

RAMP: RTL-Level Emulation with Thousand-Core-Scale Parallelism

Weigang Feng^{*1}, Yijia Zhang^{†1}, Zekun Wang², Zhengyang Wang³, Yi Wang¹, Peijun Ma², Ningyi Xu^{†1}

¹Shanghai Jiao Tong University, ²Xidian University, ³University of Toronto

[†]Co-corresponding authors: zhangyijia@sjtu.edu.cn, xuningyi@sjtu.edu.cn

Abstract—With the continuous increase in transistor counts on a single chip, the complexity of RTL verification has grown exponentially, and completing a full simulation flow often takes several months. In industrial practice, RTL simulation is typically divided into two stages: functional debugging and system verification. Functional debugging emphasizes fast compilation and is usually performed on multi-core CPUs, while system verification demands extremely high simulation speed and often relies on FPGA acceleration. However, the limited performance of CPU-based simulation has become a major bottleneck that restricts overall design productivity.

To address this challenge, we propose RAMP, a scalable multi-core RTL simulation platform that balances fast compilation with high-throughput execution. RAMP leverages a specialized architecture and compilation strategy to accelerate both combinational logic evaluation and sequential logic synchronization. For combinational logic, it adopts a balanced DAG partitioning method together with highly efficient Boolean computation cores; for sequential logic, it employs a low-latency on-chip network (NoC) to achieve efficient state synchronization across cores. Experimental results demonstrate that RAMP achieves up to 12.9× speedup over state-of-the-art multi-core simulators.

Index Terms—RTL simulation, Accelerator, Graph Partition

I. INTRODUCTION

Advances in semiconductor technology and EDA tools have enabled the integration of tens of billions of transistors on a single chip [1], [2]. This rapid growth, driven by the rise of AI, has significantly increased verification complexity. For example, NVIDIA’s A100 GPU contains 54 billion transistors [3], and verifying such designs on CPUs can take months or even years. RTL simulation—a critical phase in the verification flow—now accounts for over 24% of total development time. Meanwhile, first-pass silicon success rates have dropped by 18% over the past 12 years, with only 14% of designs succeeding on the first try in 2024 [4]. These trends highlight slow RTL simulation as a growing bottleneck in modern chip development.

For chip designers, choosing hardware for RTL simulation involves balancing compilation and simulation speed. Compilation translates RTL code into an intermediate form executable by simulation tools, while simulation verifies functionality by executing this code. Single-chip FPGAs currently offer the fastest RTL simulation by leveraging look-up tables (LUTs) to perform Boolean operations. However, FPGA compilation

remains time-intensive because placement and routing must satisfy strict timing and resource constraints. Partial reconfiguration shortens this process only when changes are tiny and floorplans stay intact. Moreover, Heavy waveform probing further consumes LUTs and BRAM, reducing max frequency and rendering full-signal visibility impractical. Once a design outgrows a single device, partitioning across multiple FPGAs introduces inter-board links that often dominate timing, sharply cutting simulation speed.

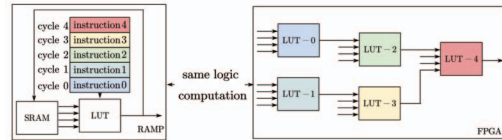


Fig. 1: RAMP accelerator completes the same logical operations as FPGA in 5 cycles, replacing routing with memory access, which gives RAMP advantages in clock frequency.

In contrast, CPU-based RTL simulation offers fast compilation but suffers from limited simulation speed on large-scale designs due to constrained computational throughput. This stems primarily from RTL simulation’s reliance on integer operations, which underutilize the CPU’s floating-point units [5], leading to poor resource efficiency. Additionally, inefficient inter-core communication impedes the frequent signal exchanges required during RTL simulation, often degrading performance when scaling to more cores [6]. Although RepCut [7] eliminates inter-core communication, its performance remains limited by cache coherence overhead, necessitating large per-core workloads to maintain efficiency.

GPU-based RTL simulation excels in multi-stimulus scenarios, where thousands of uniform test vectors can be executed in parallel through batch processing. Tools like RTLFlow [8] report up to 40× speedups over CPU-based simulators with 64K parallel runs. However, this comes at the cost of high memory consumption, often exceeding GPU capacity and limiting the number of simulation cycles. In contrast, real-world workloads are typically single-stimulus and long-cycle—for example, simulating a CPU booting an OS may require 700 million cycles [9]. In such cases, GPU simulators can be up to 10× slower than CPU-based tools like Verilator [10].

