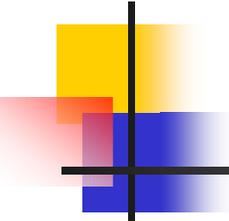


FY2006 Acquisition at Fermilab

Don Holmgren
LQCD Project Progress Review

May 25-26, 2006

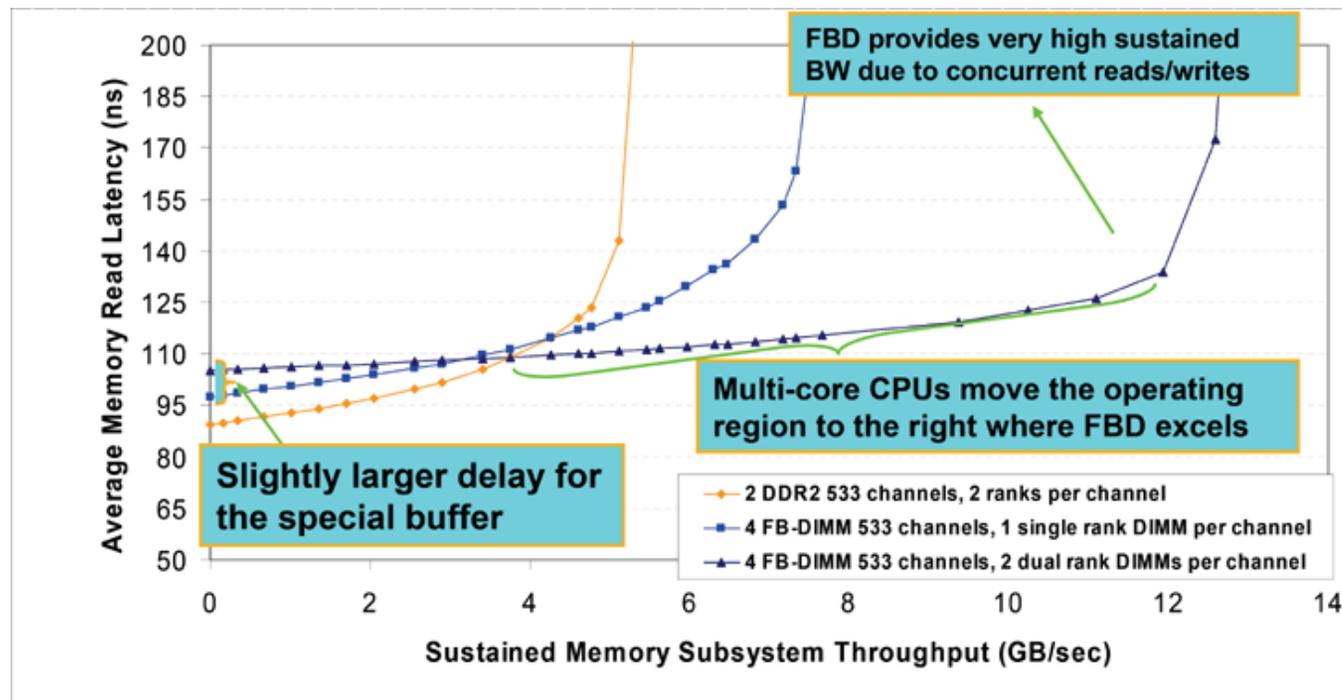
Fermilab

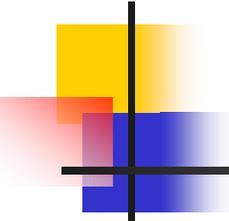


Design, Prototyping, and Projected Performance of “Kaon”

Promise of FBDIMM architecture

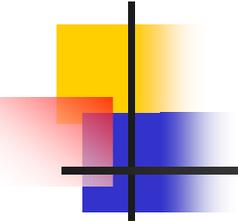
- Projected memory bandwidth performance (2004 Intel graph) was very compelling – it was clear in summer 2005 that we should wait for the February/March introduction of the first FBDIMM machines.
 - Actual introduction: May/June





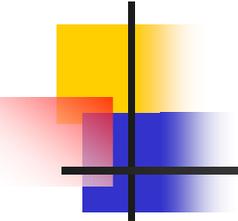
Architectures Considered

- Dual socket Xeon (“Dempsey”) with FBDIMM memory (“fully buffered DIMMs”)
 - Dual core, 1066 MHz FSB
 - Also, “Woodcrest” with 1333 MHz FSB
 - Primary option because of assumed memory bandwidth
 - Many schedule and testing issues
 - Achieved memory bandwidth did not meet expectations
- Dual socket Opteron
 - Dual core
 - DDR400 memory
 - Primary fallback (achieved best price/performance)



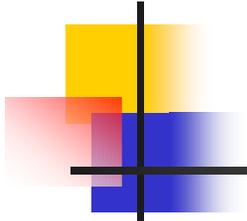
Architectures Considered cont'd

- Single Pentium with 1066 MHz FSB (front side bus)
 - Dual core
 - Also, 800 MHz FSB considered
 - Secondary fallback
- Dual socket Opteron, "Socket F"
 - Dual core
 - DDR2 memory, DDR2-667 or DDR2-800
 - Not available in time for consideration



