Methodologies for Statistical Characterization of Circuit Reliability in Advanced Silicon Processes

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

Pulkit Jain

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

Chris H. Kim, Advisor

July 2012

Acknowledgements

First, I am truly indebted and thankful Prof. Chris H. Kim for his constant support and motivation for the past five years. He has always been an inspiring figure and I will never forget the lessons of perseverance and persistence he has given me.

Next, I would like to thank the members of my final defense committee: Professors Ramesh Harjani, Sachin Sapatnekar, and Andre Mkhoyan. I appreciate you taking time out of your busy schedules to critique my work. I am also grateful to Professor Antonia Zhai for serving on my preliminary oral defense committee.

I am also thankful to SRC and Texas Analog Center of Excellence for partially funding my research.

I am highly grateful to the senior alumni of our group. Especially, I would thank John Keane and Kichul Chun for their advice and guidance throughout my stay here. I learnt so much from them. I am obliged to many of my colleagues who helped me in my understanding through all the insightful discussions as well as supported me in tapeout runs.

I would also thank all my friends living in Minneapolis especially my roommate Ketan Rajawat for the past five years. I would also thank the entire ECE staff for making all the administrative work smooth and making sure we get our supply of free food.

I am indebted to my family for their support. My late grandfather for serving as my role model throughout my childhood and for his blessings. My brother, Pallav for being my close friend. My sister, Garima for making sure I get my birthday cakes over here in Minneapolis.

Finally, I dedicate my thesis to my parents. Words fall short in describing what they have done for me.

Abstract

Rising electric fields and imperfections due to atomic level scaling create non-ideal and stochastic electrodynamics inside a transistor. These appear as reliability mechanisms such as Bias Temperature Instability (BTI), Time Dependent Dielectric Breakdown (TDDB) and Random Telegraph Noise (RTN) at transistor level, and as a convolved statistical manifestation in performance and functionality, at a circuit level. Compounded by shrinking operating margins with process variability and power constraints, these reliability issues have been propelled from device research arena to the forefront of chip design.

The first part of my thesis will explore these different reliability issues in three dedicated test chips. While device level probing has been de-facto estimation method for reliability engineers due to legacy and simplicity, the approach has become cumbersome due to time and effort needed to cover the required statistics. Conversely, we demonstrate circuit based reliability monitors which are a more scalable and representative alternative. The latter also enable superior timing resolution which is critical to record phenomenon such as BTI and RTN without measurement noise. For example, leveraging on-chip methods and intelligent timing control, we demonstrate a SRAM reliability macro with BTI estimation at three order smaller measurement times than possible using conventional approaches. On-chip logic could also be used to control test on large number of blocks resulting in a large experiment time speedup which is the basis for our TDDB macro.

The second part of my thesis will focus on 3D integration, a breakthrough technology for reducing interconnects delays and chip form factors. In particular, we measure the impact of chip stacking on power delivery and propose schemes to mitigate it through a statistical framework, fabricated in an actual 3D technology.

Overall, the ideas here can pave the way for not only accurate empirical modeling and robust guard-banding for pre-silicon phase but also post-silicon adaptive tuning. And thus we can better reap the benefits of these new silicon technologies.

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

1.1 On-chip Monitoring of Reliability

The semiconductor industry has seen unabated and unrivalled growth for the past fifty years on the lines of Moore's predictions. However, moving into the sub-µm feature regime, pedaling this growth is becoming more and more challenging. Increasing performance and lowering power has led to transistors with rising electric fields, atomistic level scaling and quantum effects such as discrete dopants and gate oxide traps. Consequently, device reliability mechanisms have become pressing concern in scaled technologies.

Temporal degradation in reliability due to 'aging' of transistor occurs through several mechanisms. Time Dependent Dielectric Breakdown (TDDB) and Bias Temperature Instability (BTI) are identified as the foremost aging concerns from a system reliability standpoint [1]. Moreover, these aging mechanisms are stochastic due to the nature of charge exchange involved leading to stochastic lifetimes. New processes such as High-k Metal Gate (HKMG), tri-gate FETs, thru-silicon via based 3D interconnects, lead to several benefits but at the same time pose new issues in reliable performance of systems. While some mechanisms have taken a back seat (e.g. hot carrier injection due to saturated clock frequencies), others are becoming more important. For instance, Positive Bias Temperature Instability (PBTI) has

become important with HKMG [2]. 3D Thru Silicon Vias (TSV) can potentially induce supply noise and mechanical stress as well as electro-migration concerns in integrated circuits [3]. Random telegraph noise which had been mostly an issue in analog circuits[4]and flash memories [5] can cause timing hazards and noise margin issues in VLSI design[6][7].

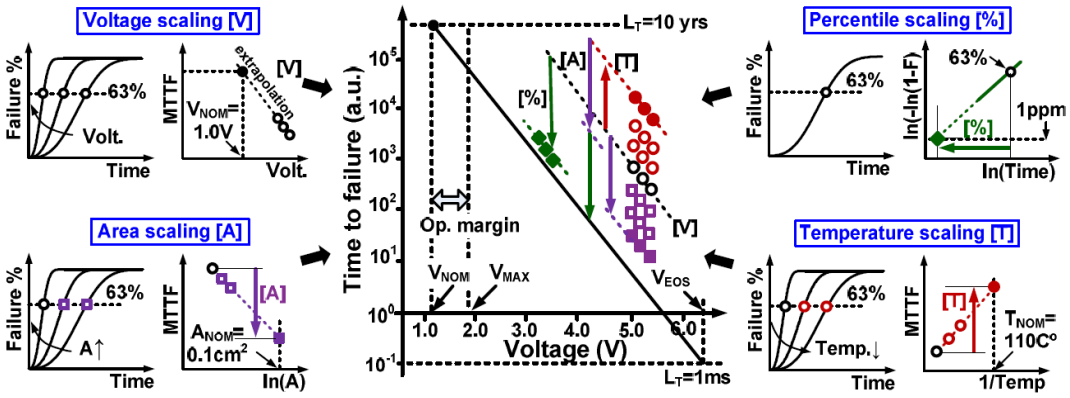
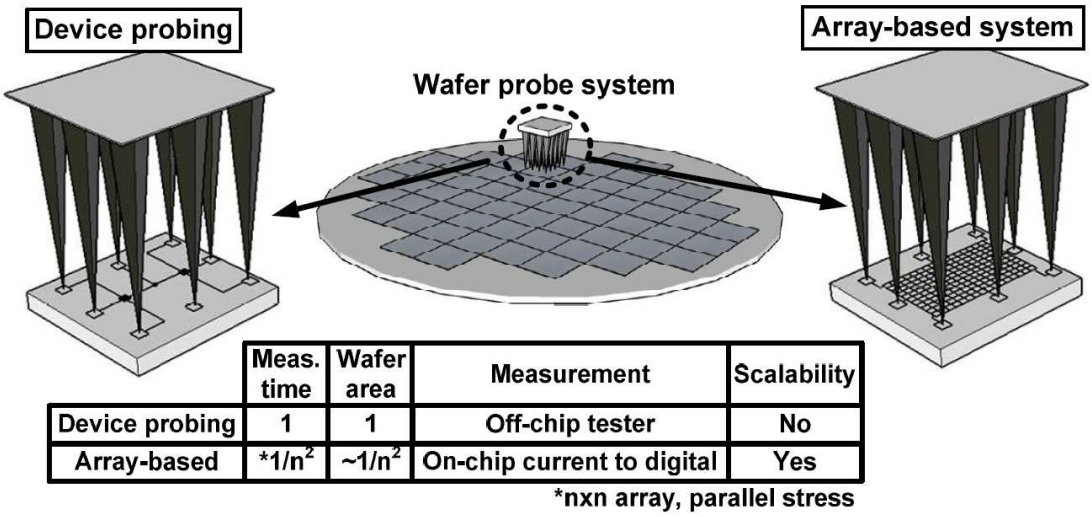


Fig. 1.1 Chip lifetime projection for TDDB based on accelerated stress involves mass data collection (e.g. up to 1000's of samples per MTTF data) to make voltage, percentile, area, and temperature projections to actual product usage conditions. The open symbols represent measurable values at accelerated condition, while the solid symbols represent the projected value.

Optimizing the fabrication process and using proper operating conditions based on accurate lifetime predictions for that process node is the most practical and effective approach for the now indispensable, design for reliability paradigm. While device level characterization has been de facto estimation method for reliability engineers due to legacy issues, modern processes have become way more diverse and complex with several functional domains, making this approach unwieldy. The operating condition on which a transistor is operated in a chip differs considerably from individual devices. Also, it is more accurate and meaningful to directly measure out macro level performance metrics such as minimum operating voltage or

maximum operating frequency of a circuit than to estimate using models derived from device level data. Fig. 1.1 gives an example of TDDB lifetime estimation flow which demonstrates that accelerated measurements take a long time and collection of *massive* statistical data from accelerated tests, as reliability is a function of a number of variables including voltage, temperature, area, dielectric thickness, and purity. Compared to device level probing methods, circuit based methods provide an efficient way to gather thousands of samples needed to correctly define a *single* Mean-Time-To-Failure (MTTF) value as illustrated in Fig. 1.2. This becomes important especially in a scenario when a few fails can cause catastrophic fails in a chip for instance a flip in a SRAM cell or violation in a critical timing path.



	Meas. time	Wafer area	Measurement	Scalability
Device probing	1	1	Off-chip tester	No
Array-based	*1/n^2	~1/n^2	On-chip current to digital	Yes

*nxn array, parallel stress

Fig. 1.2 Array based approach is an efficient way to carry out aging measurements compared to conventional probing. In the example shown above, device probing using off-chip tester with 8 probes, can test two devices at a time. On the other hand, in the array based system, using the same resources, a n by n array of devices could be tested out.

1.2 CMOS Transistor Aging Mechanisms

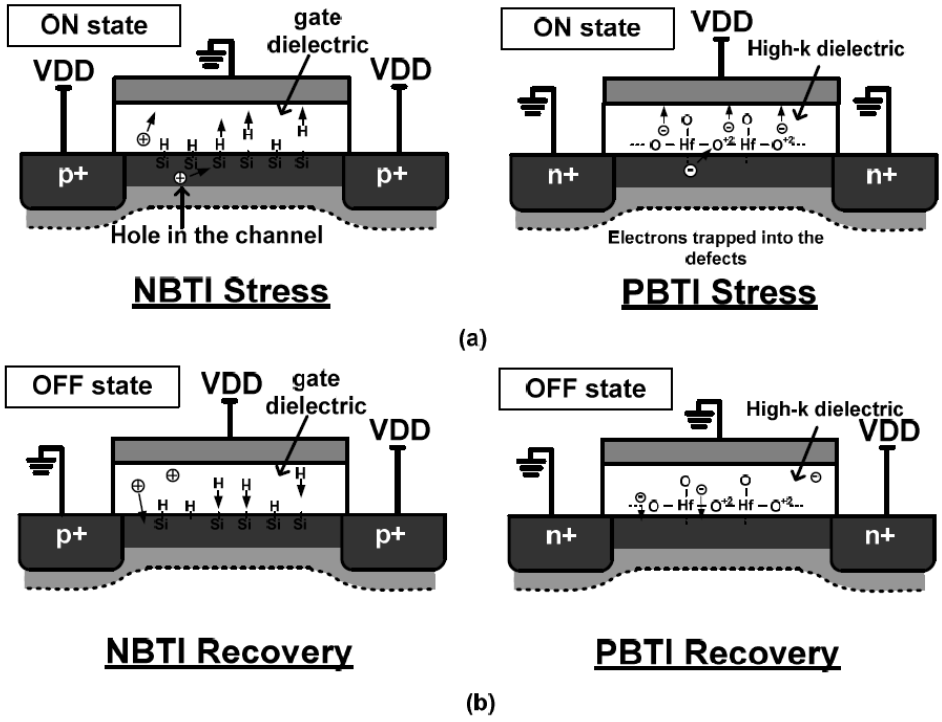


Fig. 1.3 (a) Mechanism of BTI due to charge trapping during stress phase (b) Mechanism of BTI recovery due to charge detrapping during recovery phase

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 [8].

NBTI and recently PBTI are the most critical concern for transistor reliability. NBTI has been studied since decades and has been important especially after the introduction of nitrogen into gate stacks, which reduces boron penetration and gate leakage, but leads to worse NBTI degradation [9]. 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. 3(a)) [9-10]. 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_T (Fig. 1.3(b)).

With the advent of hafnium based high-k dielectrics, even NMOS suffer from degradation [2] due to electron trapping in positively charged oxide traps, leading to similar electrostatics as NBTI and leading to V_T degradation and recovery manifestations as shown in Fig. 1.3, in what is known as PBTI.

Random Telegraph Noise can be explained by the charge trapping/detrapping mechanism similar to BTI above as seen in Fig. 1.4(a) . Either the primary or secondary carriers can gain enough energy to tunnel into the gate stack. This creates traps at the silicon substrate/gate dielectric interface, as well as dielectric bulk traps. These "traps" are electrically active defects that capture carriers at energy levels within the bandgap. This is not a permanent phenomenon and the captured carriers are emitted back into the substrate in the timescale of microseconds to milliseconds. Overall, this leads to fluctuations in device characteristics such as the threshold

voltage (V_{th}) as shown in Fig. 1.4(b). RTN is closer to the recoverable component of BTI in this sense [11] although the exact resemblance is not clear.

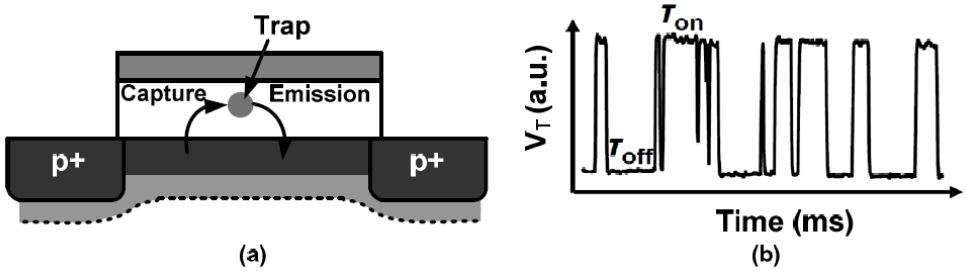


Fig. 1.4 (a) Mechanism of RTN in a transistor. Capture and emission time constants are random. (b) Typical RTN induced V_T fluctuation [25]

Finally, the traps in the gate oxide may eventually join together and form a conductive path through the stack in a process known as TDDB, or oxide breakdown [1] as shown in Fig. 1.5 (top) . 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. Fig. 1.5(bottom) shows the different TDDB modes affecting common digital circuits. While the on-state TDDB is most severe and conventionally assumed to be critical due to the entire gate area being exposed to stress, High Drain, High Source (HDHS) and High Drain (HD) OFF-state modes [1] might lead to earlier failure in circuits such as SRAM access devices that are exposed to an off-state stress for most of their lifetime. As for Input-Output (IO) devices, Electrical OverStress (EOS) and ElectroStatic Discharge (ESD) are of particular concern.

In this work we propose several characterization techniques to efficiently collect failure statistics from these important reliability mechanisms.

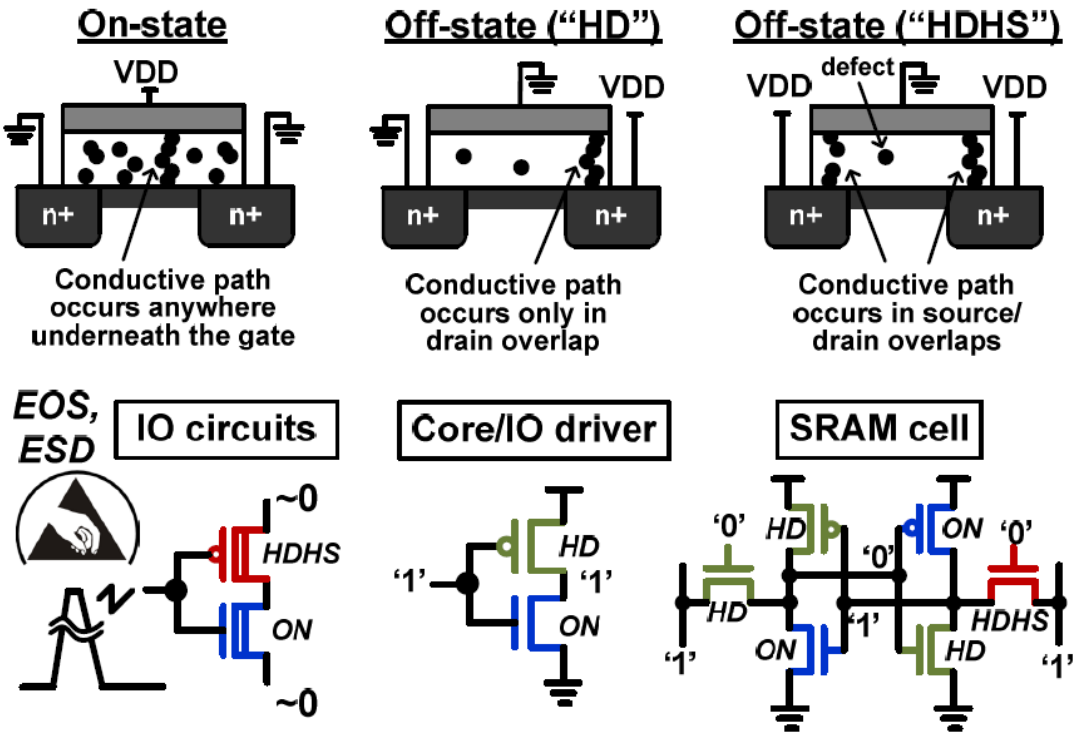


Fig. 1.5 Different occurrence of TDDB. While 'ON' and 'OFF-HD' cases are most prominent, 'OFF-HDHS' is also seen in certain cases such as SRAM access devices.

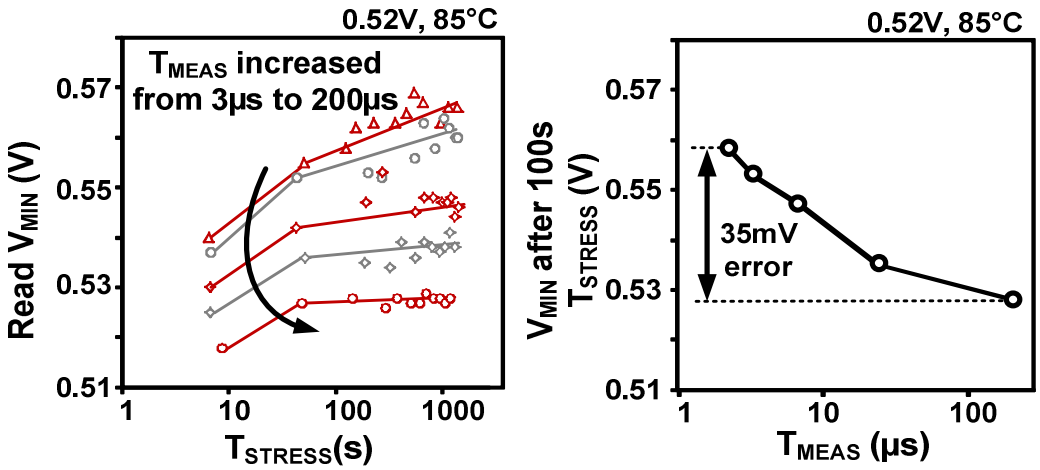
1.3 Statistical Measurement of BTI Impact in SRAM

SRAMs is the workhorse memory structure in all high performance caches. While cell mismatch due to process variations remains the foremost challenge in SRAM stability, a rising concern has been related to aging induced skew in the bit-cells [12]. The general approach is to try to explain and predict the temporal change in the dynamic and Static Noise Margins (SNM) with aging, corroborated with empirical results. The approach can herald in a bottom to up design for reliability

automated flow [13-14]. However, the validity of these noise margin models is questionable, unless the empirically obtained V_{MIN} and Bit Fail Rates (BFR) are precisely comprehended from a product. This becomes a non-trivial problem for two reasons. First, prominent aging phenomena such as Bias Temperature Instability (BTI) and oxide breakdown have a spatial distribution and variation of their own [15] and just tracking few zero time failing bits [16] would be futile. Secondly, specific to BTI, the measurements get corrupted due to phenomenon of recovery. Previous methodologies [17][18], suffer from an inherently large measurement time, T_{MEAS}, leading to incorrect comprehension of BTI statistics. In a typical Stress-Measure-Stress (SMS) routine for accelerated testing of BTI, any evaluation can be accurate only if measurements are done within the order of microseconds [14] as seen from Fig. 1.6 (left). Any larger measurement times would provide wrong evaluation of degradation metrics leading to an overly optimistic estimate of BTI effects as illustrated in Fig. 1.6(right)

In logic circuits, statistical measurement is easier due to the ability to gate on/off stress on individual blocks [14]. As a result, a small amount of data needs to be handled within T_{MEAS}, which can be temporarily stored on-chip, while parallel stressing of all blocks can be done to cut test time. However, the approach cannot be extended to SRAM/memory. Since, the supply rail is shared globally across all rows, the enormous data running into several megabits, has to be processed in parallel. Moreover, the entire data also needs to be readout off-chip as on-chip storage would

be too costly in terms of area. Considering, a typical data acquisition frequency of few megahertz, such a fast measurement becomes problematic.



(a) (b)

Fig. 1.6 (a) Longer TMEAS results in optimistic BTI data (= lower bitcell failure rate) due to the unwanted fast recovery. (b) Power law exponents measured at different TMEAS indicates a recovery time constant of ~25µs [14].

In Chapter 2, we show our proposed SRAM test structure design[19] targeting a microsecond order measurement time, an improvement of several orders over a previous on-chip approach [18] and therefore provides recovery free BTI data on a representative SRAM array.

1.4 Statistical Measurement of TDDB Impact in Circuits

TDDB is the most important reliability concern for any VLSI system. Unlike other aging phenomenon it is much more random and harder to predict. The best approach is to gather extensive measurement data in order to capture the tail cell behavior and based on that provide appropriate guard banding in supply voltage. As underscored in previous paragraphs, array based systems to monitor TDDB excel

over traditionally used device probing in terms of efficiency and scaling. A previous characterization array for TDDB [20] only considered ON-state stress in core transistors which is not enough to obtain an accurate picture of system lifetime. A combined lifetime prediction methodology is needed to take into account different modes (as shown in Fig. 1.5) in tandem with their predicted time to failures. In Chapter 3, we propose an array-based system that includes a flexible DUT cell [21][22] that can be stressed in isolation without thicker tox FETs to 4 times supply voltage. This enables accurate lifetime prediction under different ON and OFF state dielectric breakdown modes for both low voltage core and high voltage IO devices.

1.5 Statistical Measurement of RTN Impact in Logic Circuits

RTN is becoming an increasing concern in ultra-scaled technologies [23]. Erratic behavior of SRAM under RTN has especially come to scrutiny[24][25]. One potential issue that has been largely unaddressed is that RTN can cause timing hazards in logic circuits owing to the typical time constants ranging in microseconds to milliseconds [26] and abrupt shifts in V_T exceeding 25mV. An interesting approach was taken up in [7] by operating a D flip-flop in a meta-stability region to amplify RTN impact. An asymmetric RO was proposed in [59] to isolate RTN in a ring oscillator. However, no work has been reported to directly observe the impact on a traditional ROSC with high resolution. One reason has been difficulty in measuring out RTN impact, due to small shifts in frequency (0.1-1%) that are expected, overshadowed by within-die variation and supply noise. Even after alternating stress bias methods the expected shifts are very low and thus a high resolution measurement

technique is mandated. Chapter 4 describes our proposed statistical framework to directly monitor impact of RTN on ROSCs with sampling time less than 1µs for a 0.1% frequency resolution, which is at least a 10X improvement over conventional techniques [7][59]. A test chip in a 32nm silicon on insulator process features RTN measurements from 20 varieties of ROSCs, with difference in number of stages and device sizes to enable a comprehensive RTN study.

1.6 Statistical Measurement of TSV Impact on 3D ICs

3D integration is recognized as a breakthrough technology for improving interconnects performance and reducing chip form factors [29-30]. Memory bandwidth, which has become a critical performance limiter in modern processors, can be significantly increased by vertically stacking caches on top of processing cores. Extremely high memory densities have been demonstrated for stand-alone applications where multiple 2D memory chips are stacked in a single package. 3D integration technology also makes it possible to vertically integrate chips built in heterogeneous processes (e.g. logic, DRAM, flash, SiGe, InP) with slight additional cost compared to integrating monolithic chips.

The premise of 3D integrated circuits has spurred research activity at virtually all levels of the 3D design hierarchy. However, despite the recent surge in 3D IC research, there has been virtually no work from the circuit design and automation community on power delivery issues for 3D ICs. On-chip power supply noise has worsened in modern systems because scaling of the Power Supply Network (PSN) impedance has not kept up with the increase in device density and operating current

due to the limited wire resources and constant RC per wire length [31]. This situation is worsened in 3D ICs as TSVs contribute additional resistance to the supply network and the number of pins for power delivery is fundamentally limited by the footprint of the 3D chip. For example, a 3D chip with *n* tiers can only have 1/*n* the number of power supply pins compared to a single 2D chip of *k*-time footprint, which results in an *n* fold increase in the resistive and inductive parasitics. The increased IR and Ldi/dt supply noise in 3D chips may cause a larger variation in operating speed leading to more timing violations. The supply noise overshoot due to inductive parasitics may aggravate reliability issues such as oxide breakdown, negative bias temperature instability and hot carrier injection. Consequently, on-chip power delivery will be a critical challenge for 3D ICs. This is contrary to the common perception where power delivery in 3D chips was considered no different than that in conventional 2D chips.

In Chapter 5, we specifically address the TSV impact on supply noise in high performance 3D ICs. We demonstrate a test chip in a MIT Lincoln Lab's 0.15µm three stacked process to measure and address the ensuing issues [32-33].

1.7 Summary of Thesis Contributions

The remainder of this work will explore the benefits of four test chip designs that we have implemented to accurately monitor circuit reliability.

The first is the SRAM Odometer where we present a scalable test structure for recovery free evaluation of the impact of NBTI and PBTI on read/write operation in a

SRAM macro. A novel non-invasive methodology keeps the stress interrupts for measurements within a few microseconds, preventing unwanted BTI recovery, while providing a parallel stress-measure capability on 32kb sub-arrays. Measurement results in a 32nm high-κ/metal-gate silicon-on-insulator process show that proposed schemes provides 35mV better accuracy in read VMIN and 10X accuracy in BFR.

The second is a comprehensive macro, for automatically characterizing gate dielectric failure and reduces the stress time and silicon area by a factor proportional to the number of FETs to be tested. A flexible DUT cell that can be stressed in isolation without thicker tox FETs to 4 times supply voltage, enables accurate lifetime prediction under different ON and OFF state dielectric breakdown modes for both low voltage core and high voltage IO devices.

The third is a 32nm test macro to directly monitor impact of RTN on ROSCs with sampling time less than 1µs for a 0.1% frequency resolution.

The fourth is a 3D IC test chip in a MIT Lincoln Lab's 0.15µm process has been fabricated with the goal to evaluate TSV impact on supply noise as well as variability introduced in timing paths. In this work, we specifically address the power delivery issues in high performance 3D ICs, that can monolithically integrate logic and memory.

Chapter 2

A 32nm SRAM Reliability Macro for Recovery Free Evaluation of NBTI and PBTI Induced Bit Failures

2.1 Introduction

2.1.1 Importance of Measurement Time in BTI Capture

While cell mismatch due to process variations remains the foremost challenge in SRAM stability, a rising concern has been related to aging induced skew in the bit-cells [12]. The general approach is to try to explain and predict the temporal change in the dynamic and Static Noise Margins (SNM) with aging, corroborated with empirical results. The approach can herald in a bottom to up design for reliability automated flow [13-14]. However, the validity of these noise margin models is questionable, unless the empirically obtained V_{MIN} and Bit Fail Rates (BFR) are precisely comprehended from a product. This becomes a non-trivial problem for two reasons. First, prominent aging phenomena such as Bias Temperature Instability (BTI) and oxide breakdown have a spatial distribution and variation of their own [15] and just tracking few zero time failing bits [16] would be futile. Secondly, specific to BTI, the measurements get corrupted due to phenomenon of recovery. Previous methodologies [17-18], suffer from an inherently large measurement time, T_{MEAS}, leading to incorrect comprehension of BTI statistics. In a typical Stress-Measure-Stress (SMS) routine for accelerated testing of BTI, any evaluation can be accurate

only if measurements are done within the order of microseconds [15] as seen from Fig. 2.1 (left). Any larger measurement times would provide wrong evaluation of degradation metrics leading to an overly optimistic estimate of BTI effects as illustrated in Fig. 2.1(right). It has been reported that recovery can even occur during inversion mode at normal operation when the accelerating condition is removed [8-9].

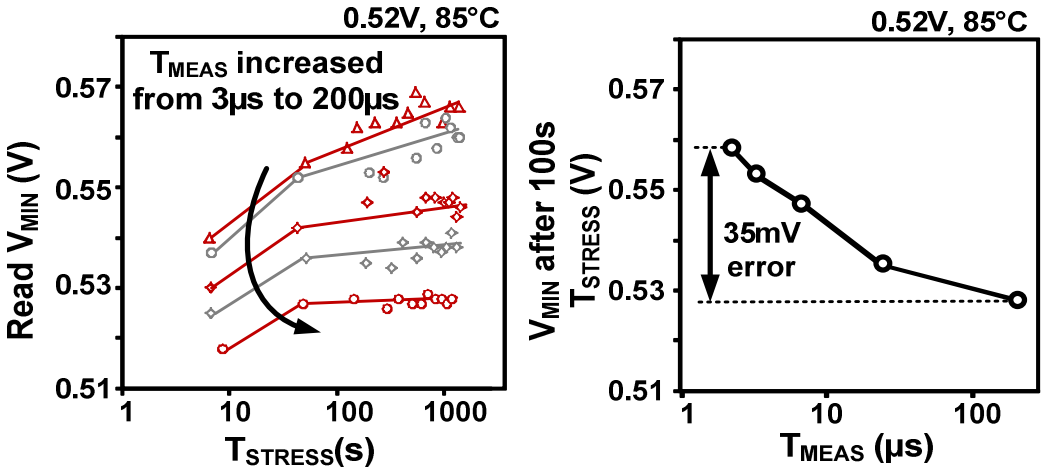


Fig. 2.1 (a) Longer TMEAS results in optimistic BTI data (= lower bitcell failure rate) due to the unwanted fast recovery. (b) Power law exponents measured at different TMEAS indicates a recovery time constant of ~25µs [4].

In logic circuits, statistical measurement is easier due to the ability to gate on/off stress on individual blocks [15]. As a result, a small amount of data needs to be handled within T_{MEAS}, which can be temporarily stored on-chip, while parallel stressing of all blocks can be done to cut test time. However, the approach cannot be extended to SRAM/memory. Since, the supply rail is shared globally across all rows, the enormous data running into several megabits, has to be processed in parallel. Moreover, the entire data also needs to be readout off-chip as on-chip storage would

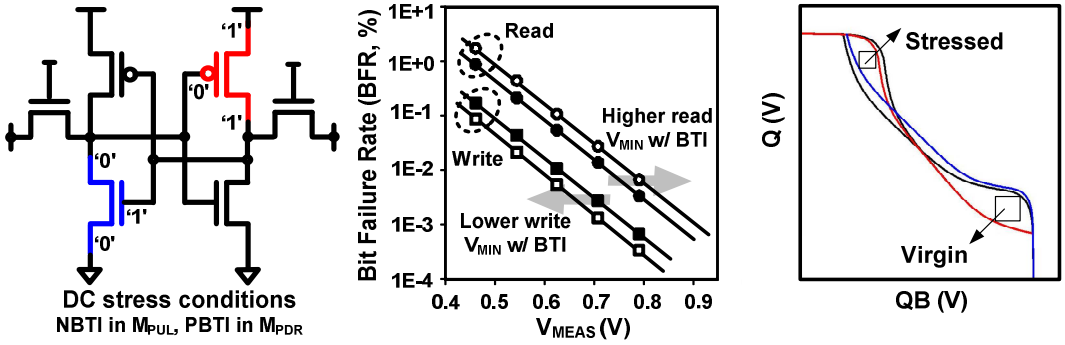
be too costly in terms of area. Considering, a typical data acquisition frequency of few megahertz, such a fast measurement becomes problematic.

Kim et al. [17], used off-chip control of supply during measurement to obtain the SRAM V_{MIN} during measurements, which takes few seconds to obtain the result, leading to extensive recovery in measurements. Recently, [18] proposed a BFR tracking approach with local data storage similar to this work for fast measurements. However, the overall approach was not scalable to full SRAM arrays and couldn't be used for progressive evaluation of BTI. Instead end-of-life estimation of degradation metric was provided, which has limited use for reliability modeling. This work, we believe, shows the first known SRAM test structure design targeting a microsecond order measurement time, an improvement of several orders over previous on-chip approaches [17-18] and therefore provides recovery free BTI data on a representative SRAM array. The main techniques proposed are 1) Pseudo-Reads consisting of WL perturbations and local data storage with deferred Stressed Readout (PR-SR), and 2) Flip-Latch-Restore approach with intermittent Scan out (FLR-S). Before delving into these techniques, we give some background on impact of BTI on SRAM.

2.1.2 BTI Impact on SRAM Reliability: General Discussion

For SRAM, the relatively low activity factor results in DC stress prevailing for the majority of the time. This leads to positive bias stress on the NMOS pull-down driver, M_{PDR} and negative bias stress on the PMOS pull-up load, M_{PUR} (Fig. 2.2(a)). In the typical first access occurring after a long DC stress, read stability degrades mainly due to the weaker driver NMOS while write stability improves due to the weaker

PMOS. This opposite behavior between read and write stability with BTI gets exhibited during dynamic operation (Fig. 2.2(b)) and static operation (Fig. 2.2(c)). Note, our goal will be to isolate out the impact on dynamic operation by the above BTI induced SNM degradation. Access time violations due to weakened drive currents or peripheral aging have not been found as critical [14].



(a) (b) (c)

Fig. 2.2 (a) SRAM static stress condition promote BTI stress in the two highlighted MOSFETs. (b) Under the influence of BTI stress, SRAM read VMIN worsens while write VMIN improves. (c) Effect of BTI on Static Noise Margin (SNM).

2.2 Proposed BTI Test Macro

2.2.1 Macro Design

Fig. 2.3 shows the proposed SRAM reliability macro. Overall, SRAM specific components are designed to be representative of a product sub-array. For reducing implementation complexity and pin count, we refrained from column multiplex or sense amplifier, and opted for a Single-Ended Sensing (SES) scheme with a slow scan based readout. A marker row with alternate hardwired '1' and '0's was used to verify correct address flow during dynamic operation. The complicated part of the BIST (Built In Self-Test), like controlling the supply switches for measurement and stress

modes, measurement times, pulse width control, read/write commands, address sequencing, etc. were handled by the on-chip Finite State Machine (FSM) and voltage controlled oscillator. The slower timings like scans and BFR readout were handled by Labview® off-chip.

