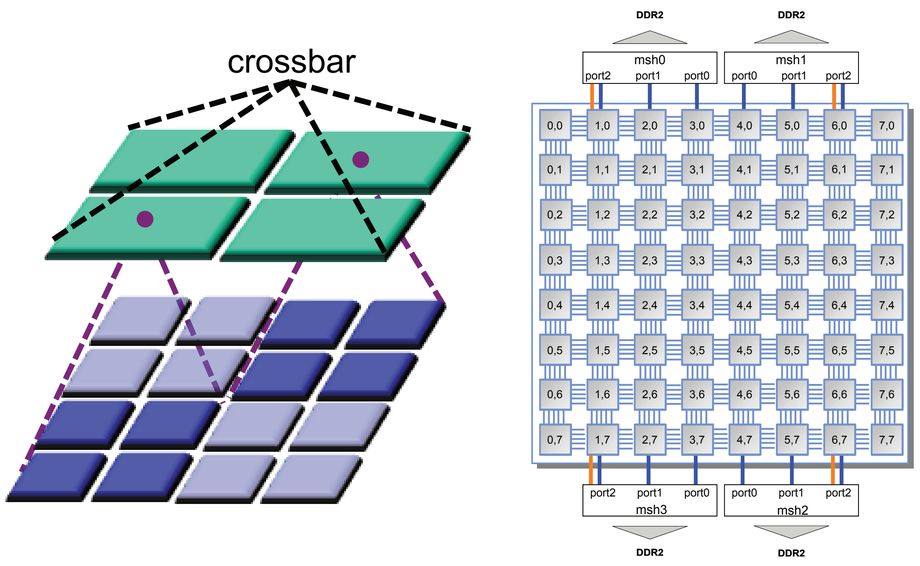
ANUMA: Asymmetry-aware Execution Placement on Manycore Chips

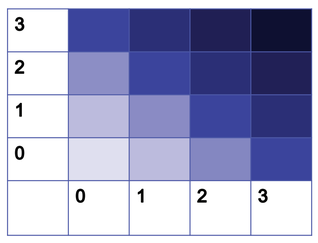
Alexey Tumanov, Joshua Wise, Onur Mutlu, Greg Ganger (CMU)

MULTIPLE CORES & MEMORY CONTROLLERS ON CHIP

- Moore's Law continues: more transistors on a chip
- Need more MCs to maintain memory bw/core
- How do we position memory controllers?
- Latency considerations:
  - › NUMA: dominated by local vs. remote latency
  - › ANUMA: gradual continuum of access latencies



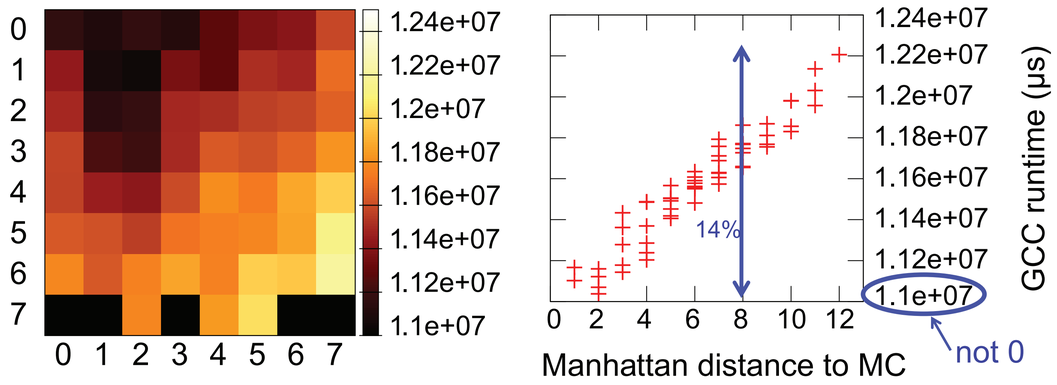
hierarchical vs. Network-on-Chip positioning of cores and memory controllers



latency drop off for ANUMA memory controller architecture

MOTIVATING EXPERIMENTS

- Observed 14% ping-pong latency differential
- MC access latency increases with Manhattan distance

