UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

A thesis presented for the degree of
Doctor of Philosophy

Algorithms and Methodologies for Interconnect Reliability Analysis of Integrated Circuits

Palkesh Jain

Advisor: Prof. Jordi Cortadella
Computer Science Department,
Universitat Politècnica de Catalunya

Co-Advisor: Prof. Sachin S. Sapatnekar
Department of Electrical and Computer Engineering,
University of Minnesota

Barcelona, Spain, 2017.

Abstract

The phenomenal progress of computing devices – from room-sized machines of 1940's to literally invisible cloud based applications, has been made possible, largely by the sustained efforts of semiconductor industry in innovating techniques for packing increasing amounts of computational circuitry into smaller and denser microchips. Indeed, gigantically-integrated-circuits today contain billions of closely-packed transistors and possibly multi-billion interconnects which enables these tiny transistors to talk to each other (needless to mention, at Gigahertz+ frequency) – all in a space of few mm^2.

Such aggressively downscaled discrete-IC-components (transistors and interconnects) silently suffer from increasing electric fields and impurities/defects during manufacturing. Compounded by the Gigahertz switching, the challenges of reliability and design integrity remains very much alive for chip designers, with Electromigration (EM) being the foremost interconnect reliability challenge.

Traditionally, EM containment revolves around three aspects, first of which is generation of EM guidelines at single-component level, whose non-compliance means that the component fails. Failure usually refers to deformation due to EM – manifested in form of resistance increase, which is unacceptable from circuit performance point of view. Second aspect deals with correct-by-construct design of the chip and lastly, it is about the final verification of EM reliability of taping-out silicon. Interestingly, chip designs today have reached a dilemma point of reduced margin between the actual and reliably-allowed current densities, versus, comparatively scarce system-failures. Consequently, this research is focussed on improved algorithms and methodologies for interconnect reliability analysis enabling accurate and design-specific interpretation of EM events.

In the first part, we present a new methodology for correct-by-construct design and verification of logic-IP (cell) internal EM verification: an inadequately attended area in the literature, unlike its counterparts in form of cell-external (signal) or power network EM. Our SPICE-correlated model helps in evaluating the cell lifetime under any arbitrary reliability specification or operating conditions, without generating additional data – unlike the traditional approaches. The model is apt for today's fab less eco-system, where there is a) increasing reuse of standard cells optimized for one market condition to another (e.g., wireless to automotive), as well as b) increasing 3rd party content on the chip requiring a rigorous sign-off. We present results from a 28nm production setup, demonstrating significant violations relaxation and flexibility to allow runtime-

level reliability retargeting.

Subsequently, we focus on an important aspect of connecting the individual component-level failures to that of the system failure. We note that existing EM methodologies are based on serial reliability 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 – in perspective, we present algorithms for EM assessment, which allow us to incorporate and quantify the benefit from system redundancies. With the skew metric of clock-grid as a failure criterion, we demonstrate that unless such incorporations are done, chip lifetimes are underestimated by over 2x.

This component-to-system reliability bridge is further extended through an extreme order statistics based approach, wherein, we demonstrate that system failures can be approximated reasonably by an asymptotic k-th component failure model, otherwise requiring costly Monte Carlo simulations. Using such approach, we can efficiently predict a system-criterion based time to failure (TTF) within existing EDA frameworks based on component level verification.

The last part of the research is related to incorporating the impact of global/local process variation on current densities as well as fundamental physical factors on EM TTF. Through Hermite polynomial chaos based approach, we arrive at novel variations-aware current density models, which demonstrate significant margins ($> 30\%$) in EM lifetime when compared with the traditional worst case approach.

The above research problems have been motivated by the decade-long work experience of the author dealing with reliability issues in industrial SoCs, first at Texas Instruments and later at Qualcomm. At TI, he led the Reliability CAD team with the charter of solving reliability issues for TI's various worldwide businesses including ASICs, DSP, Wireless and Automotives. Some of the related work which he published earlier is listed as: [JJ12, JCPA14, PJK10, Jai07].

Acknowledgement

This *journey* towards obtaining a doctoral degree, after more than a decade of professional career, wouldn't have been possible without the help, support and inspiration from a lot of different people in my life. I am grateful to be able to acknowledge their contributions here.

First and foremost, I have been fortunate enough to be guided by not one, but two thesis advisors: Professors Sachin S. Sapatnekar and Jordi Cortadella. Although a remotely done PhD, in no means has it been less rigorous than a fulltime one: with around a thousand email exchanges, and, each email warranting a high quality and meticulously thought-through response. Its indeed a privilege to get such mindshare, and, tremendous amount of patience for my writing from Sachin. His insights, insistence on simplicity of thoughts, focus on details and constant drive for precision have ensured the quality output, and, forced me to think harder about the problems and their solutions. He has been an exceptional personal mentor as well, investing a great deal of time in several emails on overseeing my general professional well-being too. Ofcourse, Sachin sets incredibly high standards of operating oneself personally and professionally. Even though it was testing on few occassions in the course of PhD, today, I truly appreciate and will consider myself accomplished to even emulate a small bit of that in my life. Thanks Sachin, for giving me this opportunity and guiding me through the very end.

The other half of the gratitude goes to Jordi. He gave me complete freedom to choose the problems and has been extremely accommodating on the part time research arrangement. I have benefitted greatly from Jordi's focus on big picture (beyond interconnects) and his uncanny system level expertise. Besides keeping an eye on the thesis progress, he was always available to help with any issues I faced, and was very prompt with feedback on the various drafts that I sent. Jordi, Sachin, working with both of you has been an honour and I look forward to our future collaborations. Thanks are also due to the research committee members and reviewers for their feedback. From the research group, would like specially thank Vivek Mishra for all the enriching brainstorming we did.

This thesis has been primarily motivated by the live reliability problems faced in industry, based on my work experience – first with Texas Instruments and later with Qualcomm. I want to express my gratitude to my colleagues, friends and management at Texas Instruments: Guru Prasad, Suravi Bhowmik, Subash Chander, Arvind NV, Gaurav Varshney, Jay Ondrusek, Vijay Reddy, Srikanth Krishnan, Anand Krishnan and Rob Baumann. Indeed, the exciting reliability journey in TI only stemmed from discussions with Frank Cano and Hugh Mair, and, my thoughts on EM wouldn't have started but for the great conversations with Young-Joon Park and Ki-Don Lee. At Qualcomm, I have been fortunate to have extremely supportive management and

must thank Venugopal Puvvada, Esin Terzioglu, Manoj Mehrotra and Rajagopal Narayanan.

Back to where it all started, I appreciate my professors at IIT Bombay who introduced me to the fascinating field of semiconductors. In particular, I'll forever remain indebted to Prof. J. Vasi for cultivating the required mindset.

Undoubtedly, work and research wouldn't have been possible without the calm and joyful backdrop which my family and friends in Bangalore, Indore and Hazaribagh provided. Can't thank them enough! Specifically to Coral - who more as friend helped on the philosophical aspects of the 'doctor of philosophy', and above all to my Mom for her unconditional love and blessings.

If there is one discernible change which has happened over the course of PhD, it is turning of our little princess, Kopal, from cuddly age of two years to that of bubbling five! Ever since she has been in my life, every day has been special. Her laughter and being around has made writing this thesis a very pleasant exercise. Lastly, words fall short in thanking my loving wife Kuhu, for her patience with me throughout the course of my doctoral work. I'm truly fortunate to have her boundless love and calm nature by my side. Undeniably, its her and Kopal, who have sacrificed the most in forms of the countless weekends, festivals and holidays over which this part-time research was primarily done and I can only promise to make up for the time gone!

When there are so many people to thank, one should thank the almighty ... here I do, for giving me such wonderful friends, colleagues, teachers and family!

Palkesh

To my Dad ... who has been the inspiration behind this, and,
to Kopal ... who has been the motivation to finish this (soon!)

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

1.1 Motivation

Computing is an indispensable part of our lives today, including our working, industrial and government interactions. The devices which perform this computing have pushed their individual performance and the power envelopes to the edge, while their counts have already surpassed the number of human beings on the planet, and by 2020, such internet-connected devices are expected to number more than 50 billion [IOT15].

This impressively sustained-and-articulated improvement in computer hardware has been supported by significant innovation in various fields, a facet of which includes periodic doubling of transistor densities in integrated circuits over the past fifty years – a phrase more commonly known as Moore's law. However, the physics, and the economics of such doubling gets tricky by the collateral reduction in the transistor gate length and interconnect wire dimensions. For example, the gate length of the state-of-the-art active device is in the range of 20nm and the minimum pitch of metal wires is 54nm, [ITR15]. Moreover, not just scaling, but several other dimensions are required to be addressed to keep pace with the desired performance-power envelope, including novel transistors (FinFETs) [HLK+00], circuits, architecture (parallelized, bigLITTLE, etc. [Gre11, Gra03]), EDA methodologies, systems and softwares.

Such large scale integration of devices and interconnects come with significant challenges revolving around the timing analysis, physical design (thermal), power delivery challenges and voltage limits regulation/management. Consequently, due to conflicting requirement of increasing performance and lowering power, we see a) narrow interconnects with much higher current densities and b) transistors with rising electric fields and quantum effects such as discrete dopants and gate oxide traps. As a result, device reliability mechanisms have become pressing concern in scaled technologies. While transistors degrade temporally due to aging caused by effects like Negative Bias Temperature Instability (NBTI), Time Dependent Dielectric Breakdown (TDDB) and hot-carrier effects,

the interconnects also degrade due to various effects, including Electromigration (EM) and stress-migration, etc.

Undoubtedly, EM is the most prominent interconnect degradation mechanism and is the central topic of research for this work. While EM is difficult to accelerate in-house on real products, it is not very uncommon to relate and attribute the field failures directly to it. Indeed, a recent product recall by Intel 6-series chipset using 65 nm technology has been attributed to Electromigration, with the consistent failure signature being "aggressive data transfers over time causing more errors" [ELE11]. Such product recalls are obviously costly, sometimes in range of $>$ \$1B, and therefore, any amount of investment upfront in designing better and robust ICs is much more economical.

Electromigration is a process in which mass transport takes place as a result of interaction between the moving electrons and the metal atoms (for example Copper or Aluminium) at high current densities [Ori10]. Also, under identical conditions of geometry, current and temperature, the rate of EM degradation depends on the specific microstructure of the metal line/via. As a result, and due to random manufacturing variations, the time-to-failure is a random variable, which brings in the stochasticity into lifetime assessment. Generally, an EM-induced failure of a metal line (wire, or interconnect) is deemed when the line-resistance changes by over a specified magnitude (for example 10%), and usually there is an upper limit on the total number of such failed wires in the design, also known as the target failure-fraction (FF), which directly relates to the defective parts per million (DPPM). Additionally, EM degradation depends very strongly on the operating conditions – namely, temperature, voltage and the overall stress time, which, along with the expected failure-fraction, constitute the reliability specification. Such specification is fundamental for IC vendors, since it is sensitive to the end-market. For example, ICs going into the wireless handheld devices rarely push EM to its limit (due to lower operating temperatures and relaxed fail/lifetime requirements), while the ICs going into automotive or server applications demand very low failure rate from EM, and that too at the harsh conditions, as seen from Fig. 1.1.

Having said this, EM is an old field of study with a strong tradition, including theoretical analysis, failure models, and full-chip estimation/checking techniques. Specifically, its verification is broadly around:

- ***Accurately computing the current densities*** in the individual wires of the chip. These wires are cell-external: signals and power nets connecting cells and cell-internal: wires within a logic-IP (standard cells) or mixed signal IP block, as seen in Fig. 1.2 [JJ12].

- ***Arriving at the current density targets*** for individual wires of a given chip. This involves careful understanding of the reliability specification, as well as the accounting of the individual wire geometry and surroundings. Additionally, while the EM reliability at component level is well specified by the foundries, it is the full chip level system reliability which gets specified to the chip designers, warranting some manipulations to derive

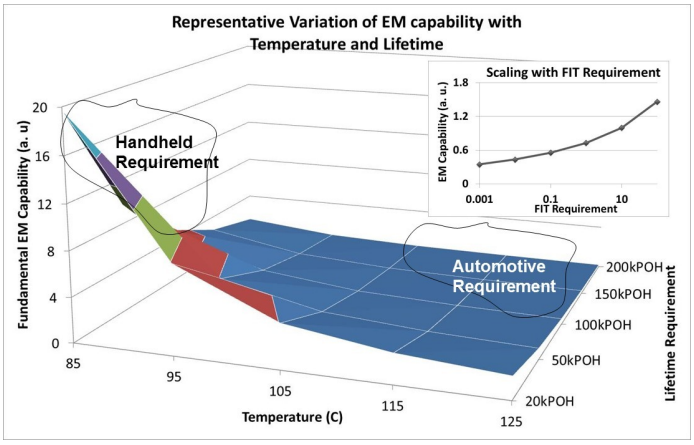


Figure 1.1: Variation of EM capability (allowed current density) across various markets.

component level targets. In short, the current targets are a function of: the system topology and the system level reliability requirements.

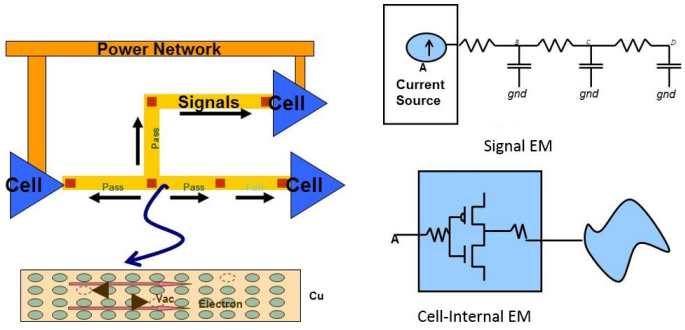


Figure 1.2: Basic Electromigration process and Electromigration in signals, power-network and cell-internal.

1.2 Existing Approaches for Electromigration

Revolving around the two themes discussed, the existing approaches for EM containment can be categorized as follows:

1.2.1 Current Computation

Cell-External: Signal and Power Analysis

The current flow computation in the cell-external signal wires and power network is routinely performed through industrial EDA tools which perform the IR drop analysis [RED15, EDI15, mag14]. The power network currents are a collateral of the IR drop analysis, although, the vectors at which the currents get computed can get slightly changed. For signals, the current computation is slightly more involved as the RMS currents through the signal wires are a function of the current waveform. However, using timing-like techniques of asymptotic waveform estimation, such a problem is well managed and is an industry practice [JJ12]. The focus of signal EM has been majorly on the clock network due to high switching activity as compared to the data nets. Due to this, the physical routing of the clock distribution is almost always on the non-default routing. Indeed, a rich field of literature exists on accurate current computation for cell-external clock and power distribution network [TMS08, Raj08, BOZD03, Lie13, YWC$^+$06].

Cell-Internal Analysis

On the contrary, the current computation for wires within the standard cells (Fig. 1.3a) haven't seen an adequate amount of research. The lack of same has been often compensated by a pessimistic design of the standard cell library which have wider than required routes and track heights. However, with the technology and area shrinking, this overdesign turns unacceptable and indeed, there is a recent string of work on cell internal EM analysis [Dod15, US14]. For example, Vaidyanathan *et al.* [VLSP14] highlights a standard cell library design methodology in which only selective routes within the standard cell are widened and how the library architecture itself is EM-aware.

Since the EM reliability of a standard cell is a function of the output load and operating frequency, some industrial implementation tools [EDI15, mag14] use a precharacterized table that models the tradeoffs in operating load and frequency, as shown in Fig. 1.3b. The intuition behind such a table (frequency versus load; f-L) is simple: the current flow inside the standard cell increases with the operating load, and hence the frequency should be lowered to meet the reliability specification. In fact, in the standard liberty file syntax, such model has been documented in standard manner [Lib16] to be readily consumed by place and route tools (Fig. 1.3c). This model has been used at the chip level to determine the safe frequency (fsafe) of an instance for any design/reliability parameter, and then make corresponding design fixes [RLC16]. Needless to say, most of the EM-critical cells are the ones that operate at higher loads, frequencies or slews. Similar model (Fig. 1.3d) has been reported by Sharma *et al.* and used on industrial designs [SKK14].

A different perspective on cell-internal modeling was presented by Panda *et al.* [OHG$^+$04] which entails pre-characterization of cell internal currents and using them for total failure rate projection. Burd *et al.* also presented an

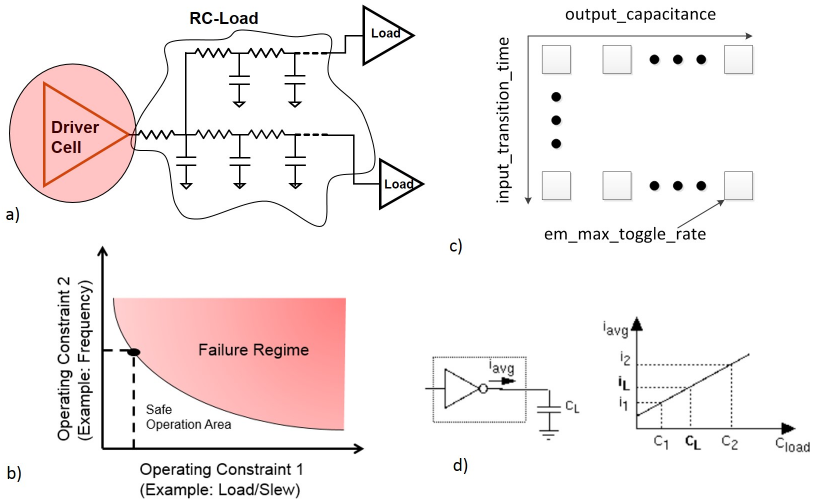


Figure 1.3: Spectrum of cell-internal EM modeling. a) Circuit schematic highlighting the problem space. b) Representative f-L modeling [JJ12]. c) EM model captured in standard liberty syntax [Lib16]. d) Model used in some of the industrial designs [SKK14].

advanced and accurate current calculation methodology based on circuit tracing and computing the effective activity of every resistor in the circuit [BAK+13].

A live example of how the reliability specifications roll-into design process can be seen from Fig. 1.4, adapted from Bickford *et al.* [BB13]. As can be seen, the reliability specifications are primary to kick-start the library design process, which trickles down to the final design database.

Conditions for Current Computation

As the current flow in the interconnect comprises of the switching and leakage current together, it is a strong function of the semiconductor process strength and statistical process variations. However, the present industry practice is to typically assume strong transistor (which enforces sharp slopes: higher RMS currents as well as higher leakage) and interconnects with worst parasitic capacitances (which increases the charging cap.). Such guidance for corner are typically provided by semiconductor foundries in their reference flows [TSM16, GF16, ICF14].

1.2.2 Current Density Target (Thresholds) Computation

Once the currents are computed appropriately, the next step is to verify the current against the guidance. Needless to say, these guidance, or specifications, become fundamental to design closure, as again inferred from Fig. 1.4, where the box highlighted in the red indicates the current density targets or thresholds.

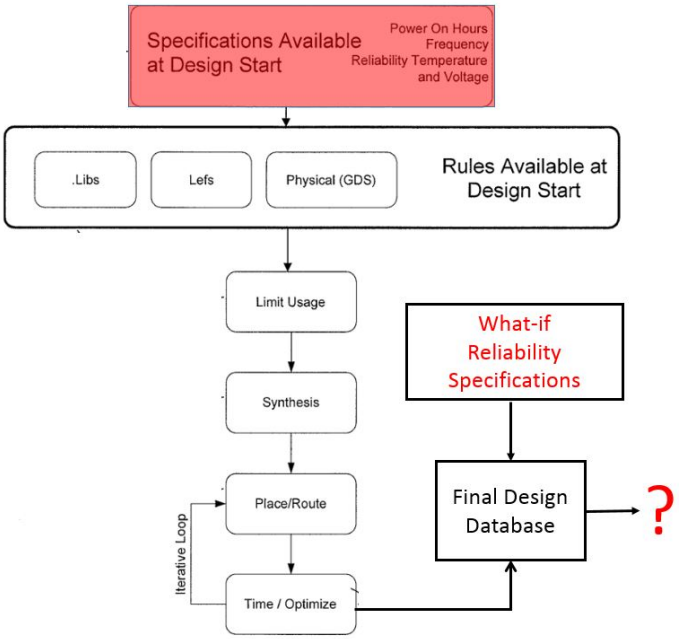


Figure 1.4: A representative design cycle in a typical semiconductor house: starting from obtaining reliability specifications, to standard cell library design to obtaining a final design database. Adapted from Bickford *et al.* [BB13].

The thresholds, in turn, are a function of several design variables, as listed below:

Topology Effects

EM verification methodologies have to deal with the dilemma that while the experimental test structures from which the current density limits are derived are discrete in nature, those limits are applied on circuits which are far more complex than a single individual interconnect. However, traditionally, designers make a simplifying approximation, in which the system is treated as a serial system, making it weakest link approximated (WLA), wherein, a failure is deemed when the first component fails [FP89]. While this approximation simplifies the checking, it completely ignores the system topology.

System Design

Another important consideration while choosing the current density thresholds is the design style itself. Indeed, an important parameter on which system reliability depends is the number of components (N_t) which are running at or near thresholds. Using serial reliability assumption, we can actually relate the chip and component level reliability (FF_{chip} and FF_0 respectively) for this case

as:

$$FF_{chip} = 1 - (1 - FF_0)^{N_t} \implies FF_0 \approx FF_{chip}/N_t \tag{1.1}$$

Since N_t is a design knob with the chip designer, a power grid style having a smaller N_t implies that FF_0 must be targeted $\approx FF_{chip}$. Alternatively, for higher N_t design style, FF_0 must be mandated to be smaller than FF_{chip}; usually 3 orders lower for typical power network comprising millions of resistors. In other words, a design which allows lots of wire to run near the current density thresholds must choose a smaller value of current threshold.

System Reliability Budget and Retargeting

Finally, the current thresholds are a strong function of the reliability requirements, namely: lifetime, joule heating criteria, temperature and the overall failure rate requirement. A relation in them is often provided by the foundries. However, as the semiconductor design houses explore newer markets for an already manufactured, functioning silicon, they are often forced to tradeoff the reliability variables [YTW14].

1.3 Limitations of Prior Art

While above methodologies have worked reasonably well till so far, with advancing technology and convoluted circuit effects, it is becoming more and more inadequate in accurately predicting EM safety. Specifically, we highlight following limitations with the prior art:

- *Cell-Internal EM Modeling:* the prior arts discussed till so far suffer from the inability to incorporate the impact of parasitic wire loading on the cell and instead abstract the reliability as a function of lumped load. Further, it cannot incorporate dependency of arbitrary switching rates on inputs pins and effects such as clock gating. Furthermore, the prior art does not have a simple way of transposing the reliability models derived from one specifications to the other. This forces a complete reliability characterization of library at new conditions [JCS16]. Indeed, if the reliability condition changes towards the later end of the design cycle (referring back to Fig. 1.4), existing capabilities lack in providing the designers with a quick answer.
- *Topological Considerations:* The prior art makes a simplifying assumption of treating the entire system as a chain, thus resorting to the weakest link approximation. It should be however noted that in circuits, variety of topologies occur, and often there is redundancy, which implies that the failure of the first component does not imply circuit failure. Additionally, for on-chip interconnects, the EM failure is context-dependent, *i.e.*, in some cases, even small changes in resistance may cause performance failures in the circuit and vice versa for others. For example, in clock-meshes

or in power grids, a large failure may be tolerated due to the inherent resilience in the circuit, where the failure of one wire may be compensated by current flow through other paths [JSC15, MS13]. Such topological considerations are valid for both: signal as well as power network and must be part of the EM verification method.

- *Variability Handling:* Interconnect currents comprise of switching and leakage, both of which are prone to statistical variations. In fact, with the advent of dark silicon, in our experience of industry chips, and as also noted by others, leakage can contribute up to 60% to the total power network current [MFT+08]. Further, such statistical process variations not only impact the current density, but also in the interconnect dimensions and the EM kinetics themselves. To our knowledge, such variations are not comprehensively addressed in prior art.

1.4 Thesis Objectives

In the previous section, we claim that chip designers use simple, deterministic and bounding approaches for current estimation as well as threshold computation, which leads to entitlement loss in modern technologies. The thesis aims to reduce such entitlement loss by capturing the essence of physics-based models into the chip design. The thesis project aims at accomplishing the following objectives:

- Algorithms and Methodology for Logic-IP internal Electromigration Management
  - ⋆ In order to overcome the limitations of prior art on resistive load modeling, we present a unique and decoupled approach in this work. Our method additionally incorporates the impact of voltage-dependent pin capacitance on EM.
  - ⋆ Our model provides an on-the-fly retargeting capability for reliability constraints by allowing arbitrary specifications (of lifetimes, temperatures, voltages and failure rates), as well as interoperability of the IPs across foundries.
- Statistical Variations and Electromigration Verification
  - ⋆ Using detailed multivariate Hermite polynomial chaos, we model the impact of **non-Gaussian global process variations** on current density and EM kinetics.
  - ⋆ As a practical application in present industrial framework, we present a direct usage of above global variations model in determining the current flow under various workloads.
  - ⋆ Incorporate the impact of **local process variations** on EM. Our results demonstrate that the wires on lower metal layers are highly susceptible to local variations arising from transistor leakage.

- Topologically Aware Stochastic Methodologies for Accurate EM Assessment in Clock and Power Grids

  ⋆ With specific application on clock grids, we present algorithms for EM assessment allowing us to incorporate and quantify the benefit from system redundancies. Taking the skew metric of clock-grid as a failure criterion, we demonstrate that unless such incorporations are done, chip lifetimes are underestimated by over 2x.

  ⋆ Above bridge between component and system reliability is further extended through an extreme order statistics based approach, wherein, we demonstrate that system failures can be approximated reasonably by an asymptotic k-th component failure model, otherwise requiring costly Monte Carlo simulations. Using such approach, we can efficiently predict a system-criterion based time to failure (TTF) within existing EDA frameworks based on component level verification.

Domain		**Prior Art**	**This Work**
Standard Cell EM Analysis	Cell-internal current estimation under arbitrary switching rates	Restricted to simple cells [Dod15, US14]. Burd *et al.* showcased complex circuit tracing for current consumption [BAK+13].	***Addressed*** through arc based current computations and demonstrated SPICE-like accuracy with significant speedup.
	Incorporating parasitic wire loading and voltage dependent pin cap	NA: leads to significant pessimism [JJ12]	
	Retargetability under arbitrary reliability constraints	NA: prior art leads to expensive recharacterization under new reliability constraints [BB13]	
Clock Distribution Network (Signal) Electromigration	Incorporating topological and contextual dependencies	NA: no prior study highlighting how redundancy impacts clock network EM	***Addressed*** and demonstrated possibilities of upto 2x pessimism reduction.
Power Distribution Network Electromigration	Global/local statistical variations into current estimation and EM kinetics	No direct studies; some in SRAM context [GMSN14] but present Monte Carlo framework	***Addressed*** through efficient polynomial chaos method and (> 30%) better EM lifetime.
	Fast topological assessment and impact of EM violations on system failure	NA: Monte Carlo based [MS13]	***Addressed*** through unique application of extreme asymptotic order statistics in non Monte Carlo manner

Table 1.1: Overall thesis contribution areas and comparison with prior art.

1.5 Organization

The document is organized as follows: in Chapter 2, we review the background and the primer material needed to build upon the theories around reliability and statistics. In Chapter 3, we address the logic-IP internal Electromigration

problem. Subsequently, in Chapter 4, we introduce the stochastic and topologically aware Electromigration methodology, followed by the detailed analysis of statistical variations and its impact on EM in Chapter 5. Later, we present the conclusions and scope for future work in Chapter 6.

Chapter 2

Background

In this chapter, we will review the basic background material, beginning with the relevant introduction of Electromigration followed by a quick recapture of the basic probability terms essential in understanding the stochastic reliability analysis. Subsequently, we review the Monte Carlo random sampling approach from an Electromigration point of view. Finally, we review briefly the clock grid structure and existing EM challenges in clock and power grids.

2.1 Electromigration basics

Electromigration is the mass transport of metal due to momentum transfer between electrons (driven by an electric field) and diffusing metal atoms. Such an event occurs when there is a flux divergence with regard to the movement of metal atoms, commonly at distortions-sites in the lattice, in the form of vacancies and/or grain boundaries [Lee13] which can be schematically represented by triple points (Fig. 2.1).

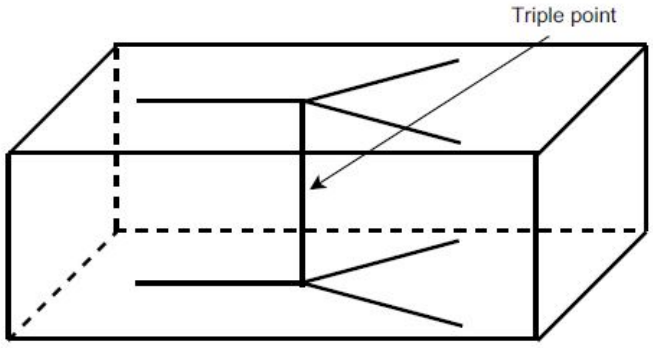


Figure 2.1: A typical triple point in a wire.

Flux divergence also arises when the flow of metal atoms into the region is

not equal to the outflow of atoms from the region [PJK10]. EM is associated with creation of either a void, or a hillock, as shown in Fig. 2.2 below.

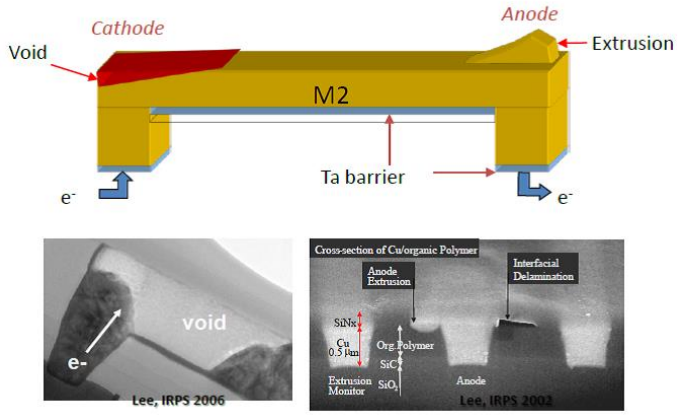


Figure 2.2: Voiding and extrusion in Copper metal lines during Electromigration [Lee13].

