LA-UR-04-7798

Approved for public release; distribution is unlimited.

Title: ASC Platforms at Los Alamos

Author(s): Steven R. Shaw, CCN-7

Submitted to: Supercomputing 2004, November 8 - 12, 2004
Pittsburgh, PA.

Los Alamos
NATIONAL LABORATORY

Abstract

This talk describes the history, current state, and future plans for ASC computational and data storage service at Los Alamos. The of the systems and services described is limited to those installed in and managed by Group CCN-7.

ASC Platforms at Los Alamos

Delivering Computational and Data Storage Services to the ASC Program

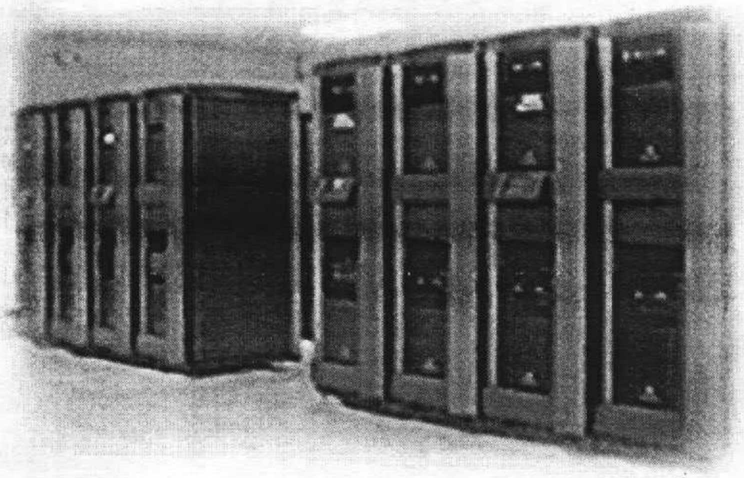
Steven R. Shaw, CCN-7

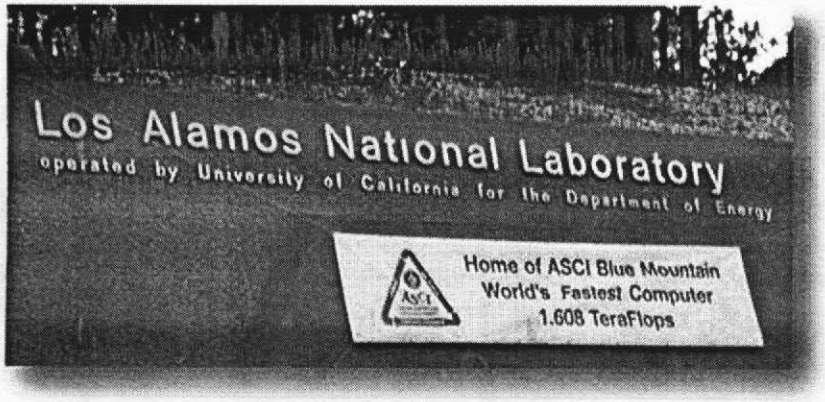
High Performance Computing Systems
Computing, Communications, and Networking Division

November 9, 2004

Agenda

- Time of Transition – when isn't it in this business?
- Blue Mountain History and Decommission
- AlphaServer (Q like) Capability and Capacity Systems – our current workhorse platforms
- Linux Based Capacity Clusters – our emerging ASC capacity platforms
- Our constant – Delivering Computational and Data Storage Services to the ASC Program




Los Alamos National Laboratory
operated by University of California for the Department of Energy
ASCI
Home of ASCI Blue Mountain
World's Fastest Computer
1.608 TeraFlops

Blue Mountain

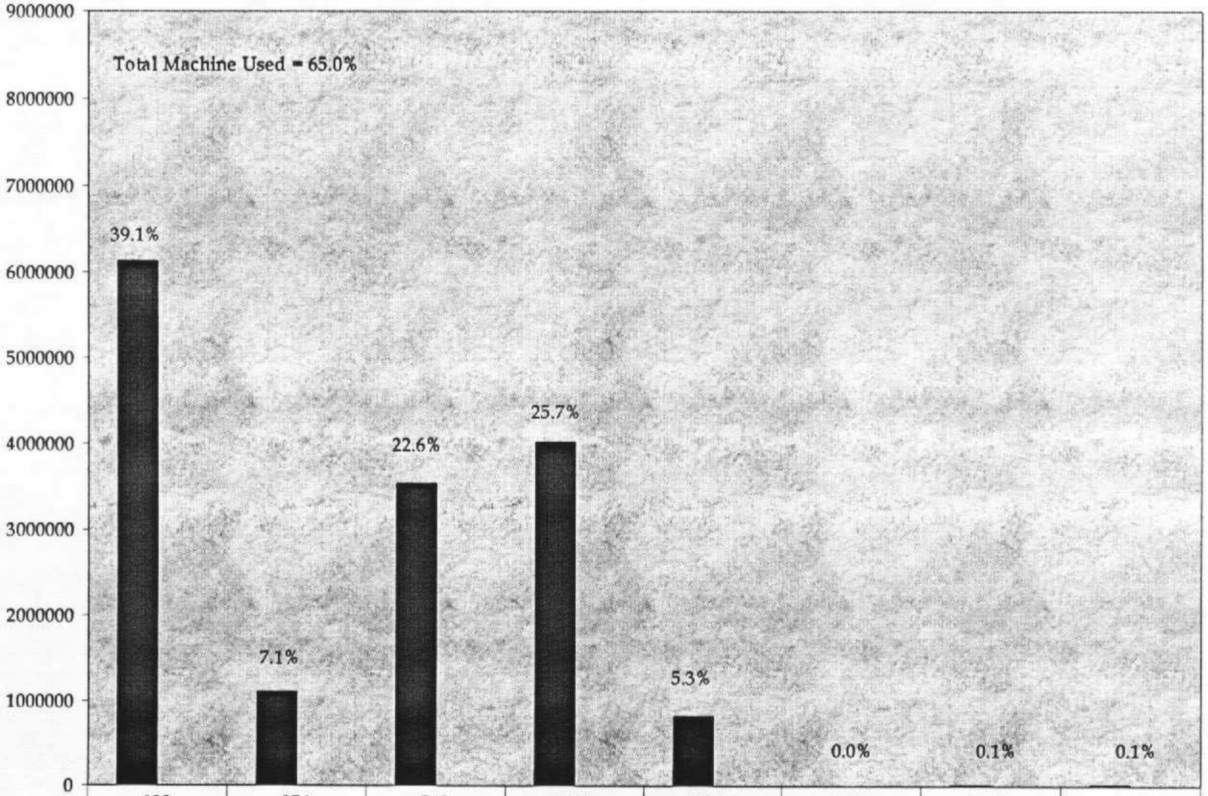
- 48 SGI Origin2000s with 6144 CPUs
- Peak Rating of 3.076TFlops

