

HPC Hardware Overview

Karl W. Schulz

Texas Advanced Computing Center
The University of Texas at Austin

UT/Portugal Summer Institute Training
Coimbra, Portugal
July 14, 2008



THE UNIVERSITY OF TEXAS AT AUSTIN
TEXAS ADVANCED COMPUTING CENTER

Summer Institute Outline

- **Monday**
 - 10:30 - 11am: Welcome & Introductions
 - 11:00 - 12pm: ~~Introduction to Parallel Computing~~
 - 12:00 - 13:00: HPC Systems Overview
 - 13:00 - 14:30: Lunch
 - 14:30 - 16:00: Introduction to MPI Programming
 - 16:00 - 16:30 Break
 - 16:30 - 17:30 TACC HPC Systems User Environment
 - 17:30 - 18:30 Lab login exercises
- **Tuesday**
 - 09:00 - 10:30 Advanced MPI
 - 10:30 - 11:00: Break
 - 11:00 - 12:00: Performance Optimization
 - 12:00 - 13:00: Lab exercises
 - 13:00 - 14:30: Lunch
 - 14:30 - 16:00 Introduction to Scientific Visualization
 - 16:00 - 16:30 Break
 - 16:30 - 17:30 TACC Visualization Systems
 - 17:30 - 18:30 Lab exercises



Summer Institute Outline

- **Wednesday**
 - 09:00 - 10:30: Programming with OpenMP
 - 10:30 - 11:00: Break
 - 11:00 - 12:00: High Throughput Computing
 - 12:00 - 13:00: Lab exercises
 - 13:00 - 14:30: Lunch
 - 14:30 - 16:00 Portuguese research presentations
 - 16:00 - 16:30 Break
 - 16:30 - 18:30 Portuguese research presentations
- **Thursday**
 - 09:00 - 10:30: Scalability Optimization & Parallel Libraries
 - 10:30 - 11:00: Break
 - 11:00 - 12:00: Debugging Parallel Applications
 - 12:00 - 13:00: Lab exercises
 - 13:00 - 14:30: Lunch
 - 14:30 - 16:00 Advanced Visualization Techniques
 - 16:00 - 16:30 Break
 - 16:30 - 17:30 Advanced Visualization Techniques
 - 17:30 - 18:30 Lab exercises



Outline

- TACC's **Lonestar** System
 - Dell blade-based system
 - InfiniBand (1st generation)
 - Intel Processors
- TACC's **Ranger** System
 - Sun blade-based system
 - InfiniBand (2nd generation)
 - AMD Processors



General Preliminary Comments

- We are going to discuss hardware and user environments in the context of systems at TACC, but the ideas are of general applicability for a number of HPC systems
- As an application programmer, you may not care about the nuts and bolts of processors/interconnect design
 - we feel your pain 😊, but we have to think about it slightly to maximize performance
 - we'll try to point out the main relevant architecture bits to keep in the back of your mind
- Please feel free to ask questions – we generally don't bite

•We provide PDF from each talk for distribution to those who are interested at the end of the day/week



Lonestar Cluster Overview



Outline

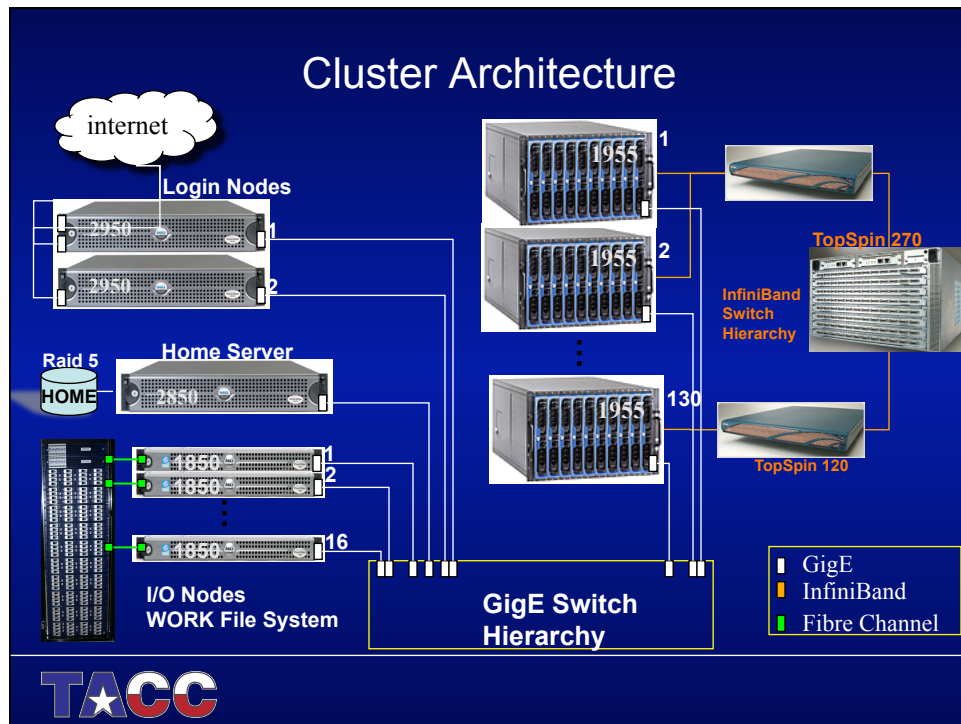
- Lonestar Cluster
 - Configuration & Diagram
 - Server Blades
- Dell PowerEdge 1955 Blade (Intel Dual-Core) Server Nodes
- 64-bit Technology
- Microprocessor Architecture Features
 - Instruction Pipeline
 - Speeds and Feeds
 - Block Diagram
- Node Interconnect
 - Hierarchy
 - InfiniBand Switch and Adapters
 - Performance



