

# A Design for Test Technique for Parametric Analysis of SRAM: On-Die Low Yield Analysis

Benjamin M. Mauck, Vishnumohan Ravichandran and Usman Azeez Mughal

Intel Corporation, Hillsboro, OR, USA

## Abstract

Parametric analysis of microprocessor SRAM through special design for test features (DFT) is used extensively by fault isolation and failure analysis engineers to find and characterize defects. Unfortunately, a growing amount of leakage on each new process is distorting these Low Yield Analysis (LYA) Testmode I-V curves, making it increasingly difficult to find and differentiate defects. The goal of this paper is to discuss the simulation and silicon results of a concept On-Die LYA (ODLYA) circuit implemented in a 65 nm CMOS process technology. ODLYA is used to curve-trace individual transistors within an SRAM cell and read out results in an automated fashion. Taking measurements on-die eliminates interconnect-dominated IR drop and leakage distortion from several levels of multiplexing. The proposed implementation enables non-destructive high-speed parametric analysis with less dependency on growing cache sizes, number of cores, and scaling process technologies.

## 1. Introduction

The process of cache testing, fault isolation and failure analysis is becoming increasingly difficult due to larger numbers of subtle defects and more complex defect mechanisms. This hardship is exacerbated by the growing percentage of dense and defect-sensitive memory relative to logic in microprocessors, an architectural necessity to raise performance per power. By examining the pass/fail status of a full suite of cache patterns on a defective die, one can infer a possible defect based on which fault model was targeted with the particular failing test [1]. More recent cache DFT such as weak-write test mode can successfully distinguish other more subtle defect types not detectable by traditional cache patterns [2]. However, to test and investigate highly complex defect mechanisms, a parametric analysis is needed. Historically this has been done by pico-probing, or non-destructively, by the low yield analysis test mode (LYA) DFT technique. Upon obtaining transistor I-V curves from reference and defective SRAM cells, simulations are performed to correlate the curves to known

defect mechanisms [3]. Because many physical failure analysis techniques are mutually exclusive, only after careful correlation is physical failure analysis completed with a high success rate.

The simplicity of the measurement circuit, along with the known family of curves, makes LYA one of the most versatile DFT techniques in cache fault isolation. [4]

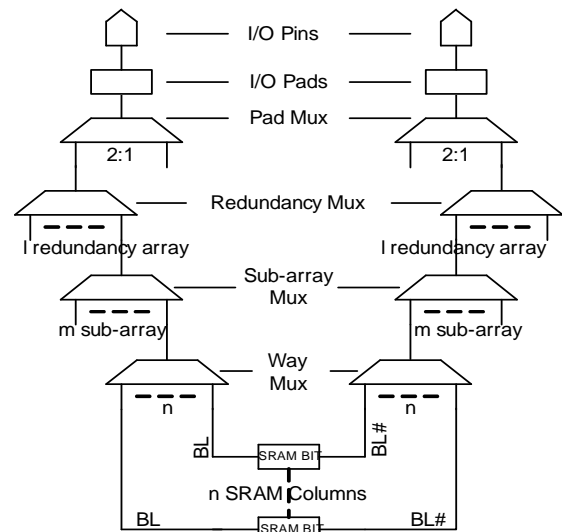


Figure 1: Path from a SRAM column to external pins as it traverses the multiplexers

Figure 1 depicts a high level circuit setup for a typical LYA implementation. Two pins on a microprocessor are connected to the bit-line and bit-line bar (BL/BL#) of a memory cell through several multiplexers (muxes). An external voltage is then applied at both pins and the output current is measured on the same pins. The muxes are used to enable any single SRAM cell to be addressed and curve traced. Depending on the desired transistor to be curve traced, either the BL or BL# is held at Vcc or Vss. The voltage on the other arm is then swept, and the current is measured.

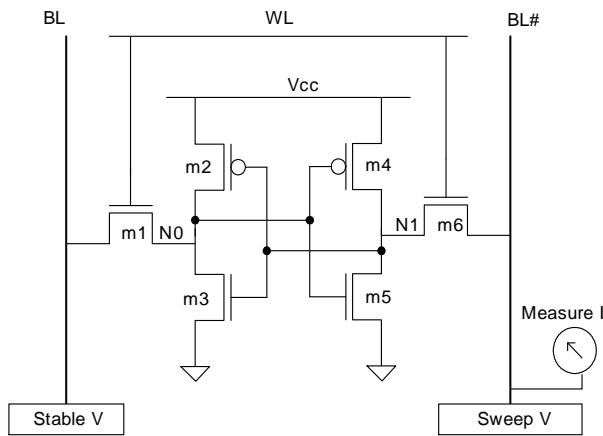


Figure 2: A typical 6T SRAM cell

If the device m5 in Figure 2 is being curve traced, then the BL is set to Vcc. This will initialize node n0 to Vcc and feed that voltage to the input of N-channel transistor m5. The voltage on BL# is now swept from Vss to Vcc, and the corresponding current is measured. This NMOS or PMOS curve trace inherently contains the leakage current embedded in it. To get the true curve for the transistor under consideration, the leakage current curve is subtracted from the total curve. The leakage current is measured using LYA  $n_0$  word-line mode, which turns off the word-line. BL# is then swept, and the current is measured. The shape of the curves obtained through sweeps of each transistor in one SRAM cell gives information on the type and strength of the defect and the strength of the cell. This information is critical for successful failure analysis.

This paper explains the limitations of the LYA technique and proposes a new parametric analysis technique to detect faults that is faster and less dependent on process scaling. ODLYA test mode is a novel design for testability (DFT) technique that is aimed at complementing the low yield analysis (LYA) DFT technique in future processes and replacing it if and when the need arises.

