Banking Chip Lifetime: Opportunities and Implementation

Zhijian Lu, John Lach, Mircea Stan, Kevin Skadron[§]
Department of Electrical and Computer Engineering, [§]Department of Computer Science
University of Virginia
Charlottesville, VA 22904
{zl4j, jlach, mircea}@virginia.edu, skadron@cs.virginia.edu

Abstract

Most existing integrated circuit reliability models assume a uniform, typically worst-case, operating temperature, but temporal and spatial temperature variations affect expected device lifetime. As a result, design decisions and dynamic thermal management techniques using worst-case models are pessimistic and result in excessive design margins and unnecessary runtime engagement (and associated performance penalties) of cooling mechanisms. By leveraging a reliability model that accounts for temperature gradients (dramatically improving interconnect lifetime prediction accuracy) and modeling expected lifetime as a resource that is consumed over time at a temperature- and voltage-dependent rate, substantial design margin can be reclaimed and runtime penalties avoided while meeting expected lifetime requirements. In this paper, we evaluate the potential benefits and implementations of this technique by tracking the expected lifetime of a system under different workloads while accounting for the impact of dynamic voltage and temperature variations.

1. INTRODUCTION

The advance of technology scaling (and the resulting increases in power density) has made thermal-related reliability one of the major concerns in modern IC design. For example, in the deep sub-micron (DSM) region, electromigration, which is temperature dependent, is widely regarded as one of the dominant failure mechanisms. Designers must therefore rely on temperature-dependent reliability models to derive the expected lifetime of their circuits, increasing design margin (e.g., wire width) as necessary to meet lifetime requirements. Traditionally, a worst-case temperature is used to evaluate the reliability of the system, which may often result in excessive design margins.

To further address thermal reliability issues, dynamic thermal management (DTM) techniques have been developed. However, these techniques also rely on worst-case assumptions, typically using a fixed temperature threshold to engage cooling mechanisms, such as frequency/voltage scaling and throttling at the expense of degraded performance. Under such pessimistic assumptions, cooling mechanisms may often be engaged (and performance penalties incurred) unnecessarily.

In this paper, we leverage an accurate reliability model we developed that takes temperature variations into account [4] to dynamically track the "consumption" of chip lifetime during operation. Generally, when temperature increases, lifetime is being consumed more rapidly, and vice versa. Therefore, if temperature is below the traditional DTM engagement threshold for an extended period, it may be acceptable to let the threshold be exceeded for a time while still maintaining the required expected lifetime. In effect, lifetime is modeled as a resource that is being "banked" during periods of low temperature, allowing for future withdrawals to maintain performance during times of higher operating temperatures. Using electromigration as an example, we show the benefits of lifetime banking by avoiding unnecessary DTM engagements while meeting expected lifetime requirements.

This paper is structured as follows. Section 2 presents an electromigration model subject to dynamic stress. Using this model, we propose a lifetime banking based dynamic reliability management method in Section 3. We implemented this method in a compact architecture-level thermal model, *Hotspot* [6], running the Spec2000 benchmarks, with the results shown in Section 4 . Finally, we conclude the paper in Section 5 .

2. DYNAMIC ELECTROMIGRATION MODEL

In this section, we briefly introduce our dynamic electromigration (EM) model and explain how we can model the expected lifetime as a resource that is consumed over time. The key to the model is to update the projected lifetime according to the actual dynamic temperature information observed during program execution. Applying this model to DTM, instead of using a fixed temperature threshold, the cooling mechanisms are engaged only when the projected lifetime falls below the required lifetime.

Black's equation is widely used to predict mean time to failure (MTF) due to electromigration:

$$T_f = \frac{A(kT)}{j^n} exp\left(\frac{Q}{kT}\right) \tag{1}$$

where T_f is the time to failure, A is a constant based on the interconnect geometry and material, j is the current density, Q is the activation energy (e.g., $1.0eV$ for copper), and kT is the thermal energy. The current exponent, n, has different values according to the actual failure mechanism. It is assumed that $n = 2$ for void nucleation limited failure and $n = 1$ for void growth limited failure [5]. However, Black's equation is suitable only for interconnects subject to constant temperature and current density.