Lonestar Cluster Overview

Hardware	Components	Characteristics
Compute Nodes Dell 1955	1,300 Nodes 5,200 Cores	2.66 GHz 4MB/Cache 8GB Mem/node
WORK File System I/O Nodes Dell 2850	24 I/O Nodes Lustre File System	100 TB
Login	1 login: lonestar	2.66 GHz, 16GB Mem
Development	20 Nodes (dev. queue)	2.66 GHz, 8GB/node
Interconnect (MPI) InfiniBand (TopSpin)	24-port leafs 96-port cores	1GB/sec P-2-P Fat Tree Topology
Ethernet (GigE)		128 MB/sec P-2-P Fat Tree Topology







Compute Node

Dell PowerEdge 1955

10 blades

7U

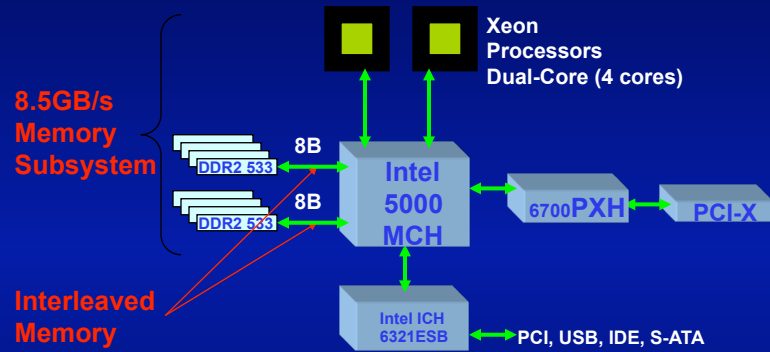
Blade

Blade

Blade:	Intel® Xeon Core Technology 4 cores per blade
Chipset:	Intel 5000P Chipset
Memory:	8GB 2:1 memory interleave (533MHz DDR2)
FSB:	1333MHz (Front Side Bus)
Cache:	4MB L2 "Smart" Cache