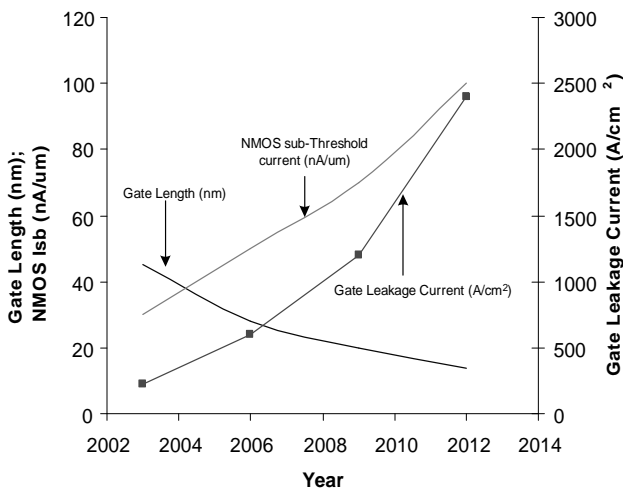


Figure 3: Leakage trend over process generations [5]

### 1.1. Limitations of LYA

Unfortunately, the leakage current of each new CMOS process generation has been drastically increasing. Gate oxide leakage has increased over each process generation and is expected to increase further in successive generations. In a 90nm CMOS technology, the gate oxide thickness is a mere 1.2nm, or just five atomic layers thick. Further reduction in oxide thickness with process generations will cause the leakage to increase exponentially. Similarly, the sub-threshold leakage and the increasing interconnect resistance (IR drop) is driving the LYA accuracy lower with every process generation. These challenges are making LYA curve traces less practical. Figure 3 shows an increasing leakage current trend for high performance devices over several years on future process generations. The data in the graph is obtained from the International Technology Roadmap for Semiconductors 2003 [5].

Overall, the leakage current increases 60% to 70% approximately every three years, and the current density increases between 35% and 50%. With this rate of increase, even with highly precise manufacturing techniques, the devices in the BL and BL# of the six-transistor SRAM cell (Figure 2) will not exactly match each other. This mismatching results in different word-line off-leakage current between BL and BL#. A similar case is also observed between BL and BL# of different columns. This phenomenon results in different curves for m3 and m5 and similarly for m2 and m4. Such distortion causes variations in the tail end of the I-V curve.

Figure 1 is an example of the path traversed from the SRAM cell to the I/O pads. It is clear that even with a full VCC applied at the pins; there is a significant voltage drop before reaching the SRAM single bit, due to the multiple levels of muxes. This reduction in the effective voltage range over which the I-V curve is obtained reduces the controllability and observability of LYA as a DFT technique. This will degrade further as the size of the cache increases, as more cores are introduced, and as the levels of muxing increase in future generations.

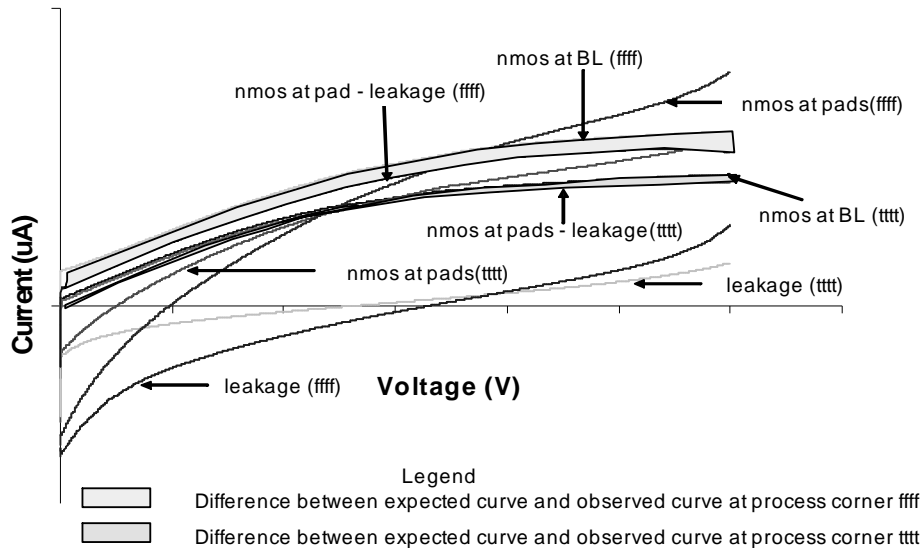


Figure 4: Leakage effects seen on NMOS I-V across process skew (tttt and ffff, 1.2V 110C)

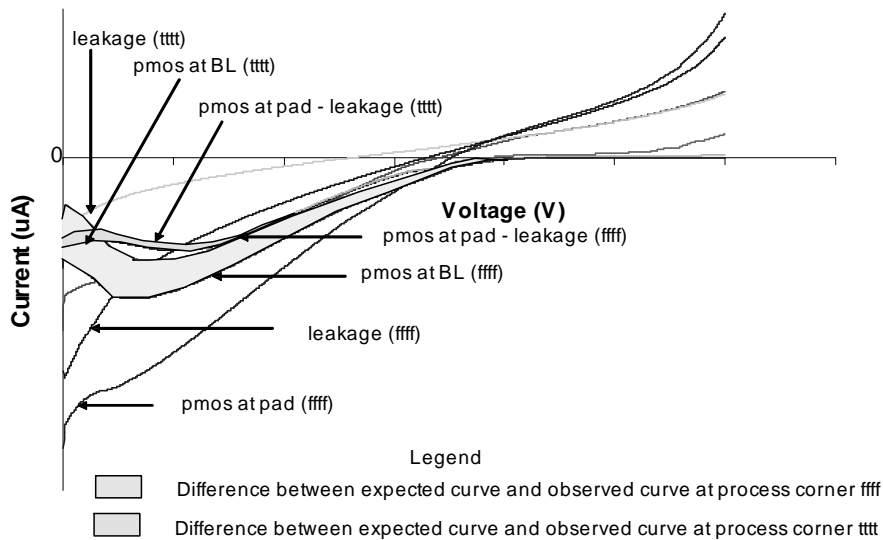


Figure 5: Leakage effects seen on PMOS I-V across process skew (tttt and ffff 1.2V 110C)