Beyond multi-core architecture and FPGAs, commercial emulators like Palladium [11] offer high performance but are costly, closed-source. In this work, we propose RAMP, a novel

This research was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0122403, and was also supported by the National Key Research and Development Program of China under Grant 2023YFB4405102.

accelerator for full-cycle RTL simulation. As illustrated in Figure 1, the key advantage of RAMP lies in its ability to achieve higher simulation speed than CPU-based approaches, while also providing faster compilation than FPGAs.

Our main contributions are summarized as follows:

- 1) We propose a netlist-level graph partitioning algorithm to evenly distribute RTL simulation tasks across multiple cores, ensuring balanced workload and improved parallelism.
- 2) We design LUT-based compute cores with efficient resource allocation and memory access, significantly improving simulation performance.
- 3) We implement a NoC-based interconnect for fast and efficient synchronization among processing cores.

II. BACKGROUND AND MOTIVATION

RTL simulation has two modes: timing-accurate and cycle-accurate. Timing-accurate mode timestamps every signal change to precisely model gate delays but is slow and mainly used for timing closure. Cycle-accurate mode samples signals only at clock edges and divides into event-driven and full-cycle methods. Event-driven simulation updates affected components based on signal activity, saving effort in sparse designs but incurring queue-management overhead. In contrast, full-cycle simulation recomputes the entire circuit each cycle, creating redundancy but enabling highly parallel execution. This paper focuses on full-cycle simulation for its simplicity and superior parallel performance, consisting of combinational logic evaluation and sequential logic synchronization phases per cycle.

A. Fiber-based partition

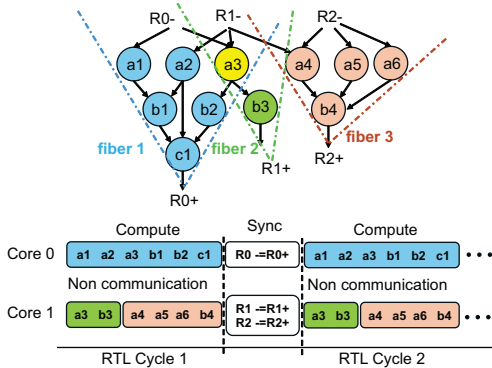


Fig. 2: A case of fiber partition. Each output node forms a fiber; fibers 2 and 3 are merged for load balancing. Shared node a3 belongs to both fiber 1 and 2, leading to redundant computation on Core 0 and Core 1.

The compilation of RTL simulation on multi-core processors often relies on graph partitioning to distribute tasks across cores. Recently, fiber-based partitioning has been proposed to accelerate this process [5], [7], [12], offering a key advantage: eliminating inter-core communication during computation. As shown in Figure 2, each core executes independently, with

synchronization needed only after all computations complete. To ensure load balancing while minimizing inter-core communication, the fiber-based partitioning algorithm proceeds in three steps: **1)** Convert the netlist into a DAG based on DFF (flip-flop) state transitions; **2)** Identify fibers by tracing backward from sink nodes; **3)** Merge fibers and assign them to individual cores for balanced computation. The graph partition algorithm in Fiber eliminates time-consuming inter-core communication, further accelerating parallel execution.

TABLE I: Compilation and simulation performance across different systems

System	Compile time	Sim. speed	Time to simulate		
			1M cycles	1B cycles	1T cycles
CPU sim.	2 mins	11 KHz	3.9 mins	1.1 days	3 years
RAMP	8 mins	2.9 MHz	8.3 mins	14 mins	4 days
FPGA $\times 2$	13 hours	1.4 MHz	13.2 hours	13.4 hours	8.7 days

B. Multi-core versus FPGA

Previous work has shown that multi-core simulators offer faster compilation, whereas FPGAs deliver higher simulation speed. Following the experimental setup in [13], we benchmark RAMP on Chronos [14], a 128-PE graph accelerator.

The CPU baseline uses Verilator, while the FPGA baseline employs two Alveo U250 boards interconnected via a 200 Gbps link. However, inter-board routing significantly extends the critical path, limiting the FPGA clock frequency to just a few MHz—even though single 7 nm parts typically achieve 100–600 MHz [15].

Thanks to its partition-aware scheduler, RAMP avoids this bottleneck and outperforms the FPGA in simulation speed. Moreover, it compiles Chronos in just 8 minutes—requiring only graph partitioning instead of placement and routing—as summarized in Table I. Nevertheless, compilation remains longer than on the CPU, since the RTL must still be converted into a LUT-based netlist. We use Yosys [16] for this step, which typically accounts for 10%–20% of total compilation time. Furthermore, while the CPU schedules few cores, RAMP must coordinate parallel scheduling across thousands of cores.

