RC23047 (W0312-121) December 29, 2003
Computer Science

IBM Research Report

The Impact of Technology Scaling on Processor Lifetime Reliability

Jayanth Srinivasan, Sarita V. Adve
Department of Computer Science
University of Illinois at Urbana-Champaign

Pradip Bose, Jude Rivers
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

The Impact of Technology Scaling on Processor Lifetime Reliability

Jayanth Srinivasan, Sarita V. Adve
University of Illinois at Urbana-Champaign
Department of Computer Science
{srinivsn,sadve}@cs.uiuc.edu,

Pradip Bose, Jude Rivers
IBM T.J. Watson Research Center
Yorktown Heights,NY
{pbose,jarivers}@us.ibm.com

Abstract

The relentless scaling of CMOS technology has provided a steady increase in processor performance for the past two decades. However, increased power densities (hence temperatures) and other scaling effects have an adverse impact on long-term processor lifetime reliability. This paper represents a first attempt at quantifying the impact of scaling on lifetime reliability due to intrinsic hard errors, taking workload characteristics into consideration.

For our quantitative evaluation, we use RAMP [20], a previously proposed industrial-strength model that provides reliability estimates for a workload, but for a given technology. We extend RAMP by adding scaling specific parameters to enable workload-dependent lifetime reliability evaluation at different technologies.

We show that (1) scaling has a significant impact on processor hard failure rates – on average, we find the failure rate of a 65nm processor to be 316% higher than a similarly pipelined, scaled 180nm processor; (2) of all the failure mechanisms, time-dependent dielectric breakdown and stress migration are the most significant, due to increasing temperatures, less than ideal voltage scaling, and reduced interconnect dimensions; and (3) with scaling, the difference in reliability from running at worst-case vs. typical workload operating conditions increases significantly, as does the difference from running different workloads. Our results imply that leveraging a single microarchitecture design for multiple remaps across a few technology generations will become infeasible; microarchitects must incorporate lifetime reliability awareness at the early design stage; and this awareness must incorporate workload-specific vs. worst-case considerations.

1 Introduction

Advances in CMOS semiconductor technology have been improving processor performance steadily over the last twenty or so years. These performance advances have been driven by aggressive scaling of device feature sizes. However, such advances are accelerating the onset of long-term hardware processor failure (or lifetime reliability) problems. This paper represents a first attempt at *quantifying the impact of scaling* on lifetime reliability of an entire processor, while consider the behavior of the workload running on the processor. Our work focuses on *intrinsic hard failures*, and considers failures due to electromigration (EM), stress migration (SM), time dependent dielectric breakdown (TDDB), and thermal cycling (TC). We do not model extrinsic hard failures and soft errors because they generally do not impact lifetime reliability, as further discussed in Section 2.

Scaling theory and practice.

Device scaling results in the reduction of feature sizes and voltage levels of transistors. Application of ideal scaling theory results in three main benefits in going from one generation to the next [4]: (a) reduction of gate delay by about 30%, resulting in an increase in operating frequency by about 43%; (b) doubling of transistor density; and (c) reduction of dynamic power per transistor by about 50% (this assumes constant electric field scaling, where the supply voltage scales down by 30% in each generation). Combining the beneficial effects of (b) and (c) implies that for the same die size, under ideal scaling considerations alone, the net chip dynamic power and power density would remain unchanged with scaling.

However, in practice, processors do not scale ideally. With real scaling in the deep sub-micron range, processor power density, and consequently temperature, have been increasing at an alarming rate, which directly affects processor lifetime reliability. The main reasons behind this increase are:

- Supply voltages are not scaling down at the ideal rate, and this prevents the dynamic power per transistor from decreasing at the ideal rate. One of the reasons behind the slowing down of supply voltage scaling is the attempt to retain competitive frequency growth by tuning up the voltage to the maximum levels allowed in a given technology generation. Also, as the gap between the threshold voltage and the supply voltage diminishes to less than a volt, basic noise immunity issues (in logic) and cell state stability issues (in SRAM macros) makes it ever harder to scale down the supply voltage. Hence, although processor area scales down ideally, power does not, resulting in higher power density and consequently higher temperatures.

- The total chip leakage power is increasing with every technology generation. Scaling down device threshold voltages (ideally) by about 15% per generation causes sub-threshold leakage current per transistor to increase by 5 times. Since the total transistor count on the die increases by about 50% per generation, the total chip leakage power increases about 7.5 times. This increase is further compounded by the exponential dependence of leakage power on temperature.

The impact of scaling on lifetime reliability and prior work.

The above non-ideal scaling coupled with the reduced feature sizes affects processor lifetime reliability in the following ways. First, all of the four failure mechanisms considered here are adversely affected by increases in temperature, with some of these mechanisms exhibiting an exponential or larger dependence on temperature. Second, the dielectric thickness of devices is fast decreasing to the point where it is approaching a few angstroms. This, coupled with the fact that there has been a general slowdown in supply voltage scaling is expected to increase the intrinsic failure rate due to gate oxide (dielectirc) breakdown (TDDB). Third, the decreasing feature size of interconnects accelerates electromigration failure rates.

The detrimental impact of scaling on intrinsic reliability in general, and gate oxide reliability in particular, has been studied extensively [7, 22, 13]. However, most of these studies have been performed at the device level, and consider individual failure mechanisms in isolation. Additionally, they are performed at fixed worst case operating points without any knowledge of the target application suite of the processor. However, an important part of understanding the impact of scaling lies in the target workload of the processor. Since the power consumed by the processor varies with the executing workload, the actual operating temperature and interconnect current densities also depend on the workload. Consequently, the failure rate of a component (or the processor as a whole) depends on the workload. Given that an application oblivious analysis of processor reliability would produce unrepresentative reliability data, it is critical to account for application characteristics.

In recent work, we have proposed an industrial-strength simulation model, called RAMP, to evaluate processor lifetime reliability for a workload, but for a given technology [20]. That work uses the workload dependence of lifetime reliability to motivate microarchitecture level mechanisms to address the growing lifetime reliability problem.

Our contributions.

To the best of our knowledge, our results represent the first quantitative evaluation of the impact of device scaling on the hard error rates and lifetime reliability of processors, from a micro-architectural perspective and incorporating workload dependence. This paper enhances RAMP by adding scaling specific parameters to enable workload-dependent lifetime reliability evaluation at different technologies. In particular, our evaluations and analysis attempt to model the scaling effects of taking one chip design, and gradually scaling that chip down from 180nm to 65nm, without any substantial modifications to the microarchitectural pipeline. Our key findings are as follows.

Our first set of results show that scaling has a significant and increasing impact on processor hard failure rates. The increase in processor temperature is one of the key reasons for this trend. In our experiments, on average, the maximum temperature reached by a 65nm processor is 15 degrees Kelvin higher than that reached by a 180nm processor. (This is based on optimistic assumptions about relative cooling costs; in reality, this number would be even higher.) The failure rate for a 65nm processor is 316% higher than the failure rate at 180nm, with similar reliability qualification. More importantly, the rate of increase of failure rate increases as we scale to smaller technologies.

Comparing the different failure mechanisms, we find that dielectric breakdown (TDDB) will provide the largest challenge followed by electromigration. The effects on stress migration are much less drastic and thermal cycling appears to be within manageable bounds.

Our results clearly demonstrate that hard failures will present a significant and increasing challenge in future technology generations. An important practical consequence is that, in contrast to current practice, leveraging a single design for multiple remaps across a few technology generations (with only minor design tweaks) will become infeasible – new microarchitectural design would be required frequently to ensure robust (and perhaps increasingly modest) increases in integration and overall performance.

Our second set of results quantify the impact of scaling on the workload-dependent nature of lifetime reliability. Our results show that failure rates computed by assuming worst-case operating conditions are increasingly pessimistic compared to those computed with real workloads, as we scale to smaller technologies. Furthermore, scaling amplifies the difference in failure rates between different applications.