- Debuted on the Top500 in June of 1999 as number 2 at 1.068Tflops

- #135 on the June 2004 list

- Decommissioning the system (except visualization servers) as we speak

Blue Mtn. YTD - Job Size
1/1/04 - 9/11/04



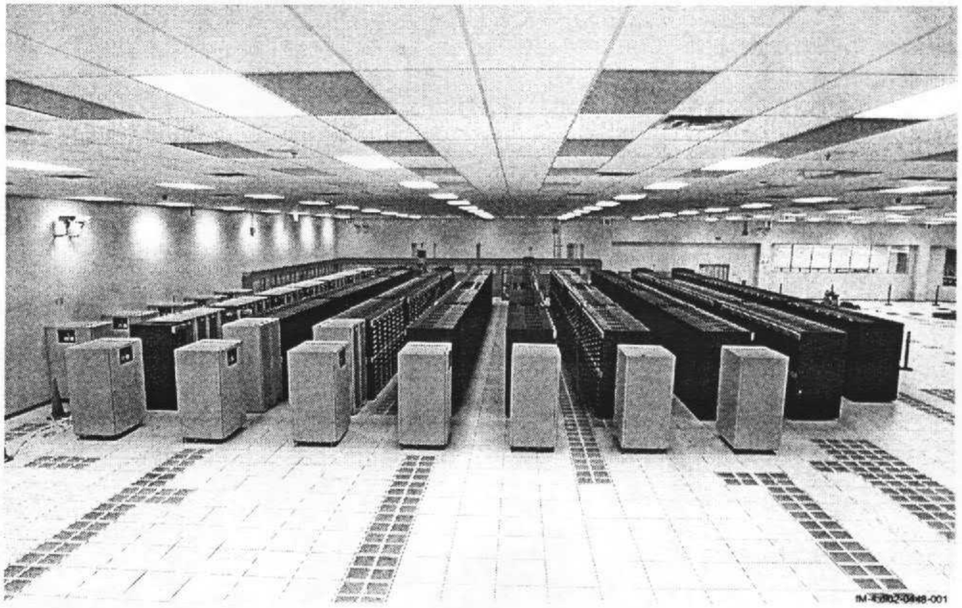
	128	256	512	1024	2048	3072	4096	6144
Proc Hrs	6131454	1114893	3546322	4022767	830100	1002	9856	9249
# jobs	569856	933	1733	906	98	2	12	10

AlphaServer Q Systems

QA (2/02) and QB (10/02)

Each System:

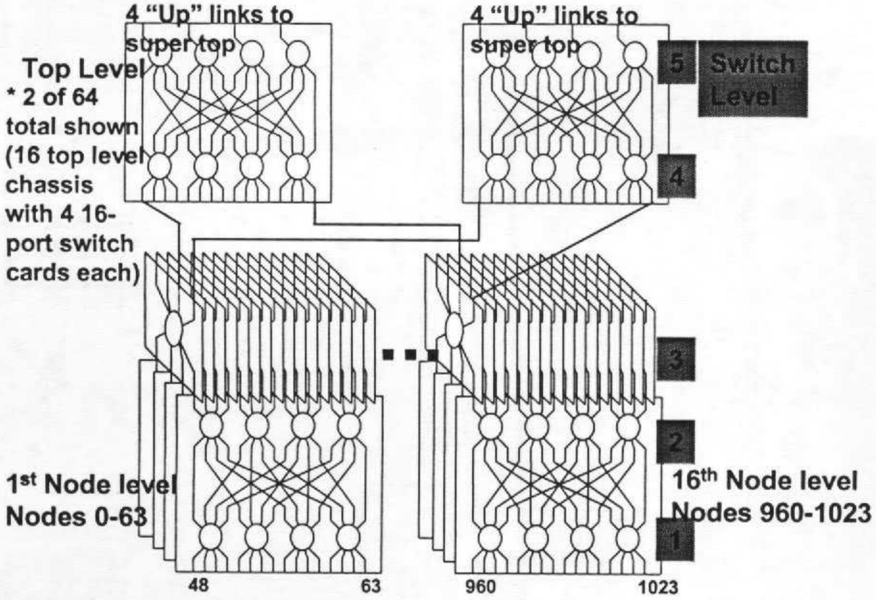
- 1024 HP ES45 nodes
- 4096 1.25GHz EV68 CPUs
- 10.24TF peak
- 11GB average memory per node (8, 16 and 32)
- 224TB unformatted local storage
- Dual Rail Quadrics Elan3



- Numbers 2 & 3 on the November 2002 Top500 at 7.727 TFlops

Connecting QA and QB

1024-Node Segment Logical Topology


4 "Up" links to super top
4 "Up" links to super top
Top Level
* 2 of 64 total shown (16 top level chassis with 4 16-port switch cards each)
5
Switch Level
4
3
2
1
1st Node level Nodes 0-63
16th Node level Nodes 960-1023
48
63
960
1023

- •Node level switches provide 64 node connections and 64 up links to top level switches

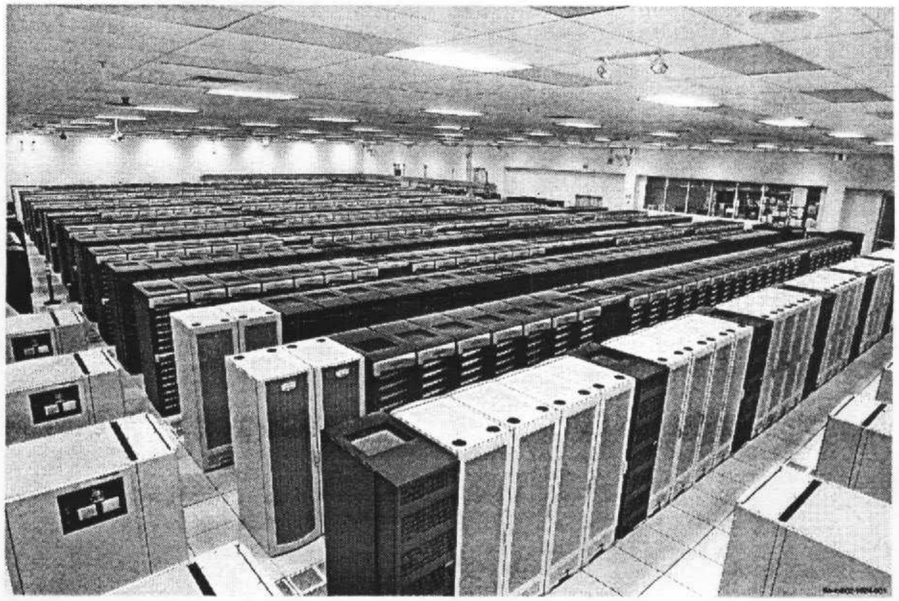
- •Top level switches interconnect node switches and provide up links to super top switches

•Super top switches interconnect two 1024-node segments to complete 2048-node topology