Networking

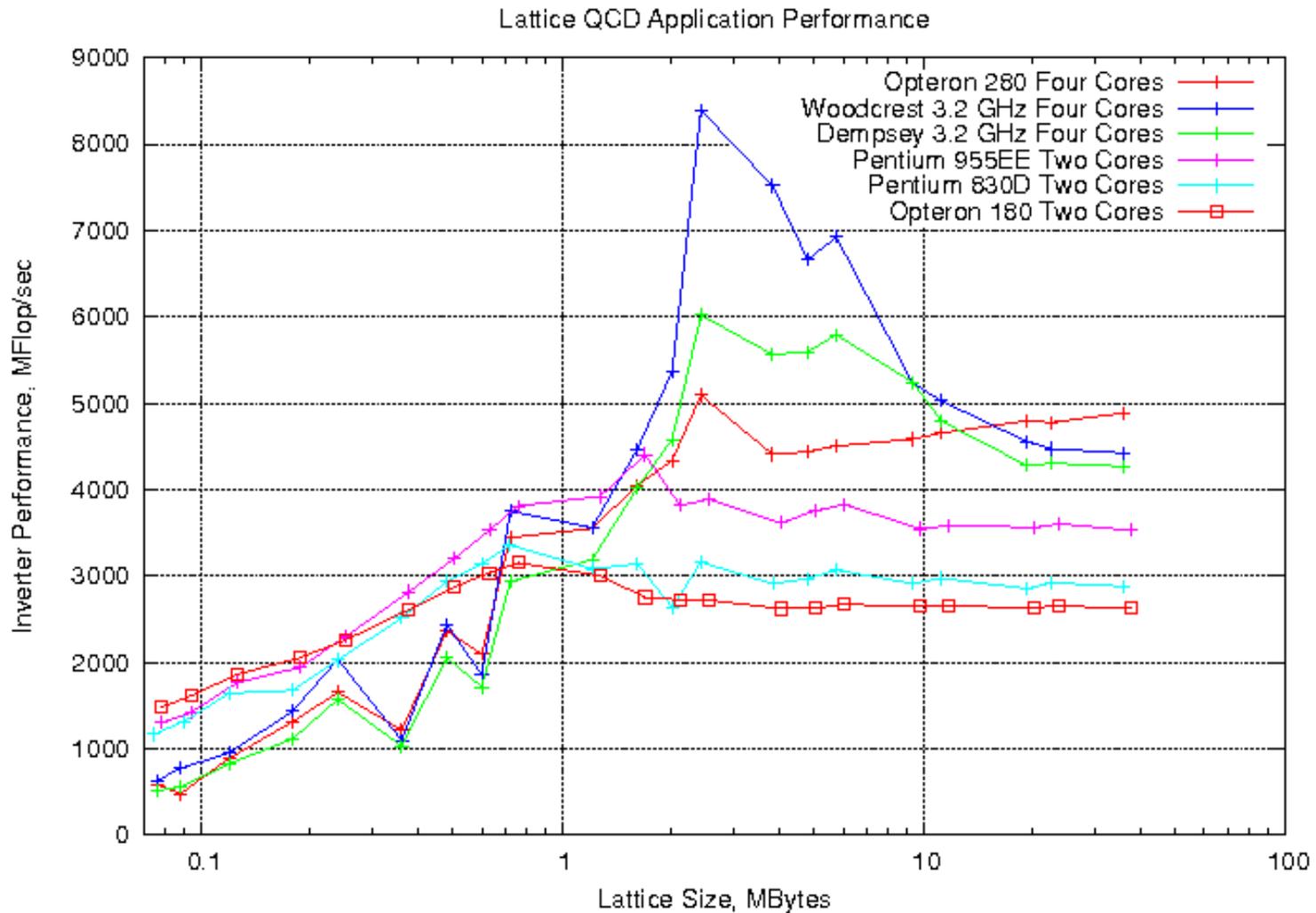
- Fast ethernet service network (for IPMI, NFS, booting)
- Infiniband
 - High performance I/O via IPoIB, possibly SDP
 - Message passing via [mvapich](#), [mpich-vmi](#)
 - Double data rate (DDR) rather than single (SDR)
 - Concerns about bandwidth required per dual-socket, dual-core node (see backup slides for more information)
 - Better possibility of DDR reuse in 3 years
- Infinipath – initially only available for Opterons
 - PCI-E availability came too late for consideration
 - Deemed too risky for this acquisition
 - Will prototype for FY07 via external SciDAC-2 project

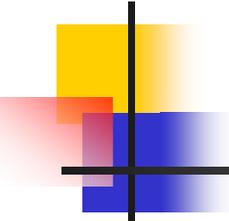


Single Node Testing Performed

- Dual core Pentium 830D (800 MHz FSB)
- Dual core Pentium 955EE (1066 MHz FSB)
- Intel dual socket , dual core systems:
 - "Dempsey", 1066 MHz FSB
 - "Woodcrest", 1333 MHz FSB
- AMD single socket, dual core
- AMD dual socket, dual core

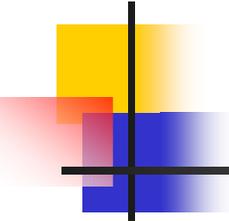
Testing (asqtad benchmark)





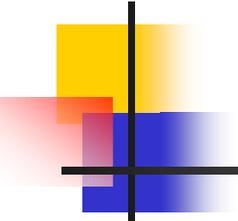
Cluster Testing Performed

- “Emerald”
 - AMD developer center
 - Dual socket, dual core Opteron 275 processors
 - 64 nodes used
 - Infiniband
 - Infinipath host channel adapters, via HTX socket on Iwill motherboards
 - Single data rate fabric
 - Single switch (144-port SilverStorm switch)



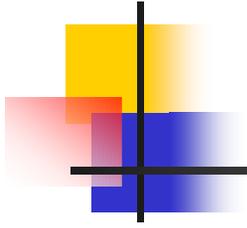
Final Design

- Based on prototyping and vendor cost estimates, we narrowed to dual socket Intel and AMD
- Non-alpha Intel “Dempsey” sample hardware was not available to us until mid-March
- Design:
 - 47U racks, 20 dual-socket servers, ethernet, 24-port DDR Infiniband leaf switch
 - Central 144-port Infiniband spine switch
 - 4 GB memory, 120 GB disk per system

