Photon: A Fine-grained Sampled Simulation Methodology for GPU Workloads

Changxi Liu[1], Yifan Sun[2], Trevor E. Carlson[1]

[1]National University of Singapore

[2]College of William & Mary

GPUs and GPU Simulators

- GPU simulators
  - Pre-silicon GPUs Architecture Exploration
  - Software profiling and optimization
  - Obtain performance characteristics for new architectures

GPUs and GPU Simulators

- GPU simulators
  - 12.5 KIPS

Simulators	GPGPUSim 3.x	gem5-APU	MGPUSim	MacSim	Multi2-Sim	Accel-Sim
Sim. Rate (KIPS)	3	N/A	28	N/A	0.8	12.5

[1]. Khairy, Mahmoud, et al. "Accel-Sim: An extensible simulation framework for validated GPU modeling." ISCA 2020.

Table source: [1]

GPUs and GPU Simulators

- GPU simulators
- Today's GPUs achieve nearly 134 TFLOPS [2]
- Over 1,000,000,000 slower than the real GPU

Simulators	GPGPUSim 3.x	gem5-APU	MGPUSim	MacSim	Multi2-Sim	Accel-Sim
Sim. Rate (KIPS)	3	N/A	28	N/A	0.8	12.5

Table source: [1]

[1]. Khairy, Mahmoud, et al. "Accel-Sim: An extensible simulation framework for validated GPU modeling." ISCA 2020.
[2]. NVIDIA, "NVIDIA H100 Tensor Core GPU Architecture," https://www.nvidia.com/en-us/data-center/h100/, 2022

Sampled GPU Simulation

	Profiling	Inter-kernel Sampling	Intra-kernel Sampling
PKA[1]	Offline	Yes, handpicked features	Stable IPC
TBPoint[2]	Offline	Yes, handpicked features	Stable IPC
Sieve[3]	Offline	Yes, handpicked features	N/A

[1]. Avalos Baddouh, Cesar, et al. "Principal kernel analysis: A tractable methodology to simulate scaled gpu workloads." MICRO 2021.

[2]. Huang, Jen-Cheng, et al. "TBPoint: Reducing simulation time for large-scale GPGPU kernels." IPDPS 2014.

[3]. Naderan-Tahan, et al"Sieve: Stratified GPU-Compute Workload Sampling." ISPASS 2023.

Sampled GPU Simulation

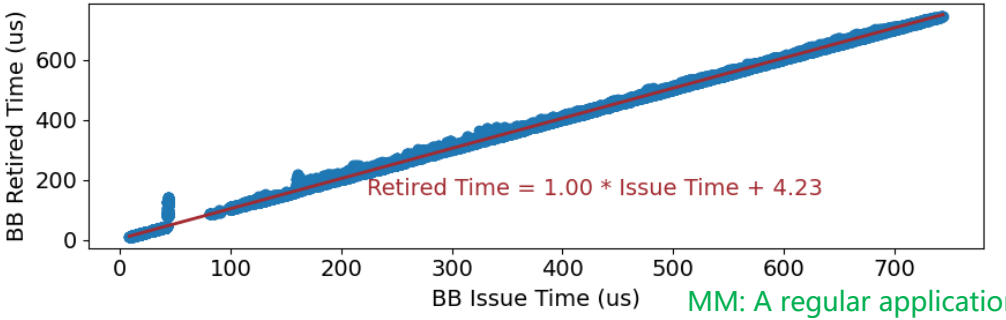
	Profiling	Inter-kernel Sampling	Intra-kernel Sampling
PKA[1]	Offline	Yes, handpicked features	Stable IPC
TBPoint[2]	Offline	Yes, handpicked features	Stable IPC
Sieve[3]	Offline	Yes, handpicked features	N/A
Photon	Online	Yes, GPU BBVs	Stable warps and basic blocks

[1]. Avalos Baddouh, Cesar, et al. "Principal kernel analysis: A tractable methodology to simulate scaled gpu workloads." MICRO 2021.

