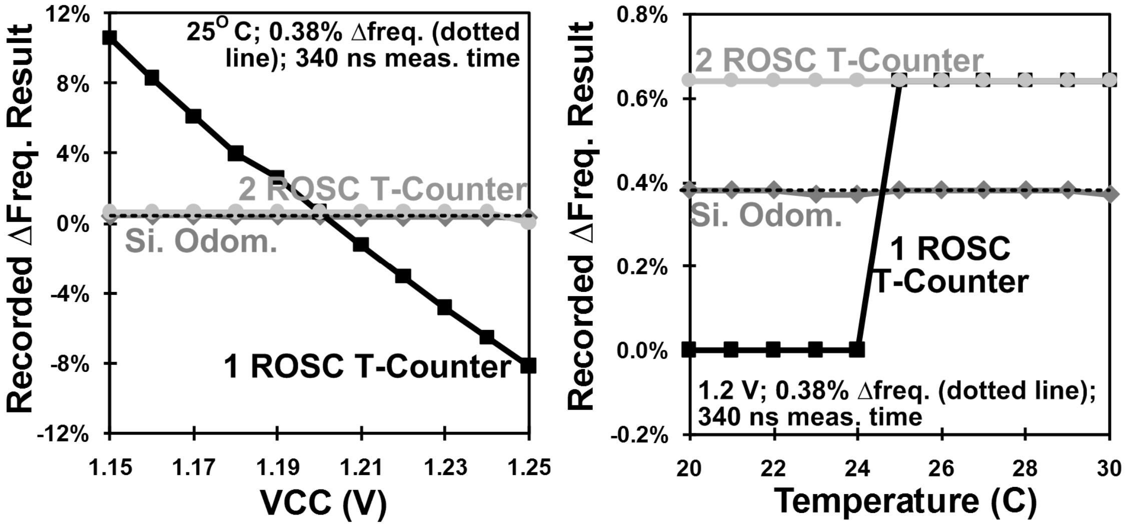


Fig. 2.6: Maximum frequency measurement resolution versus the total stress interruption time for measurements (note: lower frequency shift measured = higher resolution). A standard ROSC period counting system requires a 100X longer measurement time than the Odometer to achieve a measurement resolution of 0.01%.

In addition to the high frequency resolution, the Odometer benefits from a high immunity to voltage or temperature variations due to its differential nature. Given that the reference and stressed ROSCs are identical structures that are laid out next to each other, we assume that both will see essentially identical temporal variations, so their frequencies should be affected by roughly the same amount. The simulation results shown in Fig. 2.7 illustrate this noise immunity, and compare the Odometer results with those of the ROSC T-Counter setups. In these simulations, the stressed ROSC started out

0.64% slower than the reference when measured at nominal VCC (1.2 V) and temperature (25^{O} C), and the former structure is slowed by 0.38% due to aging (in the 1 ROSC T-Counter we only consider the 0.38% shift since there is no reference ROSC). However, if this post-stress measurement takes place under a different temperature or voltage condition, it will lead to some deviation from 0.38% in the measured value. Fig. 2.7 presents the simulated results gathered in this situation, and shows a clear benefit for the differential Odometer system. Also note that since we limited the measurement time to the ideal required by the Odometer system, the T-Counters suffer from low frequency resolution, which results in further rounding errors.

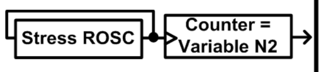
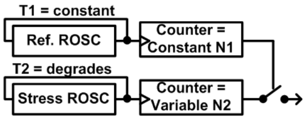
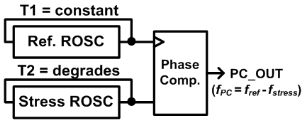


(a) **(b)**

Fig. 2.7: Simulated effects of (a) voltage, and (b) temperature variations on the Silicon Odometer, and both 1 and 2 ROSC period counter (T-Counter) systems. The values shown here are the results recorded by each system when the actual stress-induced frequency shift is 0.38%. We assume both ROSCs in the differential systems see the same variations since they are adjacent and identical in the layout. The larger rounding errors seen in the T-Counter measurements at small percentages are a result of their lower frequency resolution with short measurement times.

The Fig. 2.8 compares the three frequency measurement systems that have been discussed. While the Odometer requires additional circuits for the beat frequency detection, it achieves a significantly higher frequency measurement resolution in a shorter measurement time, and is immune to common mode environmental variations.

System	1 ROSC T-Counter	2 ROSC T-Counter	Silicon Odometer

Block Diagram		 T1 = constant
		Ref. ROSC
		Counter = Constant N1
		T2 = degrades
		Stress ROSC
		Counter = Variable N2
			 T1 = constant
			Ref. ROSC
			T2 = degrades
			Stress ROSC
			Phase Comp.
			PC_OUT
			($f_{PC} = f_{ref} - f_{stress}$)

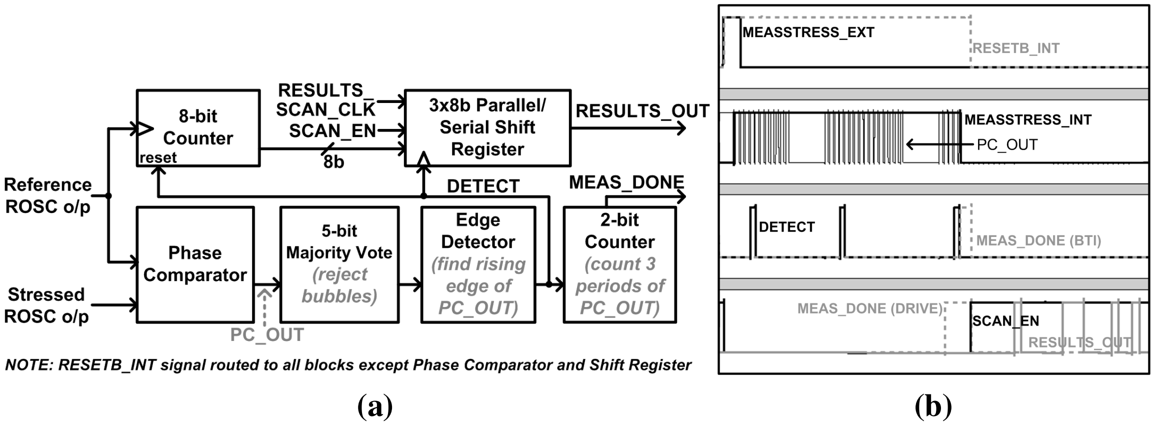
Function	Count Stress ROSC periods during externally controlled meas. time	Count Stress ROSC periods during N1 periods of Ref. ROSC	Count Ref. ROSC periods during one period of PC_OUT
Features	Simple; compact	Simple; immune to common mode variations	High resolution w/ short meas. time; immune to common mode variations
Issues	Voltage and temp. varations; meas. time vs. resolution tradeoff; requires absolute timing reference (e.g. oscilloscope)	Meas. time vs. resolution tradeoff	Requires extra circuits (e.g., Phase Comp., edge detector, etc...)
Meas. time for 1% max resolution *	30 μs	30 μs	0.3 μs
Meas. error wrt. common mode variations **	+10.18% / -8.57%	+0.26% / -0.38%	+0.06% / -0.07%

*ROSC period = 3 ns ** simulated with +/- 4% ΔVCC; 0.38% stress shift; 340 ns measurement time; error = (measured %) – (0.38%)

Fig. 2.8: Comparison of simple ROSC period counting systems with the Silicon Odometer.

2.2.4 Improved Silicon Odometer Beat Frequency Detection Circuit

In this work, we improved the beat frequency detection system by including logic which sends the circuit back into stress after three results are recorded, in order to achieve measurement times of ≤1 μs. The completion of a measurement period is flagged by the MEAS_DONE signal (Fig. 2.9(a)) when the three rising edges from the phase comparator are counted, meaning three 8b count results have been recorded. In this automated scheme, the first two counts are generally smaller than the true result due to the unpredictable starting location of the measurement at some mid-point in the phase comparator period, so they are discarded. We verify that the third count is correct during calibration by using an externally controlled longer measurement period in which the initial smaller counts are overwritten by subsequent results. In this case, all counts should be roughly identical, and equal to the third result we record during the shorter automated measurements. Moving on, the MEAS_DONE flag is sent to the FSM, which restarts stress after it is asserted by both Odometers. Using on-chip logic to control this timing allows us to avoid generating very short, accurate measurement pulses externally.



NOTE: RESETB_INT signal routed to all blocks except Phase Comparator and Shift Register

Fig. 2.9: (a) Block diagram of the improved beat frequency detection circuit. (b) Simulation results illustrating the operation of this system.

The majority voting circuit (Fig. 2.9(a)) rejects a lone ‘1’ signal in a series of ‘0’s, or vice versa. These “bubbles” can be caused by temporal variations. The edge detector is used to find the beginning of each period of the phase comparator. Its output, DETECT, is used to sample the counter output, and then to reset the counter for a new period.

Fig. 2.9(b) contains simulation waveforms illustrating the operation of this system. After the external MEASSTRESS signal is asserted, its internal counterpart is driven high by the FSM, which connects switching signals from the ROSCs to the phase comparator, and starts the measurement. After three high PC_OUT periods, we see the MEAS_DONE signal go high. As noted, these waves are from the Odometer monitoring the BTI_ROSC. The bottom line of this figure shows that the MEAS_DONE signal in the Odometer system monitoring the DRIVE_ROSC has already gone high. The combination of these two signals causes the FSM to end the measurement period and switch the parallel/serial shift registers to scan mode. An external clock is then used to scan out the results. The registers will be put back into parallel mode when MEASSTRESS_EXT is next asserted.

2.2.5 Test Setup and Procedure

A high-level pin diagram of the All-In-One Odometer system is shown in Fig. 2.10. VCO_BIAS is used to set the STRESS_CLK frequency, and MEASSTRESS_EXT is pulsed to initiate each measurement period. RECOVER_EXT is asserted to send the stressed ROSCs into recovery mode after the *next* measurement period. RESETB_EXT immediately sends the circuit into its initial startup state, where the SUPPLY node in the ROSCs is dropped to 0 V, and the FSM is left waiting for the next MEASSTRESS_EXT pulse. The RESULTS_SCAN_CLK signal is pulsed 48 times after each measurement is completed, which is indicated by a rising edge on COMPLETE. The results registers in the two Odometers are connected in series, so 48 pulses are required for the two sets of three 8b registers. Finally, VCO_CHECK is used to monitor the frequency of the VCO, and RESULTS_OUT is the scan-out port for both Odometer results.

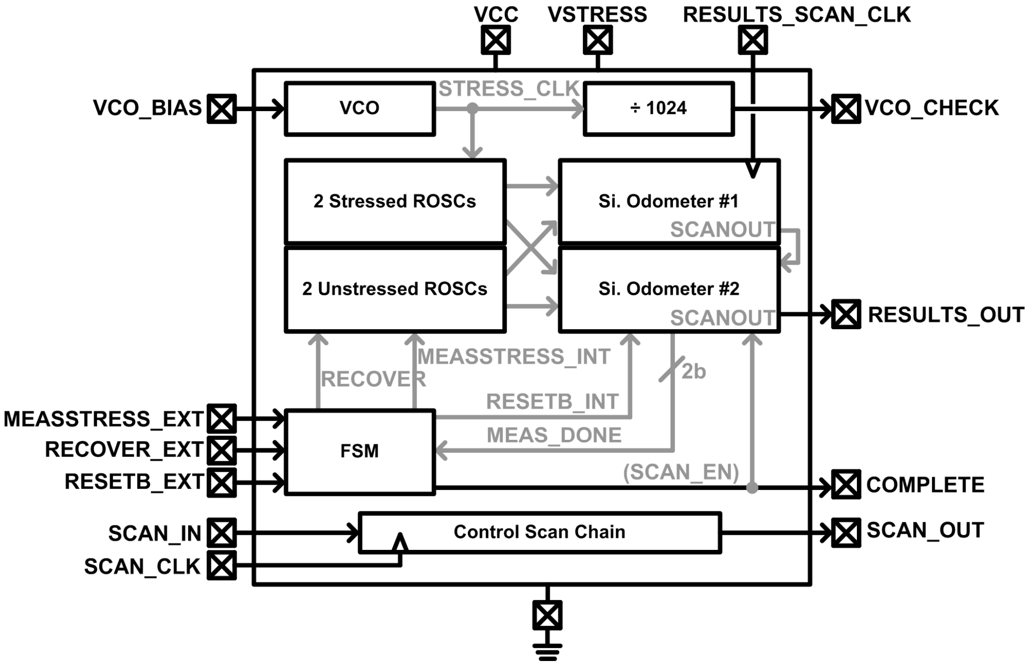
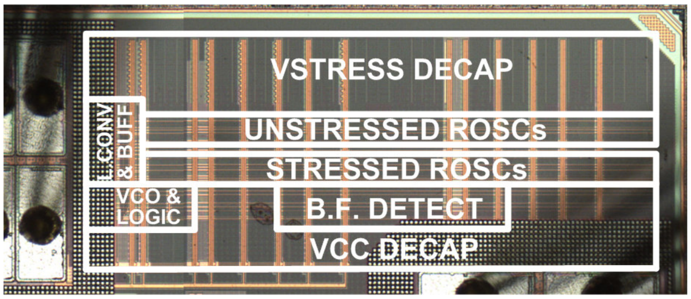


Fig. 2.10: High level pin I/O diagram with major internal signal routing.

2.3 All-In-One Odometer Test Chip Measurements

A 214x551 μm^2 test circuit was implemented in a 65 nm bulk CMOS process for concept verification. A die photo and a summary of test chip characteristics are presented in Fig. 2.11. Measurements were automated with LabVIEW^{TM} software through a National Instruments data acquisition board. Trimming capacitors were used in each 33 stage ROSC to ensure that the frequencies of the stressed structures began slightly slower than the reference frequencies (see Section 2.2.3). Trimming was also utilized to push apart the oscillating frequencies of the two sets of paired ROSCs to prevent injection locking. The DUTs were 1.5 µm/60 nm NMOS and 3 µm/60 nm

PMOS transistors in the inverter stages of the stressed ROSCs. All automated measurement times were under 1 μs, but varied according to the exact beat frequency count results, as shown in Fig. 2.6.



Technology	65 nm, 7M
Logic/IO Supplies	1.2 V / 2.5 V
Active Area	38,040 μm^2
Total Area	(214.22 X 551.43) μm^2
ΔT Resolution	< 1 ps
Measure Interrupt	< 1 μs

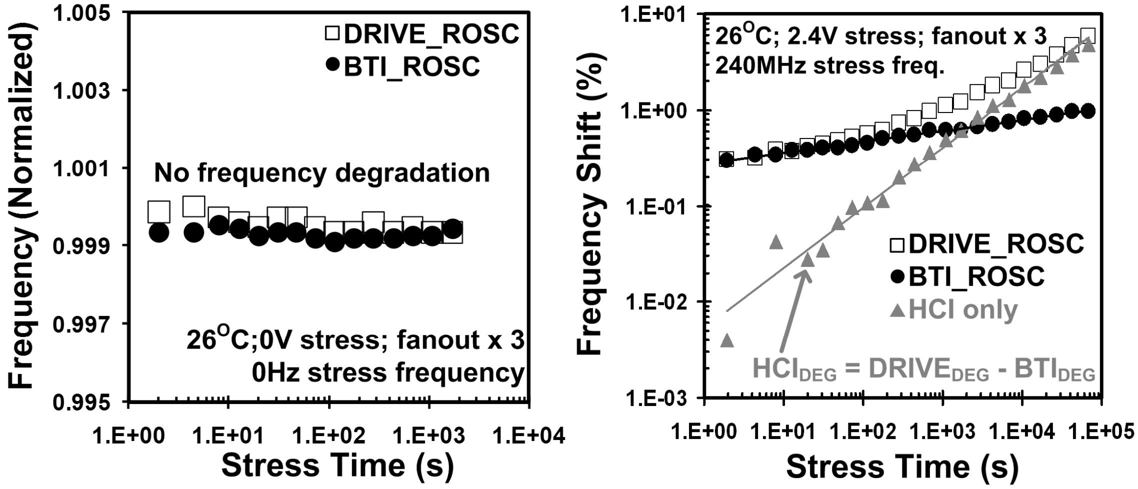
Fig. 2.11: Test chip microphotograph and summary of characteristics.

2.3.1 Circuit Verification Measurements

We first we checked the result of a 0 V stress experiment, meaning the SUPPLY node of both normally stressed ROSCs was dropped to 0 V between measurements periods by keeping the RECOVER_EXT signal high, so no aging should have taken place. The results in Fig. 2.12(a) confirm this outcome, so we can be confident that frequency shifts shown in later results are not due to aging elsewhere in this system or other circuit effects.

Fig. 2.12(b) presents example measurement results for both ROSCs under 2.4 V stress, as well as the calculated degradation due to HCI (HCI_{DEG}). As expected, both BTI and HCI degradation follow a power law behavior, although the latter is seen to saturate at long stress times. This can be explained by the finite number of bonds to be broken at the Si-SiO_2 interface and/or the self-limiting nature of HCI, where the degraded drain current produces fewer hot carriers. The power law exponent for BTI in this case was

0.12, while that of HCI was 0.63 in the range fitted on this plot. The larger value for HCI is expected, and one possible reason for this is an increasing contribution of broken Si-O bonds at the oxide interface during HCI stressing, rather than Si-H bonds [1], [23].

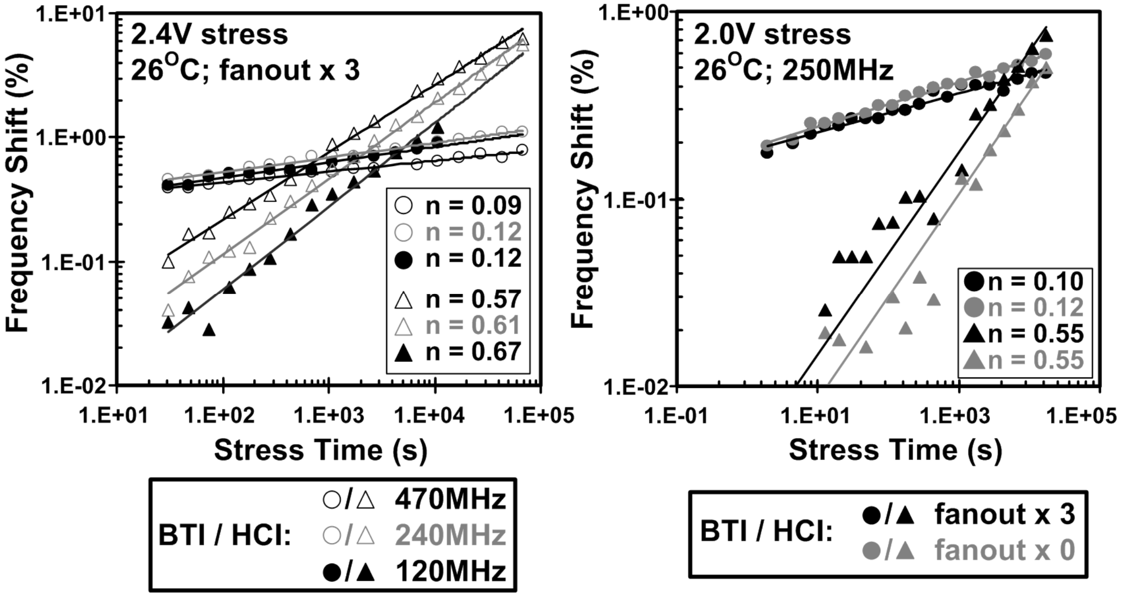


(a) (b)

Fig. 2.12: (a) Results from an experiment in which the RECOVER_EXT signal was asserted to prevent the DUTs from being stressed. As expected, no frequency shifts are observed, so shifts in subsequent stress experiments can be attributed to device aging rather than any undesired circuit effects. (b) Example measured results with AC stress conditions.

2.3.2 BTI and HCI Stress Measurements

Fig. 2.13(a) illustrates the impact of frequency on BTI and HCI. These results verify that BTI is at most weakly dependent on frequency, while HCI degrades with increased switching activity. More switching leads to an increase in current driven through the DUTs' channels, meaning more hot carriers are present. A decrease in the power law exponent of HCI was observed at higher frequencies, which is apparently due to the quick saturation of degradation in this case. In Fig. 2.13(b) we see that increased load capacitance, which causes longer transition times, accelerated HCI and had little impact on BTI. This acceleration of HCI with both increased input transition time and output load capacitance was reported in early HCI work [20], [21]. Those variables have been listed as two of the main controllable factors affecting hot carrier-induced degradation.



(a) **(b)**

Fig. 2.13: Measured frequency degradation results for (a) three stress frequencies and (b) increased load capacitance, with power law exponents (n).

Fig. 2.14(a) shows BTI's positive correlation with temperature, and that HCI aging was slightly reduced at higher temperatures due to increased phonon scattering, which reduces drain current. Both aging mechanisms degrade with voltage (Fig. 2.14(b)), and we observe a decrease in HCI's power law exponent at lower voltages. This has been explained by a possible decreasing contribution of broken Si-O bonds (in comparison to Si-H bonds) at lower voltages, closer to real operating conditions [1], [23]. Also note the crossover point when HCI begins to dominate the overall aging is pushed out in time by an order of magnitude at 1.8 V stress compared to 2.4 V. This helps to illustrate the claim that BTI becomes dominant in modern technologies operating at lower supply levels.

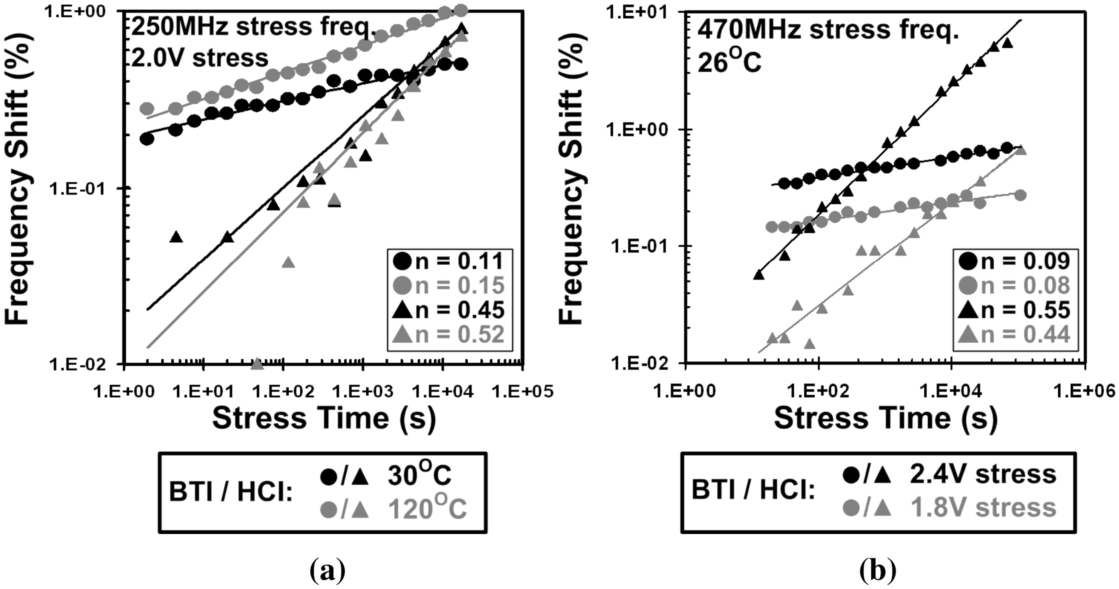


Fig. 2.14: Effect of (a) stress temperature and (b) stress voltage.

Fig. 2.15 shows a common NBTI recovery characteristic, while the HCI_{DEG} component did not improve when stress was removed. One explanation for this behavior is the Si-H bonds at the interface broken by cold carriers during BTI stress are recoverable, while hot carriers also break Si-O bonds, which do not recover [1], [23].

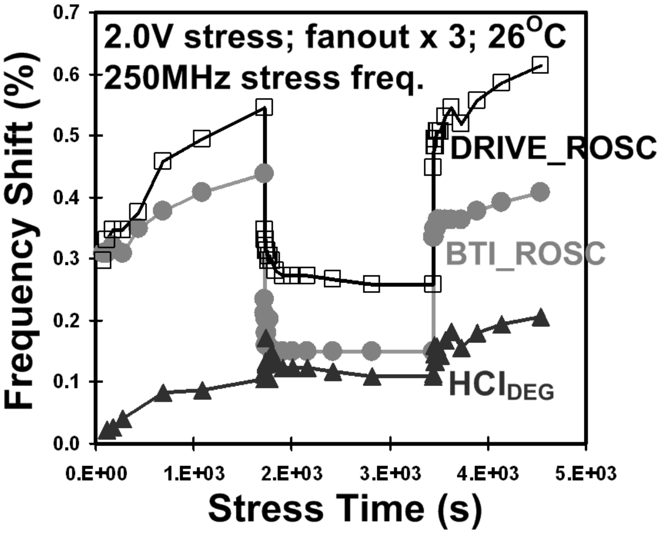


Fig. 2.15: Periodic stress/recovery characteristics. The BTI frequency curve shows a common sawtooth characteristic, while the HCI curve does not recover when stress conditions are removed.

2.3.3 TDDB Measurements in Stressed Ring Oscillators

Fig. 2.16 presents three examples of high voltage stress experiment results in which sudden jumps in ROSC frequency are interpreted as breakdown events. Thus far we have been ignoring TDDB in our results because it acts on a much longer timescale at lower stress voltages. In these experiments involving large frequency shifts, we did not use the beat frequency detection framework since it is aimed at high resolution measurements for smaller shifts. Instead, we directly read the frequency off-chip with an oscilloscope. Note that longer term experiments, or those done in future technology generations where soft breakdowns are more prevalent, will be able to make use of the Odometer system. Fig. 2.16 shows that ROSCs do continue to function after one or more breakdowns, which only lead to reduced output swing and lower frequencies, as long as subsequent logic stages in the ROSC can restore full-rail swing [24].

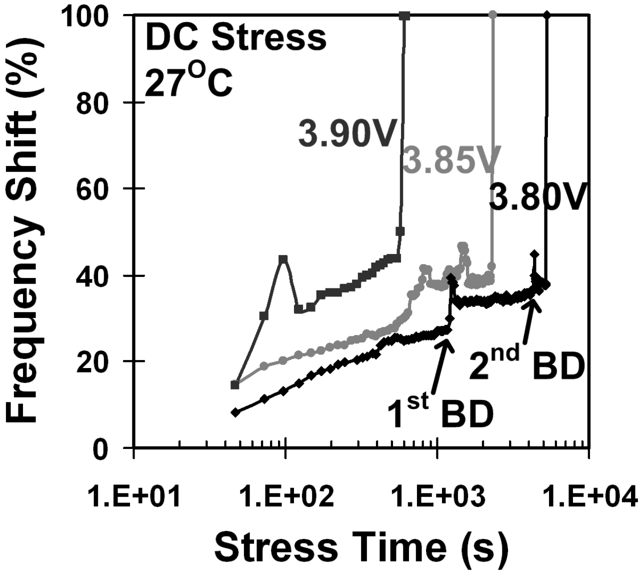


Fig. 2.16: ROSC frequency jumps attributed to TDDB before final circuit failure. The ROSCs continue to function after one or more apparent breakdowns.

2.4 Conclusions

We have implemented a test circuit in 65 nm technology that is capable of separately monitoring the frequency degradation induced by HCI, BTI, and TDDB. Sub-μs measurements are controlled by on-chip logic, and sub-ps frequency measurement resolution is achieved using the Silicon Odometer beat frequency detection system. This combination of fast measurements, which can avoid unwanted BTI recovery, along with high frequency resolution is facilitated by the Odometer framework, and is not possible in other standard measurement setups. We use a concept called "backdrive" to isolate BTI-induced aging in a ROSC gated off from the stress supply. This novel all-digital system can be used during process characterization, or for accurate real-time reliability monitoring and compensation schemes.

Chapter 3

A Statistical Silicon Odometer for Measuring Variations in Circuit Aging

3.1 Introduction to Statistical Transistor Aging

Transistor aging is the product of a finite number of trapped charges or broken bonds in exceedingly minute modern devices, so it is no longer sufficient to rely on a small set of stress tests to predict the behavior of billions of devices over the lifetime of a circuit. Statistical fluctuations in the number and spatial distribution of defects contributing to transistor degradation lead to a range of effective device "ages" at any given time [8], [9], [25]-[31]. This issue is well understood in the study of time dependent dielectric breakdown (TDDB), but has yet to be fully addressed under bias temperature instability (BTI) and hot carrier injection (HCI) stress.

In smaller transistors, the number of defects contributing to aging is reduced, making the relative impact of the creation or destruction of each of them more significant. Much like variations induced by random dopant fluctuations, aging-induced variation scales inversely with gate area [25], [26]. In addition to variations in the number of defects, the size of the step they induce in the measured parameter of interest (e.g., V_{th}), is also randomly distributed. That step size is dependent on each charge's position relative to the others [27]. Some authors have also claimed that the widely distributed time constants of the defects created during BTI stress contribute to the variation [28]. For these reasons, variations in transistor aging, particularly that due to NBTI, has received increasing attention in recent years as CMOS technology has been pushed into the deep sub-micron regime. Most of the work to this point has focused on statistical device-level measurements, and efforts to model the impact of the spread in aging characteristics on sensitive analog applications or SRAM stability.

However, little has been done to investigate the impact of varied aging characteristics on digital logic. In this paper, we present a novel measurement system that facilitates efficient statistical aging measurements involving the latter two mechanisms in ring oscillators (ROSC). The distribution of frequencies is monitored by a set of three Silicon Odometer beat frequency detection systems working in parallel. Unwanted BTI recovery during stress interruptions is avoided with measurements of down to 1 μs. Frequency shift measurement resolution ranging down to the error floor of 0.07% is achieved in combination with those quick measurements, which is not possible with standard test setups.

3.2 Prior Work in Statistical Aging Characterization

It has long been understood that TDDB is governed in part by a statistical component related to the critical density of defects required to form a conducting path through the gate dielectric. The breakdown process is well-described by Weibull statistics, and suffers from a larger spread in times to breakdown as dielectric thicknesses are scaled down with other device dimensions. BTI and HCI, in contrast, have generally not been thought of as processes that display significant spreads in their degradation characteristics. However, a growing body of work is currently investigating this topic.

In 2002, Rauch demonstrated that identical PMOS devices stressed under the same conditions will experience different amounts of degradation due to the number and spatial distribution of induced charges, and that this variance increases with stress time [29]. He claimed that random variations in both of those parameters scale inversely with gate area in a similar manner, and their effects on V_{th} variations are of roughly the same magnitude. Rauch presented results demonstrating the impact that NBTI-induced variations could have on paired PMOS device mismatch in sensitive analog applications, and observed that the induced mismatch shifts were not correlated to the initial mismatch values.

La Rosa *et al.* later investigated the impact of variations in NBTI on the stability of SRAM cells through a study of the "N Curve" characteristic [8]. The authors showed that any reliability assessment of SRAM cell stability must account for additional variation due to NBTI, rather than just the mean stress shift. Rauch then summarized both his mismatch work and this SRAM study in a 2007 paper before moving on to define the shape of the aging-induced threshold shift distributions [27]. He showed that NBTI degradation is the net result of two Poisson processes: the creation and destruction of charges. The difference between two Poissons is called a Skellam distribution, and measured results were shown to closely fit this model. Both the mean and variance of the threshold shift were found to increase with stress time, but the ratio of variance to mean, called the "dispersion factor," decreases. Rauch once more points out that the variation introduced by NBTI aging in V_{th} and the device transconductance is uncorrelated to the initial distributions.