TACC

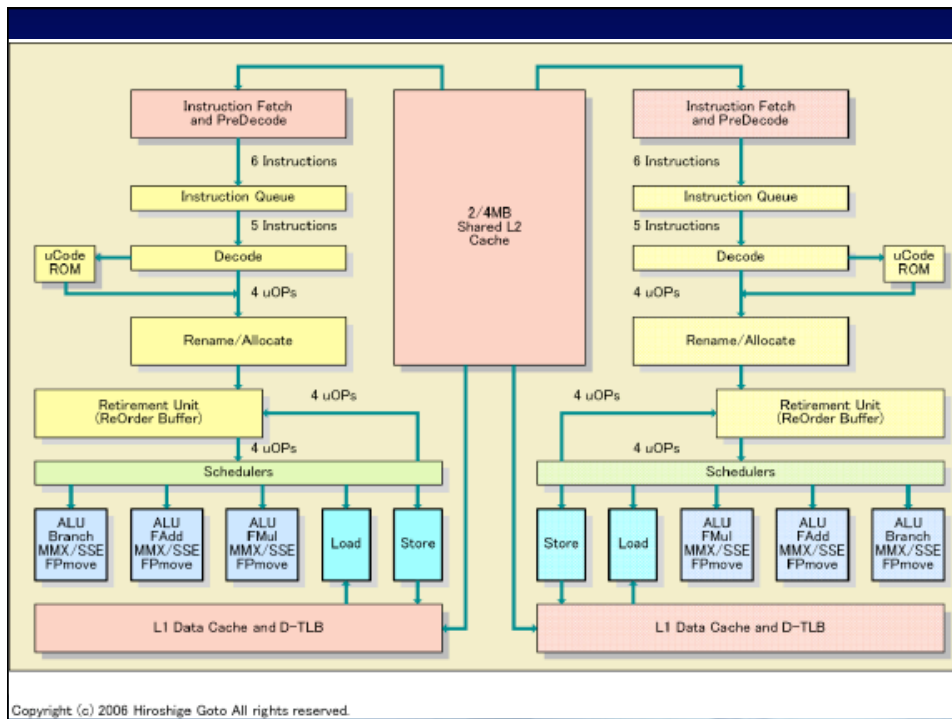
Motherboard



FSB: 1333 MHz

Memory: dual-channel, Fully Buffered 533 MHz memory

Bandwidth: 8.5GB/s Peak Throughput (10.7GB/s Front Side Bus)



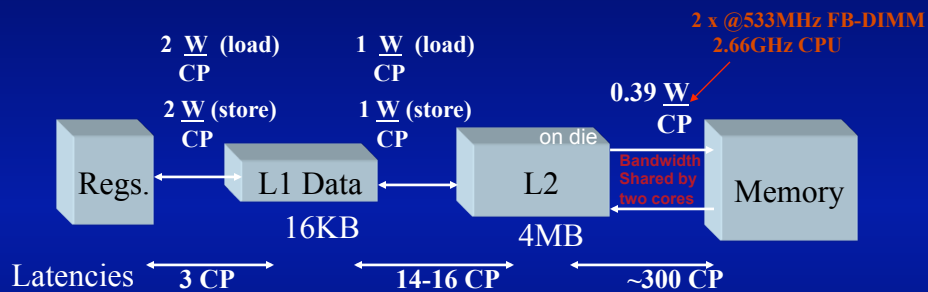
Copyright (c) 2006 Hiroshige Goto All rights reserved.

Intel Core μ Architecture Features

- Intel Core Microarchitecture (**Dual-Core MultiProcessing**)
- L1 Instruction Cache
- 14 Segment Instruction Pipeline
- Out-of-Order execution engine (**Register Renaming**)
- Double-pumped Arithmetic Logic Unit (**2 Int Ops/CP**)
- Low Latency Caches (**L1 access in 3 CP**, HW Prefetch)
- Hardware Prefetch (within a single page)
- SSE2/3/4 [Streaming SIMD Extension 2/3/4] (**4 FLOPs/CP**)



Speeds & Feeds (Xeon/Pentium 4)



