



Red Storm System Integration

Red Storm Presentations in
the ASCI Research Exhibit at SC03
James A. Ang, Coordinator & Editor

November 20, 2003
Phoenix, AZ



Topics and Presenters

Red Storm System Integration

- Jim Tomkins: System architecture
- Len Stans: SCA/CUB facility
- John Noe: Operations Overview
- Rob Leland: Application Scaling Performance
- Erik Debenedictis: Software Environment
- Philip Heermann: Results Visualization
- Luis Martinez: Interfaces to DisCom WAN
- Barbara Jennings: User Support





Facility For

Red Storm

Sandia's Supercomputer

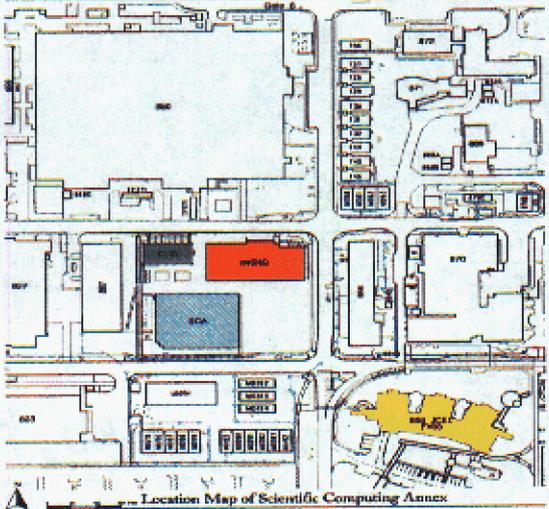
Len Stans, 9336
Sandia National Laboratories
November 2003



1

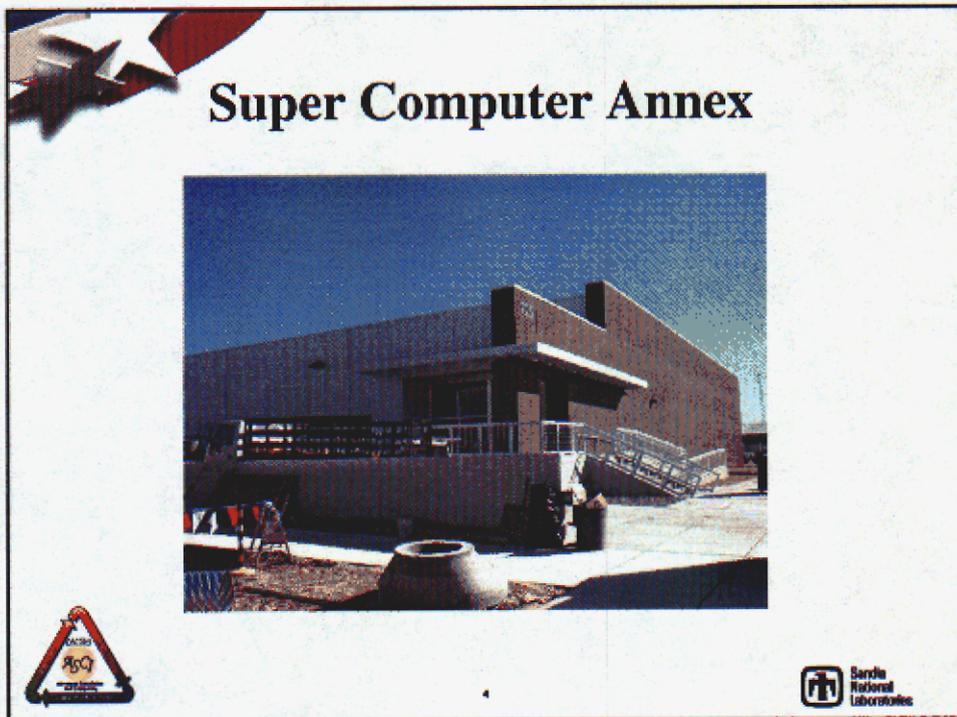
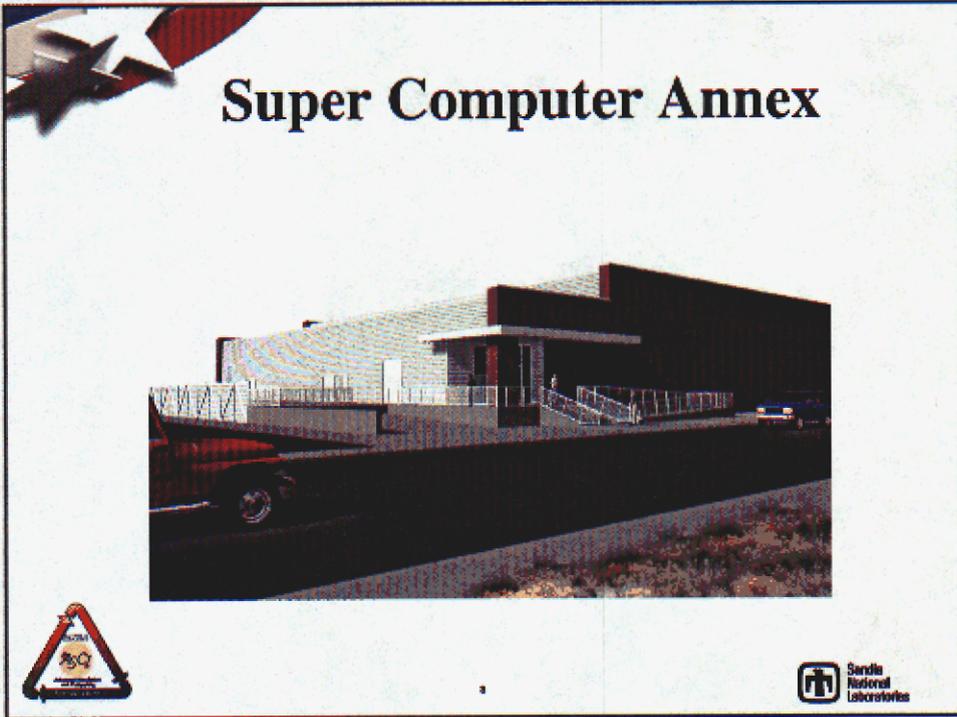


Location Of The Super Computer Annex



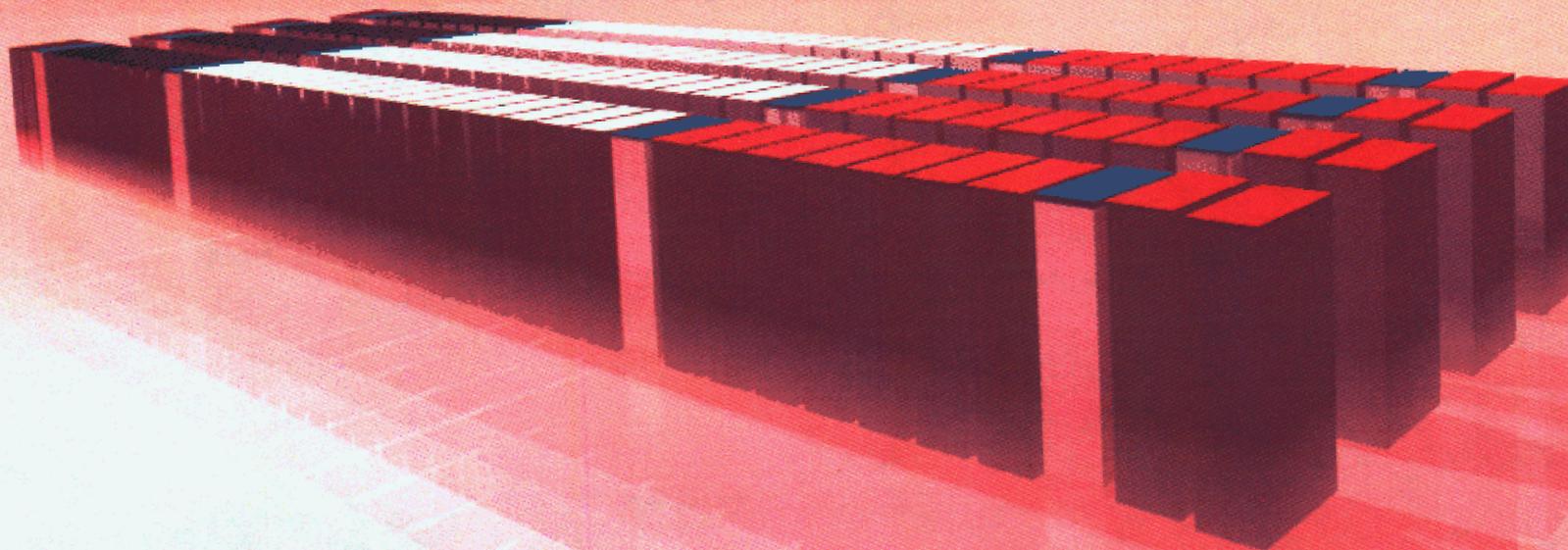
Location Map of Scientific Computing Annex

2



Red Storm Overview and Update

Jim Tomkins





Red Storm Facilities

Computer Utility Building (CUB)

- **To Supply Chilled Water for Bldg. 880 and/or the Super Computer Annex (SCA)**
- **Started Sept 2002**
- **Completed Oct 2003**
- **The CUB Provides 2000 Tons of Chilled Water to Support the Red Storm Cooling Needs and Can Be Expanded to Provide Up to 4000 Tons of Cooling When Needed**





Red Storm Facilities

Super Computer Annex (SCA)

- **Started- June 2002**
- **Completed- Oct 2003**
- **A 20,250 Sq. Ft. Facility With a Clear Span of 150' With an Additional 4,500 Sq. Ft of Space for Office and Support Equipment**
- **The Facility Provides 3.5 Mega Watts of Electrical Power That Can Be Expanded to 7 Mega Watts of Electrical Power If Needed**






A "Big" Empty Floor



7



24 Power Distribution Units



8

Outline

Architecture

Goals

Overall System

Topology and Cabinet Layout

RAS System

System Software

Performance

Overall System

Processors and Memory

Interconnect and I/O

Project Status



Red Storm

Sandia's Second ASCI Capability Platform

It's the *Next Big Thing!*



SC 2003
Nov 20, 2003

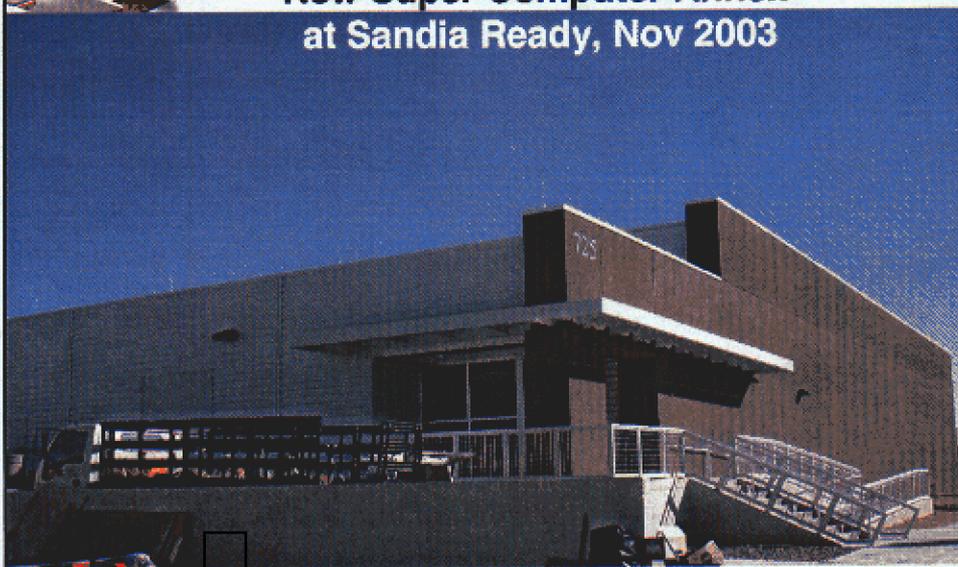
John P. Noe, Manager
Terascale Systems, Dept. 9328



Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.



