To appear in the proceedings of *The International Conference on Dependable Systems and Networks (DSN-04)*, June 2004.

The Impact of Technology Scaling on Lifetime Reliability [*]

Jayanth Srinivasan[†], Sarita V. Adve[†], Pradip Bose[‡], Jude A. Rivers[‡]
[†]Department of Computer Science, University of Illinois, Urbana-Champaign
[‡]IBM T.J. Watson Research Center, Yorktown Heights, NY
{srinivsn,sadve@cs.uiuc.edu},{pbose,jarivers@us.ibm.com}

Abstract

The relentless scaling of CMOS technology has provided a steady increase in processor performance for the past three 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 [15], 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, with SPEC benchmarks, we find the failure rate of a scaled 65nm processor to be 316% higher than a similarly pipelined 180nm processor; (2) time-dependent dielectric breakdown and electromigration have the largest increases; 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 increasingly difficult, and motivate a need for workload specific, microarchitectural lifetime reliability awareness at an early design stage.

1 Introduction

Advances in CMOS semiconductor technology have been steadily improving processor performance. These advances have been driven by aggressive scaling of device feature sizes. However, CMOS scaling is accelerating the onset of problems due to long-term processor hardware failures or lifetime reliability. This paper represents a first attempt at quantifying the impact of scaling on lifetime reliability of an entire processor, considering the behavior of the workload running on the processor. Our work focuses on *intrinsic hard failures*, and examines failures due to electromigration (EM), stress migration (SM), time-dependent dielectric (gate oxide) breakdown (TDDB), and thermal cycling (TC). We do not model extrinsic hard failures and soft errors because they generally do not impact lifetime reliability [15].

1.1 Scaling theory and practice

Device scaling results in the reduction of feature sizes and voltage levels of transistors. Under ideal scaling, gate delay decreases by 30% from one generation to the next, transistor density doubles, and dynamic power per transistor decreases by about 50% (assuming constant electric field scaling where voltage scales down by 30%) [3]. The net impact is that for the same die size, under ideal scaling, the chip dynamic power and power density remain unchanged. With real scaling in the deep sub-micron range, however, 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 ideally. This prevents the dynamic power per transistor from decreasing at the ideal rate. One reason for 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. Second, 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) make it ever harder to scale down the supply voltage. Hence, area scaling without appropriate power scaling results in higher

[*] This work is supported in part by an equipment donation from AMD Corp. and the National Science Foundation under Grant No. CCR-0209198, CCR-0205638, EIA-0224453, and CCR-0313286. A large part of the work was performed while Jayanth Srinivasan was a summer intern at IBM T. J. Watson Research Center.

power densities.

Total chip leakage power is increasing. Scaling down threshold voltages ideally causes the leakage current per transistor to increase by five times per technology generation. This increase is further compounded by the exponential dependence of leakage power on temperature.

1.2 Impact of non-ideal scaling

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 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 [8, 10, 17]. 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, 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 target workload. Thus, an application oblivious analysis of processor reliability would produce unrepresentative reliability data.

In recent work, we have proposed an industrial strength, microarchitecture level model and simulation methodology, called RAMP, to evaluate processor lifetime reliability for a workload, but for a given technology [15]. That work uses the workload dependence of lifetime reliability to motivate microarchitecture level mechanisms to address the growing lifetime reliability problem.

1.3 Our contributions

To the best of our knowledge, this paper represents 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. We enhance the RAMP reliability model by adding scaling specific parameters to enable lifetime reliability evaluation at different technologies. In particular, our evaluation 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 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. 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 gate oxide breakdown (TDDB) will provide the largest challenge followed by electromigration. The effect of scaling on stress migration and thermal cycling are much less drastic. 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 increasingly difficult.

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 steady state operation based on the highest temperature and activity factors reached by any of our applications. The difference between this worst-case failure rate and the highest actual failure rate seen by any single application went from 25% for 180nm to 90% for 65nm (computed as a percentage of the worst-case failure rate). The difference between worst-case and *average* failure rate for our applications was even more striking – 67% at 180nm to 206% at 65nm. 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 [15], is to perform reliability qualification for the expected case, backed up with dynamic application-specific responses for handling departures from the expected case.