Figures 4 and 5 show the curve trace of a NMOS and PMOS simulated for typical and fast process corners at a temperature of 110C. The curves show I-V values at different levels of the circuit. Curve “NMOS at BL” is measured at the drain of the pass transistor of the SRAM cell. This measurement is not possible to collect with a parametric tester unless the cell is pico-probed. Curve “nmos at pad” is measured at the I/O pin and is similar to that seen with real measurements. The leakage curve is measured at the I/O pad with all word lines turned off. The curve “nmos at pad – leakage” is the final curve which is obtained after a point-by-point subtraction of the two curves. The voltage on the BL is swept from  $V_{ss}$  to  $V_{cc}$ , while the BL# is maintained at fixed voltage. Figures 4 and 5 show the current measured at the pads is much higher than the current measured at the BL. This is due to the leakage current caused by the muxes. This leakage current adds to the voltage drop across the muxes shown in Figure 1. Furthermore, the difference between the final curve (curve at pad – leakage) and

the curve at BL increases significantly between the typical and fast process corners. This is expected to increase further in the future generations, drowning defects that produce an I-V curve which falls between the two curves.

Since some defects can only be differentiated by comparing the slopes of the I-V curves, the observability of these defects is lost when the I-V tail is distorted. The tail-current distortion is also misleading when determining the strength of a cell. As the leakage current increases with every process generation, the distortion is also expected to grow, thereby drowning the useful information. Furthermore, as cache size continues to grow on microprocessors, additional levels of muxes are required to address any memory cell, adding more leakage. The shift to multi-core microprocessors accelerates growth in LYA leakage, as each additional core must have the bitlines muxed all the way to the pins. Also, the increased interconnect routing causes additional IR drop, thereby reducing the effective voltage available at the cell.

Collectively, the adverse effects on the I-V curve quality degrade the usefulness of the LYA curves, and make fault isolation and failure analysis increasingly difficult. To compound the issue, increased leakage flowing through longer interconnect causes an even higher IR drop. This current component, when added to the cell current, further increases the IR drop, resulting in more distortion. Distortion can be caused by off-muxes, poor signal-to-noise ratio, and a variety of process issues, making it difficult to de-convolve, quantify, and eliminate the distortion. Although LYA test-mode is still used extensively, it is possible that the cell characteristics may eventually be irreparably convoluted by leakage distortion.

## 2. On-Die Low Yield Analysis

ODLYA test mode is a novel DFT technique designed to move the parametric testing of SRAM off the tester and into the chip. The measurement technique forces current and measures voltage, contrasted to forcing voltage and measuring current in traditional LYA. This technique is utilized to keep the measurement circuit simple and small, a necessity for implementation in a CPU. Moving the circuit closer to the memory bank reduces the IR drop to a large extent. This also improves the accuracy of the true LYA curve by lowering the amplitude and impact of the leakage current. Besides improving measurement accuracy, ODLYA is also desirable to

reduce test time by allowing many cells to be measured in parallel, and at-speed. Since ODLYA does not require a PMU on the tester, tester cost might also be reduced. Finally, by accurate and timely measurement of individual transistor characteristics, a test metric can be introduced to detect and screen the newest, most subtle defects.

Figure 6 shows the top-level circuit implementation of ODLYA. The main components of this analysis scheme are: PMOS and NMOS current mirrors (CMs), a unity gain Operational Amplifier (Op-Amp), an analog-to-digital converter (ADC), a resistive ladder, a variable resistance generator, and a scan chain.

The CMs are implemented as cascode current mirrors to provide higher gain and higher output impedance [6, 7]. The main drawback of this current mirror is the increased voltage headroom over other current mirrors. To address this issue and to keep all the transistors in the optimum operating conditions, ramped up and ramped down voltage supplies are used. These ramp supplies are provided through the two external pins shown in Figure 1 that were previously designated for traditional LYA.