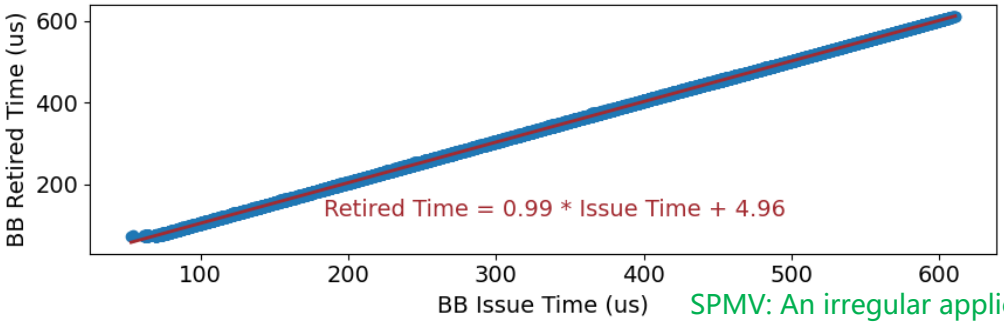
[2]. Huang, Jen-Cheng, et al. "TBPoint: Reducing simulation time for large-scale GPGPU kernels." IPDPS 2014.

[3]. Naderan-Tahan, et al"Sieve: Stratified GPU-Compute Workload Sampling." ISPASS 2023.

Observations – Basic Blocks



MM: A regular application

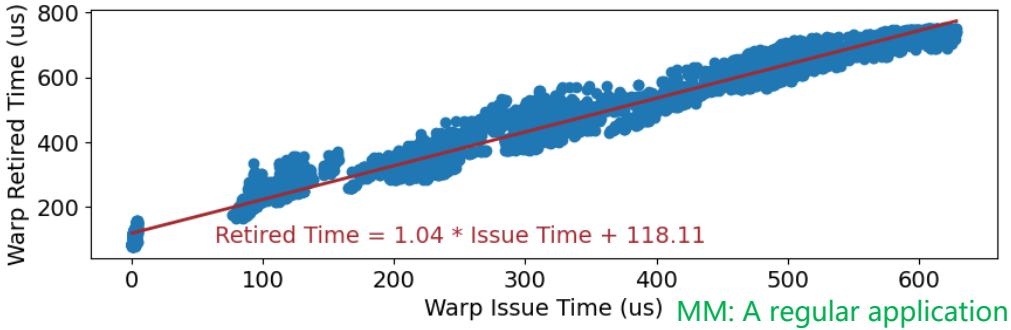


SPMV: An irregular application

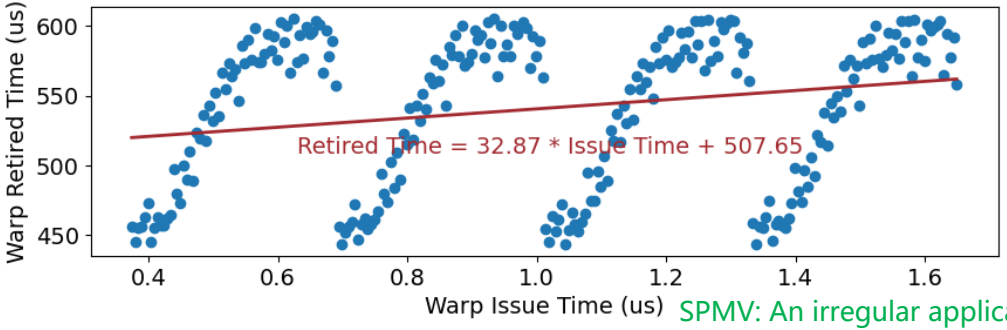
The issue and retired time of the dominating (in terms of execution time) basic blocks, which all have the same entry and exit points.

- Basic blocks' execution time can be stable over time
  - Slope value close to 1
    - The least square method

