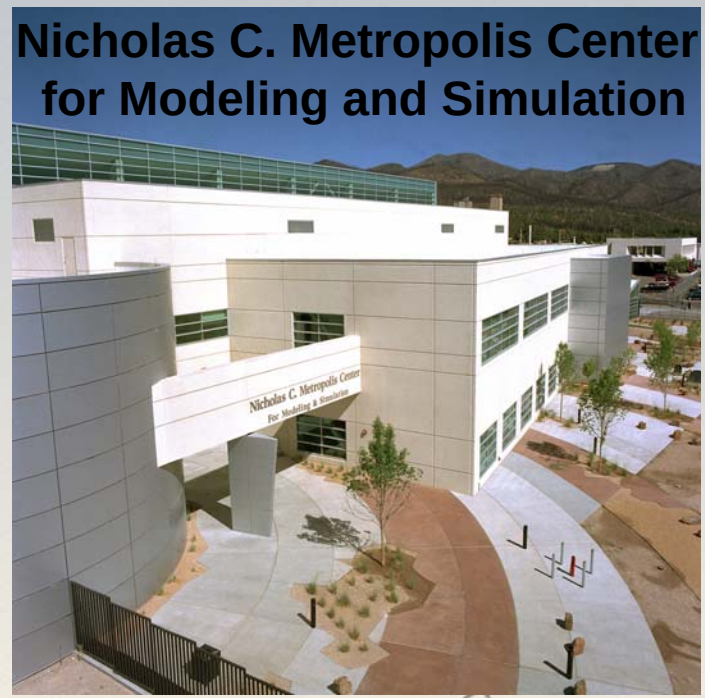
The ASCI Q System at Los Alamos National Laboratory

Conference on High Speed Computing
April 21-24, 2003



John Morrison
CCN Division Leader

Outline

1. Overview of Q System
2. Processor and Memory Subsystem
3. Interconnect
4. File System
5. Archival Storage
6. Applications Performance
7. Science Run Results
8. Summary

1. Overview of Q System

Overview

- Planned total — 20.48 TeraOPS
- Systems — 2048 AlphaServer ES45s
  - 8,192 EV-68 1.25-GHz CPUs with 16-MB cache
- Memory — 22 Terabytes
- System Interconnect
  - Dual Rail Quadrics Interconnect
  - 4096 QSW PCI adapters
  - Four 1024-way QSW federated switches
- Global Storage
  - 6144 – 72 GB Fiber Channel disk drives
  - 442-Terabytes global disk

Q is Operational for Stewardship Applications (1st 10T)

- Many ASCI applications are experiencing significant performance increases over Blue Mountain.
- Linpack performance run of 7.727 TeraOPS (more than 75% efficiency)
- Initial user response is very positive (with some issues!)

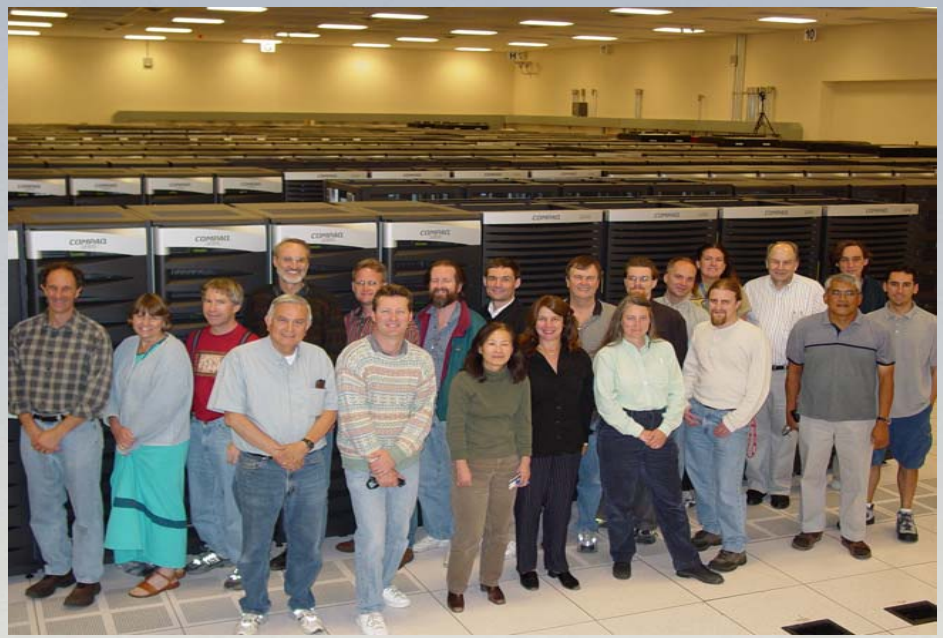
  (Users want more cycles…)
- Users from the tri-lab community are also using the system

- Available to users for Classified ASCI codes since August 2002
  - Smaller initial system available since April 2002
- Los Alamos has run its December 2002 ASCI Milestone calculation on Q

December 2002 ASCI Milestone Calculation on Q

- Shavano Project - Enhanced Primary Capability Milestone
- Summer 2002 - Q Machine needed for completing milestone
  - Performance increases of 6-10 times over Blue Mountain
- Milestone work started as soon as Q was first made available to users
- Over 1.1 million processor hours used, about 78 calendar days, without interrupts or code changes (BM would have required over 430 days)
- Series of 3-D runs using over 1000 processors, about 40 hours each
- Q Machine provided ahead of schedule to meet requirements of milestone

The second 10T of Q has completed a system checkout period



- Unclassified Science Runs provided work load
- Checkout of new software, storage firmware, file systems, LSF/RMS scale issues, Quadrics Interconnect, detailed performance testing, etc.
- Overall, helped the LANL/HP team move forward quickly in many areas
- Increased pool of expertise

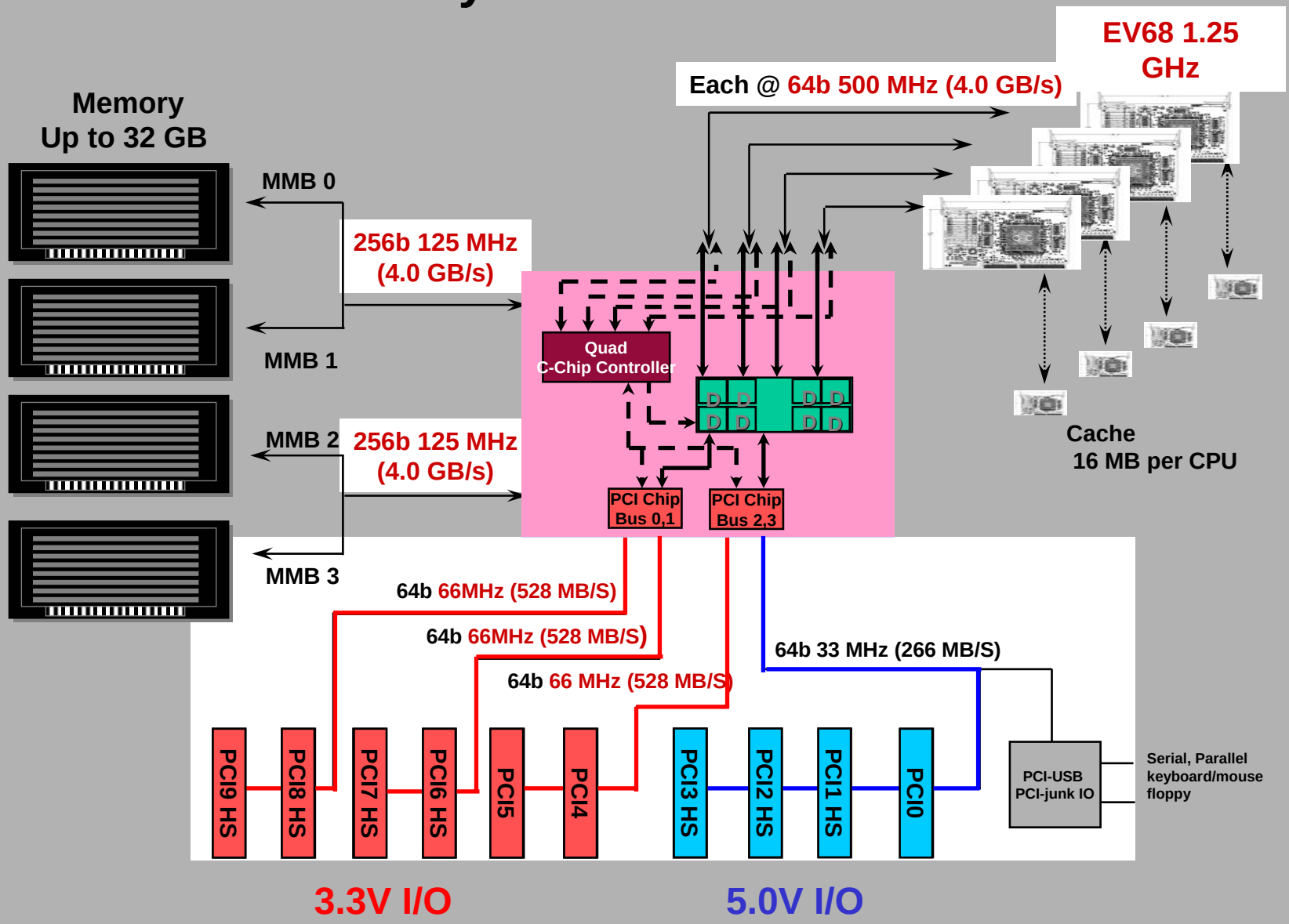
- System is now in the Secure environment
- The system is available for tri-lab use
- Demonstration runs scheduled for combined 20T system in April 2003

2. Q Processor and Memory Subsystem

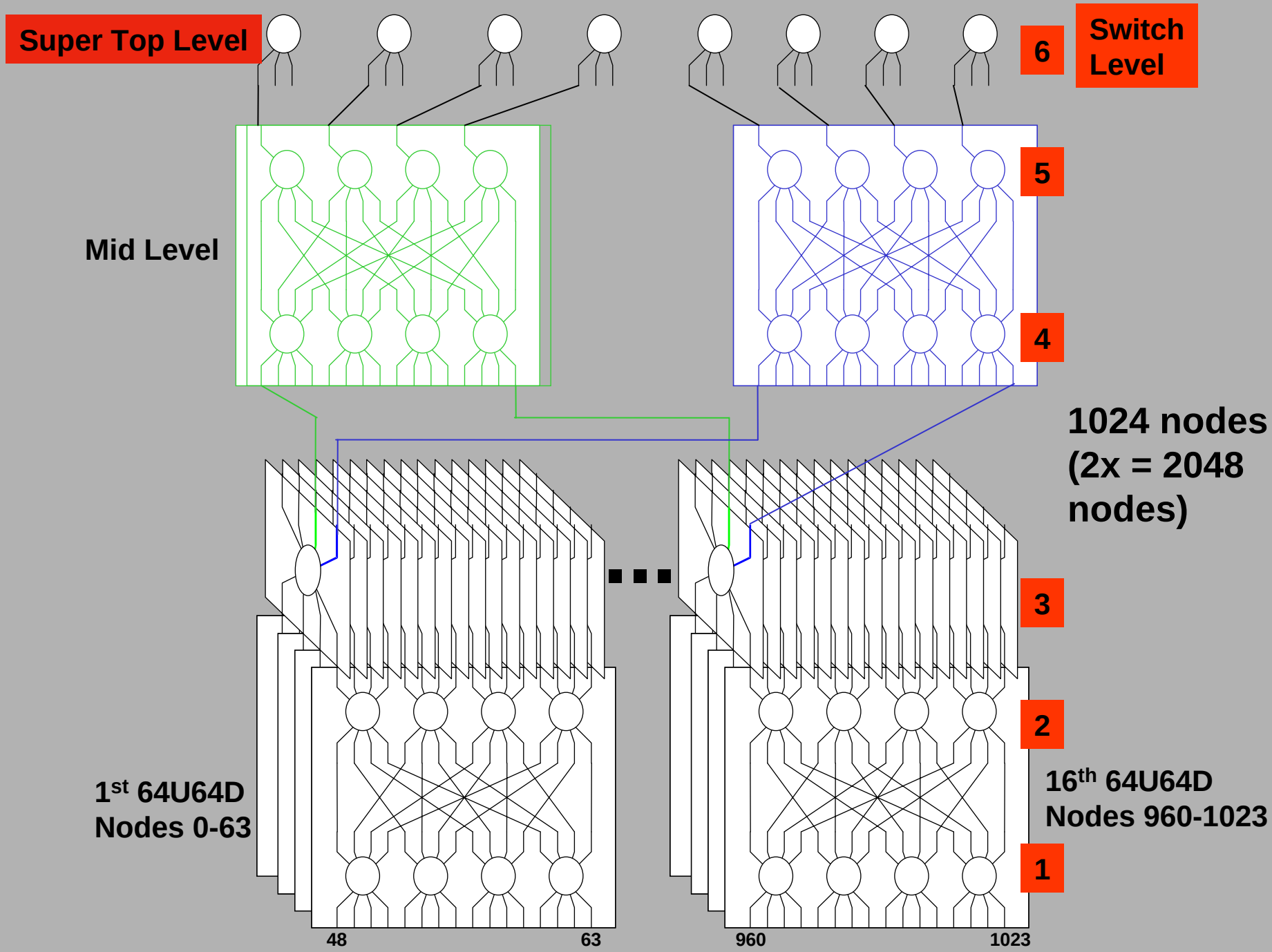
System Summary

- Alpha 21264 EV-68 processor
- AlphaServer ES45 SMP
  - 4 processor/SMP, 8/16/32 GB Memory/SMP
- Quadrics (QSW) dual rail switch interconnect
  - Fat-tree switch
  - High bandwidth (250 MB/s/rail), low latency(~5 us)
  - Will also handle File I/O traffic
- Switch-based Fibre attached Storage Arrays
  - RAID5 sets, 72 GB Drives
- AlphaServer SC and Tru64 Unix based

