Benchmarking in Production Environment

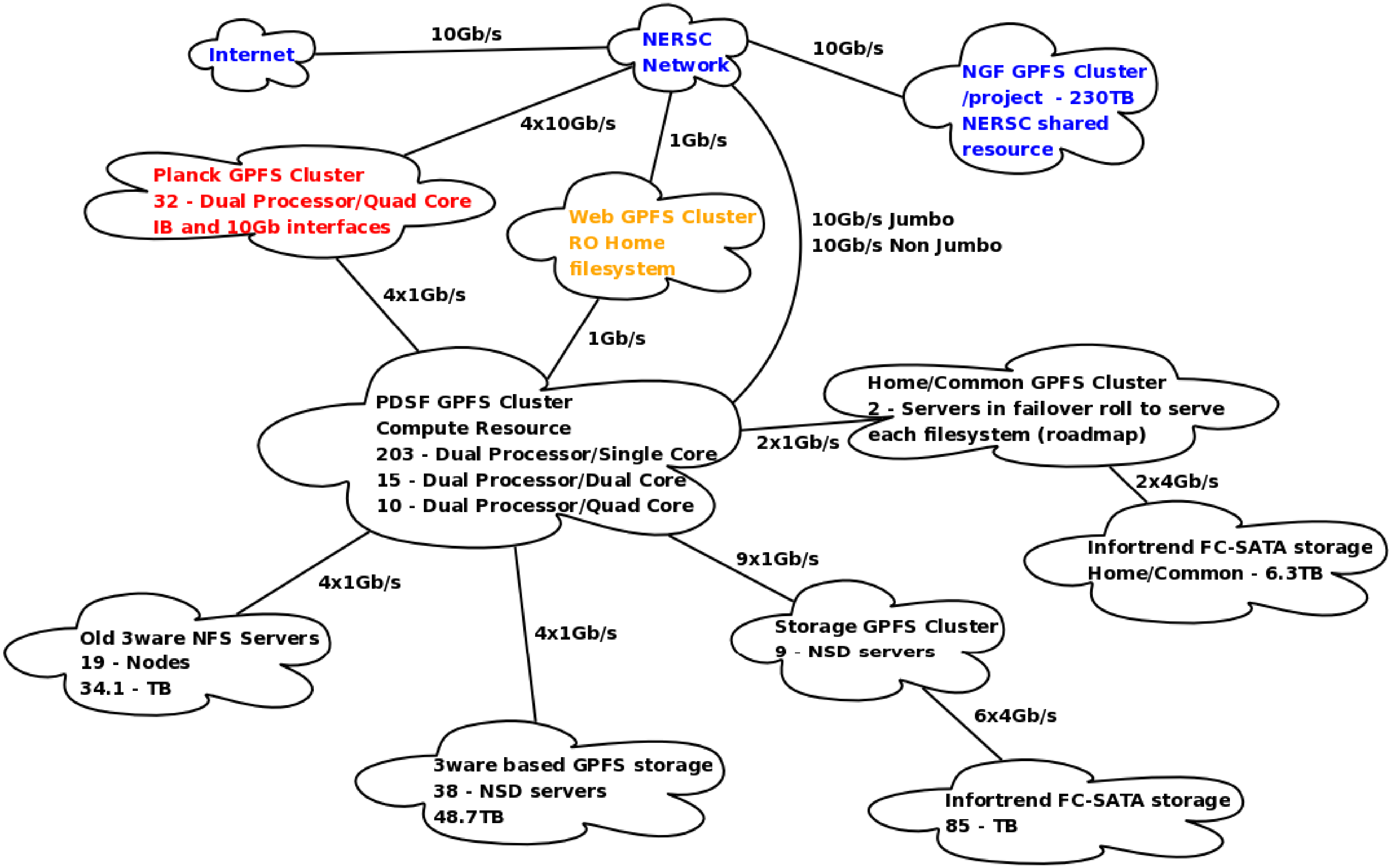
Iwona Sakrejda, Jay Srinivasan
Lawrence Berkeley National Laboratory
NERSC

HEPIX Spring Meeting
CERN 5-9 May 2008

Outline

- PDSF Production Environment
- New Hardware Acquisitions
- Benchmarking & Production
- Characteristics of Benchmarking codes
- Benchmarking Setup
- Results of Core-dependency Comparisons
- I/O tests
- Conclusions

PDSF Production Environment



Internet

NERSC Network

NGF GPFS Cluster
/project - 230TB
NERSC shared resource

Planck GPFS Cluster
32 - Dual Processor/Quad Core
IB and 10Gb interfaces

Web GPFS Cluster
RO Home filesystem

PDSF GPFS Cluster
Compute Resource
203 - Dual Processor/Single Core
15 - Dual Processor/Dual Core
10 - Dual Processor/Quad Core

Home/Common GPFS Cluster
2 - Servers in failover roll to serve each filesystem (roadmap)

