Impact of Device Scaling on Deep Sub-Micron Transistor Reliability - A Study of Reliability Trends Using SRAM

Mark White[1,2], Bing Huang[2], Jin Qin[2], Zvi Gur[2], Michael Talmor[2],
Yuan Chen[1], Jason Heidecker[1], Duc Nguyen[1], Joseph Bernstein[2]
[1] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109
[2] University of Maryland, College Park, MD 20742
Phone: 818-393-4173 Email: Mark.White@jpl.nasa.gov

Abstract. As microelectronics are scaled in to the deep sub-micron regime, users of advanced technology CMOS, particularly in high-reliability applications, should reassess how scaling effects impact long-term reliability. An experimental based reliability study of industrial grade SRAMs, consisting of three different technology nodes, is proposed to substantiate current acceleration models for temperature and voltage life-stress relationships. This reliability study utilizes Step-Stress techniques to evaluate memory technologies (0.25um, 0.15um, and 0.13um) embedded in many of today's high-reliability space/aerospace applications. Two acceleration modeling approaches are presented to relate experimental FIT calculations to Mfr's qualification data.

I. Introduction. The desire to assess the reliability of emerging technologies through faster reliability trials and more accurate acceleration models is the precursor for further research and experimentation in this field. Ramp-voltage and constant-voltage stress tests to determine voltage-to-breakdown and time-to-breakdown, coupled with temperature acceleration can be effective methods to identify and model the critical stress levels and reliability of emerging deep-sub micron microelectronics. While target product lifetimes for mil-product are generally 10 years at maximum rated junction temperature, leading edge commercial-off-the-shelf (COTS) microelectronics may be somewhat less due to reduced cost consumer electronics and reduced safety margins, including design life, as a result of increased power and thermal densities, increased performance characteristics, and device complexity. [1]

This reliability study utilizes Step-Stress techniques to evaluate some of the more recent memory technologies (0.25um, 0.15um, and 0.13um) embedded in many of today's high-reliability space/aerospace applications to substantiate current acceleration models for temperature and voltage life-stress relationships. The purpose of this study is to develop a better understanding of the impact of deep sub-micron technology scaling trends on microelectronics reliability. It also provides an independent assessment and validation of current acceleration models for users of scaled CMOS devices.

II. Failure Mechanisms & Modeling. Accelerated life testing of memories in this experiment is based on the assumption that various failure mechanisms are accelerated when elevated stress levels are applied to the operating component. The primary wear-out failure mechanisms include electromigration (EM), stress migration (SM), time-dependent-dielectric-breakdown (TDDB), thermal cycling (TC), and negative bias temperature instability (NBTI). The elevated parameters of concern are the ambient temperature [T] and the component operating voltage [V]. The models for evaluating the acceleration factors include Arrhenius for temperature, and Inverse Power or Exponential for voltage. [2] The acceleration models and parameters for various failure mechanisms remain uncertain for advanced technology CMOS devices, e.g. linearity, interactions between the stresses etc. Prior work by Srinivasan [3] and others are tabulated in Table 1, which shows the relative dependencies on temperature, voltage, and feature size of the primary wear-out failure mechanisms of interest.

Table 1. Summary of EM, SM, TDDB, and TC dependencies on temperature, voltage, and feature size.

Failure Mech.	Major temperature dependence	Voltage dependence	Feature size dependence
EM	$e^{\frac{E_{aEM}}{kT}}$		wh
SM	$\lvert T - T_0\rvert^{-m} e^{\frac{E_{aSM}}{kT}}$		
TDDB	$e^{\frac{(X+\frac{Y}{T}+ZT)}{kT}}$	$(\frac{1}{V})^{(a-bT)}$	$10^{\frac{\Delta t_{ox}}{0.22}}$
TC	$\frac{1}{\Delta T}$		

The relationship between $MTTF_{EM}$ and temperature is given by the following relationship [4]:

$$MTTF_{EM} \propto (J)^{-n} e^{E_{aEM}/kT} \qquad (1)$$

where J is the current density in the interconnect, E is the activation energy for EM, k is Boltzmann's constant, and T is absolute temperature in Kelvin. Higher operating temperatures will be seen with smaller technology nodes, therefore according to equation 1, more EM failures can be expected.