New Super Computer Annex at Sandia Ready, Nov 2003



John P. Noe, Dept. 9328, ppt





Operational Concepts for Red Storm build on success achieved with ASCI Red

We expect to follow the established progression for ASCI
Capability Platforms while continuing operation of ASCI Red:

High Performance Computing LAN - SXN

Installation and Hardware Verification/Validation

3 months (June, July, August 2004)

Initial Acceptance/System Integration

3 to 6 months (**7x performance goal**)

Limited Availability (Few codes and users)

6 to 12 months (**Sustained System Reliability**)

General Availability (More codes and users)

Remainder of system lifetime

John F. Nov, Dept. 8008 ppt



Operational Concepts for Red Storm build on success achieved with ASCI Red

You can expect a regular schedule for system time, dedicated
time, resource movement between classified and unclassified. We
will continue the weekly reliability reporting and utilization
coordination meetings with LANL and LLNL(EPR).

Center segment of Red Storm will be switched bi-weekly
beginning with Limited Availability. ~ **1 Jan 05**

Can expect frequent system dedicated time for OS
development/debugging throughout Limited Availability.

Batch job queue structure and runtime/node count limits
will be refined during Limited Availability.

Tentative schedule for **Classified Processing** is six weeks
after Limited Availability commences. ~ **15 Feb 05**

John F. Nov, Dept. 8008 ppt



Red Storm Goals

Balanced System Performance - CPU, Memory, Interconnect, and I/O.

Usability - Functionality of hardware and software meets needs of users for **Massively Parallel Computing**.

Scalability - System Hardware and Software scale, single cabinet system to ~20,000 processor system.

Reliability - Machine stays up long enough between interrupts to make real progress on completing application run (at least 50 hours MTBI), requires full system RAS capability.

Upgradability - System can be upgraded with a processor swap and additional cabinets to 100T or greater.

Red/Black Switching - Capability to switch major portions of the machine between classified and unclassified computing environments.

Space, Power, Cooling - High density, low power system.

Price/Performance - Excellent performance per dollar, use high volume commodity parts where feasible.



Sharing Red Storm in the ASCI Community

System resources will be allocated as agreed by ASCI Execs:
(Note – need to validate these percentages prior to talk)

Classified:

		Node Hrs/Month
SNL	50%	1,824,000
LLNL	25%	912,000
LANL	25%	912,000

Unclassified:

Alliances	20%	730,000
SNL	40%	1,460,000
LLNL	20%	730,000
LANL	20%	730,000

John P. Hsu, Dec 2000, ppt



Operational Concepts for Red Storm build on success achieved with ASCI Red

Multiple File Systems will be created with some space reserved for test and evaluation of configurations. (e.g. /scratch1, /Milestone1, /configtest)

File purge policy will be coordinated with early customers. Experience with Data Services clusters will help determine needs in this regard.

Red Storm will be accessible from the ASCI DisCom WAN once classified processing is approved.

John P. Hsu, Dec 2000, ppt





Customer Support for Red Storm continues Intel/Sandia model from ASCI Red

Cray, Inc will provide three on-site Computational Scientists to assist in code conversion, debugging, optimization, problem tracking and resolution. In addition to the five Parallel Systems Engineers who will perform day to day system administration and hardware support, these people will collaborate with the Tri-Lab User Support groups to address customer needs in all aspects of production use of Red Storm.

Cray, Inc. has Mike Davis and Howard Pritchard on board now working on ASCI codes. Meet them at the Cray, Inc booth.

Barbara Jennings will expand on User Support later today.

John P. Hsu, Dept. 6020 pg7



Red Storm Thor's Hammer Mjölfnir the First



John P. Hsu, Dept. 6020 pg8



Red Storm Architecture

True MPP, designed to be a single system.

Distributed memory MIMD parallel supercomputer.

Fully connected 3-D mesh interconnect. Each compute node and service and I/O node processor has a high bandwidth, bi-directional connection to the primary communication network.

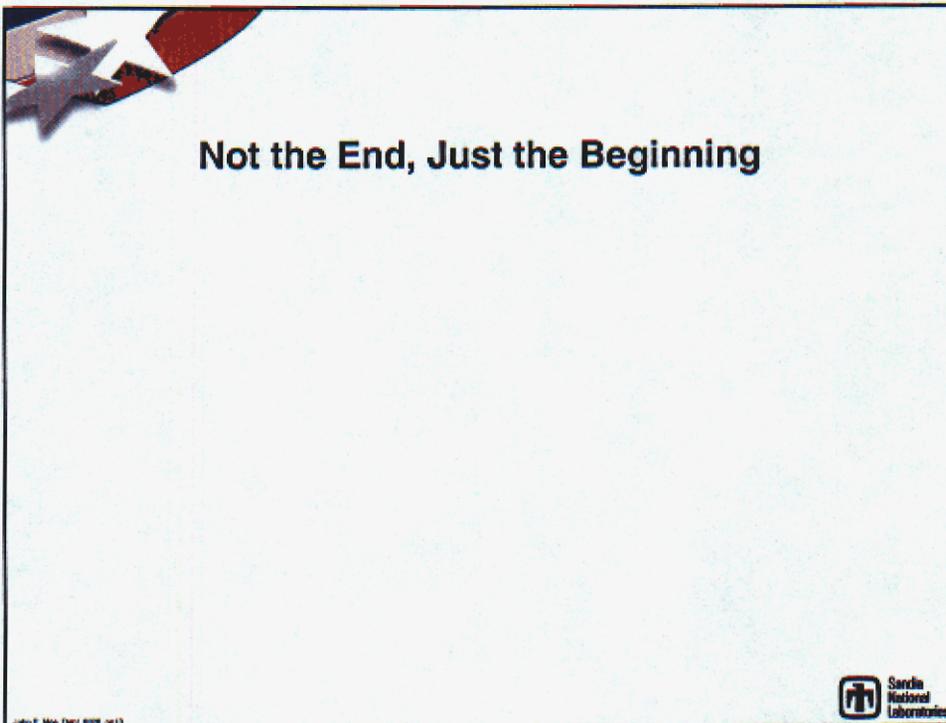
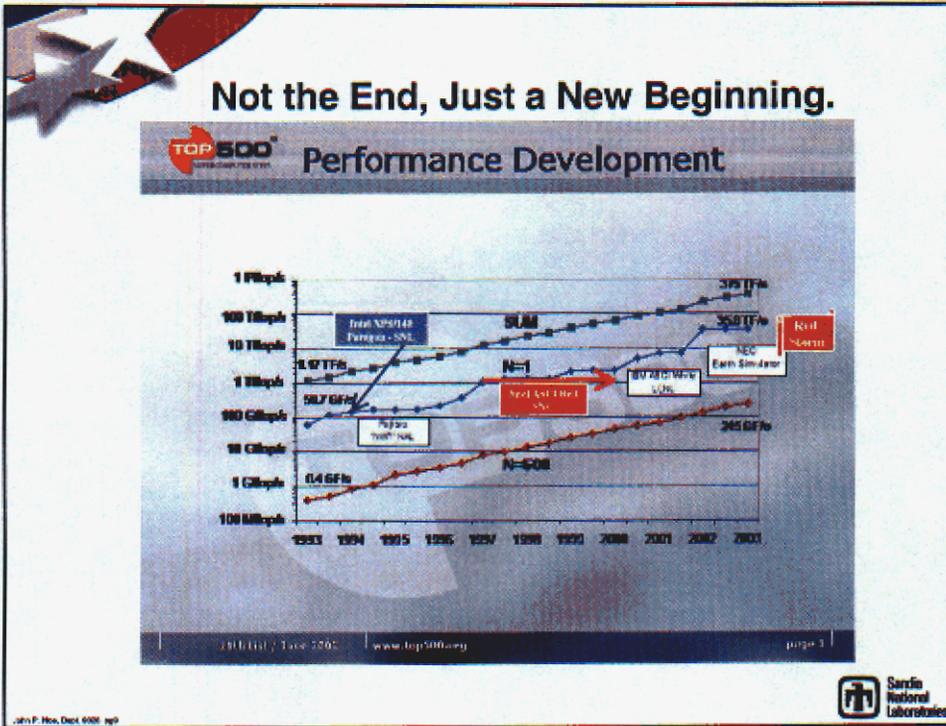
**108 compute node cabinets and 10,368 compute node processors.
(AMD Opteron @ 2.0 GHz)**

