9.9 Early Detection of Oxide Breakdown Through In Situ Degradation Sensing

Prashant Singh, Zhiyoong Foo, Michael Wieckowski, Scott Hanson, Matt Fojtik, David Blaauw, Dennis Sylvester
University of Michigan, Ann Arbor, MI

Oxide breakdown (OBD) is a major concern in high-performance design as it imposes strict limits on the supply voltage, degrading maximum performance and SRAM stability [1]. OBD is an inherently statistical process where some devices fail long before others. Hence, *a priori* lifetime prediction requires statistical estimation that leads to conservative margins in supply voltage that are unnecessary for most parts and for most operating conditions.

Dynamic (runtime) OBD monitoring [2], allows voltage margins to be reduced while the lifetime of infrequent defect-prone chips can be extended through supply voltage reduction and limits on peak temperature. To ensure that corrective measures are taken well before a device fails completely it is critical that OBD monitoring detects the *onset* of breakdown (or "soft breakdown"), when the oxide becomes leaky yet the device continues to function correctly.

Runtime OBD monitoring has been explored using dedicated sensors [3,4] that are distributed throughout the design, exposing them to the circuit temperature and voltage conditions. Since OBD is a statistical process many sensors must be averaged, resulting in a statistical prediction that still requires extensive margins to ensure a high confidence lifetime guarantee. Hence, for effective breakdown prediction and management, the onset of breakdown must be detected *in the actual devices within the circuit* instead of in separate sensors. For such *in situ* OBD detection, path delay monitors were proposed [5]. However, the delay change of a single gate that is close to failure may be obscured by other longer paths. Furthermore, we experimentally show that at the onset of soft breakdown a gate delay may increase by only 2 – 4%, and in some cases can actually improve, complicating this approach. Recently, OBD detection for a power amplifier by measuring oxide resistance using resistors and a pre-driver stage was shown [6].

This paper introduces a new *in situ* OBD detection technique that directly measures the onset of breakdown and provides an early warning of approaching failure. The technique enables dynamic reliability management and the removal of unnecessary voltage margins. The approach leverages the prevalence of MTCMOS-based designs with PMOS header switches, a common technique to reduce standby power [7] with relatively low overhead. As shown in Fig. 9.9.1, a circuit block is partitioned into sub-blocks, each connected to the power supply through a standard high-Vt MTCMOS switch and a weak PMOS device (WP) with a controllable gate voltage, Vbias. Periodically, the design is taken offline and tested for oxide degradation by sweeping the gate voltage of WP from 0 to VDD using an on-chip DAC while the virtual rail voltage is recorded using an on-chip ADC. The resulting Vbias vs. Vrail (V/V) curve is then analyzed to detect the onset of OBD. Initially, both oxide leakage and subthreshold leakage are strongly non-linear with supply voltage, resulting in a characteristic 'hockey stick' curve. However, as the device degrades, the presence of percolation paths in the oxide introduce a resistance that is linear with voltage [8], leading to a distinct flattening of the V/V curve. Based on this key behavior shift, we define a figure of merit called the Degradation Voltage Angle (DVA) that measures the angle of a straight line fitted to the V/V curve over the 90 – 10% Vrail interval. As a gate degrades, the oxide displays more linear resistive behavior and a sharp drop in DVA is observed.

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

The first chip consists of an array of 12 INV, 2 NAND, and 2 NOR gates (referred to as GUTs) as well as 5 parity circuits ranging from 64 – 1024 gates. Three modes of operation are supported: stress mode, oscillation mode (to measure performance degradation), and OBD measurement mode. To accelerate stress for testing purposes a PMOS header switch (TP) is implemented using a thick oxide transistor to transfer the stress voltage (VddST) to the virtual rail (STR_OP = 0). A thick oxide NMOS device TN protects the weak PMOS (WP) from degradation during accelerated testing (note that during regular non-accelerated operation all terminals of WP are at Vdd which prevents its degradation). During oscillation mode, VddST is brought to 1.0V and TP remains on while the circuit is placed in a feedback loop. To isolate the GUT delay, the measurement is repeated with the GUTs bypassed to allow the non-GUT portion of the delay to be subtracted. During OBD measurement TP is super-cutoff while TN is ON and Vbias is swept to record measurements.