Observations – Warps



MM: A regular application

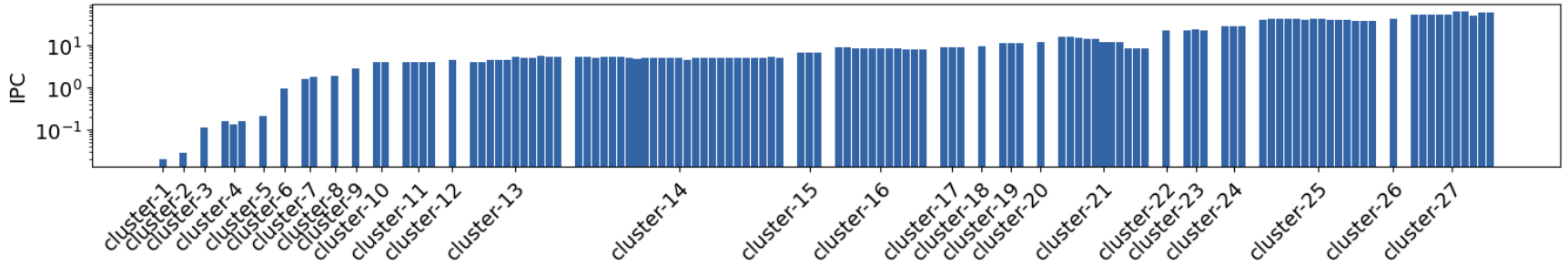


SPMV: An irregular application

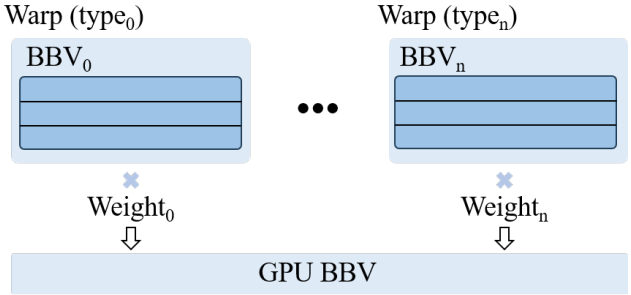
The issue and retired time of warps

- Warps for regular applications can be stable over time
  - Regular applications
    - The slope is close to 1.
  - Irregular applications
    - The slope is far away from 1.

Observations – Kernels

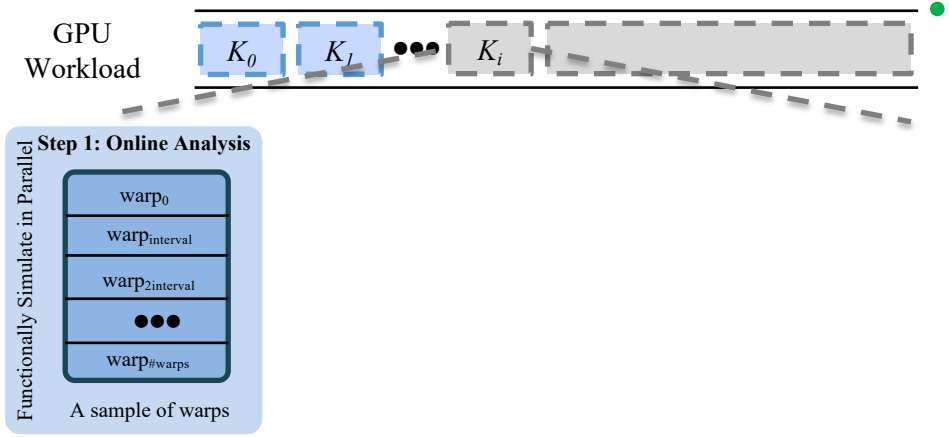


VGG-16

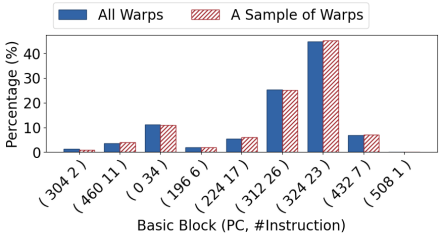


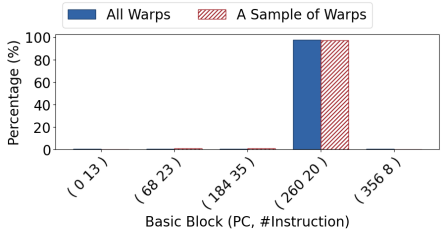
- GPU BBV
  - Concat $BBV \times Weight$ of each type of Warp
    - $weight_{type} = \frac{\#\,Warptype}{\#\,Warpall}$
- GPU Kernels with similar GPU BBV have similar IPC