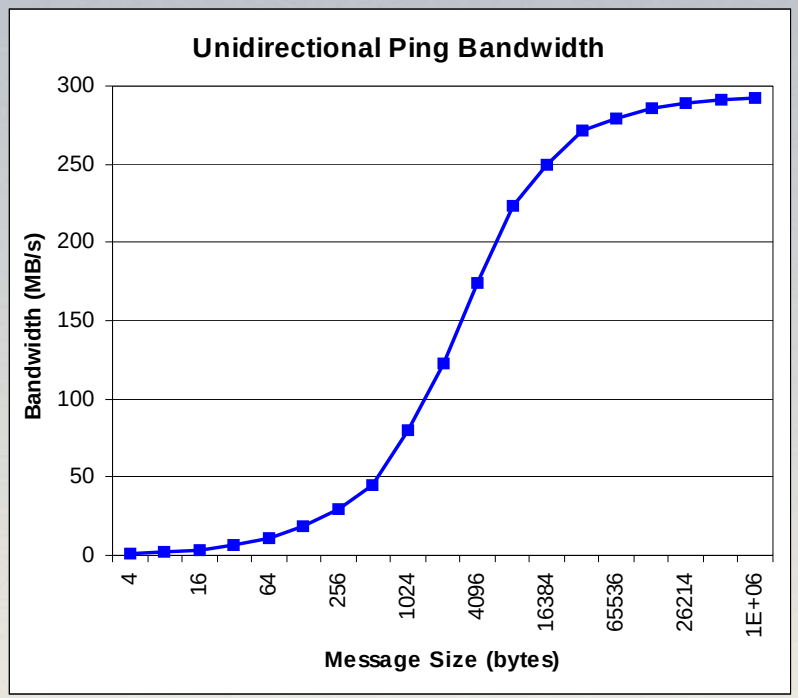
HP (Compaq) *AlphaServer ES45*
21264 System Architecture

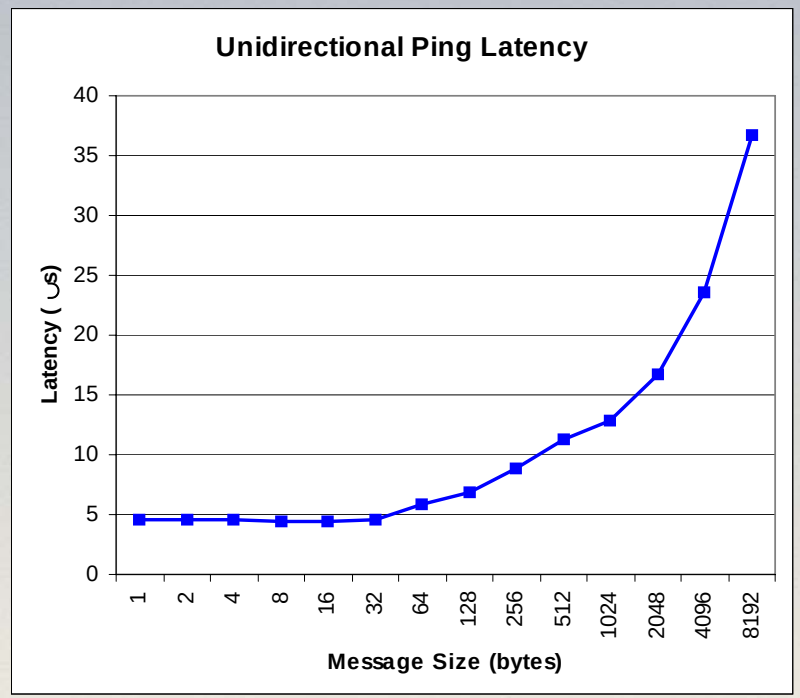


3. Q Interconnect

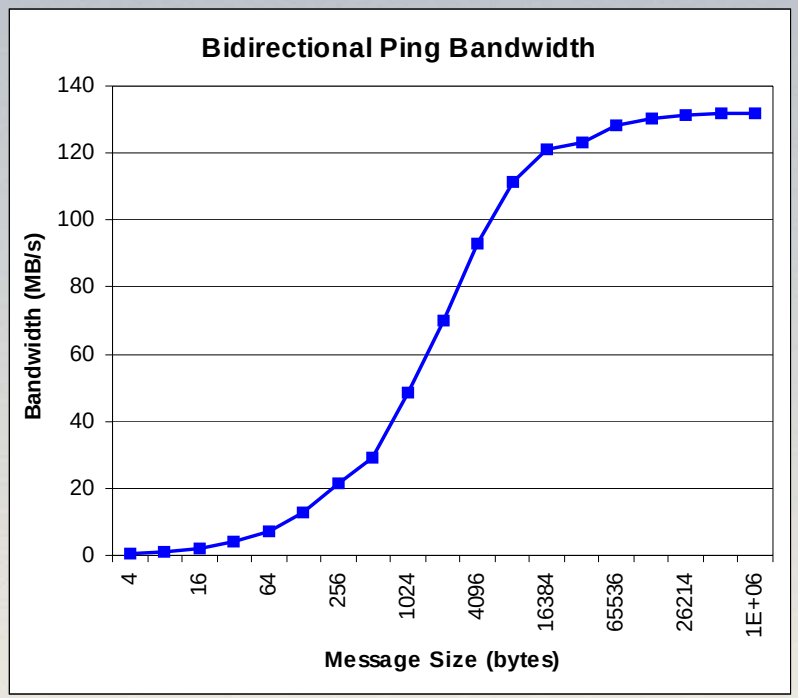

Super Top Level
6
Switch Level
Mid Level
5
4
1024 nodes (2x = 2048 nodes)
3
2
1st 64U64D Nodes 0-63
16th 64U64D Nodes 960-1023
1
48
63
960
1023

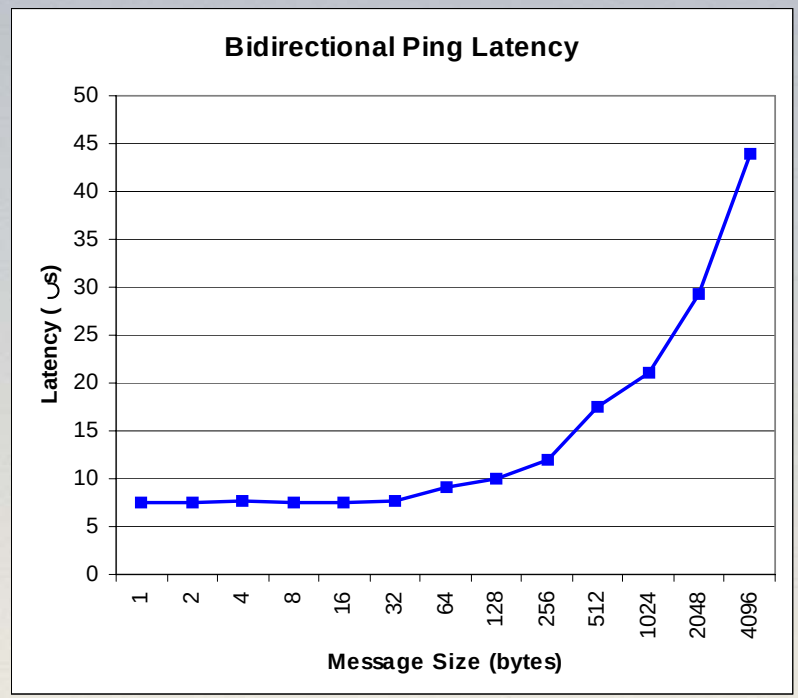
Quadrics Unidirectional Performance



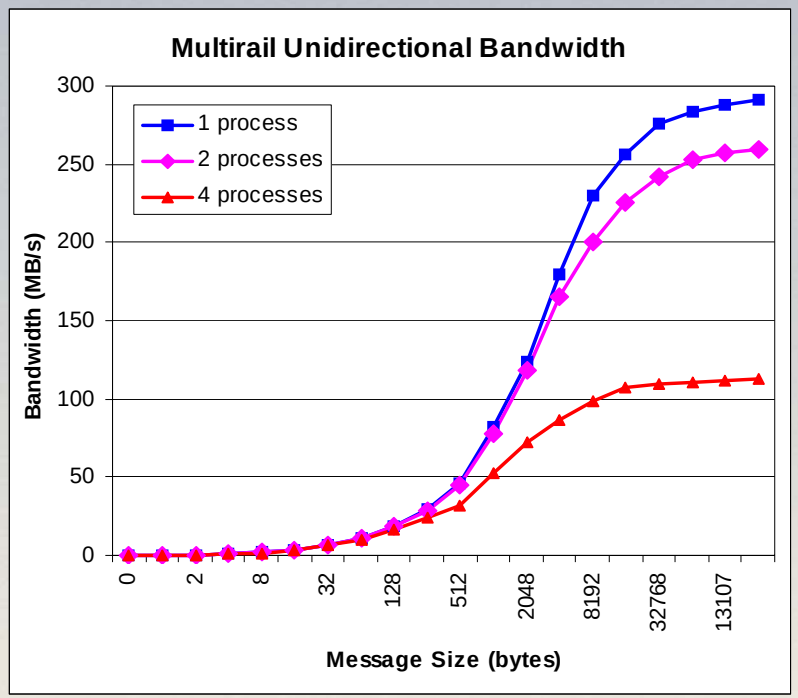


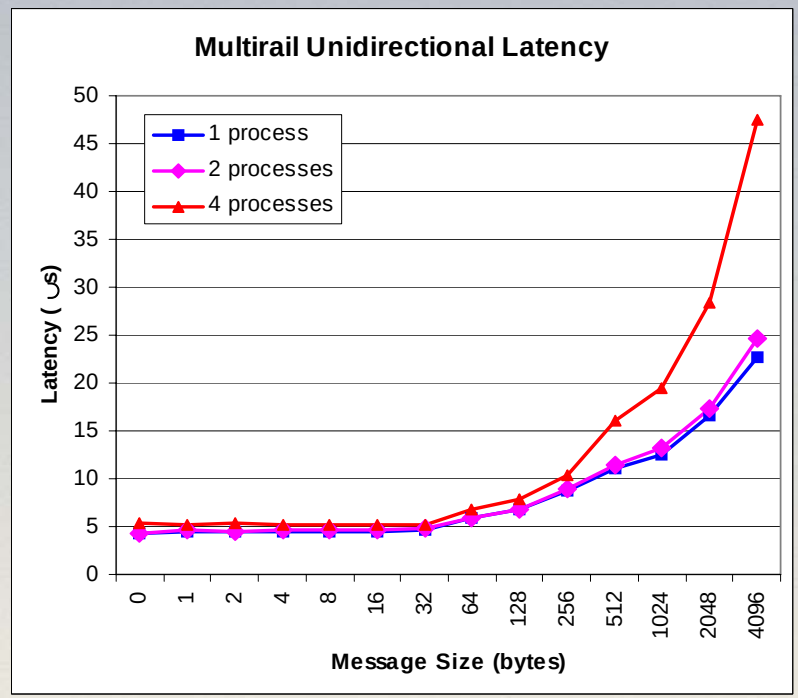
Quadrics Bidirectional Performance



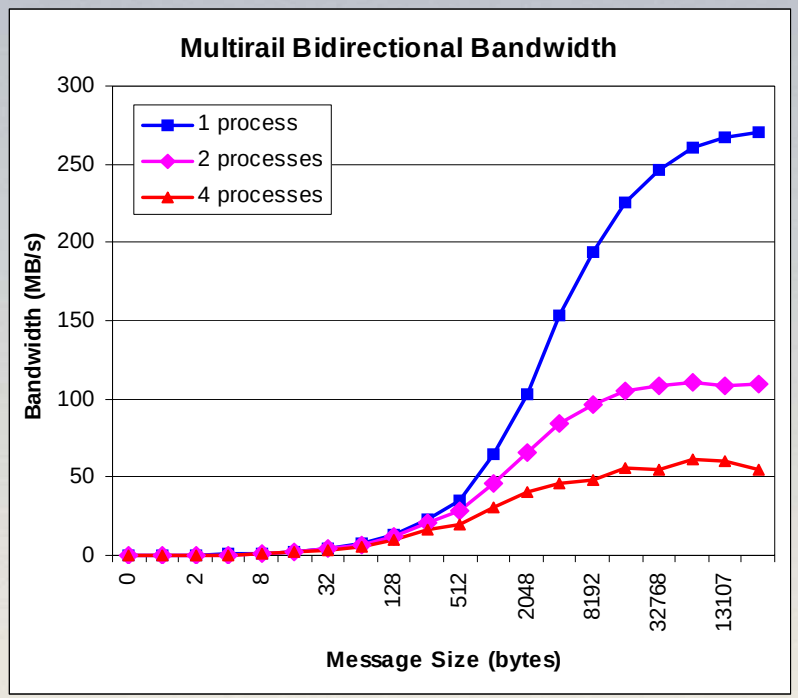


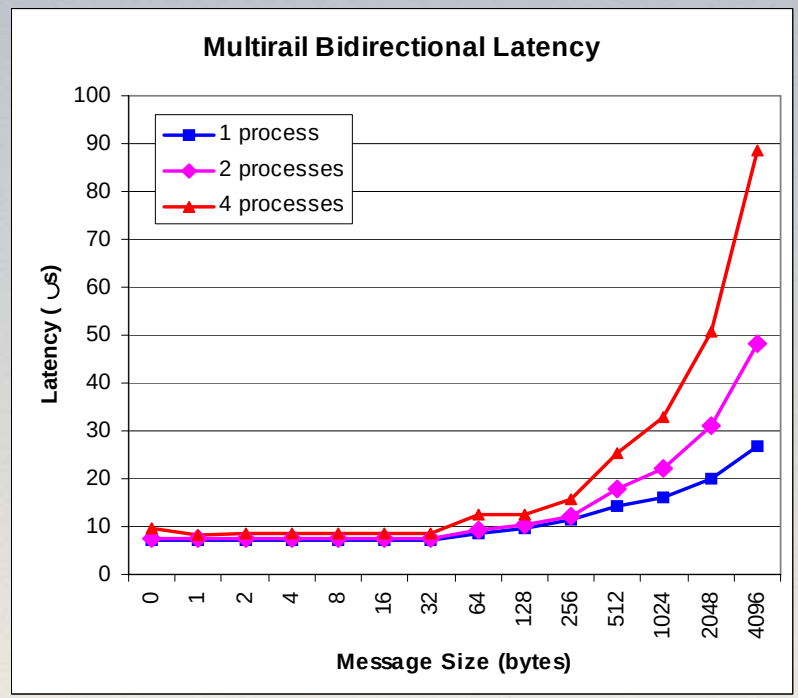
Quadrics Multirail Unidirectional Performance



