Novel techniques for timing analysis of VLSI circuits in advanced technology nodes

DATE Ph.D. Forum 2023

Dimitrios Garyfallou - ✉ digaryfa@e-ce.uth.gr
ECE Department, University of Thessaly, Greece

***Abstract*—Timing analysis is an essential and demanding verification method used during the initial design and iterative optimization of a Very Large Scale Integrated (VLSI) circuit, while it also constitutes the cornerstone of the final signoff that dictates whether the chip can be released to the semiconductor foundry for fabrication. Throughout the last few decades, the relentless push for high-performance and energy-efficient circuits has been met by aggressive technology scaling, which enabled the integration of a vast number of devices into the same die but introduced new challenges to timing analysis. To this end, this research presents several new techniques for accurate and efficient timing analysis of VLSI circuits in advanced technologies, which address different aspects of the problem, including gate and interconnect timing estimation [1]–[3], timing analysis under process variation [4], and Dynamic Timing Analysis (DTA) [5].**

I. Gate delay estimation

As process geometries shrink below 45 nm, gate delay estimation becomes even more challenging. Modern on-chip interconnects are highly resistive, while nonlinear transistor and Miller capacitances imply that signals no longer resemble smooth saturated ramps. As a result, this renders traditional models, such as the Non Linear Delay Model (NLDM), inadequate to capture the nonlinear driver waveforms, leading to significant errors in delay and slew computations. Over recent years, the semiconductor industry has adopted Current Source Models (CSMs) for accurate gate modeling. However, industrial CSMs are precharacterized into cell libraries assuming lumped capacitive loads, which poses significant challenges to the approximation of the interconnect admittance with an effective capacitance (C_{eff}) due to the significant resistive shielding effect. In fact, most related works are either computationally expensive or unable to approximate the driver output slew, which is essential in timing analysis since it also impacts interconnect delay and slew estimation. Furthermore, they require additional non-standard characterization and ignore the impact of Miller effect [6], [7].

Contributions. We present an iterative algorithm for fast and accurate gate delay estimation [2]. The proposed methodology accurately approximates the nonlinear driver output waveform and C_{eff} in multiple waveform regions while considering their interdependence. Therefore, it allows for variable analysis resolution exploiting an accuracy/runtime trade-off, enabling applicability to both optimization steps and signoff timing analysis. In contrast to prior works [6], [7], our approach is compatible with industrial CSMs (e.g., the Synopsys® CCS model) and considers the impact of Miller capacitance. Experimental evaluation on representative driver-interconnect stages [8] implemented in 7 nm Fin Field-Effect Transistor (FinFET) technology indicates that our method achieves 1.3% and 2.5% mean error against SPICE for gate delay and slew, respectively. In addition, it provides high efficiency, as it relies on closed-form formulas and converges in 2.3 iterations on average. Moreover, we investigate the impact of resistive shielding and Miller effect on gate delay and slew estimation, demonstrating that the proposed method achieves better accuracy than related schemes that assume a single C_{eff} value [6] or ignore the impact of Miller capacitance [7]. This work, along with our machine learning enhancements [3], won the first place award in 2020 and 2021 ACM TAU timing analysis contests [8].

II. Interconnect delay estimation

On the other hand, interconnect delay has become the main performance limiter in nanometer-scale designs, as it represents an increasingly dominant portion of path delay. Recent projections made by the International Roadmap for Devices and Systems (IRDS) indicate that clock frequency at nominal voltage is forecasted to mildly improve from 3.1 GHz in 2020 to 3.5 GHz in 2025, while it is predicted to be decreased to 2.9 GHz in 2034 [9]. This limited scaling is due to the increased interconnect length and routing density, which has led to increased parasitics. Although gate timing may be calculated by interpolating on precharacterized waveforms, interconnect analysis has remained a mystery. SPICE simulation offers golden accuracy but fails to meet the performance and memory requirements for full-chip analysis even when accelerated simulation methods are employed [10]), while closed-form timing metrics [11] may be quite inaccurate, especially for large RC networks with many branches, as they rely on invalid assumptions. In practice, Model Order Reduction (MOR) is employed to reduce large interconnect models and provide a good compromise between accuracy and efficiency in timing analysis, as we only need to compute the response at the output ports. However, MOR techniques [12][13] typically produce dense system matrices, which may render simulation impractical.