Huard *et al*. also investigated the impact of NBTI variation on large SRAM arrays [9]. In this work, this group expanded upon their own "composite model" of the mean shift in V_{th} due to NBTI, which they claim is due to both permanent interface trap creation and transient hole trapping/detrapping, to account for variability in device aging. The authors show that threshold shift variation is linearly proportional to the mean shift, and that the permanent part of this shift can be roughly estimated by a normal distribution when smaller sample sizes are considered. However, when large sample sizes are studied, they find, like Rauch, that the Skellam distribution is appropriate. Both Huard and Rauch claim that the non-normality of the ΔV_{th} distribution is smoothed out to a certain degree in the context of large SRAM arrays when it is convolved with the fresh Gaussian V_{th} distribution. Huard also verified that variation in the V_{th} shift is dominated by random stress-induced charge fluctuations and not process variation.

Kang *et al.* presented a statistical NBTI model based on the Reaction Diffusion (RD) framework, which only explicitly considers the creation and passivation of interface states [30]. (Note that other researchers have pointed out that while the RD model predicts the stress phase of NBTI degradation, it does not accurately describe the recovery transient [2], [9], [32], [33].) The authors claim that the standard deviation of the threshold shift increases with stress time, having a power law exponent that is half of that followed by the mean value (1/12 and 1/6 respectively). This leads to a decrease in the variance over mean ratio, or the "dispersion" as it is referred to in Rauch's work. Kang *et al.* go on to claim that the NBTI-induced circuit speed variation is reduced with increasing logic path depth due to averaging effects. They also state that random spreads in circuit performance tend to be dominated by lower granularity variations (e.g., inter-die), so NBTI variation may only account for a small portion of the overall distribution.

Pae *et al.* presented BTI variation measurements from 90 nm, 65 nm, and high-k, metal gate 45 nm technologies [26]. They emphasized the point that, although similar mean threshold voltage shifts due to BTI stress can be observed between these scaled processes, the standard deviation of the aging-induced shifts are inversely proportional to gate area, and therefore worsen with scaling. The authors also show data indicating that the fresh, time zero threshold voltage spread (primarily due to RDF), is not correlated to the BTI-induced variability, and the standard deviation of the shift increases with stress time, following power law behavior.

Tu *et al.* studied variation in the HCI- and NBTI-induced threshold voltage shifts in PMOS devices from a 90 nm technology [31]. They found that matching was more aggravated after HCI stress than BTI due to a larger standard deviation in the threshold shift. However, the authors do not describe the details of their measurement setup, and present a power law exponent of 0.255 for NBTI which indicates slow, inaccurate measurements.

Finally, in 2008, Wang *et al*. developed a statistical methodology that was claimed to predict circuit performance under the influence of process variations and NBTI [34]. The variation in the NBTI degradation mechanism itself was not considered in this work. The authors only took into account the initial process variations, which led them to the conclusion that the standard deviation in logic speed decreases with stress time, following a power law exponent of 1/6. This is contrary to all previous studies of variation in NBTI, as well as our own measurement results, which will be presented in Section 3.4.

As described in this section, most of the previous work in BTI-induced variation studies has involved modeling work, and measurements of individual devices, matched transistor pairs, or SRAM cells and arrays. In this paper, we present measured results from an array-based system that facilitates efficient statistical aging measurements in ring oscillators (ROSC). Our results indicate that the spread in the aging-induced speed shifts increases along with the mean value during stress, and we observe a decreasing standard deviation to mean ratio with stress time. In addition, no correlation is found between the stress-induced frequency shifts in the ROSCs and their fresh frequency values. Several other conclusions will be described later in this paper, demonstrating the utility of our design, and the importance of considering aging-induced variation in scaled circuits.

3.3. Statistical Odometer System Design

The statistical measurement system we have implemented consists of a 10x8 array of cells containing ROSCs to be stressed, a finite state machine (FSM) for control, a scan chain, and three Silicon Odometers with their reference ROSCs (Fig. 3.1). During tests, the whole array of ROSCs, or any one rectangular group of them, are stressed in parallel, and selected one-by-one for measurements. Alternatively, all of the ROSCs being tested can be put into a recovery state (i.e., 0 V supply), along with any cells that are not selected. During stress, the ROSC loops are opened so that their frequencies can be controlled by an on-chip voltage controlled oscillator (VCO). When each oscillator is selected for a measurement, its supply is set to the standard digital level of 1.2 V, the loop is closed, and its oscillation frequency shift is measured by the three Odometer systems.

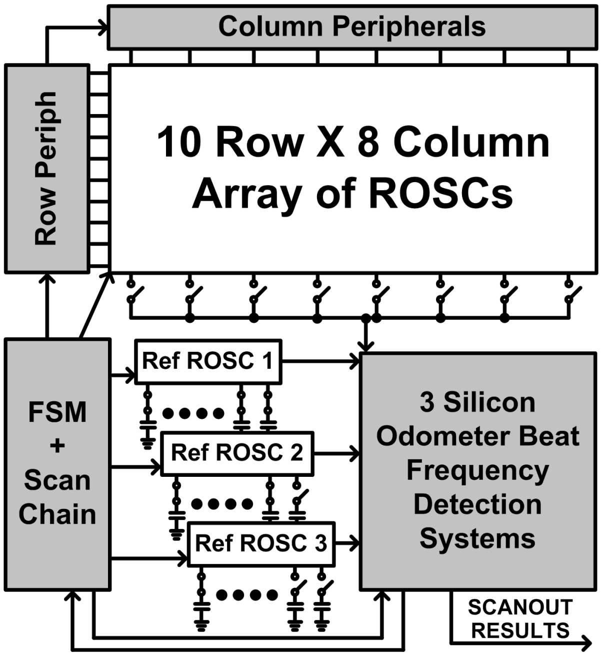


Fig. 3.1: Top level diagram. Reference ROSCs each have 15 trimming capacitors controlled by the scan chain.

3.3.1 Ring Oscillator Cell Design

Each ROSC cell in the array contains its own supply switch that sets the local virtual supply (CSUPPLY) at the stress level (VSTRESS), 1.2 V (VCC), or 0 V when the cell is not being measured or stressed (Fig. 3.2). The control logic for these switches is carefully designed to prevent direct current paths between any supply rails during transitions from one level to another. Also note that the body connection of the VCC PMOS switch is biased at the VSTRESS level. This prevents the PN junction at its drain from becoming forward biased, since the latter value is always greater than or equal to VCC.

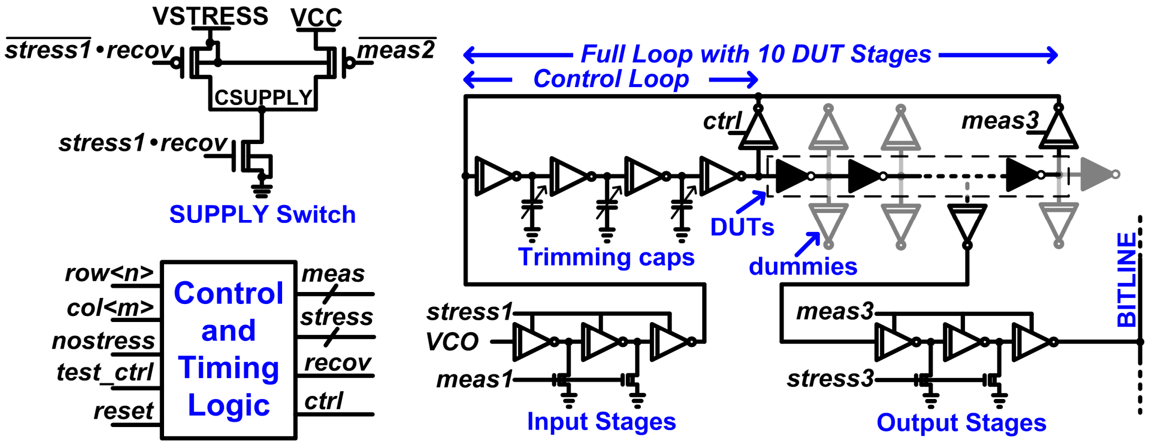


Fig. 3.2: ROSC cell design. The thin oxide logic stages under test are colored black, and all other transistors are thick oxide I/O devices (indicated by double lines).

As was just mentioned, the ROSC loops are opened during stress or recovery periods so that the devices under test do not switch at a rate determined by the cell's virtual supply. Instead, that rate is set to zero, or determined by an on-chip VCO whose high output is level shifted up to VSTRESS and fed into the array. Each cell contains the selection logic to switch a cell into measurement mode by closing the loop, and connecting one tapped output node to the bitline after the virtual supply has had time to safely settle to the VCC level. This timing is indicated in Fig. 3.2 by the order of the *meas* ("measure") and *stress* signals. The former go high during measurement periods, and the latter are driven high during stress or recovery when the cell is not selected for a measurement. First, meas1/stress1 turn off the input stages and the VSTRESS switch. Next, meas2 turns the VCC switch on, and finally meas3/stress3 close the loop and connect it to the bitline.

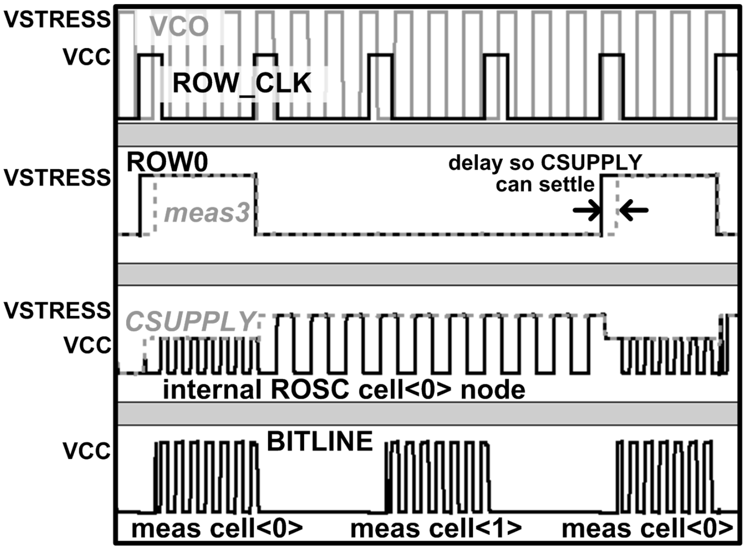


Fig. 3.3: Waveforms illustrating the basic operation of a ROSC cell. In this simulation, only two cells are included in order to demonstrate the functionality as cell<0> goes into and out of stress periods in a short simulation time.

When a cell is sent back to stress or recovery, this ordering is roughly reversed. The ROSC is first disconnected from the bitline to prevent any unwanted stress or currents in other portions of the array. At nearly the same time, the ROSC loop is opened, the VCC switch is turned off, and the input path is turned on. Finally, the VSTRESS or GND supply switch is opened to start stress or recovery, respectively. Note that several tri-state inverters and pulldown transistors are placed between the VCO input and the ROSC, as well as the ROSC and the bitline, in order to prevent any coupling when those connection paths are shut off.

The basic operation of one cell is illustrated in Fig. 3.3. In this example, only two cells are included in the simulation setup in order to demonstrate the functionality of one of them as it quickly enters and leaves stress mode. The ROW_CLK signal from the state machine starts and then stops each measurement with two consecutive pulses, as will be explained in Section 3.3.4. We see the delay between CSUPPLY dropping to VCC and meas3 closing the loop to start a measurement. This prevents unstable oscillations as that virtual supply settles after the switching event.

Ten inverters in each ROSC are 1.2V thin oxide logic devices under test (DUTs). These stages will age during stress experiments, while the rest of the stages, composed of 2.5 V thick oxide I/O transistors, are not significantly impacted. However, the thick oxide control stages contribute to the full loop delay, which must be accounted for when calculating the stressed stages' delay shift due to aging.

Therefore, in addition to the full loop that is selected with the *meas3* signal, we added a replica control path selected with *ctrl* (both are turned off to open the loop during stress). The delay of the replica path is roughly equivalent to that of the control logic in the full loop. So by first measuring the control loop frequency, and then that of the full loop during automated circuit calibration, we can calculate the percentage of the fresh full loop delay accounted for by the logic devices under test (i.e., $1 - f_{full}/f_{ctrl}$). Extracted simulations found the real delay of the DUT stages to be only 0.55% longer than the calculated value. Later, the total frequency shift of each stressed full loop measured by the Odometers is divided by the percentage of the fresh delay taken by the DUTs, in order to calculate the degradation in those thin oxide stages. All DUT stages have identical loads and layouts due to the use of dummy cells.

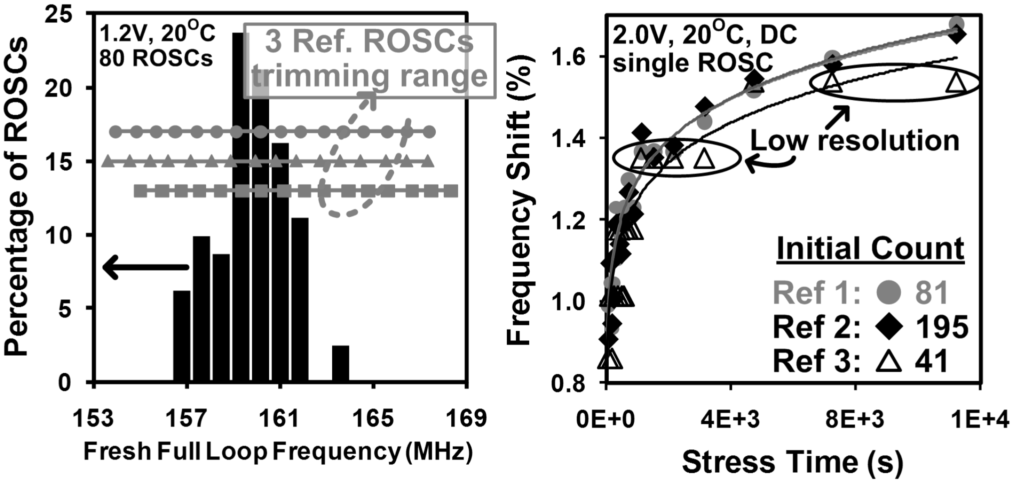
3.3.2 Silicon Odometer Beat Frequency Detection

The Silicon Odometer is an all-digital differential system that measures frequency changes in the stressed ROSCs with high resolution, theoretically ranging down to < 0.01%, and measurement times ranging down to < 1 µs. The details of the beat frequency detection circuits and calculations can be found in Section 2.2.3 and [11].

3.3.3 Multiple Reference ROSCs and Frequency Trimming

An Odometer provides high-resolution frequency shift measurements when the frequencies of the ROSC under test and reference are close. This is ensured with trimming capacitors, which are implemented with NMOS drain diffusions attached to internal nodes of the ROSCs. The sources of these devices are left floating while the gate values are driven high to increase the nodal capacitance, and hence the oscillation period. In past Odometer test circuits, each trimming capacitor on both the reference and stressed oscillators have been individually controlled with scan chain bits, allowing us maximal freedom in the frequency trimming range.

However, in the present circuit where many ROSCs are stressed in parallel and selected one-by-one for measurements, controlling the trimming bits in each stressed oscillator would be time and area consuming. Therefore, we instead hardwired nine of fifteen capacitors "on" in each of those ROSCs, while individually controlling all fifteen in the three references.



(a) (b)

Fig. 3.4: (a) Example measured fresh full loop frequency distribution for 80 cell array, with corresponding reference ROSC trimming range. (b) Measured results from all three Odometers for one ROSC under test. The ROSC under test started out slower than all references in this case. Notice the low resolution with an initial count of 41.

The Odometers associated with those references all record output counts corresponding to the beat frequency for each ROSC measurement. During post-processing, the highest-resolution degradation characteristic is selected from that set of three for each oscillator that was stressed. Fig. 3.4(a) presents an example distribution of 80 fresh full loop frequencies, along with the range covered by the three reference ROSCs under all trimming conditions. Turning on each trimming capacitor slowed a reference ROSC by roughly 900 kHz, or 0.57% of the mean fresh full loop frequency under nominal operating conditions. During calibration, the reference ROSCs are trimmed to positions within the fresh array distribution such that we maximize the resolution of the group of degradation characteristics gathered from each full stress experiment.

Fig. 3.4(b) shows a group of three degradation characteristics gathered by the references from one ROSC under test. As covered in Section 2.2.3, starting measurements with the reference and test ROSC frequencies close together, but the latter slightly slower, leads to a high resolution measurement with a monotonic decrease in the output counts. Based on equations presented in the previous publications, we know that the odometer output count is equal to the number of reference ROSC periods (N_{ref}) counted during one period of the beat frequency, during which time, one less cycle is observed in the stressed ROSC ($N_{stressed}$) [11]. Therefore, according to ($1 - N_{stressed}/N_{ref} = 1 – 40/41$), reference ROSC 3 started out 2.44% faster than the ROSC under test in the current example. This led to the low initial count of 41, and hence low resolution, which is apparent from the highly quantized outputs of the corresponding Odometer. However, reference 2 was initially only 0.513% faster than the ROSC under test, so we can select this high resolution result for our analysis instead.

Finally, note that it is also possible to use measurement results from a reference ROSC that starts out *slower* than the ROSC under test, so the output count will *increase* with stress (Section 2.2.3). Therefore, we can more easily cover the distribution of fresh frequencies in our array with only three references trimmed appropriately.

3.3.4 Test Interface and Procedure

Calibration and measurements are automated through a simple digital interface. During calibration we record the fresh "*ctrl*" and full loop frequencies from each ROSC in the array by directly reading those values with an oscilloscope. The error in this step is minimized by taking thirty samples for each reading and averaging. Next, we sweep through the trimming range in the three reference ROSCs in parallel, again averaging the results from thirty samples. After that point, the optimal trimming configurations are selected in order to cover the distribution of frequencies of the ROSCs to be tested.

A RESET signal is asserted before stress conditions are set, which prevents stress voltages from being applied to the ROSCs until after the first fresh measurement. During experiments, ROSC cells are cycled through automatically without the need to send or decode cell addresses, in order to simplify the logic and attain faster measurement times. The first cell is selected with an initialization sequence, and a single external clock signal is asserted each time that the controlling software is ready for a new measurement. The row selection signal is incremented with each measurement, and the column selection shifts after all of the cells in a row have been selected. Any cells that have not been selected for stress are kept in a 0 V no-stress state and skipped over during this process. A RECOVER signal is set high before any cell is selected for measurement which we wish to put or keep in a no-stress state.

The logic used to store the row selection signal had to minimize the time when a ROSC is taken out of stress in order to prevent unwanted BTI recovery. Although the odometer provides measurements of down to the sub-μs range, we still have to account for the time it takes to scan the results out. Therefore, two DFFs were used for each row, as shown in Fig. 3.5. The SELECT_BIT_IN signal is clocked into the first DFF$_{hold}$ with ROW_CLK during circuit initialization. The next ROW_CLK pulse starts a measurement on ROW<0> by moving the select bit to the DFF$_{sel}$. That selection bit is then sent to the DFF$_{hold}$ in ROW<1> by an automatic pulse of ROW_CLK from the FSM as soon as the actual measurement is complete, and is held there while the results are scanned out. The next ROW_CLK pulse from the controlling software to start a measurement on ROW<1>, and this process is repeated up through the rows as necessary. Note that the clock signals from the software and the FSM are input to an OR gate to create ROW_CLK.

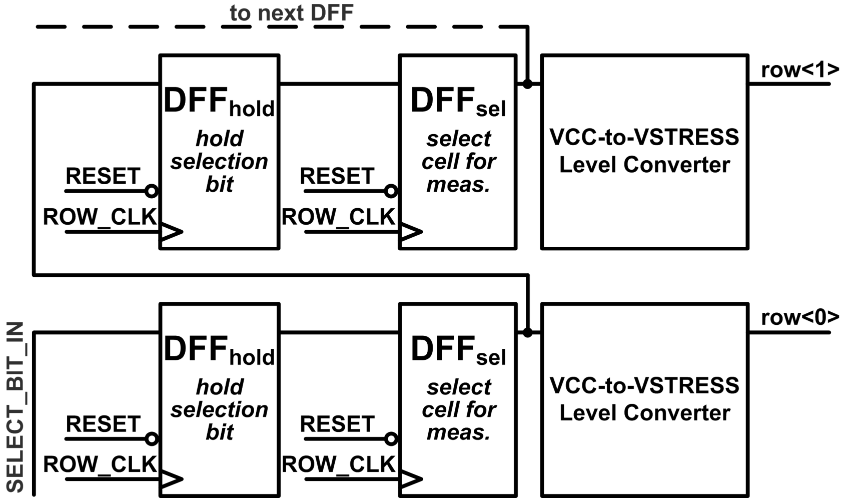
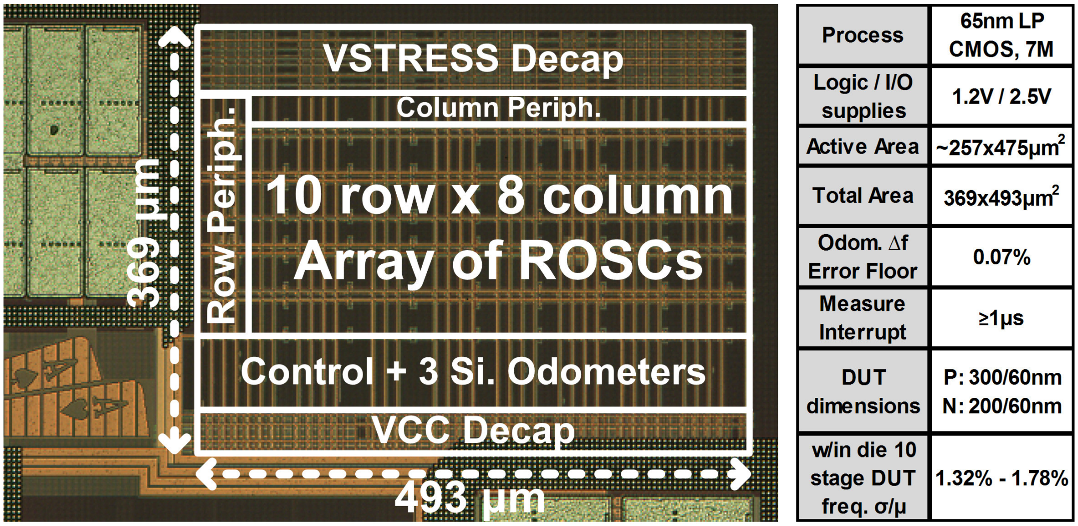


Fig. 3.5: Row selection logic (i.e., "Row Periph" from Fig. 3.1). The SELECT_BIT_IN signal is clocked into the first DFF$_{hold}$ with ROW_CLK during circuit initialization. The next ROW_CLK pulse starts a measurement on ROW<0> by moving the select bit to the DFF$_{sel}$. That selection bit is then sent to the DFF$_{hold}$ in ROW<1> as soon as the measurement is complete to reduce unwanted BTI recovery. The next ROW_CLK pulse starts a measurement on ROW<1>, and this process is then repeated through the subsequent rows.

3.4. Statistical Odometer Test Chip Measurements

A 369x493 µm^2 test circuit was implemented in 65 nm bulk CMOS for concept verification. Fig. 3.6 presents a die photo and the test chip characteristics. Measurements were automated with LabVIEW^{TM} software through a National Instruments data acquisition board. The DUTs were 200 nm/60 nm NMOS and 300 nm/60 nm PMOS transistors in the inverter stages of the stressed ROSCs. Smaller widths were not selected because the "dog bone" shape of narrower transistors introduces additional variation.



Process	65nm LP CMOS, 7M
Logic / I/O supplies	1.2V / 2.5V
Active Area	~257x475µm^2
Total Area	369x493µm^2
Odom. Δf Error Floor	0.07%
Measure Interrupt	≥1µs
DUT dimensions	P: 300/60nm N: 200/60nm
w/in die 10 stage DUT freq. σ/µ	1.32% - 1.78%

Fig. 3.6: Die photo and summary of test chip characteristics.

Automated measurement times were set to 2.5 µs unless noted otherwise, which allowed all beat frequency counts to complete correctly. In this multiple-Odometer design, where some ROSCs under test will result in high counts when a particular reference ROSC frequency is very close to its own, longer measurement times cannot always be avoided with trimming. Shorter measurement were possible, but in that case the higher count results which did not have time to complete were discarded, and the next-highest resolution output was selected for post-processing.

Therefore, we generally chose to allow the counts to complete so that we could utilize the highest resolution results. This was also done because analysis of the effects of different measurement times showed that the difference between the frequency degradation measured with 1 μs and 2.5 μs interrupts was negligible in this system, as will be described in Section 3.4.2. When an Odometer had time to complete multiple counts during the allotted measurement time, the last full count that got latched was utilized in our data analysis.

3.4.1 Measurement Error Characterization

Fig. 3.7 illustrates frequency measurement results from 0 V, no-stress experiments, so ideally there should be no shift (i.e., the normalized frequency should remain at 1.0). The characteristics of ten ROSCs are displayed, and are representative of results seen from the entire arrays measured on multiple chips. Fig. 3.7(a) was directly recorded by a 100 MHz, 1.25 GS/s oscilloscope, after the frequencies were divided down by 1024 on-chip. We see a worst case error of 0.18%, and a drift in the measured values due to some slight change in operating conditions. Fig. 3.7(b) shows a smaller worst case error of 0.07% in the frequency calculated by the Silicon Odometer, along with the fact that this differential system eliminates the effects of common mode variations.

Similar error floors were found for these systems during repeated tests, setting the lower bound on the range of frequency shifts that they can accurately measure. Finally, note that the automated oscilloscope readings required over 500 ms, while the Odometer measurements take ≥1 μs. This combination of high resolution and fast measurements is critical when measuring BTI stress, where we must avoid recovery when stress conditions are temporarily removed for readings.

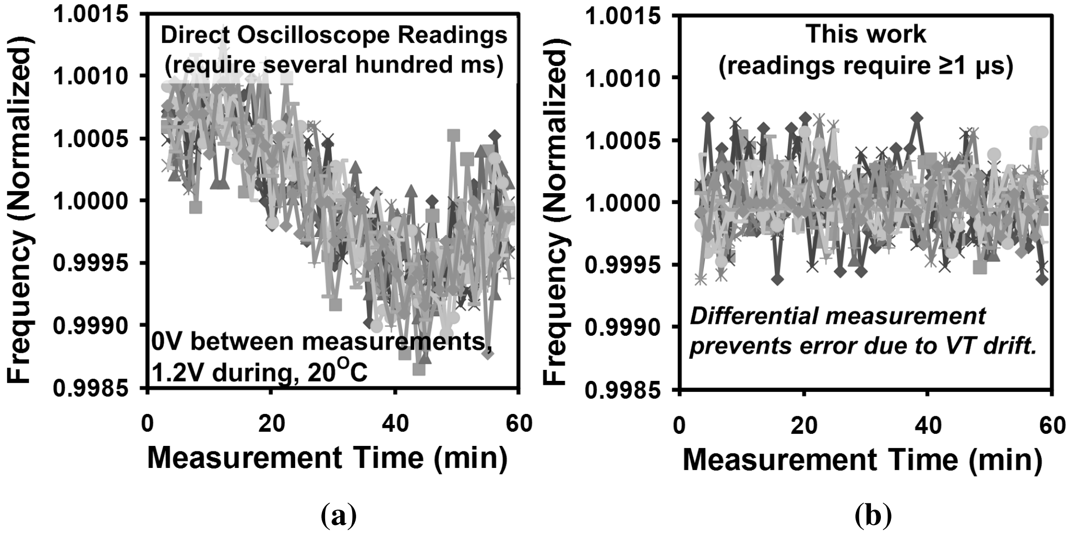


Fig. 3.7: Error (i.e., deviation from 1.0) in (a) oscilloscope and (b) faster Odometer measurements during no-stress experiments.

3.4.2 Impact of Measurement Time on BTI Results

Fig 3.8 quantifies the impact of measurement times on BTI measurements. Long interruptions take up a significant portion of the total experiment time at early measurement points. This means a large percentage of the time is spent in recovery state, which pulls down the early results and leads to a steeper degradation slope, as seen in Fig. 3.8(a). Several previous publications have clearly demonstrated this phenomenon [32], [33], [35], [36].

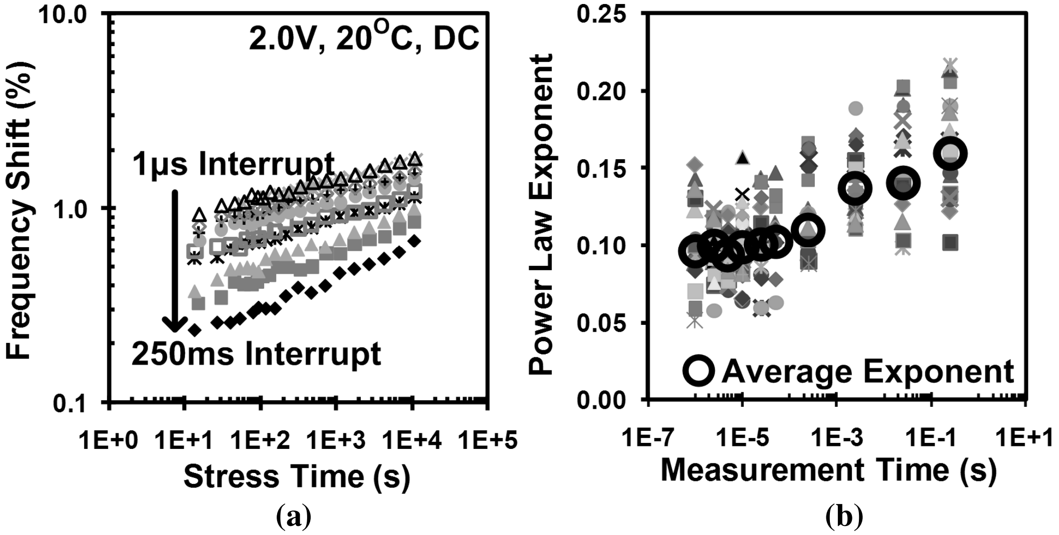


Fig. 3.8: (a) Steeper slopes for longer stress interruptions due to recovery. (b) Power law exponent vs. measurement time.

Fig. 3.8(b) shows the power law exponents that were fitted to the data from different measurement times. As times get into the millisecond range, the average exponent approaches 0.165, which has commonly been cited as the measured and theoretically correct value. It fits the RD model theory if H_2 is assumed to be the hydrogen species diffusing away from the silicon/dielectric interface [36]. However, our results, along with those from several other fast measurement techniques, show that this model and theory seem to be incomplete or incorrect [10], [32], [36].