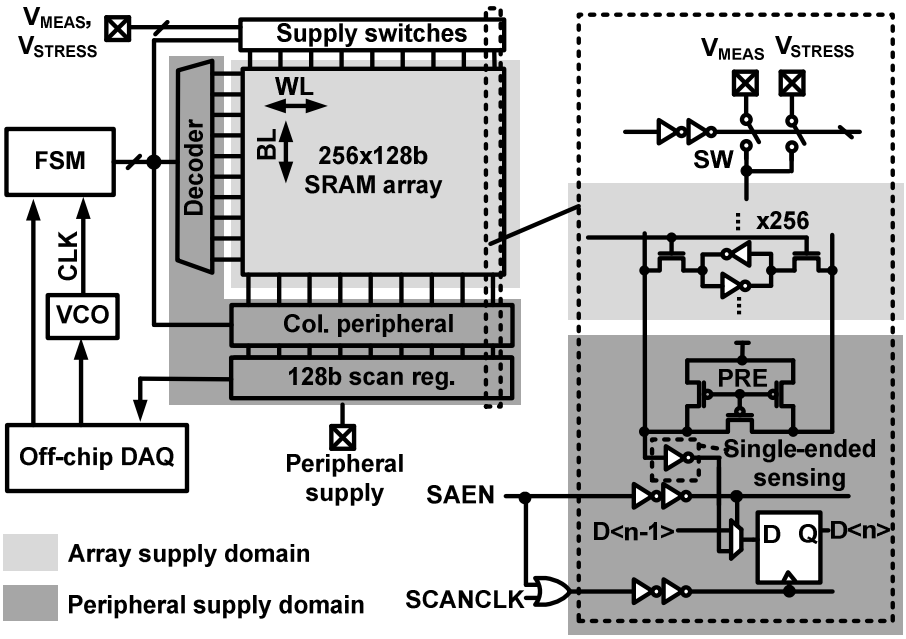
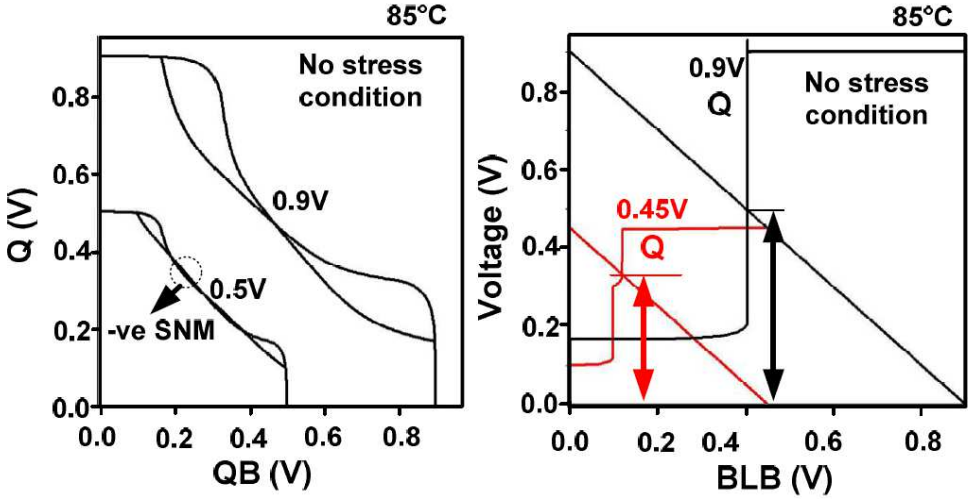


Fig. 2.3 SRAM reliability macro architecture. Bit-cell array is representative of a product sub-array and features a 128b scan and single-ended sensing for ease of test. BIST functionality is realized by an on-chip finite state machine that administers the stress-measure-stress sequence.

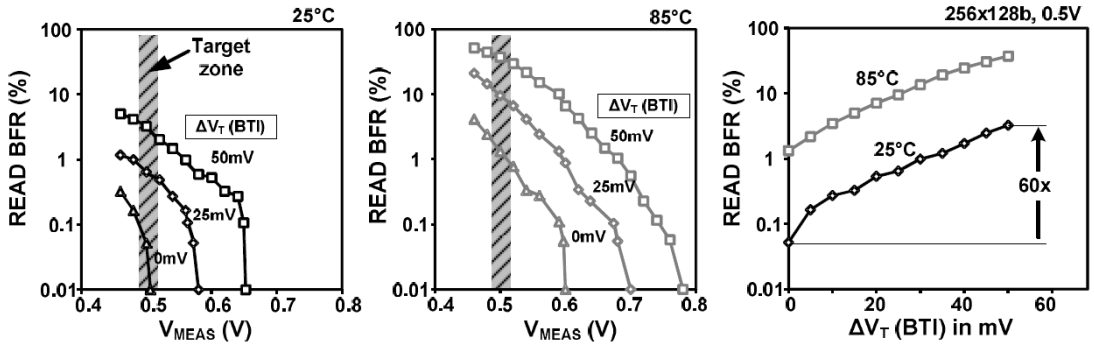
On-chip supply switches were used on a column wise granularity with delayed firing of signals to reduce current spikes during supply switching and optimize the overall switching time.

2.2.2 Operating Condition



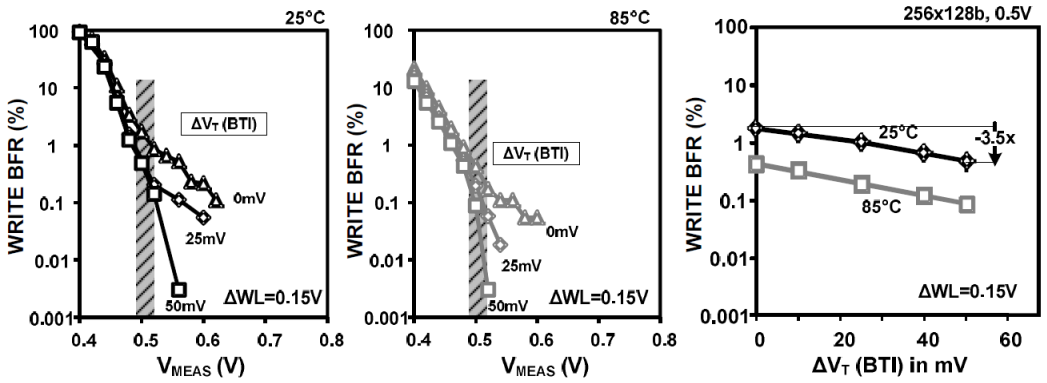
(a) (b)

Fig. 2.4 Simulations result at TT corner on a SRAM cell for (a) SNM and (b) Write margin. The cell is more prone to read fails.



(a) (b) (c)

Fig. 2.5 Read BFR simulations of a 256x128b sub-array in 32nm SOI. (a) Read BFR for different V_{MEAS} and BTI at 25°C. V_{MIN0} is around 0.5V for target a BFR value of 0.01%. BFR vs VMEAS curve becomes non-regular below 1% BFR. (b) Read BFR for different V_{MEAS} and BTI at 110°C. (c) Read BFR at V_{MEAS}=0.5V with different BTI. Around 60X increase in BFR seen with an equal V_T shift in PUR and PDL of 50mV. Above 10X increase in BFR from 25°C to 85°C



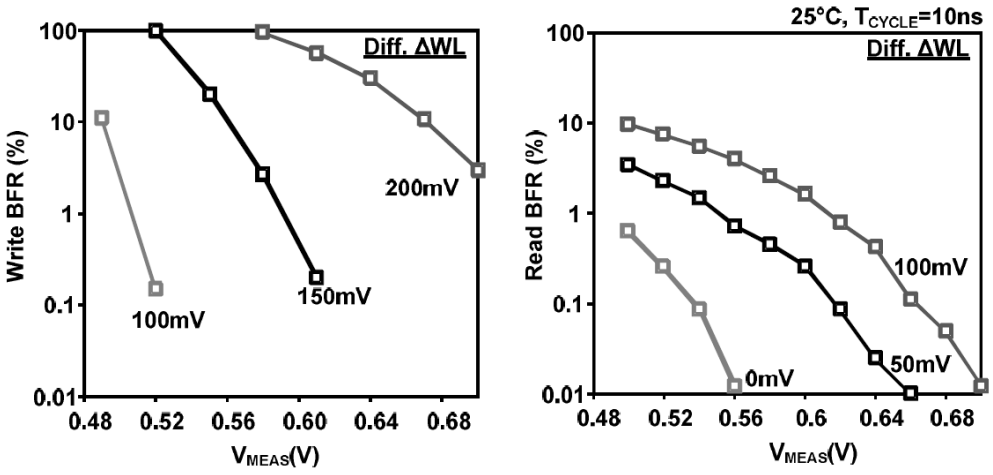
(a) (b) (c)

Fig. 2.6 Write simulations of a 256x128b sub-array in 32nm SOI. (a) Write BFR for different V_{MEAS} and BTI at 25°C. V_{MIN0} is around 0.5V for target a BFR value of 1%. (b) Write BFR for different V_{MEAS} and BTI at 110°C. (c) Write BFR at V_{MEAS}=0.5V with different BTI. Very weak dependence in WRITE BFR. This is because we are tracking 1% of the cells. Around 3.5X decrease in BFR seen with an equal V_T shift in PUR and PDL of 50mV. Above 4-5X decrease in BFR from 25°C to 85°C. SRAM cycle time for different V_{MEAS}. Cycle time is ~10ns for the target VMEAS.

Fig. 2.4 shows the simulated SNM and write margin for the SRAM cell employed in this work. The sizing was based on a high performance cell used in a commercial microprocessor. The curves demonstrated an unstable cell sized for good write margin and negative SNM at 0.5V for typical corner. However, as follows next, dynamic stability assessment showed that static simulations give a pessimistic assessment.

Figs. 2.5 and 2.6 plot simulated BFR at different operating voltages, assuming different BTI induced ΔV_T in PBTI and NBTI affected FETs. If we target a BFR around 0.01-0.1% (about 1-30 fails in 32kb), we need to be operating in the highlighted target zone i.e. the measurement voltage (V_{MEAS}) should be around 0.5V. While BTI impact will be demonstrated later, Fig. 2.7 shows the measured BFR from the test chip at different V_{MEAS} at virgin conditions which verify the trends from the

simulation. Note that in actual measurements the cells were extra robust towards read and write failure and a WL voltage control was used to obtain fails around 0.5V.



(a) (b)

Fig. 2.7 Measured (a) READ and (b) Write BFR at different V_{OP} at virgin stress conditions for calibration purposes. This validates the simulation results in Fig. 5 and Fig. 6.

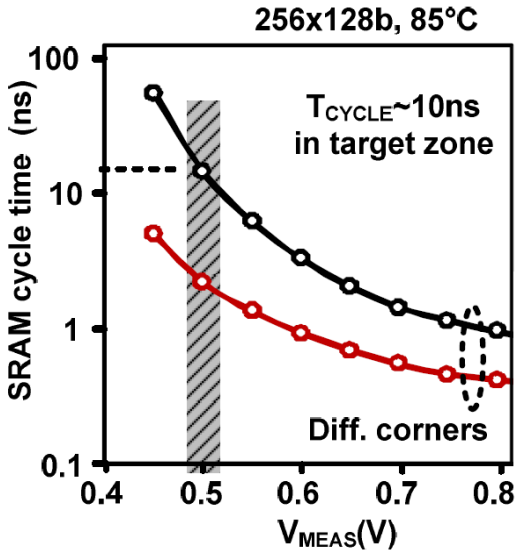


Fig. 2.8 SRAM cycle time for different V_{MEAS} at best and worst corners at 85°C. Cycle time is ~10ns for the target V_{MEAS} of 0.5V.

Fig. 2.8 depicts the simulated operating cycle time (T_{CYCLE}) dependency on V_{MEAS}. In the target zone, T_{CYCLE} is around 10ns. As will be explained in the succeeding sections, T_{MEAS} is related to T_{CYCLE} and number of SRAM rows to be traversed, and thus is roughly around 256x10ns. Note that owing to limited silicon real estate available for our project, we could at best implement a few sub-array of size 32kb. In a real megabit size cache, expecting V_{MEAS} would be around 0.7V [36], T_{CYCLE} equals around 1ns. For same sub-array size of 32kb, T_{MEAS} would be then 256x1ns.

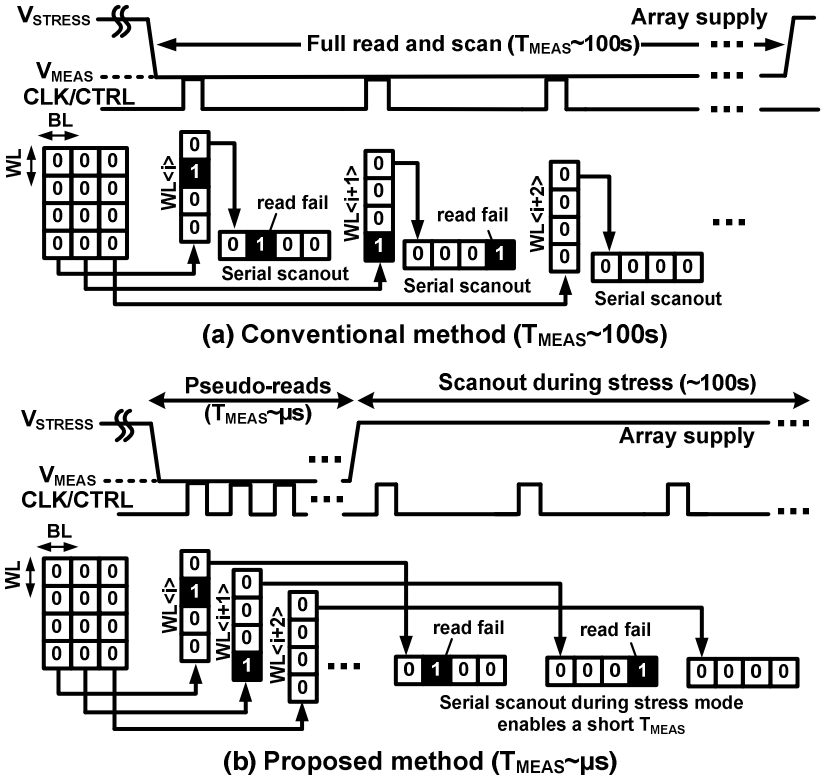


Fig. 2.9 Read BFR measurement sequence example for an array initialized to zero. (a) In the conventional method, supply is lowered to VMEAS followed by a full read and slow scan out which results in a long TMEAS (b) The proposed approach consists of a pseudo-read (=sequential WL perturbations) which stores pass/fail info in the array. The array is immediately put back into stress mode to prevent unwanted recovery followed by a full reliable read and scan out.

2.2.3 Read fail tracking with proposed PR-SR approach

Fig. 2.9 shows example timing diagrams of the conventional [17] and proposed methods. Prior to applying V_{STRESS}, all bitcells are initialized through a blanket write '0'. Next, the peripheral supply is externally lowered down to $V_{MEAS,}$ a level corresponding to a target read BFR. This completes the initialization step. Next, stress is applied in a stress-measure-stress routine with exponentially increasing stress intervals using an array supply of V_{STRESS}. In the short measure window, the array supply is lowered to V_{MEAS}, using on-chip switches with 20% of T_{MEAS} dedicated to supply switching. A pseudo-read burst consisting of up to 256 sequential WL perturbations follows next. If we consider an affected row, all cells on it that are 'weak' get a data flip, while others that are 'strong' retain their original values. Thus pass/fail information corresponding to this measurement interrupt gets stored locally in that same cell. After this, the array supply is switched back to V_{STRESS} to prevent unwanted BTI recovery. We defer the full read and off-chip data acquisition in this stressed stage as the pass/fail info is retained. Due to the long stress periods, this can be done much slowly without interrupting the overall test procedure. Note that since the array operates at a high stress voltage in this state, the chance of any cell failure occurring at this stage is remote. After the BFR has been captured and scanned out, the entire cycle is repeated. Fig. 2.12 (a) shows a simulation waveform for the FSM routine. An extension of this approach can be used to track $V_{MIN.}$ (Fig. 2.10). Instead of tracking BFR at V_{MIN0} in the above sequence, delicate control of V_{MEAS} during successive PR-DR routines, keeping target BFR roughly constant can be used to give

an indication of READ V_{MIN} as a function of stress time, T_{STRESS}. Note, that as V_{MEAS} is ramped down, a small change in BFR is used to sample out V_{MIN}.

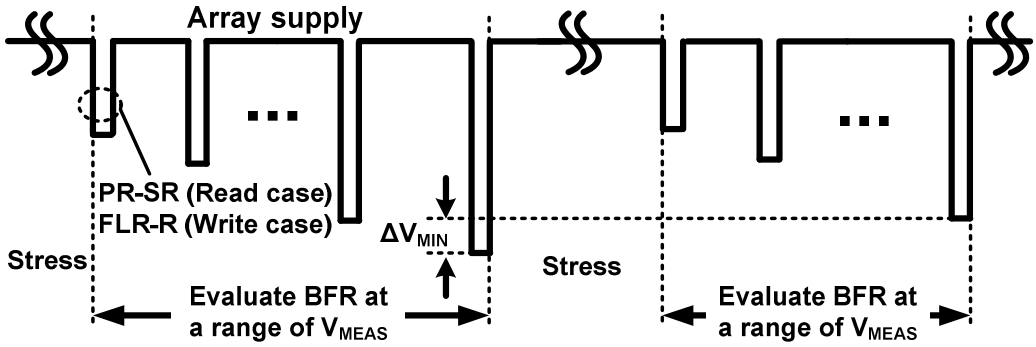


Fig. 2.10 Extension of the read BFR test sequence in Fig. 5 for read VMIN measurements with microsecond range TMEAS. Here, VMEAS is stepped down until a target BFR is reached. Similar concept can be applied for tracking write VMIN.

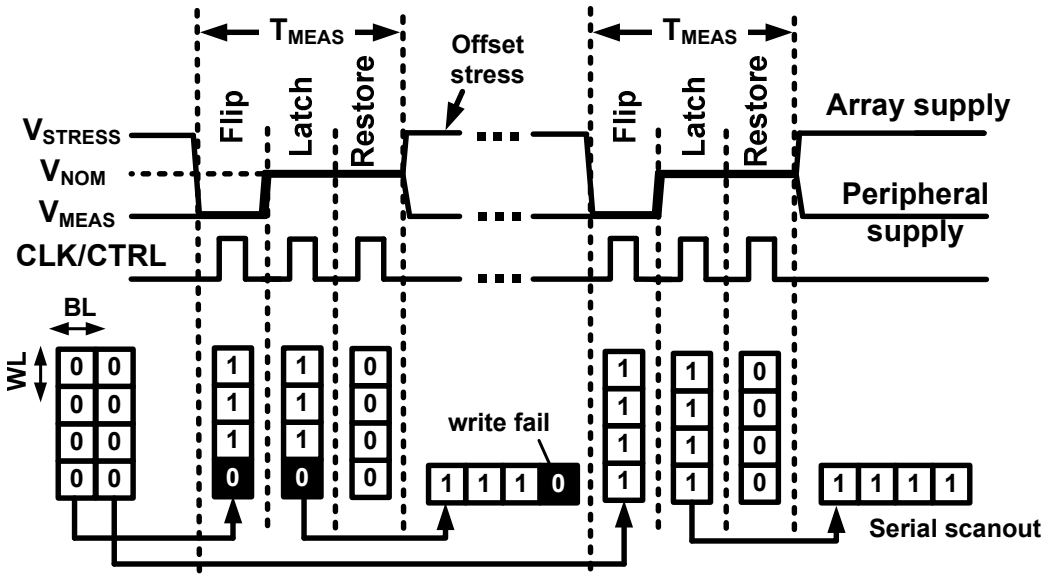
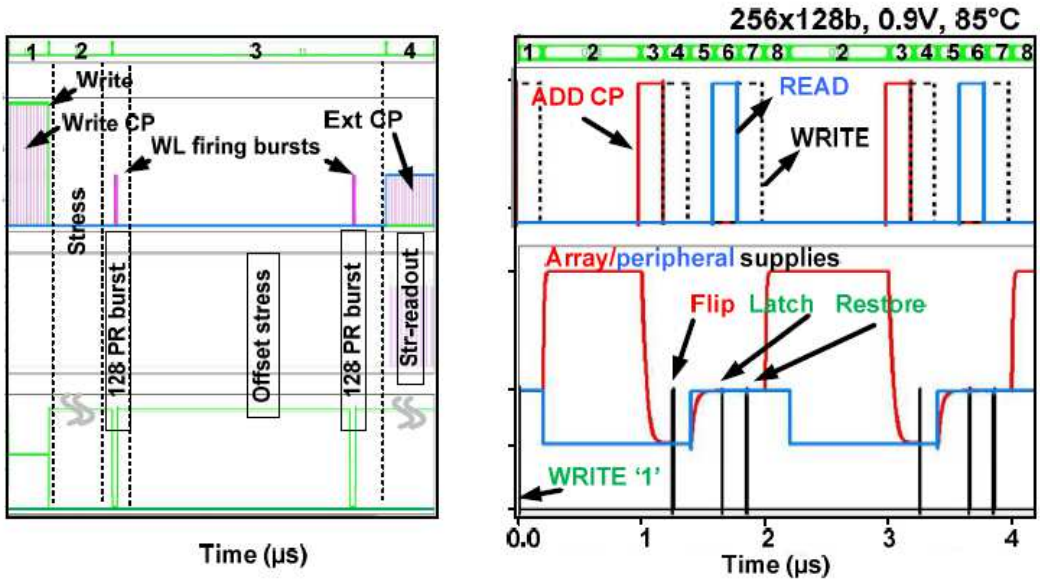


Fig. 2.11 Write BFR measurement sequence for an array initialized to zero. First, the opposite data is forced (i.e. write 1) at VMEAS. Next, supply is raised to VNOM (=0.9V) and a reliable full read at VNOM samples data into a shift register. To prevent the cells from recovering, the data is flipped back to its initial state (i.e. write 0), and the array is immediately put back to stress. Serial scan out is performed at this time.



(a) (b)

Fig. 2.12 Timing simulations for (a) Read FSM. States 1 is initialization, 2 is stress, 3 and 4 is the PR-SR sequence with pseudo read, intermittent offset stress and stressed read out (b) Write FSM employs 8 states. States 1 and 2 are for initialization and stress. States 3 to 7 are the row-wise FLR sequence with intermittent scan out in state 8. The row address gets updated in state 2.

2.2.4 Write fail tracking using proposed FLF-R approach

An approach similar to the above would not work for write case. A 'good' cell will flip easily on a write. Consequently, BTI due to the prior DC stress, would start to recover, unless an immediate second flip (or write-back) to the original state is done. Hence, the cell cannot be used as a temporary storage for BTI information, and a full readout into shift registers is needed to capture the first flip information. The ensuing timing sequence is shown in Fig. 2.11. The initialization step and stress resembles the read case. The T_{MEAS} window consists of the critical flip with array and peripheral supplies kept at V_{MEAS}, followed by a reliable read-latch and restore at V_{NOM}=0.9V. This biasing ensures that we isolate out the first flip fails. After FLR, array supply

goes to V_{STRESS} and we do a slow scan out of the data stored in the on-chip shift registers. Then, FLR-S is repeated for the next row.

The main caveat is that the latter rows would observe a somewhat AC stress behavior, which could possibly induce some error due to recovery. As claimed in [37], the error is small if we use an offset stress of $1000xT_{MEAS}$. Fig. 2.12 (b) shows a simulation waveform for the FSM routine.

2.2.5 Comparison to previous works

Table 2.1 shows a summary and comparison of this work to previous SRAM aging macros. Array based approaches score over the traditional probing based approaches in terms of scalability and representativeness. The main strength of this work from a previous array based implementation [17-18] is evaluation of bias temperature instability (BTI) without recovery induced error using the high resolution techniques described in previous sub-section. Other major improvements include minimizing switching during measurements, omitting level shifting and isolating fails due to cell noise margin from access fails.

2.2.6 Other Possible Timing Sequences Related To SRAM Aging

In this work we chose to consider the impact of BTI on typical Write, the worst case condition occurs on a cell flip and then an immediate restore [12]. A simple extension to the PR-SR timing sequence would suffice to provide BTI evaluation instead of a FLF-R approach. Fig. 2.13 shows the ensuing timing sequence which is very similar to Fig. 2.9.

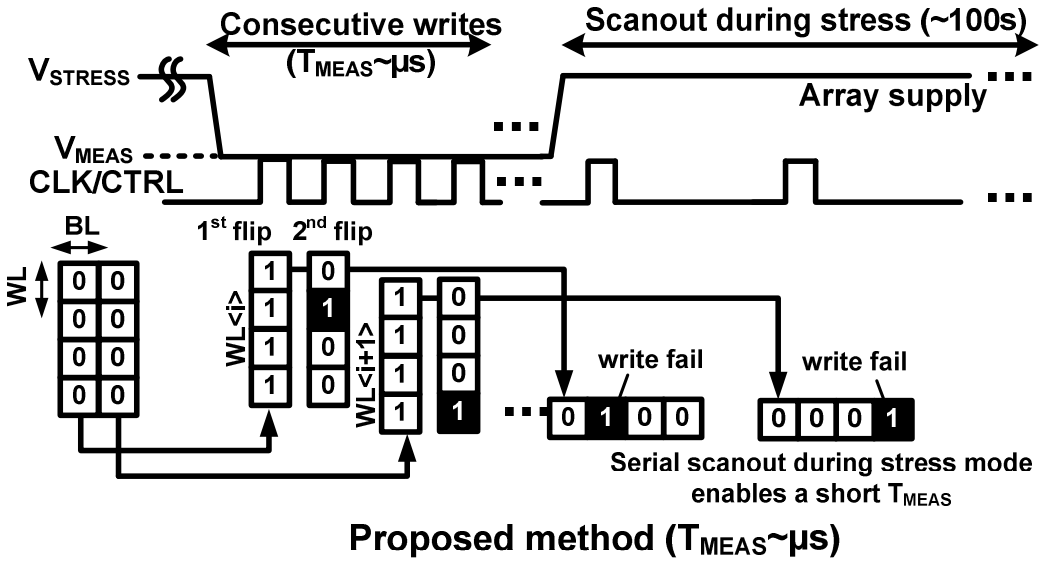


Fig. 2.13 BFR measurement sequence for a worst case for Write, consisting of long DC stress, followed by a flip and an immediate flip back to restore the cell data.

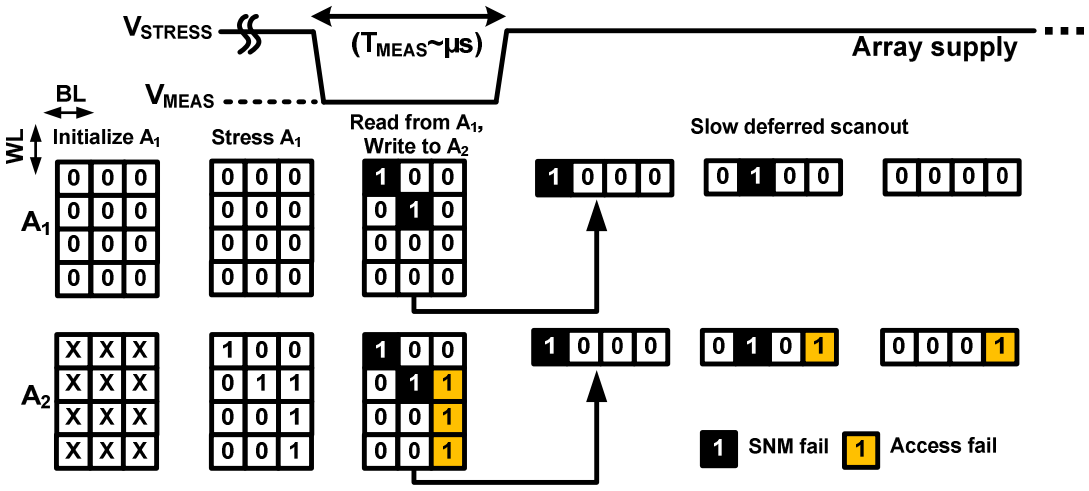


Fig. 2.14 Illustration for macro design to evaluate SNM and access time failures originating from stress in the SRAM cell separately. Two sub-arrays A_1 and A_2 are used. The supply for A_1 is switchable between V_{MEAS} and V_{STRESS}, while supply for A2 is fixed at V_{NOM}=0.9V for reliable operation. A_1 is first initialized to known data and then stressed. Next, supply is relaxed for measurement and fast unreliable read at V_{MEAS} is carried out from A_1 and written reliably into A_2. The flips stored in A_1 indicate SNM flips while those seen only in A_2 indicate access time fails while reading from A_1. Slow deferred scanout with A_1 supply at V_{STRESS} brings out these stored data off-chip.

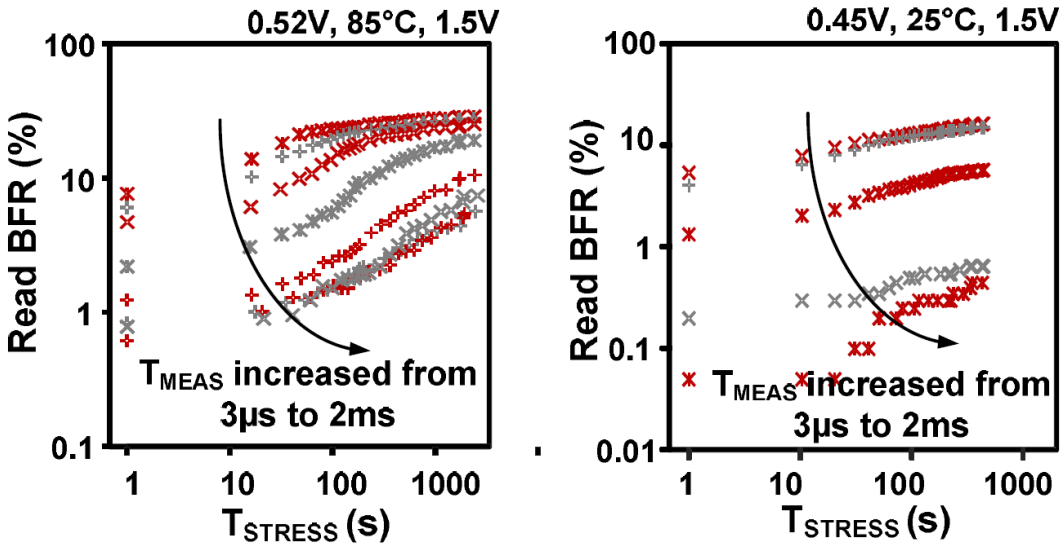
In this work, the aim was to isolate the BTI induced failures originating from the cell. For a more holistic application, we also want to look at access time impact due to reduced drive strength of the cell transistors. An approach is proposed in Fig. 2.14. Two sub-arrays A_1 and A_2 are used. The supply for A_1 is switchable between V_{MEAS} and $V_{STRESS,}$ while supply for A2 is fixed at V_{NOM}=0.9V for reliable operation. A_1 is first initialized to known data and then stressed. Next, supply is relaxed for measurement and fast unreliable read at V_{MEAS} is carried out from A_1 and written reliably into $A_{2..}$ The flips stored in A_1 indicate SNM flips while those seen only in A_2 indicate access time fails while reading from A_1. Slow deferred scanout with A1 supply at V_{STRESS} brings out these stored data off-chip.

2.3 32nm Test Chip Aging Data

2.3.1 Read Failure Data

Fig. 2.15 shows read BFR with stress time at different T_{MEAS} showing expected degradation at (a) 85°C and (b) 25°C. Over T_{STRESS}=2000s, with T_{MEAS} kept at 3µs, the BFR rises by around 10 times. Fig. 2.16 plots BFR after T_{STRESS}=10s at different T_{MEAS}. The same color curves are for repeated runs for the same chip to validate the repeatability and this is done for two chips, x and y. In general, we observe that BFR saturates consistently beyond 100µs. For low BFR, the trend is much more irregular so we use a lower VMEAS in right columns. Without using the proposed techniques, T_{MEAS} is more than few milliseconds, inculcating errors of as much as 10-100X in BFR. Fig. 2.17 shows the effect of BTI on measured V_{MIN}. Over T_{STRESS}=2000s, and T_{MEAS} =3µs, V_{MIN} changes by an amount close to 25mV. Also, by ensuring an at-least

three decade smaller T_{MEAS}, the proposed method alleviates 30mV error from the conventional methods. Note that measurements of V_{MIN} required external supply changes as shown in Fig. 2.10 leading to larger time between measurement samples. We were limited to a supply step of 1mV and a target BFR of 0.5% ensured the error due to limited resolution to relatively small.



(a) (b)

Fig. 2.15 Read BFR degradation with different T_{MEAS}. BFR at 0.52V, 85°C (upper panels) and 0.45V, 25°C (lower panels). The minimum T_{MEAS} possible by our test setup in order to cover the whole array at T_{CYCLE}=10ns is 3µs (20% allocated time for supply switching). A high BFR range (e.g. >0.1%) was chosen to obtain a smooth BRF curve.

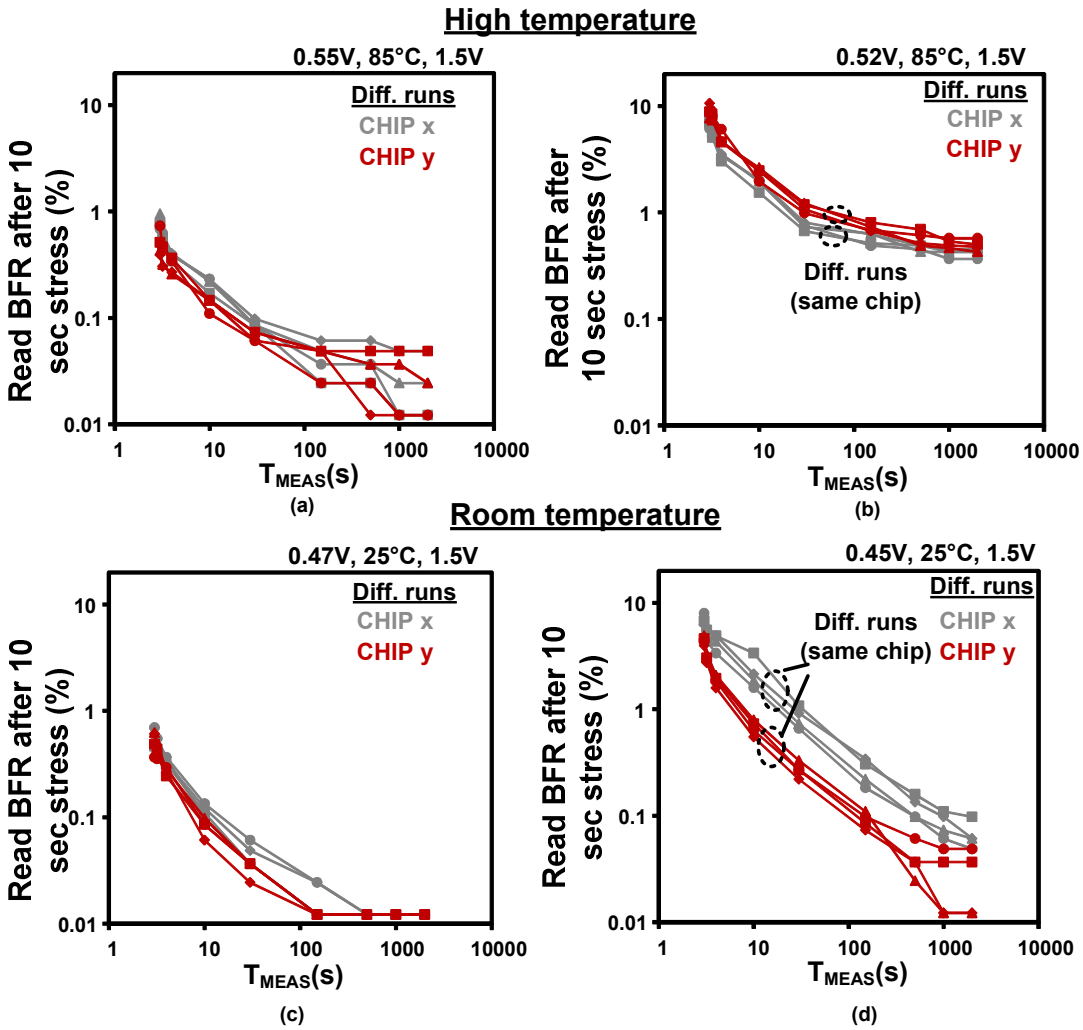


Fig. 2.16 Read BFR degradation with different T_{MEAS}. The minimum T_{MEAS} possible by our test setup in order to cover the whole array at T_{CYCLE}=10ns is 3µs (20% allocated time for supply switching). A high BFR range (e.g. >0.1%) was chosen to obtain a smooth BFR curve.