Infortrend FC-SATA storage
Home/Common - 6.3TB

Old 3ware NFS Servers
19 - Nodes
34.1 - TB

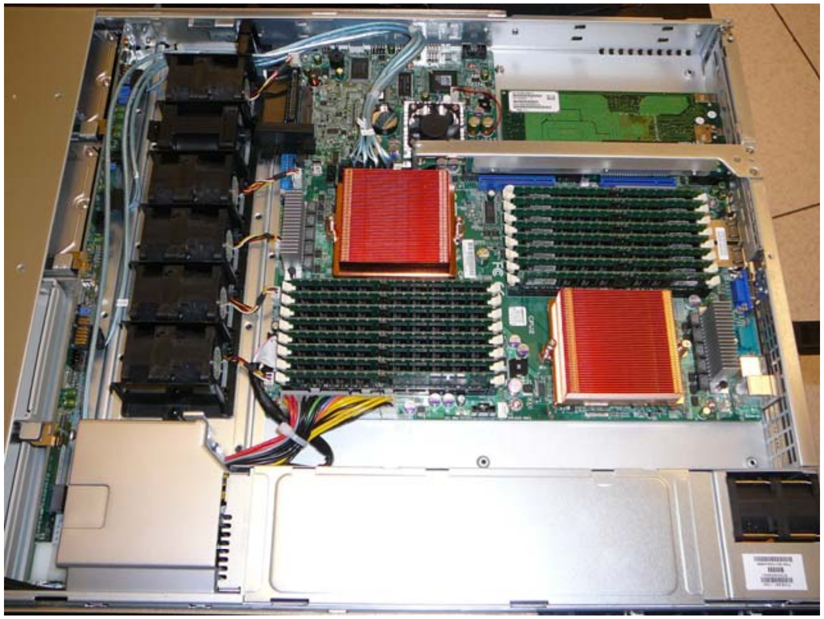
Storage GPFS Cluster
9 - NSD servers

3ware based GPFS storage
38 - NSD servers
48.7TB

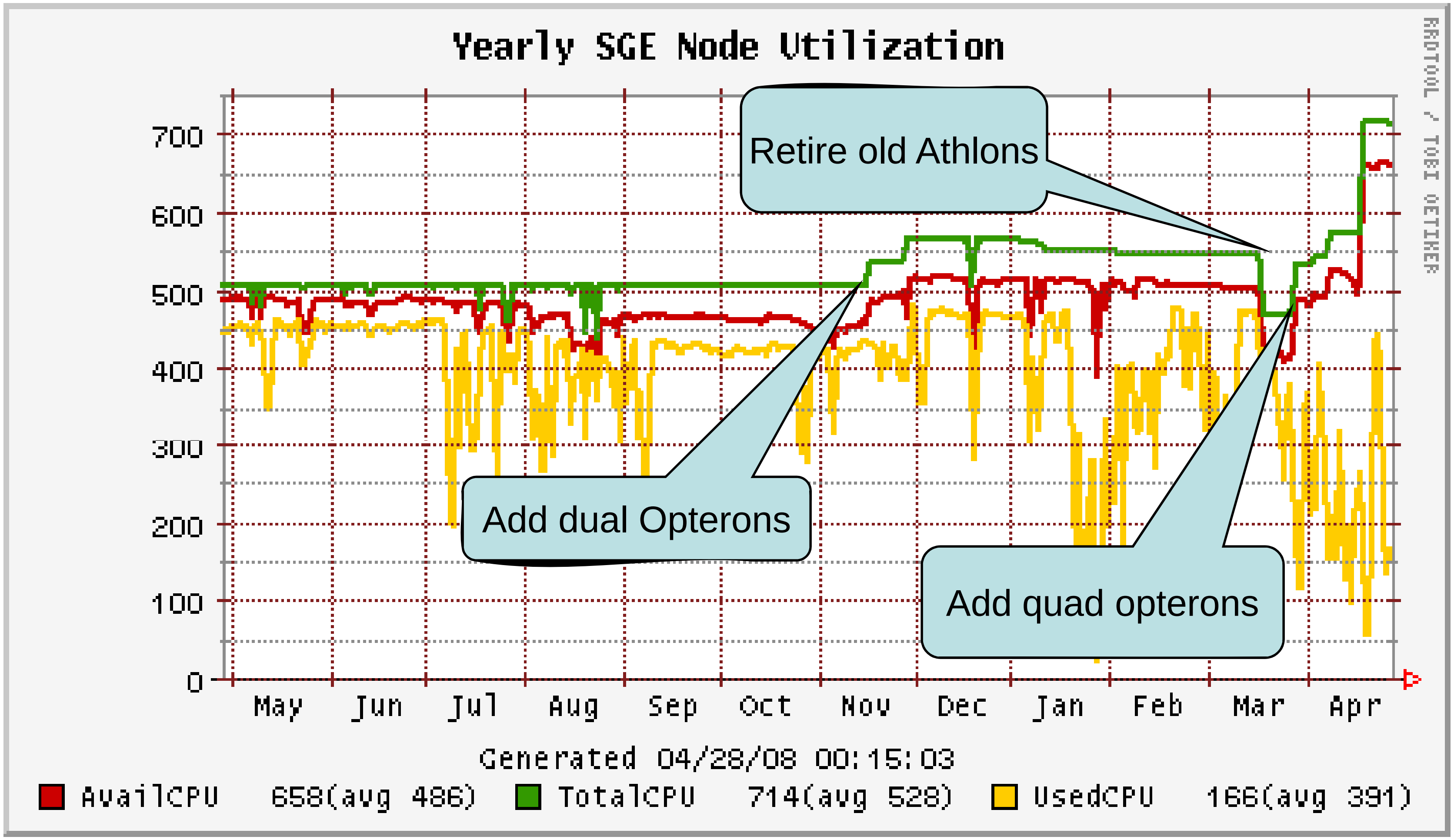
Infortrend FC-SATA storage
85 - TB

New Hardware Acquisitions

- 16 2-socket dual core 2.8GHz nodes 2GB memory/core
- 9 2-socket quad core 2.1GHz nodes 2GB memory/core
- 33 2-socket quad core 2.1GHz nodes 4GB memory/core, IB infrastructure
- 10GE network infrastructure