A void is created due to depletion of atoms (when outflow from a region is greater than the flow into it) and leads to open-circuits or unacceptable resistance increase in a line. Whereas a hillock is created due to accumulation (when the flow into a region is greater than the outflow) of atoms and usually causes short-circuits between adjacent lines and inter-level conductors.

In this work, we assume for simplicity that all interconnect failures due to EM are caused by nucleation and/or growth of voids. Based on a very simple model, Black [Bla69] was the first to derive an expression for the time to failure of a metal line subjected to electromigration. He considered that the mean time to failure, t_{50}, is inversely proportional to the rate of mass transport, R_m,

$$t_{50} \propto 1/R_m \tag{2.1}$$

and that the rate of mass transport is proportional to the momentum transfer between thermally activated ions and conducting electrons,

$$R_m \propto n_e \Delta P N_a \tag{2.2}$$

where n_e is the density of conducting electrons, ΔP is the momentum transfer from the electrons to the metal atoms, and N_a is the density of thermally activated ions. Assuming that both the electron density as well as the momentum transfer are proportional to the current density, J,

$$n_e \propto J; \Delta P \propto J \tag{2.3}$$

and that the activated ions follow an Arrhenius law with the activation energy, Q,

$$N_a \propto exp(-Q/k_B T) \tag{2.4}$$

Consequently, the mean time to failure can be represented as:

$$t_{50} = A\frac{e^{Q/k_B T}}{J^n} \tag{2.5}$$

Here, J is the DC current density through the line, Q is the activation energy, k_B is Boltzmann's constant, n is the current exponent (typically between 1 to 2), and A is a fitting parameter. It was observed that not all experimental results followed the otherwise quadratic dependency of J. However, they could be fitted by allowing a variable current density exponent, n, which must again be experimentally determined.

2.1.1 Alternative Electromigration Modeling Paradigms

At this point, it must be considered that EM failure occurs in two phases:

- Void nucleation: After a wire has been stressed, the depletion of atoms at the cathode creates a tensile stress. Once a critical stress threshold value has been crossed, the void nucleates.

- Void growth: After nucleation, further movement of metal atoms from the void results in void growth. This results in increased wire resistance due to the effectively reduced cross-section. If the void grows large enough, it may result in a break in the wire, resulting in either an open circuit or a vastly increased resistance, in cases where the current through the wire can flow through the higher-resistivity barrier layer of the copper.

Indeed, in the Black's equation (2.5), an exponent close to 1 indicates that the lifetime is dominated by the void growth mechanisms, *i.e.* the time for a void to grow and lead to failure represents the major portion of the lifetime [LK91], while a value close to 2 indicates that void nucleation is the dominant phase of the electromigration lifetime [KBT+93].

Eq. (2.5) has been used for lifetime estimation and extrapolation to operating conditions for 40 years now. However, in a recent publication Lloyd [Llo07] discussed the application of the modified equation, where nucleation and growth are explicitly accounted. If each of the processes is driven by the same driving force but exhibit different kinetics, the partition of the failure time is better represented as a joint function of driving forces as follows:

$$t_{50} = t_{nuc} + t_{growth} = \left(\frac{Ak_BT}{J} + \frac{B(T)}{J^2} \right) e^{\frac{Q}{k_BT}} \tag{2.6}$$

where A and B are constants that contain geometric information, such as the size of the void required for failure. Given any values for A and B, it follows that the relative contributions of nucleation and growth will vary as a function of the current density. At higher current densities the time to failure will proportionately be more growth than at lower current densities [Ori10].

Impact of Interconnect Topology

Additionally, it must be noted that fundamentally, void nucleation is driven by divergence of atomic flux which is typically highest at sites such as vias, contacts, or even points where the leads merge. Further, it has been reported in literature that even if the incoming atomic flux (signified by high current density) is high at such sites, the site itself may not fail due to it maintaining a low divergence; while a simple, individual-lead based Black's equation continues to predict failure for such a structure, as shown in Fig. 2.3a). This inefficiency has been recently revisited by various researchers resulting into evolution of alternative paradigms in EM checking [PJK10, ALTT04, GMSNL14].

A very clear demonstration of the topological effects can be seen through Fig. 2.3b). A two-lead system is shown here, where, the lead Y carries a current density of J, whereas the lead X a current density of $2J$. The conventional wisdom assumes the two leads are isolated in terms of EM reliability. However, they are physically connected and their atomic flux indeed interacts. Thus, the actual divergence effect at the 'via node' A and 'via node' B is not purely determined by the local current density of the each lead X and Y, respectively. The lead EM interactions also affect the divergences at the via nodes. The lead Y, on the other hand, has a current flow of J but the current keeps flowing into X with a larger amount of $2J$. The atomic flux from the lead Y does not piles up at the via node A but transfers into the lead X at faster rate due to the higher current density. This eventually increases the atomic depletion rate from the via node B and make the EM lifetime of the lead Y short. Indeed, while the conventional checks declare lead X to fail, it is lead Y which fails EM first, as also captured through the via-node vector method.

Fundamentally, such alternative methods rely on computing some form of atomic flux divergence at EM-probable sites and subsequently comparing them against set thresholds. One such method, as reported in [PJK10] is vector via-node based method, wherein the physical and directional interactions amongst various leads is incorporated to perform the reliability verification.

Notably, the fundamental inputs required to perform these effective divergence calculations still remain the individual current density in every single interconnect of the circuit, along with additional information like the circuit topology. For example, for the case considered in Fig. 2.4, all the individual

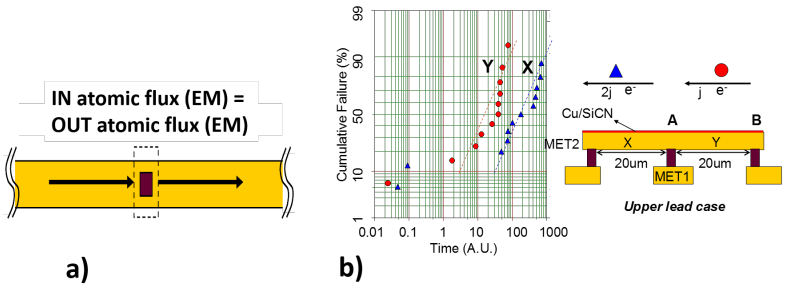


Figure 2.3: a) As the incoming and outgoing atomic flux are the same, the atomic flux divergence along the interconnect is zero, resulting in no EM damage b) Lead Y carries half current density of lead X but fails faster than lead X, contrary to the conventional EM expectation. The via-node vector EM method catches this using flux divergence [PJK10]

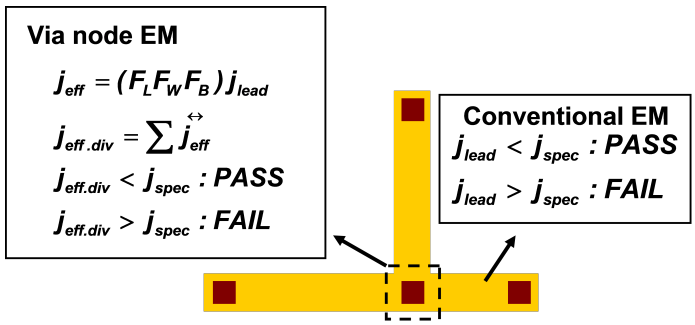


Figure 2.4: The via-node method compares the effective current density of a via node with the current density thresholds while the conventional method compares the current density. The via-nodes are most susceptible to the EM failures and the effective current density should be smaller than the thresholds. [PJK10].

lead currents are required to arrive at effective current density, along with additional factors like length, width and the interaction factors. The effective current density is in fact a product of these three factors to the individual current density, along with the vector sum. Notice that the conventional approach stop at checking individual current density against the threshold, while approaches such as these compute the effective current density, or divergence, and then compare it against the threshold. In summary, methods like these effectively capture the underlying physical phenomenon while still remaining in the realm of traditional Black's equation based verification framework.

2.1.2 Electromigration under Bipolar (AC) Currents

It must be noted that a majority of wires in the circuits carry non-DC currents, for example, the output segment of a switching inverter circuit. While a sim-

ple summation to get mathematical average current of the waveform results in close to zero estimate, it has been demonstrated that such wires do fail EM, albeit with a much higher EM lifetime [TMHM93]. Therefore, adjustments are required to (2.5). A standard approach is to compute an effective-EM current, which is a constant current value, derived from the line current waveform that gives the same lifetime for that line under the influence of EM. Such an effective current can be computed based on some assumed periodic current waveform. Additionally, it has been empirically noted that EM damage recovers with the reversal in the current-flow direction [TCH93,Lee12]. Such recovery is accounted by adjusting the current calculation in following manner:

$$J = J^{+}_{avg} - \kappa J^{-}_{avg} \tag{2.7}$$

Here, J^{+}_{avg} and J^{-}_{avg} indicate the average current density during current conduction in the positive and negative directions, respectively and κ is an empirically derived recovery factor (typically between 0.5-0.8). It has also been noted that the recovery phenomenon exhibits frequency dependence [TCH93], thereby meaning that wires with higher frequency tend to recover more than the lower one. However, this is still a subject of ongoing research [SHT15, HSK^{+}16, DFN06] and in our work, we will treat all the wires with uniform recovery factor, which is the present industry practice.

2.1.3 Accelerated Electromigration Under Wire Joule Heating

Finally, a collateral event to Electromigration is that of Joule heating, which is related to the increase in wire temperature due to the constant current flow [Hun97]. In our context, we can make a simplifying assumption that the heat flows only from the metal lead to the silicon substrate. The silicon substrate, in turn, has a maximum allowed junction reference temperature (T_{ref}). T_{ref} can be achieved during circuit design by ensuring that the package thermal impedance is able to dissipate the total power dissipation.

Using the steady state for quasi-one-dimensional (1-D) heat transport, we get:

$$J^2_{RMS} = \frac{(T_m - T_{ref})K_{ox}w_{eff}}{t_{ox}t_m w_m \rho_m(T_m)} \tag{2.8}$$

where, J_{RMS} is the RMS current density, T_m is the mean metal wire temperature, T_{ref} is the maximum allowed junction reference temperature in the silicon (for example 100 C), K_{ox} is the underlying oxide thermal conductivity, t_{ox} is the underlying oxide thickness, t_m is the metal thickness, w_m is the metal width, ρ_m is the temperature dependent metal resistivity, and w_{eff} is the effective thermal width. Above equation can be further simplified as follows, which relates

the increase in temperature (ΔT) over large scales of time, to RMS current in following manner, where c is a constant:

$$\Delta T = cJ_{RMS}^2 \tag{2.9}$$

Notice that above relation suggests that the wire temperature is a function of the RMS current, while on the other hand, EM reliability is a function of the average current, also accelerated by the wire temperature. In other words:

$$t_f \propto \frac{e^{Q/k_B T_m(J_{RMS})}}{J} \tag{2.10}$$

Pioneering work by Hunter [Hun97] combined the two effects of average EM fails and RMS-induced Joule heating in a self-consistent manner through the concept of duty cycles, making it possible to simultaneously check both conditions. If we define the effective duty cycle, r for a wire as the squared ratio of the average and the RMS current, we get:

$$r = \frac{J(T_m)^2}{J_{RMS}(T_m)^2} = J_{REF}^2 \left(\frac{e^{Q/k_B T_m}}{e^{Q/k_B T_{ref}}} \right) \frac{t_{ox} t_m w_m \rho_m(T_m)}{(T_m - T_{ref}) K_{ox} w_{eff}} \tag{2.11}$$

Since r is the squared ratio of the average and RMS currents, an unlimited number of current waveforms can have a given duty cycle. However, for a given r, there is only combination of RMS and AVG current which meets the Joule heating (ΔT) and the EM reliability together. Thus, Hunter's method then simplifies to calculating the duty cycle for any given wire and looking up the allowed (AVG, or, RMS) current density for that particular r.

2.1.4 Present Industry Approach

While above formulation (2.6) more suitably models the physics, it must be mentioned that industrially, Black's equation is still the workhorse model for EM verification [TSM16,GF16,ICF14]. As far as interconnect topologies are considered, formulations like via-node [PJK10] crisply break down the problem in current-computation and divergence calculations. Consequently, for this work, we a) keep our focus on Black's equation (2.5), b) restrict to topology-independent reliability verification and c) use a uniform recovery factor for all wires. All of above assumptions make us very relevant to the present approach in industrial EM verification frameworks from EDA vendors like Cadence, Synopsys and Ansys [EDI15, syn16, RED15], adopted all across the semiconductor industry and recommended by the semiconductor fabrication houses [TSM16, GF16, ICF14], thereby making our work directly applicable with minimal flow changes. Nevertheless, as we will show in later chapters, our formulations are flexible and could be extended to incorporate alternative paradigms of EM checking and to

interconnects with complex topologies.

2.2 EM Reliability Mathematics

A failure-event for a wire is defined as the event when the line-resistance increases by 10% (or some such number). The TTF of a line (from (2.5)) is in fact, subject to variations. Indeed, an experiment pertaining a set of wires, made of the same material, using the same manufacturing process and with identical dimensions, stressed with the same current density J at a temperature T_m for time t_0 will have a distribution of TTF. In other words, if we pick any single wire from the set, the only thing we know is the frequency distribution of the TTFs obtained from the experiment. To deal (rigorously) with such scenarios, we will be modeling the TTF of the lines using random variables.

2.2.1 TTF Distribution: Lognormal

The type of distribution function used to characterize the lifetimes arising from a wide range of failure mechanisms (including EM), can be established theoretically or empirically. There have been various proposals on the model, with maximum agreement around the usage of log-normal model and will be used throughout this work.

Indeed, the initial failure rate, $f(t)$, of each component is found to be lognormal, as represented by following relation:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\frac{\log(t)-\log(t_{50})}{\sigma}} \tag{2.12}$$

However, rather than the failure rate, it is the cumulative failure fraction fail fraction (FF) which is of higher importance. Indeed, FF follows a lognormal dependency on the time to failure (t_f, also known as stress time). The lognormal parameter (z), relates to the time-to-failure as follows, where σ is the standard deviation of the distribution, which is process-dependent:

$$z = \frac{1}{\sigma} \log \frac{t_f}{t_{50}} \tag{2.13}$$

$$FF = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \tag{2.14}$$

The FF to time relationship is also called as the cumulative failure distribution, whereas, the $f(t)$ relates the failure probability at any time instant t.

The cumulative probability distribution function (CDF: $F(t)$) is alternatively

represented by:

$$F(t) = \Phi\left(\frac{\log(t - t_{50})}{\sigma}\right) \tag{2.15}$$

with $\Phi(x)$ as the standard normal CDF.

2.2.2 TTF sample Generation

In order to statistically approach the EM reliability analysis, we would be needing to generate the TTF samples. Since $\log(t_f)$ has a normal distribution, we draw random samples from the normal distribution of z (with mean 0 and variance 1), as follows:

$$t_f = e^{t_{50}+z\sigma} \tag{2.16}$$

2.3 Monte Carlo random sampling approach

Besides the above log-normal statistics, we will be using Monte Carlo (MC) simulations to evaluate the EM lifetime characteristics of interconnects. In our Monte Carlo simulations, we generate N random lifetimes for N different interconnects and rank order these lifetimes to estimate the first time to failure [Lee03]. Such a trial is repeated sufficient number of times to get the CDF of the system. The underlying principle in the MC simulation based EM analysis is the fact that EM lifetime of a single interconnects is known to follow lognormal distributions. Thus, we can generate potential random lifetimes from the inverse of the CDF using determined or estimated lognormal parameters of single elements.

2.4 Clock grids: introduction and previous EM checking methods

Clock grids are one of the most promising approaches for clock design, when skew variation is of utmost importance, though, they come with the cost of area and power [Su02]. The skew reduction in such structures is a direct result of the high redundancy: due to presence of multiple source-to-sink paths, for every sink. Fig. 2.5 shows one such example of the clock network [Raj08].

The high frequency operation of clocks make such structures directly susceptible to Electromigration. Even though the clock mesh is designed to be a symmetric structure, still, there exists ways in which asymmetric currents can be drawn from different arms (for example, due to clock gating or process variations) and this underlines the need of accurate EM assessment.

Conventionally, a direct application of standard EM checking practices exist for clock mesh which involves a set of extraction, circuit simulation to extract

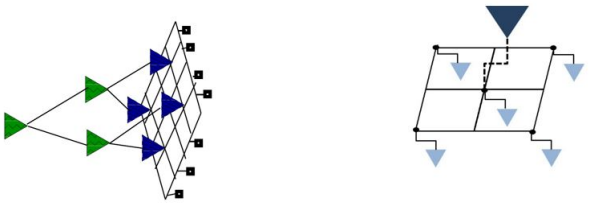


Figure 2.5: Representative clock grid, highlighting the redundant source-sink paths and the multiply driven system.

the currents and verification of the currents with the thresholds [YWC+06]. While such vanilla application of conventional EM checks on clock meshes completely ignore the natural redundancies of the clock mesh, making the analysis highly pessimistic [Raj08, TMS08]. More recently, there are also reports from foundry on circuit delay degradation due to Electromigration [HL16].

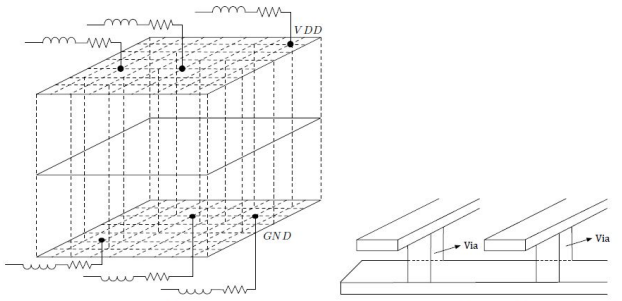
2.5 Electromigration in power grids: introduction and previous approaches

The power distribution network, commonly referred to as the 'power grid', is a multiple layer metallic mesh that connects the external power supply pins to the chip circuitry thus providing the supply voltage connections to the underlying circuit components. A typical power grid is shown in Fig. 2.6a, wherein the middle layer represents the device layer. The top and bottom layers represent the power and the ground networks, respectively, both of which could be modeled as RLC networks. Indeed, a 3-dimensional representation of the power grid is shown in Fig. 2.6b, which comprises of the various metal wires connected by the vias.

Ideally, every node in the power grid should have a voltage level equal to the supply voltage level. However, due to the parasitic behavior of grid and due to circuit activity and coupling effects, the voltage levels at the nodes drop below the supply level, commonly known as drop [Pan08]. With GHz switching, such voltage drops are approaching serious levels directly affecting the performance, reliability, and correctness of the underlying logic. Indeed, voltage drop reduction and overall power grid verification is one of the most major steps in today's design closure for large scale chips.

To make things worse, however, power grid rails are current carrying wires and therefore, also, suffer from electromigration. EM on power grids could be sometimes more serious than the one on clock tree due to the unidirectional current flow in the grid wires. The damage translates into sharp resistance increase, resulting into poor performance.

Consequently, a host of industrial EDA tools, [EDI15, mag14, RED15] check and contain the current densities in the power grid. Indeed, the power grid must be widened to accommodate the increasing current density; which is a direct



(a) an example power grid

(b) connection of metal wires of different layers

Figure 2.6: A typical power grid representation in modern SoC [Yu14].

result of increasing integrations on the chip and also with the advent of FinFET technologies, which naturally increase the current-density per unit area.

2.6 Conclusions

In this chapter, we reviewed the basics of Electromigration and the primary variables - notably, the current density, statistical distributions and also how the topological aspects of the circuits could become important for EM verification. The topological aspects become important for highly redundant circuits like clock grid and power grid, wherein, the traditional EM checking methods fail. These topics will now be addressed in depths in upcoming chapters.

Chapter 3

Logic IP-Internal Electromigration Assessment Methodology

In this chapter, we present a new methodology for SoC-level logic-IP-internal EM verification. The new framework significantly improves accuracy by comprehending the impact of the parasitic RC loading and voltage-dependent pin capacitance in the library model. It additionally provides an on-the-fly retargeting capability for reliability constraints by allowing arbitrary specifications of lifetimes, temperatures, voltages and failure rates, as well as interoperability of the IPs across foundries. The characterization part of the methodology is expedited through intelligent IP-response modeling. The ultimate benefit of the proposed approach is demonstrated on a 28nm design by providing an on-the-fly specification of retargeted reliability constraints. The results show a high correlation with SPICE and were obtained with an order of magnitude reduction in the verification runtime.

3.1 Introduction

As motivated in Chapter 2, Electromigration containment requires a) cell-external analysis for signals and power nets connecting to the cells and b) cell-internal analysis for wires within a logic-IP (standard cells) or mixed signal IP block. Recently, a great deal of innovation and improvement has been seen on the verification and design strategies for cell-external signal and power grid EM [Lie13, HYST14, JJ12, ITR15].

However, there has not been adequate focus on the robust design and reuse of the standard cells. Ensuring EM reliability for standard cells and IPs in a design implies that the exact context at which the IP is used must be bounded to guarantee its robustness in the design. This context could be stated in terms

of design limits (loads, slews, frequencies, supply voltage), or reliability (temperature, lifetime, or a failure rate specification tied to current density limits). Without rigorous assessments, a set of IPs designed for a particular reliability condition (e.g., 1.2V, 105C, 100k power-on hours (POH), 0.1% cumulative failure and 10C Joule heating (JH) limit) cannot be guaranteed to be EM-safe at another condition (e.g., 1.0V, 115C, 200kPOH, 0.01% cumulative failure and 15C JH limit).

Nevertheless, tradeoffs on these constraints are increasingly in demand in industry due to accelerated inroads of semiconductor houses into newer businesses with different reliability demands [YTW14]. For example, industrial designs demand more stringent operating conditions than traditional computing applications [JED16, Q1016]. From an EM standpoint, meeting these specifications is challenging, as seen from Fig. 1.1, which highlights the representative current density per μm^2 across various temperature and lifetime specifications. As can be seen, amongst the various environments, the current carrying capability becomes over 20x more stringent. Not only amongst different application markets, even for the same SoC itself, different complex IPs (e.g., CPU core or a DSP) can have different reliability requirements, based on their ON times and temperature specifications. The challenges increase when such reliability requirements could be only made available on-the-fly: that is, either during the final SoC verification or even after the SoC tape-out; in which cases, the original reliability targets for the IP, characterized for one application domain, may not match with the reliability requirements in a different domain.

One way to meet such diverse specifications is to approach the design in a bottom-up manner with a fresh logic-IP portfolio that meets targeted domain-specific reliability specifications. However, this is very expensive, and economic and design effort considerations often dictate that the product integration over all application domains be based on the same IP portfolio. This implies that the logic-IPs require a disciplined utilization procedure, making it important to assess their exact usage boundaries at arbitrary conditions.

A starting point towards this is to ensure that the cell is EM-safe at a specific load and frequency by selecting wire widths so that EM constraints are met. However, this only implies that a lower load and lower frequency can be considered EM-safe. The cell may (or may not) be EM-safe at a lower load and higher frequency, or a higher load and lower frequency, or a higher load and higher frequency, and this can only be uncovered through costly detailed analysis.

As an improvement, some industrial implementation tools [EDI15, mag14] use a precharacterized table that models the tradeoffs in various design/reliability parameters. Fig. 3.1a shows a representation of one such table, where x-axis represents an operating constraint of the cell (load here, but this could be slew, supply voltage, or any reliability constraint) and the y-axis represents the alternative constraint (frequency here), at a baseline reliability condition.

The intuition behind such a table (frequency versus load; f-L) is simple: the current flow in the IP increases with the operating load, and hence the frequency should be lowered to meet the reliability specification. This model

can be used at the chip level to determine the safe frequency (fsafe) of an instance for any design/reliability parameter, and then make corresponding design fixes. Needless to say, most of the EM-critical cells are the ones that operate at higher loads, frequencies or slews.

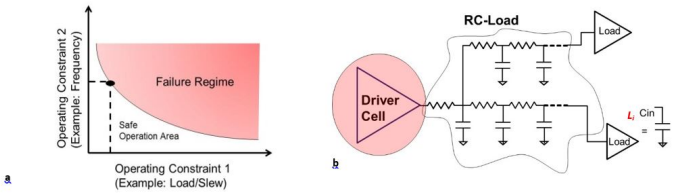


Figure 3.1: a) Traditional approach for EM verification using the safe operating region concept. b) Schematic highlighting the EM-critical cell, driving an RC load network (vis--vis safe frequency obtained for pure C load)

However, such a specification is also simplistic and with advancing technology and convoluted circuit effects, this model is inadequate in accurately predicting EM safety for several reasons. First, the frequency in Fig. 3.1a refers to the output switching frequency: for multi-input cells, the failure rate depends on the switching frequency at each input. This corresponds to a multidimensional space that is computationally expensive to characterize. Second, EM constraints are often specified in terms of average current density thresholds or RMS [Hun97]. However, having multiple relationships between operating parameters is infeasible. Lastly, while the traditional f-L model is characterized at purely capacitive loads, in reality, cells drive RC loads (Fig. 3.1b), and fast prediction of cell-internal EM safety under RC loads is an open problem.

Our goal is to address four limitations (L1L4) associated with chip-level cell-internal EM analysis:

- L1: Inability to incorporate the impact of arbitrary switching rates on inputs pins and effects such as clock gating: We overcome this by discretely characterizing the individual current components (switching or leakage). Additionally, our frequency constraints are self-consistent, which simultaneously address the average and RMS current criteria, based on formulations proposed by Hunter [Hun97].

- L2: Inability to comprehend RC loads (Fig. 3.1b) and to model voltage-dependent pin capacitance: C_{in}: We apply intelligent moment-matching-based techniques as in [JJ12], and propose a novel formulation for C_{in} estimation.

- L3: Inability to retarget reliability specifications on-the-fly for different reliability conditions: We develop the concept of equivalent stress and present closed-form formulae.

- L4: Non scalability of cell characterization data for an entire library due

to prohibitive simulation runtimes, with $\sim$ 600 simulations per cell: We perform these simulations efficiently using intelligent response modelling.

The core methodology of our work naturally enables model retargeting by separating the current density computation part from the verification, as against the tight coupling in the model of Fig. 3.1a, where the f-L curve must be characterized at each reliability condition. In our approach, the reliability conditions need to be specified in-situ: only at the design verification stage. Moreover, our model can take the operating frequency (f_{op}) of an instance as an input, or it can provide the maximum safe operating frequency as an output.

The rest of the chapter is organized as follows. Section 3.2 discusses the basics of EM and the traditional approach. We discuss the proposed model for EM verification under C-only loads and its usage for self-consistent fsafe estimation (associated with L1) in Section 3.3, while RC loads and pin capacitance issues (L2) are accounted for in Section 3.4. Next, we discuss retargeting methodology (L3) in Section 3.5, followed by the cell-response modeling (L4) in Section 3.6. Finally, production usage of the proposed methodology on a 28nm industry design is described in Section 3.7, followed by concluding remarks in Section 3.8.

3.2 EM Modeling: Basic Framework Under Purely Capacitive Loads

3.2.1 Electromigration Basics: Recap

In this section, we review the key parameters affecting EM. In our terminology, we refer to metal segments of the IP as resistors. These resistors are obtained by parasitic extraction, which retains key information such as the width, length, and the metal-level for every resistor in the netlist. Since EM is a statistical process, the time to failure for metal segments stressed in similar conditions also varies [Bla69]. Industrial markets demand low failure rates (e.g., 100 defective parts per million (DPPM) over the chip lifetime). Chip reliability engineers translate this chip-level specification to specific fail fraction (FF) targets, in units of failures-in-time (FIT), on individual resistors.

As discussed in Chapter 2, the classic Black's equation [Bla69] relates the mean time to failure (t_{50}, time to failure for half of the population) to the average current density J across the interconnect cross-section and the wire temperature T as (2.5). Black's equation predicts the time to failure, and in practice, it is predominantly used to determine the average current density thresholds to meet a target FF. It has been demonstrated that FF follows a lognormal dependency on the time to failure (t_f, also known as stress time) [Bla69]. The lognormal-transformation parameter (z), relates to the time-to-failure as in (2.14), where σ is the standard deviation of the distribution.

Indeed, (2.5) to (2.14) are an intuitive set of equations. For example, for a fixed stress time, the time to median failures, t_{50}, decreases with increasing

stress temperature ((2.5)), thereby keeping z, and eventually the FF, high. The transformation variable z helps in directly representing the cumulative failure rate with a normal cumulative distribution function [Lee03]. For example, at stress time $(t_f) = t_{50}$, z and FF consistently evaluate to 0 and 0.5 respectively.

Additionally, it must be recalled that using Hunter's formulations [Hun97], we can check for RMS as well as average EM effects in a self consistent manner. Thus, given the constraints of stress temperature, lifetime and Joule heating limit, we can arrive at the EM thresholds that should be met by all metal segments in the IP. Once we have the EM thresholds in place, we can embark on the EM verification process across various resistors in the IP.

3.2.2 Traditional Approach for Modeling EM Reliability

We begin by revisiting the traditional approach, as outlined in Fig. 3.1a. Given the physical design of the IP, EM verification requires a model that provides a tradeoff amongst various operating conditions such that within the bounds of those tradeoffs, the IP remains EM-safe. The generation of this model requires an iterative search: for example, in Fig. 3.1a, at a fixed loading and reliability condition (say, 50fF, 1.0V, 105C, 100kPOH), an iterative search over the frequency space is required to determine the maximum fsafe, where all resistors within the IP are EM-safe. This is computationally expensive since each iteration involves a SPICE-simulation-based verification. A typical optimized procedure requires ten binary search iterations at each loading condition. For a single input cell, whose operating load/slew space is covered through an 8×8 matrix in the liberty file, the number of required iterations are about $64 \times 10 = 640$ for fixed values of other parameters (supply voltage and reliability specifications). To support operation at multiple supply voltages, as well as IP reuse across application domains, this number must be multiplied by the number of use cases, resulting in a formidable characterization overhead.

