Dynamic NBTI Management Using a 45nm Multi-Degradation Sensor

Prashant Singh[1], Eric Karl[2], Dennis Sylvester[1], David Blaauw[1]
[1]University of Michigan, Ann Arbor, Michigan
[2]Intel, Hillsboro, OR

***Abstract*— We propose a low power unified oxide and NBTI degradation sensor designed in 45nm process node. The cell power consumption is 10^5 lower than a previously proposed sensor. The unified nature enables efficient reliability monitoring with reduced sensor deployment effort and area overhead. Using the sensor Dynamic NBTI Management (DNM) has been implemented for the first time. DNM trades the excess 'reliability-margin' present in the design, due to better than worst case operating conditions, with performance. For the typical case shown in this paper, DNM allows for an average boost of 90mV in accelerated supply voltage while bringing down the excess NBTI margin of 22.5mV to 8mV where the total budget for NBTI was 66mV.**

I. Introduction

With technology scaling the transistor dimensions are being pushed to the limits while the supply voltage is not scaling proportionately. This has led to high electric fields across the gate-oxide, transistor channel and interconnects. Consequently degradation mechanisms such as gate-oxide wear out, Bias Temperature Instability (BTI), electromigration and Hot Carrier Injection (HCI) have worsened. As a result it has become harder to meet the lifetime specifications in advanced process nodes. Moreover since an *a priori* lifetime prediction requires worst-case assumptions on environmental conditions (voltage, temperature), highly conservative margins are imposed on supply voltage. The inherent statistical nature of the degradation process further adds to these margins and limits the benefits from advanced process nodes. To reduce this pessimism, dynamic reliability management (DRM) has been proposed. This method involves monitoring the voltage and temperature (V, T) of the chip and then estimating its reliability using statistical models [1]. But these models become inaccurate and difficult to calibrate under time varying conditions of (V, T) and hence considerable margins remain.

Sensors that directly measure degradation improve on this by eliminating the need for models, thereby enabling greater accuracy of DRM. The degradation sensors lie alongside the core circuitry and the test devices embedded in the sensor are exposed to the same environmental conditions as the core. The measurement circuitry in the sensor then detects the degradation of the test device. Since the environmental conditions vary spatially across the chip, the sensors need to be sprinkled across the chip. Moreover, since the degradation is statistical in nature, hundreds or even thousands of sensors are required in each locality to capture the statistics of the wear-out. This imposes severe constraints on the area and power of the sensors.

Degradation sensors have been proposed to characterize and study NBTI and oxide-breakdown [2, 3, 4, 5, 6]. Reference [2, 4, 5] proposed ring oscillator based structures to measure NBTI. The test structures consisted of a pair of nearly-identical ring oscillators one of which is stressed. A measurement circuitry compares the phases of the output of the two ring oscillators and generates an amplified 'beat' frequency. Reference [3] adopts a similar approach and translates the frequency measurements to V_{th} shift using a calibration mechanism. But all these methods average out the NBTI across all the PMOS transistors of the ring oscillator, resulting in loss of statistical information, such as the sigma of the V_{th} shift, which is important for DRM. Moreover these methods require a large number of oscillator stages to produce nearly-identical baseline frequencies, which increases the area of the sensors.

To characterize gate-oxide breakdown reference [6] proposed array based test structures in which the gate-oxide of the test devices are stressed. This approach is useful to characterize hard-breakdowns but has limitations measuring the initial gate-oxide wear-out early in its life time due to spurious leakage currents which overwhelm the gate-oxide current. For DRM enabled systems it is important to capture the early on-set of gate-oxide degradation so that the system can take appropriate measures soon enough to increase the remaining life-time of the chip. Finally, in [7] separate NBTI and oxide degradation sensors were proposed which are smaller in area but consumed high power.