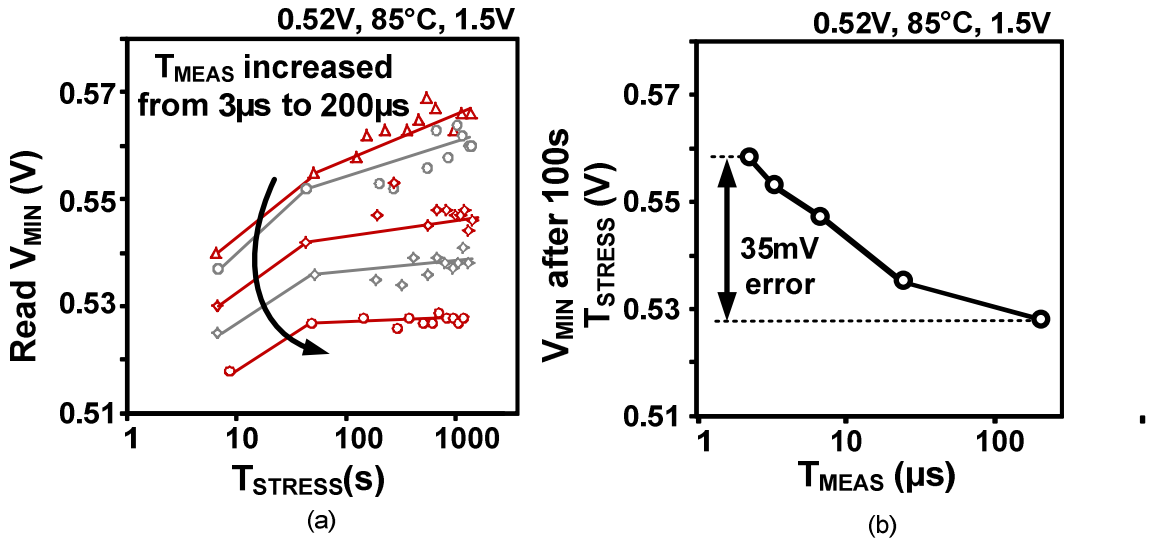


Fig. 2.17 (a) Read V_{MIN} versus T_{STRESS} for different T_{MEAS}. (b) Read V_{MIN} after a 100s stress period as a function of T_{MEAS}.

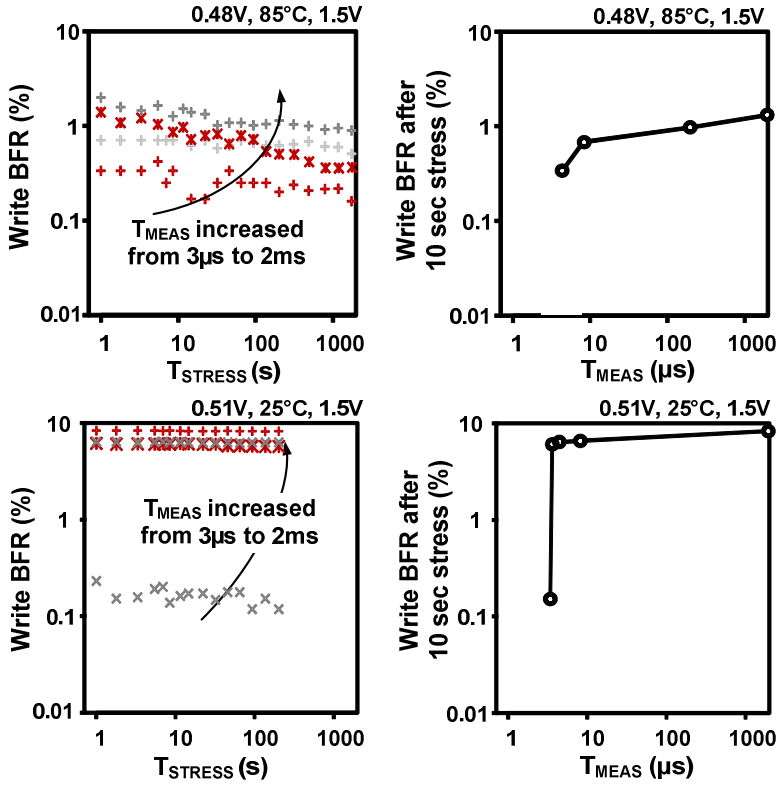


Fig. 2.18 Write BFR degradation at 0.48V, 85°C (upper panels) and at 0.51V, 25°C (lower panels). Compared to read case in Fig. 8, lower sensitivity seen towards T_{STRESS}, and higher towards T_{MEAS}. Actual stress voltage undisclosed due to confidentiality.

2.3.2 Write Failure Data

Fig. 2.18 shows the BFR evolution for write case using the test sequence in Fig. 2.6. As expected, there is an improvement seen in BFR. The sensitivity to T_{MEAS} was found to be much greater than the read, especially at 25°C, and BFR is seen to drop sharply below 3.6µs. At 85°C for T_{STRESS}=2000s, the BFR drops 2x, pointing to lower sensitivity overall to BTI stress, compared to read case. Overall, at least a 100X error in BFR is obtained from the conventional methods due to the smaller T_{MEAS}.

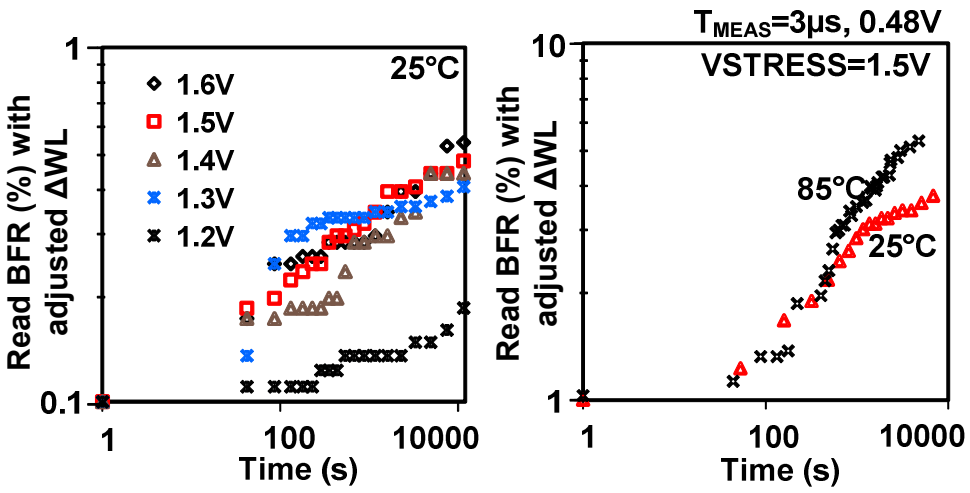
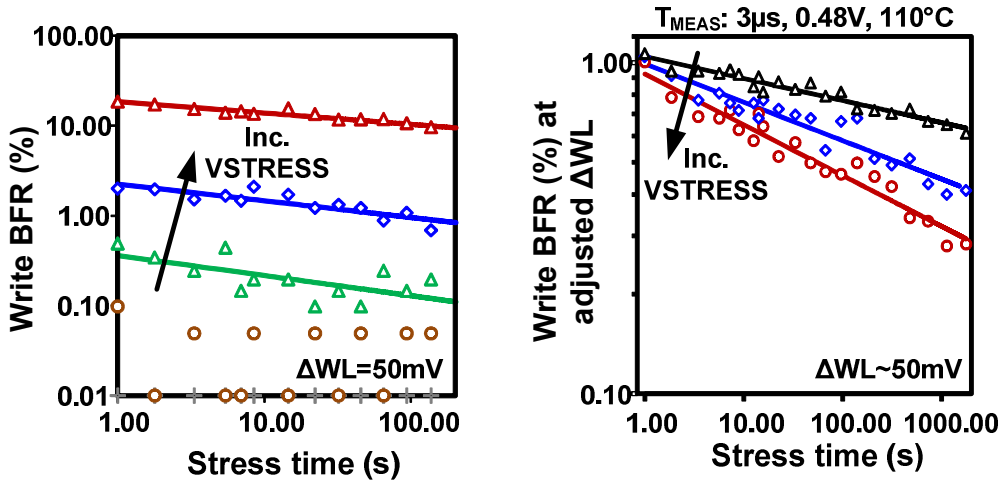


Fig. 2.19 Read BFR with different (a) V_{STRESS} and (b) Temperatures. The initial BFR varied a lot from chip to chip and to offset that WL voltage was adjusted to keep the BFR same initially at 1%

2.3.3 Different VSTRESS Results

Fig. 2.19 and 2.20 show Read and Write BFR, respectively, at different stress conditions. Note, that due to floating substrates in the SOI process employed in this work, switching between stress and measure modes may lead to some body-coupling error in the first stress time. However, as long as V_{STRESS} is same, this would be common across different measurement times, so the comparisons drawn out between

the proposed and conventional approaches are quantitatively correct [38]. While these are the artifacts of SOI process, they would be absent in regular bulk process.



(a) (b)

Fig. 2.20 Write BFR with different V_{STRESS}. Body coupling effect dominating: reverse behavior seen. It is inconclusive whether the trends predicted are due to BTI or body coupling effect due to floating bodies in this process.

2.3.4 Test Chip Feature summary

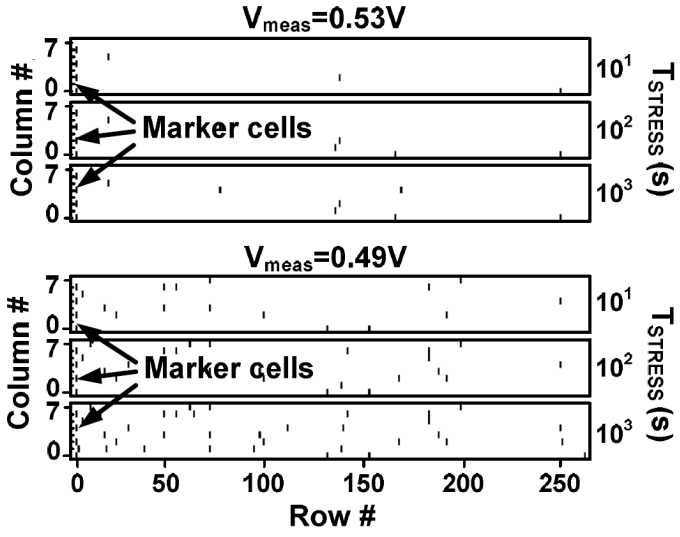
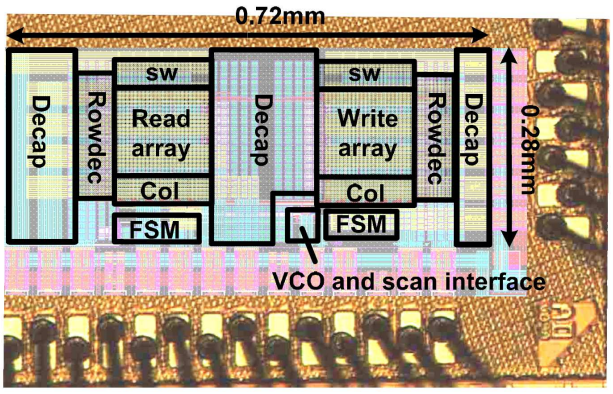


Fig. 2.21 Spatial distribution of read failures. The array initialized with data '0' in all cells. The black dots correspond to fail cells. No significant spatial correlation observable for fail bits.

Spatial distribution of the read flips in Fig. 2.21 show no apparent correlation. Fig. 2.22 shows the micro-photograph and summary of the test chip fabricated in this 32nm SOI process.



Process	0.9V 32nm HKMG SOI
Bitcell size	0.898 x 0.269μm^2
Ckt dim.	0.72 x 0.28mm^2
Density	2x32kb RD/WR macros
T_{MEAS}	>3μs @ 0.5V, 100MHz >1μs @ 0.7V, 500MHz
V_{MEAS}	0.45-0.9V
V_{STRESS}	0.9-2.5V

Fig. 2.22 Test chip micro-photograph and feature summary. Measurements were automated using a Labview™ controlled data acquisition board.

2.4 Conclusion

Recovery free evaluation of BTI in SRAM is challenging due to massive data to be captured within a few microseconds. This work provides a methodology to remove the noise in SRAM measurements due to BTI recovery. We incorporate two techniques, namely PR-DR and FLF-R, for read and write respectively on a test chip in 32nm HKMG SOI. Small T_{MEAS} of around 3μs at 0.5V, yields 35mV accuracy in read V_{MIN} and 10X accuracy in BFR over conventional approaches.

Chapter 3

An Array-Based Chip Lifetime Predictor Macro for Gate Dielectric Failures in Core and IO FETs

3.1 Introduction

Device reliability mechanisms such as bias temperature instability, hot carrier injection, and gate dielectric breakdown have become pressing concerns in scaled technologies. While parametric shifts due to the former two can be mitigated using frequency guard-banding or circuit adaptation [39-40], such techniques are ineffective against the more catastrophic dielectric breakdown where even a single instance in a chip can cause an outright system failure.

Gate dielectric breakdown can be an outcome of various kinds of stress patterns. At one end of the voltage/stress time tradeoff is the Electrical OverStress (EOS) and ElectroStatic Discharge (ESD) phenomena, which last millisecond to nanoseconds at a very high voltages, ranging from hundreds to thousands of volts [41].This particularly affects Input Output (IO) transistors during manufacturing and handling stresses and can lead to either outright or latent damage which may or may not pass through product screening. At other end is the much slower operating condition Time Dependent Dielectric breakdown (TDDB) which is of concern in both core and IO transistors.

Particularly with TDDB, optimizing the fabrication process and using proper operating conditions based on accurate lifetime predictions is the most practical and effective approach. A detailed description of a conventionally employed lifetime prediction flow is described below.

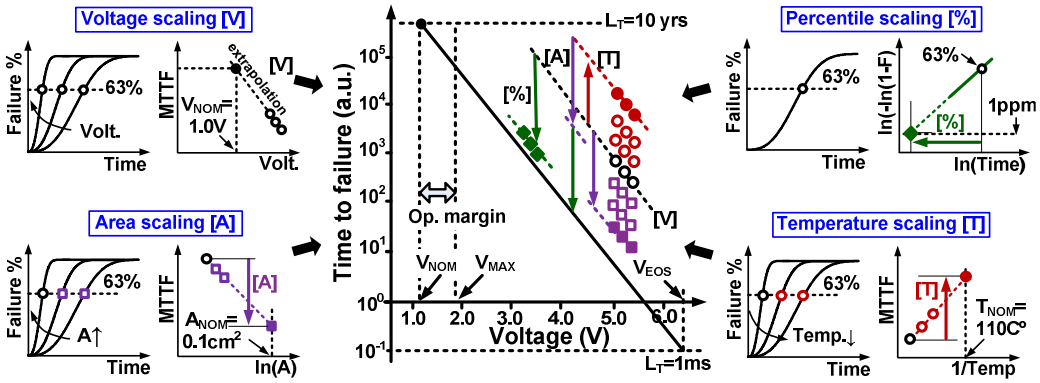


Fig. 3.1 Chip lifetime projection for TDDB based on accelerated stress involves mass data collection (e.g. up to 1000's of samples per MTTF data) to make voltage, percentile, area, and temperature projections to actual product usage conditions. The open symbols represent measurable values at accelerated condition, while the solid symbols represent the projected value.

The first step is to obtain Cumulative Distribution Function (CDF) of TTF from a massive dataset, as shown in Fig. 3.1(a). The 63% point is used to define the Mean-Time-To-Failure (MTTF). The MTTF vs voltage curve follows a power law and this is seen as a straight line fit in a log-log scale. An important property of TDDB is the failure rate follows a power law, with time constant known as Weibull slope, β. The CDF if plotted on a Weibull scale (ln(-ln(1-CDF)) is useful to graphically observe the tail cell behavior. Typically, we are want to extrapolate the results to one bad chip in a million chips (1ppm). The larger the number of samples measured, the higher the accuracy of the TTF corresponding to 1ppm. We also want to make reliable projection from the measurable small area to typical gate area in a microprocessor die

typically 0.1-1cm^2, and thus multiple points are needed for MTTF at different area. Finally, in order to further cut down the measurement time, temperature stress is employed and projection is made to operating temperatures of 25°C or 110°C. To put the different scaling projections together, we start from the voltage scaling, and one by one add the corresponding acceleration due to percentile, area and temperature and finally end up with the black solid line. The abscissa at which this solid line meets a target lifetime of say, ten years, gives the maximum operating voltage, $V_{MAX;}$ and the difference from the nominal voltage, $V_{NOM,}$ gives the operating margin.

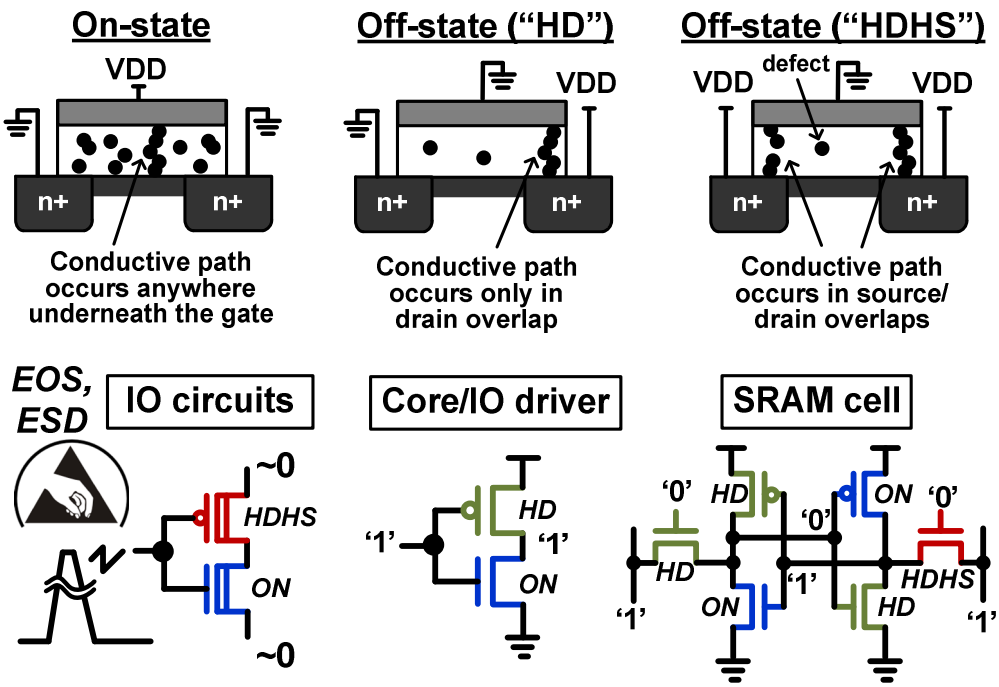
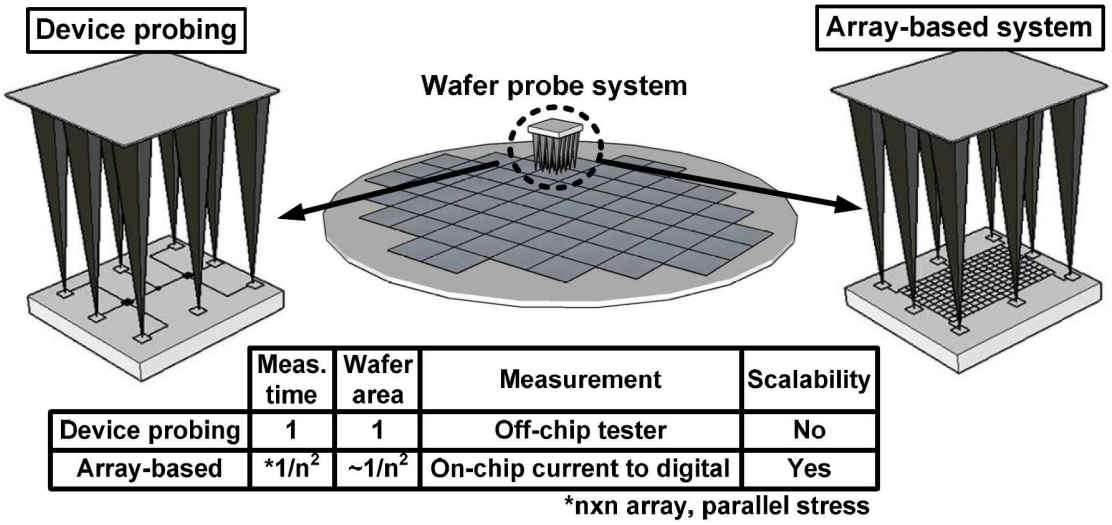


Fig. 3.2 Different occurrence of gate dielectric failure. While 'ON' and 'OFF-HD' cases are most prominent, 'OFF-HDHS' is also seen in certain cases such as SRAM access devices.

TDDB also gets exhibited in different profiles based on the biasing during operation as well as dielectric thickness. Fig. 3.2 shows the different TDDB modes affecting common digital circuits. While TDDB in transistor gates has traditionally

been studied under inversion-mode stress conditions, ultra-thin dielectrics can also suffer breakdowns in the High Drain, High Source (HDHS) and High Drain (HD) OFF-state modes [1, 42] when the channel is not inverted [1], [43], [44]. This OFF-state stress becomes particularly problematic under excessively high drain biases, such as those occurring during burn-in screening, or in certain interface circuits where a transition is made into a higher voltage domain. An important concern might be earlier failure in circuits such as SRAM access devices that are exposed to an off-state stress for most of their lifetime. On the other hand, IO devices are traditionally resistant to TDDB due to employment of thick t_{ox} devices for robustness against ESD/EOS mechanisms. However, TDDB margin targets have become an issue with extensive use of high-voltage IOs and high-power CMOS devices at interface circuits in system on chips.

From above it becomes clear that a comprehensive lifetime prediction mandates massive statistical data collection at different failure modes for accuracy. Given the need for up to thousands of samples to correctly define a *single* Mean-Time-To-Failure (MTTF) value, traditional device probing quickly becomes cumbersome. This is illustrated in Fig. 3.3. For a typical wafer probe system, the conventional approach has been to individually tap out all the terminals of a FETs and externally bias them. On the other hand, in an array based testing system, on-chip current to digital conversion provides a convenient and efficient way to parallel stress a large number of devices.



	Meas. time	Wafer area	Measurement	Scalability
Device probing	1	1	Off-chip tester	No
Array-based	*1/n^2	~1/n^2	On-chip current to digital	Yes

*nxn array, parallel stress

Fig. 3.3 Array based approach is an efficient way to carry out aging measurements compared to conventional probing. In the example shown above, device probing using off-chip tester with 8 probes, can test two devices at a time. On the other hand, in the array based system, using the same resources, a n by n array of devices could be tested out.

A previous characterization array for TDDB [20] only considered ON-state stress in core transistors which is not enough to obtain an accurate picture of system lifetime. A combined lifetime prediction methodology is needed to take into account different modes in tandem with their predicted time to failures. In this paper, we propose an array-based Chip Lifetime Predictor (CLIP) macro for efficiently collecting failure statistics under various accelerated stress conditions including ON-state and OFF-state stress modes for both low voltage core and high voltage IO devices.

In the next section, we delve into the CLIP macro design and overall test strategy. Section III and IV describe the stress cell designs with measured statistics, along with the lifetime prediction methodology using the CLIP framework. Finally, we give a conclusion in section V.

3.2 Macro Design and Test Strategy

3.2.1 CLIP Macro Design

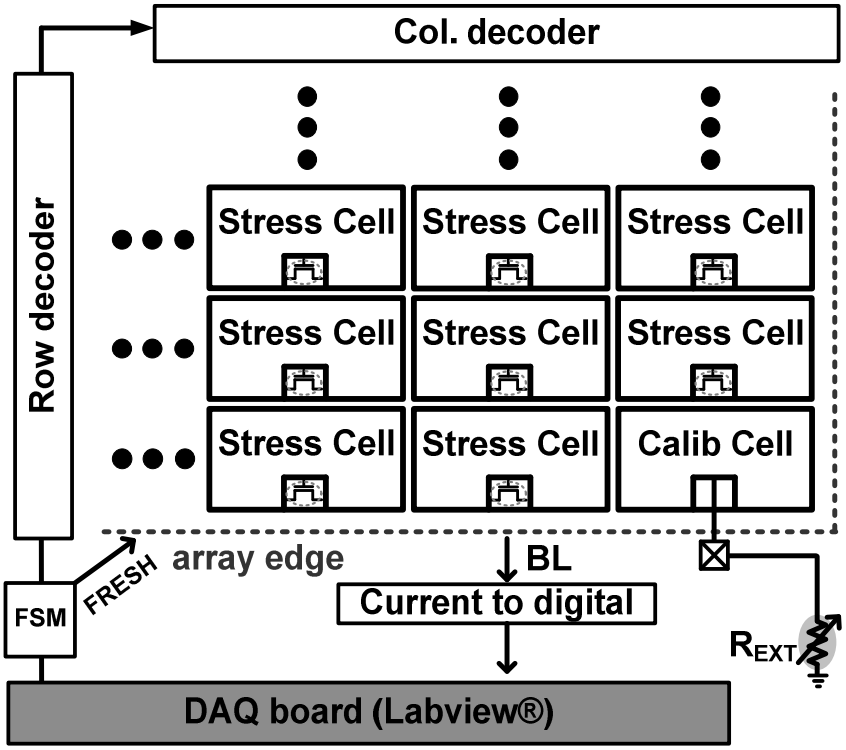
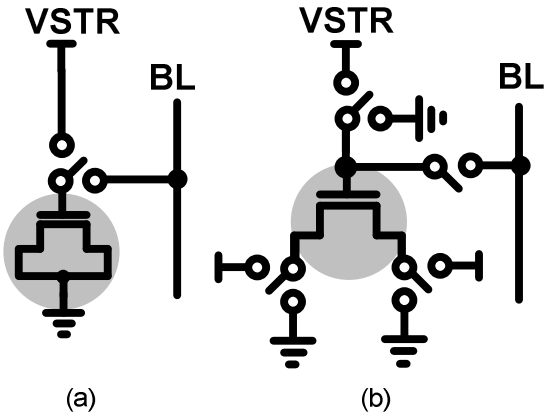


Fig. 3.4 General concept of an array-based Chip Lifetime Predictor (CLIP) macros. The column and row peripherals provide a "one hot" functionality for measuring one cell at a time, while stressing the rest in parallel.

The basic framework of the proposed CLIP macro is an array based statistical collection setup that can stress the DUTs in parallel while taking fast serial measurements controlled by a convenient scan-based interface (Fig. 3.4). This feature reduces the test time and test silicon area by a factor proportional to the number of DUTs. The gate terminal of the selected DUT is connected to the shared BL for I_G measurements. The pre-charged BL gets discharged and the progressive TDDB in the form of I_G is converted to a count by an on-chip current-to-digital converter.

The critical part is a flexible stress cell design that can be used for evaluation of the different OFF and ON-state TDDB modes with programmable control. Two different flavors of flexible stress cells are needed for IO and core cases as will be discussed in the next section. As shown in the abstraction in Fig. 3.5, the underlying principle is to connect each DUT terminal to a stress voltage using on-chip switches rather than a hardwired inflexible connection. Flexible stress conditions used for the DUT cells have been tabulated in Fig. 5(c) and will be described in the next section.



Mode	VG	VD	VS
OFF-HDHS	0	VSTR	VSTR
OFF-HD	0	VSTR	0
OFF-VST	VDD-VSTR	VDD	0
ON	VSTR	0	0

(c)

Fig. 3.5 Abstraction of different kinds of stress cells supported: (a) Conventional [7]; (b) Proposed flexible DUT; (c) Different flexible stress conditions.

3.2.2 Current to digital conversion

Reliability engineers employ both ‘hard’ and ‘soft’ increase in dielectric conduction for characterizing TDDB. It is generally accepted that thick t_{ox} devices undergo sudden hard breakdowns while thin t_{ox} devices show more slowly progressing breakdowns. We, therefore, employ two variants for current to digital conversion in this work. Fig. 3.6(a) shows the Current to Count Converter (CCC) to facilitate soft breakdown evaluation for the core case similar to the one used in [20]. We made some improvements for better noise immunity at the output probe node for

example adding a dual reference comparator and avoiding any switching activities during the probe node evaluation. Fig. 3.7(a) provides the measured calibration curve to convert the obtained count to $R_{DUT.}$

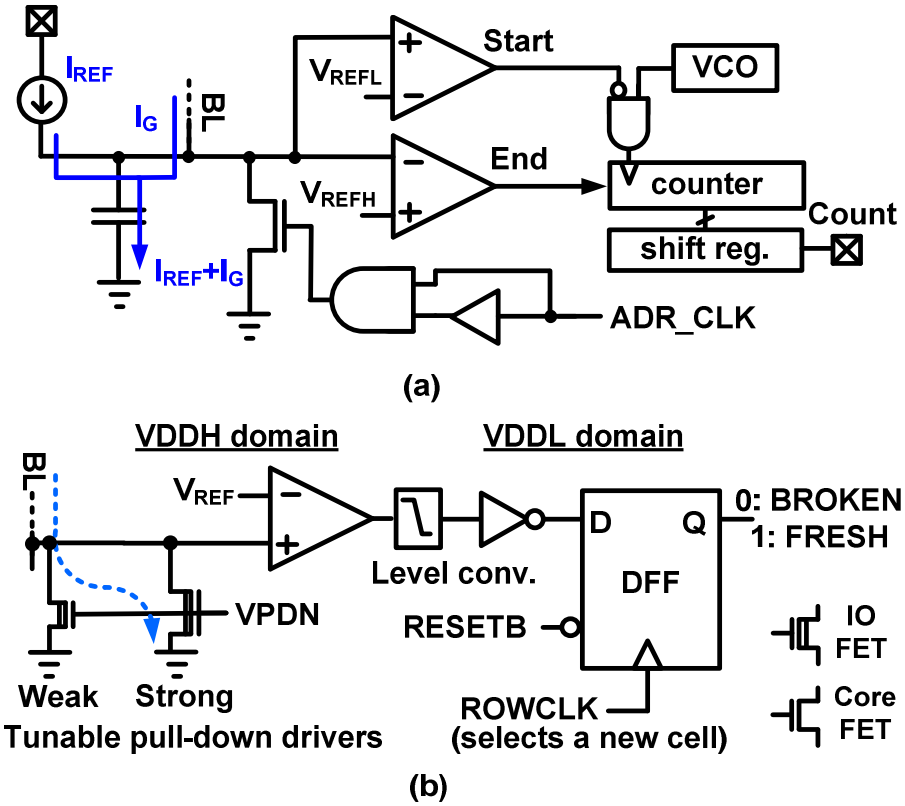


Fig. 3.6 Two flavors of current to digital blocks used (a) CCC for soft breakdown in core FETs. (b) CBC for hard breakdown in IO FETs.

Considering the high t_{ox} values in IO devices, as well as based on our preliminary findings on the core case [8], we did not expect to see progressive behavior in breakdown in our test setup. Therefore, a major simplification for higher timing resolution and ease of measurement can be done in the form of Current to Binary Converter (CBC) scheme in Fig. 3.6(b). The basic idea is to form a resistive divider between one or both of its two pull-down devices and the gate resistance of the selected DUT. If the BL node voltage falls below the V_{REF} bias, then the pre-

discharged comparator output goes to '1', and this change is latched in a DFF when the ROWCLK signal falls. A FRESH=0 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 BL 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. Using this binary (i.e., two state) approach we obtain the same amount of information—simply the time to the sudden breakdown. However, the one bit result can be recorded by a data acquisition board more quickly than a sixteen bit count result from CCC. This improves the timing resolution of the measurements, meaning there is less time between consecutive readings in each cell. Also, many researchers base their TDDB findings upon the time to the first observed breakdown—be that soft or hard [1], [45]. This compact system is sufficient to record that first event. Thus, elaborate tracking using CCC was not needed.

3.2.3 Calibration

Replica stress cells, called "calibration cells" were embedded directly in the TDDB array (Fig. 3.4). 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 resistances, and measurements were run in the calibration cell. The pull-down 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. 3.7(b)).

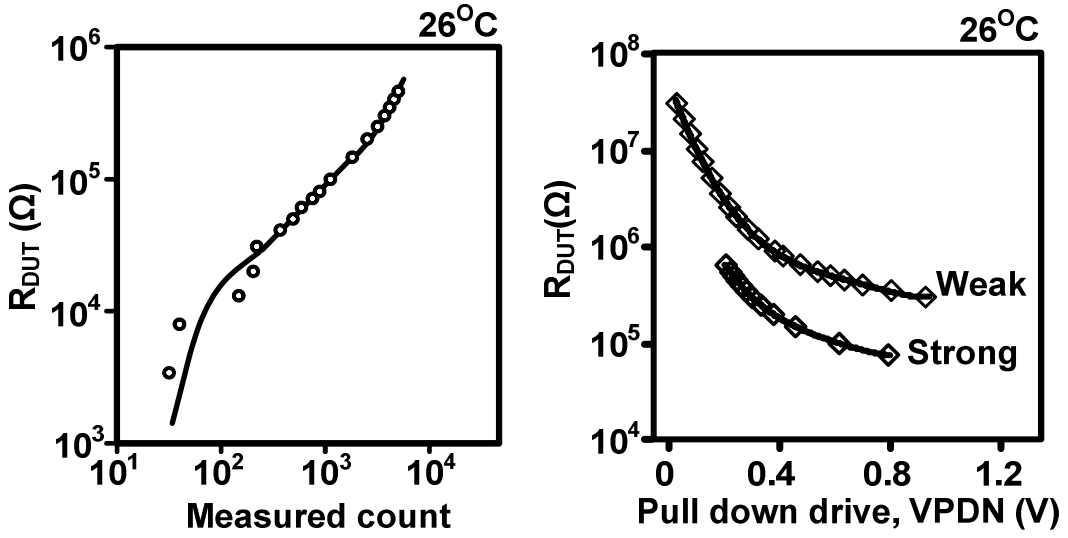


Fig. 3.7 Calibration curves using the two current to digital converters. (a) CCC case, and (b) CBC case.