AlphaServer Q Systems

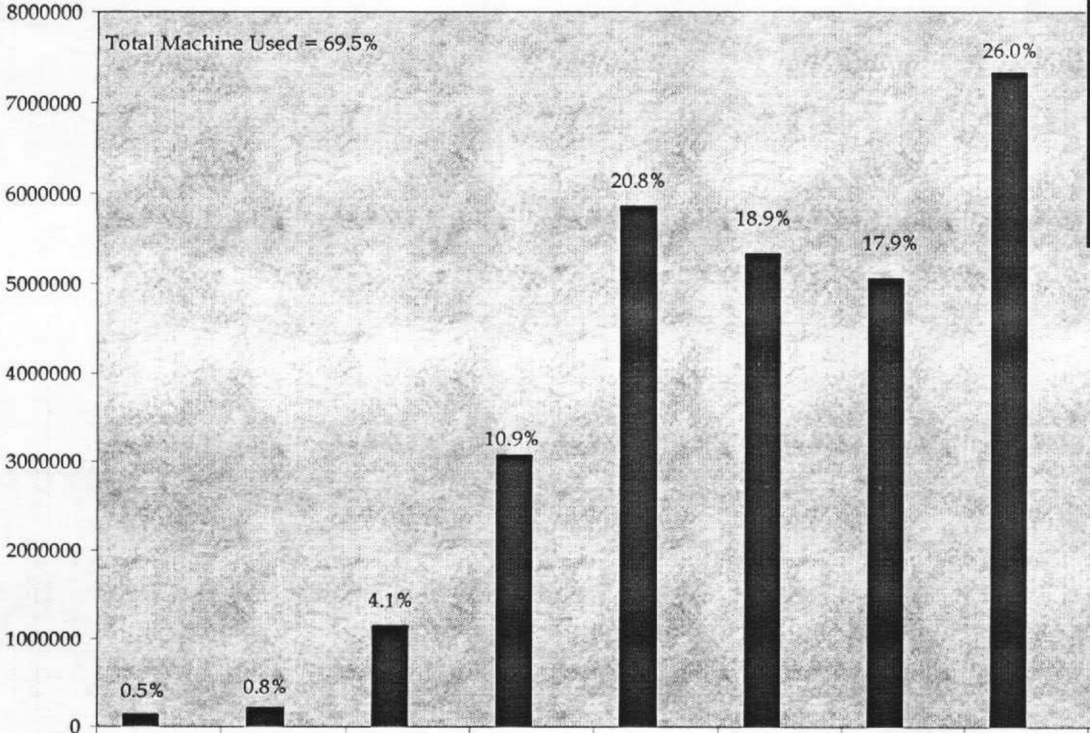
QA and QB Joined = Q

- 2048 HP ES45 nodes
- 8192 1.25GHz EV68 CPUs
- 20.48TF peak



- Number 2 on the June 2003 Top500 at 13.88Tflops

Q YTD
1/1/04 - 9/4/04



	4	32	64	128	256	512	1024	2048	3072	4096
Proc Hrs	132148	221783	1151354	3076709	5865024	5336774	5057944	7337420	48599	18163
75257	4889	7201	9940	9114	2549	2170	707	13	82	

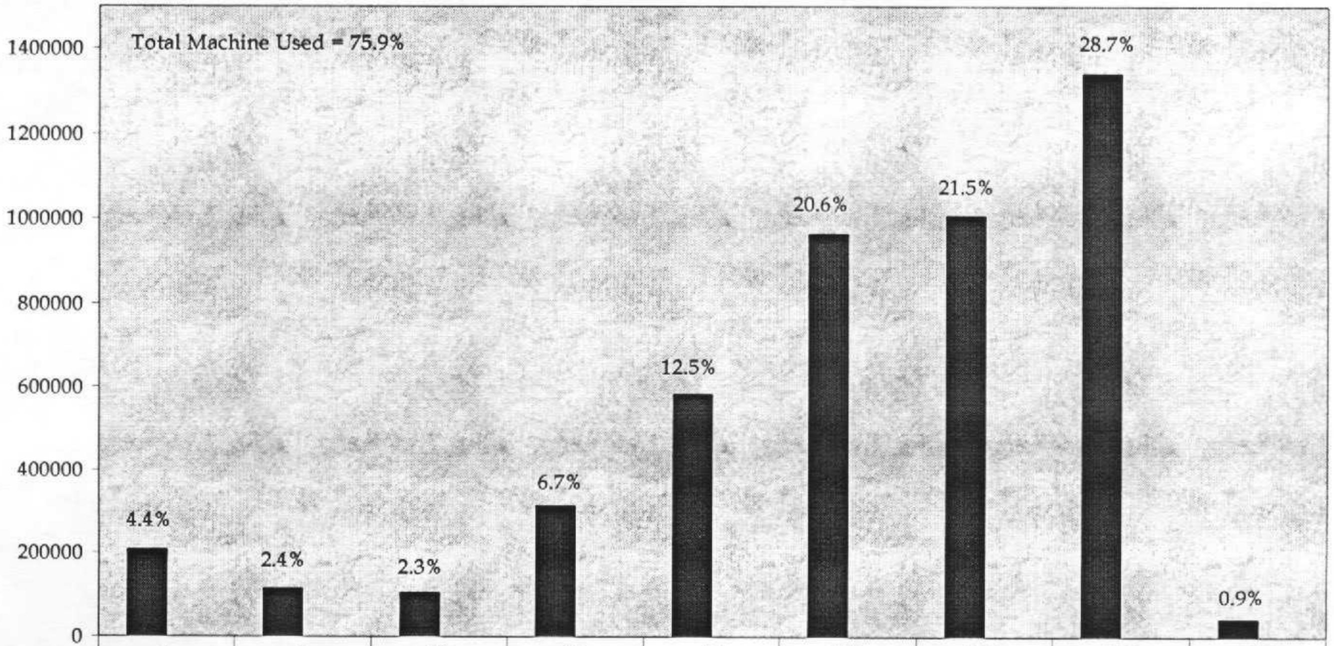
AlphaServer System in the Open Protected Network – QSC

- 256 node AlphaServer ES45 cluster
- Dual Rail QSW Elan 3 interconnect
- 16 GB/node memory per node
- 30TB Local Storage

- Usage allocations

  25% NW
  40% ASC Alliances
  35% Institutional Computing

QSC YTD
1/1/04 - 10/16/04



	4	8	16	32	64	128	256	512	1024
Proc. Hrs.	208032	114634	107107	312561	584566	963650	1007962	1342282	42771
# of Jobs	91986	4935	2827	4403	3477	2623	1235	804	135

The Need for Capacity Computing

- Successful use of ASC codes for programmatic goals
- Offload capacity work from LANL capability machines (QA & QB)
- Increased demand in capacity class workload (256 or less PEs)
- Blue Mountain plans for reduction and removal
- JASON Study Report