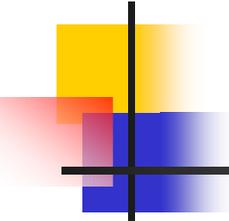


Projected Performance

- Based on scaling tests on “Emerald”:
 - Weak scaling:
 - MILC `asqtad` action with 14^4 local volume (per core) scaled at 73%, comparing 64-node (256 core) performance to single node (4 core) performance
 - `DWF:asqtad` performance on 64-node (256 core) runs was 1.39:1
 - So, to predict MILC/DWF average performance:
 - Multiply single node MILC `asqtad` by 0.73, then by 1.19, to obtain 64-node cluster performance

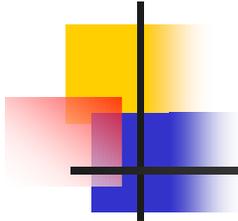


Procurement of “Kaon”



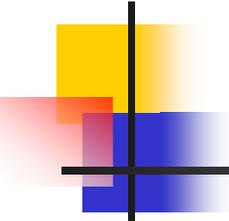
Timeline

- November/December – early prototyping
 - Pentium dual core, also early pre-alpha “Dempsey”
- February 3 – RFI release
- April 3 – RFP release
- April 28 – Bids received
- May 3 – Award recommendation to FNAL purchasing dept.
- May 26 (estimated) – PO Issued
- Late June – First Rack, by mid-August – Remaining Racks
- August – Integration/Friendly User Mode
- September 30 – goal for release to production



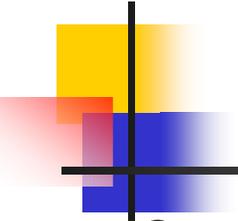
RFI (February 3)

- General description, describing:
 - 40 processors/rack (20 dual, or 40 single socket motherboards)
 - Infiniband, SDR or DDR
 - Server requirements (memory, IPMI, PCI-E, SATA)
- Recipients:
 - Dell, Koi, ACE, CSI, APPRO, Verari, Rackable, Mellanox, Cisco/TopSpin, Voltaire, IBM, HP, Insight, Raytheon, Intel, AMD
- Only informal responses solicited
 - Phone/mail conversations with all
 - Visit to Dell in Austin (Intel, other companies present)
 - Visits from Koi, APPRO, Mellanox, Cisco, Insight, Rackable



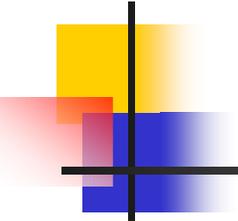
RFP

- RFP solicited bids for:
 - 20 racks, each with 20 dual socket servers, DDR Infiniband HCAs and leaf switch, ethernet, PDUs
 - 1 rack, with 144-port DDR Infiniband spine switch, two leaf switches, ethernet switch, PDUs
- Specifics:
 - Dual socket Intel (Dempsey) or Opteron (270)
 - Because of power constraints, only mid voltage Dempsey allowed
 - All bids must include LQCD benchmark results



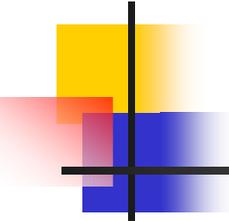
RFP, cont'd

- Specifics, cont'd
 - Opteron chipsets restricted to NVIDIA versions for which we had verified Infiniband HCA compatibility and performance
 - Specific Opteron (Asus) and Dempsey (Intel and Supermicro) motherboard part numbers were given, “or equivalent” allowed on bids (defined)
 - IPMI required, version 2.0 preferred
 - Language allows Fermilab to purchase additional full and/or partial racks



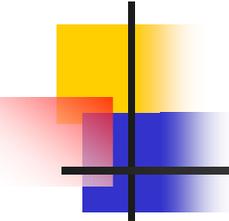
RFP, cont'd

- RFP issued by FNAL Purchasing Dept on April 3
- Recipients: CSI, ACE, Koi, Advanced Clustering Technologies, Dell, Verari, APPRO, Insight, Rackable, IBM, CDW, WSM, Raytheon, HP, Unique Digital



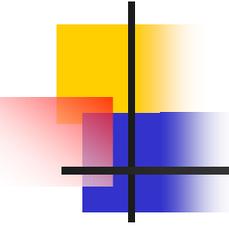
RFP, cont'd

- Bids arrived April 28 (7 companies, 12 configurations)
- Winning bid:
 - Dual Opteron 270, Asus K8N-DRE, 4GB memory
 - Mellanox/Flextronics DDR Infiniband
 - \$2650/node total
 - Includes all Infiniband, ethernet, racks, cabling, integration, 3-year warranty
 - 500 nodes, \$1.31M
 - Based on benchmarks and “Emerald” scaling, 1.93 Tflops (goal was 1.8 Tflops)
 - Will purchase 80 additional nodes (SciDAC/Supplemental) 2.2 Tflops total

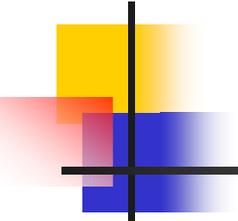


Schedule

- By policy, Fermilab encourages a staged delivery, with multiple decision points to avoid risk
- Current schedule:
 - Award (DOE approved): May 26 (estimated)
 - Visit to vendor (checklist, inspection): by June 14
 - First rack: by June 28
 - Remaining racks: by mid-August
 - Acceptance and payment: at conclusion of acceptance testing, at least 30 days after receipt of all hardware

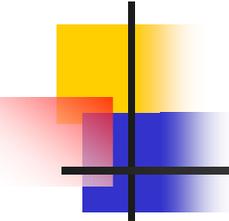


Manufacturing, Installation and Commissioning of “Kaon”