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 BL 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 FRESH 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 calibration cell also served useful as a marker or reference cell during array operation for debugging, by driving the pad from supply directly. During the test, we read out a FRESH value of 0 corresponding to the marker cell, while the rest of the array read out a FRESH=1, providing a real time check on the current address in measurement.

3.3. Core Device Breakdown Cell

3.3.1 Stress cell design

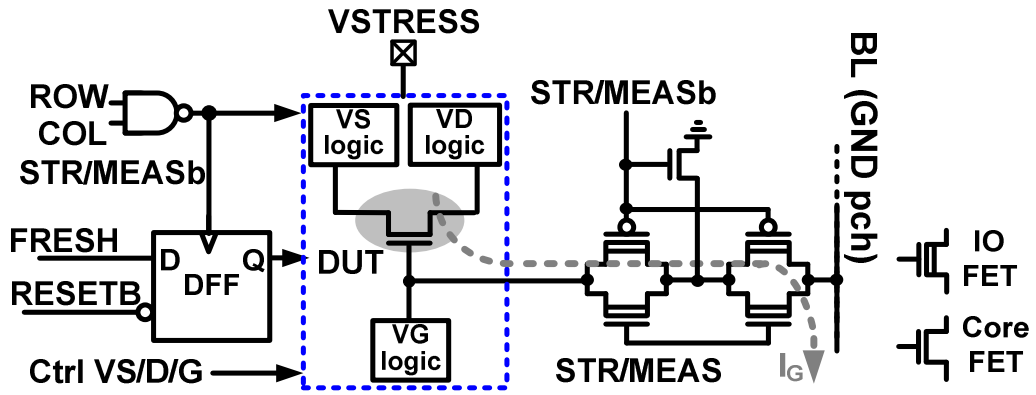


Fig. 3.8 Proposed DUT cell for core device breakdown. All FETs except DUT are thick t_{ox} devices.

Details of the core FET stress cell are shown in Fig. 3.8. The terminal voltages of the DUT are separately controlled by the STR/MEASb and VS/D/G control signals to provide programmable control of different flexible stress modes. If 'FRESH'=0, the stress is gated off to prevent excess current during the long stress experiments. A timing logic selects cells in a manner that prevents over-shoot transients on the DUT and seepage of stress voltages to non-DUT circuitry.

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 [1], [46], [47]. The HCI component becomes a more significant problem when channel lengths are scaled down, leading to increased lateral electric fields and the possibility of punch-through. 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 [1], [44], [46]. 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 [1]. The idea is to drive a high stress across the the gate and drain, but keeping the drain to source voltage same as operating voltage as shown in Fig. 3.5(c).

Each NMOS device under test had a width and length of 2 μm in order to be consistent with our previous TDDB array [20]. 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 [1].

3.3.2 Core ON-state stress results

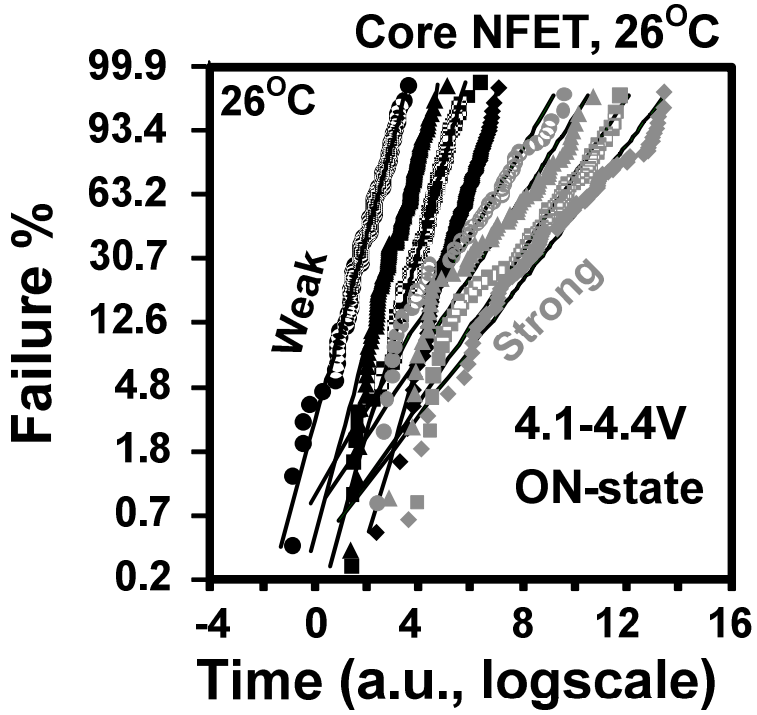


Fig. 3.9 (a) Effect of pull down strength in CBC scheme with VPDN=0.35V in the ON-state stress.

We first measured Time-To-Failure (TTF) 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. 3.9 shows that the Weibull CDFs of these ON-state results were well-behaved for stress voltages ranging from 4.1 V to 4.4 V, as expected. We also 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 [48]. 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. The Weibull slope [%]) for the 4.2 V curve was 1.444, matching well with the 1.443 value from our previous work [20]. Note that the actual TTF values must be kept confidential according to the manufacturer.

3.3.3 Core OFF-state stress resultss

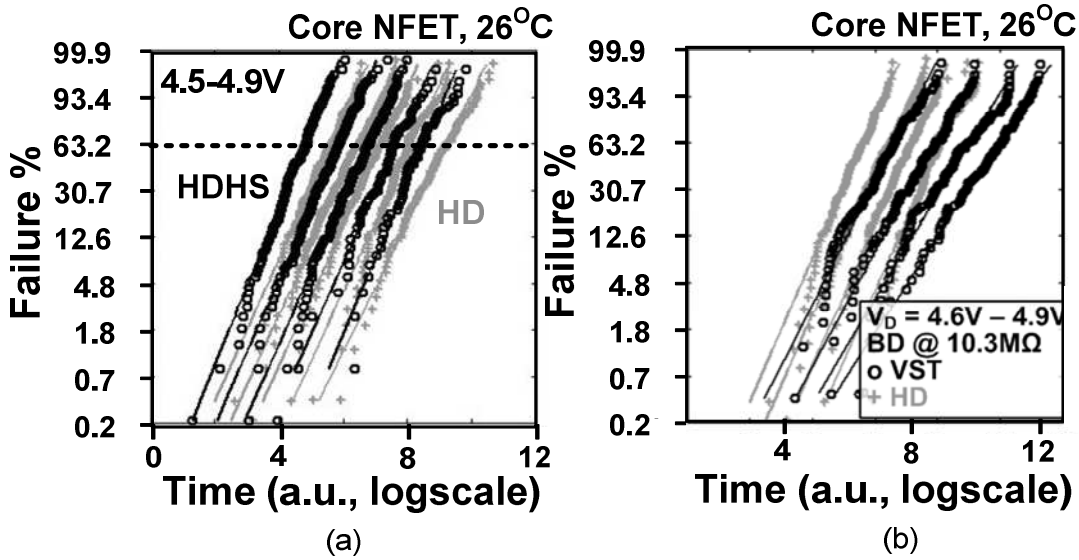


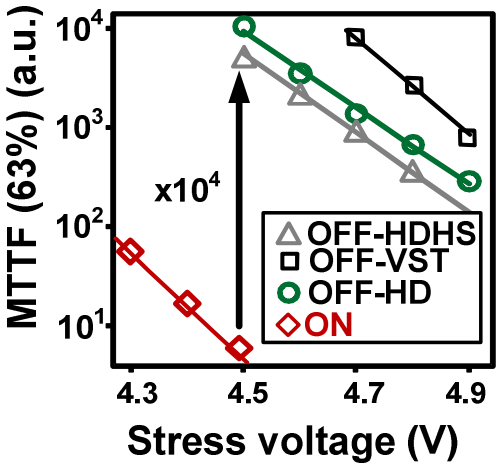
Fig. 3.10 (a) OFF-HDHS and OFF-HD (b) Off-state voltage-splitting (VST) and high drain/0 V source (HD) Weibull plots.

In Fig. 10(a), 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 TTF 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 [%] values for these OFF-state 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 above theoretical value.

Fig. 3.10(b) illustrates VST results, along with the HD stress findings. The latter has larger lateral electric field and should result in faster breakdowns, which is apparently due to a vertical field contribution from GIDL. We also observed a lower [%] 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.



(a)

Modes	[%]	[V]
Core OFF-HDHS	1.56	43.5
Core OFF-HD	1.585	41.63
Core OFF-VST	1.05	52.43
Core ON-state	1.44	51.16
IO ON-state	2.71	44.46

(b)

Fig. 3.11 (a) Relative voltage scaling in different modes (b) Relative comparison of voltage acceleration time exponent [V] and and Weibull slope [%] in different cases.

In Fig. 3.11(a), we compare the voltage acceleration characteristics for several off-state stress configurations and tabulate the slopes in Fig. 3.11(b). The HDHS had the lowest TTF due to stress on both ends of the channel. The VST results show the longest TTF. This is again presumably due to the elimination of GIDL-induced degradation. The power law exponent for the HD stress voltage acceleration was 51.16, 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. 3.7(b). More work is needed to verify the precise behavior of OFF-state degradation's relationship with voltage. Finally note that the TTF for all off-state stress conditions was around 4 orders of magnitude higher than that seen in inversion mode 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 TTF [1], [43].

3.4. IO Device Breakdown Cell

3.4.1 Stress Cell Design

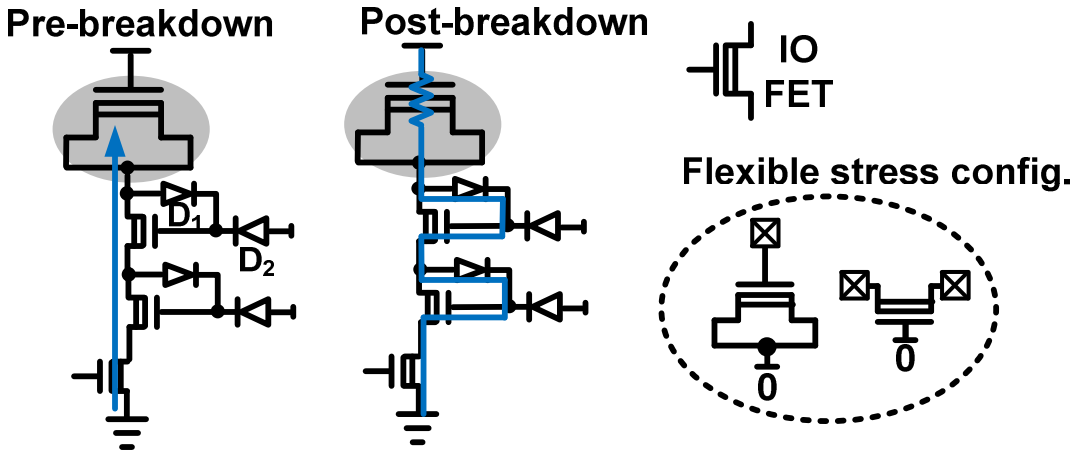


Fig. 3.12 IO stress cell in pre- and post-breakdown modes. No thicker t_{ox} devices are available so a blocking circuit was used to protect non DUT devices.

The higher stress voltage (3-4 times the IO supply) and lack of a thicker t_{ox} device complicate the design of the IO stress cell. As shown in the abstraction in Fig. 3.12, we add a blocking circuit with dynamic self-biasing through use of diodes. During the first part of the test, when the DUT is not broken (typically R_{DUT} above 1MΩ), 0 level is allowed to the DUT to provide full stress field across the DUT. A careful choice of VB is needed to ensure forward biasing diode D_1. Thus, when measuring in such a state, the intermediate devices in the current path are biased in triode keeping the parasitic resistances small. During the second part of the test, when the device breaks (suddenly or progressively) and the R_{DUT} goes below 10-100kΩ, D_2 becomes forward biased and D_1 is reverse biased. Thus, the voltage that seeps from the DUT to the

non-DUT parts sees a drop of $V_{GS, M1}+V_{T, D2}$. Overall, a stack of two blocking circuits (single stack shown in Fig. 3.12 for simplicity) was sufficient to stress the cell up to 4 times nominal supply (up to 10V). Fig. 3.13(a) shows the detailed schematic of the IO stress cell. Initially DFF is set to 1. Suppose ROW=0, COL=1 (the cell is unselected and undergoes parallel stress). In this mode, q=1 and STR/MEASb=1, turning on M_3 driving a gnd to n1. When ROW=COL=1, STR/MEASb =0, turning on TG_1 and turning off M_1 and current corresponding to R_{DUT} flows through BL. In order to prevent any floating nodes and bias the blocking circuit, a weak diode stack was used to mimic I_1. Note, unlike the core FET stress array, it is not possible to simply gate off stress to test just a small chunk of the array. A less robust technique that can be employed is to assert an FRESH = 0 to a cell to be left unstressed. During stress cycle, node (1) level is determined by various leakages and we bias it by a weak diode pull-down to raise voltage at n1, few volts above 0, substantially slowing down TDDB in that cell. Note that the bodies of all NFET devices were tied to GND while a reverse bias breakdown of the source to body junction was ruled out as long as junction biases were kept under 10V. To provide the high value of VB and VSTRESS on-chip, IO pads consisting of stacked ESD diodes were mandated.

Fig. 3.13(b) shows the simulated node voltages during stress cycle at a range of possible R_{DUT}. Before breakdown, reliably a value of 0 reaches the DUT node (1). Once a breakdown occurs, FRESH=0 needs to be asserted to prevent large current contribution from the broken DUT. This raises node n1 to almost VSTRESS (~9V). However, the blocking action allows only 2.5V to seep through to lower nodes like

n5. Fig. 3.13(c) shows the distribution of measured post-breakdown R_{DUT} which shows a bimodal behavior and as shown in Fig. 3.13(d) there was no observable spatial dependence in resistance values. The post-breakdown R_{DUT} is a function of the breakdown position [49] which is a possible reason for the above observed distribution. A criterion of 50kΩ was chosen to cover the entire range possible of post-breakdown resistances. Thus, FRESH=0 is asserted once a 50kΩ DUT resistance is recorded.

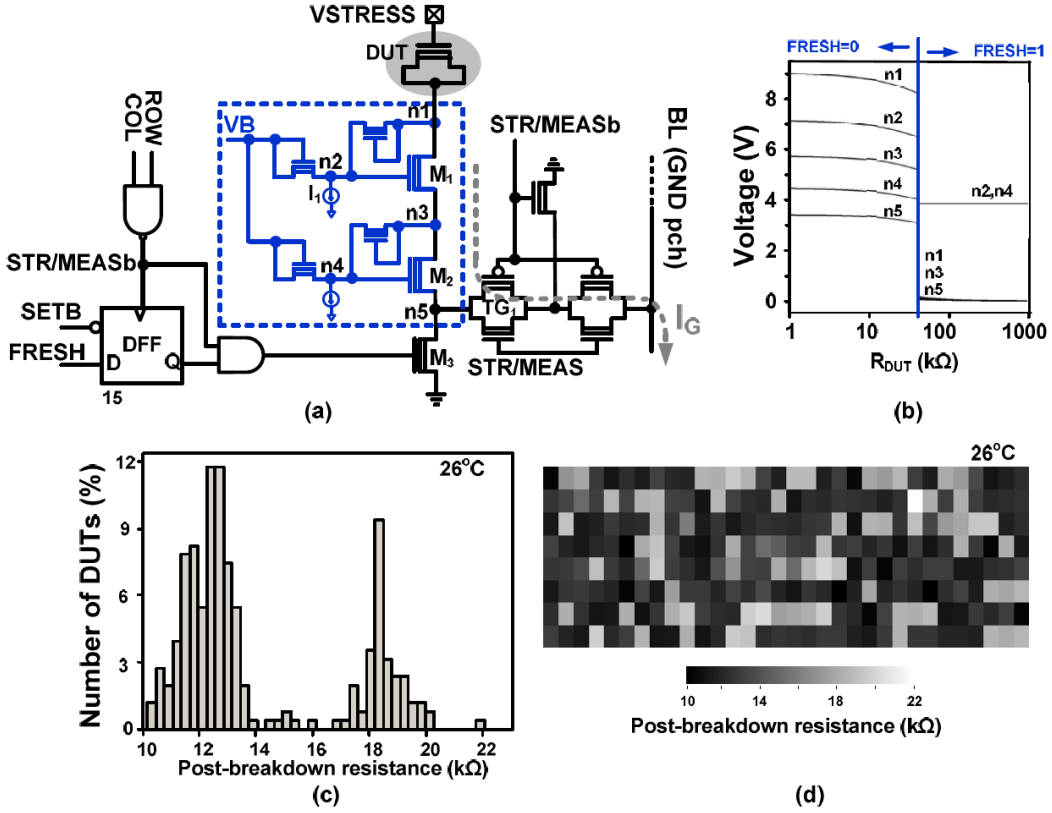


Fig. 3.13 (a) Detailed schematic of the proposed DUT cell for IO breakdown. (b) Simulation showing the various node voltages during stress cycle for a range of R_{DUT} before and after assertion of FRESH signal. (c) Histogram and (d) spatial plots of measured post-breakdown resistance. A hard breakdown resistance of 50kΩ sets the criterion for software to assert the appropriate value for the FRESH signal to cover the maximum possible post-breakdown resistance.

The previous section described a stress cell for flexibly characterizing both OFF and ON-state modes. However, a similar extension to IO breakdown cannot be supported as gating on or off VSTRESS is not possibly without incurring damage to the switches. Essentially, that means VSTRESS has to be hardwired to a cell. In order to obtain OFF-state data, we propose two different topology for the stress cell as shown in Fig. 3.12, with pad connections going to the gate for the ON-state while to the Drain/Source for the OFF-state case. However, it should be noted that for thick t_{ox} devices at the present technology nodes, the OFF-state mode is not expected to be much of an issue, since the edge tunneling currents causing OFF-state breakdown are much less in thick t_{ox} devices [1].

3.4.2 Measured IO Device Breakdown statistics

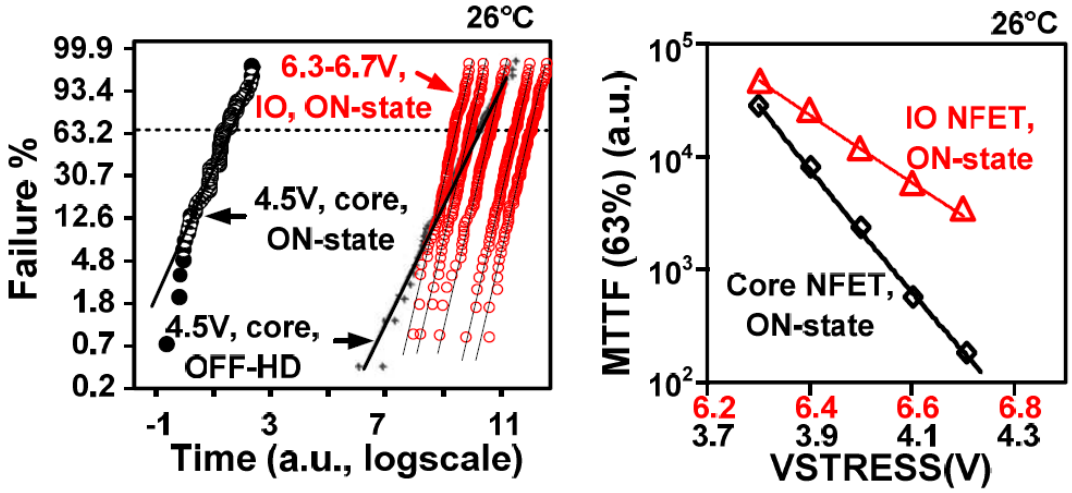


Fig. 3.14 (a) Measured breakdown data at different stress voltages for IO case. For comparison, a stress curve for the core case at 4.5V has been shown. (b) MTTF plots at different

In the IO breakdown test setup, all DUTs employed an area of 2x2 μm^2 for consistency. Preliminary tests indicated a target VSTRESS of around 6-7V for

measurable stress. So, a VB value was chosen to be 3.5V. Measured Weibull plots for IO DUTs for a range of stress voltages is shown in Fig. 3.14(a). Well-behaved results for all the stress conditions were obtained. A slight non-linearity does show up below 1% results. The MTTF was around 15-20 times of the time to first fail. The [%] value was 2.7 at 6.7V stress, which can be compared to 1.44 for the core ON case at 4.6V. This makes sense as [%] is proportional to t_{ox} and was verified from the values with similar t_{ox} in older technologies. The actual values of t_{ox} are kept confidential as per the agreement with the foundry. Overall, this means that scaling from 63% to 1ppm for the core case is projected to be 100X larger than the IO case. In Fig. 3.14(b), we plot the voltage acceleration characteristics for IO breakdown. This shows a power law exponent of 44.46 compared to a value of 51.16 for the core ON-state breakdown. The steeper slope for the core case translates into a 20X MTTF difference due to [V] scaling. MTTF for different temperatures are shown in Fig. 3.15(a). Both core and IO FETs show Arrhenius trend in the measured regime. Spatial map shown in Fig. 15(b) of the individual cell's TTF shows no obvious correlation.

3.5. Lifetime Estimation Results Using CLIP Methodology

Fig. 3.16 shows the applied CLIP methodology for different stress profiles and gate types in tandem. Measured data is denoted by open symbols for distinction from extrapolated results shown using solid symbols. We start with voltage scaling results (Fig. 3.11(a) and Fig. 3.14(b)) measured at 26°C (black) and 110°C (red). We didn't have exact gate area information in a state of the art microprocessor. Therefore, we employ area scaling from the $2x2\mu m^2$ stress cell to the effective gate areas

estimated at 0.1cm^2 and 0.01cm^2, for core and IO transistors, respectively. The failure percentile curves presented earlier (Fig. 3.14(a)) were extrapolated down to 1ppm using a Weibull fit to low percentiles. We also normalized the time scale assuming, a duty cycle of 50% between OFF and ON-state. The final result of the above four scalings is shown with green symbols. Finally, we proceed to apply the appropriate voltage acceleration factor for IO and core cases to obtain the maximum voltage, V_{MAX} for a given operating lifetime. As seen in Fig. 3.16, the IO devices meet the lifetime requirement with a guardband of 1.1V and core transistors by 0.2V. Overall, the IO devices were 100X more robust than core devices, at respective $V_{NOM.}$ The chip microphotographs and summary of the core and IO CLIP arrays are given in Fig. 3.17.

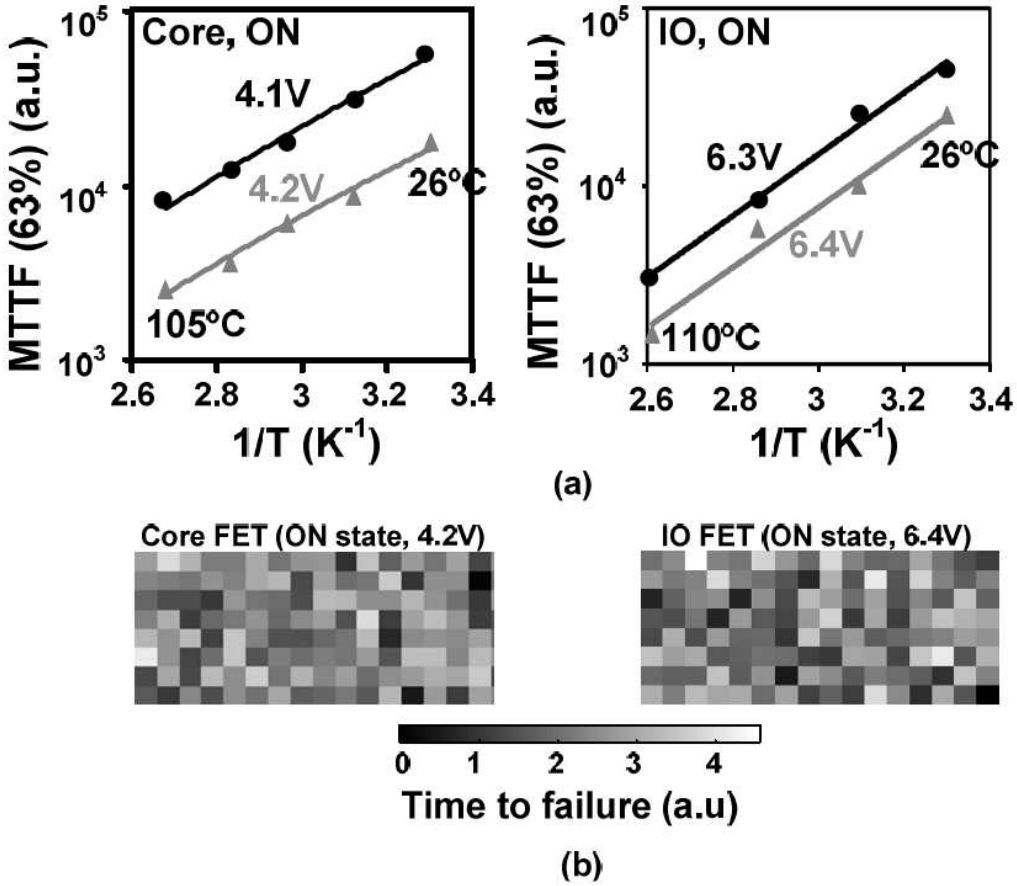


Fig. 3.15(a) MTTF for different temperatures. Both core and IO FETs show Arrhenius trend in the measured regime. (b) Spatial map of individual cell's time to breakdown.

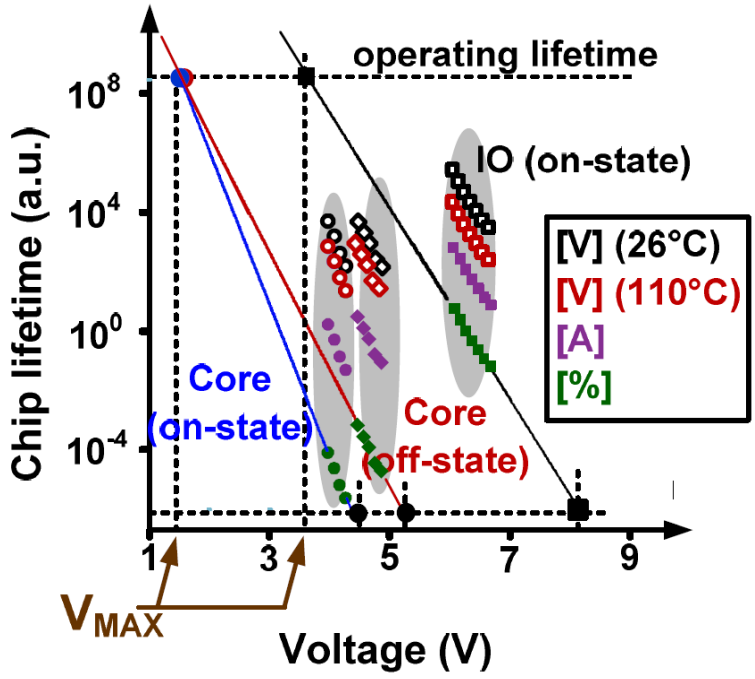
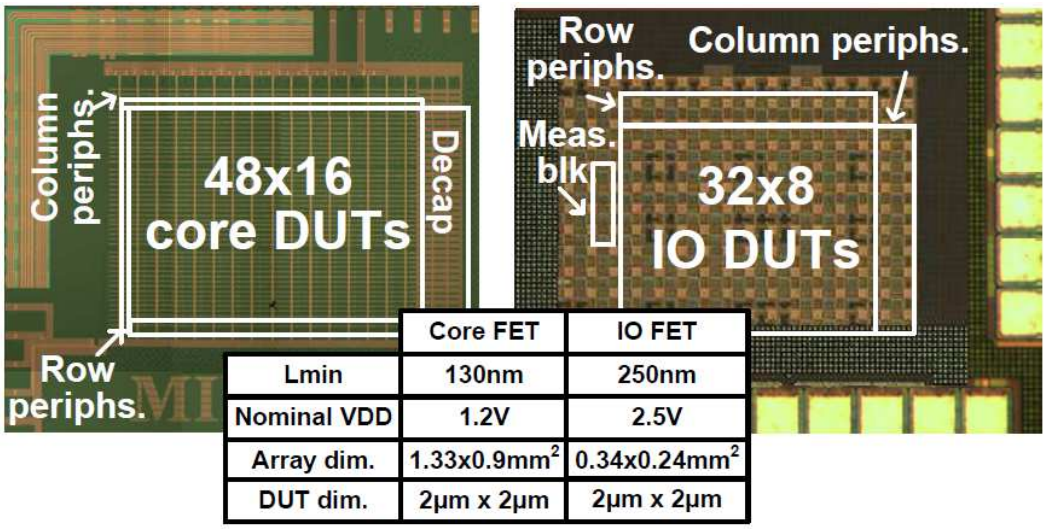


Fig. 3.16 Comparison of projected lifetimes for IO and core devices for ON and OFF (avg. of HD and HDHS) states. Voltage, area, percentile, and temperature extrapolations (solid) are performed from measured statistical data (open).



	Core FET	IO FET
Lmin	130nm	250nm
Nominal VDD	1.2V	2.5V
Array dim.	1.33x0.9mm^2	0.34x0.24mm^2
DUT dim.	2μm x 2μm	2μm x 2μm

Fig. 3.17 Test chip microphotographs of core and IO CLIP macros with chip summary.

3.6. Conclusion

Optimizing the fabrication process and using proper operating conditions based on accurate lifetime predictions are the most practical and effective ways of dealing TDDB. However, the main challenge with this approach is in the collection of massive data from accelerated tests, as TDDB is a statistical phenomenon that can only be accurately characterized from a time-to-breakdown histogram which may require up to thousands of samples for defining a single TTF data point. Moreover, TDDB is a function of a number of variables including voltage, temperature, area, dielectric thickness, and purity, making traditional device probing based methods cumbersome and time consuming. In this work, we propose an array-based gate dielectric breakdown characterization approach called CLIP, to reduce the stress time and silicon area by a factor proportional to the number of DUT cells in the array by stressing all cells in the array in parallel. The essential part is a flexible DUT cell that can be stressed in isolation without thicker t_{ox} FETs to 4 times the VDD, enabling accurate lifetime prediction under different ON and OFF state TDDB modes for both low voltage core and high voltage IO devices.

Chapter 4

Statistical Characterization of RTN in Ring Oscillators

4.1 Introduction

4.1.1 RTN: Extent of Impact

Temporal shift in threshold voltage due to capture and emission of channel carriers (Fig. 4.1) in the gate oxide exhibited as a Random Telegraph Noise (RTN) in the channel current has been known to exist for decades [50-51]. Theoretically, if S is the transistor area, C_{OX} is the gate capacitance and e is the unit charge, the shift in threshold voltage, $\Delta V_T = e/SC_{OX}$. While the issue came up earlier in flash memories [3] and analog circuits [5] due to small S and C_{OX}, it is becoming an increasing concern in ultra-scaled technologies [23].

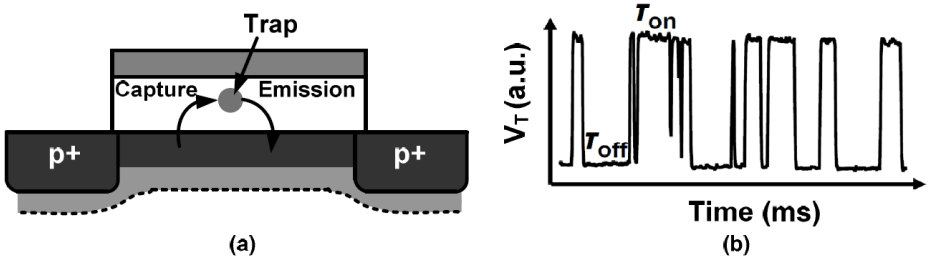


Fig. 4.1 (a) Mechanism of RTN in a transistor. Capture and emission time constants are random. (b) Typical RTN induced V_T fluctuation [23]

There is still a big gap in the understanding of physics explaining the statistics and manifestations of RTN. The observed ΔV_T has been more than what is predicted theoretically [6]. There is conflicting literature on impact of scaling on RTN and new gate dielectric materials[23][25][53]. RTN has been also linked to aging concerns like soft breakdown and bias temperature instability [11]. There has been a host of literature focusing on device level measurements to extracted the statistics and nature of traps [26,54] as well as the influence of bias voltages [55]. However, transistor level analysis is limited in usefulness as it is neither representive nor efficient in collecting large statistics. Plus, the operating conditions vary widely under circuit operation and there is a possibility that RTN behaves very differently subject to the unique circuit.

For the design community, SRAMs, with the smallest device sizes and low noise margins, are the most susceptible to RTN Process solutions have so far helped to keep the V_{MIN} impact below 50mV atleast till the 45nm nodes [56] but is projected to exacerbate at further scaling. Another important concern for RTN has been that with the typical time constants ranging in microseconds to milliseconds [26], it can cause timing hazards in logic circuits. This potential issue has been largely unaddressed in literature. [7] operated a D flip-flop in a metastability region to amplify RTN impact. An assymetric RO was proposed in [59] to isolate RTN in a ring oscillator. However, no work has been reported to directly observe the impact on a traditional ROSC with high resolution. One reason has been difficulty in measuring out RTN impact, due to small shifts in frequency (0.1-1%) that are expected. Even

after alternating stress bias methods the expected shifts are very low and thus a high resolution measurement technique is mandated. Ours is the first work to directly monitor impact of RTN on ROSCs with sampling time less than 1µs for a 0.1% frequency resolution, which is at least a 10X improvement over conventional single ROSC measurement method. A test chip in a 32nm silicon on insulator process features RTN measurements from 20 varieties of ROSCs, with difference in number of stages and device sizes to enable a comprehensive RTN study. Preliminary measurement results of shift in beat frequency has been shown.

4.2 Differential RTN Concept

4.2.2 Statistics of a Beat Signal

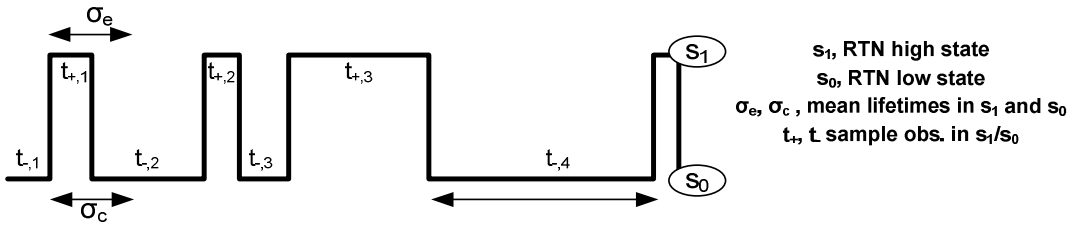


Fig. 4.2 A random telegraph signal with two states, s_1 and s_0. The observations, t_+ and t_- are the times spent in the two states and are exponentially distributed with respective means σ_e and σ_c.

