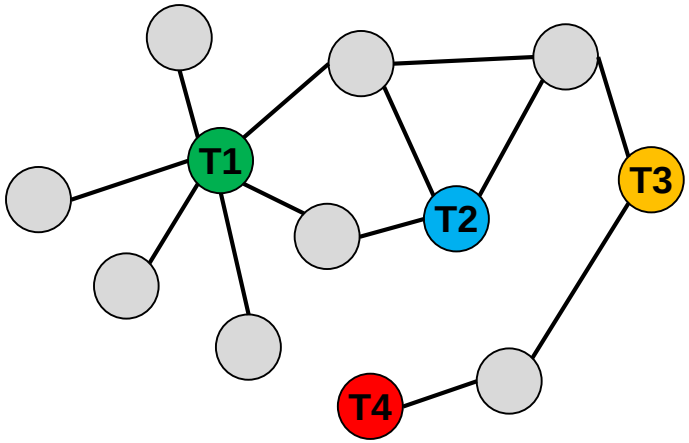
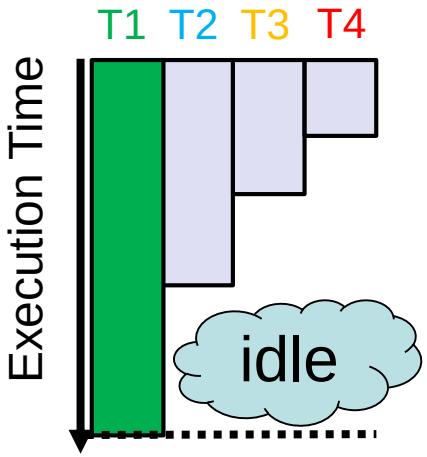
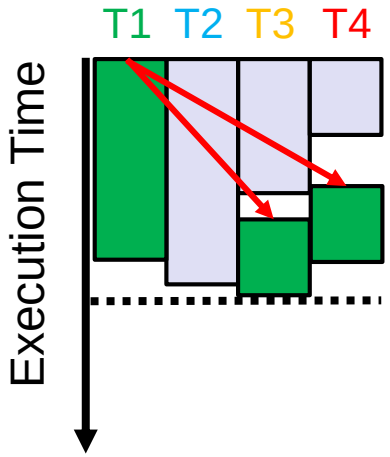
Controlled Kernel Launch for Dynamic Parallelism in GPUs

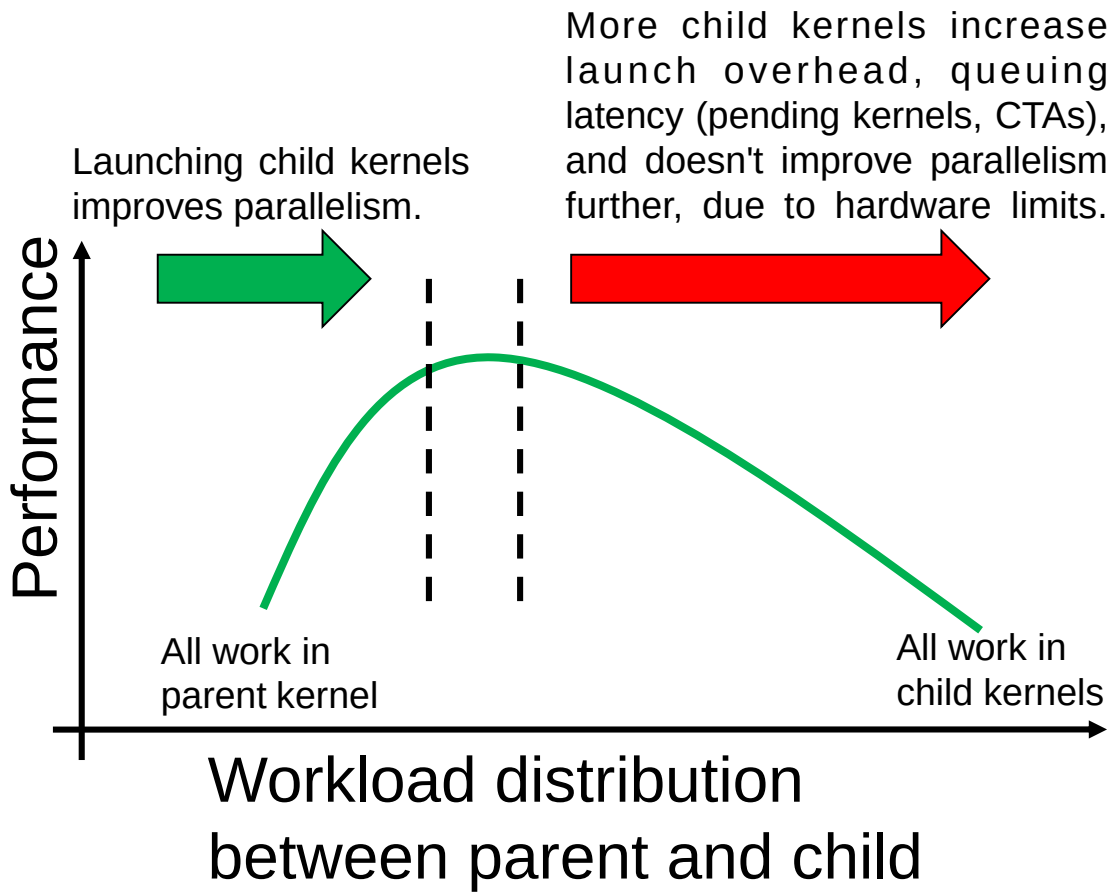
Xulong Tang, Ashutosh Pattnaik, Huaipan Jiang,
Onur Kayiran, Adwait Jog, Sreepathi Pai, Mohamed Ibrahim,
Mahmut T. Kandemir, Chita R. Das







Dynamic Parallelism in GPUs



Dynamic Parallelism in GPUs

Our proposal: SPAWN, a runtime framework, which dynamically decides how many child kernels to launch by taking into account the hardware parallelism limitations. SPAWN reduces launch overhead and queuing latency, and improves resource utilization and performance.

Performance Improvement: 69% over non-DP (All parent) and 57% over baseline-DP (Most in Child).

Session 9B: Wednesday, 12:35pm,
Salon G – 6th Floor