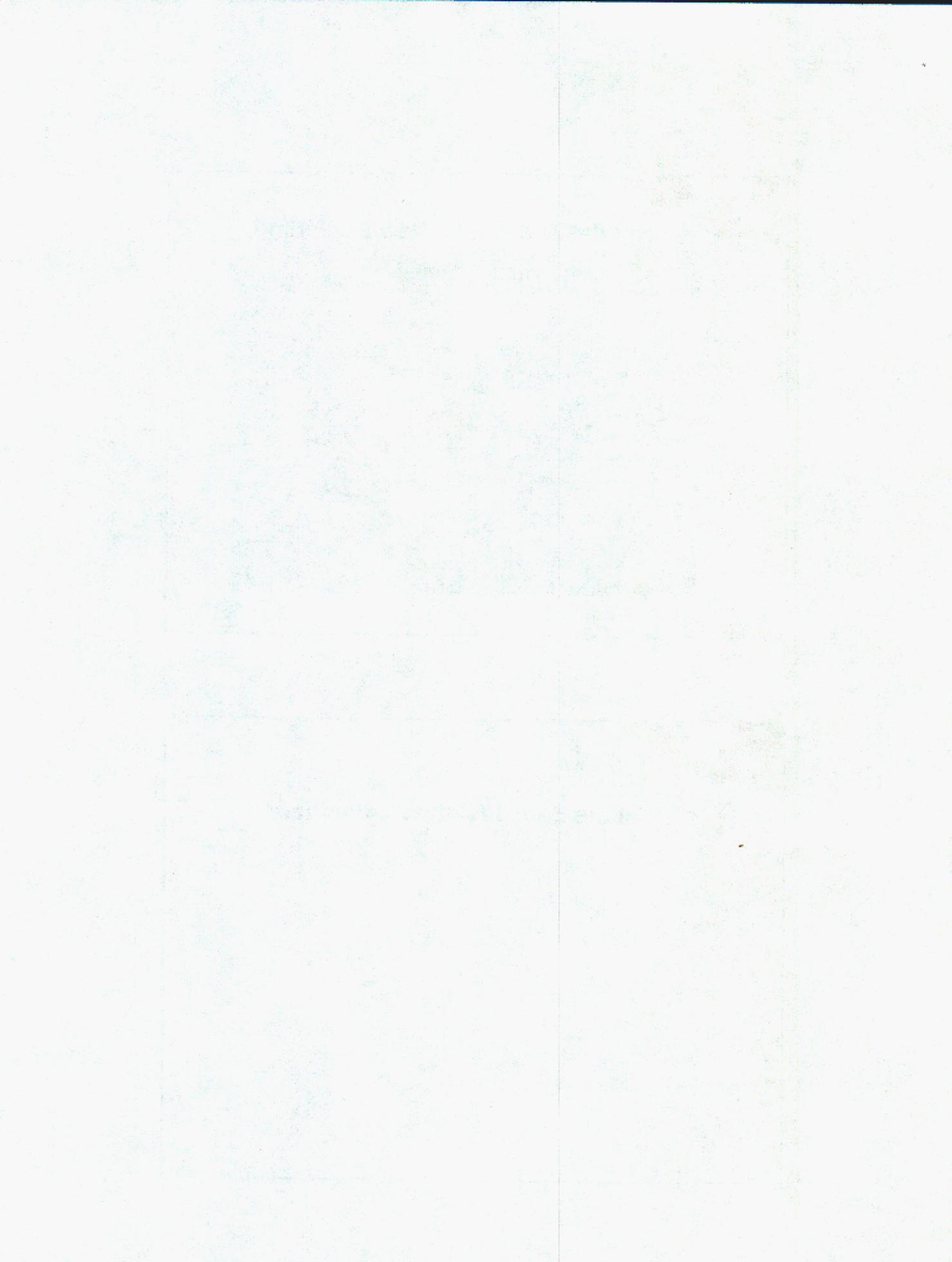
~10 TB of DDR memory @ 333 MHz

Red/Black switching - ~1/4, ~1/2, ~1/4.

8 Service and I/O cabinets on each end (256 processors for each color).

240 TB of disk storage (120 TB per color).





Red Storm Architecture

Functional hardware partitioning - service and I/O nodes, compute nodes, and RAS nodes.

Partitioned Operating System (OS) - LINUX on service and I/O nodes, LWK (Catamount) on compute nodes, stripped down LINUX on RAS nodes.

Separate RAS and system management network (Ethernet).

Router table based routing in the interconnect.

Less than 2 MW total power and cooling.

Less than 3,000 square feet of floor space.

Application Scalability

Rob Leland

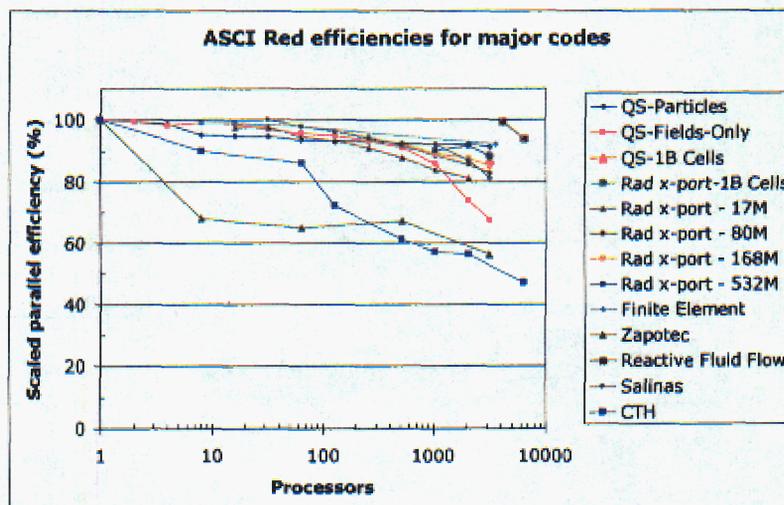
Sandia National Laboratories



Official Use Only



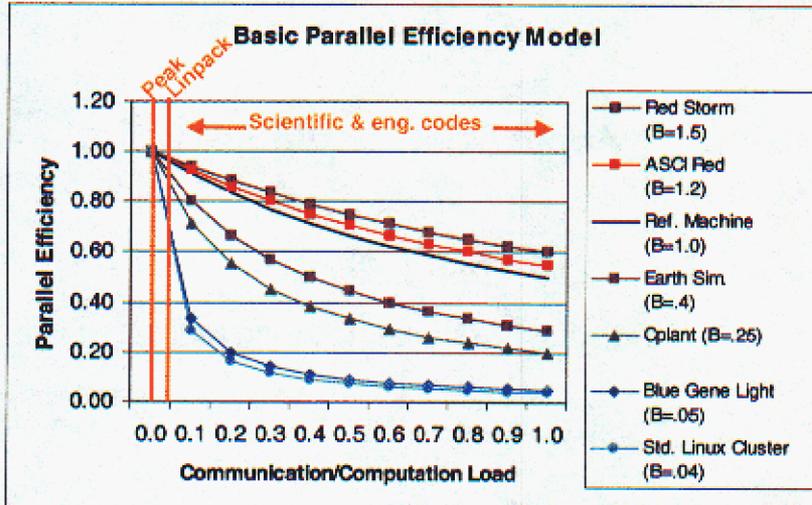
Scalable computing works



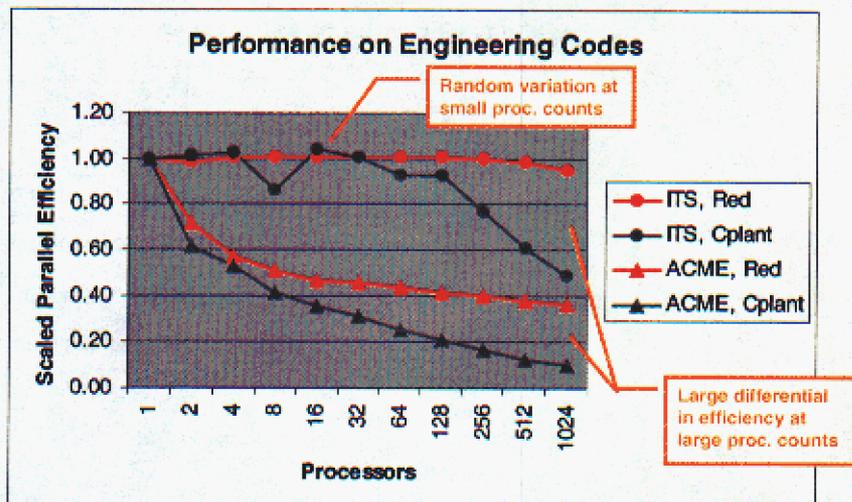
Official Use Only



Balance is critical to scalability



Scaling data for some key engineering codes



Red Storm Topology

Compute node topology:

27 X 16 X 24 (x, y, z) - Red/Black split 2688 - 4992 - 2688

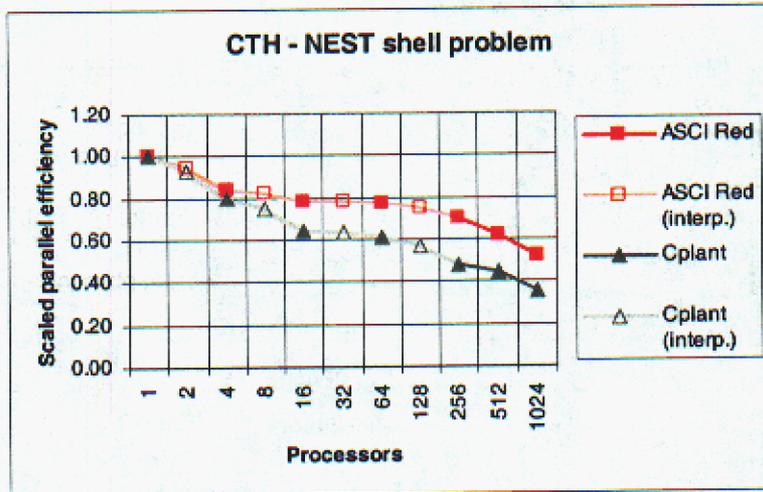
Service and I/O node topology

2 X 8 X 16 (x, y, z) on each end

Mesh is full and 2 X 16 X 16 (x, y, z) on each end

128 (256) full bandwidth links to Compute Node Mesh (384 available)