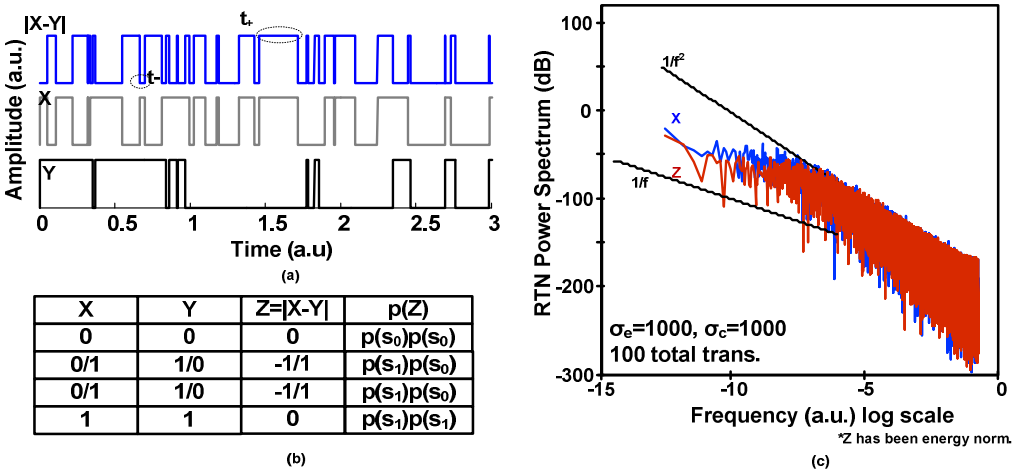
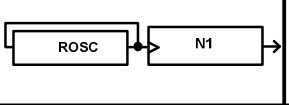
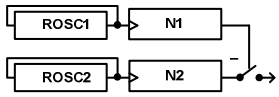
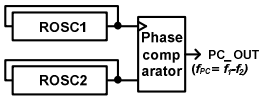


Fig. 3 Difference of two RTN signals with estimated probabilities denoted on the side. Clearly, the statistics of the resultant signal Z, differs from original signals X and Y and it is hard to get back the time distribution of X from Z. However, it can be proved using the methodology in [JAP'54], that the power spectrum density of Z is same as X and Y, for an ideal RTN stationary RTN signal. The power spectrum plots in (c) verify this behavior.

Fig. 4.2 shows a typical RTN signal with two states, s_1 and s_0. The observations, t_+ and t_- are the times spent in the two states and are exponentially distributed with respective means σ_e and σ_c. Fig. 4.3(a) shows simulated RTN signals, X and Y. The beat signal, Z= |X-Y| is also seen with estimated probability denoted in Fig. 4.3(b). It can be seen and also verified using a statistical tool that there is some loss of information in this resultant signal and it is not possible to obtain the σ_e and σ_c of X or Y from Z. However, the amplitudes are preserved and also interestingly the power spectrum of Z matches with that of X and Y as shown in Fig. 4.3(c). The methodology to prove this mathematically is given in [50].

4.2.2 Previous ROSC Based Designs

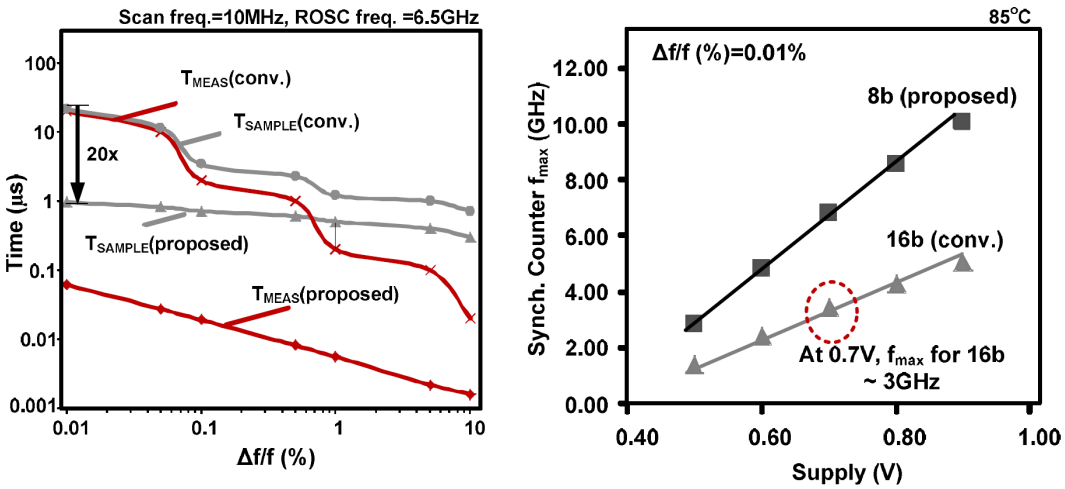
System	Single ROSC	2 ROSC, simple	2 ROSC, beat freq.
Block Diagram			
Function	Count ROSC periods during externally controlled meas. time	Count ROSC1 periods during N1 periods of ROSC2	Count ROSC1 periods during one period of PC_OUT
Features	Simple; compact	Simple; immune to common mode variations	High resolution w/ short meas. time; immune to common mode variations
Issues	Voltage and temp. varations; Sampling time vs. resolution tradeoff; requires absolute timing reference (e.g. oscilloscope)	Sampling time vs. resolution tradeoff and ease of implementation	Requires extra circuits (e.g., Phase Comp., edge detector, etc...)
T_{SAMPLE} for 0.01% max resolution *	~10ms	20 µs	0.9 µs
Counter size for 0.01% max resolution*	16b	16b	8b

*5 state ROSC freq. = 6.5GHz

Fig. 4.4 Different ROSC based designs. Column 1 is the single ROSC with a divider. The frequency is read off chip and that gives a low sampling time. Column 2 is a two ROSC based setup where the MSB of ROSC1 samples the count of ROSC2. The technique is immune to common mode variations and gives a reasonable sampling time of 20µs. However, it is associated with implementation challenges like a high speed 16b counter to accommodate the high frequency ROSCs needed. Column 3 is the proposed beat frequency based two ROSC approach which can offer a sampling time of 1µs while providing a frequency resolution of 0.01% without using a high speed counter.

Fig. 4.4 shows the different ROSC based designs possible. Column 1 is the single ROSC with a divider. The frequency is read off chip and that gives a low sampling time. Column 2 is a two ROSC based setup where the MSB of ROSC1 samples the count of ROSC2. The technique is immune to common mode variations and gives a reasonable sampling time of 20µs. However, it is associated with implementation challenges like a high speed 16b counter to accommodate the high frequency ROSCs needed. Column 3 is the proposed beat frequency based two ROSC approach which can offer a sampling time of 1µs while providing a frequency

resolution of 0.01% without using a high speed counter. Fig. 4.5(a) shows the comparison of T_{SAMPLE} improvement offered by the proposed Beat Frequency Detection (BFD) based two ROSC design compared to the conventional two ROSC design. An improvement of 20x seen for a maximum frequency resolution of 0.01% for a 5 stage ROSC running at 6.5GHz and data acquisition frequency of 10MHz. Note that a faster off-chip frequency can further improve the sampling time of the proposed setup by an order. (b) In order to get the same frequency resolution, the conventional design needs a 16b counter compared to a 8b counter needed for the BFD based design. A 16b counter has a much smaller f_{max} which is direct tradeoff for number of stages in the ROSC used impacting accurate RTN evaluation. Next we look at the working of the BFD circuit.



(a) (b)

Fig. 4.5 (a) Comparison of T_{SAMPLE} improvement offered by the proposed BFD based two ROSC design compared to the conventional two ROSC design. An improvement of 20x seen for a maximum frequency resolution of 0.01% for a 5 stage ROSC running at 6.5GHz and data acquisition frequency of 10MHz. Note that a faster off-chip frequency can further improve the sampling time of the proposed setup by an order. (b) In order to get the same frequency resolution,

the conventional design needs a 16b counter compared to a 8b counter needed for the BFD based design. An 16b counter has a much smaller f_{max} which is direct tradeoff for number of stages in the ROSC used impacting accurate RTN evaluation.

4.2.3 Beat frequency detection (BFD) [58]:

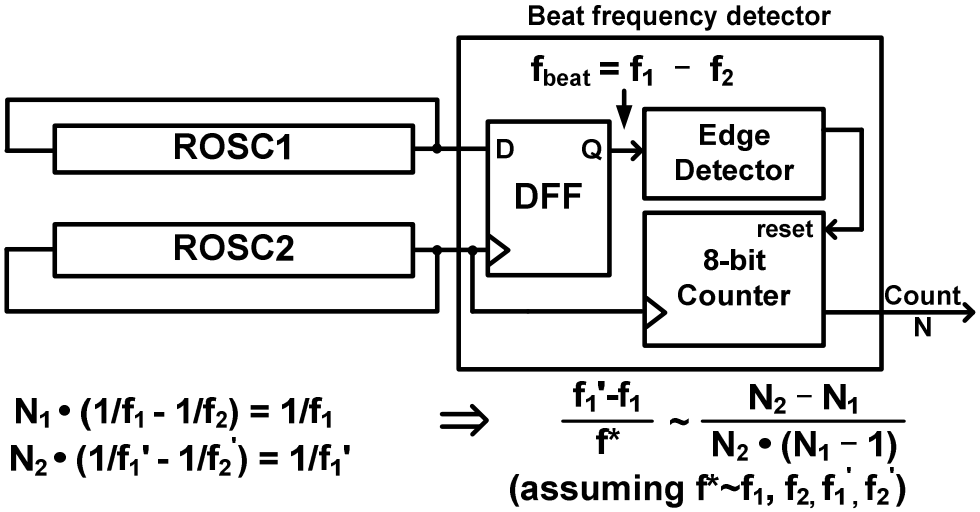


Fig. 4.6 Beat frequency odometer system used in this work. N_1 and N_2 are the counts from the counter output, recorded before and after the sampling period. Using the equations listed above, we can conveniently calculate the percentage frequency change with pico second level precision.

Fig. 4.6 gives a brief background of the BFD circuit. N_1 and N_2 are the counts from the counter output, recorded before and after the sampling period. Using the equations listed alongside the figure, we can conveniently calculate the percentage frequency change with pico second level precision.

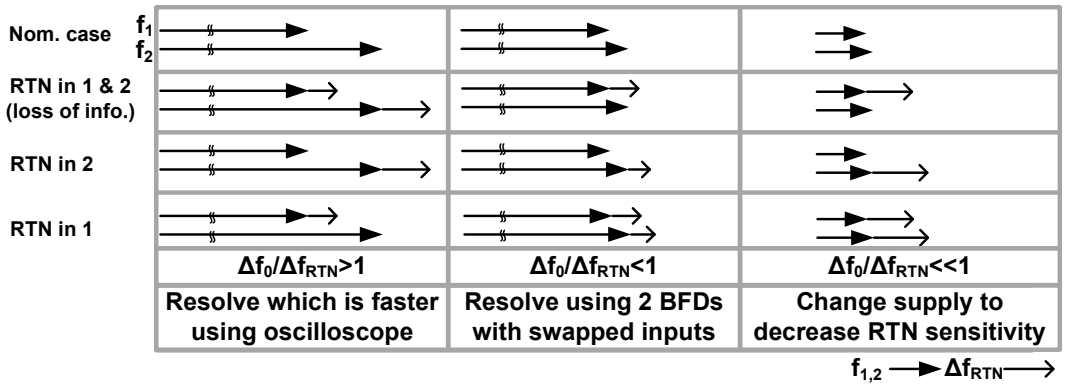


Fig. 4.7 Symbolic representation of different scenarios of RTN impacting the pair of ROSCs. Col. 1 is the easiest case to resolve, as always ROSC1 is slower, and a readout from oscilloscope easily verifies this. Col. 2 has the two ROSCs with relatively large RTN that either can be faster at any moment based on RTN in them. This is resolved using two BFDs. In BFD1, ROSC1 acts as the clock, while in BFD2, ROSC2 acts as the clock. Col 3 is the unlikely case when RTN is very large to bring the ROSCs out of trimming range. Supply control is needed to decrease the RTN sensitvity. Note that the case when we have RTN in 1 and 2 both is hard to resolve due to the differential nature of the setup.

Fig. 4.7 shows the symbolic representation of different scenarios of RTN impacting the pair of ROSCs. Col. 1 is the easiest case to resolve, as always ROSC1 is slower, and a readout from oscilloscope easily verifies this. Col. 2 has the two ROSCs with relatively large RTN that either can be faster at any moment based on RTN in them. This is resolved using two BFDs. In BFD1, ROSC1 acts as the clock, while in BFD2, ROSC2 acts as the clock. Col 3 is the unlikely case when RTN is very large to bring the ROSCs out of trimming range. Supply control is needed to decrease the RTN sensitvity. Note that the case when we have RTN in 1 and 2 both is hard to resolve due to the differential nature of the setup.

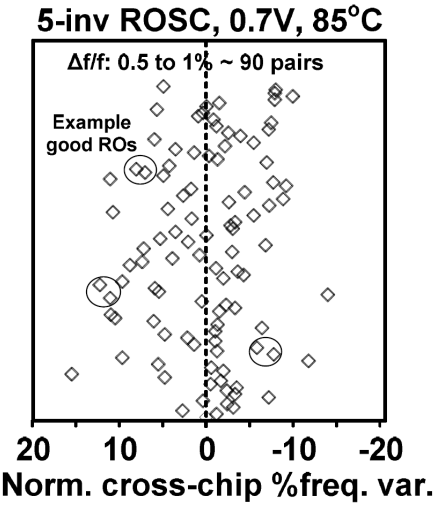


Fig. 4.8 Cross chip variation monte carlo simulations done to estimate the number of ROSC pairs needed to get the ROSCs within 0.5-1% trimming range. An optimal value of 32 ROSCs was chosen to get good statistics of about 100 good pairs.

4.3 Test Chip Description

Fig. 4.8 shows the cross chip variation using monte carlo simulations done to estimate the number of ROSC pairs needed to get the ROSCs within 0.5-1% trimming range. An optimal value of 32 ROSCs was chosen to get good statistics of about 100 good pairs. Fig. 4.9 shows the ROSC topology featuring header switches for optional stress functionality to trigger soft breakdown. The inverter stages are loaded with a fan-out 4 load. Separate array and ROSC supply and local decap at the virtual VDD employed to minimize supply noise from corrupting RTN results. MEAS goes high, when ROW=COL=1 enabling the feedback loop and selecting the ROSC for measurement. Fig. 4.10 shows the statistical framework for evaluating RTN in ROSCs. It consists of two arrays of ROSCs, RARR1 and RARR2 with 32 rows and 20 columns of ROSCs. The ROSCs are identical across rows, while different flavors

of ROSCs are provided across the columns to provided flexibility to control the number of stages and device sizes in the ROSCs. Row and column decoders are used to assert the internal MEAS signal in one of the ROSC in both the arrays. For efficient use of area, the ROSCs share the same bus, with a keeper to prevent any floating buses.

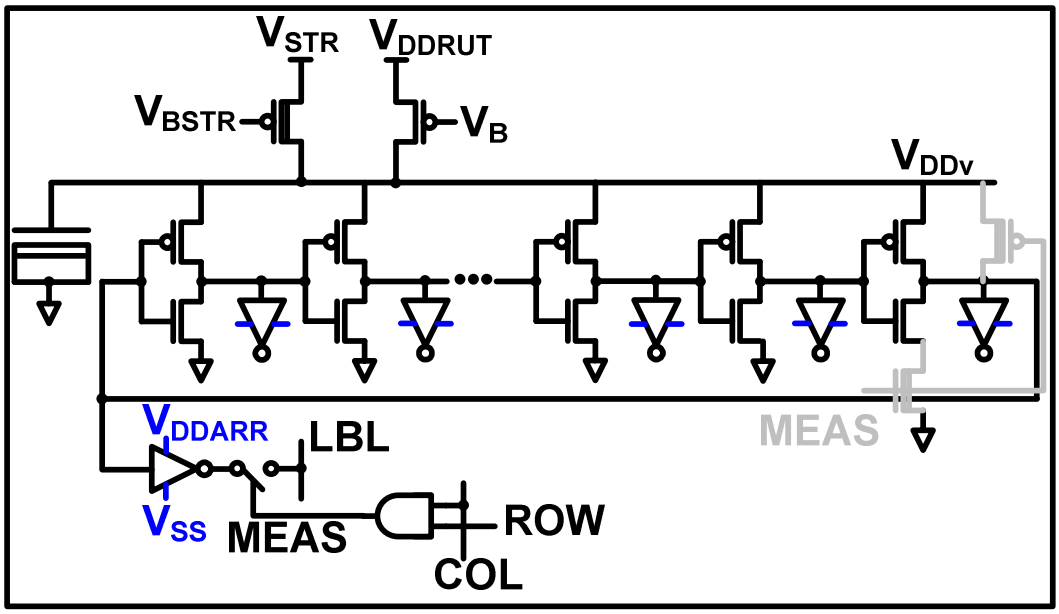


Fig. 4.9 ROSC topology featuring header switches for optional stress functionality to trigger soft breakdown. The inverter stages are loaded with a fan-out 4 load. Separate array and ROSC supply and local decap at the virtual VDD employed to minimize supply noise from corrupting RTN results. MEAS goes high, when ROW=COL=1 enabling the feedback loop and selecting the ROSC for measurement

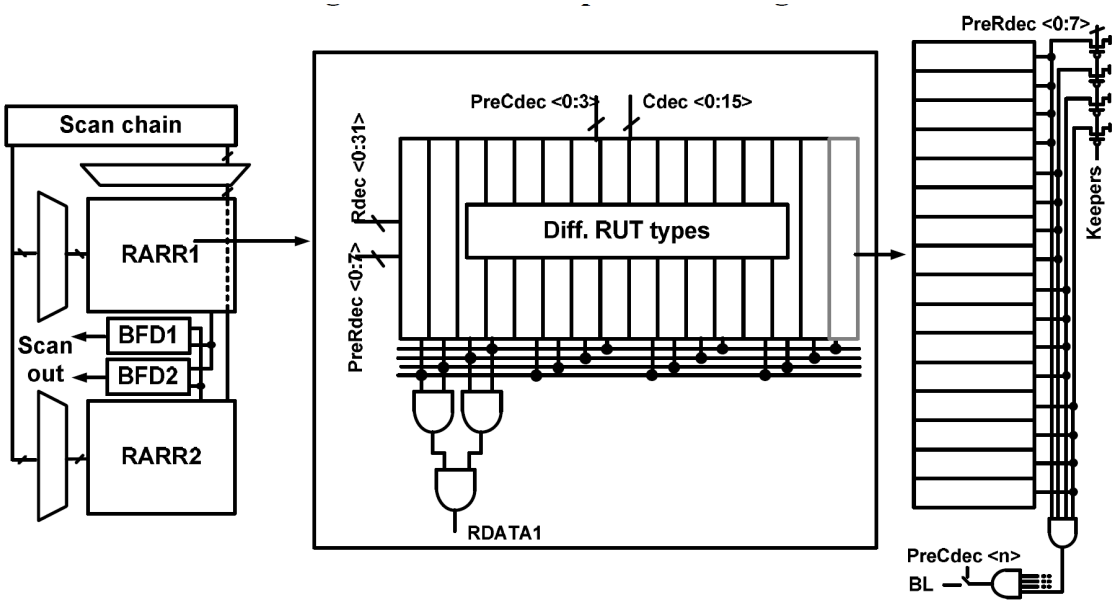


Fig. 4.10 Statistical framework for evaluating RTN in ROSCs. It consists of two arrays of ROSCs, RARR1 and RARR2 with 32 rows and 20 columns of ROSCs. The ROSCs are identical across rows, while different flavors of ROSCs are provided across the columns to provided flexibility to control the number of stages and device sizes in the ROSCs. Row and column decoders are used to assert the internal MEAS signal in one of the ROSC in both the arrays. For efficient use of area, the ROSCs share the same bus, with a keeper to prevent any floating buses.

4.3 Measurement Results

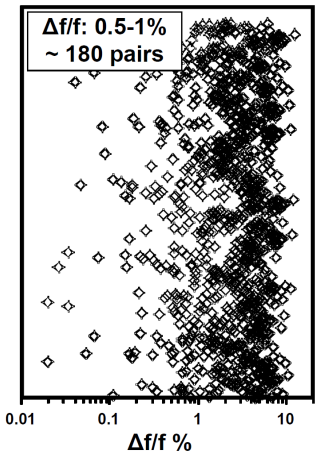
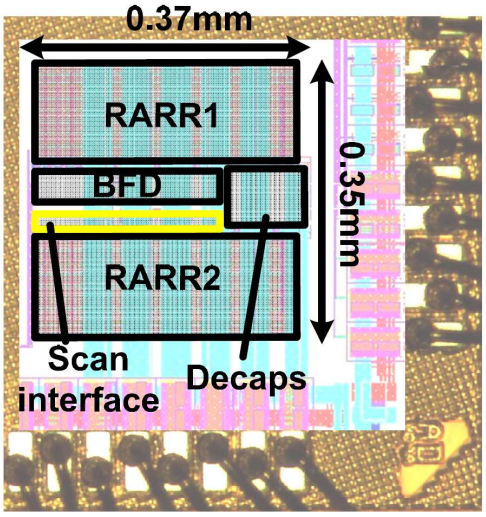


Fig. 4.11 Measured frequency variation for 32x32 combination of ROSCs. In order to trim the ROSCs, 1 of the 180 good pairs have to be chosen.

Fig. 4.11 shows the measured frequency variation for 32x32 combination of ROSCs. In order to trim the ROSCs, 1 of the 180 good pairs have to be chosen. Fig. 4.14 is the test chip microphotograph with the chip summary.



Process	0.9V 32nm HKMG SOI
Ckt dim.	0.35 x 0.37mm^2
T_{SAMPLE}	>0.9µs @ 0.9V, 10MHz
V_{DDRUT}	0.6-1.2V
V_{STRESS}	0.9-3V
ROSCs	11, 9, 7, 5, 3 stages with widths 1:1, 2:2, 1:2, 2:1

Fig. 4.14 Test chip microphotograph implemented in a 32nm HKMG SOI process. The chip feature summary is shown alongside.

4.4 Conclusions

This work proposes a 32nm test macro to directly monitor impact of RTN on ROSCs with sampling time less than 1µs for a 0.1% frequency resolution, which is at least a 10X improvement over conventional single ROSC measurement method. A

test chip features RTN measurements from 20 varieties of ROSCs, with difference in number of stages and device sizes to enable a comprehensive RTN study.

Chapter 5

Measurement, Analysis and Improvement of Supply Noise in 3D ICs

5.1. Introduction to 3D ICs

3D integration is recognized as a breakthrough technology for improving interconnects performance and reducing chip form factors [29,30]. Memory bandwidth, which has become a critical performance limiter in modern processors, can be significantly increased by vertically stacking caches on top of processing cores. Extremely high memory densities have been demonstrated for stand-alone applications where multiple 2D memory chips are stacked in a single package. 3D integration technology also makes it possible to vertically integrate

chips built in heterogeneous processes (e.g. logic, DRAM, flash, SiGe, InP) with slight additional cost compared to integrating monolithic chips.

The premise of 3D integrated circuits has spurred research activity at virtually all levels of the 3D design hierarchy. The material and process community has recently made great strides in developing high yield and low cost Through Silicon Vias (TSV) with dimensions comparable to small logic gates [30,59,60], transforming 3D integration from a laboratory exercise to a practical technology. The capability to improve TSV characteristics as traditional scaling continues to make 3D chips even more viable in future process generations. A host of techniques to deal with 3D chip design issues have been introduced by the circuit design and automation community. Thermal management is one of the most important design issues in 3D chips, as they have higher power dissipation per area and increased thermal resistance between the tiers due to the isolation layer. Various 3D architectures and interconnect models have been proposed to estimate the performance benefits, power reduction and die temperature [29, 61]. Thermal aware placement and routing algorithms for 3D ICs have been presented in a number of prior publications [62-65]. Contactless signaling between the stacked tiers using the capacitive or inductive coupling principle has been gaining traction in the circuit design community [66-68]. That work is based on the premise that by utilizing the close proximity of the circuits, TSVs between the tiers for data signals can be eliminated, which may resolve wafer alignment issues and lead to lower process complexities. At the architecture and system level, benchmark programs were used to predict the memory bandwidth improvement in various 3D architectures [69].

Despite the recent surge in 3D IC research, there has been virtually no work from the circuit design and automation community on power delivery issues for 3D ICs. On-chip power supply noise has worsened in modern systems because scaling of the Power Supply Network (PSN) impedance has not kept up with the increase in device density and operating

current due to the limited wire resources and constant RC per wire length [31]. This situation is worsened in 3D ICs as TSVs contribute additional resistance to the supply network and the number of pins for power delivery is fundamentally limited by the footprint of the 3D chip. For example, a 3D chip with *n* tiers can only have 1/*n* the number of power supply pins compared to a single 2D chip of *k*-time footprint, which results in an *n* fold increase in the resistive and inductive parasitics. The increased IR and Ldi/dt supply noise in 3D chips may cause a larger variation in operating speed leading to more timing violations. The supply noise overshoot due to inductive parasitics may aggravate reliability issues such as oxide breakdown, negative bias temperature instability and hot carrier injection. Consequently, on-chip power delivery will be a critical challenge for 3D ICs. This is contrary to the common perception where power delivery in 3D chips was considered no different than that in conventional 2D chips.

In this work, we specifically address the power delivery issues in high performance 3D ICs, that can monolithically integrate logic and memory. The only related work [70] discussed the simultaneous switching noise issues in 3D ICs, based on some compact physical models. The highlights of our work are as follows:

- A 3D test chip in a MIT Lincoln Lab's 0.15µm process has been fabricated with the goal to evaluate TSV impact on supply noise
- Compared to their 2D counterparts, we find that 3D designs have a much larger DC noise due to the added TSV resistance and reduced supply pads. The peak impedance at the resonant frequency is similar to 2D as the increase in inductive impedance is partially compensated by the increased damping from the TSV resistances.
- Low frequency supply noise is worst in the tier farthest to the supply pins (i.e. the bottom tier) while the high frequency noise is worst for the tier closest to the supply pins (i.e. the top tier).

- A multi-story power delivery (MSPD) technique is proposed for 3D chips. In this scheme, an external voltage source of kV_{DD} is applied, and power is distributed differentially between a (kV_{DD}) rail and a ((k-1)V_{DD}) rail using level conversions as required [71][72]. By recycling current between different power supply domains, the IR noise can be reduced by up to 45%, while AC noise is marginally affected.
- Design trade-offs between the number of stacked supplies, leakage power and via allocation has been analyzed in detail for the proposed multi-story power delivery scheme
- A 3D SRAM macro showcases the feasibility of the proposed scheme. The PSNs in each tier are readily separated requiring only slight modification, which makes the scheme particularly attractive for 3D chips.

The organization of this chapter is following. We first do a comprehensive TSV characterization using 3D ROSC based statistical macro in Section I. In Section II we review the power delivery idea in conventional high performance ICs. In Section III, we examine power supply integrity vis-à-vis 3D ICs based on measured TSV parameters. In section IV, we discuss the multi-story power delivery technique to address the power delivery issues in 3D chips and present measurement results for DC and AC noise. In Section V, we provide the results from a 3D SRAM, implementing the proposed scheme. Finally, section VI draws a conclusion. This work uses MIT Lincoln Lab's 1.5V, 0.15µm 3D Fully-Depleted Silicon-On-Insulator (FD-SOI) process which has 3 tiers [60].

5.2 TSV Characterization

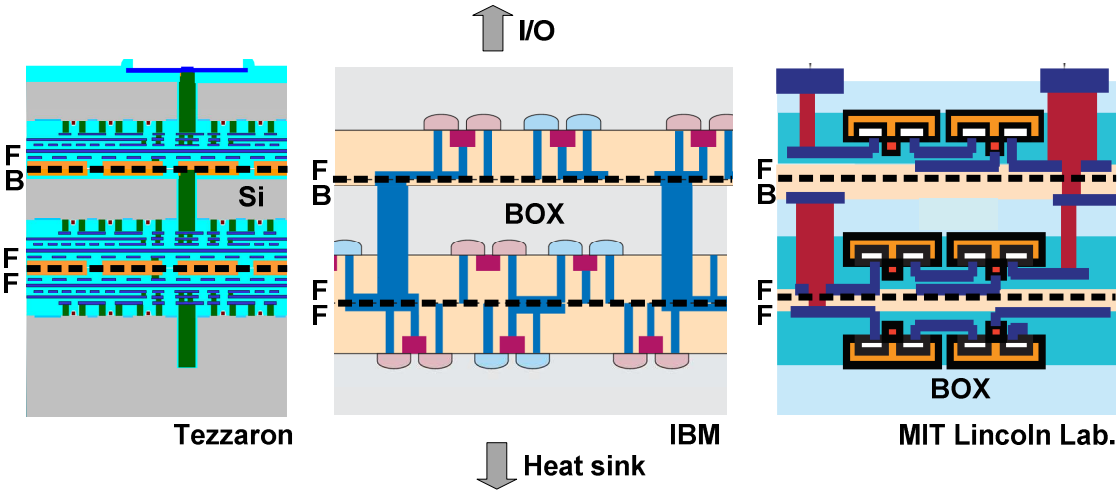


Fig. 5.1. An example of three stacked process from Tezzaron, IBM and MIT lincoln labs. Typically, if the three stacks are comprised of 2 layers of cache on processor, the latter would be placed in vicinity of the heat sink at the bottom while the I/O go through the top. The topology can be face to face or face to back in this case. For example in Tezzaron process, we take two wafers and invert one on top of the other to give a F-F arrangement. After this wafer is thinned, and the next tier is aligned F-B on the second tier. The second one is an IBM process with a via-last approach in SOI. The third figure is a MIT lincoln lab process which we used for this work. Similar to IBM, this is a fully depleted SOI process.

Fig. 5.1(a) depicts the MIT Lincoln Lab's 3D FD-SOI process vis-à-vis other state of the art processes. This process has three tiers. The bonding pads are on the top tier, while the heat sink is typically below the bottom tier. Processors or other power intensive circuits would ideally be placed on the bottom tier in close proximity with the heat sink.

The tiers are interconnected through TSVs for electrical and thermal conduction.

Conventional Techniques		Proposed
S-param extraction	Charge based	Ring osc. (ROSC) based
Low resolution.	nA external current sensitivity required. Accurate and monolithic.	High sensitivity. Monolithic and obtain loading effect on logic circuits directly
Not scalable	Scalable	Scalable

Fig. 5.2. The goal is to statistically capture the electrical behavior of TSVs. The conventional s-param based extraction approach provides direct cap

information, but has low resolution, is invasive and not scalable. The charge based method used to characterize normal vias doesn't provide impact on logic circuits in terms of frequency degradation. This work uses a ROSC based approach and identifies the impact of TSVs on logic circuit loading. It is scalable, high resolution and monolithic.

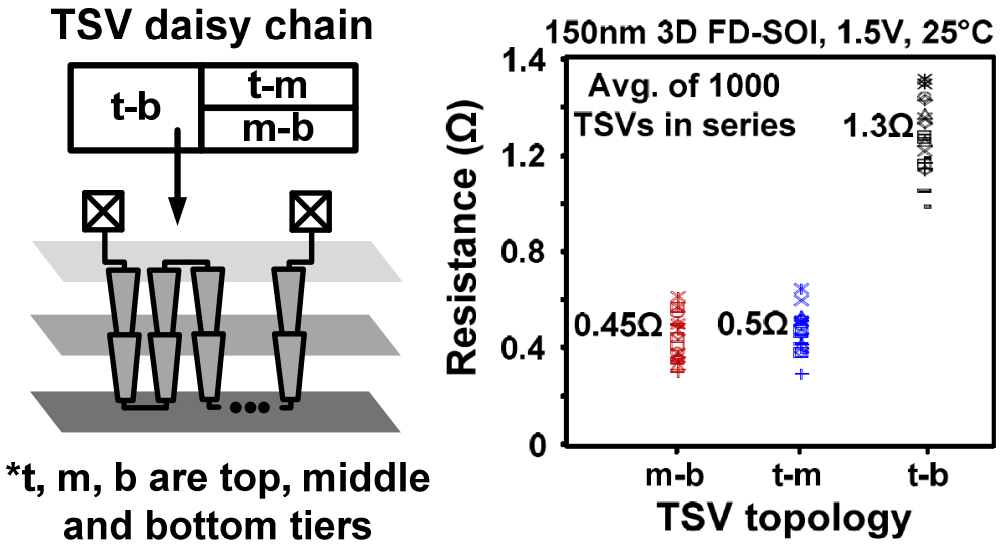


Fig. 5.3. Test structure for measuring TSV resistance (left). Measured TSV resistance distribution (right). We had 1000 chain TSVs in daisy chains with different TSV geography. Inter-TSV resistance and bonding wires calibrated out. The two tier connection is almost same resistance, while a stacked TSV resistance is more than sum of indiv.

In order to fully understand and model the impact of TSVs on the supply noise of 3D ICs, accurate characterization of the TSV parasitics must precede. Fig. 5.2 shows the advantages of the proposed 3D ROSC based statistical characterization method. Fig. 5.3 shows a simple daisy chain structure used in the test chip to obtain TSV resistance for different pair of tiers. The resistance of the stacked *t-b* TSV is slightly more than the sum of the resistances for *t-m* and *m-b* cases, which follows from the geometry of the TSV and different layers. There was no noticeable difference in the TSV resistance between the *t-m* and *m-b* layers. The performance impact of TSV capacitance on digital circuits is evaluated by a 3D ROSC array test setup in Fig. 5.4.

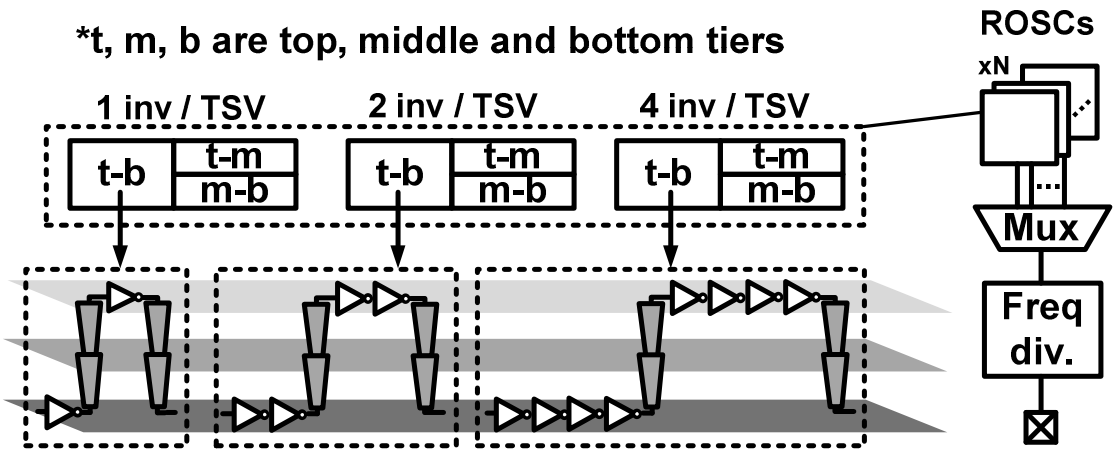


Fig. 5.4. ROSC based statistical TSV characterization block. The basic structure is a TSV loaded ROSC and we provide flexibility in terms of number of TSVs for example TSV every 1 inv, every 2 inv, every 4 inverter and none. The different topology type of TSV would be a ROSC between t-m and m-b and a stacked t-b. We have numerous such blocks. A mux selects the required ROSC, whose frequency is divided and read off externally.

