High-Performance, High-Capacity Memory and Block Characterization Enables Accurate Full-chip Timing Variation Signoff

Ken Tseng, Danny Li, Ting-Hau Hsiao, Kevin Chou
XinDA Design Automation

Introduction

Process variation library models in the Liberty Variation Format (LVF) have become commonplace in timing signoff for standard cells, yet the embedded memories and mixed-signal blocks that comprise most of the chip area still employ library modeling methodologies from several technology generations ago. This paper explores the underlying reason for this modeling gap, highlights the inefficiencies in existing methodologies, and demonstrates how a new characterization system can overcome these hurdles and bring forth accurate timing variation signoff to the entire System-on-Chip (SoC).

SoC Library Modeling Trends

Starting from 16nm FinFETs, process variation library models for standard cells have become commonplace in timing signoff. This has brought tremendous savings in timing margin, especially at ultra-low supply voltages, and also uncovered timing violations that could lead to yield loss or complete chip failure. However, modern chip designs are very SRAM-heavy with a rule-of-thumb ratio of 70/30 SRAM to logic [1], while the advance in process variation library models have so far been limited to logic cells – the ANDs, ORs, and Flip-flops. That leaves the majority of an SoC's real estate vulnerable to variation induced yield loss and simultaneously suffer from over margin, which leads to delayed timing closure and increased power consumption.

Figure 1 below shows the increasing trend in the area percentage contributed by memories and other IP blocks in an SoC, reaching 79% in 2021.

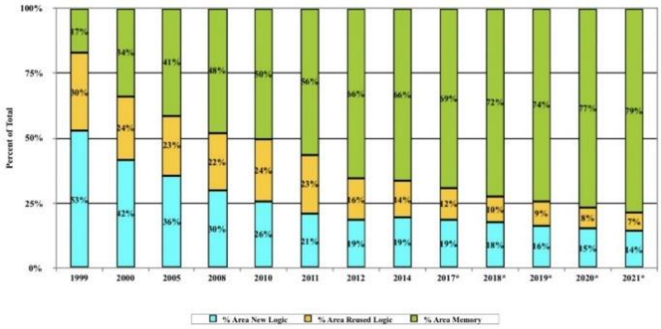


Figure 1: Silicon area consumed by logic (blue and yellow) vs. memory (green) [1]

In addition to the increasing SoC area occupied by memory and other IP blocks, the number of these macro instances are also growing rapidly. Figure 2 below shows the unmistakable trend in the total number of macro instances in an SoC, while more recent data [2] indicates this trend is accelerating. Yet the timing variation modeling has not kept pace with this trend, leaving a growing gap in SoC modeling accuracy.

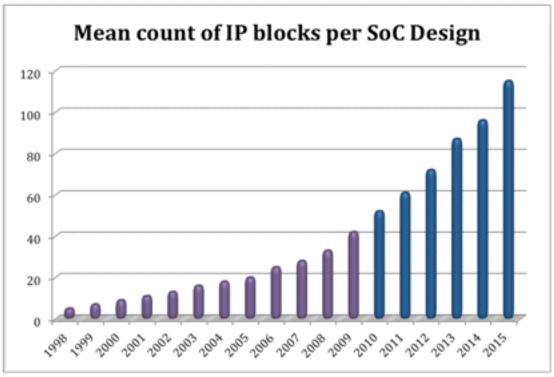


Figure 2: Mean count of IP blocks per SoC Design [3]

The root cause of this modeling gap lies in the availability of efficient EDA tools that can scale in variation modeling capacity from the small standard cells – ANDs, ORs, Flip-flops – with 10^1 to at most 10^2 transistors to the large, embedded memories exceeding 10^6 transistors.

The rest of this paper focuses on memory characterization due to its prevalence in an SoC, while the description and conclusion apply equally to non-memory IP blocks.

Memory Characterization Bottleneck

Current approaches to memory characterization belong to two major categories:

- Run FastSPICE simulation on the entire post-layout extracted netlist
- Critical path-based circuit partition and simulation

In either case, test vectors to drive the simulation and measurements must be provided by the user. Often these vectors must first establish an internal state, such as a memory cell logic value, followed by reading that written value from the primary output pins where path delay values can be measured. This results in a long multi-cycle transient simulation on a large circuit, which requires long running time of the characterization tool.

On top of that, library characterization is a repeated process where different "variations" to the circuit simulation must be measured. For example, multiple slew rates on the primary input pins and multiple capacitive loadings on the primary output pins must be applied, simulated and measured. This results in a "big loop" over a long transient simulation of a very large circuit, as shown below.