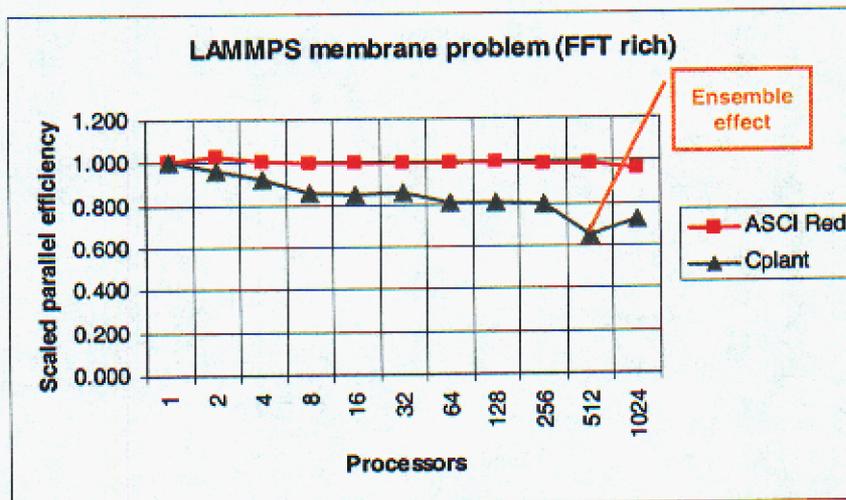
CTH scaling data



Official Use Only



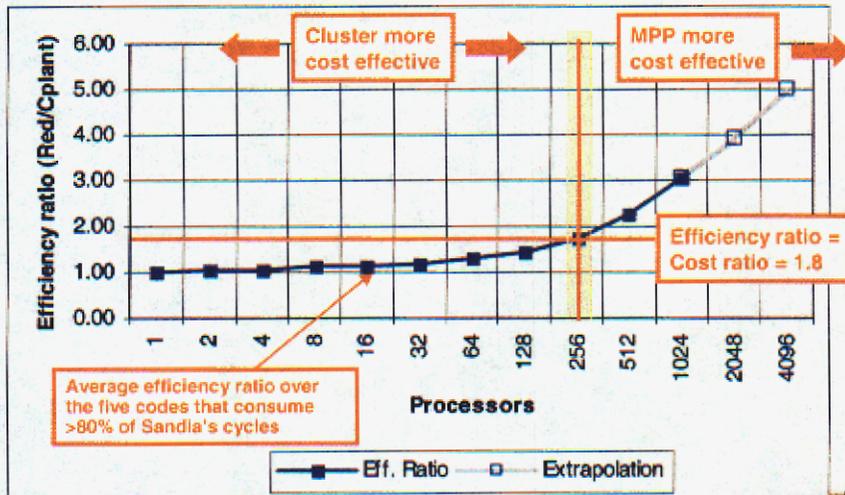
LAMMPS scaling data



Official Use Only



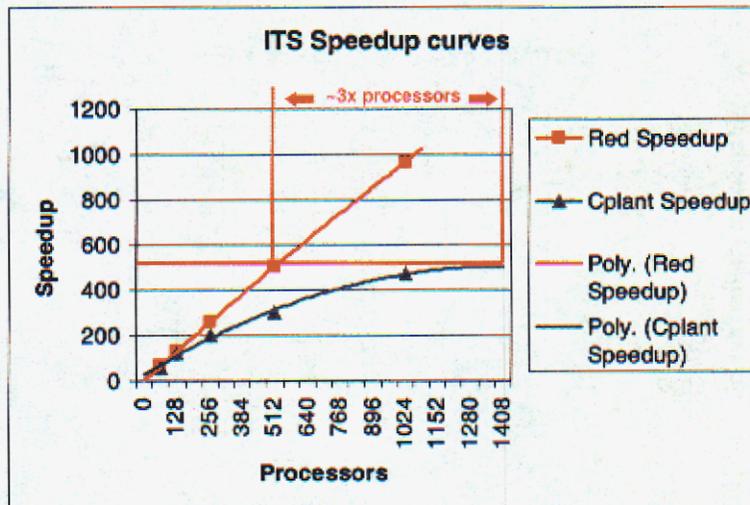
Relating scalability and cost



Official Use Only



Scalability also limits capability

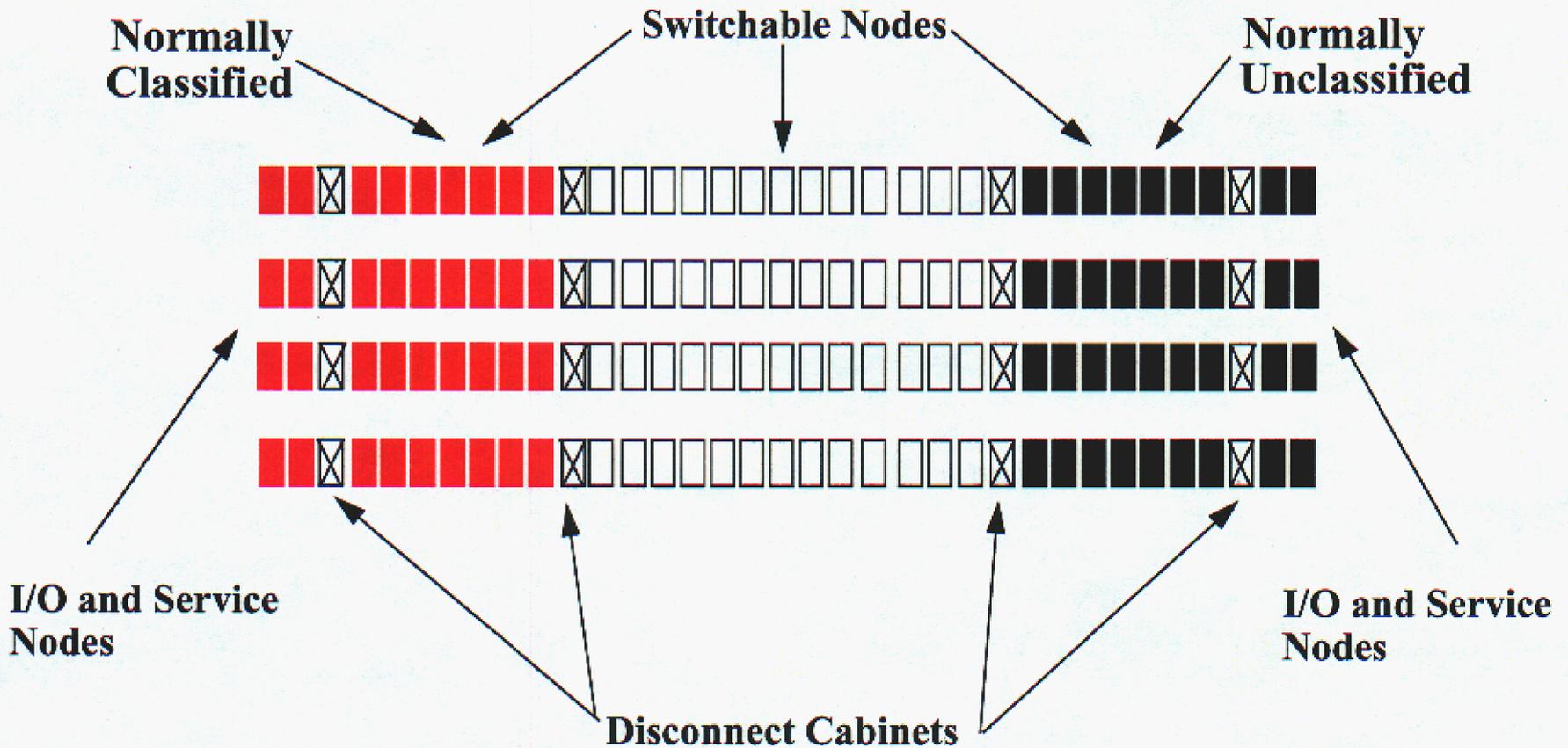


Official Use Only



Red Storm Layout

(27 X 16 X 24 mesh)



Disk storage system
not shown

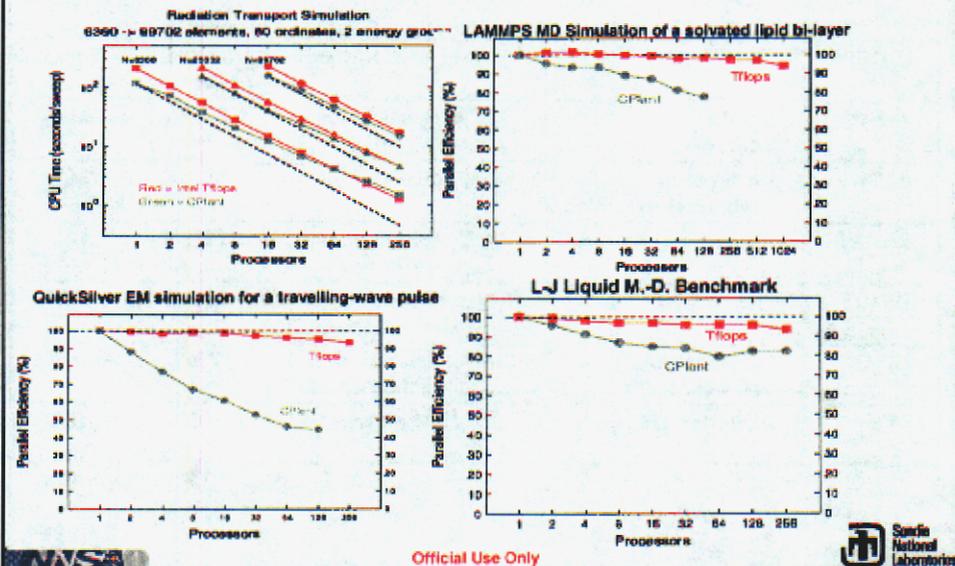
Supplementary Material



Official Use Only

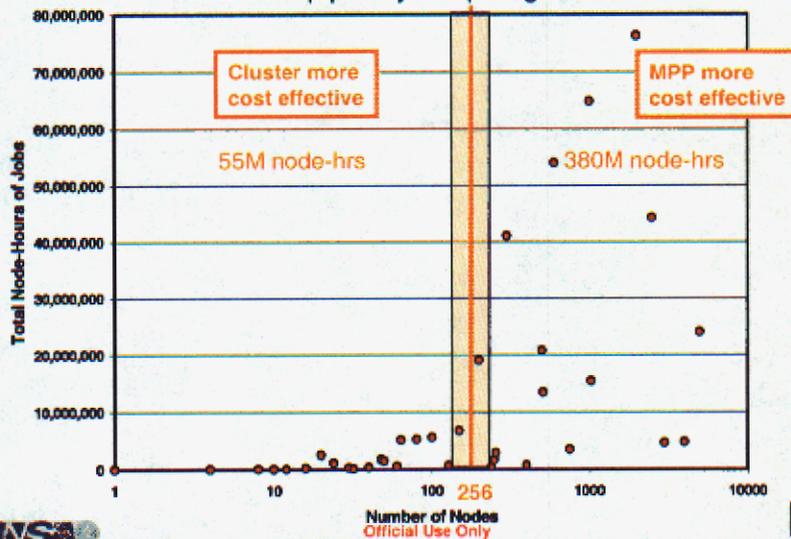


Other scaled problems



Scalability determines cost effectiveness

Sandia's top priority computing workload:



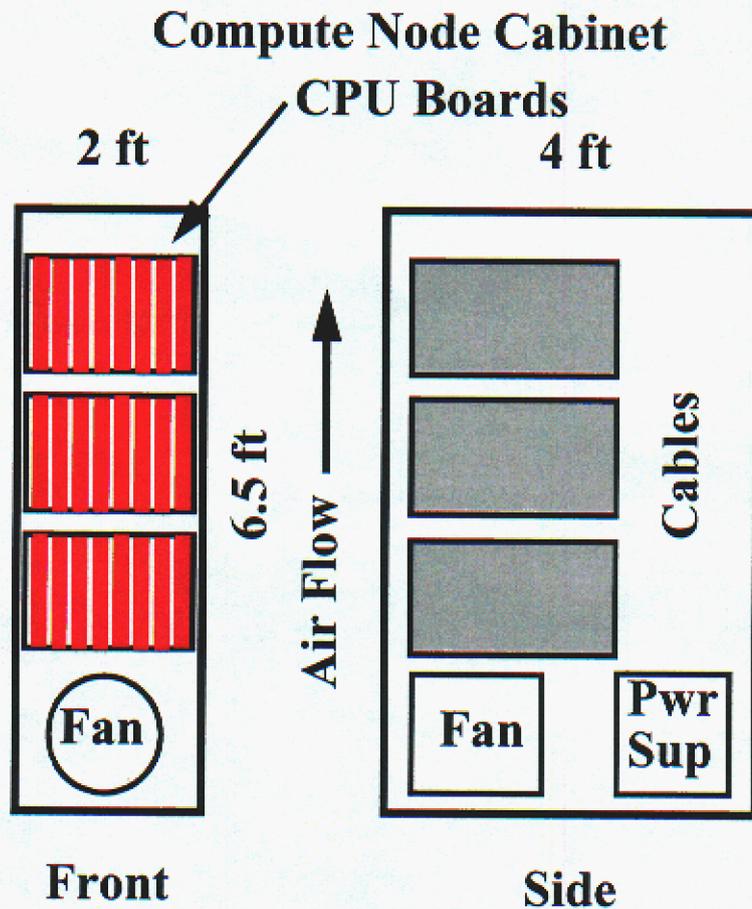
Sandia Codes

Code	Use	Numerical Method	Current Fraction	Future Fraction
Presto	Crash/ Solid dynamics	FEM, explicit time integration	34.4%	15%
Salinas	Vibration/ Structural dynamics	FEM, spectral analysis	15.8%	10%
LAMMPS	Molecular dynamics	FFT, sparse matrix methods	12.8%	10%
DSMC	Plasma dynamics	Discrete Simulation Monte Carlo	10.4%	10%
CTH	Penetration/ Solid dynamics	Control volume, explicit time integration	7.4%	10%
ITS	Radiation transport	Monte Carlo	.08%	15%

Total: 81% 70%

Official Use Only

Red Storm Cabinet Layout



Compute Node Cabinet

- 3 Card Cages per Cabinet
- 8 Boards per Card Cage
- 4 Processors per Board
- 4 NIC/Router Chips per Board
- N+1 Power Supplies
- Passive Backplane

Service and I/O Node Cabinet

- 2 Card Cages per Cabinet
- 8 Boards per Card Cage
- 2 Processors per Board
- 2 NIC/Router Chips per Board
- PCI-X for each Processor
- N+1 Power Supplies
- Passive Backplane

Balance factors

Machine (proc., network)	Node speed (Mflops/s, peak)	Link BW (Mbytes/s, peak)	B = Ratio (Bytes/Flop)
Red Storm (Opteron, custom)	4000	6000	1.5
ASCI Red (Pentium, custom)	666	800	1.2
ASCI Red (Pentium dual, custom)	666	400	.6
Cplant (Alpha, Myrinet)	932 (center) 1232 (head)	264 264	.25 (average)
Custom Linux Cluster (Opteron quad, Quadrics quad)	22,400	4000	.18
Standard Linux Cluster (Xeon dual, Myrinet)	12200	500	.04



Official Use Only



Red Storm costs

- **Capital investment** **75.5M** (67% of total; Red was 71%)
 - Hardware
 - Engineering
 - Spares with reverting ownership
 - IP to other vendors
 - Software development
- **Service (5 yrs.)** **10.4M**
 - 7x24 on site ops & maintenance, 12hrs per day prime time
 - 2 hour response time non prime time
 - 3 On site applications engineers
 - 5 Cray operations personnel
- **Operations (7 yrs.)** **26.6M**
 - Space and power
 - 2 SNL operations staff
 - 3 SNL user support staff
- **Total** **112.5M (\$FY02 @ 4% inflation)**



Official Use Only

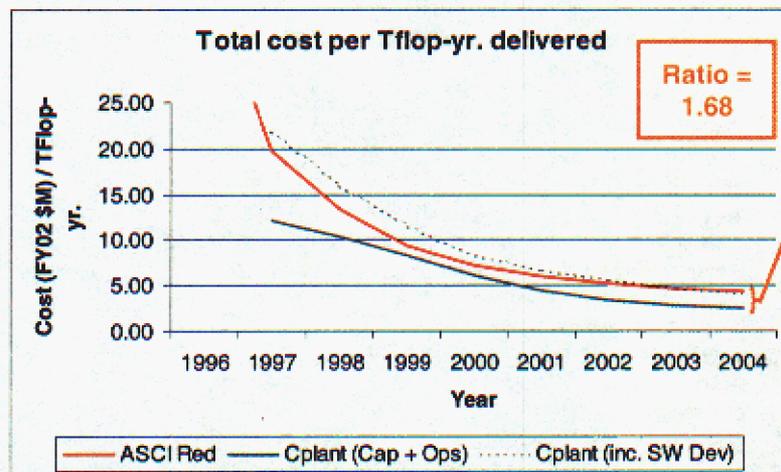


Standard Linux cluster cost projection

- **Capital investment** **24.8M**
 - Cost for 41.5Tf peak
 - Based on 7.3Tf for \$4.2M
 - Spares at same rate
- **Service (5 yrs.)** **10.4M**
 - Same standard
- **Operations (7 yrs.)** **26.6M**
 - Same standard
- **Total** **61.8M (\$FY02 @ 4%)**

Red Storm/SLC cost ratio = $112.5/61.8 = 1.82$

Historical cost data



Red Storm Performance

RAS Workstations

Separate and redundant RAS workstations for Red and Black ends of machine.

System administration and monitoring interface.

Error logging and monitoring for major system components including processors, memory, NIC/Router, power supplies, fans, disk controllers, and disks.

RAS Network - Dedicated Ethernet network for connecting RAS nodes to RAS workstations.

RAS Nodes

One for each compute board

One for each cabinet



SC 2001
November 2001

Red Storm Visualization

Philip D. Heermann
Sandia National Labs



Visualization: Data to Display

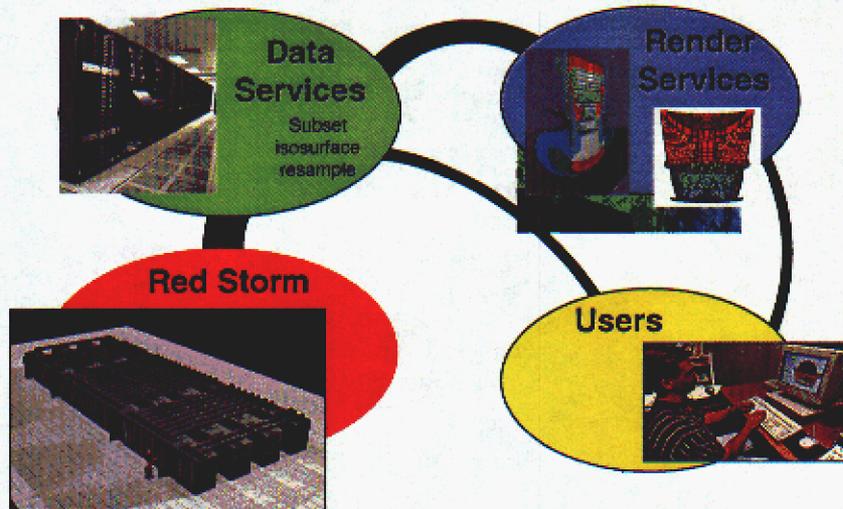
SC 2001
November 2001

Data Sources: Simulations, Archives, Experiments				Users Services: Navigation Rendering Control Advanced User Interface Collaborative Control Display Control	
Data Services:	Permutation M * N	Filtering	1D/2D Subsetting		Data Algebra x,y,z * mag/...
	Format/Representation Conversion	Data Reduction	Data Serving		
Information Services:	Feature Detection and Extraction		Data Fusion & Comparison		
	Visual Representations Generation (eg. isosurfaces)	Volume Visualization Preparation (eg. opacity assignment, resampling ...)			
Visualization Services:	Surface rendering	Volume rendering	Runtime services		
	Multi-Visualization Technique Combine	Time Sequence Generation			
Display Modalities:	Desktop Display	Theater Display	Powerwalls	Immersive Stereoscopic	



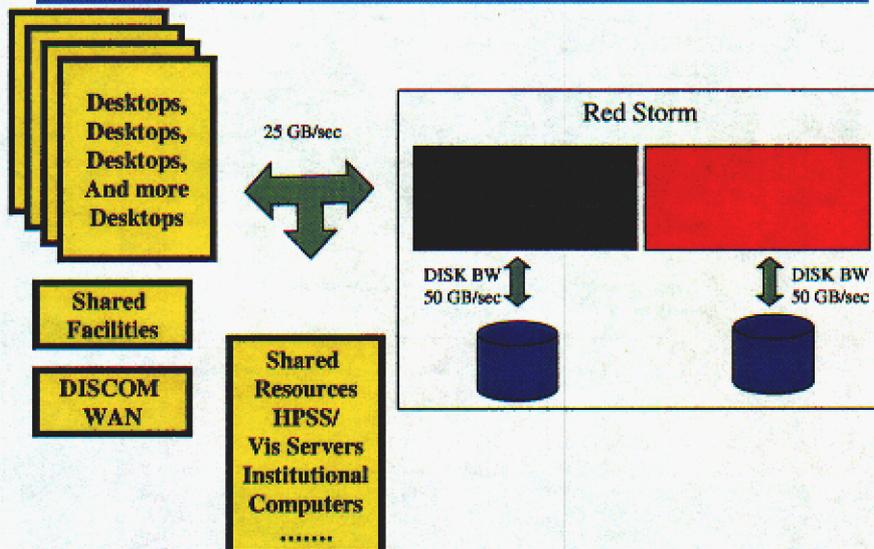
Visualization Strategy

NSF 2001
November 2001



Red Storm System

NSF 2001
November 2001



Red Storm System Software

Operating Systems

LINUX on service and I/O nodes

LWK (Catamount) on compute nodes

LINUX on RAS nodes

File Systems

Parallel File System - Lustre (PVFS)

Unix File System - Lustre (NFS)