In this paper we propose a unified NBTI and gate-oxide wear-out sensor designed in a 45nm process node. The sensor design is orders of magnitude lower power than [7] while keeping the sensor area small. The integration of NBTI and oxide degradation sensing enables efficient reliability monitoring with reduced sensor-deployment effort and overhead. Using the NBTI measurements from the sensor we implement Dynamic NBTI Management (DNM). To our knowledge this is the first time DNM (or DRM) has been demonstrated in silicon hardware.

DNM utilizes the 'reliability-margin' present in the design when its operating conditions are better than the worst PVT conditions. In our demonstration, the degradation sensors incur a slower rate of NBTI degradation at typical temperature than at high temperature. DNM exploits the resulting reliability-margin by increasing the supply voltage (and consequently the performance) while making sure that the NBTI degradation doesn't overshoot the specifications for NBTI tolerance. In our experiments we obtained an average

978-1-4244-5760-1/10/$26.00 ©2010 IEEE

increase of 90mV in supply voltage under typical conditions of temperature while reducing the excess NBTI margin from 22.5mV to 8mV where the total budget for NBTI was 66mV.

In Section II we first present the sensor design, followed by silicon sensor measurement results. In Section IV, we describe the DNM method and results.

II. Sensor Design And Operation

The proposed sensor consists of 2 DUTs (D1, D2), which are stressed and then measured for gate oxide and NBTI degradation, respectively. The other components of the circuit include: 1) muxes, to switch the sensor between stress and measurement modes, 2) a ring oscillator, which is shared between oxide and NBTI sensing circuitry, 3) a Schmitt trigger, to improve the slew of the signal originating from node *N2*, 4) level converters, to output NBTI measurements (Fig. 1).

To simplify the design we took advantage of the fact that gate-leakage in 45nm process node is comparable to the sub-threshold leakage, which simplifies the oxide sensing circuitry as node *N2* can be stressed directly using a transistor stack (S) rather through an oxide stack divider as in [7]. This also obviates the amplifier used in [7], which results in power reduction of the proposed sensor design.

The modes of operation and respective timing diagram of all the control signals and critical nodes are shown in Fig. 2. D1's gate oxide is stressed by charging *N2* to *VSTRESS* through the transistor stack S while *N1* is held at ground. During gate-oxide leakage measurement S is cut-off, allowing *N2* to be discharged through the gate leakage of D1. To prevent subthreshold leakage from affecting the discharge time, a cut-off device M1 is added to sink the stack subthreshold current. N2 drives a cross-coupled inverter-based Schmitt trigger to improve the slew of the signal originating from *N2*. The Schmitt trigger drives the ring oscillator, which is shared with the NBTI sensing circuitry. As D1's oxide degrades, its gate leakage increases and N2 is discharged faster, increasing the sensor frequency.

For NBTI sensing, D2 is stressed by muxing in a negative voltage at its gate. In the measurement mode D2 is biased in subthreshold so that any change in its V_{th} impacts the starved ring oscillator frequency exponentially. Since subthreshold circuits are extremely sensitive to temperature changes, a control header C1 is added to correct for any temperature change. A temperature calibration scheme as described in [7] is used in this work.

To record the frequencies a 20 bit counter, along with three 20 bit parallel registers are used. The counter and the storage registers collectively enable four consecutive fast frequency measurements. This feature is particularly useful to capture the phenomenon of NBTI recovery when the stress is interrupted to make measurements [8].

III. Silicon Measurements From Sensors

The sensor is implemented in a 45nm CMOS process. The test chip consisted of 16 banks, each bank containing 16 sensors, 80-bit storage which includes 20-bit counter, control and scan logic. The area of the sensor is 77.3μm^2 (6 Flip-flops). Sensor stress mode power is 8.6nW (>100,000X lower

