New Defect Behavior at 130nm and Beyond
Emerging Ideas Contribution, Extended Abstract

Rob Aitken
Artisan Components
141 Caspian Court
Sunnyvale, CA, USA 94089
aitken@artisan.com

Abstract

This paper describes some new defect behavior being observed in production at 130nm and shows how these behaviors can be accounted for by existing fault models and where these models need changes.

1 Background and motivation

With every process generation, fault models need tuning to keep pace with changing technology. In 130nm and below, existing trends are accelerated: wiring is increasingly important in measuring delay, wire aspect ratios continue to grow, metal layers increase. Also, new materials such as copper and low K dielectrics come into play. Designs grow ever larger and speeds are higher. Finally, DFM issues are blurring the distinction between faults and defects. With 90nm, statistical design is becoming mainstream [10].

There are now results from volume manufacturing in 130nm, and some interesting and previously unseen defect behaviors are emerging from these. Library development at 90nm is well underway, and these IP can also be analyzed to understand the types of defects and faults that we will see in the next few years.

The paper is divided into two main sections. Section 2 looks at technology trends, and the resulting failure mechanisms that are being seen. Section 3 expands on these results to show their impact on fault models.

2 Technology trends

Direct scaling has been primarily directed towards improved performance with reduced area. An unwanted byproduct has been an increase in power consumption of all kinds at 130nm and below. This in turn has lead to innovative solutions including copper, use of multi-threshold transistors, and multiple dynamically adjusted voltage domains. Each of these has test implications.

As an example of the scale of the changes, consider the normalized leakage trends shown in Figure 1 for a given foundry technology progressing from 180nm to 90nm. Within each generation, variance of at least three orders of magnitude is seen from slow to fast (1 to 1000 at 180nm, 10k to 10M at 90nm), and seven orders of magnitude variation exists between the worst case at 90nm and the best case at 180nm. The best case at 90nm is worse than the worst case at 180nm. This scale of change means that underlying assumptions need to be re-examined. As will be seen in later sections, leakage is causing problems in logic that have previously been observed only in memory.

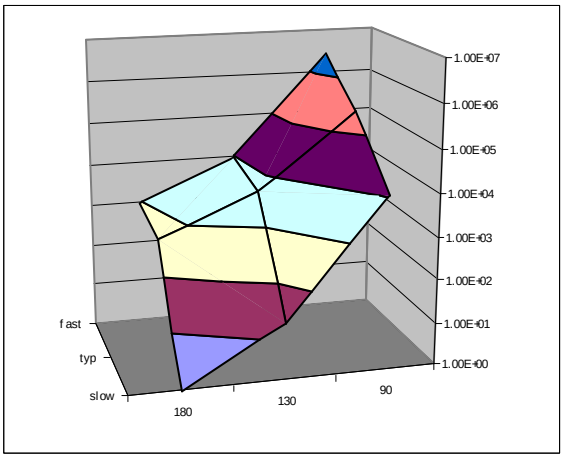

Figure 1 Leakage trends across process generations

2.1 Copper

Aluminum metallization is a subtractive process: An entire layer of metal is deposited, a mask is applied and unwanted metal is etched away. This metal etch is inherently "dirty" and results in many particles being present, some of which lead to shorts. The etching process has led to shorts being far more common faults than opens in CMOS processes for many years.

Copper, on the other hand, uses a dual *damascene*[*] process. A layer of insulator is applied to the wafer. Next troughs for wires are etched (first damascene step). Next additional troughs for vias are etched. A layer of copper is electroplated into the trenches (on top of a small tantalum barrier/seed layer). Finally, excess copper is removed via chemical/mechanical polishing (CMP).

There is no metal etch to provide a slurry of particles. As a result, additional defect mechanisms

[*] The term "damascene" comes from metal inlaid with gold or silver, an intricate craft associated with the city of Damascus.

become important. The figure below summarizes the processing differences. Note that the process is highly simplified.