Run-Time System

Logarithmic loader

Node allocator

Batch system - PBS

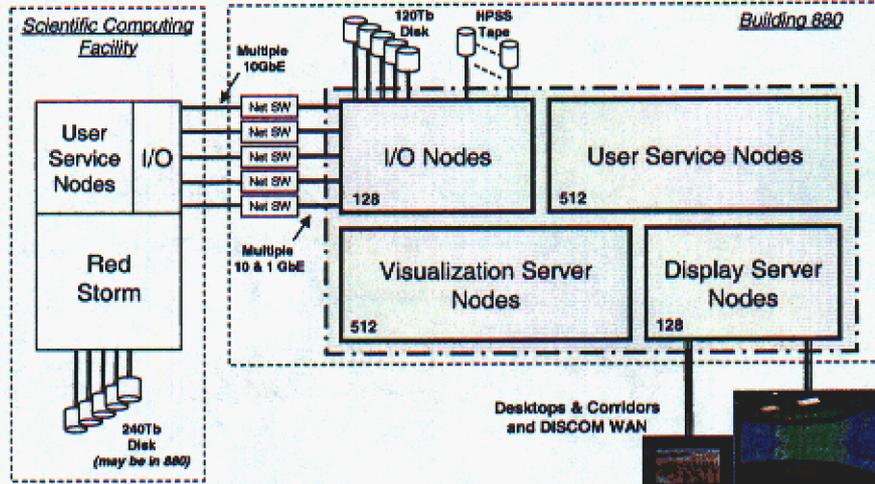
Libraries - MPI, I/O, Math

Single System View



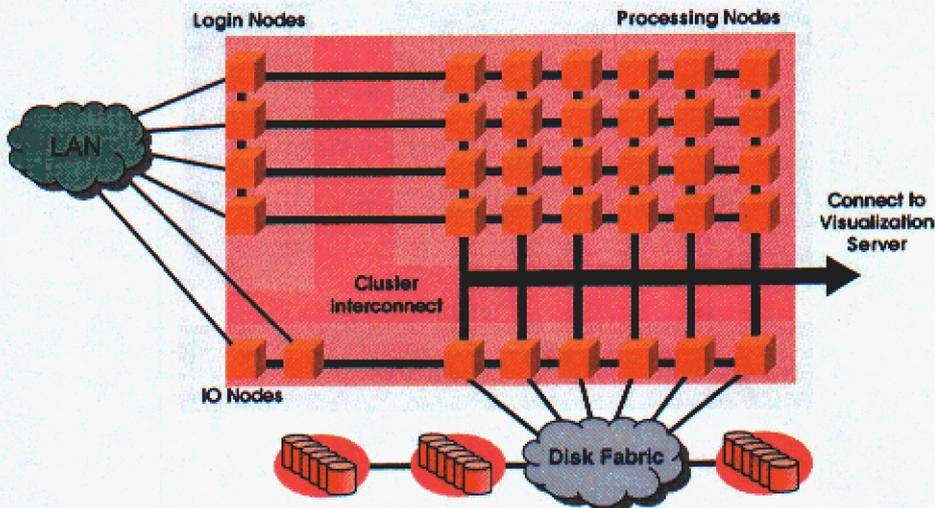
Red Storm Data Services Plan

SI 2007
November 2007



Cluster data server

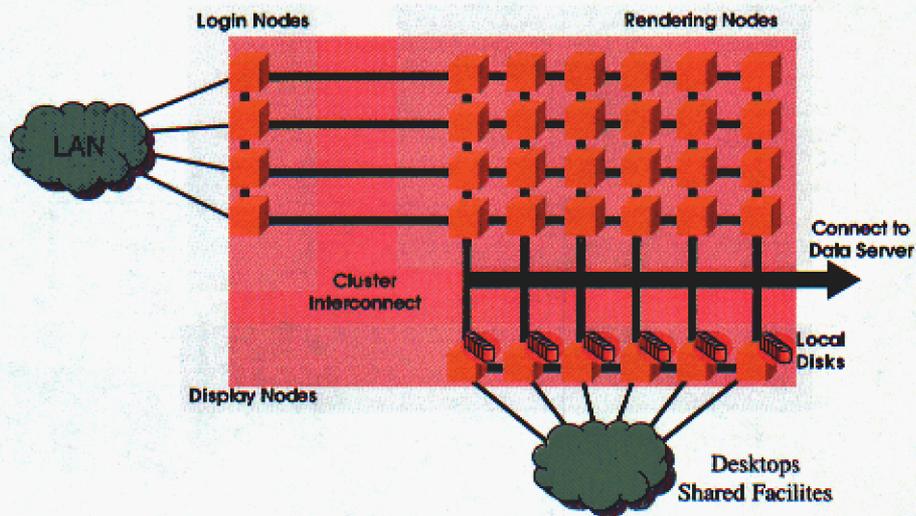
SI 2007
November 2007





Visualization cluster

NSF
November 2002



Summary

NSF
November 2002

- Leverage Specialized Clusters
 - Post processing (High performance I/O)
 - Rendering (Leverage Commodity Graphics Cards)
- Integrate Tape Archive with Data Server
- Provide Services
 - Desktop users
 - Shared Facilities

Red Storm System Software

Programming Model

Message Passing

Support for Heterogeneous Applications

Tools

ANSI Standard Compilers - Fortran, C, C++

Debugger - TotalView

Performance Monitor

System Management and Administration

Accounting

RAS GUI Interface

Single System View



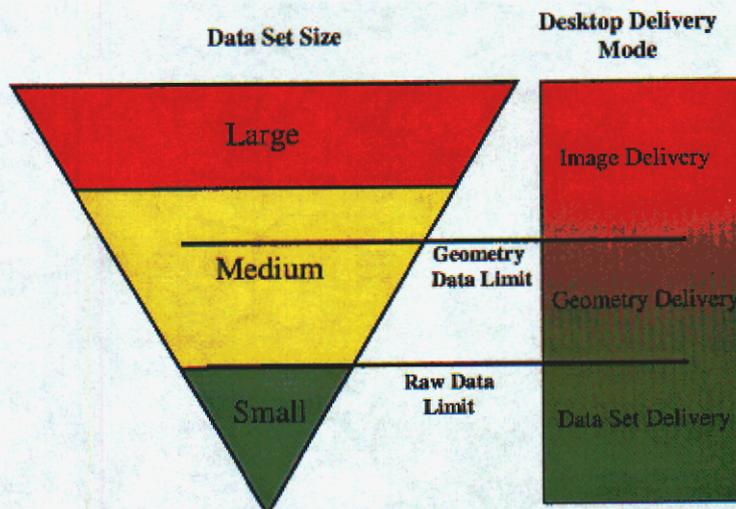
01/2003
November 2003

End



Simplified diagram

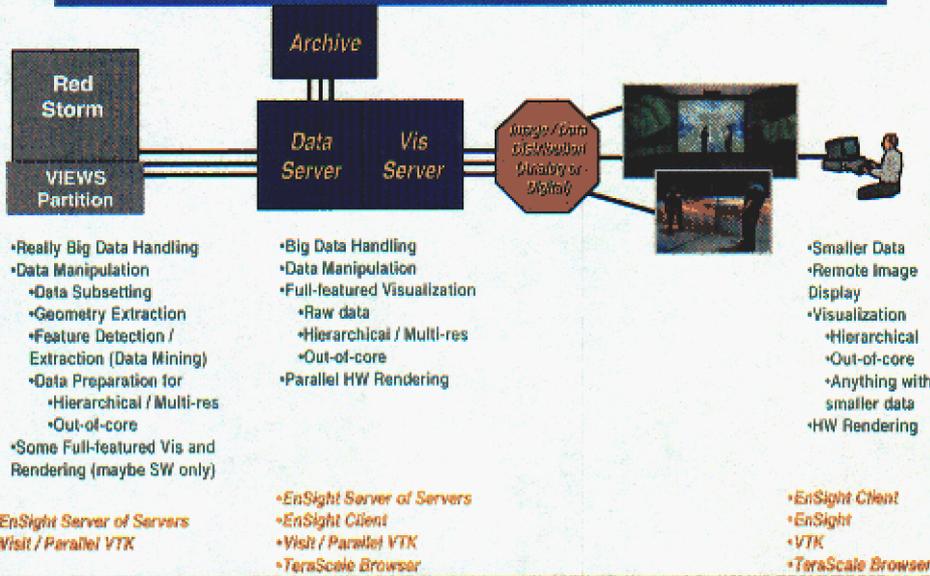
01/2003
November 2003





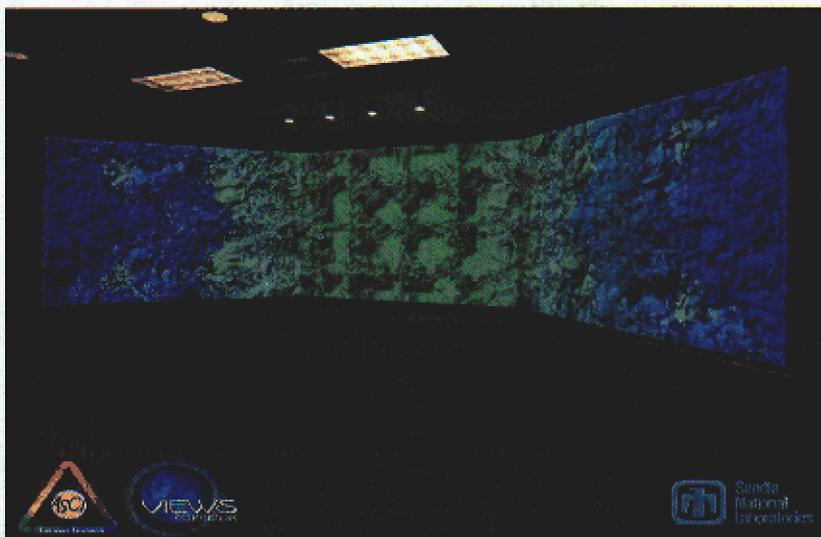
The Red Storm Big Picture

ST 2001
November 2001



Large Scale Visual Acuity Display

ST 2001
November 2001



Red Storm Performance - Overall System

Based on application code testing on production AMD Opteron processors we are now expecting that **Red Storm will deliver around 10 X performance improvement over ASCI **Red** on Sandia's suite of application codes.**

Expected MP-Linpack performance - ~30 TF.