III. DESIGN OVERVIEW

As shown in Figure 3, we propose RAMP, an accelerator for RTL simulation. RAMP is a LUT-based multi-core architecture. Through the co-design of the compiler, computing cores, and NoC, we achieve acceleration in both combinational logic computation and sequential logic synchronization.

A. Fast combinational logic computing

Compiler Design: Our frontend compilation strategy resembles that of FPGAs, using a LUT-based netlist as an intermediate representation to emulate arbitrary combinational logic. The netlist comprises three main components: LUTs, DFFs, and Block RAM (BRAM), mapped respectively onto our compute array, SRAM, and hardware BRAM. Unlike

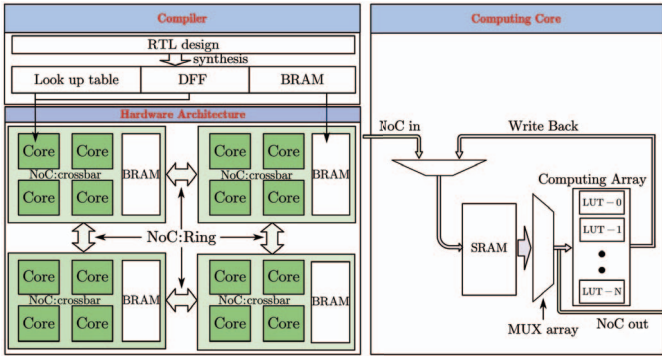


Fig. 3: The overview design of RAMP, which is a LUT-based multi-core architecture. Our compiler’s design philosophy is time-division multiplexing of LUT units.

traditional FPGA backends that map LUT nodes one-to-one to computation units, we apply time-division multiplexing, assigning multiple LUT nodes to a single unit across several cycles. Nevertheless, RAMP achieves higher simulation speeds than FPGA-based methods due to its increased simulation frequency. A key challenge in our compilation approach—DAG partitioning—is addressed in Section IV.

Computing Core Design: Each RAMP computing core consists of LUTs array and SRAM. Unlike FPGAs, which rely on interconnect wires, RAMP uses a memory-access-based communication model where LUTs exchange data through shared SRAM. This design eliminates complex wiring and long critical paths, allowing for higher operating frequencies. To hide memory access latency and improve simulation speed, we implement a compute-memory overlap strategy within each core, detailed in Section V.

B. Fast sequential logic synchronization

NoC Design: After each RAMP core simulates one RTL cycle, synchronization is performed via a hierarchical NoC. Cores are grouped into clusters with low-latency crossbar interconnects for intra-cluster communication, while inter-cluster communication uses a ring topology. This two-level design suits system needs: most synchronization occurs within clusters, and the ring’s low-latency links suffice for lower-throughput inter-cluster traffic. Details are in Section VI.

IV. BALANCED DAG PARTITION

A. Motivation

RTL DAGs effectively capture the logical flow of simulations and naturally contain many independent sub-streams, making them well-suited for parallel acceleration [17]. Prior approaches, like Verilator Multi-thread [10] and Sphinx [18], partition the DAG into subgraphs for multi-core execution. However, they often overlook data dependencies, resulting in limited speedup and poor core utilization. To address this, we introduce a fiber-based DAG partitioning method that improves parallel efficiency across cores.

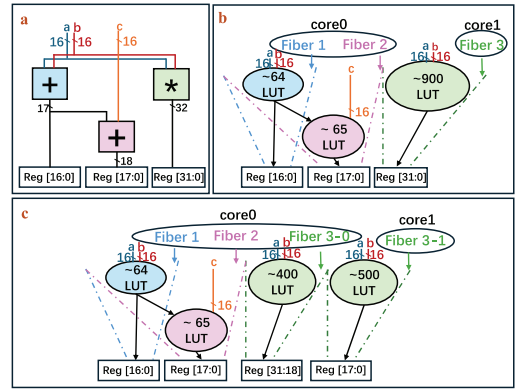


Fig. 4: Challenges in DAG partition: a) A 16-bit multiply-add circuit; b) Fiber 3 contains too many LUTs, causing imbalance when mapped across two cores; c) Splitting fiber 3 alleviates the load imbalance.

