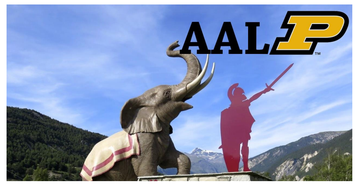
A Detailed Model for Contemporary GPU Memory Systems

Mahmoud Khairy*, Akshay Jain*, Tor Aamodt^, Timothy G. Rogers*
*Purdue University, School of Electrical and Computer Engineering
^University of British Columbia, Department of Electrical and Computer Engineering
abdallm@purdue.edu, akshayj@alumni.purdue.edu, aamodt@ece.ubc.ca, timrogers@purdue.edu
ISPASS 2019



Accelerator Architecture Lab at Purdue

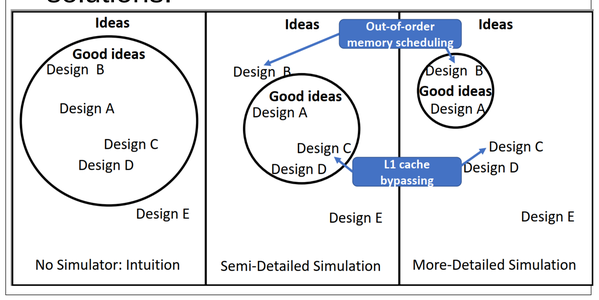
Introduction

- GPGPU-sim:
  - Widely used GPU simulator in the research community (1200+ citations).
  - The third most cited simulator in computer architecture field (after GEM5 and SimpleScalar)
- Last major update to GPGPU-Sim modeled the 9 year old NVIDIA Fermi architecture.

- But, how well does GPGPU-sim model contomprary GPU hardware?
  - Recent work [1] on validating GPGPU-Sim vs the NVIDIA Pascal architecture has demonstrated that there are several areas where a lack of detail in the memory system model creates significant error.
- We perform a module-by-module redesign of the GPU's system, demonstrating its improved correlation with real hardware.
- We develop a new *Correlator* toolset that allows users of GPGPU-Sim to easily generate counter-by-counter correlation plots.

Why Accuracy is Important?

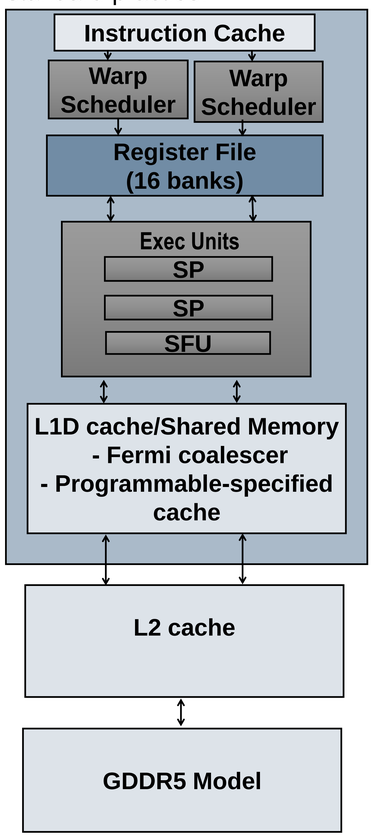
- Simulator-driven Ideas.
- Relying on inaccurate old model may lead to ineffective ideas or less optimal solutions.



