Electromigration for Microarchitects

JAUME ABELLA and XAVIER VERA

Intel Barcelona Research Center, Intel Labs – UPC

Degradation of devices has become a major issue for processor design due to continuous device shrinkage and current density increase. Transistors and wires suffer high stress, and failures may appear in the field. In particular, wires degrade mainly due to electromigration when driving current. Techniques to mitigate electromigration to some extent have been proposed from the circuit point of view, but much effort is still required from the microarchitecture side to enable wire scaling in future technologies.

This survey brings to the microarchitecture community a comprehensive study of the causes and implications of electromigration in digital circuits and describes the challenges that must be faced to mitigate electromigration by means of microarchitectural solutions.

Categories and Subject Descriptors: B.4.5 [**Input/Output and Data Communications**]: Reliability, Testing, and Fault-Tolerance—*Hardware reliability*; B.4.3 [**Input/Output and Data Communications**]: Interconnections (subsystems)—*Physical structures*; B.7.1 [**Integrated Circuits**]: Types and Design Styles—*Advanced technologies; VLSI*

General Terms: Reliability, Design

Additional Key Words and Phrases: Electromigration, degradation, wires, buses

ACM Reference Format:

Abella, J. and Vera, X. 2010. Electromigration for microarchitects. ACM Comput. Surv. 42, 2, Article 9 (February 2010), 18 pages.
DOI = 10.1145/1667062.1667066, http://doi.acm.org/10.1145/1667062.1667066

1. INTRODUCTION

Reliability is a key issue in microprocessor design because processors must provide a given performance level during a minimum time period (a product's lifetime). The current reliability level has required a huge effort and investment to achieve designs where hundreds of millions of devices are guaranteed to work during several years with a very low failure rate. Typical failure rates are in the order of tens to hundreds failures in time (a failure in time, or FIT for short, corresponds to 1 failure every 10^9 hours of operation), and only a very small fraction of those FITs is devoted to unrecoverable faults due to degradation and defects.

This work has been supported in part by the Spanish Ministry of Education and Science under grants TIN2004-03702 and TIN2007-61763, FEDER funds of the EU, and the Generalitat de Catalunya under grant 2005SGR00950.

Authors' addresses: J. Abella and X. Vera, Intel Barcelona Research Center, Intel Labs – UPC, Jordi Girona St., 29, 3rd floor, Barcelona, Spain 08034; email: jaume.abella@intel.com.

DOI 10.1145/1667062.1667066 http://doi.acm.org/10.1145/1667062.1667066

While technological evolution drives towards smaller devices (transistors and wires), the supply voltage does not scale at the same pace, leading to higher current densities and temperatures. The increased current density and temperature accelerate transistor and wire degradation and shorten the lifetime of the product (the number of FITs increases).

The main sources of degradation [Toshiba Corporation 2002; Srinivasan et al. 2004a, 2004b] affecting microprocessors are electromigration and stress migration for wires, time-dependent dielectric breakdown, hot carrier injection and negative bias temperature instability for transistors, and thermal cycling for the package and pins. This work focuses on electromigration (EM), which is the most important source of failure for wires.

EM is an undesirable consequence of driving current through wires, which causes some metal atoms to move in the direction of current in such a way that some parts of the wire become narrower and may eventually break, while in other parts of the wire, metal piles up and may eventually create shorts with neighboring wires. How fast wires degrade due to EM depends on the current density of wires and the wire geometry.

EM occurs in power/ground grids at coarse-grain (e.g., global power lines) and fine-grain (e.g., local power lines in standard cells) and clock lines. Bidirectional wires such as data buses between caches and buses between cores in a multicore processor also degrade due to EM because they may charge and discharge through different ends of the wire, whereas unidirectional buses charge and discharge through the same end, and thus, do not suffer from EM. The consequences of EM are twofold.

—Wide wires may be required. EM can be mitigated using wider wires because they have lower current densities, and hence degrade more slowly. The main drawback of using wider wires is the area overhead because larger area is required for such wires as well as some extra space between wires. This area overhead implies a significant reduction in terms of interconnect density, which is an extra constraint for the design and leads to inefficiencies. Thus, significant performance is left on the table. Furthermore, using wider wires does not eliminate EM but delays it.

—Cycle time may be impacted. EM must be tolerated to some extent in current designs, which implies that circuits must keep operating when they are somewhat degraded and their performance is lower. This is achieved by adding some guard bands in the cycle time. Thus, operating frequency is decreased.

Techniques to mitigate EM and keep it at bearable levels have been proposed in the past, mainly from the circuit and technology standpoint, but EM is still an open issue for current and future technologies. Providing technology and circuit design solutions to mitigate EM is increasingly expensive. However, microarchitecture can tackle EM degradation because high-level design solutions may be cheaper than using either new materials with lower EM or adding new constraints to circuits to make them fit EM requirements. Thus, the microarchitecture community must devote some effort to deal with EM at low cost.

This article brings a comprehensive survey about EM in digital circuits to the microarchitecture community. The objective is to motivate research on suitable solutions to EM. Our survey illustrates the benefits of balancing the current flow in each direction of wires as a potential solution to EM degradation.

The rest of the article is organized as follows. Section 2 describes the EM phenomenon. Section 3 studies the effects of EM in different types of wires. State-of-the-art techniques are presented in Section 4. Potential benefits of balancing the current in wires are shown in Section 5. Finally, Section 6 summarizes the article.

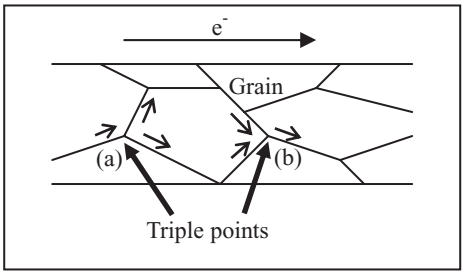

Fig. 1. Diffusion paths at triple points leading to the creation of voids (a) and hillocks (b).

2. ELECTROMIGRATION BACKGROUND

EM was discovered more than 100 years ago, but it was not a concern in bulk wires, such as those used for home circuitry. It was only when integrated circuits appeared that EM emerged as one of the most important sources of degradation. In such circuits, the current density can be nearly $10^6 A/cm^2$ (10 times higher than maximum current density in bulk wires) or even higher. With such high current density, EM, which depends on it, becomes significant. Thus, solutions for EM must be provided for integrated circuits.

The research community started working on analyzing EM and finding solutions in the late 1960s [Blech and Meieran 1967; Black 1969; D'Heurle 1971]. Several improvements were reported to mitigate EM from the circuit and technology side, and have been effective for some years, but as technology scaling continues, the EM problem must be faced again and again. Therefore, new techniques are required to deal with EM in current and future technologies, especially from the microarchitecture side where potential solutions have hardly been explored and benefits can be huge.

In this section we present the main issues involved in EM. We begin by presenting a detailed description of the EM phenomenon and how EM depends on current density. Then, we introduce the self-healing effect that allows repairing EM under certain conditions.

