NUMA Instrumentation Challenges

William C. Brantley, PhD

Advanced Micro Devices

Bill.Brantley at AMD.com

How can a parallel application's performance on a NUMA be improved by the programmer? by the compiler? by the operating system? What instrumentation is available and needed to better exploit NUMA? Mechanisms available in existing systems will be reviewed.

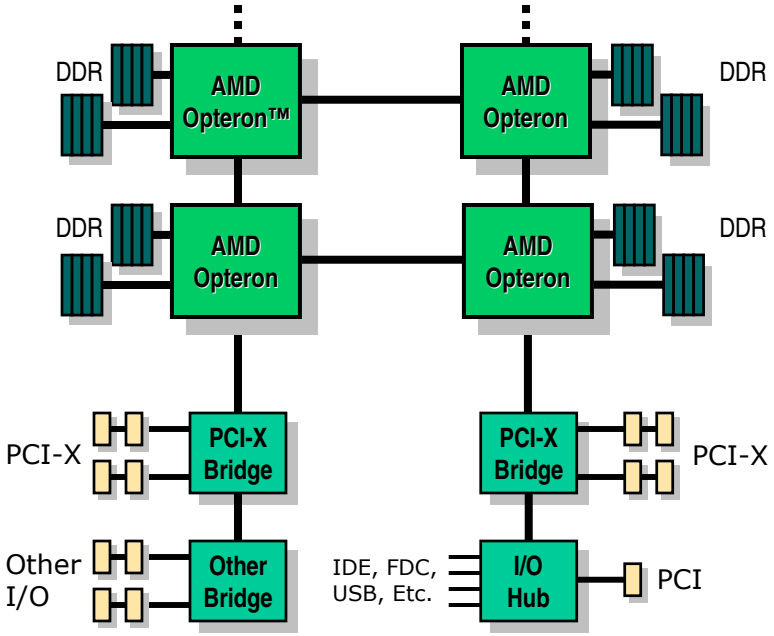
Overview

- NUMA & Memory Latency Driven
  - NUMAs are common today – Opteron
  - NUMA effect on performance
  - Measuring/locating improvements
  - Need latency sampling

- List a few other topics I hope are covered today

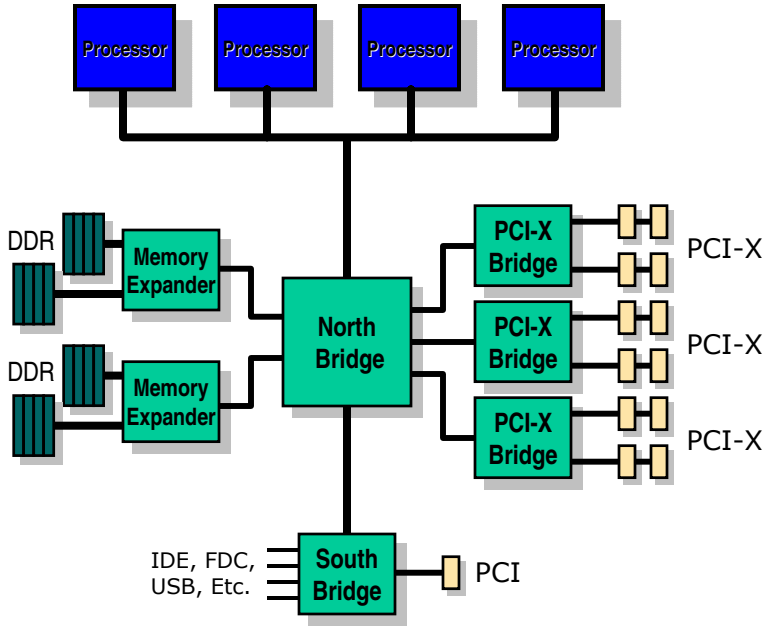
AMD Opteron™ Processor MP System Architecture