4 FLOPS/CP

Cache Line size L1/L2 = 8W/8W

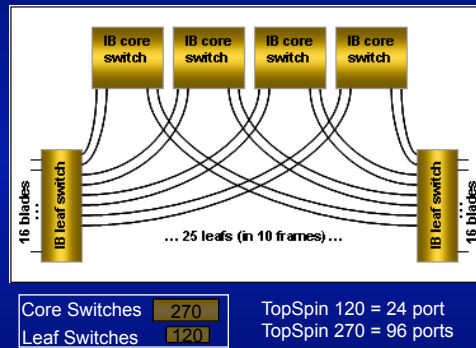
W FP Word (64 bit)
CP Clock Period



Interconnect Architecture

Network Hierarchy

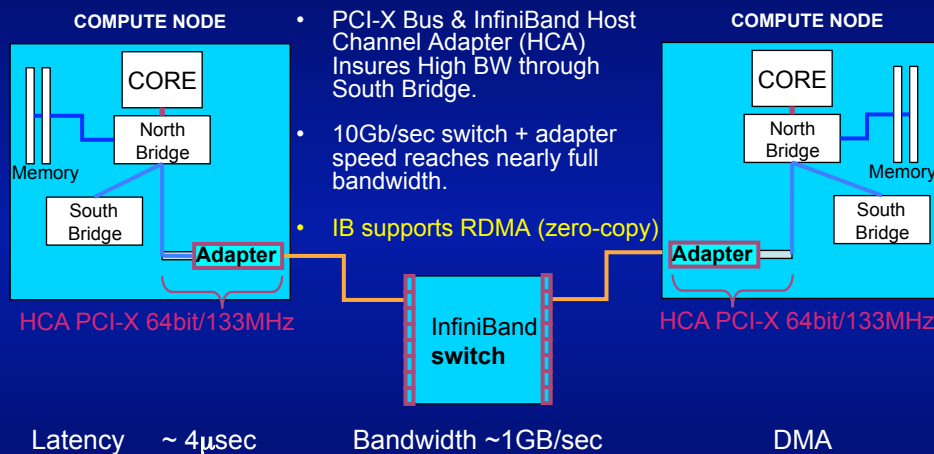
Lonestar InfiniBand Topology



Fat Tree Topology
2-1 Oversubscription

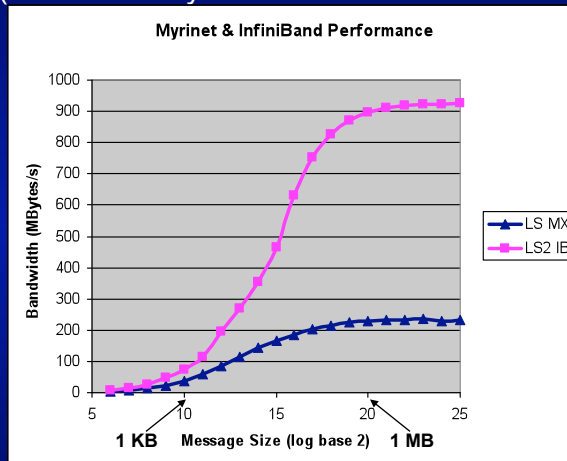


Interconnect Architecture



Interconnect Architecture

(InfiniBand/Myrinet Bandwidth on Lonestar)



HPC Hardware Overview

Ranger: AMD Quad-core System



THE UNIVERSITY OF TEXAS AT AUSTIN
TEXAS ADVANCED COMPUTING CENTER

Ranger: Introduction

- Ranger is a unique instrument for computational scientific research housed at TACC's new machine room
- Results from over 2 ½ years of initial planning and deployment efforts
- Funded by the National Science Foundation as part of a unique program to reinvigorate High Performance Computing in the United States (*Office of Cyberinfrastructure*)



ranger.tacc.utexas.edu



How Much Did it Cost and Who's Involved?

- TACC selected for very first NSF 'Track2' HPC system
 - \$30M system acquisition
 - Sun Microsystems is the vendor
 - **Very Large InfiniBand Installation**
 - ~4100 endpoint hosts
 - >1350 MT47396 switches
- TACC, ICES, Cornell Theory Center, Arizona State HPCI are teamed to operate/support the system four 4 years (\$29M)



Ranger: Performance

- Ranger debuted at #4 on the Top 500 list
- Lonestar debuted at #12 (currently ranked #38)



1	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband
2	BlueGene/L - eServer Blue Gene Solution
3	Blue Gene/P Solution
4	Ranger - SunBlade x6420, Opteron Quad 2Ghz, Infiniband
5	Jaguar - Cray XT4 QuadCore 2.1 GHz



Ranger Hardware Summary

- **Compute power - 579 Teraflops**
 - 3,936 Sun four-socket blades
 - 15,744 AMD "Barcelona" processors
 - Quad-core, four flops/cycle (dual pipelines)
- **Memory - 123 Terabytes**
 - 2 GB/core, 32 GB/node
 - ~20 GB/sec memory B/W per node
- **Disk subsystem - 1.7 Petabytes**
 - 72 Sun x4500 "Thumper" I/O servers, 24TB each
 - 40 GB/sec total aggregate I/O bandwidth
 - 1 PB raw capacity in largest filesystem
- **Interconnect - 10 Gbps /1.6 – 2.9 μ sec latency**
 - Sun InfiniBand-based switches (2), up to 3456 4x ports each
 - Full non-blocking 7-stage Clos fabric
 - Mellanox ConnectX InfiniBand (second generation)



Ranger Hardware Summary (cont.)

- **25 Management servers - Sun 4-socket x4600s**
 - 4 Login servers, quad-core processors
 - 1 Rocks master, contains software stack for nodes
 - 2 SGE servers, primary batch server and backup
 - 2 Sun Connection Management servers, monitors hardware
 - 2 InfiniBand Subnet Managers, primary and backup
 - 6 Lustre Meta-Data Servers, enabled with failover
 - 4 Archive data-movers, move data to tape library
 - 4 GridFTP servers, external multi-stream transfer
- **Ethernet Networking - 10Gbps Connectivity**
 - Two external 10GigE networks: TeraGrid, NLR
 - 10GigE fabric for login, data-mover and GridFTP nodes, integrated into existing TACC network infrastructure
 - Force10 S2410P and E1200 switches



InfiniBand Cabling for Ranger

- Sun switch design with reduced cable count, manageable, but still a challenge to cable
 - **1312 InfiniBand** 12x to 12x cables
 - 78 InfiniBand 12x to three 4x splitter cables
 - Cable lengths range from 7-16m, average 11m
- 9.3 miles of InfiniBand cable total (15.4)



Ranger Space, Power and Cooling

- **System Power:** 3.0 MW total
- **System:** 2.4 MW
 - ~90 racks, in 6 row arrangement
 - ~100 in-row cooling units
 - ~4000 ft² total footprint
- **Cooling:** ~0.6 MW
 - In-row units fed by three 400-ton chillers
 - Enclosed hot-aisles
 - Supplemental 280-tons of cooling from CRAC units
- **Observations:**
 - Space less an issue than power
 - Cooling > 25kW per rack difficult
 - Power distribution a challenge, more than 1200 circuits