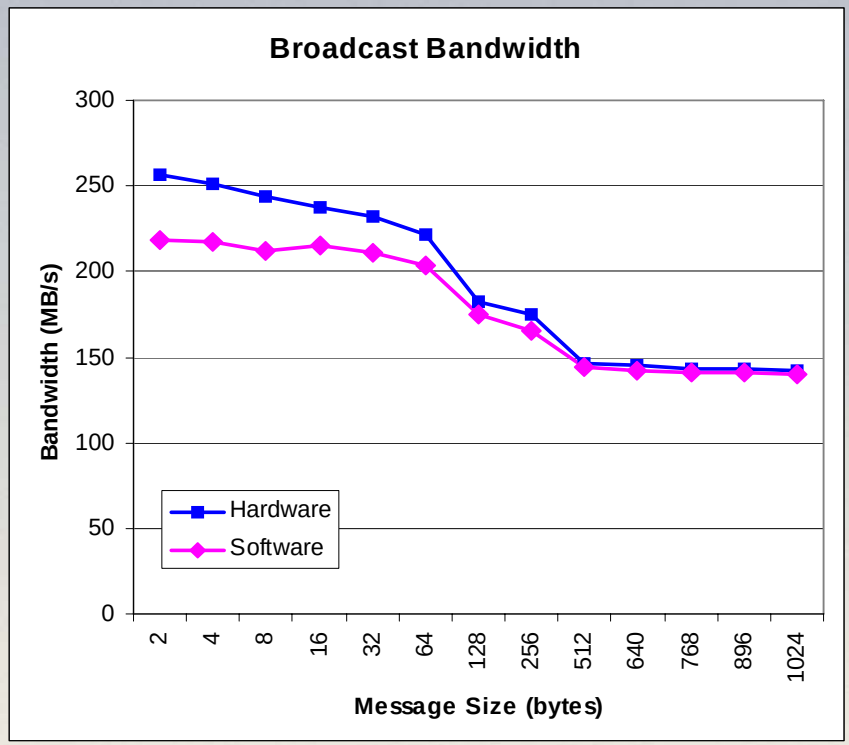


Quadrics Multirail Bidirectional Performance





Quadrics Broadcast Bandwidth



Bandwidth as a Function of Node Distance

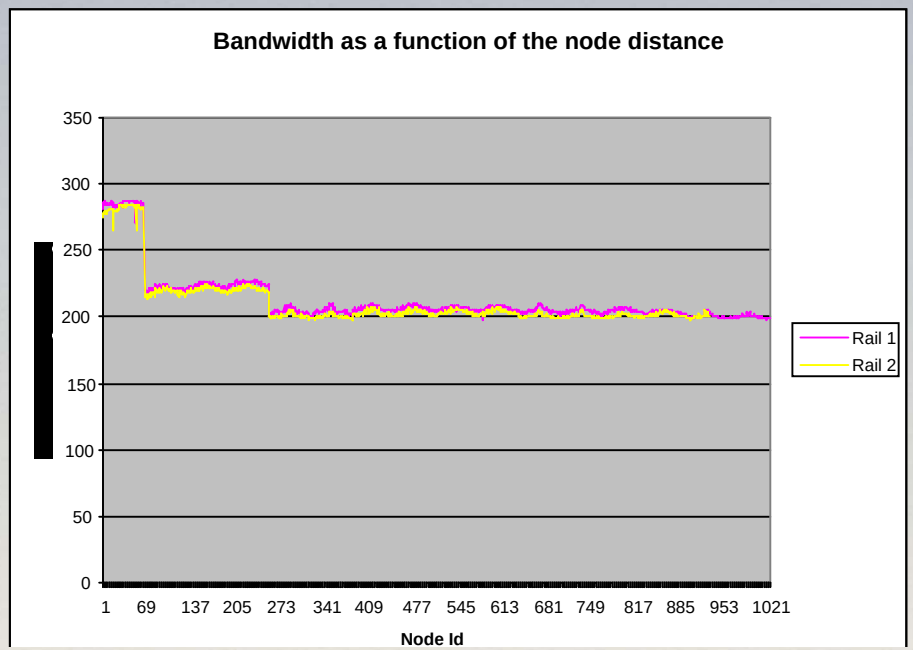


Figure 1: bandwidth observed on a ping from node 0 to all other nodes.

4. File System Performance

Contributed by Gary Grider

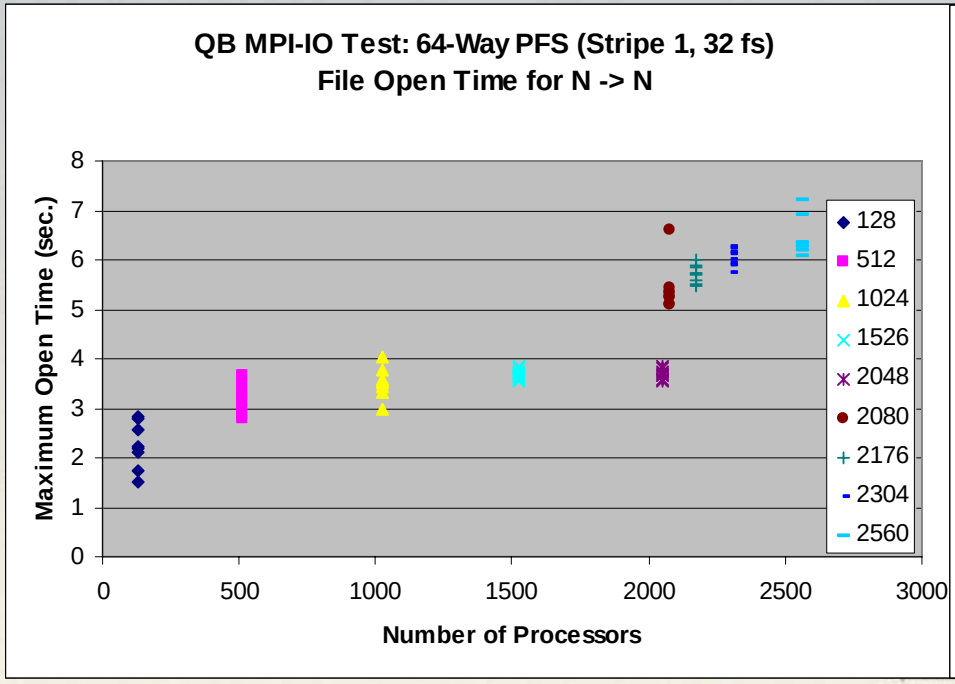
Q I/O - Parallel N to N Write

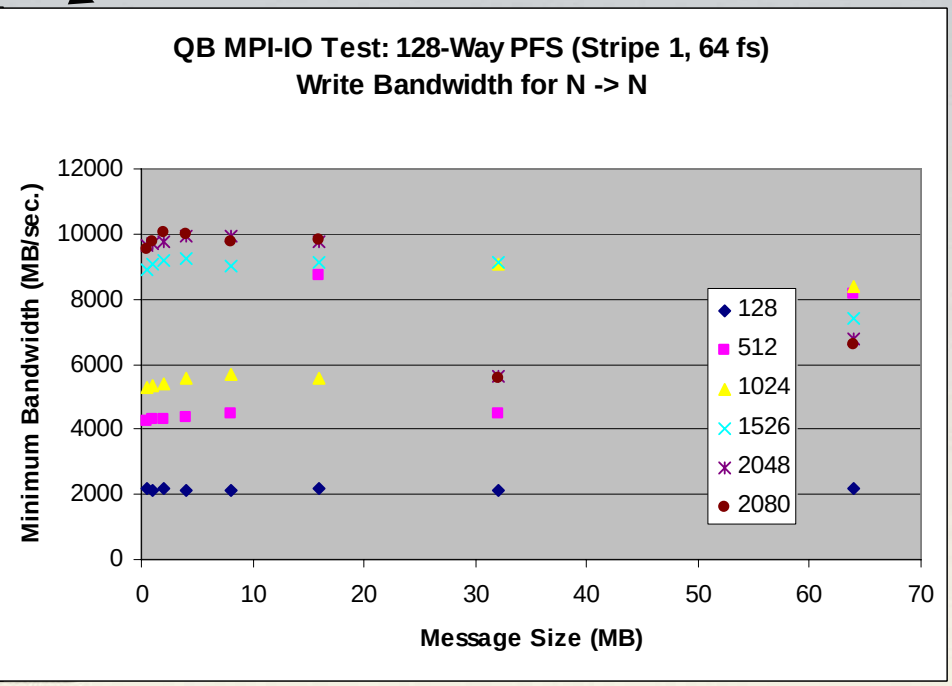
Tested on single segment of Q is 10 Tflops

Parallel metadata ops (great)

10 GB/s is screaming and with only 5 TF problem (2048 CPUs)

Double the ASCI goal of 1 GB/s per Tflop is unprecedented, especially on write with wide variety of block sizes



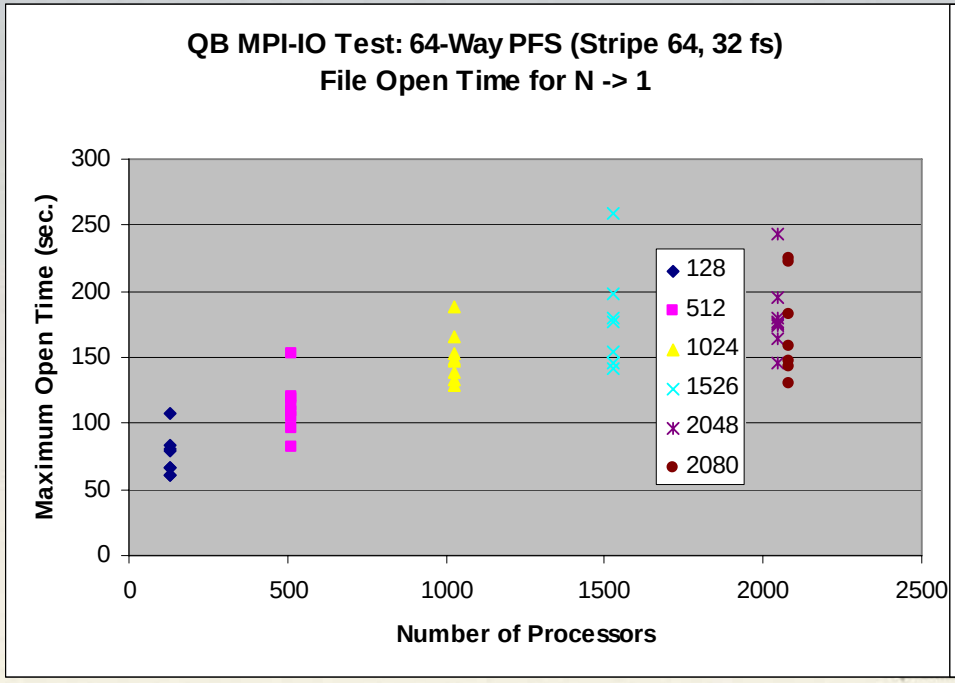


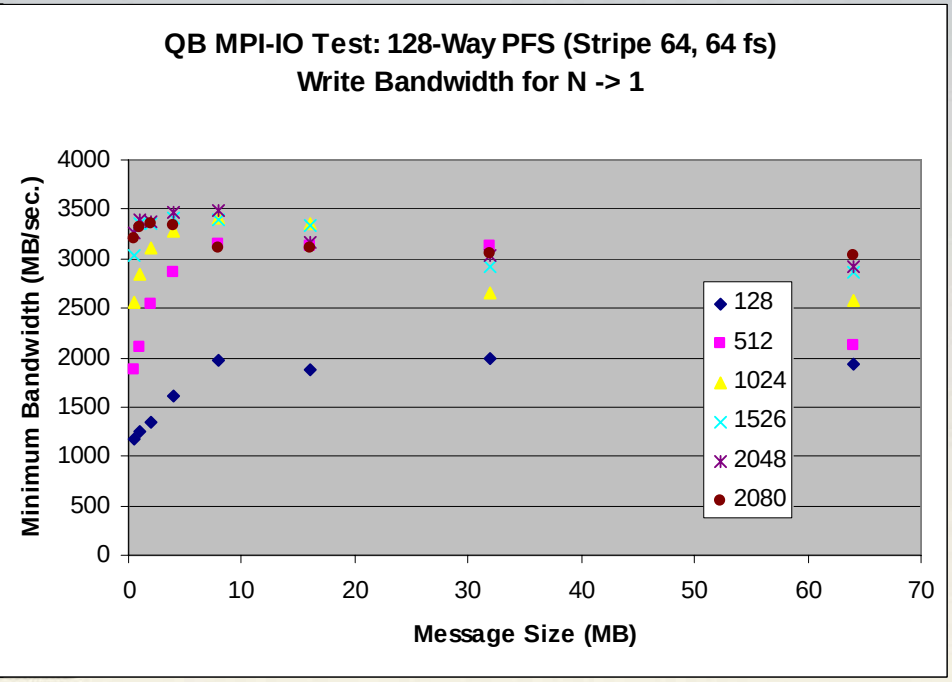
Q I/O - Parallel N to 1 Write

Tested on single segment of Q is 10 Tflops

Some overhead issues with N to 1 – working on metadata amortization to address

Almost at ASCI goal of 1 GB/s per Tflop - 3.5 GB/s on 5 Tflops (2048 procs) with wide variety of block sizes is world class and will get better soon