Methodology-Basic-Block-Sampling

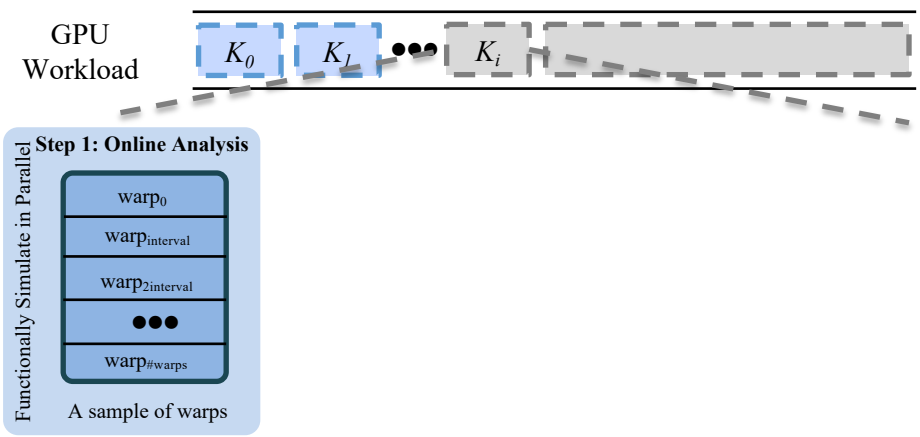


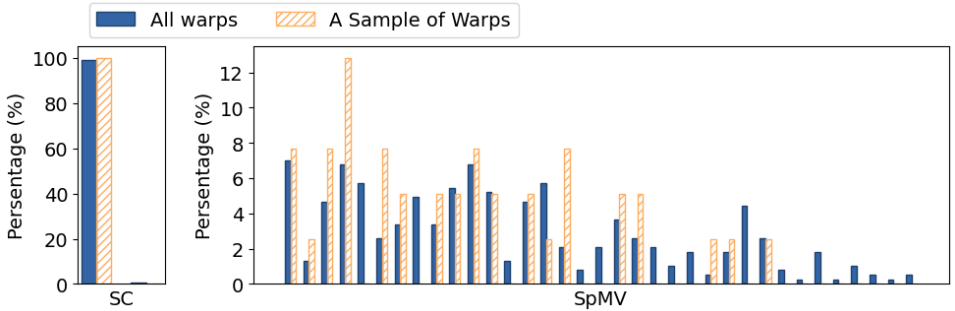
- Basic Blocks
  - The distribution of basic blocks of all warps and a sample of warps.