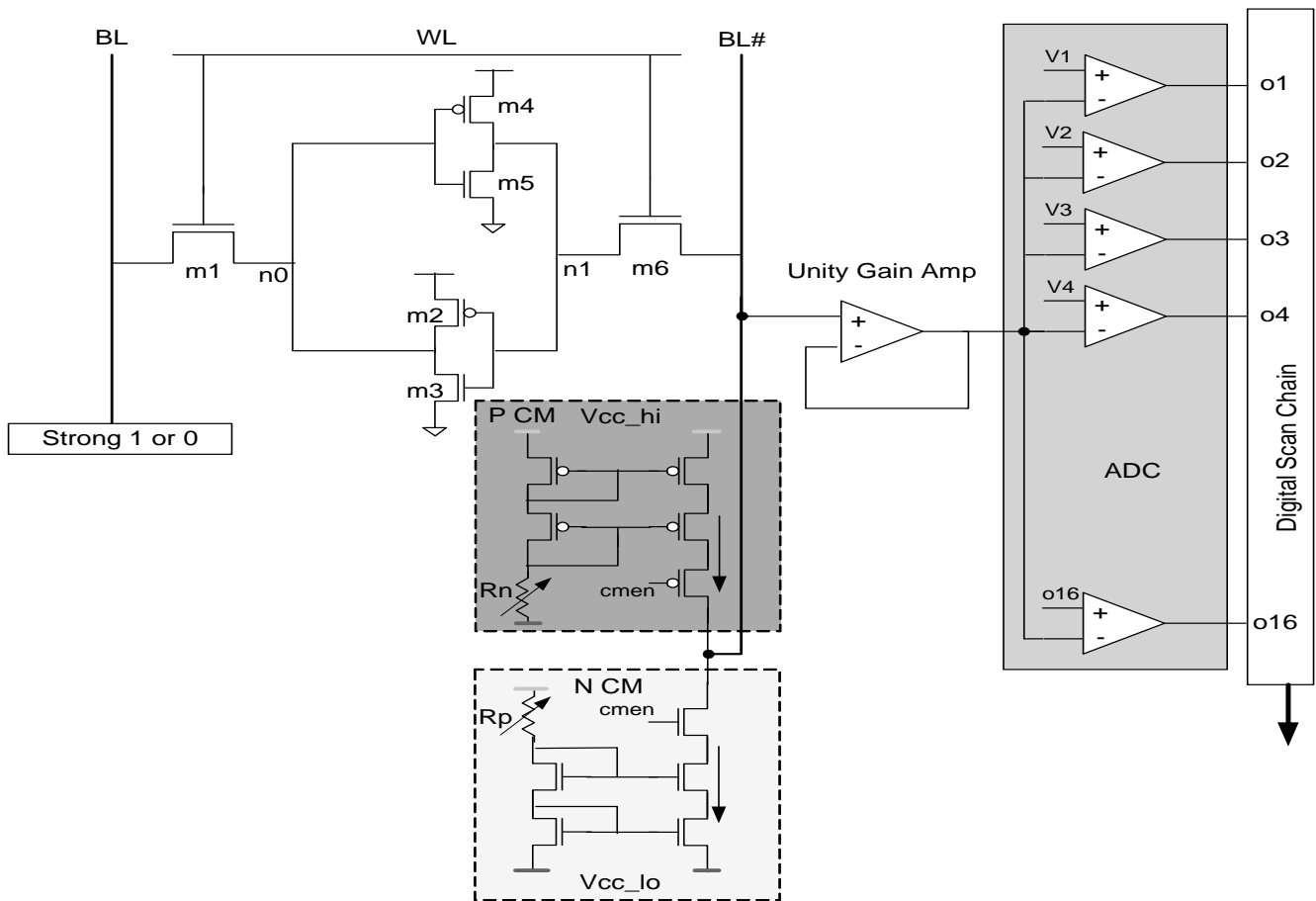
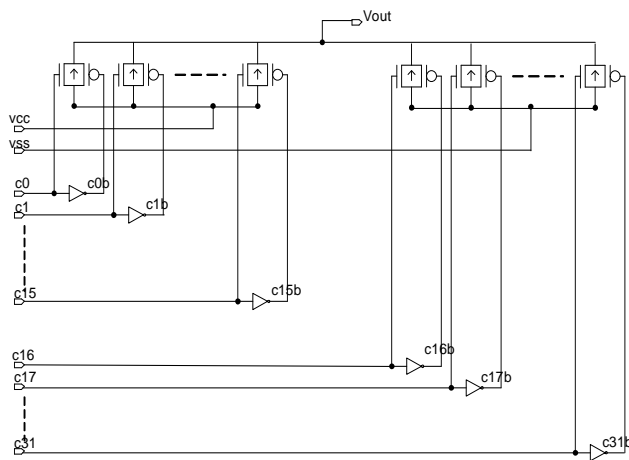


Figure 6: ODLYA setup for a single 6T-SRAM cell



**Figure 7: Variable Resistor for Current Mirror**

The variable resistors used to sweep the current are a tree of pass transistors, shown in Figure 7. The variable resistors  $R_n$  and  $R_p$  shown in Figure 6 are implemented using the tree of pass transistors as shown in Figure 7. The resistance is varied by changing the number of pass transistors that are turned on. A scan chain is used to set the resistance value. A finite state machine can also be implemented to slowly traverse through this tree of pass gates to generate a resistance every  $N$  bus clock cycles, long enough for the rest of the circuit to measure the corresponding voltage.

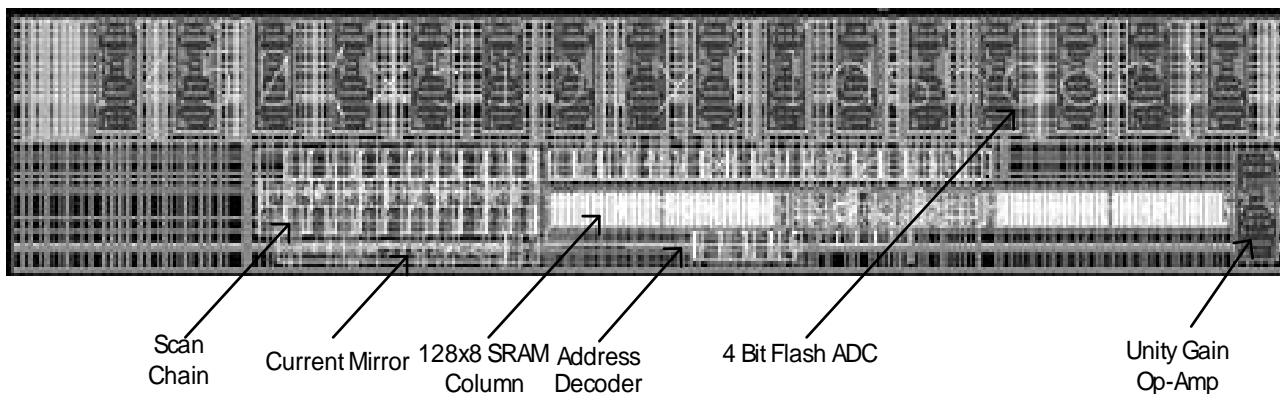
The unity gain Op-Amp shown in Figure 6 is used to generate a copy of the voltage on BL or BL#. The Op-Amp is shown only on BL# for simplicity. However, in implementation, a mux selects the input of the Op-Amp from either BL or BL#. The purpose of the unit-gain Op-Amp is to minimize loading on the measured node. To save power, the Op-Amp is implemented as a two-stage amplifier. The Op-Amp is also operated using ramped

supplies to provide more voltage headroom. In normal SRAM operation, all circuitry associated with ODLYA is disabled.