AlpahServer Capacity Systems
CA, CB, CC and CX

CA, CB, and CC – Each system

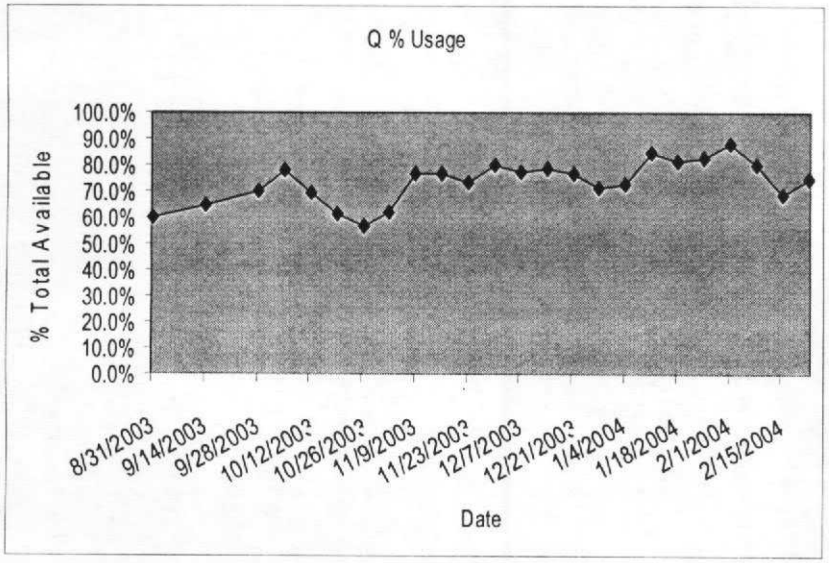
128 Nodes HP ES45 – 512 1.25 GHz CPUs
4GB memory per node
13.9TB local storage
Single rail QSW Elan3 interconnect
1.28 TFlop peak

CX – Optimized for serial processes

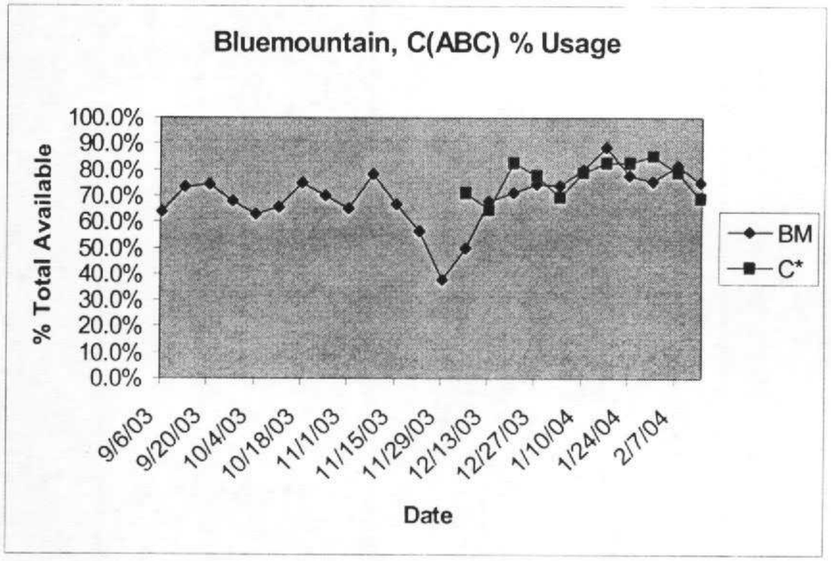
32 Nodes HP ES45 – 128 1.25Ghz CPU
6.7 TB local storage
Single rail QSW Elan3 interconnect
320 GFlop peak