Specifically, we computed the worst-case failure rate assuming a steady state operation based on the highest temperature and activity factor reached by any of our applications. The difference between this worst-case failure rate and the actual maximum failure rate seen by any application went from 25% for 180nm to 90% for 65nm (computed as a percentage of the maximum failure rate). The difference between steady-state worst case and actual average failure rate was even more striking – 67% at 180nm to 206% at 65nm.

We also see that the range of failure rates across different applications in our suite grows from 62% at 180nm to 104% at 65nm (the range is the difference between the maximum and minimum failure rates, reported as a percentage of the average failure rate across all applications).

Thus, at technologies with smaller feature sizes, reliability qualification for worst case operating conditions will result in significantly and increasingly over-designed processors. A promising approach, proposed in [20], is to perform reliability qualification for the expected case, backed up with dynamic application-specific responses for handling departures from the expected case. Our quantification unequivocally shows that these or alternate mechanisms will be increasingly important as we scale to lower feature sizes.

2 Background

As mentioned in Section 1, we use a simulation methodology called RAMP, described in [20], to calculate lifetime reliability of processors from a microarchitectural viewpoint. RAMP represents the first architectural-level methodology for evaluating processor lifetime reliability, and uses state-of-the-art analytic models for important intrinsic failure mechanisms. RAMP was developed based on extensive discussions with back-end, front-end, and reliability groups at IBM T.J. Watson Research Center.

RAMP's design and implementation details are discussed in detail in [20]. RAMP currently models four main wear-out intrinsic failure mechanisms experienced by processors – electromigration (EM), stress migration (SM), gate-oxide breakdown or time dependent dielectric breakdown (TDDB), and thermal cycling (TC) [2, 1]. RAMP implements the failure models at a microarchitectural structure level, for a *given technology generation*. The standard reliability metric used in the analytical models in RAMP is MTTF, which is the average expected lifetime of the processor.

RAMP should be used in conjunction with a timing simulator to determine workload behavior, a power simulator to calculate cycle by cycle power consumption, and a thermal simulator which provides the continuous temperature profile of different structures on the processor. The specific simulators we use in this paper are described in Section 4.

As mentioned in Section 1, we use RAMP to only study *intrinsic hard failures*, which are wear-out based failures that are intrinsic to, and depend on, the materials used to make the processor, and are related to process parameters, wafer packaging and processor design. We do not evaluate the impact of scaling on extrinsic hard failures, which arise from process and manufacturing defects. After manufacturing, using a technique called burn-in [14], processors are tested at elevated operating temperatures and voltages in order to accelerate the manifestation of extrinsic failures. As a result, extrinsic hard failure rates tend to be very low in shipped processors and do not significantly affect processor lifetime-reliability. Finally, we do not discuss soft errors in this paper. Although soft errors can cause errors in computation and corruption to data, they do not fundamentally damage the processor and are not viewed as a lifetime reliability concern.

Next, we review the individual failure models in RAMP.

2.1 Electromigration

Electromigration is well understood, and extensive research has been performed by the material science and semiconductor community on modeling and understanding the effects of electromigration [1, 7].

Electromigration in processor interconnects is due to the mass transport of conductor metal atoms in the interconnects due to momentum transferred by the electron current. This can lead to the formation and growth of voids at sites of depletion leading to open circuits, increased interconnect resistance, and other problems. At the site of metal atom pile up, extrusions can form causing shorts between adjacent metal lines causing circuit failure.

The model for the MTTF due to electromigration, $MTTF_{EM}$, used in RAMP is [20]:

$$MTTF_{EM} \propto (J)^{-n} e^{\frac{E_{a_{EM}}}{kT}} \tag{1}$$

where J is the current density in the interconnect, $E_{a_{EM}}$ is the activation energy for electromigration, k is Boltzmann's constant, and T is absolute temperature in Kelvin. n and $E_{a_{EM}}$ are constants that depend on the interconnect metal used.

RAMP models electromigration at the granularity of a microarchitectural structure. The value of J for a structure is equal to the product of the activity factor of the structure, p, and the maximum allowed interconnect current density for that technology generation. The value of p for a structure is obtained from the timing simulator. The maximum allowed interconnect current density is a technology specific parameter and its absolute value was not relevant to the work in [20]. The numbers we use in this paper are discussed in Section 4.6. Also, RAMP models copper interconnects and uses a value of 1.1 for n and 0.9 for $E_{a_{EM}}$ [1].

2.2 Stress Migration

Stress migration which is similar to electromigration, is a phenomenon where the metal atoms in the interconnects migrate due to mechanical stress. Stress migration is caused by thermo-mechanical stresses which are caused by differing thermal expansion rates of different materials in the device [1].

The model for the MTTF due to stress migration, $MTTF_{SM}$, used in RAMP is [20]:

$$MTTF_{SM} \propto |T_0 - T|^{-m} e^{\frac{E_{a_{SM}}}{kT}} \tag{2}$$

where T is the absolute temperature in Kelvin, T_0 is the stress free temperature of the metal (the metal deposition temperature), and m and $E_{a_{SM}}$ are material dependent constants.

As mentioned, RAMP models copper interconnects and uses a value of 2.5 for m and 0.9 for $E_{a_{SM}}$ for stress migration. RAMP assumes that sputtering (versus vapor deposition) was used to deposit the interconnect metal and uses a value of 500K for T_0 [9].

2.3 Time-dependent dielectric breakdown (TDDB)

Time-dependent dielectric breakdown, or gate oxide breakdown, is another well studied failure mechanism in semiconductor devices. The gate oxide (or dielectric) wears down with time, and fails when a conductive path forms in the dielectric. When a conducting path forms between the gate and the substrate, it is no longer possible to control current flow between the drain and the source with a gate electric field, effectively rendering the transistor device useless [3, 13, 22].

The model for the MTTF due to TDDB used in RAMP [20] is based on recent experimental work performed by Wu et at. at IBM [22] [1]:

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

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

Based on the experimental data collected by Wu et al. [22], the values used in RAMP for the TDDB model are $a = 78$, $b = -0.081$, $X = 0.759ev$, $Y = -66.8evK$, and $Z = -8.37e - 4ev/K$.

2.4 Thermal Cycling

Fatigue failures can occur due to temperature cycling. Permanent damage accumulates every time there is a cycle in temperature eventually leading to failure. Fatigue due to thermal cycling is most pronounced in the package and die interface (for example, solder joints) [1].

The package goes through two types of thermal cycles – large cycles which occur at a low frequency (a few times a day) due to effects like powering up and down, and small cycles which occur at a much higher frequency (a few times a second) due to variations in application behavior. The effect of small thermal cycles at high frequencies has not been well studied by the packaging community, and validated models are not available, and is hence not modeled in RAMP.

The model for the MTTF due to large thermal cycles is based on the Coffin-Manson equation [20] and is:

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

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

As mentioned previously, RAMP only models cycling fatigue in the package, since that is where the impact of cycling is most pronounced. For the package, the value of the Coffin-Manson exponent, q, is 2.35 [1].

2.5 Sum-of-failure-rates (SOFR) Model

RAMP uses the sum-of-failure-rates (SOFR) model [21] to calculate overall reliability of the processor [20]. The SOFR model is used to combine the effects of different failure mechanisms, across different structures, over time.

[1] Although RAMP models a fixed technology generation, it includes the dependence on voltage to account for techniques like dynamic voltage scaling found in recent processors.

Combining failure mechanisms, in general, requires knowledge of the lifetime distributions of the mechanisms, and is generally difficult to do. The SOFR model, which is a standard model used by the industry, makes two assumptions to address this problem: (1) The processor is a series failure system – in other words, the first instance of any structure failing due to any failure mechanism would cause the entire processor to fail; and (2) each individual failure mechanism has a constant failure rate (equivalently, every failure mechanism has an exponential lifetime distribution). The failure rate (also known as the hazard function), $h(t)$ at a time t, is defined as the conditional probability that a component will fail in the interval $(t + \delta t)$, given that it has survived till time t. A constant failure rate implies that the value of $h(t)$ will remain fixed, and will not vary with the component's age; i.e., $h(t) = \lambda$. This assumption is clearly inaccurate – a typical wear-out failure mechanism will have a low failure rate at the beginning of the component's lifetime and the value will grow as the component ages (the probability that a component will fail will increase, the older the component gets). However, this assumption is often used in the industry for lack of better validated models. Our future work includes integrating time dependence in to our failure models.