It consists of sets of 9 inter-tier communicating ROSCs. They connect between *t-m, m-b* and *t-b* tiers each with intermittent TSV connection every 1, 2 or 4 inverters, with 5, 10 or 20µm TSV pitches, respectively. The divided frequency output of the selected ROSC module is read off chip.

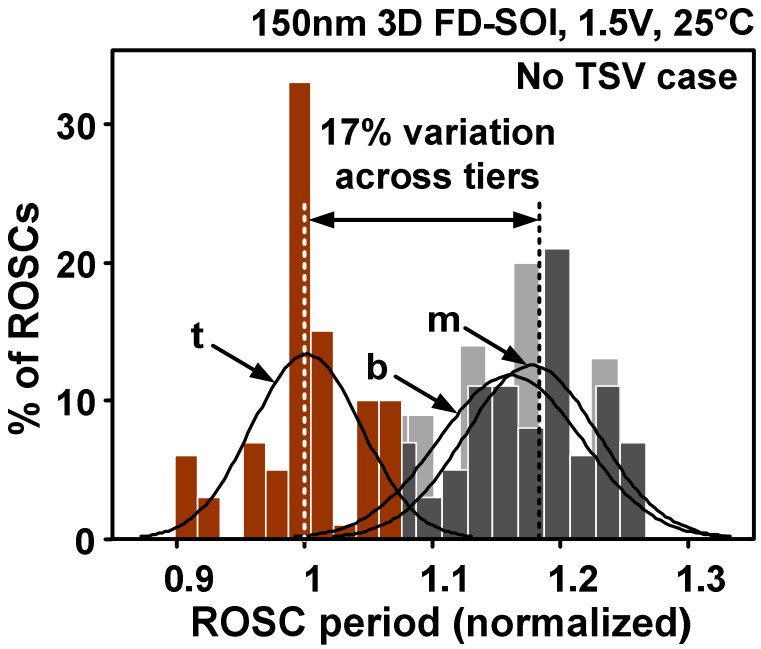


Fig. 5.5 Inter-tier systemic variation. The bottom and middle tier were found to be on an average 17% slower than the top tier. We have to calibrate out this variation before separating TSV impact.

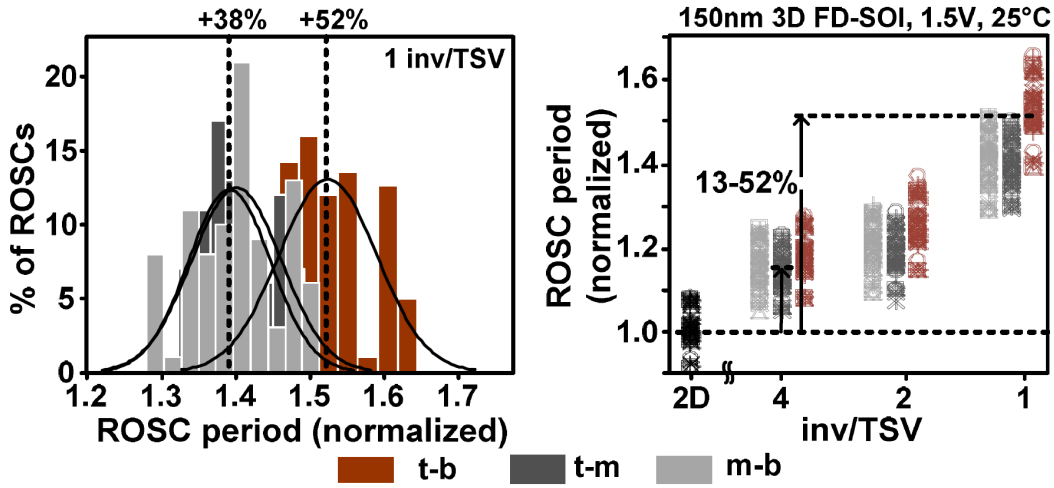


Fig. 5.6 Measured 3D-ROSC characteristics. Here we show results for TSV loaded 3D ROSCs, for the various possible TSV topology. Inter-tier variation calibrated out as the noTSV and with TSV cases were locally well correlated. The first figure is for a TSV every 1 inv case and we observe a 38-52% shift in ROSC period from a 2D case. In general, it changes with number of inv/TSVs. For example a 13-52% change is obtained for the stacked TSV case from a 2D case.

Fig. 5.5 makes a comparison of 2D ROSCs in *t*, *m* and *b* tiers. The bottom and middle tier were found to be on an average 17% slower than the top tier. We have to calibrate out this variation before separating TSV impact. Fig. 5.6 shows results for TSV loaded 3D ROSCs, for the various possible TSV topology. Inter-tier variation calibrated out as the noTSV and with TSV cases were locally well correlated. The first figure is for a TSV every 1 inv case and we observe a 38-52% shift in ROSC period from a 2D case. In general, it changes with number of inv/TSVs. For example a 13-52% change is obtained for the stacked TSV case from a 2D case.

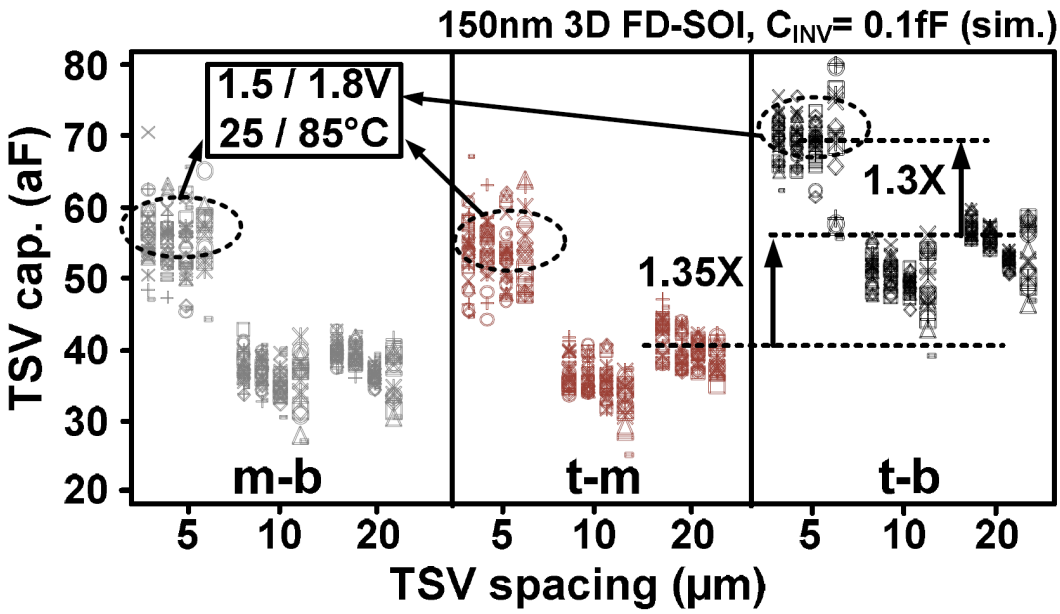
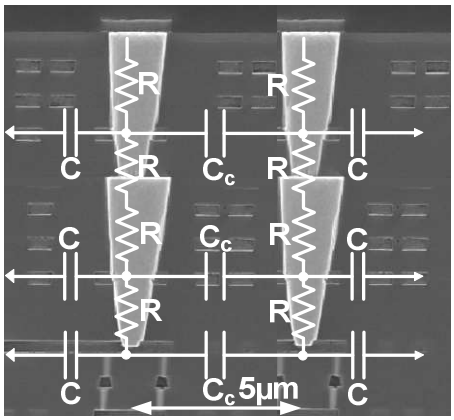


Fig. 5.7 Extraction of TSV capacitance using $C_{TSV}=N_{INV/TSV}C_{INV}\Delta T_{3D}/T_{2D}$. The 3D ROSC period is related to TSV cap., by this simple equation where TSV cap is expressed in a linear relationship with $N_{INV/TSV}$, inv load, a shift in 3D ROSC freq and 2D ROSC freq. Here I have plotted TSV cap with three variables-TSV spacing, topology of TSV and the same across differnet voltages and temperature. For small TSV spacing, there is a strong coupling or mech. stress effect observed and we quantify it to be about 30%. The stacked TSV cap is about 35% more than the individual caps. Also, the TSV cap is not affected by voltage and temperature, which points out that the metric is justified.

Fig. 5.7 plots TSV capacitance using $C_{TSV}=N_{INV/TSV}C_{INV}\Delta T_{3D}/T_{2D}$. The 3D ROSC period is related to TSV cap., by this simple equation where TSV cap is expressed in a linear relationship with NINV/TSV, inv load, a shift in 3D ROSC freq and 2D ROSC freq. Here I have plotted TSV cap with three variables-TSV spacing, topology of TSV and the same across different voltages and temperature. For small TSV spacing, there is a strong coupling or mech. stress effect observed and we quantify it to be about 30%. The stacked TSV cap is about 35% more than the individual caps. Also, the TSV cap is not affected by voltage and temperature, which points out that the metric is justified.



150nm 3D
FD-SOI,
1.5V, 25°C

R	0.33 Ω
C	40aF
C_c	6aF*

*For 5µm pitch

Fig. 5.8 Proposed TSV model. The model takes care of TSV resistance, TSV sidewall capacitances and coupling capacitance between TSVs. The values are the average R and C estimated from the measured statistical data. Rigorous quantification of the electrical behavior of a TSV is significant for understanding the supply noise dynamics in a 3D chip.

Fig. 5.8 shows the cross-sectional scanning electron microscope photograph [60] of a stacked TSV connecting the back metal of the top tier with the top level metal of the bottom tier. An analytical TSV model is superimposed on this. The model takes care of TSV resistance, TSV sidewall capacitances and coupling capacitance between TSVs. The values are the average R and C estimated from the measured statistical

data. Rigorous quantification of the electrical behavior of a TSV is significant for understanding the supply noise dynamics in a 3D chip, which is taken up in the next section.

5.3 Power Delivery: General Idea

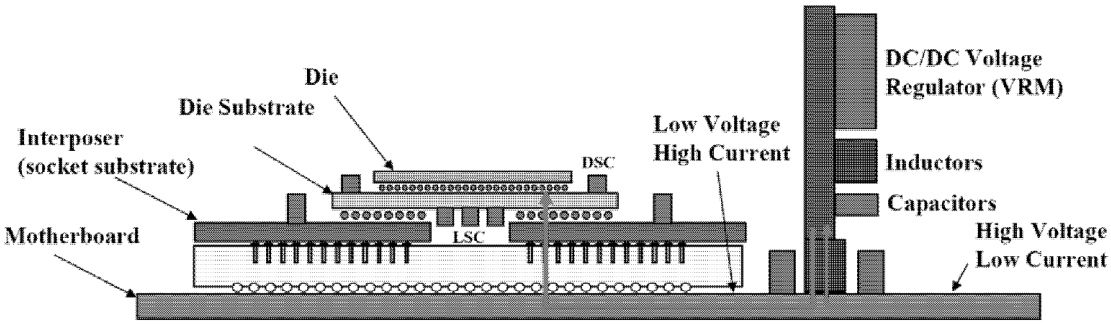


Fig. 5.9. Conventional power delivery architecture using a voltage regulator mounted on the motherboard [18].

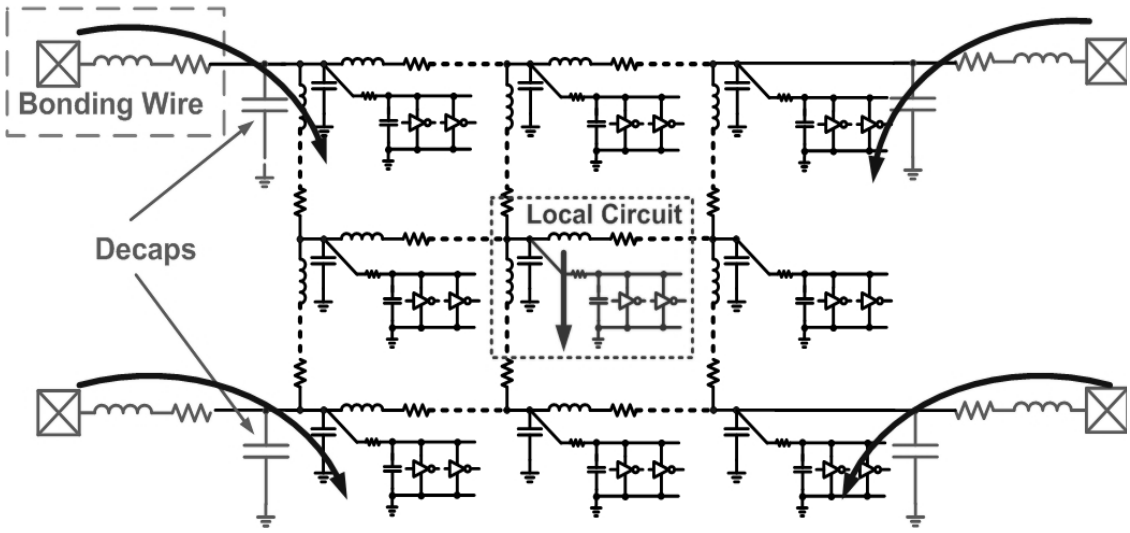


Fig. 5.10. Section of a power supply grid model used to simulate AC noise [19]. A small signal noise source at the local circuit is used to perturb the entire supply grid.

Conventional power delivery methods to high performance ICs employ a DC-DC converter (known as a voltage regulator module) mounted on the motherboard, with external interconnects providing the power to the processor chip as depicted in Fig. 5.9 [73]. The supply that reaches the processor has IR and Ldi/dt drop across the package constituting the supply noise. With scaling, while the larger currents are aggravating IR drop, the faster transients, due to faster clock rate are worsening the Ldi/dt drop. Worse, if these fast transients happen at the circuit's natural resonant frequency of excitation, large droops on supply are triggered. With these increased levels of noise and reduced noise margins, as V_{DD} levels scale down, reliable power delivery to power-hungry processor chips has become a major challenge.

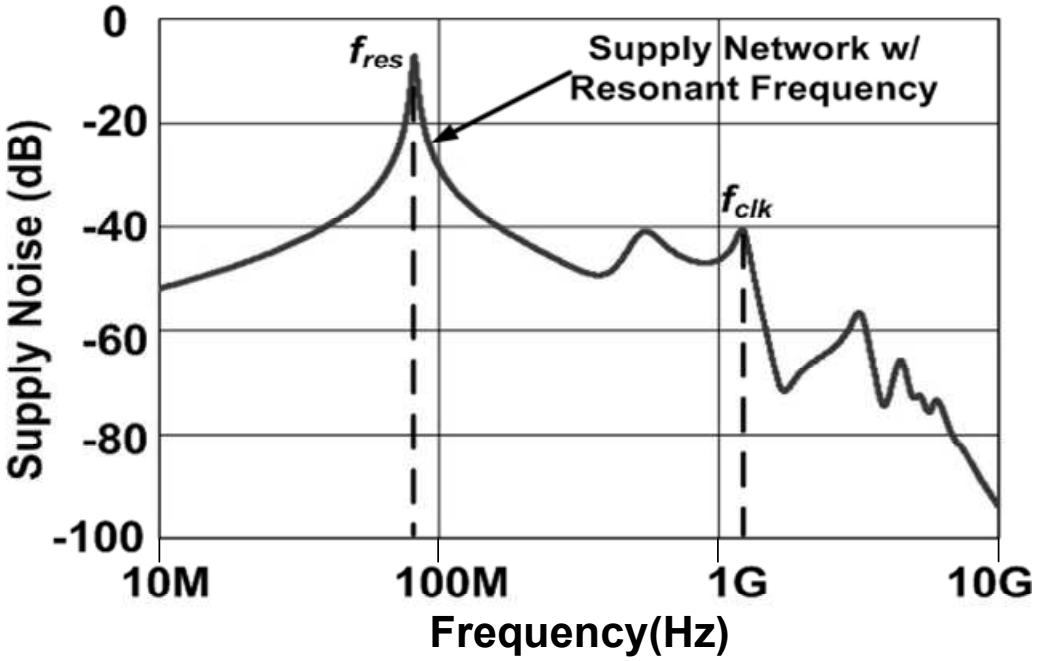


Fig. 5.11. Simulated supply noise spectrum from power grid model in Fig. 5.10

Fig. 5.11 shows the supply noise spectrum obtained from a typical power delivery grid [19] shown in Fig. 5.10. The DC component of the noise is given by IR drop

across the package and power grid. The first peak in the figure corresponds to the resonant frequency, given by $f_{res} = 1/2\pi\sqrt{LC}$ which typically appears in the range of 50-300MHz. An excitation at this frequency can be triggered during microprocessor loop operations or wakeup. Several other peaks are seen in the figure, due to switching at clock frequency and its higher harmonics, or due to local resonance: the corresponding noise is typically an order less in magnitude than the resonant peak. The noise at a particular frequency is estimated by multiplying the impedance with the current component at that frequency [75-76].

Next, we focus on power delivery, specifically to 3D ICs and analyze the PSN noise problem in this regime.

5.4 Power Delivery in 3D ICs

The TSV resistance encountered in the supply path imposes new challenges in 3D power delivery vis-à-vis the conventional 2D case. First, the lower tiers experience worsened power supply noise due to the increased resistance in the power network. Furthermore, power intensive circuits have to be placed at bottom tier, which makes reliable power delivery even more difficult.

5.4.1 Frequency Response of PSN: 2D vs 3D

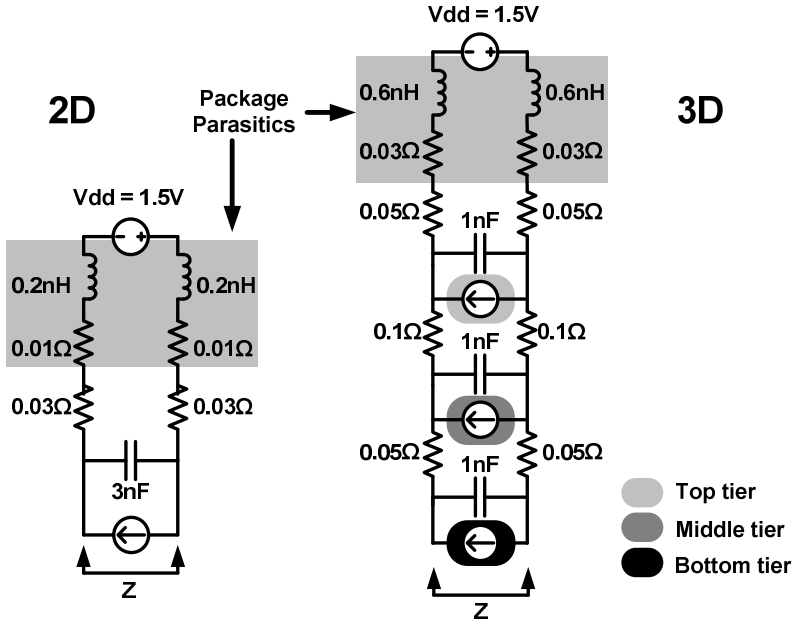


Fig. 5.12. Simplified PSN models for comparing impedance response in 2D and 3D.

We discussed the supply noise behavior in 2D chips in previous section. In 3D chips, it needs a reassessment in light of reduced pin count and extra TSV resistances. A methodology to obtain a power grid model for a 3D IC is developed in [70]. By tying this model with the parameters of the specific 3D process used in this work, along with the quantitative understanding of the 2D power supply noise, we can to get some useful heuristic circuit models for comparing the 2D and 3D power delivery scenarios. These are shown in Fig. 5.12. We analyze the case when the resistance in 3D supply path is dominated by the TSVs and model ten of them here. There are a few assumptions made. First, the overall chip capacitance (3nF in typical 2D case) is split equally between the three 3D tiers. This assumes equal footprint across the tiers of a 3D die. Second, due to the reduced footprint of the 3D die, the number of power

pins would be third of the 2D case, leading to 3X increase in package parasitic inductance and resistance.

Since, noise at the bottom tier is predictably worst, we compare its impedance response to the conventional case. The normalized impedance comparison is shown in Fig. 5.13, which illustrates the following:

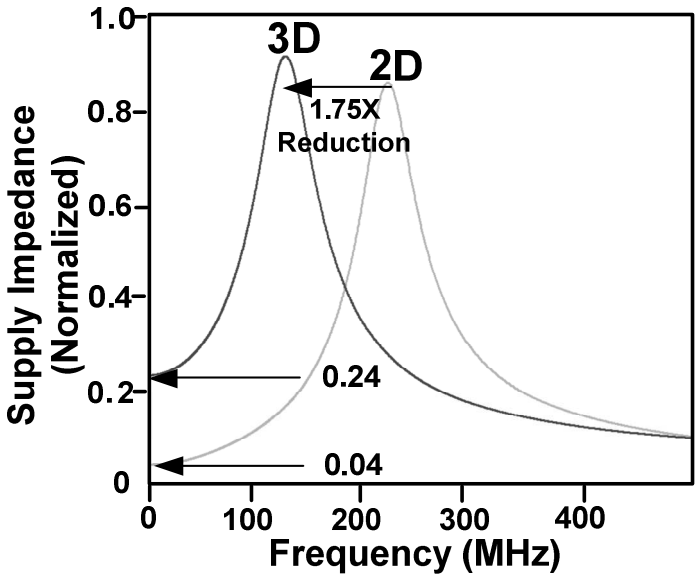


Fig. 5.13. Supply impedance response comparison between 2D and 3D. 6X more DC noise is observed in 3D IC compared to 2D IC, as well as a 1.75X reduced f_{res}

- Low frequency behavior: The capacitors and inductors are open and short circuited, respectively. Therefore, the 2D model has an impedance of 2(0.01+0.03) = 0.08Ω, while the 3D model has an impedance of 2(0.03+0.05+0.1+0.05) = 0.46Ω. This indicates that for the same amount of current, the 3D chip will have 0.46/0.08~6X more IR drop compared to 2D.
- High frequency behavior: The impedance values die to zero at high frequencies. The degradation is faster in 2D due to lower inductance.

- Resonant behavior: The resonant frequency is decreased as inductance has gone up. Unlike resonant frequency, the resonant peak is strongly dependent on resistive damping. Thus, the increased inductance in 3D (due to the smaller footprint) is counteracted by the increased damping provided by the larger resistance drop, yielding comparable peaks.

There are host of existing techniques like active decaps, controlled wakeup, resistive damping, etc. to deal with the worst case noise induced due to resonance [76,78,79], and can be easily extended to 3D ICs. Localized decaps help in combating local high frequency noise. In fact, in 3D ICs with many tiers, it might be even cost effective to have dedicated decap-tiers. On the other hand, DC noise is a more challenging problem and often more crucial for circuit designers.

5.4.2 Impedance Response of Power Supply in Each 3D IC Tier

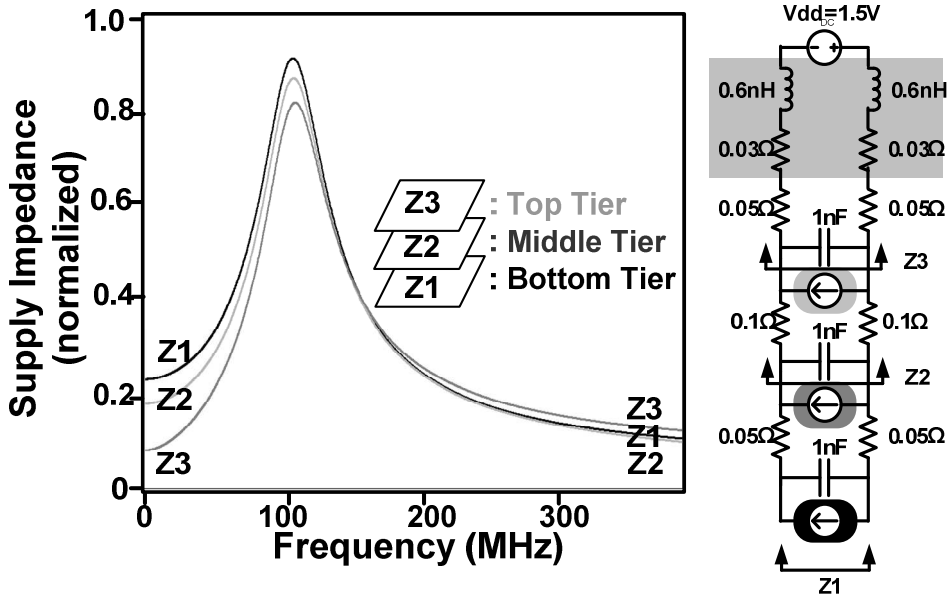


Fig. 5.14. Supply impedance response of the three tiers in a 3D IC. Impedance at the bottom tier is largest.

To understand the supply noise behavior at different tiers, we show the comparative AC impedance in Fig. 5.14, simulated using the adjacent test circuit. The key results are as follows:

- Low frequency impedance: As expected, the low frequency impedance, equivalent to the path resistance to the respective tier, shows a worsening trend for the lower levels, owing to the TSVs.
- High frequency impedance: Although a little counter-intuitive, the top tier has the maximum impedance while the middle tier, the minimum. The middle tier is in effect shielded away by the top and bottom tier decap. The effective resistance for the middle tier is then (0.1+0.1) || (0.05+0.05) = 0.066Ω, while for the top tier it is (0.1+0.1) = 0.2. The above trend is more noticeable at high frequencies beyond the resonance peak.
- Resonant behavior: Since the shielding effect mentioned above is not significant at mid-frequencies, the resonance peak follows the lower frequency trend with bottom tier being the worst case. However, there is a reduced noise offset as noted from the simulated curves. Also, since the effective capacitance is almost the same for all tiers, the resonant frequencies are all identical.

In summary, the impedance is worst for the bottom tier from low to mid frequencies around resonance, but beyond that, the top tier has slightly larger impedance. Since thermal constraints dictate that the bottom tier is likely to contain circuit blocks with large current consumption, the supply noise in the bottom tier (i.e.

product of current and impedance) will become a significant concern for 3D implementations.

5.4.3 In-situ Supply Noise Measurement

In order to capture the supply noise behavior across different tiers in a 3D IC, the test setup in Fig. 5.15 was constructed. It consists of noise sensing (*Nsen*) and generation (*Ngen*) modules in each tier. 80 TSVs for the noisy supply DVDD were put at a pitch of 5µm occupying an area of ~2500µm^2. The *Ngen* module consists of a programmable number of units, each consisting of a clock gated switch to control the current drawn from DVDD. A VCO sweeps the clock frequency. The *Nsen* differentially captures the AC noise between DVDD and DGND in frequencies ranging from 1MHz to 500MHz with a gain of 10dB. We first capture the noise spectrum in different tiers of a 3D IC by providing current excitation to that tier. From Fig. 5.6 (left), *t* shows a resonant peak while the peak noise in other two tiers is markedly reduced, -53% for *b* and -72% for *m*. This reveals significant shielding effect from tiers *t* and *b* on tier *m*. In addition, the TSVs provide decap as well as resistive damping, although those would be second order effects based on the characterized values.

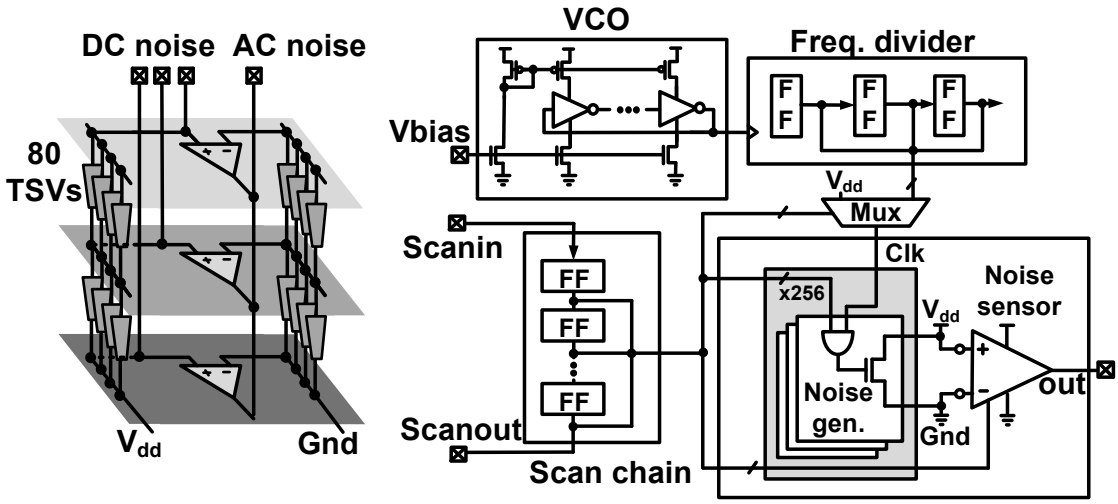


Fig. 5.15. DC and AC supply noise measurement setup. It consists of a a supply grid on each tier, with 80 stacked TSVs providing the electrical connection between tiers. The supply is fed externally through the pads on top tier. On each tier the DC noise is tapped by Kelvin probing. AC noise is sensed by an on-chip opamp based supply noise sensor. The right figure shows the details of the noise generation block. It is basically a programmable control to the supply noise current magnitude and frequency through a clock gated switch. The number of noise gen. gated on, decide the magnitude while the clock decides the noise frequency. An adjustable VCO provides a fine tuning of frequency while the divider provides the coarse tuning.

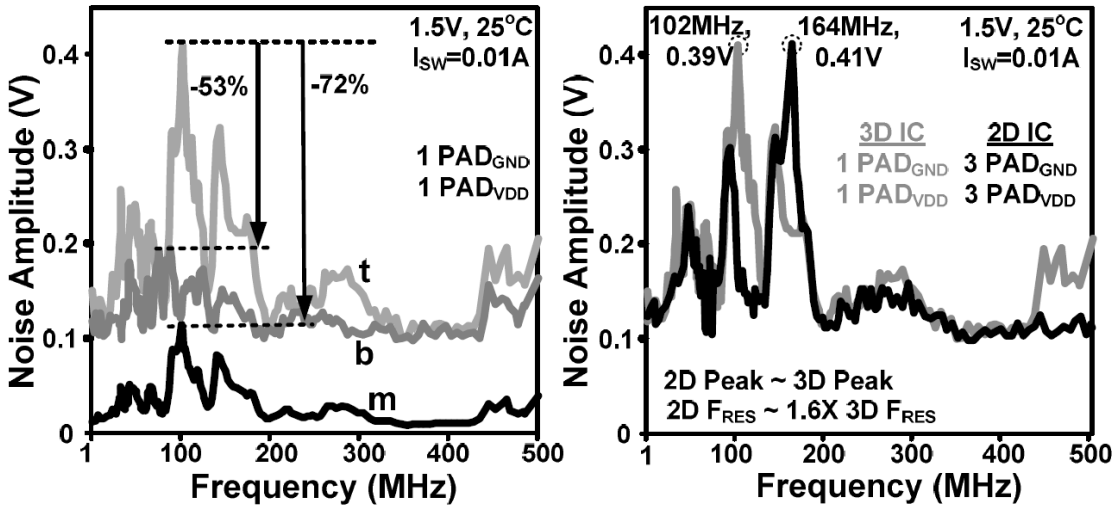


Fig. 5.16. Measured AC noise in different tiers (left) and between 3D and 2D configurations (right).

Next we compare the noise spectrum of the top tier *t* with that of a 2D circuit. From Fig. 5.16 (right), there is no significant difference in the peak noise amplitude. The higher frequency noise amplitude is also same. However, there is a shift in the resonant frequency owing to a lesser inductance with the fewer pads. Measurements indicate that the AC noise problem is largely unaffected from the 2D case for tier *t*. IR noise on the other hand, is directly affected by the reduced pads and the TSVs in the supply path. Fig. 5.17 (a) shows the comparison between 2D and 3D (tier *b*). For a 100mV supply drop, the I_{MAX} is 69% less in 3D. For the test setup, most of the supply drop contribution is due to the 3X fewer pads available as seen in Fig. 5.17(b). The contribution of TSVs to the supply drop in each tier can also be seen in Fig. 5.17(b). The *t-m* path has more TSV drop than the *m-b* path owing to the *f2b* arrangement in the former versus *f2f* in the latter as seen in Fig. 5.17(c). Projections can be made off this measured data to obtain the required TSVs and pads for achieving a particular I_{MAX}, shown in Fig. 5.18. For example, for a 7A current, 400 TSVs and 125 pads would be required for less than a 100mV IR drop.

5.4.4 Remarks on the Supply Noise Measurement Macro

The analysis in this section provides some quantitative understanding of power delivery in 3D ICs. It should be pointed out that the numbers presented here are tied to a specific process, and will change depending on the process. For example, if the technology allows TSVs with much lower resistance or area, then the impedance bottleneck in a path may be due to the supply pads, and the PSN models should account for it. Thus, as far as supply noise reduction is concerned, increasing

TSV count indiscriminately is not useful. Also, DC noise is something that can't be simply solved by adding extra decaps, which cater to resonant noise only. One approach has been to incorporate localized voltage regulator and power planes on a dedicated tier [70,73]. Clearly, this would not be very attractive unless there are many stacked tiers to amortize the extra cost. This work proposes an alternative architectural level regulation scheme that can solve the power delivery problem.

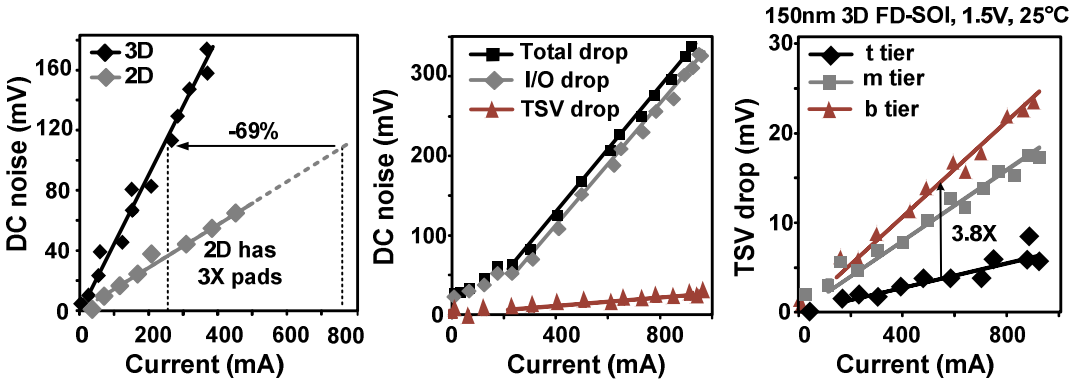


Fig. 5.17. DC noise comparison between the three tiers and between a 2D and 3D cases. Across the tiers, we see a 3.8X drop between the top and bottom tiers. However, the total IR noise in 3D IC for this process, is mostly dominated by the I/O drop. To compare 2D and 3D, we take into account the reduced footprint of 3D and thus assume 3X pads for 2D case. For a 10% VDD noise, the max current for a 3D is thus reduced about 70% from a 2D chip. This points the criticality of DC noise in 3D and shows that DC noise is more I/O dominant rather than TSVs for this process.