New Hardware Acquisitions

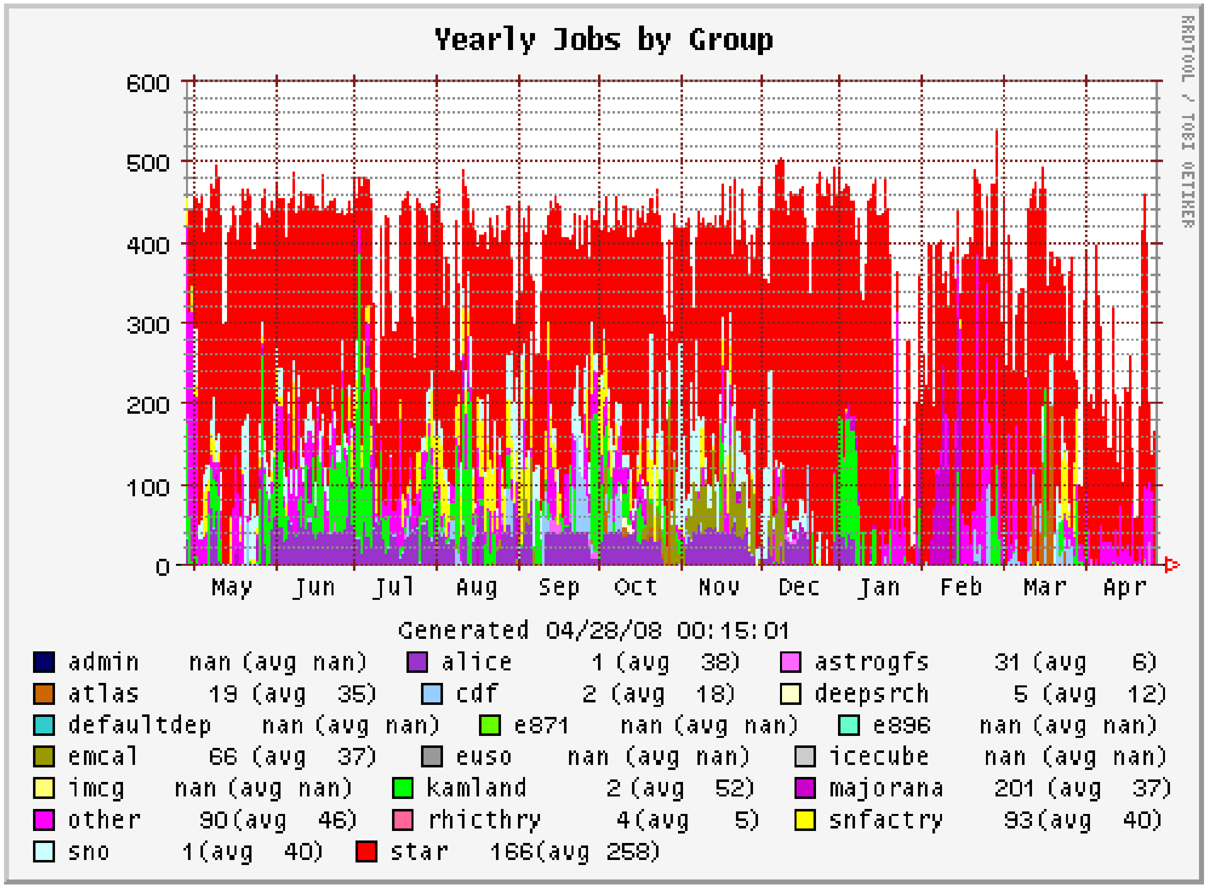


Retire old Athlons

Add dual Opterons

Add quad opterons

PDSF Production Environment



Widely varying workload composition and intensity.

Benchmarking setup

- Selected nodes from the compute setup drained and designated as test nodes, but remaining in their racks - slightly different path to storage.
- Production continuing on the whole cluster - limited control of network and storage load.
- Facility located downtown Oakland - limited ability to control power distribution

This used to be our 12 kV line

- User groups asked to provide (submit held) jobs that were representative to their workload and “as identical as possible”.
- Jobs release:
  - Simultaneous
  - Uniform - jobs from 1 group running at any given time
  - Job multiplicity = 2xcore multiplicity
- Varying quality of storage where input and output files were located.