Installed in October 2003

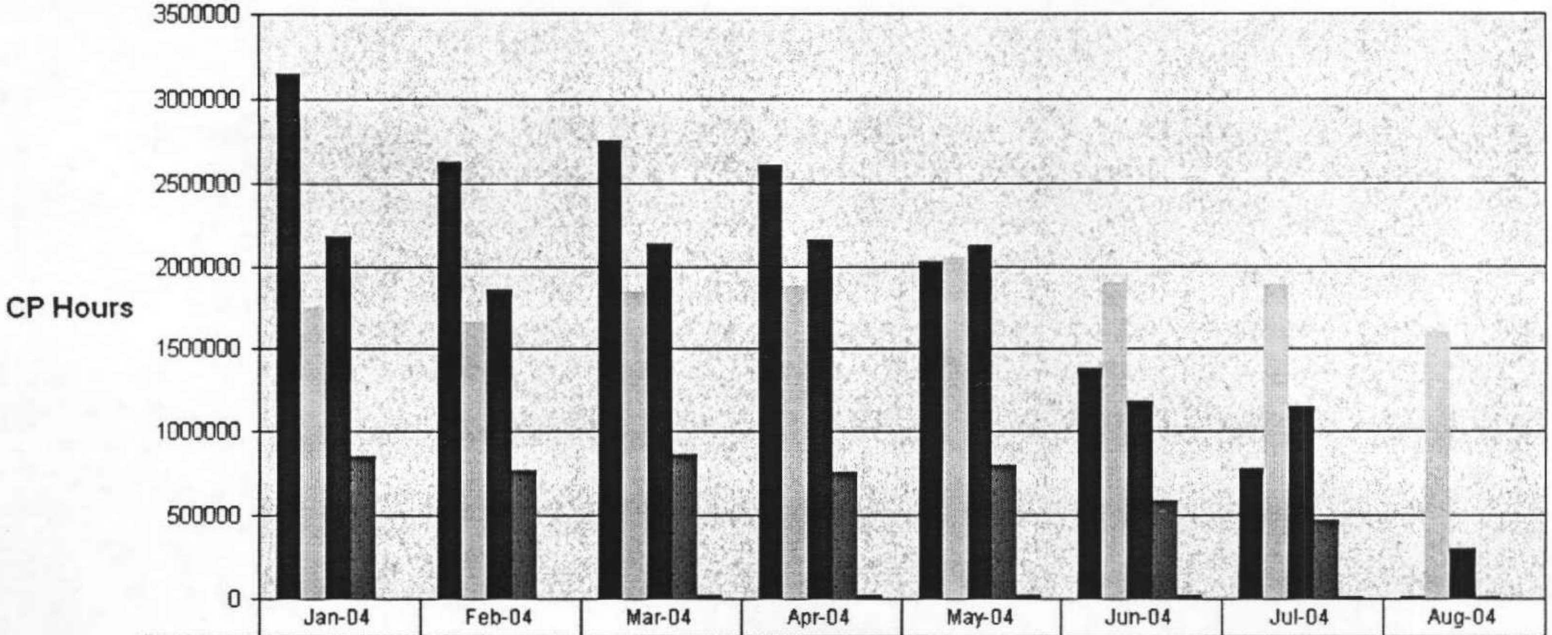
Capacity quickly consumed



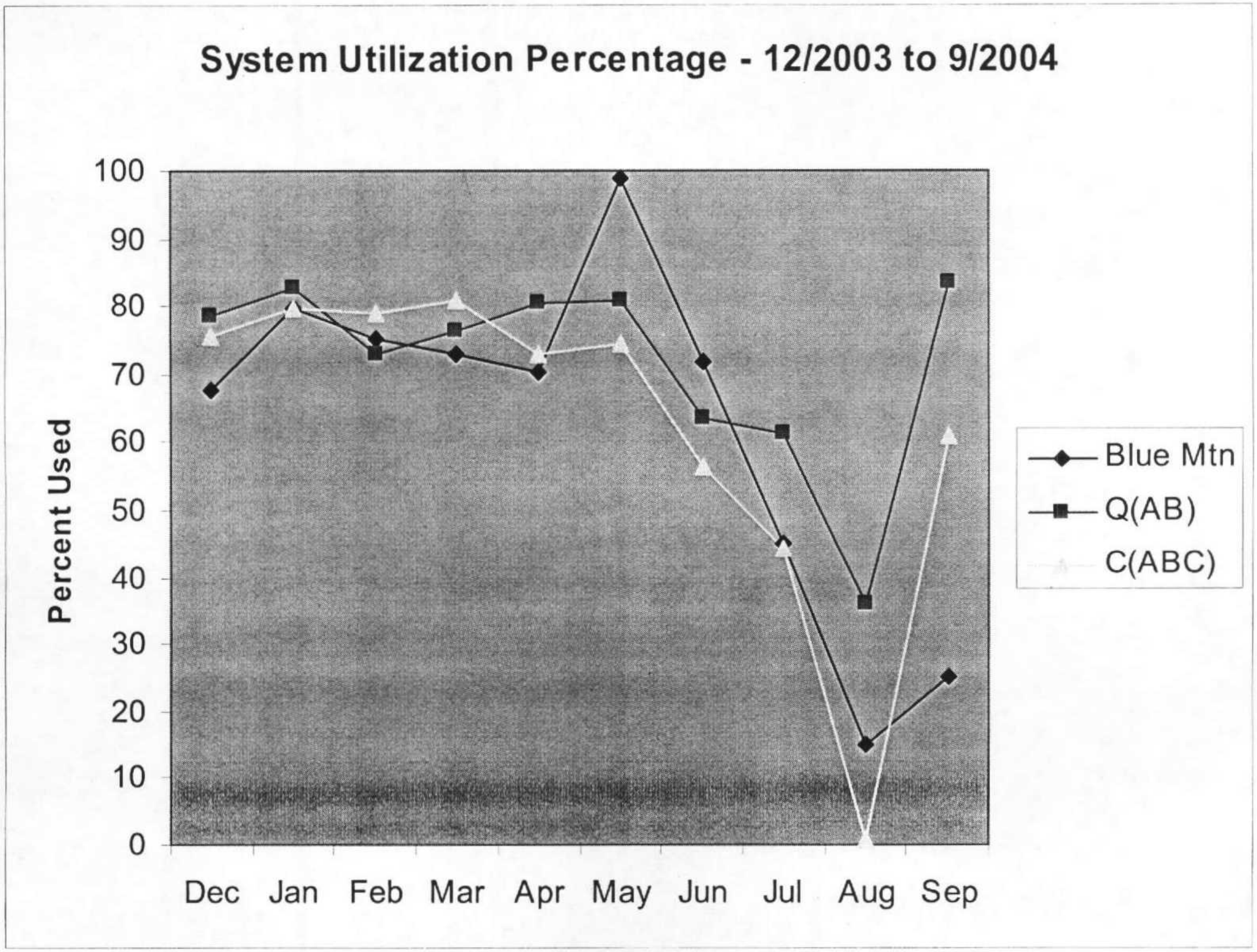
Caveat, of course all systems are not equal



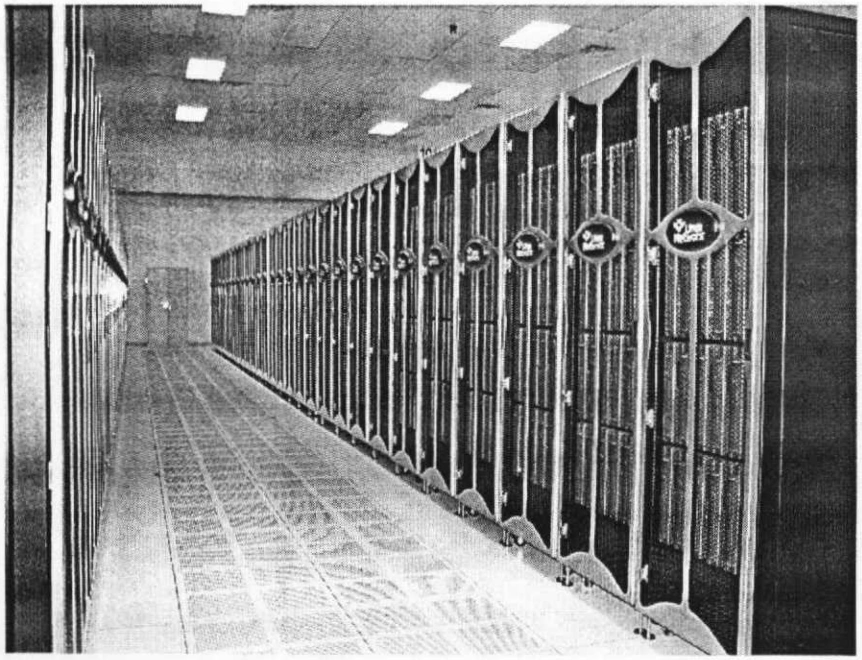
Secure Machine Usage
Jan - Aug 2004



	Jan-04	Feb-04	Mar-04	Apr-04	May-04	Jun-04	Jul-04	Aug-04
■ Blue Mtn	3150191	2622232	2751790	2599520	2032089	1390516	776660	7360
■ QA	1747716	1663880	1841707	1876199	2049583	1895418	1891953	1601432
■ QB	2176074	1857190	2135016	2152536	2123147	1181852	1155085	295740
■ C(ABC)	855378	763458	865411	755576	798033	585756	473055	7118
■ CX	0	0	23649	24688	17066	21465	9453	75

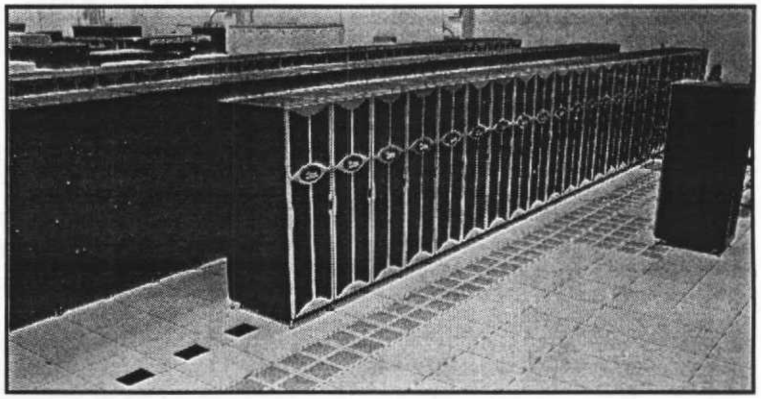

System Utilization Percentage - 12/2003 to 9/2004
Percent Used
100
90
80
70
60
50
40
30
20
10
0
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Blue Mtn
Q(AB)
C(ABC)