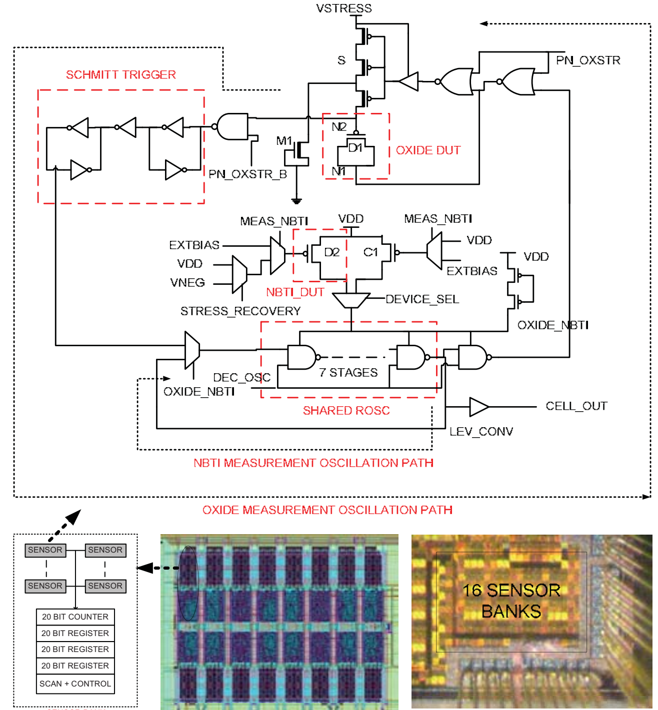

Fig. 1. (Top) Sensor circuit, (Left Bottom) Sensor bank Architecture, (Center Bottom) Chip Layout, (Right Bottom) Die shot

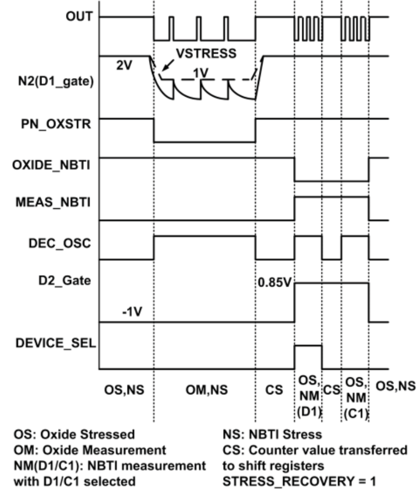

Fig. 2. Timing diagram of all the control signals and corresponding sensor modes of operation

than [7]) while measurement mode power is 84.7nW. Hence the power overhead of laying out thousands of sensors would only be a few hundreds of μW at maximum, which is a small fraction of power relative to a microprocessor core. The area and power overhead of the storage and other logic can be amortized by inculcating more sensors in a bank.

Fig. 3a shows the sensor output during gate oxide stress measurement results. The sensor captures the initial gradual increase in the gate-oxide leakage as the oxide wears out.

This data would be used by the DRM system to detect onset of gate-oxide wear-out or *soft breakdown*, and raise an alert.
Soft breakdown was defined as a 10% increase in gate leakage. In nominal conditions soft breakdown will occur over years. This is followed by a phase where the leakage remains relatively constant with occasional fluctuations. The fluctuation in gate-leakage occurs because defects are continuously being injected and neutralized during the process of degradation [9]. This noisy behavior of the gate-leakage is also indicative of the worn-out oxide as it allows more defect formation. At this point the DRM system would take measures to reduce the rate of wear-out. Finally there is a *hard breakdown* of the oxide marking the end of life of transistor, as a result of which there is a step increase of 17X in gate leakage and sensor frequency. The DRM system would ensure that hard breakdown does not occur in the core circuitry, by introducing some pessimism in its analysis while adjusting the supply voltage and temperature limit on the core.

Fig. 3b shows the distribution of time available to the DRM system after soft-breakdown detection (SBD), normalized to total life time. In this experiment 256 sensors were stressed out of which only 70 had hard breakdown at the end of the experiment. For 95% of the oxides soft breakdown occurred at less than half of their life time. In remaining 5% of oxides, it can occur as late as up to the last 10% of their lifetime.