Contributions. We propose a sparsity-aware MOR methodology for efficient timing analysis of VLSI interconnects with many ports [1]. Contrary to well-established MOR techniques [12], our method produces sparse reduced-order models by applying key congruence transformations on the original interconnect model and then exploiting the correspondence between Laplacian matrices and graphs. Moreover, the generated models can be straightforwardly realized into equivalent compact RC networks and be used in several other steps of the design flow. Although our approach may be more suitable for timing signoff, it can also be incorporated in iterative optimization flows to improve the convergence rate by providing a fast and accurate estimation of the critical paths. Evaluation on complex interconnects of the ISPD 2012 benchmarks implemented in 45 nm technology show that a high sparsity ratio (over 97%) of the reduced matrices can be achieved, resulting in 4× speedup over [12] and 30× speedup over SPICE simulation of the original model. At the same time, a reasonable delay accuracy of 3% is maintained, while Elmore delay may deviate up to 288%.

III. Statistical timing analysis

In an ideal world, fabrication would be a predictable process and all chips would meet their timing specifications. However, manufacturing is getting more complex due to shrinking physical dimensions. Equipment imprecision and process limitations lead to extensive variations of transistor and interconnect parameters, which critically affect timing and may result in up to 30% variation in operating frequency, causing timing failures [14]. Conventionally, designers apply corner-based analysis [15] to consider the impact of process variation. However, this approach is too slow as the number of variability sources proliferates with process scaling or inaccurate due to the assumption that the worst-case delay resides at the corners of the design parameters. Another approach that has gained excessive research interest is Statistical Static Timing Analysis (SSTA), which is performed either by probability distribution propagation or by Monte Carlo (MC) simulation [16]. The former method is significantly faster but less accurate than the latter, as it relies on simplistic assumptions about the underlying timing models and distributions.

Contributions. We introduce a novel statistical methodology based on MC simulation and Extreme Value Theory (EVT) to estimate the worst-case timing of VLSI circuits under variations in gate and interconnect parameters [4]. In contrast to corner-based or traditional SSTA approaches [15], [16], our approach can be applied regardless of the underlying timing models or any assumption about the distribution of the examined timing parameters, such as Arrival Time (AT) or slack. Experimental results on ISCAS85/89 circuits implemented in 45 nm technology demonstrate that the estimated maximum AT on path endpoints can be within 5% of the actual value (with a 99% confidence level), at a cost of few (2000-3000) MC trials, providing a speedup of six orders of magnitude over exhaustive MC simulation. As a result, our method is very appealing for integration into any level of timing analysis abstraction (from transistor-level to gate-level).

IV. Dynamic timing analysis

To ensure reliable operation, designers adopt timing guardbands that force circuits to operate at a lower frequency, providing sufficient margins to mitigate errors induced by Process, Voltage, Temperature, and Aging (PVTA) variations [17]. However, such margins are overly pessimistic since they are estimated according to the most critical paths identified through multi-corner Static Timing Analysis (STA), ignoring the dynamically activated critical paths existing due to workload variability. Recently, it has been demonstrated that 99% of statically estimated critical paths are triggered by less than 10% of all possible input vectors, while there is a very low possibility to experience the worst-case input conditions [17]. These findings have turned the attention of many studies to the exploitation of Dynamic Timing Slack (DTS) that may exist within any path depending on the dynamically changing processed data [18]–[20]. Such studies may have revealed extensive DTS but rely on Graph-Based DTA (GB-DTA) [20], which inherently makes worst-case assumptions and still ignores some data-dependent timing properties. This may cause significant DTS underestimation, leading to unexploited frequency scaling margins and incorrect error estimation.

Contributions. We present an accurate framework that reveals available DTS underestimated by prior works ISQED20. Contrary to GB-DTA that relies on worst-case assumptions, our approach exploits gate-level Event-Driven DTA (ED-DTA) to consider the data-dependent timing properties of activated paths. Evaluation on ISCAS85/89 and AxBench designs in 45 nm technology show that ED-DTA achieves an average of 2.35% and up to 194.51% DTS improvement over a conventional GB-DTA method based on delay-annotated gate-level simulation (i.e., ModelSim). Considering only the critical activated paths, the average DTS improvement is increased to 11.2%. Compared to existing frequency scaling schemes [18], ED-DTA enables further clock frequency increase by up to 10.42%. Finally, it reveals that timing errors may be up to 2.94× less than those estimated by existing works [19].