Fig. 9.9.2 shows typical measured V/V curves as an inverter is progressively stressed, as well as its delay degradation and the proposed DVA metric. Defining a 15% drop in DVA as the 'onset of failure' point, oxide degradation is detected when the delay increases by 3%, illustrating the difficulty of detecting OBD through delay change alone. Eventually the gate delay abruptly increases to 20X and then fails completely. Impact of stress on delay shows non-monotone behavior, at times resulting in faster gate delays. This is expected and is caused by the suppression of voltage swing under certain failure modes [8]. The corresponding degradation is captured by the DVA.

Since the background leakage of non-failing gates in a block can overwhelm the change in behavior of a failing gate, we tested the DVA metric across temperature and block sizes. For a 64 gate parity circuit, a 100°C temperature change induces a 10X change in leakage while the DVA changes by only 7%, which is less than the DVA threshold of 15% (Fig. 9.9.3). The excellent robustness of DVA metric eliminates the need to calibrate or compensate for temperature, which would significantly increase complexity of the approach. The sensitivity of DVA to parity tree block size indicates that the time to failure detection is compromised beyond block sizes of 512 gates since the change in gate leakage due to OBD is masked by the growing background leakage (Fig. 9.9.4). It must be noted that in a typical design majority of the timing paths are non-critical; hence even though the failure detection would be delayed due to a larger number of gates, the performance degradation at the time of detection may not necessarily be significant.

The second test chip applies the approach to a 16-bit, 8-tap FIR filter consisting of 7K gates (Fig. 9.9.5). The FIR is divided into 360 blocks of ~20 gates placed into 36 rows and 10 columns using an automated design flow. To monitor each of the 360 virtual rails (VRs), a low leakage 360x1 two-stage mux is used. Since the VRs are driven by small leakage currents it is extremely important to isolate the selected VR. To this end, a unity gain buffer mirrors the voltage seen on the selected VR onto the other VRs. The thick oxide devices in the HU result in a large overhead in this case, but are only required for accelerated testing. In a regular design with only thin oxide devices the design area overhead is 17% compared to a design without MTCMOS and 5% compared to a standard MTCMOS design. The overhead can be reduced by increasing the block size, which reduces the number of HUs and VRs. Fig. 9.9.6 shows the block failures over time and their spatial distribution. Out of 360 blocks, only 141 blocks were stressed and monitored. The first detected block failure corresponds to a performance degradation of the FIR by 0.5% which allows for enough time to take measures to prevent functional failure. The get the spatial map of failure times of blocks frequent performance and DVA measurements were made on a die to show the gradual progress in failure. Blocks with Time > 300sec did not fail by the end of the experiment.

Acknowledgements: The authors would like to thank NSF and GSRC for funding support and ST Microelectronics for fabrication support.

References:

[1] J. H. Stathis, "Gate Oxide Reliability for Nano-Scale CMOS," *Int. Conf. on Microelectonics*, pp. 78-83, 2006.

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

[3] E. Karl, P. Singh, D. Blaauw, D. Sylvester, "Compact In-Situ Sensors for Monitoring Negative-Bias-Temperature-Instability Effect and Oxide Degradation," *Proceedings of ISSCC*, 2008, pp.410-411.

[4] J. Keane, S. Venkatraman, P. Butzen, C. H. Kim, "An array-based test circuit for fully automated gate dielectric breakdown characterization," IEEE Custom Integrated Circuits Conference (CICC),Sept. 2008, pp. 121-124.

[5] A. Drake *et al.*, "A distributed critical-path timing monitor for a 65 nm high-performance microprocessor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2007, pp. 398–399.