The discrepancy with that traditional RD theory has been explained by demonstrating that NBTI appears to be the result of two component processes: (1) a slower creation of interface traps (i.e., donor-like states resulting from broken Si-H bonds at the Si-SiO_2 interface) and subsequent diffusion of H_2 into the gate dielectric, along with (2) a faster hole trapping/detrapping process in the dielectric bulk [2], [27], [32], [36], [37]. The former recovers on the scale of tens of microseconds or more, and the latter begins to recover within a microsecond or less. Therefore, slow measurement techniques requiring milliseconds or seconds miss much of the fast hole detrapping process, so the observed results are primarily due to the slower interface trap creation mechanism. This process has been claimed to have a higher power law exponent, which has in fact been show to lie in the 0.165 range, while fast hole trapping has a much smaller exponent (≤ 0.1) [36]. Therefore, fast measurement techniques are required to directly observe the full degradation due to NBTI. It should be noted that this point is still disputed by several researchers who claim that NBTI is primarily due to interface trap generation and annealing, so the RD model can describe this aging process with a high degree of accuracy [38].

Our measurements show that stress interruptions of tens of microseconds or less are required to observe the average power law exponent of ~0.1 under the listed DC stress conditions (Fig. 3.8). Ji *et al.* also found that measurement times of ≤ 40 µs did not show any measureable recovery [35], although other authors have claimed that this process can be recorded down to ≤100 ns [32], [36]. The significant spread in the range of exponents observed at each measurement time (roughly +/- 0.05) illustrates the variation in the aging process, and the importance of characterizing a statistically significant sample set.

3.4.3 DC Stress Results

PDFs of fresh DUT frequencies are shown in Fig. 3.9 with the resulting distributions after 3.1 hours of DC stress (11,200 s). The primary degradation mechanism at work in these experiments was NBTI, since PBTI is not significant when high-k dielectrics are not used, and there was no switching during stress. In Fig. 3.10(a) we see that there was no significant correlation between the fresh ROSC frequency and the stress-induced shift. This lines up with previous findings that the stress-induced V_{th} mismatch in PMOS pairs was uncorrelated to the initial mismatch [29], and that the initial spread in the V_{th} is not correlated to that caused by aging [26]. Fig. 3.10(b) shows the average (μ) frequency shifts and the standard deviation (σ) of the shifts vs. stress time. The σ increases with stress [25], [26], [30], roughly following a power law with an exponent (n) of just under 1/2 that of the μ shift. Therefore, the σ/μ ratio of the shift decreases with stress time [27].

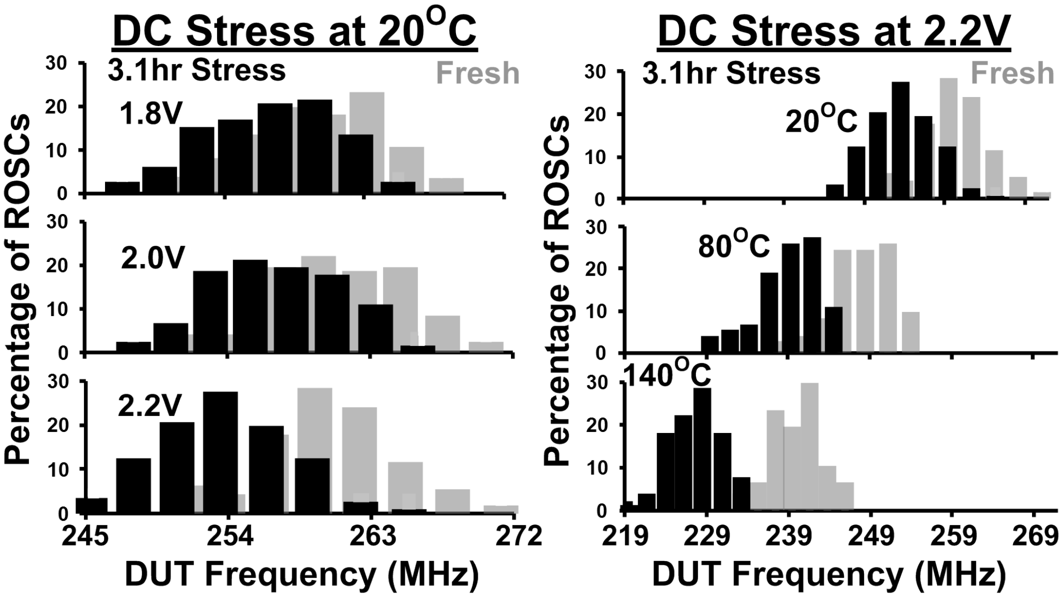
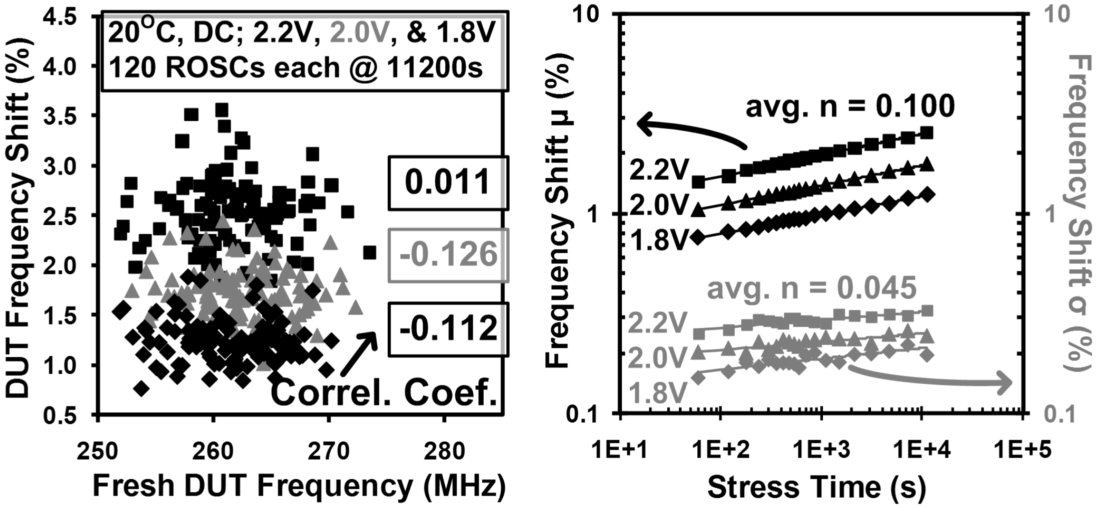


Fig. 3.9: Shift in frequency distributions after 3.1 hour stress.

The σ of the calculated frequency, on the other hand, did not show a clear trend with stress time. This value was seemingly random, and was very poorly fitted by the power law (R-squared values of only ~0.03-0.30), with the exponent of this weak fit ranging from -0.002 to 0.028, meaning the σ value remained generally flat during stress. That behavior is expected because the spread in the fresh frequency is larger than that of the spread in the aging-induced shifts, and the increase in that latter value is modest during stress, as seen in Fig. 3.10(b). This trend stands in contrast to results from previous work that claimed a decrease in the σ of a path delay with stress time showing a power law exponent of 1/6 [34].



(a) (b)

Fig. 3.10: (a) No significant correlation of the frequency shift with fresh frequency. (b) Mean and standard deviation of Δf.

3.4.4 Temperature Dependence of DC Stress-induced Degradation

Fig. 3.11 displays the degradation characteristics of the μ and σ of the frequency shifts at high temperatures. The power law exponents of these values increase at higher temperatures, and that of σ is just over 1/2 that of the μ characteristic. Varghese *et al.* stated that a linear dependence of n on temperature points to dispersive temperature dependence rather than Arrhenius activation, and that this phenomenon is simply an artifact of long measurement times [39]. They showed results indicating that the temperature dependence of n disappears with on-the-fly measurements. Other recent work by Liu *et al*. showed n increasing with temperatures up to roughly 110°C, where the value saturated at 0.18, even when using fast and on-the-fly methods [40]. Like many issues surrounding BTI degradation, this matter is unsettled, but our results from a large sample set support the dispersive transport model.

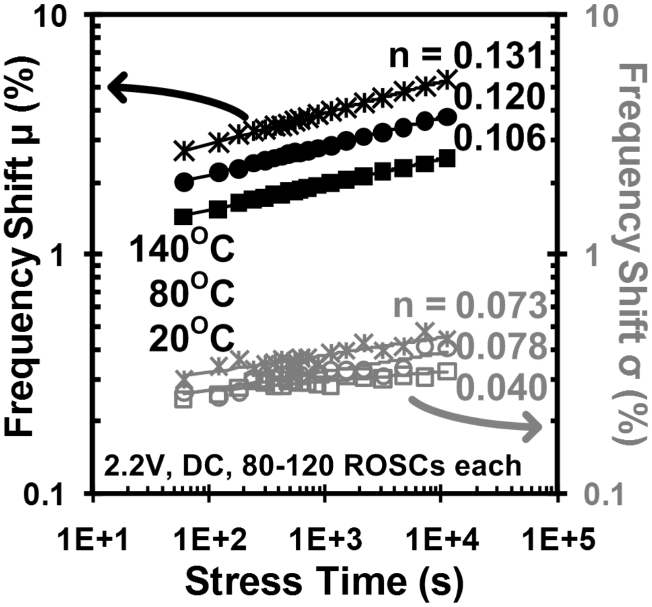


Fig. 3.11: Mean and standard deviation of the measured frequency shift at increasing temperatures.

3.4.5 AC Stress and Stress/Recovery Characteristics

Fig 3.12(a) shows a drop in total degradation of roughly half at low frequencies, compared with DC stress, due to the recovery that takes place during each half cycle for all NMOS. As the frequency is raised, HCI plays a larger role in the aging due to the increased switching activity. This leads to a larger n, which is a signature of HCI [1]. At high voltages, we see that HCI eventually dominates the overall aging of the DUTs when the AC stress lines cross the DC characteristic. However, we have shown in previous work that this crossover point is highly dependent on voltage, and NBTI is dominant at lower stress voltages, closer to those found in real operation (Section 2.3.2).

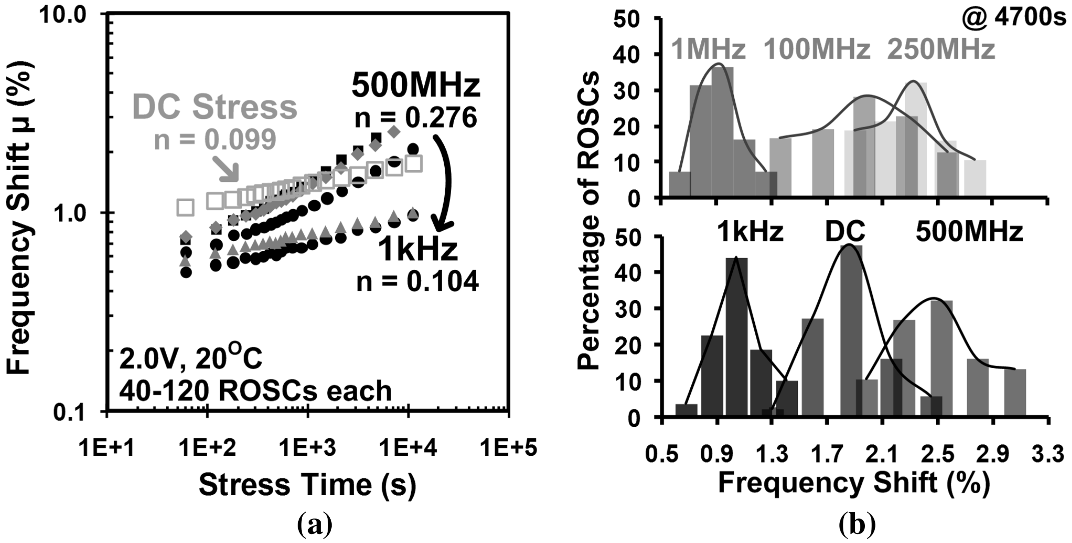


Fig. 3.12: (a) Mean AC stress results compared with DC. (b) PDFs of the frequency shift at the 4700s point in part (a).

Note that this analysis is not equivalent to those found in pure AC BTI experiments, where no current flows through the channel, even during switching. Fernandez *et al.* found pure NBTI to be frequency independent up to 2 GHz [41]. Li *et al.* stated that only the slow component of NBTI is frequency independent, while the degradation due to the fast component increases with frequency up to ~10 kHz, after which it becomes frequency independent [36].

PDFs of these shifts at the 4700 s point are presented in Fig. 3.12(b). As in the DC case, no significant correlation was found between the fresh DUT frequency and the total frequency shift. However, we would expect a stronger positive correlation if the ROSCs were operated in a closed-loop free-running mode during stress, since more switching activity leads to more HCI degradation. The σ of the frequency shift again increases with stress, at a rate that increases with frequency. Although this curve only roughly follows a power law (i.e., R-squared value of 0.85 at 500 MHz), the exponent values are around one half of that of the μ shift. For example, the n for the 500 MHz μ shift is 0.2763, while the n for the σ of the shift is 0.1618. Finally, the σ of the calculated frequency remained nearly flat on average, as explained for DC stress in Section 3.4.3.

Stress/Recovery curves taken from four ROSCs simultaneously are presented in Fig. 3.13. The bottom point of the recovery phases increases with each period, as more damage accumulates. The fast and significant recovery we observe after stress conditions are removed has been detected with other fast measurement setups, and some authors claim it cannot be correctly described by the RD model [9], [10]. They state that one must use a fast hole trapping/detrapping model to find a theoretical explanation for these dynamics. However, other researchers still state that the RD model can be used with little error [38]. More work will be required to resolve this issue.

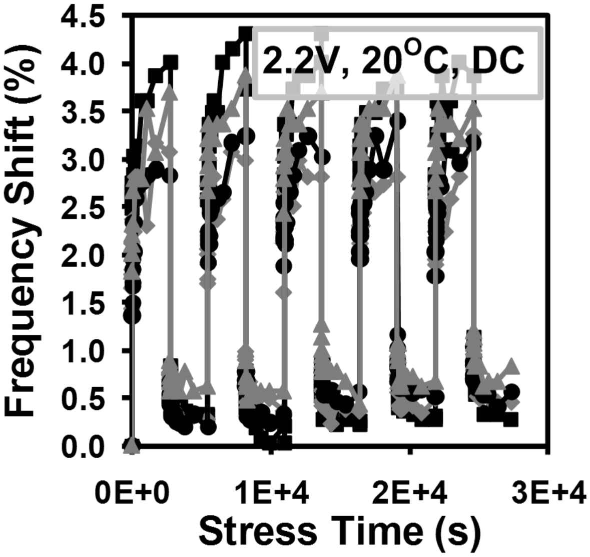


Fig. 3.13: Stress/Recovery curves taken from four ROSCs simultaneously. The bottom point of the recovery phases increases with each period, as more permanent damage accumulates.

3.5 Conclusions

We have implemented the first measurement system that facilitates efficient statistical aging measurements involving BTI and HCI in ring oscillators (ROSC). Measurement results from a test chip built in 65 nm technology show that the differential Silicon Odometer beat frequency detection system can measure frequency shifts with an error of ≤ 0.07%, and stress interruptions of ≥1 μs. We illustrate the positive correlation between measurement times and the n of the degradation curve under DC BTI stress, demonstrating the need for fast measurement techniques. The n of the frequency shift after BTI stress is found to increase with temperature, pointing to dispersive temperature dependence, rather than Arrhenius activation. Low frequency stress leads to reduced degradation when compared with DC due to BTI recovery, but the degradation rate increases with frequency and the addition of HCI degradation, which eventually dominates the total aging. Statistical results show that fresh frequency and the AC or DC stress-induced frequency shift are uncorrelated, both the mean and standard deviation of that shift increase with stress, and the ratio of this standard deviation/mean decreases with stress time. These findings point to the utility of our proposed system for process characterization, and important trends in the aging of vanishingly small modern transistors.

Chapter 4

A DLL-Based On-Chip NBTI Sensor for Measuring PMOS Threshold Voltage Degradation

4.1 Introduction to The DLL-Based NBTI Sensor

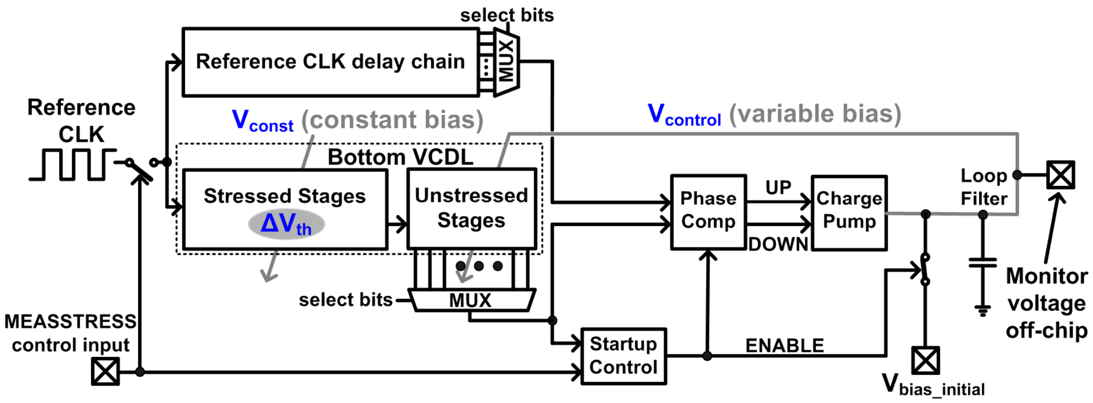
Negative Bias Temperature Instability (NBTI) is one of the most critical device reliability issues in sub-130 nm CMOS processes. In order to better understand the characteristics of this mechanism, accurate and efficient means of measuring its effects must be explored (see Chapter 1 for further details on NBTI and previously published measurement methods). In this work, we describe an on-chip NBTI degradation sensor using a delay-locked loop (DLL), in which the increase in PMOS threshold voltage (V_{th}) due to NBTI stress is translated into a control voltage shift in the DLL for high sensing gain. The proposed sensor is capable of supporting both DC and AC stress modes, and avoids the use of high speed off-chip signals. Results from a test chip fabricated in a 130 nm bulk CMOS process show an average gain of 10X in the operating range of interest, with measurement times in tens of microseconds possible for minimal unwanted threshold voltage recovery. NBTI degradation readings across a range of operating conditions are presented to demonstrate the flexibility of this system.

4.2 DLL-Based NBTI Sensor Circuit Details

4.2.1 System-Level Overview

A block diagram of the measurement system is displayed in Fig. 4.1(a). This NBTI sensor is primarily composed of an analog DLL and an on-chip reference clock (CLK) generator. The DLL contains a voltage controlled delay line (VCDL), and the circuitry needed to adjust the control voltage ($V_{control}$) which locks the VCDL output into phase with the delayed reference CLK. This control unit includes a phase comparator, a charge pump, and a loop filter. A startup control circuit is also included to prevent false locking and improve lock times by resetting the phase comparator when the DLL is shut down, and pinning $V_{control}$ to a bias in middle of its locking range ($V_{bias\_initial}$) until the DLL is switched on with the MEASSTRESS signal.

The VCDL consists of a chain of delay stages that are placed under NBTI stress, in series with a number of unstressed stages. The latter number can be adjusted with a MUX during calibration to move $V_{control}$ into the optimal gain region. Note that an adjustable delay line was also added in the reference CLK path, so this DLL locks the VCDL output directly into phase with the reference input to the phase comparator (i.e., there is not a 360^{O} phase difference between the reference and the VCDL output as is the case in other common designs [42], [43]).



(a)

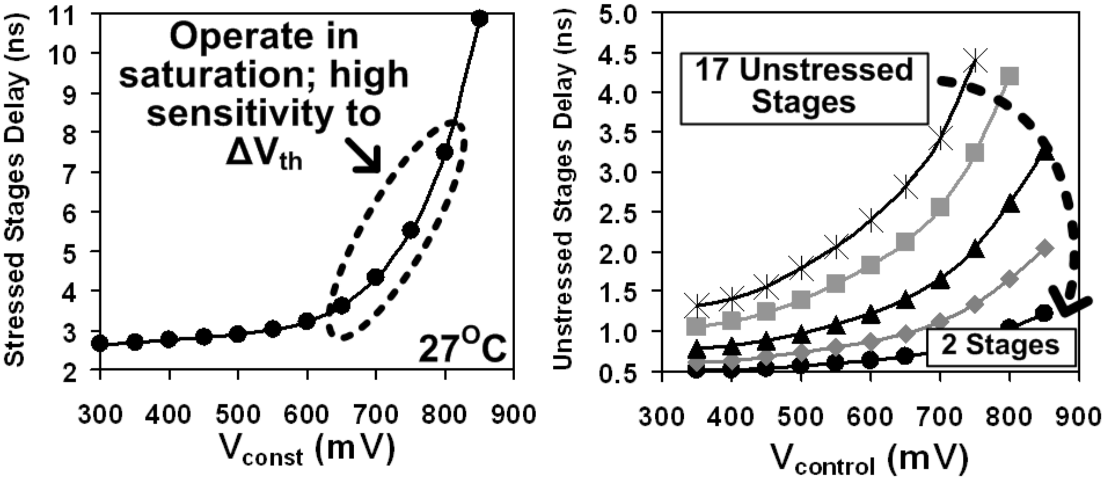
I_{CH}	Charge Pump Current	81.5931 μA
K_{VCDL}	VCDL Gain	9.7936 ns/V
F_{REF}	Ref. CLK Frequency	125 MHz
C	Loop Filter Capacitance	100.006 pF
ω_N	Loop Bandwidth (rad/s)	0.9988 M
ω_N/ω_{REF}	ratio guideline from [45]	< 0.1
ω_N/ω_{REF}	calculated ratio	0.00127

$$\omega_N = I_{CH} * K_{VCDL} * F_{REF} * (1/C)$$
$$\omega_{REF} = 2*\pi*F_{REF}$$

(b)

Fig. 4.1: (a) Block diagram of the proposed NBTI degradation sensor. (d) Loop bandwidth calculation and comparison to the reference frequency.

During stress periods, the DLL is deactivated while stress conditions are applied to each DUT in the stressed stages. When a measurement is started, stress conditions are removed and the DLL is activated so that $V_{control}$ can settle and be recorded off chip. The stressed stages are biased by the constant V_{const} during measurements, so this portion of the VCDL will slow down after NBTI stressing due to the V_{th} degradation, whose impact on delay is directly proportional to that of an increasing V_{const} bias (Section 4.2.3, Fig. 4.2(a)). Note that V_{const} was supplied from off-chip, but can also be driven by an unstressed replica DLL [44], or another on-chip bias generator. The unstressed stages are biased by $V_{control}$, which decreases to speed these buffers up (Fig. 4.2(b)) during measurements to compensate for the slower stressed stages. This system was designed to achieve a maximum gain of over 15X from the increase in $|V_{th}|$ to the corresponding decrease in $V_{control}$, which is characterized in a simple calibration step, as covered in Section 4.2.3.



(a) (b)

Fig. 4.2: (a) Stressed stages total delay vs. V_{const} (see Fig. 4.5). (b) Unstressed stages total delay vs. $V_{control}$ for a varied number of unstressed stages.

DLLs are often preferred over phase locked loops in applications such as frequency synthesis for a variety of reasons, including their unconditional stability when the loop bandwidth (ω_N) is held a decade or more below the operating frequency [45]. The DLL used in the proposed NBTI sensor is not employed in a clocking network where a high loop bandwidth might be required, so we designed to stay well below this one decade guideline with a ratio of ~0.00127 when the reference CLK frequency was 125 MHz, as seen in Fig. 4.1(b). Due to the fact that the DLL acts as a single-pole low-pass filter with a cutoff frequency of ω_N to shifts in the reference CLK, setting a low value for this parameter also enhances the system's jitter performance [43], while still allowing locking at a rate proportional to ω_{REF}/ω_N cycles [45]. This results in a roughly 7 to 20 μs locking time in our design.

4.2.2 Selected System Components

Selected system components are pictured in Fig. 4.3. The unstressed stages are capacitor-loaded delay buffers (Fig. 4.3(a)). As $V_{control}$ drops, the output load of each stage is lowered due to the smaller V_{gs} value on the Mncap transistor, thereby decreasing the line delay. The effectiveness of lowering this value as a means of decreasing the delay rapidly decreases as it approaches the threshold voltage of Mncap (Fig. 4.2(b)). Next, the phase comparator (Fig. 4.3(b)) asserts equal short duration output pulses for in-phase signals to avoid a dead-band region [42]. An ENABLE signal was added to this design to reset the comparator during stress periods.

In order to take advantage of the short-duration pulses created by the phase comparator for in-phase signals when the DLL is active, the charge pump (Fig. 4.3(c)) output ("OUT") does not change when both input signals from the phase comparator are asserted for equal periods [42]. Note that in order to avoid large drain-source voltages in the devices adjacent to OUT, which could lead to additional unwanted shifts in $V_{control}$, and therefore a phase offset in the VCDL output, it is best to operate with this value centered at ~VCC/2. In the 1.2 V technology used for this implementation, we designed to stay in the 450 mV – 850 mV range.

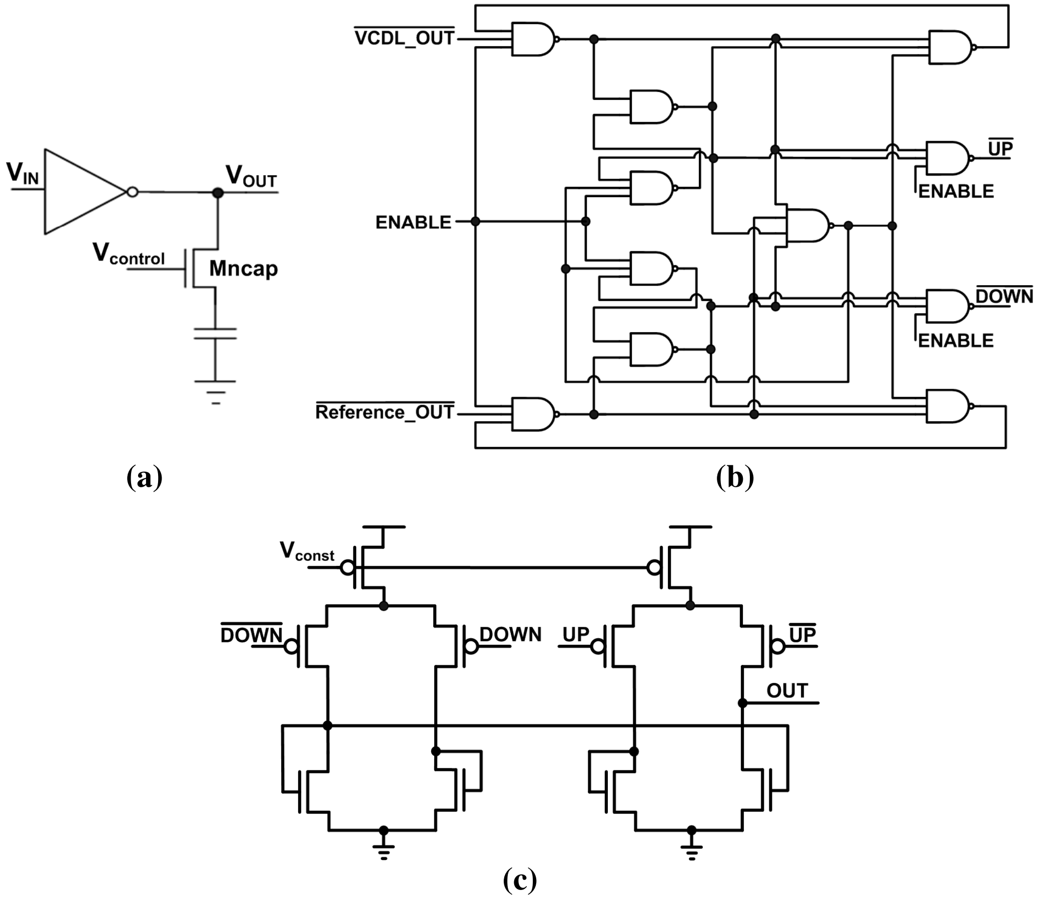
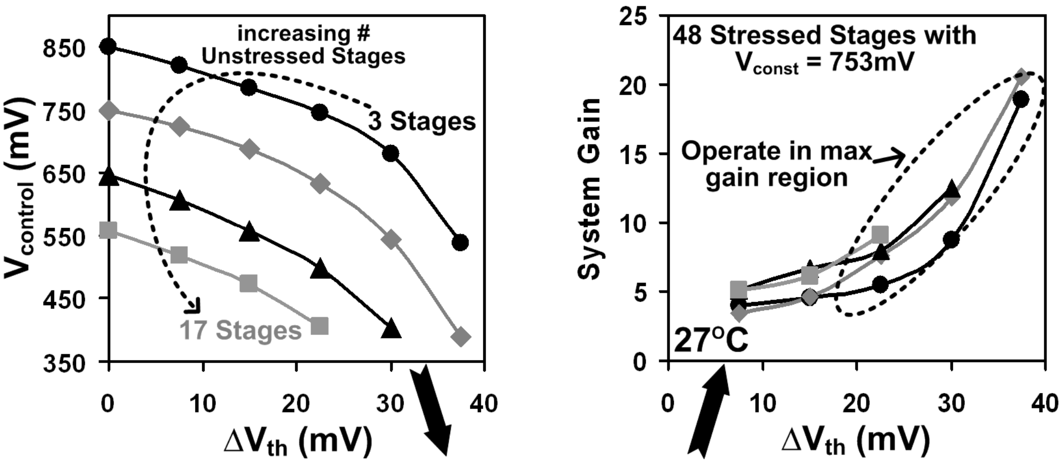


Fig. 4.3: (a) The unstressed stages are adjustable capacitor-loaded buffers. (b) The phase comparator [42] with added ENABLE signal. (c) The charge pump [42].

4.2.3 System Gain and Calibration

Our design was tuned to attain maximal gain from ΔV_{th} in the stressed stages to the decrease in $V_{control}$, since the latter value will be measured off-chip and translated into the threshold shift. The simulation results in Fig. 4.4(a) illustrate the translation of ΔV_{th} into $\Delta V_{control}$ for a varied number of unstressed stages at an equivalent nominal delay point. The corresponding system gain plot in Fig. 4.4(b) is created with the simple equation in Fig. 4.4(c). Note that during measurements, the gain plot of interest will be gain vs.

$V_{control}$, when we have selected one particular number of unstressed stages to use (see Fig. 4.9(b)).



(a) (b)

$$\boxed{\text{System Gain }(V_{th\_2}) = \frac{V_{control\_1} - V_{control\_2}}{V_{th\_2} - V_{th\_1}}}$$

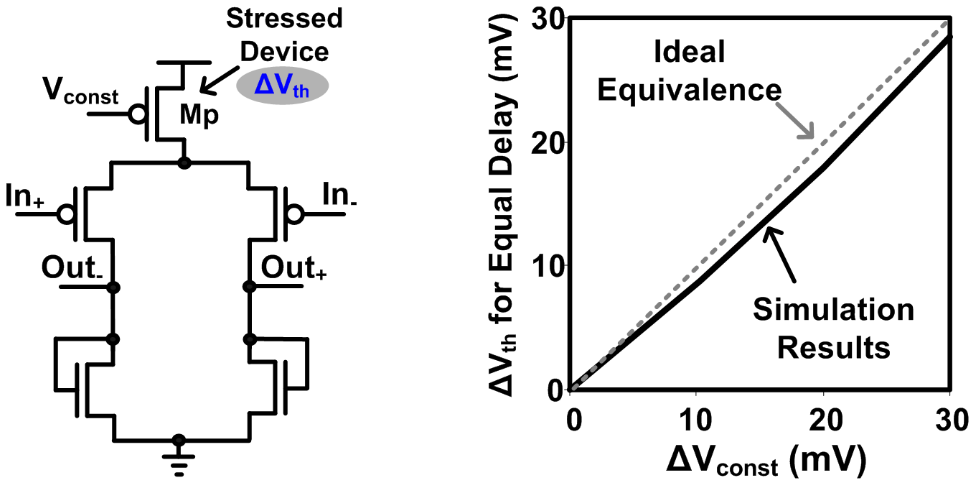
(c)

Fig. 4.4: (a) Simulated translation of ΔV_{th} to $\Delta V_{control}$ for equivalent VCDL delay. (b) The ΔV_{th} to $\Delta V_{control}$ gain plots corresponding to part (a). (c) Simple equation used to calculate the system gain at each point. Note that during measurements when we have selected a particular number of unstressed stages to use, the gain plot of interest will be system gain vs. $V_{control}$ (see Fig. 4.9(b)).