The output of the unity-gain Op-Amp is fed into a 4-bit ADC. A 4-bit ADC is chosen for finer measurement [8]. There is a trade-off between fine measurement and silicon area. A different type of ADC can be designed if area is the primary limitation. Accuracy can also be adjusted, depending on the desired operation frequency. The reference voltages,  $v_1$ ,  $v_2$ ,  $v_3$  etc., for the ADC's are generated using a resistive ladder. The lower and higher amplifiers in the ADC can be tuned to distinguish fine steps in voltages (output of unity-gain Op-Amp) while the middle part of the ADC is made coarser. This is done to make finer voltage measurements near the turn-on portion of the transistors (tail of the I-V curve).

If the input voltage is less than the reference voltage, the output of that ADC amplifier is low. Once the measured voltage becomes greater than the reference voltage, the output voltage of the ADC amplifier switches to high. The output of the ADC is fed into a scan chain. The switching ADC bit determines the measured voltage.

For example, if  $m_5$  (N-channel) is the device under test, the P-channel CM is enabled by setting the  $cm_{en}$  signal in Figure 6 to low. This cuts off the N-channel CM from the cell under test. The BL is then set to  $V_{cc}$ , and the current on the BL# is swept by traversing through the tree of pass gates. The corresponding voltage on BL# is then copied, using the unity-gain Op-Amp, and then measured with the ADC.



**Figure 8: Layout of ODLYA circuit**

Figure 8 shows the layout of the implemented ODLYA. Most of the area is consumed by the 4-bit flash ADC. Choosing a smaller ADC with optimized layout can reduce the overall area and power consumption of the circuit. Since the measurement circuits are on-die, the implementation will always require more silicon area than traditional LYA, but the placement of the circuit is somewhat flexible. Therefore from a performance standpoint, ODLYA is no more obtrusive than regular LYA.

The ODLYA DFT output forms a new family of R-V curves, similar to I-V curves. The R (resistance) part of the curve corresponds to the resistance set for the CM in use. For most cases, it relates to the current being forced. The type of defect is then determined using the shape of the R-V curves. Since current is forced using a current mirror, it is not possible to determine the exact current flowing without extensive measurement circuits. Thus, resistance is plotted against the

corresponding voltage measurement. Forcing current and measuring voltage is used to reduce the IR drop as much as possible. Forcing voltage and measuring current on-die can be accomplished, but requires a more complex circuit, increased silicon area and reduced accuracy. When a voltage is forced on-die, the Op-Amp driving the voltage will also drive a current. This combines with the memory cell current and makes it difficult to differentiate between the two.

### 3. Results

#### 3.1. ODLYA Test Process

The ODLYA presented in this paper is implemented on a test chip with a 65nm process as a proof of concept. The circuit is designed to curve trace 16 SRAM cells of a memory slice. The 16 addressable SRAM cells have a few good 6T cells and a few deliberately introduced defects. The inserted defects are: n1 to VSS short, n1 to VCC short, and n0 to n1 short. To obtain the curve trace of a single cell, the wordline of the desired cell is turned on by scanning-in the address bits. The control signals are held at a particular value to enable a curve trace of the desired transistor in the selected cell. This varies depending on the transistor (m2, m3, m4 and m5) to be curve traced. The 32-bit variable resistor value is scanned-in to set the current forced in the current mirror. The ADC output is then scanned-out after the output stabilizes. The measurement steps are repeated for each resistance value to complete one full R-V sweep. The process is then repeated on every transistor in the desired cell(s). Since the parallel load of the scan chain is not enabled until all 32 bits of the resistor are scanned-in, the operation of scanning-in the next set of resistor values can be performed in parallel with scanning-out the output.

The circuit also has I/O features to enable independent testing of variable resistor scan chain, current mirror operation, and the output portion of the ODLYA. The variable resistor part of the circuit can also be used for other circuit functions. The variable resistor is a ladder of identical pass transistors mirrored to form a voltage divider. The same resistance value can be produced with multiple combinations of the arms, and process variation between the arms can then be monitored by measuring the output voltage. If multiple ODLYA circuits are implemented in a cache, variable resistor measurements across the die might serve as a process monitor.

#### 3.2. Simulation Results

Most of the defects in SRAM can be modeled using either series or shunt resistors. The shape of the NMOS and PMOS curves depends on the type of the defect. Not all curves are affected by all defects, but each defect can be detected on at least one of M2/M3/M4/M5 sweeps. However, the type of detected defect can be identified only after studying sweeps of all the transistors in a cell.

