
**GPU’s performance is dominated by critical warp’s performance**

![Normalized Execution Time Graph]

- **Warps in a Thread Block**
  - **Original**
  - **Ideal**

CAWA Criticality-Aware Warp Acceleration

Coordinated Warp Scheduling and Cache Prioritization

Which warp is the critical warp?

CPL Criticality Prediction Logic

How to accelerate critical warp execution?

- gCAWS: greedy Criticality Aware Warp Scheduling
- CACP: Criticality Aware Cache Prioritization
CAWA: Coordinated Warp Scheduling and Cache Prioritization for Critical Warp Acceleration of GPGPU Workloads

Shin-Ying Lee  Akhil Arunkumar  Carole-Jean Wu
Arizona State University