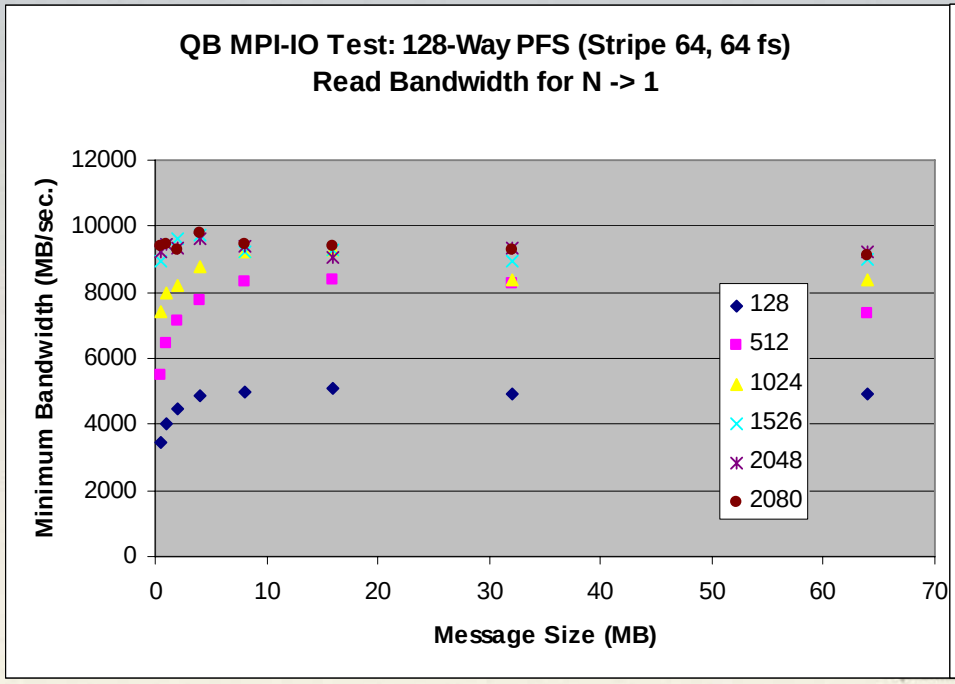


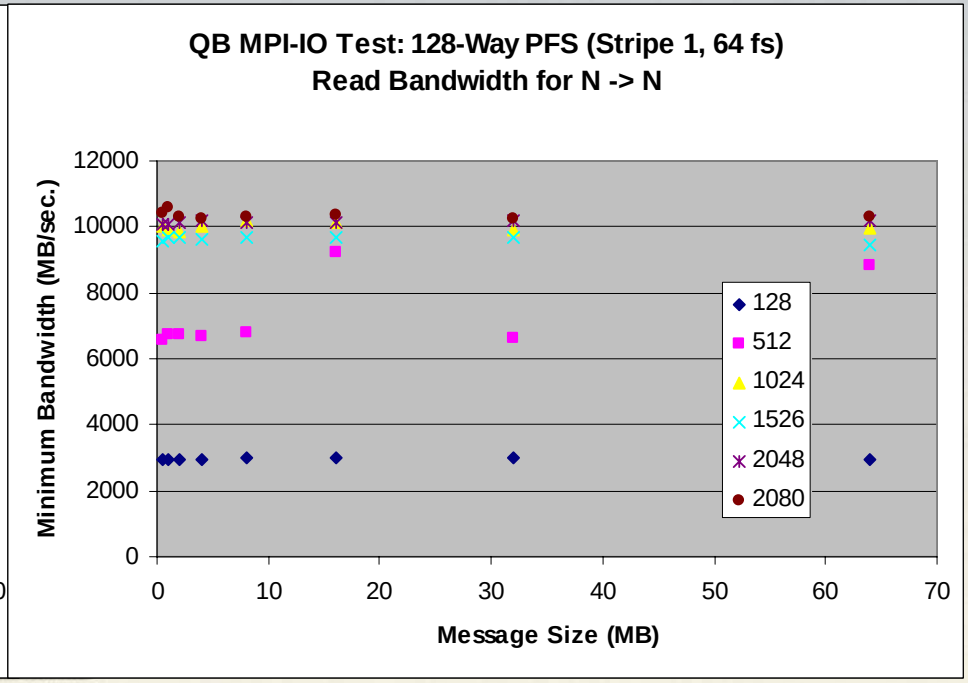


Parallel Reading

Twice the ASCI goal of 1 GB/s per Tflop – 10 GB/s per 5 Tflops (2048 procs) on reading N to N and N to 1 with wide variety of block sizes

Again, unprecedented, truly screaming Parallel I/O, thanks to lots of work, dedicated I/O testbed for Q, and good support





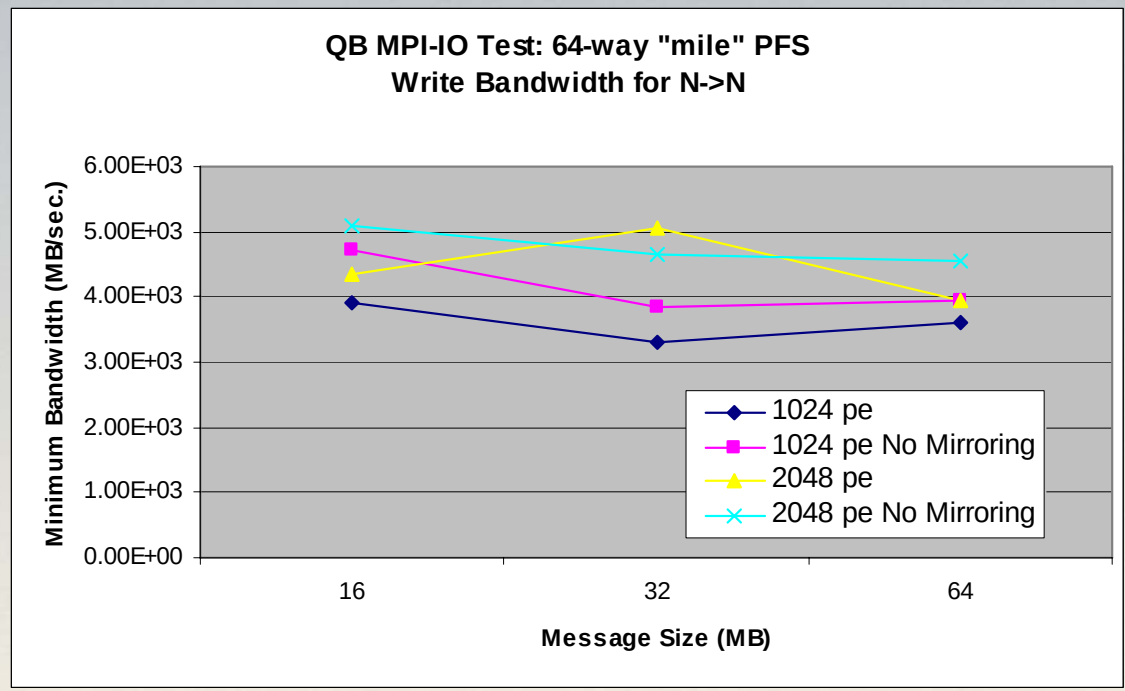
What Are Real Applications Seeing on Parallel Workloads on Q?

Multiple narrower file systems were created for high availability specifically for the latest application milestone

Application uses N to N parallel I/O model

Users getting > 2 GB/s typically with milestone application

I/O including defensive I/O was small percentage of run time



Q I/O Summary

Scalable I/O and File Systems

- Why? - Without scalable I/O, we don't do simulation
- How? - Tri-lab, Leverage, Partnerships, Leverage, Leverage

Q

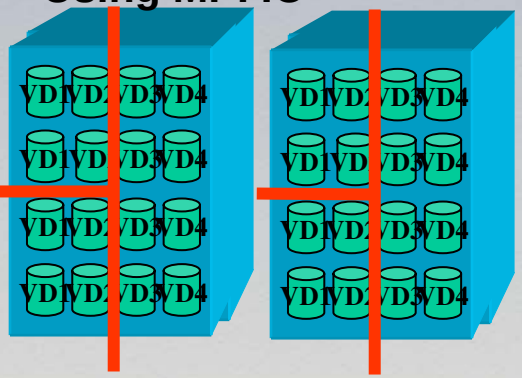
- Parallel performance at or near world-class
- Remaining issues with availability, small file, single process serial file, and NFS client performance
- All issues being worked actively with reasonable near-term goals

5. Archival Storage Performance

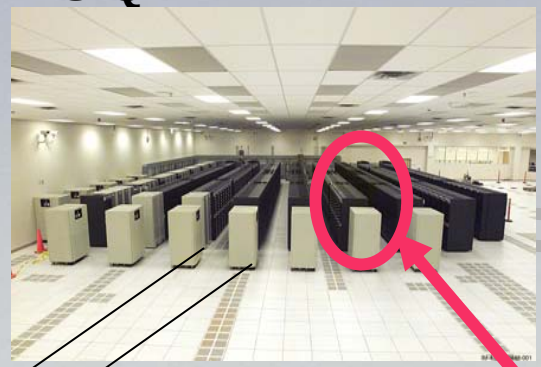
Contributed by Ray Miller

Q Simulation Environment

300 MB/s for 16 stripe
500-900 MB/s for 64 stripe
3000 MB/s for 128 stripe
Using MPI IO

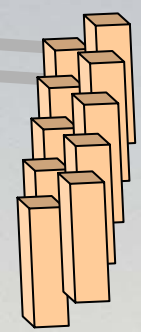


FS-QA



Distance

Rendering platform
network connected to
visualization platform

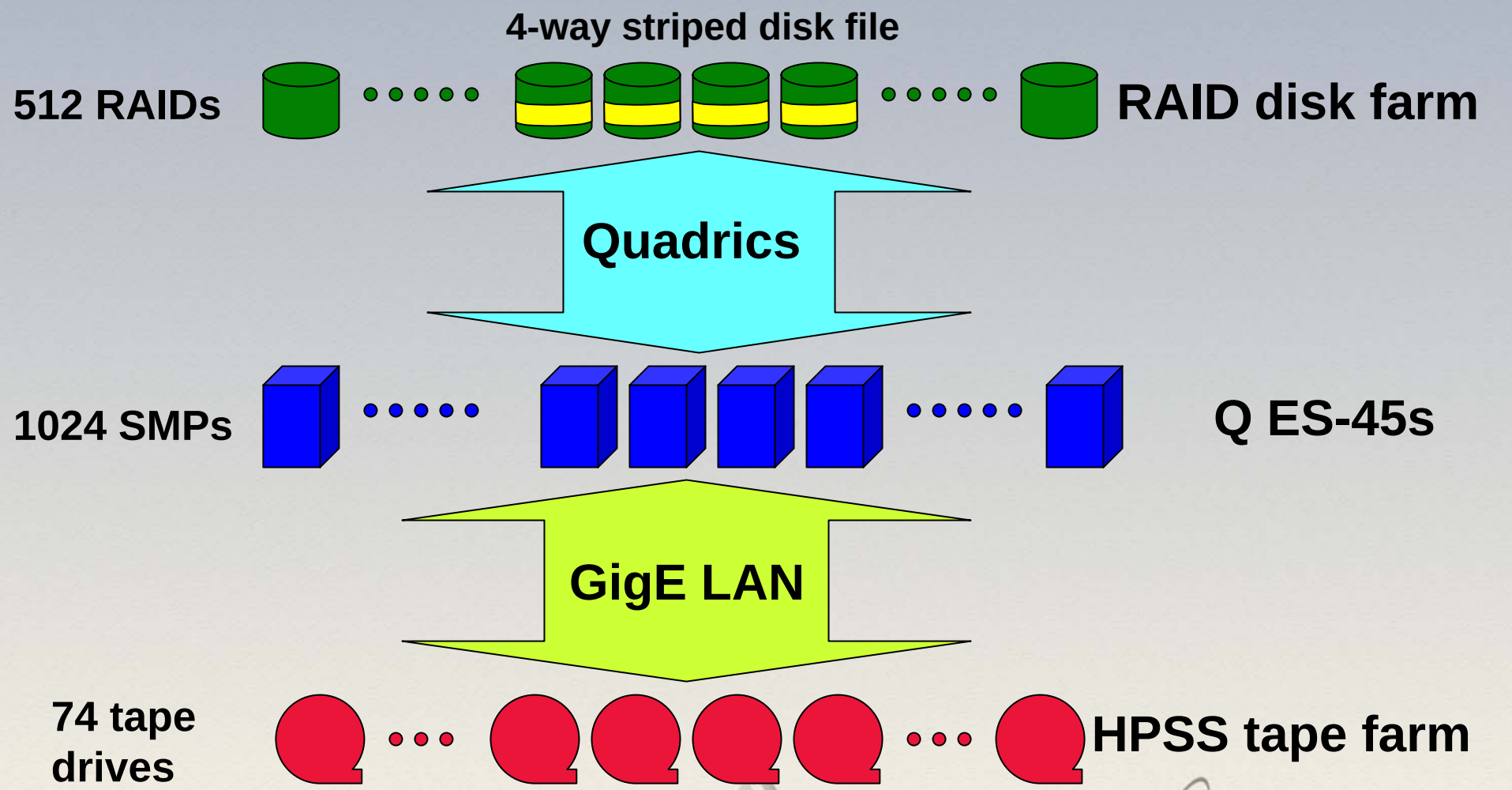


HPSS
High Performance Storage System

4-way stripe
(new 9940B tapes)
450 Gbyte file
Stored at 166 MB/s
Spring 2003 16-way stripe
~600 MB/s

Visualization nodes

Archival Data Flow To/From Q Segment



Bottlenecks

Description	Limitation
Single RAID	40-65 MB/s
Single node	125 MB/s
Number of GigE nodes	High demand for scarce resource
Node level 12-port card	4 GB/s (500 MB/s)
GigE router	50% of Link Speed (2 GB/s)
Single tape	30-70 MB/s
> 4-way tape	Reliability

Balancing the Data Path

Description	Action Taken	Resulting Limits
Single RAID	Default 4-way stripe	160-240 MB/s 256-way stripe
Single node	Multinode PSI Mover software	Not a bottleneck
Number of GigE nodes	GigE adapter on each node	Not a bottleneck
Node level 12-port card	Interleaved node connections	Up to 16 Gb/s (2 GB/s peak)
GigE router	Direct connect Q switch & HPSS switch	100% of link speed, or 8 Gb/s (1 GB/s)
Single tape	Stripe across 4 tapes	120-280 MB/s
> 4-way tape	Implement 16-way mirroring	480-1120 MB/s

LANL Developed *Multinode* PSI Mover

- PSI archives a single file or all files in a directory
- PSI archives to disk, a single tape, or 4-way tape
- PSI uses nodes given to maximize bandwidth
- For example, a write of a striped disk file of >1 GB in size would use 4 Q nodes, if 4 were available, to write to a 4-way 9940B tape file on HPSS and additional nodes, if available, to archive small files to disk, and medium files to single tape in parallel; all with a single invocation of PSI.