2 Background

As mentioned in Section 1, we use a model and simulation methodology called RAMP, described in [15], to calculate lifetime reliability of processors from a microarchitectural viewpoint. RAMP represents the first microarchitecture-level methodology for evaluating processor lifetime reliability, and uses state-of-the-art analytic models for important intrinsic failure mechanisms. Its design and implementation are discussed in detail in [15]. It currently models four main intrinsic failure mechanisms experienced by processors – electromigration (EM), stress migration (SM), gate-oxide or time dependent dielectric breakdown (TDDB), and thermal cycling (TC) [1, 2]. It implements the failure models at a microarchitectural structure level (e.g., caches, ALUs, instruction window, etc.), for a *given technology generation*. The standard reliability metric used in the analytical models in RAMP is MTTF (mean time to failure), which is the average expected lifetime of the processor.

RAMP should be used in conjunction with a timing simulator to determine workload behavior, and a power and thermal simulator for power and temperature profiles. This is dicussed further in Section 4.

Next, we review the individual (structure level) failure models in RAMP, assuming steady state operation at a fixed operating point. We then review how RAMP combines the different failure models, across all chip structures, while accounting for temporal variations within an application.

Electromigration. This failure mechanism is well understood, and extensive research has been performed by the material science and semiconductor community on modeling and understanding its effects [1, 8]. Electromigration in processor interconnects is due to the mass transport of conductor metal atoms in the interconnects. Sites of metal atom depletion can lead to increased resistance and open circuits. At the site of metal atom pile up, extrusions can form causing shorts between adjacent metal lines.

The model for the MTTF due to electromigration [1, 8], $MTTF_{EM}$, used in RAMP [15] is:

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

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 (1.1 and 0.9 respectively for the copper interconnect modeled in RAMP [15]).

RAMP models MTTF 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.

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

The model for the MTTF due to stress migration [1], $MTTF_{SM}$, used in RAMP [15] is:

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

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 (2.5 and 0.9 respectively for the copper interconnects modeled in RAMP [15]). RAMP assumes that sputtering (versus vapor deposition) was used to deposit the interconnect metal and uses a value of 500K for T_0 [6].

Time-dependent dielectric breakdown. Also known as gate oxide breakdown, this 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 [10, 17]. The model for the MTTF due to TDDB used in RAMP [15] is based on recent experimental work performed by Wu et at. at IBM [17]:

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

where T is the absolute temperature in Kelvin, a, b, X, Y, and Z are fitting parameters, and V is the voltage [1].

Based on the experimental data collected by Wu et al. [17], the values used in RAMP for the TDDB model are $a = 78$, $b = -0.081$, $X = 0.759ev$, $Y = -66.8evK$, and $Z = -8.37e - 4ev/K$.

Thermal cycling. Permanent damage accumulates every time there is a cycle in temperature in the processor, eventually leading to failure. Fatigue due to thermal cycling is most pronounced in the package and die interface (e.g., at solder joints) [1]. 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. The model for the MTTF due to large thermal cycles is based on the Coffin-Manson equation [1] and is:

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

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.

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].

[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 the models. To calculate the overall MTTF of the processor, RAMP needs to combine the effects of the different failure mechanisms, across all chip structures, and over time. In general, this is difficult and requires knowledge of the lifetime distributions of the different failure mechanisms. RAMP addresses this problem by using the sum-of-failure-rates (SOFR) model [16]. The SOFR model, which is a standard model used in industry makes two assumptions: (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). 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. However, this assumption is often used in the industry for lack of better validated models. The above two assumptions imply [16]: (1) The MTTF of the processor, $MTTF_p$, is the reciprocal 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, $MTTF_p = \frac{1}{\lambda_p} = \frac{1}{\sum_{i=1}^{j}\sum_{l=1}^{k}\lambda_{il}}$ where λ_{il} is the failure rate of the i^{th} structure due to the l^{th} failure mechanism (which is the reciprocal of the corresponding MTTF).

Further, the MTTF models so far assume fixed operating conditions (in particular, fixed temperature, activity factor, and voltage). However, when an application runs, the temperature, activity factor, and voltage all vary with time. We assume that we can account for the impact of this variation by: (1) calculating an instantaneous λ_{il} based on instantaneous T, V, and p (measured over a reasonably small time granularity); and (2) using an average of these values to determine the actual failure rate when running the application (this averaging over time is similar to the assumption used in the SOFR model which averages over space).

The standard method of reporting constant failure rates for semiconductor components is in Failures in Time (FITs) [16], which is the number of failures seen per 10^9 device hours: $MTTF_p = \frac{1}{\lambda_p} = \frac{10^9}{FIT_p}$. We will use FITs as our metric when reporting results.