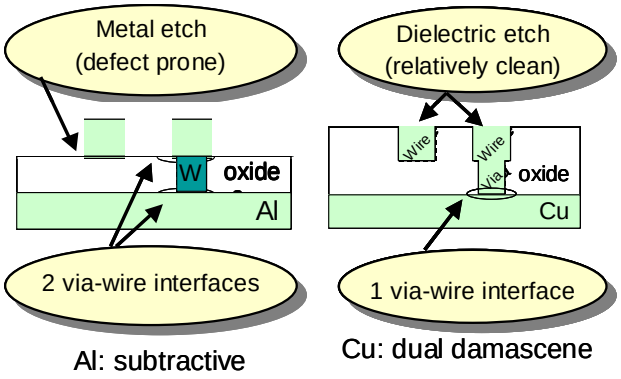

Figure 2 Comparison of Al and Cu processing

In addition to a metal etch, aluminum processing also features two via/wire interfaces (tungsten to aluminum). Copper has only a single via/wire interface, since both vias and wires are deposited together.

The polishing process for copper leads to four major defect mechanisms: *under polishing*, which leaves copper or slurry remaining where it is not wanted, which can lead to shorts, *dishing*, where over polishing has occurred, and *erosion*, where insulator has been polished as well as copper. In each case, the defect leaves a surface that is poorly planarized and susceptible to defects at higher levels (e.g. wide metal dishing on one layer can cause shorts in the layer above as wires fold in on each other). The fourth mechanism, scratches, result in combinations of shorts and opens, spread across relatively large areas.

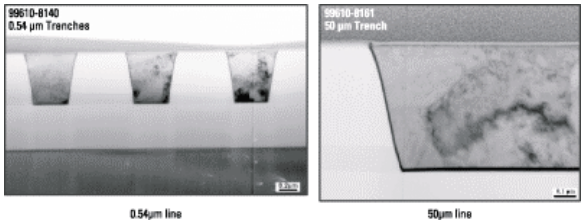

Figure 3 Dishing and Erosion (courtesy Applied Materials)

In addition to CMP-induced defects, voids (also known as partial opens) are another major defect mechanism. These defects often have an easily recognizable signature: delay faults at low temperatures. This happens because of thermal expansion of the copper into the void. As temperature increases, the void shrinks and the two metals make better contact, as shown in Figure 4.

The copper void defect mechanism shows that while high temperatures are the worst operating condition for defect-free circuits, defective circuits may have worst-case behavior at room temperature, or even lower. If possible, delay tests should be applied at two temperatures to account for void defects and process variation at tolerance limits.

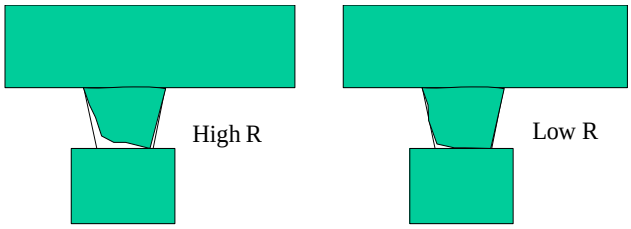

Figure 4 Temperature dependence for copper voids: Low T at left, high T at right

ATPG for void-related defects should emphasize via crossings, especially those where only a single via has been used (redundant vias are the standard design-for-manufacturability – DFM – approach to via voids).

2.2 Optical defects

At 130nm and below, many features of a layout are now below the wavelength of light used during photolithography (typically 248nm or 193nm). As a result, optical proximity correction (OPC) is applied to masks to ensure accurate printing. In many ways, OPC is similar to the use of serif fonts printing, which developed to ensure adequate ink coverage (e.g. "**X**" versus "**X**"). OPC can add additional features (e.g. hammerheads or outriggers) or shrink features.

OPC leads to two major classes of defect: mask errors and DRC escapes. In a mask error, OPC fixes result in a short when features interfere with one another. In general, mask defects are catastrophic and result in 100% yield loss. Reticle inspection helps reduce the number of these, but inspection programs have been confused by OPC features that look like defects but are legitimately needed for the design[9].

In a DRC escape, OPC is meant to fix a particular feature, but can't, due to neighboring circuitry. Consider the layout of figure SDS: hammerheads should be added to each gate, but nearby circuitry prevents this from being possible on the leftmost gate. As a result, the gate is foreshortened. In some cases, catastrophic failure will result, but often only weaker than usual gates will be formed. The resulting channel is shorter in on portion and narrower as well. This defect will typically behave as a delay fault, and can often be detected by transition fault tests.

2.3 Design-related defects

While process technology is the main driving force between new defect types, the impact of design methodology should not be ignored. At the most

obvious level, decisions made by routing tools determine which wires are likely to short and which aren't. Even adjacent wire spacing can be controlled. This section explores two more subtle mechanisms, both relating to low power design.