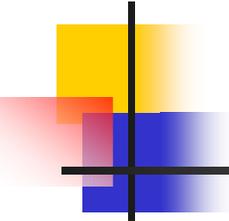
Steps to First Rack Integration

- Using a prototype system in hand, we will do the following:
 - Prepare an installation image, based on Scientific Linux 4.2 for x86_64
 - Installation image will include OpenIB stack, as well as binaries and scripts for LQCD codes to be used to verify and benchmark performance
 - Installation image will create a working cluster within one rack
 - For example, subnet manager will be included
 - Benchmarks will include single and cluster jobs



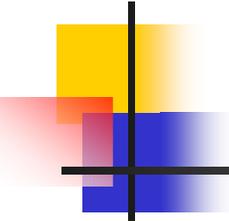
First Rack Integration

- During visit to vendor site, Fermilab personnel will:
 - Deliver installation images and instructions
 - Deliver manufacturing checklist, which must be executed for each delivered computer and rack
 - Inspect manufacturing and integration facility
 - Determine shipping procedures, including property tagging and reporting
 - Teach vendor how to install images, configure systems, verify single and cluster node operation
 - **Winning vendor is local, and sold “Pion” cluster to FNAL**



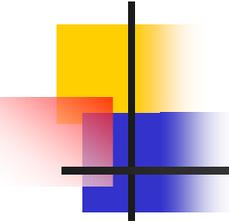
Testing of First Rack

- After delivery to Fermilab (or at vendor, if local):
 - Verification of power consumption
 - Verification of vendor-measured performance
 - Stability testing:
 - Multiple repeated reboots
 - Full memory testing
 - Application testing, including full I/O load
 - After testing:
 - Modifications will be made to installation and configuration procedures, tagging, shipping, as needed



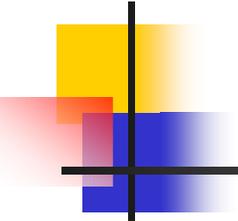
Integration of Full Cluster

- Vendor will be responsible for delivering full racks which match the first rack and meet all performance requirements
- All racks to be tested under our application load before delivery
- We will make individual racks available to friendly users as early as possible
 - Physics production can start with first rack in July
- Fermilab will integrate racks (connect IB and gigE uplinks, configure subnet manager, integrate with head node and batch scheduler)



Acceptance Testing

- Standard procedure:
 - Initially, all nodes will rotate through LQCD applications tests, memory tests, reboot and/or power cycle tests (1 to 2 weeks)
 - Release of full cluster to friendly user mode as soon as stable
 - 95% uptime for 30 days is required for full acceptance



