Stochastic and Topologically Aware Electromigration Analysis for Clock Skew

Palkesh Jain
Customer Enablement Group,
Qualcomm Technologies Inc., India, Bangalore
palkesh@qti.qualcomm.com

Sachin S. Sapatnekar
Department of ECE
University of Minnesota
sachin@umn.edu

Jordi Cortadella
Department of Computer Science
Universitat Politècnica de Catalunya
jordi.cortadella@upc.edu

Abstract— **An important link between individual component-level EM failures and the failure of the associated system is established in this work. Conventional EM methodologies are based on the weakest link assumption, which deems the entire system to fail as soon as the first component in the system fails. With a highly redundant circuit topology – that of a clock grid – we present algorithms for EM assessment, which allow us to incorporate and quantify the benefit from system redundancies. We demonstrate that unless such an analysis is performed, chip lifetimes are underestimated by over 2x.**

Keywords- electromigration, clock, clock-grid, skew, delay-degradation

I. INTRODUCTION

The phenomenal growth of mobile and wireless systems has been marked by increasing levels of integration of computational components on smaller and denser microchips. Indeed, integrated circuits today contain billions of closely-packed transistors and multi-billion copper interconnects that enable these transistors to communicate. Such aggressively downscaled components (transistors and interconnects) suffer from increasing electric fields and impurities/defects during manufacturing. Compounded by the gigahertz switching, chip designers face significant challenges of reliability and design integrity, with electromigration (EM) being the foremost interconnect reliability challenge.

EM in interconnects occurs due to the movement of metal atoms, activated by momentum transfer from collisions with free electrons [1]. When bounded by a blocking boundary such as a barrier layer, this movement causes a depletion of atoms at the cathode end and a surplus at the anode; this depletion eventually leads to void nucleation and subsequent growth [2]. Since the critical stress for void nucleation is very small for copper dual damascene (CuDD) structures, voids can form early in the lifetime of a design [3].

At the level of *individual components* (metal segments), EM is a fairly well-understood phenomenon, both in terms of the failure criteria (*e.g.*, 10% resistance change) as well as the time-to-failure (TTF) via Black's equation [2]. Additionally, EM recovers with a reversal in the current flow direction, and the average current is computed using an empirical recovery factor (typically between 0.6 and 0.8) [3]. To ensure EM robustness, foundries specify current density limits on wires.

At the *system level*, EM analyses in industry are based on the weakest-link approximation (WLA), wherein the system is deemed as a failure when first component fails [4]. The conventional method for managing EM revolves around containing the current densities in interconnects. These interconnects could be cell-external – signals and power nets connecting cells – or cell-internal, wherein they are wires within a logic-IP (standard cell) or a mixed-signal IP block.

Practically speaking, since the primary determinant for EM is the current flowing through the interconnect, it is in circuits such as clock network – which carry high amounts of current over the chip's lifetime – that EM is a serious concern. In fact, much of the chip-level signal EM analysis is focused on ensuring safety of clock nets, even though they are physically routed at non-default widths due to delay considerations. Pushing the performance envelope of the clock under the constraints of variability and skew has been a critical challenge, and approaches based on clock grids have remained popular since they enable ultra-high frequency and clock signal delivery with minimal skew [5], [6]. Clock grids show high tolerance to variations due to their inherently high redundancy, with multiple source-to-sink paths for every sink.

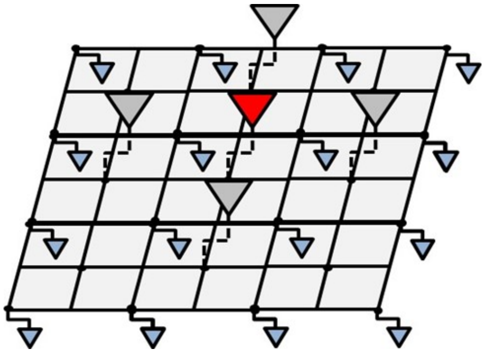

Fig.1 A one-level clock grid schematic showing several buffers arranged in redundant configuration

Thus, although the high frequency and high current characteristics of clock grids make them vulnerable to EM, their highly redundant interconnect structure breaks the WLA assumption of traditional EM containment approaches [7].