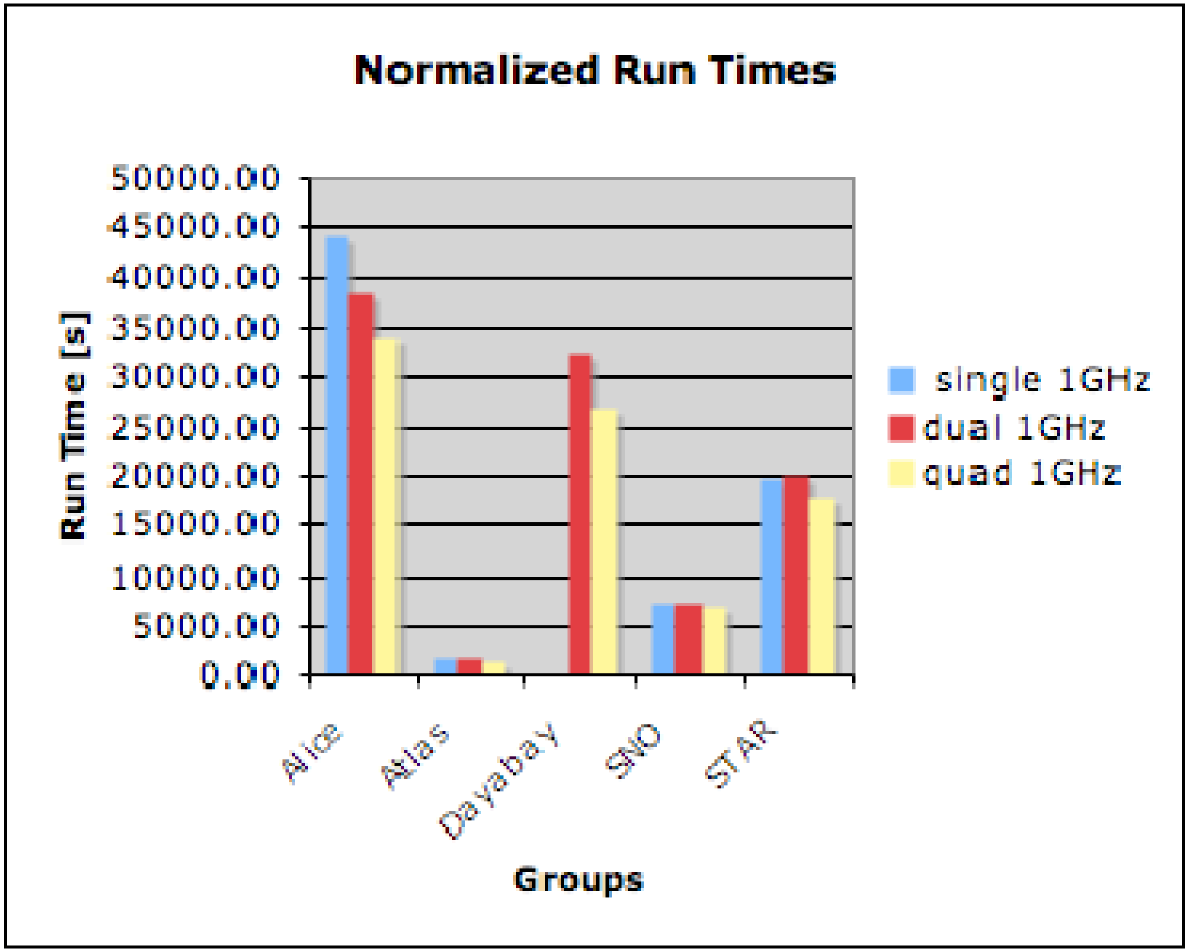
Description of user codes

- Alice - MC. Memory footprint - 1.2GB
- Atlas - Analysis code (end user activity). The full job consists of environment setup and the actual program execution. All jobs are identical and read in the same set of input files from (about 14 GB). Memory footprint 330MB.
- Dayabay - Memory footprint 265MB.
- SNO - The code (called SNOMAN) is used by the SNO group to generate MC simulations of the detector response to various signals and is also used in the final pass of the data analysis. The jobs that I had submitted had SNOMAN configured to simulate neutrino signals in the SNO detector. The total size of the input files is on average of 42 Mb. SNOMAN outputs its results into an ntuple file and a root file. Memory footprint - 720MB
- STAR - simulation and reconstruction (embedding) . Memory footprint - 700MB

Results

Note: Alice code half way through testing increased memory footprint by 25%

Dayabay - user accidentally deleted jobs he designated for testing before single-core jobs were run.

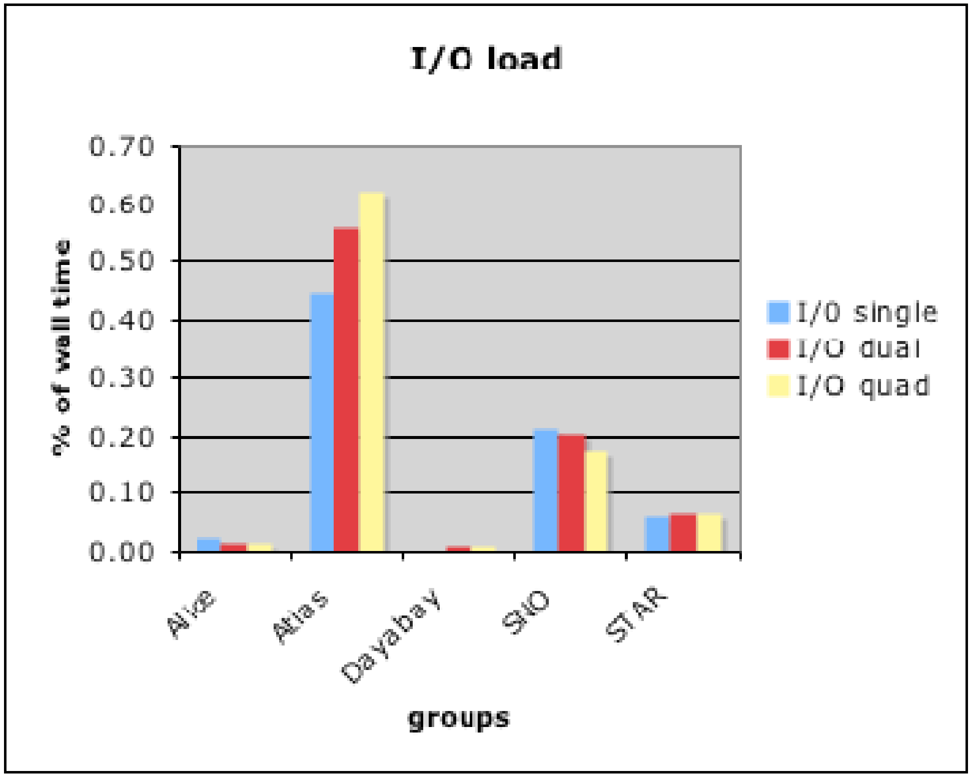


Hardware Specs

Model	248	2220	2352
Speed[GHz]	2.205	2.8	2.114
Core Count	1	2	4
Revision		f3	b3
Memory Controller	2.2GHz	2.8GHz	1.8GHz
L1 Cache	64K+64K	64k+64k	64k+64k
L2 Cache	1MB	1MB	512KB
L3 Cache			2048KB

Comparison of wall clock runtimes

- Multi-core systems need higher bandwidth to the racks to efficiently run “analysis” type of jobs.
- We are going to do channel-bonding between switches in the racks and the core network switch
- We are looking at the 10GE infrastructure

