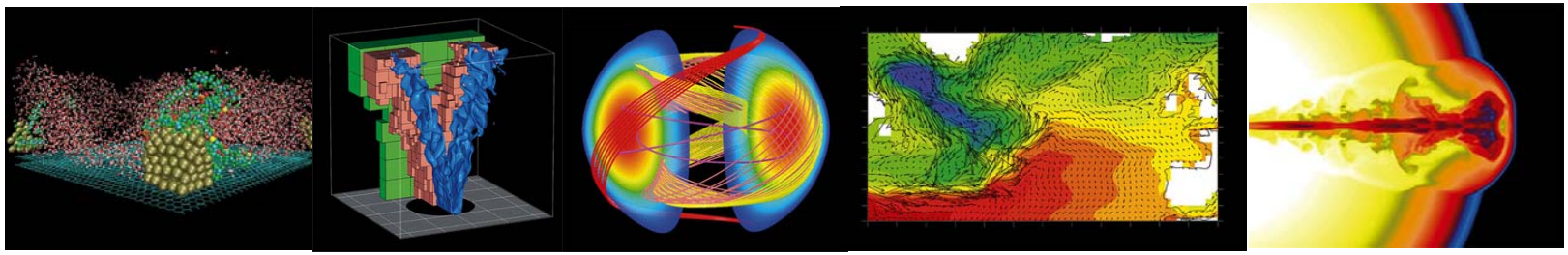


User and Performance Impacts from Franklin Upgrades

Yun (Helen) He
National Energy Research Supercomputing Center

Cray User Group Meeting
May 4-7, 2009

BERKELEY LAB

NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

Outline

- Franklin Introduction
- Benchmarks
- Quad Core Upgrade
- CLE 2.1 Upgrade
- IO Upgrade
- Summary

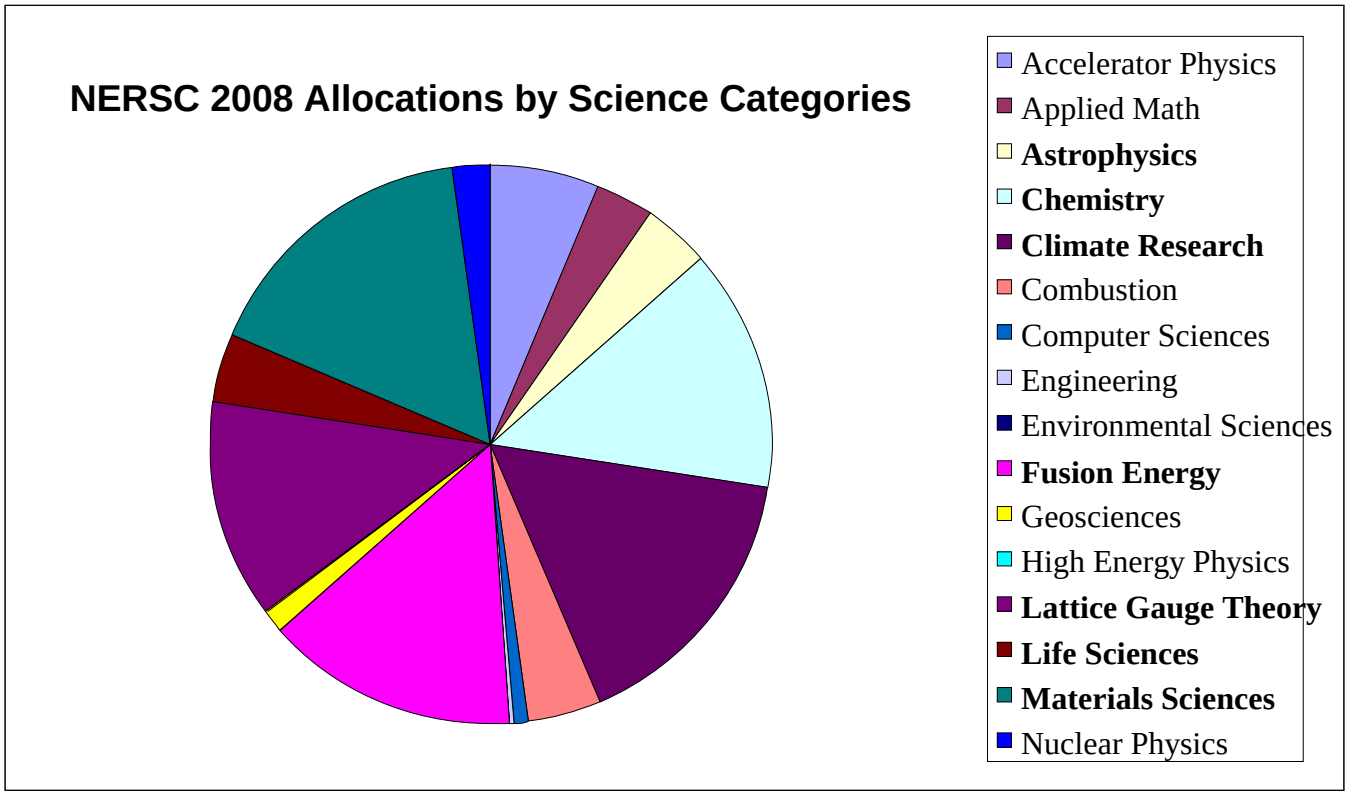
Franklin's Role at NERSC

- NERSC is the US DOE's keystone high performance computing center.
- Franklin is the "flagship" system at NERSC serving ~3,100 scientific users in different application disciplines.
- Serves the needs for most NERSC users from modest to extreme concurrencies.
- Expects significant percentage of time to be used for capability jobs on Franklin.

Kernel Benchmarks

- **Processor: NAS Parallel Benchmarks (NPB)**
  - **Serial: NPB 2.3 Class B**
    - **best understood code base**
  - **Parallel: NPB 2.4 Class D at 64 and 256 ways**
    - **Class D is not available with 2.3**
- **Memory: STREAM**
  - **Measures sustainable memory bandwidth**
- **Interconnect: Multipong**
  - **Maps out switch topology latency and bandwidth**
- **I/O: IOR**
  - **Exercise one file per processor or shared file accesses for common set of testing parameters**

Application Benchmarks Represent 85% of the Workload



- **Large variety of applications.**
- **Different performance requirements in CPU, Memory, Network and IO.**

Applications Summary

Application	Science Area	Basic Algorithm	Language	Library Use	Comment
CAM3	Climate (BER)	CFD, FFT	FORTRAN 90	netCDF	IPCC
GAMESS	Chemistry (BES)	DFT	FORTRAN 90	DDI, BLAS	
GTC	Fusion (FES)	Particle-in-cell	FORTRAN 90	FFT(opt)	ITER emphasis
MADbench	Astrophysics (HEP & NP)	Power Spectrum Estimation	C	Scalapack	1024 proc. 730 MB per task, 200 GB disk
MILC	QCD (NP)	Conjugate gradient	C	none	2048 proc. 540 MB per task
PARATEC	Materials (BES)	3D FFT	FORTRAN 90	Scalapack	Nanoscience emphasis
PMEMD	Life Science (BER)	Particle Mesh Ewald	FORTRAN 90	none	

Quad Core Upgrade

- Upgraded during July to October 2008.
- Done in multiple phases in order to have maximum system availability and job throughput.
- Deliver >= 75% of the original computing power on the “production” system throughout the upgrade.
- Limited to 4 upgrade phases to minimize resource commitment and interruption for users.
- Reduce risk of problems by gradually increasing number of upgrade columns during each phase.
- 7-day full system production stabilization time between phases.
- Burn-in time (check out for failed nodes) for the upgraded modules and friendly user time on the “test” system.