Fig. 4 shows little correlation between pre-stress oscillation frequency and time to hard failure. The pre-stress frequency is indicative of the initial gate leakage, or gate oxide thickness. An outlier with high frequency or low oxide thickness will fail early, however similar failure times are observed even for nominal oxide thickness sensors. This confirms the significant randomness inherent in the formation of oxide defects that lead to failure [10] and supports the need for a larger number of low-power and compact reliability sensors on a chip to construct statistical bounds on expected lifetime.

Fig. 5a shows typical saw-tooth curve for NBTI degradation of device D2. The measurement time in all NBTI measurements is 100μs, which is same as in [7]. Fig. 5b shows ΔV_{th} of D2 due to NBTI, determined using sensor frequency measurements under different accelerated conditions of temperature and voltage. Comparing these threshold shifts in 45nm with results in [7] (in 130nm) under similar stress conditions confirms that scaling has significantly increased NBTI effects.

The pre-NBTI V_{th} differs among devices due to process variation. Fig. 6a shows correlation between pre-NBTI relative V_{th} and ΔV_{th} post-stress due to NBTI. Higher V_{th} devices are more likely to have large V_{th} shifts compared to low V_{th} devices. This indicates that slow corner chips degrade at a higher rate than nominal or fast corner chips. Furthermore, slower chips may operate at higher voltages to meet performance, further accelerating their degradation. Fig. 6b shows results for an implemented reliability scheme where a single active period (DC stress) is divided into ten active periods, each separated by a short sleep (recovery) time with an active: sleep ratio of 25:1. The recovery intervals reduce the degradation rate and hence can improve the overall performance and reliability of the chip.

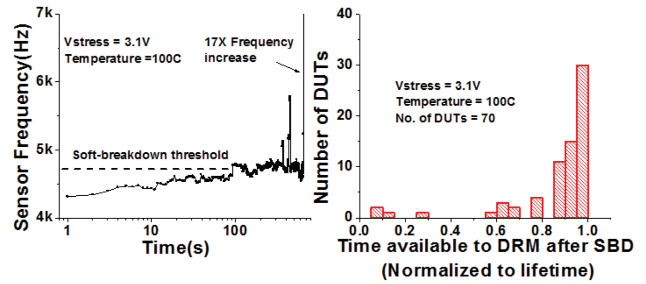

Fig. 3a. (Left) Gate-oxide stress and measurement results from the sensor.
Fig. 3b. (Right) Early soft breakdown detection gives sufficient time for DRM

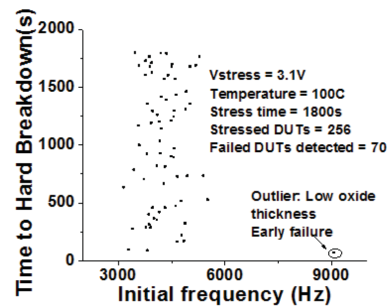

Fig. 4. No correlation observed between initial gate leakage and time to breakdown. This shows inherent random nature of oxide breakdown.

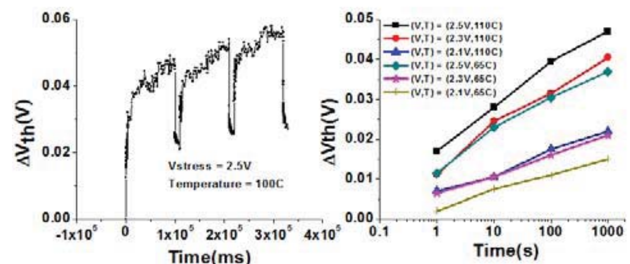

Fig. 5a. (Left) NBTI stress and recovery measurements from the sensor (On:Off = 5:1)
Fig. 5b. (Right) NBTI measured under different stress conditions of voltage and temperature. As expected the degradation shows power law dependence on time.