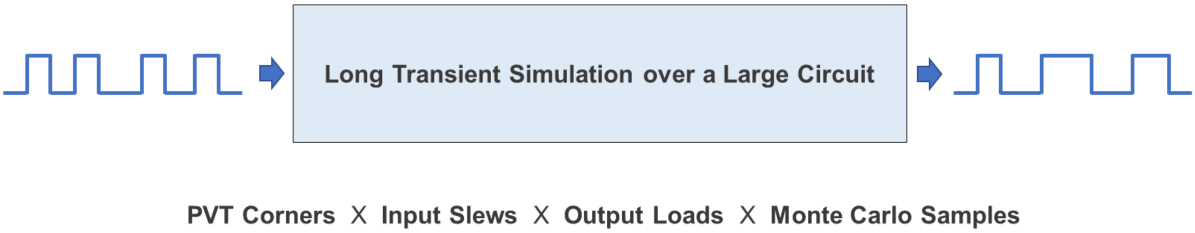

Long Transient Simulation over a Large Circuit
PVT Corners X Input Slews X Output Loads X Monte Carlo Samples

Figure 3: Traditional memory characterization with a "big loop" over a large simulation

In addition, each PVT (Process, Voltage, Temperature) corner must be separately simulated. The total turnaround time for a complete set of memory libraries over hundreds of PVT corners is very large even with the availability of large server farm with thousands of CPUs. Recent advance in corner interpolation techniques, if employed, only reduces the number of simulated corners by a factor of around 2.

It is with this background that accurate process variation models have not been practically considered for memories, because it would require running some flavor of Monte Carlo on top of the aforementioned huge set of large simulations. It would further increase characterization turnaround time by several orders of magnitude!

A New Approach

This is the time when one should step back and re-examine the problem, to discover the root cause of the inefficiencies in the current methodology. We have been utilizing this core engine (FastSPICE simulation of a large circuit) for each additional variation requirement (slews, loads, PVT corners, process variation, etc.), without considering whether each variation requirement calls for a complete re-run of the entire core engine. Specifically, there are two main types of variations that occur in library characterization – global variation and local variation. Global variation affects the entire design, while local variation affects a small portion of the design. In this paper, the terms "global" and "local" variations are not limited to process variation but apply to any circuit or environmental variations to a circuit simulation. These variations are classified below:

- Global variations – PVT corners
- Local variations – Input slews, output loads, random process variations

An optimal characterization approach would apply global simulations to global variations, and local simulations to local variations.

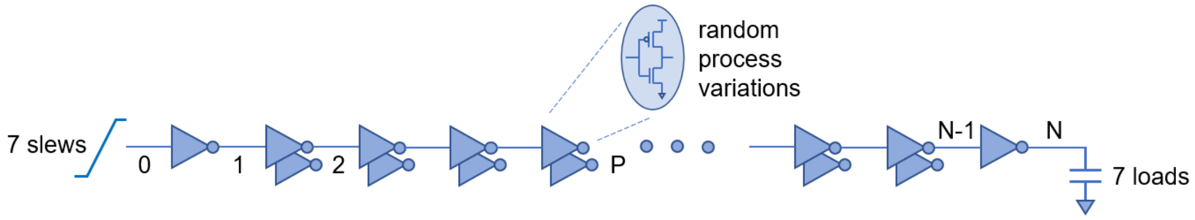
Local simulations for local variations bring tremendous runtime savings because the results affect only several neighboring circuit stages from the variation source, which is a miniscule fraction of the entire design size. Only the affected circuit stages need to be re-simulated to capture voltage waveform changes and timing shifts due to these variations.

The 3 types of local variations – slews, loads, random process – and their local circuit impact are illustrated with the example circuit in Figure 4. It is a 7nm inverter chain with fanout of 2 at each stage, and each wire consists of 3 large RC segments to model a typical wire load. It is a simple circuit, yet representative of the general circuit behavior of local variation effects becoming muted after several circuit stages.

Input Slew Variation

Library lookup tables often consist of 7 input slews covering the entire range of input waveform possibilities. Traditional characterization approaches simulate the entire circuit 7 separate times to capture their unique timing behavior. However, the wide range of input slews converge after several circuit stages resulting in the same waveform shape. A characterization system that automatically recognizes this and performs local re-simulation only when necessary will achieve close to 7 times reduction in simulation effort without compromising accuracy.

In this example, 7 input slews ranging from 1ps to 144ps are applied to the circuit at Stage 0. As shown in the left table of Figure 4 below, the waveform shapes converge after 4 circuit stages with 7.4ps slew in all cases.