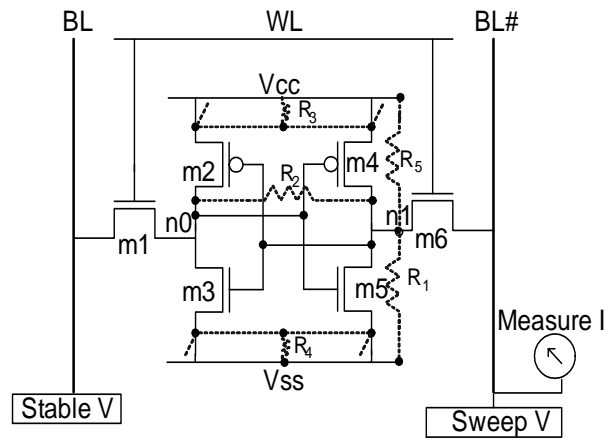


Figure 9: Potential defects in an SRAM cell shown as resistors R1, R2, R3, and R4

Figure 9 shows a SRAM cell with several possible defects. The resistors R1, R2 and R5 are examples of a series type defect while R3 and R4 represent shunt type defects. The shape of the curve is largely affected by the value of the resistor. If the value is very large, then a series defect acts as an open, but if the resistor value is very small, the shunt defect is modeled as a resistive short. The resistors R1, R2, R3, R4 and R5 are labeled in the following graphs as n1-Vss short (R), n0-n1 short (R<sub>2</sub>), P-channel contact, P-channel contact, and n1-Vcc respectively.

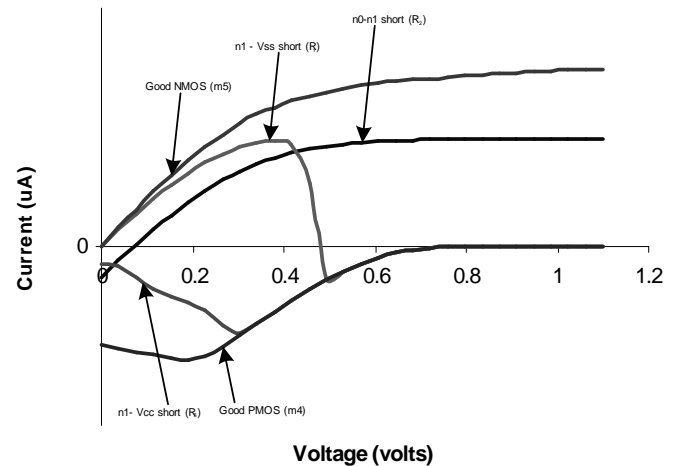
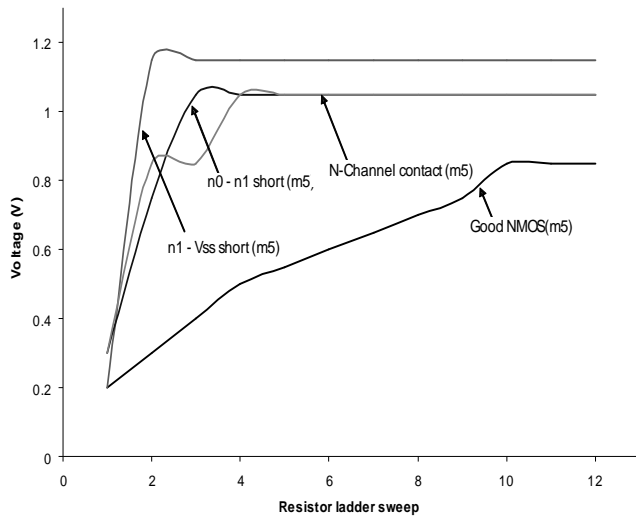
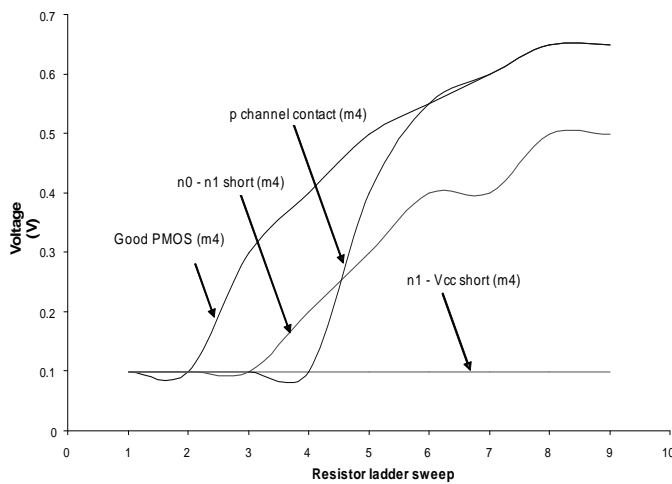


Figure 10: Comparison of ideal PMOS/NMOS I-V

Figure 10 displays the regular LYA curve trace of good transistors m5 and m4. It also compares the curve trace of the same transistors (m5 and m4) in SRAM cells with defects as shown in Figure 9. It can be seen that the shape of the curve is different for different types of defects. For instance, if there is a n1-Vcc short then the bit flips while sweeping the NMOS on node n0 (node opposite to short). This type of defect makes it impossible to write a logical high to the node n1. Another type of defect is a n0-n1 short (R<sub>2</sub>) where the bit does not flip. This symmetrical defect is characterized by an NMOS curve whose magnitude is much lower than that of the ideal NMOS.



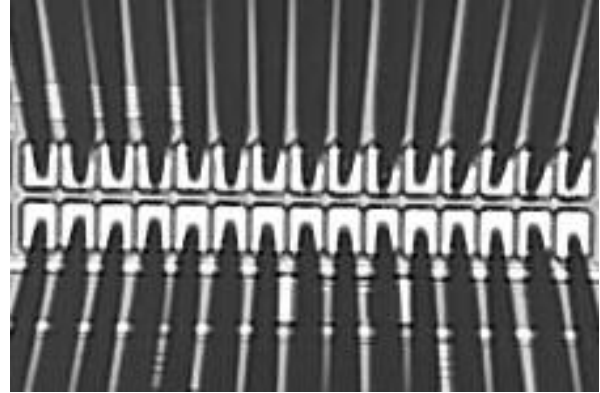
**Figure 11: Simulated NMOS ODLYA R-V curves for normal and defective SRAM cells**