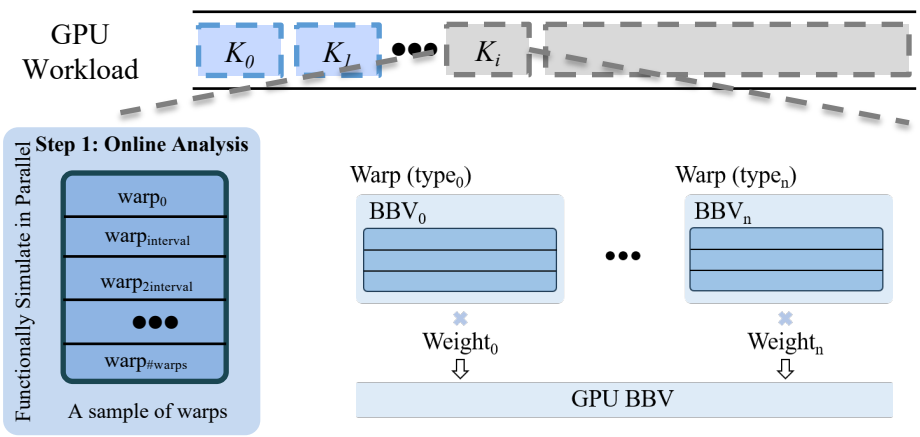
Methodology-Basic-Block-Sampling

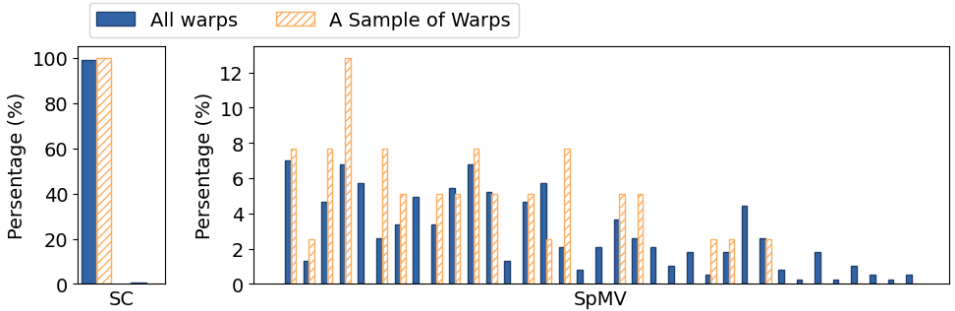




- Basic Blocks
  - The distribution of basic blocks of all warps and a sample of warps.
- Warps
  - The distribution of different warp types of all warps and a sample of warps.