The above two assumptions imply [21]: (1) The MTTF of the processor, $MTTF_p$, is the inverse of the total failure rate of the processor, λ_p; and (2) the failure rate of the processor is the sum of the failure rates of the individual structures due to individual failure mechanisms. Hence,

$$MTTF_p = \frac{1}{\lambda_p} = \frac{1}{\sum_{i=1}^{j} \sum_{l=1}^{k} \lambda_{il}} \tag{5}$$

where λ_{il} is the failure rate of the i^{th} structure due to the l^{th} failure mechanism.

The standard method of reporting constant failure rates for semiconductor components is in Failures in Time (FITs) [21], which is the number of failures seen per 10^9 device hours – $MTTF_p = \frac{1}{\lambda_p} = \frac{10^9}{FIT_{value}}$. We will use FITs as our metric when reporting results.

As mentioned previously, the time distribution of processor reliability is also important. It is important to understand that the processor FIT value alone does not portray a complete picture of processor reliability, and incorporating time dependent failure models is an important area of future work.

Finally, the proportionality constants used in the individual failure mechanism models (Equations 1, 2, 3, 4) have to be provided to RAMP. These constants depend on many factors like the materials used for design, yield, etc. High values for the proportionality constants imply more reliable processors, which comes at a higher cost. Cheaper systems will have smaller values for the proportionality constants. For a specific set of proportionality constants, and by using architectural and performance parameters obtained from an architectural timing simulator, a power simulator and a temperature simulator, RAMP provides FIT values for simulated applications.

3 Impact of Scaling on Intrinsic Failure Mechanisms

In this section, we discuss the impact of scaling on the failure mechanisms discussed in Section 2. As mentioned in Section 2, RAMP models reliability for a given technology generation. In this section, we discuss our methodology of extending RAMP to model different technology generations. For each failure mechanism modeled in RAMP, we examine the parameters that change with scaling and derive models for failure rates at different technology generations.

3.1 Electromigration

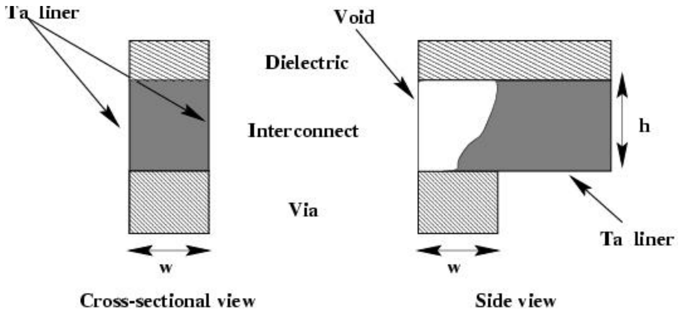


Figure 1. Electromigration in copper interconnects.

For years, copper doped aluminum had been the semiconductor industry's interconnect metal of choice because of its ease of integration into the manufacturing process, low resistivity, and cheap availability. In the past few years, reliability problems due to electromigration and a need for even lower resisitivities prompted the industry to consider using only copper for interconnects. Copper has lower resistivity (and hence lower interconnect delay) than copper doped aluminum and is much more resilient to electromigration (Experimental results by Hu et al. [7] show that the electromigration lifetime of copper interconnects is 50 to 1000 times as high as copper doped aluminum interconnects).

Copper interconnects are typically fabricated using a damascene processing method. In these structures, the top surface of the copper damascene line is covered with a dielectric film, while the bottom surface and two sidewalls are sealed with a tantalum (Ta) liner [8]. The tantalum liner prevents electromigration along the surfaces it covers. However, the top surface of the line can not be covered with tantalum due to manufacturing constraints. As a result, electromigration in copper is dominant at the top interface layer between the interconnect and the dielectric [7]. This is illustrated in Figure 1.

If the effective thickness of the interface layer is δ, and the interconnect width is w, then the electromigration flux is constrained to an area δw. If the height of the interconnect is h, then the interconnect current flows through an area wh. The relative amount of atomic flux flowing through the interface region is proportional to the interface area to interconnect area ratio $\frac{\delta w}{wh} = \frac{\delta}{h}$.

Electromigration voids are found to occur most commonly at the interface between the interconnects and the metal vias [7]. Although large vias are favorable from a reliability perspective, large vias incur area overheads and the interconnect density is reduced. As a result, the width of the via is kept the same as the width of the interconnect, w. Electromigration failure is considered to have occurred when the void formed grows larger than the width of the via, w (in that case, there is no path to conduction between the void and the interconnect other than the liner, and hence, resistance goes up causing circuit failure). Mean time to failure due to electromigration, $MTTF_{EM}$, is then proportional to the width of the via, w, and inversely proportional to the relative amount of flux passing through the interface region, $\frac{\delta}{h}$ [7]. Combining these terms with the previous equation for electromigration (Equation 1):

$$MTTF_{EM} \propto (J)^{-n} e^{\frac{E_{a_{EM}}}{kT}} wh \tag{6}$$

Thus, when a scaling factor of κ is applied to Equation 1, electromigration lifetime reduces by κ^2 due to w and h (both w and h scale by κ while δ remains constant). Additionally, as discussed in Section 2, the value of J for a structure is equal to the product of the activity factor of the structure, p, and the maximum allowed interconnect current density for that generation. This maximum allowed current density changes with scaling. The values we use are given in Table 4, and justified in Section 4.6.5. Finally, $MTTF_{EM}$ reduces further due to the higher temperatures seen on chip with scaling.

3.2 Stress Migration

The only impact of scaling on stress migration that we model is the dependence on temperature, which is already modeled in Equation 2. Temperature affects stress migration failure rate in two ways: there is an exponential dependence on temperature which is detrimental to reliability, and there is the $|T - T_0|^{-m}$ term from Equation 2 which has a positive effect on reliability. However, the exponential term usually overshadows the other term, resulting in a decrease in reliability with temperature. Scaling has no other direct impact on stress migration.

However, there are some indirect effects due to scaling. Scaling requires the increased use of low-k dielectrics for the inter-level dielectric layers. These materials tend to be porous, and brittle. Despite the use of low-k dielectrics by some manufacturers like IBM for their 0.13μm copper process the dielectrics are thought to have some reliability problems [10]. The thermal expansion rates of these new low-k dielectrics also tends to be significantly different from the interconnect thermal expansion rate [10]. A consequence of the different expansion rates and the brittle nature of the dielectrics causes higher failure rates due to thermo-mechanical stresses in stress migration. However, since our experiments assume that our scaled processors all use the same type of interconnect metal and dielectric material, we do not model any of the above effects on scaling.

3.3 Time-dependent dielectric breakdown (TDDB)

Scaling has a profound effect on gate oxide reliability. Effects of scaling on TDDB already modeled in RAMP in Equation 3 are the detrimental effect of increasing temperatures and the beneficial effect of decreasing supply voltage. Due to less than ideal scaling as described earlier, however, the benefit seen from lower supply voltage is also less than ideal. Further, the decrease in voltage does not compensate for the reduced reliability caused by the higher temperatures and the other scaling parameters as described below.

First, decreasing gate oxide thickness with scaling decreases reliability. The fundamental problem is related to increasing gate leakage current, I_{leak}. With ultra thin gate oxides with thickness, t_{ox}, less than 4nm, I_{leak} is further increased due to an added gate tunneling current component, which is also increasing [13]. The mean time to failure due to gate oxide breakdown is directly proportional to the value of the I_{leak}. I_{leak} increases by one order of magnitude for every 0.22nm reduction in gate oxide thickness [13]. As a result, if gate oxide thickness reduces by Δt_{ox} with scaling, then $MTTF_{TDDB}$ reduces by $10^{\frac{\Delta t_{ox}}{0.22}}$, where the reduction in gate oxide thickness, Δt_{ox}, is expressed in nanometers.