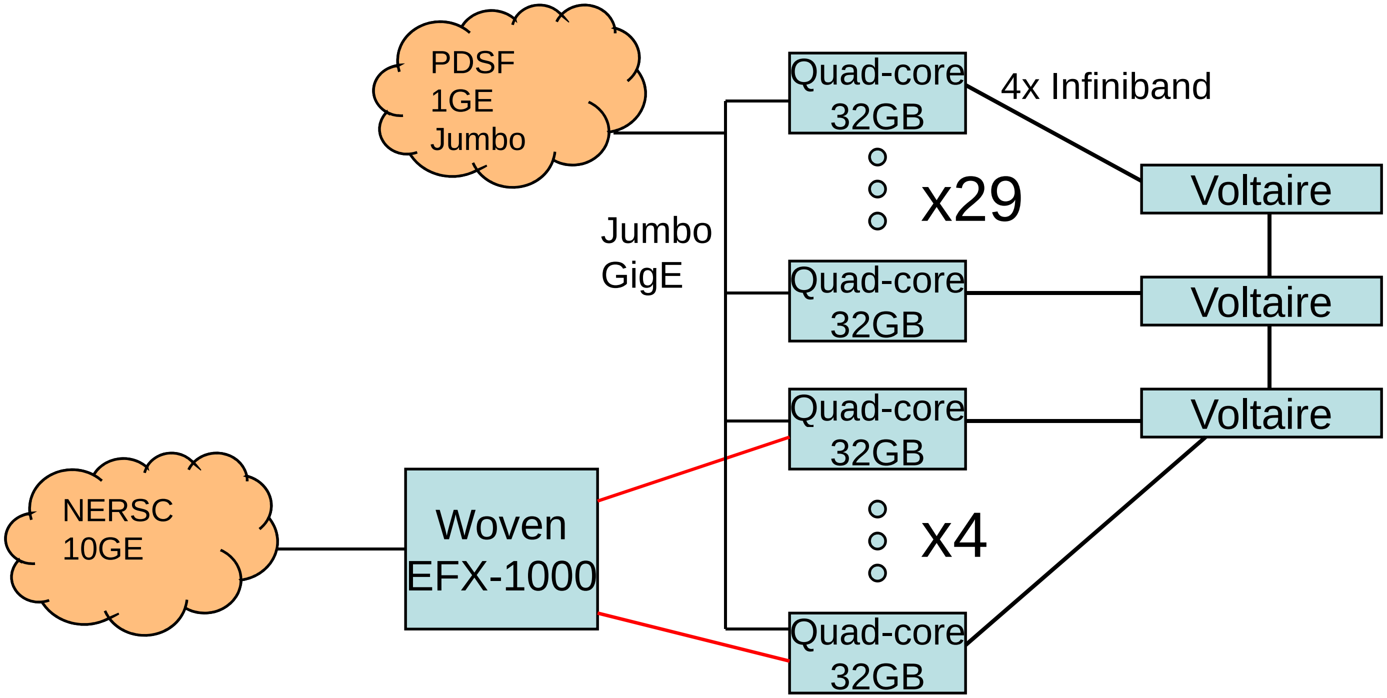


PDSF workload runs well on multi-core systems with adequate amount of memory per core. Quad-core performs slightly better thanks to re-designed architecture of memory caching