External Power and Cooling Infrastructure



Switches in Place



TACC

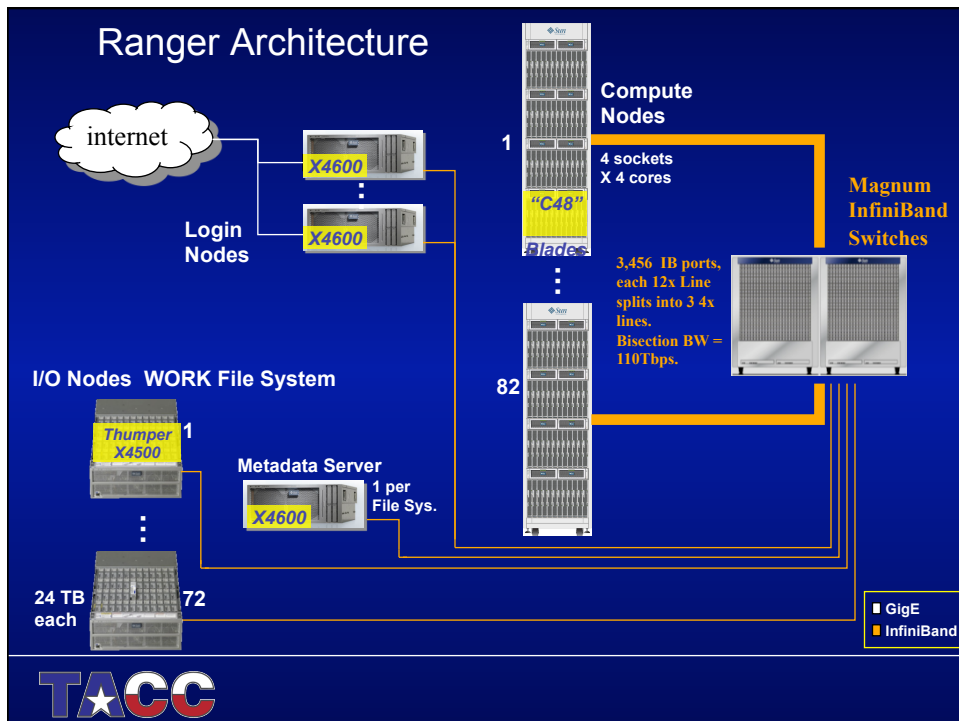
InfiniBand Cabling in Progress



TACC

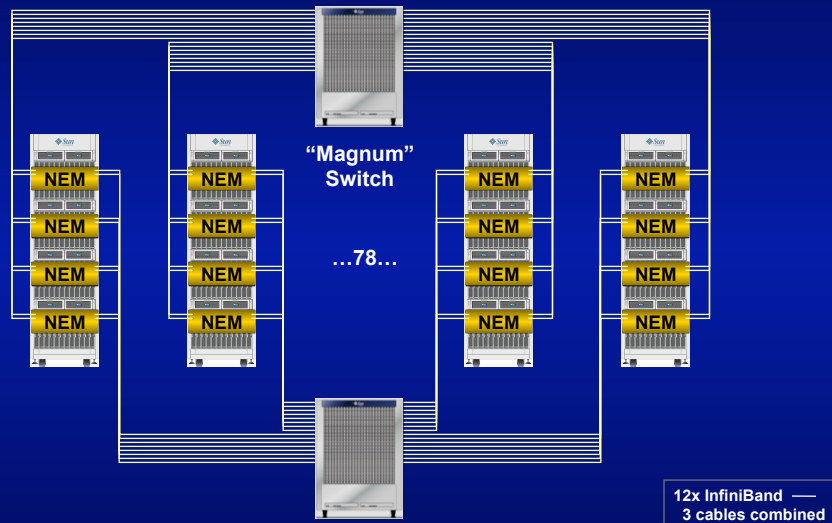
Ranger Features

- AMD Processors:
 - HPC Features → 4 FLOPS/CP
 - 4 Sockets on a board
 - 4 Cores per socket
 - HyperTransport (Direct Connect between sockets)
 - 2.3 GHz core
 - Any idea what the peak floating-point performance of a node is?
 - $2.3 \text{ GHz} * 4 \text{ Flops/CP} * 16 \text{ cores} = 147.2 \text{ GFlops Peak Performance}$
 - Any idea how much an application can sustain?
 - *Can sustain over 80% of peak with DGEMM (matrix-matrix multiply)*
- NUMA Node Architecture (16 cores per node, **think hybrid**)
- 2-tier InfiniBand (NEM – “Magnum”) Switch System
- Multiple Lustre (Parallel) File Systems