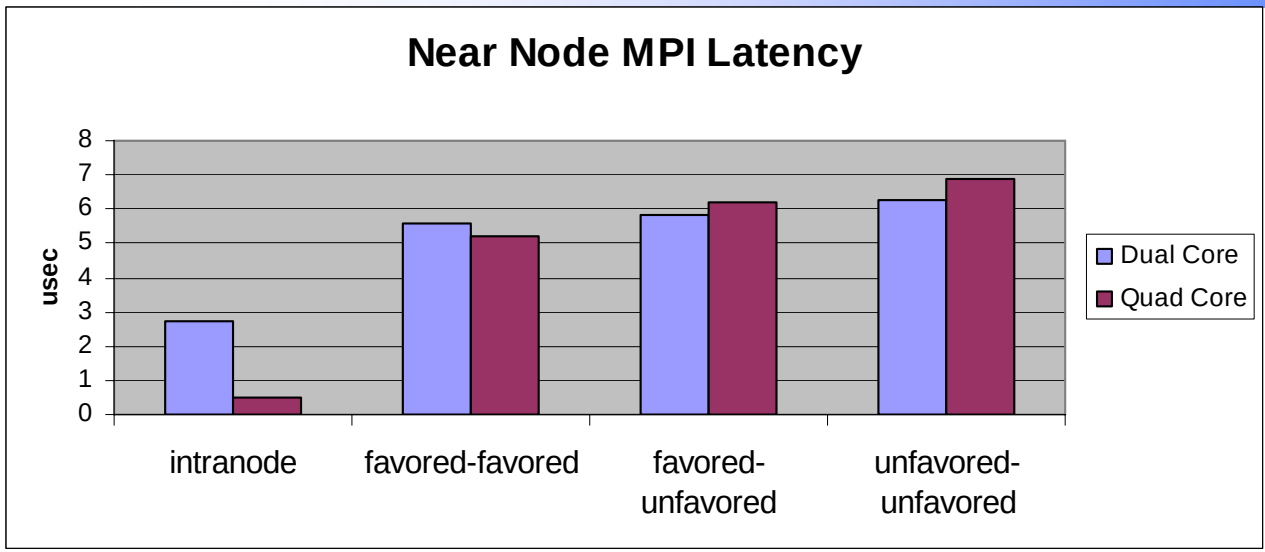
Franklin Configuration Change

	Dual Core Franklin	Quad Core Franklin
Compute nodes	9,660	9,660
Cores per node	2	4
Total compute cores	19,320	38,640
Processor core type	Opteron 2.6 GHz Dual Core	Opteron 2.3 GHz Quad Core
Theoretical peak per core	5.2 GFlop/sec	9.2 GFlop/sec
System theoretical peak (compute nodes only)	101.5 TFlop/sec	356 TFlop/sec
Physical memory per compute node	4 GB	8 GB
Memory usable by applications per node	3.75 GB	7.38 GB

User Impacts

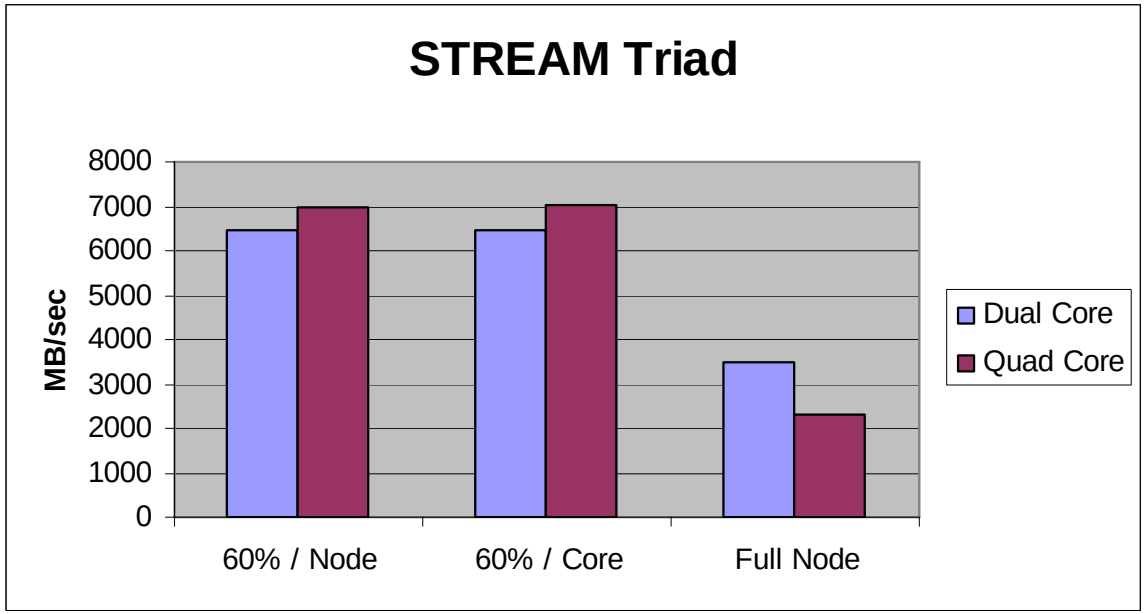
- CPU clock rate was reduced, but memory speed improved.
- Overall application performances (NERSC Sustained System Performance) are about the same.
- Quad core nodes were free during the early upgrade phases, and were only charging 2 cores/node for AY 2008.
- Average queue wait time longer when total nodes reduced. The situation became better after upgrade completed.
- During some phases, Franklin was a mixed dual core and quad core system. Users run on either dual or quad core nodes. Franklin was set to be quad core default when majority of compute cores were quad core.
- Users could experiment more with hybrid MPI/OpenMP.
- Higher UME rates in phase 3 upgrade, causing more job failures.

MPI Latency



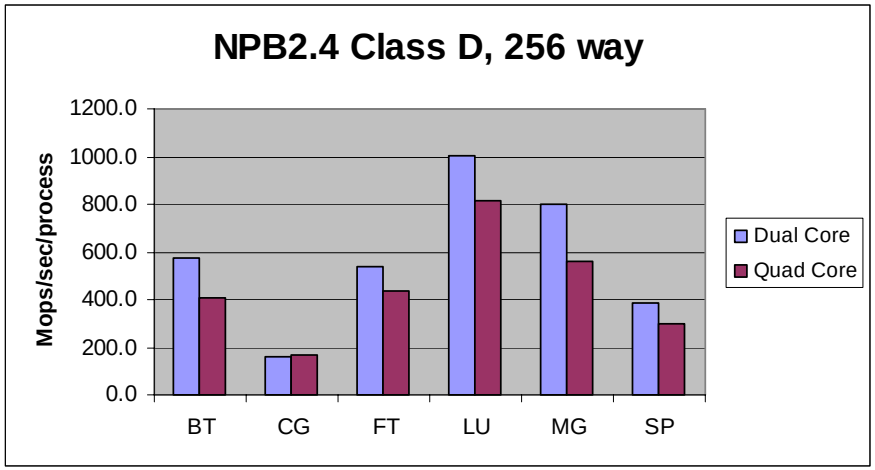
- One favored core per node.
  - 1 unfavored core for a dual core node
  - 3 unfavored core for a quad core node.
- Intranode improvement mainly from xt-mpt/3.
- Far node latency is about 1.9 us extra with the 3-D torus Franklin full configuration.
- 35 hops (between farthest nodes) * 0.053 us (per hop latency) = 1.855, rounded up to 1.9.