Intrinsic Failure mechanism	Temperature dependence	Voltage dependence	Other feature size dependence
EM	$e^{\frac{E_{a_{EM}}}{kT}}$		wh
SM	$\lvert T-T_0\rvert^{-m}e^{\frac{E_{a_{SM}}}{kT}}$		
TDDB	$e^{\frac{(X+\frac{Y}{T}+ZT)}{kT}}$	$(\frac{1}{V})^{(a-bT)}$	$10^{\frac{\Delta t_{ox}}{0.22}}$
TC	$\frac{1}{T^q}$		

Table 1. Summary of impact of scaling on MTTF of different failure mechanisms. The parameters above are explained in Section 3. Scaling voltage has a positive effect on TDDB. All other scaling effects are negative.

Second, for the current and future range gate oxide thicknesses, $MTTF_{TDDB}$ is inversely proportional to the total gate oxide surface area [13, 22].

Combining the scaling effect of voltage, gate oxide thickness, area and temperature, if we scale down from process 1 to process 2, which have supply voltages V_1 and V_2, gate oxide thicknesses, t_{ox1} and t_{ox2}, total gate oxide areas of A_1 and A_2, at temperatures T_1 and T_2, the ratio of mean time to failures, $MTTF_1$ and $MTTF_2$ is given by:

$$\frac{MTTF_1}{MTTF_2} = 10^{\frac{(t_{ox1}-t_{ox2})}{0.22}} \times \frac{V_2^{(a-bT_2)}}{V_1^{(a-bT_1)}} \times \frac{A_1}{A_2} \times \frac{e^{\frac{(X+\frac{Y}{T_1}+ZT_1)}{kT_1}}}{e^{\frac{(X+\frac{Y}{T_2}+ZT_2)}{kT_2}}} \tag{7}$$

where X, Y, Z, a and b are empirically determined constants.

3.4 Thermal Cycling

Like stress migration, the only impact of scaling on thermal cycling we model is the impact of temperature, which is already modeled in RAMP. Scaling has no other direct impact on thermal cycling.

There are some indirect effects on thermal cycling due to scaling. Fewer thermal cycles are required to cause failure in low-k dielectrics because of the increased porosity and brittleness. The adhesive properties of dielectrics (in particular low-k dielectrics) also degrades with scaling, increasing susceptibility to thermal cycling failure. However, again, since our experiments assume that our scaled processors all use the same type of interconnect metal and dielectric material, we do not model these effects of scaling.

3.5 Summary of impact of different parameters on intrinsic failure rates

Table 1 summarizes the impact of different parameters on the intrinsic failure mechanisms. It shows that temperature has an exponential detrimental impact on EM and SM (despite the $|T - T_0|$ in SM), a more than exponential impact on TDDB, and a less than exponential impact on TC. Electromigration is also detrimentally impacted by smaller values of w and h, and TDDB is adversely affected by reducing t_{ox}. Finally, a positive effect of scaling is observed in TDDB due to lower supply voltages. (Note that lower voltages also help with temperature, but not enough because of increasing power density.)

4 Experimental Methodology

4.1 Architecture Modeled and Performance Simulation Methodology

The base processor simulated is a 180nm out-of-order 8-way superscalar processor, conceptually similar to a single core 180nm POWER4-like processor [15]. Table 2 summarizes the base 180nm processor modeled. Note

Technology Parameters	
Process technology	180 nm
V_{dd}	1.3 V
Processor frequency	1.1 GHz
Processor core size (not including L2 cache)	81mm^2 (9mm x 9 mm)
Leakage power density at $383K$	0.04 W/mm^2
Base Processor Parameters	
Fetch/finish rate	8 per cycle
Retirement rate	1 dispatch-group (=5, max) per cycle
Functional units	2 Int, 2 FP, 2 Load-Store (Agen), 1 Branch, 1 LCR
Integer FU latencies	1/7/35 add/multiply/divide (pipelined)
FP FU latencies	4 default, 12 div. (pipelined)
Reorder Buffer size	150
Register file size	120 integer, 96 FP
Memory queue size	32 entries
Base Memory Hierarchy Parameters	
L1 (Data)	32KB
L1 (Instr)	32KB
L2 (Unified)	2MB

Table 2. Base 180nm POWER4-like processor.

that some of the microarchitectural parameters like cache sizes, assumed in our base model, are different from real values in the POWER4 processor. Also, although we model the performance impact of the L2 cache, we do not model its reliability. This is because the temperature of the L2 cache is much lower than the processor core [15], resulting in very low L2 intrinsic failure rates. Hence, we concentrate on intrinsic failures in the core.

The processor is modeled using a trace-driven research simulator called Turandot [16], developed at IBM T.J. Watson Research Center. The modeled microarchitecture is conceptually similar to a single core 180nm POWER4-like processor [15]. As described in [16], Turandot was calibrated against a pre-RTL, detailed, latch-accurate processor model. Despite the trace-driven nature of Turandot, the extensive validation methodology provides high confidence in the results.

4.2 Power Simulation Methodology

To estimate power dissipation, we use the PowerTimer toolset developed at IBM T.J. Watson Research Center [6]. This toolset, in its default form, is built around the Turandot cycle-accurate performance simulator referred to in the previous section. The power models that are built into the Turandot-based PowerTimer are based on circuit accurate power estimations from the 180nm POWER4 processor [15]. The power analysis has been performed at the macro level using CPAM, a circuit-level power analysis tool [18]. Multiple macros are combined to form microarchitectural structures. PowerTimer combines the circuit accurate power estimates from over 400 macros into 60 primary microarchitectural structures. PowerTimer uses microarchitectural activity information obtained from the performance simulator, Turandot, to provide per-cycle power estimates. For our simulations, we use realistic clock gating assumptions in PowerTimer, in tune with actual data available from current generation (post-POWER4) microprocessors.

4.2.1 Leakage Power

Leakage power is calculated based on modeled structure areas. For the base 180nm process modeled, a leakage power density of 0.04 W/mm^2 at 383K is used. This value is based on simulation-based estimates for processors like the POWER4 , and assumes standard leakage power control techniques like the use of high-threshold devices in non-critical logic paths and arrays.

We also model the impact of temperature on leakage power using the technique in [11]. At a temperature T, the leakage power, $P_{leakage(T)}$, is given by:

$$P_{leakage(T)} = P_{leakage(383K)} \times e^{\beta(T-383)} \tag{8}$$

where β is a curve fitting constant. The value of β we use is taken from [11].

4.3 Temperature Simulation Methodology

We use the HotSpot tool [19] to derive temperature estimates from power. The chip floorplan fed to HotSpot resembles a single core of a 180nm POWER4-like processor, of size $81mm^2$ ($9mm$ x 9 mm). The microarchitectural structures modeled using Turandot and PowerTimer are combined into 7 structures. Based on each structure's area, HotSpot calculates thermal resistance and capacitance values. These thermal resistances and capacitances are combined into an RC network. HotSpot dynamically solves this RC network to produce temperature measurements at the granularity of 1μ second (using power information from PowerTimer).

4.3.1 Heat Sink Temperatures

As explained in [19], the RC time constant of the processor heat sink is significantly larger than the RC time constant of individual structures. Since the RC time constant of the heat sink is so large, there is not enough time for there to be significant changes in heat sink temperature during simulation runs. Hence, it is critical that HotSpot be initialized with accurate heat sink temperatures.

As a result, all simulations are run twice - the first run is used to obtain average power consumption values for every structure on chip. These average power values are then used to calculate the initialization temperature of the heat sink. Once the heat sink is initialized, the second run produces accurate temperature results.