Archival Storage Aggregate Performance

Limitations:

- Tape drives
  - 36-9840 @ 13 MB/s → 468 MB/s
  - 20-9940B @ 45 MB/s→ 900 MB/s
  - *16-9940B @ 45 MB/s → 720 MB/s (June 03)*
- Disk drives → 50 MB/s
- Aggregate disk + tape total → 2.138 GB/s

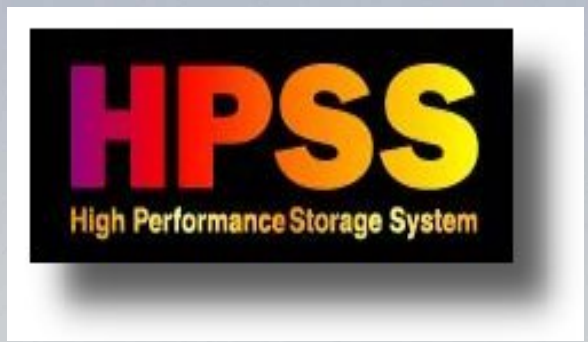
Conclusions:

- *Probably will need more direct connections from Q to HPSS once 16-way tape archive is implemented (June '03)*

Actual Q to HPSS Transfers

- **Safety code transfer 1**
  - 450 GB size file, 6 disk components, 4-way tape
  - Transfer rate: 166 MB/s
- **Safety code transfer 2**
  - 66 GB size file, 6 disk components, 4-way tape
  - Transfer rate: 268 MB/s (2.2x compression & 96% of max tape rate)
- **CCN-7 test transfer**
  - 3 32-GB-size files, 4 disk components/file, 8 Q nodes,
  - 3 4-way tapes (12 tapes)
  - Transfer rate: 478 MB/s (1.33x compression)

Archival Storage Summary



LANL has:

- **Increased from 10 MB/s to 30 MB/s native**
- **Implemented 3 X ~150MB/s streams (4-way stripe)**
- **Direct connected Q & HPSS to double throughput**
- **Increased from 20 GB/tape to 200 GB/tape native, 300 GB w/ compression**
- **Provided multi-node HPSS PSI mover with the capability to not only maximize archival bandwidth but also to greatly simplify the user's archival process.**

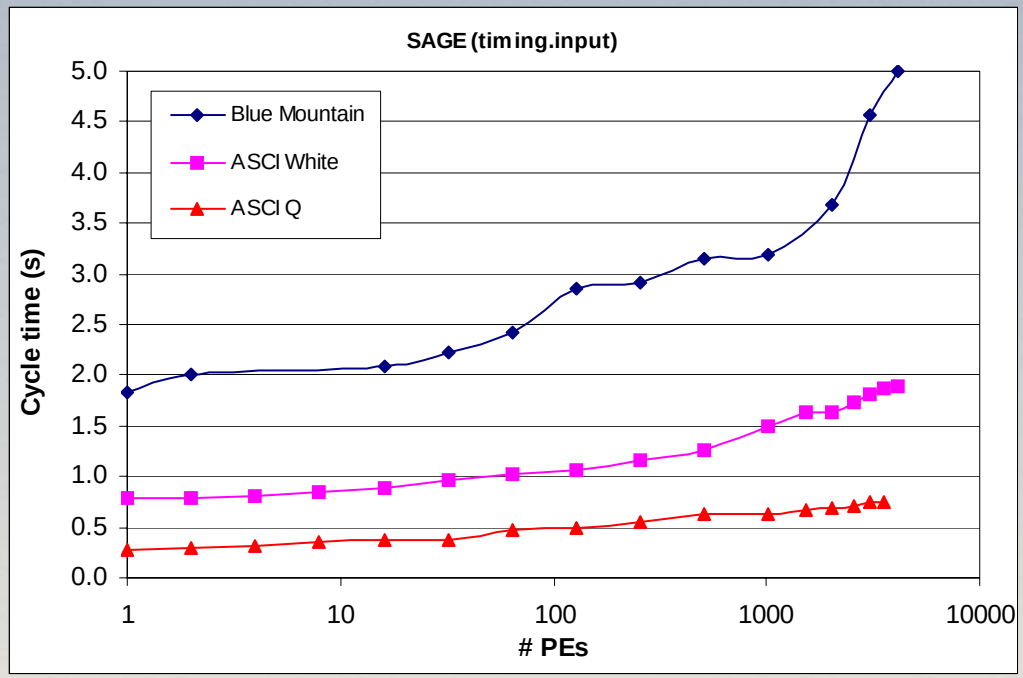
LANL will:

- **Make available 1.5 PB of high speed, high volume, 9940B tapes by FY03, Q3.**
- **Archive a 400 GB file @ at least 600 MB/s by FY03 Q2 using mirroring. Can implement 2 of these 16-way mirrored systems if needed.**
- ***This is a complex problem, but we have done the analysis. We will meet the S&CS Requirements***

6. Application Performance

Contributed by Petrini, Kerbyson, Hoisie.

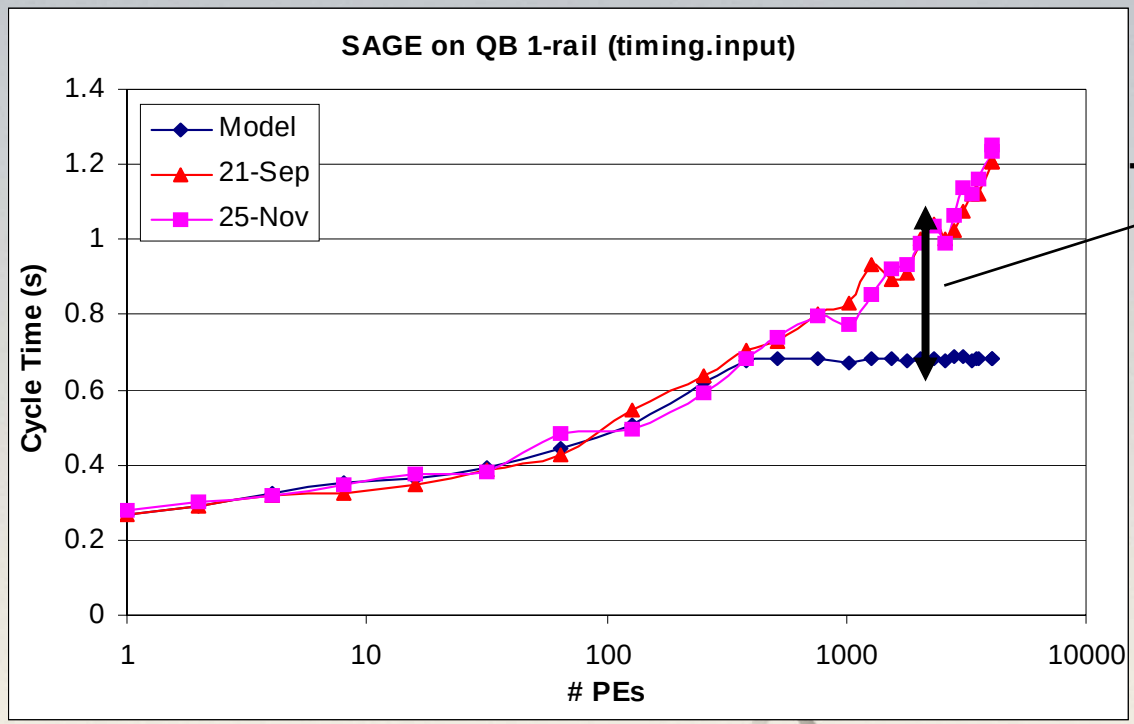
Performance Comparison
Q vs White vs Blue Mountain



Cycle-time : lower is better
Weak-scaling of SAGE (problem per processor is constant)
-> ideal cycle-time is a constant for all PEs (but have parallel overheads)

Modeled and Measured Performance

- **Unique capability for performance prediction developed in the Performance and Architecture Lab (PAL) at Los Alamos**
- **Latest two sets of measurements are consistent (~70% longer than model)**



There is a difference: <u>why</u> ?

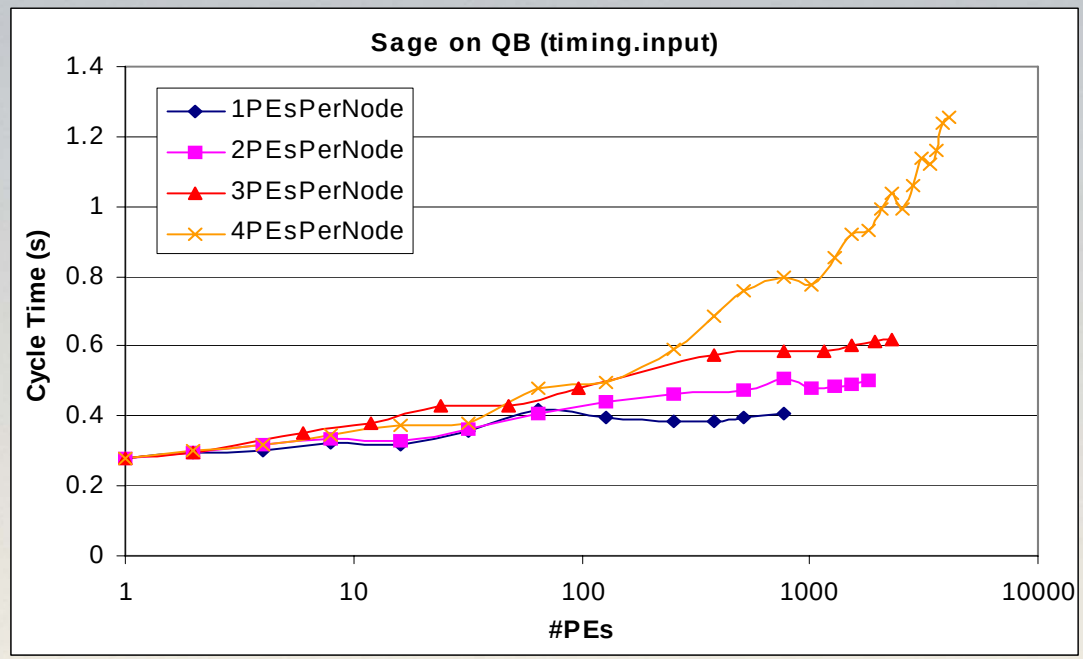
Lower is better!

Using Fewer PEs per Node

Test performance using 1, 2, 3 and 4 PEs per node

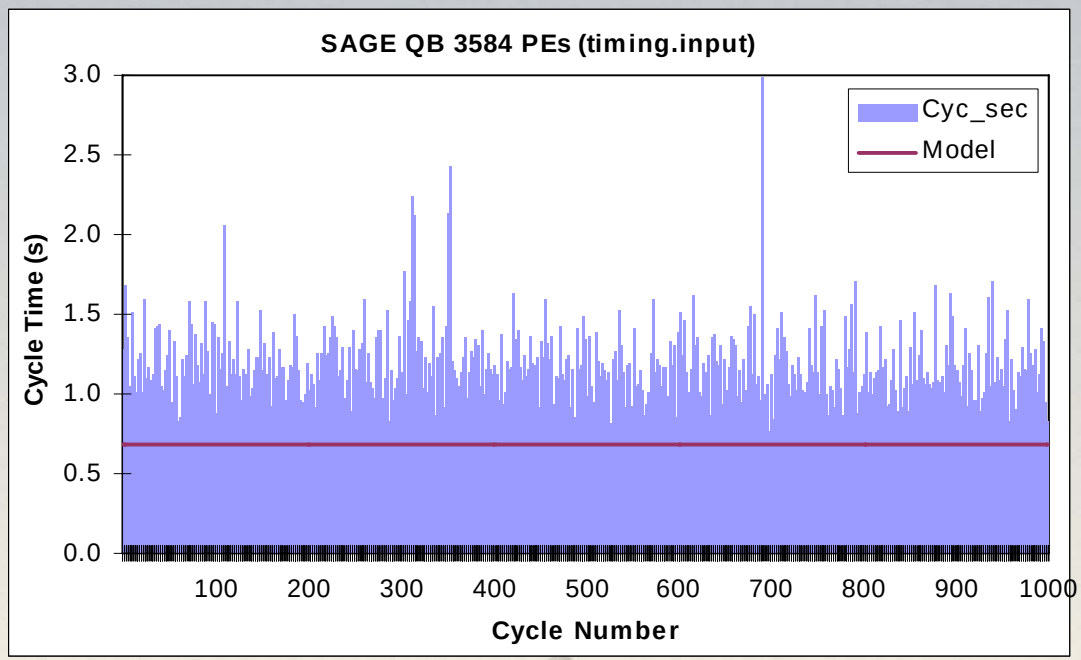
Reduces the number of compute processors available

Performance degradation appears when using all 4 procs in a node!



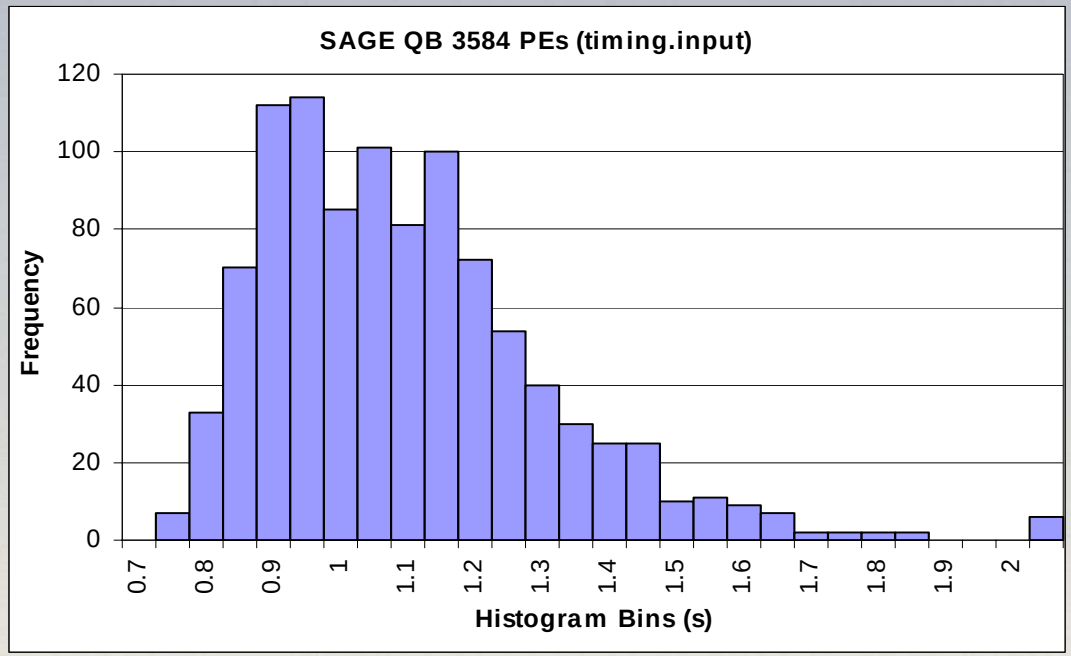
Performance Variability

- Lots of noise on the nodes: daemons and kernel activity
- This noise was analyzed, quantified, modeled, and included back in the application model
- This system activity has structure: it was identified and modeled
- Cycle-time varies from cycle to cycle



Performance Variability (2)