2.3.1 Multi-Vt Libraries

Transistor off current and drive current both depend on threshold voltage, as shown in Figure 5 (based on [1]). A low threshold device has somewhat higher saturation current, and thus higher performance, but at the cost of substantially higher off current. By using both high and low VT transistors in a design, leakage and performance can be traded off more finely than by using exclusively one or the other. For consistency, we will use "HVT" for the high threshold voltage devices and "RVT" for the regular VT devices. A common approach to get the reduced leakage of HVT devices with the performance of RVT devices is to mix them at the gate level – some gates will use each type [2]. Making this work easily in a tool path requires that the HVT and RVT gates be artwork compatible (same cell height, routing pitch, etc.). Ideally, they should be the same artwork with different implants, in order to allow drop-in substitution after placement. Using these techniques can save 70-80% of leakage power, even for high performance designs [3]. Two variants are discussed here, but processes supporting 4 or 5 variants are currently available.

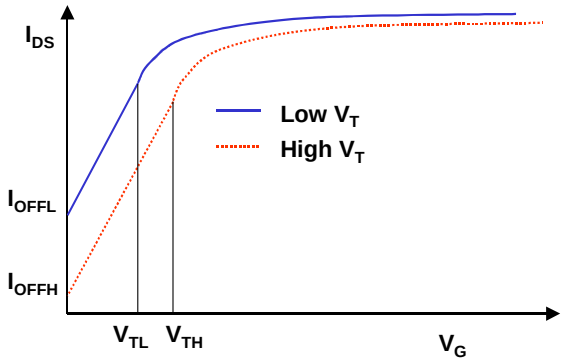

Figure 5 Relationship between off current and drive current [1]

Multi-Vt cells can increase the chances of "Byzantine Generals" (see section 3.1) behavior of bridging faults because of varying thresholds among gates and varying drive currents. HVT devices tend to have higher variability, and their weaker drive currents give them higher critical resistances than RVT devices

2.3.2 Multiple Voltage Domains

Power consumption varies as the square of voltage. Low power design approaches, especially for battery-powered applications, are increasingly using variable power supply voltages. Two basic approaches are possible: Open loop, where voltage levels are predetermined based on characterization and expected power consumption, and closed loop, where the system dynamically adjusts its voltage to the lowest level able to sustain its present level of activity (i.e. embedded CPU voltage will be higher during peak activity times, such as displaying an MPEG video, and will be reduced during lower activity times (voice transmission), and can be reduced even further or even shut off during sleep modes [4].

Keeping voltage domains separate is best accomplished by using level shifting circuitry, but when they interact new defect behavior is possible. For example, bridging fault models usually make the simplifying assumption that a bridging fault is not activated if the involved nodes are at the same logic value. However, consider the circuit in Figure 6. The P transistor in the transmission gate has both its gate and source at logic 1 in the low voltage domain (1_L), so the gate is off in the fault free case. A short (or even capacitive coupling) with the higher voltage domain leads to a voltage on the source above the VDD value of 1_L, denoted by 1_H. This causes the difference between the gate and source voltages of the transistor to be greater than its threshold value ($V_{sg}>V_{th}$), turning the transistor on. The faulty value is now free to propagate as any with other fault. Isolation of level shifters can help reduce this problem, but will not totally eliminate it. These shorts will be more of a concern for diagnosis than for bridging ATPG, which should try to sensitize them the conventional way (opposite values).

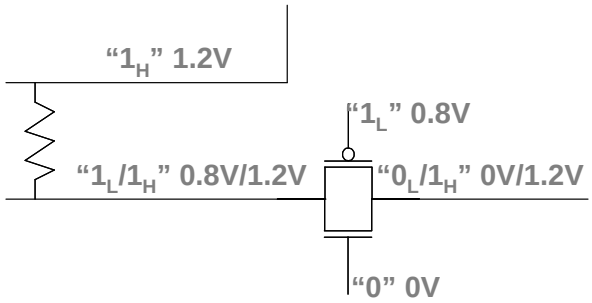

Figure 6 Faulty behavior with multiple VDD values bridged

Defects in the level shifters themselves cannot be avoided, but careful attention needs to be paid in layout in order to ensure that they result in catastrophic (easily measurable) failures, rather than subtle problems that might escape detection during manufacturing test.

3 Fault Models

Experimental results consistently show that most defective chips are detected by most tests [5]. These defects are not of concern when updating fault models. Instead, subtle faults are important. One class of subtle faults is those defects that are rarely activated [6].