As illustrated in Fig. 4.5(b), ΔV_{th} (in this case, a shift in the nominal threshold voltage value, VTH0, in the BSIM parameter file) is directly proportional to ΔV_{const} in the buffer structure used for the stressed stages (Fig. 4.5(a)). That is, V_{th} must be changed by the same amount as V_{const}, with the other held constant, in order to cause an equivalent stage delay shift. This relationship can be derived from the standard saturation current equation, where we see that the two values of interest have the same effect on the drain current, and hence, the buffer delay as well since delay is proportional to C_{load}*VCC/I_D, where:

$$I_D = \frac{1}{2}\mu_p C_{OX}\left(\frac{W}{L}\right)\left(\left[VCC - V_{const}\right] - \left|V_{th}\right|\right)^{\alpha}.$$

Therefore, the system gain from ΔV_{th} to $\Delta V_{control}$ can be checked during calibration by sweeping V_{const}, while recording the corresponding change in $V_{control}$, as described next. Note that this delay stage could also be used in the Silicon Odometer framework [11] in order to take advantage of that system's precision and digital nature, while isolating NBTI stress in the delay stages' PMOS header devices, rather than simply raising the supply voltage of standard inverters to get a general stress measurement.



(a) (b)

Fig. 4.5: (a) Stressed stage buffer design. (b) ΔV_{th} of Mp in the stressed buffers is directly proportional to ΔV_{const}. This relationship allows us to calibrate the sensor, as shown in Fig. 4.9.

In order to calibrate this sensor, we first estimate the maximum expected V_{th} degradation based on published data. In the 130 nm process used for this work, we estimated a maximum degradation of ~20-30 mV with 2.4 V stress [41], [46], [47], and this number was then refined based on our initial measurements. Next we determine a range of acceptable V_{const} bias points in the saturation region which will keep the stressed stages' total delay in the indicated portion of the plot in Fig. 4.2(a) throughout a worst-case degradation. The stressed buffers have a high sensitivity to $\Delta V_{control}$, and therefore ΔV_{th}, in this region. With the initial V_{const} bias point set, we sweep this parameter from that point using an off-chip source for a varying number of unstressed stages at the desired measurement temperature, and record the change in $V_{control}$ for each sweep.

As seen in Fig. 4.4(a), there is a larger range of $V_{control}$ biases which will still allow a phase lock with the delayed reference CLK when fewer unstressed stages are used. If an excessive number is selected, the minimum delay (reached roughly when $V_{control}$ approaches the NMOS threshold voltage) is not sufficiently low to compensate for the maximum projected degradation in the stressed buffers' delay. However, rather than design with a fixed short delay line, we allow the number of unstressed stages to be varied in order to account for any effects that are not captured correctly in simulations. When $V_{control}$ moves to lower values to speed up the unstressed portion of the VCDL, it has a small effect on the stage delay (Fig. 4.2(b)), so a large bias change is needed to compensate for the degradation in the stressed stages. Based on the calibration results, we select an optimal number of stages and the PMOS header bias for maximum gain across the projected V_{th} degradation range. The resultant $V_{control}$ vs. V_{const} curve defines the translation of the final measured $V_{control}$ values into the PMOS threshold degradation characteristic that is sought in NBTI measurements. Note that while the calibration for the initial chip involves an exploration of the possible operating space, subsequent test chips should only require one sweep of V_{const} to extract the required translation curve for each temperature of interest, barring significant sensor-to-sensor variation. In our experiments involving fifteen operational test chips with two sensor instances each, no adjustments were required after system parameters were selected for the first tested instance. Each DLL tested after this locked correctly across the entire range of V_{const}. In addition, note that it is not expected that the DUTs will experience any appreciable aging during the short calibration period, as all voltage drops across any pair of terminals are less than |VCC|.

4.2.4 DLL Locking Time and Measurement Delay

The DLL in our application is required to shut down and start up quickly and reliably for each measurement. The startup control circuit pictured in Fig. 4.1(a) helps to ensure that the DLL will fall into the proper phase lock each time, and improves lock times at a fixed ω_N. It accomplishes these tasks by resetting the phase comparator when the DLL is shut down, and pinning $V_{control}$ to a bias in middle of its locking range ($V_{bias\_initial}$) until the DLL is switched on. This circuit enables the DLL control unit to begin comparing its two input clock signals only after the MEASSTRESS signal goes high and one full output pulse is detected from the VCDL. Due to this timing feature, and the fact that the reference input to the phase comparator is also delayed by roughly one CLK period, we must simply make sure that the first pulse of this reference input rises before the first VCDL output pulse falls.

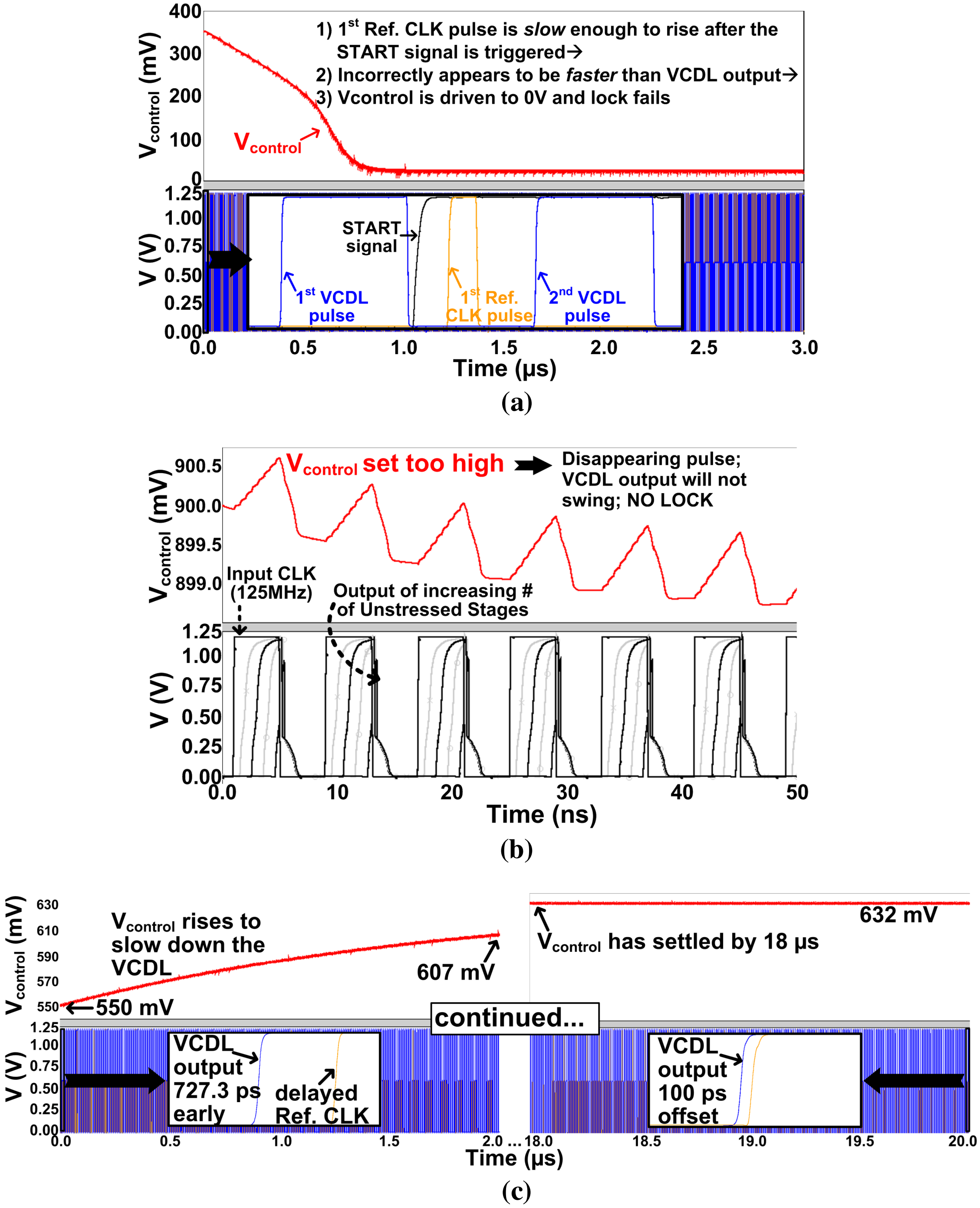


Fig. 4.6: (a) Failed phase lock due to the first delayed reference CLK pulse at the phase comparator input being excessively late. (b) Failed lock due to high initial value of $V_{control}$. (c) Correct phase lock simulation (with a time gap in the plot due to plot file sizes). Lock is achieved within 18 µs in this example, which is representative of standard operation.

Fig. 4.6(a) illustrates the consequence of failing to meet this constraint—the late delayed reference CLK pulse appears to be arriving at the phase comparator earlier than the VCDL output, so $V_{control}$ is driven low to compensate. In this example, $V_{control}$ is driven all the way to 0 V, and even a harmonic phase lock is not possible. We can prevent this by selecting a smaller number of reference CLK delay buffers, a larger number of unstressed stages, and/or a higher initial value for $V_{control}$.

Selecting a proper initial value for $V_{control}$ ($V_{bias\_initial}$ in Fig. 4.1(a)) can prevent harmonic locking, or in the worst case even a complete failure of the DLL to operate as shown in Fig. 4.6(b). In that simulation, an initial bias of 900 mV causes and excessively long rise time in the unstressed stages in comparison with the total reference CLK period. This effect leads to a shrinking pulse width in the later stages of the unstressed delay chain. If enough unstressed stages are selected, the input CLK pulse may not propagate all the way through the delay chain, and the VCDL output will remain at 0 V, preventing the DLL from operating. In contrast, the simulation results shown in Fig. 4.6(c) demonstrate the ability of the DLL to quickly lock when a good initial bias, number of unstressed buffers, and reference CLK delay chain length are chosen. Simulations showed that with the application of these proper initial conditions, locking times are less than ~20 μs, which sets the lower limit for our measurement time.

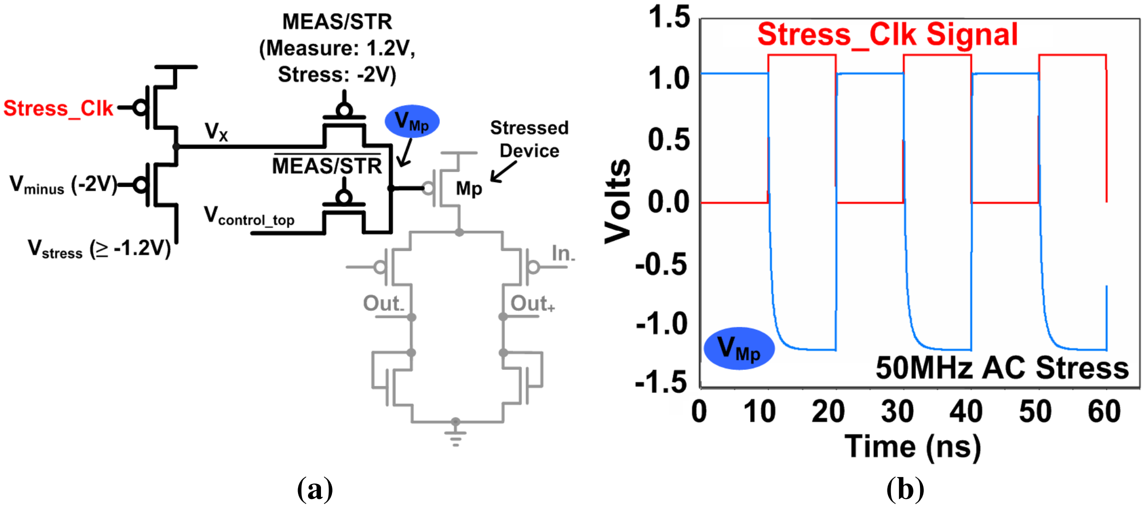
The main issues that may prevent DLL locking are concisely summarized here: (1) the total VCDL delay is too short with respect to the reference CLK, or (2) a high value of $V_{bias\_initial}$ in combination with too many unstressed stages prevents a full swing at the VCDL output. The steps taken to prevent this during system calibration are as follows:

(1) Start with the minimum available number of unstressed stages. (2) Choose an initial number of reference CLK delay stages and sweep through V_{const} values to check for DLL locking. (3) If the DLL fails to lock across the desired operating range, reduce the number of reference CLK stages and repeat the previous step. (4) If the DLL fails to lock with the minimum number of reference CLK stages, begin increasing the number of unstressed VCDL stages until a lock is achieved in the V_{const} range of interest. As this number is increased, the DLL may fail to lock at high values of $V_{bias\_initial}$, but this will be observed during the calibration sweeps and that range can be easily avoided during measurements.

4.2.5 Stress Switch Design for AC Stress Measurements

The circuitry illustrated in Fig. 4.7(a) can be used in conjunction with the stressed VCDL stages in order to facilitate AC stress measurements. The Stress_Clk signal shown in that figure can be held constant at either zero or VCC (1.2 V). The former is applied during the measurement period to bias V_X at 1.2 V, and the latter is used during DC stress measurements. Alternatively, Stress_Clk can swing between these values at frequencies up to 50 MHz during a stress period to test for the frequency dependency of NBTI degradation (Fig. 4.7(b)). This frequency limit is imposed by a degraded falling transition time in relation to the total Stress_Clk period in this 130 nm technology.

V_{minus} is set at -2 V in order to pass V_{stress} signals ranging down to -1.2 V, due to the fact that PMOS transistors conduct weak low voltages. The use of PMOS devices was necessary with negative voltages on the source and drain terminals so as not to forward bias PN junctions between those diffusion areas and the substrate. This PMOS-based setup creates a stress condition ranging down to V_{gs} = -2*VCC, and therefore would not negatively impact our results since creating this stress bias is our goal. The PMOS biased at V_{minus} is always in strong inversion, so the width of the device stacked above it is skewed up (10 µm compared to 0.5 µm) in order to drive the internal node voltage (V_X) back up to ~1.2 V. The MEASSTRESS signal switches between -2 V during stress periods and 1.2 V when the DLL is activated for measurements. Note that V_{minus}, V_{stress}, and MEASSTRESS are driven from off-chip, while Stress_Clk is controlled from off-chip, but the alternating signal for AC stress measurements is created on-chip with a voltage controlled oscillator.

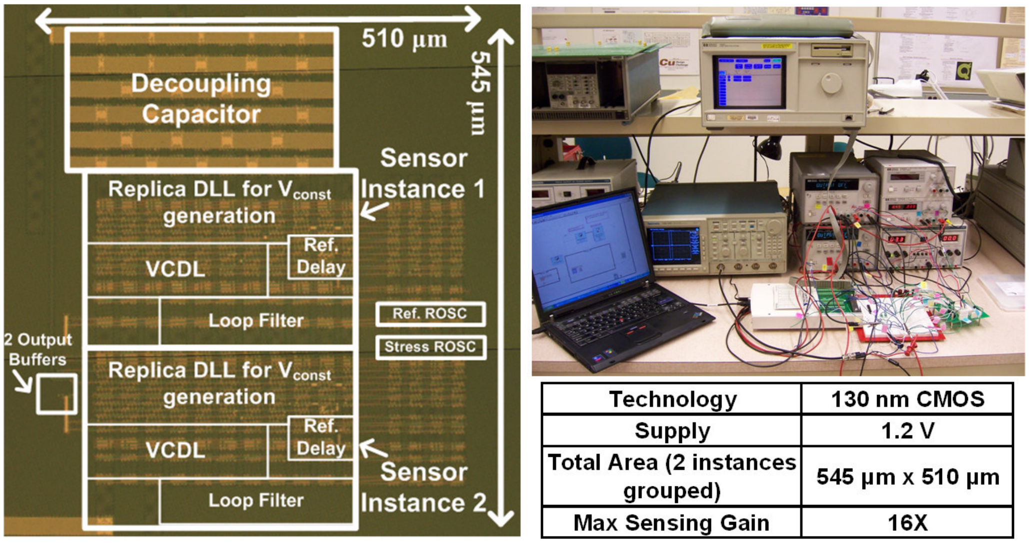


(a) (b)

Fig. 4.7: (a) Stress switch capable of driving signals at V_{Mp} ranging down to -1.2 V. This structure facilitates DC and AC stress conditions. (b) Simulated AC stress waveforms generated from an extracted netlist of this stress switch. V_{Mp} values do not swing up fully to 1.2 V due to V_{minus} remaining at -2 V even during this high duty cycle.

4.3 DLL-Based NBTI Sensor Test Chip Measurements

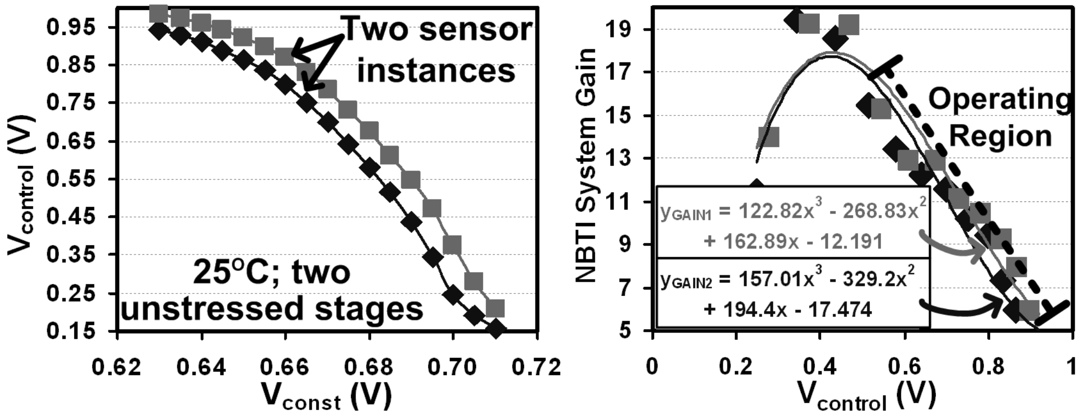
The proposed NBTI sensor was fabricated in a 1.2 V, 130 nm CMOS process. The dimensions of the PMOS DUTs are W/L = 6 μm/260 nm. Automated MEASSTRESS signal pulses and other control signals are generated with LabVIEW^{TM} software, which allows us to take fast measurements of $V_{control}$. A chip microphotograph, a picture of the measurement setup, and the test chip characteristics are shown in Fig. 4.8.



Technology	130 nm CMOS
Supply	1.2 V
Total Area (2 instances grouped)	545 µm x 510 µm
Max Sensing Gain	16X

Fig. 4.8: Chip microphotograph, measurement lab setup including the LabVIEW^{TM} software interface, and summary of the test chip characteristics.

As covered in Section 4.2, our sensor is calibrated using a set of variables prior to applying stress, including the adjustable number of stages in the reference CLK delay structure and in the unstressed stages of the VCDL. After finding the optimal point for maximum gain within our preferred operating space, we extract the $V_{control}$ vs. V_{const} (and therefore V_{th}) curve as illustrated in Fig. 4.9(a). Next, we calculate the gain in each increment and create a gain vs. $V_{control}$ plot (points in Fig. 4.9(b)). Finally, we fit a third-order polynomial to this gain plot (solid lines in Fig. 4.9(b)), which will later be used to translate the measured $\Delta V_{control}$ during stress experiments to a ΔV_{th} characteristic. This is accomplished by plugging each sequentially measured pair of $V_{control}$ values into equations (1) and (2) in Fig. 4.9.



(a) (b)

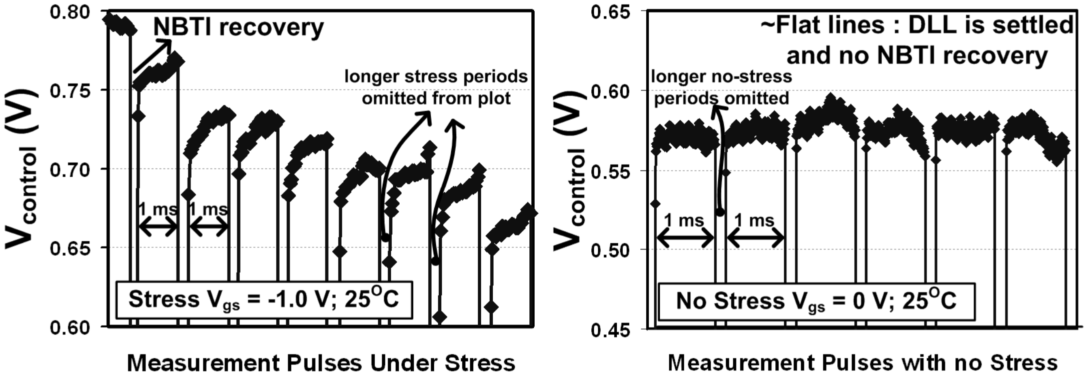
$$\text{NBTI System Gain }(V_{control\_2}) = \frac{V_{control\_1} - V_{control\_2}}{V_{const\_2} - V_{const\_1}} \Rightarrow \boxed{\begin{array}{c} y_{GAIN}(V_{control\_X}) = \\ \text{CurveFit}(V_{control\_1} : V_{control\_N}) \end{array}} \quad (1)$$

$$V_{DIFF} = V_{control\_1} - V_{control\_2} \Rightarrow V_{AVG} = V_{control\_2} + \frac{V_{DIFF}}{2} \Rightarrow \boxed{\Delta V_{th} = \frac{V_{DIFF}}{y_{GAIN}(V_{AVG})}} \quad (2)$$

Fig. 4.9: (a) Measured calibration curves. (b) A polynomial is fit to the corresponding gain plots (derived from equation (1) in this figure), and subsequently used to translate $\Delta V_{control}$ readings during stress experiments into ΔV_{th} for each sensor.

During constant stress experiments, three measurements were made per decade of time (on a seconds scale) since NBTI is known to follow a power law behavior. Even during the fast 1 ms measurement pulses with a sampling rate of 35 kHz, which is sufficient to track changes in tens of microseconds, recovery can be clearly observed as $V_{control}$ quickly rises from its initial settling value after the DLL is activated (Fig. 4.10(a)).

In order to confirm that this rising value is due to NBTI recovery rather than a long control voltage settling time, $V_{control}$ was also measured on a fresh sensor that had not undergone any stress. As illustrated in Fig. 4.10(b), the control voltage settles almost immediately in that case, even when V_{const} is adjusted such that $V_{control}$ is near the lower end of its operating range.

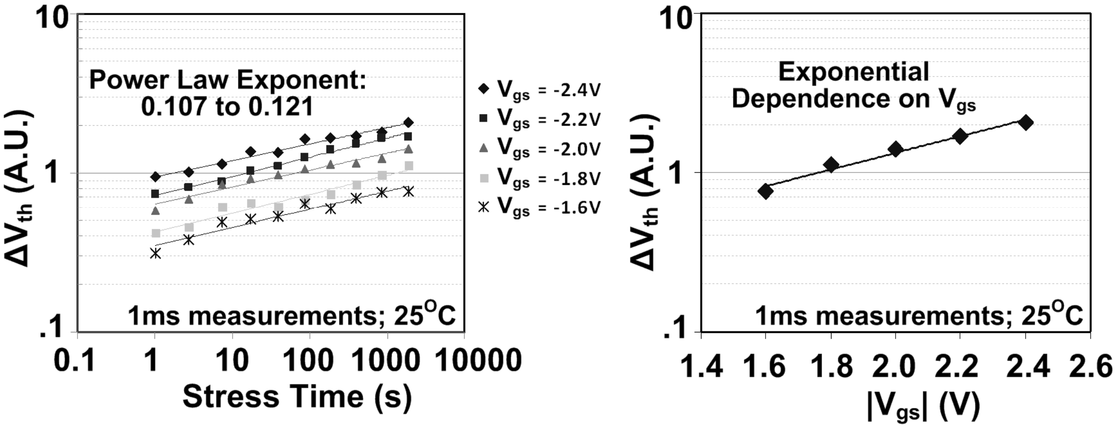


(a) (b)

Fig. 4.10: (a) Measurements taken during V_{gs} = -1.0 V stress. A µs-order NBTI recovery is apparent as $V_{control}$ rises quickly to slow down the stressed stages while their threshold voltage recovers. (b) Fresh DLL readings (with V_{gs} = 0 V between measurements) remain relatively flat.

Therefore we infer that the fast rise in $V_{control}$ seen throughout the 1 ms measurement periods is due to V_{th} recovery in the stressed DUTs. That measurement time was chosen based on our experience with the test equipment, and maintained for consistency, although the results in Fig. 4.10 show that our design is capable of providing readings in tens of microseconds. In these particular results, we see that the maximum required measurement time is roughly 57 µs or less, since a steady $V_{control}$ value is available by the second measured point with a 35 kHz sampling rate. Note that several papers published after this design was fabricated have reported faster measurement times [10], [11], [48], showing that any future implementation of the DLL system presented here should have faster locking times (< 1 μs). This is possible with any number of improvements found in DLL design literature.

Constant stress experiment results are displayed in Fig. 4.11. Stress voltages were varied from -1.6 V to -2.4 V over a period of 1850 seconds. A power law exponent ranging from 0.107 to 0.121 is observed in Fig. 4.11(a), matching well with recently published findings [10], [11]. These exponents are smaller than the 0.25 value attributed to a H diffusion limited process, or 0.16 when H_2 is the diffusing reactant, and have been attributed to a faster charge trapping/detrapping process that cannot be correctly observed with slower measurement methods [10].



(a)

(b)

Fig. 4.11: (a) Constant stress experiment results. (b) The threshold degradation after 1850 seconds of stress plotted versus the stress voltage. NBTI degradation is exponentially dependent on this value.

Fig. 4.11(b) shows the exponential dependency of the threshold degradation on the stress voltage. Measurements were also obtained at high temperatures, showing accelerated degradation due to this thermally activated process (Fig. 4.12(a)). Fig. 4.12(b) illustrates the effects of taking measurements at the end of a longer 2 second measurement pulse, and hence, after excessive unwanted threshold voltage recovery. As observed in past publications [49], the power law exponents of results from extended measurement interruption periods are markedly higher.

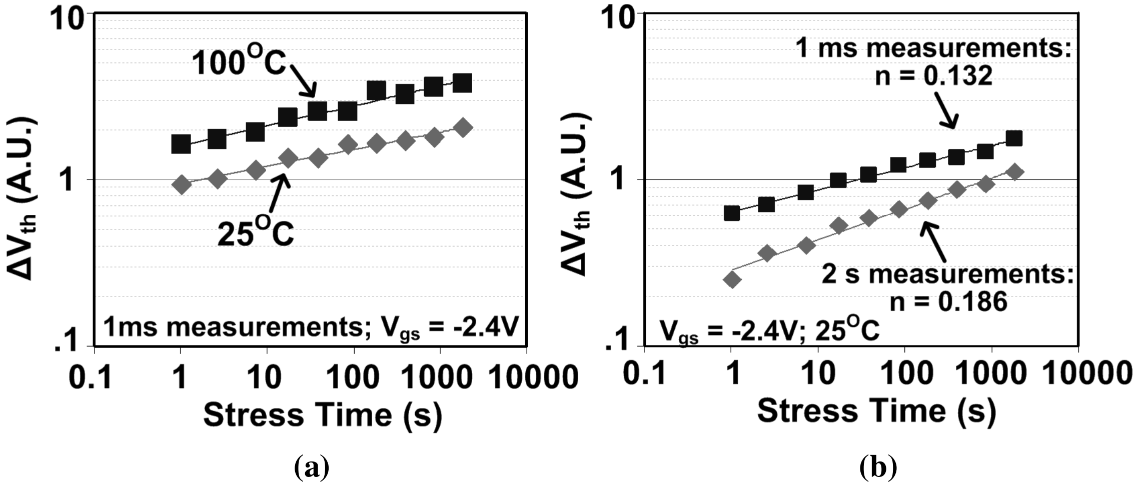
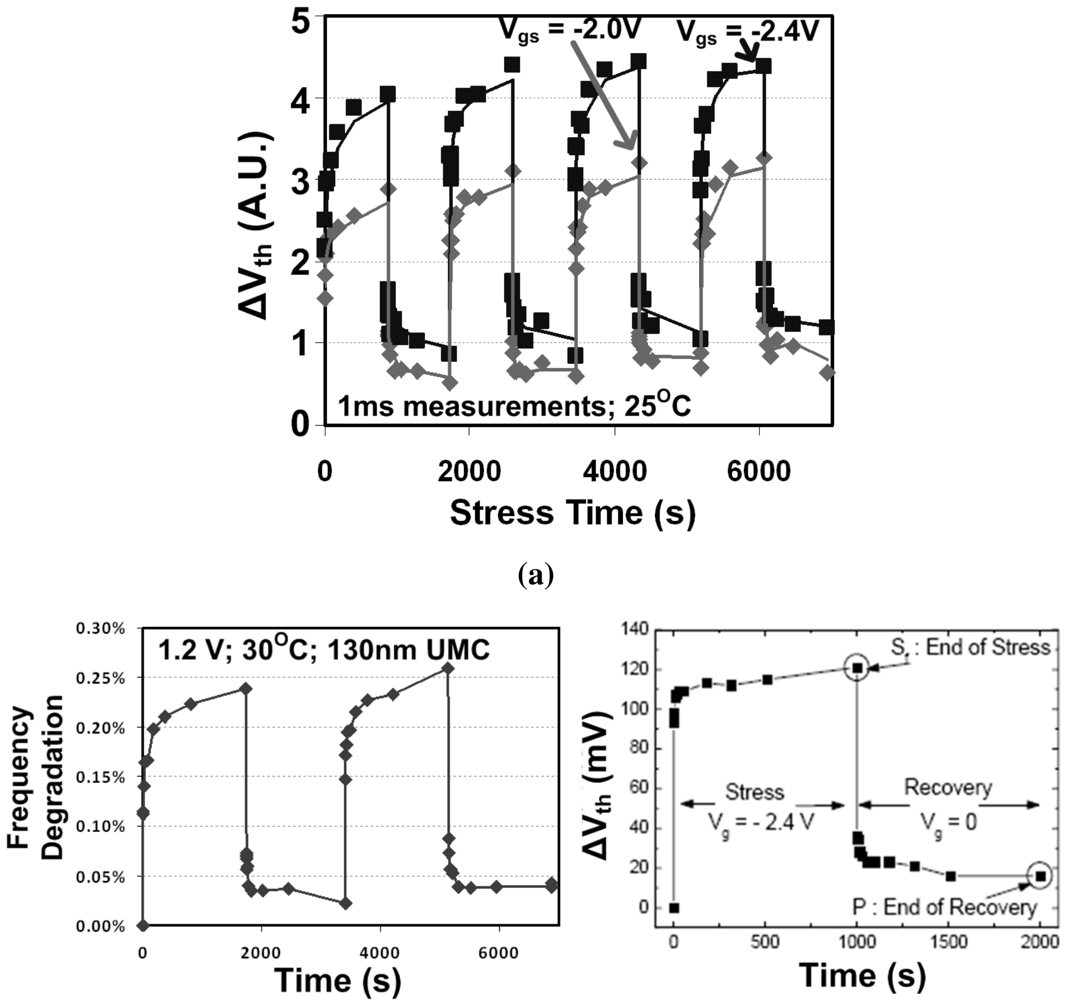


Fig. 4.12: (a) Comparison of NBTI stress measurements at 25^{O}C and 100^{O}C. (b) Comparison of 1 ms and 2 second measurement pulse results. A larger power law exponent is observed with longer measurement times, as found in [49].