In our previous work [4], we derived a model to predict interconnect lifetime due to electromigration under simultaneous dynamic thermal and current stresses. We found that Black's equation is still valid, but one should use reliability-equivalent temperature $T_{equivalent}$ and current density $j_{equivalent}$ as defined in the following:

$$j_{equivalent} = \frac{E\left[j(t)\left(\frac{exp(\frac{-Q}{kT(t)})}{kT(t)}\right)\right]}{E\left[\frac{exp(\frac{-Q}{kT(t)})}{kT(t)}\right]} \qquad (2)$$

$$\frac{exp\left(\frac{-Q}{kT_{equivalent}}\right)}{kT_{equivalent}} = E\left[\frac{exp(\frac{-Q}{kT(t)})}{kT(t)}\right] \qquad (3)$$

where $E[\,]$ is the expected-value function, $j(t)$ and $T(t)$ are time-dependent current density and temperature functions, respectively. Substitute into Black's equation (1) the above two expressions for reliability-equivalent current density/temperature and using $n = 1$ for dual-damascene copper interconnect (widely used in modern chip manufacturing) [3], we derive the MTF under time-varying current density and temperature stresses:

$$\begin{aligned} T_f &= \frac{A(kT_{equivalent})}{j_{equivalent}} exp\left(\frac{Q}{kT_{equivalent}}\right) \\ &= \frac{A}{E\left[j(t)\left(\frac{exp(\frac{-Q}{kT(t)})}{kT(t)}\right)\right]} \end{aligned}$$

Or equivalently, by eliminating the expected-value function, we can express the MTF in an integral form:

$$\int_0^{t_f} j(t)\left(\frac{exp(\frac{-Q}{kT(t)})}{kT(t)}\right) dt = C \qquad (4)$$

where C is a constant determined by the structure of the interconnect.

Equation (4) models interconnect time to failure (i.e., interconnect lifetime) as a resource consumed by the system over time. Function $r(t) = \left[j(t)\left(\frac{exp(\frac{-Q}{kT(t)})}{kT(t)}\right)\right]$ can be regarded as the consumption rate. In DSM copper technology, void growth failure is the major EM induced failure mechanism [3], and $r(t)$ represents the void growth rate in this case. Due to the temporal behaviors of system workloads, interconnect current density and temperature are time-dependent. Equation (4) provides a model to capture the effect of transient behaviors on system lifetime.

3. DYNAMIC RELIABILITY MANAGEMENT BASED ON LIFETIME BANKING

Recently, many DTM techniques [6, 7] have been proposed to ensure that a chip will never operate above some temperature threshold. However, these techniques do not explicitly study the effects of transient behaviors on system reliability, and instead implement a temperature upper-bound at the expense of degraded performance. By modeling lifetime as a resource to be consumed over time, we can manipulate chip lifetime directly at runtime. In this section, we present a simple dynamic reliability management (DRM) scheme built on conventional DTM techniques. Other runtime reliability management schemes based on Equation (4) will be investigated in future work.

Reliability-aware runtime management

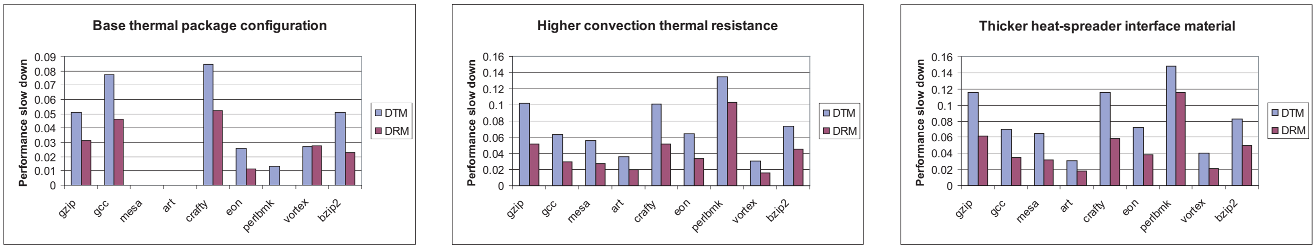
When a chip is designed, usually an expected lifetime (e.g., 10 years) is specified under some operating conditions (e.g., temperature, current density, etc.). We use $r_{nominal}$ to denote the lifetime consumption rate under the nominal conditions. During runtime, we monitor the actual operating conditions regularly, calculate the actual lifetime consumption rate $r(t)$ at that time instance, and compare the actual rate with the nominal rate $r_{nominal}$ by calculating $\int(r_{nominal} - r(t))dt$, which we call the "lifetime banking deposit". When $r(t) < r_{nominal}$, the chip is consuming its lifetime slower than the nominal rate. Thus, the chip's lifetime deposit is increased. When $r(t) > r_{nominal}$, the chip is consuming its lifetime faster than the nominal, and the lifetime banking deposit will be reduced. According to Equation (4), as long as the lifetime deposit is positive, the expected lifetime will not be shorter than that under the nominal consumption rate $r_{nominal}$.