AMD Opteron™ System



- Up to 8 processors without glue logic
- Each processor adds memory
- Each processor adds additional HyperTransport™ buses for PCI-X and other I/O bridges
- Fewer chips required

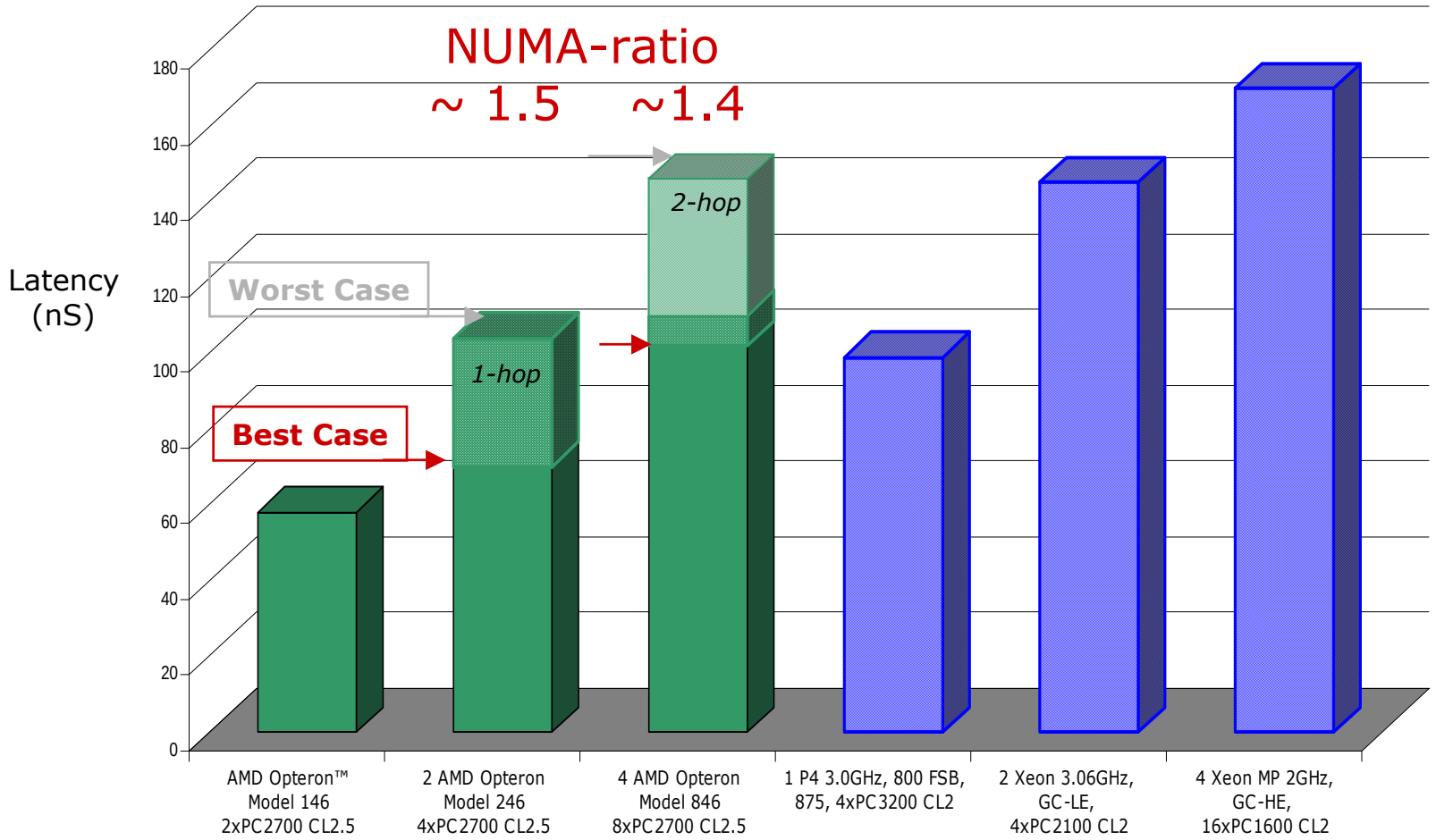
Typical MP System



- Maximum of 4 processors
  - Processors compete for FSB bandwidth
- Memory size and bandwidth are limited
- Maximum of 3 PCI-X bridges
- More chips required

Low Memory Latency

ScienceMark 2.0 Beta, 512-Byte Stride



All benchmarks run on Microsoft® Windows® Server 2003 Enterprise Edition in 2003.