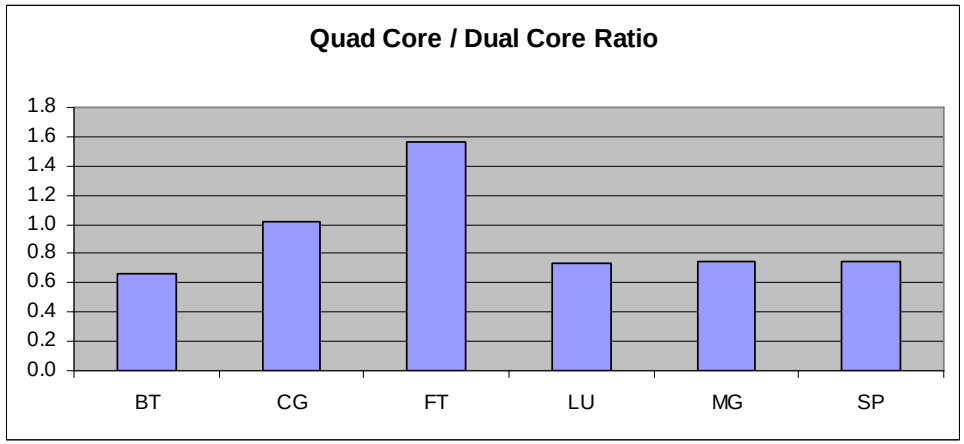
STREAM Benchmark Performance



- Measures sustained memory bandwidth
- Quad core higher:
  - Single core, use 60% node memory
  - Single core, use 60% core memory
- Quad core lower:
  - All cores, use 60% node memory

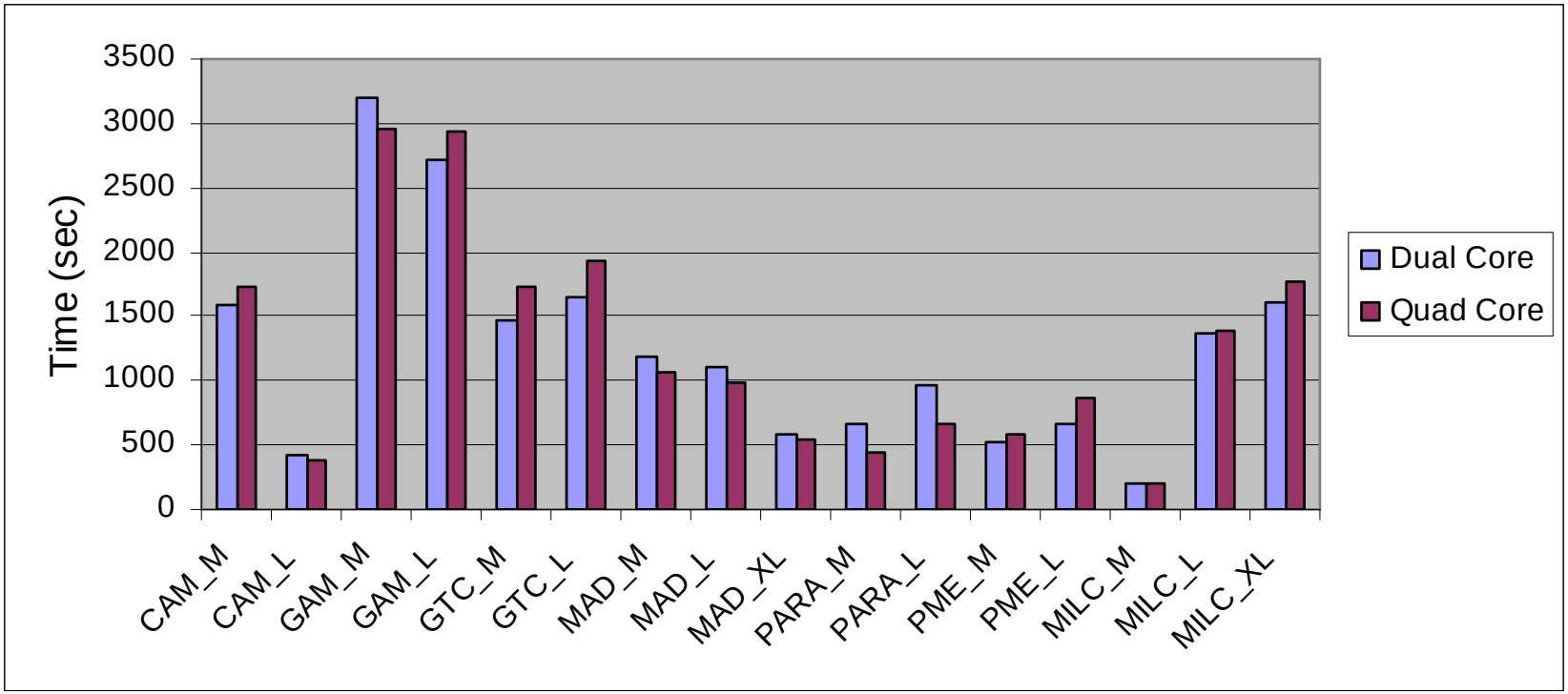
NPB Benchmark Performance





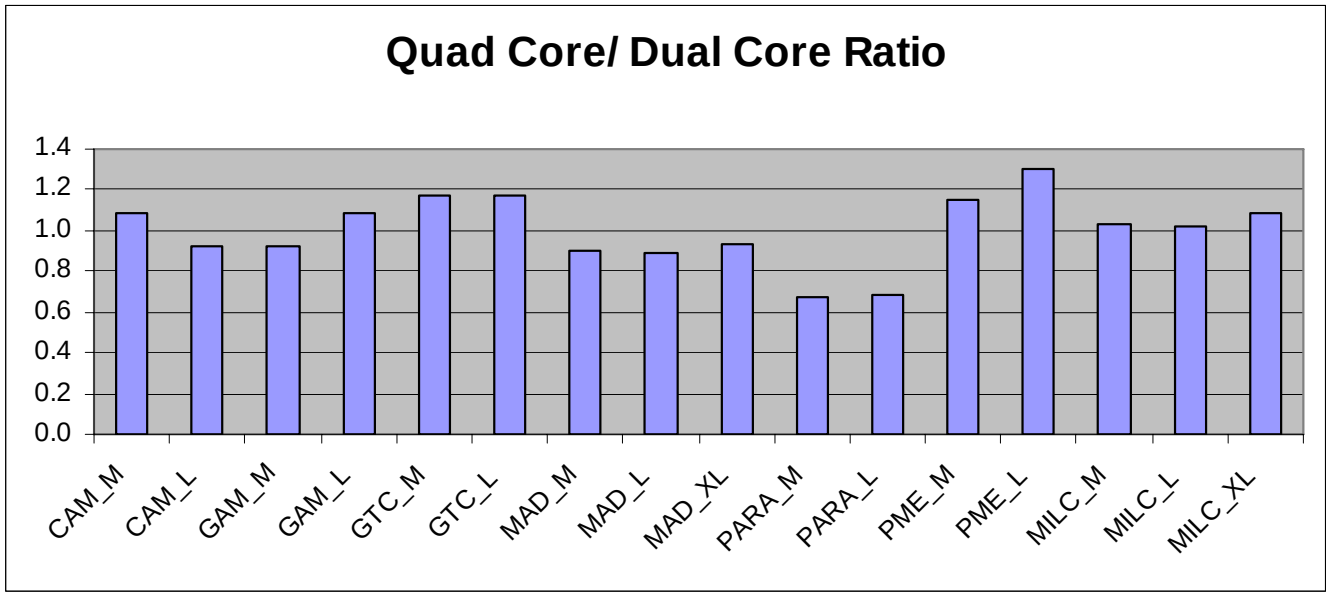
- NAS Parallel Benchmarks (NPB)
  - Serial: NPB 2.3 Class B
  - Parallel: NPB 2.4 Class D at 64 and 256 ways
- Quad core mostly slower except for CG.
- Dual core version highly tuned.