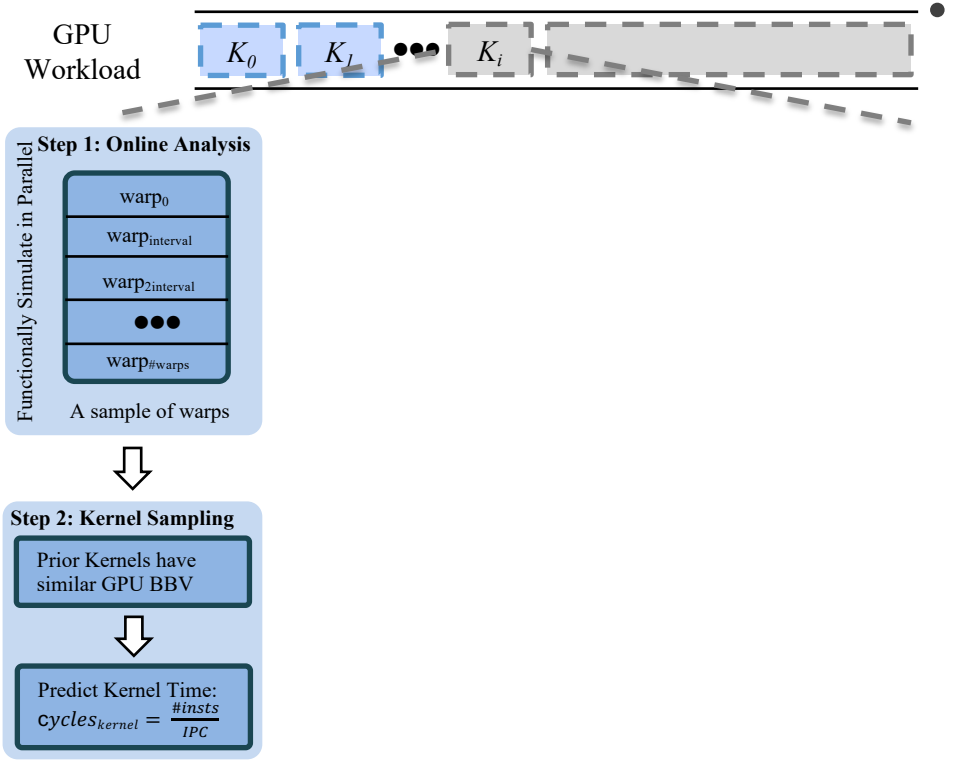
Methodology-Basic-Block-Sampling





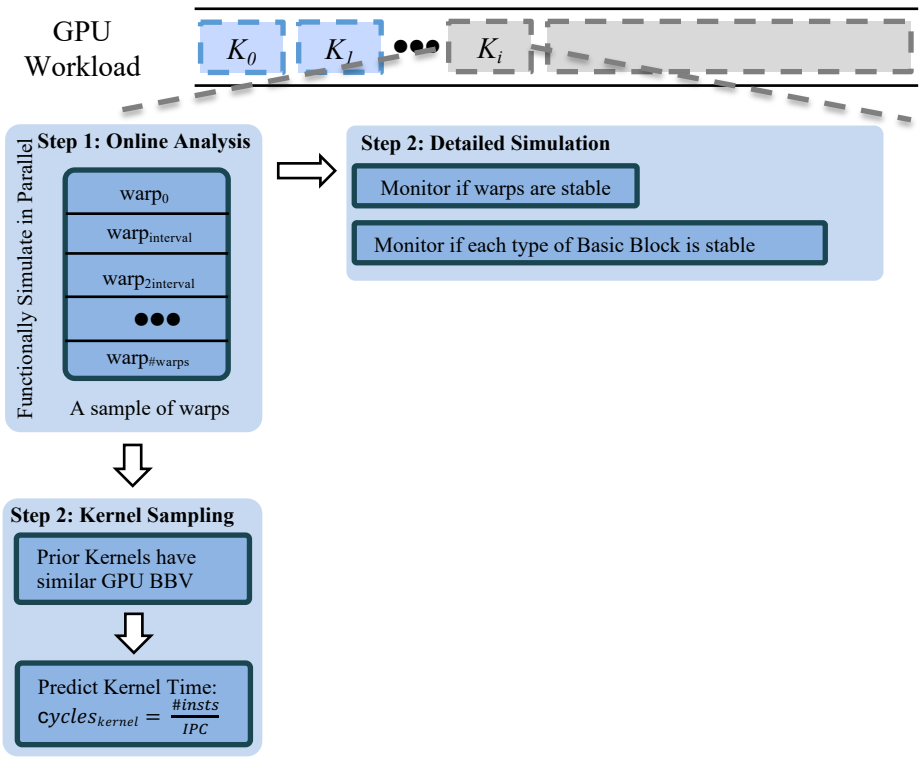
- Basic Blocks
  - The distribution of basic blocks of all warps and a sample of warps.
- Warps
  - The distribution of different warp types of all warps and a sample of warps.
- Kernels
  - GPU BBV is combined with BBVs for each type of warp from a sample of warps.

Methodology-Basic-Block-Sampling



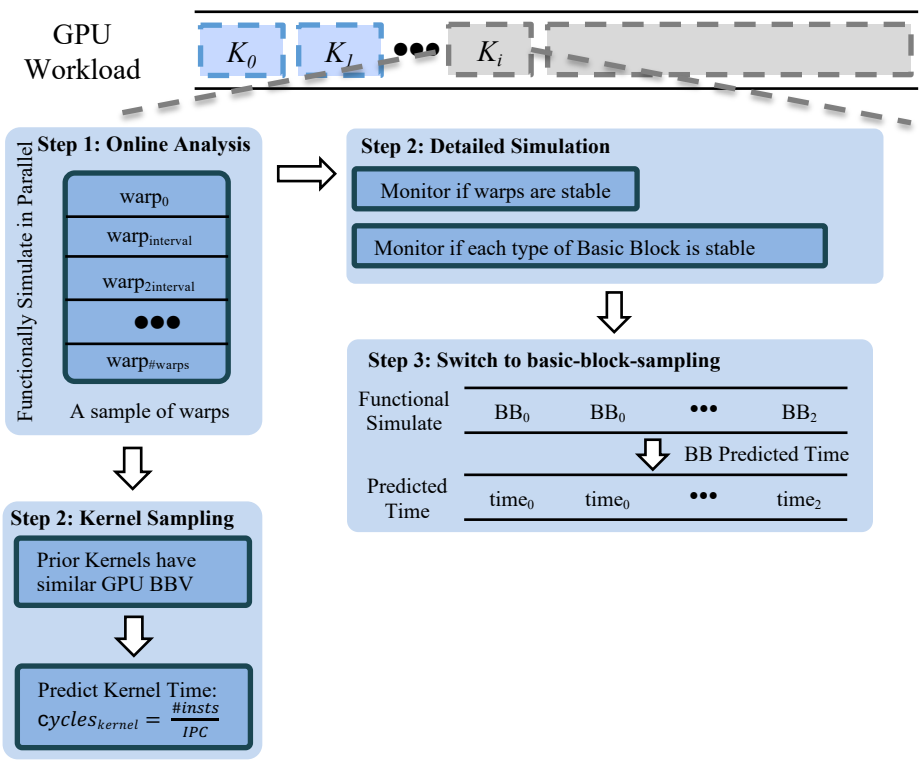
- Kernel-Sampling
  - Prior Kernels have similar GPU BBV
    - Skip the simulation of the kernel
    - predict the simulation time with prior similar kernels' IPC
  - Prior Kernels **do not** have similar GPU BBV
    - basic-block-sampling and warp-sampling