References

[1] D. Garyfallou, C. Antoniadis, N. Evmorfopoulos, and G. Stamoulis, "A Sparsity-Aware MOR Methodology for Fast and Accurate Timing Analysis of VLSI Interconnects," in *International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 89–92, 2019.

[2] D. Garyfallou *et al.*, "Gate Delay Estimation With Library Compatible Current Source Models and Effective Capacitance," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 5, pp. 962–972, 2021. [GitHub: UTH-Timer].

[3] D. Garyfallou, A. Vagenas, C. Antoniadis, Y. Massoud, and G. Stamoulis, "Leveraging machine learning for gate-level timing estimation using current source models and effective capacitance," in *Great Lakes Symposium on VLSI (GLSVLSI)*, p. 77–83, 2022. [GitHub: UTH-Timer].

[4] C. Antoniadis, D. Garyfallou, N. Evmorfopoulos, and G. Stamoulis, "EVT-based worst case delay estimation under process variation," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1333–1338, 2018.

[5] D. Garyfallou, I. Tsiokanos, N. Evmorfopoulos, G. Stamoulis, and G. Karakonstantis, "Accurate Estimation of Dynamic Timing Slacks using Event-Driven Simulation," in *International Symposium on Quality Electronic Design (ISQED)*, pp. 225–230, 2020.

[6] R. Puri, D. S. Kung, and A. D. Drumm, "Fast and accurate wire delay estimation for physical synthesis of large ASICs," in *the 12th Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 30–36, 2002.

[7] P. Feldmann *et al.*, "Driver waveform computation for timing analysis with multiple voltage threshold driver models," in *the 45th Design Automation Conference (DAC)*, pp. 425–428 , 2008.

[8] ACM TAU Timing Analysis Contest. Accessed: Jan. 1, 2023. [Online]. Available: https://www.tauworkshop.com/

[9] IEEE IRDS - 2020 Update. Accessed: Jan. 1, 2023. [Online]. Available: https://irds.ieee.org/editions/2020/

[10] D. Garyfallou, N. Evmorfopoulos, and G. Stamoulis, "A Combinatorial Multigrid Preconditioned Iterative Method for Large Scale Circuit Simulation on GPUs," in *International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, pp. 209–212, 2018.

[11] K. Agarwal, D. Sylvester, and D. T. Blaauw, "Simple metrics for slew rate of RC circuits based on two circuit moments," in *the 40th Design Automation Conference (DAC)*, pp. 950–953, 2003.

[12] K. Kerns and A. Yang, "Stable and efficient reduction of large, multiport RC networks by pole analysis via congruence transformations," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 16, no. 7, pp. 734–744, 1997.

[13] C. Chatzigeorgiou, D. Garyfallou, G. Floros, N. Evmorfopoulos, and G. Stamoulis, "Exploiting Extended Krylov Subspace for the Reduction of Regular and Singular Circuit Models," in *the 26th Asia South Pacific Design Automation Conference (ASP-DAC)*, pp. 773–778, 2021.

[14] P. Gupta *et al*, "Underdesigned and opportunistic computing in presence of hardware variability," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 32, no. 1, pp. 8–23, 2013.

[15] H. Zhang, T. Chen, M. Y. Ting, and X. Li, "Efficient design-specific worst-case corner extraction for integrated circuits," in *the 40th Design Automation Conference (DAC)*, pp. 386–389, 2009.

[16] D. Blaauw, K. Chopra, A. Srivastava, and L. Scheffer, "Statistical Timing Analysis: From Basic Principles to State of the Art," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 27, no. 4, pp. 589–607, 2008.

[17] R. G. Dreslinski, M. Wieckowski, D. T. Blaauw, D. Sylvester, and T. N. Mudge, "Near-threshold computing: Reclaiming moore's law through energy efficient integrated circuits," *Proceedings of the IEEE*, vol. 98, no. 2, pp. 253–266, 2010.

[18] J. Constantin, L. Wang, G. Karakonstantis, A. Chattopadhyay, and A. Burg, "Exploiting dynamic timing margins in microprocessors for frequency-over-scaling with instruction-based clock adjustment," in *the Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 381–386, 2015.

[19] X. Jiao *et al.*, "CLIM: A cross-level workload-aware timing error prediction model for functional units," *IEEE Trans. on Computers*, vol. 67, no. 6, pp. 771–783, 2018.

[20] H. Cherupalli, R. Kumar, and J. Sartori, "Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems," in *the 43rd Annual International Symposium on Computer Architecture (ISCA)*, pp. 671–681, 2016.