[6] M. Acar, A.J. Annema, B. Nauta," Digital Detection of Oxide Breakdown and Life-Time Extension in Submicron CMOS Technology," IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, Feb. 2008, pp. 530–633.

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

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

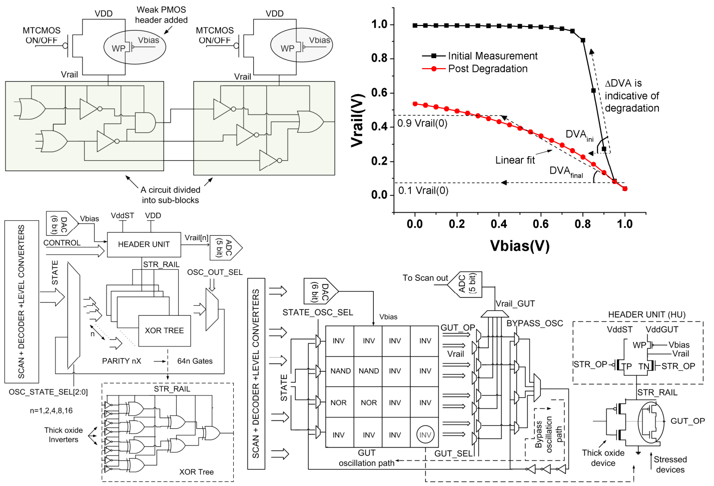

Figure 9.9.1: (Left Top) Circuit concept (Right Top) Simulated OBD detection (Left Bottom) Sensing circuit implemented on parity blocks of different gate count and (Right Bottom) logic gates (INV, NAND, NOR)

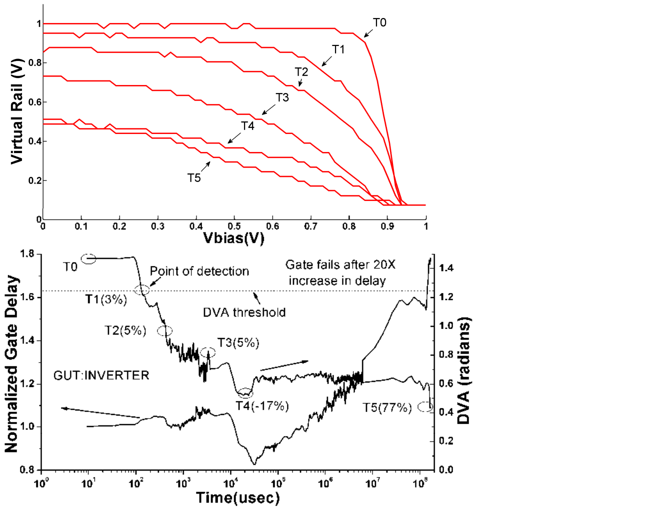

Figure 9.9.2: (Top) Measured V/V curve for an INVERTER (Bottom) Delay/DVA measurements for the same INVERTER. Stress voltage = 4V, Temperature =125C

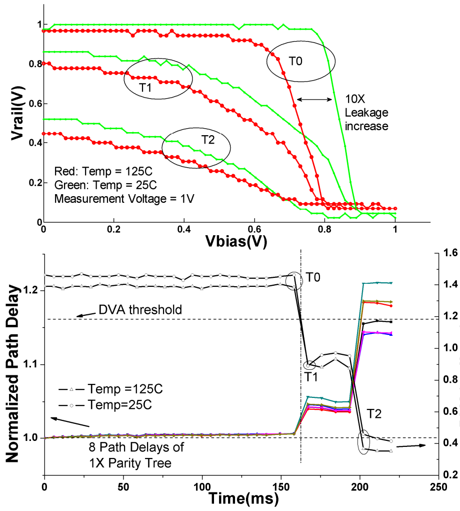

Figure 9.9.3: (Top) V/V curves at 25C and 125C at different degradation times T0 (initial), T1 and T2 for 64 gate parity block (Bottom) Corresponding delay/DVA measurements. Stress voltage = 4V