Successive stress and recovery measurements are presented in Fig. 4.13 (a). A fast and significant recovery is observed, matching the characteristics seen in Figures 4.13(b) and 4.13(c) [10], [11]. This behavior has been attributed to the charge trapping/detrapping process, and was shown to be incompatible with a slower diffusion limited process [10], in which the recovery process would be slower and stress time dependent. However, it should be noted that this assertion stands in contrast to other recent work, where the fast recovery is said to be compatible with the R-D model in which interface trap creation and passivation is the underlying cause for NBTI transient effects [50].



(a)

(b)

(c)

Fig. 4.13: (a) Stress/Recovery curves demonstrate fast recovery when V_{gs} = 0 V. Note that ΔV_{th} does not fully recover in ~1000 seconds at 25^{O}C. This behavior was also found by (b) Kim [11] and (c) Shen [10] with high-speed measurement techniques, as well as Varghese [50] with on-the-fly measurements.

4.4 Conclusions

We have described a fast and efficient on-chip NBTI degradation sensor using a DLL. The shift in PMOS threshold voltage due to stress is amplified and directly translated into the control voltage of that DLL, which facilitates an easy characterization of NBTI in the

DUTs. The proposed measurement system is capable of measuring the effects of accelerated DC and AC stress by simply monitoring that control voltage with standard lab equipment. No expensive probe stations are required, and if designed with an emphasis on minimizing area, this system could be used as an on-chip real-time aging monitor to control an aging compensation mechanism such as clock frequency adjustments. A test chip was fabricated in a 1.2 V, 130 nm CMOS process for concept verification. The hardware implementation demonstrated a maximum system gain from the PMOS threshold degradation to a DLL control voltage drop of up to 16X in the operating range of interest, with an average gain of ~10X. Simulations as well as measurements show that this system is capable of taking threshold voltage readings in tens of microseconds in order to avoid unwanted recovery. NBTI degradation measurements were presented for a range of stress voltages, temperatures, and measurement times, demonstrating the flexibility of our design.

Chapter 5

An Array-Based Test Circuit for Fully Automated Inversion Mode Gate Dielectric Breakdown Characterization

5.1 Introduction to TDDB and the Array-Based Measurement System

While scaling CMOS device dimensions allows designers to pack more, and faster, transistors on a die, it also leads to an increased susceptibility to variations and reliability mechanisms. One such reliability issue is time-dependent dielectric breakdown (TDDB) in gate stacks, as was briefly explained in Chapter 1. This mechanism causes a conductive path to form through a gate dielectric layer placed under electrical stress, leading to parametric or functional failure. Breakdown has been a cause for increasing concern as gate dielectric thicknesses are scaled down to the one nanometer range, because a smaller critical density of traps is needed to build a conducting path through these thin layers, and stronger electric fields are formed across gate insulators when voltages are not scaled as aggressively as device dimensions.

In addition, the time to breakdown (T_{BD}) distributions for thinner gate dielectrics have a larger statistical spread over time [5], [51]. This can lead to large errors when extrapolating accelerated stress experiment results to realistic operating conditions and low failure percentiles in order to make device reliability predictions. The scaling of the physical dimensions of gate stacks can now be slowed or reversed with the introduction of high-k dielectrics, but TDDB remains a critical aging mechanism in those materials, and is currently being studied by device physicists [4], [6].

Although many of the physical details behind TDDB are still under debate, the percolation model is widely used to describe the gradual accumulation of electrical defects through a stressed oxide, which eventually form a current conduction path resulting in breakdown (Fig. 5.1) [51]. This model has been extended and modified to deal with ultra-thin oxides [52], [53]. Some studies have used the time to the first breakdown event (defined as an increase in gate current to some pre-determined level) to extrapolate predicted device lifetimes from accelerated stress experiments [5], [54]. A range of currents can be detected after this first event, and the distinction between current paths with low and high conduction levels led to the classification of “soft” and “hard” breakdowns. However, the definitions of those terms are contentious, and some authors claim that all breakdowns are more correctly described as progressive in nature [55].

In addition, it has become apparent that transistors can continue to function in certain cases after one or more breakdowns (Fig. 5.1(b)), and the progressive, post-breakdown current evolution must also be taken into consideration to obtain less pessimistic lifetime projections [55]-[57]. This is particularly true when operating at lower voltages and with thinner dielectrics, making an observable progressive breakdown current more likely before final device failure.

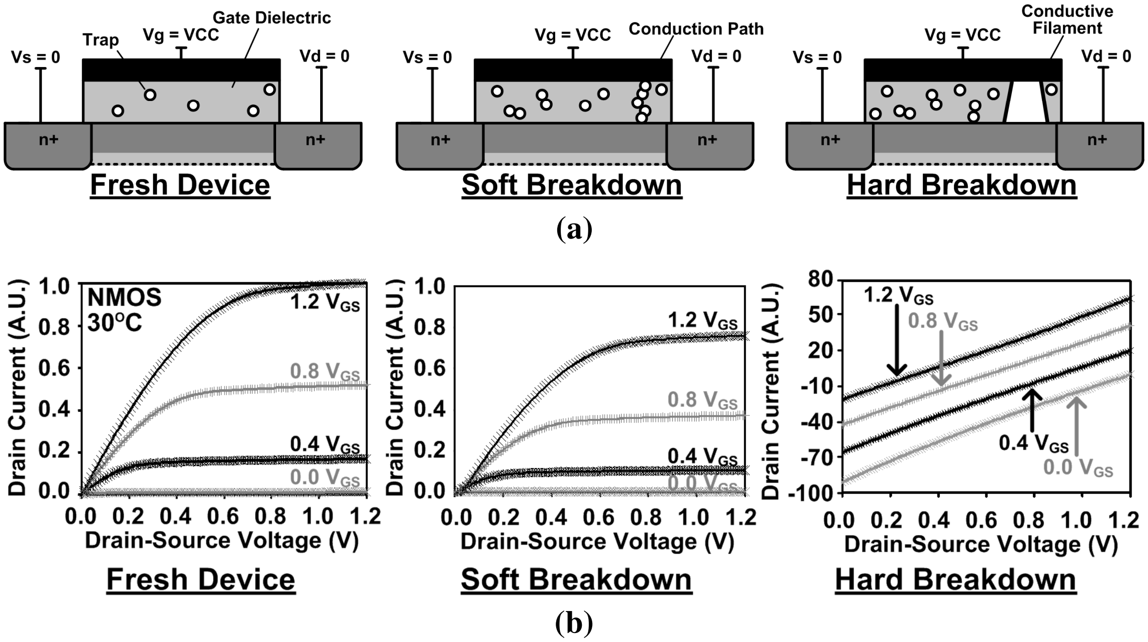


Fig. 5.1: (a) Cross sections of an NMOS transistor under inversion mode stress experiencing a progressive breakdown. (b) Measured I-V curves from device probing experiments on a NMOS device in a 130 nm bulk technology. In this case constant gate voltage stress was stopped when a "soft" breakdown was observed (i.e., a gate current that was still orders of magnitude below the specified current compliance). After the complete hard breakdown, the device fails to exhibit the desired characteristics.

TDDB is a function of a number of variables, including the gate voltage and oxide thickness as mentioned earlier, as well as temperature, device area, and dielectric materials and purity. Several models have been used to describe the relationship between the time to failure due to breakdown and these variables, but additional work is needed to more fully characterize TDDB in general so that the correct predictive models can be selected. The specific breakdown behavior of each new CMOS process must also be thoroughly tested during the process characterization phase in order to obtain a detailed understanding of the technology reliability.

Most of the previously published TDDB measurement results were gathered from individual device probing experiments. The equipment used in those tests can be expensive, and testing each device individually leads to long experiment times. However, one on-chip circuits-based method to monitor gate dielectric wear-out was recently proposed [15]. In this case, results were provided in the form of a frequency shift of a Schmitt trigger oscillator which is modified by the increasing gate leakage through a pair of stressed PMOS devices. This design can provide some indication of the wear-out behavior of stressed transistors, but does not facilitate a direct reading of gate resistance degradation, or any other specific device characteristics (i.e., the end result is oscillator degradation with no suggestion of how to translate this into another parameter).

In this paper, we present a circuit design that performs automated measurements in a test array to directly gather the inversion mode (i.e., “on-state”) breakdown characteristics that define this statistical process. The proposed circuit can monitor a progressive decrease in gate resistance, or simply an abrupt failure, often referred to as a hard breakdown. This structure greatly reduces the required process characterization time, which may involve continuously monitoring the current through a single device under test (DUT) per experiment with a parametric test system.

Given the need for up to thousands of samples to correctly define the Weibull slope of the T_{BD} distribution [5], [58], that serial testing process quickly becomes cumbersome. Therefore, in the circuit presented here, DUTs are stressed in parallel and we continuously loop through the array, temporarily removing stress conditions in one cell at a time and measuring each DUT's gate current. In addition, the array format is a convenient method to study any spatial correlation of TDDB without requiring elaborate test setups. Test array structures are gaining popularity as an efficient way to gather process technology information, since individual device probing is not convenient when large numbers of readings are required [59], [60].

5.2 Breakdown Characterization Array Design

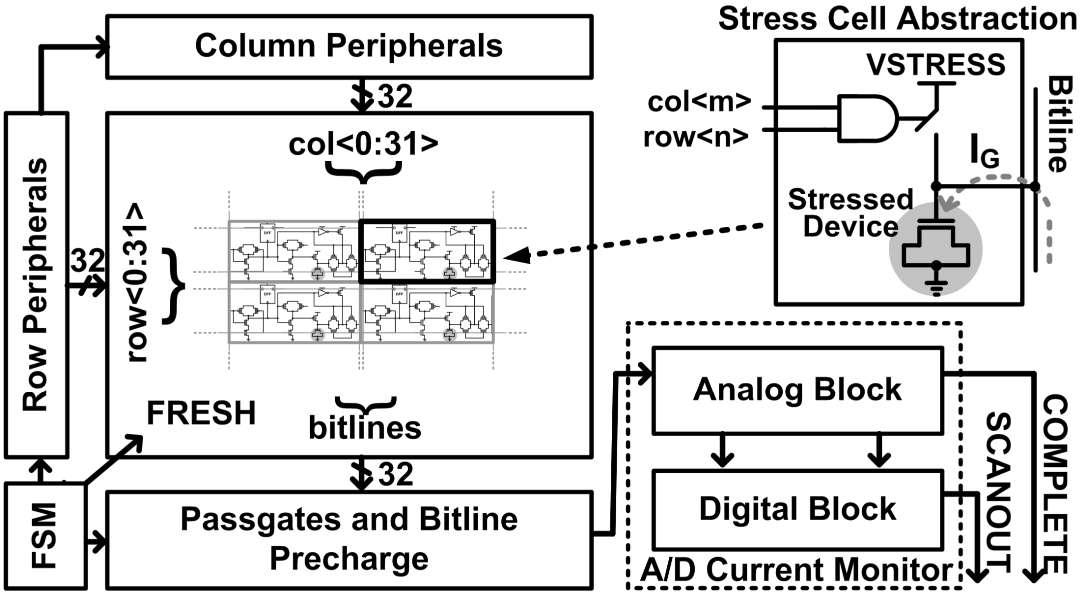
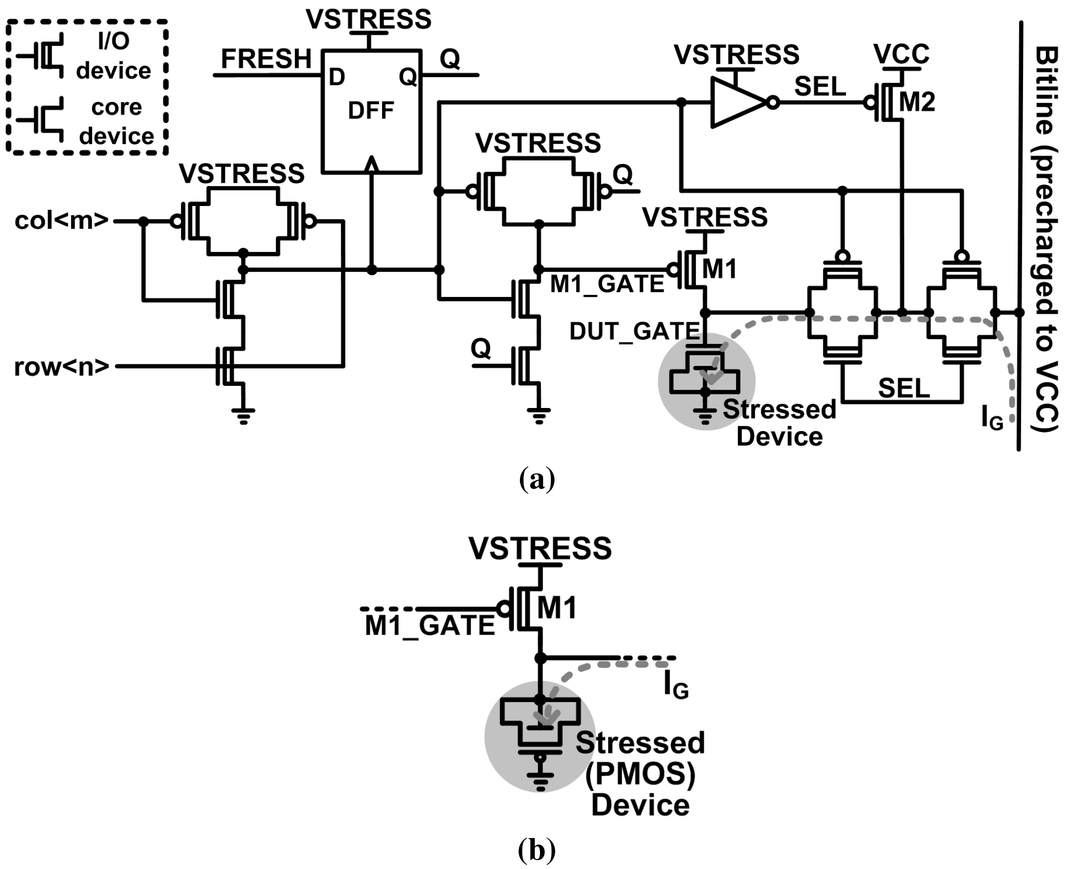


Fig. 5.2: Top level diagram of the 32x32 array for fully automated gate dielectric breakdown characterization.

The proposed test circuit design consists of a 32x32 array of structures we call "stress cells" that contain the DUTs, whose gate currents (I_G) are periodically measured using an A/D current monitor and on-chip control logic (Fig. 5.2). After an initialization sequence, cells are cycled through automatically without the need to send or decode cell addresses, in order to simplify the logic and attain faster measurement times. A single external clock signal is asserted each time that the controlling software is ready for a new I_G measurement. Although we chose to simplify and speed up the circuit in this manner, we do have the ability to select any one portion of the array for measurements while turning off stress in the rest of the test cells, as will be discussed later. The finite state machine (FSM) in Fig. 5.2 controls the initialization sequence timing, as well as that of the subsequent measurements. The row and column peripheral circuits contain D flip-flops (DFFs) and multiplexers used to select a particular cell, as well as level shifters to boost signals from the 1.2 V (VCC) digital supply domain to the stress voltage (VSTRESS) level, which is used as the supply voltage within the array.

5.2.1 Stress Cell Design

The stress cell structure shown in Fig. 5.3(a) was implemented to facilitate the accelerated stressing of the DUTs, by using thick oxide I/O transistors in the supporting circuitry to avoid excessive aging or breakdown in these other devices. (The dual-oxide requirement for the present design is commonly met by modern processes, but we currently have work underway to implement stressing circuits with a single oxide thickness.) Transistor M1 drives the DUT gate to VSTRESS if the cell has been turned on for a stress test. At the same time, M2 holds the node between the two pictured transmission gates at VCC, which matches the bitline precharge level, until the cell is selected for a measurement. The row<n> and col<m> signals are used to execute this selection event by setting both to the logic high level. At that time, devices M1 and M2 are turned off, and the transmission gates connecting the gate of the DUT to its bitline are turned on. After these steps are taken, the gate current through the stressed DUT is measured by the A/D current monitor.



(a)

(b)

Fig. 5.3: (a) NMOS stress cell with bitline leakage compensation and stress/no-stress capability. (b) A PMOS stress cell would be identical to that seen in (a), with the change illustrated here. Note that the PMOS DUT requires its own isolated nwell.

The FRESH signal is used to permanently gate off stress on a broken DUT when a high gate current is detected, in order to avoid excessive current draw from the VSTRESS supply. After a sufficiently high breakdown current is measured in a selected cell, FRESH is set to 0 V by the controlling software before row<n> goes low, which latches a logic low value on Q. This isolates the DUT from VSTRESS by turning off device M1. When a cell is not being selected for measurement and Q is still high, M1 is turned on and the DUT is placed under constant voltage stress. Note that the M1 devices should be sized to model a realistic gate driver. The current limitation of the driving stages in TDDB experiments, such as M1 in this case, have been reported to strongly influence post-breakdown characteristics [61]. However, it was stated in that same work that the time until the first breakdown is not changed by the strength of these drivers.

PMOS transistors could be tested within this same framework by changing the DUT configuration as shown in Fig. 5.3(b). In this case, the PMOS DUT would be contained in an isolated nwell, and the drain and source would be connected to its body contact. The gate terminal would stay grounded, while the other three terminals are stressed or left floating for current measurements. In the first implementation presented here, we used only NMOS devices for simplicity and consistency.

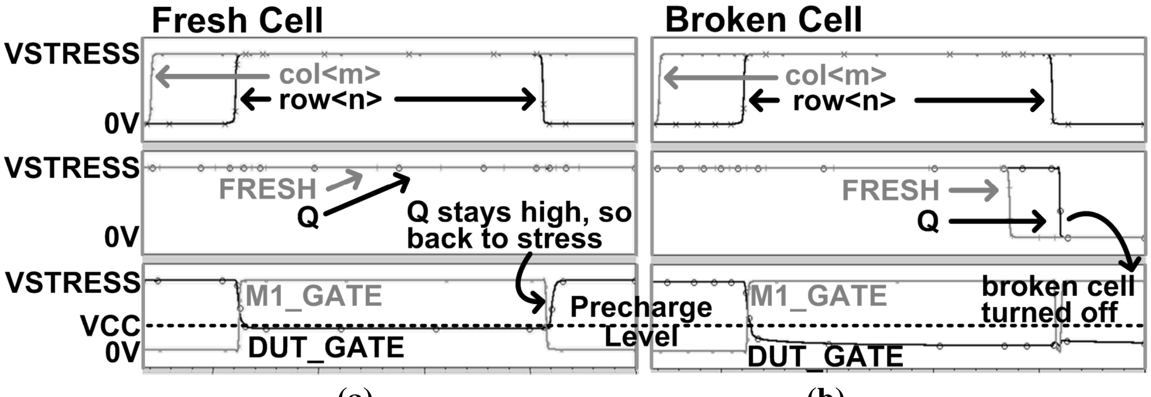


Fig. 5.4: Simulated stress cell operation corresponding to Fig. 5.3(a). (a) Illustrates a measurement taking place in a fresh (i.e., pre-breakdown) cell. (b) Illustrates a measurement in a broken cell.

Simulation waveforms demonstrating the measurement procedure for a fresh cell (i.e., DUT with low I_G) and a broken cell (i.e., DUT with high I_G) are presented in Figs. 5.4(a) and (b), respectively. In the unbroken, or "fresh" cell, the DUT_GATE node voltage drops to the 1.2 V precharge level before slowly decaying to ~V_{REF} (a value defined in Section 5.2.2), and then being charged to VSTRESS again when the row<n> signal drops to 0 V. When the "broken" cell is accessed for a measurement the DUT_GATE node discharges to 0 V through the breakdown path. In this case, the FRESH signal is set to 0 V before the cell is deselected, so M1_GATE remains high, and no further stressing occurs in this cell. We do not expect that the voltage transients on the DUT_GATE node during the measurement transitions should significantly impact the breakdown process since no voltage overshoots are observed. Also, several reports have found that time to breakdown simply increases with shorter stress duty cycle, rather than being negatively impacted by the switching activity [62], [63]. Finally, the drop from VSTRESS to VCC on the DUT_GATE does not impact the gate resistance measurement because the bitline is held at VCC by the precharge device until after the cell is selected.

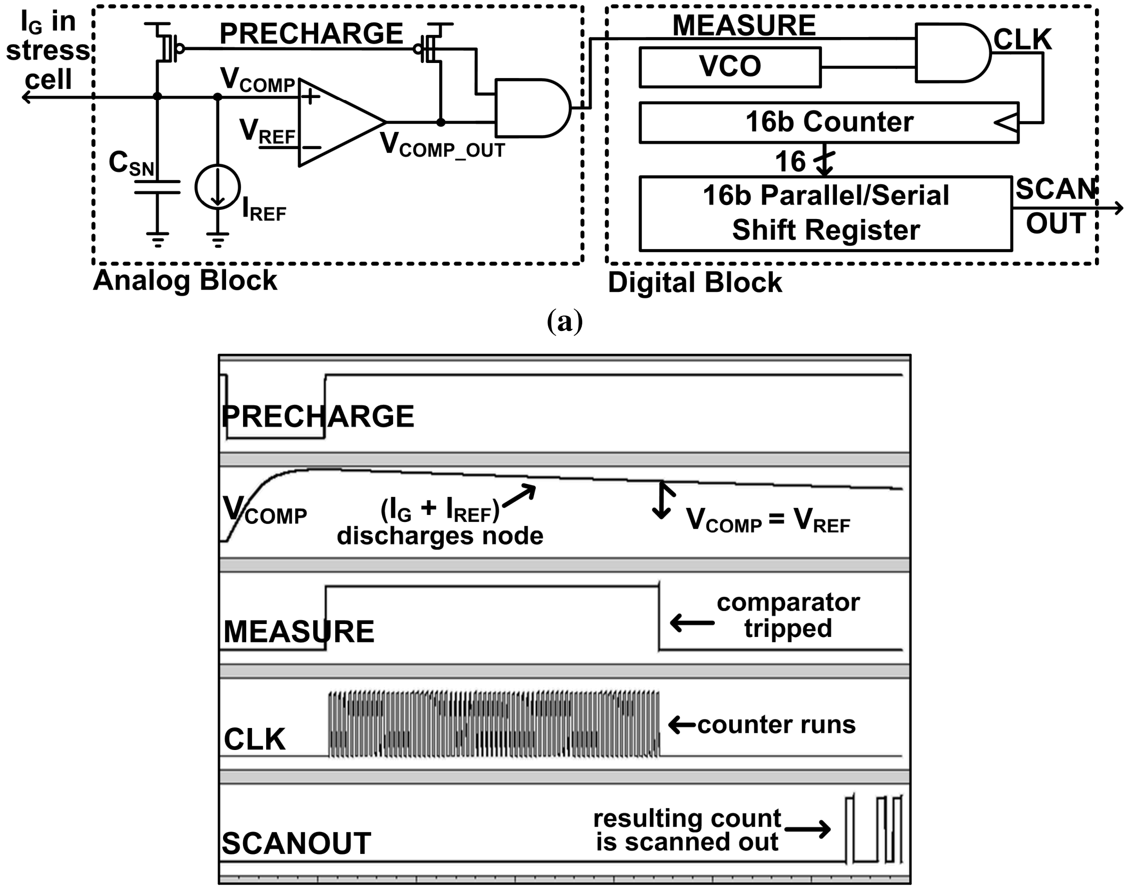
The FRESH signal can also be used during circuit initialization to gate off stress in a range of unused cells which may be tested at a later time, or cells that are already broken from a previous experiment. This feature allows us to measure any one portion of the array during a single test, rather than the entire 1024 cells, if so desired. In the cell range selection step, we leave VSTRESS at 1.2 V during the first loop through the entire array, while setting FRESH low for those cells that we do not wish to measure during subsequent loops. The thick oxide I/O transistors in the stress cells operate correctly at this low voltage level, and we expect that no appreciable gate dielectric degradation will occur in the DUTs at their nominal supply voltage over the course of a few seconds. After this initialization loop, VSTRESS is raised to the stressing voltage, and measurements proceed as usual in the selected portion of the array.

Two transmission gates were placed between the gate of each DUT and its bitline, with the internal node held at VCC when the cell is not selected, in order to keep leakage between all unselected stress cells and the A/D current monitor low and consistent. Simulations show that the total leakage sourced by all 1023 unselected cells in the array during a measurement is limited to ~108 nA. This worst-case leakage on the discharge path occurs when the selected DUT's gate node has discharged to ~1.1 V (the V_{REF} level) at 30 ºC and VCC = 1.2 V.

5.2.2 A/D Current Monitor

The analog block shown in Fig. 5.5(a) contains a comparator whose precharged output drops to 0 V when the precharged input voltage (stored on an 80 pF metal capacitor, C_{SN}) falls below the reference voltage (V_{REF}) level. That discharge rate is determined by I_G in the selected cell, plus an external reference current (I_{REF}) that is also used for calibration purposes. The digital block (Fig. 5.5(a)) contains a 16 bit counter that runs at a rate set by a voltage controlled oscillator (VCO), from the end of the precharge event until the comparator's output falls, indicating that a measurement is complete. Therefore, lower I_G (i.e., a larger gate resistance, R_{GATE}) translates into a higher count, and vice versa.

The final count result is latched into a parallel/serial shift register and scanned off-chip after the software interface detects a completion signal, which is asserted by the analog block. The results are stored in a convenient spreadsheet format for post-processing. A calibration technique described in the Section 5.3.1 makes it possible to translate these resulting counts into gate resistance values, so we can monitor a progressive breakdown process in each of the stressed devices by running measurements in a continuous loop through the array. The simulation waveforms presented in Fig. 5.5(b) illustrate the basic outline of this measurement procedure.



(a)

(b)

Fig. 5.5: (a) The A/D current monitor used to translate the gate current through a DUT (I_G) into a 16 bit digital count. (b) Simulation of this A/D conversion.

5.2.3 Peripheral Circuits and Operational Flow

The first two rows of the row peripheral block from Fig. 5.2 are illustrated in Fig 5.6(a). These circuits are identical to those in the column peripherals, but the latter are only clocked once after each time an entire column of cells has been selected individually. As mentioned earlier, cells are cycled through automatically at each pulse of the internal clock (Φ_{INT}) without the need to send or decode cell addresses, in order to simplify the logic and attain faster measurement times. The all_stress and no_stress select signals for the 3-way MUX are used during circuit initialization, or to hold the array in a steady state where either all cells or no cells are stressed.

Fig. 5.6(b) shows an I/O diagram for the FSM, which uses three bit state encoding, and the corresponding state transition diagram is presented in Fig. 5.6(c). State transitions, or moves to the next cell to be tested during the *MEASURE* state, occur with the assertion of an external clock signal (Φ_{EXT}) and depend on the current state and FSM inputs. Note that internal signals not shown explicitly in this transition diagram are set to 0 V. When the external reset signal ($RESET_{EXT}$) is asserted at any point during operation, the measurement system enters the *RESET* stage, where all DFF outputs are driven to 0 V and stress is turned off in all cells.

The latter is accomplished by setting the no_stress signal high, which drives all peripheral MUX3 outputs to VCC, thereby selecting all cells to turn off all M1 transistors (Fig. 5.3(a)). After the next assertion of Φ_{EXT}, the unstressed array will wait in the *NO STRESS* state until $STRESS_{EXT}$ is set to logic high, meaning that we wish to enter a constant stressing state for all cells (*ALL STRESS*), or $\text{START}_{\text{EXT}}$ is asserted indicating that we want to start normal stress/measurement operation (*MEASURE*). In either case, the $\text{FRESH}_{\text{EXT}}$ signal is first set high before deselecting all cells in order to clock all of the DFFs within the stress cells (Fig 5.3(a)), and set all Q values in those cells high. All cells are then deselected by setting all_stress high, which drives all peripheral MUX3 outputs to 0 V. If $\text{START}_{\text{EXT}}$ is asserted, we continue from this step into normal stress/measurement operation.

As stated earlier, VSTRESS is left at the nominal digital supply voltage of 1.2 V during the short circuit initialization period in order to prevent accelerated stressing of the DUT gates. It is then held at 1.2 V throughout the first measurement loop if we wish to keep only a selected portion of the cells on for stress by appropriately asserting the $\text{FRESH}_{\text{EXT}}$ signal. Immediately after this startup phase, VSTRESS is raised to the stressing voltage, and measurements proceed as usual in the selected portion of the array.

The on-chip phase of the measurement after each cell selection event required ~100 μs or less, as determined by the values of C_{SN}, I_{REF}, I_G, and additional leakage currents. However, the timing bottleneck was the results scanout routine executed by the controlling software. This portion of each measurement led to a total measurement time of several hundred milliseconds. Therefore, in order to keep a reasonable timing resolution of roughly 15 seconds or less between sequential checks of each stressed cell (depending on the VSTRESS value), we limited the number of cells tested in any one run. For example, experiments often covered a 5x5 portion of the array. As explained earlier, the FRESH signal in each cell was set such that the stress cells not used during any particular experiment were turned off. In the future, an improved software interface or different data acquisition board could be used to greatly reduce the results scanout time.