3.1 Bridging faults

Shorts, even in copper technology, remain a major source of defects. Some of these are due to classic effects, such as particle contamination, while others result from CMP scratches, optical effects, and more[7]. Shorts involving many lines are typically catastrophic, so while they are vital to yield calculation, they are less important for fault modeling. Many models have been proposed over the years, but we will consider the dominance model (strongest node always wins), the voting model [11] (winner depends on relative drive strengths of opposing values), and analog models (including so-called Byzantine generals behavior, where downstream gates interpret an intermediate shorted voltage differently, some as 0, some as 1).

3.1.1 Voting model

The voting model is most appropriate in cases where drive strengths are near equal, as shown in the table below (vote values, which relate to W/L ratios, but are scaled down for PFETs, are shown for each drive strength). In this technology, an X1 P drive is virtually always dominated by an X2 N drive, but there are some exceptions (e.g. a NAND3X1 with all three PFETs active versus an INVX2), so the voting model is still necessary. This is not the case with the X4 drive strength, which dominates all X1 drives. The PFETs are generally weaker, so voting is still needed even for an X4 PFET fighting an X1 NFET. Methods for applying voting to complex gate structures are given in [11] and [12].

	P drive	**X1**	**X2**	**X4**	**X6**	**X8**	**X12**
N drive	**Vote**	2	4	8	12	16	24
X1	4	V	V	V	V	D1	D1
X2	8	V	V	V	V	V	D1
X4	16	D0	V	V	V	V	V
X6	24	D0	D0	V	V	V	V
X8	32	D0	D0	D0	V	V	V
X12	48	D0	D0	D0	D0	V	V

Table 1 Dominance and Voting behavior for shorts of different drive strengths

Notice that for the most part, NFETs dominate, with a 0 result for a bridge.

As an example, a bridging fault between two small NAND gates (NAND2X1) is considered in a particular 0.13 technology. The behavior depends on the resistance of the defect (this is a slight variant of the well-known critical resistance concept):

Bridge resistance	**Behavior (1.2V, 25C)**
<3000 ohms	Voting model
3000-3700 ohms	Analog/Byzantine
>3700 ohms	Delay fault

Table 2 Example bridge behavior at 1.2V, 25C

The Byzantine behavior, while real (as shown in Figure 7, where one downstream gate interprets an intermediate voltage as high and another interprets it as low), is highly dependent on defect resistance and supply voltage. Running the same test at multiple voltages can eliminate the problem, as shown below:

Bridge resistance	**Behavior (1.3V, 25C)**
<2500 ohms	Voting model
2500-3000 ohms	Analog/Byzantine
>3000 ohms	Delay fault

Table 3 Example bridge behavior at 1.32V, 25C

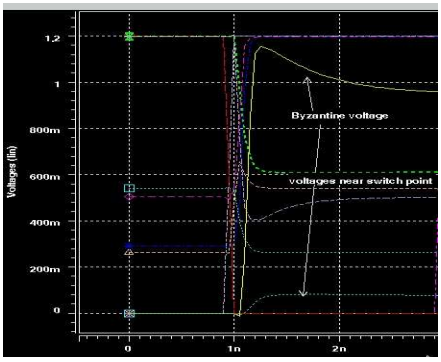

Figure 7 Byzantine bridging behavior at 130nm

Does this mean that the Byzantine Generals have been demoted and are no longer of interest for determining bridging fault behavior? Yes, and no. Multiple voltage tests will eliminate delay-independent Byzantine behavior. However, in addition to gate switching voltages, Byzantine behavior can be caused by gate switching times. Long settling times for bridging voltages lead to large slews on gate inputs, which in turn result in different switching times on gate outputs. If path slack is low enough that values are latched before switching is complete, Byzantine behavior can result (see Figure 8). In theory, increasing

clock speed could eliminate this problem, but in practice running such tests is difficult. As a result, diagnosis of short-induced delay faults should consider Byzantine behavior a strong possibility.

Also, as was seen with the copper voids, resistance is not a fixed property for defects. Voltage changes result in localized temperature changes around a defect (in general, higher voltages lead to higher currents, and these lead to higher local temperatures). The results are not always predictable. Bitline to ground shorts have been observed to cause both operating voltage floors and operating voltage ceilings, in some cases on the same RAM instance (but different die).