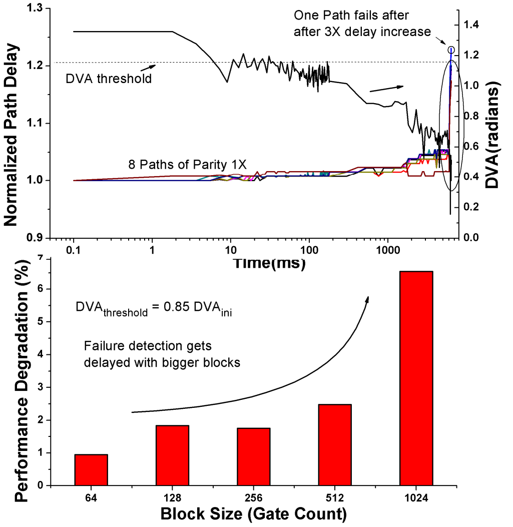

Figure 9.9.4: (Top) Delay/DVA measurements for a 64 gate parity block (Bottom) Failure detection for parity blocks of different sizes. Stress voltage=4V,Temperature=125C

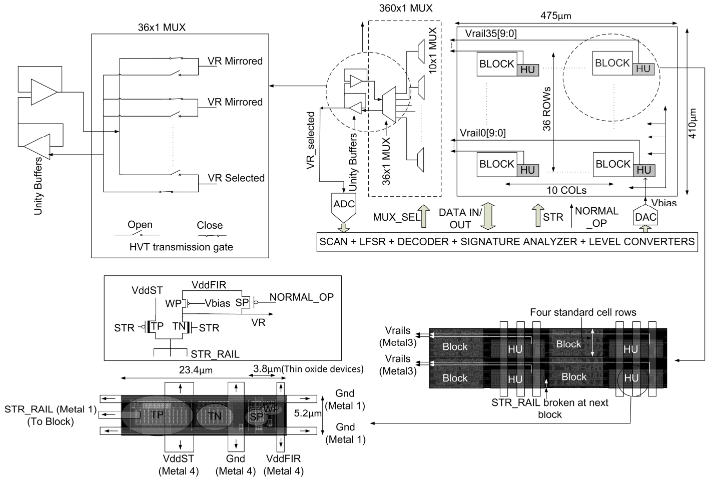

Figure 9.9.5: Sensing circuit implemented on 16 bit, 8-tap FIR filter. 141 blocks out of 360 blocks were stressed and monitored.

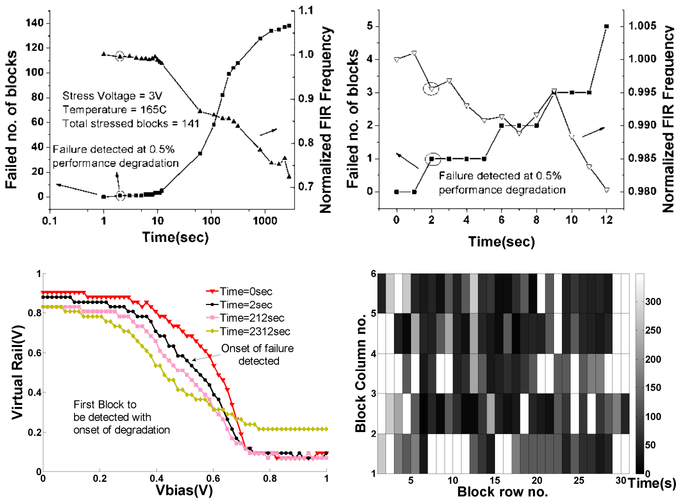

Figure 9.9.6: (Left Top, Right Top) Number of failed blocks and performance degradation rate (Left Bottom) V/V curves of first block to fail (Right Bottom) Spatial map of the failure time of blocks for a die.