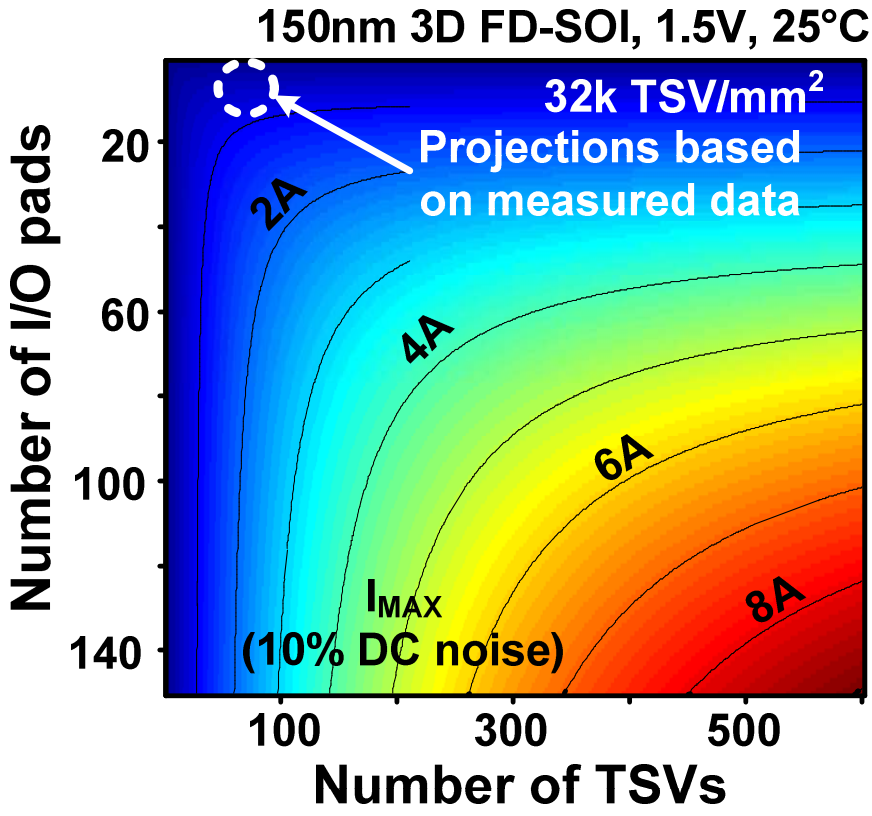


Fig. 5.18. I/O and TSV count dependency. Here we extrapolate the measured data for larger current and different TSV and I/O counts. For example, for a 4 A load, 140 pads would require atleast 150TSVs.

5.5 Multi-Story Power Delivery

5.5.1 Basic Idea

Multi-story power delivery (MSPD) is based on the idea of current recycling [71], originally proposed to mainly reduce the power dissipation in the PSN. Gu et al. [72], later demonstrated its effectiveness in reducing PSN noise. Several design issues related to separation of supply modules has made the promising technique so far impractical in traditional circuit design. Our work [32] demonstrates that MSPD is potentially much more pertinent and feasible for 3D chips because of the inherent split configuration and severer supply noise and power densities encountered.

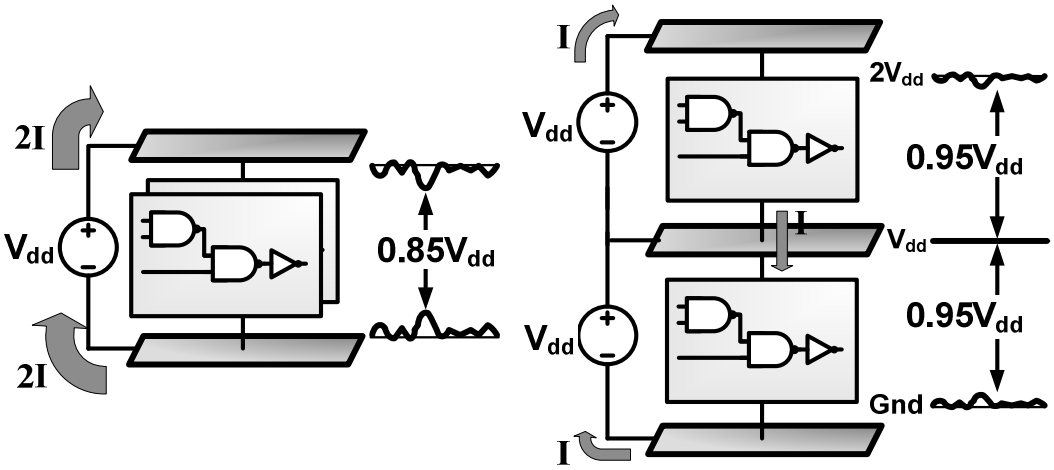


Fig. 5.19. A conv. single story case draws power at VDD and supplies current I. The same circuit if split across two stories, can be powered by a I/2 current at a 2VDD voltage [16][17]. The reduced current reduces the supply noise in I/O path. Plus, if the two stories are ensured to be balanced, the middle node is quiet. The technique is fraught with implementation challenges in a conventional bulk process, however, 3D Ics with there inherent split topology can offer readily separable power supply networks.

Fig. 5.19 explains the basic concept of multi-story power delivery. A conventional single-story structure is shown in Fig. 5.19 (a), where all circuits draw current from a single power source. Fig. 5.19 (b) shows the multi-story supply structure with sub-circuits operating between two supply stories. (Note that here; "story" is only an abstraction to illustrate the nature of the power delivery scheme, as opposed to the 3D IC architecture, where circuits are physically stacked on top of each other.) In this scheme, current consumed in the "$2V_{DD}$-V_{DD} story" is subsequently recycled in the "V_{DD}-Gnd story". Due to this internal recycling, half as much current is drawn compared to the conventional scheme, with almost the same total power consumption. A reduced current is beneficial since it cuts down the supply noise. Thus, in the best case, if the currents in the two sub-circuits are completely balanced,

the middle supply path will sink zero current. This results in minimal noise on that rail, as also illustrated in Fig 19(b).

A conventional single-story structure is shown in Fig. 5.20(a). Here, the resistance, *r*, would be the vertical path resistance which is inversely proportional to number of pads (and TSVs in 3D IC). The total switching current is denoted by *I*. With this model, we calculate the worst case DC noise and power dissipation in the PSN as $2 \cdot I \cdot r$ and $2 \cdot I^2 \cdot r$, respectively. Fig. 5.20(b) shows the electrically equivalent MSPD model employing *m*-stories. The net current is distributed in *m* equal *I/m* current blocks. Due to the increased number of supply stories, the overall effectiveness of this scheme should be seen with a fixed number of supply routes. Thus if *r* was the resistance for say *N* V_{DD} connections, it will be $0.5(m+1)\ r$ for $2N/(m+1)$ V_{DD} connections.

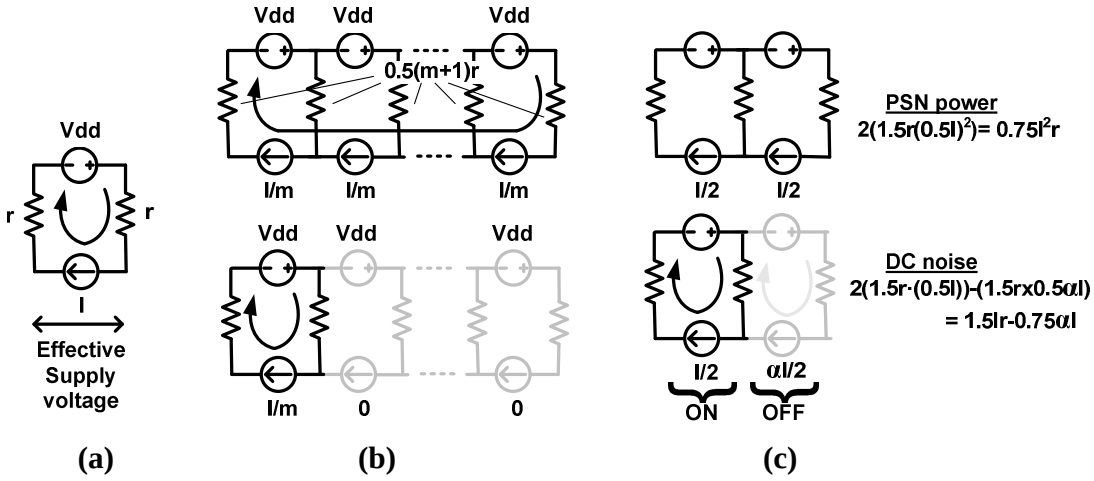


Fig. 5.20 (a) Single-story (b) Multi-story (shaded denote the off supplies) (c) Two-story PSN (left). Worst case for DC noise (right). α denotes the ratio of off current to on current.

Some heuristic results from the above proposed topology are useful in subsequent sections and we summarize them below.

- The best case is the equal simultaneous switching scenario. This happens when all stories are drawing current equal in magnitude as well as phase. Then, middle supply paths contribute no supply noise due to zero current flow in them.
- Unlike the conventional scheme in Fig. 5.20(a), the worst case condition for noise occurs when only one story is switching, while others are not as depicted in Fig. 5.20(b) (gray). The worst case noise comes out to be $I \cdot R \cdot (1+1/m)$.
- In the two-story structure of Fig 20(c), we also consider a leakage fraction, a in the off stories, assuming it to be 25-50% of the on-current in contemporary technologies. It is evident from Fig. 5.20(c), that the leakage current opposes the regular current flow and reduces the worst case drop across the common supply path. As calculated below the Fig., compared to the single-story scheme of Fig. 5.20(a), we get a DC supply noise reduction of 44% and the worst case PSN power decrease of 62.5%, assuming an α of 0.5.
- The maximum power dissipation occurs when every alternate story is off. However, in the two story case, it would be the same as the case when both stories are on. Note that here we are referring to the relatively small portion of the total power supplied to the core that gets wasted in the PSN.

Fig. 5.21 shows the plot of the worst case DC noise and PSN power versus number of stories m. Clearly, the curve shows great returns in terms of power and noise for

m=2, beyond which the returns diminish. Considering the overhead for partitioning the circuit and generating multiple power supplies, a two-story network is preferable.

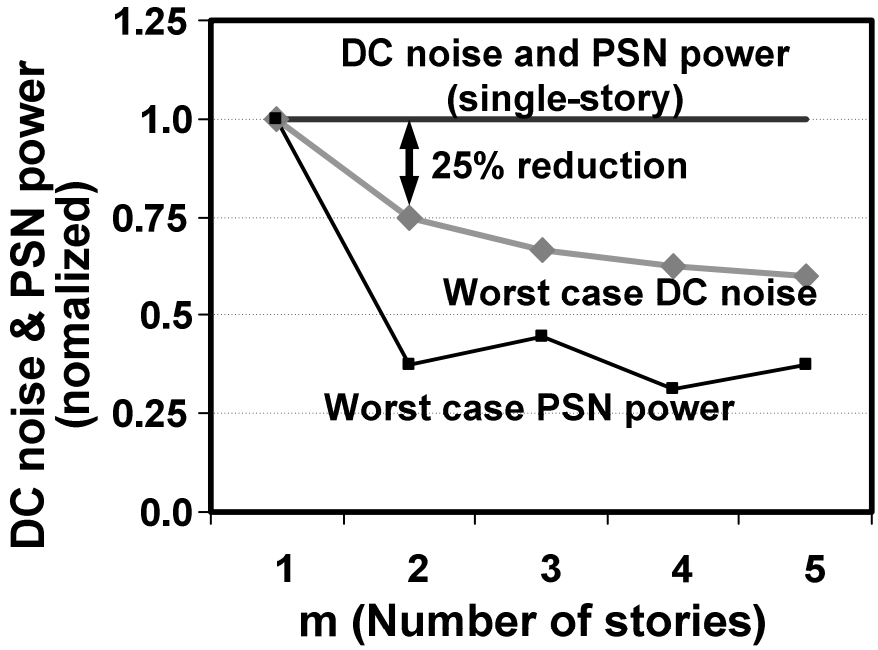


Fig. 5.21. DC noise and power consumption for different number of stories, m. DC noise exhibits a diminishing reduction with m. Clearly, m=2 provides best returns, considering the implementation overhead of additional stories.

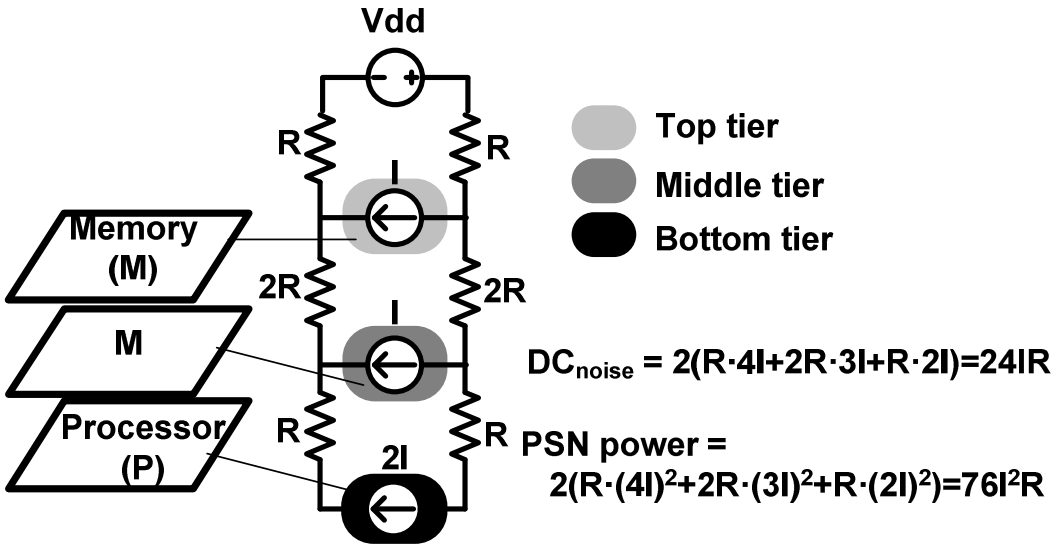


Fig. 5.22. PSN model of M-M-P architecture in a single-story 3D IC design. The equations for PSN power and DC noise at bottom tier are shown alongside.

Next, we try to extend the idea of MSPD to the realm of 3D ICs. As we discussed earlier, their inherent split configuration makes them ideal for MSPD implementation

5.5.2 Multi-story PSN for a Memory-Memory-Processor Architecture

Fig. 5.22 is the DC 3D IC model for a memory (M), memory (M), processor (P) stacked configuration. To model the difference between M and P blocks, the latter is assumed to draw twice the current of the former. We denote the two currents by I and 2I, respectively. The tier-tier path impedance is denoted in terms of R. Note that R is inversely proportional to number of vertical paths comprised of TSVs and supply pads. The equations for the worst case power dissipation in the supply nets and the worst case DC noise are depicted alongside the figure.

Considering the benchmark model for a 3D IC in Fig. 5.22, the application of MSPD can lead to a variety of different electrically equivalent architectures, depicted in Fig. 5.23-26. Here, the tier-tier per-path impedance is denoted in terms of R. Note that MSPD requires another supply rail, implying number of supply rails have increased by a factor of 3/2. If we assume that all structures are normalized to a fixed number of supply paths, each supply rail in the latter will have two-thirds the number of dedicated paths. This will correspond to a 3/2 fold impedance. Now we will consider each of these structures in detail and comment on their applicability.

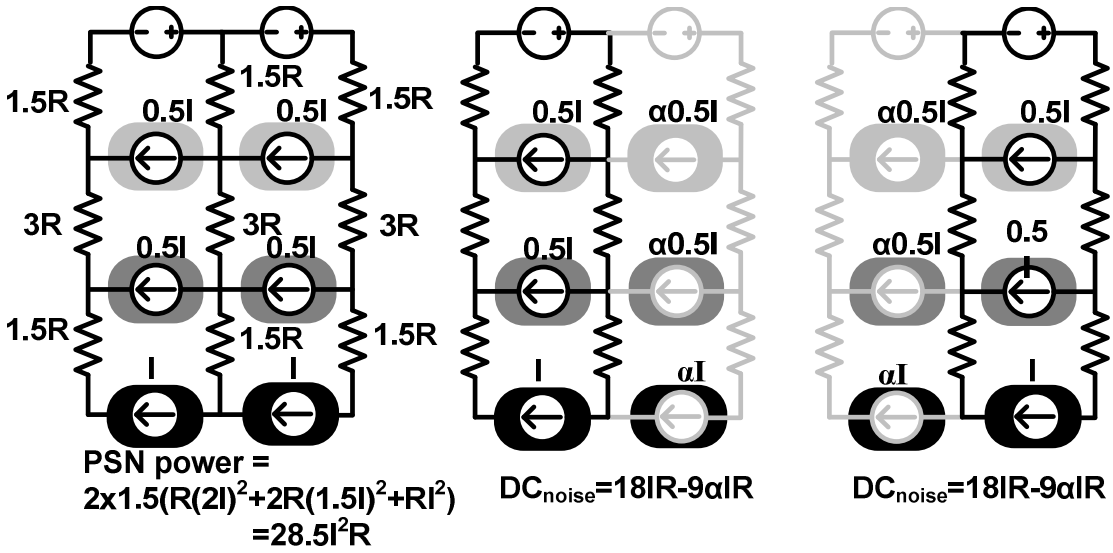


Fig. 5.23. Balanced two-story power delivery scheme in M-M-P architecture. The grayed portions emphasize the off-stories in the worst case supply noise situations. A factor 1.5 is incorporated in the resistance values to account for reduced number of TSVs per supply path. α denotes ratio of off current to on current.

Balanced MSPD: A fine-grained application of MSPD to each tier in a 3D IC can yield the balanced MSPD configuration of Fig. 5.23(a). Here, the power supply domain of each tier has been split into two equal stories, with the current from one story being recycled to the other, within and across the different tiers.. Fig. 5.23(b) and (c) show the two worst case possibilities, with the faded figure showing the off part conducing only leakage current. Thus at 50% leakage (α=0.5), we get a 44% reduction in DC supply noise, while a 62.5% decrease in PSN power calculated from the resistive dissipation in Fig. 5.23(a). The base case for comparison is the topology in Fig. 5.22. Note that these results are identical to ones from Fig. 5.20(c)

The balanced MSPD scheme leverages its inherent balanced topology to obtain maximal reductions in supply noise levels. However, the presence of multiple

supply rails poses a problem for designers especially in a regular bulk process, since NMOS devices on each tier have to share the same body bias. We now propose an alternate scheme that solves this problem.

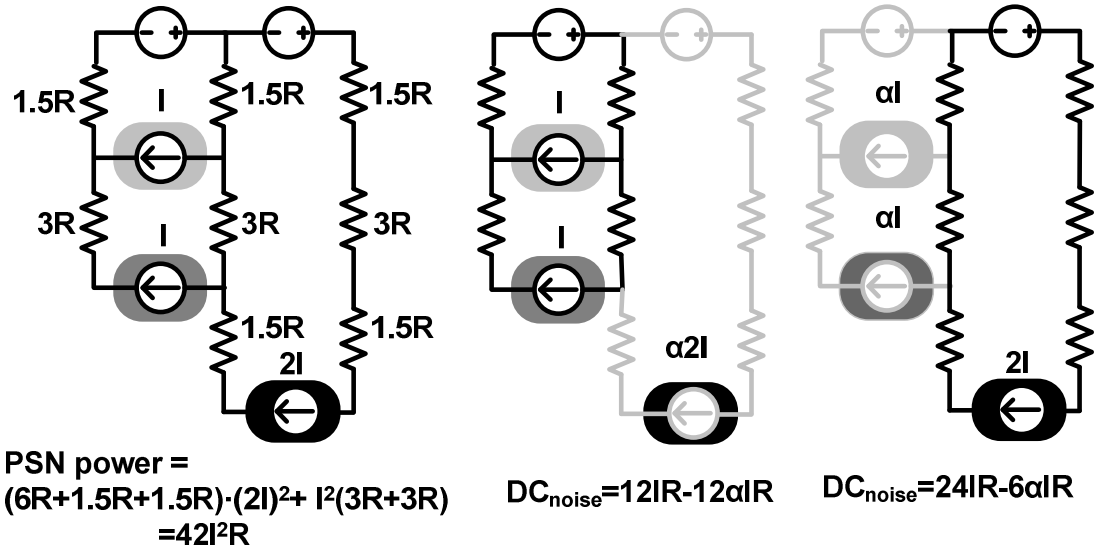


Fig. 5.24. Coarse two-story scheme in a M-M-P architecture. Each tier has just one set of supplies, while current is transferred to a different story lying in a different tier. DC_{noise} equations show the unevenness in the two worst cases.

Coarse-grain MSPD in M-M-P stack: Fig. 5.24(a) is a coarse-grain MSPD approach exploiting the readily-segregated tiers in a 3D IC. The operating current is recycled between the processor in the bottom tier and the memories in the other two tiers. Here, in spite of maintaining current recycling from a higher supply story to a lower one, each tier has only one dedicated story and single body, which can greatly simplify the implementation of this scheme. The worst case for PSN power, represented by Fig. 5.13(a), yields a value of $42 \cdot I^2 \cdot R$, a reduction of 45%.compared to the base single-story case. By analyzing the two cases in Fig. 5.24(b) and (c), separately for IR drop we find the worst case noise is given by:

$$DC_{noise} = \max((24IR - 6\alpha IR), (12IR - 12\alpha IR))$$

Here α denotes the ratio of off-current to on-current in a particular story. Typically, this could be anything between 25-50%. At 0% α, it equals $24 \cdot I \cdot R$, which shows little improvement from the single-story case. At higher leakage currents, the effectiveness is better than the balanced model but is still limited by the skew of the DC noise in the two worst case possibilities, as seen in the above equation.

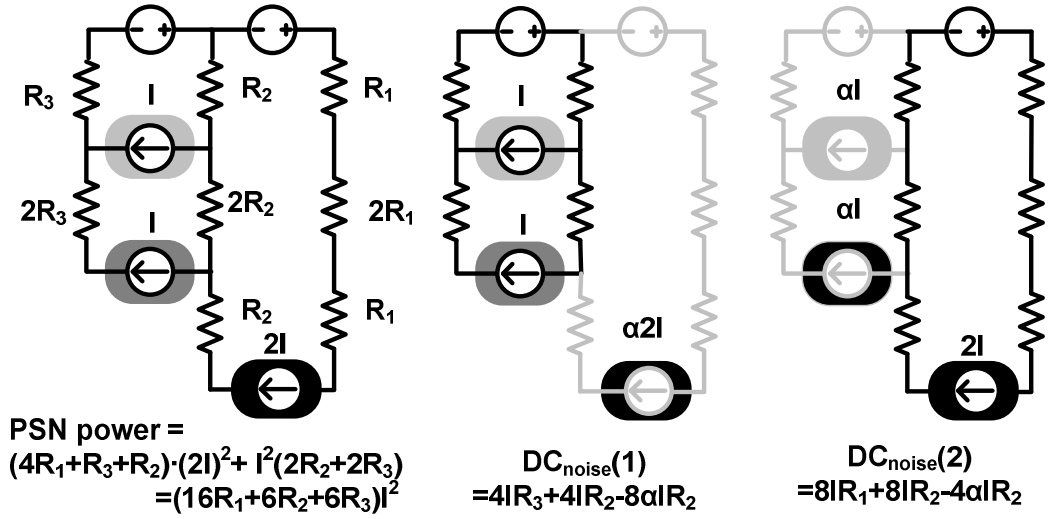


Fig. 5.25. Coarse two-story scheme in an M-M-P architecture with TSV redistribution. Equations in Fig. 5.13 showed skewed worst cases. Redistributing TSVs between different supply rails to make them even can optimize the overall worst case noise.

TSV Redistributed Coarse-grain MSPD: There is some scope for improvement in the coarse-grain scheme by redistributing the TSVs for different supply paths to optimize the overall worst case. Fig. 5.25(a) shows the same circuit with a non-uniform via distribution, using the variables R_1, R_2 and R_3 which are not necessarily equal. The worst case for PSN power is Fig. 5.14(a). The two extreme cases with the worst case DC_{noise} are depicted in Fig. 5.25(b) and (c). Thus, we formulate the optimal via

distribution condition for minimal DC noise as a choice of R_1, R_2 and R_3 for which *max ($DC_{noise}(1)$, $DC_{noise}(2)$)* is minimized with the fixed TSV constraint:

$$N_1 + N_2 + N_3 = 2N \quad or \quad \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} = \frac{2}{R}$$

Intuitively, the optimization should converge towards making the two worst cases equal. The DC noise results are presented in Table 5.1 for different α values. Thus, the proposed optimized scheme offers a 22-34% improvement in DC noise. Simultaneously, it would decrease the PSN power by as much as 37% (for α=50%).

It should be noted that the above optimization was done to decrease the IR drop. Another criterion could be to minimize the PSN power expression shown in Fig. 5.25(a). Hence, we reformulate the *TSV optimization criterion for minimizing the power supply network as a choice of N_1, N_2, N_3 (or R_1, R_2, R_3) for which $F=16/N_1+6/N_2+6/N_3$ is minimized with the constraint that $N_1+N_2+N_3=2N$.* We substitute $N_1=2N-N_2-N_3$ into the expression for F, take partial derivatives with respect to N_2 and N_3 and equate to zero. We obtain $N_1=0.89N$ ($R_1=1.12R$) and $N_2=N_3=0.55N$ ($R_{2=}R_3=1.8R$). This yields an improvement in PSN power efficiency by 48% but degrades the supply noise.

It is important to emphasize that the balanced topology of Fig. 5.23 is more preferable against the coarse topology of Fig. 5.24 or 25 for the M-M-P architecture being considered here. The latter topology tries to balance the processor current with the memory current in the upper tiers. This may not yield significant noise benefit in the case when the processor current is much larger than memory current, making the

two worst conditions for DC noise too skewed to seek any advantage from via optimization. However, the situation is different in a M-P-P or M-M-M stack.

Coarse-grain MSPD in an M-P-P stack: Fig 26(a) is a representation of a M-P-P stack in a conventional 3D IC. Fig. 5.26(b) is an application of the coarse MSPD idea to this stack. The implementation is easy, since the different tiers can be readily separated as independent memory sub-blocks with different supply stories. The analysis follows that of the topology in Fig. 5.22, except there is better balance between the middle and bottom tier currents. Thereby, little further optimization is required for noise. With α=0.5, this scheme offers a 40% and 52% reduction in noise and PSN power, respectively (The comparison was made with a slightly altered benchmark owing to different stack arrangement). Removing heat from the middle tier is a challenge [5], unless better cooling techniques are incorporated. The stack ordering can also be changed (for example changed to P-M-P), if future cooling techniques allow sufficient outflow of heat from top and bottom tiers. Alternatively, if we have an M-M-M stack, application of coarse MSPD scheme can promise significant noise reduction, but the situation is less critical due to smaller currents.

Leakage percent, α	Via distribution (in fraction)	DC_{noise} red. (%)	R1, R2, R3 (in R units)
0	0.86, 0.86, 0.28	23	1.16,1.16, 3.5
25	0.95, 0.79, 0.27	28	1, 1.27, 3.7
50	1, 0.75, 0.25	34	1, 1.33, 4

Table 5.1. DC noise optimization criterion at different leakages.

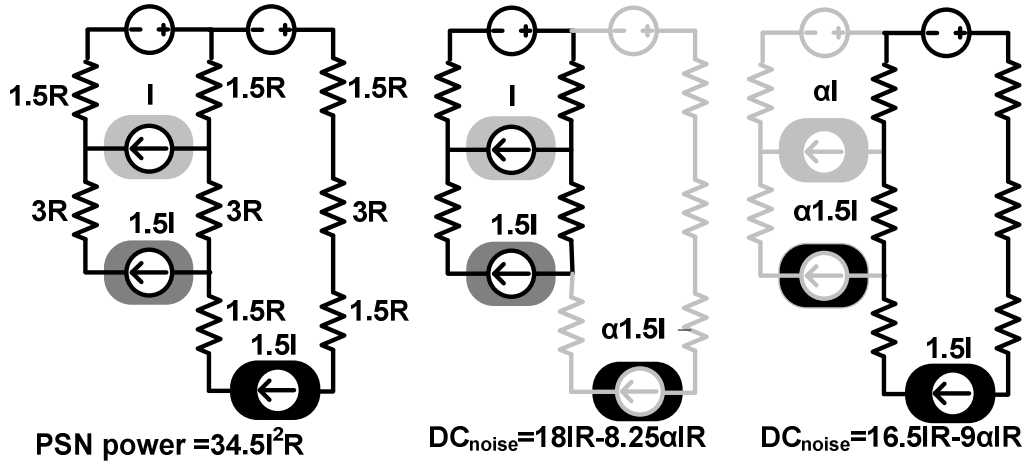
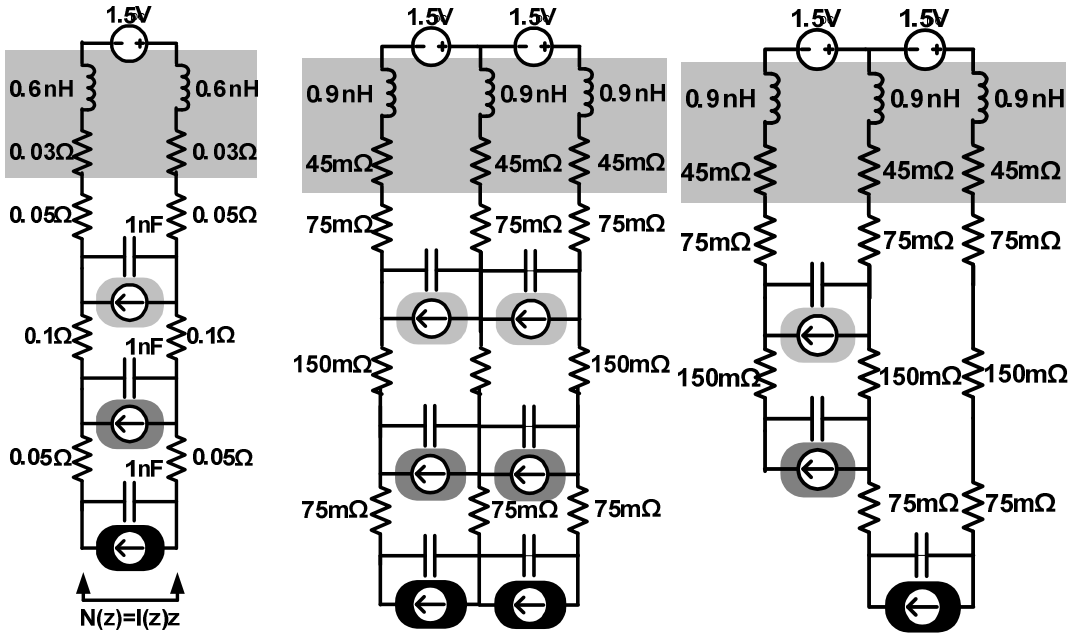


Fig. 5.26. Coarse two-story structure in an M-P-P stack. The coarse MSPD idea is particularly attractive for the M-P-P structure rather than an M-M-P one, and can provide better current recycling between stories across different tiers without TSV redistribution.

5.5.3 Impact of MSPD on AC Supply Noise

We had stressed upon DC noise until now based on the reasoning developed in section II, whereby DC rather than AC noise is portrayed to be the greater issue in 3D ICs. MSPD works well against DC noise and the worst case happens when one-story is on, while the second is off. However, the net supply noise is a superposition of noise with currents at all frequencies including DC. In principle, MSPD helps in cutting down AC noise on the common middle node, provided the currents meet in-phase. Otherwise, if the currents are out of phase, they add up and flow through the middle node, exacerbating the worst case noise situation. Thus, even though the magnitudes of currents may be equal, they could be offset in phase preventing proper current balancing. Next, we seek to understand this issue quantitatively through some simple AC models.

Here, we extend the modeling approach developed in Section II, to MSPD schemes. Fig. 5.28 shows the ensuing models obtained. Fig. 5.28(a) is the conventional 3D IC. The worst case supply noise is assumed to occur when there is simultaneous switching noise in all three tiers. Thus, equal small signal current sinusoidal excitation is provided to all blocks, and the supply is monitored at the bottom tier. Fig. 5.28(b) is the balanced MSPD structure, while Fig. 5.28(c) is the coarse MSPD topology. The total decap is kept the same.



(a) **(b)** **(c)**

Fig. 5. 27. AC analysis of MSPD (a) Benchmark 3D IC presented earlier (b)Balanced MSPD 3D IC (c) Coarse MSPD 3D IC. To obtain the MSPD models, we multiply a factor of 1.5 to the values of the various parasitic components in the benchmark model. This accounts for reduced number of TSVs per supply path.

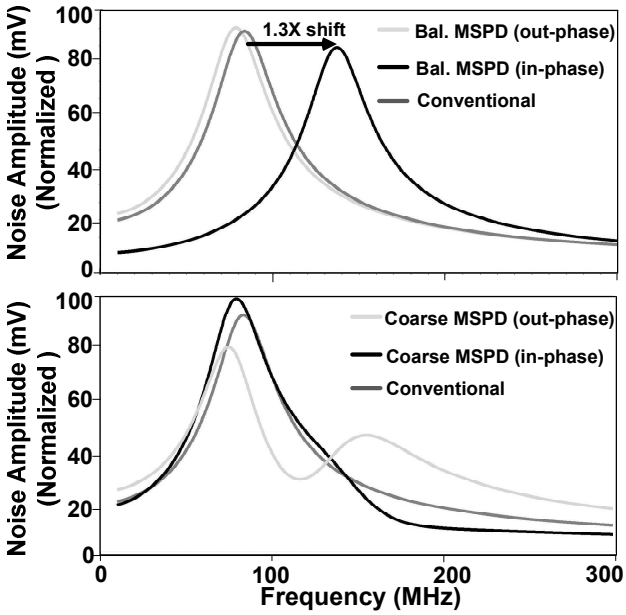


Fig. 5.28. AC Noise Spectrum for balanced MSPD and coarse MSPD compared against the conventional 3D IC case. Both the in-phase and out-phase cases are shown.

Scheme	Application	Implementation	DC_{noise} reduction	Power reduction
Balanced PSN	M-M-P	Easy in SOI Difficult in bulk	Best	Best
Coarse PSN	M-M-P	TSV/supply pad redistribution required	Good	Good
Coarse PSN	M-P-P	Simpler in bulk and SOI, thermal issues for middle tier	Reasonable	Reasonable

Table 5.2. Overview of various MSPD schemes.