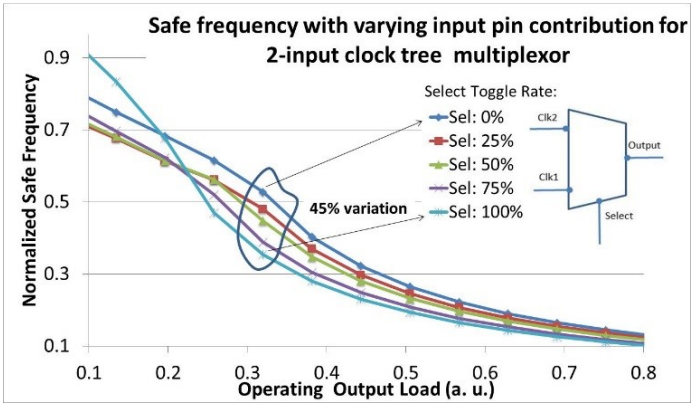


Figure 3.2: f_{safe} plot for a 2-input clock-multiplexor cell. Both input clocks switch at 100%, while the select pin chooses one of them, with varying likelihoods.

While this may even be tractable for single-input cells, for multiple-input cells, this characterization becomes challenging, not just from a computational point of view, but also from the fundamental modeling (L1) viewpoint. To illustrate this idea, consider the example of a two input clock-tree mux IP block that is used to alternate amongst clocks for downstream propagation. The user may examine typical workloads and use cases and provide an EM analysis tool with information about the switching rates of the input pins of a block. In this experiment, both the input pins (Clk1, Clk2) switch at 100%, but the select pin is toggled to allow passing of first and second clocks in varying amounts (going from 0% to 100% in steps of 25%).

The f-L plots for the five cases are shown in Fig. 3.2, and show a variation of up to 45% in f_{safe} estimates, depending on how often Clk1 or Clk2 is selected over the lifetime, but the traditional model will choose the pessimistic fsafe over all cases. Such an asymmetric response can only be captured by the traditional model by individually generating and storing the f-L data for various input excitations, which is expensive both in terms of computation during characterization and storage of the f_{safe} tables. Further, effects like clock gating and power gating are not straightforward to handle in the traditional model.

Another significant drawback (L2) with the traditional model is the fact that it has been generated using a lumped C load, while real applications involve RC loading. Due to resistive shielding effects, a direct application of traditional model to assess reliability of instances that drive RC loads turns out to be severely pessimistic. Finally, we also note that the traditional model is locked to a particular reliability specification (supply voltage, temperature, lifetime and failure rate target), and is incapable in allowing a tradeoff on these (L3), unless, the f-L data is regenerated along these vectors, which becomes computationally unaffordable for the entire library (L4). With the above background and detailed understanding of the traditional model (including generation, usage and associated limitations), we now look at building the proposed model, which can address the various limitations.

3.3 Addressing L1: Incorporating Arbitrary Switching And Clock Gating In Frequency Estimation

3.3.1 Library Level Current Characterization

In order to build the model which can help predict the reliability of an IP for arbitrary switching scenarios, we begin in an ab initio manner by trying to classify the current flow in the IP as either leakage or switching current. We observe that for a combinational IP with m inputs, 2^m distinct static states (various combinations of input pins at logic 1 or 0) are possible. Each of these states can have different leakage flow. Additionally, based on the IP functionality, there could be several paths (later referred to as arcs) from an input pin resulting in

an output transition. Every such output transition, causes a switching current flow in the IP-internal resistors (belonging to the resistor-set $\Re$).

Thus, first step in our approach is to discretely characterize the current flow: average and RMS, both through every resistor R in the IP (resistor-set $\Re$), in every legal logical state (for leakage current) or arc through the cell (for switching current). Such a characterization will be used to compute the eventual effective current density through any resistor of the cell as a weighted summation of the current densities in unique scenarios, coupled with the information of arc switching rates and probabilities of legal state occurrences.

The salient feature of our characterization is that it remains independent of the reliability condition, which is actually an input during chip-level verification. As the leakage current density in the cell depend only on the static states of inputs, we can easily obtain the current density through R by cycling through all possible input states in SPICE (note that average and RMS remain the same due to DC nature of the waveform). On the other hand, switching current densities are tied to a particular input-pin to output-pin combination (also referred to as a timing arc), through a fixed cell-internal path, with other inputs in non-controlling states enabling the transition. For example, for a three-input AOI gate (Y = !(A + BC)) shown in Fig. 3.3, the output Y can fall because of a rise on A in three different states of BC, namely, 00, 01 and 10. Hence, for this particular A $\rightarrow$ Y arc, the current density must be computed through R for these three logical states of BC. We can leverage the simulation framework of industrial timing characterization systems [mag14], to obtain information about all such arcs and states through the cell. For a particular arc i and associated non-controlling state k, we denote the time duration over which this current density is calculated as s_{ik}. A similar convention is followed by $J_{avg,R_{ik}}$ and $J_{rms,R_{ik}}$ to define the average and RMS current densities through R. As we leverage the timing characterization framework, we do not recompute s_{ik}, but reuse it from the timing analysis step [EDI15, ALT14]. Moreover, s_{ik} is typically greater than the delay itself, and therefore accurately captures the tail effects.

3.3.2 Effective Current Estimation for a Chip-Level Instance

After characterizing the leakage and switching current densities for various arcs and states, we now present the calculations for the effective average and RMS densities in the circuit.

Effective Leakage Current Density Through a Resistor Across All States

For an m-input gate, let the leakage current density through resistor R for a state k (of 2^m states) in the positive [negative] direction be denoted by $L_{+R_k}[L_{-R_k}]$. Then, the average effective leakage current density ($L_{avg,R}$) covering all the states and incorporating recovery ((2.7)) would be:

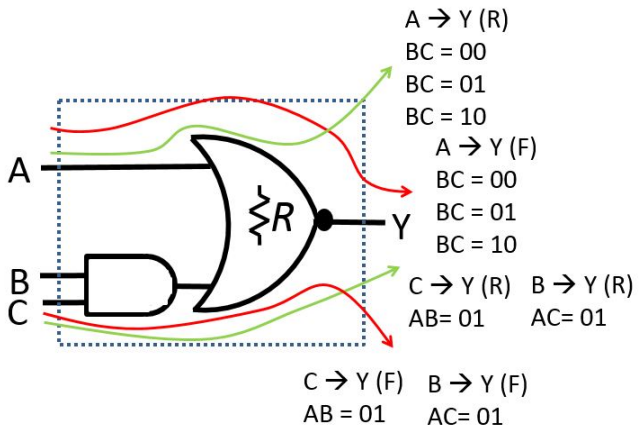


Figure 3.3: Showing all possible timing arcs through a 3-input AOI gate for current characterization.

$$L_{avg,R} = \sum_{k=1}^{l} P_k^+ L_{+R_k} - \kappa \left(\sum_{i=1}^{2^m - l} P_k^- L_{-R_i} \right) \tag{3.1}$$

Here l is the number of states with positive current density, and $P_q^+[P_q^-]$ is, the probability of occurrence of state q in which the current flows in the positive [negative] direction. These probabilities are a function of the duty cycle at the inputs of the gate.

The RMS effective leakage current density is given by:

$$L_{rms,R}^2 = \sum_{k=1}^{l} P_k^+ L_{+R_k}^2 + \sum_{i=1}^{2^m - l} P_k^- L_{-R_i}^2 \tag{3.2}$$

Effective Switching Current Density Through a Resistor Across All Switching Arcs

In similar spirit, the effective-average-switching current density ($J_{avg_{sw},R}$) through R is given by:

$$J_{avg_{sw},R} = \sum_{i=1}^{\text{all arcs}} \left(\sum_{k=1}^{\text{all states}} P_{ik} J_{avg,R_{ik}} \frac{s_{ik}}{T_{clk}} \right) \tag{3.3}$$

Here, P_{ik} and T_{clk} are the design-level parameters the switching probability of the particular arc, and the switching period respectively. The scaling factor, s_{ik}/T_{clk}, translates the characterized current density ($J_{avg,R_{ik}}$), which was av-

eraged during the characterization over the switching duration s_{ik}, to the entire clock period. This scaling factor accounts for the fact that the current is inactive during the remainder of the clock period.

Similar calculations for RMS current density ($J_{rms_{sw},R}$) yield:

$$J^2_{rms_{sw},R} = \sum_{i=1}^{\text{all arcs}} \left(\sum_{k=1}^{\text{all states}} P_{ik} J^2_{rms,R_{ik}} \frac{s_{ik}}{T_{clk}} \right) \tag{3.4}$$

Effective Average and RMS Current Densities

After computing the effective switching and leakage current densities independently, we must now compute the effective average and RMS current densities. In a normal design flow, the chip level probabilistic activity propagation tools already provide the effective switching rate ($f_{ik} = P_{ik}/T_{clk}$) for any given arc i and associated non-controlling state k, along with the state probabilities (P_q^+) in (3.1) for all gates of the design.

Since equations (3.1) to (3.4) discretely describe the leakage and switching current densities, we can sum them to derive the effective average current density ($J_{avg,R}$) and add them in an RMS manner to derive the effective RMS current density ($J_{rms,R}$) for any resistor R in the cell.

We compute the average and RMS current densities by consolidating (3.1) to (3.4) as:

$$\begin{aligned} J_{avg,R} &= J_{avg_{sw},R} + L_{avg,R} \\ J^2_{rms,R} &= J^2_{rms_{sw},R} + L^2_{rms,R} \end{aligned} \tag{3.5}$$

It must be mentioned here that RMS formulations work under the assumption that the different current density (leakage and switching) are non-overlapping. This strictly is not true; however, we find that this assumption leads to very marginal errors. Next, we look at incorporating clock gating in the formulations.

Incorporating Clock Gating

Clock gating is a widely-used technique for reducing the dynamic clock power by disabling the clock signal to the idle parts of the circuit thereby also directly affecting the reliability of the signals in the gated domain [TMS08]. In order to assess the reliability impact of clock gating, we notice that as a phenomenon, clock gating can occur in an arbitrary way over the lifetime of the chip. For instance, the clock could be gated for a fixed number of cycles, after every specific period of activity, in a repeated manner. Such uniform gating is akin to a direct reduction in the operating frequency and can be readily approximated by specifying the activity-rate-adjusted frequency in (3.6).

However, the cases when the clock gating is non-uniform, or is uniform only in the intervals, are nontrivial and require equivalent reliability-lifetime calculations. The key determinant in such calculations is the thermal time constant of Joule heating in interconnect (typically in several microseconds for

copper [BM01]), which signifies the duration after which the interconnect responds to the RMS current in the form of a temperature rise. Hence, if the time interval between successive clock gating events is larger than the thermal time constant, then, the full current (without activity correction), should be ideally used for RMS and average density estimations, for the appropriate durations.

We will defer treatment for non-uniform clock gating to Section 3.5.2 (subsequent to incorporation of arbitrary reliability specifications), and focus the formulations now only for the uniform case. This makes the solution similar to setting a pin specific activity rate on the cell. Hence, if a 1GHz clock tree element remains gated-high for 25% of the lifetime, we would note the corrected f_{ik} as 750MHz in (3.6), and state probability as 0.375 (assuming 50% duty cycle for clock). The computation procedure can thus be captured as:

Algorithm 3.1 Current density computation through every resistor of a cell.

Input: SPICE setup (with all resistors), timing characterization setup
Output: $J_{avg,R_{ik}}$ and $J_{rms,R_{ik}}$
for each library cell; every load/slew in the 8×8 matrix **do**
    simulate for every legal input state combination (k)
    for each resistor R of the cell **do**
        store average leakage density $L_{R,k}$ ((3.1))
    end for ▷ every state
    for every legal switching scenario (arc i) **do**
        for each resistor R **do**
            store $J_{avg,R_{ik}}$ and $J_{rms,R_{ik}}$
        end for
    end for ▷ every arc
end for ▷ end cell characterization
for each instance in the design **do**
    estimate f_{ik}, P_q^+, s_{ik} for all input pins, arcs and states
    for each resistor R of the instance at chip level **do**
        query-and-add $J_{avg,R_{ik}}$, $J_{rms,R_{ik}}$ and $L_{R,k}$ as in (3.6)
        store $J_{avg,R}$, $J_{rms,R}$ at given condition (f_{ik}, P_q^+, s_{ik})
    end for ▷ for every resistor of the instance
end for ▷ for every instance

It can be inferred that current density computation through every resistor requires a number of simulations equal to the unique number of arcs and leakage states, which are anyways part of the timing characterization. Let n_i be the number of inputs in the cell and n_R be the total number of resistors in the cell. Now, the upper bound on total number of timing arcs through a cell can be computed as $n_i \times 2^{n_i}$, since for every single input-output transition, the non-controlling inputs $(n_i - 1)$ could be in 2^{n_i-1} states, thus making the upper bound on total number of arcs as $2 \times 2^{n_i-1} \times n_i$. Consequently, the overall complexity of EM characterization is then given by $\mathcal{O}(n_R n_i 2^{n_i})$. We would like to reiterate that since we are harping on the timing characterization tool to derive all the

timing arcs and leakage states of the cell, this part of the implementation cost is not considered by us. The additional cost is that of current measurements and storage which follows above complexity.

3.3.3 Instance Safe Frequency Estimation at Chip Level

Once we have estimated the currents in the cell, the EM checking procedure can subsequently be approached in two manners, as noted in Section 3.2 earlier:

- Predict the safety of the cell (pass or fail), given a full set of operating conditions of the cell.
- Calculate a set of safe operating parameters for the cell under a partial set of operating conditions. For example, if the frequency, slew and supply voltage are given, the safe load may be computed.

The first is rather trivially obtained from the above discussion, since equation (3.6) and Algorithm 3.1 lend themselves readily to allow substitution of the exact operating conditions, and subsequent verification of currents (through all resistors) against the foundry EM thresholds.

In real designs, however, the actual operating frequency of the instance can depend on the design and is unavailable while characterizing an IP library that is used across a wide variety of designs and application domains. It is necessary to work the problem backwards by recommending a maximum fsafe based on other parameters. *In contrast to the f-L data of Fig. 3.1a obtained by iterated binary-search SPICE simulations, our approach here provides closed-form solutions for the* f_{safe}.

It must be noted that potentially, every resistor in the cell could have unique frequency dependence, and therefore, the maximum f_{safe} procedure must find the minimum safe frequency over all resistors in the instance.

Let $J_{avg,th}(T,t)$ and $J_{rms,th}(\Delta T)$ represent the current density limits for average and RMS current densities respectively, as a function of stress temperature, stress time, and maximum heating constraint. Further, note that in eqs. (3.1) to (3.4), the dependence on the frequency $f = 1/T_{clk}$ appears only in the expressions for the average and RMS switching current densities. By setting the left-hand sides of equation (3.6) to be no larger than the threshold densities and combining them with eqs. (3.1) to (3.4), we can constrain the RMS or average-limited frequencies $f_{max,AVG,R}$ and $f_{max,RMS,R}$, respectively) for each intra-cell resistor R in following manner:

$$f_{max,AVG,R} = \frac{J_{avg,th}(T,t) - L_{avg,R}}{\sum_{i=1}^{\text{all arcs}} \left(\sum_{k=1}^{\text{all states}} P_{ik} J_{avg,R_{ik}} s_{ik} \right)} \tag{3.6}$$

$$f_{max,RMS,R} = \frac{J^2_{rms,th}(\Delta T) - L^2_{rms,R}}{\sum_{i=1}^{\text{all arcs}} \left(\sum_{k=1}^{\text{all states}} P_{ik} J^2_{rms,R_{ik}} s_{ik} \right)}$$

Since all parameters on the right-hand sides of the above equations are known for each resistor in each instance, we can now apply the self-consistent formulations [Hun97] to estimate the safe parameter (frequency) of the resistor. The entire process has to be approached iteratively, as shown in Algorithm 3.2, to determine the safe operating frequency for an instance, which can be then used as a design constraint. The safe frequency for a resistor is the lower of the two values in (3.7) and the safe frequency fsafe for a cell instance is the smallest safe frequency over all resistors in the instance. The entire process is also highlighted through the flowchart in Fig. 3.4.

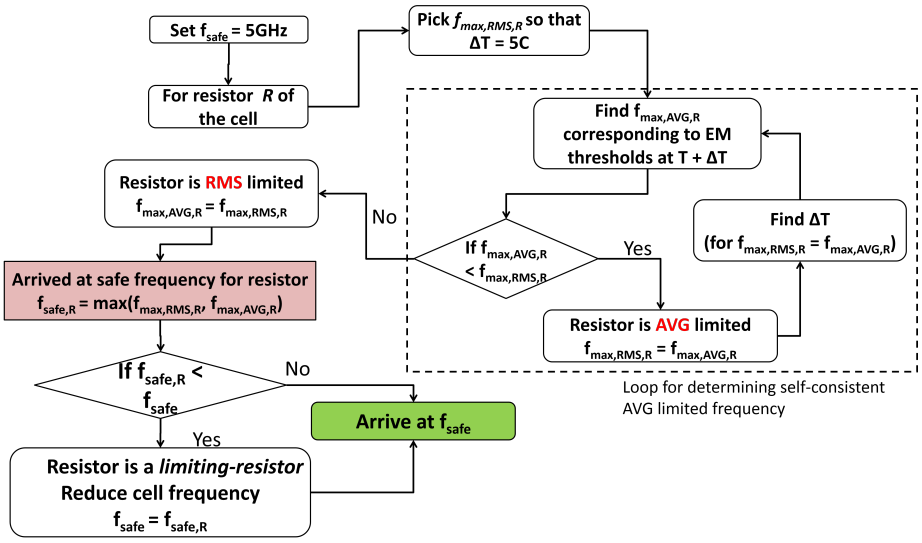


Figure 3.4: Flowchart outlining the safe frequency estimation procedure for a cell.

Note that the safe frequency estimation for a standard cell is an iterative procedure when the cell is limited by average EM instead of the RMS. This is expected, since average EM is a function of temperature, which in turn is a function of the wire RMS currents. On the other hand, if RMS is the limiting factor, then, the safe frequency gets automatically derived just using Joule heating criterion. In either case, the overall complexity of the safe frequency estimation algorithm can be computed by splitting the procedure in two parts: current computation and finding the worst resistor. Using previously defined notations, the first part has a complexity of $\mathcal{O}\left(n_R n_i 2^{n_i}\right)$, whereas, the second part is $\mathcal{O}(n_R)$ complexity, thus making the overall complexity as: $\mathcal{O}\left(n_R n_i 2^{n_i} + n_R\right)$, which can be approximated as $\mathcal{O}\left(n_R n_i 2^{n_i}\right)$.

Next, to evaluate and verify our procedure, we revisit the two-input clock tree mux from the earlier discussion around Fig. 3.2. Fig. 3.5 provides the f_{safe} plot for this case, for a fixed operating condition and output load, showing the results of binary-search-based SPICE simulation, our approach, and the traditional method that chooses the fsafe pessimistically over all switching conditions. We

see that the proposed model fits the SPICE behavior very well and can model the arbitrary switching rates on different pins, as against the large pessimism in the traditional approach.

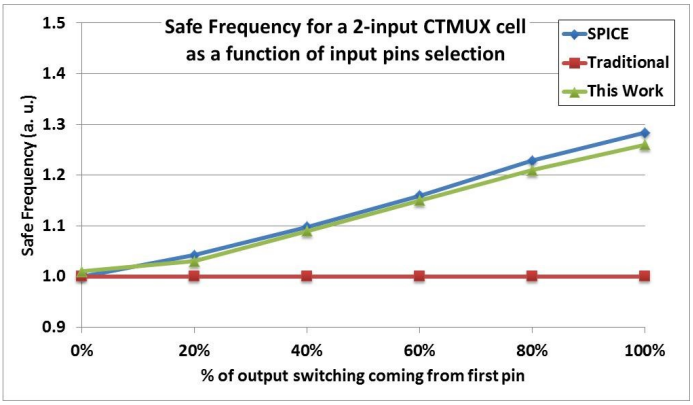


Figure 3.5: Evaluation of the f_{safe} for the circuit in Fig. 3.2, at a selected load point. The f_{safe} varies based on the extent of switching coming from first or second pin. The proposed model completely captures the behavior, but the traditional is excessively pessimistic.

While the results shared in Fig. 3.2 were from a single cell, consolidated results from the entire 28nm design library will be shared later in Section 3.7. It must also be noted that thus far, we have demonstrated Black's equation ((2.5)) based EM verification. However, as our methodology aptly decouples the current density computation and the verification part, it easily lends itself to other EM verification schemes, such as the via-node based scheme or modified Black's approach ((2.6)) as discussed earlier in Section 2.1.1.

3.4 Addressing L2: Modeling The Impact Of Arbitrary RC Loading

The model developed so far is capable of covering following parameters: lumped capacitive load (C load), slews, multi-input gates, arbitrary switching rate and clock gating. This is directly relevant at the chip level, when the IPs are used at arbitrary frequencies and under clock gating. Next, we look at incorporating RC load into the assessment.

3.4.1 Overview of Prior Work

In the last section, we used the operating lumped load as one of the metrics for EM reliability. To a great extent, the C load model in itself can be used for accurate estimation of average current density, and is largely independent of resistive effects [JJ12], provided the rail-to-rail swings for the output net. On the other

Algorithm 3.2 Self-consistent safe frequency estimation of the cell

Input: T_{clk},$J_{avg,R}$, $J_{rms,R}$ from Algorithm 3.1 and the average and RMS Electromigration thresholds from foundry. Design Database.
Output: f_{safe} of the cell
for every instance of the design; start with high f_{safe} **do**
    for every resistor R of the instance in $\Re$ **do**
        start with a JH limited estimate of $f_{max.RMS,R}$
        JH: estimate ΔT for this RMS current (from (2.9))
        estimate $f_{max,AVG,R}$ using $J_{avg,th}(T + \Delta T)$ (3.7)
        if $f_{max,AVG,R} < f_{max,RMS,R}$ **then** ▷ R is AVG-limited
            $f_{max,RMS,R} = f_{max,AVG,R}$
            goto: JH
        else
            $f_{max,AVG,R} = f_{max,RMS,R}$ ▷ R is RMS-limited
        end if ▷ Found self consistent f_{safe} for R
        $f_{safe,R} = \min(\ f_{max,AVG,R}, f_{max,RMS,R})$
        if $f_{safe,R} < f_{safe}$ **then** ▷ R is cell-limiting
            $f_{safe} = f_{safe,R}$
        end if
    end for ▷ for every resistor in $\Re$
    return f_{safe}
end for ▷ for full design

hand, RMS current densities depend not only on the total charge transferred but also on the duration of transfer, and are thereby directly impacted by the resistive effects of RC loads on the cell [Sap04, McC89, QPP94, WM09, CW03, GS12].

The effect of RC load (Fig. 3.1b) for signal EM reliability was addressed earlier in [JJ12]. It was further established that resistive shielding cannot be accounted using the traditional C_{eff} approach, derived from timing constraints. Hence, a current-criterion-based moment matching was devised to come up with a C_{eff}, by performing the RC tree traversal along with the basic timing information. The method was then shown to be very accurate for signal current density estimation.

3.4.2 Prior Work: Limitations

We notice that there are at least two limitations of the prior work associated both with the traditional model (of Fig. 3.1a) and the model proposed in the last section.

Firstly, RC loads affect the current flow in all segments: cell-external as well as cell-internal. While the cell-external problem was solved in [JJ12], the cell-internal piece of the problem has remained unsolved. In fact, it was proposed to simulate the entire active network with the actual distributed load (at transistor level) through SPICE. As we will later see (Section 3.6), the number of such

simulations required for a block/SoC could run in thousands, becoming a major computational and logistical overhead.

Secondly, we notice that not just the effective capacitance, but also the lumped capacitance depends on the current waveform shape making the load capacitance itself as voltage-dependent. This can be explained by the fact that the pin capacitance has a strong voltage dependence: as seen from Fig. 3.6, where $C_{in}(V)$ goes from 1X at 10% operating voltage to about 2.8X at full operating voltage [SABR12]. Hence, even though there is no explicit dependence of the of the average current on the network resistance, it implicitly exists because of the dependence on $C_{in}(V)$. Therefore, assuming a fixed value of C_{in} for performing current calculations on a net becomes very pessimistic, not only for the RMS currents, but also for the average. We now attempt to solve both problems.

3.4.3 Proposed Solution: RC Loading and C_{in} Modeling

We begin by observing that the basic challenge for cell-internal EM arises from the fact that the characterization of the fundamental currents (through Algorithm 3.1) must be performed using the single C load values, but the data must be applied to instances that drive RC loading. Hence, we require a good proxy of the RC load, which can be used to query the characterized data. Extending the concepts developed in [JJ12], if we use C_{eff} to query only the RMS component of the current from the precharacterized data, an accurate match can be achieved. Indeed, we do see a reduction in error (compared with SPICE), if we use C_{eff} for RMS estimation and C load for average, as compared to the case of using C load for both. As we will later see in Fig. 3.8, the mean error of about 2X in RMS estimation reduces to about 20% with the C_{eff} incorporation; however, there still are outliers in the 50

Next, we also compared the currents derived from the case, when the load cells were modeled as a $C1/C2$ combination, where $C1$ represents the pin capacitance from 0-50% swing of the voltage, and $C2$ from 50-70% [CCS05]. This model has recently become popular, as it turns out to be very useful in constructing the current waveforms from a regular STA, or a crosstalk delay perspective. However, we notice that at an individual load pin level itself, an accurate RMS match cannot be obtained with a $C1/C2$ model, as it fails in capturing the tail effects of capacitance. Note that the basic motivation for such a model is matching the transistor operation in linear region, whereas the tail effects are also important from the current estimation perspective [JJ12, LVFA09].

Hence, we propose calculating an effective C_{in} ($C_{in,eff}$) from the multi-piece $C_{in}(V)$ table (Fig. 3.6; typically 8 points). Since C_{in} is a function of the voltage waveform, which in turn is a function of C_{in}, the entire computation must be carried out in an iterative manner. Accordingly, in the k-th iteration, we make use of the starting current waveform (as incident on the load cell L_i of Fig. 3.1b).

Such a current waveform ($I_{L_{i,k}}(t)$) , is obtained through a single $C_{in,k}$ and uses a double exponential model with estimated parameters $A_{0,k}, T_{a,k}$ and

$T_{b,k}$. The estimation of these parameters is performed by RC-tree traversal and moment matching technique with assumption on the waveform shape at the driver (a mixture of ramp/exponential) [JJ12]. The current waveform is modeled as:

$$I_{L_{i,k}}(t) = A_{0,k}\left(e^{-t/T_{a,k}} - e^{-t/T_{b,k}}\right) \tag{3.7}$$

Subsequently, the voltage waveform $V_{L_{i,k}}(t)$, as seen on the load pin, can be generated as an area under the curve of this current waveform, using a constant $C_{in,k}$, as follows:

$$V_{L_{i,k}}(t) = \frac{1}{C_{in,k}} \int_0^t I_{L_{i,k}}(t')dt' \tag{3.8}$$

This voltage waveform can then be used along with the varying C_{in}: $C_{in}(V)$ table, to reconstruct a new current waveform $I'_{L_{i,k}}(t)$ as:

$$I'_{L_{i,k}}(t) = C_{in}(V)\frac{dV_{L_{i,k}}(t)}{dt} \tag{3.9}$$

Note that only an update in the current waveform at the load pin is required, since we are interested in the current specifically at this point. Assuming the duration of this current waveform as d (approximated by the corresponding 0-100% slew at the load pin obtained through STA), its RMS is given by:

$$\text{RMS for } I'_{L_{ik}}(t) = \sqrt{\frac{1}{d}\int_0^d I'_{L_{i,k}}(t)dt} \tag{3.10}$$

Note that for the next iteration, we require an updated value for C_{in}. Hence, we make use of the RMS current through $I'_{L_{i,k}}(t)$, to derive a single effective C_{in}, assuming an equivalent triangular current waveform (with d being the delay at the load pin). For such a triangular waveform, the RMS current expression is standard:$\sqrt{\frac{4}{3}\frac{CV}{d}}$, where C is the equivalent load. In order to obtain an equivalent pin capacitance which can match the RMS current of $I'_{L_{i,k}}(t)$, we equate (3.10) to the RMS current of triangular waveform, to get below capacitance (to be used for next iteration) as:

$$C_{in,k+1} = \frac{d}{V}\sqrt{\frac{3}{4}\int_0^d I'_{L_{i,k}}2(t)dt} \tag{3.11}$$

We expect convergence in 2-3 iterations, though, for our work, we have made

only a single update to starting C_{in}. In similar way, the average current case can be approached, and virtually, for every individual RC-network, we can compute a new effective C_{in}, which when used, matches the average current computation accurately. While this means that we must ideally compute two separate capacitances for C_{in}: namely $C_{in,eff,RMS}$and $C_{in,eff,AVG}$, our experiments indicate acceptable errors for the average case, and hence for this work, we do iterative computation only for RMS matching.

In summary, we accurately incorporate the impact of the voltage-dependent input pin capacitance as well as the impact of parasitic RC loading on the cell-internal current densities, by:

- Making an initial estimate of the current at the driving points and iterating with the loads voltage-dependent pin capacitance to arrive at the final current flow.
- Estimating the effective capacitance (C_{eff}), which matches the final current flow in the network.
- Using this C_{eff} to query the precharacterized cell-internal RMS current density database and C load for querying AVG cell-internal current densities.

Note that in absence of this method, we would have used C load to query the cell-internal AVG as well as RMS current densities, which is very pessimistic. Note also that the formulations for incorporating voltage-dependent pin capacitance automatically improve the accuracy of cell-external current densities as well.

3.4.4 RC Loading and Cin Model Validation: Results

We perform validation at multiple levels of the RC and C_{in} modeling approaches. First, we validate the C_{in} approach at the load-level circuit, followed by the validation of the combined RC loading and C_{in} modeling in the driver-load pair case (Fig. 3.1b), followed finally by the results from several driver instances (driving unique RC loads).