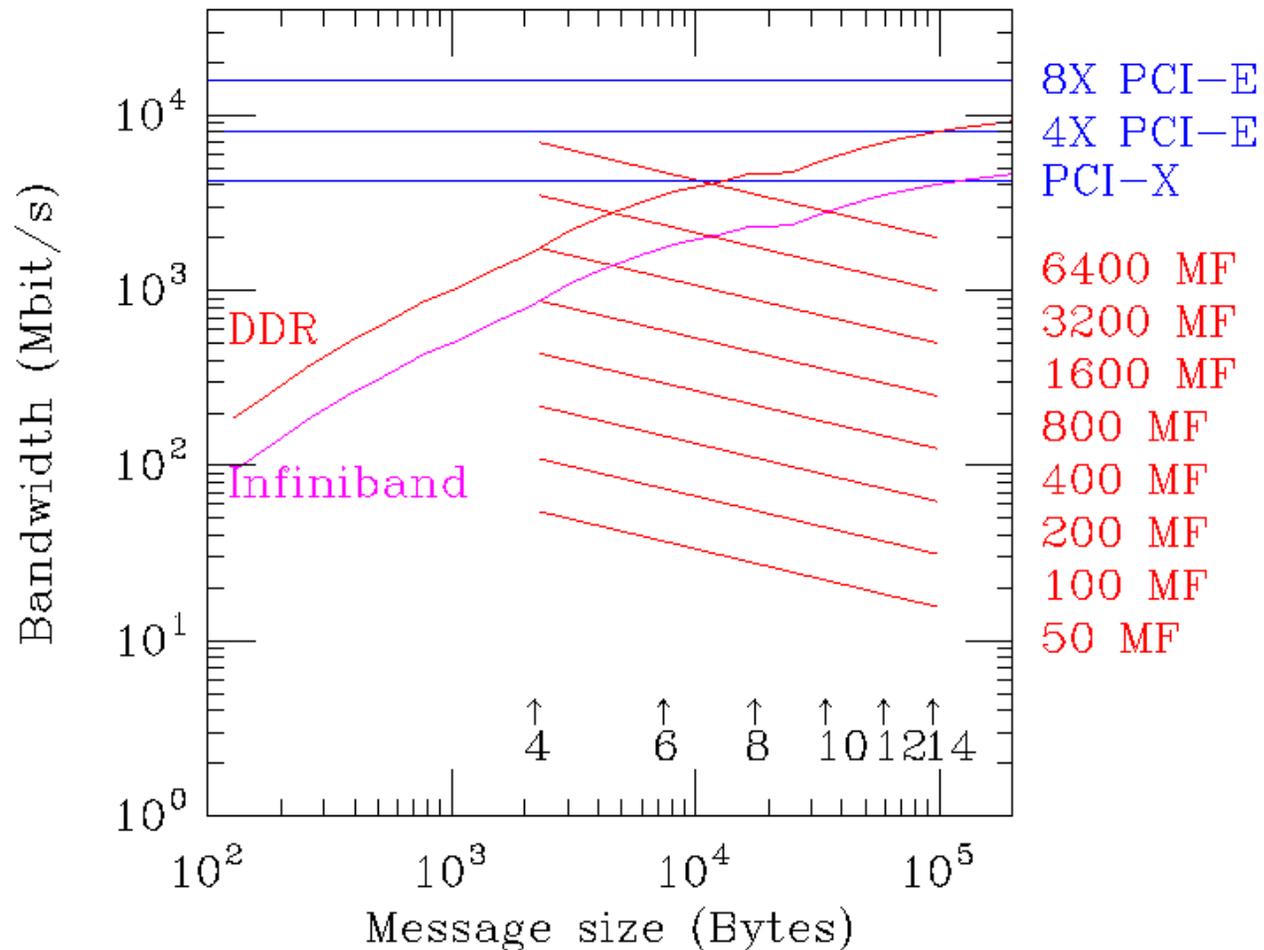
Facility

- “Kaon” will be housed in a refurbished computer room adjacent to “Pion” and “QCD” clusters
 - GPP project to complete mid-June (on schedule)
 - Facility provides:
 - 9.4 Kwatt/rack cooling (Leibert overhead fan coils)
 - Space for 52 racks
 - Served by 1680 kVA transformer