**Figure 12: Simulated PMOS ODLYA R-V curves for normal and defective SRAM cells**

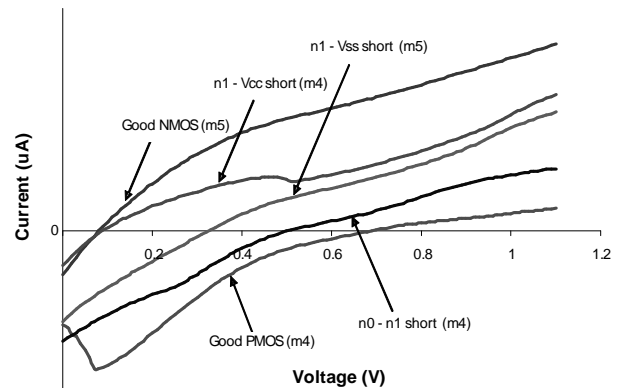
Figures 11 and 12 show the simulation results for the NMOS and PMOS transistors that are curve traced using ODLYA for the process skew tttt at 110°C. The value of the resistance is obtained from a look-up table which lists resistance values corresponding to the scan chain's input. These resistance values correspond to the current being forced by the CM. The output curves for a good NMOS and PMOS are shown in Figure 11 and Figure 12 respectively. The output curves for several defects are shown in the same graph. Simulation results show differences in the curve between each type of defect. If there is a defect in the memory cell, then the slope and/or intercept is different from the good cell. Furthermore, the signature with which the curve hits the rail is different for each defect. Certain defects are identified by a combination of more than one curve trace. In such cases, the characteristics of different transistors of the same memory cell are studied to identify the defect

### 3.3. Silicon Results



**Figure 13: Wafer level test setup**

Silicon results are measured at the wafer level, using an IMS ATS tester connected to a 2x15 probecard. A Keithley parametric analyzer is used to take the analog measurements, such as regular LYA sweeps. Figure 13 shows the 30 probes touching the pads for wafer level testing, with the ODLYA circuit visible in the upper left.



**Figure 14: Comparison of silicon PMOS/NMOS I-V**

Initial silicon functionality of the scan chains, control logic, and SRAM operation was demonstrated using several short patterns. The tester was then halted while addressing each SRAM cell, and the regular LYA sweeps were performed. Figure 14 shows the measured LYA I-V curves for a good NMOS and PMOS, as well as for each of the defects. Results are similar to the simulation I-V curves seen in Figure 10, but not exactly the same, as the resistance values for layout-inserted defects were only estimated for simulation. Leakage is minimal by design, as BL and BL# are connected directly to I/Os. Even with the direct connection to the bitlines, some level of leakage is evident and can be seen in Figure 14 by the shifted Y-intercept. This is not a realistic way of obtaining regular LYA curves, as many levels of muxes would be required to address a single cell in a typical microprocessor. However, these curves are included to better understand differences between I-V and R-V curves for various defects, and to help in interpreting the R-V curves.