The difference between conventional DTM and our DRM lies in the case when the chip's instantaneous consumption rate is larger than its nominal rate. In DTM, the lifetime consumption rate is never allowed to be larger than the nominal. In DRM, before we engage thermal management mechanisms we first check to see if the chip currently has a positive lifetime balance. If enough lifetime has been banked, the system can afford to run at a lifetime consumption rate larger than the nominal rate. Otherwise, we apply some DTM mechanism to lower the consumption rate, thus preventing a negative lifetime balance. In this study, we use dynamic voltage/frequency scaling as the major DTM mechanism. DRM need only monitor the actual lifetime consumption rate and to update the lifetime banking deposit. The computation overhead is negligible compared to that of DTM.

Current density scaling

Our dynamic reliability model (4) assumes that we can monitor both temperature and current density at runtime. For simplicity, we can use the maximum temperature measured across the chip to calculate the consumption rate. However, the variability of current density across the chip makes it much harder to track in real-time. Thus, we use the worst-case current density specified at design time in our calculations.

As mentioned above, in this study, we apply dynamic voltage/frequency scaling to prevent negative lifetime balances.



(a) (b) (c)

Figure 1. Performance comparison of traditional DTM and the proposed DRM. (a) Using base thermal configuration of the chip. (b) Using higher convection resistance. (c) Using thicker heat spreader interface material.

Thus, the worst-case current density should be scaled according to the voltage/frequency setting used. The relationship between current density, supply voltage and clock frequency can be modeled by transferred charges per clock cycle [1]:

$$j \propto \frac{CV}{T} = CVf$$

where C is the effective capacitance. Therefore, the worst-case current density can be scaled by the product of voltage and frequency.

4. EXPERIMENTS AND ANALYSIS

In this section, we show the simulation results for both single- and multi-program workloads using the dynamic reliability management technique presented here. We compare these results with those obtained using traditional threshold-based DTM.

Experimental set-up

We ran a set of programs from the Spec2000 benchmark suite on a processor simulator (SimpleScalar [2]) with the characteristics similar to a 0.13μm Alpha 21364. We simulated each program for a length of 5 billion instructions. We recorded both dynamic and static (leakage) power traces, which were then fed as inputs to the chip-level compact thermal model *Hotspot* [6] for trace-driven simulation. We used 110°C as the temperature threshold for both DTM and DRM. We implemented both DTM and DRM in the Hotspot model. Both schemes used a feed-back controlled dynamic voltage/frequency scaling mechanism to guard the program execution. For example, in DTM, when the actual temperature is above a certain temperature threshold, a controller is used to scale down the frequency/voltage, ensuring the program will never run at a temperature higher than 110°C. On the other hand, our DRM scheme use 110°C as the nominal temperature for the nominal lifetime consumption rate. When DRM observes that the lifetime banking deposit is close to zero, it invokes the DTM policy to guarantee that the MTF will not be shorter than that under the nominal consumption rate. Our DRM scheme converges to threshold-based DTM when the program never runs at less than the nominal temperature and the DTM policy is always engaged. In our trace driven simulations, we included the penalty due to frequency/voltage switching, which is about 10us in many real systems [6]. Furthermore, since leakage power is strongly dependent on temperature, we scaled the leakage power trace according to the actual temperature during the trace-driven simulation. Finally, we recorded the simulated times for fixed length power traces as the system performances under the two runtime management techniques.

Single-program workload

Figure 1 shows the performance penalty for both DTM and DRM with different thermal package set-ups. Only those benchmarks subject to performance penalties due to runtime management are shown here. As clearly shown in the figure, performance penalty with our DRM scheme is less than that with the DTM scheme in most benchmarks. Figure 2 replots the results from Figure 1 using the average value. Also shown in Figure 2 are the standard deviations across the benchmarks represented by the error bars. On average, our DRM technique reduces the performance penalty by about 40% of that due to DTM.

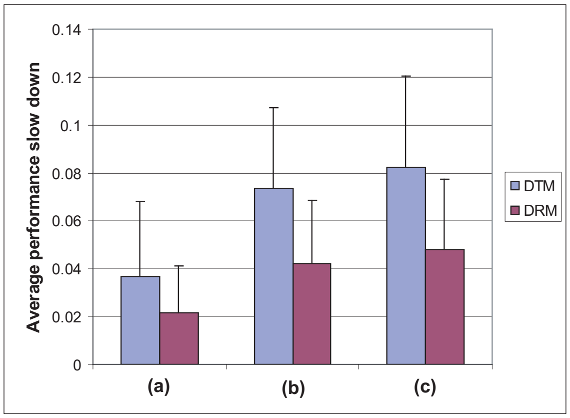


Figure 2. Average performance comparison of DTM and DRM with different thermal package set-ups. (a) Using base thermal configuration of the chip. (b) Using higher convection resistance. (c) Using thicker heat spreader interface material.