Application Benchmarks Performance



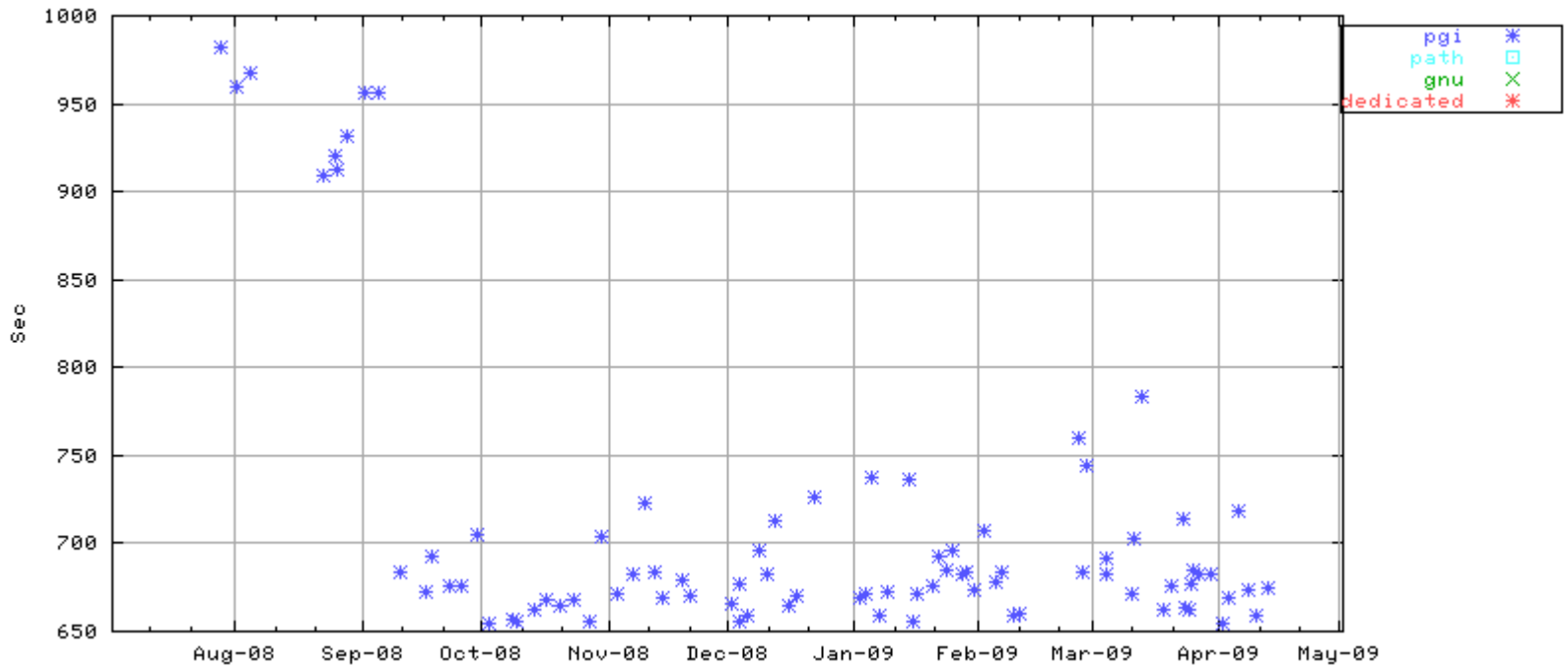
- **Medium: 64 cores, except CAM_M 56 cores.**
- **Large: 256 cores, except GAMESS_large 384 cores.**
- **XL: Madbench 1,024 cores, MILC 2,048 cores.**

Application Benchmarks Performance (cont'd)



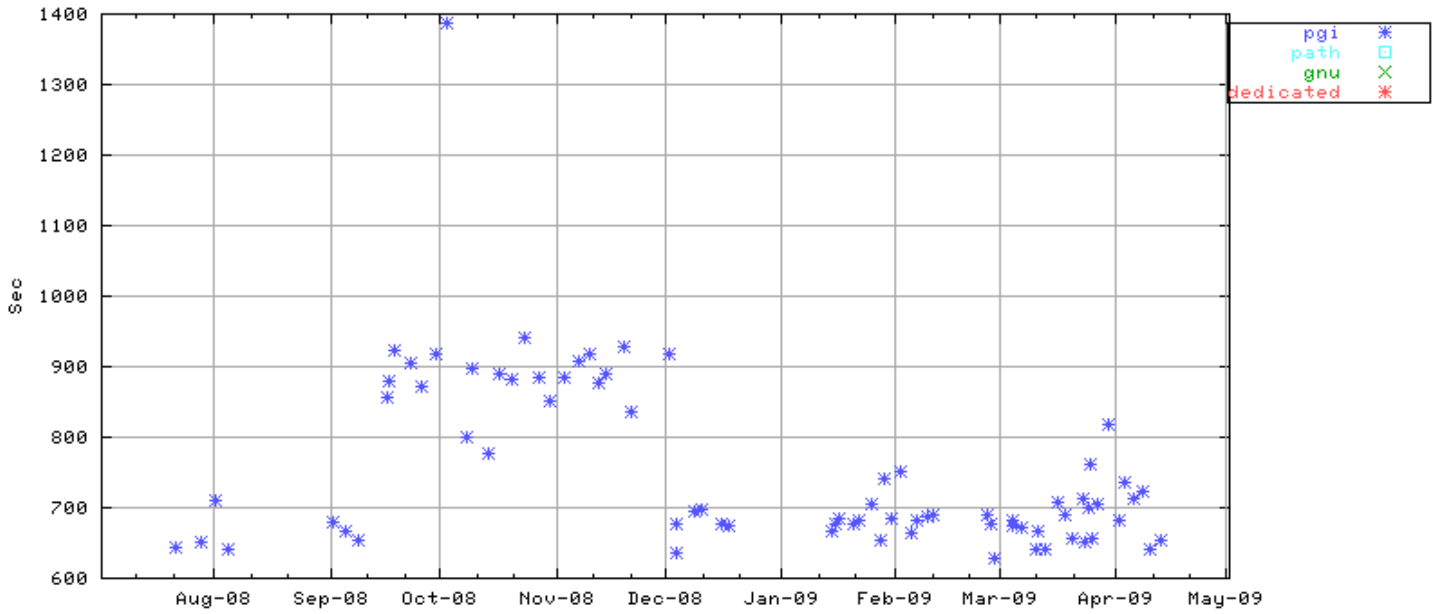
- Some applications are faster (Madbench, PARATEC)
- Some are slower (GTC, PMEMD).
- Most applications differ within 20% except PMEMD_large.
- PARATEC: >35% faster. Taking advantage of SSE128 optimization.
- MILC: quad core extra tuning.

Paratec_Large



- **Taking advantage of SSE128 optimization.**

PMEMD_Large



- Large amount of short communication messages.
- Sensitive to latency and memory caching effect.
- Performance degraded after Quad Core upgrade.
- Performance improved after CLE 2.1 upgrade.

CLE 2.1 Upgrade

- Upgraded on Dec 3-4, 2008.
- The quad core migration system was used by Cray and selected NERSC users as the CLE 2.1 test bed before upgrade.
- Major enhancements from CLE 2.0 to CLE 2.1 are:
  - Service nodes OS upgraded to SLES10 Linux Service Pack 1
  - Lustre File System upgraded from 1.4 to 1.6 release.
  - Kernel now includes Non-Uniform Memory Access.
  - Comprehensive System Accounting (CSA) open source software is supported.
  - OS now supports 2 MB huge pages as well as the default 4KB small pages.
  - System Resiliency Enhancements: System admin tools include new feature to recover from system or node failures.