Capacity Computing Vision- Beyond Q



- Augment existing HPC systems with Linux-based capacity clusters
- Lower integration and operational costs by leveraging internal resources and open source software
  - CCN takes on role of system integrator
  - Successful collaborative relationships established with LAMPI, Panasas FS, Science Appliance and third-party software agents
- Establish a standard production system and user environment and use ASC Tier 1 SQA project to tie them together

Lightning Capacity System Overview



System Hardware

- 1408 dual-processor AMD Opteron nodes (11.26 TeraOps peak, ~5.6 TB memory)
    - One Arima Rio Works HDAMA system board with AMD 8111 and 8131 chipsets
    - Two 2.0 GHz 64-bit processors with 1 MB L2 cache/node
    - Four GB of memory/node
    - One 120-GB disk drive/node
    - One ICEBOX controller/node for hardware monitoring
    - Scalable to 2048 nodes (scalable design plans for interconnect)
- Myrinet Interconnect (latency ~7 usec, bandwidth ~250 MB/sec)
- Gigabit copper network to network services such as NFS, Panasas
- A copper-based 10/100 network for system monitoring system reboot, etc.

System Software

Linux

Science Appliance software

Beoboot, LinuxBios, Bproc, Supermon

Compilers

Message Passing

LAMPI

MPICH as risk mitigation

Debugging - TotalView

Archival storage - HPSS

Resource management - Load Sharing Facility (LSF)

Lightning Project Status

- The Lightning project involved the successful implementation of both a new Object-based storage file system (Panasas) and a new operating system model (BProc).
- 30 days after delivery of the hardware, a LINPACK test was run that achieved sixth spot on November 2003 Top 500 list at 8.051 TFlops
- Completed a milepost demonstrating that key ASCI application code teams can make effective use of a Linux-based software environment on a secure capacity-computing platform.
- We will complete a Production Computing milepost by 11/15 to demonstrate that the Lightning cluster is available to the targeted LANL ASCI user community in 32-bit mode.
- Developed new optimization, debugging and configuration management environments.
- The Lightning Project Team was awarded a NNSA Defense Programs Awards of Excellence

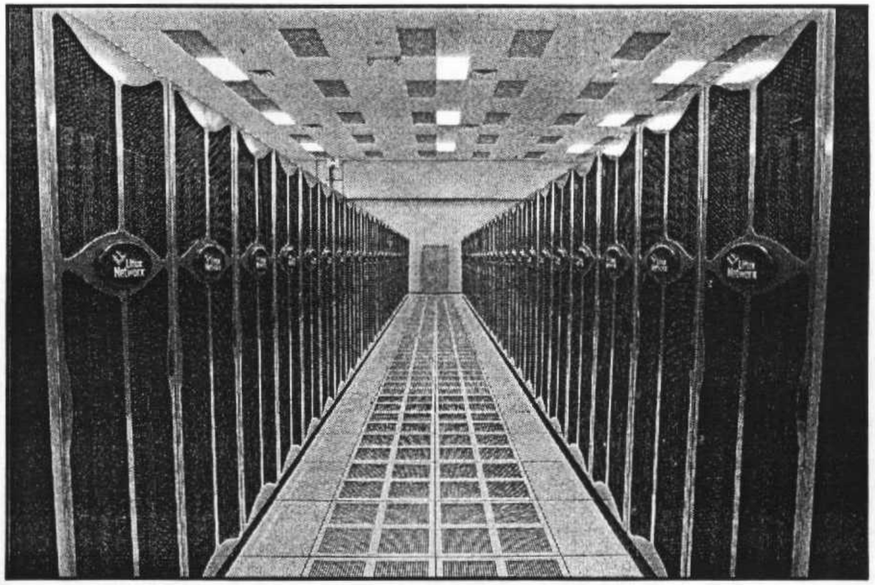
Linux Based Bproc Capacity Cluster Plans

Hardware on Lightning (Q1 FY05)
- Install additional 256 Nodes
- Increase Panasas to 200TB
- GIG-E reconfiguration and FS (Panasas) integration

