Application Slowdown Model: Quantifying and Controlling Impact of Interference at Shared Caches and Main Memory

Lavanya Subramanian, Vivek Seshadri, Arnab Ghosh, Samira Khan, Onur Mutlu
Carnegie Mellon University, Intel Labs, University of Virginia

Problem: Shared Resource Interference
Impact of Shared Resource Interference
Our Goal
Provide high and predictable performance in the presence of shared resource interference

Our Approach
1. Build a model to estimate slowdowns
2. Leverage our model for slowdown-aware resource management

Observation: Proxy for Performance
For a memory bound application, Performance ∝ Cache access rate

Challenge: Estimating Cache Access Rate Alone
Minimize memory bandwidth contention: Using priority (Subramanian et al., HPCA 2013)
Challenge: Estimating Cache Access Rate Alone
Quantify shared cache capacity contention: Using auxiliary tag stores (Pomerene et al., 1989)

Our Goal
For a memory bound application, Allocation memory bandwidth proportional to slowdowns

Leveraging the Application Slowdown Model
Slowdown-aware cache capacity partitioning
Coordinated Resource Allocation Schemes
Providing Slowdown Guarantees
• Cache allocation with the goal of meeting a slowdown bound
• Allocate just enough cache space to critical application
• Allocate remaining cache space to other applications

Previous work: Reduce miss counts;
Our proposal: Reduce slowdowns

Slowdown-aware memory bandwidth partitioning

Allocation memory bandwidth proportional to slowdowns

Significant fairness benefits across different channel counts