Input slew variation converges after 4 stages

Stage	Stage Slew (ps)	
	Min In Slew	Max In Slew
0	1	144
1	4.2	29.2
2	6.9	10.4
3	6.7	7.0
4	**7.4**	**7.4**
5	**7.0**	**7.0**

Random process variation converges after 2 stages

Stage	Stage Slew (ps)	
	Nominal	Variation
P – 1	5.0	5.0
P	6.4	7.8
P + 1	7.3	7.1
P + 2	**6.9**	**6.9**
P + 3	**7.6**	**7.6**

Output load variation only affects last stage

Stage	Stage Slew (ps)	
	Min Out Load	Max Out Load
N – 2	**7.4**	**7.4**
N – 1	**7.0**	**7.0**
N	6.1	655.9

Spice simulation of 7nm INV chain with FO2, each wire has 3 large RC segments.

Figure 4: 7nm circuit path demonstrating variation locality

Output Load Variation

In addition to 7 input slews, library lookup tables often consist of 7 output loads covering the entire range of output loading possibilities. Traditional characterization approaches simulate the entire circuit 7 separate times to capture their unique timing behavior. Coupled with 7 input slew variations, 7 x 7 = 49 total simulation combinations are required.

However, similar to input slew variation, the circuit affected by output load variation is very localized. In fact, as shown in the right table of Figure 4 above where the output load value ranges from 1e-18 to 1e-13 Farads and the corresponding output slew ranges from 6.1ps to 655.9ps, the previous stage (Stage N – 1) slews are identical and only the last stage (Stage N) must be re-simulated to capture the output load timing effect without loss of accuracy.

Random Process Variation

Random process variation is a challenge for existing memory characterization methodologies. Monte Carlo simulations must be performed on the entire circuit or its critical paths, resulting in orders of magnitude runtime penalty. Exacerbating the problem is low voltage or near-threshold voltage design where even more samples are required, even with so-called “fast Monte Carlo” techniques, to accurately capture the “tails” of the heavily skewed timing distribution. Furthermore, automotive and other high reliability applications require variation analysis at a high sigma requiring even more samples and more runtime.

This is a huge additional burden for the current set of characterization methodologies. As a result, designers are forced to bypass accurate models and instead apply a simple timing margin that could be inaccurate and tends to over-estimate most circuit paths yet under-estimate other circuit paths. Obviously, a new methodology is required to bring memory and block characterization into the same accuracy fold as standard cells.

Similar to input slew variation, random parameter variation can be simulated at the local “gate” level with the resulting variation in voltage waveform propagated to the fanout stages. And similar to input slew variation, the circuit impact converges quickly. In the middle table of Figure 4 above where Stage P contains the source of the random process variation, the variation impact converges with 6.9ps slew at Stage P + 2, or 2 circuit stages after the variation source.

Local simulation of process variation not only provides significant speedup, but also enables higher modeling accuracy than global simulation approaches. Since local simulation is very efficient, the library characterizer can afford many more simulations to fully explore the variation space of a given logic gate. This is especially important for near-threshold voltage library characterization where the variation effect on timing is highly non-linear. Highly accurate gate-level variation models can be built for every gate in all critical paths in the design, at a fraction of the simulation cost of traditional global simulation and sampling methods. Once the gate-level variation models are built, the subsequent random samples on every transistor on the critical paths are evaluated and propagated without requiring additional simulation. Hence, the sampling evaluation is extremely efficient allowing for a large number of samples on each critical path to further increase variation modeling accuracy. As a result, this method not only excels at the regular 3-sigma model requirement, but is also especially suitable for high sigma library characterization that requires a huge number of samples on each library arc, slew and load.

In addition to the above improvements, there is yet another benefit to the local simulation approach. In ultra-low voltage designs where the timing distribution is very skewed, fast Monte Carlo approaches often have to perform numerous re-simulations to improve accuracy at the model predicted timing distribution tails. In a local simulation approach, only the circuit stages with extreme parameter variations must be re-simulated, resulting in another large factor of runtime savings.

To summarize, a characterization system that performs local variation simulation and builds accurate variation model at the gate level, propagates variation waveform to the fanout stages, performs Monte Carlo sampling without simulation and, if necessary, performs tail re-simulation selectively at the gate level will be able to characterize accurate timing variation libraries at near-threshold supply voltage at a high sigma and with a short turnaround time.

Introducing Charlie

Charlie is a new memory and block characterization tool that produces nominal and process variation (LVF) libraries at a fraction of the runtime as existing tools that produces only nominal libraries. It achieves this by automatically applying local simulations to local variation sources, and automatically propagating the variation effects throughout the circuit.