Finally, to calculate absolute FIT rates, 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, and yield. High values for the proportionality constants imply more reliable processors, which comes at a higher cost. Conversely, cheaper systems will have low values for the constants.

3 Impact of Scaling on Failure Mechanisms

This section explores the impact of scaling on the failure mechanisms discussed in Section 2. We examine the parameters that change for different technology generations and extend RAMP to incorporate their impact.

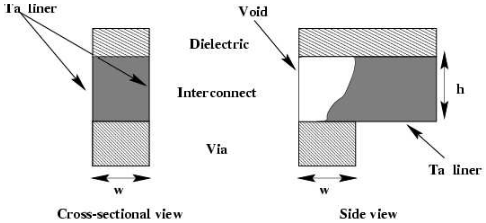


Figure 1. EM in copper interconnects.

Electromigration. The detrimental impact of increasing temperatures on electromigration due to scaling is already modeled in RAMP [15]. However, scaling also reduces interconnect dimensions which has a negative impact on electromigration.

Due to the need for low interconnect resistivity and high electromigration reliability, the semiconductor industry has recently shifted to using copper interconnects [8] (as against copper doped with aluminum). 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 cannot 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 [8]. 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, $\frac{\delta w}{wh} = \frac{\delta}{h}$ [8].

Electromigration voids are found to occur most commonly at the interface between the interconnects and the metal vias [8]. Electromigration failure is considered to have occurred when the void formed grows larger than the width of the via, w (which is the same as the interconnect width). Hence, mean time to failure due to electromigration, $MTTF_{EM}$, is proportional to the width of the via, w, and inversely proportional to the relative amount of flux passing through the interface region, $\frac{\delta}{h}$ [8]. Thus, when a scaling factor of κ is applied, 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 (in Equation 1) is equal to the product of the activity factor of the structure, p, and the maximum allowed interconnect current density for that technology generation. This maximum allowed current density changes with scaling. The values we use are given in Table 4, and justified in Section 4.6.

Stress migration. The main impact of scaling on stress migration 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.

There are indirect scaling effects on stress migration due to the use of new low-k dielectrics which tend to be porous and brittle [13]. However, 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.

Time-dependent dielectric breakdown. 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. Gate oxide reliability depends on other scaling parameters as described below.

First, 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}. I_{leak} increases by one order of magnitude for every 0.22nm reduction in gate oxide thickness [10]. 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. Second, for current and future range gate oxide thicknesses, $MTTF_{TDDB}$ is inversely proportional to the total gate oxide surface area [17].

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, 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:

$$\frac{MTTF_1}{MTTF_2} = 10^{\frac{(t_{ox1} - t_{ox2})}{0.22}} \times \frac{V_2^{(a-bT_2)}}{V_1^{(a-bT_1)}} \times \frac{A_1}{A_2} \times \frac{e^{\frac{(X + \frac{Y}{T_1} + ZT_1)}{kT_1}}}{e^{\frac{(X + \frac{Y}{T_2} + ZT_2)}{kT_2}}} \tag{5}$$

Failure Mech.	Major temperature dependence	Voltage dependence	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.

where X, Y, Z, a and b are empirically determined constants, described in Section 2.

Thermal cycling. Like stress migration, the main impact of scaling on thermal cycling modeled in RAMP is the impact of temperature. Scaling has no other direct impact on thermal cycling. There are indirect scaling effects on thermal cycling due to the use of new low-k dielectrics which have inferior adhesive properties [13], increasing susceptibility to thermal cycling failure. However, 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.

Summary of impact of different parameters. Table 1 summarizes the impact of different scaling related 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 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 [11]. Table 2 summarizes the base 180nm processor modeled. 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 [11], resulting in very low L2 intrinsic failure rates. The processor is modeled using a trace-driven research simulator called Turandot [12], developed at IBM T.J. Watson Research Center. As described in [12], 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 dynamic power dissipation, we use the PowerTimer toolset [5] developed at IBM T.J. Watson Research Center, which works in its default mode with Turandot. The

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 rate	8 per cycle
Retirement rate	1 dispatch-group (=5, max)
Functional units	2 Int, 2 FP, 2 Load-Store 1 Branch, 1 LCR
Integer FU latencies	1/7/35 add/multiply/divide
FP FU latencies	4 default, 12 divide
Reorder buffer size	150
Register file size	120 integer, 96 FP
Memory queue size	32 entries
Base Memory Hierarchy Parameters	
L1 D/L1 I/L2 unified	32KB/32KB/2MB
Base Contentionless Memory Latencies	
L1 D/L2/Main memory	2/20/102 cycles

Table 2. Base 180nm POWER4-like processor.

power models are based on circuit accurate power estimations from the 180nm POWER4 processor [11], and assume realistic clock gating.

