Macsim Mini: A Lightweight Cycle-Level GPU Simulator for Architecture Education

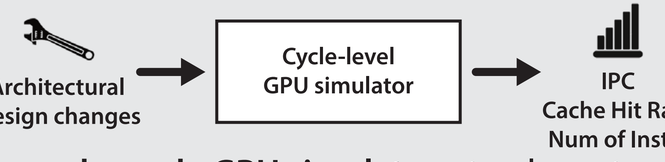
Euijun Chung*, Saurabh Singh*, Huanzhi Pu, Yuxiao Jia, Anurag Kar, Sam Jijina, Scott Madeira, Hyesoon Kim

HPArch (High Performance Architecture) Lab @ **Georgia Tech**

** Equal Contribution*

1. Motivation

- **Cycle-level simulators** are proven teaching tools; students learn μarchitecture by modifying the code.



- **Research-grade GPU simulators**: too large to navigate in one semester (>60K LoC).
- **Functional models**: too abstract to expose μarchitectural trade-offs, e.g., warp scheduling, exec. units.

Key Insight: **Memory system** dominates GPU performance trends in design space exploration.
→ Model **memory** in detail, keep **compute** simple.

2. Simulator Design

Goal: cycle-level simulation with HW (SASS) traces.
Design Principle: Memory in detail**, compute** simple.

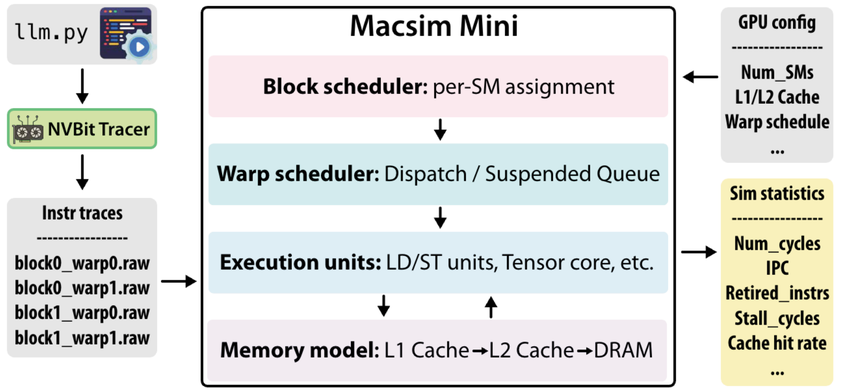
- **Modeled:** warp/block scheduling, L1/L2 cache, intra-warp memory coalescing, tensor cores, DRAM.
- **Simplified:** fixed-latency compute pipeline with RAW dependencies.
- **Omitted:** sub-core, NoC, MSHR, address translation.

Table I. Comparison of Accel-Sim and Macsim Mini.

	Accel-Sim [10]	Macsim Mini (ours)
Granularity	Cycle-level	Cycle-level
Simulation speed (MIPS)*	1.46	14.1
Lines of code	64.2 K	3.8 K
Accuracy vs. real HW	High	Trend-accurate

*Arithmetic mean of MIPS (Million Instructions Per Second), measured on Intel Xeon E5-2696 v4 CPU with the Rodinia benchmark suite [2].

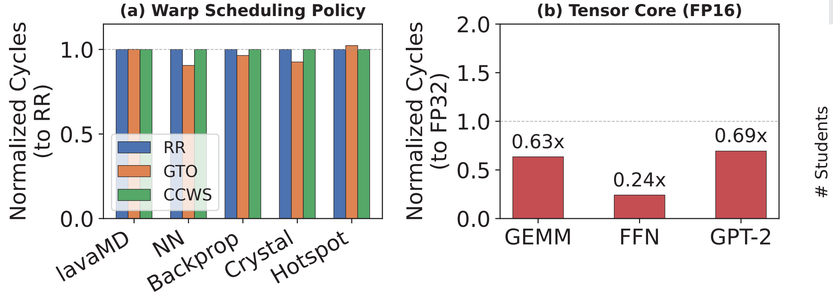
3. Simulation Flow



- **NVBit SASS tracer:** per-warp instruction traces.
- **Two-level** warp scheduler [Narasiman '11].
- **Modular DRAM model:** fixed-latency / lightweight row & bank model / Ramulator.

4. Evaluation Result

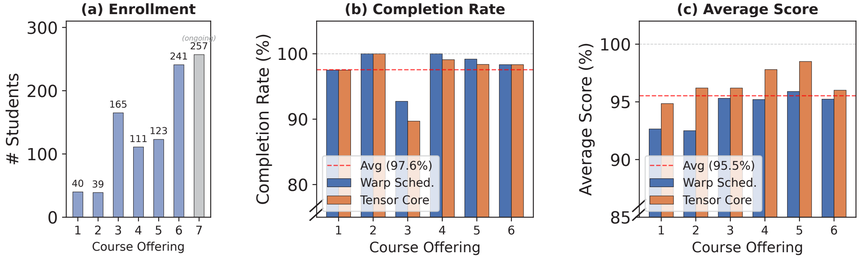
- **Task 1:** Round robin (RR) / Greedy-then-oldest (GTO) / Cache-Conscious (CCWS) warp scheduling.
- **Task 2:** Tensor core dispatch (FP16 vs. FP32).
- **Benchmarks:** Rodinia suite & GEMM, FFN, GPT-2.



5. Course Integration at Georgia Tech

Macsim Mini has been deployed at Georgia Tech's graduate GPU architecture course for 7+ semesters.

- **976 students** used Macsim Mini since Spring 2024.
- **Completion rate** of **97.6%** and **average score** of **95.5%**; Macsim Mini is accessible for most students.



6. Takeaway

Macsim Mini is a *compact*, *accurate*, and *extensible* cycle-level GPU simulator for education purposes.

- **3.8K LoC:** students modify within one semester.
- **Trend-accurate:** reproduces published behaviors.
- **Modular:** easily swap DRAM models or adding execution units.
- **Future work:** Adding sub-core, address translation, and TMA (Tensor Memory Accelerator) modules.

Macsim Mini lowers the barrier to cycle-level GPU simulation for education and research.