Further, grids are multiply driven by several buffers (Fig. 1) connected to a common clock source: these redundant drivers reduce clock skew and lower load/delay variations. Fig. 2

shows a schematic of a single clock grid stage. Additionally, failures in the supply network of the clock grid may cause delay shifts, but the supply network is also redundant due to its mesh structure and can withstand some failures.

WLA ignores all of these redundancies and does not consider the sensitivity of the system functionality to failing wires: a system may operate well even after a component fails. Under system-level failure criteria, we show that unless redundancies are considered, circuit lifetimes are underestimated by over 2X.

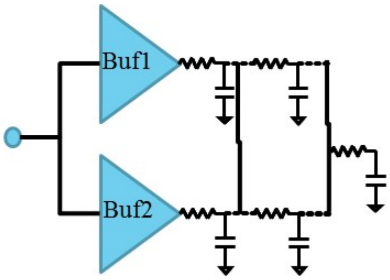

Fig. 2 A single stage of the clock grid with multiple buffers driving the wire segments.

Instead of the WLA, a better criterion for system failure is based on determining when the system becomes non-functional, or when a critical system specification is violated. The purpose of our work is to incorporate the system redundancies and analyze failures at a higher abstraction level than the individual interconnect. We first motivate the benefits of system redundancy in terms of TTF-margins using analytical formulations in section II. Next, we develop a failure abstraction model at the circuit level (e.g., 10% delay degradation) in section III, and demonstrate it on a redundant clock grid structure in section IV.

II. Analytical Approach For Systems With Redundancy

A. *Basics of Electromigration*

EM is widely computed using the Black's equation [2], which describes EM-induced using the relation:

$$t_{50} = A\, J^{-n}\, e^{Q/k_B T} \qquad (1)$$

Here t_{50} is the time-to-failure for half of the experimental population, A is a constant depending on the material properties, J is the current density through the wire and the current-exponent n is empirically determined to be between 1 and 2. Q is the activation energy, while k_B is the Boltzmann constant and T being the wire temperature. For bidirectional current flow in the wires, we must adjust the calculations to accommodate for partial EM recovery. This can be done by modifying J (which is actually a temporal average) with the help of the recovery factor, $\mathfrak{R}$, that is empirically obtained [3]:

$$J = J_{avg}^{+} - \mathfrak{R} J_{avg}^{-} \qquad (2)$$

Here, J_{avg}^{+} and J_{avg}^{-} indicate the average current density in the positive and negative directions, respectively. The temperature T must incorporate the wire temperature rise, ΔT, which depends on the RMS current, J_{RMS}, as:

$$\Delta T = c\, J_{RMS}^{2} \qquad (3)$$

Eq. (3), with c as a fitting parameter, follows directly from heat conduction principles. Typically, a limit on the maximum temperature rise due to Joule heating is a design constraint, and this automatically places limits on RMS current densities.

B. *Reliability Calculations for Changing Stress*

The EM failure statistics of each component depends on its current. Under redundancy, after the first component fails, *current-crowding* is seen in other components, altering their failure statistics.

The initial failure rate, $f(t)$, of each component is lognormal:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{\ln t - \ln t_{50}}{\sigma}\right)} \qquad (4)$$

The cumulative probability distribution function (CDF) is therefore given by:

$$F(t) = \Phi\left(\frac{\ln t - \ln t_{50}}{\sigma}\right) \qquad (5)$$

with Φ (x) as the standard normal CDF.

Consider now a system comprising two components (as in Fig. 3(a)), where both the components initially carry a current density J_1 (Fig. 3(b)). When one of them fails at time t_1, the current in the surviving component changes to J_2.