We begin by showing the load-level comparison first, for effective C_{in} estimation (versus SPICE) for different C_{in} models (Fig. 3.7). This comparison is at the load circuit level, where we apply a voltage waveform at the load pin and the load cell is modeled as: a) single C_{in}, b) a two-piece voltage dependent capacitor in SPICE [HSP12], and c) a single $C_{in,eff}$ (obtained from equations (3.7) to (3.11)).

As also discussed earlier, since C_{in} is a function of the starting input voltage waveform, we have computed the errors for different types of input voltage waveforms (the x-axis represents waveforms going from fully ramp to fully exponential), whose shape is controlled with the coefficient a in below equation, with T_r being the rise time:

$$V_{in}(t) = a\left(1 - e^{-t/T_r}\right) + (1-a)\,t/T_r = 1 - ae^{-t/T_r} \tag{3.12}$$

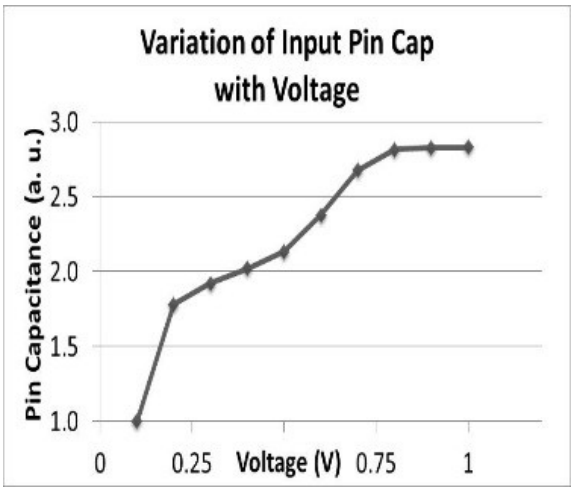


Figure 3.6: Variation of the input pin cap with voltage.

Hence, setting a to zero in above equation, results in a fully saturated ramp input waveform, whereas setting a to unity makes it complete exponential. Such a formulation is a good representation of the various input waveforms which can be incident on the load pin. As we can see from Fig. 3.7, the traditional approach of single C_{in} leads to almost 2x error with SPICE. The error reduces using $C1/C2$ model, however, it still remains unacceptable and the effective capacitance computation approach from an eight-piece piecewise-linear table fits the SPICE results in a better way. We can also see that because of increased tail effects in the exponential input voltage waveform, much higher error exists in all models for a completely exponential case.

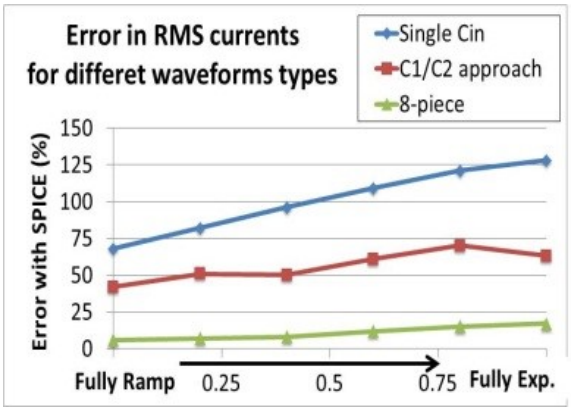


Figure 3.7: Error in the RMS estimates (versus SPICE) for various Cin modeling approaches and waveform types (x-axis; going from fully ramp to fully exponential)

Fig. shows the maximum error from several instances (which drive different

RC loads; plotted on the x-axis). While the left y-axis shows the errors, the C_{eff}/C-load ratio, an indicative of the extent of resistive load the instance is driving, is plotted on the right y-axis. The exact set of instances and their driving RC load information is obtained from a 28nm production design. We show the comparison of: the traditional case (using lumped load for current querying), C_{eff} model alone and the combined $C_{eff} + C_{in}$ model.

Overall, amongst all cases, we find about 2X mean error in RMS current estimation with the usage of lumped load, which drops to about 21% mean with the usage of C_{eff} model, and further down to about 7% mean error with the combined usage of C_{eff} and C_{in}model. This is a significant improvement as compared to the traditional cases. Moreover, for instances and designers desiring high accuracy, a SPICE simulation based verification can still be used. However, the number of such SPICE simulations will be drastically reduced as compared to the traditional flow. We also see that for instances driving severely resistive loads (indicated by the ratio of C_{eff} to C load), the original error with C load usage is very high, with outliers that cross 50% error. The final algorithm for estimating the accurate currents through cell-internal segments for arbitrary loading is shown in Alg. 3.3.

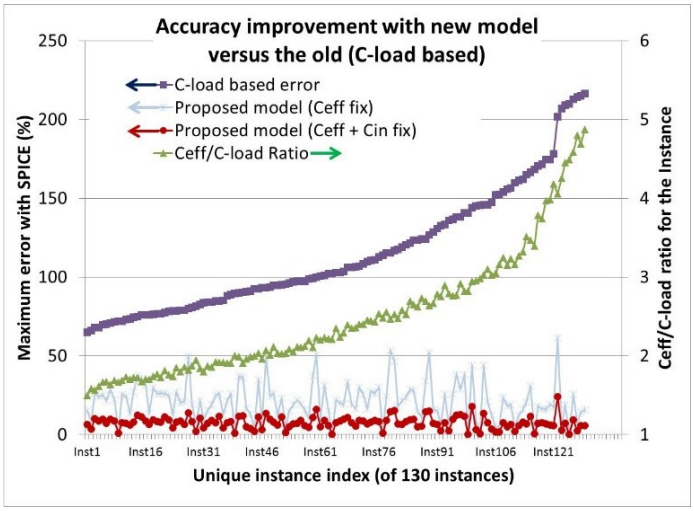


Figure 3.8: Maximum error in RMS current estimation across several instances driving different kinds of RC loading (indicated by the Ceff/C-load ratio) at the design level.

In our observation, for every instance, it takes about two to three iterations to compute the effective pin capacitance. Overall, the complexity of above procedure is a function of the cell properties (number of resistors and total number of inputs), as well as the number of fanouts for a given instance (n_F). Notice that the effective load computation is a precursor procedure to the currents computation. Thus, using previously defined notations, we can note the complexity of the procedure as: $\mathcal{O}\left(n_F + n_R n_i 2^{n_i}\right)$.

Algorithm 3.3 Accurate EM verification considering RC loads

Input: SPEF, Capacitance(Voltage) for all cells
Output: Pass/Fail results for instances after EM verification
for every instance in the design **do**
    obtain the timing information (load, slew and the SPEF)
    for every load cell, L_i, of the instance, in k-th iteration **do**
        set $C_{in,k}$, compute $I_{L_{i,k}}(t)$, using (3.7)
        use $I_{L_{i,k}}(t)$ to construct $I'_{L_{i,k}}(t)$ using C_{in}(V) eq. (3.7) to (3.11)
        recalculate $C_{in,eff}$ and iterate till acceptable accuracy
    end for
    compute network C_{eff} by using $C_{in,eff}$ for every load [JJ12]
    for every resistor R_j of the instance **do**
        use Cload to query average, C_{eff} to query RMS currents
        verify currents against thresholds
    end for ▷ flag pass if all resistors pass
end for ▷ for every instance

Thus, we have examined the impact of RC loading on the EM reliability, and demonstrated significant improvement in accuracy with the proposed method. Such an improvement helps directly in reduction of violations at chip level, when the model of Section 3.3, is used at a given C load and C_{eff} from the network.

3.5 Addressing L3: On-The-Fly Retargeting of Reliability For Arbitrary Specifications

The formulations of previous sections were all dependent on the library data, characterized at one set of operating conditions, and the foundry EM thresholds at a specified reliability condition. However, as described in Section 3.1, there is an increasing need for on-the-fly reliability retargeting, at design verification stage, as the IP library is used under different reliability conditions. As noted earlier, meeting this goal is impractical under the traditional methodology, as it requires a new characterizations of the entire IP library (Fig. 3.1a) at each new condition.

The core methodology of this work enables the ability to perform this retargeting efficiently, since the current density computation part is separated out from the verification part (whereas these are tightly coupled in the traditional approach). We begin with the fundamental relation between EM lifetime and the lognormal variable. From equations (2.5) and (2.14), taking logarithm, we obtain:

$$\sigma z = \log(t_f) - \log(A) + n\log(J) - \frac{Q}{k_B T} \tag{3.13}$$

Now, if we have two different sets of stresses, denoted by subscripts a and b, each is described by the same fitting parameter A, but other terms in (3.13) may differ. Naturally, their reliability is related as follows (by substituting the parameter A):

$$\sigma z_{cond,b} = \sigma z_{cond,a} + \log\left(\frac{t_b J_b^n}{t_a J_a^n}\right) - \frac{Q}{k_B}\left(\frac{1}{T_b + \Delta T_b} - \frac{1}{T_a + \Delta T_a}\right) \quad (3.14)$$

Here, the variables t, J, T, and ΔT represent the stress time, current densities, stress temperature and Joule heating, respectively, while the subscripts a and b refer to the two different conditions.

This equation is a powerful representation of the scaling factors that can either be used to assess:

- the required tradeoffs in new reliability conditions to meet the same fail fraction levels, or,
- the actual fail fractions at the new reliability conditions.

For example, we can directly use above equation to find the equivalent stress time (t_b) that causes the same reliability loss as benchmark condition, but with increased current densities. In order to do so, we must set $z_{cond,b} = z_{cond,a}$, since the reliability loss has to be equated and obtain the equivalent lifetime t_b as a function of (t_a, J_a, and J_b). Obviously, if $J_b \quad J_a, t_b$ will be estimated to be lower than t_a. The basic idea here is that EM aging is either accelerated or decelerated by the change in operating conditions. Hence, if the circuit is aged under a new condition b, then the equations here help transpose the stress to the known (characterized) condition a, with one of the stress parameter of condition a, and may become either more severe or benign than condition b.

We now look at the application of the retargeting concepts, based on (3.14), to some of the case studies, followed by application to non-uniform clock gating. Unlike uniform clock gating, which was previously treated with generic activity reductions in section 3.3.2, non-uniform clock gating requires a more accurate sliding window based analysis, wherein, every frame potentially becomes a new reliability condition.

3.5.1 Case Studies Incorporating Reliability Retargeting

Case I: Variations in Temperature If the use temperature and/or POH specification are different from the original conditions, then it is straightforward to address this by using (3.14) to determine new current density thresholds, and then updating the f_{safe} in (3.7). Such a modification only affects the average, and not the RMS reliability, which depends not on the temperature specification, but on the ΔT constraint for the RMS rule: if this changes, we can update $J_{rms,th}(\Delta T)$ and then the safe frequency estimate in (3.7).

A second situation is the common industry scenario when the stress profile is provided by the user as a temperature profile, as the series $\{(J_1, T_1, t_1)$,

(J_2, T_2, t_2), .. , $(J_m, T_m, t_m)\}$, i.e., from time t_{k-1} to time t_k, it experiences current stress J_k at temperature T_k. If the baseline stress is characterized for J_0 at temperature T_0, then can relate the k-th stress vector to the baseline stress at (J_0, T_0) with an equivalent stress time $t_{k,0}$. In other words, the stress at temperature T_k is transposed to an equivalent stress time at temperature T_0. Consequently, our stress retargeting scheme will map the entire stress to $(J_0, T_0, t_{eq,0})$, where:

$$t_{eq,0} = \sum_{k=1}^{m} t_{k,0} \tag{3.15}$$

Case II: Variation in Operating Voltage If the eventual use voltage of the library is different from the characterization voltage, current scaling must be performed. Such a scaling is straightforward in our framework, since the leakage and switching related components are separately stored, as described in eqs. (3.3) and (3.4). Based on our experiments, we see that a linear scaling works very well for voltage scaling, while an exponential model is required for leakage. Note that this scaling must be performed for every discrete component of the currents for every resistor in the circuit (eq. (3.3), (3.4)).

A second situation (arising due to power management scenarios like dynamic voltage frequency scaling (DVFS) [HM07]), is when the voltage is represented as a series: $(V_1, t_1)((V_m, t_m)$. In such a case, we can follow the scaling procedure to obtain a series of currents, which can then be dealt in the same way as the earlier case.

Case III: Variation in Failure Rate Specification The $J_{avg,th}(T, t)$ in eq. (3.7) is really a function of the fail fraction FF, which in turn is a function of z (eq. (2.14)). Therefore, z_{target} is the inverse function of FF_{target}. Hence, if the FF specified by the end user changes from, say, 0.1% to say 0.01% cumulative, it can be readily translated to z, translated to a current density limit using eq. (3.14), and then used in (3.7) for verification.

Fig. 3.9 shows a graphical representation of such a retargeting using the proposed model from a representative cell. For ease of exposition, we represent our model at a fixed slew, as in Fig. 3.1a.

As can be seen:

- Curve (a) represents the reliability at the baseline condition. If the FF requirement of the design changes and drops to 10% of the original, the curve slides down to (b) due to reduction in EM capability at tighter FF requirement. The drop is not drastic as this specific IP is RMS-current-limited, rather than being limited by the average current.

- Similarly, if the use voltage has a 150mV overdrive over the characterized value, the reliability is represented by curve (c), which shows degraded reliability due to increased current flow.

- Similar behavior is seen in curve (d) if the Joule heating (RMS current specification) is tightened by 5C.

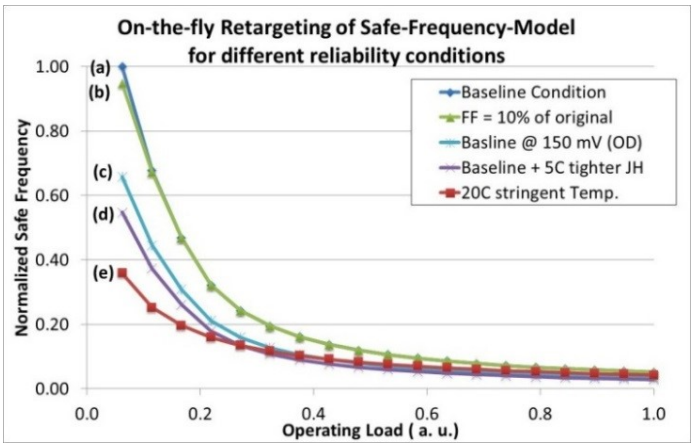


Figure 3.9: Demonstrating on-the-fly retargeting of the basic frequency-load curve (Fig. 3.1a) with changes in the constraining criteria (at a fixed slew point).

- Finally, if the temperature requirement becomes 20C higher, design closure becomes more challenging with the reliability being now represented by the curve (e) almost 3X tightening.

The case study in Fig. 3.9 is handled very naturally in our approach. We reiterate that handling them in the traditional approach require a complete re characterization of the fsafe model at various conditions. Next, to validate the retargeting methodology, we directly compare the curves of Fig. 3.9 to the curves of traditional methodology (obtained by the actual characterization at the exact condition). As earlier, we present results from a single representative cell.

We show the percentage error for two conditions, (c) and (e) in Fig. 3.10. For (e), where the temperature specification is altered, the required retargeting only affects the verification part (as the current density limits are scaled), which incurs little error. For (c), the retargeting is due to 150mV overdrive, where we use a more approximate current-scaling model. The error here, although high, is acceptable, considering the fact that it is in a lower load regime (usually a low-current, EM-safe zone).

Summary: Thus, as discussed in Sec. 3.1, EM degradation of a standard cell is a function of the operating parameters, namely, the load, slews and the frequency. Obviously, for a given cell, a higher operating load would mean that the allowed safe frequency must be lowered for reliable operations. In this section, we showed that as the reliability criteria changes, even without a re characterization, we can recommend new values of safe frequency of a cell. This is of value, since it allows the designer to assess their design at newer reliability conditions without actually fixing violations – but – trading off the reliability with lower operating frequencies. In the upcoming results section (Sec. 3.7), we will share more results and benefit from this modeling.

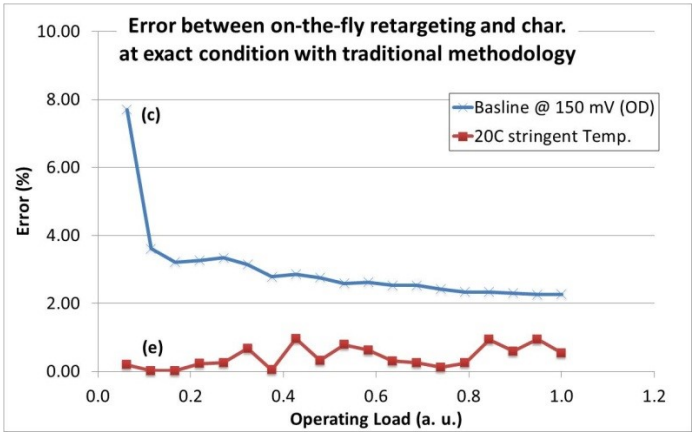


Figure 3.10: Validation of retargeting methodology versus SPICE for two conditions, (c) and (e), of Fig. 3.7.

3.5.2 Incorporating Non-uniform Clock Gating

The case studies of previous subsection were helpful in outlining a general thought process on approaching the problem, when the eventual use scenario is different from the baseline one. We now consider the extension of those principles to the problem of clock gating. In our previous treatment of clock gating (Section 3.3.2), we assumed uniformity, wherein, the activity rate of the clock tree elements can be factored into an effective frequency. However, in reality, the clock gating is an arbitrary process, and hence, a more accurate assessment of reliability should incorporate the associated non-uniformities. In order to do so, a key input required is an activity profile of the design over several clock cycles, from which we can extract the clock activities in smaller durations. Fig. 3.11 shows a representation of such an activity profile, which will be required for the analysis [PTP12]. We noted previously that the thermal time constant is a key determinant in addressing the non-uniform clock gating [CCF+07, ZKS08], and any change in current profile (of a larger duration), should be handled individually, and cannot be combined as a time-weighted summation.

Hence, we follow a sliding window approach (with the duration as the thermal time constant), wherein the complete clock activity profile is scanned in a step-by-step manner. For every single time window scanned, we compute the effective activity rate, which can then be used to compute the cell-internal current densities through every resistor, arcs and states of the cell. Eventually, for a resistor R, in the i-th arc and k-th state, we can represent the current densities as: $(J_{avg,R_{ik},0}, T_0)$,....$(J_{avg,R_{ik},m}, T_m)$ .., where the index m refers to the index of the sliding window. Clearly, every window can have a unique activity rate, and thereby a different RMS and average current density. For instance, in Fig. 3.11, the sampling windows S_a, S_b, S_c and S_d correspond to a 75%, 50%, 100% and 67% activity rate respectively. This current stress can then be collapsed

into a single equivalent stress, based on the concepts developed in the earlier discussions and using eq. (3.14).

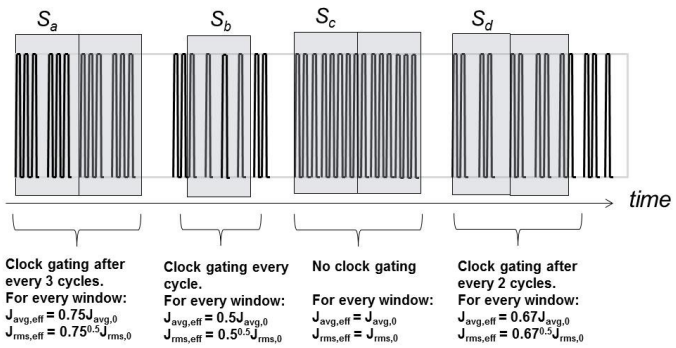


Figure 3.11: Representative clock activity profile for a large duration. Different sampling windows show different activity rates (and corresponding J_{avg}, J_{rms}).

Indeed, for a variety of examples considering clock gating, we can notice a significant difference in the reliability. Fig. 3.12 shows the normalized reliability (of the clock tree element), based on the extent and the nature of the clock gating. For this experiment, we considered a single clock tree element, which underwent different kinds of clock gating, though all amounting to a net 50% duration of gating in the chip lifetime. For example, the third column in the Fig. 3.12 corresponds to a case where the clock remains 25% uniformly gated (meaning gated every one in four cycles) the first half of stress time, followed by 75% gated in the other half; and so on, for the other cases. After reliability computations based on eq. (3.14), we plot the equivalent stress times for all cases, considering the (50%,50%) case as the baseline.

As we can see, for the same clock gating duration amongst all cases (50%), the worst case reliability occurs for the case in which the full-throttle events are clustered together thereby meaning maximizing the average current, as well as JH together. On the other hand, if the clock gating is completely uniform, the JH is lowered, causing the least reliability loss. For the sake of completeness, we also note that for the same case, a free running clock corresponds to an equivalent lifetime of about 3X as compared to the (50%,50%) case. Thus, we can capture the algorithm to incorporate the exact clock gating impact as in Algorithm 3.4.

Overall, the complexity of above procedure is a function of various parameters of cell: the number of resistors, number of inputs, as well as the number of timing-windows (n_W) in which the given clock gating profile is split into. Using previous notations, we can derive the complexity for a given instance as: as $\mathcal{O}\left(n_W n_R n_i 2^{n_i}\right)$. It must be mentioned that such a profiling data is hard to come by in real designs. Therefore, in absence of information, it is recommended to either assume no clock gating, or assume clock gating in the non-uniform manner (corresponding to the first in Fig. 3.12), where the full-throttle and clock

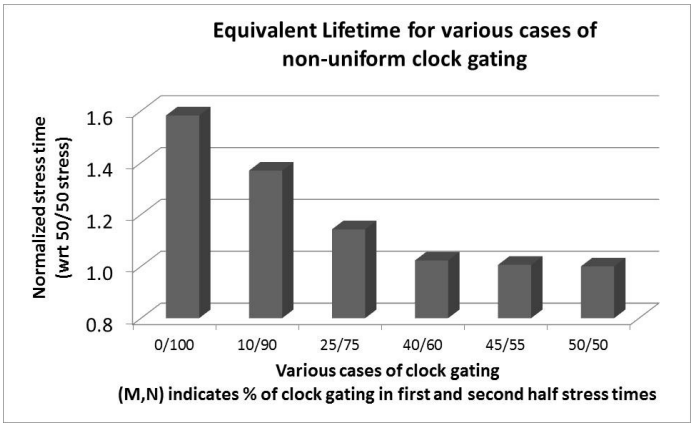


Figure 3.12: Variation in reliability based on the extent of uniform clock gating in first half and second half of the stress time.

gating periods are separated out. Having looked at the various determinants of cell-EM reliability and ways to incorporate them in our model, we now look at expediting the data generation process during the characterization stage.

3.6 Addressing L4: Accelerated Data Generating Using Cell Response Modeling

In this section, we look at ways to speed-up the characterization. As discussed in Section 3.2.2, for the traditional methodology, the safe frequency estimation at a single load-slew point requires about ten SPICE simulations with an optimized binary iterative search, translating to about 640 SPICE simulations per cell for covering an 8x8 load/slew matrix. Indeed, complete data generation for a production 28nm library, consisting of a few thousand cells, can run into days of effort. Such high runtimes for just a single baseline reliability condition make the process of EM characterization prohibitive under the traditional model. Although the efficiencies suggested earlier in this work can greatly reduce this overhead, it is still essential to use the baseline operating condition and characterize the currents using equations (3.1), (3.2), (3.3), (3.4): a process that can be very compute-intensive when carried out for all load/slew conditions. Hence, we look at the ways of expediting the characterization process. Our approach is based on the use of response modeling.

In the retargeting discussion of last section, we noticed that the traditional methodology is inflexible, since it commingles the processes of current computation and EM verification. For the same reason, it also does not lend itself for application of response modeling. The challenge here is twofold:

- From the circuit point of view, operating parameters such as load and slew non-uniformly affect the individual RMS and average resistor currents in

Algorithm 3.4 Incorporating non-uniform clock gating

```
1: Input: Clock gating profile, characterized cell and timing info
2: Output: Javg, R and Jrms, R
3: for each clock tree instance in the design do
4:     obtain timing information: free-running frequency f, slew s
5:     for every window_m of clock profile do
6:         compute activity rate for the window_m and reuse f, s
7:         for each resistor R of the instance do
8:             query-and-add various current components
9:             components from: (3.1), (3.2), (3.3), (3.4)
10:            find equivalent time for R's window_m stress at baseline condition
   (free running): t_eq,R,m,0
11:            add t_eq,R,m,0 to t_eq,R
12:        end for
13:    end for                                          ▷ all windows
14:    for each resistor R of the instance do
15:        use t_eq,R to find resistor pass/fail
16:    end for                                  ▷ report pass for all R passing
17: end for
```

various arcs and states.

- At the same time, the reliability specifications like lifetime and fail-fraction requirements non-uniformly influence the average and RMS current-density thresholds.

Both of above eventually cause the average and RMS-limited frequencies (discussed in the self-consistent estimation in Algorithm 3.2) to be asymmetrically impacted, thereby making the traditional frequency-level abstraction as non-scalable across load/slews (illustrated for example in Fig. 3.13) and reliability specifications (discussed earlier in Fig. 3.9).

On the other hand, a key feature of our approach is to keep characterization and verification disjoint: current characterization is performed at library level (independent of the reliability conditions), whereas the reliability conditions are only required during the design-verification stage. This virtue of our framework, naturally presents opportunity for model building during the characterization phase and accelerate the data generation process.

The fundamental components of our model are the average and RMS currents through all resistors under all states and arcs, for every load/slew point (eqs. (3.1), (3.2), (3.3), (3.4)). As noted in Section 3.4 earlier, average current flow through the resistor is purely a function of the total charge transferred (lumped load), while the RMS current also has an inverse relationship with slew [JJ12]. Based on these observations, we attempt to model the current (J_R) through any given resistor R in the IP as a polynomial function of output

loads/inputs slews:

$$J_R = a_0 + b_1 L + b_2 L^2 + c_1 s + c_2 s^2 + d_1 L s + d_2 L^2 s^2 \tag{3.16}$$

Here, a_i, b_i and c_i are fitted coefficients, and L and s are the loads and slews, respectively. We identify seven critical points (in the 8×8 load/slew matrix) that help shape up the polynomial model: the four corners, (1,1), (1,8), (8,1), (8,8), and a few internal points (2,4), (4,4) and (4,2), where the indices represent the index of the load and slew, respectively, in the table. The parameter fitting is then performed, based on (3.16), providing a model to predict the currents at any arbitrary load/slew point. Note that the response modeling must be performed for every current component (of eqs. (3.1), (3.2), (3.3), (3.4)) of the resistor R. The number of models corresponds to the total number of unique arcs and states of the cell. For example, for a single input clock-tree inverter, we would require a total of four simulations: two to cover the arcs (input rise to output fall, and vice versa), and two to cover the static states (input high and input low).

We now examine the validation of the response model for a representative IP cell in Fig. 3.13; the results from the entire library will be presented in an end-to-end manner in the next section. For various load/slew points on the x-axis, we first develop the characterization data based on full SPICE simulations (using eqs. (3.1), (3.2), (3.3), (3.4)). Subsequently, the model from (3.16), is built using the simulation data from seven sampled points, and later evaluated at each of the 64 load/slew points. The normalized f_{safe} is plotted on the left y-axis, and the error between model and SPICE, on the right y-axis.

The non-monotonic behaviour of f_{safe} with load/slew can be readily observed from this plot. Such non monotonicity arises from the fact that at different load/slew indices, the metal segments which limits the EM performance of the cell varies. For example, at a fixed load condition (say 100fF), a lower input slew ($\sim$ 25ps) would mean large RMS current in the output signal resistors, while a smaller short-circuit current in the power-ground resistors. On the other hand, a higher input slew ($\sim$ 200ps) means vice versa. Thus, for sharp input slews, the output signal resistors may often limit the cell reliability (due to RMS constraint), while, at the sluggish slews, the cell-internal power-ground resistors may be limiting (due to the average constraint). Such an interplay finally leads to a non-monotonic fsafe behaviour of the cell with load/slews.

Our methodology, however, works only at the currents level, and hence, remains unaffected by the reliability constraints which bring in the non-monotonicity. Using the representation from (3.16) (for every resistor, per arc state), we can readily obtain the currents at the chosen load/slew condition, and can subsequently, use those currents to compute the safe frequency of the cell by using Algorithm 3.2, which additionally requires the reliability condition. Consequently, we can cover all the load/slew points to get the safe frequency plot, and as we can see, the response modeling approach works reasonably well in predicting the currents and fsafe, with an acceptable marginal error.

Next, the runtime impact for a single cell is summarized in Table 3.1. As

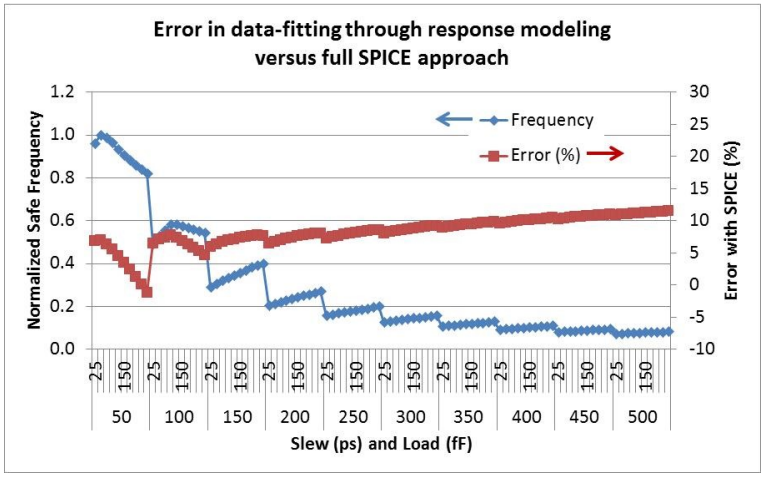


Figure 3.13: Comparison of the response modeling approach ((3.16)) with full SPICE (red). fsafe obtained through response modeling (blue).

we see, the characterization runtime for our approach drops down significantly as compared to the traditional methodology, which takes $\sim$ 10 minutes to generate the safe frequency data for all load/slew points, whereas the proposed methodology (response modeling) is completed in about a minute. It must be mentioned here that the number of simulations required in the traditional methodology grows linearly with the number of design/reliability conditions required (as discussed in Section 3.5). For example, every change in the voltage, stress temperature or lifetime requires a new characterization. On the other hand, the proposed methodology comprehends the design/reliability conditions on-the-fly, using the same database (Section 3.5), hence further keeping down on the number of simulations required. Thus, the methodology in this section directly addresses the limitation L4 and significantly speeds up characterization.

