Energy Efficient GPU Transactional Memory via Space-Time Optimizations

Wilson W. L. Fung
wwlfung@ece.ubc.ca

Tor M. Aamodt
aamodt@ece.ubc.ca

University of British Columbia

TM on GPU: Energy Concern

- TM on GPU: Simple Irregular Parallelism on GPUs
  - nBody 5M Bodies
  - Difficulty: Undergrad, PhD
  - Risk: Medium, High
  - TM: Post-Grad, Medium

- Aggregate Control Messages
  - TM Reduces Risk in Software Development:
    - Predictable Dev Time
    - No Deadlocks
    - More Maintainable Code

- 1640s 5.2s
- Key Idea: Use globally synchronous on-chip timers to record when each word in memory is last written.

Warp Level Transaction Management

- Key Idea: Manage Transactions in a Warp as a Whole Entity
  - Enables exploits of Spatial Locality:
    - Scalar Control Messages
      - SIMT Core TX1, TX4
      - CU
    - Aggregated Messages
      - SIMT Core TX1, TX2, TX3, TX4
      - Last Level Cache

- Intra-Warp Conflict Resolution
  - Naive Conflict Resolution
    - O(T^2 x (R+W)^2)
  - 2-Phase Parallel Conflict Resolution
    - Insight: Fixed priority for conflict resolution enables parallel resolution

- Spatial Locality
  - Owner ID < My ID: Abort
  - Owner ID = My ID: Pass
  - Owner ID = NULL: Pass

- Stored in Shared Memory (On-Chip Per-Core Scratchpad)

Temporal Conflict Detection

- Motivation: Skip Value-Based Conflict Detection for Conflict-Free Read-Only Transactions
  - Examples:
    - TX1
      - if (C = 0)
      - B = B + 1;
    - TX2
      - int K;
      - K = X + Y;

- Key Idea: Use globally synchronous on-chip timers to record when each word in memory is last written.

GPGPU-Sim 3.2.1 + GPUWatch

- HT-(H/M/L) – Hash Table Construction
- ATM – Bank Transactions
- BH-(H/L) – Barnes Huts (N-Body)
- CL/CLc – Cloth Simulation
- CC – Maxflow/Mincut Graph
- AP – Data Mining

- Results
  - Energy Usage
    - 2X → 1.3X