384 Node System in the Open Protected Network to augment / offload QSC

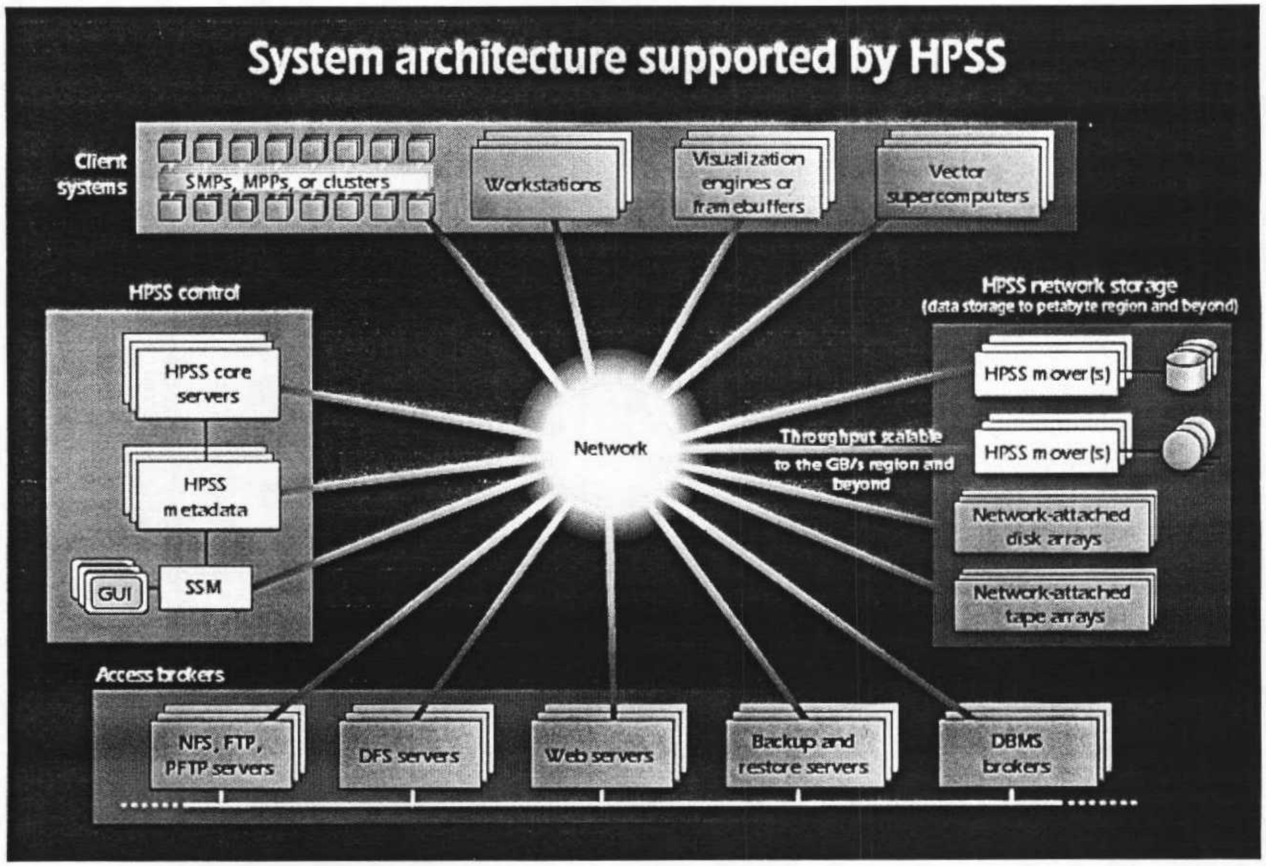
Full 64 bit support

Leverage knowledge gained from integration and operation of non ASC Clusters (Pink, TLC, others)



Data Storage Services – HPSS

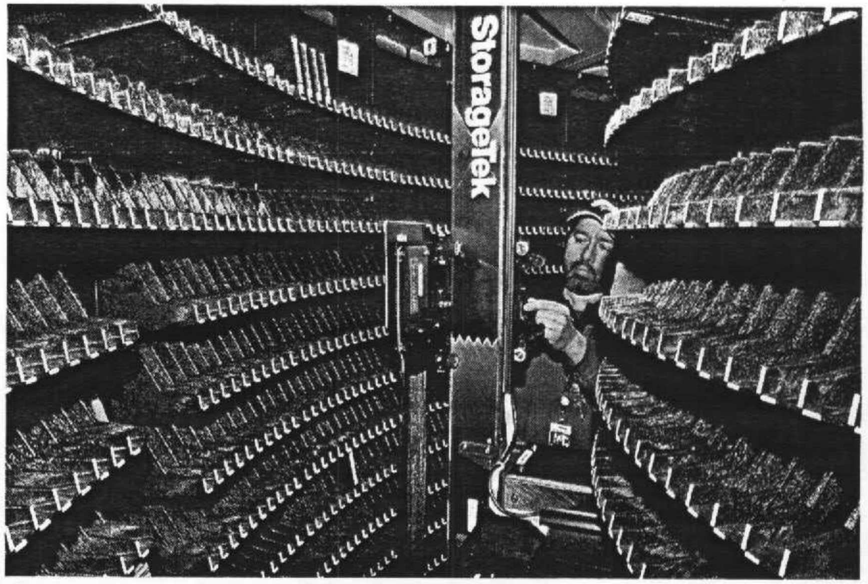
The Los Alamos HPSS configuration includes a variety of worker machines that are served by HPSS via different network paths.


System architecture supported by HPSS
Client systems
SMPs, MPPs, or clusters
Workstations
Visualization engines or framebuffers
Vector supercomputers
HPSS control
HPSS core servers
HPSS metadata
GUI
SSM
Network
Throughput scalable to the GB/s region and beyond
HPSS network storage (data storage to petabyte region and beyond)
HPSS mover(s)
HPSS mover(s)
Network-attached disk arrays
Network-attached tape arrays
Access brokers
NFS, FTP, PFTP servers
DFS servers
Web servers
Backup and restore servers
DBMS brokers

HPSS Design Goals

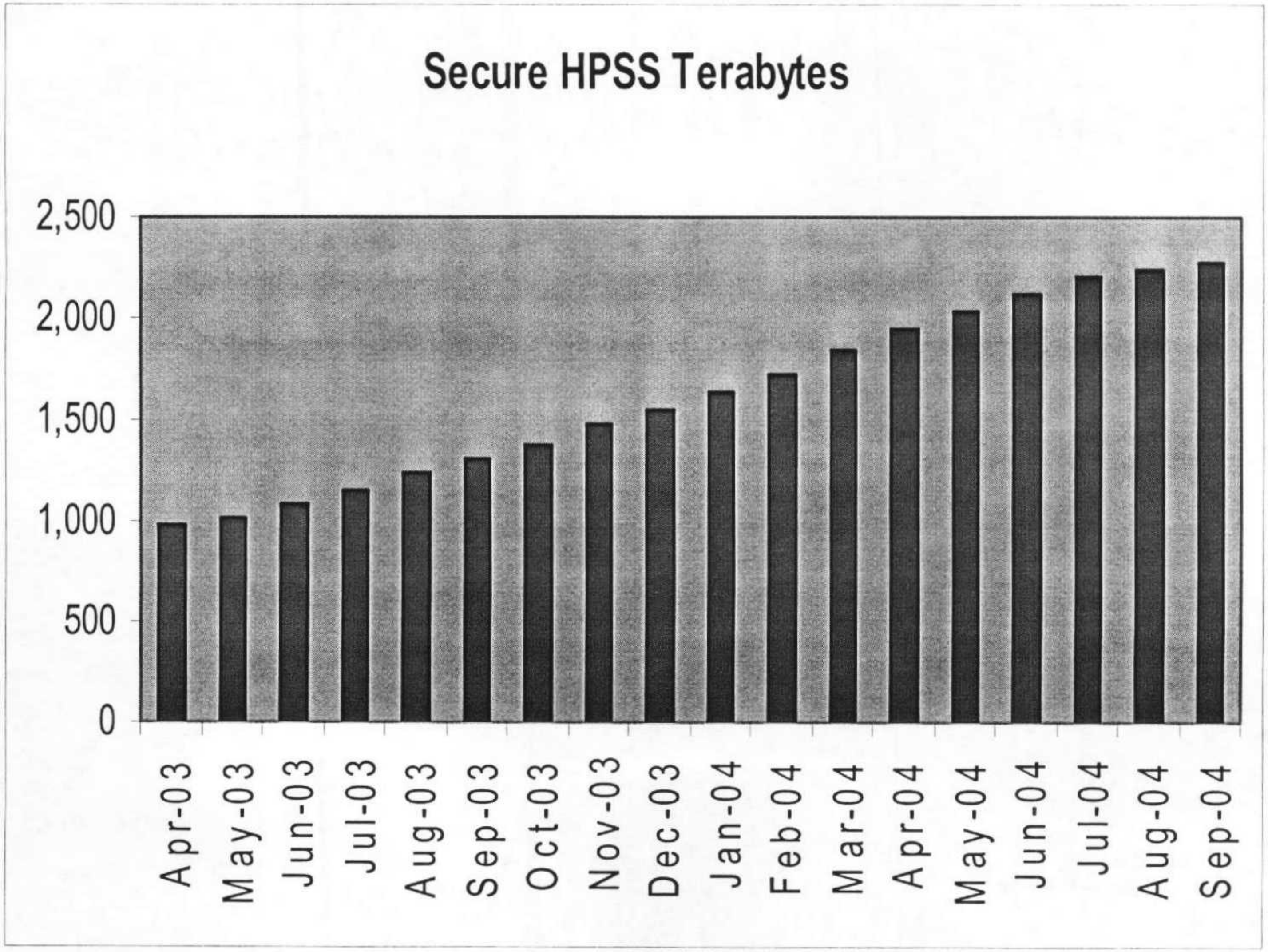
- Serve users whose requirements for storing massive amounts of data generated during simulation runs can not be met by other storage systems.