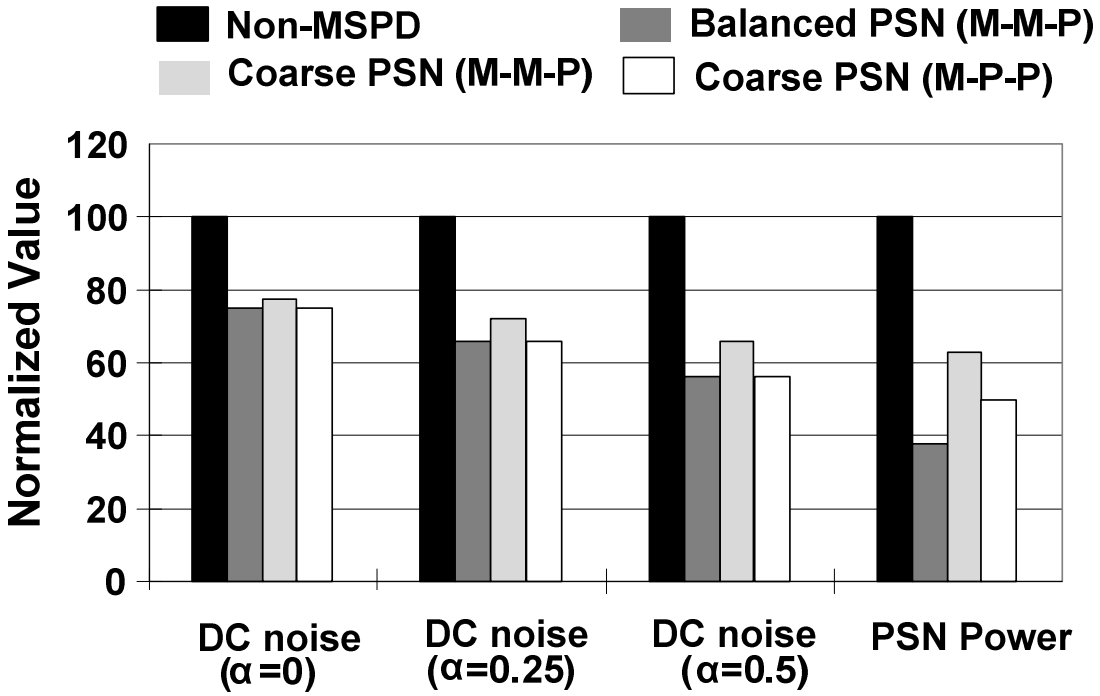


Fig. 5.29. DC noise and PSN power of different schemes. DC noise for balanced PSN in M-M-P and for coarse PSN in M-P-P demonstrate best improvements. Note that the noise and power values are normalized against the corresponding non-MSPD scheme.

5.5.4 Summary of MSPD Schemes

Table 5.2 presents an overview of the entire section, demonstrating that MSPD can promise substantial PSN noise reduction, with a caveat on implementation feasibility. Note that coarse MSPD based schemes exploits the inherent split configuration of 3D ICs and should be the topology of choice for easier implementation in a bulk process. While the balanced MSPD scheme, is preferable in a SOI implementation. The main difficulty in the balanced MSPD implementation is need of isolated bulks if different stories need to be integrated on the same tier. This is easy in a silicon-on-insulator process but practically impossible in a bulk process.

Also, even if they are in different tiers as in coarse MSPD scheme, any inter-story communication requires some level converters.

A more quantitative depiction is shown in Fig. 5.29. Here, the three schemes are compared against each other, vis-à-vis their respective PSN DC noise and power dissipation. The values are normalized with the corresponding nominal non-MSPD 3D IC model, shown in black (give them the same color). Clearly, the MSPD technique promises a DC noise reduction of 20-40% with 50% leakage. It is again interesting to note that DC noise is reduced with leakage in a MSPD scenario. Measured data shown for a 3D-MSPD scheme in Fig. 5.30 reveals a 1.4-2.8x boost in I_{MAX}.

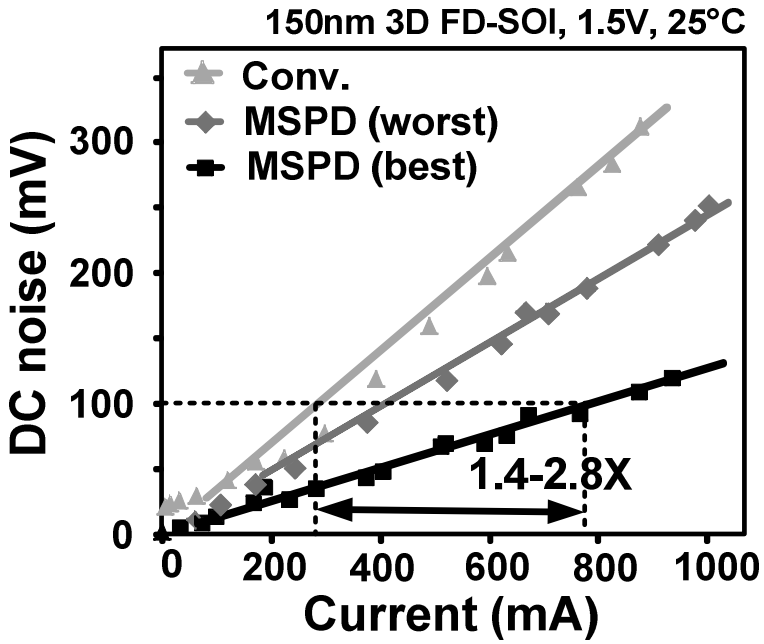


Fig. 5.30 Measured DC noise benefit of MSPD from a conv. case. For fair comparison, we assume equal number of pads and TSVs. Even in the worst case, a 1.4X max. current is obtained

Some remarks can be made as follows:

- *Low frequency*: We see that MSPD in the best case of in-phase currents, promise some noise reduction, around 13% in balanced case. The out of phase case yields a slight worsening of low frequency noise.
- *Resonant frequency:* For balanced MSPD schemes versus the conventional case, we see almost identical behavior for out-phase noise. Even the peaks are identical. This can be understood, if we consider that we can invert one of the story and superpose on the other, and we still end up with the conventional structure. The shift in the resonant frequency for in-phase case is due to effective reduction in inductance, as the middle supply path is virtually invisible. For coarse MSPD, the in-phase component has two dominant resonances; attributed namely to part of the current that gets recycled, and to the other part flowing into the middle supply path. The former gives the smaller peak, at higher frequency, while latter the larger peak at the smaller frequency.
- *High frequency:* All of them follow the same trends. The structures with a larger effective inductance, L, degrade slowly.

In general, AC noise in MSPD vis-à-vis conventional case is not excessively degraded, even for the worst case out-phase scenario. It does however complicate the analysis with its phase dependent behavior. In real case, the relation between the phases of the two stories is dictated by the nature of the circuits, and is expected to lie between the two extreme scenarios.

5.6 Layout Considerations in MSPD Implementation

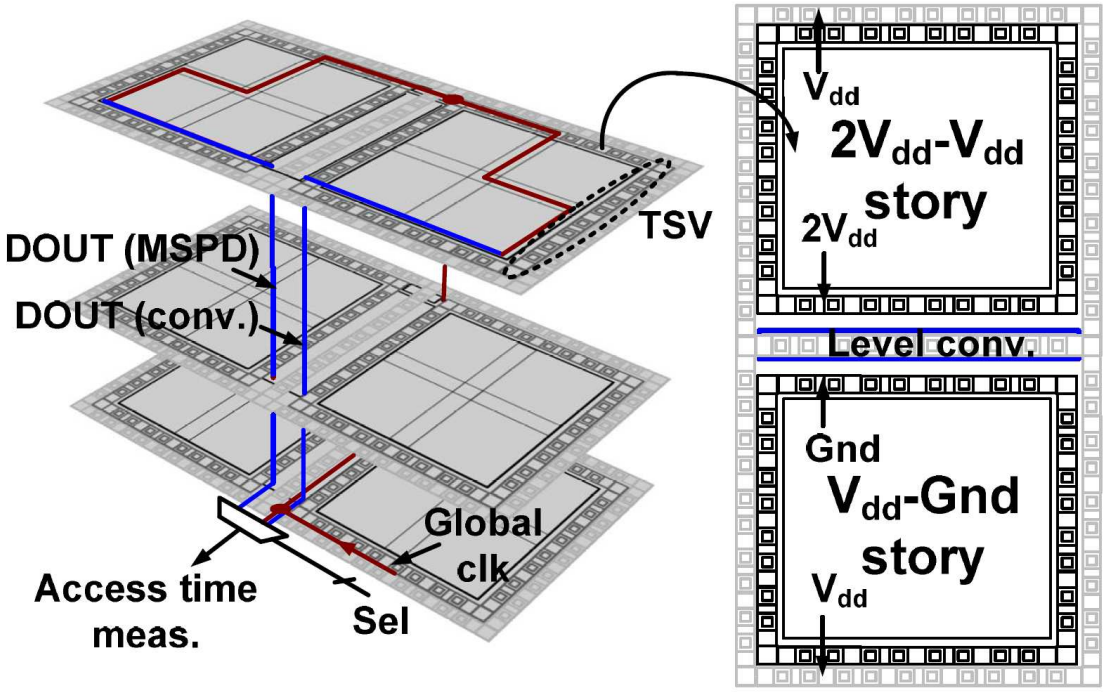


Fig. 5.31. MSPD demonstrated on a 3D SRAM. All tiers were made identical except for the I/O and control path on top tier for modular implementation. Stacked TSVs connect to the tiers on the periphery. Each tier consists of two stories. We resorted to a concentric supply ring structure for the two stories. The 2VDD and GND are dedicated to respective story while VDD is shared.

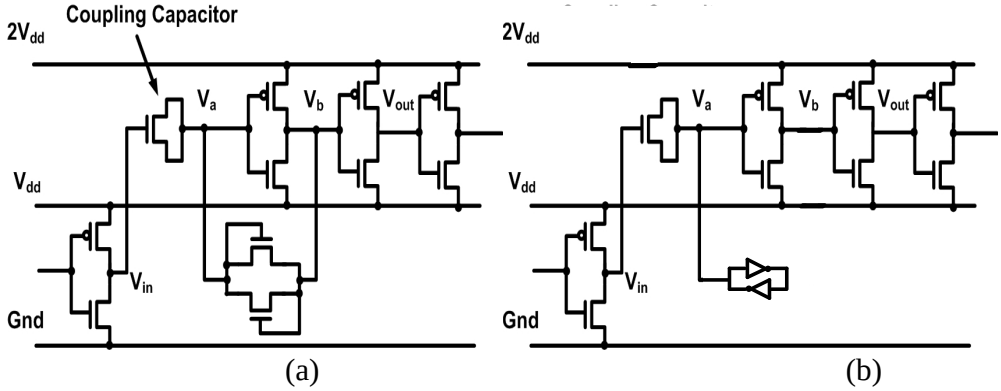


Fig. 5.32. Capacitive coupling circuits for inter-story data transfer (a) Half-keeper design [72] (b) Full keeper design

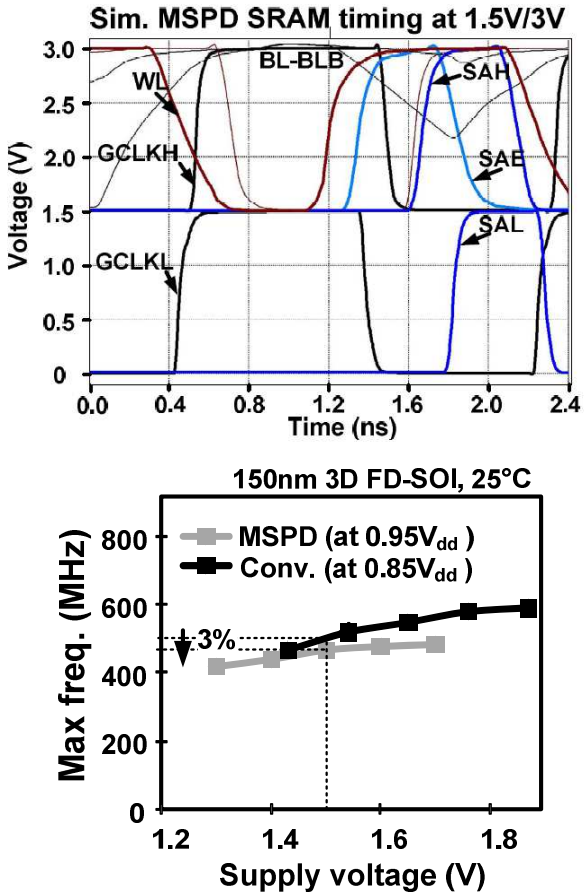


Fig. 5.33. (a) Simulation waveshot of the MSPD SRAM timing. The external signals are at nominal VDD while the internal ones are level up converted. To evaluate, the performance impact of MSPD we should take into account the difference in supply noise plus the overhead due to level converters. Simulations show a 19% FMAX improvement and 15% latency improvemnets assuming 10% extra DC noise in conv. (b) Measurements however didn't reflect this and we obtained an anomolous 3% degradation in MSPD performance at the same normalized conditions, which we suspect is due to process variation.

For demonstrating the feasibility of the proposed scheme, we laid out a 3x128kb 3D SRAM, using MITLL-0.15µm FD-SOI design kit. Since, this was a SOI process where the transistor bodies are isolated; the balanced MSPD scheme (from Fig. 5.25) was suited for implementation, rather than the coarse one (from Fig. 5.26) and has been illustrated in Fig. 5.31 (a). Each tier was split up into V_{DD}-Gnd and $2V_{DD}$-V_{DD} stories. Each story was a bank of four 16kb SRAM arrays and was powered by

appropriate supply rails (highlighted in figure for visibility) that are laid in a concentric ring topology. TSVs are placed on these rails for power supply connection to arrays in the lower tiers. TSVs were also used for inter-tier signal transfer, at an array level granularity.

Supply voltage	1.5V
Array size	393kb, 6T cell
Array area	2.1mm x1.9mm
F_{MAX} and power	470MHz,10.7mA
MSPD area overhead	3.5%*
MSPD F_{MAX} overhead	3%**
DC noise reduction	2.8X (same power)

*Sub-array granularity of level conv
** Suspected cause process variation

Table 5.3. 3D-MSPD SRAM measurement summary. The area overhead is about 3% for this single sub-array implementation and is expected to go down further. The scheme boosts DC noise 2.8X from a conv. case, that can bring about substantial performance improvement.

A read access cycle simulation for the $2V_{DD}$-V_{DD} domain SRAM bank is shown in Fig. 5.32(a). For this story, all external input and output signals are in the V_{DD}-Gnd domain, while the internal ones in the $2V_{DD}$-V_{DD} domain. Capacitive coupling based level up/down converters were used for signal translation between these two domains. Fig. 5.32 shows two kinds of level up-converters employed. Fig. 5.32(a) was chosen for clock and scan signals, while Fig. 5.32(b) for data signals. The former, with the diode-based half keeper, although area-efficient, consumes larger standby power and requires initialization at internal node V_a, if used for non-toggling data signals, unlike the latter.

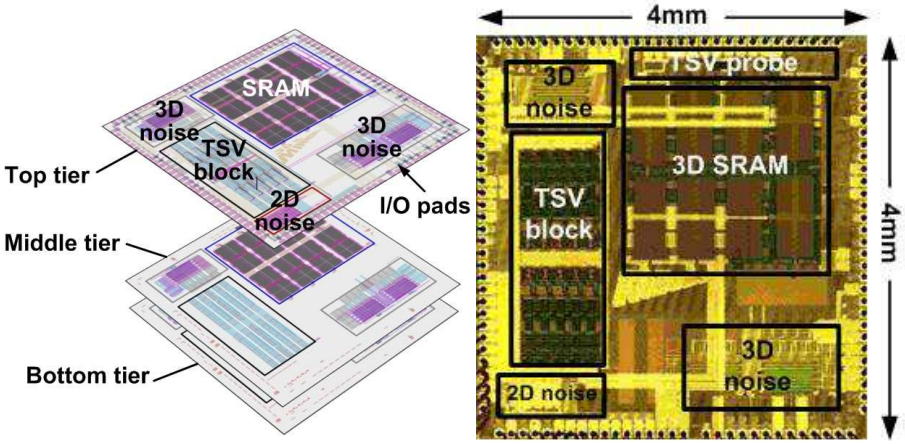


Fig. 5.34. Die microphotograph. The chip looks somewhat blurry as we are looking into the back side of the top tier. From the breakdown of the layouts, we can see that the overall design was predominantly very modular, to avoid designing each tier seperately. The point of departure for the top tier was 2D noise sensing block and I/O pads.

To evaluate, the performance impact of MSPD we should take into account the difference in supply noise plus the overhead due to level converters. Simulations show a 19% FMAX improvement and 15% latency improvements assuming 10% extra DC noise in conv. (b) Measurements however didn't reflect this and we obtained an anomolous 3% degradation in MSPD performance at the same normalized conditions, which we suspect is due to process variation. Table 5.3 shows a summary of measured and simulated data obtained from the SRAM chip. The die microphotograph is shown in Fig. 5.34. The chip looks somewhat blurry as we are looking into the back side of the top tier. From the breakdown of the layouts, we can see that the overall design was predominantly very modular, to avoid designing each tier seperately. The point of departure for the top tier was 2D noise sensing block and I/O pads.

5.6. Conclusions

Supply noise measurements from a 3D IC have been presented for the first time. IR noise rather than Ldi/dt noise is shown to be dominant due to the fewer supply pins and the additional resistance from the through-silicon vias (TSVs). Kelvin probing for IR noise reveals that the effect of pins is significantly more than TSVs. A novel multi-story power delivery is demonstrated for a 393kb SRAM suppressing the IR noise by 30-70%.

Chapter 6

Conclusion

On-chip reliability monitors score over traditional device probing based approaches in scalability and test time and effort. For BTI and RTN characterization, they provide orders better timing resolution which cannot be obtained through conventional means. Three such reliability monitors were described namely for gauging the impact of BTI, TDDB and RTN in circuits.

First the SRAM aging macro was described. Recovery free evaluation of BTI in SRAM is challenging due to massive data to be captured within a few microseconds. This work provides a methodology to remove the noise in SRAM measurements due to BTI recovery. We incorporate two techniques, namely pseudo read with stressed deferred readouts and flip-latch-restore with intermittent row-wise scanout, for read and write respectively on a test chip in 32nm HKMG SOI. Small T_{MEAS} of around 3µs at 0.5V, yields 35mV accuracy in read V_{MIN} and 10X accuracy in BFR over conventional approaches.

Next, we proposed an array-based gate dielectric breakdown characterization approach called CLIP, to reduce the stress time and silicon area by a factor proportional to the number of DUT cells in the array by stressing all cells in the array in parallel. The essential part is a flexible DUT cell that can be stressed in isolation

without thicker t_{ox} FETs to 4 times the VDD, enabling accurate lifetime prediction under different ON and OFF state TDDB modes for both low voltage core and high voltage IO devices.

In the second part of the thesis, we investigated power delivery issues for 3D ICs. We presented supply noise measurements from a 3D IC have been presented for the first time. IR noise rather than Ldi/dt noise was shown to be dominant due to the fewer supply pins and the additional resistance from the through-silicon vias (TSVs). Kelvin probing for IR noise revealed that the effect of pins is significantly more than TSVs. A novel multi-story power delivery was demonstrated for a 393kb SRAM suppressing the IR noise by 30-70%.

Bibliography

[1] 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 Journal of Research and Development, pp. 287-298, March/May 2002.

[2] 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 International Symposium on the Physical and Failure Analysis of Integrated Circuits, pp. 1-6, 2008.

[3] Tan, Y.C., Tan, C.M., Zhang, X.W., Chai, T.C. and Yu, D.Q., “Electromigration performance of Through Silicon Via (TSV) – A modeling approach,” Microelectronics.

Reliability Journal., Vol. 50, Issues 9-11 (2010), pp. 1336-1340.

[4] C. M. Compagnoni, R. Gusmerodli, A. S. Spinelli, A. l. Lacaita, M. Bonanomi, A. Visconti, “Statistical Model for Random Telegraph Noise in Flash Memories”, IEEE Transactions on Electron Devices, vol. 55, No. 1, 2008

[5] X. Wang, P. Rao, A. Mierop, A. Theuwissen, “Random Telegraph Signal in CMOS Image Sensor Pixels”, IEEE International Electron Devices Meeting, 2006.

[6] M. Agostinelli, “Erratic Fluctuations of SRAM Cache Vmin at the 90nm Process Technology Node”, IEEE International Electron Devices Meeting, 2005.

[7] K. Ito, T. Matsumoto, S. Nishizawa, H. Sunagawa, K. Kobayashi, H. Onodera, "The Impact of RTN on Performance Fluctuation in CMOS Logic Circuits", IEEE International Reliability Physics Symposium, 2011.

[8] 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.

[9] 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 Transactions On Electron Devices, vol. 56, no. 9, pp. 1943-1952, September 2009.

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

[11] T. Grasser et al., "Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise", IEEE International Electron Devices Meeting, 2009.

[12] A. Bansal et al., "Impacts of NBTI and PBTI on SRAM static/dynamic noise margins and cell failure probability", Microelectronics Reliability Journal, Elsevier, 2009

[13] V. Huard et al., "NBTI degradation: From transistor to SRAM arrays," IEEE International Reliability Physics Symposium, 2008

[14] V. Huard et al., "Managing SRAM reliability from bitcell to library level," IEEE International Reliability Physics Symposium, 2010

[15] J. Keane et al., “An Array-Based Odometer System for Statistically Significant Circuit Aging Characterization”, IEEE Journal of Solid-State Circuits, Oct. 2011

[16] A. T. Krishnan et al., “SRAM Cell Static Noise Margin and VMIN Sensitivity to Transistor Degradation”, IEEE International Reliability Physics Symposium, 2006

[17] T. Kim et al., "An SRAM Reliability Test Macro for Fully Automated Statistical Measurements of Degradation," IEEE Custom Integrated Circuits Conference, 2009

[18] S. Drapatz et al., "Impact of fast-recovering NBTI degradation on stability of large-scale SRAM arrays," European Solid-State Device Research Conference,, 2010

[19] P. Jain, A. Paul, X. Wang C.H. Kim, " A 32nm SRAM Reliability Macro for Recovery Free Evaluation of NBTI and PBTI Induced Bit Failures", submitted to IEEE International Electron Devices Meeting, 2012.

[20] J. Keane, S. Venkatraman, P. Butzen, and C.H. Kim, "An Array-Based Test Circuit for Fully Automated Gate Dielectric Breakdown Characterization", IEEE Transactions of VLSI Systems 2009.

[21] P. Jain, J. Keane, C.H. Kim, "An Array-Based Chip Lifetime Predictor Macro for Gate Dielectric Failures in Core and IO FETs ", European Solid-State Device Research Conference, Sep. 2012 (to appear)

[22] J. Keane, “On-Chip Circuits for Characterizing Transistor Aging Mechanisms in Advanced CMOS Technologies”, Ph.D. dissertation, University of Minnesota, Minneapolis, MN, U.S.A., 2010.

[23] N. Tega et al., "Impact of HK / MG stacks and future device scaling on RTN," IEEE International Reliability Physics Symposium, 2011.

[24] N . Tega, “Study on Variability in Transistor Characteristics due to Random Telegraph Noise”, IEEE Workshop on Variability Modeling and Characterization, 2009

[25] N. Tega et al., " Impact of threshold voltage fluctuation due to random telegraph noise on scaled-down SRAM," IEEE International Reliability Physics Symposium, 2008.

[26] W. Feng et al., "Fundamental origin of excellent low-noise property in 3D Si-MOSFETs ~ Impact of charge-centroid in the channel due to quantum effect on 1/f noise”, IEEE International Electron Devices Meeting, 2011.

[27] T. Grasser et al., “Switching oxide traps as the missing link between negative bias temperature instability and random telegraph noise”, IEEE International Electron Devices Meeting, 2009.

[28] T. Nagumo et al., “New Analysis Methods for Comprehensive Understanding of Random Telegraph Noise”, IEEE International Electron Devices Meeting, 2009.

[29] K. Banerjee, S. Souri, P. Kapur, K. Saraswat, “3-D ICs: A Novel Chip Design for Improving Deep Sub-micrometer Interconnect Performance and Systems-on-chip Integration”, Proceedings of the IEEE, pp. 602-633, May 2001.

[30] K. Bernstein, P. Andry, J. Cann, P. Emma, D. Greenberg, W. Haensch, M. Ignatowski, S. Koester, J. Magerlein, R. Puri, A. Young, "Interconnects in the Third Dimension: Design Challenges for 3D ICs," Design Automation Conference, pp.562-567, June 2007.

[31] R. Mahajan, R. Nair, V. Wakharkar, J. Swan, J. Tang, and G. Vandentop, “Emerging Directions for Packaging Technologies”, Intel Technology Journal, pp. 62–75, 2002.

[32] P. Jain, T. Kim, J. Keane, C. H. Kim, “A Multi-story Power Delivery Technique for 3D Integrated Circuits”, IEEE International IEEE Symposium on Low Power Electronics and Design, pp. 57-62, Aug. 2008.

[33] P. Jain, D. Jiao, X. Wang, C.H. Kim, "Measurement, Analysis and Improvement of Supply Noise in 3D ICs", VLSI Circuits IEEE Symposium, Jun. 2011

[34] J. Keane, T. Kim, and C.H. Kim, "An On-Chip NBTI Sensor for Measuring PMOS Threshold Voltage Degradation", IEEE Trasactions on VLSI Systems, June 2010

[35] C. R. Parthasarathy, M. Denais, V. Huard, G. Ribes, E. Vincent, A. Bravaix, "New Insights into Recovery Characteristics Post NBTI Stress," IEEE International Reliability Physics Symposium, 2006

[36] S. Natarajan et al., “32nm Logic Technology Featuring 2nd -Generation High-k + Metal Gate transistors, Enhanced Channel Strain and 0.171μm2 SRAM Cell Size in a 291Mb Array”, IEEE International Electron Devices Meeting 2009

[37] H. Reisinger et al., “Analysis of NBTI degradation- and recovery-behavior based on ultra fast VT-measurement,” IEEE International Reliability Physics Symposium, 2006

[38] Bernstein et al., “SOI Circuit Design Concepts”, Springer 2007

[39] E. Saneyoshi, K. Nose, M. Mizuno, “A Precise-tracking NBTI-degradation Monitor Independent of NBTI Recovery Effect”, IEEE International Solid State Circuits Conference, 2010

[40] E. Karl, P. Singh, D. Blaauw, D. Sylvester, “Compact In-situ Sensors for Monitoring Negative-bias-temperature-instability effect and Oxide Degradation”, IEEE International Solid State Circuits Conference, 2008

[41] G. Langguth, C. Russ, W. Soldner, B. Stein, and H. Gossner, “ESD Challenges in Advanced CMOS Systems on Chip”, IEEE International Conference on IC Design and Technology, 2010

[42] S. Pae et al. ," Reliability Characterization of 32nm High-k and Metal-gate Logic Transistor Technology", IEEE International Reliability Physics Symposium, 2010

[43] N. Dumin, K. Liu, and S.-H. Yang, “Gate Oxide Reliability of Drain-Side Stresses Compared to Gate Stresses”, IEEE International Reliability Physics Symposium, 2002

[44] K. Hofmann, S. Holzhauser, and C. Kuo, “A Comprehensive Analysis of NFET Degradation Due to Off-State Stress”, IEEE International Integrated Reliability Workshop, 2004

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

[46] S. Chang, C. Chen, C. Wang, and K. Wu, “A New Off-State Drain-Bias TDDB Lifetime Model for DENMOS Device”, IEEE International Reliability Physics Symposium, 2009

[47] 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 International Reliability Physics Symposium, 2008

[48] 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, 2008

[49] B. Kaczer, R. Degraeve, A. Keersgieter, K. Mieroop, V. Simons, and G. Groeseneken, “Consistent Model for Short-Channel nMOSFET After Hard Gate Oxide Breakdown”, IEEE Transactions on Electron Devices, Mar 2002

[50] S. Machlup, “ Noise in semiconductors: Spectrum of a Two-parameter Random Signal”, Journal of Applied Physics, Vol 25, No. 1, 1954

[51] K. K. Hung, P.K. Ko, C. Hu, Y. C. Cheng, “Random Telegraph Noise of Deep Submicrometer MOSFETs”, IEEE Electron Device Letters, Vol. 11, No. 2, 1990

[52] C. M. Compagnoni, R. Gusmerodli, A. S. Spinelli, A. l. Lacaita, M. Bonanomi, A. Visconti, “Statistical Model for Random Telegraph Noise in Flash Memories”, IEEE Transactions on Electron Devices, vol. 55, No. 1, 2008

[53] W. Feng et al., "Fundamental origin of excellent low-noise property in 3D Si-MOSFETs ~ Impact of charge-centroid in the channel due to quantum effect on 1/f noise”, IEEE International Electron Devices Meeting, 2011.

[54] S. Realov and K. Shepard, “Random Telegraph Noise in 45-nm CMOS: Analysis Using an On-Chip Test and Measurement System”, IEEE International Electron Devices Meeting, 2010.

[55] S. O. Toh, T-J. K. Liu, B. Nikolic, “Impact of random telegraph signaling noise on SRAM stability”, IEEE Symposium on VLSI Technology, 2011.

[56] S. O. Toh, Y. Tsukamoto, Z. Guo, L. Jones, T-J. K. Liu, B. Nikolic, “Impact of Random Telegraph Signals on Vmin in 45nm SRAM”, IEEE International Electron Devices Meeting, 2009.

[57] S. Fujimoto, I. Mahfzul, T. Matsumoto, H. Onodera, “Inhomogeneous Ring Oscillator for WID Variability and RTN Characterization”, IEEE IEEE International Conference on Microelectronic Test Structures (ICMTS), 2012.

[58] T. Kim, R. Persaud, and C.H. Kim, "Silicon Odometer: An On-Chip Reliability Monitor for Measuring Frequency Degradation of Digital Circuits", IEEE Journal of Solid-State Circuits, Apr. 2008

[59] K. Takahashi, M. Sekiguchi, “Through Silicon Via and 3-D Wafer/Chip Stacking Technology”, IEEE Symposium on VLSI Circuits, pp. 89-90, June 2006.

[60] J. A. Burns, B. F. Aull, C. K. Chen, C. Chang-Lee, C. L. Keast, J. M. Knecht, V. Suntharalingam, K. Warner, P. W. Wyatt, D-R.W. Yost, "A Wafer-scale 3-D Circuit Integration Technology," IEEE Transactions On Electron Devices, pp.2507-2516, Oct. 2006.

[61] H. Hao, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal, W. R. Davis, "Exploring Compromises Among Timing, Power and Temperature in Three-dimensional Integrated Circuits," IEEE Design Automation Conference, pp.997-1002, June 2006

[62] B. Goplen, S. Sapatnekar, “Placement of 3D ICs with Thermal and Interlayer Via Considerations”, IEEE Design Automation Conference, pp. 626-631, June 2007.

[63] J. Cong, Y. Zhang, “Thermal via Planning for 3-D ICs”, IEEE International Conference on Computer Aided Design, pp. 745-752, Nov. 2005.

[64] C. Abebei, H. Mogal, K. Bazargan, “Three-dimensional Place and Route for FPGAs”, IEEE Asian Pacific Design Automation Conference, pp. 713-718, Jan. 2005.

[65] S. Das, A. Chandrakasan, R. Reif, “Design Tools for 3-D Integrated Circuits”, IEEE Asian Pacific Design Automation Conference, pp. 53-56, Jan. 2003.

[66] A. Fazzi, R. Canegallo, L. Ciccarelli, L. Magagni, F. Natali, E. Jung, P.L. Rolandi, R. Guerrieri, "3D Capacitive Interconnections with Mono- and Bi-Directional Capabilities," IEEE International Solid State Circuits Conference, pp. 356-608, Feb. 2007.

[67] Q. Gu, Z. Xu, J. Ko, “Two 10Gb/s/pin Low-Power Interconnect Methods for 3D ICs”, IEEE International Solid-State Circuits Conference, pp. 448-449, Feb. 2007.

[68] K. Kanda, D. D. Antono, K. Ishida, H. Kawaguchi, T. Kuroda, T. Sakurai, "1.27Gb/s/pin 3mW/pin Wireless Superconnect (WSC) Interface Scheme," IEEE International Solid-State Circuits Conference, pp. 186-487, Feb. 2003.

[69] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory,” IEEE International IEEE Symposium on Computer Architecture, pp. 130-141, June 2006.

[70] H. Gang, M. Bakir, A. Naeemi, H. Chen, J.D. Meindl, "Power Delivery for 3D Chip Stacks: Physical Modeling and Design Implication,” IEEE Electrical Performance of Electronic Packaging, pp. 205-208, Oct. 2007.

[71] S. Rajapandian, K. Shepard, P. Hazucha, and T. Karnik, “High-Tension Power Delivery: Operating 0.18μm CMOS Digital Logic at 5.4V”, IEEE International Solid-State Circuits Conference, pp. 298-299, Feb 2005.

[72] J. Gu, C. Kim, “Multi-Story Power Delivery for Supply Noise Reduction and Low Voltage Operation”, IEEE International IEEE Symposium on Low Power Electronics and Design, pp. 192-197, Aug. 2005.

[73] J. Sun, J. Lu, D. Giuliano, T.P. Chow, R. J. Gutmann, "3D Power Delivery for Microprocessors and High-Performance ASICs," IEEE Applied Power Electronics Conference, pp.127-133, Feb.- March 2007.

[74] N. Nanju, T. Budell, C. Chiu, E. Tremble, I. Wemple, "The Effects of On-chip and Package Decoupling Capacitors and an Efficient ASIC Decoupling Methodology," IEEE Electronic Components and Technology Conference, pp. 556-567, June 2004.

[75] A. Waizman, "CPU Power Supply Impedance Profile Measurement Using FFT and Clock Gating", IEEE Electrical Performance of Electronic Packaging, pp. 29-32, Oct. 2003.

[76] E. Hailu, D. Boerstler, K. Miki, Q. Jieming, M. Wang, M. Riley, "A Circuit for Reducing Large Transient Current Effects on Processor Power Grids," IEEE International Solid State Circuits Conference, pp. 2238-2245, Feb. 2006.

[77] C. Keast, B. Aull, J. Burns, N. Checka, C. Chen, C. Chen, M. Fritze, J. Kedzierski, J. Knecht, B. Tyrrell, K. Warner, B. Wheeler, D. Shaver, V. Suntharlingam, D. Yost, "3D Integration for Integrated Circuits and Advanced Focal Planes", Fermilab Colloqium, MIT Lincoln Laboratory, Feb. 2007.

[78] J. Xu, P. Hazucha, M. Huang, P. Aseron, F. Paillet, G. Schrom, J. Tschanz, Z. Cangsang, V. De, T. Karnik, G. Taylor, "On-Die Supply-Resonance Suppression Using Band-Limited Active Damping," IEEE International Solid State Circuits Conference, pp.2238-2245, 2007.

[79] J. Gu, H. Eom and C. H. Kim, "A Switched Decoupling Capacitor Circuit for On-Chip Supply Resonance Damping," IEEE Symposium on VLSI Circuits, pp. 126-127, 2007.