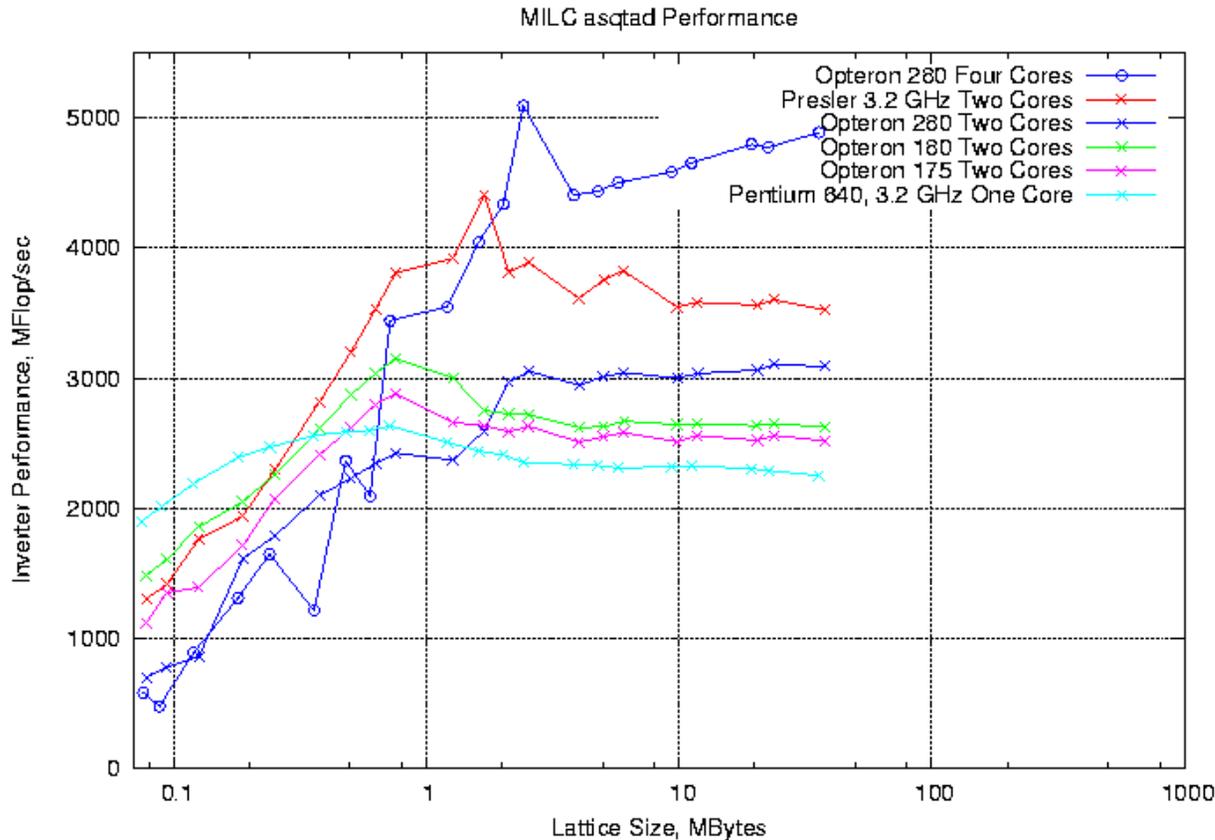
Backup Slides

SDR vs. DDR Infiniband

Communications Requirements



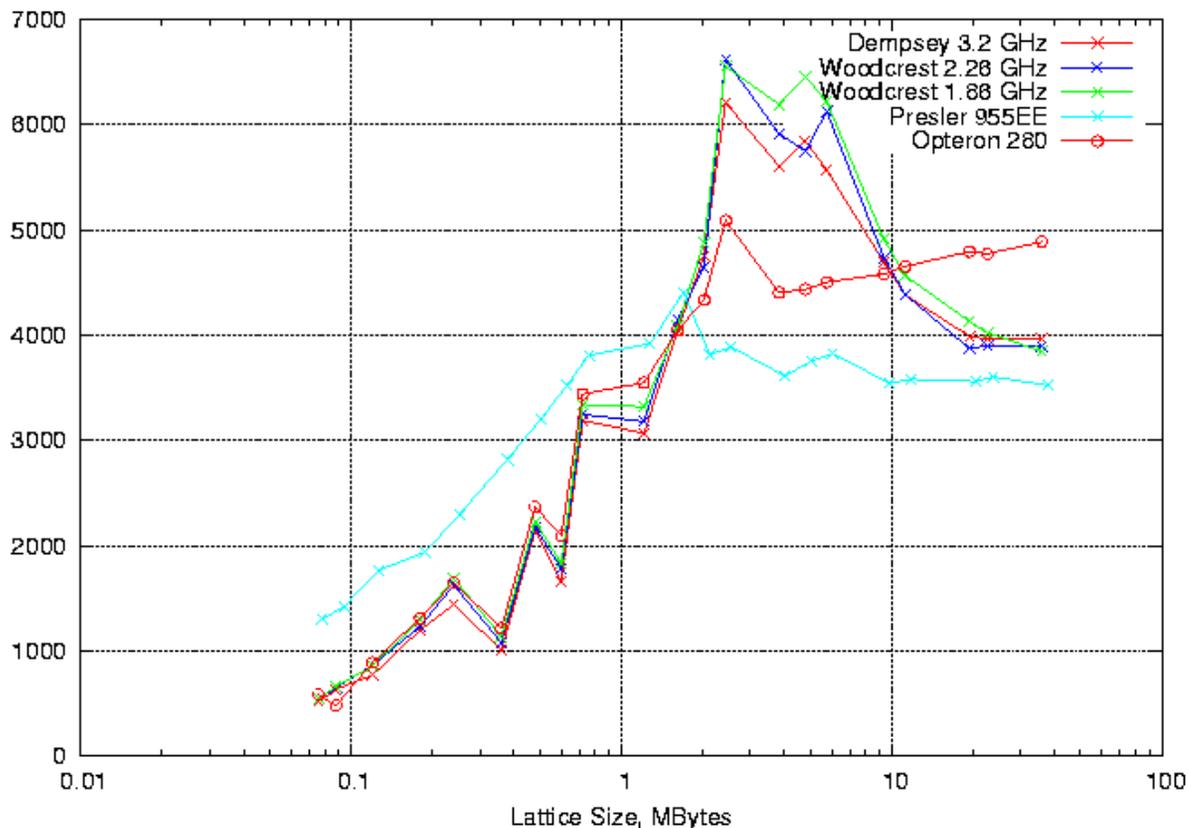
Testing (asqtad benchmark)



- “Presler” – 1066 MHz FSB, dual core
- Aggregate performance per node shown (sum of n processes, one per core)
- Opteron 175/180 shows that performance doesn't scale with clock
- Pentium 640 – used on Fermilab “Pion” cluster