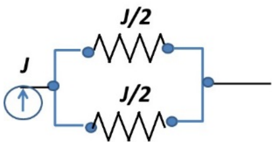

Fig. 3(a) Schematic showing a parallel two-component system

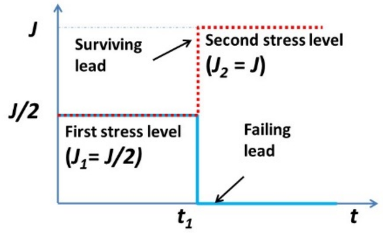

Fig. 3(b) Current profile evolution, with first failure occurring at time t_1.

To analytically approach this, we notice that until t_1, the reliability CDF of each component is described by:

$$F_1(t) = \Phi\left(\frac{\ln t - \ln t_{50,1}}{\sigma}\right) \qquad (6)$$

where $t_{50,1}$ is the MTTF for J_1, as in Fig. 4.

For a general component that carries current corresponding to second stress level (J_2), the reliability is represented by a CDF, $F_2(t)$, and the associated $t_{50,2}$. Now for our case of Fig. 3(b), the CDF trajectory for the surviving component at t_1 therefore must change from F_1 to F_2. To ensure continuity of the CDF curve after the step jump in the current, we shift $F_2$by time δ_1 to ensure continuity with F_1 at time t_1, i.e.,

$$F_2(t_1 - \delta_1) = F_1(t_1) \qquad (7)$$

This equivalence physically implies that the curve follows the trajectory of F_2, starting at the same fraction of the failed population under the two stresses, but that the failure rate increases after t_1. For example, for a ξ_{ij} fail probability (y-axis in Fig. 4), the TTF changes from t_{ijh} (if only first stress were applicable) to t_{ijk} (after change of stress). The effective CDF curve (Fig. 4) is

$$F_1(t) = \Phi\left(\frac{\ln t - \ln t_{50,1}}{\sigma}\right)\ 0 \le t \le t_I$$
$$F_2(t-\delta_1) = \Phi\left(\frac{\ln(t-\delta_1) - \ln t_{50,2}}{\sigma}\right)\ t \ge t_I \qquad (8)$$

Note that using the continuity at t_I, we derive the time shift δ_1. For a system where components undergo a change in stress multiple times, we can generalize the formulation to account for k changes in current, from J_1 to J_2 ... to J_k:

$$\delta_1 = t_1\left(1 - \frac{t_{50,2}}{t_{50,1}}\right)\ldots$$
$$\delta_k = \left(t_k - \sum_{i=1}^{k-1}\delta_i\right)\left(1 - \frac{t_{50,k}}{t_{50,k-1}}\right) \qquad (9)$$

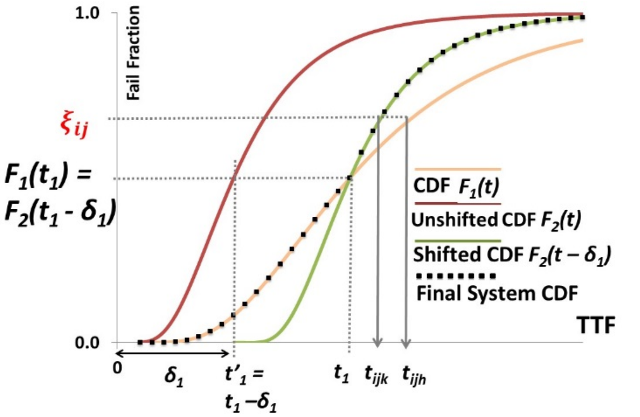

Fig. 4 Analytically estimated CDF evolution of a single component when it undergoes a stress change. The dotted line is the effective CDF, when stress change occurs at t_I.

C. Reliability Calculations for System with Redundancy

We now apply this idea and basic formulation to analyze the system reliability for the structure in Fig. 3(a). We define the system to be functional as long as there is a valid electrical connection between the two terminals of the parallel system. Now, if both components are from the same process population (Fig. 3(b)), the reliability of the case when both are simultaneously functional is given by:

$$R_{11}(t) = \left(1 - F_1(t)\right)^2 \qquad (10)$$

where $F_1(t)$ is as defined in Section IIB.