3.7 Production Design Analysis

We now examine the final application of the proposed methodology in an industrial scenario, discussing the setup and the workflow. For the production design analysis, we took a 28nm high performance block (2mm$\times$2mm; $\sim$600K

Methodology	Simulations required per cell	Overall compute
Traditional (Fig. 3.1a)	$640\times n$ (n: number of design and reliability conditions	10 mts. per cell
Proposed (equation (3.16))	28 (7 load-slew $\times$ 4 unique sims)	50 sec

Table 3.1: Runtime comparisons with proposed and traditional methods, for a single cell.

instances, $\geq$ 10M transistors), operating at 1GHz clock frequency. The block is part of a large industry SoC. We compare the characterization as well as the final data application, with entire flow being outlined in Fig. 3.14 for the proposed method.

The new method, in essence, is a three-step process:

- (a) IP characterization at a baseline reliability conditions
- (b) determining the reliability constraints for this design, and,
- (c) integration into the timing/implementation tool.

Note that the true retargeting flexibility of the proposed approach comes in form of (b), which is a runtime-level input to verification that is completely detached from (a). The flow of (c) uses a standard industrial design methodology.

3.7.1 Library Characterization

The entire library of a few thousand cells was characterized in two ways: (a) a full SPICE-based approach, where the traditional fsafe table was generated at a baseline condition, and (b) the methodology proposed in this work. Parallelized and multithreaded SPICE simulations (using Cadence Spectre) were used. The runtime for (a) was about 800 CPU hours of raw simulation, excluding extraction, whereas the methodology in (b) completes in about 80 CPU hrs. For (b), the production characterization framework for timing was used to arrive at the various arcs and logical states for switching and leakage current characterization.

3.7.2 Final Reliability Verification

The final application of the library-generated data was performed in the timing tool (Encounter Timing System), through a custom developed scriptware, which reads in both the characterization data types. The timing analysis of the design was performed at the baseline condition, to arrive at the slews and probabilistic switching rates through all the input pins. In the traditional approach, the scriptware steps through the timing information of every instance in the design and compares the queried fsafe (from the traditional model) to the operating frequency. Note that since this approach suffers from the problems discussed earlier (specifically, L1 and L2), a final full SPICE simulation (with the RC loading of the driver instance) is required after the initial results from the frequency comparisons. A total of about 600K instances were analyzed in this way, and finally, the instances with the frequency ratio > 1 (around 4500), were simulated further. The excitations for the SPICE simulations were a simple 1010 transition (at operating frequency), since all the instances were single input clock tree inverters, buffers and ICG cells (only eight unique cells). The final set of violations after the full SPICE simulations came down to 426.

On the other hand, in the new approach, the scriptware additionally implements Algorithms 3.1-3.3, and based on the chip level design and reliability

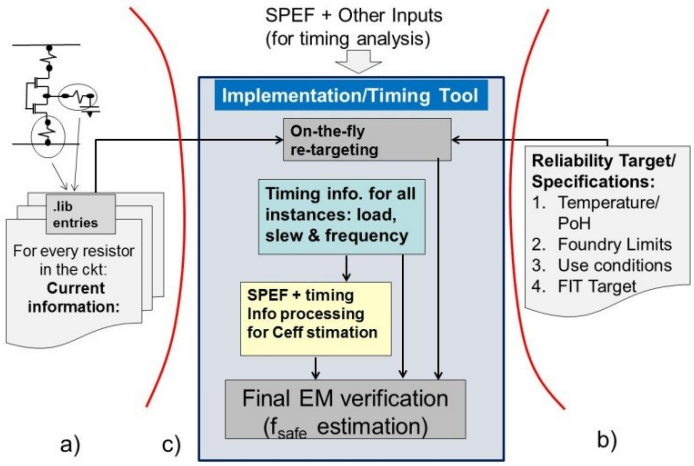


Figure 3.14: Overall methodology and data-flow diagram for the proposed method.

conditions specifications (lifetime/temperature/voltage/fail fraction), the equations are updated on-the-fly for the final frequency comparison of every instance.

Finally, we plot the population distribution of frequency ratios in Fig. 3.15. We consider five cases: a), b) corresponding to analysis with traditional and proposed methods at baseline reliability conditions respectively and c), d), e) corresponding to the analysis at retargeted reliability conditions of a tighter JH limit, an overdrive case and a high temperature requirement, respectively. For every method, we plot the ratio of f_{op} to f_{safe}, which signifies the EM criticality for that instance. Hence, an instance with f_{op} greater than f_{safe} is deemed as EM failure and must be acted upon for fixing (either by load reduction or replacement). The y-axis shows the distribution of number of instances in design with a particular f_{op}-f_{safe} ratio . We document the total number of violations from various analyses in Table 3.2.

Methodology – SPICE Verification

As we can see from Fig. 3.15 and Table 3.2, the proposed approach reports a total of 442 violations, 421 of which overlaps with the traditional methodology (+ SPICE). The remaining: false (21) and escaped (5) violations from the new approach were found to be relatively less critical, with frequency ratios in the range of 1.14 to 0.9. Thus, the new approach agrees well with SPICE.

Methodology – Retargeted Reliability Calculation

As discussed earlier in Sec. 3.1, EM degradation of a standard cell is a function of the operating parameters, namely, the load, slews and the frequency. Obviously, for a given cell, a higher operating load would mean that the allowed safe frequency must be lowered for reliable operations. For traditional approaches, as

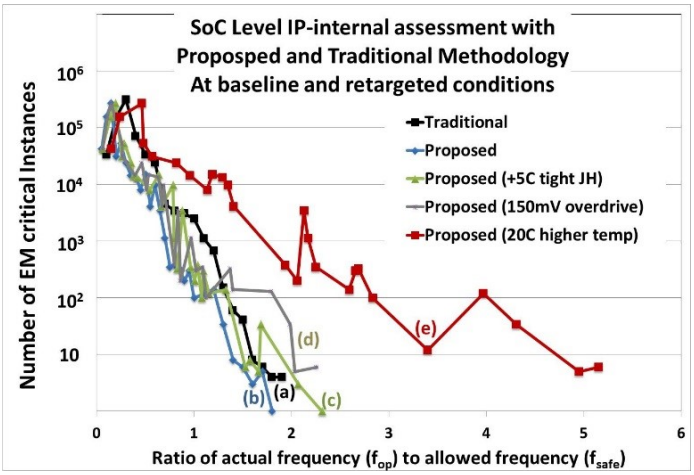


Figure 3.15: Distribution plot for a 28nm block ($\geq$ 600K instances), highlighting the number of EM-critical instances and violations (with f_{op}/f_{safe} ratio ≥ 1) for a), b) baseline reliability analysis with traditional and proposed methods; c), d), e): retargeted reliability condition analysis with proposed methodology.

the reliability targets are fixed during the library-design process, a re-verification of the design at new reliability conditions costs a huge turn-around time. Even if one goes through it, the outcome of such reliability analysis is eventually in form of the design-instances which fail EM and the recommended safe limits at new operating conditions. Indeed, for a library vendor like ARM [Dod15], delivering the traditional EM model [Lib16] of a 28nm-like technology node, with about 3000 cells at one operating conditions (10 year/105C reliability, for example) could take almost one month of cycle time. Now, if the semiconductor design house wants to reverify the design at a different operating condition (5 years, 125C for example), there is no easy way but to ask the library vendor to redeliver the library models. Often, such what-if exercises (refer back to Fig. 1.4) are done several times during the semiconductor business development cycle and such large turnaround times are deterrent in applications warranting different reliability targets.

On the contrary, our method allows project of safe operating conditions at

Analysis Type	Reliability Conditions	Violations
Traditional (SPICE)	Baseline	426
Proposed Methodology	Baseline	442 (421)
	Baseline + 5C tigher JH	1297
	Baseline + 150mV overdrive	1093
	Baseline + 20C higher temperature requirement	56945

Table 3.2: Overall comparison of traditional versus proposed methodology. Traditional method was run only at baseline condition due to runtime issues, whereas the proposed method could run at various reliability conditions.

arbitrary reliability conditions using a decoupled approach of current computation and threshold verification. We now demonstrate the final retargetability of our proposed approach, as evident by the curves c), d) and e) in Fig. 3.15. These are analyzed at new reliability conditions. Run c), corresponding to an additionally tight constraint of 5C lower JH, results in almost 3X increased violations, due to tighter RMS limits. Run d), which is at overdrive conditions results in a similar violation profile. However, run e), which corresponds to a 20C higher stress temperature run results in a plethora of violations. Such a run is a close proxy to a direct application of consumer IPs to an automotive space! *Clearly, without a recharacterization, such verification is not possible through the traditional approach.*

Finally, based on the stage of the chip-design execution, design community has multiple ways to act upon this EM verification feedback. Although a detailed solution to developing EM fixes is beyond the scope of this work, we provide some pointers in the rest of this paragraph. In many cases the harshness of reliability criterion softens due to a lower lifetime requirement for instance, in infotainment category chips [AUT15]. Alternatively, an avoidance strategy can be followed upfront, wherein, based on the logic, high drive-strength cells are used to drive large fanout points. However, this requires careful consideration since unwarranted improvement in drive-strength is associated with sharp output-slew reduction resulting in increased RMS currents. On the other hand, a forceful lowering of drive-strength for instances with timing slack causes slew degradation resulting in increased short-circuit currents. A better approach may be through fanout-load or activity reduction, which predictably reduces the current flow.

3.8 Conclusion

In summary, an accurate and retargetable methodology for IP-internal EM verification was presented in this work. Generic switching rates for various pins of the IP are comprehended, including aspects of clock gating. Significantly high accuracy, with respect to SPICE, was achieved by incorporating the impact of arbitrary parasitic loading, and, an intelligent way of coming up with the effective pin capacitance of load cells. The methodology was shown to be highly flexible, in terms of allowing on-the-fly retargeting for the reliability. Finally, the complete data generation process at library level is expedited by application of cell response modeling. Results on a 28nm production setup were shared, to demonstrate significant relaxation in terms of violations, along with close correlation to SPICE. We shared various cases of runtime-level reliability retargeting, by specifying varying reliability conditions for the production block verification.

The methodology presented in this work is most suitable in a third-party-IP context. The need is only underlined further with the increasing porting of designs from one business segment to a different one, which requires on-the-fly assessment of the reliability of all the components.

Chapter 4

Stochastic and Topologically Aware Electromigration Assessment Methodology

In this chapter, we will establish an important link between individual component-level EM failures and the failure of the associated system. 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.

4.1 Introduction

In the previous chapters (2,3), we successfully abstracted the Electromigration verification challenge for large circuits (containing thousands of resistors) as a component-level problem. Such an abstraction greatly simplifies the problem, since 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 [Bla69]. Indeed, the conventional method for managing EM revolves around containing the current densities in interconnects, which 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.

However, such a simplification comes at a cost. The system now becomes a weakest link approximated (WLA) one, wherein, a failure is deemed when the

first component fails [FP89].

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 chips 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 [QRKG12, Su02]. Clock grids show high tolerance to variations due to their inherently high redundancy, with multiple source-to-sink paths for every sink.

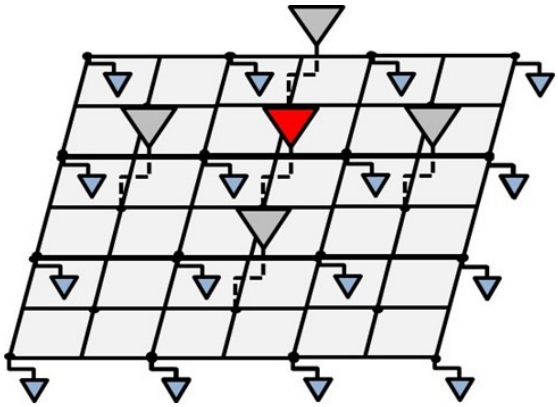


Figure 4.1: A one-level clock grid schematic showing several buffers arranged in redundant configuration

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 [TMS08].

Further, grids are multiply driven by several buffers (Fig. 4.1) connected to a common clock source: these redundant drivers reduce clock skew and lower load/delay variations. Fig. 4.2 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.

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

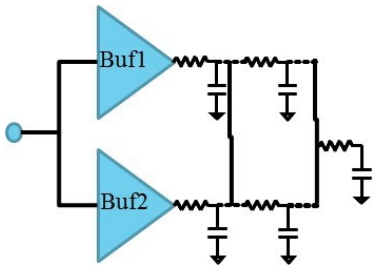


Figure 4.2: A single stage of the clock grid with multiple buffers driving the wire segments.

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 4.2. Next, we develop a failure abstraction model at the circuit level (e.g., 10% delay degradation) in section 4.3, and demonstrate it on a redundant clock grid structure in section 4.4.

We would like to reiterate that a primary objective of this study is to highlight how redundancy impacts signal EM, particularly the clock network. An abstraction of a highly redundant topology, that of a clock mesh in Fig. 4.1 serves as a good circuit to study how a system could tolerate redundancy and still not fail. Furthermore, since skew is the primary metric of the clock mesh, we make an attempt to relate our results to the skew degradation. It is the aim of this study to eventually provide benefit to the designers, who, otherwise would be fixing wires which are inconsequential to the eventual circuit performance.

4.2 Analytical Approach for Systems With Redundancy

4.2.1 Basics of Electromigration

We continue to use the popular abstraction of EM in form of Black's equation [Bla69], which describes EM lifetime using the relation (2.5), discussed in Chapter 2. For bidirectional current flow in the wires, we continue to adjust the calculations to accommodate for partial EM recovery by modifying J (which is actually a temporal average) with the help of the recovery factor, κ, that is empirically obtained [Lee12], as in (2.7). Additionally, the wire heating (ΔT) assumes an inherent dependence on the RMS current, J_{RMS}, as (2.9).

4.2.2 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. Indeed, such a scenario of "load-sharing" has been commonly dealt in mechanical engineering aspects [Bir07, MV04, HX08] and inspirations from them have been derived to address the needs with respect to current calculations.

The initial failure rate, $f(t)$, of each component is lognormal, as given by: (2.12), whereas the cumulative failure probability is directly given by (2.15).

Consider now a system comprising two components (as in Fig. 4.3), where both the components initially carry a current density J_1 (Fig. 4.4). When one of them fails at time t_1, the current in the surviving component changes to J_2.

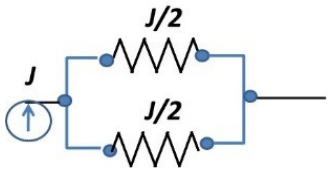


Figure 4.3: Schematic showing a parallel two-component system

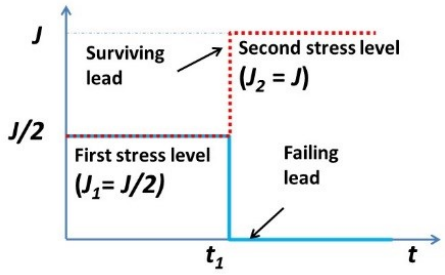


Figure 4.4: Current profile evolution, with first failure occurring at time t_1.

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

$$F_1(t) = \Phi\left(\frac{\log(t) - \log(t_{50,1})}{\sigma}\right) \tag{4.1}$$

where $t_{50,1}$ is the MTTF for J_1, as in Fig. 4.5.

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. 4.4, 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 [FCF13], [Cha13], i.e.,

$$F_2(t_1 - \delta_1) = F_1(t_1) \tag{4.2}$$

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. 4.5), the TTF changes from t_{ijh} (if only first stress were applicable) to t_{ijk} (after change of stress). The effective CDF curve (Fig. 4.5) is

$$F_1(t) = \Phi\left(\frac{\log(t) - \log(t_{50,1})}{\sigma}\right) \tag{4.3}$$

$$F_2(t - \delta_1) = \Phi\left(\frac{\log(t - \delta_1) - \log(t_{50,2})}{\sigma}\right) \tag{4.4}$$

Note that using the continuity at t_1, 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 $J_2 \cdots$ to J_k:

$$\delta_1 = t_1\left(1 - \frac{t_{50,2}}{t_{50,1}}\right) \tag{4.5}$$

$$\delta_k = \left(t_k - \sum_{i=1}^{k-1}\delta_i\right)\left(1 - \frac{t_{50,k}}{t_{50,k-1}}\right) \tag{4.6}$$

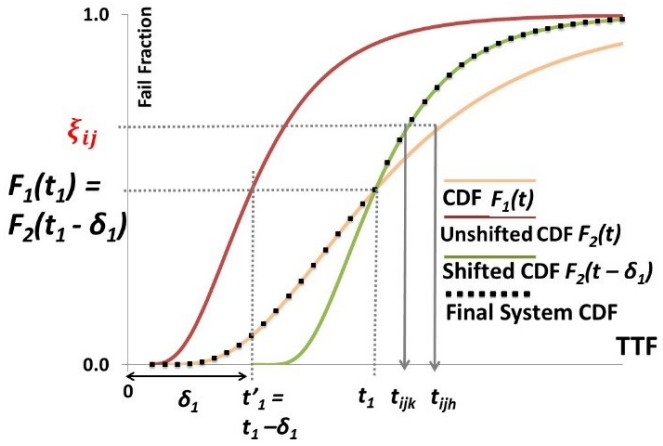


Figure 4.5: Analytically estimated CDF evolution of a single component when it undergoes a stress change. The dotted line is the effective CDF, when stress change occurs at t1.

We would like to highlight that while the present demonstration is on the basis of Black's equation, alternative approached discussed in Section 2.1.1 such as the modified Black's approach ((2.6)) or via-node based methods are also applicable in our framework, as they only change the definition of t_{50} in (4.6).

4.2.3 Reliability Calculations for System with Redundancy

We now apply this idea and basic formulation to analyze the system reliability for the structure in Fig. 4.3. 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. 4.4), the reliability of the case when both are simultaneously functional is given by:

$$R_{11}(t) = (1 - F_1(t))^2 \tag{4.7}$$

where $F_1(t)$ is as defined in Section 4.2.2.

Next, the reliability for the case when the first component fails at an arbitrary time t_1, 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_1 and $(t1 + \Delta t_1)$ 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_1, the failure statistics of the surviving component are given by the CDF $F_2(t - \delta_1)$, from (4.2). Thus, the concurrent multiplicative probability of the second component working when the first has failed is

$$(1 - F_2(t - \delta_1))\, f_1(t_1)\Delta t_1 \tag{4.8}$$

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

$$R_{12}(t) = \int_{t_1=0}^{t_1=t} (1 - F_2(t - \delta_1))\, f_1(t_1) dt_1 \tag{4.9}$$

The effective failure probability therefore is given by

$$F_{parallel}(t) = 1 - (R_{11}(t) + 2R_{12}(t)) \tag{4.10}$$

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. 4.6 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

Additionally, for this two-component system, another alternative is to use a single component of twice the width to carry the entire current, $2J_1$. 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. 4.6, 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.

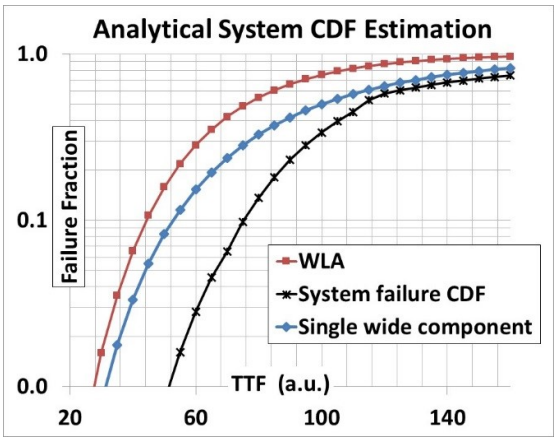


Figure 4.6: CDF for a system with redundancy, arrived using analytical formulations ((4.10)). Shown are the CDFs using the weakest link approximation (WLA), and the CDF for a single wide component.

Further, such a benefit from redundancy scales with the extent of parallelism, as illustrated in Fig. 4.7. Typically in input/output buffers and chip level power/ground networks, the wires are often required to be wide (1μm), 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. 4.7, the benefit from redundancy monotonically increases.

4.3 Monte Carlo Framework for System Reliability Estimation

The analytical two-component example in section 4.2 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

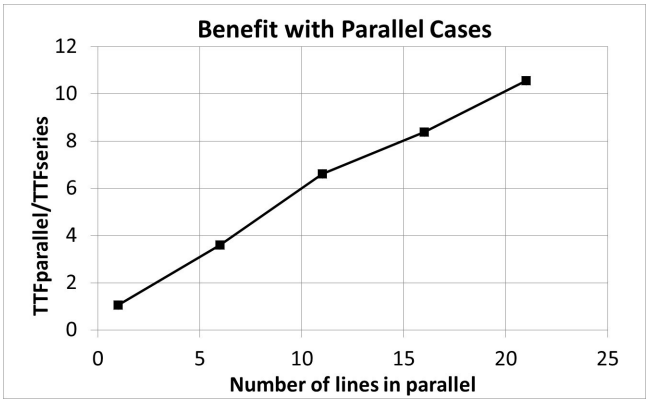


Figure 4.7: Showcasing the increasing benefit of redundancy with the number of components arranged in parallel configuration.

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 4.2.2. The component failure CDF is an unshifted lognormal before the first failure, and must be modified using equations (4.4), (4.6) 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 in Algorithm 4.1.

The WLA analysis is also conducted using stochastic MC analysis, but the first component failure is assumed to cause circuit failure. It is indeed easy to see that the WLA based TTF is available in step 9) of the algorithm. We would like to note that the complexity of above algorithm is a function of total number of resistors in the circuit (N_R), total number of Monte Carlo trials desired (mc_{limit})

Algorithm 4.1 Monte Carlo based approach for stochastic EM analysis

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

and the number of iterations (n_{iter}) which the circuit requires per MC trial to fail. Additionally, for every MC trial, we must generate a set of (N_R) random numbers, which has a complexity of $\mathcal{O}(N_R)$. Now, within each iteration of every MC trial, we require the TTF calculation for the resistors and their sorting to create an EM event on the resistor with least TTF. Thus, the overall complexity of the algorithm can be estimated as: $\mathcal{O}(mc_{limit}N_R + mc_{limit}n_{iter}N_R)$ which can be approximated to $\mathcal{O}(mc_{limit}n_{iter}N_R)$.

4.3.1 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. 4.8). Here, the only candidate EM sites are the intra-cell power/signal resistors.

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 [JJ12, Lee03].

Using the MC framework, we generate the circuit failure CDF for the 32x buffer, shown in Fig. 4.9.

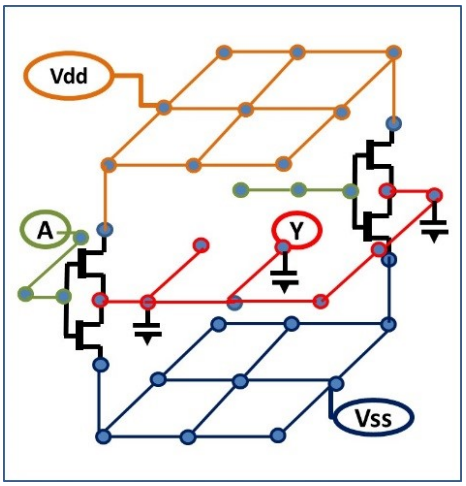


Figure 4.8: A simple high-drive (32x) buffer driving lumped load. Shown are Vdd, Vss, input and output resistors (sites for EM), analyzed stochastically.

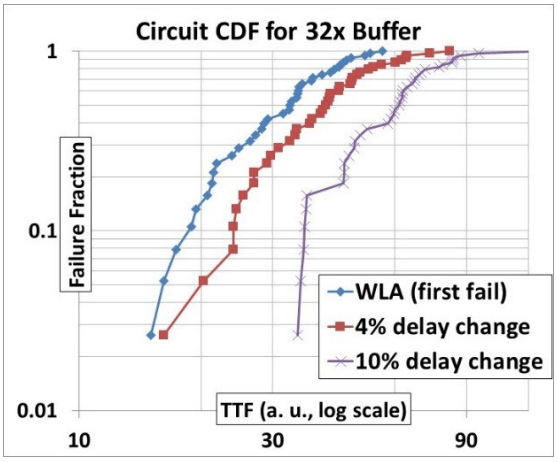


Figure 4.9: Circuit CDF showing the failure rate evolution in a single 32x drive buffer circuit (of Fig. 4.8), driving lumped load.

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.

We repeat this for the circuit failure CDF (Fig. 4.10) 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,

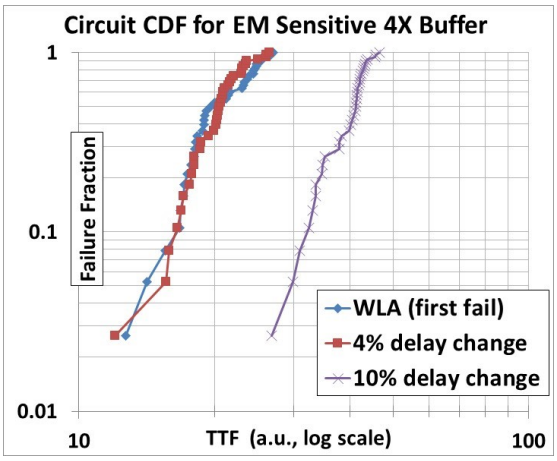


Figure 4.10: CDF for a low-drive 4x circuit, where circuit redundancies reduce, leading to an early delay-based EM failure.

we see that its failure CDF is closer to WLA.

4.3.2 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. 4.2, 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. 4.11 below.

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.

4.4 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. 4.1), 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

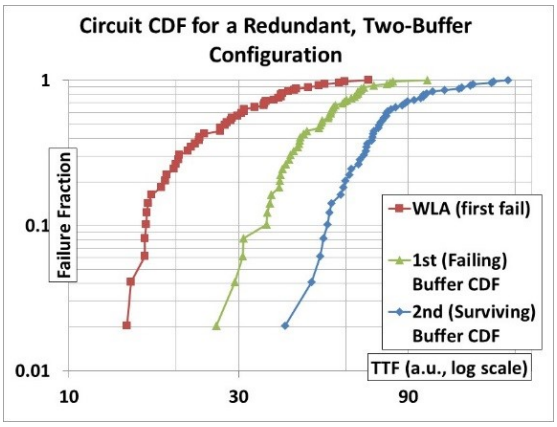


Figure 4.11: CDF for a system with two buffers arranged in a redundant configuration (as in Fig. 4.2). Significant margin is shown between TTF and the failure of first buffer. Margin builds up with addition of one more redundant buffer.

likelihood of EM failure is negligible and we focus on failures that may occur in within-cell wires or in the power grid (Fig. 4.8).

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:

- When all of the five neighboring buffers degrade by less than 2%.
- When all of the five neighboring buffers degrade in a similar, bounded manner (e.g., between 2%-4%, 4%-7%, or 7%-10%).
- 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. 4.9), as Fig. 4.12 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).

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:

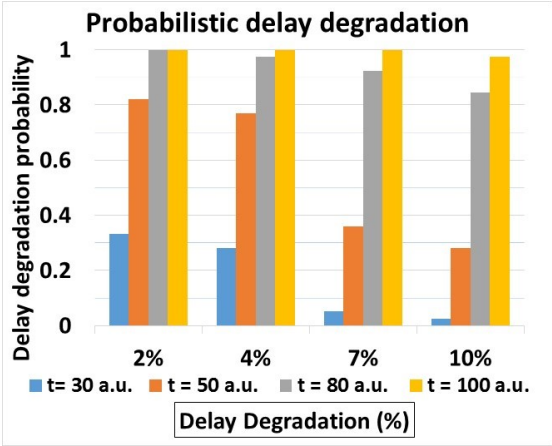


Figure 4.12: Probabilistic delay degradation with time (column-cluster represent various times).

$$\begin{cases} P_1 = (1 - F_{2\%})^5 \\ P_2 = (F_{2\%} - F_{4\%})^5 + (F_{4\%} - F_{7\%})^5 + (F_{7\%} - F_{10\%})^5 \\ P_3 = \binom{5}{2} (1 - F_{4\%})^3 (F_{7\%} - F_{10\%})^2 + \binom{5}{1} (1 - F_{2\%})^4 F_{10\%} \end{cases} \tag{4.11}$$

where $F_{x\%}$ represents the CDF of each buffer, representing the probability that the delay degradation is more than x%, and P_1 to P_3 are pass-probabilities for above cases. The combined probability therefore is given as follows:

$$F_{skew} = 1 - (P_1 + P_2 + P_3) \tag{4.12}$$

Associated CDF is as in Fig. 4.13. 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.

4.5 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,

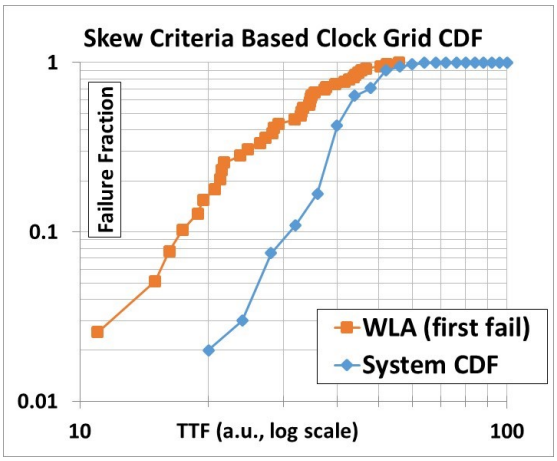


Figure 4.13: Skew-criteria based CDF of the clock-grid. For a 10% FF, about 2X margin exists between WLA and skew-criteria based failures.