For the sake of comparing technology generations, we maintain a constant heat sink temperature for each application with scaling (different applications have different heat sink temperatures, which remain constant with scaling). In order to simulate this, we vary the thermal resistance of the heat sink with technology generation in HotSpot. It should be noted that this could result in potentially conservative failure rate estimates for advanced technologies. In an actual scaling scenario, maintaining constant heat sink temperature might not be feasible from a heat sink cost perspective, resulting in higher heat sink temperatures (and resultant higher peak temperatures) for advanced technology generations.

4.4 Reliability Calculation

Based on temperature estimates obtained from HotSpot and power estimates obtained from PowerTimer, RAMP calculates FIT values, for every structure on chip at 1 μsec intervals (for all failure mechanisms). A running average of these FIT values is maintained which provides the final FIT value of the structure (the sum of which will give the processor FIT value).

As mentioned previously, the proportionality constants in the failure mechanism equations in RAMP are dependent on various factors and vary with acceptable cost. To determine the value of these constants, we use an

Type	Application	IPC	180nm processor power (W)	Maximum Temperature reached (K)
Spec2K Float	ammp	1.06	26.08	334.5
	applu	1.17	26.94	336.4
	sixtrack	1.38	27.32	336.0
	mgrid	1.71	27.78	338.1
	mesa	1.75	29.21	339.3
	facerec	1.79	29.60	339.4
	wupwise	1.66	30.50	341.2
	apsi	1.64	30.65	341.0
SpecFP average		**1.52**	**28.51**	
Spec2K Int	vpr	1.38	26.93	335.5
	bzip2	2.31	27.71	336.6
	twolf	1.26	28.44	337.5
	gzip	1.85	28.69	337.1
	perlbmk	2.25	30.59	340.7
	gap	1.76	31.24	341.2
	gcc	1.24	31.73	340.9
	crafty	2.25	31.95	342.4
SpecInt average		**1.79**	**29.66**	

Table 3. Average IPC, power consumption, and the maximum temperature of the hottest structure on the 180nm base processor for our workload, ordered by increasing power.

approach similar to [20]) as follows. Current processors are expected to have an MTTF of around 30 years – this implies that the total FIT value of the processor should be around 4000 (10^9/30years). We assume that each failure mechanism contributes equally to the total FIT value at reliability qualification. Hence, we assume that reliability qualification is performed for the 180nm processor such that the average FIT value of each individual failure mechanism across all the applications is 1000, giving the system a total average FIT value of 4000. The failure rate at other technology points is calculated relative to $180nm$.

Finally, as mentioned in Section 4.1, although we model the performance of the L2 cache, we do not model its reliability.

4.5 Workload Description

Our experimental results are based on an evaluation of SPEC2K benchmarks. We report experimental results based on PowerPC traces of 16 SPEC2K benchmarks (8 SpecInt + 8 SpecFP). The SPEC2K trace repository used in this study was generated using the Aria trace facility in the MET toolkit [17], and was generated using the full reference input set. Sampling was used to limit the trace length to 100 million instructions per program. The sampled traces have been validated with the original full traces for accuracy and correct representation [12].

Table 3 summarizes the benchmarks studied, including the IPC, average power consumption, and maximum temperature reached on the 180nm base processor. The power values include leakage power consumption. As can be seen, for our processor, SpecInt has a higher average IPC and marginally higher power consumption than SpecFP.

4.6 Scaling Methodology

Tech gen nm	V_{dd} V	Frequency GHz	Relative Capacitance	Relative Area	t_{ox} $\AA$	Interconnect cur density $\frac{mA}{m^2}$	Leak. power $\frac{W}{mm^2}$	Average Total Power (Dyn+Leak) (W)	Relative Total Power Density
180	1.3	1.1	1.0	1.0	25	9.0	0.040	29.1	1.0
130	1.1	1.35	0.7	0.5	17	6.0	0.10	19.0	1.31
90	1.0	1.65	0.49	0.25	12	4.0	0.25	14.7	2.02
65 (0.9V)	0.9	2.0	0.4	0.16	9	4.0	0.54	14.4	3.09
65 (1.0V)	1.0	2.0	0.4	0.16	9	4.0	0.60	16.9	3.63

Table 4. Scaling parameters used.

We study the failure rate for our POWER4-like processor for five technology generations, ranging from $180nm$

to $65nm$. The scaling parameters used are listed in Table 4. All scaling is done with respect to $180nm$, as the performance and power simulator are calibrated for this technology point. A scaling factor of 0.7 is assumed from $180nm$ to $90nm$. For $90nm$ to $65nm$, a scaling factor of 0.8 is used, based on the assumption that a scaling factor of 0.7 will be difficult to maintain in technology generations after 90nm. Next, we discuss each column in Table 4.

4.6.1 Voltage and frequency scaling

With an ideal scaling factor of 0.7, [4] states the best-case device frequency scaling per generation would be about 43%. However, in the most recent era of technology remaps, ideal frequency scaling is constrained, primarily due to tighter design rules caused by increased wiring pressures. Also, in doing progressive scaling of the same microarchitecture over multiple technology generations, it is hard to achieve ideal frequency boosts without significant investment in re-tuning all the circuit delay paths in the machine. Hence, we assume conservative 22% frequency scaling per generation. (If we assume 43% scaling, power and temperature figures will be higher resulting in even higher failure rates.)

With ideal scaling, the supply voltage should scale by a factor of 0.7 every generation. However, since threshold voltage scaling is limited to a few mV per technology generation, as the supply voltage becomes smaller, the reduction obtained by scaling slows down, such that the device overdrive (supply voltage minus threshold voltage) does not drop by more than a factor of 0.7. Additionally, due to the increase in leakage power with threshold voltage scaling, further scaling limits are imposed on threshold voltage and supply voltage. The supply voltage values in Table 4 are carefully chosen to match up with the scaled frequencies, while adhering to threshold voltage scaling that would reflect the leakage power density assumptions shown.

Also, we simulate two 65nm processors. One processor assumes that the voltage scales down from 90nm to 65nm to a value of 0.9 V. However, as the supply voltage approaches the threshold voltage, scaling voltage appropriately is becoming increasingly difficult. Basic noise immunity issues (in logic) and cell state stability issues (in SRAM macros) make it difficult to operate reliably at voltages below 1.0 V. As a result, we also simulate a 65nm processor which runs at 1.0 V, which we believe is more realistic (no voltage scaling takes place between 90nm and 65nm). The two different technology points are represented as 65nm (0.9V) and 65nm (1.0V) in our results.

4.6.2 Capacitance scaling

The capacitance value for each technology generation is proportional to the scaling factor used for that generation. The 180nm processor is assumed to have a relative capacitance value of 1.0.

4.6.3 Area scaling

The area of the processor for each technology generation is proportional to the square of the scaling factor used for that generation. The 180nm processor is assumed to have a relative area of 1.0.

4.6.4 t_{ox} scaling

The values of t_{ox} used were obtained from the high performance logic parameters in the ITRS roadmap [2]. As can be seen, changes in t_{ox} are proportional to the scaling factor.

4.6.5 Interconnect current density

In order to compensate for decreasing interconnect dimensions, the current through interconnects is also reduced. In order to compensate for the decrease in electromigration reliability with scaling, designers have been reducing interconnect current density every technology generation. We assume a 33% reduction in interconnect current density every technology generation [4]. However, this implies less and less current for devices, and limits are being reached. It is expected that interconnect current density can not be reduced beyond the 90nm technology point. Hence the values for 90nm and 65nm are the same.

4.6.6 Power scaling

The leakage power densities used for each technology point assume aggressive leakage control techniques are used [5]. The total power consumption (dynamic + leakage power), averaged across all applications, is also given (based on simulations). Finally, the relative total power density (which is the ratio of the total power consumption and area), averaged across all applications, is also given. As can be seen from Table 4, up to 90nm, scaling reduces the total power consumption of the core [2]. Beyond 90nm, the increase in leakage power negates the benefits from reduced dynamic power, resulting in a net increase in total power. Also, it can be seen that the average power density goes up steadily with scaling (because voltage is not scaling down ideally and leakage power is going up).