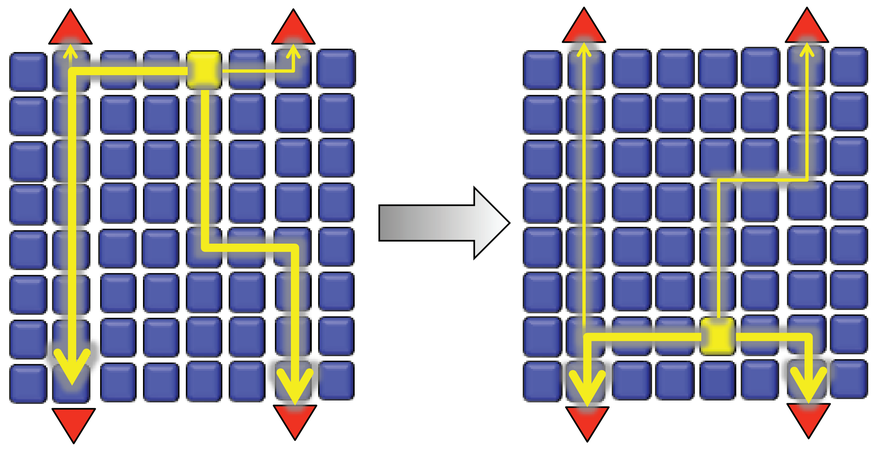


FUTURE WORK

- More experimentation to understand benefits
  - › As a function of application properties
- ANUMCA: considering cache affinity
- Optimizing for application-level goals
  - › SLAs, performance goals
  - › Multi-threaded applications

PROBLEM STATEMENT

- MC access asymmetries result in bad placement
- How do we detect it?
- How do we fix it?



THREAD PLACEMENT SOLUTION

MC access tracking

- Page fault on memory access
- Each fault is attributed to specific MCs
- Adjustable sampling frequency (20Hz)
  - › Trades off overhead vs. accuracy

Thread placement

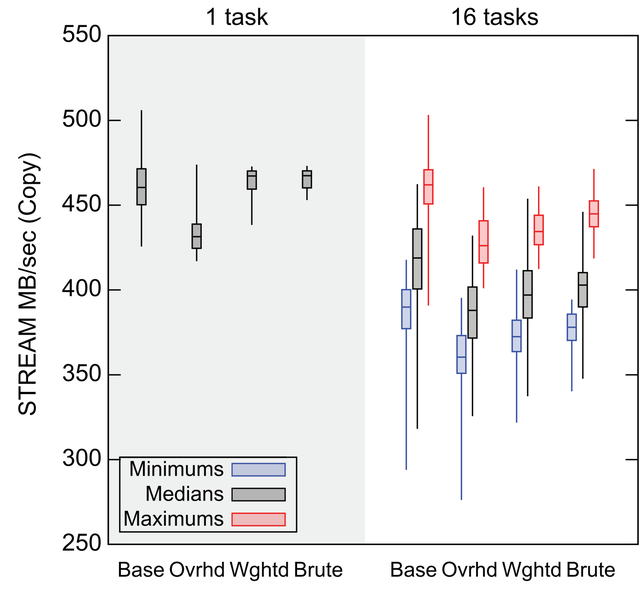
- Place threads nearest to MCs they use
  - › Greedy order w.r.t. memory intensity
- Weighted geometric placement algorithm

EXPERIMENTAL SETUP & RESULTS

- TilePro64, 8x8 board, standard linux host
- Placement policies compared:
  - › Base: linux scheduler
  - › Overhead: base + stats collection
  - › Weighted: our approach
  - › Brute: brute force approach

Result Takeaways

- Asymmetry-aware placement helps
- Stats collection overhead is high
  - › Calls for per-core, per-MC stats counters



Reference
Alexey Tumanov, Joshua Wise, Onur Mutlu, Greg Ganger. "Asymmetry-aware execution placement on manycore chips". In Proc. of SFMA'13, EuroSys'13, Apr 14-17, 2013.

Intel Science & Technology Center for Cloud Computing
NSERC CRSNG
Carnegie Mellon University
Georgia Tech
intel
PRINCETON UNIVERSITY
UC Berkeley
UNIVERSITY of WASHINGTON