The relationship between SM and temperature is given by the following relationship [4]:

$$MTTF_{SM} \propto \left|T_o - T\right|^{-m} e^{E_{aSM}/kT} \qquad (2)$$

where T is the absolute temperature in Kelvin, T_o is the stress free temperature of the metal (the metal deposition temperature), and m and E are material dependent constants.
Temperature affects stress migration failure rate in two ways:

1. There is an exponential dependence on temperature which is detrimental to reliability.
2. There is the $|T - T_o|^{-m}$ term in equation 2 which has a positive effect on reliability.

The exponential term overshadows the other term, which means $MTTF_{SM}$ decreases, and therefore reliability decreases, with increasing temperature.

0-7803-8992-1/05/$20.00 ©2005 IEEE

The relationship between $MTTF_{TDDB}$ and temperature is given by [4]:

$$MTTF_{TDDB} \propto \frac{1}{V}^{a-bT} e^{\frac{X+\frac{Y}{T}+ZT}{kT}} \quad (3)$$

where T is the absolute temperature in Kelvin, a, b, X, Y, and Z are fitting parameters, and V is the voltage.

Decreasing gate oxide thickness with scaling decreases reliability due to increasing gate leakage and tunneling current, I_{leak}. The mean-time-to-failure due to gate oxide breakdown is directly proportional to the value of $I_{leak.}$ and increases by one order of magnitude for every 0.22nm reduction in gate oxide thickness [4]. As a result, if gate oxide thickness reduces by Δt_{ox} with scaling then $MTTF_{TDDB}$ reduces by $10^{\Delta tox/0.22}$ where the reduction in gate oxide thickness, $\Delta t_{ox,}$ is expressed in nanometers. For ultra-thin gate dioxides (< few nanometers), $MTTF_{TDDB}$ is inversely proportional to the total gate oxide surface area. According to equation 3, TDDB is also adversely affected by temperature, in which the dominating term is the exponential.

Permanent damage accumulates every time there is a cycle in temperature in VLSI devices, eventually leading to failure. Fatigue due to thermal cycling has the most impact at the package and die interface. The package goes through two types of thermal cycles: large cycles which occur at a low frequency (due to powering up and down), and small cycles which occur at a much higher frequency (due to variations in application behavior). The effect of small thermal cycles has not been well studied and validated models are not available [4].

The relationship between large thermal cycles and temperature is given by [4]:

$$MTTF_{TC} \propto \left(\frac{1}{T_{average} - T_{ambient}} \right)^{q} \quad (4)$$

where $T_{ambient}$ is the ambient temperature in Kelvin, $T_{average} - T_{ambient}$ is the average large thermal cycle a structure on the chip experiences, and q is the Coffin-Manson exponent, an empirically determined material-dependent constant.

Like EM and SM, the main impact of scaling on TC is the impact of temperature. Scaling has no other direct impact on thermal cycling.

NBTI is an effect that surfaced as gate oxide thickness was scaled. Gate oxide thickness for the 130-nm technology node has already resulted in sensitivity to NBTI. As the scaling of MOSFETs continues, NBTI becomes a more prominent issue in more current VLSI technology. It may become one of the ultimate limiting factors since NBTI is more severe than hot carrier stress for ultra-thin oxides at low electric fields [5]. The NBTI effect is more severe for PMOS FETs than NMOS FETs due to the presence of holes in the PMOS inversion layer that are known to interact with the oxide states [6]. In CMOS devices, the NBTI-induced threshold voltage shift will occur over a period of months or years, depending on the operating conditions of the device. Clearly, this means serious reliability issues for devices in terms of endurance and retention. NBTI is most problematic for high-performance or high-reliability devices, and analog/mixed-signal devices are more susceptible than digital devices.

III. Experimentation. Memory devices are excellent candidates for experimentation to demonstrate the accuracy and appropriateness of analytical models that have been proposed to characterize the life-stress relationship of present-day microelectronic devices. Volatile Static Random Access Memory (SRAM) devices are arranged in a matrix array and storage of data occurs within memory cells. These cells typically include between 4-6 transistors that form the inverter circuits and flip-flops, which are capable of assuming two states. Because the matrix array is designed for repetitive write-read cycles, large amounts of performance reliability data may be obtained through experimentation with relatively small quantities of commercial SRAM devices; technologies may be compared and contrasted with experimentation of a range of technology nodes.

A step-stress accelerated test technique was implemented to evaluate 1Mb (0.25um), 4Mb (0.15um) and 16Mb (0.13um) SRAM devices of similar cell designs configured in 128K x 8b, 256K x 16b, and 1M x 16b words respectively. Reference Tables 2 and 3. Devices were subjected to repetitive Write/Read cycles consisting of four data values for each memory cell or address at each stress step. Voltage was held constant while temperature was stepped-up, and then temperature was held constant while voltage was stepped-up. As stress conditions increased (voltage and temperature), bit failure times were read and recorded until devices catastrophically failed.