5 Results

5.1 Temperature analysis

Figure 2 shows the average temperature of the hottest structure on chip for each application for each technology generation. Also shown is the heat sink temperature, averaged over all applications (recall that we adjust the heat sink thermal resistance such that the average heat sink temperature remains constant with scaling).

As can be seen, while the heat sink temperature remains nearly constant with scaling, the peak temperature increases. On average, from 180nm to 65nm (1.0V), the peak temperature of the hottest structure on chip increased by 15 degrees Kelvin. The peak temperature for each application increases with scaling because the power density on chip (as seen in Table 4) is increasing. Finally, in an actual scaling scenario, maintaining a constant heat sink temperature might not be feasible from a cost perspective. In the situation where the heat sink temperature is not constant, but also increases with scaling, the range seen in peak temperatures would be even larger.

The results also show that there is a significant range in temperatures across applications. There is high correlation between application power and temperature and some correlation with IPC. The hottest applications (wupwise for SpecFP and crafty for SpecInt) also have the highest power consumptions (and high IPCs) in Table 3. The same holds for the coolest applications in our suite (ammp for SpecFP and vpr for SpecInt). Also, on average, SpecInt applications have marginally higher temperatures than the SpecFP applications. This, again, is because of the

[2] We again emphasize that this is for remapping of the same core.

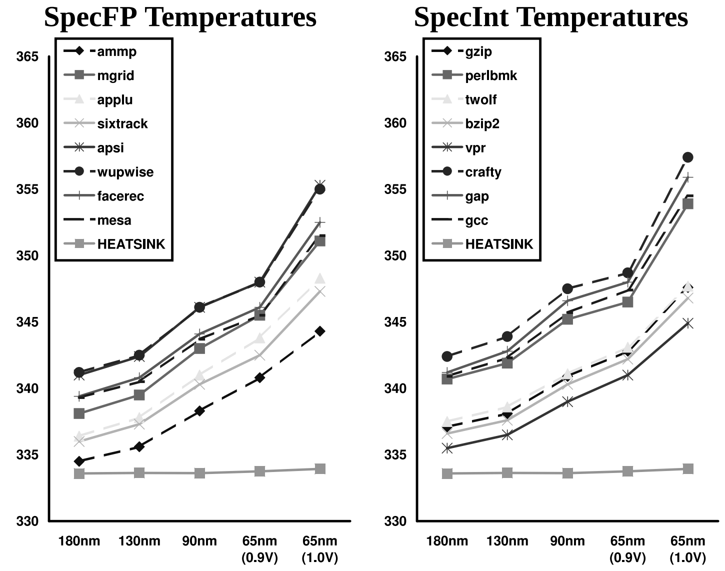
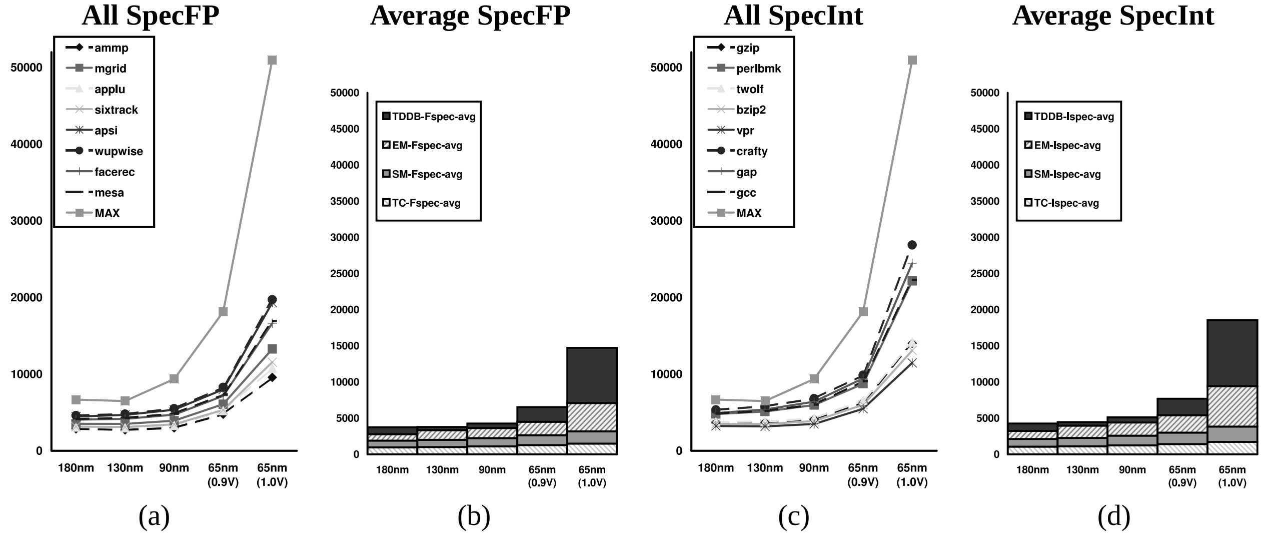


Figure 2. Maximum temperature at the hottest structure in Kelvin (shown on the vertical-axis) reached for each application for the different technology points (horizontal-axis). The heat sink temperature averaged over all the applications is also shown.

higher average power consumption of the SpecInt applications for our system (as can be seen in Table 3). The distinct segregation of SpecInt applications into two groups arises from their power dissipation characteristics. As can be seen in Table 3, four SpecInt applications have power consumptions distinctly lower than the other four SpecInt applications. SpecFP on the other hand has a uniform spread of power distributions.

5.2 Total FIT value scaling



(a) (b) (c) (d)

Figure 3. (a) and (c) show the total processor FIT value for each application. (b) and (d) show the FIT value averaged across all (SpecInt or SpecFP) applications, and the relative contribution of individual failure mechanisms to the total FIT value.

Figure 3 presents the data for this section. Parts (a) and (c) show the scaling behavior of the total processor FIT value for each application, for SpecInt and SpecFP respectively. They also show FIT values calculated based on worst case conditions over all the applications. To compute these values, we found the highest activity factor (p)

and the highest temperature across all applications and used them for the entire run. Note that this is worst case conditions only for the applications studied – it is possible that the maximum FIT value of the processor can be even higher. Parts (b) and (d) of Figure 3 show the FIT value averaged across all the applications, with scaling (for SpecInt and SpecFP respectively). At each technology generation, each FIT bar has also been broken down into the individual contributions by each failure mechanism.

5.2.1 Increase in Total FIT value

As can be seen, there is a marked rise in the total FIT value with technology scaling. On average, the total FIT value of the SpecFP applications increased by 274% from 180nm to 65nm (1.0V). The increase seen in SpecInt was larger at 357%. Also, at each technology point, the average FIT value of SpecInt applications was higher than SpecFP applications. This is because of the higher power consumptions seen in the integer applications. Figure 3 (b) and (d) give the contributions of the individual failure mechanisms, which are further analyzed in Section 5.3.

There is a significant difference in FIT value from 65nm (0.9V) to 65nm (1.0V). As discussed in Section 4.6.1, many architectural structures can potentially not operate reliably at voltages lower than 1.0V. However, as can be seen, maintaining a constant voltage from 90nm to 65nm leads to a large rise in FIT values. On the other hand, if the voltage does scale down from 90nm to 65nm, the increase in FIT value seen from 180nm to 65nm (0.9V) is brought down to (a still significant) 70% for SpecFP and 86% for SpecInt.

5.2.2 Workload dependence on FIT value

In Figures 3 (a) and (c), when considering the workload dependence on the total FIT value, there are two points of note. First, the worst case FIT value is distinctly higher than the FIT value of any individual application. More significantly, this difference increases with scaling. Specifically, the worst case FIT value was 25% higher than any application for 180nm and 90% higher than any application at 65nm (computed as a percentage of the maximum failure rate). More striking was the difference between the worst case FIT value and the average application FIT value – 67% at 180nm and 206% at 65nm.