Next, the reliability for the case when the first component fails at an arbitrary time t_I, and the second component works successfully till time t, must be computed in steps. The probability for the first component to fail between time t_I and $(t_I + \Delta t_1)$ is $f_1(t_1)\Delta t_1$, where $f_1(t)$ is the density function associated with $F_1(t)$. After the current redistribution at t_I, the failure statistics of the surviving component are given by the CDF $F_2(t-\delta_1)$, from eq. (8). Thus, the concurrent multiplicative probability of the second component working when the first has failed is

$$[1 - F_2(t-\delta_1)]f_1(t_1)\Delta t_1. \qquad (11)$$

Integrating over all possible failure times from 0 to t, the reliability for this case is:

$$R_{12}(t) = \int_{t_1=0}^{t_1=t}[1 - F_2(t-\delta_1)]f_1(t_1)dt_1 \qquad (12)$$

The effective failure probability therefore is given by

$$F_{parallel}(t) = 1 - [R_{11}(t) + 2R_{12}(t)] \qquad (13)$$

Such a formulation directly enables us to compare the EM reliability of components connected in parallel topology, versus a single narrow or a wide component. Indeed, for a given CDF for a single component, Fig. 5 compares the CDF for the system failure using this analysis with the WLA case. Note that for a single narrow or a single wide component case, WLA is rightly applicable. However, traditional approach even applies WLA for the parallel system, and it is clear that such an application leads to pessimistic estimates of failure-times. For an exemplary failure fraction of 10%, the TTF is computed to be 35% lower. Thus, WLA could lead to overdesign as designers strive to fix failures that will not happen.

Additionally, for this two-component system, another alternative is to use a single component of twice the width to carry the entire current, $2J_I$. Such a component has the same current density as the parallel leads and its failure probability is the single component CDF, F_1, in Fig. 5, which is significantly worse. Qualitatively, this margin arises from EM stochasticity, since the probability of two narrower components failing simultaneously is smaller than that of a single wide failing.

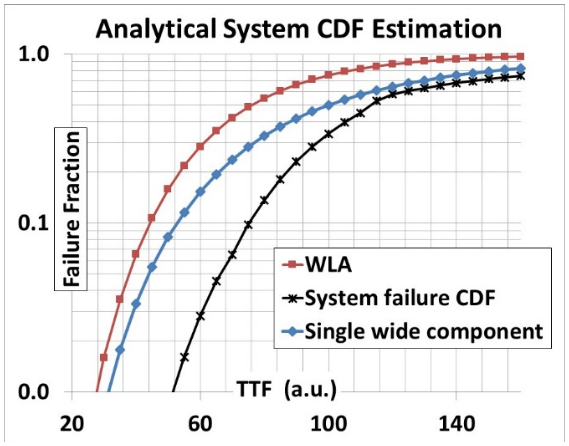

Fig. 5 CDF for a system with redundancy, arrived using analytical formulations (eq. (13)). Shown are the CDFs using the weakest link approximation (WLA), and the CDF for a single wide component.

Further, such a benefit from redundancy scales with the extent of parallelism, as illustrated in Fig. 6. Typically in input/output buffers and chip level power/ground networks, the wires are often required to be wide (> 1um), to support carrying large currents. Such wires can be laid out as a single wide structure (within the maximum width constraint by foundry), or as a parallel connection of several narrow components, wherein the narrow components must adhere to the minimum design rule constraint (DRC) spacing specified by the foundry.

The experiment is set up so that the width of the single wide wire matches the sum of the widths of the narrow wires. This means that the wire parasitics for both cases are roughly the same, but the set of narrow wires occupies a larger area due to the DRC spacing constraints. As we can see from Fig. 6, the benefit from redundancy monotonically increases.