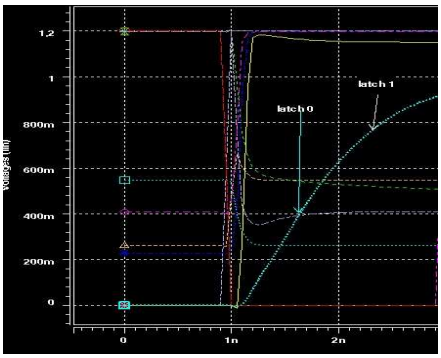

Figure 8 Byzantine delay behavior

3.2 Leakage faults

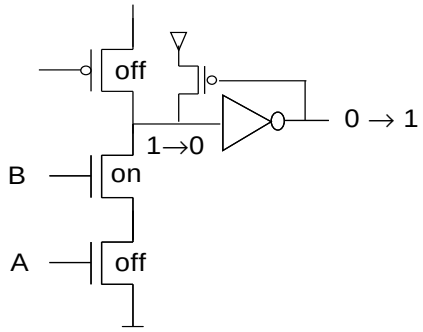

Figure 9 Leakage induced faulty latch behavior

In 130nm technology, leakage is a significant problem, especially in the fast process corner. Leakage is still dominated by subthreshold leakage, although gate leakage is a significant factor at lower temperatures. Leakage is not directly controlled by many foundries, but instead is a byproduct of other parameters (e.g. channel length). Because of this, it is possible for leakage on single transistors to be well beyond specified process limits. As an example, a latching circuit that uses a weak pullup transistor as a holding device has been observed to fail in silicon due to leakage on its input, as shown in Figure 9. The transistors connected to A and B are comparatively large, and off-state leakage on A was more powerful than the pullup drive of the PFET. This eventually pulled the input of the inverter to zero, which changed its output to a 1.

3.3 Open faults

As mentioned earlier, partial opens due to copper voiding are a common defect mechanism at 130nm. The two Shmoo plots shown in Figure 10 are characteristic of this type of failure. Room temperature delay testing is clearly needed.

Other open failures at 130nm behave similarly to those in previous generations, and similar detection methods apply, with the exception of CMP scratches, which produce complex patterns of shorts and opens, although these are usually readily detectable in logic and do not require specific ATPG.

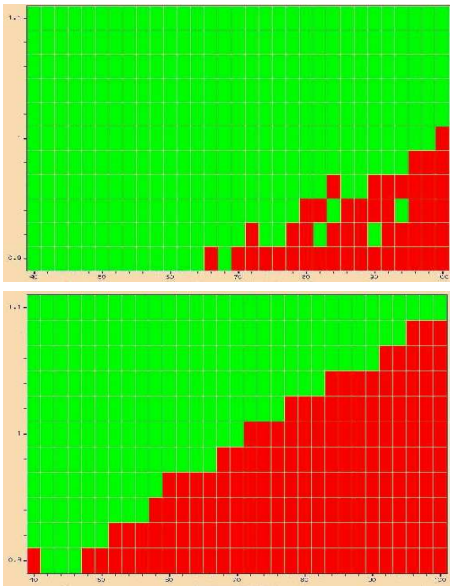

Figure 10 Voltage versus timing shmoos at 90C (top) and 25C (bottom) showing copper void behavior

3.4 Delay faults

It has long been known that logic synthesis will result in additional near critical paths and hence additional delay faults [8]. As designs shrink, two additional issues come to the fore: capacitive defects and signal integrity effects. Let us look in more detail at the former.

3.4.1 Capacitive defects

Originally, bridging faults were considered to have zero resistance. Later resistance models were added. Now, defect capacitance must also be considered. Intuitively, the increasing importance of wire capacitance in delay calculations is due to the increasing relative capacitance of wiring versus gates as a result of CMOS scaling. In order to measure timing

behavior accurately, parasitic capacitances between wires must now be extracted for almost all designs. Increasing frequency and decreasing slack means that the capacitive properties of defects can significantly affect their behavior (RC time constants of defects are well within the range of standard parasitics, and strongly influence timing).