Underlying goals of this experiment were to:

- Calculate the FIT based on the test statistics without the physical models
- Validate the models and parameters upon failure investigation
- Segregation and data analysis
- Calculate the FIT using those models
- Compare and contrast to Mfr's FR data
- Determine if experimental results support lifetime reliability predictions across scaled technologies

A comparison of the results will then introduce more accurate statistical models and/or data fitting into existing physical failure model approaches, e.g. Inverse Power, Exponential, etc.

Table 2. Step-Stress Conditions (a)

Stress Conditions	Temp [°C]	V/Vnom	Time [hrs]
stress level 1	125	1.3	96
stress level 2	140	1.3	96
stress level 3	140	1.4	96
stress level 4	155	1.4	96
stress level 5	155	1.5	96
stress level 6	165	1.5	96
stress level 7	165	1.6	96
stress level 8	165	1.7	96

Table 3. Step-Stress Conditions (b)

Stress Conditions	Temp [°C]	V/Vnom	Time [hrs]
stress level 1	155	1.3	288
stress level 2	165	1.3	288
stress level 3	155	1.4	288
stress level 4	165	1.4	288
stress level 5	155	1.5	288
stress level 6	165	1.5	288
stress level 7	165	1.6	288

IV. Discussion & Results. Table 4 shows expected bit failure rates comparing Inverse Power and Exponential Voltage acceleration models and the manufacturer's life test data. Cumulative weighted test times were calculated for all stress operation levels. Total equivalent operating times were calculated for both Exponential and Power Law Models, and failure rate (λ) was calculated at 55°C and nominal operating voltage. Evaluation of the failure rate was conducted at 60% confidence using Reliasoft Alta software for maximum likelihood estimation with the assumption of constant failure rate. Cumulative weighted times were calculated to represent all the stress operation levels. Two basic assumptions were made: Case 1 reflects the assumption that there is only one dominating failure mechanism and the others are neglected; Case 2 reflects the assumption that there is no dominating failure mechanism, and that all are equally likely.

According to the assumptions outlined in Case 1 and Case 2, two models were applied: (a) Multiplication of AF's (temp. and voltage) using both Exponential and Power Law Models: AF_1 = AFt * AFv(e) and AF_2 = AFt * AFv(p); and (b) A proposed weighted sum model of the AF's where AF_3 = (AFt + AFv(e))/2 and AF_4 = (AFt + AFv(p))/2. These equations are expanded as follows:

$$AF1 = \frac{\lambda(T_2,V_2)}{\lambda(T_1,V_1)} = AF_T \cdot AF_V = \exp\left(\frac{E_a}{k}\left(\frac{1}{T_1}-\frac{1}{T_2}\right)\right)\exp(\gamma_1(V_2-V_1)) \quad (5)$$

$$AF2 = \frac{\lambda(T_2,V_2)}{\lambda(T_1,V_1)} = AF_T \cdot AF_V = \exp\left(\frac{E_a}{k}\left(\frac{1}{T_1}-\frac{1}{T_2}\right)\right)(V_2/V_1)k) \quad (6)$$

$$AF3 = \frac{\lambda(T_2,V_2)}{\lambda(T_1,V_1)} = (AF_T + AF_V)/2 = (\exp\left(\frac{E_a}{k}\left(\frac{1}{T_1}-\frac{1}{T_2}\right)\right) + \exp(\gamma_1(V_2-V_1)))/2 \quad (7)$$

$$AF4 = \frac{\lambda(T_2,V_2)}{\lambda(T_1,V_1)} = (AF_T + AF_V)/2 = (\exp\left(\frac{E_a}{k}\left(\frac{1}{T_1}-\frac{1}{T_2}\right)\right) + (V_2/V_1)k))/2 \quad (8)$$

Equations (7) and (8) may be expanded for n independent failure mechanisms where and λ_{LTFMi} represents the i^{th} failure mode at accelerated conditions, and λ_{useFMi} represents the i^{th} failure mode at normal conditions. AF then may be expressed as (9) assuming failure modes have equal frequency of occurrence during the *use conditions* [7]:

$$AF = \frac{\lambda_{use_{FM1}} \cdot AF_1 + \lambda_{use_{FM2}} \cdot AF_2 + .. + \lambda_{use_{FMn}} \cdot AF_n}{\lambda_{use_{FM1}} + \lambda_{use_{FM2}} + .. + \lambda_{use_{FMn}}} = \frac{\sum_{i=1}^{n} AF_i}{n} \quad (9)$$