2.1. Electromigration Phenomenon

Metal wires are imperfect due to missing atoms (vacancies), impurities, boundaries between crystals of metal with different orientations (grain boundaries), and so on. Such imperfections make electrons flowing through wires collide with metal atoms which are pulled along in the direction of the current flow, causing what is known as EM. Additionally, high temperatures and current densities increase the likelihood that metal atoms will move. When metal atoms are pulled along, there are two undesirable consequences: voids and hillocks. On the one hand, voids appear in those parts of the wire where metal atoms were before being pulled along. Eventually, those parts of the wire may break. On the other hand, metal atoms are piled up in some other parts of the wire, forming hillocks. Eventually, those hillocks may create shorts with neighboring wires.

Highest EM happens in grain boundaries, particularly in the so-called "triple points" [Lienig and Jerke 2005]. As shown in Figure 1, when we have more outbound than inbound transport paths in a given grain boundary, that location is very likely to form a void (Figure 1(a)), whereas otherwise it is very likely to form a hillock (Figure 1(b)) (more inbound than outbound paths).

This kind of diffusions is avoided with the so-called "bamboo" wires [Scorzoni et al. 1991; De Munari et al. 1995], where the grain diameter exceeds the wire width; but this

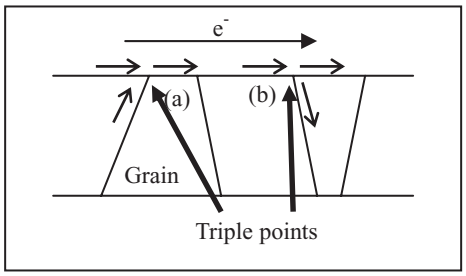

Fig. 2. Diffusion paths at grain boundaries of a "bamboo" wire leading to the creation of voids (b) and hillocks (a).

design is feasible for very narrow wires only. In this case, most grain boundaries involve only two grains (see Figure 2). Although bamboo wires significantly reduce EM, they do not eliminate it because grain boundaries are not completely perpendicular with regard to the electron flow, and hence, significant EM is still observed in such diffusions. As shown in Figure 2, depending on the orientation of diffusions, they are likely to form voids (Figure 2(b)) or hillocks (Figure 2(a)).

2.2. Electromigration Modeling: The Effect of Current Density

Black's law [Black 1969; D'Heurle 1971; Cadence 2002a] accurately describes the mean time to failure (MTTF)[1] due to EM with the following formula:

$$MTTF = A \cdot j^{-n} \cdot e^{\frac{Q}{k \cdot T}} \tag{1}$$

where A is a technology dependent constant that needs to be empirically determined, j is the current density in the interconnect, n is a constant that depends on the metal used for the interconnect (typically between 1 and 2, 1.1 for copper [JEDEC 2002]), Q is the activation energy for EM, k is Boltzmann's constant, and T is the absolute temperature in Kelvin. j is computed as follows [D'Heurle 1971]:

$$j = \frac{C \cdot V_{DD}}{W \cdot H} \cdot f \cdot p \tag{2}$$

where C stands for capacitance, W (H) stands for width (height) of the wire, V_{DD} is the supply voltage, f is the clock frequency and p the switching probability.

Technological evolution leads to significantly smaller wire width and height, whereas V_{DD} scales moderately. Hence, current density increases, shortening the lifetime due to EM. From the formula, it is clear that increasing the width of wires would reduce the current density, which would reduce EM and extend the lifetime of the wire. While this solution is cheap for vias (connections between wires at different metal layers), since they can be widened or even replicated with very low area cost [Atakov et al. 1998; Nguyen et al. 2001], increasing the width of wire stripes has significant area overhead. The area overhead is caused by the extra area needed for the wider metal stripes and the extra gap required between stripes. This area overhead has a significant impact on

[1]MTTF is defined either as the mean time to failure or the median time to failure. Using one or the other definition has some statistical implications affecting the final value of MTTF. However, in this article we work with relative and normalized values, and hence both definitions provide the same results.

the design because it reduces the interconnect density, and hence fewer wires can be placed, which may lead to much more inefficient designs. The power supply and ground grids are especially susceptible to this issue because they require a large number of wires driving high current to feed the whole chip.

Furthermore, we must take into account that EM is a self-accelerating source of failure because as some metal atoms are pulled along, some parts of the wire become narrower. Hence, current density increases in those parts, accelerating EM. Additionally, since current density is higher, higher temperatures are generated, leading to further acceleration of EM. Such an effect is known as self-heating, and it is a hard constraint in wire-sizing because very high current densities make EM degradation grow exponentially.

2.3. Self-Healing Effect

EM is an important source of degradation, which can be undone by two different self-healing effects. The first one affects those wires whose product $j \cdot L$ (where j and L stand for current density and wire length respectively) is below a technology-dependent threshold [Blech 1976]. Metal atoms are pulled along in the direction of the current flow, but the mechanical stress in the hillock area due to the accumulation of metal atoms causes a reversed migration process. This reversed migration can reduce or even compensate the effective material flow. Hence, EM is self-healed in those wires, which are known as "immortal wires". In general, this only happens for very short wires driving very low current, which is a very small percentage in current designs. In particular, it was recently shown that this effect may only happen in wires whose length is up to some tens of microns (e.g., 100 μm [Lienig 2006], 50 μm [Arnaud et al. 1998]).

The second self-healing effect happens when current flows in both directions of a wire [Blech and Meieran 1967]. This self-healing effect can be easily explained by recalling Figure 1. We can see that current flows from left to right and that the triple point in Figure 1(a) is a source of void formation because it has one inbound and two outbound transport paths. If the current is reversed and flows from right to left, then Figure 1(a) has two inbound and one outbound paths, thus becoming a source of hillock formation. In general, those grain boundaries that are very likely to form voids when current flows in one direction are very likely to form hillocks when current flows in the opposite direction. Hence, whenever the same amount of current flows in each direction at similar temperatures (MTTF depends on temperature, as shown in Eq. 1), EM is practically self-healed. Furthermore, if EM healing is performed in a timely way, the self-acceleration effect of EM is avoided. Although wire repair is significant, full recovery is not achieved because not all the atoms that were dislocated are dragged back to their original location. This self-healing effect of EM extends wire lifetime by several orders of magnitude [Liew et al. 1990; Tao et al. 1996] (depending on the metal being used, it can be more than 1,000 or 10,000 times longer) even for high operating frequencies [Tao et al. 1996]. In the rest of this survey we will refer to this second self-healing effect as *the self-healing effect*.