NUMA Effect on Performance

- Variation in run-times
  - But, remember even worst case is better than bus.
- Mask effect by interleaving in hardware or OS
  - Average avoids peaks and valleys -- mediocre for all
- How to exploit NUMA?
  - OS allocate data local to process
    - With allocate-on-first-touch policy

      => initialize data in same thread that uses the data later.

What Could Be Measured?

- Indirect measures:
  - Run time variation
  - Local & remote memory reference counts
  - Reference counter/CPU/cache line (e.g., SGI Origin® 2000)
    - Enables OS page migration, but too much overhead to break even
  - Average memory latency = (cycles of total latency)/(references)
  - Ability to get physical location of a process' pages (Per Ekman patch)
  - Reference trace + post-process
- Direct measures:
  - Sampled memory reference latency
    - E.g., Itanium2®

Viewing Sampled Latency

- Latency profile by data address/object
  - Which objects should be moved?
  - Only target address & latency needed
  - Could help find false sharing
- Latency profile by instruction address
  - Which part of my program is being delayed?
  - Difficult if IA not captured with target address & latency – (interrupt occurs much after event)

Itanium2® Memory Latency Sampling

- Event Address Registers for I/D Misses
  - Instruction & data miss events
  - Both instruction & data addresses captured
- Event filtering
  - Instruction address range or opcode
  - Data address range
  - Latency threshold by powers of 2
    - Allows measuring just L3 misses

Publications Addressing Latency, NUMA, and Instrumentation

- Doug O'Flaherty, Michael Goddard, "AMD Opteron® Processor Benchmarking for Clustered Systems", http://www.sun.com/amd/39497A_HPC_WhitePaper_2xCli.pdf, 7/15/2003.
- Nathan Robertson, Alistair P. Rendell: "OpenMP and NUMA Architectures I: Investigating Memory Placement on the SGI Origin® 3000." International Conference on Computational Science 2003: 648-656
- Bryan R. Buck, Jeffrey K. Hollingsworth: "Using Hardware Performance Monitors to Isolate Memory Bottlenecks." Conf. on High Performance Networking and Computing, Proc. 2000 ACM/IEEE conf. on Supercomputing, 2000.
- Mustafa M. Tikir, Jeffrey K. Hollingsworth: "Using Hardware Counters to Automatically Improve Memory Performance." Proceedings of SC'04, Nov. 2004.
- Bryan R. Buck, Jeffrey K. Hollingsworth: "Data Centric Cache Measurement on the Intel® Itanium® 2 Processor." Proceedings of SC'04, Nov. 2004.
- "Intel® Itanium® 2 Processor Reference Manual For Software Development and Optimization." Intel order number 251110-003 , May 2004.

Itanium2® Memory Latency Experience?

- How effective is Itanium2's latency sampling?
- Audience feedback?

Virtualization

- E.g., Vmware & Xen virtual machine monitors
- Secure Computing Platform
  - e.g., La Grande & Presidio
- Importance increasing
- PMU must be virtualized
- How to measure virtual machine monitor?

Trademarks

AMD®, the AMD Arrow logo, and combinations thereof, Opteron® and AMD-K8 are trademarks of Advanced Micro Devices, Inc.

HyperTransport is a licensed trademark of the HyperTransport Consortium.

Itanium®, Itanium2® are registered trademarks of Intel.

Origin® 2000 is a registered trademark of Silicon Graphics, Inc.

Other names used are for identification purposes only and may be tradesmarks of their respective owners.