B. Challenge and Insight

Partitioning DAGs into fibers for parallel acceleration presents two primary challenges:

- 1) **Load imbalance between cores:** To reduce inter-core communication, prior works (e.g., RepCut [7], Manti-core [12], Parendi [5]) partition DAGs into independent fibers. We follow this approach but schedule at the LUT node level rather than arithmetic operations. As shown in Figure 4, traditional DAGs treat operations like addition and multiplication equally, whereas in our framework, they differ greatly in LUT counts—often by an order of magnitude. Naive assignments (e.g., merging fiber1 and fiber2 in core0 and assigning fiber3 to core1) can cause significant load imbalance.
- 2) **Redundant computation:** Fibers naturally form tree-like structures, often leading to overlapping nodes between adjacent fibers (Figure 4(b)). If fiber1 and fiber2 are mapped to different cores, up to 64 LUT nodes may be redundantly computed (Figure 4(c)), degrading simulation performance.

We make two key observations: (1) Fiber granularity is often too coarse and can be refined through further partitioning and merging to improve load balance; (2) Load imbalance and redundant computation are not strictly correlated, requiring a multi-objective optimization. Based on these insights, we propose an efficient fiber-partitioning algorithm that jointly optimizes load balance and redundancy.

C. Approach

Our method for minimizing computational redundancy while maintaining load balance across cores consists of two steps:

- 1) **Determining the required number of cores:** During synthesis, we compute the total number of LUT nodes and divide it by the target core count to derive the ideal per-core load, denoted as N_{balance} . Fibers exceeding this threshold are split by incrementally selecting register vector bits until a sub-fiber exceeds N_{balance} ; the previous

bit is then chosen as the split point. For example, in Figure 4(c), fiber3 is divided into two sub-fibers with 400 and 500 LUTs, respectively, achieving balanced core utilization.

- 2) **Fiber merging heuristic:** To optimize fiber assignments for both load balance and minimal redundancy, we use a multi-start hill climbing algorithm [19]. Starting from a random assignment, we define a cost function that considers both imbalance and redundancy. Fibers are iteratively moved between groups: beneficial moves are kept, others reverted. The process continues until no further improvement or a maximum iteration count is reached. Multiple restarts help avoid local optima and yield a near-optimal fiber assignment.

Let $\mathcal{H} = \{H_1, H_2, \dots, H_k\}$ be the merged fiber set, where N is the number of LUTs in the original netlist, and $\alpha, \beta \in \mathbb{R}$ are user-defined parameters. Minimize the following objective:

$$\text{minimize } \alpha \cdot \text{Extra}(\mathcal{H}) + \beta \cdot \text{Imbalance}(\mathcal{H})$$

Compute extra computation overhead (Extra) and load imbalance (Imbalance):

$$\text{Extra}(\mathcal{H}) = \left| \sum_{i=1}^k |H_i| - |N| \right|$$

$$\text{Imbalance}(\mathcal{H}) = \sqrt{\frac{1}{k} \sum_{i=1}^k \left(|H_i| - \frac{\sum_{j=1}^k |H_j|}{k} \right)^2}$$

We demonstrate the effectiveness of our partitioning algorithm using the JPEG benchmark [20] as a case study. Compared to directly applying hill climbing with individual registers as fiber root nodes, our method preserves register vector semantics in the RTL and performs iterations using register vectors as roots. As a result, under the same number of iterations, it reduces the Extra metric by 26.8 \times and the Imbalance metric by 30.8 \times , highlighting the importance of a good initial solution.

V. EFFICIENT IN-CORE PARALLEL COMPUTATION

A. Motivation

The previous section focused on fiber-based partitioning to enhance parallelism across cores. However, parallel efficiency within each core is equally critical: poor in-core scheduling that causes frequent LUT stalls can significantly increase computation time—potentially reducing simulation speed by half if excessive idle cycles occur.

This section therefore addresses the problem of optimizing parallel execution within individual compute cores.

B. Challenge and Insight

To improve in-core computation efficiency, we face two major challenges:

Computation Resource Allocation: Each fiber assigned to a core is a LUT-based computation graph comprising thousands of nodes—far exceeding the number of available physical LUT units. The complex data dependencies within each fiber pose

significant challenges for efficient time-multiplexing of these nodes onto a limited set of LUT resources, while maintaining high hardware utilization.

Memory Access Design: While fiber-based partition removes inter-core communication, intra-core LUT nodes still exchange data via SRAM-based shared memory. Due to LUTs’ 4-input, 1-output structure, SRAM demands many read ports, causing asymmetric bandwidth needs. Typical single- or dual-port SRAMs [21] can’t meet this. For instance, 4 LUT units require 16 read ports; using dual-port SRAMs means at least 8 SRAM blocks, increasing area and leaving many write ports underused.