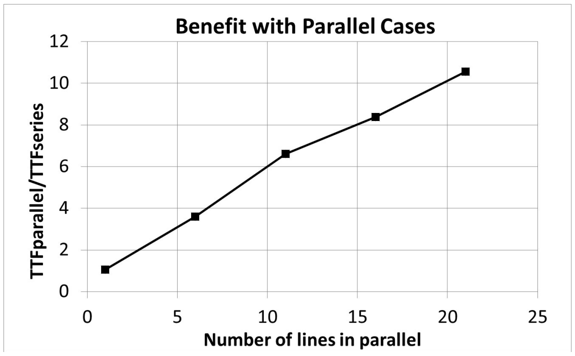

Fig. 6. Showcasing the increasing benefit of redundancy with the number of components arranged in parallel configuration.

III. Monte Carlo Framework For Sytem Reliability Estimation

The analytical two-component example in section II is a useful illustration, but complex circuits do not admit analytical solutions and the failure criteria involve more complicated metrics. Consequently, we resort to numerically modeling the EM stochasticity using a Monte Carlo (MC) analysis. Each MC trial models a cascade of EM events to successively degraded states. In each trial, a TTF sample is generated for each component, based on the component failure CDFs. Starting from the lowest TTF, each iteration in a trial includes the next lowest TTF. Just like in previous section, an EM event on a component is modeled by catastrophic increase in its resistance, essentially an open circuit. Consequently, every such EM failure causes:

1) ***Current crowding*** – which changes the wire failure CDFs and also causes additional Joule heating in the surviving components
2) ***Changes in circuit performance*** (here, the delay) due to EM failures, which could impact clock grid metrics such as skew.

Moreover, while some EM events may result in functional failure, others may result in only a small performance change due to redundancies in the circuit.

We incorporate both the effects through MC in our model. The former is well comprehended using the formulations of section IIB. The component failure CDF is an unshifted lognormal before the first failure, and must be modified using Eqs. (8), (9) subsequently. The latter effect of circuit performance change in each iteration is computed by conducting a SPICE-based delay analysis.

The iterations in an MC trial stop when the cumulative impact of the failures makes the circuit delay degradation unacceptable (*e.g.*, 10%). The corresponding time instant becomes the TTF of the circuit. Note that depending on the circuit functionality and layout, multiple component failures may be required to reach circuit failure. Eventually, a large number of such trials is conducted (which depends on the desired confidence level for estimation-error to be lower than specified) to obtain the circuit failure CDF. For this work, we keep a limit of 100 on the MC trials. The final algorithm is summarized as below:

Algorithm 1 Monte Carlo based approach for stochastic EM analysis

```
Input: Original SPICE netlist of the CUT (circuit-under-test),
testbench for currents, delay measurement; random number generator
Output: CDF of the circuit (probabilistic TTF)
Variable: $mc_i$: number of the Monte Carlo iteration
1.  Set $mc_{limit}$ based on desired accuracy
2.  for ($mc_i$ =0; $mc_i$++; $mc_i < mc_{limit}$) {
3.  t=0 SPICE simulation of CUT → currents through all resistors
4.  use random number generator to assign TTF for all resistors.
5.  rank order the resistors in the TTF manner; EM event on
    resistor with least TTF.
6.      while (circuit-delay degradation < specification) {
7.          recalculate the new current flow in the resistors
8.          TTF-rank order resistors; EM on resistor with least TTF
9.      }
10.     report circuit-TTF
11. }
12. rank-order various TTF to generate circuit CDF
```

The WLA analysis is also conducted using stochastic MC analysis, but the first component failure is assumed to cause circuit failure. It’s easy to see that the WLA based TTF is available in step 3) of the above algorithm.

A. *Monte Carlo Framework Based Clock Buffer Reliability Analysis*

We now apply the MC framework to a single 28nm 32x-drive clock buffer from an industry library, driving a lumped load at 1GHz frequency (Fig. 7). Here, the only candidate EM sites are the intra-cell power/signal resistors.