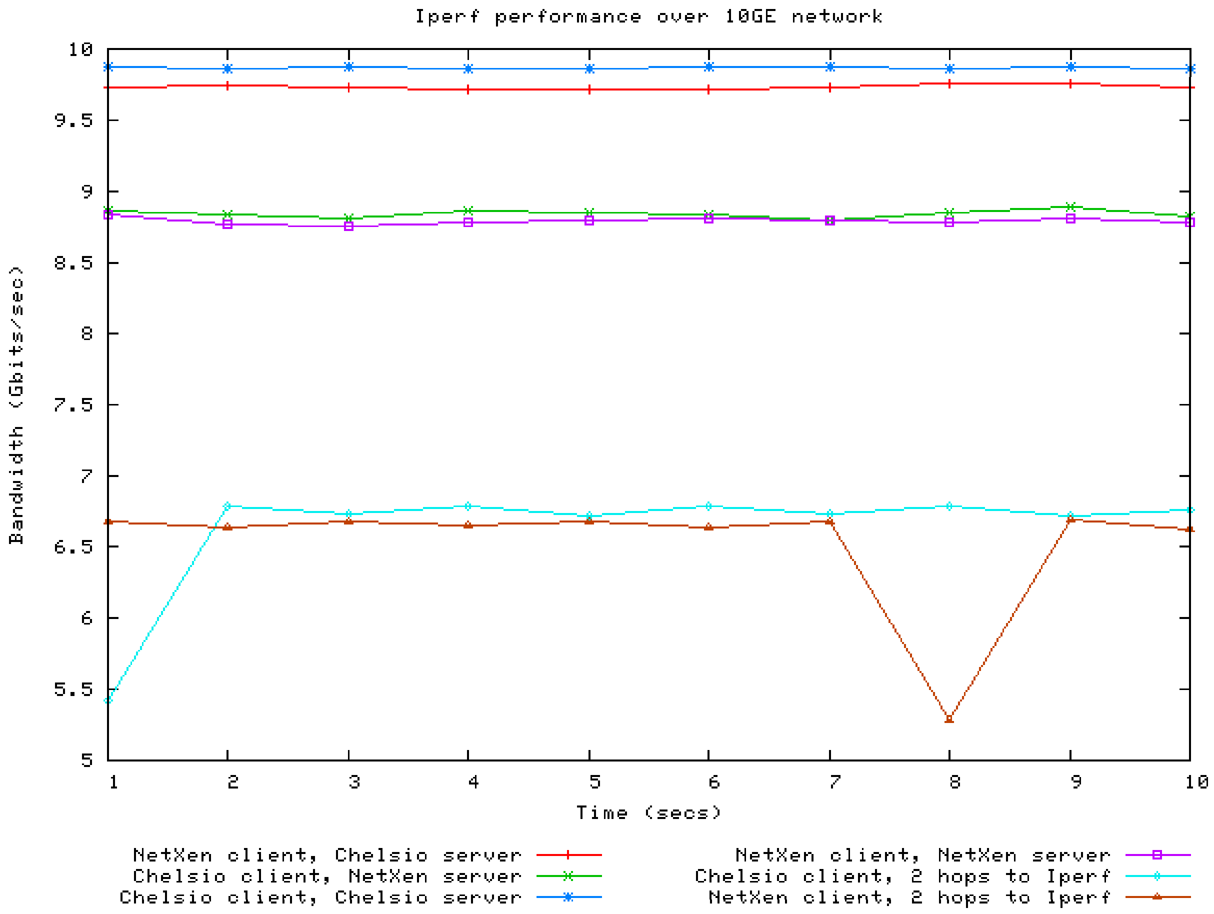
I/O results from Planck nodes

- We have a “sub-cluster” for a group of users who want to run MPI codes.
- The cluster has Infiniband interconnect (low-latency internode communication) as well as 10GE connectivity (high-bandwidth connection to the global filesystem)
- Testing 10GE switch from Woven Systems
- Testing Chelsio and NetXen NICs
- 29 nodes gateway through 4 10GE nodes

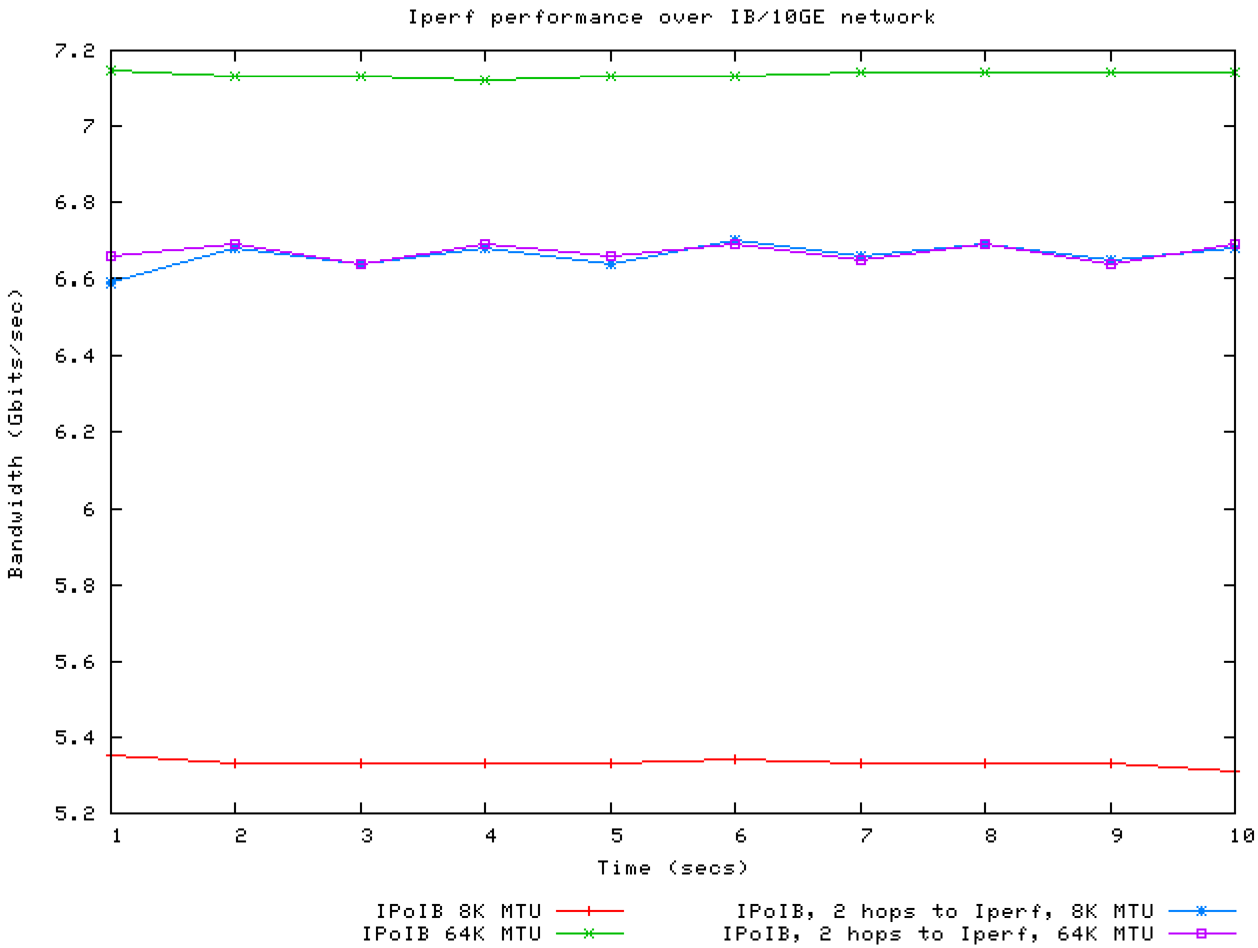
Infiniband "sub-cluster"