Charlie automatically partitions a post-layout netlist into its smallest electrically cohesive building blocks (often similar to a logic "gate"), automatically generates exhaustive local simulation vectors, automatically propagates waveforms and variations through the design, automatically traces critical timing paths including the memory array, and produces all modern library data including LVF, current source timing and current source noise models. It also generates detail timing reports for critical paths and also variation reports that pinpoint the most significant contributors to timing variation.

Figure 5 below shows Charlie performance for 3 types of memories – single-port, dual-port and ROM – of various sizes. Charlie LVF characterization overhead is less than double of Charlie nominal characterization, which in turn is an order of magnitude faster than existing solutions. Therefore, it takes less time for Charlie to produce an LVF library than for existing solutions to produce a nominal library.

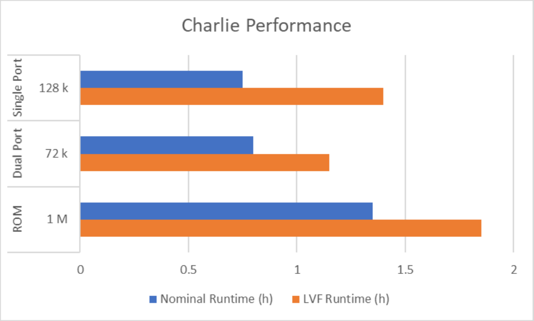


Figure 5: Charlie runtime for nominal (blue) and LVF (orange) characterizations

Charlie achieves huge performance gain by automatically running global simulations for global variations and local simulations for local variations. Table 1 below compares Charlie's usage of global and local simulations versus existing characterization solutions.

Variations	Competitor Simulation	Charlie Simulation
PVT Corner	Global	Global
Input Slew	Global	Local
Output Load	Global	Local
Random Process	Global	Local
High Sigma Tail Re-sim	Global	Local

Table 1: Global versus local simulation usage in memory characterization

In addition to the performance gains, Charlie's accuracy compares favorably to Monte Carlo simulation. Charlie LVF accuracy at low supply voltage is shown below, with the reference established by Monte Carlo simulation of an SRAM clock to output delay arc. The accuracy metrics are nominal, mean and standard deviation (LVF moments) and early and late sigmas (LVF sigmas). Overall accuracy is well within 1% while delivering 500 times speedup.

Tool	Nominal	Moment LVF		Sigma LVF		Runtime
		Mean Shift	Std Dev	Early	Late	
Monte Carlo	5110.0	38.1	165.6	134.6	230.3	53 hours
Charlie	5145.4	39.1	196.3	135.8	250.7	6 minutes
Accuracy	0.6%	0.0%	0.6%	0.0%	0.4%	**500x speedup**

Table 2: Charlie LVF runtime and accuracy versus Monte Carlo

In addition to the local simulation efficiency benefits, simulating at the gate level also eliminates repeated simulation of the same circuitry that is common between different critical paths. A common example is the clock tree which is present in many critical paths and are repeatedly simulated by existing characterization solutions leading to wasted simulation effort. Charlie maintains a gate level simulation database from local simulations of the common circuitry so that the simulation results are shared between all critical paths. While on a per-path basis Charlie provides 500x speedup over Monte

Carlo, on a complete library with many paths the speedup easily exceeds 3 orders of magnitude due to simulation sharing.

Summary and Benefits

Charlie extends the LVF characterization capacity envelope from small standard cells to large memories and mixed-signal blocks. It does this at a fraction of the turnaround time of existing memory characterization tools. Charlie enables accurate timing variation models for the key missing pieces of an SoC, which in turn allows accurate full-chip timing variation signoff that reduces over-margin and simultaneously covers under-margin especially in ultra-low voltage domains and high sigma timing signoff.

The modern SoC comprises hundreds of IP blocks with two-thirds of the chip area occupied by memories. Your SRAMs, ROMs and mixed-signal blocks deserve the same timing variation accuracy as your standard cells.

References

[1] Trend of memory and logic area for SoC (Source: Semico Research)

[2] The Complexity of Block-Level Placement @ 56th DAC, June 2019, by Tom Dillinger https://semiwiki.com/eda/cadence/259749-the-complexity-of-block-level-placement-56thdac/

[3] Past, present and future count of IP blocks per device (Source: Semico research)

XinDA

Re-imagine
Transistor Level Circuit Analysis

XinDA Design Automation is an Electronic Design Automation company headquartered in Hong Kong. XinDA re-imagines transistor level circuit analysis to create innovative software solutions that enable orders of magnitude productivity gains in memory and mixed-signal block library characterization.