User Impacts

- Overall application performances (NERSC Sustained System Performance) are about the same.
- Required all user codes to be recompiled with MPT3.
- NERSC implemented an aprun wrapper to make sure no MPT2 codes are executed on the compute nodes that would cause system issues.
- NWCHEM rebuilt with MPT3 compiled Global Array library.

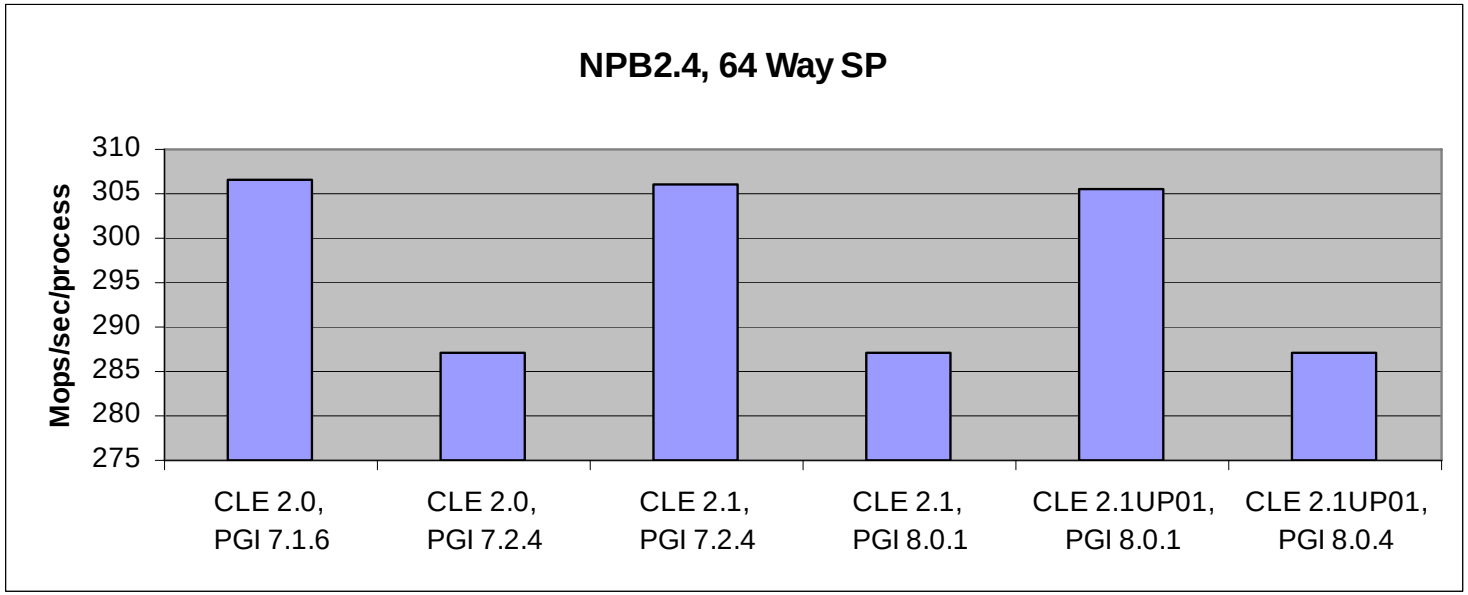
MPI Latency

- **Under CLE 2.0, measured latency between two nearest nodes could land in three different buckets:**
  - **one favored core and three unfavored cores within each quad core node.**
  - **favored/favored core pairs: average 5.46 *usec***
  - **favored/unfavored core pairs: average 6.09 *usec***
  - **unfavored/unfavored core pairs: average 6.74 *usec***
- **Significant latency changes in CLE 2.1 resulted from underlying portals software change.**
  - **No more favored/unfavored cores in each quad core node**
  - **Latency between any core pairs in two nearest nodes more uniform and averaged 6.46 usec.**

STREAM and NPBs Benchmark Performance

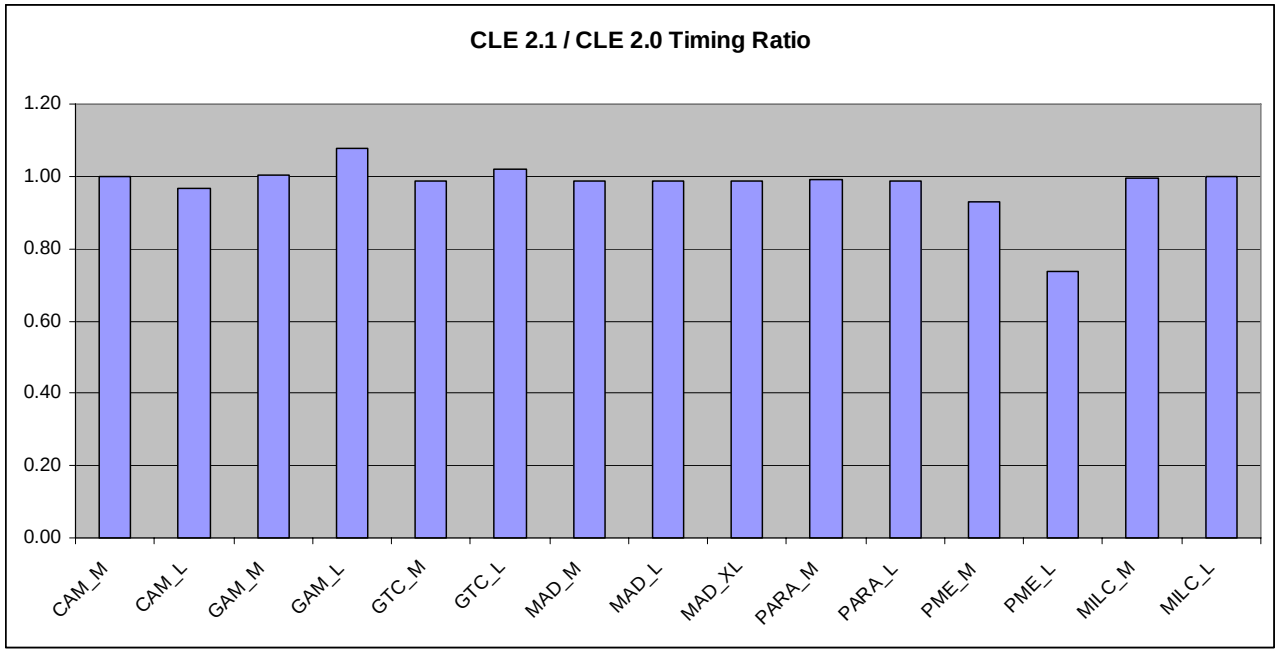
- Almost no performance difference in the memory benchmark STREAMS TRIAD operation:
  - 60% memory of each node
  - 60% memory of each core
  - full node
- No noticeable performance differences for all NPB benchmarks, except the NPB 2.4, 64-way SP.
  - The SP performance has been seen to be very sensitive to compiler options and user environment changes.