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.

Chapter 5

Fast Stochastic Analysis of Electromigration in Power Distribution Networks

In this chapter, we present a fast and stochastic analysis methodology for electromigration (EM) assessment of power distribution networks. We examine the impact of variability on EM time-to-failure (TTF), considering altered current densities due to global/local process variations as well as the fundamental factors that cause the conventional EM TTF distribution. Through novel variations-aware current density model based on Hermite polynomial chaos, we demonstrate significant margins in EM lifetime when compared with the traditional worst-case approach. On the other hand, we show that the traditional approach is altogether incompetent in handling transistor-level local variations leading to significantly optimistic lifetime estimates for lower metal level interconnects of PDN. Subsequently, we attempt to bridge the conventional, component-level EM verification method to the system level failures, inspired by the extreme order statistics. We make use of asymptotic order models to determine the TTF for the k^{th} component failure due to EM, and demonstrate application of this approach in developing IR drop aware system-level failure criteria.

5.1 Circuit-level electromigration verification

Electromigration (EM) in copper interconnects is caused by the current-driven movement of metal atoms and remains the foremost challenge to interconnect reliability. The flow of a contemporary industrial EM verification cycle is represented in Fig. 5.1, which outlines a comparison of specified EM limits on the current density in an interconnect (J_{th}) against the calculated actual current density (J) in the circuit. The procedure is based upon the characterization of failure data on serially interconnected test structures [Lee03], where failure

is defined by the first break in any element of the serial structure. The time to failure (TTF) of the structure primarily depends on the current density and stress temperature, empirically related through Black's equation [Bla69, LK91], and characterization is performed under accelerated aging conditions of high voltage and temperature, in a regime where the failure fraction (FF) is high ($\sim 0.1 - 0.5$). To capture the stochastic nature of EM, the TTF is obtained over several test structures as a function of the current density, and then modeled as a lognormal distribution for the failure of a single wire.

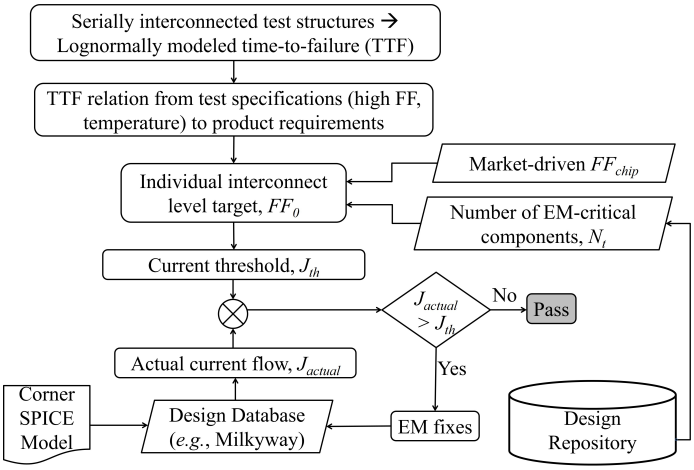


Figure 5.1: A schematic of the traditional EM verification flow.

To apply this characterized distribution to compute the EM TTF distribution in a manufactured product, it is important to account for the differences between the product use case and the characterization scenario described above. First, the product operates at a nominal temperature and voltage that is lower than the accelerated stress conditions during characterization. The TTF data based on the lognormal distribution is therefore scaled to these use case values [JCS16]. Second, while characterization addresses the failure of one wire, a chip typically consists of millions of interconnects, and the prevalent EDA approach requires the acceptable failure fraction, FF_{chip}, at the chip level to be translated to the component-level failure fraction, FF_0, on individual wires. Since $FF_0 \ll 1$, the chip- and component-level FFs can be related as [LK91]:

$$FF_{chip} = 1 - (1 - FF_0)^{N_t} \implies FF_0 \approx FF_{chip}/N_t \tag{5.1}$$

Here, N_t is the number of EM-critical resistors [LMB+11]. In a typical industry flow [ALN+15, LCBY14], the value of N_t is estimated at point of design definition, based on design experience and historical data. Using the computed value of FF_0, the lognormal PDF is used to determine the current density threshold, J_{th}, for each wire in the design.

Traditionally, amongst various interconnects, the on-chip power delivery net-

work (PDN) has been the primary EM concern. Industrial EDA tools verify these structures in a design by computing the actual current, J, in each wire at a given process corner, for a known clock frequency, application, and parasitics. Immortal wires are first filtered out using the Blech criterion [Ble76]. For each of the remaining wires, J is compared against J_{th}. In case the threshold is violated, EM fixes are invoked through PDN optimization procedures such as wire widening and current flow reduction.

5.2 Limitations of existing EM methodologies

As an increasing number of PDN wires becomes susceptible to EM in scaled technologies, several significant limitations in the methodology of Fig. 5.1 become more acute, resulting in the incorrect identification of EM-critical wires. The objective of this work is to identify these factors, as outlined in the remainder of this section, and to propose a new EM verification methodology that addresses these issues.

5.2.1 Statistical variations in J

The PDN carries both leakage and switching currents, both of which are susceptible to statistical process variations. Variations in switching current are moderate and are captured by a Gaussian, while leakage currents have a much wider non-Gaussian spread owing to the exponential dependency of leakage on threshold voltages [TN13]. In traditional designs, the switching current often limits the EM TTF, but as leakage currents become more significant, their impact can become dominant in some scenarios, particularly due to their large statistical spread. Finding the worst-case (WC) corner for current evaluation is difficult: using the timing WC corner can lead to unwarranted pessimism (up to 2×, as we will demonstrate in Table 5.1). Prior work has largely neglected statistical variations, barring a few studies that assume Gaussian variations [GMSN14, BCS+14, Lie13] and may not appropriately model leakage.

To demonstrate the impact of leakage current in on-chip power grids, we consider the example in Fig. 5.2, which represents an industrial octacore chip. The chip is shown to operate under four different workloads, in each of which a different numbers of cores is in active mode (shown using a solid outline), or in an idle or power-gated state (shown using a dotted outline). For each workload, we report the ratio of the total leakage current to the total switching current at the nominal process corner. Following common design practice, all cores share the PDN to contain the cost and complexity [big11]. This sharing results in a mix of leakage and switching currents in upper metal layers, which are illustrated through current contours overlaid on the octacore layout. For example, under workloads b) and c), the active quad cluster sees identical activity, but the cores in the other quad cluster are either idle or power gated, altering the leakage:switching current ratios.

To examine the effect of process variations, we consider the impact of global

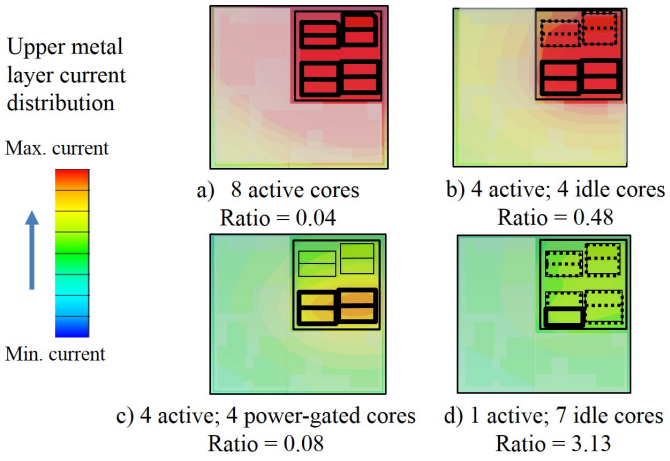


Figure 5.2: An octacore SoC, with the eight CPUs shown on the upper right, under various workloads. Depending on whether the CPUs are in active, idle, or power-gated mode, the ratio of total active power to total leakage power may vary, and the nominal current in the power grid (shown by the contours) may show different distributions.

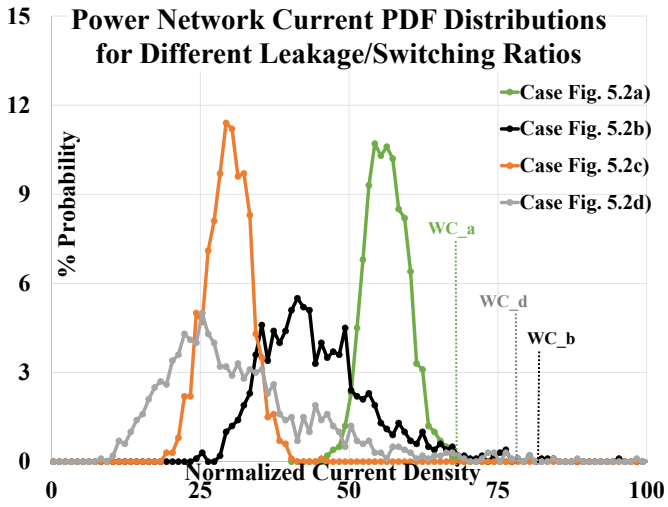


Figure 5.3: Current density PDFs in a power network for various cases mapping to Fig. 5.2.

process variations on these workloads through Monte Carlo (MC) simulations. These variations are further aggravated for local within-die variations. It can be seen that although workload a) has the highest nominal current density, workloads b) and d) have a much larger variance due to their leakage-dominant

nature. At a 99.7% yield point, the current density of 66.3 units for case a) is overshadowed by the value of 83.4 units for case b), implying that workload b) is the case that limits EM lifetime. On the other hand, at the $\sim 95\%$ yield point, the largest current density corresponds to the switching-dominated workload a).

The above data indicate that the worst estimation due to statistical variations depends not just on the workload, but also on the yield requirement. The traditional WC model is neither workload-dependent nor yield-configurable. Traditional WC SPICE models are often targeted to correlate with either 3σ transistor drive-current for the timing corner, or transistor leakage for the leakage corner, but not both, as required for EM verification. An EM-specific analysis is warranted because process variations alter both the total interconnect currents and the underlying failure kinetics of EM [HL14], neither of which is captured by the timing or leakage WC corner. Moreover, incorporating workload dependency in a WC model, i.e., deriving a unique 3σ-matching artificial process point from the current spread for every workload, is logistically impractical.

5.2.2 Outline of the proposed methodology

In the first step in Fig. 5.1, the characterization of the lognormal relies on test data collection at the foundry based on chain-like serially-interconnected test structures [Lee03]. However, the actual circuit topologies that are evaluated are much complex. To simplify the analysis of the PDN, the traditional flow uses the weakest link approximation (WLA), wherein the entire PDN is deemed to have failed when the first component fails [FP89]. By breaking down system level reliability to a component-level problem, the WLA has enabled EDA methodologies for PDN verification to scale up to millions of wires. Unfortunately, this simplification does not grasp that system-level failure occurs beyond the point of first component failure, e.g., on-chip interconnects such as the PDN can satisfy IR drop constraints even after multiple EM failures due to its inherent redundancy [JSC15, MS13]. A Monte Carlo based approach has been proposed to solve this problem [MS13], modeling a cascade of EM events leading the system to successively degraded states, but is computationally prohibitive for large systems. Therefore, there is a strong need to bridge the gap between verification scalability and system-level feedback.

In this work, we present a new EM verification procedure that modifies the traditional approach by addressing these limitations. We augment the traditional methodology in Fig. 5.1 through additional steps represented by the highlighted blocks in Fig. 5.4. Our technical contributions as follows are summarized as follows:

- We incorporate system redundancy by applying the theory of order statistics and address system failure criteria using an asymptotic failure model to determine the TTF for the k^{th} component failure due to EM. Our approach is cognizant of the typical current-day EDA framework and arrives at a modified component level FF target incorporating a known extent of system redundancy. For a given system, the extent of redundancy is computed

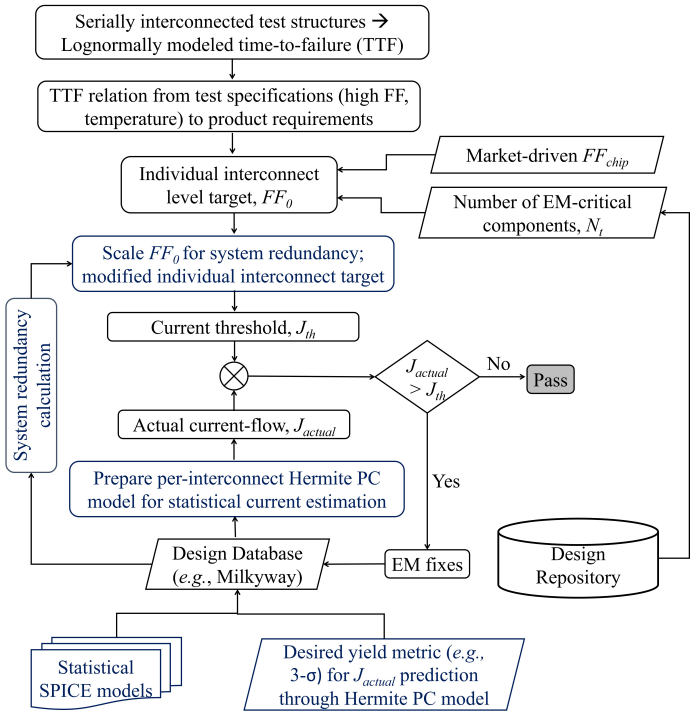


Figure 5.4: The proposed EM verification flow, where the highlighted regions indicate modifications to the traditional flow (Fig. 5.1).

one-time as average number of sustainable failures (k) through Monte Carlo means during the early design phase. As we rely on Monte Carlo simulation only for one-time, it makes our method faster against performing such simulations for every iteration till the PDN optimization is achieved. Thus, our approach efficiently bridges the gap between component and system reliability for a given system.

- To incorporate the aggravated impact of non-Gaussian statistical process variations under arbitrary workloads, we statistically derive a value for the worst-case TTF. Our model is based on multivariate Hermite polynomial chaos.

The remainder of the chapter is organized as follows. We recapitulate the mathematical basis for the current-day deterministic EM methodology in Section 5.3. Section 5.4 details the multivariate Hermite polynomial based method for incorporating global variations. Next, Section 5.5 describes a formulation for incorporating system redundancies based on extreme order statistics. Finally, Section 5.6 shows a list of experimental results and their analysis in an industrial context and Section 5.7 concludes the chapter.

5.3 Modeling EM and wire currents

5.3.1 TTF modeling

We now recapitulate the traditional, deterministic, EM modeling approach. EM failures are accelerated by two operational parameters for a circuit: the current density, J, and the temperature, T. Black's equation [Bla69] specifies the mean TTF, t_{50}, as:

$$t_{50} = \frac{A}{J^m} \tag{5.2}$$

where m is the current exponent, and a typical value is $m = 1$ [LK91]. Based on [Hau05],

$$A = \frac{L_c\, kT}{eZ\, \rho\, D_{eff}} e^{\frac{E_a}{kT}} \tag{5.3}$$

Here, E_a is the activation energy, k is Boltzmann's constant, eZ is the effective electron charge, ρ is the resistivity, L_c is the critical void length that causes a failure, and D_{eff} is the effective diffusivity, a constant.

The 50% fail fraction of (5.2) is too high for real applications. Since t_{use} is a lognormal function of random variable z_{use} [Lee03], we represent a realistic fail fraction FF at time t_{use} as:

$$FF(z_{use}) = \Phi(z_{use}) \text{ where } z_{use} = \frac{\ln t_{use}/t_{50}}{\sigma} \tag{5.4}$$

where Φ is the standard Gaussian CDF. From (5.2) and (5.4), with $m = 1$,

$$t_{use} = \frac{A}{J} e^{\sigma z_{use}} \tag{5.5}$$

Finally, we will relate the value of z and t under two different sets of stresses, a and b, with currents J_a and J_b and temperatures T_a and T_b. Typically, condition a corresponds to the reference condition provided by the foundry, and condition b is the use condition at which EM reliability is evaluated. Substituting (5.2) for both cases into the corresponding expression for z in (5.4), we obtain:

$$z_b = z_a + \frac{1}{\sigma}\left[\ln\left(\frac{t_b J_b}{t_a J_a}\right) - \frac{E_a}{k}\left(\frac{1}{T_b} - \frac{1}{T_a}\right)\right] \tag{5.6}$$

This expression relates the lognormal random variable, z, related to the failure fraction, with the time to failure, t, under two stress conditions a and b. As we will see, in Section 5.5, we formulate the shift in z required to correctly model system redundancy, and then map that solution back to an effective TTF using this equation.

5.3.2 Evaluating the PDN

A critical ingredient of TTF modeling is based on determining the current through each wire. Consider a PDN with N nodes, and let the conductance matrix of the power grid be represented by $\mathbf{G}$. Given a vector of excitations, $\mathbf{I}$, representing the current consumption at the N nodes, the behavior of the PDN is described by:

$$\mathbf{GV} = \mathbf{I} \tag{5.7}$$

where $\mathbf{V}$ is the vector of node voltages in the PDN.

Let I_k denote the k^{th} element of $\mathbf{I}$. The gate currents that contribute to I_k comprise the switching ($I_{S,k}$) and leakage ($I_{L,k}$) components, which are given by:

$$I_{S,k} = \sum_i \alpha_i C_i V_i f_i \tag{5.8}$$

$$I_{L,k} = \sum_i I_{0,i} e^{\beta_i V_t} \tag{5.9}$$

Here, α_i is the switching activity factor, V_i is the the supply voltage V_k, f_i is the operating frequency, and C_i is the equivalent switching capacitance for gate i. $I_{0,i}$ is the base leakage, and β_i is the sensitivity to the threshold voltage V_t for gate i. In both equations, the summations are taken over all gates whose currents contribute to node k. We write the switching current equation in terms of its nominal current, $I_{S,k}^{nom}$ without variations and its components due to global and local variations, $I_{S,k}^{g}$ and $I_{S,k}^{l}$, respectively, as

$$I_{S,k} = I_{S,k}^{nom} + I_{S,k}^{g} + I_{S,k}^{l} \tag{5.10}$$

Similarly,

$$I_{L,k} = I_{L,k}^{nom} + I_{L,k}^{g} + I_{L,k}^{l} \tag{5.11}$$

The switching and leakage currents are subject to process variations. The key process parameters that undergo global/local statistical variations are the capacitances and the threshold voltages, which are modeled by Gaussian distributions. These impact of the global and local variations on each component of current is described next.

Switching current: In the expression for $I_{S,k}$ in (5.8), the process-dependent term is C_i, which follows a normal distribution under global variations. This implies that its global component, $I_{S,k}^{g}$, is a weighted sum of normal distributions, which is also normally distributed. We represent $I_{S,k}^{g} \sim \mathcal{N}(\mu_{S,k}^{g}, {\sigma_{S,k}^{g}}^2)$.

Since the switching current is affected linearly by statistical local on-die variations, as seen in (5.8), the weighted sum of zero-mean Gaussian-distributed switching current terms used to obtain the local component of switching current, $I_{S,k}^{l}$, has zero mean. Further, since the summation that computes $I_{S,k}^{l}$

corresponds to the addition of a large number of gate switching currents, the variance of this sum virtually vanishes due to cancellation effects. Therefore, this variance is negligible as compared to the nominal switching current and can be neglected, i.e., $I^l_{S,k} \approx 0$.

Leakage current: Owing to global normal variations in threshold voltage, $I_{L,k}$ becomes a weighted sum of lognormal distributions. The sum of lognormal random distributions can be approximated by another lognormal distribution using a moment matching approach, similar to the widely-used Wilkinson approximation [CR00]. Using this approximation, the global leakage component, $I^g_{L,k}$, of the current at node k of the PDN is given by (5.9), and can be written as

$$I^g_{L,k} \approx I^g_{0,k} e^{\beta_k V_t^g} \tag{5.12}$$

where the terms $I^g_{0,k}$ and β_k are obtained from Wilkinson's formula. Note that in case dual V_t is used, there will be two terms in the above expression, one corresponding to each value of V_t, and the subsequent analysis is very similar.

For the leakage current component associated with local variations, denoted by $I^l_{L,k}$, since the variations are independent, it is easy to show that the variance of the sum of a large number of such terms is negligible, and that the effect of summation of a large set of independent lognormals is to shift the mean. Its local leakage component, $I^l_{L,k}$ is given by

$$I^l_{L,k} = \sum_i E[I_{0,i} e^{\beta_i V_t^l}] = I^l_{0,k} \tag{5.13}$$

When the number of leakage components to be added is smaller (e.g., on lower metal layers of the power grid), the evaluation of the sum of these lognormals proceeds using Wilkinson's method, in a manner similar to the global variation case except that the lognormals are all uncorrelated.

5.3.3 Modeling the distributions of wire current densities

To compute the current density, J_M, for any metal segment M between the nodes (i, j) of the grid one can solve (5.7) as

$$J_M = (V_i - V_j)/(R_M w_M) \tag{5.14}$$

where R_M is the branch resistance and w_M is the segment width. Technically, the current density is J_M/t_M, where t_M is the segment thickness, but we use this simpler definition since t_M is constant for a metal layer and can be incorporated into a maximum current density limit. Using (5.7), the voltage at the node i is given by:

$$V_i = \sum_{k=1}^{N} q_{ik} I_k \tag{5.15}$$

where q_{ik} is the $(i,k)^{\text{th}}$ entry of $\mathbf{G}^{-1}$. This leads to:

$$J_M = \sum_{k=1}^{N} q_{ij,k} \frac{I_k}{R_M w_M} \tag{5.16}$$

where $q_{ij,k} = q_{ik} - q_{jk}$. Note that the computation of $\mathbf{G}^{-1}$ is impractical, but we use this notion for ease of exposition. As we will soon see, this computation is not necessary.

In order to assess the impact of these variations on the segment current density J_M, we must first separately compute the leakage and switching current density components ($J_{S,M}$ and $J_{L,M}$, respectively), and subsequently determine the impact their statistics on the distribution of J_M. Noting the individual compositions of leakage and switching currents for I_k, we can represent J_M as follows:

$$J_M = \underbrace{\sum_{k=1}^{N} q_{ij,k} \frac{I_{S,k}}{R_M w_M}}_{J_{S,M}} + \underbrace{\sum_{k=1}^{N} q_{ij,k} \frac{I_{L,k}}{R_M w_M}}_{J_{L,M}} \tag{5.17}$$

Next, we show how $J_{S,M}$ and $J_{L,M}$ can be evaluated without the expensive step of explicitly inverting $\mathbf{G}$. To compute $J_{S,M}$, we begin by solving (5.7) under switching current excitations only, i.e., for the case where $I_k = I^g_{S,k}$, since $I^l_{S,k} = 0$). Through LU factorization, the system (5.7) can be rewritten as $(\mathbf{LU})\,\mathbf{V} = \mathbf{I}$, where $\mathbf{G} = \mathbf{LU}$, and the solution can be obtained through:

$$\mathbf{Ly} = \mathbf{I} \tag{5.18}$$

$$\mathbf{UV} = \mathbf{y} \tag{5.19}$$

Forward substitution in (5.18) obtains each $y_k, 1 \leq k \leq N$, as

$$y_k = \left(I_k - \sum_{l=1}^{k-1} L_{kl} y_l \right) \Big/ L_{kk} \tag{5.20}$$

Since $I^g_{S,k}$ is a Gaussian, the evaluation of each y_k involves the summation of some constants and/or Gaussians, and therefore it is easy to represent each y_k as a Gaussian.

The backward substitution step in (5.19) evaluates voltages as

$$V_k = \left(y_k - \sum_{l=k+1}^{N} U_{kl} V_l \right) \Big/ U_{kk} \tag{5.21}$$

As before, each step involves a summation of Gaussians and/or constants, and therefore each voltage can be obtained as a Gaussian distribution. Using (5.14), the switching component of each branch current density is thus expressed as a

Gaussian distribution.

A similar approach can be used to find $J_{L,M}$ in each wire. The excitation is now set either to $I_k = I^l_{L,k}$, a constant (on the upper metal layers), or to the lognormal sum (on the lower metal layers), and the first component of $J_{L,M}$ is computed for each wire as a constant. Next, setting $I_k = I^g_{L,k}$, a lognormal, we perform forward substitution, using Wilkinson's method to approximate the sums of lognormals as a lognormal at each step. The same approach is used in backward substitution, and this leads to expressing the second component of $J_{L,M}$ as a lognormal.

Finally, adding up $J_{S,M}$ and $J_{L,M}$, each branch current distribution is expressed as a sum of a Gaussian, a lognormal, and a constant.

5.4 Modeling wire current variation

To assess the impact of variations in a single wire, they must be incorporated into (5.5). We consider the electrical and physical parameters that affect the EM lifetime, namely, (a) the switching and leakage current variations, as discussed in the previous section, driven by shifts in the threshold voltage, V_t, and interconnect parasitics, and (b) EM kinetics. These are modeled as set of uncorrelated random variables.

In some contexts where the impact of perturbations is relatively small, such as statistical timing analysis, it is common to use a first-order Taylor series expansion to capture the performance impact of variations, but for EM lifetime estimation, the existence of exponential terms implies that first-order expansions are inadequate and a higher order Taylor series expansion is mandated. An alternative approach to incorporate higher order expansions and non-Gaussian variations is to treat the varying electrical and physical parameters as a continuous stochastic process. This process can be represented as an infinite series of orthogonal polynomial chaos (PC) in a Hilbert space of random variables, truncated later by finite-dimensional projections while minimizing the error. This results in a response expression as a multidimensional polynomial in the random variables that represent the variations [GS03]. Indeed, using these polynomials, much higher order expansions to capture nonlinear terms are efficiently possible when compared to perturbation techniques. Orthogonal PC based methods have been previously applied for analyzing IC performance in [VWG06,MFT$^+$08], but have not yet been employed for reliability analysis.

In this work, we employ the Hermite PC scheme and Galerkin procedure [GS03] to convert the stochastic reliability problem to a set of deterministic problems, later solved through standard matrix manipulations. This leads to the mean and variance estimation of interconnect EM lifetime, and correspondingly, a worst-case estimate (e.g., at the 3σ value).

5.4.1 Hermite PC based model

The basic principle of Hermite PC based approach is to use a series of orthogonal polynomials (of orthonormal Gaussian random variables) to facilitate stochastic analysis. Let $\xi = [\xi_1, \xi_2, \cdots, \xi_n]$ denote a vector of n independent unit Gaussian random variables, modeling the variations in V_t, interconnect, and EM kinetics. As a result, t_{use} is also a random variable that is a function of ξ, and the impact of the random variable ξ on (5.5) can be explicitly shown as:

$$t_{use}(\xi) = \frac{A(\xi)}{J(\xi)} e^{\sigma(\xi) z_{use}} \tag{5.22}$$

Based on the principle of orthogonal polynomials, t_{use} can be approximated by a truncated Hermite PC expansion:

$$\hat{t}_{use}(\xi) = \sum_{k=0}^{P} t_k H_k^n(\xi) \tag{5.23}$$

where $H_k^n(\xi)$ is an n-dimensional Hermite polynomial with deterministic coefficients t_k. The number of terms, P, is related to n [GS03]. For a single variable, ξ_1, we can represent:

$$H_0^1(\xi_1) = 1; H_1^1(\xi_1) = \xi_1, H_2^1(\xi_1) = \xi_1^2 - 1; \tag{5.24}$$

The weighting function for Hermite polynomials is the Gaussian probability density function and they are orthogonal with respect to this weighting function [Zer11] in the Hilbert space:

$$< H_i(\xi), H_j(\xi) > = < H_i^2(\xi) > \delta_{ij} \tag{5.25}$$

where δ_{ij} is the Kronecker delta and $< *, * >$ denotes the inner product. The coefficients t_k can be evaluated by the projection operation onto the Hermite PC basis. From (5.23), the mean $\mu_{t_{use}}$ and variance $\sigma^2_{t_{use}}$ of $\hat{t}_{use}(\xi)$ are:

$$\mu_{\hat{t}_{use}} = t_0 \;\; ; \;\; \sigma^2_{\hat{t}_{use}} = \textstyle\sum_{k=1}^{P} t_k^2 E[H_k^2] \tag{5.26}$$

We define the error, $\Delta(\xi)$, as the difference between the exact t_{use} and its value from (5.22), i.e., $\left(\hat{t}_{use}(\xi) - \frac{A'}{J(\xi)}\right)$, where $A' = Ae^{\sigma(\xi) z_{use}}$. This implies that

$$J(\xi)\Delta(\xi) = \left(J(\xi)\hat{t}_{use}(\xi) - A'\right) \tag{5.27}$$

To obtain t_k, we use Galerkin's method, which states that for the best approximation of $t_{use}(\xi)$, the error is orthogonal to the polynomials [GS03], i.e.,

$$< J(\xi)\Delta(\xi), H_k(\xi) >= 0, k = 0, 1, \cdots, P \tag{5.28}$$

This approach transforms the stochastic analysis to the deterministic task of computing the Hermite PC coefficients.

Next, we consider how to represent the process-dependent parameters – the leakage current $J_L(\xi)$, the switching current $J_S(\xi)$, and the term $A(\xi)$ that captures the EM kinetics – as functions of the underlying orthonormal Gaussians.

We represent the lognormal **leakage current** in a wire, J_L, as a function of the normally-distributed V_t. The global variations in V_t have mean and variance as (μ_L, σ_L^2), and in turn, are modeled using the unit normal random variable $\xi_L \sim \mathcal{N}(0,1)$. Therefore:

$$J_L(\xi) = J_0 e^{\beta(\mu_L + \xi_L \sigma_L)} \tag{5.29}$$

For the lognormal relationship of leakage current with threshold voltage, we use a second-order Hermite polynomial:

$$J_L(\xi) = \sum_{k=0}^{2} J_{Lk} H_k^n(\xi) \tag{5.30}$$

$$= J_{L0}\left(1 + \beta\sigma_L\xi_L + \frac{1}{2}\beta^2\sigma_L^2(\xi_L^2 - 1)\right) \tag{5.31}$$

Here, $J_{L0} = J_0 e^{\beta(\mu_L + \sigma_L^2/2)}$, $J_{L1} = J_{L0}\beta\sigma_L$, and $J_{L2} = J_{L0}\beta^2\sigma_L^2/2$, where (μ_L, σ_L) come from the V_t distribution.