Second, Figures 3 (a) and (c) also show that there is a large range in FIT values across applications. FIT values for applications correlate well with application temperature. The hottest applications (from Figure 2) also have the highest FIT values, and the order of the curves in Figures 2 and 3 remains the same. This is because, at any *given technology point*, the only difference in the FIT values of applications arises from temperature differences and from differences in the value of J (through the activity factor, p). Also, the slope of the FIT value curves is steeper than the slope of the temperature curve. This is because of the more than linear dependence of FIT values on temperature (as can be seen in the temperature column in Table 1). Additionally, the range in FIT values across applications also increases with scaling. The range across all applications (SpecFP + SpecInt) increases from 2479 FITS (which is 62% of the average FIT value) at 180nm to 5095 (which is 72% of the average FIT value) at 65nm (0.9V) to 17272 (which is 104% of the average FIT value) at 65nm (1.0V).

Our results indicate that future reliability qualification mechanisms should be targeted at specific applications or classes of applications. If an application oriented reliability qualification methodology is not used, the processor would be severely over-designed for most applications. This issue is also discussed in [20] where an application-aware dynamic reliability management scheme is proposed. Our quantification unequivocally shows the increasing importance of this and other application-aware reliability approaches, as processors are designed with increasingly smaller feature sizes.

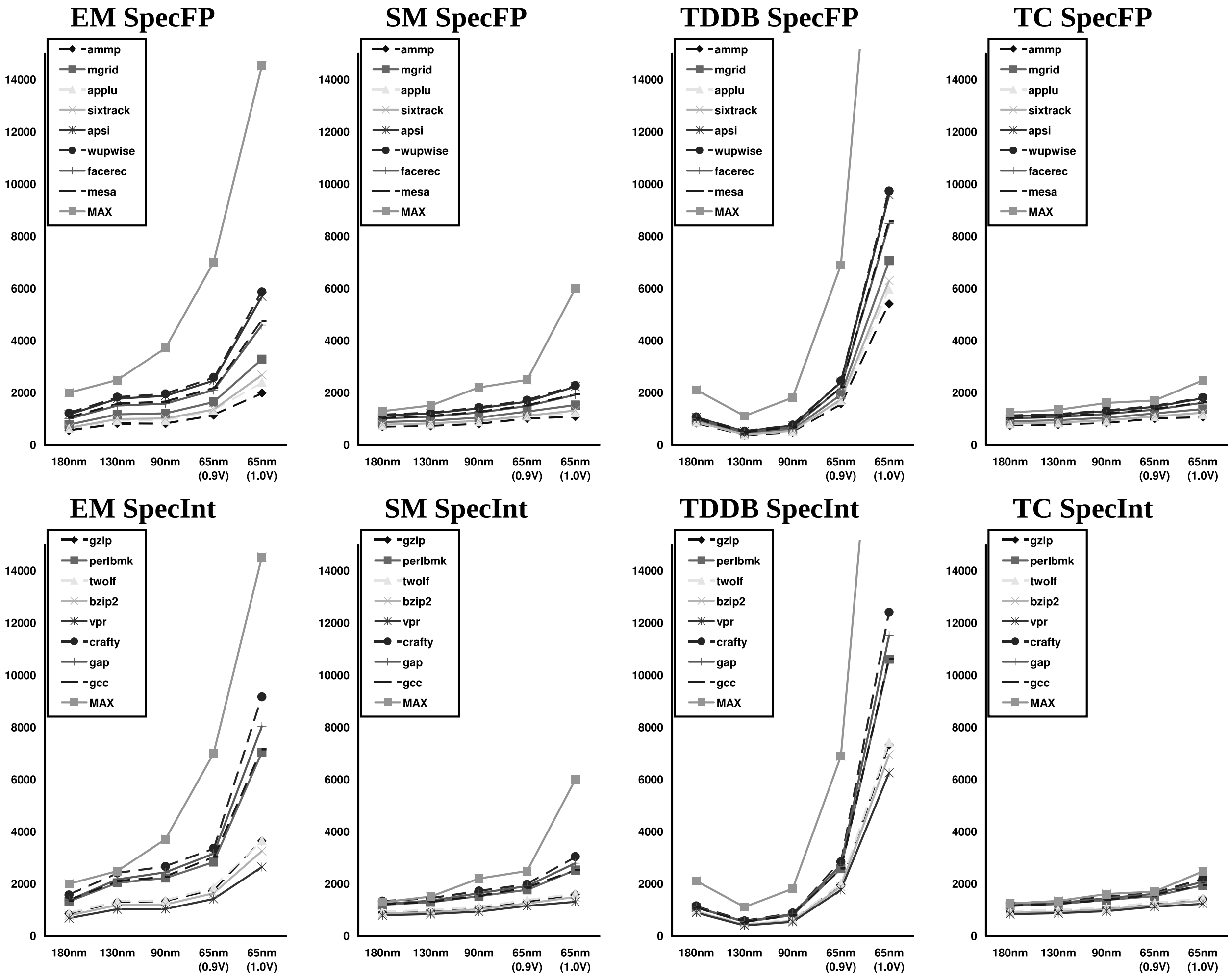


Figure 4. Failure rates for each failure mechanism for SpecFP and SpecInt. The worst case FIT value curve for each mechanism is also shown.

5.3 Individual failure mechanisms

Next, we examine scaling behavior in individual failure mechanisms. Figure 4 shows the scaling behavior for individual failure mechanisms for all the applications. The failure rates for each individual mechanism based on worst case operating conditions is also given.

5.3.1 EM scaling

As can be seen in Figure 4, scaling has a significant impact on electromigration failure rate. The electromigration failure rate increases by 303% on average for SpecFP benchmarks and 447% on average for SpecInt benchmarks between 180nm and 65nm (1.0V). The increase is 97% for SpecFP and 128% for SpecInt between 180nm and 65nm (0.9V).

As can be seen from Table 1, the increase is due to temperature as well as a reduction in interconnect dimensions (w and h). The temperature dependence is underscored by the difference in FIT values between 65nm (0.9V) and 65nm (1.0V) (where the only distinction is from temperature).

As discussed in Section 5.2, there is a large range in FIT values across applications. This range is amplified in electromigration. There is also a large difference between worst case failure rates and application failure rates.

5.3.2 SM scaling

For SM, there is a 76% increase between 180nm and 65nm(1.0V) and a 43% increase between 180nm and 65nm (0.9V) for SpecFP on average. The corresponding values for SpecInt are 106% and 52%.

Scaling impacts stress migration through an increase in temperature. The exponential dependence of stress migration failure rate on temperature (as shown in Table 1) can be seen in Figure 3. Like electromigration, the large jump in FIT value between 65nm(0.9V) and 65nm (1.0V) can be attributed to the exponential impact of temperature. However, this increase is smaller than the increase seen in electromigration due to the $|T - T_0|$ term in the stress migration equation (Equation 2). This term improves reliability with scaling, but its impact is overshadowed by the exponential relationship.

For the same reason, the range in FIT values across applications in stress migration is smaller than in electromigration. However, there still is an increase in the range with technology scaling. Similarly, the difference between application FIT values and worst case FIT values is also smaller than in electromigration.

5.3.3 TDDB scaling

As can be seen in Table 1, TDDB FIT value depends heavily on the values of V and t_{ox} used. There is also a more than exponential dependence on temperature. The negative effect of t_{ox} combined with temperature result in a decrease in TDDB reliability with scaling, despite the positive effect of voltage scaling. This is compounded by the non-ideal scaling of voltage.

As a result, these factors contribute to the huge increase in FIT value from 180nm to 65nm (1.0V) (667% on average for SpecFP and 812% for SpecInt). The increase from 180nm to 65nm (0.9V) is less severe, but still significant (106% for SpecFP and 127% for SpecInt).

