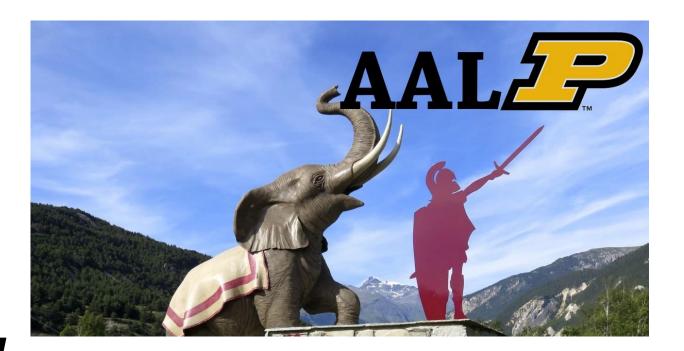


# A Detailed Model for Contemporary GPU Memory Systems

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** 



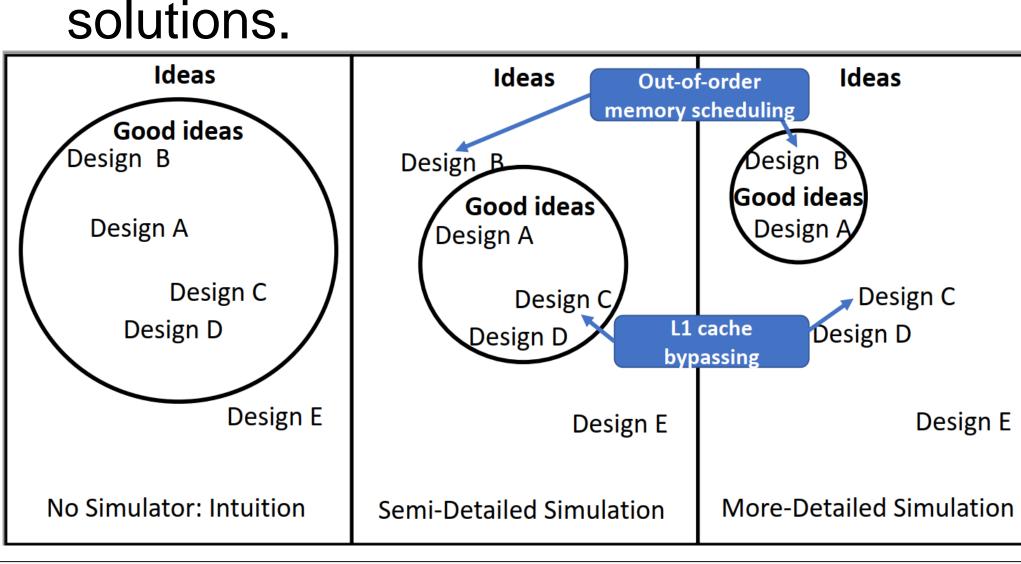
**Accelerator Architecture Lab at** Purdue

### 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



1200 kernels).

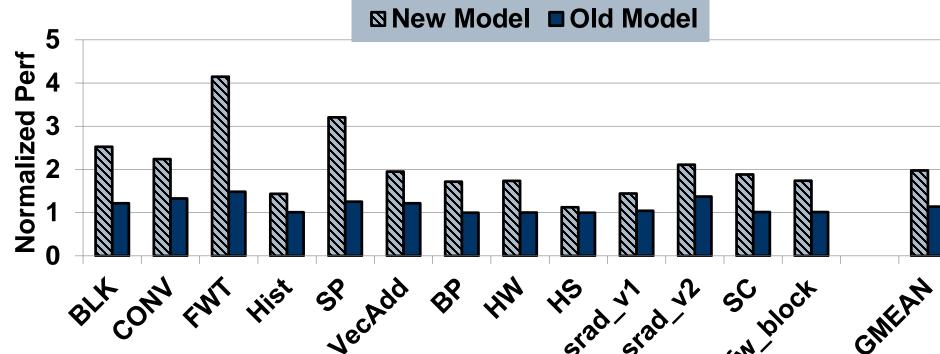
#### **New Volta-based Model** Old Fermi-based Model The new model characteristics were obtained via publicly Note: "old model" for Volta was obtained by available documents and patents from Nvidia and via scaling the Fermi resources similar to running extensive in-house micro-benchmarks. standard practice. **Instruction Cache Instruction Cache** Subcore Warp Warp Warp Warp Warp model Scheduler | Scheduler | Scheduler | Scheduler | Scheduler Scheduler disaggregate Register Register Register Register Register File **Register File** File File File (16 banks) (2 banks) **Exec Units Exec Units Exec Units Exec Units Exec Units** +DP, INT, SP SP Tensor SP SFU SFU SFU SFU SFU Tensor Tensor Tensor Tensor INT L1D cache/Shared Memory High L1D Cache / Shared Memory - Fermi coalescer Throughput \ Volta coalescer (8 threads coalescer) - Programmable-specified +Sectored, +Adaptive cache (128 KB), L1 Cache cache +Streaming cache +Banked L2 Cache +Sectored, +memory copy engine model L2 cache cache +New Lazy\_Fetch\_on\_Read write policy, +partition-camping-aware hashing **HBM** High-**GDDR5 Model** + HBM Model, +dual-bus interface bandwidth + Read/Write buffers Memory **Hardware Correlation**