In order to put in perspective the potential and overhead of both approaches, that is, of widening the wires and balancing current flow in each direction, we compare their impact on EM. We performed an experiment based on the EM model in Section 2.2, and plotted the results in Figure 3. We show the relative lifetime of a range of wire widths for three different metals; we consider $n \in \{1, 1.1, 2\}$. As explained in Section 2.2, n is typically between 1 and 2, and 1.1 is the value for copper [JEDEC 2002]. If we assume that the lifetime of a wire extends by 1,000 (the lower bound in Liew et al. [1990]) when current is perfectly balanced in both directions, the width of the wire

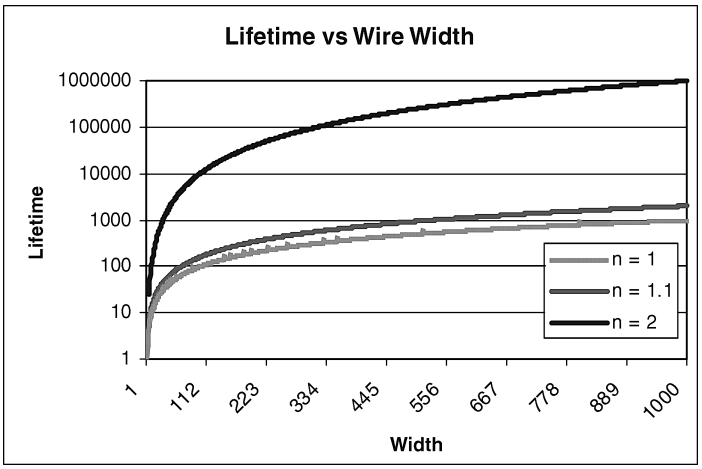

Fig. 3. Lifetime versus wire width for different metals (values of n).

must be increased by at least a factor of 32 to guarantee the same lifetime for n=2. In the case of copper, the width must be increased by at least a factor of 534! Achieving such benefits in terms of area savings (or lifetime extension) requires balancing the current in each direction of the wire. However, the mechanisms to do so may have some overheads that must not offset the benefits provided by balancing the current. Any solution that keeps the current flowing in both directions balanced guarantees a lifetime 1,000 times longer than the base case with the same wire width, or an area saving of more than 99.8% (533/534) with respect to a copper connection with the same lifetime if EM is the only constraint for wire width. Obviously, nobody will increase the wire width that much to extend the lifetime due to EM by 1,000, since other sources of failure like oxide breakdown, negative bias temperature instability (NBTI), and so on, will become, relatively, much more significant than EM; but the example is useful to illustrate the potential benefits of microarchitectural solutions that exploit self-healing as a solution for EM.

3. WIRES AFFECTED BY ELECTROMIGRATION

Different types of wires can be found in digital circuits of a microprocessor. Some of them are affected by EM whereas some others are not. The main types of wires are as follows.

(a) *Unidirectional wires*. In such wires a signal is sent from one end (A) of the wire to the other end (B), but B is not allowed to send any signal to A (see Figure 4(a)). When the signal sent does not change the previous value of the wire, it does not drive any current, but if the value changes (from 0 to 1 or from 1 to 0), some current is driven. When A sends a "1" after sending a "0" some current is charged through A and flows to B, but it is not discharged. This signal can stay high for a while, but extra current is not driven. After some time, A sends a "0", which implies that the current previously sent to B discharges through A, compensating the EM produced when the signal moved from 0 to 1. Hence, we can see that unidirectional wires driving signals are not affected by EM significantly, unless the amount of current that is charging and discharging is significantly different, which can be addressed by properly sizing the devices plugged in to these wires. The only open issue regarding EM for unidirectional wires is the self-heating effect, which arises when the current

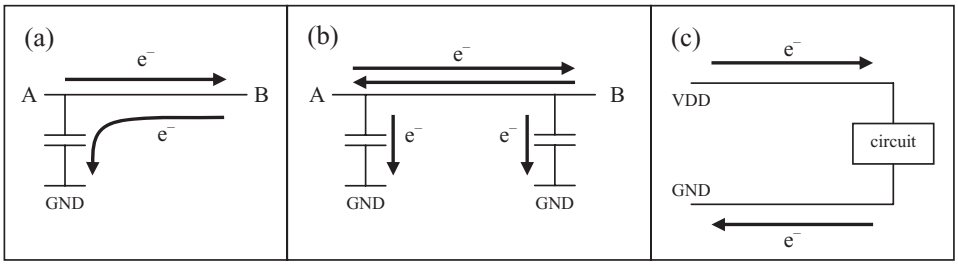

Fig. 4. Current flow in (a) unidirectional wires, (b) bidirectional wires and (c) power/ground wires.

density in a given wire is very high, and ends up accelerating EM degradation exponentially. Due to the high current they drive, clock networks are one of the most important candidates to suffer from the self-heating effect if they are not sized properly. In any case, note that self-heating is an issue for any wire, regardless of current flow direction.

(b) *Bidirectional wires*. In such wires both ends can send signals (see Figure 4(b)). Hence, it may happen that *A* sends a "1" to *B* after observing a "0" charging some current through *A*, and then *B* sends a "0" to *A* discharging through *B* the current sent from *A*. As we can see, the current charged through one end can be discharged through any end (the other one or itself), and hence, depending on the activity pattern, the wire may experience EM degradation. Moreover, devices plugged in to each end of the wire may have different current requirements in such a way that the amount of current sent in each direction is always different.

(c) *Supply power and ground grids*. Wires in such grids are unidirectional, but they do not send signals (see Figure 4(c)). The supply power grid drives the current required by the circuits in order to operate. Later, such current flows back through the ground grid. Since these grids feed all structures in the chip, the amount of current that they drive is very high, and hence their wires must be significantly widened to deal with EM. The extra area required for the wide wires and the extra gap between them reduce severely the available areas in some metal layers, which is a hard constraint for the design of the circuits in the chip, thus leading to inefficient designs in performance and power.

4. STATE-OF-THE-ART APPROACHES

EM is a well-known problem, and has been studied in the past. In this section we present the main areas of research: EM analysis and evaluation, technological solutions, preverification approaches, verification, and postverification solutions.

4.1. Electromigration Analysis and Evaluation

Lienig [2006] has studied the implications of EM at the physical level and how to deal with EM systematically. In particular, Lienig studied the impact of various materials such as copper and aluminum, as well as temperature, wire length and width, and via arrangements. Buerke et al. [2000] studied the impact on EM of grain orientation as well as void and hillock formation. Maiz [1989] analyzed how current direction and pulse frequency impact EM.

The efficiency of accelerated tests as well as degradation and recovery mechanisms in vias have also been studied [Sun et al. 2002; Zitzelsberger et al. 2003].

4.2. Technological Solutions

So far EM has been considered an unavoidable source of degradation in circuits [Young and Christou 1994; Steele et al. 1998; Cadence 2002b; Ajami et al. 2005; Lienig and Jerke 2005; Wang and Marek-Sadowska 2005; Lienig 2006], and there are some design rules that are used to deal with EM, especially in power/ground grids. Young and Christou [2004] have analyzed the causes for EM and described some design criteria for avoiding EM, such as carefully choosing the proper geometries for wires and vias. By avoiding the high current density of particular wires and vias, EM is mitigated. Similar considerations regarding geometry and arrangements for wires and vias have been proposed by Lienig [2006].

4.3. Preverification Approaches

Dasgupta and Karri [1996] have proposed techniques to reduce the activity in some wires so as to mitigate EM degradation. Such solutions can be used to mitigate EM in bidirectional buses, but only if there are multiple instances of a given bus in such a way that they can choose which one to use. This is not often the case, since it is common to have a single instance of this kind of bus (i.e., a bus between DL0 and UL1, between UL1 and memory, and between different cores).