This section identifies layered topological sorting as key to tackling these challenges. Based on this, we develop an efficient fiber-to-LUT mapping and optimize each core’s memory architecture, boosting parallel execution efficiency of multiple LUTs within a core.

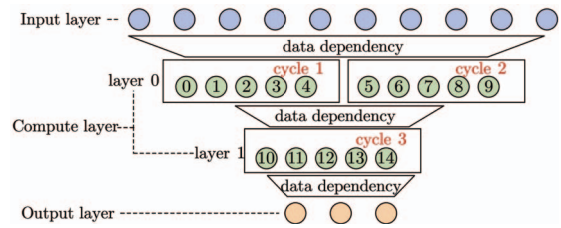


Fig. 5: Layered topological sorting exposes two independent layers; nodes in layers 0 and 1 can run in parallel. The core’s five LUT units bound the number of nodes executed per cycle.

C. Approach

1) *Computation resource allocation:* To address the challenge of computation resource allocation, an effective scheduling strategy is required to map LUT nodes in a DAG onto the available LUT computation units.

We apply layered topological sorting to the LUT-DAG, placing each node in the first layer whose predecessors are already assigned. Nodes in the same layer are independent. A sliding-window schedule, whose width equals the number of LUT units per core, is used. For example, in Figure 5, the core with five units processes nodes 0–4 in cycle 1, 5–9 in cycle 2, and so on.

2) *Memory Access Design:* To support high-speed access, we design an asymmetric multi-port SRAM matching the 4-input, 1-output nature of LUTs, which requires high read bandwidth. As shown in Figure 6, data is accessed via an address port with fixed-width words (e.g. 8, 16, or 32 bits). Since adding read ports can reduce memory frequency [22]–[24], we use multiple SRAM blocks to balance read bandwidth and access speed.

Specifically, we adopt a 5R1W SRAM [23], providing five read ports and one write port per block, supporting 1.5GHz access with 32-bit data words. To enable fine-grained access, a multiplexer selects specific bits from each word, allowing efficient single-bit reads required by computations.

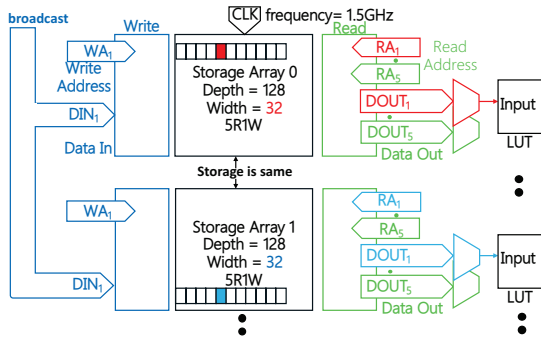


Fig. 6: In our implementation, we use four 5R1W SRAMs, providing a total of 20 ports, which can be connected to five 4-input LUTs.

VI. LOW-LATENCY NOC DESIGN

A. Motivation

Each RTL simulation cycle consists of two phases: combinational computation and sequential synchronization. Previous sections optimize computation, enabling each core to process its DAG subgraph efficiently. Before the next cycle, however, cores must synchronize to ensure consistent signal sharing. This synchronization often accounts for 20–50% of total cycle time [12], becoming a key performance bottleneck. Reducing this latency is critical for faster RTL simulation.

B. Challenge and Insight

Synchronization depends on inter-core interconnects for any-to-any data transfer and involves two key steps: injecting results from SRAM into the NoC and transmitting them to other cores. Each step presents its own challenges.

Sparse data injection challenge: Simulation results are sparsely scattered in SRAM, making them hard to read, pack, and inject into the NoC efficiently. Due to complex sub-DAG dependencies, register vector values are stored as individual bits across memory, rather than in contiguous locations, complicating data extraction for communication.

Physical constraints challenge: In large-scale NoCs connecting thousands of cores, maintaining low latency under area and power constraints is difficult. Topologies like mesh and ring support local communication but require many routers, increasing hop counts and delays [25].

To address this, we introduce two techniques: **1) Reusing memory access circuitry** as the interface between compute cores and the NoC, activated post-computation to mitigate sparse injection; **2) Clustering communication-intensive cores** after compiler partitioning, where intra-cluster synchronization uses a high-speed crossbar and inter-cluster communication relies on a low-speed ring, reducing power and area overhead, as shown in Figure 3.

C. Approach

1) Reusing Memory Access Circuits: To address sparse data injection, we repurpose memory circuitry as the interface between the core and NoC once computation completes. The

circuitry efficiently locates scattered bits in SRAM, enabling retrieval of an entire register vector in one cycle. After computation, cores synchronize through this interface, ensuring low-latency transfer and efficient coordination.