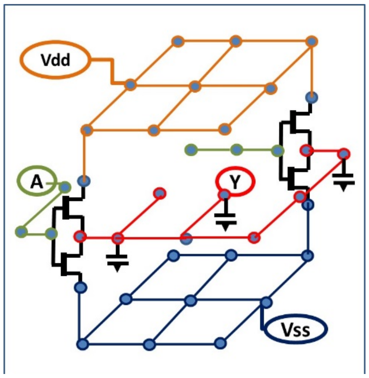

Fig. 7. A simple high-drive (32x) buffer driving lumped load. Shown are Vdd, Vss, input and output resistors (sites for EM), analyzed stochastically.

Note that the redundancy in this cell arises from: (a) parallel M1-M2 lines connected to the supply, so that an EM event in one metal level may still allow the cell to be functional (b) failure in the output line can result in a lowering of the cell power (e.g., from 32x to 30x), which alters the delay but maintains functionality.

It must be noted that while such cell-internal segments (on M1/M2) are much smaller in length, the Blech length benefit is typically not applicable as these segments carry purely AC current [8][9].

Using the MC framework, we generate the circuit failure CDF for the 32x buffer, shown in Fig. 8.

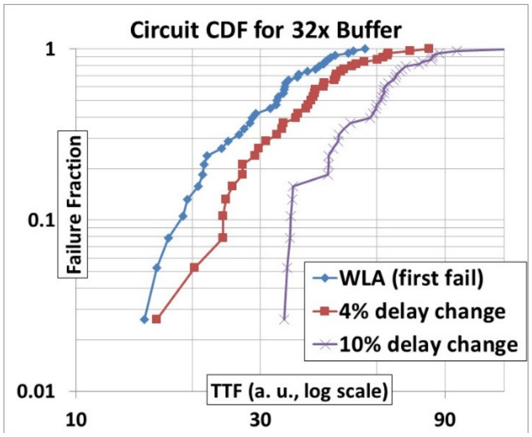

Fig. 8. Circuit CDF showing the failure rate evolution in a single 32x drive buffer circuit (of Fig. 7), driving lumped load

The framework is exercised under varying extents of acceptable delay degradations (shown here for 4% and 10%). Here, a relaxed specification implies acceptability of several EM events in the circuit. For a 10% fail fraction, the benefit from the inherent circuit redundancies is apparent in form of 2X margin in TTF over WLA.

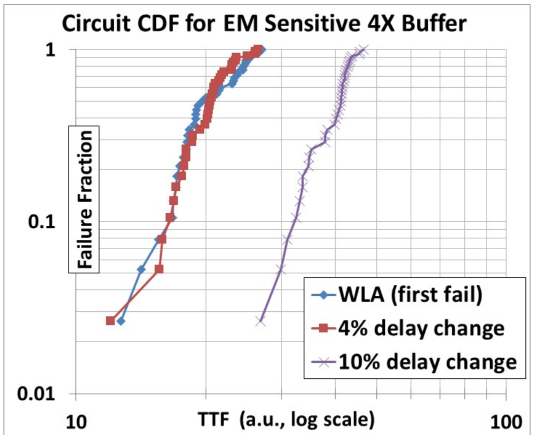

Fig. 9. CDF for a low-drive 4x circuit, where circuit redundancies reduce, leading to an early delay-based EM failure

We repeat this for the circuit failure CDF (Fig. 9) for a 4x-drive buffer driving a correspondingly lowered target load at 1GHz. Since this circuit has fewer redundancies, and correspondingly lower margins due to tighter layout, we see that its failure CDF is closer to WLA.

B. *Monte Carlo Framework Based Analysis of Buffers in Redundant Configuration*

We now look at the failure evolution in the case when two high-drive buffers are arranged in a redundant configuration. We again note that while WLA predicts complete system failure as soon as the first metal fails, in reality, even a delay degradation in a single buffer does not necessarily mean system failure, when several buffers are arranged in a redundant configuration. Indeed, if a buffer delay increases, its switching burden is placed on the alternate buffer, thereby moderating the impact. In Fig. 2, if *Buf1* degrades, then *Buf2* compensates for it. We study this particular configuration through the MC framework and present the CDF of the a) the individual buffer and b) the redundant buffer configuration in Fig. 10 below.