Methodology-Basic-Block-Sampling



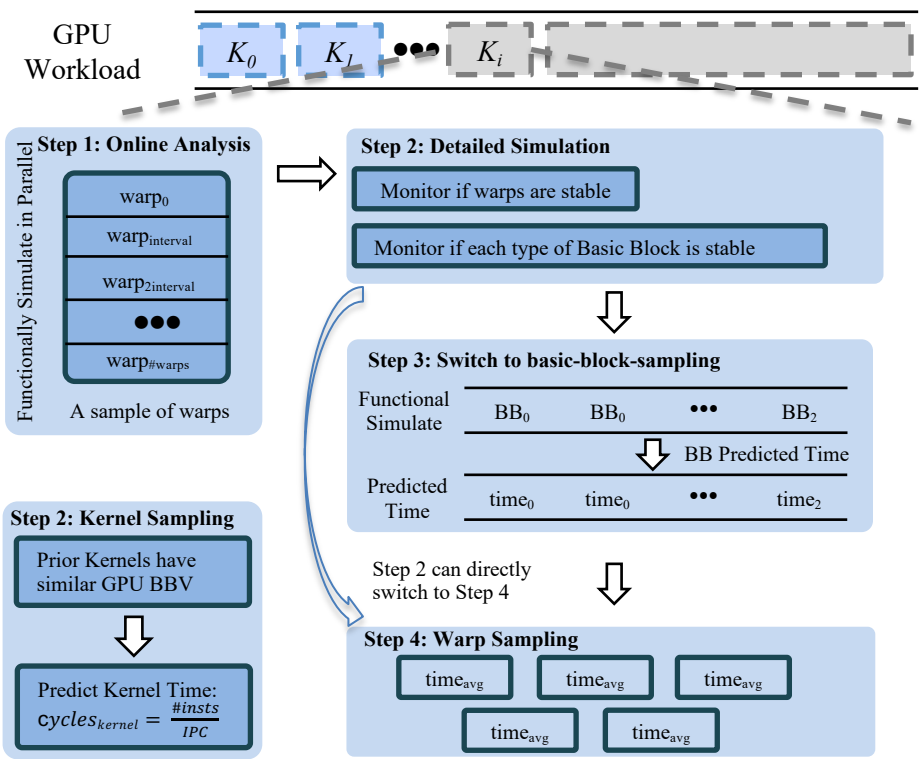
- Warp-Sampling
  - Check if warps are stable.
- Basic-Block-Sampling
  - Check if each type of basic block is stable.
  - Check the percentage of stable basic blocks
- Stable
  - The slope value of the last N warps (basic blocks) is close to 1.

Methodology-Basic-Block-Sampling



- Basic-Block-Sampling
  - Functional simulate
  - Predict the simulation time of warps using the predicted time of each type of basic block.