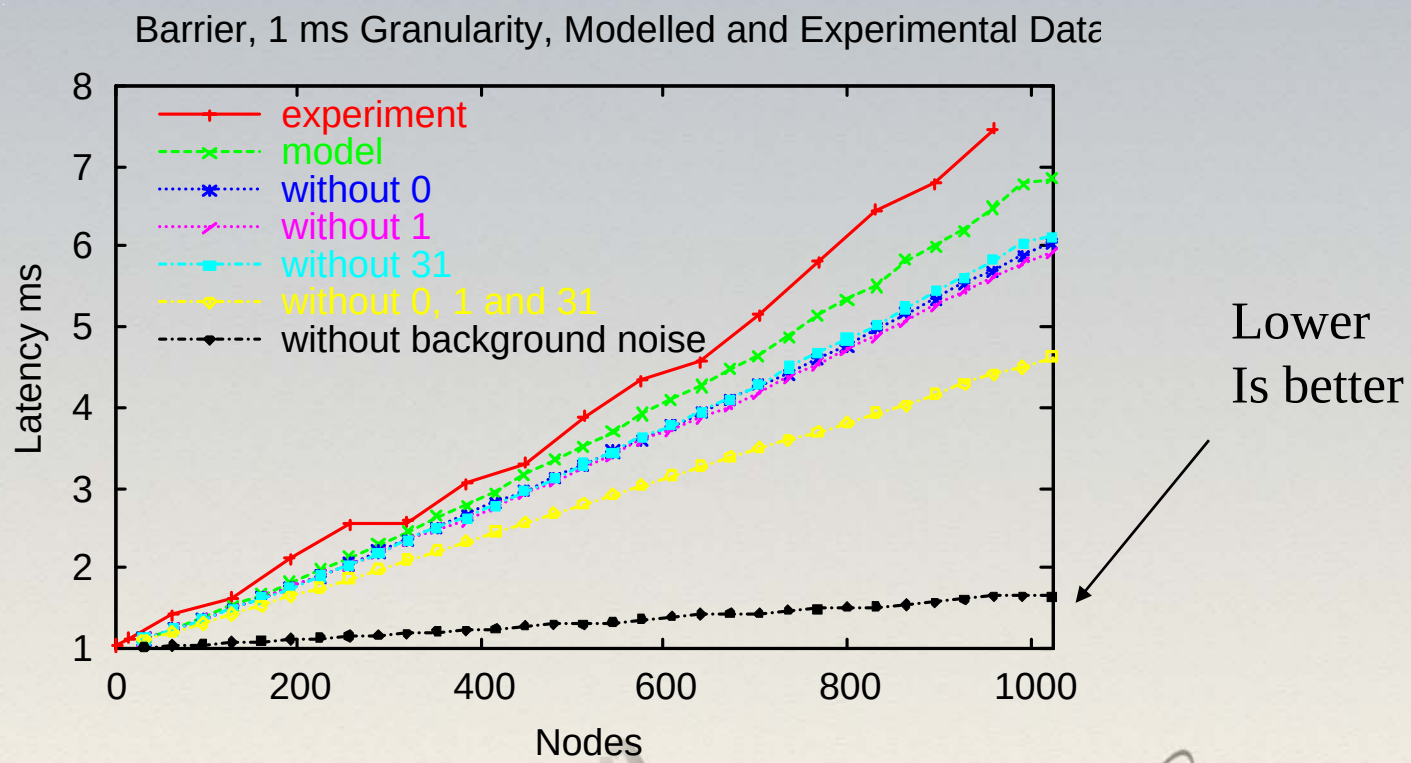
- Histogram of cycle-time over 1000 cycles
- Minimum cycle-time is very close to model! (0.75 vs 0.70)



Performance is variable (some cycles are not affected!)

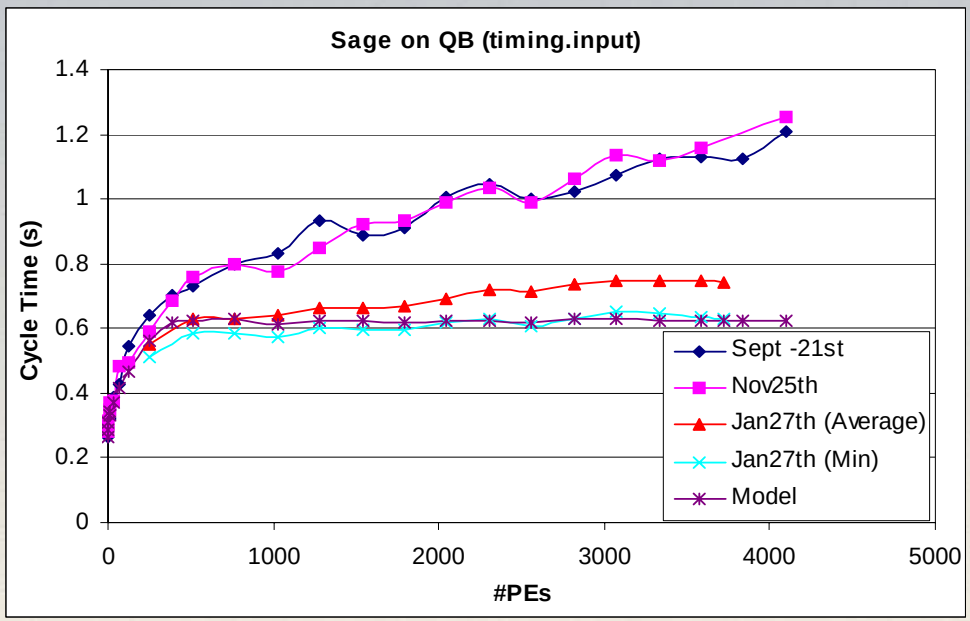
Modeled and Experimental Data

- **The model is a close approximation of the experimental data**
- **The primary bottleneck is the noise generated by the compute nodes (Tru64)**



Performance After System Optimization

After system mods (both kernel and daemons and Quadrics RMS: right on target! After these optimizations, Q will deliver the performance that it's supposed to. Modeling works!



Summary on Performance

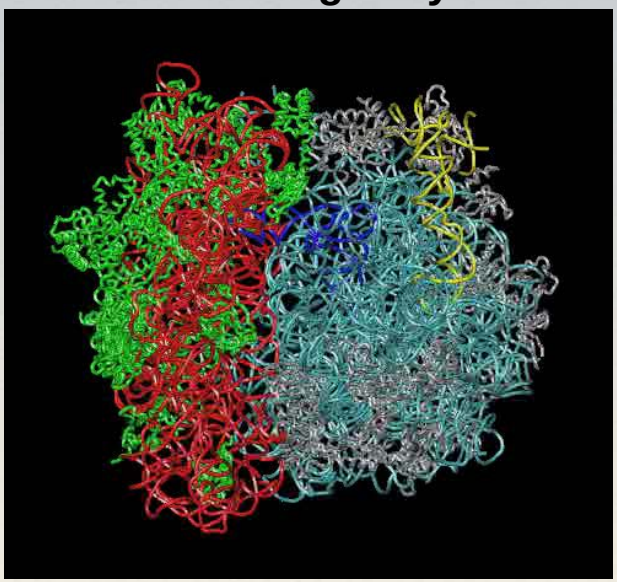
- Performance of Q machine is meeting and exceeding performance expectations
- Performance modeling integral part of Q machine system deployment
- Performance testing done at each major contractual milestone
- FS-QB used in the unclassified environment for performance variability testing.
- Approach is to systematically evaluate and implement recommendations of performance variability testing

7. Science Runs on Q

Science Runs on Q

1. The first million-atom simulation in biology: molecular mechanism of the genetic code. (*Kevin Sanbonmatsu*)

   This work will define a new state-of-the art in biomolecular simulation, paving the way for other researchers in the community to study larger, biologically relevant modules.



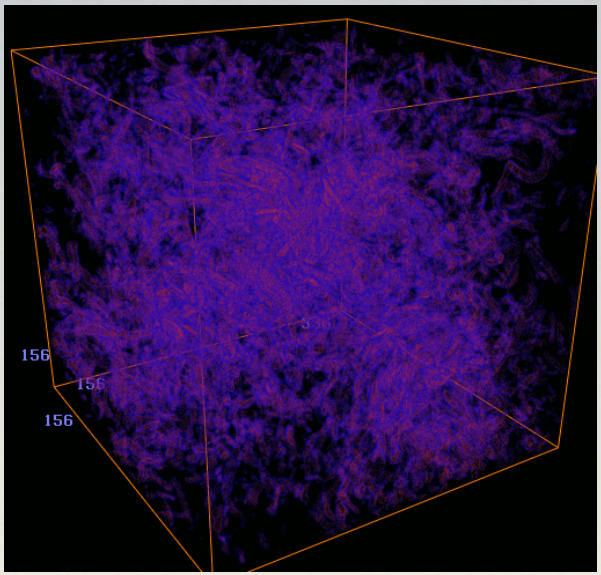
Conclusions

- Simulation > 5 times larger than largest to date.
- Core of the ribosome is more stable than outer regions.
- Identified possible pivot point for ratcheting motion during translocation.

Science Runs on Q

2. Direct Numerical Simulation of a recent decaying Navier-Stokes turbulence experiment. (*Darryl Holm, Mark Taylor)*

This work will study the fundamental properties of turbulence by performing simulations of the incompressible Navier-Stokes equations at record-setting resolution



Turbulence Simulation

ASCI Q can easily run 2048^3, representing a 256- fold increase in computing power for this application

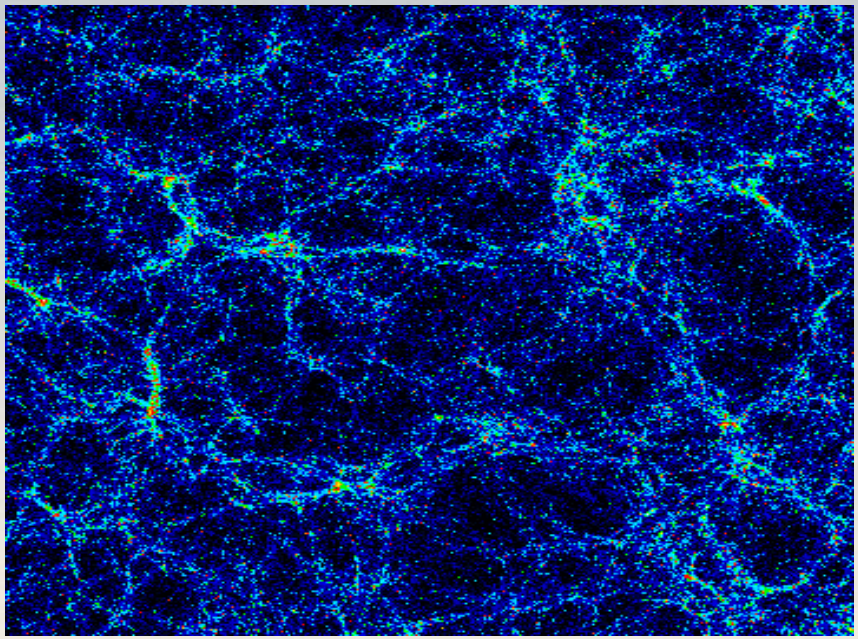
2048^3 crosses an important threshold in resolution, providing for the first time invaluable flow information at high Reynolds number for fully developed turbulence

The data from this 2048^3 simulation of the recent wind tunnel experiment, Kang, Chester and Meneveau (2002), is expected to generate tremendous interest and excitement in the turbulence community.

Science Runs on Q

3. Computational astrophysics. *(Mike Warren)*

This work will model the formation of large-scale extragalactic structures in unprecedented detail and compare with observations such as the Sloan Digital Sky Survey.

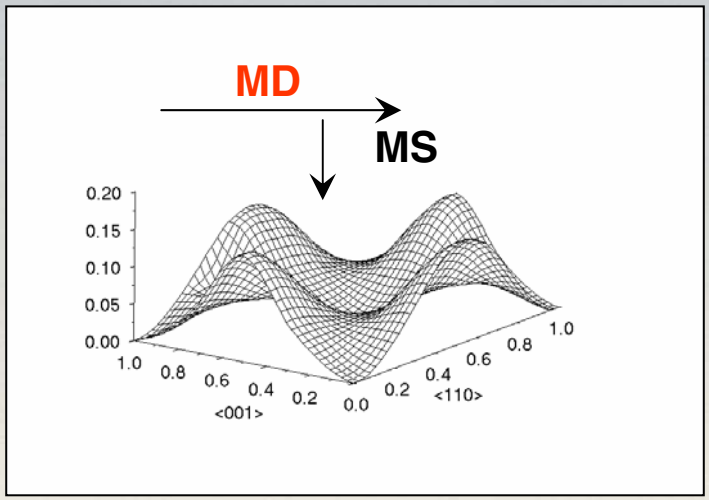


- **1.4 billion particles in gravitational simulation of galaxy formation.**
- **Spatial resolution is 10 times better than previously published simulation.**
- **The image represents the density of dark matter in the universe. Our galaxy lives in a clump of dark matter similar to the size of the medium-sized objects in the picture.**
- **Simulated over 100 different cosmological models during their runs on QB.**

Science Runs on Q

4. Quantum-Based Molecular Dynamics Simulations of High-Z Metals. ***(Randy Hood, Lin Yang, LLNL)***

This work involves quantum-based simulations of the structural and thermodynamic properties of high-Z transition and actinide metals at high temperatures and pressures.



PPMD simulation is an application with compute/communication ratio of 60/40

- 16 electrons/node, 1024 CPUs

CPU time (seconds) per time step

QB	2089.4
MCR	4190.1
ASCI White	6865.7

Science Runs on Q

5. Simulation of the K-T impact event at Chicxulub (the "Dinosaur Killer").
   (Galen Gisler)

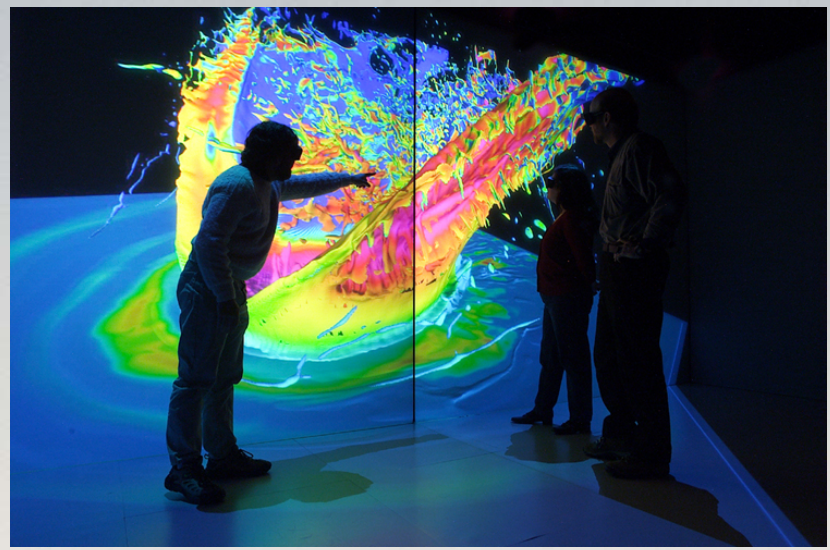
This work will study in detail the impact with the earth which is widely accepted to have initiated the sequence of mass extinctions at the Cretaceous-Tertiary (K-T) boundary 65 million years ago.