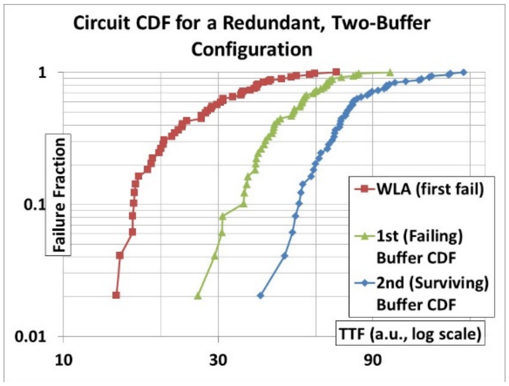

Fig. 10. CDF for a system with two buffers arranged in a redundant configuration (as in Fig. 2). Significant margin is shown between TTF and the failure of first buffer. Margin builds up with addition of one more redundant buffer.

As we can see, even in this case, the system continues to work even after the first resistor fails or after the first buffer fails entirely. Note that for this analysis, the failure criterion is the degradation in the slack.

IV. Clock Skew Estimation

We now apply the MC framework to analyze a clock grid structure. In the clock grid, the redundancies lie within the cells, in the power grid, and in the clock grid itself that is driven by multiple buffers. We consider a one-level clock grid (Fig. 1), with an exemplary buffer and its four identical neighbors to the north, south, east, and west, implemented with 28nm proprietary libraries, at 1GHz. In our example, wire widths in the clock grid are large so that the likelihood of EM failure is negligible and we focus on failures that may occur in within-cell wires or in the power grid (Fig. 7).

A primary figure of merit for a clock grid is the skew, or difference in arrival times at sink nodes in the grid. For our system, we translate the skew criterion to a delay criterion, and constrain the allowable degradation of a buffer and its neighbors. We enumerate a set of ways in which the skew specification can be met even after the buffers degrade:

1) When all of the five neighboring buffers degrade by less than 2%.

2) When all of the five neighboring buffers degrade in a similar, bounded manner (*e.g.*, between 2%-4%, 4%-7%, or 7%-10%).
3) When a buffer degrades by over 10% and all of its neighbors degrade by no more than 2%, or when a buffer and one neighbor degrade by over 7% and others by under 4%.

This is not an exhaustive list of all cases where the system operates correctly. Thus, failure analysis based on these criteria is pessimistic.

In order to proceed, we reproduce the probabilistic delay degradation CDFs of individual buffers (Fig. 8), as Fig. 11 below. This data from the individual buffer enables us to estimate the failure probability, at any given time, with any given failure criteria (say x% delay degradation).

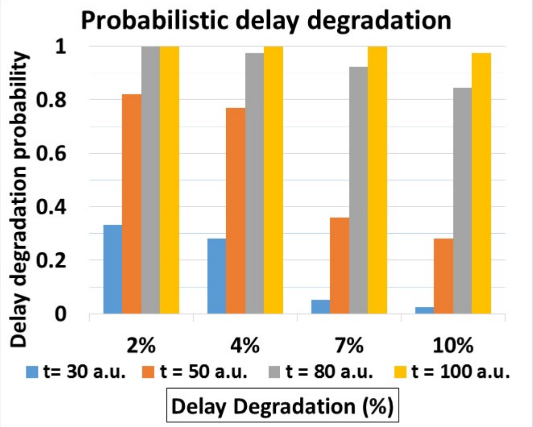

Fig. 11. Probabilistic delay degradation with time (column-cluster represent various times)

Consequently, we can use these relationships to arrive at the failure probabilities for the individual cases enumerated above and therefore for the effective skew-failure probability as:

$$\begin{aligned}
P_1&: \quad (1-F_{2\%})^5 \\
P_2&: \quad (F_{2\%}-F_{4\%})^5+(F_{4\%}-F_{7\%})^5 \\
&\qquad +(F_{7\%}-F_{10\%})^5 \\
P_3&: \quad {}^5C_2(1-F_{4\%})^3(F_{7\%}-F_{10\%})^2 \\
&\qquad +{}^5C_1(1-F_{2\%})^4F_{10\%} \\
&F_{skew}=1-(P_1+P_2+P_3)
\end{aligned} \qquad (14)$$

where $F_{x\%}$ represents the CDF of each buffer, representing the probability that the delay degradation is more than x%, and P_1-P_3 are pass-probabilities for above cases. The associated CDF is as in Fig.12.