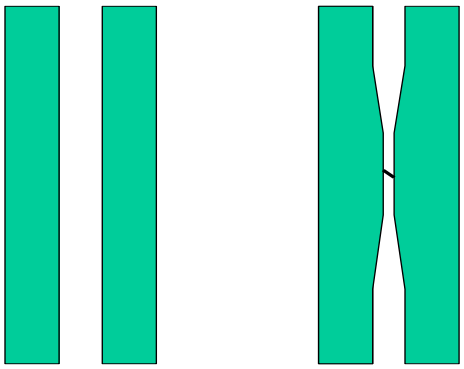

Figure 11 Proposed mechanism for capacitive faults (reduced separation gives higher C)

The leakage failure discussed in section 3.2 is an example of a delay failure that operates correctly at full speed but fails at lower speeds (similar in behavior, but not mechanism, to the well-known retention faults in SRAM). Another failure with this behavior is a high resistance, high capacitance bridge. The physical motivation for these is shown in Figure 11, and a SPICE simulation is shown in Figure 12. This type of behavior, which has been observed in silicon, although not in silicon that is within process specifications (at least to the author's knowledge), shows that low speed tests are needed, in addition to full speed tests, to ensure high quality.

4 Conclusions and Future Directions

Test is best accomplished by combining multiple approaches. Effective testing should include test at multiple supply voltages and multiple frequencies in order to identify new defect mechanisms such as copper voids, and leakage defects, while retaining coverage of classical behaviors such as shorts. Design issues such as multiple threshold transistors and multiple operating voltage domains need to be included in test planning, and this will become increasingly important in 90nm and below, with new techniques such as back biasing becoming important. Similarly, and with statistical design, the distinction between defective behavior and out of spec behavior will become blurrier, requiring additional interaction between DFM, DFT and fault modeling.

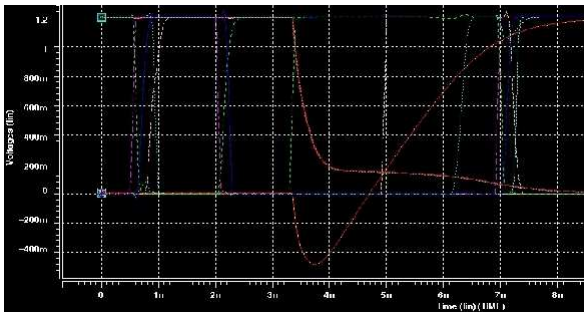

Figure 12 SPICE simulation showing capacitive inverse delay fault

5 References

[1] Jaume Segura, Charles F. Hawkins, *CMOS Electronics: How It Works, How It Fails*, IEEE Press (Wiley), 2004.

[2] L. Wei, Z Chen, M. Johnson, K. Roy and V. De, "Design and Optimization of Low Voltage High Performance Dual Threshold CMOS Circuits", pp. 489–494, *Proc. of Design Automation Conf*., 1998.

[3] R. Aitken et al, "Library Modeling for Effective Leakage Management", to appear, *Synopsys Users Group*, San Jose CA, 2004.

[4] D. Monticelli, "Solving the Real Challenges of Low-Power SOC Design in 90 Nanometers", *DesignCon 2004*, Santa Clara CA, February 2004.

[5] P. Nigh et al, "An Experimental Study Comparing the Relative Effectiveness of Functional, Scan, IDDq and Delay Fault Testing*", Proc. VLSI Test Symp.*, pp. 459-464, Monterey CA, April 1997.

[6] J. Dworak et al, "Enhanced DO-RE-ME Based Defect Level Prediction Using Defect Site Aggregation – MPG-D", *Proc. International Test Conference*, pp. 930-939, 2000.

[7] W. Maly et al, "Deformations of IC Structure in Test and Yield Learning", *Proc. International Test Conference*, pp. 856-865, 2003.

[8] T.W. Williams et al, "The Interdependence Between Delay-Optimization of Synthesized Networks and Testing", *Proc. Design Automation Conf.*, pp. 87-92, 1991.

[9] F M. Schellenberg, "Sub-Wavelength Lithography Using OPC", Semiconductor Fabtech Journal, 9th edition, March 1999.

[10] C. Visweswariah, "Death, Taxes, and Failing Chips", *Proc. Design Automation Conf*., pp. 343-347, 2003.

[11] J.M. Acken, S.D. Millman, "Accurate Modeling and Simulation of Bridging Faults", *Proc. Custom Integrated Circuits Conference*, pp 17.4.1 – 17.4.4, 1991.

[12] P.C. Maxwell and R.C. Aitken, "Biased Voting: A Method for Simulating CMOS Bridging Faults in the Presence of Variable Gate Logic Thresholds", *Proc. Int. Test Conf*., pp. 63 – 72, 1993.