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. We also model the impact of temperature on leakage power using the technique described in [7]. At a temperature T, the leakage power, $P_{leakage(T)} = P_{leakage(383K)} \times e^{\beta(T-383)}$, where β is a curve fitting constant. The value of β we use (0.017) is taken from [7].

4.3 Temperature simulation methodology

We use the HotSpot tool [14] 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), not including the L2 cache. HotSpot models temperature at a microarchitectural structure granularity. We combine the microarchitectural structures on the POWER4-like core into 7 distinct structures, and use HotSpot to produce temperature measurements at the granularity of 1μ second (using power information from PowerTimer).

As explained in [14], the RC time constant of the processor heat sink is significantly larger than the RC time constant of individual silicon structures. Hence, we cannot run our simulations long enough for the heat sink to reach its steady state temperature. Therefore, it is critical that HotSpot be initialized with the right heat sink temperature. For this purpose, we run all simulations twice. The first run is used to obtain average power consumption values for each structure on chip. These average values are then fed into the steady state temperature model to calculate a steady state heat sink temperature. This temperature is then used to initialize the second simulation run, which gives the correct temperature for the silicon structures.

For the heat sink thermal resistance at 180nm, we use 0.8 W/K [14]. As the processor is scaled, this resistance will increase as area shrinks. For comparing technology generations, we scale the heat sink thermal resistance such that a constant heat sink temperature is maintained for each application (different applications have different heat sink temperatures, which remain constant with scaling).

SpecFP app.	IPC	180nm power (W)	SpecInt app.	IPC	180nm power (W)
ammp	1.06	26.08	vpr	1.38	26.93
applu	1.17	26.94	bzip2	2.31	27.71
sixtrack	1.38	27.32	twolf	1.26	28.44
mgrid	1.71	27.78	gzip	1.85	28.69
mesa	1.75	29.21	perlbmk	2.25	30.59
facerec	1.79	29.60	gap	1.76	31.24
wupwise	1.66	30.50	gcc	1.24	31.73
apsi	1.64	30.65	crafty	2.25	31.95
Average	**1.52**	**28.51**	**Average**	**1.79**	**29.66**

Table 3. Average IPC and power consumption for the 180nm base processor for our workload.

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 instantaneous FIT values is maintained which provides the final FIT value of the structure. The sum of these will give the processor FIT value.

As mentioned previously (Section 2), 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 approach similar to [15] as follows. Current processors are expected to have an MTTF of around 30 years [1] – this implies that the total FIT value of the processor should be around 4000 ($\frac{10^9}{30years}$). 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. This gives the proportionality constants that can be fed back into RAMP to get the failure rate at other technology points.

4.5 Workload description

We report experimental results for PowerPC traces of 16 SPEC2K benchmarks (8 SpecInt + 8 SpecFP). 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 [9]. Table 3 summarizes the benchmarks studied, including the IPC and average power (dynamic + leakage) consumption. As can be seen, for our processor, SpecInt has a higher average IPC and marginally higher power consumption than SpecFP.

Tech gen nm	V_{dd} V	Frequency GHz	Relative Capacitance	Relative Area	t_{ox} $\mathring{A}$	Interconnect cur density $\frac{mA}{m^2}$	Leakage power $\frac{W}{mm^2}$	Average Total Power (Dynamic+Leakage) (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. Scaled parameters used.

4.6 Scaling methodology

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.

Voltage and frequency scaling:

With ideal scaling of 0.7, [3] states that best-case frequency scaling per generation would be about 43%. However, while 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. The supply voltage values in Table 4 are carefully chosen to match up with the scaled frequencies, while also satisfying the leakage power density assumptions. 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. The two different technology points are represented as 65nm (0.9V) and 65nm (1.0V) in our results.

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.

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.

