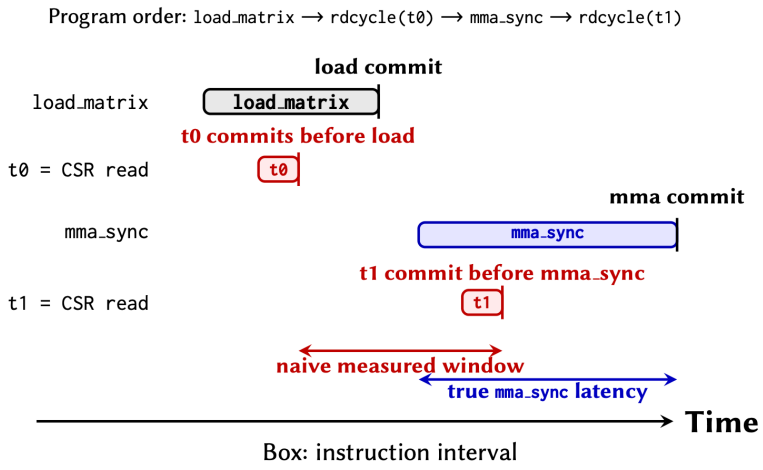
CycleFence: Precise Cycle-Level Profiling for RISC-V GPUs

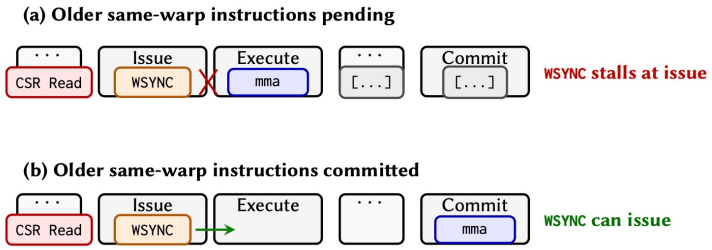
Xinle Song, Blaise Tine

Problem Statement:
Accurately profiling AI workloads on RISC-V GPUs is challenging because naive cycle-counter reads can **misalign** with long-latency operations, causing measured latency to deviate from true execution time.



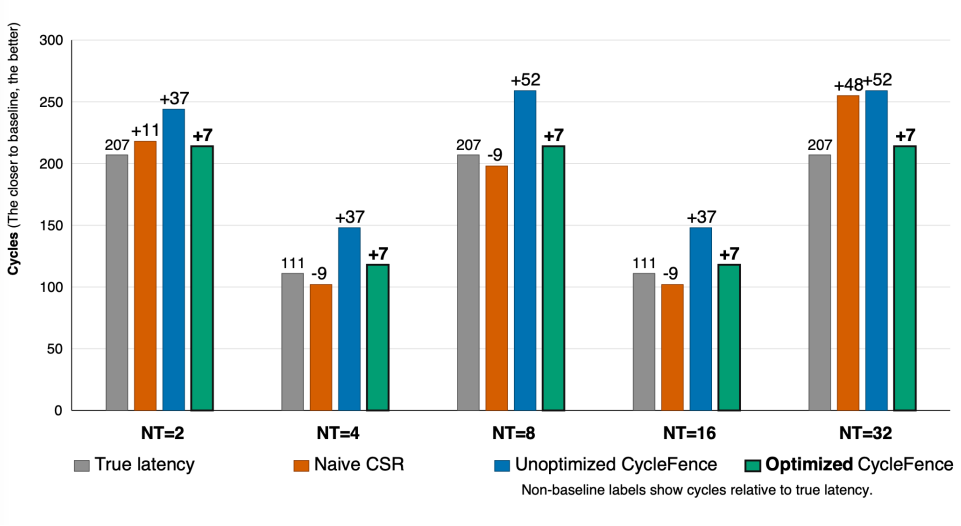
Methodology:

- **Warp-local serialization:** WSYNC stalls the issuing warp until all older same-warp instructions have committed.
- **Low-overhead timestamping:** Branch-free begin(), end(), and diff() intrinsics anchor timestamp reads to the true profiling window while minimizing overhead.



Results:

- **Naive CSR is unreliable:** timestamps can drift around long-latency `mma` operations.
- **WSYNC fixes the inaccuracy:** timestamps wait for older instructions to commit.
- **Optimized CycleFence is precise:** consistent **+7 cycle overhead**.
- **Much lower overhead:** avoids the unoptimized path's **37–52 cycle** extra cost.



UCLA Samueli School of Engineering