2) Two-Level NoC Synchronization: We implement synchronization via a two-level NoC: the first level adopts a crossbar topology, and the second level employs a ring topology. Core allocation is guided by the METIS [26] partitioning method based on inter-subgraph communication volume, exploiting spatial locality of data transfers. For comparison, we also design a mesh-based NoC. Simulation results show that, under the mesh network, synchronization cost is 1.96× higher and area overhead 5.37× larger than the proposed two-level topology.

VII. EVALUATION

A. Experimental Setup

The RTL simulation time cost of RAMP is measured using the following tools. **Architecture Simulator:** We developed a cycle-accurate simulator with two components—the computing core and the NoC—defining computation and data transfer interfaces. It measures computation and synchronization cycles per RTL cycle. **CAD Tools:** The hardware is designed in Verilog and synthesized using Synopsys Design Compiler with the TSMC 28nm standard cell library. The critical path delay of the slowest logic module is 0.53 ns, including setup and hold times. However, the overall operating frequency is ultimately constrained by the SRAM access latency of 0.67 ns, and RAMP operates at 1.5 GHz accordingly. The total power consumption remains below 40W. **Compilation:** Benchmarks are synthesized via Yosys, producing LUT-based netlists compiled into RAMP instructions using Python and C++. These instructions drive the architecture simulator for performance evaluation.

B. Hardware Features

In RAMP, each LUT has an instruction depth of 512. With 5 LUTs per core, one core simulates 2560 LUTs. The two-level 36×36 core topology totals 1296 cores, enabling simulation of 3.31 million LUTs and the area is 449.36mm² given by Design Compiler. Consequently, **RAMP** attains an emulation-capacity density of 0.135 mm²/kLUT₄. Commercial FPGAs fabricated in the same process node provide roughly 0.28 mm²/kLUT₆. According to a Xilinx report, one LUT₆ can replace approximately 1.6 LUT₄; therefore, the FPGA figure converts to 0.28 / 1.6 ≈ 0.175 mm²/kLUT₄. Hence, RAMP delivers a capacity-density improvement of (0.175 - 0.135) / 0.175 ≈ 23% over FPGAs at the same technology node.

C. Benchmark

VTA: Versatile Tensor Accelerator — a flexible, open-source deep learning accelerator [27]. **RV32R:** A 16 in-order, pipelined RISC-V processor implemented in Chisel [28]. **MC:** A Monte Carlo-based FPGA engine for stock option pricing [29].

D. Baseline

In the RTL simulation acceleration experiment, our work addresses the slow performance of functional-debug simulations by targeting single-stimulus, long-cycle workloads commonly encountered in real-world designs, and we thus compare our approach primarily against conventional CPU-based simulators and multi-core DSA specialized for simulation acceleration. We use Verilator [10] and Manticore [12] as baselines. Verilator is an efficient open-source tool that converts Verilog to C++ for faster simulation, supporting both single- and multi-threaded modes on CPUs. Manticore is a multi-core DSA that speeds up simulation through parallelism, offering higher efficiency.

	CPU	Manticore	RAMP
Cores	32	225	1296
GHz	2.5-3.4	0.475	1.5
MiB	105.5	18.45	32.59

TABLE II: Different Hardware Platforms. Cores, GHz, MiB denote physical core count, clock frequency, SRAM capacity.

E. Results

DUT	Simulation Frequency (KHz)				Speedup over SOTA
	Verilator-ST	Verilator-MT	Manticore	RAMP	
VTA	32.4	94.9	278	2964	10.66X
RV32R	97.3	73.3	221	2852	12.90X
MC	26.6	68.9	423	3104	7.33X

TABLE III: Simulation performance of Verilator, Verilator-MT, Manticore, and RAMP.

RTL simulation time cost. In Table III., we compare RAMP with Verilator and Manticore, respectively. For Verilator, both single- (Verilator-ST) and multi-threaded (Verilator-MT) ones are compared. We align the number of threads in Verilator-MT with the configuration used in the paper [12], setting it to 5, which is the fastest configuration. RAMP is $45 \times$ faster than Verilator because it scales to many more cores, shown in Table II, which enables significantly higher computational throughput. For Manticore, RAMP is $12.9 \times$ faster due to the fine-grained compute units and the effective scheduling algorithm in our design. Our design enables parallel acceleration by leveraging the fine-grained nodes in the netlist. By contrast, Manticore compiler produces a coarse-grained graph, which limits exposed parallelism and hinders acceleration through core scaling. The comparisons show the effectiveness of RAMP.