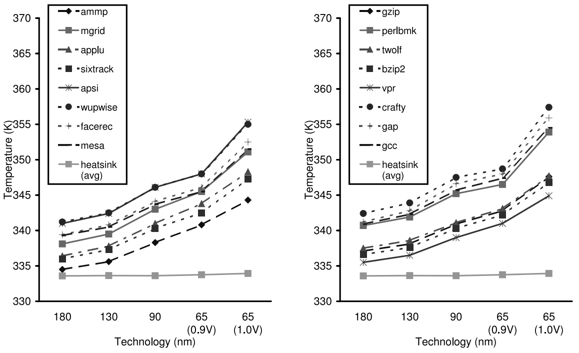
t_{ox} **scaling:** The values of t_{ox} used were obtained from the high performance logic parameters in the ITRS roadmap [2]. As can be seen, changes in t_{ox} are proportional to the scaling factor.

Interconnect current density: In order to compensate for decreasing interconnect dimensions and the resultant impact on electromigration reliability, designers have been reducing interconnect current density every technology generation. We assume a 33% reduction in interconnect current density every technology generation [3], until 90nm. Beyond this point, it is expected that interconnect current density can not be reduced further.

Power scaling: The leakage power densities used for each technology point assume aggressive leakage control techniques [4]. Table 4 also gives the total power consumption (dynamic + leakage power), based on simulations, and the relative total power density (which is the ratio of the total power consumption and area), averaged across all applications. Up to 90nm, scaling reduces the total power consumption of the core. However, the average power density goes up steadily with scaling (due to non-ideal voltage reduction and increases in leakage power).

5 Results

5.1 Temperature analysis



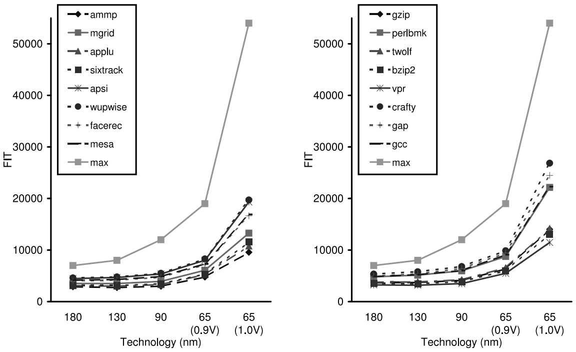
(a) SpecFP temperatures (b) SpecInt temperatures

Figure 2. Maximum temperature reached by any structure. The heat sink temperature, averaged across all applications, is also shown.

We start by presenting results for temperature since they affect reliability so significantly. Figure 2 shows the maximum temperature reached by any 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 this temperature remains constant with scaling). As can be seen, while the heat sink temperature remains nearly constant with scaling, the temperature of the hottest structure increases. On average, from 180nm to 65nm (1.0V), the temperature of the hottest structure on chip increased by 15 degrees Kelvin. Application temperatures increase because the power density on chip (as seen in Table 4) is increasing with scaling.

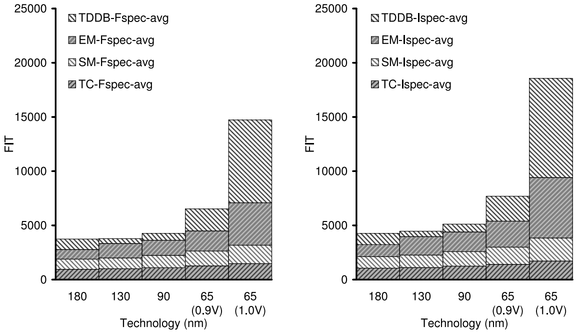
The results also show that there is a significant range in temperatures across applications. There is high correlation between application power and temperature and some correlation with IPC. The hottest applications (wupwise and apsi for SpecFP and crafty for SpecInt) in Figure 2 also have the highest 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).

5.2 Total FIT value scaling



(a) Individual SpecFP (b) Individual SpecInt

Figure 3. Total processor FIT value for each application.