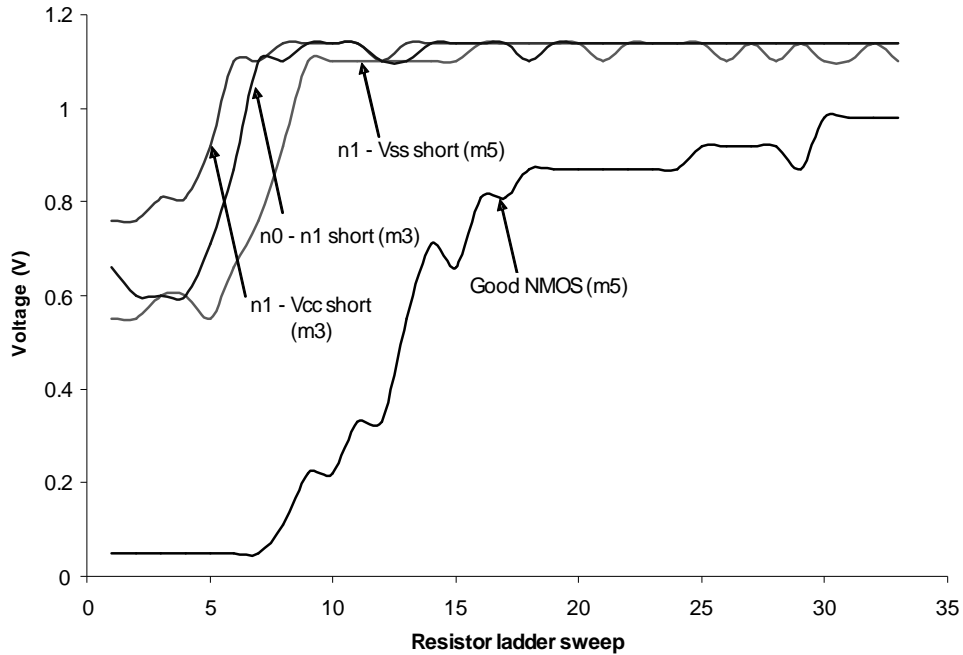


Figure 15: Silicon NMOS ODLYA R-V curves for normal and defective SRAM cells

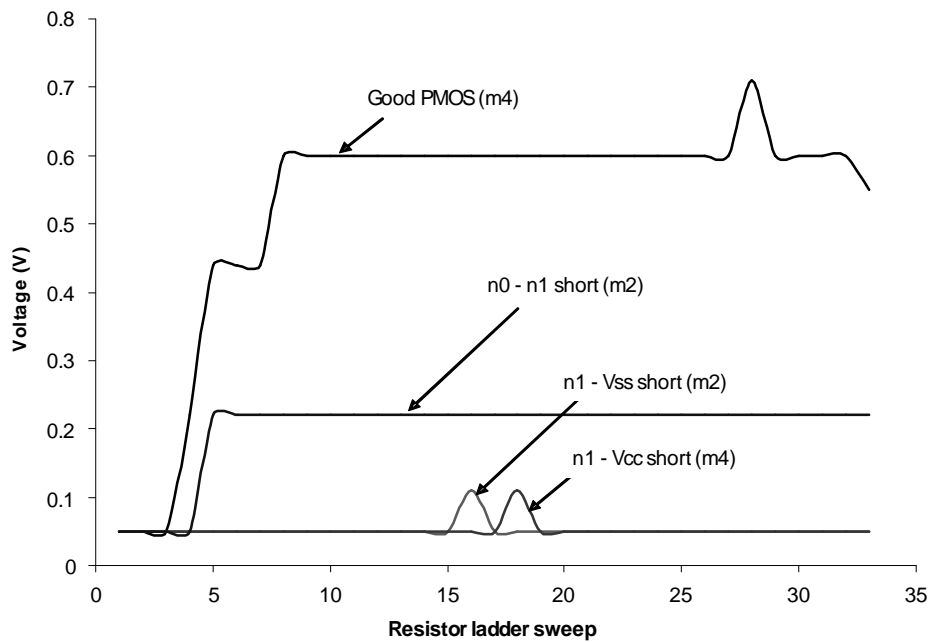


Figure 16: Silicon PMOS ODLYA R-V curves for normal and defective SRAM cells

Two patterns of approximately 18K vectors each were developed to measure the R-V curves for every resistance value, and the results are plotted. Figures 15 and 16 show the measured R-V curves for good NMOS and PMOS devices, as well as for each defect. The silicon results are similar to the simulation results found in Figures 11 and 12. For simplicity, only the curves relevant to detecting each type of defect are presented. This is evident in

Figure 16, where differences between n1-Vss and n1-Vcc curves are only seen in the noise; however the differentiation is made by which transistor shows the bad behavior (m2 for a n1-Vss short vs. m4 for a n1-Vcc short).

From this data, it is evident that ODLYA is able to detect and differentiate all inserted defects on the 65nm test-chip. Furthermore, it is able to do so at a much higher

frequency than regular LYA, and can report results digitally. In this implementation, direct comparisons can be made between the resulting LYA and ODLYA curves. In a microprocessor implementation, a very large amount of leakage would obfuscate the LYA curves due to many levels of muxing. Therefore, the LYA curves shown are highly optimistic in terms of leakage. However, the ODLYA curves in a CPU implementation would be similar to those presented here, allowing much simpler differentiation of defects.

A significant amount of noise existed in the testing environment due to a temporary breadboard connection between the tester and probe-station, as well as several DFT features in the ODLYA circuitry which allowed several other tester channels to mux directly into BL and BL#. The noise is evident on the ODLYA curves only, as all these pins were not driven (high-Z) or disconnected during the LYA measurements. Future designs can eliminate this noise by removal of the extraneous DFT control logic used to test other features on this chip

#### 4. Conclusion

This paper addresses the problems associated with performing non destructive off-chip parametric analysis (LYA) on next generation SRAM cells and presents a design-for-test technique to enable parametric analysis of SRAM on-die. The proposed ODLYA circuit is placed close to the measured component, minimizing leakage issues, enabling at-speed parallel testing, and reducing test time. A proof of concept circuit is implemented on a 65nm process and tested at the wafer level. Silicon results validate simulation in showing ODLYA R-V curves to be effective in detecting and differentiating defects. Further enhancement to the ADC can save silicon area, reduce power and increase resolution. Even higher degrees of accuracy and precision can be achieved by implementing a new variable resistor. Since the resulting SRAM R-V curves are stored digitally on-chip, further automation of defect analysis is possible, such as comparing these curves with a known library of curves and matching the closest defect.

#### 5. Acknowledgments

The authors would like to thank the following individuals for their contributions: Ron Wright, Chee How Lim, Songmin Kim, Greg Taylor, Steven Chen, Tom B Nguyen, Michael Rifani, Rachael Parker, Matt Kirsch, Tom Schwabel, and Shaw Cham Su.

#### 6. References

- [1]. van de Goor, A.J. “*Testing Semiconductor Memories: Theory and Practice*”, ComTex Publishing, Verzijl Gouda, The Netherlands 1999.
- [2] Meixner, Anne. “Weak Write Test Mode: An SRAM Cell Stability Design for Test Technique”, *IEEE International Test Conference*, 1997: 1043.
- [3] Mulder, Randal, Sam Subramanian, Ed Widener, and Tony Chrasteky. “Improved SRAM 6T Cell Failure Analysis using MCSpice Bit Cell Defect Modeling”, *International Symposium for Testing and Failure Analysis*, 2003: 363-370.
- [4] Chinnaswamy, Kumar, Ma, Lin, Pandyan, Gunjan and Chen, Wenliang. “Design of Cache LYA Access Under High Leakage Process,” *Intel Design and Test Technology Conference* 2001.
- [5] [www.sematest.com](http://www.sematest.com), March 2004.
- [6] Rajput, S.S and Jamuar, S.S. “Low voltage, low power, high performance current mirror for portable analogue and mixed mode application,” *IEEE Proc.-Circuits Devices Syst.*, Vol.148, No.5, October 2001: 273- 278.
- [7] Behzad Razavi. “*Design of Analog CMOS Integrated Circuits*,” Tata McGraw-Hill Edition 2002.
- [8] Gendai, Yuji, Yoshihiro Komatsu, Shinya Hirase, and Masato Kawata. “An 8b 500MHz ADC,” *IEEE International Solid-State Circuit Conference*, 1991: 172 – 174.