The **switching current** in a wire, J_S, is altered linearly with the normally-distributed global capacitance variations, with mean and variance as (μ_S, σ_S^2), modeled as $\mu_s + \sigma_s\xi_s$ where $\xi_s \sim \mathcal{N}(0,1)$.

$$J_S(\xi_S) = \sum_{k=0}^{1} J_{Sk} H_k^n(\xi) = J_{S0} + J_{S1}\xi_S \tag{5.32}$$

Here, $J_{S0} = \mu_S$ and $J_{S1} = \sigma_S$.

The impact of **EM kinetics** is felt in the form of global variations in the term $A' = Ae^{\sigma z_{use}}$, caused by the process-dependent elements of the prefactor A in (5.3), and the variance, σ of the lognormal [HL14]. We model these variations in EM kinetics as lognormal, wherein the underlying normal distribution has mean and variance as (μ_K, σ_K^2) and is modeled through the unit random variable $\xi_K \sim \mathcal{N}(0,1)$. Thus, if $A_0' = Ae^{z_{use}(\mu_K + \sigma_K^2)}$, for a specified z_{use}, we can represent this nonlinear dependence using the Hermite polynomial:

$$A'(\xi_K) = \sum_{k=0}^{2} A_k H_k^n(\xi) \tag{5.33}$$

$$= A_0'\left(1 + z_{use}\sigma_K\xi_K + \frac{1}{2}z_{use}^2\sigma_K^2(\xi_K^2 - 1)\right) \tag{5.34}$$

where A_0, A_1, and A_2 can be deduced from the above.

5.4.2 Hermite PC: Coefficient estimation

Next, we assess the eventual influence of ξ_L, ξ_S, and ξ_K on the EM lifetime through the Hermite PC representation for t_{use}, represented in the second order

form as:

$$\hat{t}_{use}(\xi) = t_0 + t_1\xi_L + t_2\xi_S + t_3\xi_K + t_4(\xi_L^2 - 1) + t_5(\xi_S^2 - 1) + t_6(\xi_K^2 - 1) + t_7\xi_L\xi_S + t_8\xi_S\xi_K + t_9\xi_K\xi_L \tag{5.35}$$

To compute the Hermite PC coefficients, we apply (5.28):

$$\begin{aligned}
&< J(\xi)\Delta(\xi), 1 > = 0; && < J(\xi)\Delta(\xi), \xi_L > = 0 \\
&< J(\xi)\Delta(\xi), \xi_S > = 0; && < J(\xi)\Delta(\xi), \xi_K > = 0 \\
&< J(\xi)\Delta(\xi), \xi_L^2 - 1 > = 0; && < J(\xi)\Delta(\xi), \xi_S^2 - 1 > = 0 \\
&< J(\xi)\Delta(\xi), \xi_K^2 - 1 > = 0; && < J(\xi)\Delta(\xi), \xi_L\xi_S > = 0 \\
&< J(\xi)\Delta(\xi), \xi_S\xi_K > = 0; && < J(\xi)\Delta(\xi), \xi_K\xi_L > = 0
\end{aligned}$$

Next, we substitute $J(\xi) = J_L(\xi_L) + J_S(\xi_S)$, and use (5.27) and (5.35). Comparing the coefficients of like terms on both sides of each of the 10 inner products above, we obtain:

$$\widetilde{J}\ \widetilde{t}_{use} - \widetilde{A} = 0 \tag{5.36}$$

$$\text{where} \quad \widetilde{A} = [A'_0, 0, 0, A'_1, 0, 0, A'_2, 0, 0, 0]^T$$
$$\widetilde{t}_{use} = [t_0, t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9]^T$$

Here, $\widetilde{J}$ is the 10×10 matrix:

$$\begin{pmatrix}
J_{00} & J_{L1} & J_{S1} & 0 & 2J_{L2} & 0 & 0 & 0 & 0 & 0 \\
J_{L1} & J_{00} + 2J_{L2} & 0 & 0 & 2J_{L1} & 0 & J_{S1} & 0 & 0 & 0 \\
J_{S1} & 0 & J_{00} & 0 & 0 & 2J_{S1} & 0 & J_{L1} & 0 & 0 \\
0 & 0 & 0 & J_{00} & 0 & 0 & 0 & J_{S1} & 0 & J_{L1} \\
2J_{L2} & 2J_{L1} & 0 & 0 & 2J_{00} + 8J_{L2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & J_{S1} & 0 & 0 & 2J_{00} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & J_{00} & 0 & 0 & 0 \\
0 & J_{S1} & 0 & J_{L1} & 0 & 0 & 0 & J_{00} + 2J_{L2} & 0 & 0 \\
0 & 0 & J_{S1} & 0 & 0 & 0 & 0 & 0 & J_{00} & 0 \\
0 & 0 & 0 & J_{L1} & 0 & 0 & 0 & 0 & 0 & J_{00}
\end{pmatrix}$$

where ($J_{00} = J_{L0} + J_{S0}$). Thus, depending on coefficients of the leakage distributions (J_{L0},J_{L1} and J_{L2}), switching current distribution (J_{S0}, J_{S1}) and the EM kinetics distributions (A_0, A_1, and A_2), the matrix can be solved to find the $\widetilde{t}_{use}$, which solely defines the statistics of EM lifetime.

The TTF formulations in (5.26) can be used to determine a realistic worst-case estimate of the EM lifetime due to global variations, given by the computed $\mu_{t_{use}} - 3\sigma_{t_{use}}$ value.

5.4.3 Relevance to alternative EM checking paradigms

We would like to reiterate that the formulations developed so far have been based on void-growth dominated Black's equation (5.2): the *de facto* model applied across industry. However, it is also extendible to alternative EM check-

ing paradigms such as the modified nucleation based Black's approaches [Ori10, Llo07] or flux divergence based approaches [PJK10, GMSNL14].

For example, incorporating modified Black's equation requires alteration to the starting representation as: $t_{use}(\xi) = \left(\frac{A(\xi)}{J(\xi)} + \frac{B(\xi)}{J^2(\xi)}\right) e^{\sigma(\xi) z_{use}}$, where the void-nucleation related current term gets added. Consequently, the Hermite PC coefficients must be reworked in (5.37). On the other hand, flux divergence based approach like via-node vector method is directly applicable on (5.22) and (5.37), with a single change that the current density J, changes from an individual wire density to the effective current density at the via-node [PJK10].

5.5 EM Under circuit redundancy

The lognormal model is empirically driven and is based on the weakest link failure data from a set of test structures, i.e., it assumes that the first wire failure results in system failure However, the applicability of the weakest link model has been often questioned [LK91,HL14], since real multimillion interconnect systems have significant redundancy and survivability even after the first component failure. Some studies on power grid and signal interconnects have proposed TTF models where failure is declared when a critical system parameter (e.g., skew or IR drop) exceeds a specification [JSC15,MS13]. Notably, such exceedances occur after multiple individual components fail, highlighting the need to incorporate redundancy into EM TTF estimations.

One way to address the problem is to perform MC simulations, which can model a cascade of EM events that lead the system to successively degraded states, but this is computationally prohibitive. Alternatively, if we could develop a framework to accurately predict the time to the k^{th} component failure in a system, it could be utilized for high-level assessments of the reliability benefit due to system redundancies, or even in deriving the EM guidelines. Interestingly, this problem statement fits into the *Order Statistics* branch of EVT [Gum12, DN70], which is widely used in financial risk management. In this work, we explore its usage to predict the time to k^{th} component failure and now recapitulate some of the required mathematical concepts.

Order Statistics Distribution: If we draw n samples from a population described by a random variable z, i.e., $Z = [Z_1, Z_2, \cdots, Z_n]$, and rearrange them in increasing order as $[Z_{1,n} \leq Z_{2,n} \leq \cdots \leq Z_{n-1,n} \leq Z_{n,n}]$, then the k^{th} term, $Z_{k,n}$, of this sequence is the k^{th} order statistic. For finite n and k, we can represent the k^{th} order statistic as:

$$F_{k,n} = P(Z_{k,n} \leq z) = \sum_{i=k}^{n} \binom{n}{i} F(z)^i \left\{1 - F(z)\right\}^{n-i} \tag{5.37}$$

where $F(z)$ is the CDF of z. Intuitively, this is the probability of *at least* k (out of n) draws of Z_i to be less than or equal to z. Two common cases of order statistics are:

- **first minima**, or the first order ($F_{1,n} = 1-(1-F(z))^n$), where one draw out of n is $\leq z$.
- **first maxima**, or the last order ($F_{n,n} = F(z)^n$), where all draws out of n must be $\leq z$.

The k^{th} order statistic, $F_{k,n}$, corresponds to the k^{th} minima. In reliability terminology, $F(z)$ corresponds to the interconnect failure CDF. Consequently, the *minima* maps to the series system, where the system failure occurs at *first* component failure, whereas *maxima* maps to the parallel, where the system fails at the *last* component failure. The first *minima* forms the basis of the weakest link based EM checking practices in industry [LMB+11], whereas *maxima* is not of interest to our problem.

To assess the applicability of order statistics on TTF estimation, we generate and analyze the TTF of an ensemble of hundred independent interconnects through Monte Carlo means. We proceed in following manner:

- We first assign 1000 random FF values (generated through normal distribution) per interconnect, which are used to estimate the lognormal TTF for every interconnect.
- These hundred TTF distributions are subsequently ordered.
- The first, second, third and fourth time to failures across all hundred distributions are collected to obtain the distribution of first four failures of the ensemble.
- We plot these TTF from various failure orders on a Gumbel-form scale, $\ln(-\ln(1-CDF))$ versus $\ln(t)$.

For above case, the outcome is illustrated through Fig. 5.5, where the x-axis indicates the normalized logarithmic TTF and the y-axis is the Gumbel-form failure probability, $\ln(-\ln(1-CDF))$. Indeed, when we examine the plot of first failure (also the worst, represented in grey color) across all the hundred interconnects, we notice that it follows a linear trend with a high correlation coefficient. The second, third and fourth TTF distributions also exhibit similar behaviour. This observation is indeed a signature of ordered behaviour and a strong motivation for its applicability on TTF estimation, as also anticipated through [HL14].

Asymptotic Order Statistics: While (5.37) represented the k^{th} order statistics for finite n, MC simulation is far too expensive. We will resort to asymptotic EVT as $n \to \infty$ and $k \ll n$, both of which are relevant in dealing with power grid system of multi-million interconnects. This also has the benefit of providing closed-form analytical solutions.

For any given CDF, $F(z)$, an asymptotic maxima distribution, $Z_{n,n}$, is said to exist *iff*

$$F(a_n z + b_n)^n \to G(z) \tag{5.38}$$

where a_n and b_n are fitting constants and $G(z)$ is a function. For such cases, $F(z)$ is also said to be in the domain of maximal attraction of $G(z)$. Now, a normal CDF, $F(z)$, satisfies the convergence criteria, and the limiting minima

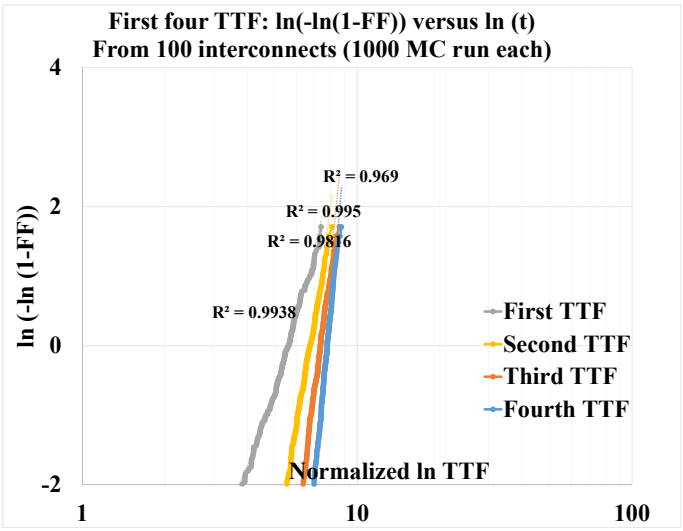


Figure 5.5: Time to k^{th} failure on a Gumbel plot demonstrating applicability of order statistics.

function is given by the Gumbel form [DN70]:

$$\hat{G}(z) = 1 - \exp(-e^z) \tag{5.39}$$

The above relation is for the asymptotic minima, but we are interested in the asymptotic k^{th} minima ($\hat{G}^{(k)}(z)$), for which we reuse the results from Gumbel [Gum12] as:

$$\hat{G}^{(k)}(z) = 1 - [\exp(-ke^{z_k})] \sum_{j=0}^{k-1} k^j \frac{e^{z_k j}}{j!} \tag{5.40}$$

where $z_k = (z - b_k)/a_k$, with a_k and b_k being the empirically derived fitting constants. The reader is referred to [Gum12] for a detailed derivation. It can be verified easily that for $k = 1$, the ordered distribution evaluates to that of the asymptotic minima, $\hat{G}(z)$.

In our methodology, we obtain initial guidance of the extent of system redundancy as an input from the designer or through system-level statistical Monte Carlo simulations. These simulations provide us the system failure CDF as well as an approximation of the number of failures the system can sustain before the system requirement breaks, which becomes our guidance for the value of k. Notice that even though the formulation (5.40) assumes that all the underlying CDFs, $F(z)$, are identical, its applicability to TTF estimation of PDN interconnects is still motivated by the fact that the PDN weaknesses are mostly found in clusters containing interconnects of similar nature. Thus, with the help

of appropriately derived fitting constants a_k and b_k to approximate the failure CDF obtained through a set of Monte Carlo simulations, we can fit the results to a Gumbel distribution to arrive at order statistics based CDF formulation.
Application for TTF Estimation:
If the system can tolerate k failures, then the k^{th} minima, $\hat{G}^{(k)}(z)$, represents the failure CDF for the system, and this can be used for reliability prediction under system redundancy as follows:

1. Given a customer requirement, FF_{chip} on the acceptable fail fraction for the chip, (5.1) can be applied to translate it to a component-level fail fraction, FF_0.

2. Based on the value of k, the number of wire failures that can be tolerated before system falure, the corresponding Gumbel distribution, $\hat{G}^{(k)}(z)$, is used to map FF_0 to $z_{use} = [\hat{G}^{(k)}]^{-1}(FF_0)$.

3. Using (5.4), the foundry failure specification, specified as $(z_{ref}, J_{ref}, t_{ref}, T_{ref})$ is translated to z_{use}, as computed above, t_{use}, the lifetime specification on the chip, and T_{use}, the operating temperature specification for reliability evaluation of the chip. This results in a current limit, J_{th}, on the wire.

4. For each wire, the actual current, J_{actual}, through the wire is compared against J_{th} to verify whether it passes EM verification or not.

Alternatively, if the lifetime associated with a current, J_{actual}, is to be computed, we may use (5.4) to translate the reference values, along with z_{use} and T_{use}, to obtain the lifetime, t_{use}.

5.6 Results

We now present the results from the methods developed so far. For consistency, all of our results and implementation are based on the standard IBM power grid benchmark circuits [Nas08]. We take the technology constants from an industry environment, and we normalize the data for confidentiality. We proceed as follows:

- Firstly, we validate the Hermite PC based framework to model the statistical variability in Section 5.6.1. We benchmark our results against statistical SPICE simulations (involving transistor global/local variability). The outcome of this statistical analysis is the 3σ current density for every resistor, which can be then used for verification against the EM thresholds.
- Subsequently, we take the current densities of individual resistors and incorporate system redundancy through order statistics based approach in 5.6.2. We benchmark our results against system-level statistical reliability simulations [JSC15].

5.6.1 Statistical Variability Estimation

Power Network: Experimental Setup

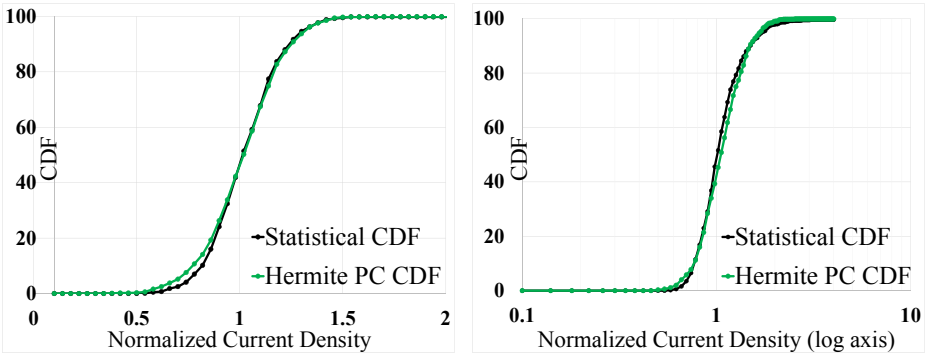
The IBM PG grid benchmarks are representative of different classes of industrial designs and vary over a reasonable range of complexity. The original intent of these PG grids was to benchmark improvements in the capability of simulation tools for PG grid simulation. Therefore, they only contain grid parasitics and voltage/current sources, instead of transistor level subcircuits of standard cells. The grids represent various blocks on the chip, and each block contains number of current sources as a representative of the switching instances of standard-cells. These current sources are DC and correspond to the cumulative current requirement of the standard-cell (including switching and leakage contributions).

In contrast with this original intent, our primary purpose is to study the impact of statistical variations (in the form of transistor V_t variations and switching capacitances) on EM lifetime, and therefore, we must separate the current-flow per standard cell in switching and leakage components. This is achieved by assigning Gaussian and lognormal distributions to the switching and leakage components, respectively, of the current sources. The nominal value of the cumulative current requirement of the standard-cell matches the original IBMPG grid value, and the ratio between nominal values of leakage and switching contributions is globally controlled. We utilize the HSPICE Monte Carlo variability simulation setup, in which the parameter governing leakage variation of the standard-cell could be treated as a globally (and identically) distributed or locally (and independently) distributed parameter. In this way, through statistical SPICE simulations, we can arrive at the current distribution spread in every resistor of the power grid, considering global and/or local variations. Subsequently, we derive (a) Hermite PC model and (b) WC based estimates to perform comparison against the statistical results.

Evaluation of the Hermite PC approach

Through statistical SPICE simulations on the IBMPG2 grid, with global variations enabled, we directly obtain the current density CDFs for these resistors. These CDFs are plotted through black solid lines in Fig. 5.6 for two cases, representing the switching-dominated case and a case with significant leakage contributions. All data is normalized with respect to its median value.

We validate the results of our Hermite PC model under global variations against these simulations. We follow the procedure outlined earlier in Section 5.3, and compute the nominal values of leakage and switching currents per resistor. For the resistors in the power grid, we extract the $q_{ij,k}$ coefficients in (5.16). As outlined earlier, these are used to derive the variance of leakage and switching current density per wire using Wilkinson's method, noting that these variations are identical and fully correlated. Subsequently, using (5.37), we set up the Hermite PC model and evaluate it to produce the CDFs of the current density, as also plotted through Fig. 5.6 in green solid lines. As we can see, for



(a) Gaussian current-density distribution

(b) Non-Gaussian current density distribution

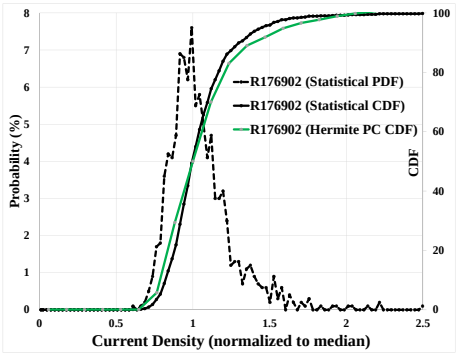
Figure 5.6: Current density PDFs and CDFs derived through statistical SPICE simulations and Hermite PC based approach for Gaussian and non-Gaussian cases.

both the cases, the Hermite PC based CDF is in close agreement with the CDF derived through statistical SPICE simulations.

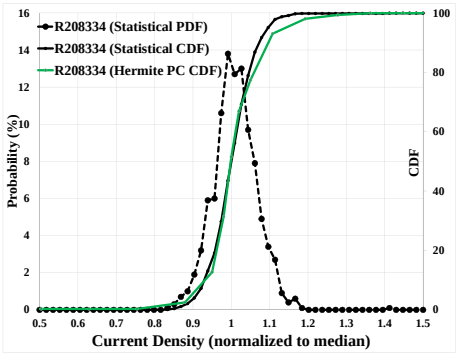
Next, to evaluate the Hermite PC based model under local variations, we first set the leakage variability parameter in HSPICE to be locally and independently distributed. Thus, through statistical SPICE simulations, we can directly obtain the current density CDFs for various resistors in the power grid under local variations. For illustration purposes, we pick resistors corresponding to lower, middle and upper metal layers of the PG grid, and their CDFs are plotted using black solid lines in Fig. 5.7 where the data for every resistor is normalized with respect to its median value. Table 5.1

As discussed earlier in Section 5.4, the current density in a given resistor is a weighted summation of the individual gate currents ((5.16)), wherein the position of the resistor in the power grid affects the weightages. Unlike global variations, which are fully correlated, the local variations are independent and uncorrelated. Thus, a resistor in lower metal layer, which sees the current-flow primarily from adjacent standard-cells, experiences a relatively high spread in its distribution, as shown through Fig. 5.7a. On the other hand, as we move to upper metal layers, the number of influencing standard-cells for a given resistor keeps increasing, resulting in a much tighter distribution shown through Figs. 5.7b and 5.7c (Note that the x-axis range grows progressively smaller in these figures). The CDFs obtained from the Hermite PC model are shown in Fig. 5.7 through green solid lines. For all the three cases, the Hermite PC based CDF correlates very well with the CDF derived through statistical SPICE simulations.

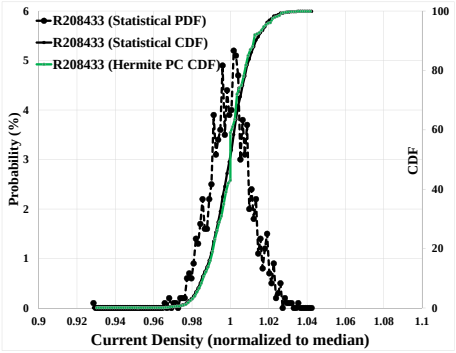
The influence of wires in various metal layers is further illustrated when we graphically plot the $q_{ij,k}$ coefficients in (5.16) for these resistors in Fig. 5.8. We normalize and rank order the coefficients for the three resistors and plot the top 100 values. As we can see, for resistor R176902, which is on lower metal layer,



(a) Statistical currents for lower metal layer resistor: R176902



(b) Statistical currents for mid metal layer resistor: R208334



(c) Statistical currents for upper metal layer resistor: R208433

Figure 5.7: Current density PDFs and CDFs derived through statistical SPICE simulations and Hermite PC based approach for three resistor cases, corresponding to lower, mid and upper metal layers, incorporating local variations. Distributions becomes narrower as the resistors move to upper metal layers.

there are only few cells that influence its value, as signified by fewer coefficients with high values. On the other hand for resistor R208433 on a middle metal layer, there are more cells with high coefficients. Using these coefficients in Wilkinson's method, and noting that these variations are now uncorrelated, we can compute the leakage and switching variances per resistor, eventually leading to the Hermite PC based model using (5.37). Thus, the impact of local variations is more prominent on lower metal layers and Hermite PC based model rightly comprehends this behaviour.

Next, we estimate the current density from a timing-based WC approach, which assumes a strong transistor (optimized to 3σ transistor current) and worst-case capacitance. We set the individual standard-cell current sources to their corresponding timing-based WC values and perform the current mea-

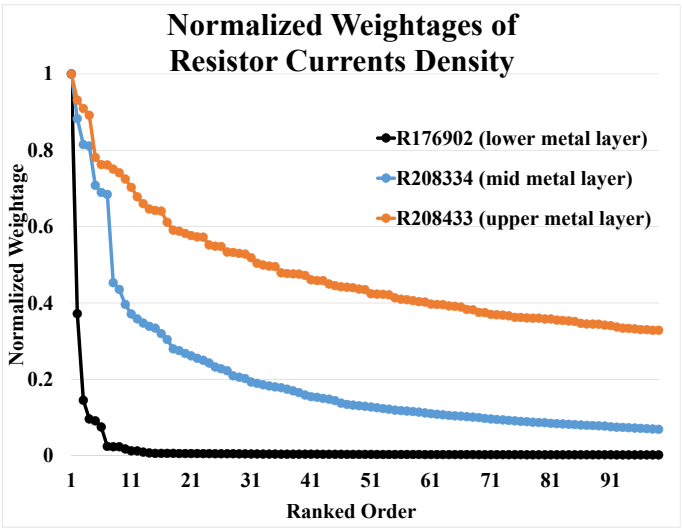


Figure 5.8: Normalized and ranked-order sensitivity of different resistors to individual current sources. x-axis indicates the identifier for one amongst several current-sources in the design.

Design Entity	Distribution Spread (with local variations) 3σ/Mean	t_{use} Improvement (with global variations) 3σ/WC
R176902 (lower metal layer)	2.38	1.37
R208334 (mid metal layer)	1.15	1.32
R208433 (upper metal layer)	1.03	1.21

Table 5.1: Comparison of our analytical EM lifetime prediction against a timing-based WC approach.

surements for the entire power grid for a given ratio of leakage and switching current per cell. For the same scenario, we also compute the lifetime from the Hermite PC model and the results are as tabulated in the last column of Table 5.1, where, we can see that Hermite PC based lifetime estimates are over 1.3× better than the WC ones.

In summary, our Hermite PC model agrees very well with statistical SPICE estimates for comprehending global as well as local variations. For the cases studied, our model reduces pessimism in EM lifetime estimates as compared to WC approach for global variations, whereas the timing-based WC approach is altogether incompetent in modeling local variations. Additionally, our model can be directly applied to project the lifetime for a given failure fraction requirement, thus enabling the yield tradeoffs which provides an opportunity to designers to cope with outlier violations.

Grid name	Total voltage sources	Total interconnects	Average failures required for system failure (rounded)	Normalized mean TTF (hrs)		
				WLA	System MC	Proposed model
IBMPG1	100	14,750	3	1.000	1.626	1.507
IBMPG2	120	101,718	8	0.571	0.999	1.097
IBMPG3	494	677,388	10	0.593	1.089	1.120
IBMPG4	312	780,699	8	0.696	1.220	1.352
IBMPG5	100	495,756	7	0.568	0.970	1.073
IBMPG6	132	797,711	9	0.612	1.186	1.160
IBMPGNEW1	494	698,595	18	0.754	1.566	1.577
IBMPGNEW2	494	1,157,849	16	0.807	1.590	1.866

Table 5.2: Comparison of our analytical EM lifetime prediction against Monte Carlo and WC approach, performed on different power grid benchmarks. The failure criterion is 50mV higher voltage drop on any node as compared to its $t = 0$ value.

5.6.2 Application of Order Statistics

We now present results based on the methodology developed in Section 5.5 on the IBM power grid benchmark circuits [Nas08], wherein our intention is to mimic the system failure through order statistics based prediction. We first discuss the methodology for progressive interconnect degradation in the PG grid resulting in increased IR drop. We define system failure when the IR drop increases by 50mV at any node in the grid from its $t = 0$ value.

The PG benchmark circuits are solved to obtain the currents and IR drop at every node; the currents are scaled to create a 100mV IR drop at $t = 0$. We obtain the system failure time through MC simulation, similar to the method in [JSC15]. Each MC simulation involves:

- computing $t = 0$ currents and assigning normally distributed random FF to resistors
- using the current-flow, operating temperature and FF per resistor to derive the TTF for every resistor ((5.6))
- rank-ordering of resistors to open-circuit the resistor with least TTF
- recomputing currents and IR drop, and
- iterating till system criterion failure criterion is met.

We perform 100 MC simulations to obtain the statistical distribution, with different numbers of iterations per simulation warranted by different PG benchmark circuits to breach the system failure criterion.

Specifically looking at IBMPG1, from the MC runs, the TTF for the first, third and fifth (out of ∼15K) failing resistors from every MC simulation can be obtained. Next, using the $t = 0$ currents data and the transformation from z to t, we obtain the ordered statistics model from (5.40). The TTF distributions for both the MC-based and the ordered statistics model are plotted in Fig. 5.9, where the x-axis is the normalized time. The region of interest corresponds to small FF values (< 0.25), and in this region, order-based failure estimations are in good agreement with MC estimations.

Additionally, from the MC simulations, we also extract the system TTF,

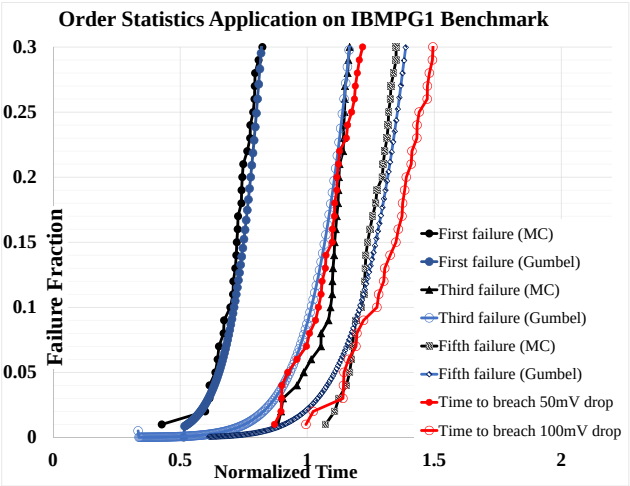


Figure 5.9: Application of Order Statistics Based EM Prediction on PG benchmark, IBMPG1.

i.e., the time at which, due to progressive EM degradation, any node in the PG grid suffers a drop that is 50mV higher than its $t = 0$ value. This data is also plotted in Fig. 5.9. It is clear that this system-level criterion is violated *much* later than the first component failure. Moreover, the system failure CDF can be approximated by the k^{th} component failure CDF. Here, the third failure CDF could be used to approximate the TTF to a 50mV drop. Notice that the order statistics based model (5.40) is guided one-time through the MC means to derive the average number of components which must fail before the system failure. This is an important result, since even though MC based system TTF estimation provides an accurate picture of failures, performing full MC analysis on a production PG grid is computationally prohibitive for every iteration cycle till the PDN optimization is achieved.