The proposed weighted sum Exponential Model (7) best correlates the manufacturers published data (7-20 FIT) to the experimental data (19.482 FIT), normalized to 55C and nominal Vdd operating conditions. Reference Table 4. The accuracy of an estimate is given by its standard error and confidence interval. The estimates approximate the true parameter values and the confidence intervals for model parameters indicate the uncertainty in the statistical estimates by their width. Statistical confidence bounds do not account for model uncertainty. Therefore, sensitivity analysis is important in any quantitative analysis involving uncertainty and for assessing the effects of model uncertainty. In this experiment, model uncertainty was addressed by analyzing different model assumptions and different models to determine the best fit scenario between the test results, prior SRAM test results, and the manufacturer's failure rate qualification data. Maximum Likelihood methods were used to provide the estimates and confidence limits for the model parameters.

Table 4. Step-Stress Accelerated Test Results Compared to Manufacturer's Data

Test level	Cumulated test time	Equivalent op. time @55deg&nominal voltage			
		Case1 (Multiplication)		Case2 (Weighted Sum)	
		AFv Exp. Model (1)	AFv Power law (2)	AFv Exp. Model (1)	AFv Power law (2)
stress level 1	576	32464923.04	237589693.1	310353.6276	2170970.594
stress level 2	384	43090951.76	315354698.1	217390.3382	1457801.649
stress level 3	384	434116546.9	3918127282	1998870.897	17871738.22
stress level 4	384	824942335.4	7445532987	2017841.11	17890708.43
stress level 5	384	8310819403	77740152267	19965232.78	186422071.3
stress level 6	384	12452806266	1.16485E+11	19985188.96	186442027.5
stress level 7	335.8	1.09721E+11	9.14211E+11	175611815.3	1462841979
stress level 8	133.6	4.39858E+11	2.85782E+12	703819229.5	4572690225
Total equiv. time:		5.71677E+11	3.97817E+12	923925922.4	6447787521
Failure rate @55C &Vnom (FIT)		**0.031**	**0.004**	**19.482**	**2.792**
Failure rate reported by Manuf: 7 – 20 FIT					

Case 1 – refers to assumption a.
Case 2 – refers to assumption b.
(1) - Voltage Acceleration Factor according to Exponential. Model (γ = 7)
(2) - Voltage Acceleration Factor according to Power Law Model (k=34)
(3) – Mfr's FIT reported at 60% CL. ALT comparison also at 60% CL.

Examination of the component failure times show that at specific times, large numbers of bit failures were recorded. The failures that were recorded at the same time represent a single failure event which was reflected on multiple addresses and therefore, counted as a single failure for reliability evaluation. Hard and soft failures were treated equally in the reliability evaluation because once a soft failure has occurred in a high-reliability, remote application, e.g. an un-repairable system, the address corresponding to the failure are circumvented and not used in future write cycles. Table 5 shows technology node and stress conditions vs. accumulated time to failure of 0.1% of the bits in a device.

Table 5. Technology node and stress conditions vs. time to failure of 0.1% of the bits in a device.

Tech. Node	Vratio (Vapp/Vnom)	Temp C	Time (Hrs) to 0.1% Device-Bit Failures
0.13	1.4	165/155	588
	1.5		
0.15	1.6	165	528
0.25	1.7	165	768

Failure rate, Vratio stress, and temperature are plotted over time in Figures 1, 2 and 3 for three different technology nodes.

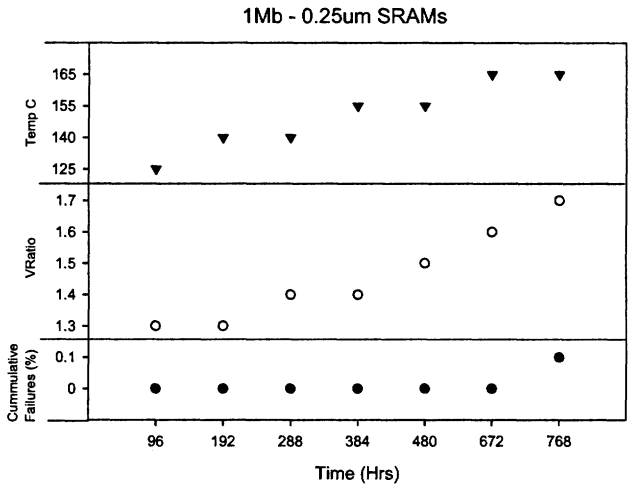

Figure 1. Time-to-fail (0.1%) as a function of Vratio and Temp (C)