Unlike the other failure mechanisms, the change in TDDB FIT values does not completely follow the change in temperature. This is because of the voltage dependence of TDDB. Hence, although the temperature increases from 180nm to 130nm, the drop in voltage between these two technology points reduces the FIT value. The beneficial impact of voltage is highlighted by the large difference between the FIT values at 65nm (0.9V) and 65nm (1.0V) (the difference is magnified further due to the temperature difference between the two points).

Like electromigration, the absolute FIT values of applications amplifies the range in FIT values and difference from the worst case value across all the applications.

5.3.4 TC scaling

There is a 52% increase between 180nm and 65nm (1.0V) and a 32% increase between 180nm and 65nm (0.9V) for SpecFP on average. The corresponding values for SpecInt are 66% and 36%.

Like stress migration, scaling impacts the FIT value of thermal cycling through an increase in temperature. However, unlike stress migration which has an exponential dependence on temperature, thermal cycling varies as the power of q, which is the Coffin-Manson exponent (as seen in Table 1). In our experiments, we used a value of 2.35 for q. Hence, although there is an increase in FIT value due to temperature with scaling, the increase is less steep than stress migration. The range in FIT values across applications is also smaller than that seen in stress migration. The difference between the worst case FIT values and application FIT values are also small.

5.3.5 Summary

In summary, all the failure mechanisms see an increase in FIT value with scaling. A large part of this is due to the detrimental impact of temperature scaling. TDDB shows the steepest increase due to its greater than exponential dependence on temperature and reduced gate oxide thickness. Electromigration follows closely, again primarily because of its exponential dependence on temperature and due to reduced interconnect dimension. Stress migration and thermal cycling are much less drastic.

6 Conclusions and Discussion

The long term reliability of microprocessors has been taken for granted up until the 0.18 micron CMOS technology timeframe. While the growth in on-chip integration levels is expected to remain close to the forecasts ordained by the venerable Moore's Law for the better part of this decade, the reliability implications of future scaling have not been quantified with rigor so far. Intuitively, due to the onset of a period of escalating power densities (and therefore, temperatures), the possible effect on hard failure rates is something that chip vendors and researchers alike are clearly worried about.

Early warnings in the form of increased incidences of server outages (caused by overheating) are already on the rise. Even temperature-independent effects of scaling, such as the likelihood of dielectric (oxide) breakdown as the thickness scales down to a few atoms, are issues that are high on the ”pain list” of future design teams. Worrying about hard failure rates and reliability is not something that early-stage microarchitecture definition teams engaged in during past designs; but the writing on the wall is clear: one cannot ignore the implications of future scaling even in early-stage design. Recent research in power-aware microarchitectures (and related modeling) has yielded many new ideas of managing power consumption and temperatures, without significant performance degradation. The next frontier in such technology-aware microarchitecture research, is to understand and then exploit the early-stage design tradeoffs between power, performance and reliability.

In this paper, we take the first step in building up the basic understanding (at the architect's level) of the reliability implications of scaling in the deep-submicron era. The model used, RAMP, which is the first of its kind, is easily integratable into early-stage, workload-driven power-performance simulators. By using RAMP in conjunction with a representative, current-generation superscalar processor simulator, we have been able to quantify the aforesaid concerns about reliability, in the context of actual workloads. The results reported, point to potentially large and sharp drops in long-term reliability, especially beyond 90 nm. Of the failure modes that were modeled, time-dependent dielectric breakdown (TDDB) appears to present the steepest challenge, while thermal cycling effects promise to remain largely within manageable bounds. Electromigration-related failures are also projected to be a major problem, while stress migration lies in between TDDB and electromigration. All of these results have been interpreted in some detail in the preceding text. In each case, our quantitative modeling also conclusively illustrates how increased scaling is increasing the difference between failure rates assuming worst-case conditions vs. typical operating conditions, as well as amplifying the differences among different applications. These results serve to illustrate the increasing importance of architecture-level workload-aware reliability management approaches.

The information obtained from this study is extremely valuable in understanding what failure modes to worry about most in future processor design. The challenge in future work, will be to invent solutions at the highest level of design specification, that will counter the projected reliability trend. It is true, of course, that in reality, progress will also be made at the lower levels (including in the underlying CMOS technology) to ensure that the effects of scaling do not turn out to be quite as drastic as would appear from the data presented in this paper. Yet, in our opinion, such work as presented in this paper, will spawn or hasten solutions at all levels of design abstraction: from microarchitecture down to physical design and layout.

7 Acknowledgments

The authors would like to thank Chao-Kun Hu of IBM T.J. Watson Research Center for his extensive help with the electromigration scaling model. They would also like to thank Barry Linder, also of IBM T.J. Watson Research Center and Ernest Wu of IBM Burlington for their help with the TDDB scaling model.

References

[1] Failure mechanisms and models for semiconductor devices. In *JEDEC Publication JEP122-A, Jedec Solid State Technolgy Association*, 2002.

[2] Critical Reliability Challenges for The International Technology Roadmap for Semiconductors. In *International Sematech Technology Transfer Document 03024377A-TR*, 2003.

[3] W. Abadeer et al. Key Measurements of Ultrathin Gate Dielectric Reliability and In-Line Monitoring. In *IBM Journal of Research and Development*, 1999.

[4] S. Borkar. Design challenges of technology scaling. In *IEEE MICRO*, 1999.

[5] P. Bose. Power-efficient microarchitectural choices at the early design stage. In *Keynote Address, Workshop on Power-Aware Computer Systems*, 2003.

[6] D. Brooks et al. Power-aware microarchitecture: Design and modeling challenges for the next-generation microprocessor. In *IEEE Micro*, 2000.

[7] C.-K.Hu et al. Scaling effect on electromigration in on-chip cu wiring. In *International Electron Devices Meeting*, 1999.

[8] D. Edelstein et al. A high performance liner for copper damascene interconnects. In *International Interconnect Technology Conference*, 2001.

[9] E.Eisenbraun et al. Integration of cvd w- and ta-based lines for copper metallization. In *MKS white paper, http://www.mksinst.com/techpap.html*, 2000.

[10] E.T.Ogawa et al. Leakage, breakdown, and tddb characteristics of porous low-k silica based interconnect materials. In *International Reliability Physics Symposium*, 2003.

[11] S. Heo, K. Barr, and K. Asanovic. Reducing power density through activity migration. In *Proc. of Intl. Symp. on Low Power Electronics Design*, 2003.

[12] V. Iyengar, L. H. Trevillyan, and P. Bose. Representative traces for processor models with infinite cache. In *Proc. of the 2nd Intl. Symp. on High-Perf. Comp. Architecture*, 1996.

[13] J.H.Stathis. Reliability limits for the gate insulator in cmos technology. In *IBM Journal of Research and Development*, 2002.

[14] N. P. Mencinger. A mechanism-based methodology for processor package reliability assessments. In *Intel Technology Journal*, Q3,2000.

[15] C. Moore. The power4 system microarchitecture. In *Microprocessor Forum*, 2000.

[16] M. Moudgill, P. Bose, and J. Moreno. Validation of turandot, a fast processor model for microarchitectural exploration. In *IEEE International Performance, Computing, and Communications Conference*, 1999.

[17] M. Moudgill, J. Wellman, and J. Moreno. Environment for powerpc microarchitectural exploration. In *IEEE Micro*, 1999.

[18] J. S. Neely et al. Cpam: A common power analysis methodology for high-performance vlsi design. In *9th Topical Meeting on the Electrical Performance of Electronic Packaging*, 2000.

[19] K. Skadron et al. Temperature-Aware Microarchitecture. In *Proc. of the 30th Annual Intl. Symp. on Comp. Architecture*, 2003.

[20] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers. The Case for Microarchitectural Awareness of Lifetime Reliability. In *UIUC CS Technical Report No. UIUCDCS-R-2003-2391, (Submitted for publication)*, 2003.

[21] K. Trivedi. Probability and statistics with reliability, queueing, and computer science applications.

[22] E. Y. Wu et al. Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate dioxides. In *Solid-state Electronics Journal*, 2002.