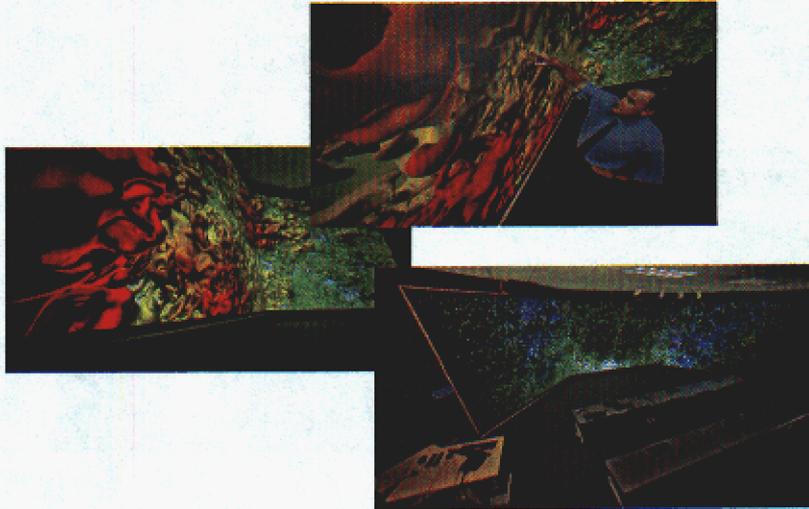
Aggregate system memory bandwidth - ~55 TB/s

Aggregate sustained interconnect bandwidth > 100 TB/s



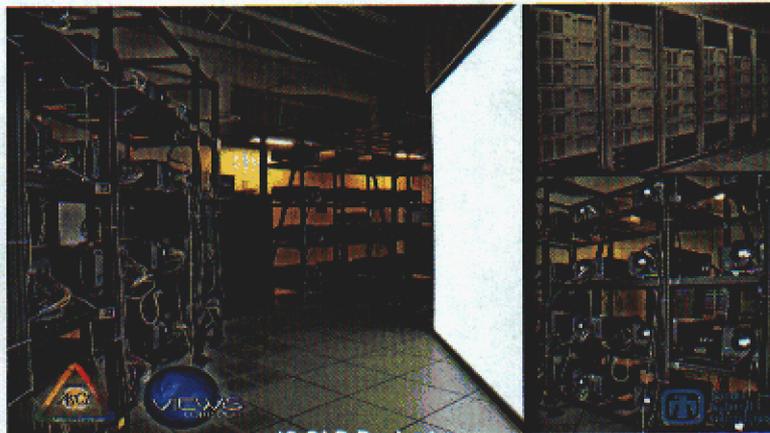
Large Scale Visual Acuity Display

AT 2001
November 2001



Rendering and Display System

AT 2001
November 2001





Scalable Rendering & Display

SC 2001
November 2001

- 62 MegaPixel Resolution Display
- 1 Billion Poly/S (to single display)
- 60 Million Poly/S (to 62MP display)

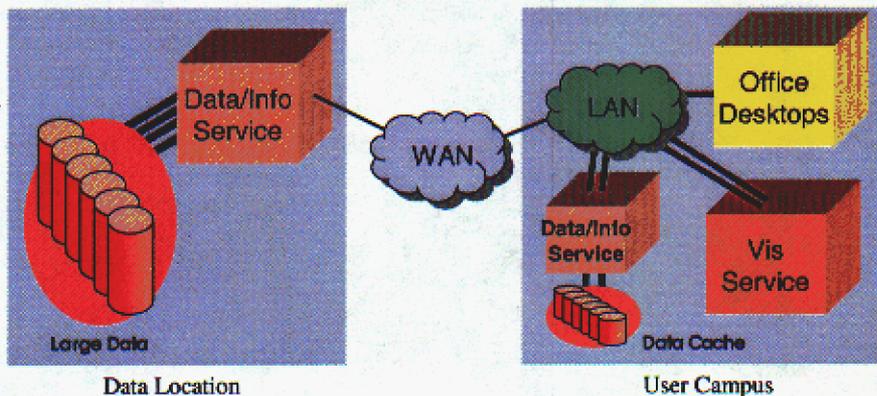
Major Progress toward replacing SMP
Visualization Servers



Desktop system

SC 2001
November 2001

Visualization Desktop Delivery System



Red Storm Performance - Processors and Memory

Processors

AMD Opteron (Sledgehammer)

2.0 GHz

64 Bit extension to IA32 instruction set

64 KB L1 instruction and data caches on chip

1 MB L2 shared (Data and Instruction) cache on chip

Integrated dual DDR memory controllers @ 333 MHz

Integrated 3 Hyper Transport Interfaces @ 3.2 GB/s each direction

Node memory system

Page miss latency to local processor memory is ~80 nano-seconds.

Peak bandwidth of ~5.3 GB/s for each processor.



Red Storm Data Network

To

DISCOM

Luis Martinez, 9336
Sandia National Laboratories
November 2003



Overview

- Data I/O Requirements
- Design Goals
- Red Storm Data Network
- Connectivity To SNLA Production Data Network
- Internet Connectivity
- Connectivity to DISCOM
- Questions





Service and I/O Requirements

- **Service and I/O cabinets** (16 total- 8 red, 8 black)
 - 16 (2 AMD) Processor modules per cabinet
 - 128 total processor modules (64 red, 64 black)
- **Network Bandwidth**
 - 25 GB/sec (200 Gigabits/sec)
- **Login nodes**
 - 64 ea. 1.0 Gigabit Ethernet NICS per service partition



Design Goals

- **Keep design to a single switch per Service Partition**
 - Cost
 - Throughput limited by size of switching fabric not size of available uplinks
 - Single, simpler configuration
- **Provide multiple 10 gigabit access to Red Storm**



Red Storm Performance - Interconnect and I/O

Interconnect performance

Latency <2 μ s (neighbor) <5 μ s (full machine)

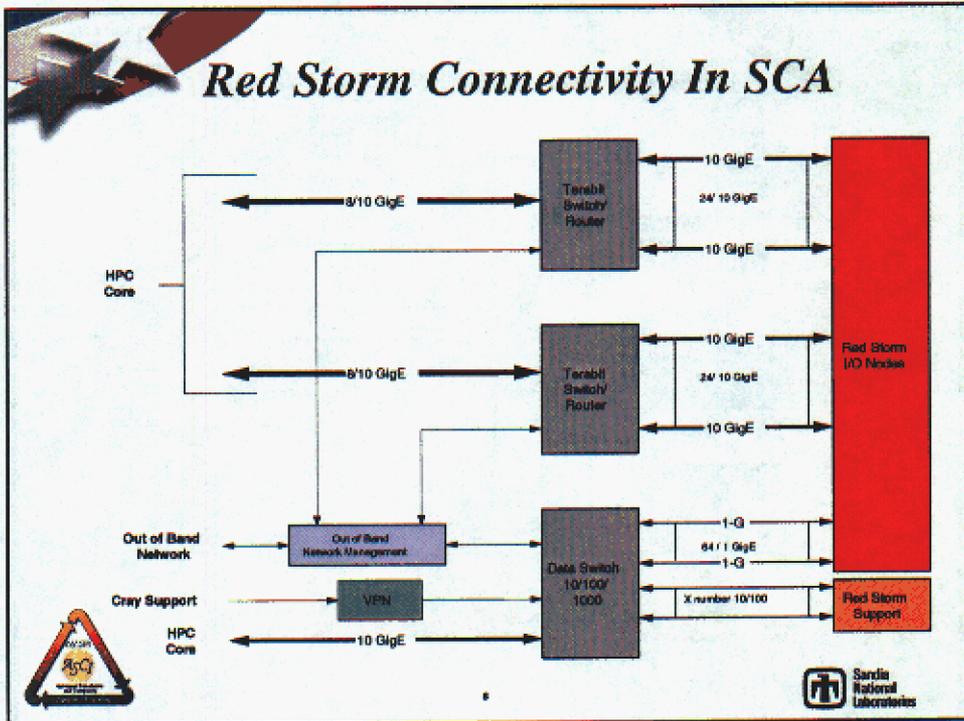
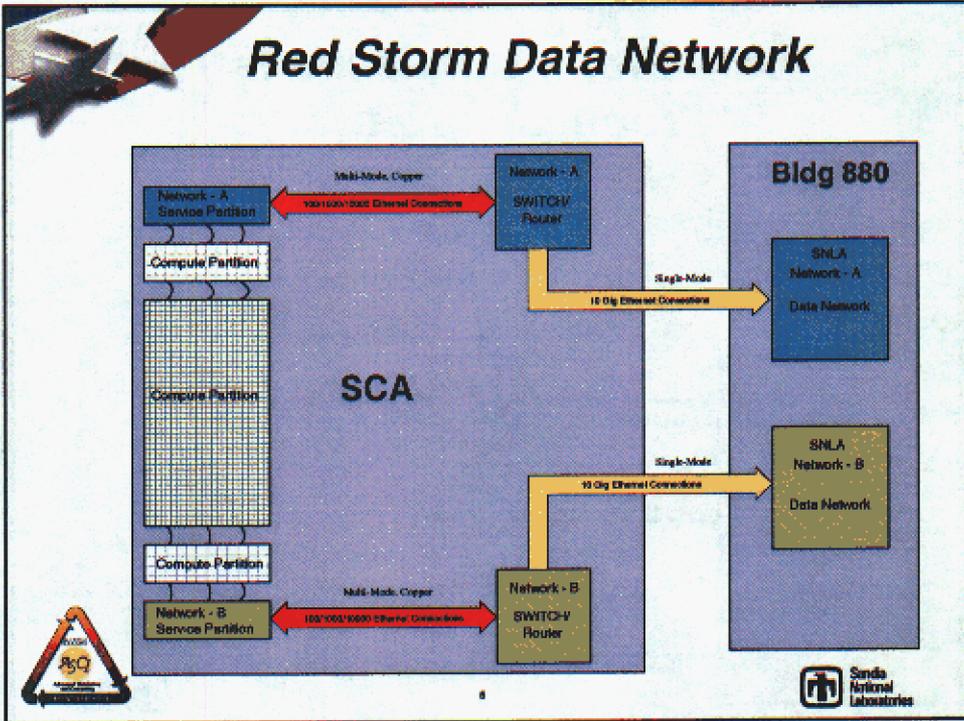
Peak Link bandwidth ~3.84 GB/s each direction