Differential signaling buses [Chiarulli et al. 2005] can be used instead of conventional buses (single-ended buses) because they do not suffer from EM as much as conventional buses. Unfortunately, differential signaling buses are significantly slower than conventional buses, since the number of wires required may be up to twice the number for conventional buses and their power and area overhead is significant due to the logic for sensing and outputing their data. Therefore, differential signaling buses are suitable for off-chip communication where they enable low operating voltage, but they are not suitable for high-performance processors where the latency of buses is critical for performance.

4.4. Verification

Design rules to mitigate EM systematically are based on measuring how much current drives each wire and redistributing current when possible [Steele et al. 1998; Cadence 2002b; Ajami et al. 2005; Lienig and Jerke 2005; Wang and Marek-Sadowska 2005]. Such a verification process is set up to detect whether current constraints are violated in any location once the layout is ready and whether an EM hazard may happen in any wire or via. Once this process finishes, corrective actions must be taken until those violations are fully eliminated. In general, these corrective actions require modifying the design, which has an impact on performance, power, and area.

4.5. Postverification Solutions

Different solutions have been proposed to reduce EM violations in vias and wires. Solutions for vias [Atakov et al. 1998; Nguyen et al. 2001] focus on replicating and resizing vias to mitigate EM. They do not solve the EM problem, but allow delaying EM for vias with little overhead, and hence EM becomes a significant concern for metal stripes only.

Conventional solutions for EM in wires are based on increasing the width of wires. Increasing the width of a wire by a factor of K increases its lifetime by a factor in the range $[K{:}K^2]$, depending on the value of n in Eq. 1. For instance, a wire of width $2 \cdot W$ has a lifetime between 2 and 4 times longer than a wire of width W. Increasing the width has a significant overhead in terms of area due to the size of the wire itself and

Table I. Wire pitch, supply voltage, operating frequency, gate capacitance and chip area for different technologies

Technology	90 nm	65 nm	45 nm	32 nm	22 nm
Metal 1 pitch (nm)	214	152	108	76	54
Intermadiate wiring pitch (nm)	275	195	135	95	65
Global wiring pitch (nm)	410	290	205	140	100
Supply voltage (V)	1.00	0.90	0.81	0.76	0.72
Operating frequency (GHz)	3.4	4.2	5.1	6.2	7.4
Gate capacitance (F/μm)	8.79E-16	6.99E-16	7.35E-16	6.18E-16	5.25E-16
Chip area (mm^2)	140	140	140	140	140

the extra space that must be left between adjacent wires. In general, the power and ground grids require a very large width increase to deal with EM, and hence the area overhead is huge. Some work [Xuan 2004] proposes analyzing circuits to widen the most vulnerable wires only. Although this technique shows promising results, it is useless for circuits where all wires (or the vast majority of them) behave the same because all (or the vast majority) must be widened, which is the case for the most vulnerable wires in the chip, that is, the power/ground grids, the clock network, and bidirectional buses. Other work [Venkateswaran et al. 2005] proposes setting up some spare wires next to the most vulnerable ones to replace them when they become faulty. Again, such a solution is ineffective for those circuits where all wires behave the same, such as power/ground grids, the clock network, and bidirectional buses.

While all those solutions mitigate EM to some extent, either they do not suffice in avoiding EM or they come at that high cost. Thus, further solutions overcome such limitations are required, and balancing the current shows high potential for mitigating EM degradation at low cost.

5. TECHNOLOGY SCALING

In this section we examine the benefits in lifetime and interconnect density that can be achieved by balancing the amount of current flowing in each direction. We focus on some wires in digital circuits where such a potential microarchitectural solution can provide benefits, that is, bidirectional wires and power/ground grids. We start by analyzing the impact of EM for different technologies, and then provide case studies for the power/ground grids and bidirectional wires between different cache levels.

5.1. Impact of Technology Scaling on Lifetime

Based on Eqs. 1 and 2 in Section 2.2, and Intel [Borkar 2006] and SIA projections [Semiconductors Industry Association (SIA) 2003, 2005], we studied the lifetime degradation due to EM from generation to generation. Table I shows the wire pitch different types for wires for different technology generations, as well as the expected supply voltage, operating frequency, gate capacitance, and chip area. Metal 1 stands for the lowest metal layer, intermediate wiring corresponds to some metal layers just on top of metal 1, and global wiring corresponds to the uppermost metal layers. We can see that wire pitch decreases at a rate of 0.7 for all metal layers, so from this point on we focus on metal 1, although the conclusions we draw apply to all metal layers. Supply voltage cannot be scaled at the same pace, and just scales at a decreasing rate [Borkar 2006] (0.92 on average), while gate capacitance scales at an 0.88 average rate. On the other hand, frequency cannot increase as much as projected by SIA [Semiconductors Industry Association (SIA) 2003, 2005] due to power budget limitations and decreased oxide thickness scaling, so we use more moderate projections [Borkar 2006] with near-linear frequency scaling. The area of the processor is assumed to remain

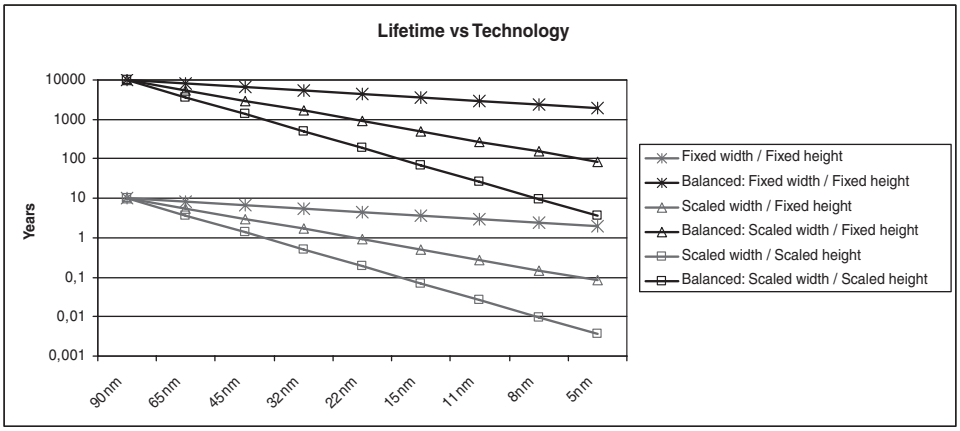

Fig. 5. Wire lifetimes for different technologies.

constant (e.g., 140 mm^2) according to SIA projections [Semiconductors Industry Association (SIA) 2003, 2005].

Based on these values we compute the relative MTTF of each generation with respect to the initial one (10 years for 90nm) for copper connections ($n = 1.1$ in Eq. 1) [JEDEC 2002; Srinivasan et al. 2004b; Shin et al. 2007]. Note that we neglect parameters such as wire length, among others. Our study applies to any wire independently of its length. Therefore, if a given wire suffers small EM degradation due to its short length, it will not be considered for optimization. However, our results on the impact of technology scaling apply to any wire whose MTTF due to EM is of concern.