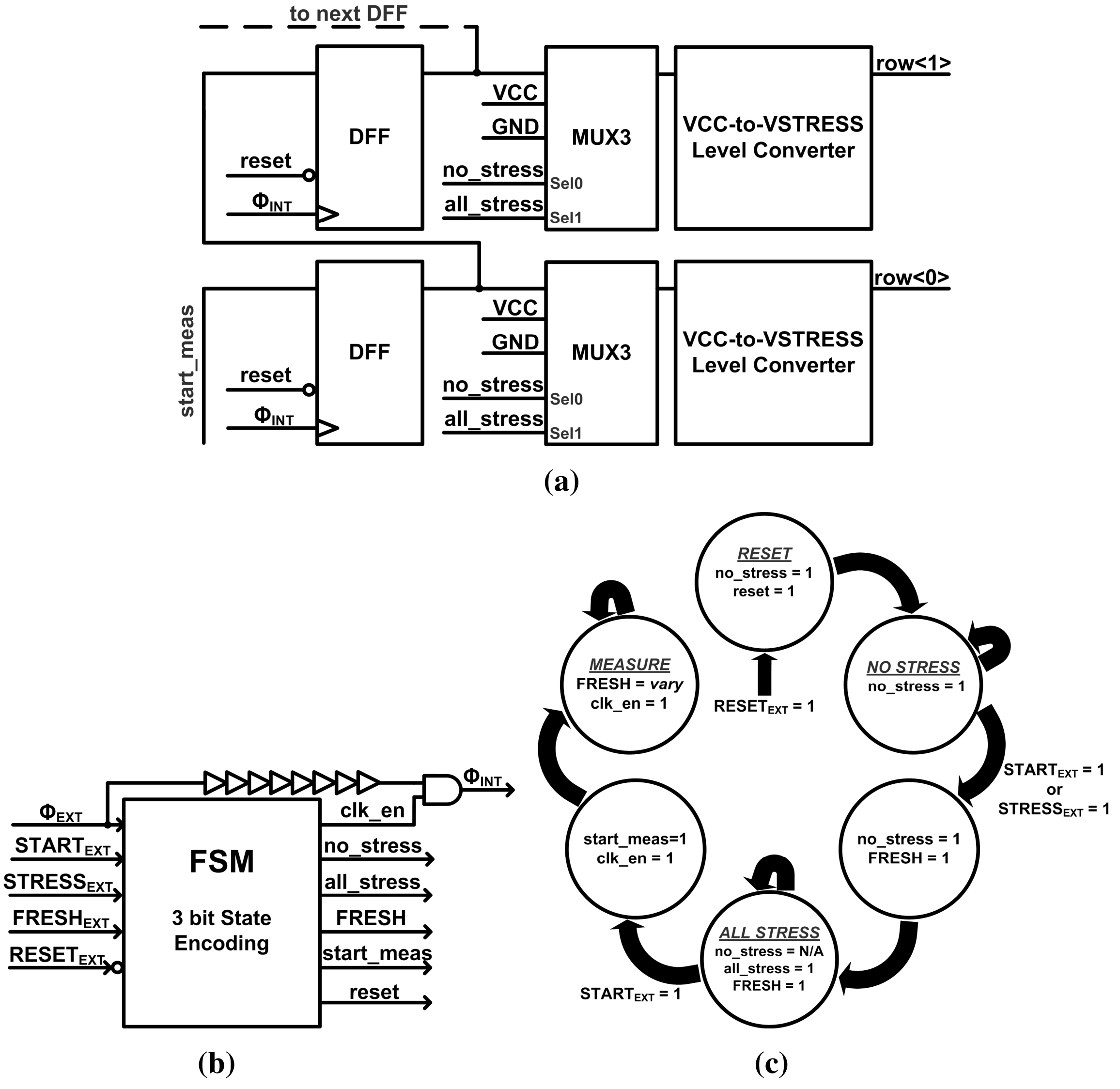
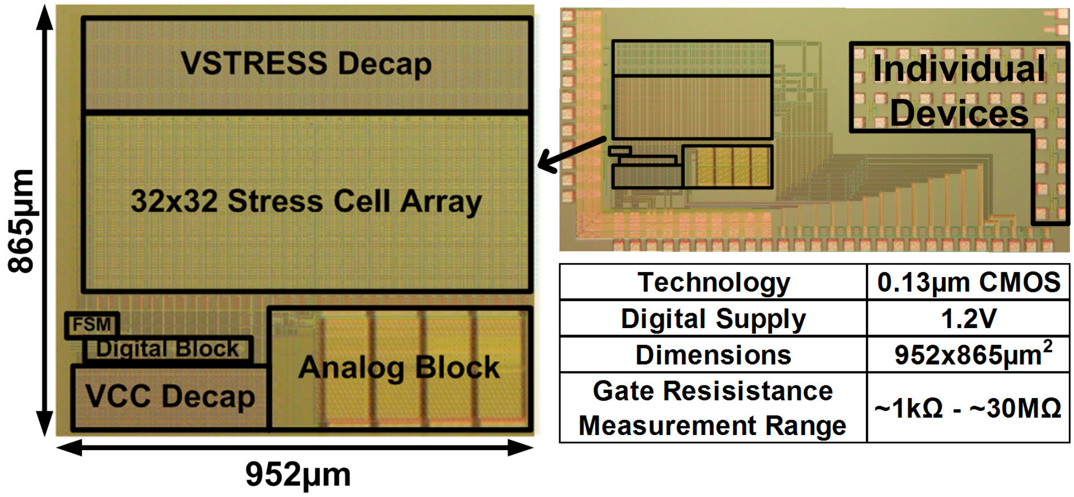


Fig. 5.6: (a) Block diagram of the first two rows of the row peripherals from Fig. 5.2. The column peripherals are identical to this, but are only clocked once after each time an entire column of cells has been accessed. (b) I/O diagram of the finite state machine (FSM) which uses three bit state encoding. (c) State transition diagram. Transitions occur with each assertion of the external clock signal (Φ_{EXT}). Note that all internal signals not show explicitly in each state are set to 0 V.

5.3 Inversion Mode TDDB Array Test Chip Measurements

A test chip was fabricated in a 1.2 V, 130 nm bulk CMOS process. Each NMOS device under test had a width and length of 2 µm. Information about the gate dielectric construction is confidential, but the thickness is within a reasonable range for this technology node. Automatic measurements were completed with LabVIEW™ software and a National Instruments data acquisition board, which was connected to a laptop through a USB port. A microphotograph of the chip and a summary of the circuit characteristics are shown in Fig. 5.7. In the picture on the right of this figure, which captures a larger portion of the total chip, we point out a number of individual devices that were fabricated to verify that the results from our array match well with probing measurements. The probing experiments were completed with an HP semiconductor parameter analyzer and a Signature probe station.



Technology	0.13µm CMOS
Digital Supply	1.2V
Dimensions	952x865µm^2
Gate Resisistance Measurement Range	~1kΩ - ~30MΩ

Fig. 5.7: Microphotograph and summary of the test chip characteristics. The individual devices reserved for probing experiments are labeled to the right of the TDDB array measurement system.

5.3.1 Test Chip Calibration Procedure

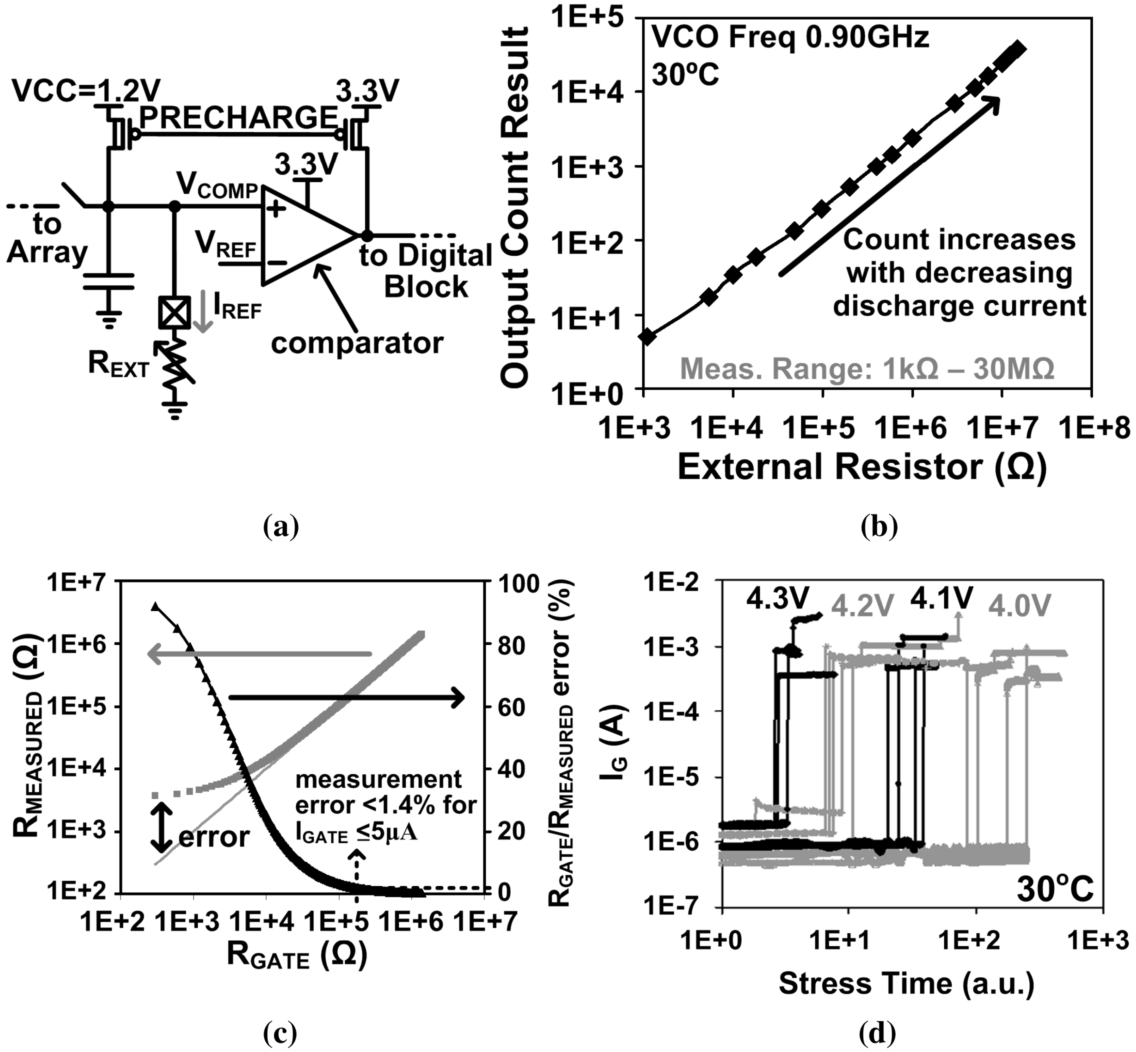
The calibration procedure and measurement results are illustrated in Figs. 5.8(a) and 5.8(b), respectively. In order to obtain the final count vs. total discharge path resistance characteristic, the A/D current monitor was gated off from the breakdown array, and an adjustable external resistor (R_{EXT}) was attached to the I_{REF} path. Therefore the total discharge path resistance (R_{TOTAL}) in this case is simply the value of the external resistor. During this calibration procedure, the A/D current monitor is run normally, as it would during stress measurements, but with a range of known R_{TOTAL} values. This leads to a calibration curve, like that shown in Fig. 5.8(b).

After the calibration is completed, each output count recorded during stress measurements can be translated into a gate path resistance by using the calibration curve, and the simple equation $R_{TOTAL} = R_{EXT} \| R_{GATE}$. Throughout measurements, R_{EXT} is held at a known constant value, and R_{TOTAL} is taken from the calibration curve at the point with an equivalent output count result (i.e., the stress measurement count result matches a certain calibration count result), so R_{GATE} is the only unknown.

The range of gate resistances that this array-based system is able to record is roughly bounded from above by the value of R_{EXT}, since the smaller value of $R_{EXT} \| R_{GATE}$ will dominate this equation, as well as the size of the counter that the VCO clocks during measurements. As explained in Section 5.2.2, a larger R_{GATE} leads to longer C_{SN} discharge times, and hence higher count results. Therefore, measuring high values of R_{GATE} requires a sufficiently large counter. The lower bound of the measurement range is set by the speed of this VCO, because a higher clock rate is required to maintain sufficient resolution with faster discharge times. These bounds should be appropriately adjusted at design time, as well as during calibration. As we show in Fig. 5.8(b), the present design achieves an R_{GATE} measurement range of roughly 1 kΩ through 30 MΩ (following the plotted trend through a count of 2^{16}) with the VCO clocked at 900 MHz.

The resistance of the transmission gates located on the path from V_{COMP} to the DUT gates is not accounted for in this calibration procedure, and therefore introduces error that becomes more severe as the DUT gate resistance drops into the hard breakdown region. That is, our measured R_{GATE} results will be larger than the correct value because of the additional transmission gate resistances. However, due to the relatively high value of R_{GATE} during the progressive degradation stage leading up to the final hard breakdown, this error is small in the region of interest, as shown in Fig. 5.8(c). The measurement error is less than 1.4% for R_{GATE} values of 240 kΩ and greater, corresponding to gate currents up to 5 μA at a sensing voltage of 1.2 V. Several authors have indicated that the soft to progressive breakdown regimes are within this current limit [56], [57].

A more detailed calibration path that duplicates the additional transmission gate resistances and other non-idealities could be included in future test chips to eliminate this small error. For example, the circuit could include replica cells embedded within the measurement array for calibration. An external resistor could then be attached directly to the node within those cells where a DUT gate would regularly be located. This procedure would exactly duplicate the normal measurement routine so that all leakages and parasitics are accounted for.

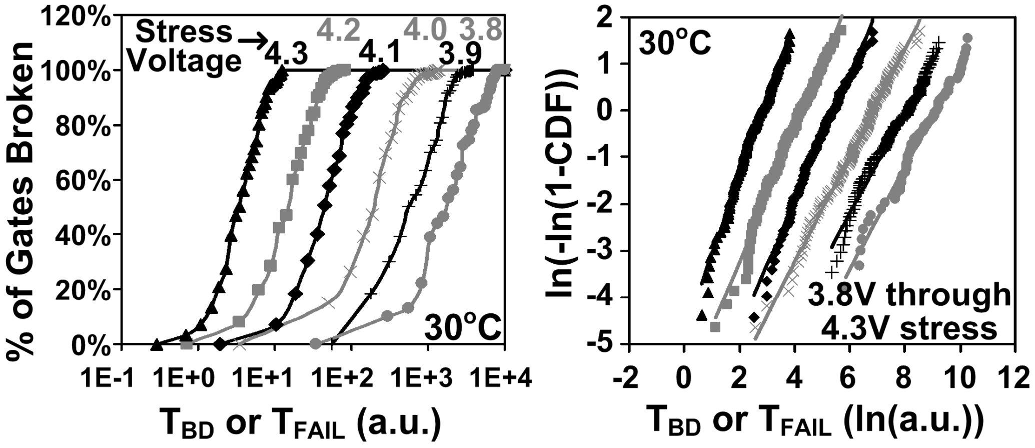


(a) (b)

(c) (d)

Fig. 5.8: (a) Measurement calibration setup. (b) Measured calibration results. (c) The resistance of the transmission gates located on the path from V_{COMP} to the DUT gates is not accounted for in this calibration procedure, but only introduces a small measurement error in the progressive breakdown region. (d) Individual device probing results indicate that in the stress voltage range of interest, with a sampling rate of 4 Hz, we expect to observe hard breakdowns in the majority of our experiments.

However, as seen in the direct device probing results of Fig. 5.8(d), we typically did not observe progressive dielectric breakdown in the CMOS process used here. This data was recorded during accelerated measurements with stress voltages of 4 V, when recording four measurements per second. Therefore, although the proposed design is capable of monitoring progressive breakdowns, we were specifically looking for hard breakdowns in our automated array measurements. These events were defined as a sudden and sustained decrease in the scanned out discharge time count of roughly two orders of magnitude, when the VCO clocking the counter in the A/D current monitor was running at 900 MHz.



(a) (b)

Fig. 5.9: Measured T_{BD} CDFs on (a) a standard percentage scale and (b) a Weibull scale.

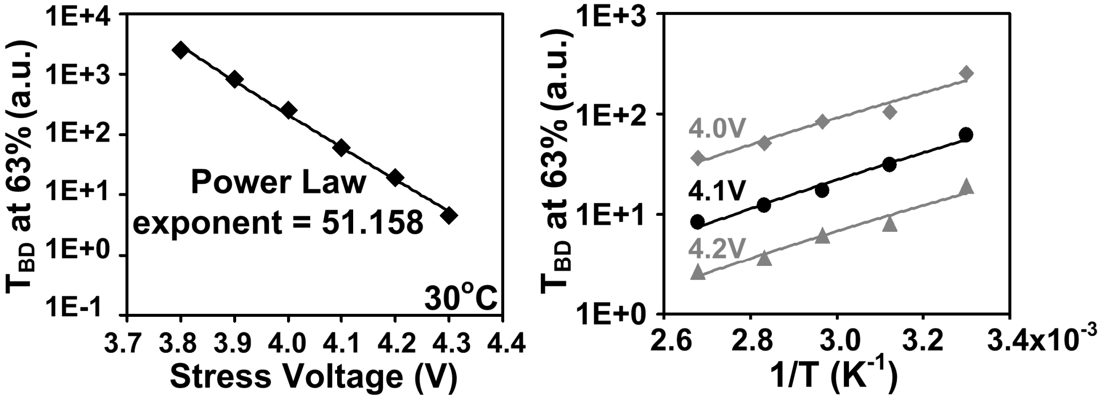
5.3.2 Measured Breakdown Distributions

Cumulative distribution functions (CDF) of the time to breakdown, both on a standard percentage scale as well as the Weibull scale, are displayed in Figs. 5.9(a) and 5.9(b), respectively. That data was gathered at 30 ºC, with stress voltages from 3.8 V to 4.3 V in this 1.2 V process. TDDB follows Weibull statistics because this mechanism has a weakest-link character, since there are a large number of spots in each gate where the first breakdown can occur, and the breakdown process proceeds independently at each of them. The first breakdown at any of those locations leads to device degradation or failure though, so it can be thought of as the "weakest link." When we have a weakest-link process, extreme value distributions are the first functions we try to fit to measured data. Since in the case of time to breakdown the distribution is bounded from below at time zero, specifically we use a Weibull distribution [64].

The Weibull slope factor (β) for 4.2 V stress was 1.443, with that factor slightly decreasing for lower stress voltages, and increasing at 4.3 V. We generally expect that the Weibull slope should be dependent on gate dielectric thickness, but not voltage, so this slight difference was not expected. However, the slope values are still in good agreement with other published data [5], [61]. This trend is also observed to some degree in the results presented by Röhner, although not mentioned explicitly [61]. Finally, in a recent publication by Tous exploring breakdown in ultra-thin gate oxides, an explanation was provided for steeper distribution slopes at lower ranges of T_{FAIL} on the Weibull plot [58]. This phenomenon was attributed to the non-Weibull shape of T_{FAIL} for very thin oxides, which is only correctly observed with sufficiently large test sample sizes (well over 100). For these reasons, the small variation in our measured breakdown distribution slopes seems reasonable, and may have a theoretical justification.

5.3.3 Voltage and Temperature Acceleration of TDDB

The exponential relationship of the Weibull characteristic life (time at which 63% of the devices have failed) with voltage is illustrated in Fig. 5.10(a). The power law exponent is ~51, which is slightly larger than that reported in previous work where the time to the first breakdown event (soft or hard) was recorded [65]. Note that Wu et al. provided a physics-based explanation for voltage acceleration power law factors in the 40-50 range in that paper.



(a) (b)

Fig. 5.10: (a) Voltage acceleration of T_{BD} at the 63% point. (b) T_{BD} at the 63% point vs. the inverse of the temperature in Kelvins.

The measured dependency of the time to breakdown on stress temperature is shown in Fig. 5.10(b) for a range of voltages. In this temperature range of 30 ºC to 100 ºC, TDDB follows Arrhenius behavior with only small errors. Although the temperature dependence of breakdown is often modeled using Arrhenius behavior, non-Arrhenius dependence has also been reported at temperatures over 100 ºC, particularly for thin gate dielectrics [66], [67]. At any rate, the temperature acceleration of TDDB imposes more severe limits on modern CMOS designs where device density and high clocking rates lead to increased local heating.

5.3.4 Area Scaling Property of TDDB

In addition to showing that the CDFs form straight lines on a Weibull scale in order to justify the use of these statistics, we can also check that the process follows the area scaling property of this extreme value distribution (equation in Fig. 5.11). (Although it has long been established that TDDB follows Weibull statistics, we address this issue to illustrate the important concept of area scaling.) In our case since all DUTs are the same size, the measured numbers for different areas were obtained by combining the results for a given number of spatially adjacent DUTs. We then selected the smallest time to breakdown from each group, due to the weakest-link character of dielectric breakdown.

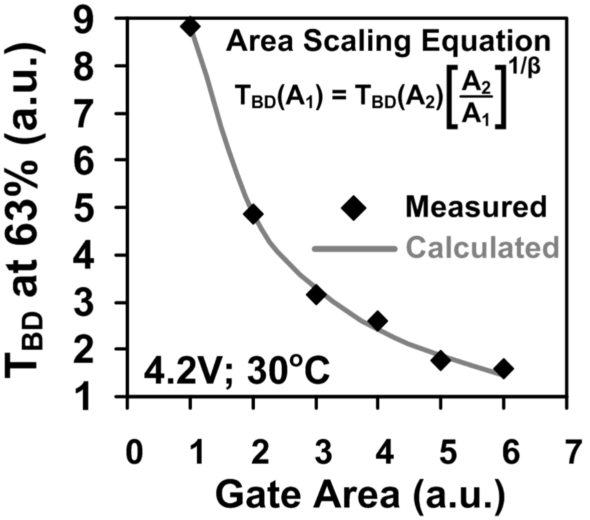


Fig. 5.11: Area scaling data computed from the combined measurement results of spatially adjacent stress cells, compared with theoretical results [5].

The results shown in Fig. 5.11 indicate that our measured data matches well with the theoretical area scaling equation [5], [64], [65]. The scaling property is also used in other studies to define the Weibull slope parameter with a high degree of accuracy. That is done by measuring the time to breakdown for devices with a large area ratio, and then using the equation shown in Fig. 5.11 where the only unknown is β.

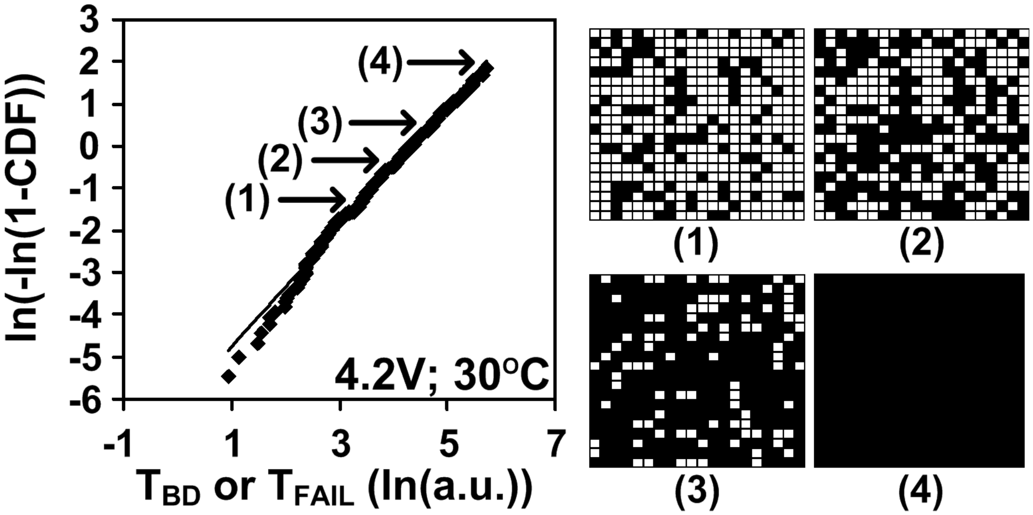


Fig. 5.12: Spatial distribution of TBD in a 20x20 stress cell array at four time points on the Weibull scale CDF. Cell locations are filled in once their DUT gates have broken down.

5.3.5 Spatial Distribution of Time to Breakdown

Test arrays such as ours, where a large number of devices are closely spaced, facilitate investigations of any spatial correlation in the process or characteristics being studied. For example, spatial correlation of gate oxide thicknesses could lead to a correspondingly correlated breakdown process [68]. The spatial distribution of T_{BD} in a 20x20 portion of a test array stressed at 4.2 V is plotted in Fig. 5.12, along with the corresponding Weibull distribution. The four spatial diagrams correspond to the four divisions of the Weibull plot representing 25% of the cells each. No spatial correlation is apparent from these plots, and it is possible to check our conclusion with a quantitative measure of that phenomenon by calculating the local and global Moran's I statistics [69], [70].

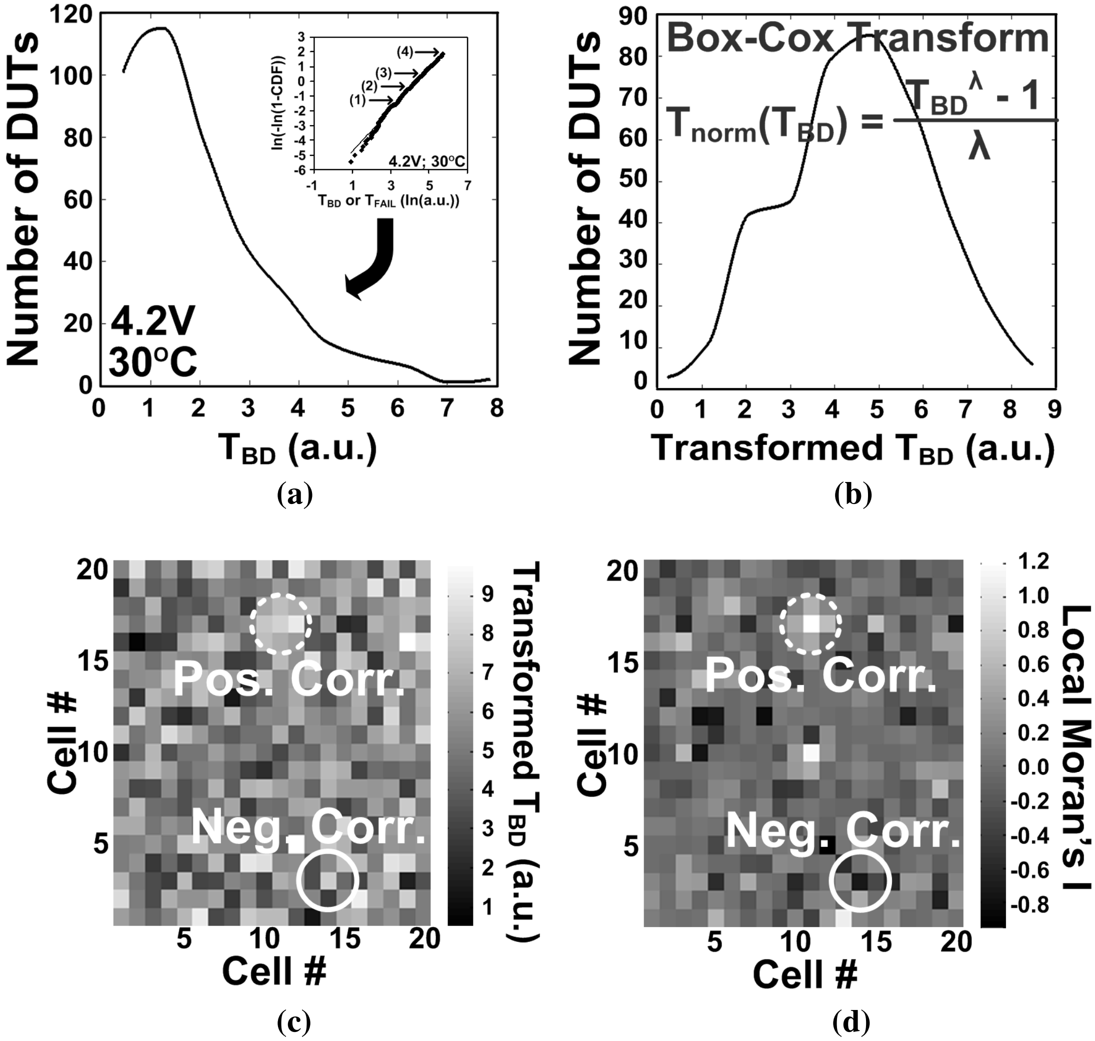


Fig. 5.13: (a) Histogram of time to breakdown in a 20x20 portion of a test array stressed at 4.2 V along with the corresponding Weibull plot from Fig. 5.12 (inset). (b) Histogram after the Box-Cox transformation is applied to create a normal distribution of T_{BD} data ($\lambda = 0.2833$). (c) Spatial diagram of the 20x20 array of cells with colors indicating each location's transformed T_{BD} (in arbitrary units matching those in part (b)). (d) Local Moran's I for each cell location. Light colors in this last plot indicate positive correlation (i.e., "clustering") while darker colors indicate negative correlation (i.e., "dispersion").

However, this method works under the null hypothesis that the input data are normally distributed random variables, which we have seen is not the case for T_{BD} distributions. This is made clear in Fig. 5.13(a) where we plot the histogram of the data used to create Fig. 5.12. The null hypothesis is common in statistical data analysis tools, so mathematicians have developed a number of methods to transform non-normal distributions to the normal form. The equation for the Box-Cox transformation, which can be used to transform Weibull distributions for this purpose, is shown in Fig. 5.13(b) [71]. This operation is defined by the λ in that equation, which in our case was calculated with the MATLAB® "boxcox" function. The exact value found was 0.2833, and the resulting histogram is shown along with the transform equation. We verified the symmetry of this new data set with a "triples test" [72], [73]. The transformed data is also shown in a spatial plot in Fig. 5.13(c) with arbitrary units, matching those in Fig. 5.13(b). The area of this 20x20 array is roughly 555 μm x 225 μm in the physical implementation.

A sliding 3x3 contiguity matrix in the queen configuration was used to calculate the local Moran's I statistics [69]. This matrix defines the neighborhood around each value that is used to calculate spatial correlation, and the "queen" term is an analogy to chess. In this case, correlation with all eight nearest neighbors surrounding one cell is computed, and the results are plotted in Fig. 5.13(d). Lighter colors in this last plot indicate stronger positive correlation (i.e., "clustering") while darker colors indicate negative correlation (i.e., "dispersion"). Examples of both extremes are indicated. It is apparent that positive spatial correlation corresponds to cell locations in Fig. 5.13(c) that are surrounded by similar T_{BD} values, or similar colors in this plot format. The opposite is true for negative correlation. No strong correlation trend is observed, and the global Moran's I for this example was -8.907e-4, indicating negligible spatial correlation of T_{BD}. No significant difference is observed in the results when a larger contiguity matrix is used.

5.4 Conclusions

We have presented a circuit design for the efficient characterization of gate dielectric breakdown. The proposed system consists of a large array of test cells that facilitate the accelerated stressing of the DUTs without significant aging or breakdowns in the supporting circuitry. An A/D current monitor translates the gate current of each device into a convenient 16 bit digital count that is scanned off chip for post processing. Although in the technology used here, we generally only observed hard breakdowns, this design is capable of tracking a progressive decrease in a gate resistance with a high degree of accuracy down to the start of the hard breakdown region. Our automated array-based design would greatly reduce testing times, as up to thousands of samples are needed to correctly define the statistical characteristics of TDDB. Specifically, when compared with individual device probing, our proposed system can cut the test time down by a factor proportional to the number of devices under test, since all of these transistors are stressed in parallel in our circuit. A range of test chip measurements from a 32x32 array implemented in a 1.2 V, 130 nm bulk CMOS process were presented to demonstrate the functionality and flexibility of this design.