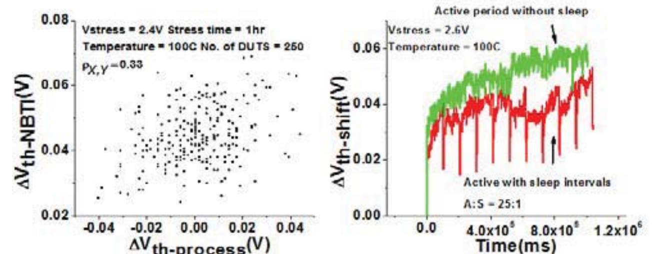

Fig. 6a. (Left) A weak positive correlation is observed between initial V_{th} and NBTI degradation
Fig. 6b (Right) Short sleep intervals result in 10-15% reduction in ΔV_{th} and performance improvement

IV. DYNAMIC NBTI MANAGAMENT

Typically process engineers impose a strict upper limit on the operating voltage of the circuits depending on the maximum tolerable threshold voltage shift at the end of lifetime ($\Delta V_{th\text{-}LT\text{-}MAX}$). Worst temperature conditions are assumed to find this limit which is a pessimistic assumption for a typical chip. Since NBTI degradation varies strongly with temperature change, there is a significant reliability margin which remains untapped at the end of the lifetime (T_{LT}). DNM trades these underlying margins with performance by increasing the supply voltage of the chip while making sure that the minimal required life time of the chip is met. In our experiment with accelerated conditions (V, T) T_{LT} = 3hrs and $\Delta V_{th\text{-}LT\text{-}MAX}$ was chosen to be ($\mu+3\sigma$) point of ΔV_{th} distribution (66 mV) obtained at 100C and 2.2V stress voltage.

In our DNM implementation we treated 32 sensors as actual sensors while the remaining 224 sensors were considered as core devices. The sensor-subset (32 sensors) is sampled periodically to obtain their ΔV_{th}. The computed ΔV_{th} values are fit to a power-law model in time:

$$\Delta V_{th}(t) = K\, t^{n}, \tag{1}$$

where K and n are fitting coefficients. We used a generic power law model because to our knowledge no work has been published which models NBTI with dynamic stress voltage variation. The fit-window is positioned over the slow-saturation regime of $\Delta V_{th}(t)$ curve because that regime gives more information about the long-term prognosis of NBTI. Fig. 7a shows the effect of the size of fitting- window on the accuracy of predicted ΔV_{th} at the end of life-time ($\Delta V_{th\text{-}LT\text{-}PRED}$). As shown, a sampling window of 200 samples, as used in this paper, results in an accurate fit.

The $\Delta V_{th\text{-}LT\text{-}PRED}$ data from each of the sensors in the sensor-subset is fitted to a Gaussian distribution and extrapolated to ($\mu+3\sigma$), to predict the maximum ΔV_{th} ($\Delta V_{th:\mu+3\sigma}$) for all of the core devices (with 99.7% confidence). If $\Delta V_{th:\mu+3\sigma} > \Delta V_{th\text{-}LT\text{-}MAX}$, then the supply voltage (in our case an accelerated supply voltage), is decremented by 100mV, and vice-versa. Fig. 7b shows the distribution of $\Delta V_{th\text{-}LT\text{-}PRED}$ compared to $\Delta V_{th\text{-}LT\text{-}MAX}$. Since $\Delta V_{th:\mu+3\sigma}$ is less than $\Delta V_{th\text{-}LT\text{-}MAX}$, DNM increments the stress voltage by 100mV. After this, the DNM algorithm waits for at least 100 samples to compute the new fit, so that $\Delta V_{th}(t)$ curve has stabilized. The algorithm waits to get a good fit to the data from all the sensors before it reevaluates $\Delta V_{th:\mu+3\sigma}$. Fig. 8a shows the readings from one of the sensors whose stress voltage is controlled by DNM. It also shows the corresponding model fit and the stress voltage scaling governed by DNM. The data processing and the DRM algorithm are implemented in MATLAB.