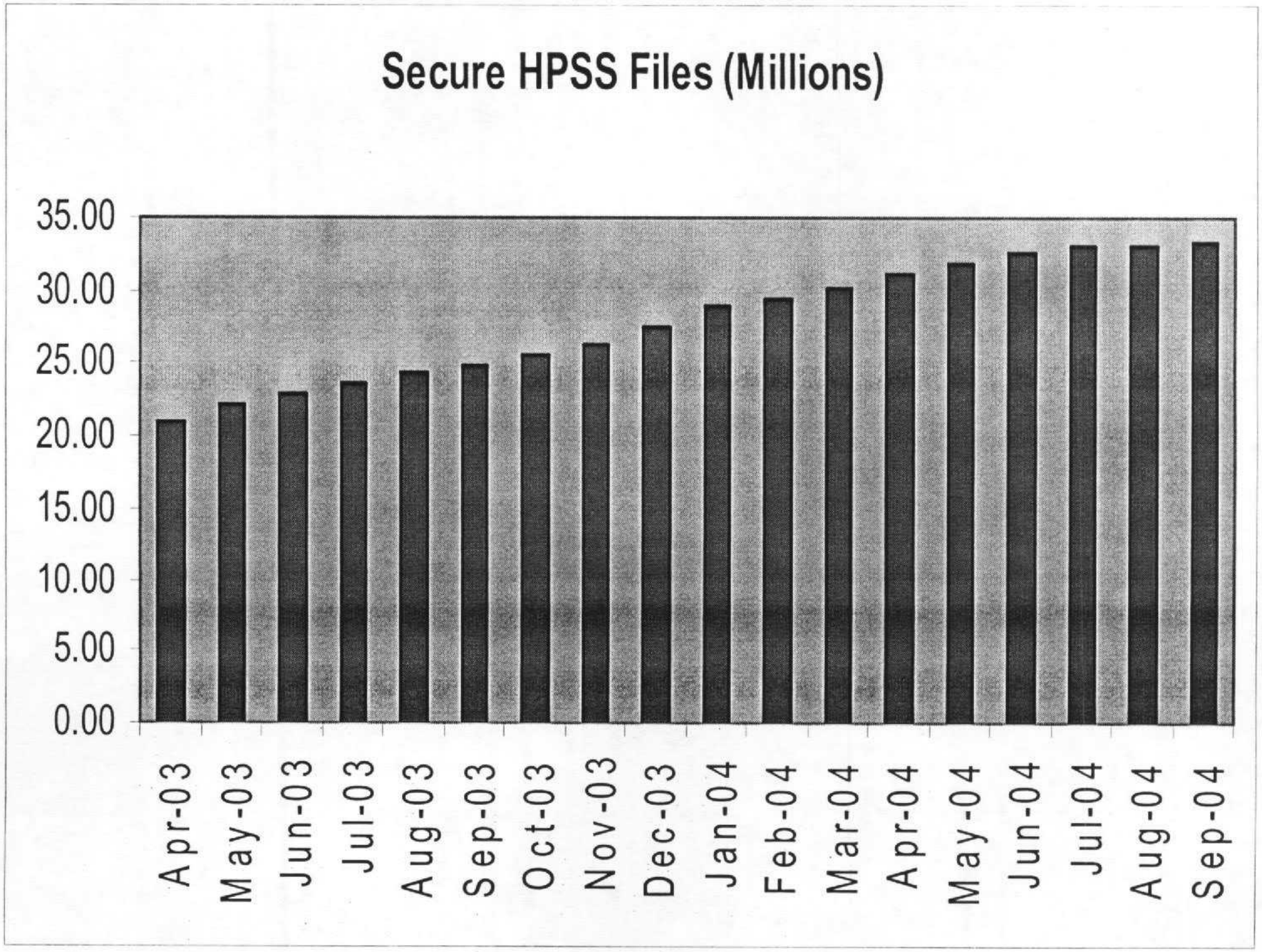
- A storage system that can store extremely large files, has high data-transfer rates, has storage capacity for petabytes of data, and is scalable in a cost-effective way.



Los Alamos HPSS Status and Plans

- Archival storage contains ~ 2,500 terabytes of data in 34 million files for 1,300 ASC users. (FY05 – est. growth of 1,200TBs)
- Continuous modernization has kept HPSS viable with our changing infrastructure and increased user demand for both capacity and performance. (FY05 - HPSS v6.2)
- HPSS is used as a tri-lab resource and provides a centralized, secure repository for classified data.
- Planning & development underway to add even more parallelism to HPSS in order to decrease the total time to archive for ASC users. (FY05 – PSI enhancements)
- Research is underway to integrate HPSS with highly-parallel, objected-based file systems with the goal of improved end-to-end performance. (FY05 – U of M contract)


Secure HPSS Terabytes
2,500
2,000
1,500
1,000
500
0
Apr-03
May-03
Jun-03
Jul-03
Aug-03
Sep-03
Oct-03
Nov-03
Dec-03
Jan-04
Feb-04
Mar-04
Apr-04
May-04
Jun-04
Jul-04
Aug-04
Sep-04


Secure HPSS Files (Millions)
35.00
30.00
25.00
20.00
15.00
10.00
5.00
0.00
Apr-03
May-03
Jun-03
Jul-03
Aug-03
Sep-03
Oct-03
Nov-03
Dec-03
Jan-04
Feb-04
Mar-04
Apr-04
May-04
Jun-04
Jul-04
Aug-04
Sep-04

Other sessions of interest to learn more ASC Platforms at Los Alamos

- The ASCI/DOD Scalable I/O History and Strategy - Jamez Nunez and Gary Grider
- Bringing Clustermatic Systems into Production – Poster in the Los Alamos Booth – Andrew Shewmaker and Harvey Wasserman
- LANL's MPI and Open MPI – Rich Graham
- High Performance Storage for Cluster Computing: What do we do Next? - Garth Gibson (Panasas)
- LANL ASC Visualization Corridor Update- Dave Modl
- The Ribosome Comes to Life on ASC Q: Movement of RNA into the Ribosome - Dr. Kevin Sanbonmatsu
- Javelina: Analysis Tool for Studying Application Based Source Test Coverage - David R. Kent
- The Eclipse Parallel Tools Project - Greg Watson
- Unified Data Model - Parallel I/O Library for Simulations and Data Management - William Dai