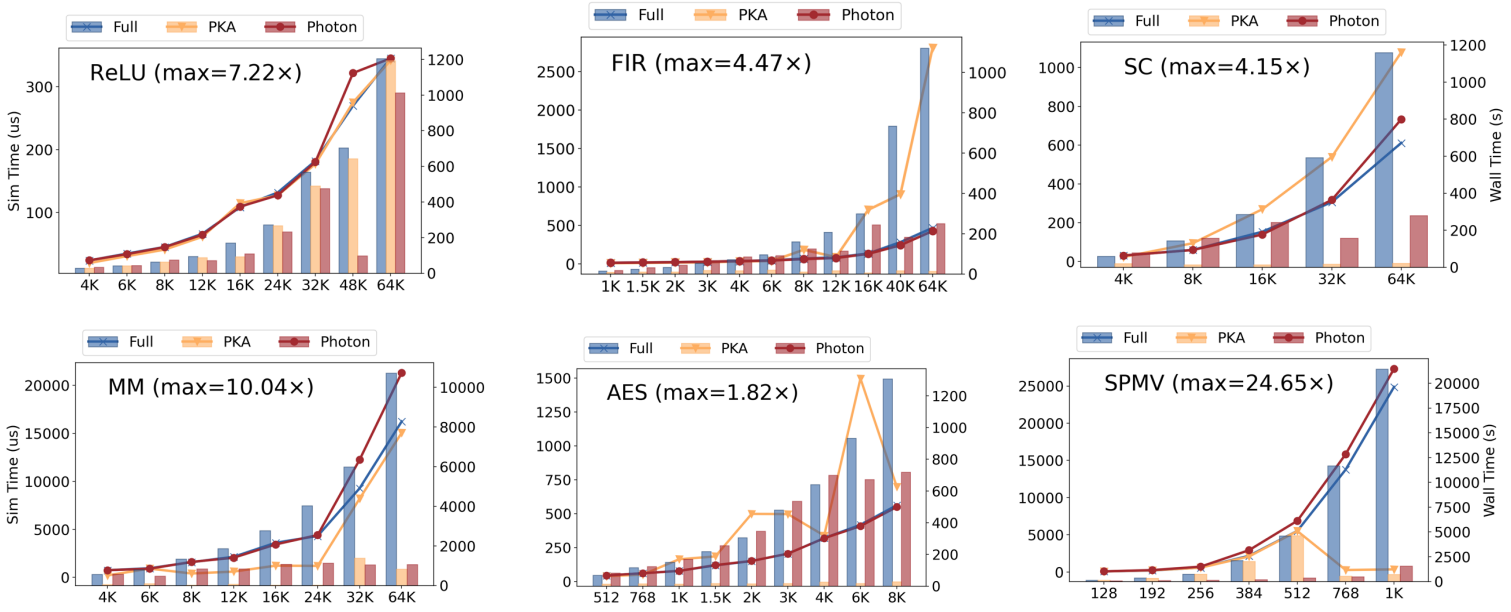
Methodology-Basic-Block-Sampling



- Basic-Block-Sampling
  - Functional simulate
  - Predict the simulation time of warps using the predicted time of each type of basic block.
- Warp-sampling
  - Predict the simulation of warps using the average warp runtime
  - Only the warp scheduler enables

Experiments

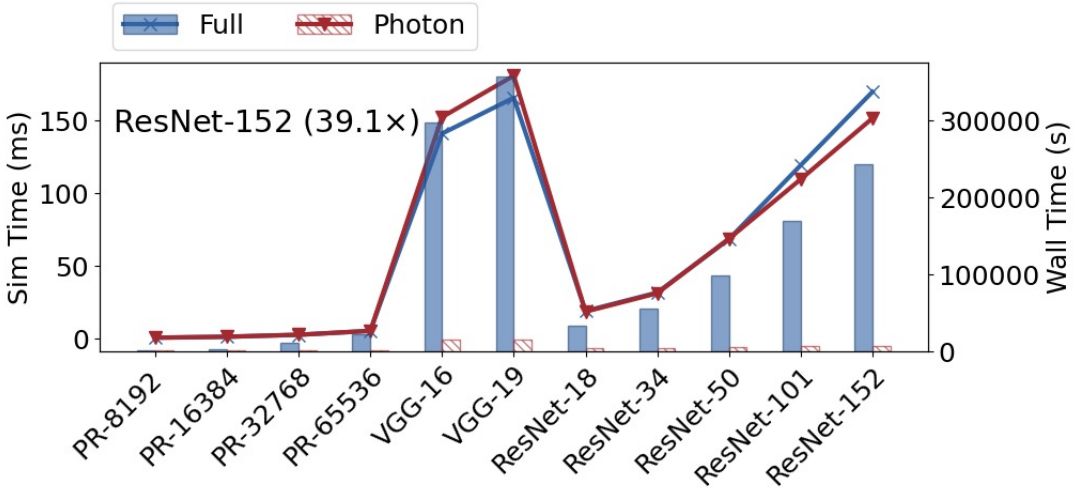
- Photon achieves up to 24.65 × speedup (average speedup 1.87 ×) with an average simulation error of 6.83%



kernel execution time (left y-aixs with lines); Wall time (right y-axis with bars)

Experiments

- Photon reduces the simulation time needed to perform one inference of ResNet-152 with batch size 1 from 7.05 days to just 1.7 hours with a low sampling error of 10.7%.



kernel execution time (left y-aixs with lines); Wall time (right y-axis with bars)

Photon: A Fine-grained Sampled Simulation Methodology for GPU Workloads

Changxi Liu[1], Yifan Sun[2], Trevor E. Carlson[1]

[1]National University of Singapore

[2]College of William & Mary