NPB 2.4, Class D, 64 Way SP



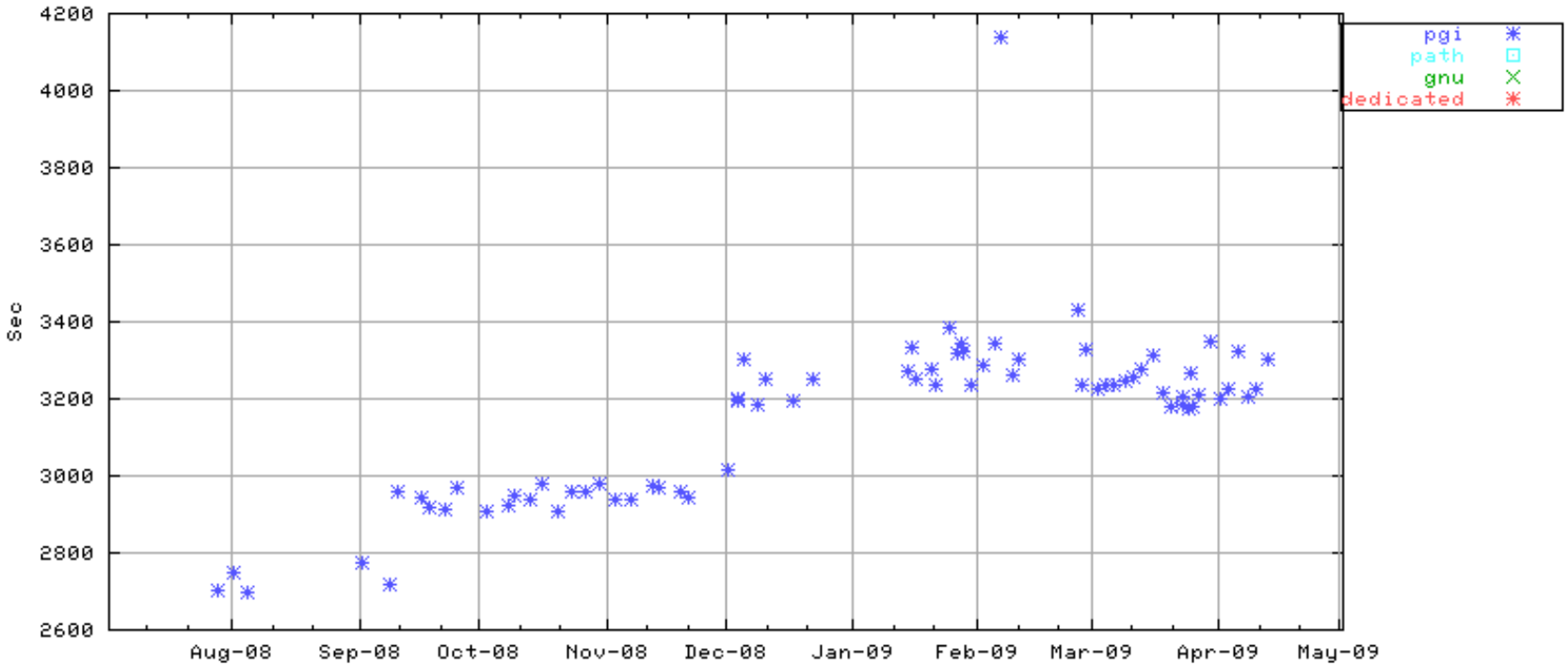
- **Performance swings between 306 and 287 Mops/sec/process with the OS level and compiler version changes.**

Application Benchmarks Performance



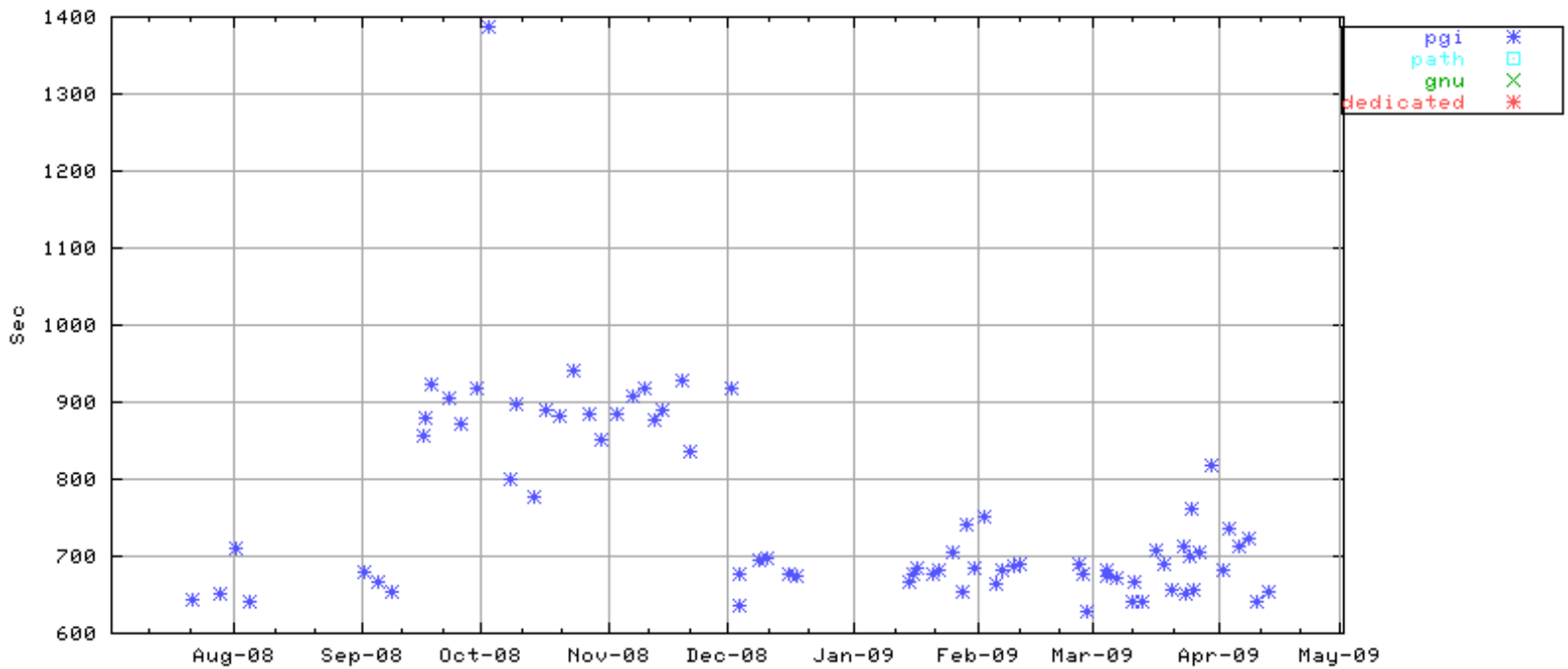
- Most applications see within 3% performance differences.
- GAMESS Large is 8% slower.
- PMEMD Medium is 7% faster
- PMEMD Large is 26% faster.

GAMESS_Large



- **Slowdown due to the message passing library (ddi2) used in the application is not quad core optimized.**

PMEMD_Large



- Large amount of short communication messages.
- Sensitive to latency and memory caching effect.
- Performance degraded after Quad Core upgrade.
- Performance improved after CLE 2.1 upgrade.

IO Upgrade

- Completed during mid March to early April 2009.
- Upgraded the interactive network adapters (PCI to PCI-e) to improve network performance between the interactive nodes and other NERSC systems, including NGF (/project).
- Doubled the number of I/O service nodes and upgrade their networking cards (PCI to PCI-e) to improve scratch IO performance.
- Separation of login and batch management (MOM) nodes.
- Reformat /scratch file system. Introduce a new /scratch2 file system.
- Installed service nodes for Data Virtualization Services (DVS) to be able to export NGF (/project) directly to compute nodes later this year.