Compilation time cost. In Table IV, compared with Manticore, RAMP achieves notable speedup. This is because the fine-grained mapping between nodes and compute units in RAMP enables simpler, more effective parallelization, accounting for the largest portion of Manticore’s compilation time.

Instruction analysis. By conducting instruction analysis in Figure 7, we find the acceleration achieved by RAMP can mainly be attributed to small proportion of idle instructions. As idle instructions arise mainly from load imbalance, this observation indicates that our fiber-based graph partition

DUT	Compilation Time (s)		
	Verilator	Manticore	RAMP
VTA	2m33	15m29s	7m23s
RV32R	1m56s	5m57s	2m17s
MC	1m13s	12m57s	5m26s

TABLE IV: Compilation-time comparison of Verilator, Manticore, and RAMP.

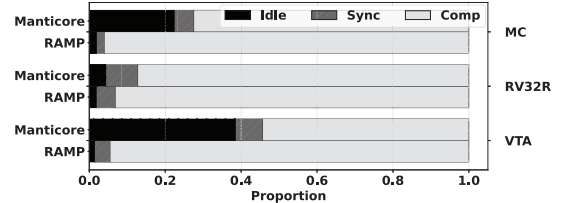


Fig. 7: Comparison of Proportion of Idle, Synchronization, and Computation Instructions between RAMP and Manticore.

algorithm effectively achieves load balancing across cores. As a result of few idle and synchronization instructions, the simulation speedup of RAMP can be roughly estimated to be proportional to the number of LUTs, showing the computational scalability of our design.

VIII. DISCUSSION AND FUTURE WORK

RAMP offers a practical approach for RTL acceleration. While the current SRAM design is functionally correct, it incurs notable area and power overhead from architectural complexity. We also explored implementing multi-port memory using DFFs in Verilog, but under the TSMC 7nm process the critical path reached 10.3 ns, limiting frequency below 100 MHz and rendering the solution impractical. Although FPGA experiments were not performed, this aligns with the Manticore benchmark suite, where all test cases fit on a single FPGA. In such settings, FPGAs achieve higher frequencies than RAMP but at the cost of longer compilation—an overhead our design avoids. We therefore evaluate end-to-end efficiency under a consistent benchmark framework. Achieving the high energy efficiency envisioned by RAMP will ultimately require custom multi-port memory blocks, which we leave for future work.

According to public data, the emulation speed of current processor-based emulators such as Palladium is about 100 K–2 M cycles per second [30]. Our approach already exceeds this performance, though further studies on larger designs are needed to fully validate scalability.

IX. CONCLUSION

This paper presents RAMP, an LUT-based accelerator for full-cycle RTL simulation that combines balanced DAG partitioning with deeply pipelined compute cores to accelerate combinational logic, and integrates a low-latency on-chip network to streamline sequential synchronization. Evaluated on benchmark designs, RAMP achieves up to $12.9 \times$ speedup over state-of-the-art multi-core simulators, demonstrating the effectiveness of its co-designed hardware and compilation flow.