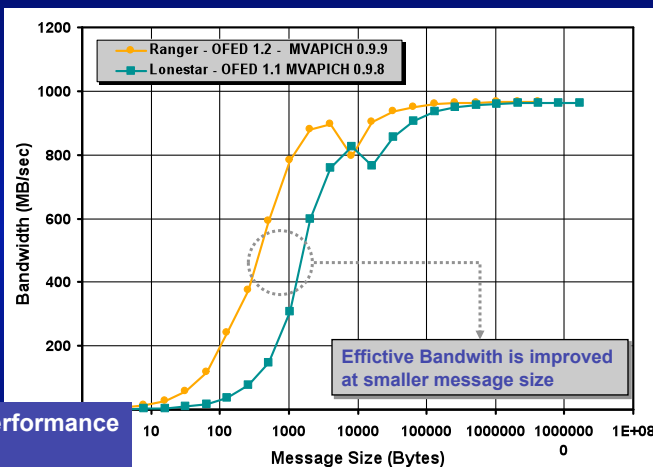


Interconnect Architecture

Ranger InfiniBand Topology



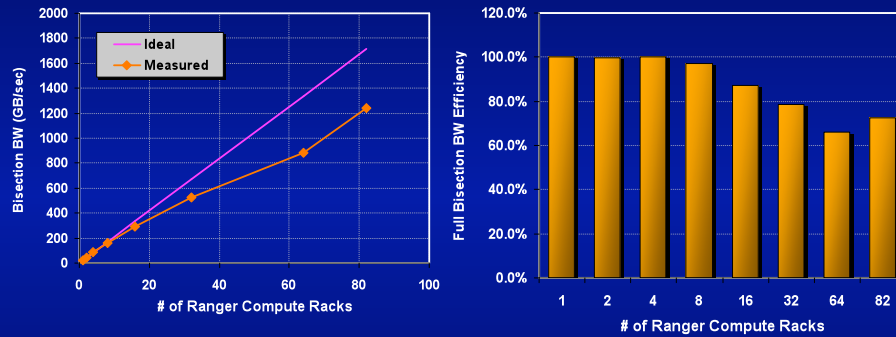
MPI Tests: P2P Bandwidth



•Point-to-Point MPI Performance (Measured)

- Shelf MPI Latencies: ~1.6 μ s
- Rack MPI Latencies: ~2.0 μ s
- Peak Bandwidth: ~965 MB/s

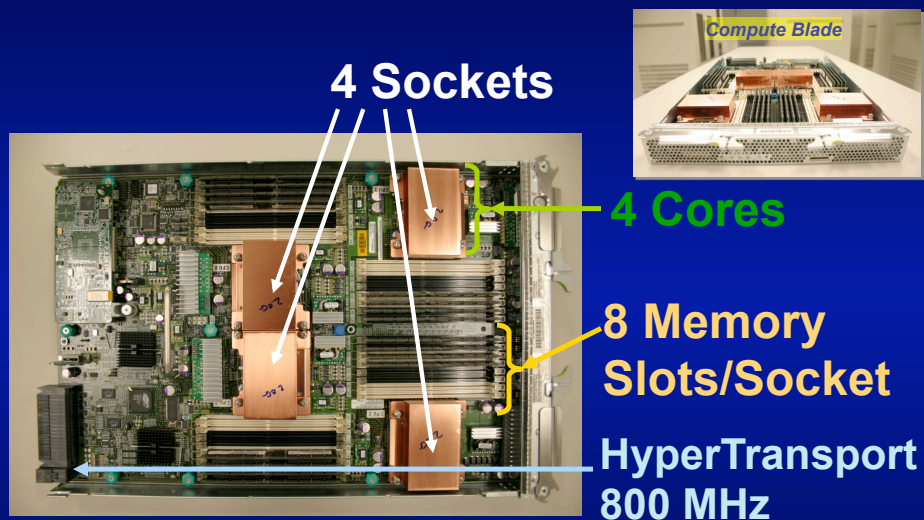
Ranger: Bisection BW Across 2 Magnums



- Able to sustain ~73% bisection bandwidth efficiency with all 3936 nodes communicating simultaneously (82 racks)