In two months of running on ASCI machine QB (ending 1/21/2003), we have generated output and restart dumps totaling:

60˚ runs	1.59 Tbytes
45˚ runs	9.69 Tbytes
30˚ runs	1.74 Tbytes
90˚ runs	0.10 Tbytes
Total CPU hours	0.9 million

Science Runs on Q

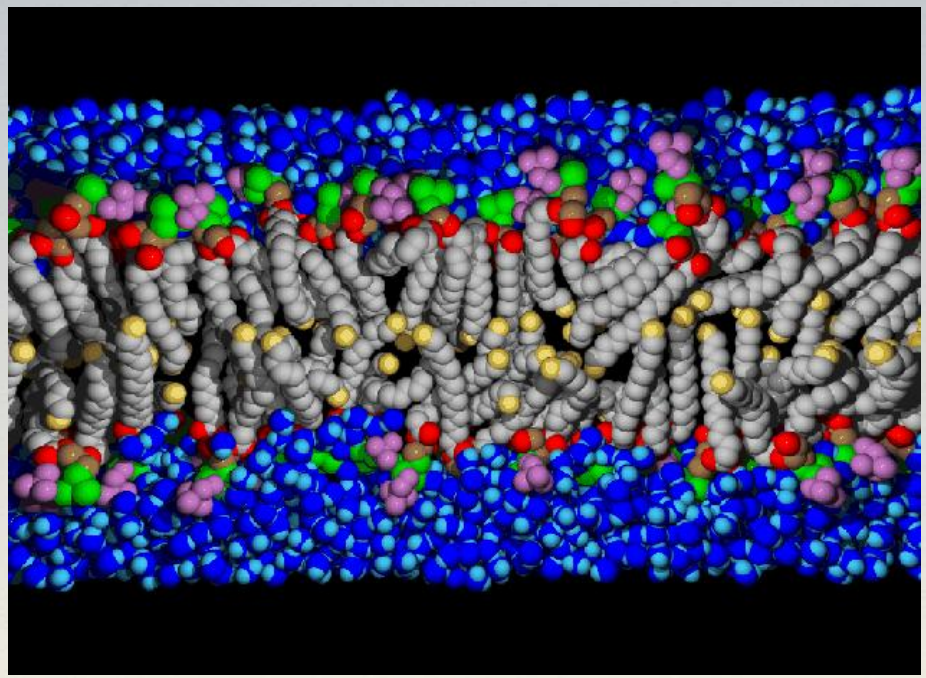
5. Simulation of the K-T impact event at Chicxulub (the "Dinosaur Killer").
(Galen Gisler)



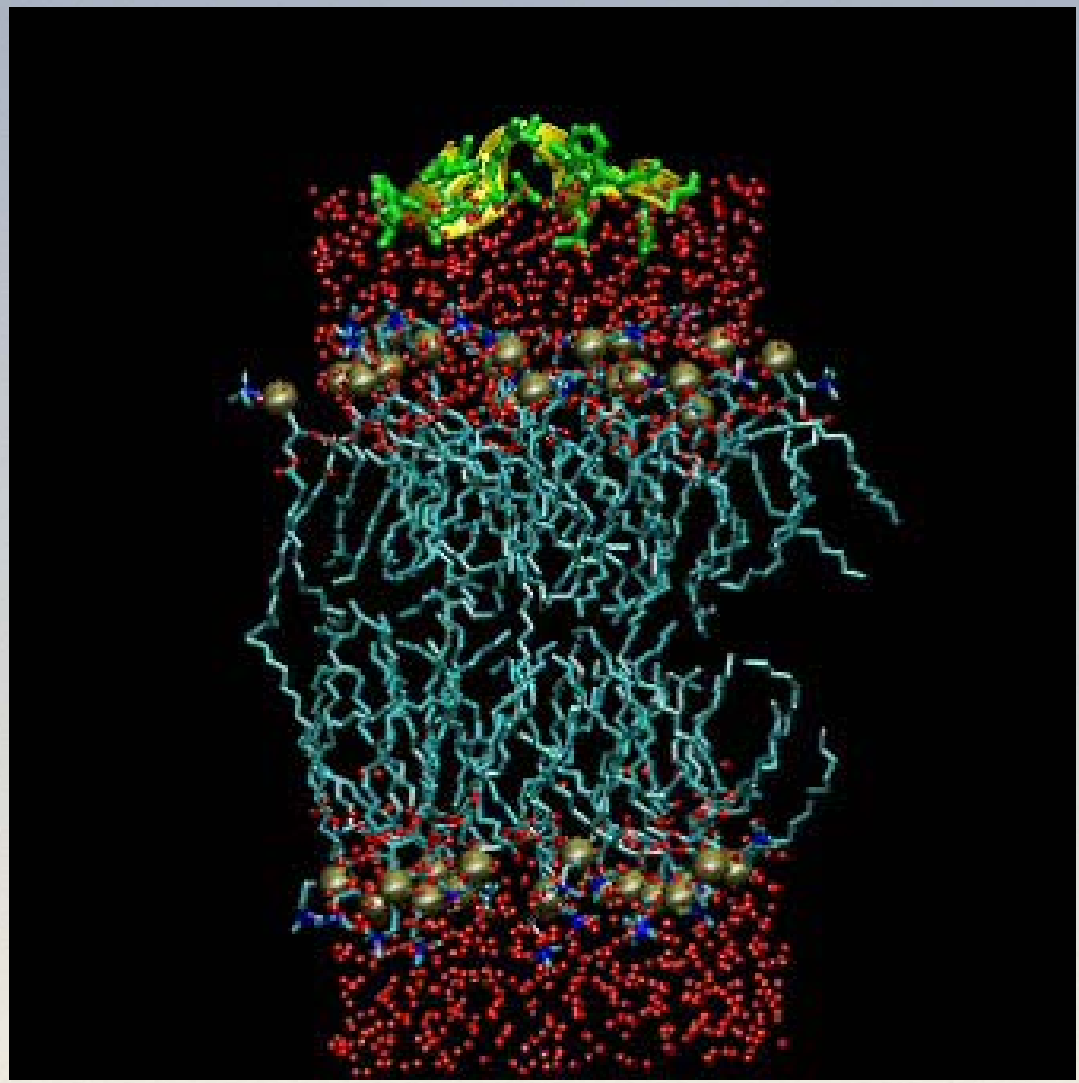
Science Runs on Q

6. Atomic simulations of the protein folding and insertion into a lipid membrane *(Angel Garcia)*

Lipids – protein interactions (here DPPC)



- **New sampling methods allow for the simulation of protein folding from physical principles**
- **–Highly parallel**
- **–We have performed the first atomic simulations, with explicit treatment of solvent, of the thermodynamics of:**
  - **Helix-coil transition (Fs peptide, A21 peptide)**
    - **–Garcia and Sanbonmatsu, PNAS (2002)**
  - **Protein folding (Protein A)**
    - **–Garcia and Onuchic, submitted to Science**
  - **Peptide insertion in a membrane**
    - **–Nymeyer, Woolf, and Garcia, In preparation**



Science Runs Summary

- Scientists still analyzing data, papers being written
- Some results already being presented at conferences and have been submitted for publication.
- Running on Q accelerated each of the science runs significantly
- Data generated will be made available to scientists around the world
- Quote from scientist, “These runs on the Q machine have changed my career.”
- In conjunction with the Laboratory’s 60th Anniversary we will be having scientists present the results of their work on the Q machine.
- “Rewarding” for CCN staff in providing institutional computing

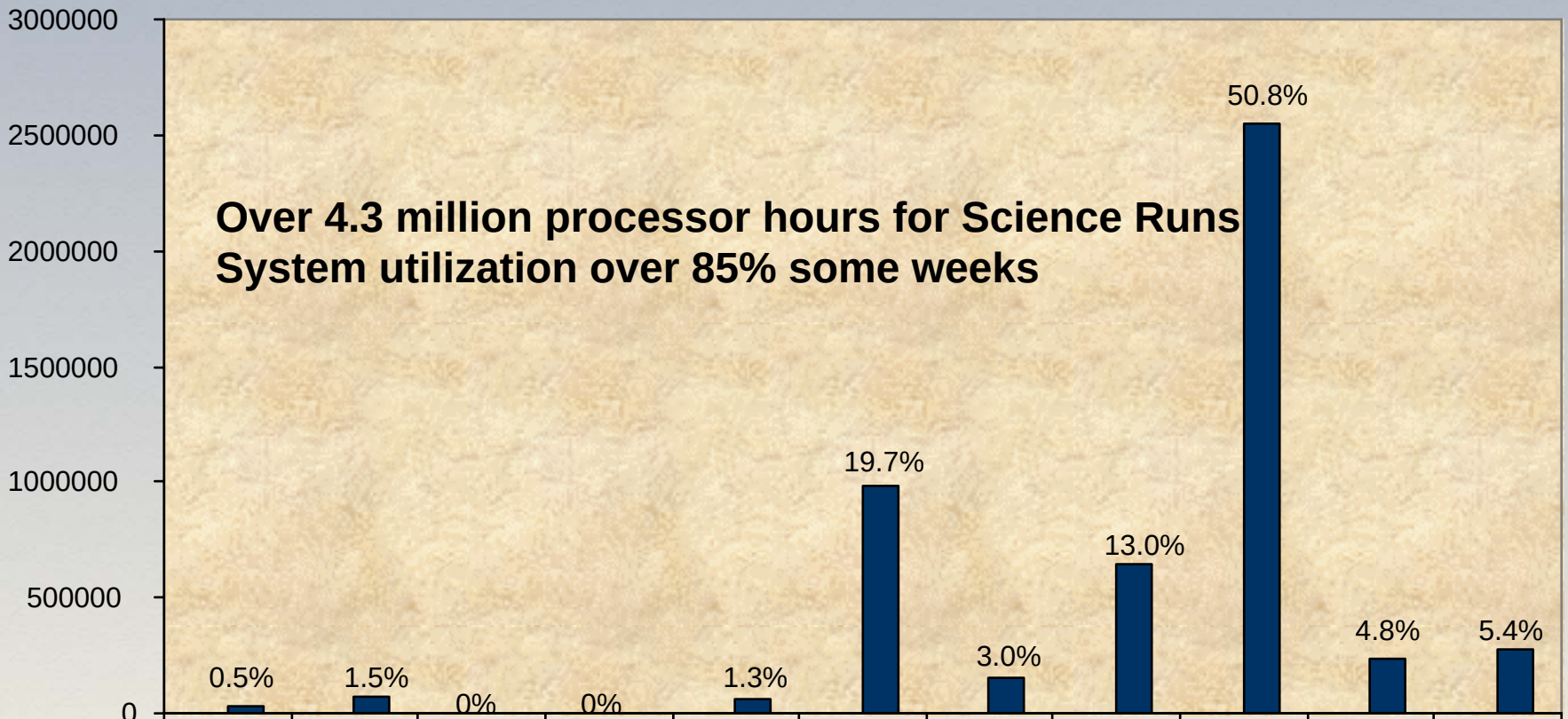
8. ASCI Q Summary

Summary

- Significant progress has been made on the deployment of the ASCI Q machine at Los Alamos.
  - Ahead of schedule for milestone work
    - Deployed two 10T systems in less than 6 months
  - Have overcome significant obstacles and challenges
  - CCN staff have been involved in many of the technical solutions
- System stabilization, performance improvements, and system connectivity work remains.
- The ASCI Q machine will provide significant capability and capacity computing to the program over the next several years.