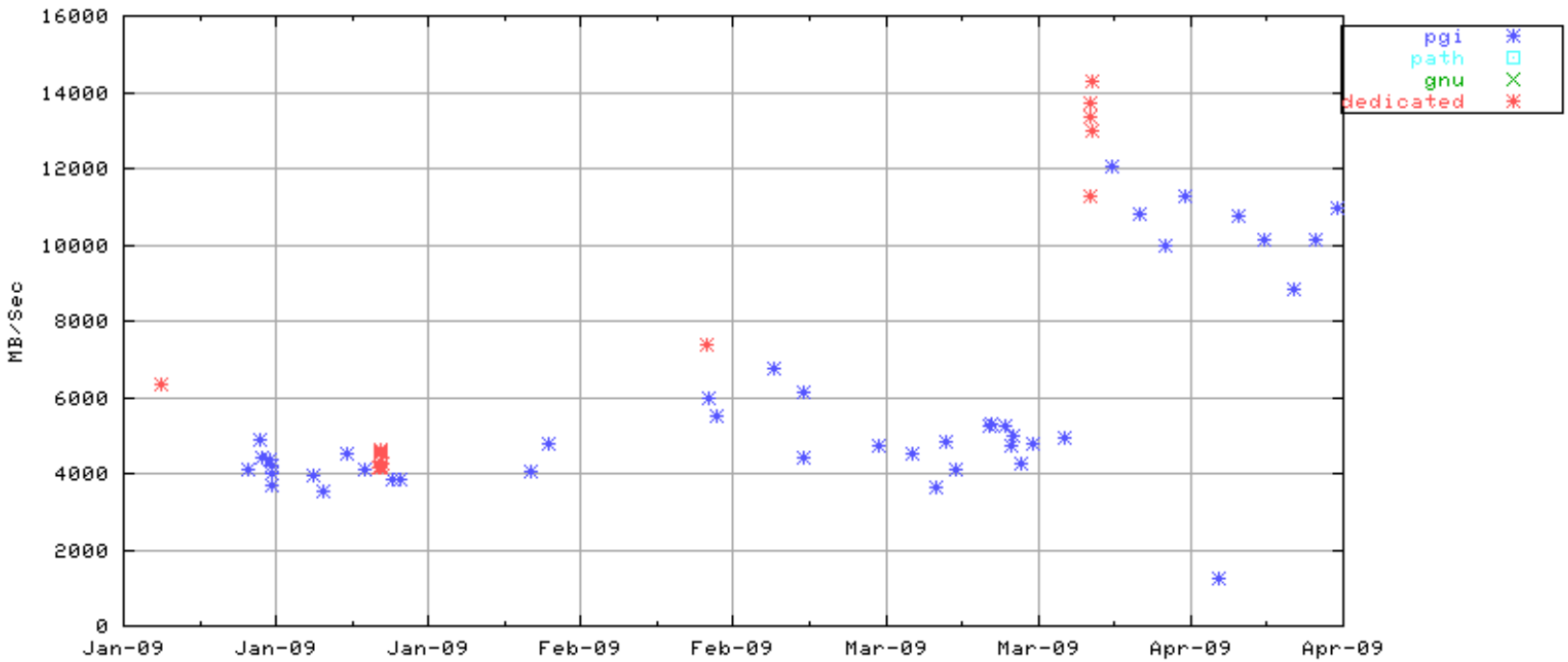
User Impacts

- Significant IO performance improvement.
- Heavy IO applications performance improvement.
- Weekly maintenances during upgrade.
- Users' data lost due to /scratch file system reformat (with enough advance notices)
- Having two file systems allows less impact on one file system when there is IO contention on the other one.
- Better interactive response on the login nodes.
- Separation of login and MOM nodes prevents more user jobs failures due to login node crashes.

IO Configuration Changes

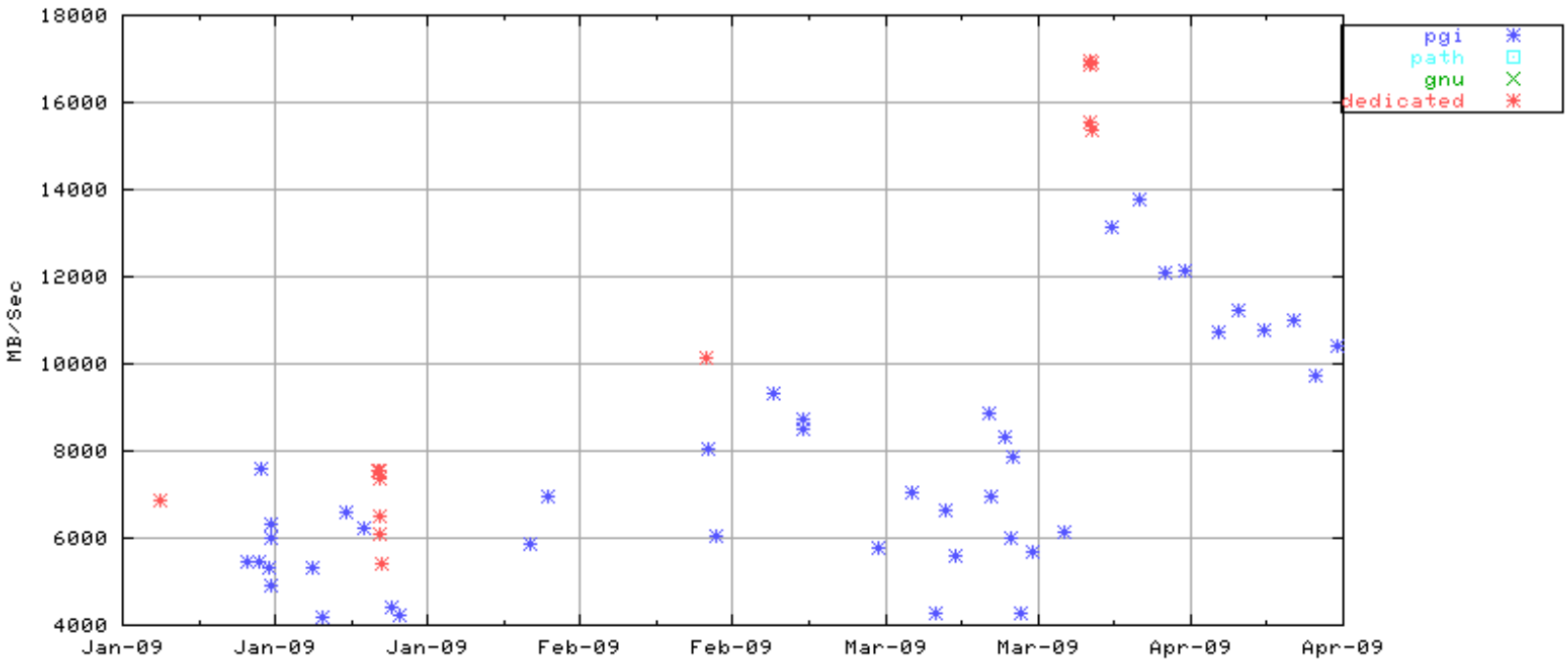
	Before IO Upgrade	After IO Upgrade
Compute Nodes	9,680	9,592
Login Nodes	10	10
MOM Nodes	16 (also serve as login nodes)	6 (distinct)
I/O Server Nodes	32	56
DVS Server Nodes	0	20
File Systems	/scratch	/scratch and /scratch2
Storage	346 TB	420 TB (210 TB each)
OSS	20	48 (24 each)
OST	80	96 (48 each)