Using our DRM scheme, we can explicitly trade-off reliability with performance by targeting different lifetimes. As shown in Figure 3, DRM can reduce performance penalty at the expense of reduced expected lifetime.

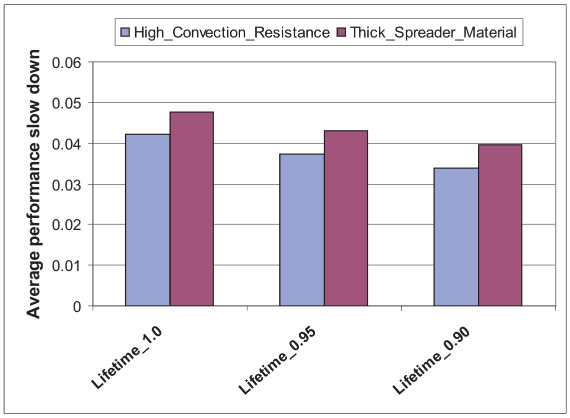


Figure 3. DRM performance at different targeted lifetimes.

When compared with the conventional threshold-based DTM, a distinct feature of DRM is its ability to "remember" the effects of previous behaviors. If the lifetime balance is high due to previous deposits, DRM will be more tolerant to higher operating temperatures for longer time intervals, thus reducing the performance penalties due to conventional DTM slow down mechanisms. In summary, the advantage of DRM over DTM is largely dependent on the temperature profile of the workload. Larger temperature variations result in larger savings due to DRM.

Multi-program workload

Another interesting program execution scenario is multiple workloads with context-switching. When a hot benchmark and a cold benchmark are executed together, the average operating temperature should be between the individual benchmarks' operating temperatures. Figure 4 shows the temperature profiles of such two benchmarks, and Figure 5 illustrates the effects of different context-switch time intervals on temperature fluctuation.

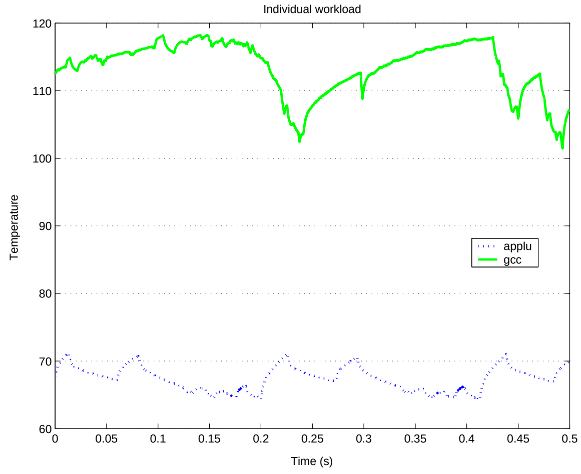


Figure 4. Temperature profiles of a hot benchmark and a cold benchmark.

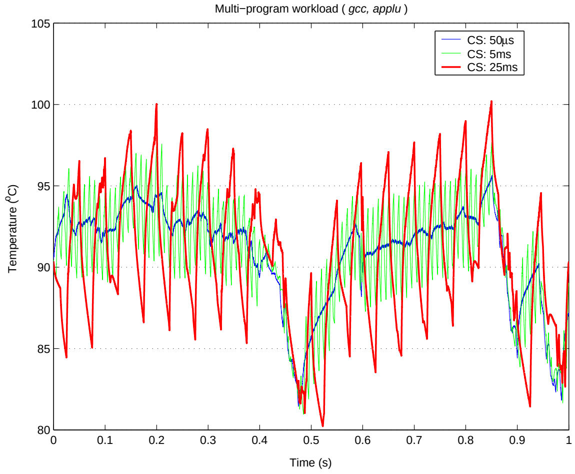


Figure 5. Temperature profiles of a multi-program workload with different context-switch intervals.

As we expect, the smaller the context-switch interval, the less temperature fluctuation, with the thermal package of the chip working as if a low-pass filter. When the context-switch interval is increased, individual benchmarks can show their hot/cold properties, and the temperature variation in the workload becomes obvious. In order to investigate how multi-program workloads affect the performance of DTM and DRM, we reduced the temperature threshold of the targeted lifetime from 110°C to 90°C. Figure 6 shows the performance penalties of DTM and DRM for this multi-program workload with different context-switch intervals. DRM outperforms DTM with the same thermal package configurations. As a comparison, the performance of DTM with a more expensive thermal package (three-fold smaller convection thermal resistance) is also shown in the figure. As the context-switch interval increases, the performance of DRM becomes closer to that of DTM with a more expensive thermal package.