(a) Average SpecFP (b) Average SpecInt

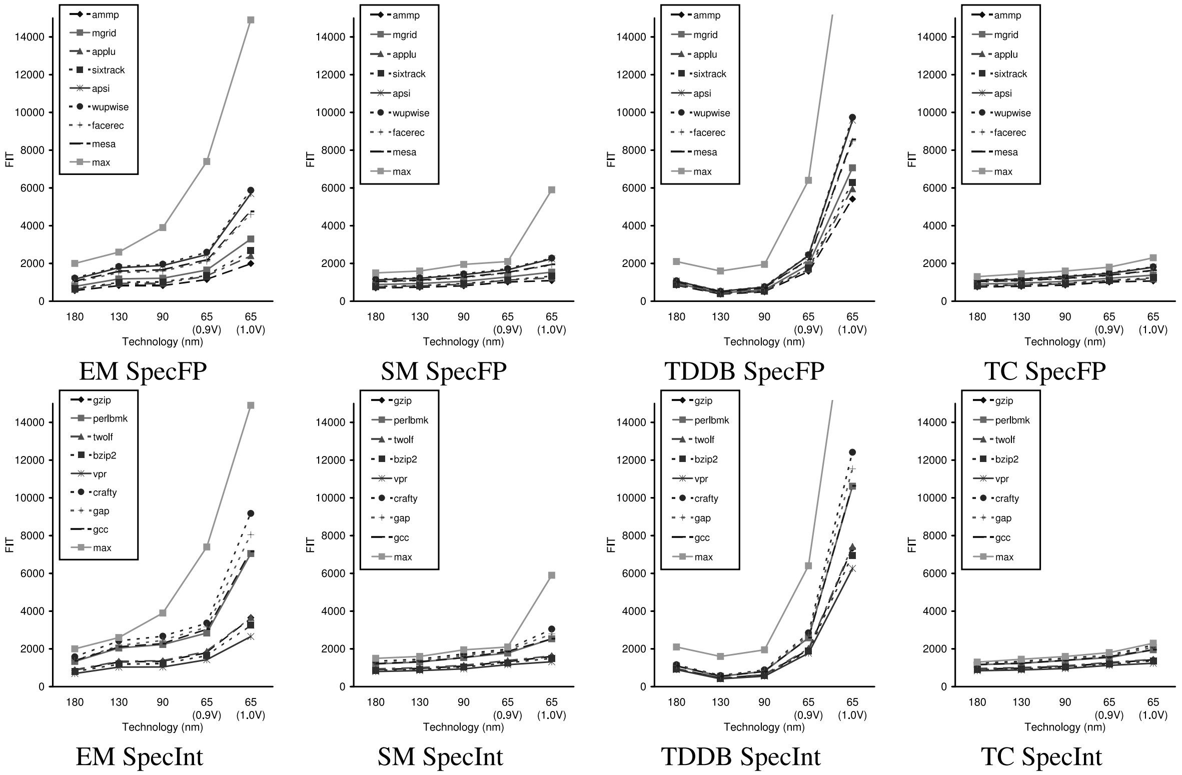
Figure 4. FIT value averaged across (SpecInt or SpecFP) apps, and the relative contribution of each mechanism.

Figures 3 and 4 present the data for this section. Figures 3 (a) and (b) show the scaling behavior of the total processor FIT value for each application, for SpecFP and SpecInt respectively. They also show FIT values calculated based on worst-case conditions (labeled as max) over all the applications. To compute these worst-case 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. Figures 4 (a) and (b) show the FIT value averaged across all the applications, with scaling (for SpecFP and SpecInt respectively). At each technology generation, each FIT bar has also been broken down into the individual contributions by each failure mechanism, which will be discussed in Section 5.3.

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 increases by 274% from 180nm to 65nm (1.0V). The increase seen in SpecInt was larger at 357%. Also, at each scaled 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. There is a significant difference in FIT value from 65nm (0.9V) to 65nm (1.0V). As discussed in Section 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.

Workload dependence of FIT value: In Figure 3, 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, compared to the application with the highest FIT value, the worst-case FIT value is 25% higher for 180nm and 90% higher at 65nm (computed as a percentage of the highest FIT seen by any application). 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, Figure 3 also show that there is a large range in FIT values across applications, and this range increases with scaling. 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). However, 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). Thus, the range in FIT values across applications also increases with scaling. The range across all applications (SpecFP + SpecInt) increases from 2479 FIT (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



EM SpecFP SM SpecFP TDDB SpecFP TC SpecFP

EM SpecInt SM SpecInt TDDB SpecInt TC SpecInt

Figure 5. Failure rates for each failure mechanism for SpecFP and SpecInt. The worst-case FIT value curve (labeled max) for each mechanism is also shown.

(which is 104% of the average FIT value) at 65nm (1.0V).

Our results indicate that future reliability qualification mechanisms should be application-aware; otherwise, the processor could be severely over-designed for most applications. Our previous work proposed dynamic reliability management as an application-aware reliability approach [15]; the quantification here unequivocally shows the increasing importance of such approaches.