Chapter 6

A Flexible Array-Based Test Circuit for Inversion or Off-State TDDB Characterization

6.1 Introduction to Off-State TDDB

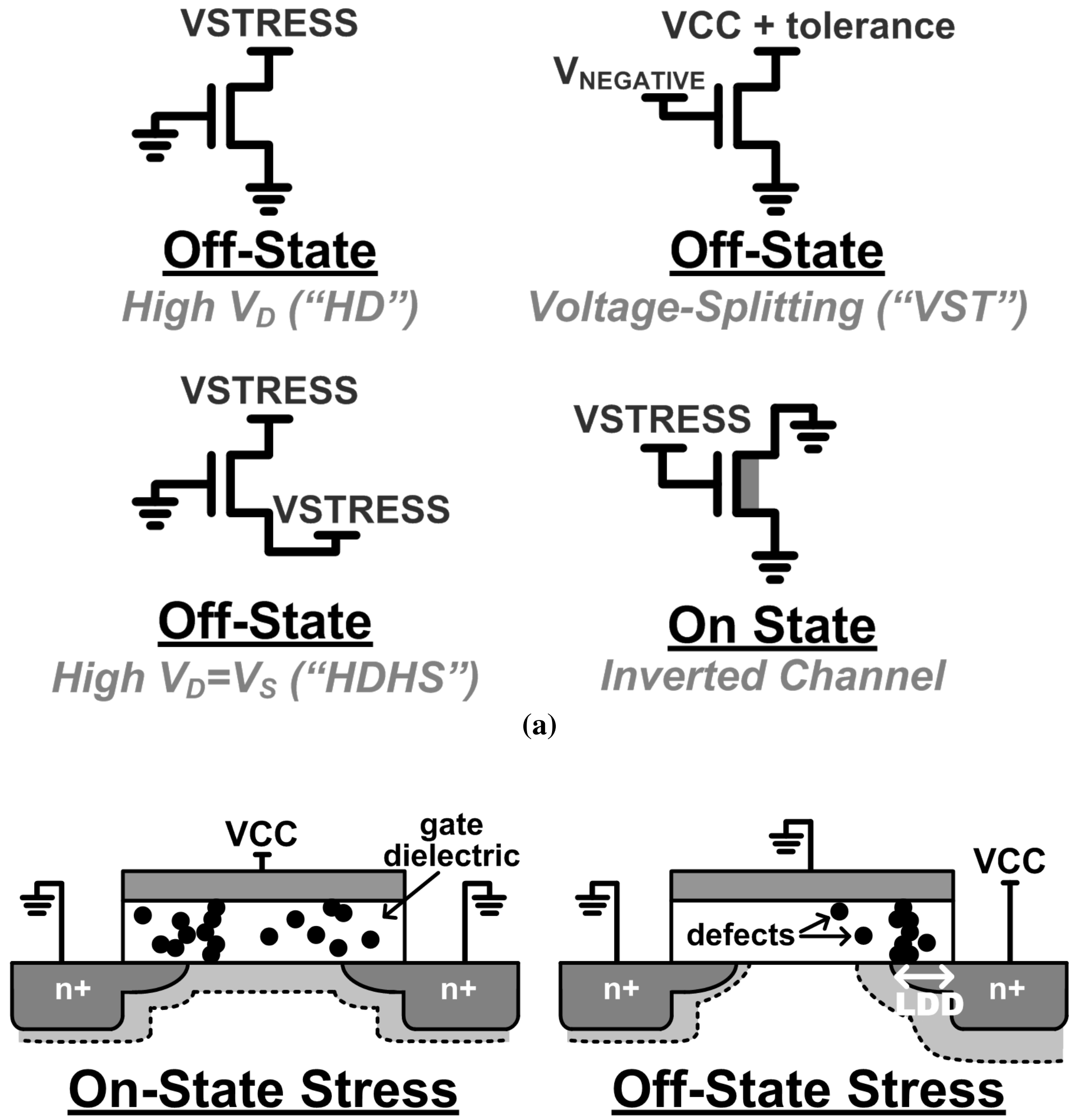
As gate dielectric thicknesses are scaled down to the range of just several atomic layers, they become more susceptible to reliability mechanisms that threaten their insulating properties. While time dependent dielectric breakdown (TDDB) in transistor gates has traditionally been studied under inversion-mode stress conditions, ultra-thin dielectrics can also suffer breakdowns in the off-state when the channel is not inverted [22], [74], [75]. This off-state stress becomes particularly problematic under excessively high drain biases, such as those occurring during burn-in screening, or in certain I/O circuits where a transition is made into a higher voltage domain.

Large voltages on the drain accelerate the "intrinsic" breakdown process we generally observe in inversion (on-state) mode by activating hot carrier injection (HCI) from the source, as well as gate-induced drain leakage (GIDL). Both of these mechanisms have been found to contribute to the defect generation of TDDB [22], [76], [77]. The HCI component becomes a more significant problem when channel lengths are scaled down, leading to increased lateral electric fields and the possibility of punchthrough.

In addition to realistic situations in which high drain biases might be found in modern circuits, test engineers must also be aware of the effects of this bias in accelerated stress tests. Since unrealistically high voltages on a transistor's drain in off-state lead to additional damage from HCI and GIDL, one cannot make accurate lifetime reliability projections for off-state TDDB based on this simple stress configuration [22], [75], [76]. In order to address this problem, Wu *et al.* proposed a "voltage-splitting technique" (VST) which they claim results in only intrinsic TDDB stress, while still facilitating fast stress test times [22].

Fig. 6.1(a) illustrates high drain off-state stress, as well as the voltage-splitting technique, and two of the other stress modes facilitated by our test circuit. Fig. 6.2(b) shows that while breakdown in the on-state mode can take place throughout the gate area, off-state is focused in the gate-drain overlap region. The VST configuration in Fig. 6.1(a) utilizes a negative gate voltage, and a roughly nominal supply voltage on the drain of the device under test (DUT). More specifically, the drain bias is higher than the supply voltage by only the amount safely tolerated by the technology, which is generally around 10%. This creates accelerated stress conditions between the drain and gate without inducing the excess HCI and GIDL that comes with high drain voltages. Note that the effects of the accumulation mode stress due to the negative gate voltage on the channel outside of the drain overlap region were found to be negligible by comparison.

This is because tunneling current depends exponentially on voltage, which is lower away from the drain, and only linearly on area.



(a)

(b)

Fig. 6.1: (a) Schematic views of on- and off-state stress configurations. (b) Cross sections with lightly doped drain (LDD) overlap regions included. Gate dielectric stress takes place throughout the full gate area during on-state stress, but only in the overlap area in off-state.

Wu *et al.* verified that VST results were due to the intrinsic breakdown process in a number of ways. For one, they found that the Weibull time-to-breakdown (T_{BD}) curves under VST were well behaved, or straight, compared with a distorted characteristic for high drain stress in short channel devices. They also compared the charge-to-breakdown findings (calculated by integrating the gate current during stress) from inversion mode and VST, after normalizing the latter results to the total gate area that is stressed in inversion. This analysis showed that the normalized charge-to-breakdown values for both configurations were nearly identical, leading them to conclude that any differences between those breakdown processes are only due to the effective areas being stressed and the total tunneling current.

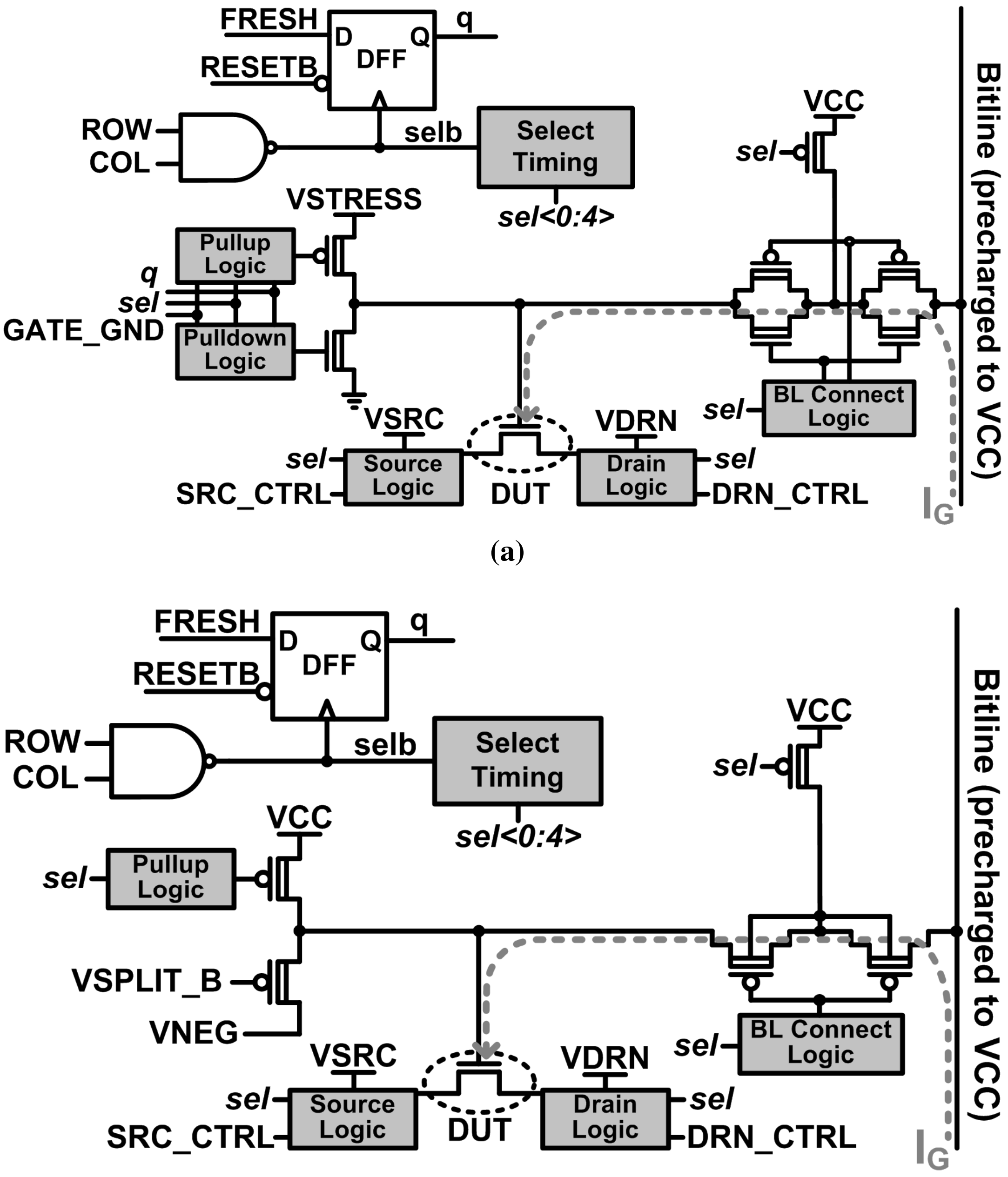
In this work, we build upon our previous array-based test circuit used for the efficient statistical characterization of only inversion mode TDDB, by implementing two new stress cell designs that facilitate inversion mode as well as several off-state tests. One of the cells in the new array allows us to perform VST experiments so that we can easily test the intrinsic breakdown process in the off-state. In addition, we use a simple binary (i.e., two state) measurement setup in this new design, unlike the measurement circuit included in the previous test circuit which was capable of tracking progressive decreases in DUT gate resistance. The new, faster measurement block is used only to indicate when the gate resistance of each device under test has fallen below a user-defined breakdown threshold level. Finally, “calibration cells” that are identical to stress cells are included directly within the array, so that the calibration routine takes into account all of the parasitics and leakage currents that affect real measurements. This updated test circuit provides the flexibility required to gather TDDB statistics under a range of stress configurations, and with a simple digital interface.

6.2 Flexible TDDB Characterization Array Design

The general structure of this new test array is identical to that described in Chapter 5 (Section 5.2), but updates and improvements will be examined in this section. The new test circuit design consists of a 48x16 array of "stress cells" that contain the DUTs, whose gate currents (I_G) are periodically measured using our binary breakdown measurement block and on-chip control logic. All devices in the cell other than the DUT are thick oxide I/O transistors which will not age appreciably or breakdown during tests aimed at the thin oxide devices. As in the previous design, cells are cycled through automatically without the need to send or decode cell addresses, in order to simplify the logic and attain faster measurement times. A single external clock signal is asserted each time that the controlling software is ready for a new I_G measurement.

6.2.1 General Stress Cell Design

The first type of stress cell included in our updated TDDB characterization array is simply called the "general cell" because it facilitates either on- or off-state stress (Fig. 6.2(a)). The DUT's gate terminal can be held at VSTRESS or GND, while the drain and source are independently controlled and can be at either of those values, or floating. The body connection is tied to GND. This enables standard on-state inversion stress, or various off-state modes with the exception of voltage-splitting.



(a)

(b)

Fig. 6.2: (a) The general stress cell design facilitates standard on-state inversion stress, or various off-state modes, with the exception of voltage splitting. (b) The voltage splitting stress cell facilitates its namesake experiments by allowing us to drive a negative voltage on the DUT gate while independently controlling the source and drain terminals. It can also be used for other off-state tests when the gate is grounded.

We also added improved timing logic to correctly control the selection process in the cells in a manner that prevents over-shoot transients on the DUT and other undesirable events. For example, this logic ensures that the DUT gate is held at the steady VCC level before the transmission gates to the bitline are turned on for a measurement to prevent erroneous results, and that those gates are fully off before stress is reapplied so that stress voltages are contained within the cells.

6.2.2 Voltage-Splitting Stress Cell Design

The second type of stress cell in this circuit is called the “voltage-splitting cell,” because it allows us to drive a negative voltage on the gate so that we can perform VST stress experiments (Fig. 6.2(b)). Due to the negative voltage on the DUT gate here, we can only use PMOS devices on all adjacent nodes so that we do not forward-bias any PN junction diodes. (That is, if a voltage of roughly -0.7 V or lower is driven on the drain or source of an NMOS, the grounded p-type body to n-type diffusion junction would be forward biased.) The use of a PMOS driver limits the negative gate voltage we can achieve since PMOS conduct weak low voltages. Extracted simulations showed that the lowest value we could drive through the stressing PMOS was -2.7 V, with the driver’s own gate (VSPLIT_B) held at -4 V. Therefore, the positive drain bias applied during stress had to make up the difference between 2.7 V and the total magnitude of the voltage drop (e.g., V_D = 1.8 V for a 4.5 V stress experiment).

This means that the drain bias was slightly higher than the desired level specified in Section 6.1 (VCC + tolerance ≈ 1.32 V), but still well below the large values required to cause excessive HCI and GIDL [22]. Note that we can also set the gate of the DUTs in these voltage-splitting cells at GND, and thereby perform other types of off-state experiments. The selection timing logic in this cell charges the DUT gate back up to VCC before turning on the pass gates for a measurement, so that a breakdown is not incorrectly recorded from a fresh DUT.

6.2.3 Binary Breakdown Measurement Block

The measurement block included in our original inversion-mode TDDB characterization array was designed to monitor the progressive breakdown processes, meaning the gradual decrease in DUT gate resistance. In this work we present a simpler design with a two state output that only indicates whether a measured DUT's gate resistance has fallen below a user-defined breakdown level (Fig. 6.3). This was done for several reasons.

First, this circuit is implemented in the same technology as our original TDDB array, in which we found that sudden hard breakdowns were more prevalent than slowly progressing breakdowns. Therefore, using the new binary (i.e., two state) approach generally provides the same amount of information—simply the time to the sudden breakdown. Second, the one bit result can be recorded by a data acquisition board more quickly than the sixteen bit result that was scanned out in the previous design. This improves the timing resolution of the measurements, meaning there is less time between consecutive readings in each cell. Third, many researchers base their TDDB findings upon the time to the first observed breakdown—be that soft or hard [22], [78]. This compact system is sufficient to record that first event. However, if one prefers to track progressive breakdowns, the previous design has already been described in detail, and can be utilized or expanded upon.

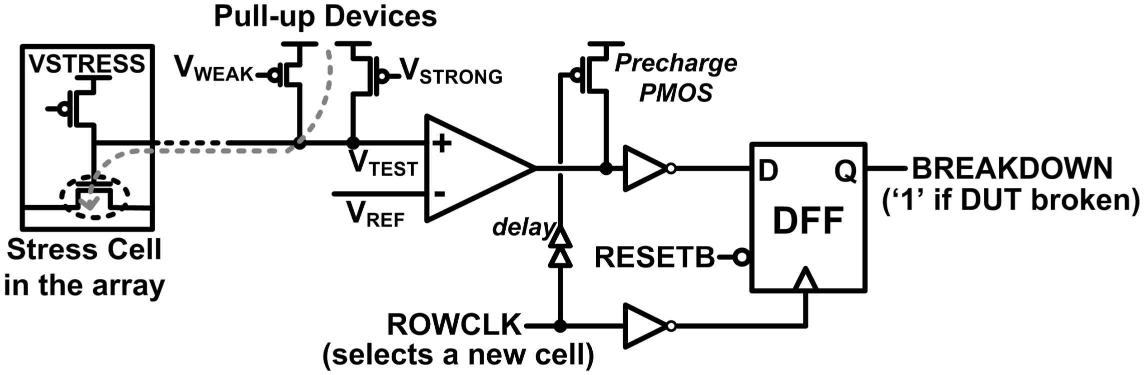


Fig. 6.3: "Binary" (i.e., two state) breakdown measurement setup. A resistive divider is formed between the pullup device(s) and the DUT. If $V_{TEST} < V_{REF}$, the comparator output falls, and BREAKDOWN = '1' after ROWCLK drops to end a cell measurement.

The new measurement block, pictured in Fig. 6.3, forms a resistive divider between one or both of its two pull-up devices and the gate resistance of the selected DUT. If the V_{TEST} node voltage falls below the V_{REF} bias, then the precharged comparator output drops, and this change is latched as a '1' in a DFF when the ROWCLK signal falls. That '1' output indicates that a breakdown has occurred. The level at which this breakdown is triggered is set by the strength of the pull-up device(s). The "strong" device biased by V_{STRONG} has a wide channel, and hence a low resistance, so it can hold the V_{TEST} node above V_{REF} even as the DUT's gate resistance drops to relatively low values. The "weak" device has a narrower channel, and is used to set higher breakdown trip points because of its larger source-to-drain voltage drop. The exact breakdown point is modulated by the pull-up device gate biases, which are determined during circuit calibration. Note that any number of pull-ups can be implemented and then used in parallel or alone to cover the breakdown resistance values targeted by an experiment.

6.2.4 Embedded Calibration Cells and Measured Characteristics

In our previous inversion mode TDDB array, calibration was completed by gating the measurement block off from the array, and attaching a range of known resistance values to a reference current path. Measurements were run as they would normally, allowing us to match each known resistance value with a result output. However, this calibration routine did not take into account the different leakage currents or the parasitic resistance and capacitance that affected real stress experiment measurements within the array itself. This led to errors, as explained in Section 5.3.1.

In this work, we avoided that calibration error by embedding replica stress cells, called "calibration cells" directly in the TDDB array (Fig. 6.4). These calibration cells were identical to the stress cells, but they did not have DUTs. Instead, a metal interconnect path was routed from the DUT gate node out to a pad. During calibration, a known range of resistances were attached to that pad in order to mimic a range of DUT gate resistances, and measurements were run in the calibration cell. The pull-up bias values were swept for each resistance during calibration, in order to find the bias at which a breakdown would be indicated by the measurement block (Fig. 6.5).

For example, with a 2.04 MΩ resistance in the general stress cell design, V_{WEAK} biases below 0.957 V (with V_{STRONG} off at 1.2 V) held the V_{TEST} node above V_{REF}, so no breakdown was indicated. This is because the sufficiently low PMOS pull-up biases kept that device's resistance low. However, once V_{WEAK} was raised to 0.957 V or above, V_{TEST} dropped below V_{REF}, so a high value would be latched on the BREAKDOWN output bit. (Note that the exact values sometimes varied between different chips.) The pull-up bias values were swept through multiple times for each resistance value during calibration, and the results were averaged to eliminate measurement error. The standard deviation in the trip point biases ranged from less than 0.5 mV at higher biases corresponding to breakdowns in the MΩ range, to roughly 2.5 mV at harder breakdown level of 76 kΩ in the general cell.

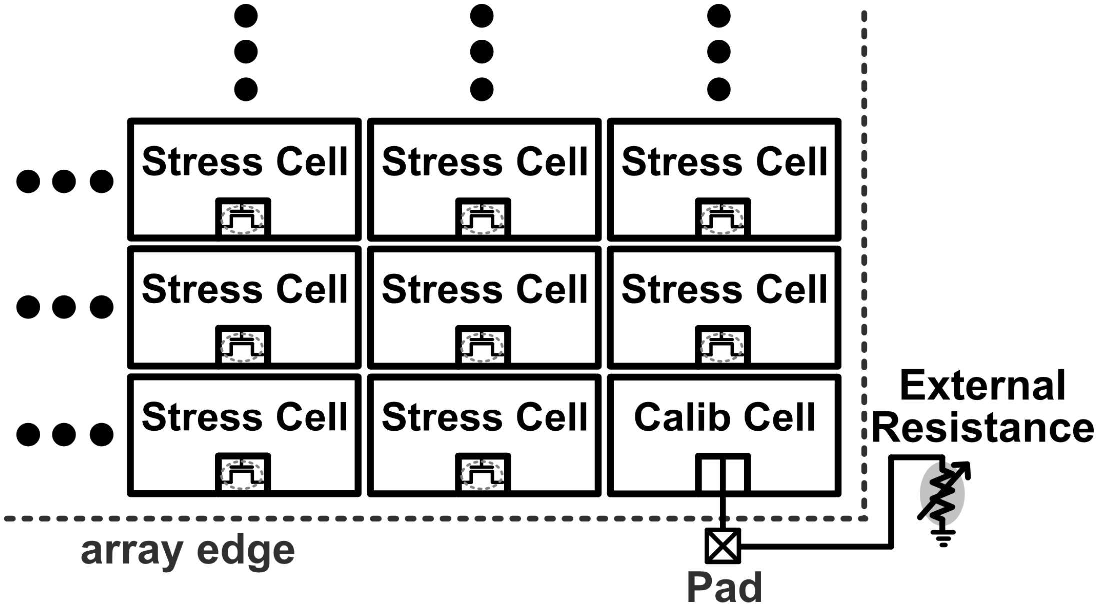
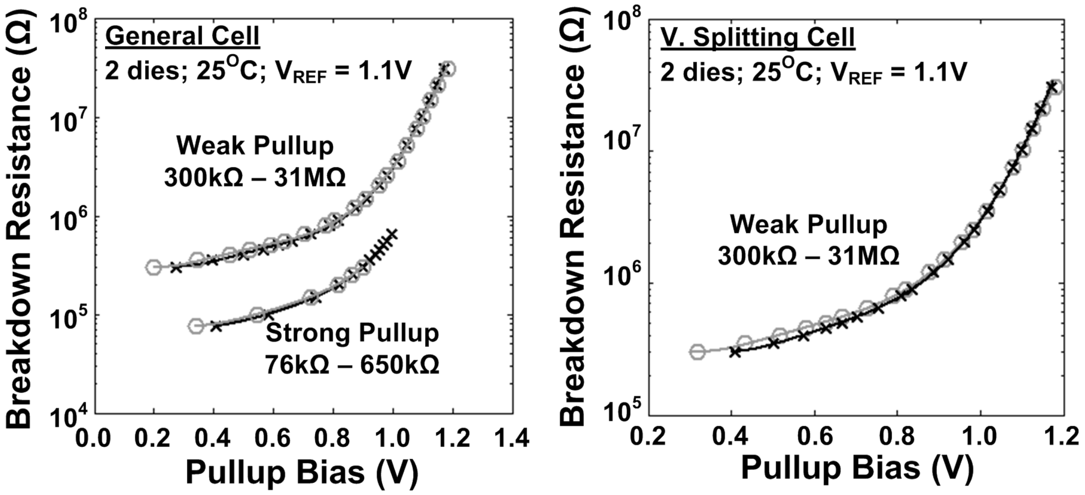


Fig. 6.4: Calibration cells were embedded within the test array. These cells are identical to the stress cells, but the DUT is removed, and a metal interconnect path is routed from the normal DUT gate node to a pin. Including the replica calibration cell within the test array leads to a more representative calibration procedure that captures the effects of the parasitic RC values and leakage currents in that structure.

Calibration cells were included for both the general voltage-splitting cells. Example measured results from the calibration routine on two dies are presented in Fig. 6.5. In the general cell, we were able to measure breakdowns at levels ranging from 76 kΩ to 31 MΩ, with the lowest values being covered by the stronger pull-up device. A range of 300 kΩ to 31 MΩ was measured by the weaker pull-up device for the voltage-splitting cell, but notice the corresponding V_{WEAK} bias levels are higher than those for the general cell. This is because PMOS pass gates are used to access the DUTs in the voltage-splitting cells, rather than transmission gates, so there is more resistance on the measurement path (Section 6.2.2). This extra resistance holds the V_{TEST} node higher unless the measurement block's PMOS pull-up resistance is also increased. Lower breakdown resistance values could not be read reliably with the VST cells because, again, the PMOS passgates drive weak low values, such as those found on the DUT gate node after a hard breakdown. Therefore, the strong pull-up device was not needed for these cells.



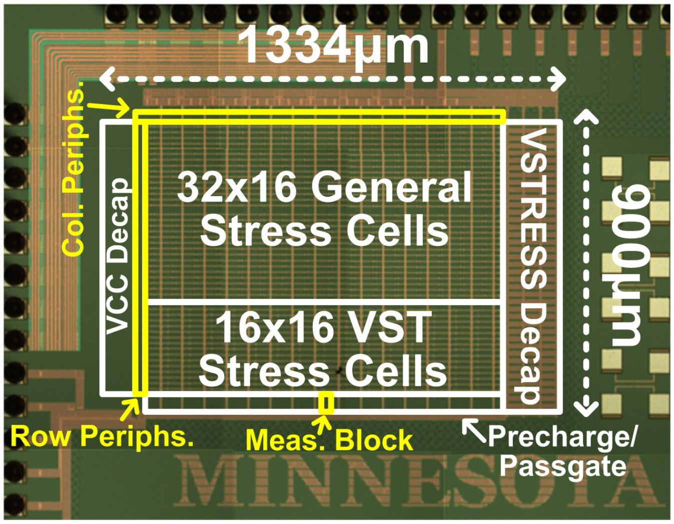
(a) **(b)**

Fig. 6.5: Measured calibration curves using the circuit from Fig. 6.3 and the calibration cells explained in Fig. 6.4.

6.3 Flexible TDDB Array Test Chip Measurements

A test chip was fabricated in a 1.2 V, 130 nm bulk CMOS process. Each NMOS device under test had a width and length of 2 µm in order to be consistent with our previous TDDB array. However, future studies of off-state TDDB should include shorter channel lengths as well, since that parameter strongly impacts the degradation characteristics with high drain stress. As stated earlier, HCI and the lateral field component of GIDL both enhance TDDB in short channel devices [22].

Information about the gate dielectric construction is kept confidential by the manufacturer, but the thickness is within a reasonable range for this technology node. Automatic measurements were completed with LabVIEW® software and a National Instruments data acquisition board, which was connected to a laptop through a USB port. A microphotograph of the chip and a summary of the circuit characteristics are shown in Fig. 6.6.

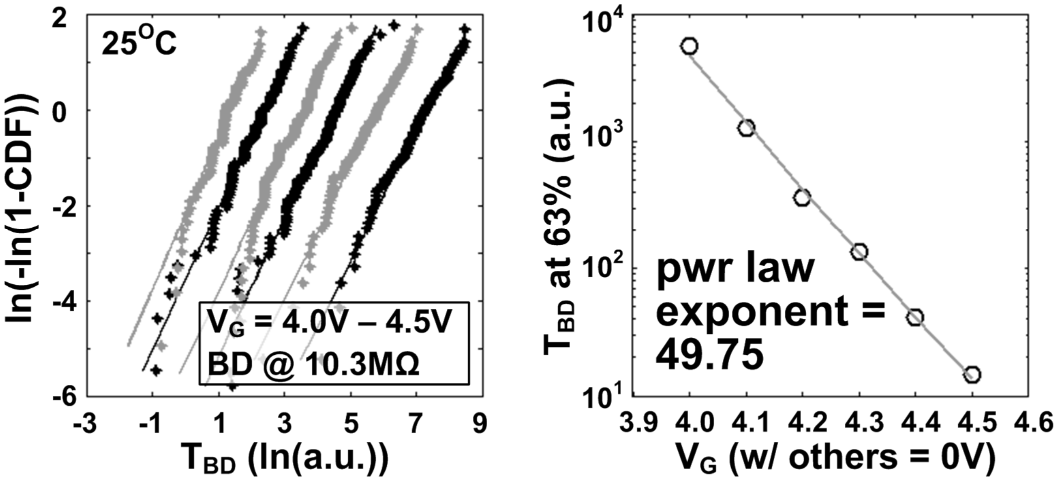


Process	130nm bulk CMOS, 8M
Logic / I/O supplies	1.2V / 3.3V
Total Area	1334 x 900µm^2
Gen. Cell Meas. Range	76kΩ - 31MΩ
VST Cell Meas. Range	300kΩ - 31MΩ
NMOS DUT dimensions	2µm x 2µm

Fig. 6.6: Die photo and test chip characteristics.

6.3.1 Inversion Mode Stress Results

We first measured T_{BD} in inversion mode. A high breakdown resistance point of 10.3 MΩ was chosen to detect breakdowns early in the degradation process in case any progressive TDDB takes place. Fig. 6.7(a) shows that the Weibull CDFs of these on-mode results were well-behaved for stress voltages ranging from 4.0 V to 4.5 V, as expected. The Weibull slope (β) for the 4.2 V curve was 1.444, matching well with the 1.443 value from our previous work. Again, this value slightly increased with higher stress voltages and decreased at lower stresses, as explained in Section 5.3.2. The power law exponent of 50 for the voltage acceleration in Fig. 6.7(b) also matches well with the value of 51 found in the original TDDB array.



(a) (b)

Fig. 6.7: Standard inverstion (i.e., "on-state") stress results.

In Fig. 6.8, we see the effects of setting a harder (i.e., lower) breakdown resistance threshold. The hard breakdown curves display a bend early in their evolution, and then a low Weibull slope if only the points after that bend are fitted. Tous et al. explained that while the time to first breakdown and the progressive breakdown times follow Weibull statistics, the time to final failure (a convolution of those two times) does not [58]. The authors provide a theoretical basis for a bend in the time-to-final failure's characteristics on a Weibull plot which may explain our results.

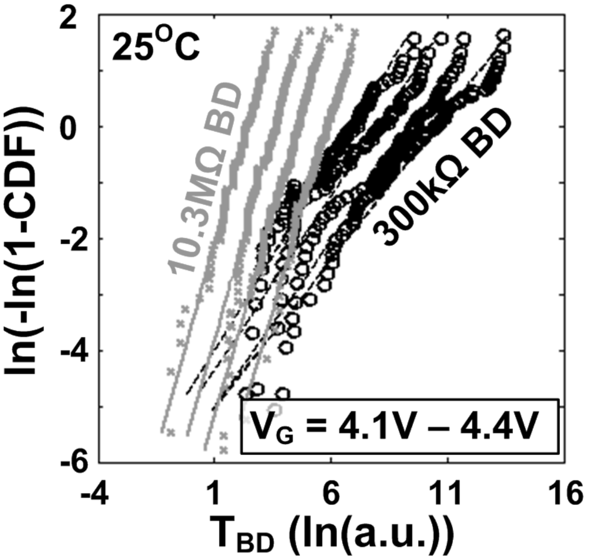


Fig. 6.8: Inversion-mode stress to a "soft" breakdown trigger (10.3MΩ), and a harder breakdown (300kΩ). There is a distinct bend in the hard breakdown CDFs, with a faster breakdown process acting early on, followed by a slow-down.