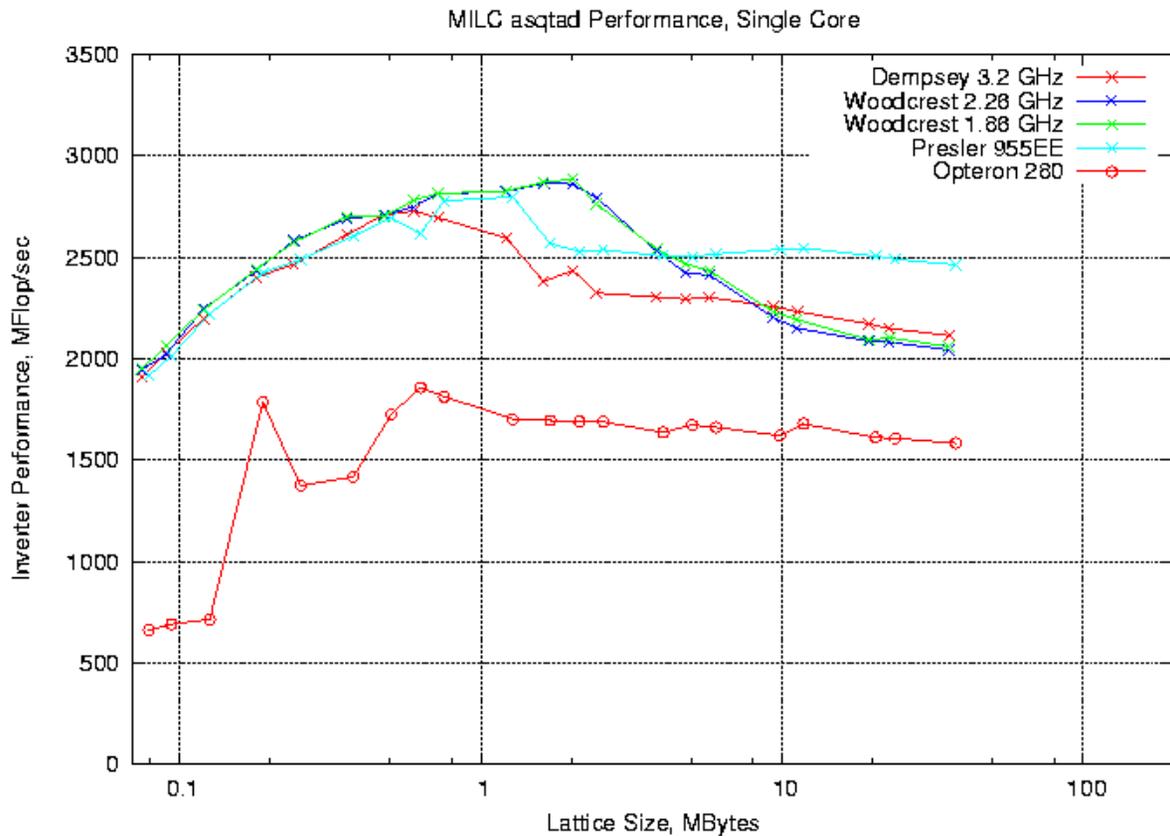
Testing (asqtad) – cont'd

MILC asqtad Performance, Multicore



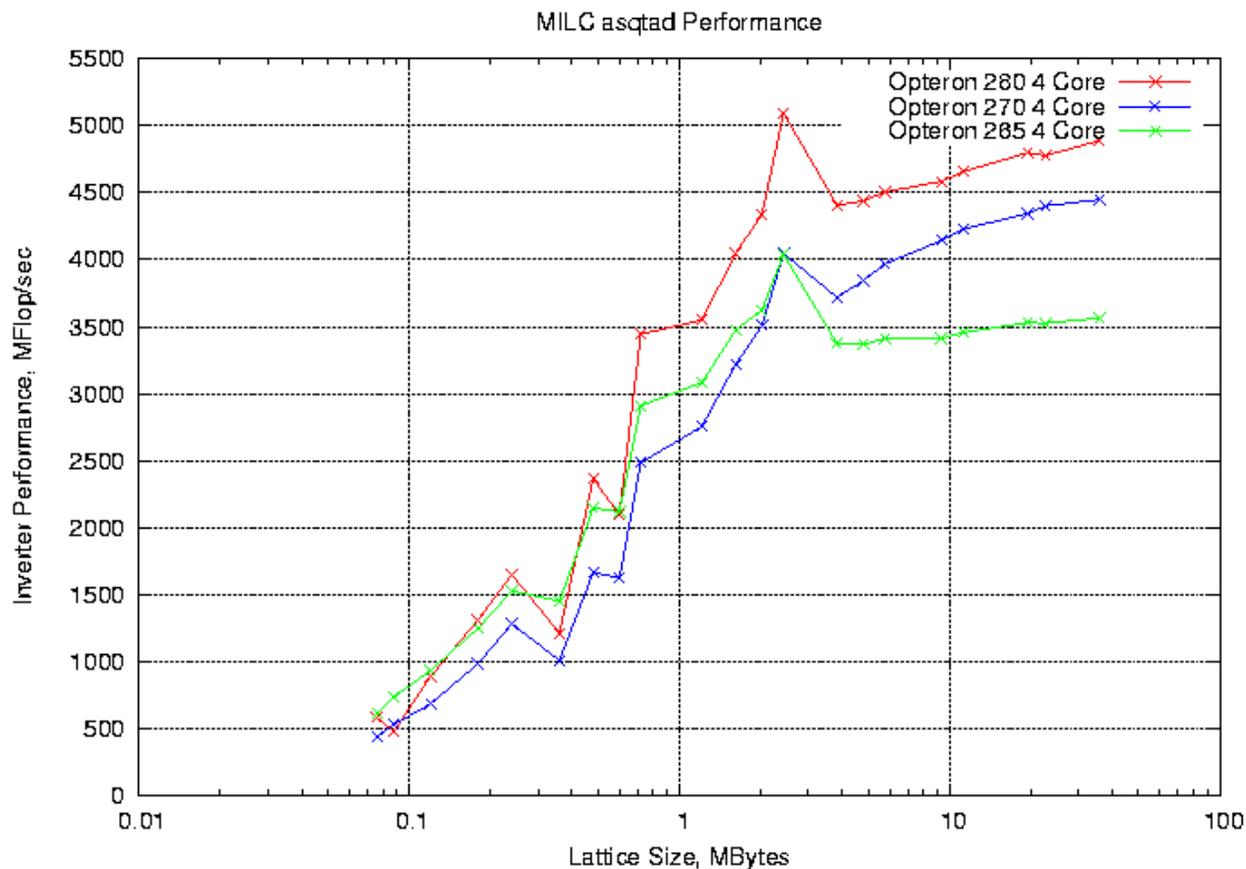
- Aggregate performance of n processes
- “Dempsey”, “Woodcrest” use FBDIMM memory architecture (1066 MHz)
- “Woodcrest” uses new Intel Microarchitecture
- FBDIMMs are clearly not solving the memory bandwidth issue for dual Xeons
- Opteron rising performance results from declining surface/volume ratio \rightarrow decreasing communications overhead

Testing (asqtad) – cont'd



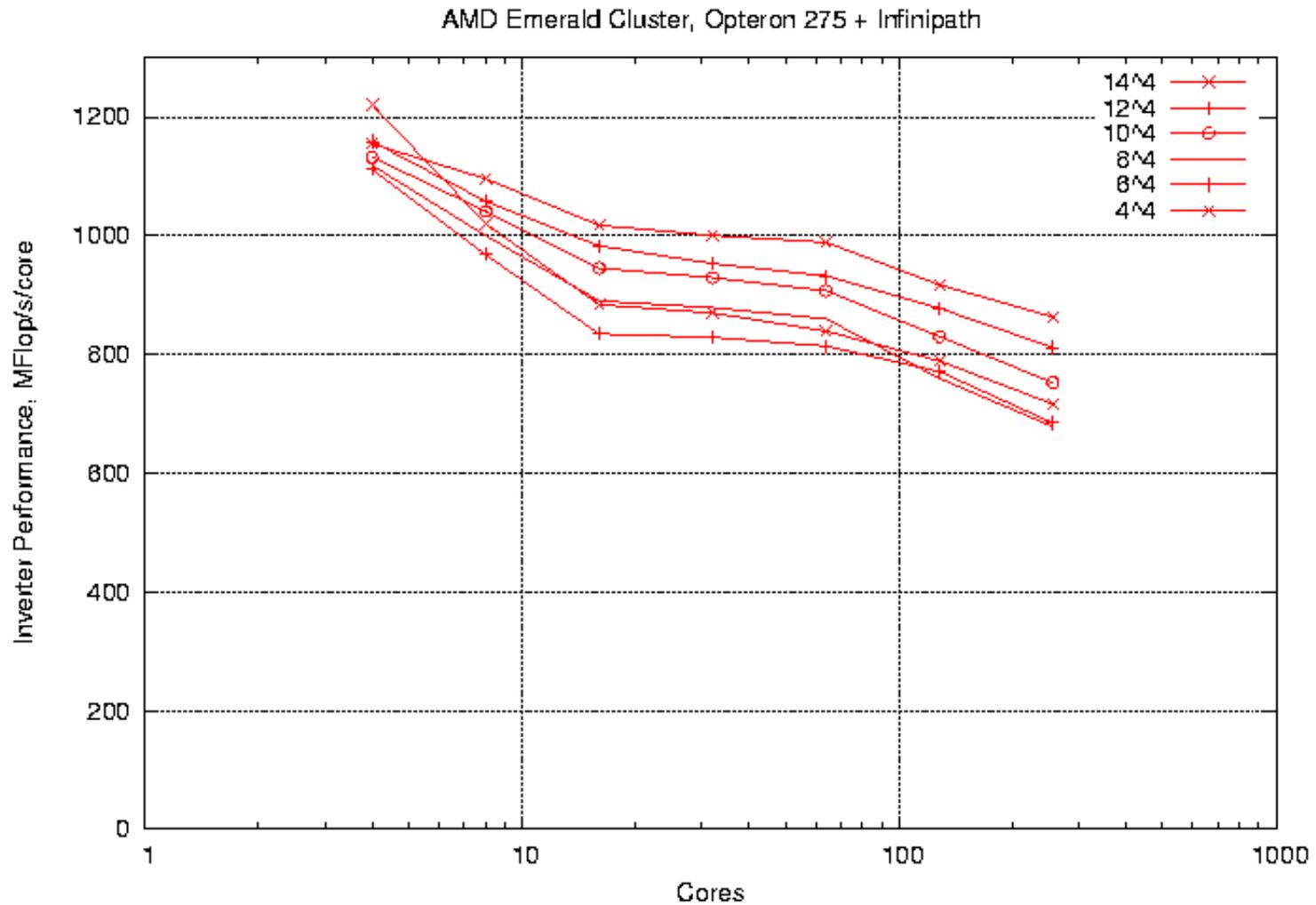
- Single core performance
- “Presler” at 3.73 GHz, 1066 FSB has the best performance, but is too costly