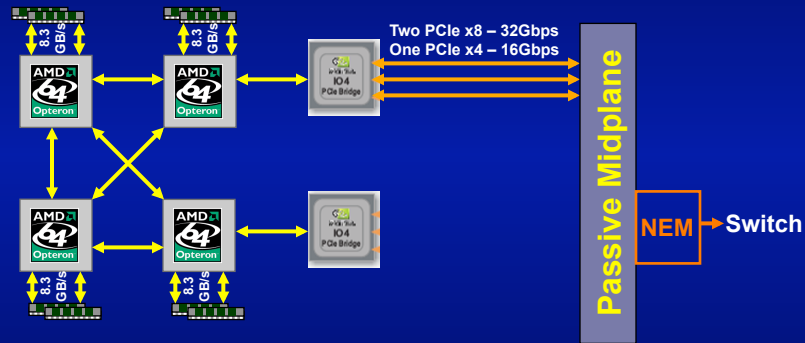


Sun Motherboard for AMD Barcelona Chips



Sun Motherboard for AMD Barcelona Chips

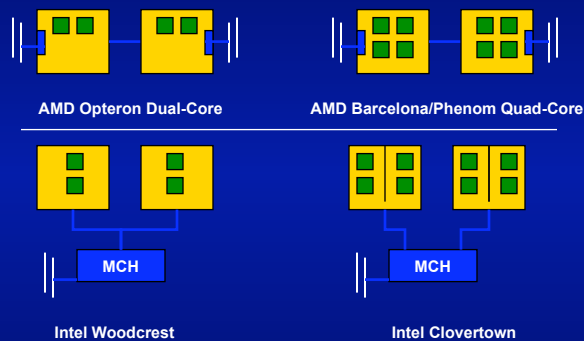
A maximum neighbor NUMA Configuration for 3-port HyperTransport.



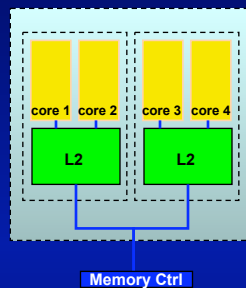
HyperTransport Bidirectional is 6.4GB/s, Unidirectional is 3.2GB/s.
Dual Channel, 533MHz Registered, ECC Memory



Intel/AMD Dual- to Quad-core Evolution

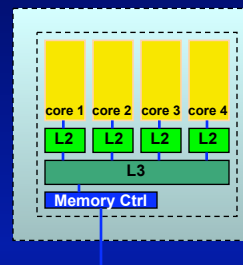


Shared & Independent Caches



Intel Quad-core

All L2's are not independent.

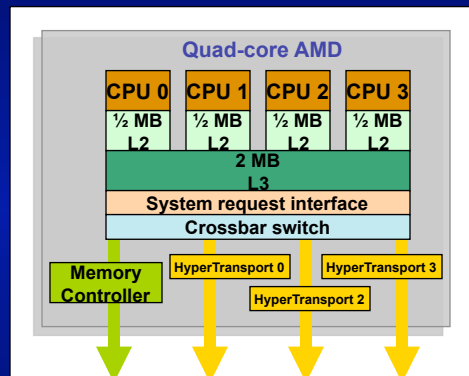


AMD Quad-core

All L2's are independent.



AMD Barcelona Chip: Cache Sizes

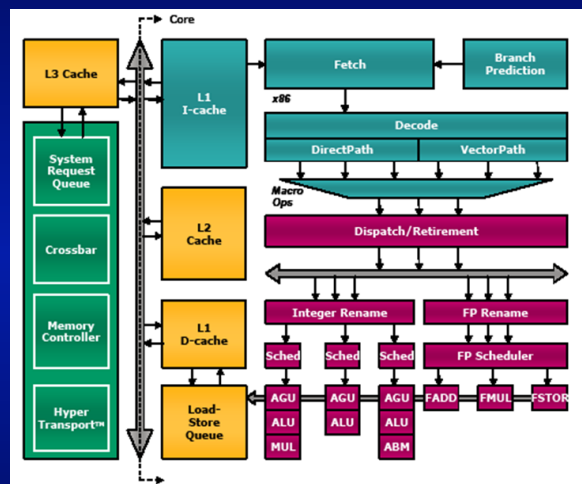


Other Important Features

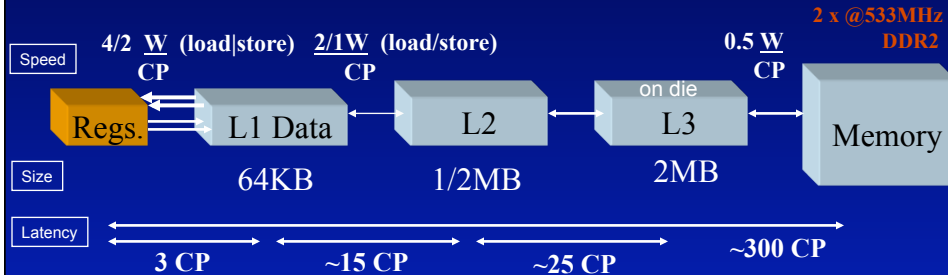
- AMD Quad-core (K10, code name Barcelona)
- Instruction fetch bandwidth now 32 bytes/cycle
- 2MB L3 cache on-die; 4 x 512KB L2 caches; 64KB L1 Instruction & Data caches.
- SSE units are now 128-bit wide --> single-cycle throughput; improved ALU and FPU throughput
- Larger branch prediction tables, higher accuracies
- Dedicated stack engine to pull stack-related ESP updates out of the instruction stream