Iperf Tests using 10GE network

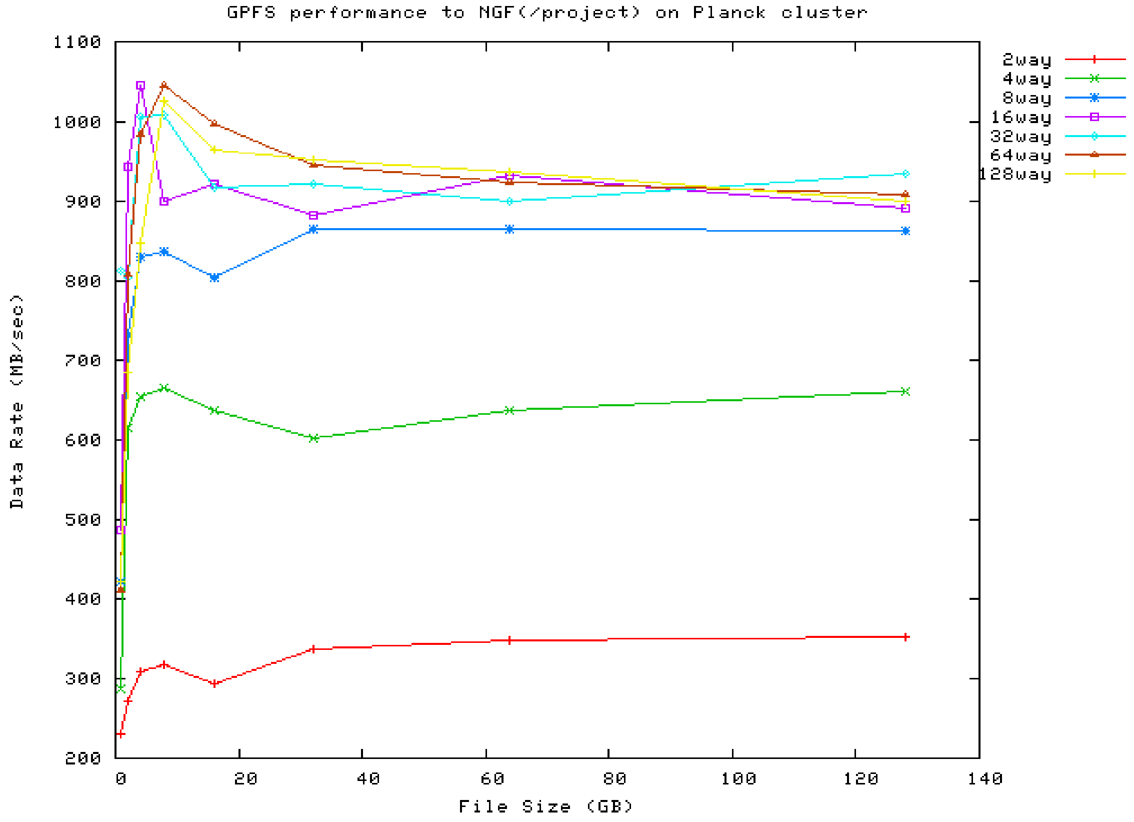


Iperf Tests using 10GE/IB network



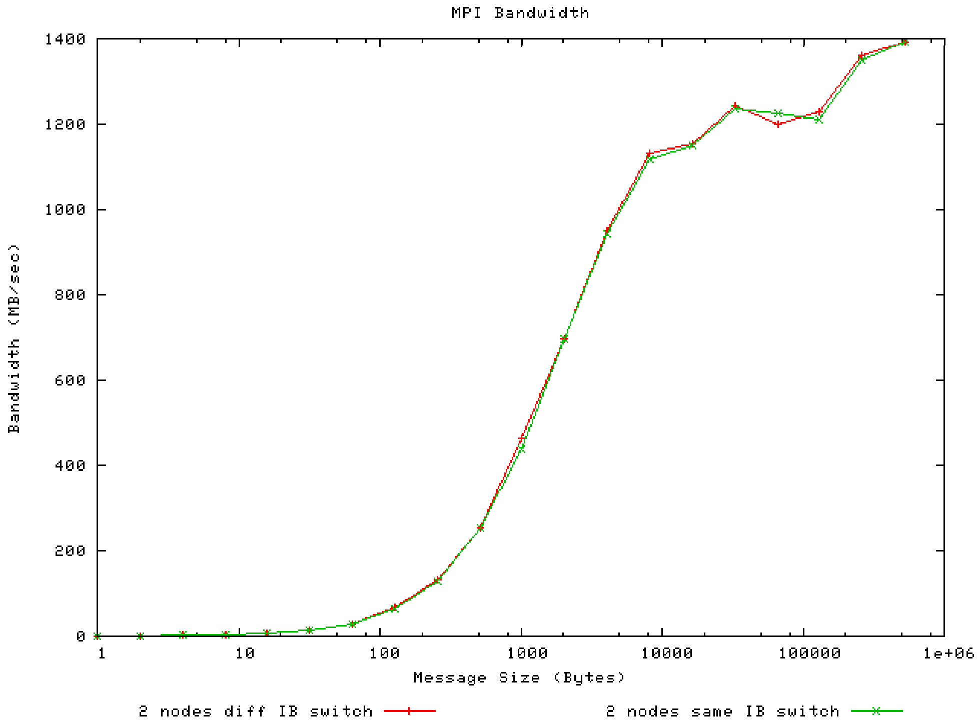
I/O tests using 10GE/IB network

- Posix writes to single file from multiple processes
- Similar performance on reads

