Preemptive Thread Block Scheduling with Online Structural Runtime Prediction for Concurrent GPGPU Kernels
Sreepathi Pai¹, R. Govindarajan² and Matthew J. Thazhuthaveetil²
¹The University of Texas at Austin  ²Indian Institute of Science, Bangalore

Concurrent Kernels
Current GPUs support concurrent execution of kernels
- GPU Concurrency is space-sharing not time-sharing
- Works only for small kernels
- Large kernels execute serially in FIFO order

The Problem with FIFO Scheduling
As an example, consider FIFO scheduling on 2-program workloads
- FIFO behaves like Shortest Job First (SJF) or Longest Job First (LJF)
- FIFO throughput is lower (average 15%) than SJF

Replacing FIFO with SRTF
- GPU Kernels execute as discrete grids of thread blocks
- Thread blocks are independent units of execution
- Kernels can be pre-empted easily at thread block boundaries
But SRTF requires runtime of kernels.

Predicting Runtimes by Harnessing Regularity
SGEMM on one Streaming Multiprocessor of a GPU

Structural Runtime Prediction
Structural Prediction treats the execution of n thread blocks as n repeated executions of the same code.
- Observe runtime for first few thread blocks of kernel
- Use model and observations to predict runtime for kernel

Simple Slicing Predictor
$$\text{Pred. Cycles} = \frac{\text{Active Kernel Cycles} + (\text{Total Blocks} – \text{Done Blocks}) \times t}{\text{Resident Blocks}}$$
where:
- Active Kernel Cycles: executed cycles
- t: duration of thread block
- Total Blocks: total thread blocks
- Done Blocks: blocks until now
- Resident Blocks: kernel residency
- Pred. Cycles: prediction

Accuracy of Predictor
For single-gpu, the predictions are between 0.48x to 1.08x of actual runtime after observing 1 thread block.

SRTF Thread Block Scheduler (TBS)
Our TBS samples t for waiting kernels, predicts their runtime and pre-empts running kernel per SRTF policy.

Results on all 2-program ERCBench workloads

Conclusion
- SRTF is superior in terms of system throughput, turnaround time and fairness.
- SRTF improved STP by 1.18x and ANTT by 2.25x compared to FIFO
- SRTF also outperformed resource-sharing MPMax by 1.16x (STP) and 1.3x (ANTT).
- SRTF bridges 49% of the gap between FIFO and SJF, approaching to within 12.64% of SJF’s throughput.