Exploring the Contemporary GPU System

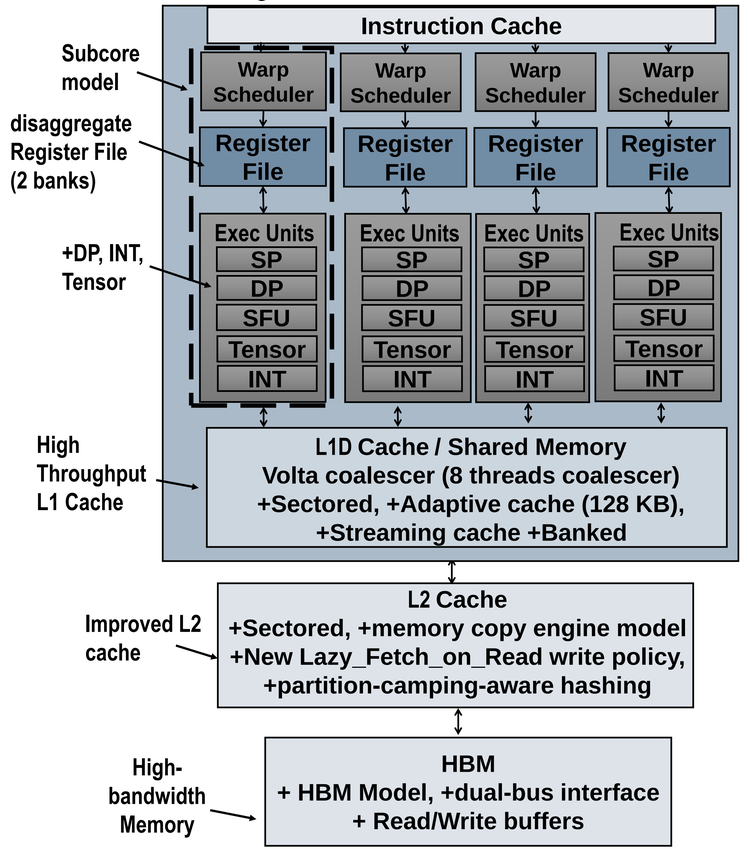
Old Fermi-based Model

- Note: "old model" for Volta was obtained by scaling the Fermi resources similar to standard practice.

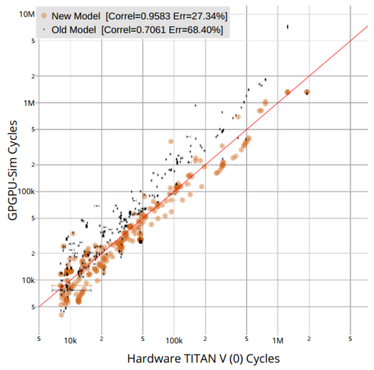


New Volta-based Model

The new model characteristics were obtained via publicly available documents and patents from Nvidia and via running extensive in-house micro-benchmarks.



Hardware Correlation

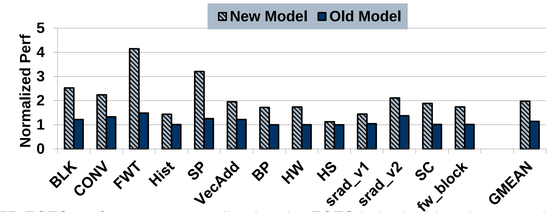


Execution cycles in simulation (old model vs new model) relative to hardware (over 1200 kernels).

Statistic	Means Abs Error		Correlation		
	Old Model	New Model	Old Model	New Model	
L1 Reqs	48%	0.5%	92%	100%	
L1 Hit Ratio	41%	18%	89%	93%	
L2 Reads	66%	1%	49%	94%	1% error in L2 behavior (66x read error reduction)
L2 Writes	56%	1%	99%	100%	
L2 Read Hits	80%	15%	68%	81%	
DRAM Reads	89%	11%	60%	95%	7X Error Reduction in DRAM reads
Execution Cycles	68%	27%	71%	96%	2.5X Error Reduction in Exec time

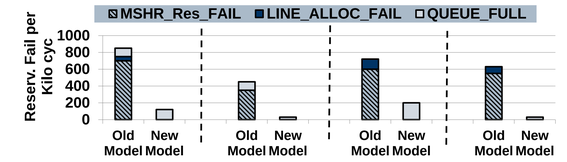
Design Decision Case Study

- The memory scheduling policy has a more dramatic impact on performance in the new model than in the old model.

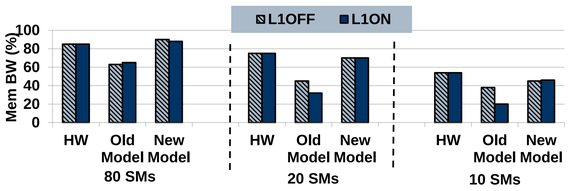


FR FCFS performance normalized to the FCFS in both old and new model

- L1 cache throughput bottleneck is mitigated in modern GPUs.



L1 cache reservation fails per kilo cycles for some microbenchmarks



BW utilization of STREAM workload for TITANV HW, Old Model and New Model when L1 cache is turned On and Off

Conclusion

- This paper presents the most accurate open-source model of a contemporary GPU to date.
- We refer the reader to [2] for a complete analysis of our model.

Now integrated into the GPGPU-Sim dev branch [3]

References

[1] A. Jain, M. Khairy, and T. G. Rogers, "A quantitative evaluation of contemporary gpu simulation methodology," *SIGMETRICS* 2018
[2] M. Khairy, A. Jain, T. M. Aamodt, and T. G. Rogers, "Exploring modern GPU memory system design challenges through accurate modeling," *http://arxiv.org/abs/1810.07269*, 2018
[3] GPGPU-sim Github, dev branch, https://github.com/gpgpusim/gpgpusim_distribution/tree/dev