AMD 10h Processor



Speeds & Feeds (Barcelona)



Cache States: MOESI (Modified, Owner, Exclusive, Shared, Invalid)
MOESI is beneficial when latency/bandwidth between cpus is significantly better than main memory

4 FLOPS/CP

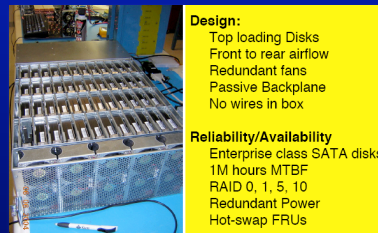
Cache Line size L1/L2 = 8W/8W

W FP Word (64 bit)
 CP Clock Period

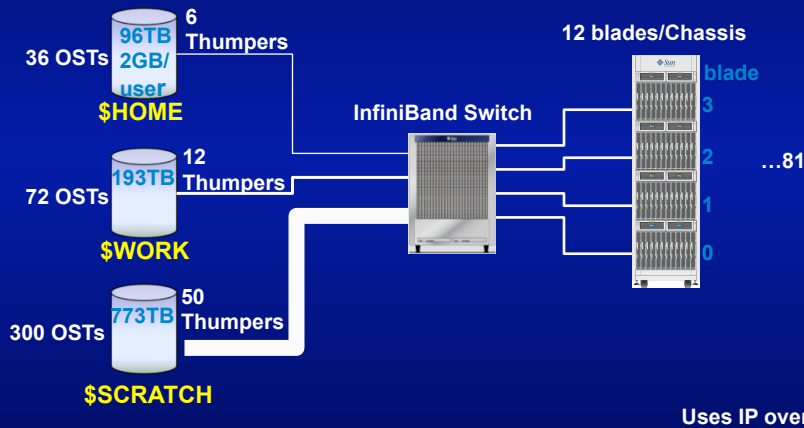


Ranger Disk Subsystem - *Lustre*

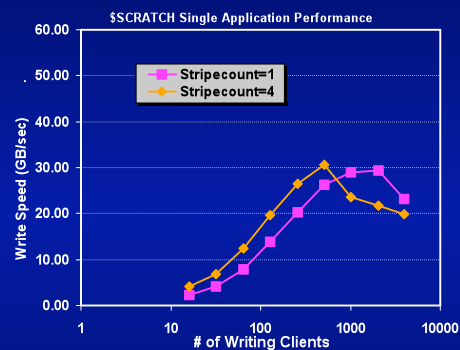
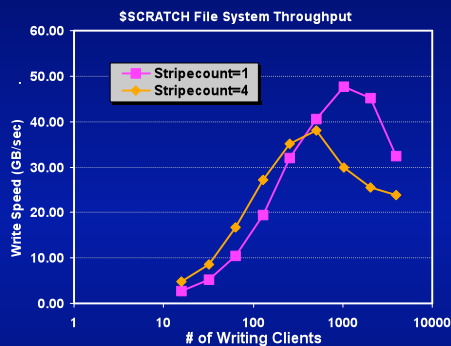
- Disk system (OSS) is based on Sun x4500 "Thumper"
 - Each server has 48 SATA II 500 GB drives (24TB total) - running internal software RAID
 - Dual Socket/Dual-Core Opterons @ 2.6 GHz
 - **72 Servers Total: 1.7 PB raw storage** (that's 288 cores just to drive the file systems)
- Metadata Servers (MDS) based on Sun Fire x4600s
- MDS is Fibre-channel connected to 9TB Flexline Storage
- Target Performance
 - Aggregate bandwidth: 40 GB/sec



Ranger Parallel File Systems: Lustre



I/O with Lustre over Native InfiniBand



- Max. total aggregate performance of 38 or 46 GB/sec depending on stripecount (Design Target = 32GB/sec)
- External users have reported performance of ~35 GB/sec with a 4K application run



Lonestar Related References

- www.tomshardware.com/
- www.topspin.com
- <http://developer.intel.com/design/pentium4/manuals/index2.htm>
- <http://www.tacc.utexas.edu/services/userguides/lonestar/>



Ranger Related References

Guides: www.tacc.utexas.edu/services/userguides/

Forums:

AMD: <http://forums.amd.com/devforum>

PGI: <http://www.pgroup.com/userforum/index.php>

Developers

AMD: <http://developer.amd.com/home.jsp>

AMD Reading: http://developer.amd.com/rec_reading.jsp