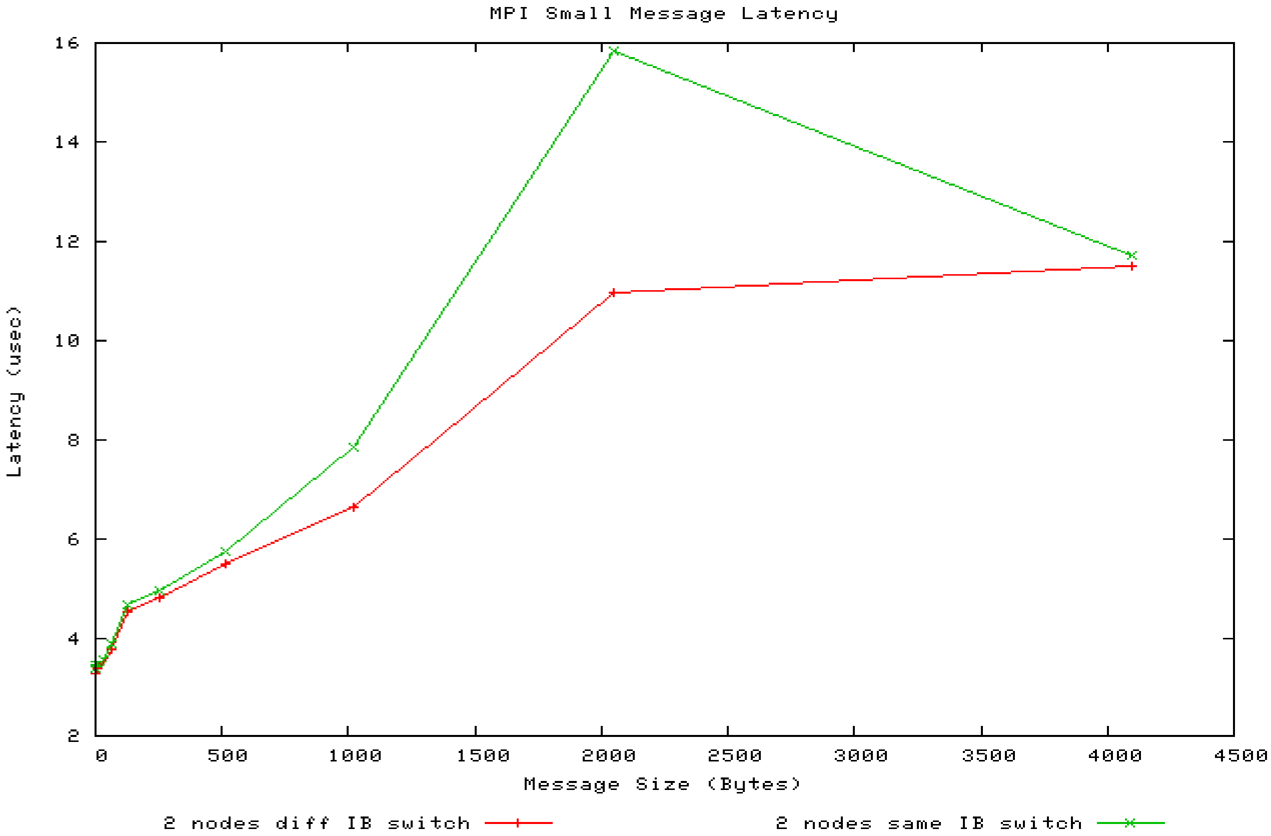


MPI performance

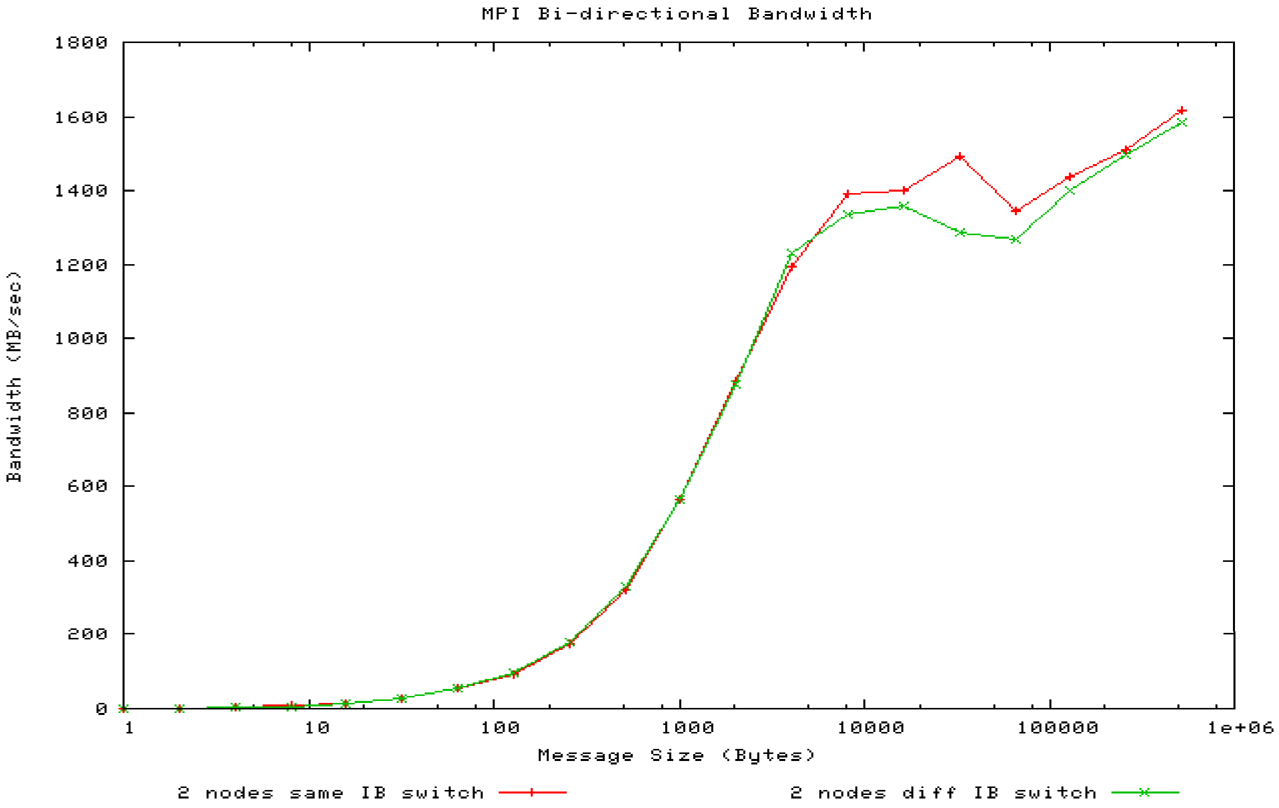
Bandwidth and latency tests using OpenMPI-1.2.5



MPI performance



MPI performance



Summary

- Indication of adequate performance of the quad-core AMD architecture for the purpose of our workloads.
- Need to upgrade network infrastructure to match the higher per rack core density.
- More testing needed……