Backup Slides

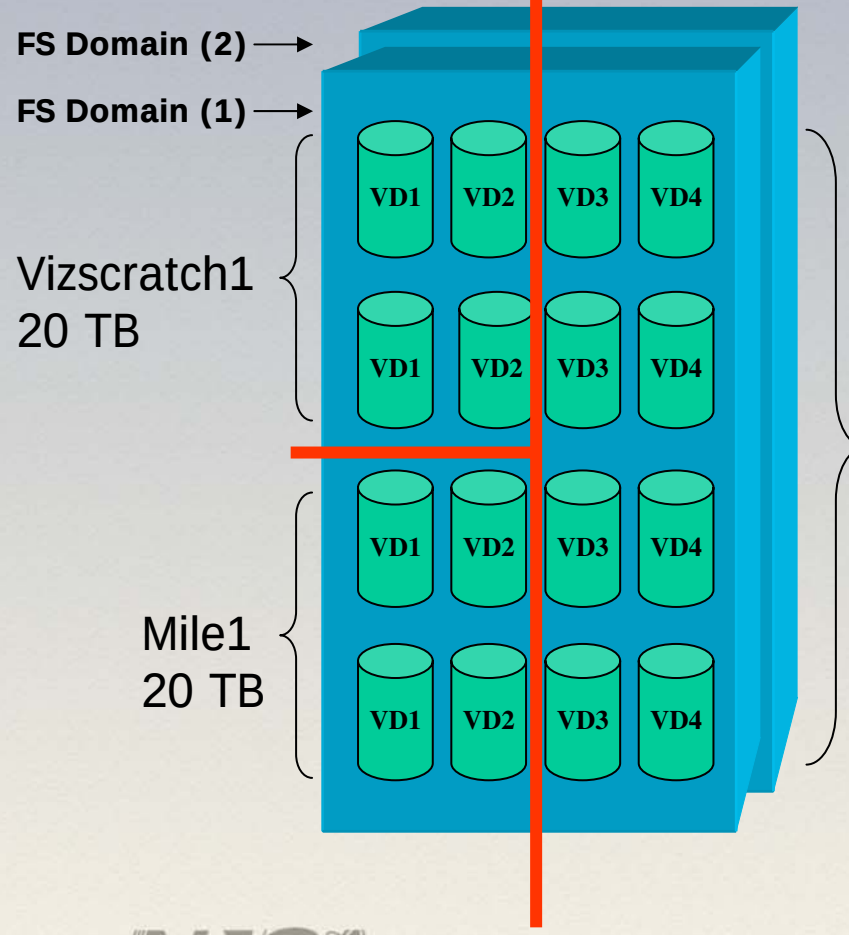
QB Final
11/1/02 - 1/22/03



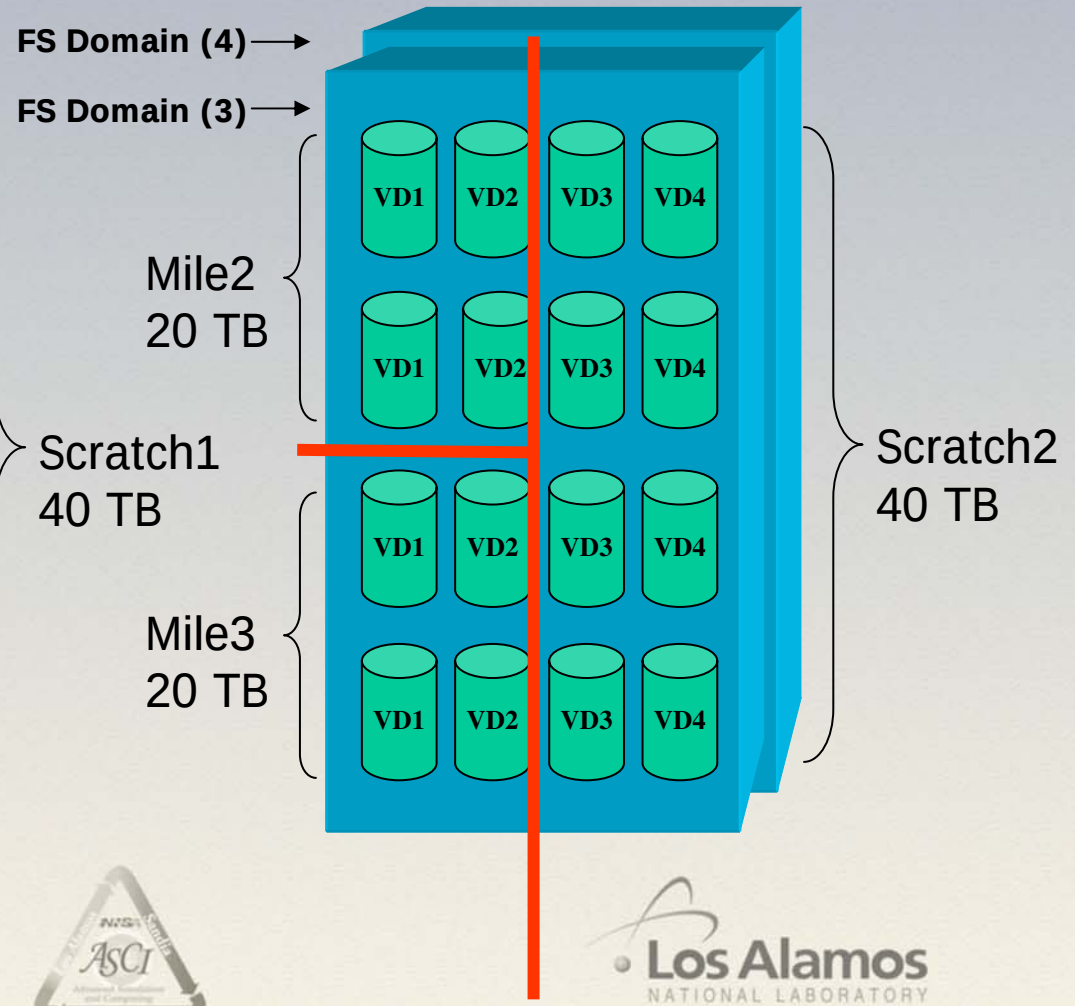
	4	8	16	32	64	128	256	512	1024	2048	4096
# jobs	4694	4382	142	82	549	925	254	1062	1031	277	559
Proc Hrs	26052	75050	1266	684	63379	987403	150619	649458	3E+06	239710	272862

PFS File Systems for QA QB QC

Current on QA



Needs to be reconfigured on QA



Q was deployed in stages

Two 10 TeraOPS segments

- FS-QA
- FS-QB

All identical configurations

Each 1/2 of total system

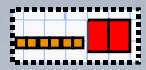
- Operate independently
  - High Performance file I/O
- Or as a single machine
  - With high-performance file serving within segment
  - 6th level Quadrics Network connecting segments (25%)

Plus 256 node ES45 system in the unclassified partition (QSC)

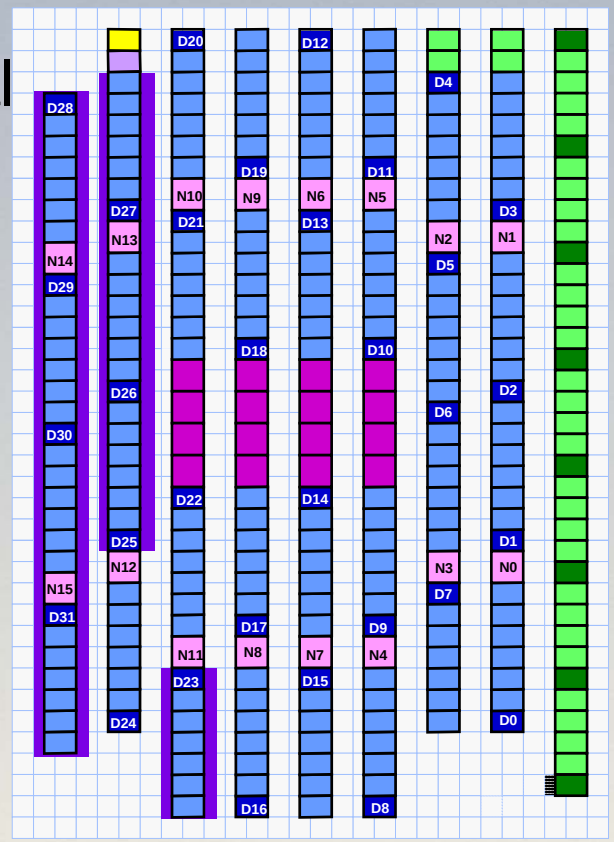
Performance Variability

- System performance being checked against a prediction model using applications typical of ASCI workload
- Adolfy Hoisie's team in CCS and system personnel in CCN
- System performance variability issues
  - Checking all system components
  - Communication bottlenecks, computational bottlenecks
  - Computational noise, in the form of background system tasks, is creating most of the performance variability issues
- Testing last weekend with HP by removing some daemons, etc. confirmed these results.
- Will integrate these findings into operational system

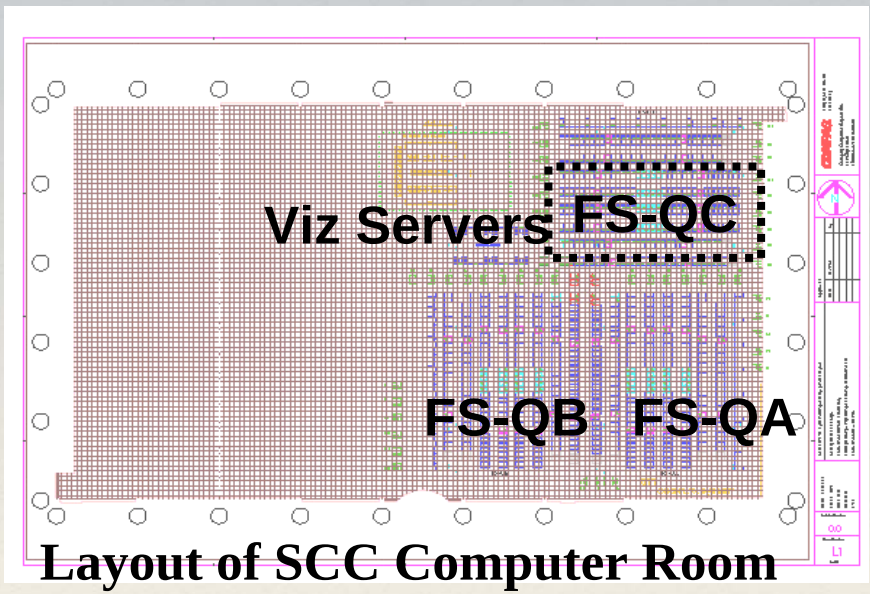
FS-QA, 10 TF, 11 TB memory, 221 TB disk



FS-QB - Identical System

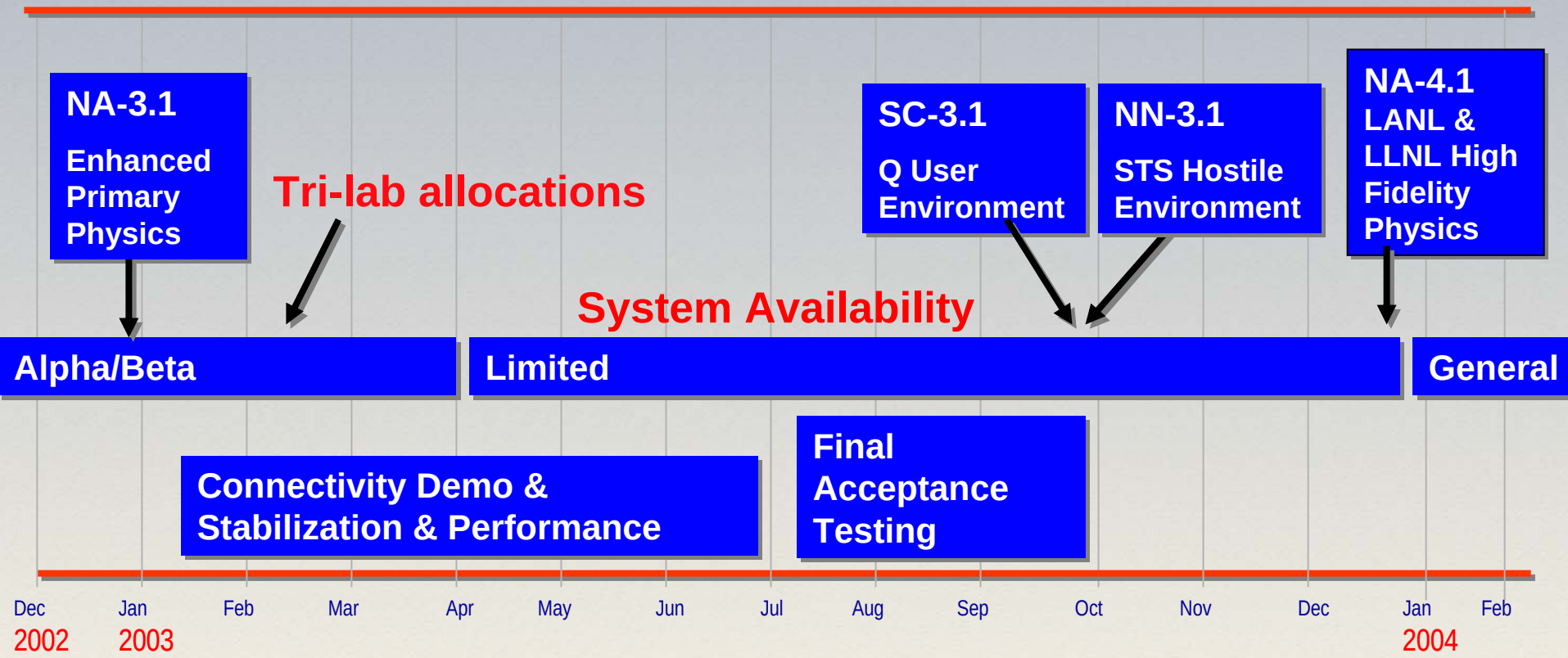


FS-QA FLOOR LAYOUT

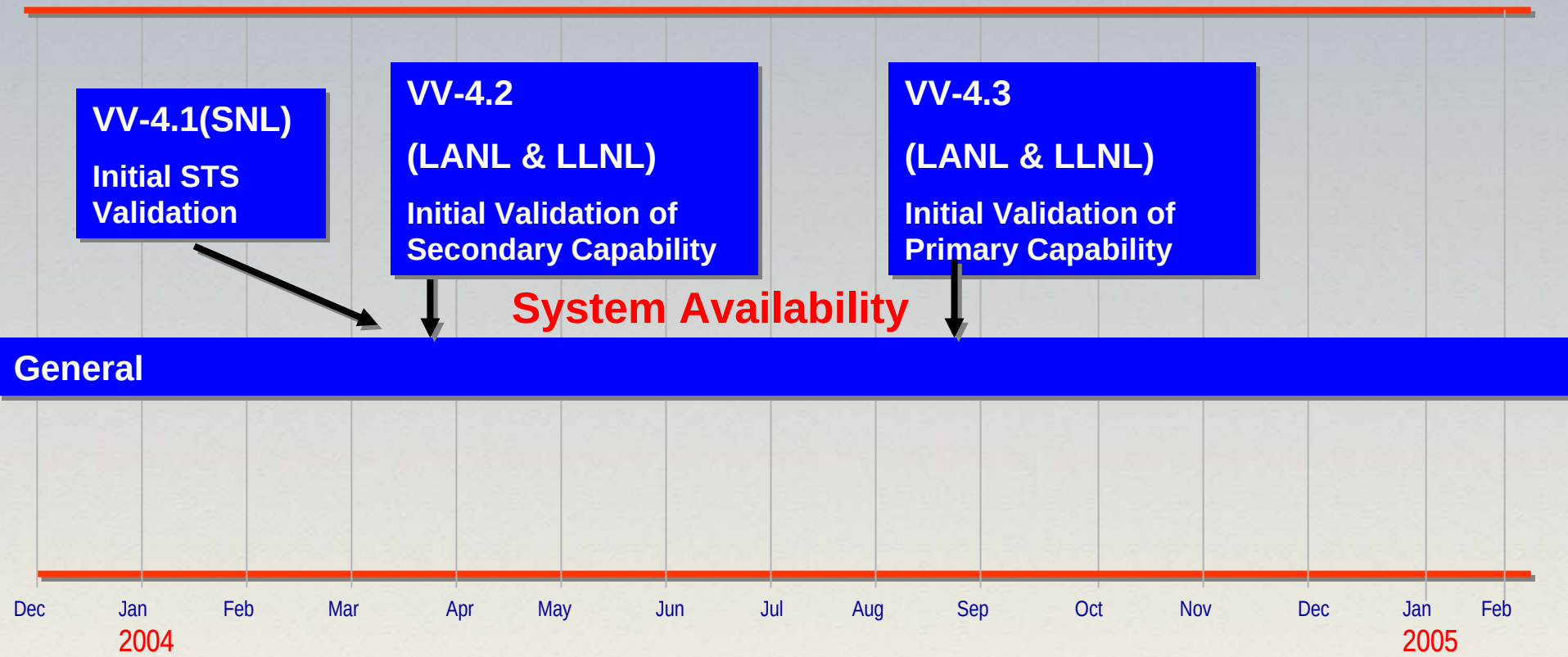


Layout of SCC Computer Room

Q **Commissioning** and Utilization Plans



Q **Commissioning** and Utilization Plans



Schedule Impacts

*** Will work with Tri-Lab community to minimize impact**