IOR Aggregate Read



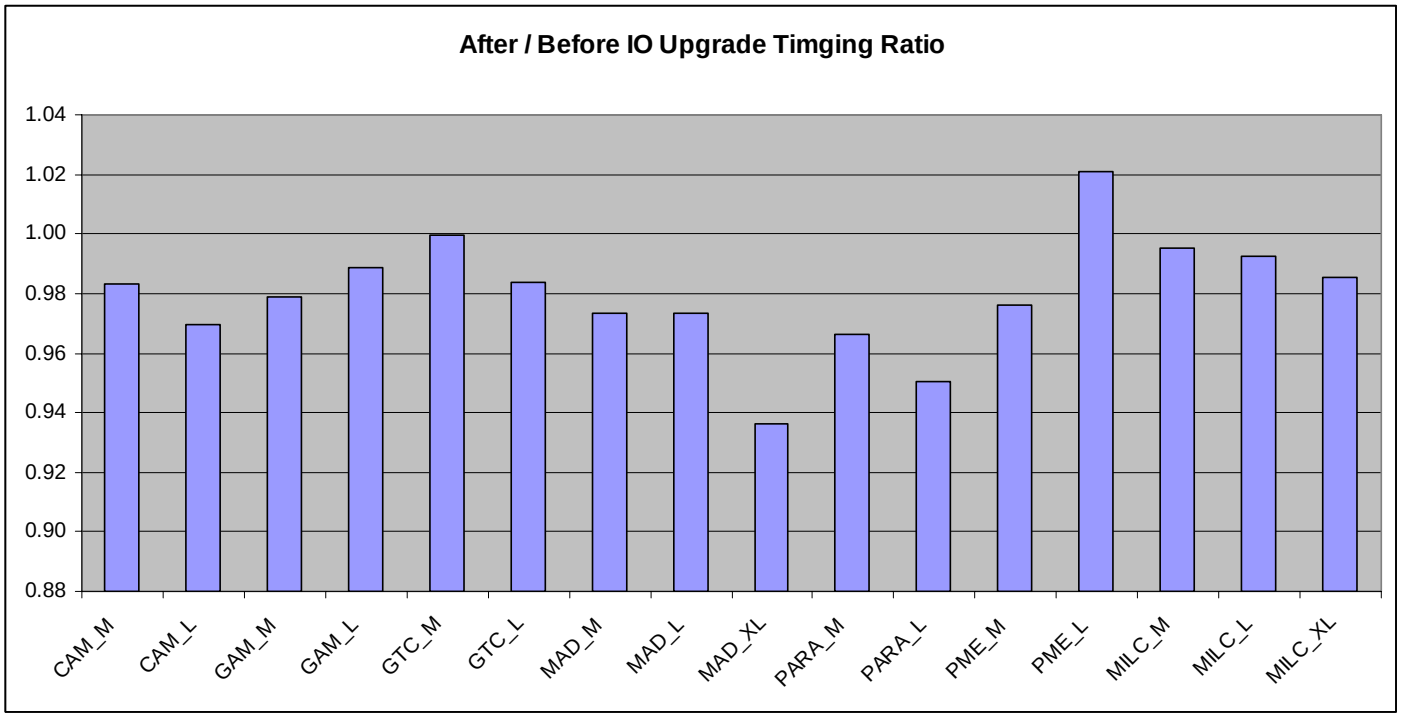
- Both the dedicated and production performance improved.
- Dedicated Aggregate Read improved from 7 to 14 GB/sec.

IOR Aggregate Write



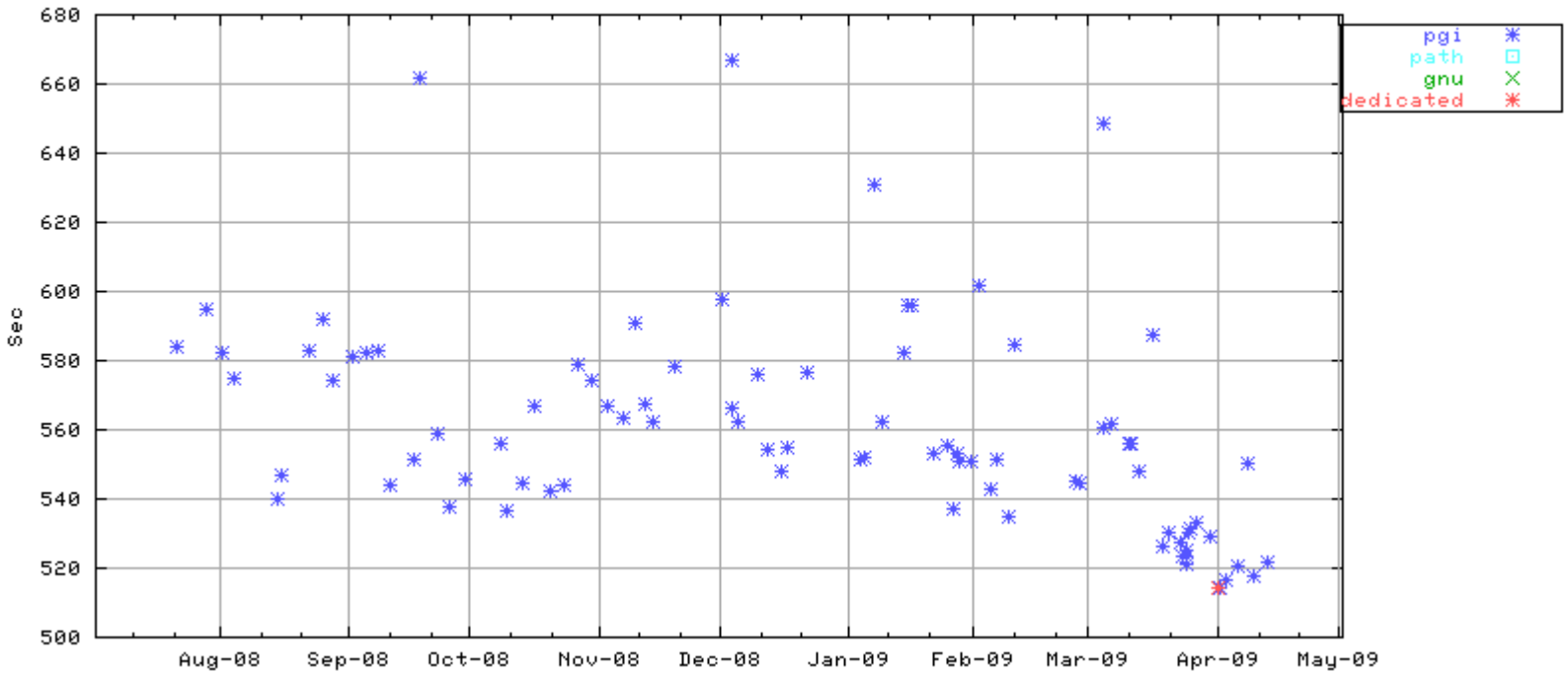
- Both the dedicated and production performance improved.
- Dedicated Aggregate Write improved from 10 to 17 GB/sec.

Application Benchmarks



- Most applications see slight (1~3%) performance improvement.
- MADBench Xlarge is 6% faster.
- Paratec Large is ~5% faster.
- PMEMD Large is 2% slower.

MADBench_Xlarge



- **MADBench is a heavy IO application.**

Summary

- Franklin has undergone three major upgrades during the last year.
- Service interruptions were minimal during upgrades
  - Users had free to half charging discounts
- Although users had to adapt to programming environment changes, the end results are worthwhile:
  - Doubled the system size and deliverable computing cycles
  - More than doubled the aggregate IO performance
  - More stable system
  - More potential system functionalities:
    - DVS, Checkpoint/Restart

Acknowledgement

- Cray support teams (on-site and remote) for Franklin.
- NERSC management, User Services Group, Computational Systems Group, File Systems Group, and N5 Procurement Team.
- Valuable feedbacks from NERSC users.
- Author is supported by the Director, Office of Science, Advanced Scientific Computing Research, U.S. Department of Energy.