REFERENCES

- [1] R. Saleh, S. Wilton, S. Mirabbasi, A. Hu, M. Greenstreet, G. Lemieux, P. P. Pande, C. Grecu, and A. Ivanov, "System-on-chip: Reuse and integration," *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1050–1069, 2006.
- [2] S. Saponara, M. S. Greco, and F. Gini, "Radar-on-chip/in-package in autonomous driving vehicles and intelligent transport systems: Opportunities and challenges," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 71–84, 2019.
- [3] J. Choquette and W. Gandhi, "Nvidia a100 gpu: Performance & innovation for gpu computing," in *2020 IEEE Hot Chips 32 Symposium (HCS)*. IEEE Computer Society, 2020, pp. 1–43.
- [4] H. Foster *et al.*, "The 2022 wilson research group functional verification study," *Siemens. com*, 2022.
- [5] M. Emami, T. Bourgeat, and J. Larus, "Parendi: Thousand-way parallel rtl simulation," *arXiv preprint arXiv:2403.04714*, 2024.
- [6] S. Emami, "Highly parallel rtl simulation," Ph.D. dissertation, EPFL, 2024.
- [7] H. Wang and S. Beamer, "Reput: Superlinear parallel rtl simulation with replication-aided partitioning," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, 2023, pp. 572–585.
- [8] D.-L. Lin, H. Ren, Y. Zhang, B. Khailany, and T.-W. Huang, "From rtl to cuda: A gpu acceleration flow for rtl simulation with batch stimulus," in *Proceedings of the 51st International Conference on Parallel Processing*, 2022, pp. 1–12.
- [9] D. Kim, *FPGA-Accelerated Evaluation and Verification of RTL Designs*. University of California, Berkeley, 2019.
- [10] W. Snyder, "Verilator," <https://www.veripool.org/verilator/>, n.d., accessed: 2021-05-10.
- [11] Cadence, "Cadence palladium," https://www.cadence.com/en_US/home/tools/system-design-and-verification/emulation-and-prototyping/palladium.html, 2022.
- [12] M. Emami, S. Kashani, K. Kamahori, M. S. Pourghannad, R. Raj, and J. R. Larus, "Manticore: Hardware-accelerated rtl simulation with static bulk-synchronous parallelism," in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4*, 2023, pp. 219–237.
- [13] F. Elsabbagh, S. Sheikhha, V. A. Ying, Q. M. Nguyen, J. S. Emer, and D. Sanchez, "Accelerating rtl simulation with hardware-software co-design," in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, 2023, pp. 153–166.
- [14] M. Abeydeera and D. Sanchez, "Chronos: Efficient speculative parallelism for accelerators," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 1247–1262.
- [15] N. Childs, "How Intel agilex[®] fpga is enabling resource and power efficient 4k, 8k video processing solutions," Intel Corporation, White Paper, 2023. [Online]. Available: <https://www.intel.com/content/dam/www/central-libraries/us/en/documents/agilex-fpga-video-processing-solutions-white-paper.pdf>
- [16] C. Wolf, "Yosys open synthesis suite," <https://yosyshq.net/yosys/>, 2013.
- [17] H. Qian and Y. Deng, "Accelerating rtl simulation with gpus," in *2011 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2011, pp. 687–693.
- [18] R. Pu, Y. Sun, P.-H. Ho, F. Yang, L. Shang, and X. Zeng, "Sphinx: A hybrid boolean processor-fpga hardware emulation system," in *2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE, 2023, pp. 1–9.
- [19] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine learning*, vol. 65, pp. 31–78, 2006.
- [20] UltraEmbedded, "High throughput jpeg decoder," https://github.com/ultraembedded/core_jpeg, 2023.
- [21] H. Noguchi, S. Okumura, Y. Iguchi, H. Fujiwara, Y. Morita, K. Nii, H. Kawaguchi, and M. Yoshimoto, "Which is the best dual-port sram in 45-nm process technology?—8t, 10t single end, and 10t differential—," in *2008 IEEE International Conference on Integrated Circuit Design and Technology and Tutorial*. IEEE, 2008, pp. 55–58.
- [22] D.-P. Wang, H.-J. Lin, C.-T. Chuang, and W. Hwang, "Low-power multiport sram with cross-point write word-lines, shared write bit-lines, and shared write row-access transistors," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 3, pp. 188–192, 2014.
- [23] S.-F. Hsiao and P.-C. Wu, "Design of low-leakage multi-port sram for register file in graphics processing unit," in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2014, pp. 2181–2184.
- [24] R. Ohara, M. Kabuto, M. Taichi, A. Fukunaga, Y. Yasuda, R. Hamabe, S. Izumi, and H. Kawaguchi, "A 1w/8r 20t sram codebook for deep learning processors to reduce main memory bandwidth," *IEICE Technical Report; IEICE Tech. Rep.*, vol. 123, no. 144, pp. 41–44, 2023.
- [25] N. Abeyratne, R. Das, Q. Li, K. Sewell, B. Giridhar, R. G. Dreslinski, D. Blaauw, and T. Mudge, "Scaling towards kilo-core processors with asymmetric high-radix topologies," in *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2013, pp. 496–507.
- [26] G. Karypis, "Metis," <https://github.com/KarypisLab/METIS>, 2025, accessed: 2025-04-22.
- [27] T. Moreau, T. Chen, L. Vega, J. Roesch, E. Yan, L. Zheng, J. Fromm, Z. Jiang, L. Ceze, C. Guestrin *et al.*, "A hardware–software blueprint for flexible deep learning specialization," *IEEE Micro*, vol. 39, no. 5, pp. 8–16, 2019.
- [28] D. Kim, "riscv-mini," <https://github.com/ucb-bar/riscv-mini>, 2023.
- [29] X. Tian and K. Benkrid, "Design and implementation of a high performance financial monte-carlo simulation engine on an fpga supercomputer," in *2008 International Conference on Field-Programmable Technology*. IEEE, 2008, pp. 81–88.
- [30] L. Scheffer, L. Lavagno, and G. Martin, *EDA for IC system design, verification, and testing*. CRC press, 2018.