5. CONCLUSION

Operating temperature variations have a major impact on the expected lifetime of an integrated circuit. By taking such variations into account, we can model lifetime as a resource that is consumed over time at a temperature- and voltage-dependent rate. Future work will show how this model can be used to reclaim the design margin that is introduced when existing worst-case temperature reliability models are used.

In this paper, we detailed the use of the temperature variability and lifetime resource models to develop a novel DRM technique that reduces the performance penalties associ-

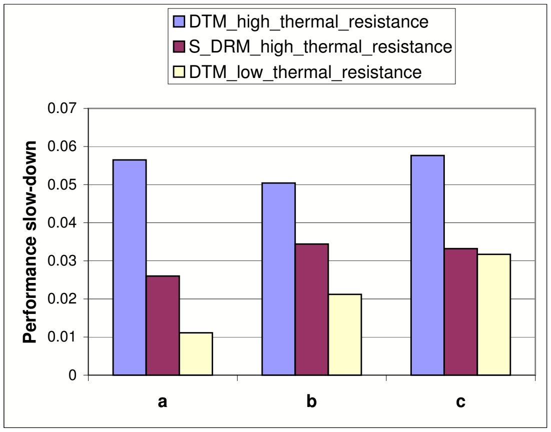


Figure 6. Performance comparison of DTM and DRM on a multi-program workload with different context-switch intervals (50μs, 5ms, and 25ms).

ated with existing DTM techniques while maintaining the required expected chip lifetime. When the operating temperature is below a nominal temperature (i.e., the threshold temperature used in DTM techniques), lifetime is being consumed at a slower than nominal rate, effectively banking lifetime for future consumption. A positive lifetime balance allows the nominal temperature to be exceeded for some time (thus consuming lifetime at a faster than nominal rate) instead of automatically engaging DTM and unnecessarily suffering the associated penalties.

Results, using interconnect electromigration as the temperature-dependent failure mechanism, revealed that DRM provides performance improvements over traditional threshold-based DTM without sacrificing expected lifetime. In addition, we showed the relationship between performance and expected lifetime, revealing how one can be traded off for the other, thus providing another design optimization and runtime management dimension. Future work will include evaluating the effectiveness of DRM on more benchmark suites and incorporating other thermal related failure mechanisms into the DRM framework. In addition to using target lifetime instead of a fixed temperature as the setpoint for DTM, lifetime banking can be used in other ways, like reliability-aware overvolting and overclocking for improved performance. Further research in these directions will also be conducted.

ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation under grant Nos. CCR-0105626, CCR-0133634, and EHS-0410526, and grants from Intel MRL and the Southeastern Center for Electrical Engineering Education. We would also like to thank the anonymous reviewers for their helpful comments.

REFERENCES

[1] Blaauw, D. T., Oh, C., Zolotov, V., and Dasgupta, A. Static electromigration analysis for on-chip signal interconnects. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems 22*, 1 (January 2003), 39–48.

[2] Burger, D., and Austin, T. M. The simplescalar tool set, version 2.0. *Computer Architecture News 25*, 3 (June 1997), 13–25.

[3] Davis, J. A., and Meindl, J. D., Eds. *Interconnect Technology and Design for Gigascale Integration*. Kluwer, 2003.

[4] Lu, Z., Huang, W., Lach, J., Stan, M., and Skadron, K. Interconnect lifetime prediction under dynamic stress for reliability-aware design. In *Proc. of International Conference on Computer Aided Design* (November 2004), pp. 327–334.

[5] Park, Y.-J., Andleigh, V. K., and Thompson, C. V. Simulations of stress evolution and the current density scaling of electromigration-induced failure times in pure and alloyed interconnects. *J. Appl. Phys. 85*, 7 (April 1999), 3546–3555.

[6] Skadron, K., Stan, M. R., Huang, W., Velusamy, S., Sankaranarayanan, K., and Tarjan, D. Temperature-aware microarchitecture. In *Proc. of the 30th International Symposium on Computer Architecture* (June 2003), pp. 2–13.

[7] Srinivasan, J., Adve, S. V., Bose, P., and Rivers, J. A. The case for lifetime reliability-aware microprocessors. In *Proc. of the 31st International Symposium on Computer Architecture* (June 2004), pp. 276–287.