We show in Figure 5 three different scenarios: (i) the wire width remains fixed and does not shrink as transistors do; (ii) the wire width scales at the same rate as transistors do; and (iii) both wire width and height scale at the same rate as transistors do. For each scenario we plot the expected lifetime for (i) the base case and (ii) the case where current flow in both directions is balanced. We assume that such a balance extends the lifetime by a factor of 1,000, which is the most conservative value in the literature [Liew et al. 1990; Tao et al. 1996].

In all cases we assume that the gate capacitance (C in Eq. 1) scales at a rate of 0.88, without considering that the area of transistors fed shrinks. This means that we take the capacitance scaling per area unit into account, but we consider that the area of the transistors fed remains constant. This is fully accurate for the power/ground grids because transistor geometry scales, but there are more transistors, and therefore the total area fed remains practically constant (basically, this area is the chip size, which is expected to remain constant [Semiconductors Industry Association (SIA) 2003, 2005]). In the case of bidirectional wires, capacitance may reduce if the number of transistors does not grow and they scale. In this case, the lifetime does not decrease as fast as shown in the figure, but new bidirectional wires (e.g., communication between different cores), which will require some current balancing, are expected to have higher load capacitances. Even in the ideal case where the capacitance scales perfectly and the lifetime is larger, such wires still need some current balancing to allow the wire width and height to scale.

We can observe in the figure that lifetime decreases gradually in the scenario where current is imbalanced and the width of wires remains constant (it does not scale). Such a scenario is undesirable because the width of the wires increases with respect to the size of the transistors, which means that there is the same number of wires for an increased

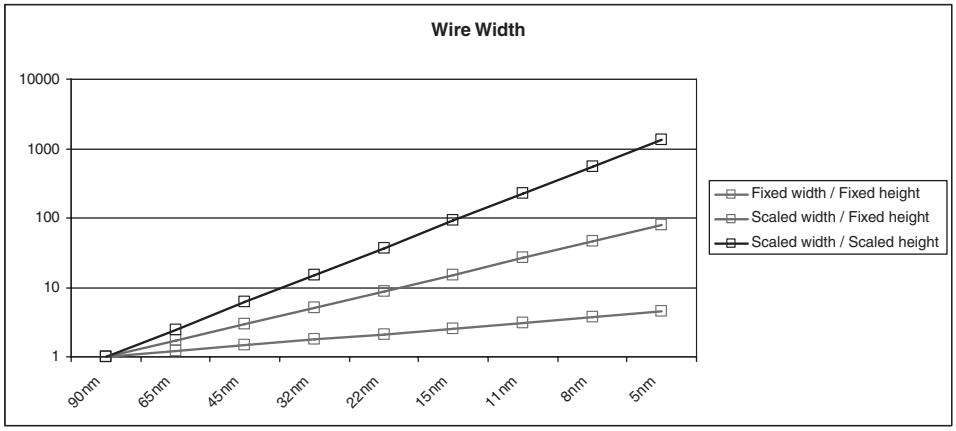

Fig. 6. Wire widths for different technologies.

number of transistors. Real technology scaling requires the wire width to scale at the same pace as transistor geometry. In this case, we observe that the lifetime rapidly decreases. If wire height also scales, then the situation is aggravated. For instance, for 22nm technology, scaled width wires last for 1 year if height does not scale, and only for 2 months if height scales.

On the other hand, current balancing increases lifetime by several orders of magnitude, and hence lifetime is not expected to be a concern until we move beyond 8nm for the case where wire width and height scale. If it is the case that the wire height does not scale, balancing current is enough to deal with EM beyond 5nm. Finally, if we consider the case where width and height do not shrink, then balancing the current solves the EM problem for more than 30 technology generations, which is much more than CMOS is expected to reach.

We have also evaluated what the cost would be in terms of area if we wanted to keep the lifetime constant. We plotted the relative wire width required for such a constant lifetime in Figure 6. For instance, for 22nm technology, if we keep the width and height constant, we still have to increase the width by a factor of 2.1 to keep the lifetime constant. This wire width corresponds to a wire width increase of a factor of 8.8 with respect to the case where the width scales (the same width but compared against the scaled width wire). Finally, if wire width and height scale, the width must be increased by a factor of 36.9 to keep the lifetime constant for 22nm technology. As shown in Figure 5, if the current is balanced, wires do not have to be widened, since a lifetime of over 100 years is guaranteed. Hence, balancing current saves 52%, 89%, and 97% in area in the metal layers respectively, while the lifetime is still higher than the target lifetime.

5.2. The Case of the Power/Ground Grid

Power supply and ground grids involve large amounts of wires that occupy a vast area of some metal layers [Choi et al. 2002; Su et al. 2004]. Such grids are typically designed as interleaved identical stripes of metal at different metal layers, whose width decreases when moving to lower metal layers [Choi et al. 2002; Cadence 2002b; Cai et al. 2005]. Hence, at any metal layer, power and ground wires are identical and distributed in the same way. Figure 7 shows an example of a power/ground grid [Cai et al. 2005]. We see that layers $N + 1$ and N are designed in the same way as layers $N + 3$ and $N + 2$. The

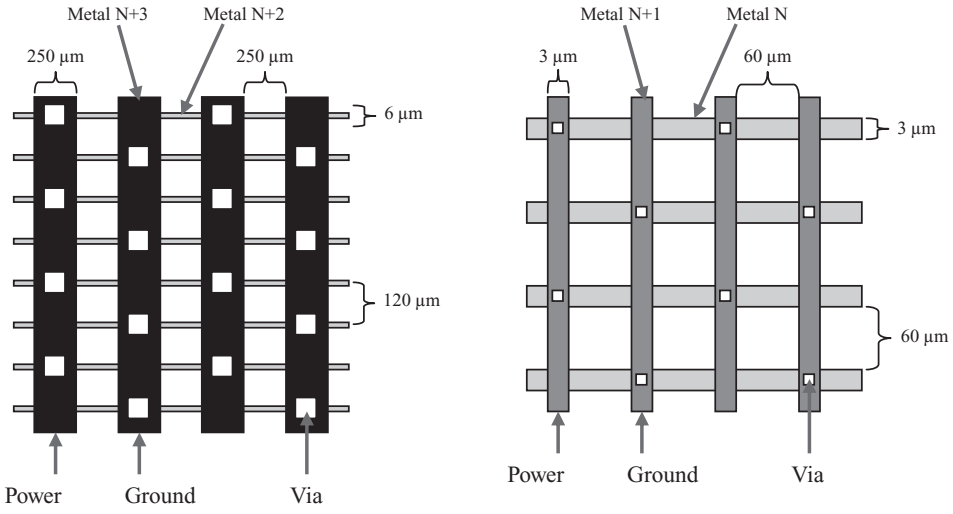

Fig. 7. On-chip power/ground grid structure.

only difference between these two pairs of metal layers is the wire width and period.