**Exploring the Contemporary GPU System** 

New Model [Correl=0.9583 Err=27.34%] Old Model [Correl=0.7061 Err=68.40%]	Statistic	Means Abs Error		Correlation		
		Old Model	New Model	Old Model	New Model	
1M 5	L1 Reqs	48%	0.5%	92%	100%	
GPGPU-Sim Cycles  The state of	L1 Hit Ratio	41%	18%	89%	93%	
D 5	L2 Reads	66%	1%	49%	94%	1% error in L2
2	L2 Writes	56%	1%	99%	100%	behavior (66x read error reduction)
10k	L2 Read Hits	80%	15%	68%	81%	
Hardware TITAN V (0) Cycles	DRAM Reads	89%	11%	60%	95%	7X Error Reduction in DRAM reads
Execution cycles in simulation (old model s new model) relative to hardware (over 200 kernels)	Execution Cycles	68%	27%	71%	96%	2.5X Error Reduction in Exec time

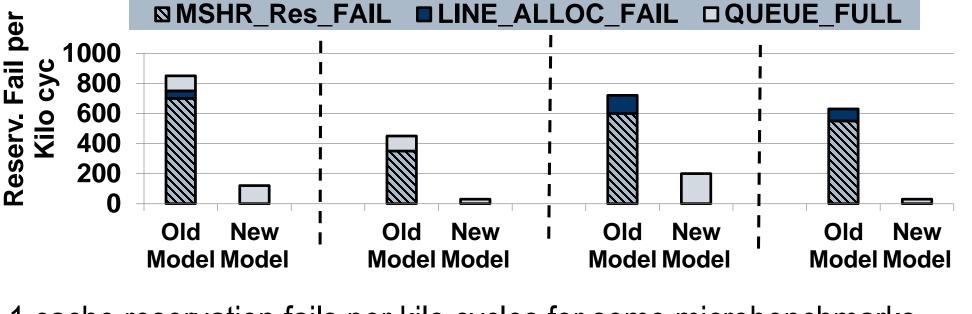
## **Design Decision Case Study**

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

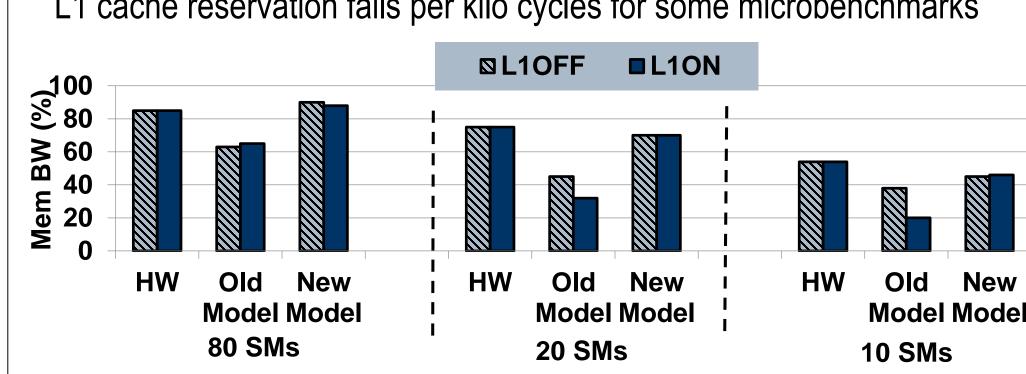


FR FCFS performance normalized to the FCFS in both old and new model

 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