5.3 Individual failure mechanisms

Next we examine scaling behavior of individual failure mechanisms, illustrated in Figure 5. The figure also provides data on the worst-case operating conditions (as seen during an application run, and measured as in Section 5.2).

EM scaling: Scaling has a significant impact on electromigration failure rate – going from 180nm to 65nm (1.0V), the failure rate increases by 303% on average for SpecFP and 447% on average for SpecInt. Going from 180nm to 65nm (0.9V), the increase is 97% for SpecFP and 128% for SpecInt. 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 for total FIT values, there is a large range in FIT values across applications for electromigration and also a large difference between worst-case and application FIT values.

SM scaling: For SM, there is a 76% increase in FIT values going from 180nm to 65nm (1.0V) and a 43% increase going from 180nm to 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 5. Like electromigration, the large jump in FIT value between 65nm(0.9V) and 65nm (1.0V) is entirely due to the exponential impact of temperature. However, this increase is smaller than the increase seen in electromigration due to the $|T - T_0|^{-m}$ term in the stress migration equation (Equation 2). This term improves reliability with scaling, but its impact is overshadowed by the exponential relationship. There is a large range in FIT values across applications for stress migration and also a large difference between worst-case and application FIT values. However, the magnitude of these differences is less than that seen in electromigration.

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 results in an overall 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).

TC scaling: There is a 52% increase in TC FIT values going from 180nm to 65nm (1.0V) and a 32% increase going from 180nm to 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 is also small.

6 Conclusions

Advances in CMOS semiconductor technology, driven by aggressive device scaling, have been steadily improving processor performance. However, CMOS scaling is resulting in escalated power densities and processor temperatures, and accelerating the onset of problems due to long-term processor hardware failures or lifetime reliability.

In this paper, we take a first step in establishing the basic understanding (at the architect's level) of the reliability implications of scaling in the deep-submicron era. Our results 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) and electromigration appear to present the steepest challenge. Our results also illustrate how 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.

Our results present two broad implications. First, it will become increasingly difficult to leverage a single microarchitectural design for multiple remaps across a few technology generations. Second, the need for workload specific, microarchitectural lifetime reliability awareness is illustrated.

7 Acknowledgments

We would like to thank Chao-Kun Hu of IBM for his help with the electromigration scaling model, and Barry Linder and Ernest Wu, also of IBM, for their help with the TDDB scaling model.

References

[1] Failure Mechanisms and Models for Semiconductor Devices. In *JEDEC Publication JEP122-A*, 2002.

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

[3] S. Borkar. Design Challenges of Technology Scaling. In *IEEE MICRO*, Jul-Aug 1999.

[4] P. Bose. Power-Efficient Microarchitectural Choices at the Early Design Stage. In *Keynote Address, Workshop on Power-Aware Computer Systems*, 2003.

[5] D. Brooks et al. Power-aware Microarchitecture: Design and Modeling Challenges for the next-generation microprocessor. In *IEEE Micro*, 2000.

[6] 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.

[7] S. Heo et al. Reducing Power Density Through Activity Migration. In *Intl. Symp. on Low Power Elec. Design*, 2003.

[8] C. K. Hu et al. Scaling Effect on Electromigration in On-Chip Cu Wiring. In *International Electron Devices Meeting*, 1999.

[9] 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.

[10] J.H.Stathis. Reliability Limits for the Gate Insulator in CMOS Technology. In *IBM Journal of R&D, Vol. 46*, 2002.

[11] C. Moore. The POWER4 System Microarchitecture. In *Microprocessor Forum*, 2000.

[12] M. Moudgill et al. Validation of turandot, a fast processor model for microarchitectural exploration. In *IEEE Intl Perf., Computing, and Communications Conf.*, 1999.

[13] E. T. Ogawa et al. Leakage, Breakdown, and TDDB Characteristics of porous low-k silica based interconnect materials. In *International Reliability Physics Symposium*, 2003.

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

[15] J. Srinivasan et al. The Case for Microarchitectural Awareness of Lifetime Reliability. In *Proc. of the 31st Annual Intl. Symp. on Comp. Arch.*, 2004.

[16] K. Trivedi. Probability and Statistics with Reliability, Queueing, and Computer Science Applications. Prentice Hall, 1982.

[17] 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.