The benefits of balancing current in power/ground grids may be huge in terms of lifetime and area savings. Recalling Figures 5 and 6, 52% of the area of those wires can be saved for 22nm technology if the wire width and height do not scale, and the lifetime can be extended by a factor of 436 with respect to the target lifetime simultaneously. If wire width is scaled, then 89% of the area is saved and the lifetime is extended by a factor of 91. Finally, if both wire width and height are scaled, the area savings are 97% and the lifetime is extended by a factor of 19. We must keep in mind that those are the improvements in area and lifetime assuming that wire width is constrained only by EM. It may happen that some wires of the power/ground grids are widened beyond EM requirements just because of current constraints or other reasons such as IR drops, electrical overstress, and so on. In other words, we can shrink those wires if our only constraint is EM, but we may need to keep them wide to drive enough current to prevent electrical overstress. Hence, the improvements depend on the concrete design. For the sake of simplicity, in the rest of this article we refer to all wire scaling constraints but EM as current constraints.

To illustrate how significant the improvements in terms of area and lifetime can be, recall the example in Figure 7. Metal layers N and $N + 1$ have 3μm-wide wires for power and ground, with a period of 60μm. We assume that those wires cannot be made narrow due to current constraints. If we move to the following technology, they must be 3.6μm wide for the same wire height, or 5.2μm for scaled wire height because of EM. Assuming that the gap between wires must be at least as wide as the wires, the base case uses 10% of the area of those metal layers. When moving to the following technology node, the area devoted to the power/ground grids in those metal layers is 12% if the height does not scale, and 17% if it does. On the other hand, if we manage to balance the current flowing in each direction of the wires, the area devoted to power/ground grids in those metal layers can be reduced to 7% or remain constant at 10% if needed due to current constraints.

We illustrate the whole evaluation space in Figure 8, assuming that wire width cannot be made narrow due to EM constraints. The first two columns show the case where current is balanced and the wire can shrink because there are no current constraints

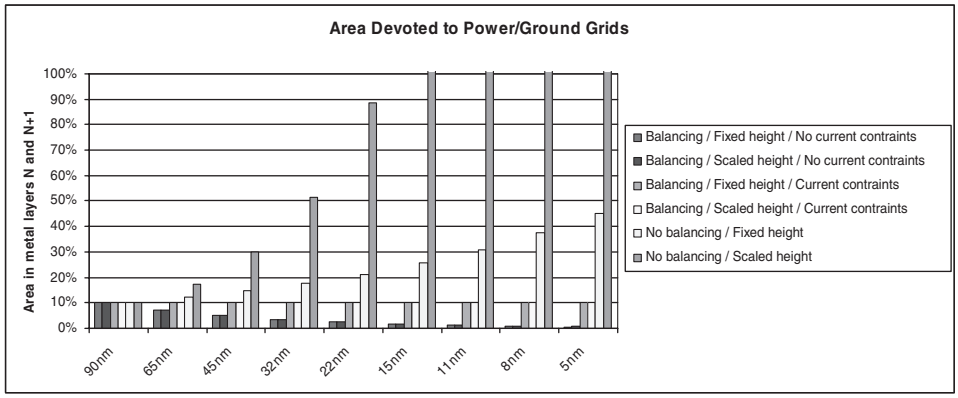

Fig. 8. Area devoted to the power/ground grids.

and lithography allows scaling them that much. If the wire height is fixed, we can scale the width for all technology generations, otherwise, the wire width increases after 8nm technology, but its area is below 1% even for 5nm technology. The third and fourth columns represent the same scenarios as the first and second columns, but in this case we cannot decrease the wire width due to current constraints (on top of the EM constraints). Hence, the wire width is the maximum between the initial 10% and the width allowed if current is perfectly balanced. Since the area required is below 10% to meet EM constraints in all scenarios when current is balanced, we do not need to increase the wire width further. The two last columns represent the case where no balancing is performed. We can observe that the wire width must be increased in both cases, leading to situations where this design is unfeasible because power/ground grids require more than 100% of the area (15nm for scaled height). Even in those cases where area requirements are below 100%, a large fraction of the area must be devoted to the power/ground grids, reducing the area available for other wires.

Although balancing current in power/ground grids can mitigate EM, any solution in this direction must face several issues.

—*Device isolation.* Power and ground wires are designed to drive current in a single direction. Injecting current in the opposite direction in those wires requires isolating all devices from power/ground wires being repaired. Otherwise, those devices might be damaged.

—*Current distribution.* Another issue is how to feed power/ground wires to repair them. If the power distribution network is disabled for repair, it is not clear how to provide current in the proper locations to repair the wires. If only a subset of the wires is disabled for repair, the rest of the power/ground grids can be used to provide current. However, isolating some parts of the grid is tricky due to the large number of wires connected to each one of the power/ground wires.

—*Keeping state.* The devices connected to power/ground wires under repair may not be fed by other sources of current. Therefore, such devices must be turned off during power/ground wire repair. However, if those devices store some critical data (e.g., the state of the program being run), a mechanism to copy such a state to another location and restore it afterwards is required.

—*Characteristics of the repair process.* Last but not least, the amount of current that has to be injected to repair wires and how long repair has to last must be tracked somehow. Depending on the granularity used for repair (either large or small parts

Table II. Workloads

Benchmark suite	# traces	Description
Encoder (ENC)	62	Audio/video encoding
SpecFP2000 (SFP2K)	41	Floating-point specs 2000
SpecINT2000 (SINT2K)	33	Integer specs 2000
Kernels (Kernels)	53	VectorAdd, FIRs
Multimedia (MM)	85	WMedia, photoshop
Office (Office)	75	Excel, Word, Powerpoint
Productivity (PROD)	45	Internet contents creation
Server (Server)	55	TPC-C
Workstation (WS)	49	CAD, rendering
Spec2006 (SP2K6)	33	Spec 2006

of the grids each time), the overhead to track the information may change. Similarly, depending on the accuracy of the information (e.g., fully accurate for perfect repair, or approximate for near-optimal repair) it may be costly to retrieve the information required.

As shown, balancing current has a huge potential to mitigate EM in power/ground grids. How it can be done from the microarchitecture side is an open issue, and overheads to exploit, either fully or partially, the benefits of balancing current may kill many solutions. However, enabling technology scaling for some generations by mitigating EM with microarchitectural techniques is a promising path to follow.

5.3. The Case of Intercache Buses

In this section we examine the potential benefits in terms of lifetime of perfect current balancing in those bidirectional buses set up to communicate different cache levels: (i) the first-level data cache (DL0) and the second-level cache (UL1), and (ii) the UL1 and the memory interface of the chip. Our analysis compares the lifetime of perfect current balancing with respect to both the worst-case situation (any access makes the current always discharge through the same wire end), and the real-case. In order to perform such experiments, we had run simulations of traces from a wide set of workloads (see Table II) in a IA32 Intel simulator to collect bus activity patterns. The simulator resembles the Intel® Core™ Microarchitecture [De Gelas 2006], although results are mostly independent of the processor configuration. In our experiments, all wires in a given bus have identical wire widths. Such width is arbitrary because relative MTTF benefits hold for any wire width, based on the assumption that such width is constrained only by EM.