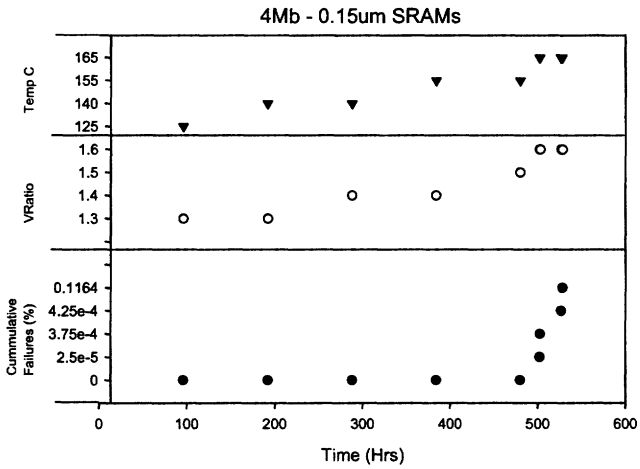

Figure 2. Time-to-fail (0.1%) as a function of Vratio and Temp (C)

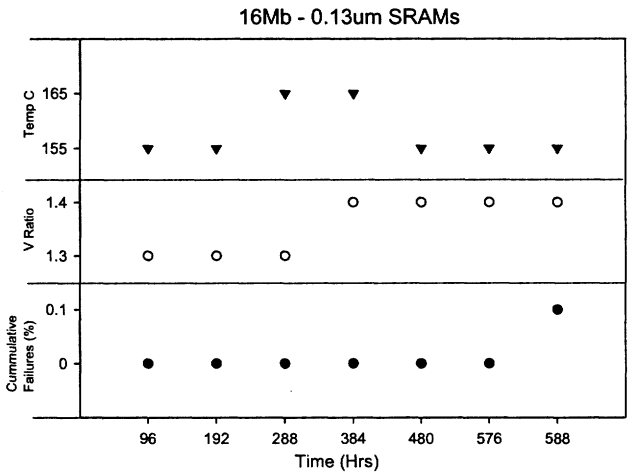

Figure 3. Time-to-fail (0.1%) as a function of Vratio and Temp (C)

V. Conclusion. An experimental based reliability study of industrial grade SRAMs consisting of three different technology nodes was conducted to substantiate current acceleration models for temperature and voltage life-stress relationships. Two different acceleration models were tested to relate experimental FIT calculations to Mfr's qualification data; the weighted sum exponential model best correlated. While time-to-fail across technology nodes were generally of similar magnitudes, the V stress ratio (increased V dependency) appears to be a primary failure mechanism driver with smaller technology nodes. Experimental results do support reduced lifetime reliability predictions as technologies are scaled. Failure analysis to identify root cause failure mechanisms and further experimentation with 90 and 65-nm technology nodes is warranted.

VI. Acknowledgements. The work described in the paper was conducted at the Jet Propulsion Laboratory, California Institute of Technology in collaboration with the University of Maryland.

REFERENCES

[1] White, M, et al., "Impact of Junction Temperature on Microelectronics Device Reliability and Considerations for Space Applications," IEEE IRW (2003).

[2] Meeker W. and Escobar L., "Statistical Methods for Reliability Data", John Wiley and Sons, c 1998.

[3] Srinivasan, Adve, Bose, Rivers. "The Impact of Technology Scaling on Lifetime Reliability." International Conference on Dependable Systems and Networks. June 2004.

[4] "Failure Mechanisms and Models for Semiconductor Devices." JEDEC Publication JEP122-A, 2002.

[5] Zhu, Suehle, Bernstein, and Chen. "Mechanism of Dynamic NBTI of pMOSFETs." Integrated Reliability Workshop Final Report, 2004 IEEE International, 2004.

[6] Peters, Laura. "NBTI: A Growing Threat to Device Reliability." Semiconductor International. http://www.reed-lectronics.com/semiconductor/article/CA386329

[7] Talmer, et al. "Competing Failure Modes in Microelectronic Devices and Acceleration Factors Modeling," Intl. Symposium on Stochastic Models in Reliability, Safety, Security and Logistics, Feb. 2005 Proceedings, Israel.