A Case for Interconnect-Aware Architectures

Naveen Muralimanohar, Rajeev Balasubramonian
School of Computing, University of Utah

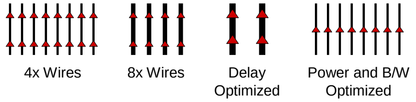
Work supported in part by NSF award CCF-0430063 and NSF CAREER award CCF-0545959

Abstract

In future multi-core chips, a large fraction of chip area will be dedicated for cache hierarchies and interconnects between various cache components. Large caches will likely be partitioned into many banks, connected by an on-chip network. We show results for a design space exploration of non-uniform cache architecture (NUCA) organizations. We make a case for a heterogeneous interconnect architecture where each link is composed of different wire types.

Key Observations

- Network parameters (wire/router type) have a major impact on cache access latency
- Parts of a message are more important than others
- Wires can be designed in a variety of ways

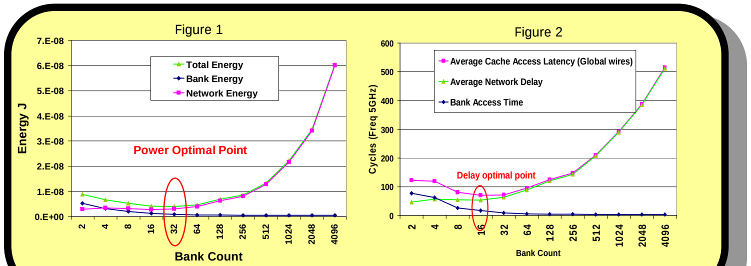


Contributions

Design Space Exploration: CACTI tool extensions to model average network and bank delay/power for each cache bank size

Cache Access Optimizations: Low latency wires carry a subset of the address to initiate prefetch out of a cache bank

Hybrid Architectures: The address network employs low-latency wires and a combination of buses and point-to-point links

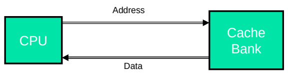


Design space exploration results – energy and latency as a function of number of banks

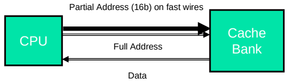
Leveraging Interconnect Choices for Performance Optimizations

Early Look-up

A partial address is sent on fast wires to start indexing into the cache while the remaining address is sent on slower wires



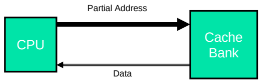
Access Time = Address transmission + Bank access + Data transmission



Partial address is enough to initiate bank look-up
Transmission of full address and bank look-up happen in parallel

Aggressive Look-up

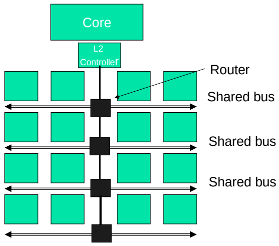
The partial address is sent on fast wires to perform a partial tag match and aggressively send all matched blocks to the cache controller



Cache access with partial tag match

Hybrid Network

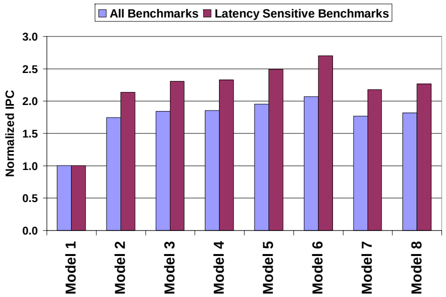
The aggressive look-up approach with low-latency wires allows the address network to employ fewer routers – a bus broadcasts the address to all banks in a row. The data network continues to employ a grid network.



Results

Models Simulated

Model 1 – 512 banks, B-wires
Design Space Exploration: Model 2 – 16 banks, B-wires
Model 3 – 16 banks, B,L-wires, Early Lookup
Model 4 – 16 banks, B,L-wires, Aggressive Lookup
Hybrid Model: Model 5 – 16 banks, B,L-wires
Model 6 – 1 cycle address transfer (upper bound)
Model 7 – 16 banks, L-wires for both add. and data
Model 8 – 16 banks, Address transfer on L-wires



Methodology

Simplescalar simulator, SPEC2k benchmarks
Network parameters: grid topology, virtual channel flow control, partially adaptive network, four VCs per physical channel, router pipeline delay of three cycles
Wire model: ITRS 2005, analytical equations (Ho et al.)

Future Work

- Comprehensive tool to identify optimal cache organizations
  - explore various link and router types
  - models for network contention (especially in CMPs)
- Exploiting heterogeneity within routers

For more details and related work (HPCA 2005, ISCA 2006), visit:
Draft paper: http://www.cs.utah.edu/~rajeev/pubs/draft06.pdf
Related papers: http://www.cs.utah.edu/~rajeev/research.html