Testing (asqtad) - Opteron

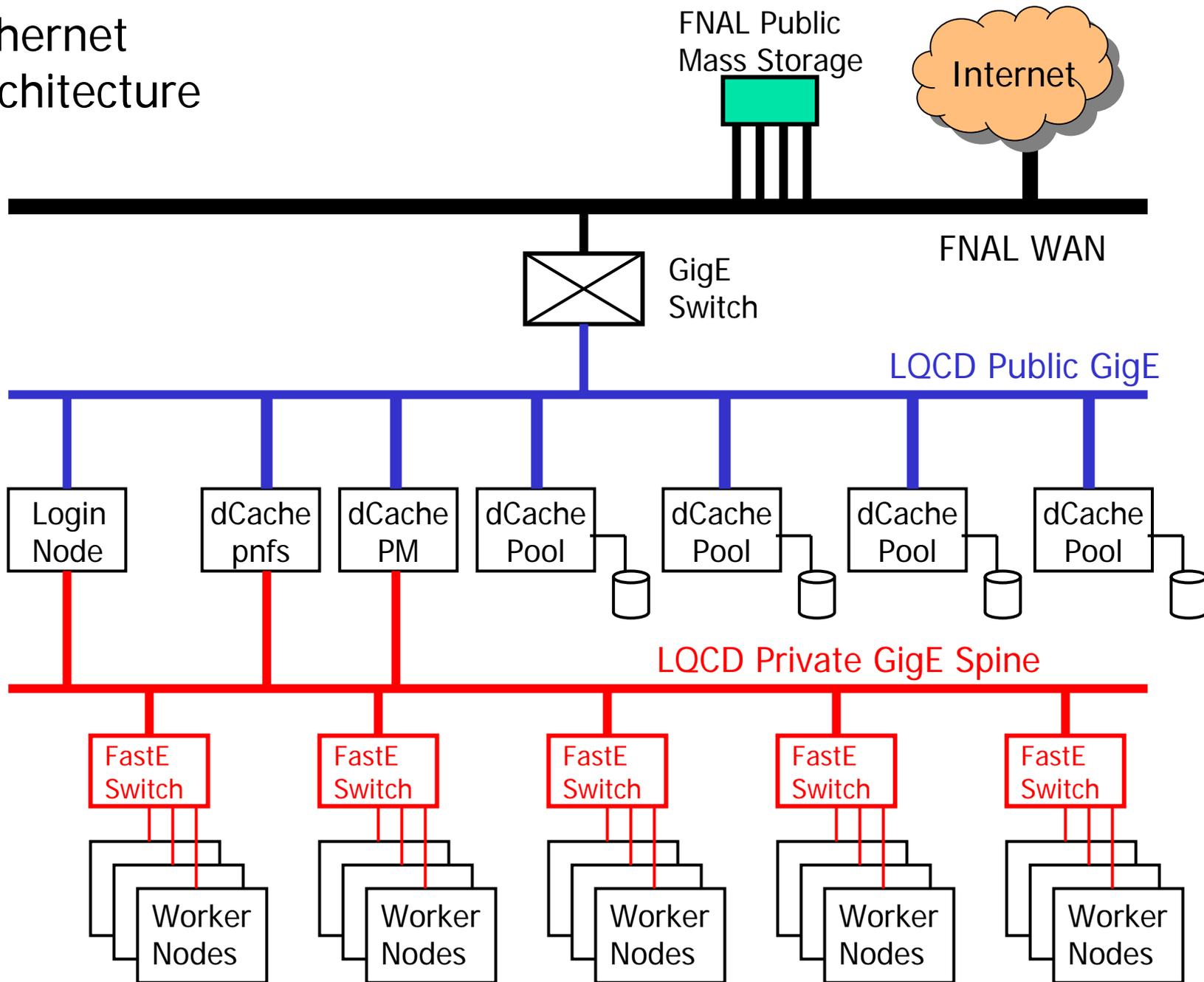


- Another example showing that Opteron performance does not scale with clock speed
- Best price/performance obtained from mid-speed step processor

Cluster Testing (cont'd)



Ethernet Architecture



Infiniband Architecture