Fig. 8b shows the measured $\Delta V_{th\text{-}LT}$ distribution after NBTI stress at a worst-case temperature of 100C, a typical temperature of 55C without DNM and at 55C with DNM. DNM allows for an average boost of 90mV in the accelerated supply voltage while reducing the excess NBTI margin of 22.5mV to 8mV where the total budget for NBTI was 66mV.

V. CONCLUSION

DRM aims to trade the unused reliability margins present in a chip (due to static reliability margining) with performance. We proposed a unified NBTI and oxide wear-out sensor in 45nm process node which enables efficient DRM. The low area and power consumption (> 10^5 times lower then a previous sensor) of the sensor enables their use in large numbers. To our knowledge, a sensor based DNM was tested in silicon for the first time. For the typical case, the proposed DNM allows for an average boost of 90mV in the accelerated supply voltage while reducing the excess NBTI margin of 22.5mV to 8mV where the total budget for NBTI was 66mV.

ACKNOWLEDGEMENTS

We thank ST Microelectronics for their fabrication support.

REFERENCES

[1] E. Karl *et al*, DAC, pp. 1057-1060, 2006.

[2] M. Ketchen, M. Bhushan, and R. Bolam, Proc. IEEE Int. Conf. Microelectronic Test Structures, 2007, pp. 42-47.

[3] J. Keane, T.-H. Kim, and C. H. Kim, ISLPED, August 2007, pp. 189–194.

[4] T-H. Kim, R. Persaud, C.H. Kim., IEEE Journal of Solid-State Circuits, vol. 43, Issue 4, pp. 874–880, 2008.

[5] J. Keane, D. Parsaud, C.H. Kim, Symp. VLSI Circuits, pp. 108-109, 2009.

[6] J. Keane *et al*, CICC,Sept. 2008, pp. 121-124.

[7] E. Karl, P. Singh, D. Blaauw, and D. Sylvester, ISSCC, pp. 410-411, 2008.

[8] S. Rangan *et al*., *IEDM*, pp. 14.3.1-14.3.4, 2003.

[9] J.C. Reiner, Integrated Reliability Workshop, pp. 37-40, 2004

[10] J. H. Stathis, J. of App. Physics, vol. 86, no. 10, 1999, pp. 5757–5766.

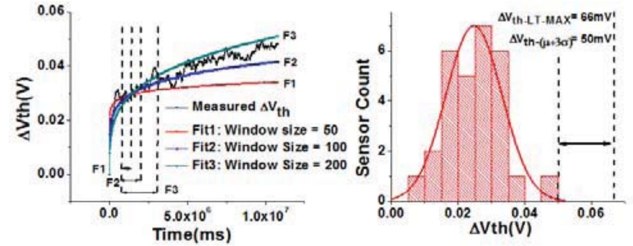

Fig. 7a (Left) Sampling window of 200 samples gives an accurate model fit. Fig. 7b (Right) Distribution of $\Delta V_{th\text{-}LT\text{-}PRED}$ at the first instance of NBTI evaluation by DNM. Since $\Delta V_{th:\mu+3\sigma} < \Delta V_{th\text{-}LT\text{-}MAX}$, supply voltage is incremented by 100mV

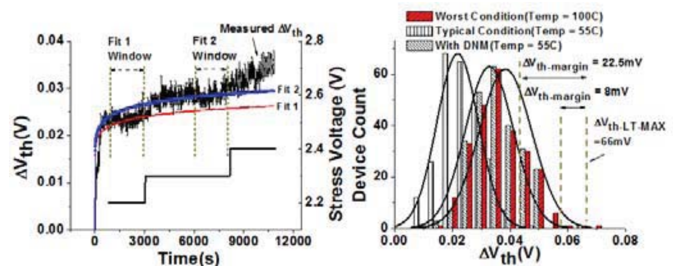

Fig. 8a.(Left) One of the 32 sensors' readings, with model fit and supply voltage scaling in a DNM implementation.
Fig. 8b.(Right) ΔV_{th} distribution at T_{LT} for 100C, 55C and with DNM at 55C