As stated in previous sections, the lifetime of the case where current is balanced is expected to be around 1,000 times longer than that for the worst case. Figures 9 and 10 show the results for DL0/UL1 and UL1/Memory buses respectively, normalized with respect to the worst case (in logarithmic scale). Lifetime for the real case is computed by tracking how many times current is discharged through each end of each wire. Then, the maximum value for any wire in the bus is used as j in Eq. 1 and normalized with respect to the lifetime in the worst-case scenario, where any bus transaction would discharge all wires in the bus through the same end.

While the real case is around 36 times better than the worst case on average, it is far from the lifetime of perfect current balancing, which offers a lifetime 1,000 times longer [Liew et al. 1990; Tao et al. 1996] than the worst case (28 times longer than the realcase). Even if the worst case and real case buses are affordable in terms of area, balancing the current can be used to reduce the delay guard band due to EM, leading to higher operating frequencies or lower latencies for the affected components.

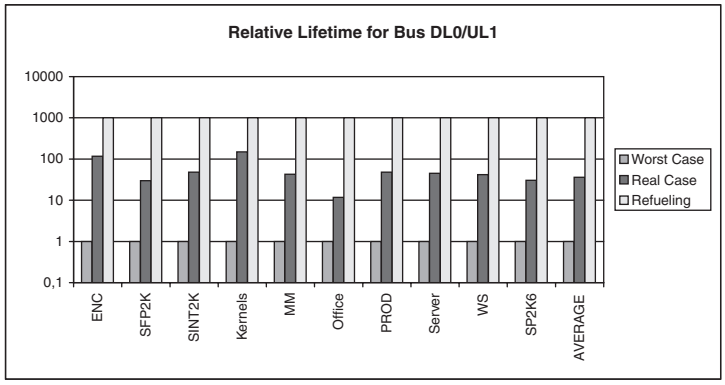

Fig. 9. Relative lifetimes for the worst-case, real-case, and perfectly balanced scenarios for the DL0/UL1 bus.

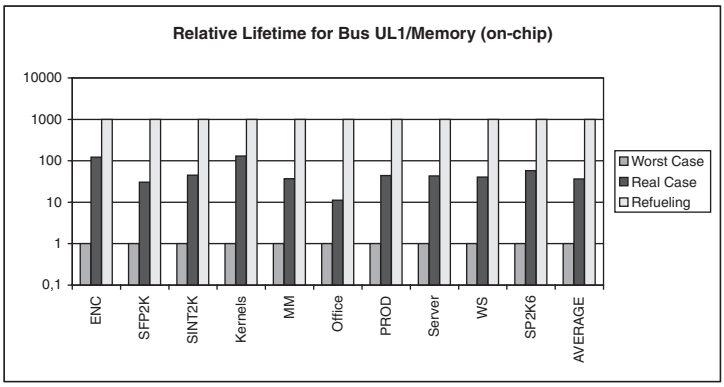

Fig. 10. Relative lifetime for the worst-case, real-case, and perfectly balanced scenarios for the UL1/Memory bus.

Specific results for each group of benchmarks show that *Kernels* and *Encoder* groups are the ones with the highest real-case lifetimes because their activity is highly self-balanced. *Office* and *SpecFP2000* are the groups with shortest lifetimes because their activity is highly imbalanced. There is not a clear correlation between the amount of activity in the buses and lifetime. For instance, *SpecFP2000* is the group with the highest activity (59 L1 accesses and 54 memory accesses in 1,000 cycles) and its lifetime is low (29 years). However, the *Kernels* group also shows very high activity (47 L1 accesses and 34 memory accesses in 1,000 cycles) and its lifetime is the highest one (149 years). Conversely, *Office* has low activity (31 L1 accesses and 29 memory accesses in 1,000 cycles) and its lifetime is very low (12 years) when compared to *Kernels*, which has much more activity. Therefore, we conclude that what really matters is not the amount of activity but the current imbalance produced by such activity. For instance, *Office* benchmarks operate on heterogeneous data that cause many discharges of the wires in the bus because bit values change very often. Thus, some of the wires discharge current through the same end most of the time, and their degradation is high.

In contrast to the case of the power/ground grids, balancing current in bidirectional wires has few draw packs. The main issues that must be considered are as follows.

—*Bus isolation.* Buses must be made unavailable for operation during bus repair. In some cases it may be easy to implement such a mechanism if those buses are made unavailable for other reasons (e.g., when they are busy) because the hardware required is already in place.

—*Activity injection.* Since bidirectional wires allow current to flow in both directions, repairing them has few problems. The only concern is how much current must be injected in each particular wire to repairs it. Similarly to the case of the power/ground grids, a mechanism tracking how long the repair will take is required.

—Area and power trade-offs. Finally, buses occupy less area than power/ground grids, so the cost of the hardware to balance current in buses might offset the area savings in the metal layers. Thus, any solution must ensure small use of power and area.

As shown, the area devoted to bidirectional buses can be shrunk in future technologies if it is constrained by EM. However, any solution to balance current must have low hardware overhead because the benefits are moderate. Only if such a constraint is met will microarchitectural solutions be effective.

6. SUMMARY

Electromigration is a threat to continuous wire scaling since metal atoms can be dragged away from their original locations, thus creating voids and hillocks which may cause wires to break or to short with other wires. Fortunately, degradation caused by electromigration can be recovered by driving current in the opposite direction in such a way that those atoms that were dragged away in a given direction are dragged back to their original location.

Some of the most vulnerable wires due to electromigration are power/ground grids and bidirectional buses. Providing solutions for some future technology nodes from the circuit design and process technology side may be very expensive, and the microarchitecture community must contribute new approaches to mitigate electromigration.

We have shown that balancing current flow in each direction in wires may provide huge benefits in terms of both lifetime and area, and hence, cheap microarchitectural solutions must be devised to achieve those benefits and remove electromigration from the large set of processor design constraints.

ACKNOWLEDGMENTS

We wish to thank James W. Tschanz and Osman Unsal for reviewing early drafts of this paper, and the anonymous reviewers for their insightful comments.

REFERENCES

AJAMI, A. H., BANERJEE, K., AND PEDRAM, M. 2005. Scaling analysis of on-chip power grid voltage variations in nanometer scale ULSI. *Analog Integrat. Circuits Signal Process. 42,* 3, 277–290.

ARNAUD, L., TARTAVEL, G. L., AND ULMER, P. W. 1998. Analysis of Blech product threshold in passivated AlCu interconnections. In *Proceedings of the IEEE International Interconnect Technology Conference*. IEEE, Los Alamitos, CA, 289–291.

ATAKOV, E. M., SRIRAM, T. S., DUNNELL, D., AND PIZZANELLO, S. 1998. Effect of VLSI interconnect layout on electromigration performance. In *Proceedings of the 36th International Reliability Physics Symposium (IRPS'98)*, IEEE, Los Alamitos, CA, 348–355.

BLACK, J. R. 1969. Electromigration failure modes in aluminum metallization for semiconductor devices. *Proc. IEEE 57,* 9, 1587–1594.