Even in this case, the benefit from system redundancies – in form of multiple buffers – is apparent, as WLA turns out to be significantly pessimistic. Our method brings out >2X margin in TTF, wherein system failure is attributed in a more accurate manner of the skew. Such a margin can be further improved, by accurately incorporating the arrival times at each sink node, along with the logical correlation.

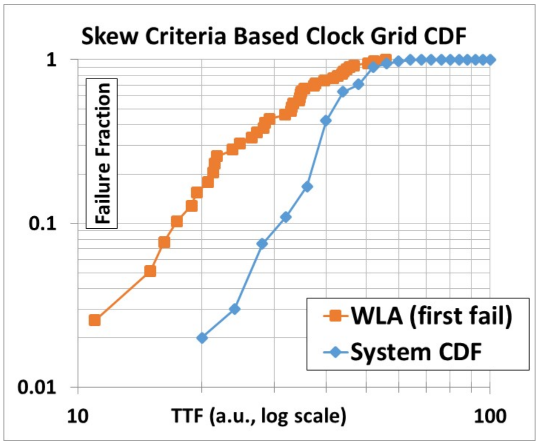

Fig. 12. Skew-criteria based CDF of the clock-grid. For a 10% FF, about 2X margin exists between WLA and skew-criteria based failures.

V. Conclusion

A novel method of assessing EM is presented in this work, which exploits the inherent randomness of the phenomenon, along with the system redundancies and connects component failure to the system impact. We use Monte Carlo based framework to model the stochasticity of EM and SPICE based methods to continuously monitor the system level impact of EM events. Using this method, we demonstrate > 2X margin wrt WLA based TTF estimates on circuits like high-drive clock buffers and assess significant margin between WLA and skew-criteria based TTF of a clock grid.

References

[1] L. R. De Orio, C. Hajdin, and S. Selberherr, “Physically based models of electromigration: From Black’s equation to modern TCAD models,” *Microelectronics Reliability*, vol. 50, no. 6, pp. 775-789, June 2010.

[2] J. Black, “Electromigration failure modes in aluminum metallization for semiconductor devices,” *Proceedings of the IEEE*, vol. 57, no. 9, pp. 1587-94, September 1969.

[3] K.-D. Lee, “Electromigration recovery and short lead effect under bipolar-and unipolar-pulse current,” *IEEE International Reliability Physics Symposium*, 2012, pp. 6B-3.1-6B-3.4.

[4] D. F. Frost, and K. F. Poole, “Reliant: a reliability analysis tool for VLSI interconnect,” *IEEE Journal of Solid-State Circuits*, vol. 24, no, 2, April 1989.

[5] H. Qian *et al.* “Subtractive router for tree-driven-grid clocks,” *IEEE Transactions on Computer Aided Design*, vol. 31, no. 6, pp. 868-877, June 2012.

[6] H. Su, *Design and Optimization of Global Interconnect in High Speed VLSI Circuits*, PhD Dissertation, Dept. of Electrical Engineering, University of Minnesota, 2002.

[7] A. Todri and M. Marek-Sadowska, “A study of reliability issues in clock distribution networks,” *IEEE International Conference on Computer Design*, 2008, pp. 101-106.

[8] P. Jain and A. Jain, “Accurate current estimation for interconnect reliability analysis,” *IEEE Transactions on VLSI Systems*, vol. 20, no. 9, pp. 1634-1644, September 2012.

[9] K.-D. Lee, *Electromigration Critical Length Effect and Early Failures in Cu/oxide and Cu/low k Interconnects*, PhD Dissertation, Dept. of Electrical Engg., University of Texas at Austin, 2003.