6.3.2 Off-State Stress Results

In Fig. 6.9, we compare off-state high drain results (all other terminals at 0 V, called "HD" stress), with those from high drain *and* source experiments (HDHS). The latter display an earlier T_{BD} because twice the area in each DUT gate is stressed in this case (i.e., the source and drain overlap regions). Weibull processes such as dielectric breakdown follow Poisson area scaling as described in Section 5.3.4, so we can use that equation to calculate the expected ratio of characteristic life parameters for both distributions (i.e., the time at which 63% of the devices have failed, denoted by α) as follows using the 4.6 V results:

$$\frac{\alpha_{HD}}{\alpha_{HDHS}} = \left(\frac{AREA_{HDHS}}{AREA_{HD}} \right)^{1/\beta} = (2)^{1/1.505} = 1.585$$

Note that the β values for these offstate stress conditions were slightly higher than those measured in inversion, and the value used in this equation is the average of those found for HD and HDHS. The actual characteristic life ratio from our results is 1.597, which matches fairly well with the theoretical value. Note that the actual α and T_{BD} values must be kept confidential according to the manufacturer.

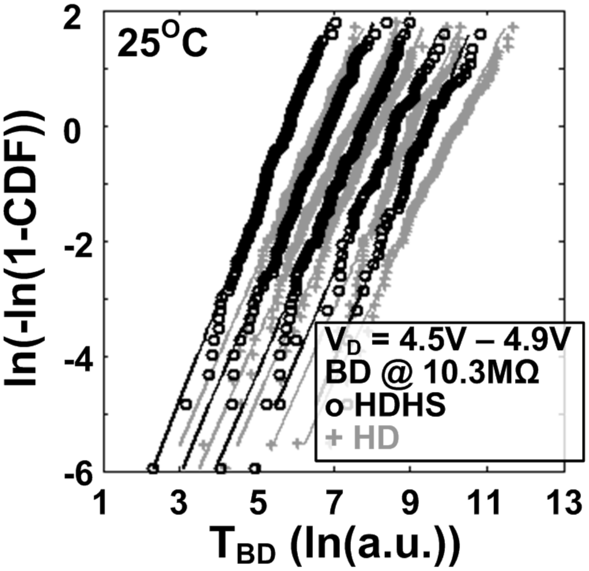


Fig. 6.9: Off-state high drain/0 V source (HD) and highdrain/high source (HDHS) Weibull plots.

Fig. 6.10 illustrates VST results, along with the high drain stress findings. Although our DUTs are long-channel devices, so the lateral electric field does not degrade the HD stress results with respect to those from VST, the former conditions still result in faster breakdowns possibly due to a vertical field contribution from GIDL. We also observed a lower β value for VST compared with HD results (e.g., 1.03 versus 1.48 at 4.6V stress). This was not expected based on Wu's work, and one possible explanation is that he tracked the time to the *first* breakdown—be that soft or hard. They may have used sensitive lab equipment to detect the individual breakdown events, so gate resistances of even higher than 10.3 MΩ were used to indicate the onset of TDDB. This is also possible in our array-based system, particularly if a very weak pull-up device is implemented, but would need to be investigated further in future work.

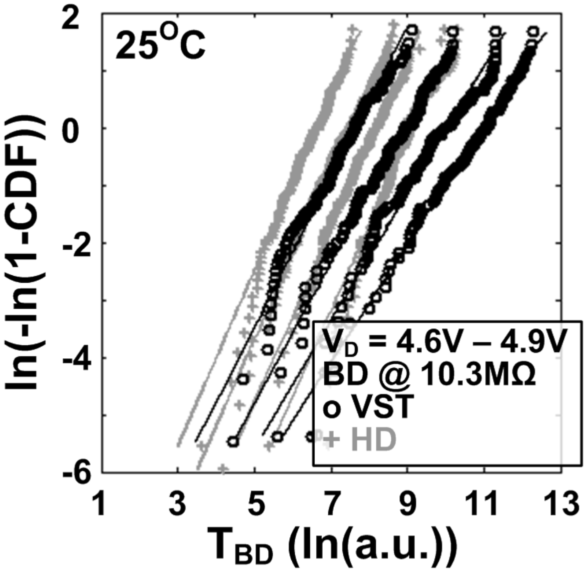


Fig. 6.10: Off-state voltage-splitting (VST) and high drain/0 V source (HD) Weibull plots.

In Fig. 6.11, we compare the voltage acceleration characteristics for several off-state stress configurations. The HDHS had the lowest T_{BD} due to stress on both ends of the channel. Simple HD stress was very similar to the case in which we floated the source (rather than setting it to 0 V), which is to be expected in long channel devices. Floating the source (HDFS) eliminates the impact of the lateral electric field in the degradation process, but the contribution of that field to the TDDB is negligible when channel lengths are sufficiently long. The VST results show the longest T_{BD}. This is again presumably due to the elimination of GIDL-induced degradation.

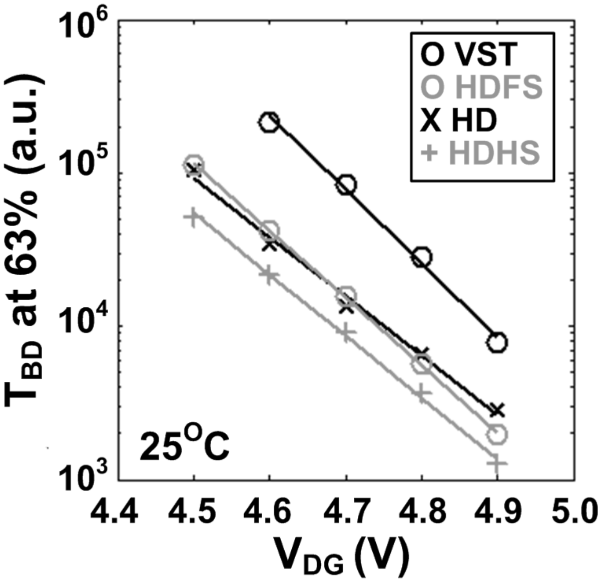


Fig. 6.11: Off-state voltage acceleration plots.

The power law exponent for the HD stress voltage acceleration was 41.63, while that of VST was 52.43. Wu et al. found lower exponents for VST stress conditions than inversion mode, so this latter value was expected to be smaller than the 49.75 shown in Fig. 6.7(b). More work is needed to verify the precise behavior of off-state degradation's relationship with voltage. Finally note that the T_{BD} for all off-state stress conditions was around 4 orders of magnitude higher than that seen in inversion mode (Fig. 6.7(b)) at

4.5 V stress. This gap is also larger than that found in the original VST work. However, we are using a different technology which could result in significantly improved off-state reliability. For example, the gate oxide thickness could be thicker at the edges or the drain overlap region may be shorter in this technology, leading to longer off-state T_{BD} [22], [74].

6.4 Conclusions

In this work, we implemented a more flexible TDDB characterization array design that facilitates inversion mode as well as several off-state stress tests. Calibration cells were embedded within the array in order to accurately replicate real measurements when characterizing this system. A simple measurement block with a one bit output indicating whether a selected DUT gate resistance has fallen below a user-defined breakdown resistance level was presented. Measured results indicate that voltage-splitting stress lead to longer T_{BD} values compared with high drain stress, even in long channel devices, due to the elimination of GIDL-induced degradation. Off-state T_{BD} values were several orders of magnitude higher than those from inversion-mode stress in the devices tested here. However, breakdowns are still clearly observed in the off-state, so this must be considered in cases where devices are left “off” for long periods of time, or high drain biases are applied. The array-based system presented here enables fast and efficient statistical measurements without expensive probing stations or other test equipment, and can be utilized in future studies of either on- or off-state TDDB degradation.

Chapter 7

Conclusions

The parametric shifts or circuit failures caused by Hot Carrier Injection (HCI), Bias Temperature Instability (BTI), and Time Dependent Dielectric Breakdown (TDDB) have become more severe with shrinking device sizes and voltage margins. These mechanisms must be studied in order to develop accurate reliability models, which are used to design robust circuits. Another option for addressing aging effects is to use on-chip reliability monitors that can trigger real-time adjustments to compensate for lost performance or device failures. For example, the system clock could be slowed down as performance degrades, rather than adding a large frequency guardband at the beginning of the circuit's life. The need for efficient technology characterization and aging compensation is exacerbated by the rapid introduction of process improvements, such as high-k/metal gate stacks and stressed silicon.

In this thesis I presented five unique test chip designs that we have implemented in order to demonstrate the benefits of utilizing on-chip logic and a simple test interface to automate transistor aging characterization experiments. In addition to avoiding the use of expensive probing equipment, implementing on-chip logic to control the measurements enables much better timing resolution. This is critical when interrupting stress to record NBTI measurements, as this mechanism is known to recover within microseconds or less. We will also saw that the Silicon Odometer beat frequency detection system allows us to measure ring oscillator frequency shifts with resolution ranging down to a theoretical limit of less than 0.01%. That mix of speed and accuracy is not possible with standard off-chip equipment.

In Chapter 2, the Odometer framework was used in conjunction with a custom ROSC design to separate the effects of HCI and BTI stress. High voltage experiments also allowed us to observe the impact of TDDB on those ROSCs. Chapter 3 introduced an array of ROSCs whose frequency degradation due to BTI and HCI was monitored with a small set of Odometers. That system facilitated efficient statistical measurements, and the results indicated that variations in the aging mechanisms themselves must be considered during technology characterization.

A DLL-based test circuit for measuring NBTI was presented in Chapter 4. In that design, the stress-induced threshold shift in PMOS headers was translated into a VCDL control voltage shift for an average sensing gain of 10X. As with our other BTI measurements, results from this test circuit showed that the previously proposed BTI degradation power law exponents of 1/6 and 1/4 were incorrect, and resulted from excessively slow measurement techniques.

Next, in Chapters 5 and 6 we presented two test chip designs for the efficient characterization of gate dielectric breakdown. The proposed systems each consist of a large array of stress cells that facilitate the accelerated stressing of the DUTs without significant aging or breakdowns in the supporting circuitry. The first implementation performed on-state (i.e., strong inversion) TDDB experiments, while the latter also allowed us to test several off-state breakdown modes. Our automated array-based designs greatly reduce testing times, as up to thousands of samples are needed to correctly define the statistical characteristics of TDDB. Specifically, when compared with individual device probing, our proposed system can cut the test time down by a factor proportional to the number of devices under test, since all of these transistors are stressed in parallel in our circuits.

Taking advantage of these benefits while obtaining accurate CMOS aging information with on-chip circuits would allow manufacturers to avoid excessively wasteful overdesign and frequency guardbanding based on pessimistic degradation projections. The resulting performance improvements will become increasingly valuable as traditional CMOS scaling slowly grinds to a halt in the coming years. We expect this to lead to continued interest in the work presented here, as researchers strive to develop more reliable transistors and complex digital systems.

Bibliography

[1] H. Kufluoglu, “MOSFET Degradation Due to Negative Bias Temperature Instability (NBTI) and Hot Carrier Injection (HCI) and its Implications for Reliability Aware VLSI Design,” Ph.D. dissertation, Purdue University, West Lafayette, IN, U.S.A., 2007.

[2] D. Ielmini, M. Manigrasso, F. Gattel, and M. G. Valentini, “A New NBTI Model Based on Hole Trapping and Structural Relaxation in MOS Dielectrics,” *IEEE Trans. On Electron Devices*, vol. 56, no. 9, pp. 1943-1952, September 2009.

[3] T. Grasser and B. Kaczer, “Evidence that Two Tightly Coupled Mechanisms are Responsible for Negative Bias Temperature Instability in Oxynitride MOSFETs,” *IEEE Trans. on Electron Devices*,” vol. 56, no. 5, pp. 1056-1062, May 2009.

[4] R. Degraeve, M. Aoulaiche, B. Kaczer, P. Roussel, T. Kauerauf, S. Sahhaf, and G. Groeseneken, “Review of Reliability Issues in High-k/Metal Gate Stacks,” *IEEE Int. Symp. on the Physical and Failure Analysis of Integrated Circuits*, pp. 1-6, 2008.

[5] E. Y. Wu, E. Nowak, A. Vayshenker, W. L. Lai, and D. L. Harmon, “CMOS Scaling Beyond the 100-nm Node with Silicon-Dioxide-Based Gate Dielectrics,” *IBM Jour. of Research and Development*, pp. 287-298, March/May 2002.

[6] B. H. Lee, “Unified TDDB Model for Stacked High-k Dielectrics,” *IEEE Int. Conf. on IC Design and Technology*, pp. 83-87, 2009.

[7] S. Kumar, C. Kim, and S. Sapatnekar, "Impact of NBTI on SRAM Read Stability and Design for Reliability", *IEEE Int. Symp. on Quality Electronics Design*, pp. 210-218, March 2006.

[8] G. La Rosa, W. Ng, and S. Rauch, “Impact of NBTI Induced Statistical Variation to SRAM Cell Stability,” *IEEE Int. Reliability Physics Symposium*, pp. 274-282, 2006.

[9] V. Huard, C. Parthasarathy, C. Guerin, T. Valentin, E. Pion, M. Mammasse, N. Planes, and L. Camus, “NBTI Degradation: From Transistor to SRAM Arrays,” *IEEE Int. Reliability Physics Symp.*, pp. 289-300, 2008.

[10] C. Shen, M.-F. Li, C. Foo, T. Yang, D. Huang, A. Yap, G. Samudra, and Y.-C. Yeo, “Characterization and Physical Origin of Fast Vth Transient in NBTI of pMOSFETs with SiON Dielectrics,” *IEEE Electron Devices Meeting*, pp. 1-4, December 2006.

[11] T. H. Kim, R. Persaud, and C. H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits,” *IEEE Jour. of Solid-State Circuits*, vol. 43, no. 4, pp. 874-880, April 2008.

[12] M. Denais, C. Parthasarathy, G. Ribes, Y. Rey-Tauriac, N. Revil, A. Bravaix, V. Huard, and F. Perrier, "On-the-fly Characterization of NBTI in Ultra-Thin Gate Oxide PMOSFET's," *IEEE Int. Electron Devices Meeting*, pp. 109-112, December 2004.

[13] M. Denais, A. Bravaix, V. Huard, C. Parthasarathy, C. Guerin, G. Ribes, F. Perrier, M. Mairy, and D. Roy, "Paradigm Shift for NBTI Characterization in Ultra-Scaled CMOS Technologies," *IEEE Int. Reliability Physics Symp*., pp. 735-736, March 2006.

[14] T. Grasser, W. Gös, V. Sverdlov, and B. Kaczer, "The Universality of NBTI Relaxation and Its Implications for Modeling and Characterization," *IEEE Int. Reliability Physics Symp*., pp. 268-280, April 2007.

[15] E. Karl, P. Singh, D. Blaauw, and D. Sylvester, "Compact In-Situ Sensors for Monitoring Negative-Bias-Temperature-Instability Effect and Oxide Degradation," *IEEE Int. Solid-State Circuits Conf*., pp. 410-411, 2008.

[16] P. Singh, Z. Foo, M. Wieckowski, S. Hanson, M. Fojtik, D. Blaauw, and D. Sylvester, "Early Detection of Oxide Breakdown through In Situ Degradation Sensing," *IEEE Int. Solid-State Circuits Conf*., pp. 190-191, 2010.

[17] E. Saneyoshi, K. Nose, and M. Mizuno, "A Precise-Tracking NBTI-Degradation Monitor Independent of NBTI Recovery Effect," *IEEE Int. Solid-State Circuits Conf*., pp. 192-193, 2010.

[18] F. Gebara, J. Hayes, J. Keane, S. Nassif, and J. Schaub, “Delay-Based Bias Temperature Instability Recovery Measurements for Characterizing Stress Degradation and Recovery,” U. S. Patent Application 12/142,294, Filed June 19, 2008.

[19] J. Keane, D. Persaud, and C. H. Kim, “An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI, and TDDB,” *IEEE VLSI Circuits Symp.*, pp. 108-109, 2009.

[20] W. Jiang, H. Le, J. Chung, T. Kopley, P. Marcoux, and C. Dai, “Assessing Circuit-Level Hot-Carrier Reliability,” *IEEE Int. Reliability Physics Symp.*, pp. 173-179, 1998.

[21] K. Quader, E. Minami, W. Huang, P. Ko, and C. Hu, “Hot-Carrier Reliability Design Guidelines for CMOS Logic Circuits,” *IEEE Custom Integrated Circuits Conf.*, pp. 30.7.1-30.7.4, 1993.

[22] E. Wu, E. Nowak, and W. Lai, “Off-State Mode TDDB Reliability for Ultra-Thin Gate Oxides: New Methodology and the Impact of Oxide Thickness Scaling,” *IEEE Int. Reliability Physics Symp.*, pp. 84-94, 2004.

[23] S. Mahapatra, D. Saha, D. Varghese, and P. B. Kumar, “On the Generation and Recovery of Interface Traps in MOSFETs Subjected to NBTI, FN, and HCI Stress,” *IEEE Trans. on Electron Devices*, vol. 53, no. 7, pp. 1583-1592, July 2006.

[24] B. Kaczer, R. Degraeve, M. Rasras, K. Van de Mieroop, P. Roussel, and G. Groeseneken, “Impact of MOSFET Gate Oxide Breakdown on Digital Circuit Operation and Reliability,” *IEEE Trans. on Electron Devices*, pp. 500-506, vol. 49, no. 3, March 2002.

[25] M. Agostinelli, S. Pae, W. Yang, C. Prasad, D. Kencke, S. Ramey, E. Snyder, S. Kashyap, and M. Jones, “Random Charge Effects for PMOS NBTI in Ultra-Small Gate Area Devices,” *IEEE Int. Reliability Physics Symp.*, pp. 529-532, 2005.

[26] S. Pae, J. Maiz, C. Prasad, and B. Woolery, “Effect of BTI Degradation on Transistor Variability in Advanced Semiconductor Technologies,” *IEEE. Trans. on Device Materials and Reliability*, vol. 8, no. 3, pp. 519-525, September 2008.

[27] S. Rauch, “Review and Reexamination of Reliability Effects Related to NBTI-Induced Statistical Variations,” *IEEE Trans. on Device Materials and Reliability*, vol. 7, no. 4, pp. 524-530, December 2007.

[28] B. Kaczer, T. Grasser, J. Martin-Martinez, E. Simoen, M. Aoulaiche, P. Roussel, and G. Groeseneken, “NBTI from the Perspective of Defect States with Widely Distributed Time Scales,” *IEEE Int. Reliability Physics Symp.*, pp. 55-60, 2009.

[29] S. Rauch, “The Statistics of NBTI-Induced VT and β Mismatch Shifts in pMOSFETs,” *IEEE Trans. on Device and Materials Reliability*, vol. 2, no. 4, pp. 89-93, December 2002.

[30] K. Kang, S. Park, K. Roy, and M. Alam, “Estimation of Statistical Variation in Temporal NBTI Degradation and its Impact on Lifetime Circuit Performance,” *IEEE/ACM Int. Conf. on Computer-Aided Design*, pp. 730-734, 2007.

[31] C. Tu, S. Chen, A. Chuang, H. Huang, Z. Jhou, C. Chang, S. Chou, and J. Ko, "Transistor Variability after CHC and NBTI Stress in 90 nm pMOSFET Technology," *IEEE Electronics Letters*, vol. 45, no. 16, pp. 854-856, July 2009.

[32] T. Grasser, P. Wagner, P. Hehenberger, W. Goes, and B. Kaczer, "A Rigorous Study of Measurement Techniques for Negative Bias Temperature Instability," *IEEE Trans. on Device and Materials Reliability*, vol. 8, no. 3, pp. 526-535, September 2008.

[33] H. Reisinger, R.-P. Vollertsen, P.-J. Wagner, T. Huttner, A. Martin, S. Aresu, W. Gustin, T. Grasser, and C. Schlunder, "A Study of NBTI and Short-Term Threshold Hysteresis of Thin Nitrided and Thick Non-Nitrided Oxides," *IEEE Trans. on Device Materials and Reliability*, vol. 9, no. 2, pp. 106-114, June 2009.

[34] W. Wang, V. Reddy, B. Yang, V. Balakrishnan, S. Krishnan, and Y. Cao, "Statistical Prediction of Circuit Aging under Process Variations," *IEEE Custom Integrated Circuits Conf.*, pp. 13-16, 2008.

[35] Z. Ji, J. Zhang, M. Chang, B. Kaczer, and G. Groeseneken, "An Analysis of the NBTI-Induced Threshold Voltage Shift Evaluated by Different Techniques," *IEEE Trans. on Electron Devices*, vol. 56, no. 5, pp. 1086-1093, May 2009.

[36] M.-F. Li, D. Huang, Ch. Shen, T. Yang, W. Liu, and Z. Liu, "Understand NBTI Mechanism by Developing Novel Measurement Techniques," *IEEE Trans. on Device and Materials Reliability*, vol. 8, no. 1, pp. 62-71, March 2008.

[37] Y. Wang, “Effects of Interface States and Positive Charges on NBTI in Silicon-Oxynitride p-MOSFETs,” *IEEE Trans. on Device and Materials Reliability*, vol. 8, no. 1, pp. 14-21, March 2008.

[38] S. Mahapatra and M. Alam, “Defect Generation in p-MOSFETs Under Negative-Bias Stress: An Experimental Perspective,” *IEEE. Trans. on Device and Materials Reliability*, vol. 8, no. 1, pp. 35-46, March 2008.

[39] D. Varghese, D. Saha, S. Mahapatra, K. Ahmed, F. Nouri, and M. Alam, “On the Dispersive versus Arrhenius Temperature Activation of NBTI Time Evolution in Plasma Nitrided Gate Oxides: Measurements, Theory, and Implications,” *IEEE Int. Electron Devices Meeting*, pp. 684-687, 2005.

[40] W. Liu, D. Huang, Q. Sun, C. Liao, L. Zhang, Z. Gan, W. Wong, and M.-F. Li, “Studies of NBTI in pMOSFETs with Thermal and Plasma Nitrided SiON Gate Oxides by OFIT and FPM Methods,” *IEEE Int. Reliability Physics Symposium*, pp. 964-968, 2009.

[41] R. Fernández, B. Kaczer, A. Nackaerts, S. Demuynck, R. Rodriguez, M. Nafria, and G. Groeseneken, “AC NBTI Studied in the 1 Hz – 2 GHz Range on Dedicated On-Chip Circuits,” *IEEE Int. Electron Devices Meeting*, pp. 337-340, December 2006.

[42] J. Maneatis “Low-Jitter Process-Independent DLL and PLL Based on Self-Biased Techniques,” *IEEE Journal of Solid-State Circuits*, vol. 31, no. 11, pp. 1723-1732, November 1996.

[43] H. Chang, J. Lin, C. Yang, and S. Liu, “A Wide-Range Delay-Locked Loop with Fixed Latency of One Clock Cycle,” *IEEE Journal of Solid-State Circuits*, vol. 37, no. 8, pp. 1021-1027, August 2002.

[44] J. Keane, T. Kim, and C. H. Kim, “An On-Chip NBTI Sensor for Measuring PMOS Threshold Voltage Degradation,” *IEEE Int. Symp. on Low Power Electronics and Design*, pp. 189-194, August 2007.

[45] A. Chandrakasan, W. J. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*, New York, NY: IEEE Press, 2001, pp. 235-260.

[46] G. Chen, M. Li, C. Ang, J. Zheng, and D. Kwong, “Dynamic NBTI of PMOS Transistors and its Impact on Device Lifetime,” *IEEE Electron Device Letters*, vol. 23, no. 12, pp. 734-736, December 2002.

[47] T. Yang, M. F. Li, C. Shen, C. Ang, C. Zhu, Y. Yeo, G. Samudra, S. Rustagi, M. Yu, and D. Kwong, “Fast and Slow Dynamic NBTI Components in p-MOSFET with SiON Dielectric and their Impact on Device Lifetime and Circuit Application,” *IEEE Symp. on VLSI Technology*, pp. 92-93, June 2005.

[48] C. Schlunder, W. Heinrigs, W. Gustin, and H. Reisinger, “On the Impact of the NBTI Recovery Phenomenon on Lifetime Prediction of Modern p-MOSFETs,” *IEEE Int. Integrated Reliability Workshop*, pp. 1-4, October 2006.

[49] J. Li, M. Chen, P. Juan, and K. Su. “Effects of Delay Time and AC Factors on Negative Bias Temperature Instability of PMOSFETs,” *IEEE Int. Integrated Reliability Workshop*, pp.16-19, October 2006.

[50] D. Vargese, G. Gupta, L. M. Lakkimsetti, D. Saha, K. Ahmed, F. Nouri, and S. Mahapatra, “Physical Mechanism and Gate Insulator Material Dependence of Generation and Recovery of Negative-Bias Temperature Instability in p-MOSFETs,” *IEEE Trans. on Electron Devices*, vol. 54, no. 7, pp.1672-1680, July 2007.

[51] R. Degraeve, G. Groeseneken, R. Bellens, J. Ogier, M. Depas, P. Roussel, and H. Maes, “New Insights in the Relation Between Electron Trap Generation and the Statistical Properties of Oxide Breakdown,” *IEEE Trans. on Electron Devices*, vol. 45, no. 4, pp. 904-911, 1998.

[52] J. Suñé, E. Wu, and S. Tous, “A Physics-Based Deconstruction of the Percolation Model of Oxide Breakdown,” *Microelectronic Engineering*, vol. 84, issue 9-10, pp. 1917-1920, 2007.

[53] A. Krishnan and P. Nicollian, “Analytical Extension of the Cell-Based Oxide Breakdown Model to Full Percolation and its Implications,” *IEEE Int. Reliability Physics Symp.*, pp. 232-239, 2007.

[54] Y. Lee, N. Mielke, M. Agostinelli, S. Gupta, R. Lu, and W. McMahon, “Prediction of Logic Product Failure Due to Thin-Gate Oxide Breakdown,” *IEEE Int. Reliability Physics Symp.*, pp. 18-28, 2006.

[55] J. Stathis, “Gate Oxide Reliability for Nano-Scale CMOS,” *IEEE Int. Conf. on Microelectonics*, pp. 78-83, 2006.

[56] J. Suñé, E. Wu, and W. Lai, "Statistics of Competing Post-Breakdown Failure Modes in Ultrathin MOS Devices," *IEEE Trans. on Electron Devices*, vol. 53, no. 2, pp. 224-234, 2006.

[57] A. Kerber, "Lifetime Prediction for CMOS Devices with Ultra Thin Gate Oxides Based on Progressive Breakdown," *IEEE Int. Reliability Physics Symp.*, pp. 217-220, 2007.

[58] S. Tous, E. Wu, and J. Suñé, "A Compact Model for Oxide Breakdown Failure Distribution in Ultrathin Oxides Showing Progressive Breakdown," *IEEE Electron Device Letters*, vol. 29, no. 8, pp. 949-951, 2008.

[59] L. Pang and B. Nikolic, "Impact of Layout on 90nm CMOS Process Parameter Fluctuations," *IEEE Symp. on VLSI Circuits*, pp. 69-70, 2006.

[60] K. Agarwal, F. Liu, C. McDowell, S. Nassif, K. Nowka, M. Palmer, D. Acharyya, and J. Plusquellic, "A Test Structure for Characterizing Local Device Mismatches," *IEEE Symp. on VLSI Circuits*, pp. 67-68, 2006.

[61] M. Röhner, A. Kerber, and M. Kerber, "Voltage Acceleration of TBD and its Correlation to Post Breakdown Conductivity of N- and P-Channel MOSFETs," *IEEE Int. Reliability Physics Symposium*, pp. 76-81, 2006.

[62] M. Nafria, D. Yelamos, J. Suñé, and X. Aymerich, "Frequency Dependence of Degradation and Breakdown of Thin SiO2 Films," *Quality and Reliability Engineering Int.*, vol. 11, no. 4, pp. 257-261, 1995.

[63] E. Rosenbaum and C. Hu, "High-Frequency Time-Dependent Breakdown of SiO_2," *IEEE Electron Device Letters*, vol. 12, no. 6, pp. 267-269, June 1991.

[64] D. Wolters and J. Verwey, "Breakdown and Wear-Out Phenomena in SiO2 Films," in *Instabilities in Silicon Devices*. Amsterdam, The Netherlands: Elsevier, 1986, ch. 6.

[65] E. Wu, A. Vayshenker, E. Nowak, J. Suñé, R. Vollertsen, W. Lai, and D. Harmon, "Experimental Evidence of T_{BD} Power-Law for Voltage Dependence of Oxide Breakdown in Ultrathin Gate Oxides," *IEEE Trans. on Electron Devices*, vol. 49, no. 12, pp. 2244-2253, December 2002.

[66] D. DiMaria and J. Stathis, "Non-Arrhenius Temperature Dependence of Reliability in Ultrathin Silicon Dielectric Films," *Applied Physics Letters*, vol. 74, no. 12, pp. 1752, 1999.

[67] B. Kaczer, R. Degraeve, N. Pangon, and G. Groeseneken, "The Influence of Elevated Temperature on Degradation and Lifetime Prediction of Thin Silicon-Dioxide Films," *IEEE Trans. on Electron Devices*, vol. 47, no. 7, pp. 1514-1521, 2000.

[68] K. Chopra, C. Zhuo, D. Blaauw, and D. Sylvester, "A Statistical Approach for Full-Chip Gate-Oxide Reliability Analysis," *IEEE/ACM Int. Conf. on Computer-Aided Design*, pp. 698-705, 2008.

[69] L. Anselin, "Local Indicators of Spatial Association—LISA," *Geographical Analysis*, vol. 27, no. 2, pp. 93-115, 1995.

[70] F. Hebeler, "Moran's I," Internet: http://www.mathworks.com/matlabcentral/fileexchange/13663, June 20, 2007 [Nov. 10, 2008].

[71] G. Box and D. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society*. Series B (Methodological), vol. 26, no. 2, pp. 211-252, 1964.

[72] R. Randles, M. Flinger, G. Policello, and D. Wolfe, "An Asymptotically Distribution-Free Test for Symmetry Versus Asymmetry," *Journal of the American Statistical Association*, vol. 75, no. 369, pp. 168-172, 1980.

[73] J. van der Geest, "Triplestest," Internet: http://www.mathworks.com/matlabcentral/fileexchange/19547, May 8, 2008 [Nov. 12, 2008].

[74] N. Dumin, K. Liu, and S.-H. Yang, "Gate Oxide Reliability of Drain-Side Stresses Compared to Gate Stresses," *IEEE Int. Reliability Physics Symp.*, pp. 73-78, 2002.

[75] K. Hofmann, S. Holzhauser, and C. Kuo, "A Comprehensive Analysis of NFET Degradation Due to Off-State Stress," *IEEE Int. Integrated Reliability Workshop*, pp. 94-98, 2004.

[76] S. Chang, C. Chen, C. Wang, and K. Wu, "A New Off-State Drain-Bias TDDB Lifetime Model for DENMOS Device," *IEEE Int. Reliability Physics Symp.*, pp. 421-425, 2009.

[77] P. Liao, C. Chen, J. Young, Y. Tsai, C. Wang, and K. Wu, "A New On-State Drain-Bias TDDB Lifetime Model and HCI Effect on Drain-Bias TDDB of Ultra Thin Oxide," *IEEE Int. Reliability Physics Symp.*, pp. 210-214, 2008.

[78] Y.-H. Lee, N. Mielke, W. McMahon, Y.-L. Lu, and S. Pae, “Thin-Gate-Oxide Breakdown and CPU Failure-Rate Estimation,” *IEEE Trans. on Device and Materials Reliability*, vol. 7, no. 1, March 2007.