BLECH, I. A. 1976. Electromigration in thin aluminum films on titanium nitride. *J. Appl. Phys. 47,* 4, 1203–1208.

BLECH, I. A. AND MEIERAN, E. S. 1967. Direct transmission electron microscope observation of electrotransport in aluminum thin films. *Appl. Phys. Lett. 11,* 8, 263–266.

BORKAR, S. 2006. Extending and expanding Moore's law – Challenges and opportunities. In *Proceedings of the 2nd Workshop on System Effects of Logic Soft Errors (SELSE'06)*.

BUERKE, A., WENDROCK, H., AND WETZIG, K. 2000. Study of electromigration damage in Al interconnect lines inside a SEM. *Crystal Res. Technol. 35,* 6–7, 721–730.

CADENCE. 2002a. Electromigration for designers: An introduction for the non-specialist. White paper, Cadence.

CADENCE. 2002b. Power grid verification. White paper, Cadence.

CAI, Y., PAN, Z., TAN, S. X.-D., HONG, X., HOU, W., AND WU, L. 2005. Relaxed hierarchical power/ground grid analysis. In *Proceedings of the Conference on Asia South Pacific Design Automation (ASP-DAC'05)*. ACM, New York, 1090–1093.

CHIARULLI, D. M., BAKOS, J. D., MARTIN, J. R., AND LEVITAN, S. P. 2005. Area, power, and pin efficient bus transceiver using multi-bit-differential signaling. In *Proceedings of the International Symposium on Circuits and Systems (ISCAS'05)*. IEEE, Los Alamitos, CA, 1662–1665.

CHOI, J., WAN, L., SWAMINATHAN, M., BEKER, B., AND MASTER, R. 2002. Modeling of realistic onchip power grid using the FDTD method. In *Proceedings of the IEEE International Symposium on Electromagnetic Compatibility (EMC'02)*. IEEE, Los Alamitos, CA, 238–243.

DASGUPTA, A. AND KARRI, R. 1996. Electromigration reliability enhancement via bus activity distribution. In *Proceedings of the 33rd Annual Conference on Design Automation (DAC'96)*. ACM, New York, 353–356.

DE GELAS, J. 2006. Intel Core versus AMD's K8 architecture. AnandTech. http://www.anandtech.com.

DE MUNARI, I., SCORZONI, A., TAMARRI, F., AND FANTINI, F. 1995. Activation energy in the early stage of electromigration in Al-1% Si/TiN/Ti bamboo lines. *Semiconductor Sci. Technol. 10,* 3, 255–259.

D'HEURLE, F. M. 1971. Electromigration and failure in electronics: An introduction. *Proc. IEEE 59,* 10, 1409–1418.

JEDEC. 2002. Failure mechanisms and models for semiconductor devices. JEDEC Publication JEP122-A.

LIENIG, J. 2006. Introduction to electromigration-aware physical design. In *Proceedings of the International Symposium on Physical Design (ISPD'06)*. ACM, New York, 39–46.

LIENIG, J. AND JERKE, G. 2005. Electromigration-aware physical design of integrated circuits. In *Proceedings of the 18th International Conference on VLSI Design (VLSID'05)*. (held jointly with the *4th International Conference on Embedded Systems Design*). IEEE, Los Alamitos, CA, 77–82.

LIEW, B. K., CHEUNG, N. W., AND HU, C. 1990. Projecting interconnect electromigration lifetime for arbitrary current waveforms. *IEEE Trans. Electron Devices 37,* 5, 1343–1351.

MAIZ, J. A. 1989. Characterization of electromigration under bidirectional (BC) and pulsed unidirectional (PDC) currents. In *Proceedings of the 27th International Reliability Physics Symposium (IRPS'89)*. IEEE, Los Alamitos, CA, 220–228.

NGUYEN, H. V., SALM, C., MOUTHAAN, T. J., AND KUPER, F. G. 2001. Modelling of the reservoir effect on electromigration lifetime. In *Proceedings of the 8th International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA'01)*. IEEE, Los Alamitos, CA, 169–173.

SCORZONI, A., NERI, B., CAPRILE, C., AND FANTINI, F. 1991. Electromigration in thin-film interconnection lines: Models, methods and results. *Material Sci. Rep. 7,* 4–5, 143–220.

SEMICONDUCTORS INDUSTRY ASSOCIATION (SIA). 2003. International technology roadmap for semiconductors.

SEMICONDUCTORS INDUSTRY ASSOCIATION (SIA). 2005. International technology roadmap for semiconductors.

SHIN, J., ZYUBAN, V., HU, Z., RIVERS, J., AND BOSE, P. 2007. A framework for architecture-level lifetime reliability modeling. In *Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)*. IEEE, Los Alamitos, CA, 534–543.

SRINIVASAN, J., ADVE, S. V., BOSE, P., AND RIVERS, J. A. 2004a. The case for lifetime reliability-aware microprocessors. In *Proceedings of the 31st Annual International Symposium on Computer Architecture (ISCA'04)*. IEEE, Los Alamitos, CA, 276.

SRINIVASAN, J., ADVE, S. V., BOSE, P., AND RIVERS, J. A. 2004b. The impact of technology scaling on lifetime reliability. In *Proceedings of the International Conference on Dependable Systems and Networks (DSN'04)*. IEEE, Los Alamitos, 177–186.

STEELE, G., OVERHAUSER, D., ROCHEL, S., AND HUSSAIN, S. Z. 1998. Full-chip verification methods for DSM power distribution systems. In *Proceedings of the 35th Annual Conference on Design Automation (DAC'98)*. ACM, New York, 744–749.

SU, H., HU, J., SAPATNEKAR, S. S., AND NASSIF, S. R. 2004. A methodology for the simultaneous design of supply and signal networks. *IEEE Trans. Comput.-Aid. Des. Integrat. Circuits Syst. 23,* 12, 1614–1624.

SUN, Y., ZHOU, P., KIM, D. Y., GOODSON, K. E., AND WONG, S. S. 2002. Recovery of open via after electromigration in Cu dual damascene interconnect. In *Proceedings of the 40th International Reliability Physics Symposium (IRPS'02)*. IEEE, Los Alamitos, CA, 435–436.

TAO, J., CHEN, J. F., CHEUNG, N. W., AND HU, C. 1996. Modeling and characterization of electromigration failures under bidirectional current stress. *IEEE Trans. Electron Devices 43,* 5, 800–808.

TOSHIBA CORP. 2002. Semiconductor reliability handbook (discrete devices).

VENKATESWARAN, N., BALAJI, S., AND SRIDHAR, V. 2005. Fault tolerant bus architecture for deep submicron based processors. *SIGARCH Comput. Archit. News 33,* 1, 148–155.

WANG, K. AND MAREK-SADOWSKA, M. 2005. On-chip power supply network optimization using multigrid-based technique. *IEEE Trans. Comput.-Aid. Des. Integrat. Circuits Syst. 24,* 3, 407–417.

Received April 2007; revised September 2008; accepted October 2008