Exploiting GPU Peak-power and Performance Tradeoffs through Reduced Effective Pipeline Latency

Syed Gilani (AMD)
Nam Sung Kim (UW-Madison)
Michael Schulte (AMD Research)
Problem

- Thread-level parallelism (TLP) limited by
  - Thread synchronization patterns
  - Memory access patterns
  - Data dependencies
  - Limited hardware resources

- Low TLP exposes pipeline latencies
  - Data-forwarding networks are power hungry
Contributions

• Limited forwarding for a few recently executed instructions

• Reduce impact of pipeline latency on performance
  • Low voltage pipelines with negligible impact on performance

• Mean speedups of 23% (SP/Int) and 33% (DP) within the same power-budget