Next, we look at larger PG grid benchmark circuits for applicability of this principle. For one such grid, IBMPGNEW1, the voltage drop maps at $t = 0$ is shown at left in Fig. 5.10, which shows the inherent drop of the circuit before any wire failures. As the stress builds and we take the system through progressive degraded states, at the point of system failure, there will be at least one node in the PG grid whose voltage drop is 50mV higher as compared to its $t = 0$ value. The voltage drop map at that time instant is shown at right in Fig. 5.10. Notice that prior to this point, the circuit is functional and is able to tolerate several EM failures.

The results for the other power grid benchmarks are presented in Table 5.2. As noted before, we perform 100 MC simulations to obtain the statistical distribution, with different numbers of iterations per simulation warranted by different PG benchmark circuits to breach the system failure criterion. Here, we list the average number of failures required per PG grid to violate the system crite-

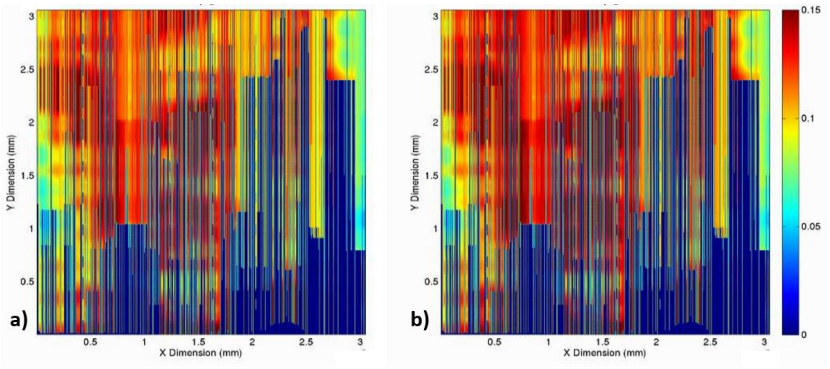


Figure 5.10: Voltage drop maps of the power grid, IBMPGNEW1 (left) at $t = 0$, showing the inherent IR drop of the circuit with no wire failures (right) after the circuit undergoes 20 EM events, after which there is at least one node whose voltage drop is 50mV higher as compared to its $t = 0$ value. The IR drop scale is described at right.

rion (rounded off to nearest integer). We also report the normalized mean TTF (hrs) from WLA, system MC and proposed order based method. Recall that as the order method requires only one-time estimation of system redundancy through MC means, it is computationally superior as compared to performing costly MC sims for every iteration of design-closure till the PDN optimization is achieved. As we can see from the Table 5.2, the lifetime estimates from the proposed method agrees well with the more expensive MC-based computation and reduce significant pessimism as compared to the WLA method. It must be mentioned, however, that due to topological differences, the number of interconnects required to create a system failure are different for various grids (column 4 in Table 5.2), and that is an input to the order statistics based TTF generation procedure.

Lastly, we share the results of EM thresholds based verification from the proposed order based method, using the guidance of average number of failures required per grid. The default thresholds are targetted for a tighter reliability specification so as to expose the EM violations. To derive the thresholds from order approach, we follow the steps enumerated earlier in Sec. 5.5, and the results are as shown in Table 5.3. While Table 5.2 establishes the accuracy of order based method against the system level MC simulations, we now see the translation of lifetime benefit in form of reduced number of violations. Indeed, for IBMPG2 circuit, a designer must *only* fix 85 (of 372) violations and still, safe operation of the circuit is guaranteed as per the given system criterion. Thus, using the order statistics model, we demonstrate that the system time-to-failure can be approximated analytically as well as its application for threshold based EM verification in a conventional context.

Grid name	Number of EM violations	
	Default thresholds	Order model
IBMPG2	372	85
IBMPGNEW2	116	41

Table 5.3: Application of order model for threshold based verification of circuits.

5.7 Conclusion

A fast and stochastic methodology for electromigration analysis of power distribution networks was been presented in this chapter. The impact of statistical global/local process variation on EM TTF assessment has been examined. Through novel variations-aware current density models, we demonstrate significant margins in EM lifetime when compared with the traditional worst-case approach. On the other hand, we show that the traditional approach is altogether incompetent in handling transistor-level local variations leading to significantly optimistic lifetime estimates for lower metal level interconnects. Additionally, we show that the traditional component-level model is inadequate in predicting the system failure since system failure often occurs after multiple components fail. Through an extreme order statistics based approach, we have demonstrated that system failures, in form of IR drop exceedances, can be approximated reasonably by an asymptotic k^{th} component failure model. As our method requires only one-time estimation of system redundancy through Monte Carlo means, it is computationally superior as compared to performing costly MC simulations for every iteration of design-closure till the PDN optimization is achieved. Thus, our approach efficiently bridges the gap between component and system reliability for a given system.

Chapter 6

Conclusions

Algorithms and methodologies for improving interconnect reliability analysis of modern integrated circuits were presented in this research work. Improvements in specifically three aspects were done:

1. As a first part, a new methodology for correct-by-construct design and verification of logic-IP (cell) internal EM verification was presented. We proposed sophisticated models for estimating the current-flow within any cell-internal metal segment under any arbitrary chip level specification of voltage, loads, slew, and activities. Our method achieves SPICE-like accuracy by incorporating the impact of arbitrary parasitic loading, and, an intelligent way of coming up with the effective pin capacitance of load cells. The methodology was shown to be highly flexible, in terms of allowing on-the-fly retargeting for the reliability. Finally, the complete data generation process at library level is expedited by application of cell response modeling. Results on a 28nm production setup were shared, to demonstrate significant relaxation in terms of violations, along with close correlation to SPICE. We shared various cases of runtime-level reliability retargeting, by specifying varying reliability conditions for the production block verification.

2. A second part of this research focusses on an important aspect of connecting the individual component-level failures to that of the system failure. We noted that existing EM methodologies are based on *serial* reliability 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 in perspective, we presented algorithms for EM assessment, which allow us to incorporate and quantify the benefit from system redundancies. We demonstrated that unless such incorporations are done, chip lifetimes are underestimated by over 2x! Using such a reliability criteria, we further studied the skew evolution in clock grids and demonstrated lifetime benefit.

3. Finally, the impact of statistical global/local process variation on EM time-to-failure (TTF) assessment was examined in the last part of the work. Through novel variations-aware current density models, we demonstrate significant margins ($> 30\%$) in EM lifetime when compared with the traditional worst case approach. Additionally, we show that the traditional component-level model is vastly inadequate in predicting the system failure. Through an extreme order statistics based approach, we demonstrated that system failures can be approximated reasonably by an asymptotic k^{th} component failure model, which otherwise requires costly Monte Carlo simulations.

Above contributions take forward the status-quo in interconnect reliability verification in a significant manner by presenting systematic algorithms to incorporate systemic redundancies into lifetime assessment, as well as the connect between component and system failures.

Scope for Future Work

Although being a traditionally known design challenge, accurate reliability estimation of integrated circuits continues to gain traction as designers seek to regain the performance loss by adhering to blanket and strict reliability guidelines. This thesis highlighted, with rigorous examples, such instances and proposed systematic methodologies for incorporating them in design analysis. This thesis also demonstrated application of ideas from seemingly distant fields, for example, redundancy and finance, leading to new ways of attacking and solving reliability problems.

In the same spirit, there are several avenues to further extend this work. A natural extension of methodologies developed in the cell-internal aspect of the research is to compiler memories and analog circuits. While compiler memories are highly systematic, the constraints at which they were designed, and get used, are rarely the same leading to significant re-verification cost.

On redundancy aware electromigration verification aspect, it would be interesting to extend the framework developed to other reliability effects, primarily the oxide breakdown. Similar to EM, oxide breakdown is a fundamentally statistical phenomenon and further, the impact of breakdown on the eventual circuit performance is context sensitive. This makes direct application of EM framework to oxide breakdown assessment highly plausible.

Other reliability effects, namely, hot-carrier injection (HCI) and bias temperature instability (BTI) are not that suitable for direct application due to their dominant deterministic nature.

Lastly, the variation aware EM methodology, developed in Chapter 5 can be directly extended for Signal EM assessment. The problem undergoes multiple changes from Power EM to Signal EM. The long leakage tail diminishes for Signal EM, however, the dependence on RMS currents increases.

Bibliography

[ALN+15] J.-G. Ahn, M. F. Lu, N. Navale, D. Graves, P.-C. Yeh, J. Chang, and S. Y. Pai. Product-level reliability estimator with budget-based reliability management in 20nm technology. In *Proceedings of the IEEE International Reliability Physics Symposium*, pages 6B–2.1–6.B–2.5, 2015.

[ALT14] Altos tool user-manual. `http://www.cadence.com`, 2014.

[ALTT04] Syed M Alam, Gan Chee Lip, Carl V Thompson, and Donald E Troxel. Circuit level reliability analysis of Cu interconnects. In *Quality Electronic Design, 2004. Proceedings. 5th International Symposium on*, pages 238–243. IEEE, 2004.

[AUT15] Automotive Processors Overview. `http://www.ti.com/lsds/ti/processors/dsp/automotive_processors/overview.page`, 2015.

[BAK+13] Tom Burd, Yuri Apanovich, Srinivasaraghavan Krishnamoorthy, Vishak Kumar Venkatraman, and Anand Daga. Interconnect and transistor reliability analysis for deep sub-micron designs, January 15 2013. US Patent 8,356,270.

[BB13] John E Barwin and Jeanne PS Bickford. Method of managing electromigration in logic designs and design structure thereof, October 15 2013. US Patent 8,560,990.

[BCS+14] Y. Ban, C. Choi, H. Shin, Y. Kang, and W. H. Paik. Analysis and optimization of process-induced electromigration on signal interconnects in 16nm FinFET SoC (system-on-chip). In *SPIE Advanced Lithography*, pages 90530P–1–90530P–11, 2014.

[big11] High-level Considerations for Power Management of a big.LITTLE System. `http://www.infocenter.arm.com/help/topic/com.arm.doc.dai0424a/DAI0424A.pdf`, 2011.

[Bir07] Alessandro Birolini. *Reliability engineering*, volume 5. Springer, 2007.

[Bla69] James R Black. Electromigration failure modes in aluminum metallization for semiconductor devices. *Proceedings of the IEEE*, 57(9):1587–1594, 1969.

[Ble76] I. A. Blech. Electromigration in thin aluminum films on titanium nitride. *Journal of Applied Physics*, 47(4):1203–1208, 1976.

[BM01] Kaustav Banerjee and Amit Mehrotra. Coupled analysis of electromigration reliability and performance in ULSI signal nets. In *Proceedings of the 2001 IEEE/ACM international conference on Computer-aided design*, pages 158–164. IEEE Press, 2001.

[BOZD03] David T Blaauw, Chanhee Oh, Vladimir Zolotov, and Aurobindo Dasgupta. Static electromigration analysis for on-chip signal interconnects. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 22(1):39–48, 2003.

[CCF+07] Jeonghwan Choi, Chen-Yong Cher, Hubertus Franke, Henrdrik Hamann, Alan Weger, and Pradip Bose. Thermal-aware task scheduling at the system software level. In *Proceedings of the 2007 international symposium on Low power electronics and design*, pages 213–218. ACM, 2007.

[CCS05] CCS technology, 2005.

[Cha13] S. Chatterjee. *Redundancy-Aware Electromigration Checking for Mesh Power Grids*. PhD thesis, University of Toronto, 2013.

[CR00] P. Cardieri and Theodore S. Rappaport. Statistics of the sum of lognormal variables in wireless communications. In *Proceedings of the Vehicular Technology Conference Proceedings*, volume 3, pages 1823–1827, 2000.

[CW03] John F Croix and DF Wong. Blade and razor: cell and interconnect delay analysis using current-based models. In *Design Automation Conference, 2003. Proceedings*, pages 386–389. IEEE, 2003.

[DFN06] L Doyen, X Federspiel, and D Ney. Improved bipolar electromigration model. In *2006 IEEE International Reliability Physics Symposium Proceedings*, pages 683–684. IEEE, 2006.

[DN70] Herbert Aron David and Haikady Navada Nagaraja. *Order statistics*. Wiley Online Library, 1970.

[Dod15] Jim Dodrill. The Frontiers of Robust Circuit Design in Sub28nm Process Technologies. Tutorial, ISQED 2015, 2015.

[EDI15] Encounter design tool user manual. `www.cadence.com`, 2015.

[ELE11] A matter of scale: Electromigration. `http://engineerblogs.org/2011/03/a-matter-of-scale-electromigration/`, 2011.

[FCF13] M. Fawaz, S. Chatterjee, and Najm F. Vectorless framework for power grid electromigration checking. In *Proceedings of the 2013 IEEE/ACM international conference on Computer-aided design*. IEEE Press, 2013.

[FP89] David F Frost and Kelvin F Poole. Reliant: a reliability analysis tool for VLSI interconnect. *Solid-State Circuits, IEEE Journal of*, 24(2):458–462, 1989.

[GF16] Global Foundaries Reference Signoff Flow. `http://www.globalfoundries.com/services/globalsolutions-ecosystem`, 2016.

[GMSN14] Zhong Guan, Malgorzata Marek-Sadowska, and Sani Nassif. Statistical analysis of process variation induced SRAM electromigration degradation. In *Quality Electronic Design (ISQED), 2014 15th International Symposium on*, pages 700–707. IEEE, 2014.

[GMSNL14] Zhong Guan, Malgorzata Marek-Sadowska, Sani Nassif, and Baozhen Li. Atomic flux divergence based current conversion scheme for signal line electromigration reliability assessment. In *Interconnect Technology Conference/Advanced Metallization Conference (IITC/AMC), 2014 IEEE International*, pages 245–248. IEEE, 2014.

[Gra03] Ananth Grama. *Introduction to parallel computing*. Pearson Education, 2003.

[Gre11] P Greenhalgh. big. little processing with arm cortex-a15 and cortex-a7. *Citado na*, page 46, 2011.

[GS03] Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. Courier Corporation, 2003.

[GS12] Saket Gupta and Sachin S Sapatnekar. Compact current source models for timing analysis under temperature and body bias variations. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 20(11):2104–2117, 2012.

[Gum12] Emil Julius Gumbel. *Statistics of extremes*. Courier Corporation, 2012.

[Hau05] M. Hauschildt. *Statistical analysis of electromigration lifetimes and void evolution in Cu interconnects*. PhD thesis, University of Texas at Austin, Austin, TX, USA, 2005.

[HL14] Andrea E Hubbard and JR Lloyd. Study of the variation in electromigration performance from lot to lot variations and the implications for design rule generation. In *Integrated Reliability Workshop Final Report (IIRW), 2014 IEEE International*, pages 123–126. IEEE, 2014.

[HL16] M.H Hsieh and Y.H. Lee. Timing Characterizations of Device and CPU-like Circuit to Ensure Process Reliability. In *Reliability Physics Symposium (IRPS), 2016 IEEE International*, pages 4C–1. IEEE, 2016.

[HLK+00] D. Hisamoto, Wen-Chin Lee, J. Kedzierski, H. Takeuchi, K. Asano, C. Kuo, Erik Anderson, Tsu-Jae King, J. Bokor, and Chenming Hu. FinFET-a self-aligned double-gate MOSFET scalable to 20 nm. *Electron Devices, IEEE Transactions on*, 47(12):2320–2325, Dec 2000.

[HM07] Steven Herbert and Diana Marculescu. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In *Low Power Electronics and Design (ISLPED), 2007 ACM/IEEE International Symposium on*, pages 38–43. IEEE, 2007.

[HSK+16] Xin Huang, Valeriy Sukharev, Taeyoung Kim, Haibao Chen, and Sheldon X-D Tan. Electromigration recovery modeling and analysis under time-dependent current and temperature stressing. In *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 244–249. IEEE, 2016.

[HSP12] HSPICE User Manual. `http://www.synopsys.com`, 2012.

[Hun97] William R Hunter. Self-consistent solutions for allowed interconnect current density. II. Application to design guidelines. *Electron Devices, IEEE Transactions on*, 44(2):310–316, 1997.

[HX08] Guang Bo Hao and Li Yang Xie. Damage equivalent method of fatigue reliability analysis of load-sharing parallel system. In *Advanced Materials Research*, volume 44, pages 853–858. Trans Tech Publ, 2008.

[HYST14] Xin Huang, Tan Yu, Valeriy Sukharev, and Sheldon X-D Tan. Physics-based electromigration assessment for power grid networks. In *Proceedings of the 51st Annual Design Automation Conference*, pages 1–6. ACM, 2014.

[ICF14] Intel Custom Foundary: Reference Reliability Signoff Flow Using Ansys Tools. `https://newsroom.intel.com/news-releases/ansys-and-intel-collaborate-to-deliver-power-em-and-reliability-sign-off-reference-flow-for-customers-of-intel-custom-foundry/`, 2014.

[IOT15] Internet of Things. `http://share.cisco.com/internet-of-things.html`, 2015.

[ITR15] ITRS Roadmap. `http://www.itrs.net`, 2015.

[Jai07] P. Jain. Design-in-reliability: Interconnects perspective. Invited talk, 1st International Design for Variability Workshop, 2007.

[JCPA14] P. Jain, F. Cano, B. Pudi, and N.V. Arvind. Asymmetric aging: Introduction and solution for power-managed mixed-signal socs. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 22(3):691–695, March 2014.

[JCS16] Palkesh Jain, Jordi Cortadella, and Sachin S Sapatnekar. A Fast and Retargetable Framework for Logic-IP-Internal Electromigration Assessment Comprehending Advanced Waveform Effects. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(6):2345–2358, 2016.

[JED16] JEDEC Reliability Standards. `http://www.jedec.org`, 2016.

[JJ12] P. Jain and A. Jain. Accurate current estimation for interconnect reliability analysis. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 20(9):1634–1644, Sept 2012.

[JSC15] Palkesh Jain, Sachin S Sapatnekar, and Jordi Cortadella. Stochastic and topologically aware electromigration analysis for clock skew. In *Reliability Physics Symposium (IRPS), 2015 IEEE International*, pages 3D–4. IEEE, 2015.

[KBT+93] MA Korhonen, P Bo, KN Tu, Che-Yu Li, et al. Stress evolution due to electromigration in confined metal lines. *Journal of Applied Physics*, 73(8):3790–3799, 1993.

[LCBY14] B. Li, C. Christiansen, D. Badami, and C.-C. Yang. Electromigration challenges for advanced on-chip Cu interconnects. *Microelectronics Reliability*, 54(4):712–724, 2014.

[Lee03] Ki-Don Lee. *Electromigration Critical Length Effect and Early Failures in Cu/oxide and Cu/low k Interconnects*. PhD thesis, University of Texas at Austin, Austin, Texas, USA, 2003.

[Lee12] Ki-Don Lee. Electromigration recovery and short lead effect under bipolar-and unipolar-pulse current. In *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pages 6B–3. IEEE, 2012.

[Lee13] Ki-Don Lee. CMOS BEOL Metal Reliability (Electromigration and Stressmigration). Tutorial at the IEEE International Reliability Physics Symposium, 2013.

[Lib16] Liberty: Format Specification and Documentation. `https://www.opensourceliberty.org/`, 2016.

[Lie13] Jens Lienig. Electromigration and its impact on physical design in future technologies. In *Proceedings of the 2013 ACM international symposium on International symposium on physical design*, pages 33–40. ACM, 2013.

[LK91] JR Lloyd and J Kitchin. The electromigration failure distribution: The fine-line case. *Journal of applied physics*, 69(4):2117–2127, 1991.

[Llo07] JR Lloyd. Blacks law revisited: Nucleation and growth in electromigration failure. *Microelectronics Reliability*, 47(9):1468–1472, 2007.

[LMB+11] Baozhen Li, Paul McLaughlin, Jeanne Bickford, Peter Habitz, Dileep Netrabile, and Timothy Sullivan. Statistical evaluation of electromigration reliability at chip level. *Device and Materials Reliability, IEEE Transactions on*, 11(1):86–91, 2011.

[LVFA09] David D Ling, Chandu Visweswariah, Peter Feldmann, and Soroush Abbaspour. A moment-based effective characterization waveform for static timing analysis. In *Proceedings of the 46th Annual Design Automation Conference*, pages 19–24. ACM, 2009.

[mag14] Magma-Talus tool user manual. `http://www.synopsys.com`, 2014.

[McC89] S. P. McCormick. *Modeling and Simulation of VLSI Interconnections with Moments*. PhD thesis, MIT, Cambridge, MA, USA, 1989.

[MFT+08] Ning Mi, Jeffrey Fan, Sheldon X-D Tan, Yici Cai, and Xianlong Hong. Statistical analysis of on-chip power delivery networks considering lognormal leakage current variations with spatial correlation. *Circuits and Systems I: Regular Papers, IEEE Transactions on*, 55(7):2064–2075, 2008.

[MS13] Vivek Mishra and Sachin S Sapatnekar. The impact of electromigration in copper interconnects on power grid integrity. In *Proceedings of the 50th Annual Design Automation Conference*, page 88. ACM, 2013.

[MV04] Adamantios Mettas and Pantelis Vassiliou. Application of quantitative accelerated life models on load sharing redundancy. In *Reliability and Maintainability, 2004 Annual Symposium-RAMS*, pages 293–296. IEEE, 2004.

[Nas08] Sani R Nassif. Power grid analysis benchmarks. In *Proceedings of the 2008 Asia and South Pacific Design Automation Conference*, pages 376–381. IEEE Computer Society Press, 2008.

[OHG+04] Chanhee Oh, Haldun Haznedar, Martin Gall, Amir Grinshpon, Vladimir Zolotov, Pon Ku, and Rajendran Panda. A methodology for chip-level electromigration risk assessment and product qualification. In *Quality Electronic Design, 2004. Proceedings. 5th International Symposium on*, pages 232–237. IEEE, 2004.

[Ori10] R. L. D. Orio. *Electromigration Modeling and Simulation*. PhD thesis, Technischen Universitt Wien, 2010.

[Pan08] S. Pant. *Design and Analysis of Power Distribution Networks in VLSI Circuits*. PhD thesis, University of Michigan, 2008.

[PJK10] Young-Joon Park, P. Jain, and Srikanth Krishnan. New electromigration validation: Via-Node Vector Method. In *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pages 698–704, May 2010.

[PTP12] PTPX Power Estimation Tool. `http://www.synopsys.com`, 2012.

[Q1016] Q100 specification. `www.aecouncil.com/AECDocuments`, 2016.

[QPP94] Jessica Qian, Satyamurthy Pullela, and Lawrence Pillage. Modeling the 'effective capacitance' for the RC interconnect of CMOS gates. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 13(12):1526–1535, 1994.

[QRKG12] Haifeng Qian, Phillip J Restle, Joseph N Kozhaya, and Clifford L Gunion. Subtractive router for tree-driven-grid clocks. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 31(6):868–877, 2012.

[Raj08] A. Rajaram. *Synthesis of Variation Tolerant Clock Distribution Networks*. PhD thesis, University of Texas at Austin, Austin, Texas, USA, 2008.

[RED15] Ansys-Apache Redhawk tool user manual. `www.ansys.com`, 2015.

[RLC16] Sankar Ramachandran, Y.-K. Lai, and F.-T. Chan. Standard Cell Power and Signal EM Qualification Methodology. Designer's Track Presentation: Power Management in Advanced IC Design, Design Automation Conference, 2016.

[SABR12] Debjit Sinha, Soroush Abbaspour, Adil Bhanji, and Jeffrey M Ritzinger. Method of employing slew dependent pin capacitances to capture interconnect parasitics during timing abstraction of vlsi circuits, January 24 2012. US Patent 8,103,997.

[Sap04] Sachin Sapatnekar. *Timing*. Springer Science & Business Media, 2004.

[SHT15] V Sukharev, X Huang, and SX-D Tan. Electromigration induced stress evolution under alternate current and pulse current loads. *Journal of Applied Physics*, 118(3):034504, 2015.

[SKK14] Pramod Sharma, Madhur Kashyap, and Narayanan Kannan. Capacitive cell load estimation using electromigration analysis, September 23 2014. US Patent 8,843,873.

[Su02] H. Su. *Design and Optimization of Global Interconnect in High Speed VLSI Circuits*. PhD thesis, University of Minnesota, Minneapolis, MN, USA, 2002.

[syn16] Synopsys PrimeRail tool user manual. `http://www.synopsys.com`, 2016.

[TCH93] Jiang Tao, Nathan W Cheung, and Chenming Hu. Electromigration characteristics of copper interconnects. *IEEE Electron Device Letters*, 14(5):249–251, 1993.

[TMHM93] LM Ting, JS May, WR Hunter, and JW McPherson. AC electromigration characterization and modeling of multilayered interconnects. In *Reliability Physics Symposium, 1993. 31st Annual Proceedings., International*, pages 311–316. IEEE, 1993.

[TMS08] Aida Todri and Malgorzata Marek-Sadowska. A study of reliability issues in clock distribution networks. In *Computer Design, 2008. ICCD 2008. IEEE International Conference on*, pages 101–106. IEEE, 2008.

[TN13] Y. Taur and T. H. Ning. *Fundamentals of modern VLSI devices*. Cambridge University Press, Cambridge, UK, 2013.

[TSM16] TSMC Reference Signoff Flow. `http://www.tsmc.com/english/dedicatedFoundry/services/reference_flow.htm`, 2016.

[US14] S. Udupi and K. Sahni. Comprehensive Full-Chip Methodology To Verify Electro-Migration and Dynamic Voltage Drop On High Performance FPGA Designs In The 20nm Technology. Technical Presentation, DesignCon 2014, 2014.

[VLSP14] Kaushik Vaidyanathan, Lars Liebmann, Andrzej Strojwas, and Larry Pileggi. Sub-20 nm design technology co-optimization for standard cell logic. In *Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design*, pages 124–131. IEEE Press, 2014.

[VWG06] Sarma Vrudhula, Janet Meiling Wang, and Praveen Ghanta. Hermite polynomial based interconnect analysis in the presence of process variations. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 25(10):2001–2011, 2006.

[WM09] Cheng C Wang and Dejan Markovic. Delay estimation and sizing of cmos logic using logical effort with slope correction. *Circuits and Systems II: Express Briefs, IEEE Transactions on*, 56(8):634–638, 2009.

[YTW14] Chimin Yuan, Dave Tipple, and Jeff Warner. Optimizing standard cell design for quality. In *SPIE Advanced Lithography*, pages 90530O–90530O. International Society for Optics and Photonics, 2014.

[Yu14] T. Yu. *Power Grid Verification and Optimization*. PhD thesis, University of Illinois at Urbana-Champaign, 2014.

[YWC+06] C Yeh, G Wilke, Hongyu Chen, S Reddy, H Nguyen, Takashi Miyoshi, W Walker, and Rajeev Murgai. Clock distribution architectures: A comparative study. In *Proceedings of the 7th International Symposium on Quality Electronic Design*, pages 85–91. IEEE Computer Society, 2006.

[Zer11] Elhadj Zeraoulia. *Models and Applications of Chaos Theory in Modern Sciences*. CRC Press, 2011.

[ZKS08] Yong Zhan, Sanjay V Kumar, and Sachin S Sapatnekar. *Thermally-aware design*. Now Publishers Inc, 2008.

List of Publications

We attach a list of current publications that have been made in scope of this research effort:

1. **P. Jain**, S. S. Sapatnekar and J. Cortadella, A Retargetable and Accurate Methodology for Logic-IP-internal Electromigration Assessment, in Proc. of the 20th Asia and South Pacific Design Automation Conference(**ASPDAC**), 2015.

2. **P. Jain**, S. S. Sapatnekar and J. Cortadella, Stochastic and topologically aware Electromigration analysis for clock skew, selected for presentation at the 2015 IEEE International Reliability Physics Symposium (**IRPS**), April, 2015.

3. **P. Jain**, J. Cortadella and S. S. Sapatnekar, A Fast and Retargetable Framework for Logic-IP-internal Electromigration Assessment Comprehending Advanced Waveform Effects, IEEE Transactions on VLSI (**TVLSI**), 2016.

4. **P. Jain**, V. Mishra and S. S. Sapatnekar, Fast Stochastic Analysis of Electromigration in Power Distribution Networks, **under review for IEEE Transactions on VLSI (TVLSI)**.

Biography

Palkesh Jain graduated from the **Indian Institute of Technology Bombay**, in 2004 with Bachelors and Masters in Electrical Engineering. His Masters research, under the guidance of Prof. Juzer Vasi, was on Soft Errors and Total-Dose Radiation Reliability in CMOS devices. The project culminated into the first successful radiation-tolerant quarter micron tape-out from IIT Bombay at TSMC through MOSIS.

From campus, he joined the ASIC group at **Texas Instruments India**, where he defined and developed, several of the GHz enabling reliability methodologies from 130nm till 28nm technology node. In particular, his work on reliability assessment under statistical-variations and power-management has been used on generations of speed-binned designs.

Subsequently, he joined the Yield and Product Engineering team at **Qualcomm India** in 2014, where he is involved with system level power and thermal management methodologies. He holds 15 US patents (granted/ pending) and is pursuing his on-job doctoral studies at the Universitat Politècnica de Catalunya.

Palkesh is married to Kuhu, and in spare time, loves helping in her mosaic art-work. They are also kept busy by their four year old princess, Kopal!

This dissertation was typeset with LaTeX[1] by the author.

[1] LaTeXis a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TEX Program.


statistics
electromigration
reliability
interconnects
EVT
model
clock
novel
verification
lifetime
safe
system
Black's
MonteCarlo
redundancy
methodologies
application
demonstrated
assessment
breakdown
highly
failure
design
presented
developed
research
methodology
Signal
systematic
aspect
failures
component
impact
circuits
work
oxide
retargeting
algorithms
redundancies
component-level
integrated
extend
incorporate
leading
variation
thesis
interconnect
traditional
approach
arbitrary
shared
terms
production
analysis
proposed
compiler
incorporating
demonstrate
current
memories
process
performance
models
direct
based
done
part
level
aware
Finally