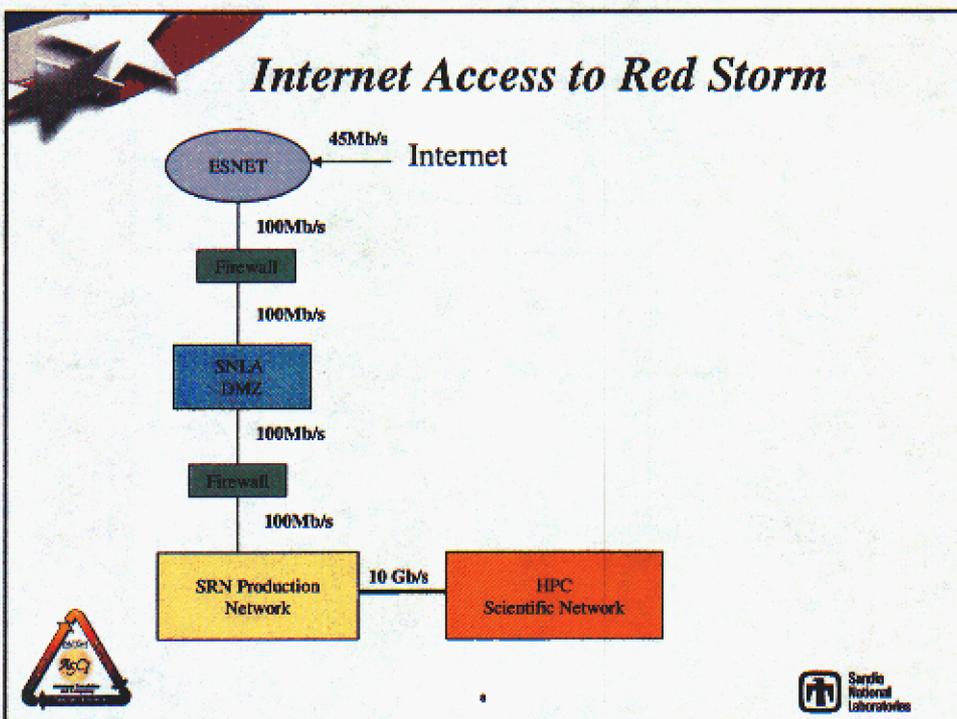
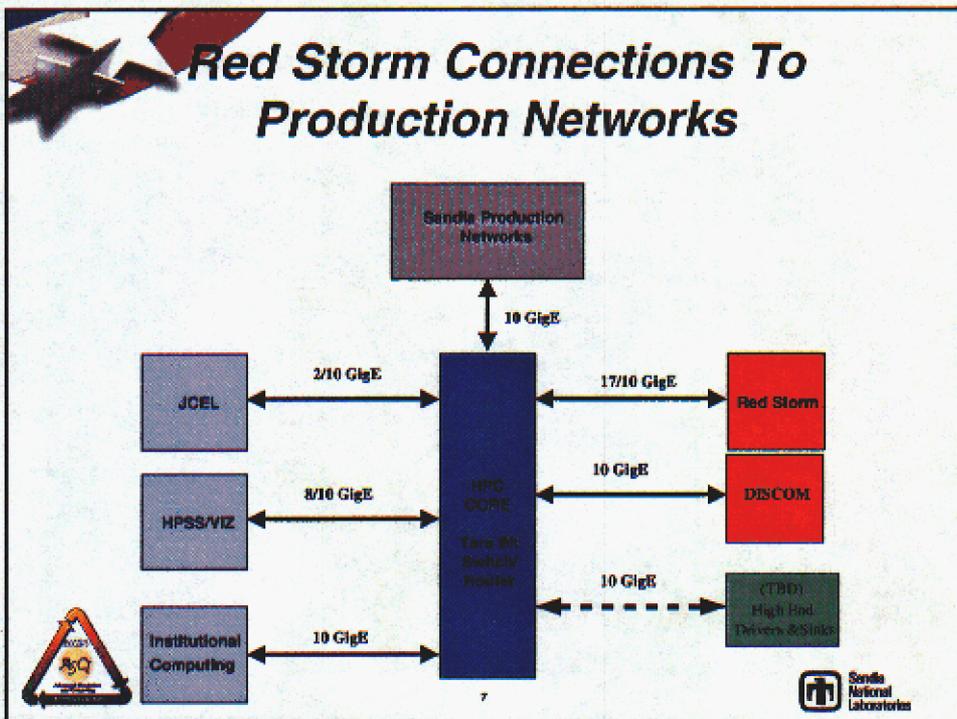
Bi-section bandwidth ~2.95 TB/s Y-Z, ~4.98 TB/s X-Z, ~6.64 TB/s X-Y

I/O system performance

Sustained file system bandwidth of 50 GB/s for each color.

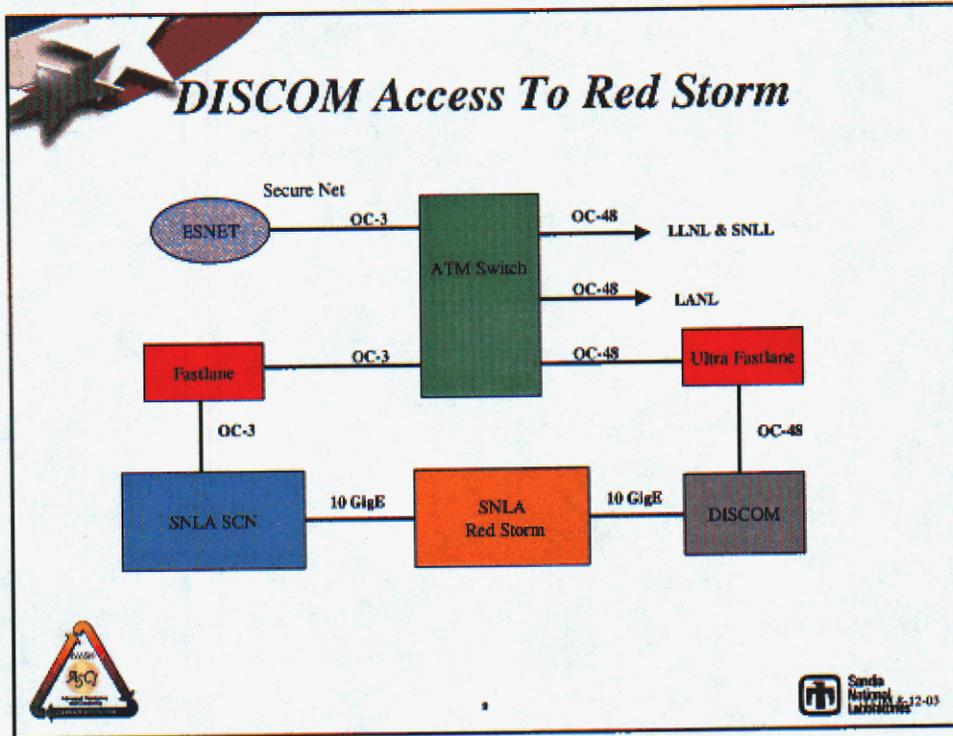
Sustained external network bandwidth of 25 GB/s for each color.





Comparison of **ASCI Red** and **Red Storm**

	ASCI Red	Red Storm
Full System Operational Time Frame	June 1997 (Processor and Memory Upgrade in 1999)	Q3 of 2004
Theoretical Peak (TF)	3.15	41.47
MP-Linpack Performance (TF)	2.379	>14 (est)
Architecture	Distributed Memory MIMD	Distributed Memory MIMD
Number of Compute Node Processors	9,460	10,368
Processor L1 Caches L2 Cache	Intel P II @ 333 MHz 16 KB Data & Instruction 512 KB Shared	AMD Opteron @ 2.0 GHz 64 KB Data & Instruction 1.0 MB Shared
Total Memory	1.2 TB	10.4 TB (up to 80 TB)
System Memory B/W	2.5 TB/s	55 TB/s
Disk Storage	12.5 TB	240 TB
Parallel File System B/W	1.0 GB/s each color	50.0 GB/s each color
External Network B/W	0.2 GB/s each color	25 GB/s each color



Questions ?

Logos for the Sandia National Laboratories and the 30th anniversary of the Laboratory are present at the bottom of the slide.

Comparison of **ASCI Red** and **Red Storm**

	ASCI Red	Red Storm
Interconnect Topology	3-D Mesh (x, y, z) 38 X 32 X 2	3-D Mesh (x, y, z) 27 X 16 X 24 2 X 16 X 16 each end
Interconnect Performance MPI Latency Link B/W Min Bi-section B/W	15 μ s 2 hop, 20 μ s max 800 MB/s 51.2 GB/s	2.0 μ s 2 hop, 5 μ s max 7.68 GB/s 2.95 TB/s
Full System RAS RAS Network RAS Processors	10 Mbit Ethernet 1 for each 32 CPUs	100 Mbit Ethernet 1 for each 4 CPUs
Operating System Compute Nodes Service and I/O Nodes RAS Nodes	Cougar TOS (OSF1) VX-Works	Catamount (Cougar) LINUX LINUX
Red Black Switching	2260 - 4940 - 2260	2688 - 4992 - 2688
System Foot Print	~2500 sq ft	~ 3000 sq ft
Power Requirement	850 KW	~2.0 MW



Sandia National Laboratories

High Performance Computing Customer Support

**Barbara Jennings
Scientific Computing Systems**

Slide 1



Designing High Performance Computing Customer Support

- **HPC is an Unconventional Environment**
- **Developing a New Culture**
- **The Knowledge-centered Culture**
- **Knowledge creation is dependent on sharing knowledge**
- **HPC Customers require customized level of support**
- **Our plan**

Slide 2





The High Performance Computing environment is unconventional.

**In the evolving world of HPC our customers are the
pioneers developing:**

- **New platforms**
- **New operating systems**
- **New code**
- **New scientific capabilities**
- **New developments daily**

Slide 3



How to exist in a changing environment.

Develop a knowledge centered culture.

- **Create a Socio-Technical environment**
 - Identifying the experts
 - Enable Problem Solving
 - Provide Tools
- **Capture Knowledge**
 - Capturing technology changes and user's experience
 - Capturing the solutions - Intrinsically
- **Disseminate Knowledge**
 - Web based shared repository
 - Consistent approach to access information among platform providers
 - On-line training

Slide 4



Red Storm Project

Hardware Status

Cabinet Design is complete - Prototypes have been built
Red/Black Switch design is complete - On display in ASCI Booth
NIC/Router (Seastar) Chip Final Net List has been released
Prototype boards with FPGA for NIC/Router are being tested

System Software Status

System Software is being developed and tested on IA32 and Opteron clusters

Contracts for development work are in place with Etnus (TotalView) and CFS (Lustre) for software development.

System is scheduled for completion in Q3 of 2004.



What is a knowledge-centered culture?

An environment of trust where everyone's contribution is valued and consisting of:

- **Learning**
 - Learning increases performance
- **Mentoring**
 - Providing a path for the new comers
- **Collaboration**
 - Collective brain power enables users to maximizing the resource potential
- **Sharing Ideas**
 - Value of knowledge increases with its accessibility and the frequency that is is shared

Slide 5



What is knowledge sharing?

Knowledge Creation is of limited value if it is not shared.

- **Knowledge Creation**
 - Learning new things
- **Knowledge Dissemination**
 - Skills transfer
 - Teaching new things to others
- **Knowledge Storage and Retrieval**
 - Must be intrinsic to everyday operations
 - Easily accessible
- **Knowledge Application**
 - Getting the job done

Slide 6





Unconventional Users require unconventional support.

- **Environmental Complexity**
 - Ever changing environment
- **Customer Expectations**
 - Require an expert at the other end & staff that work in the area and are familiar with the environment to provide support
- **Programmatic Need**
 - Need to use platforms and services at SNL as well as the other NNSA Labs
 - Need may be an immediate concern to national security
- **Remote Customer Base**
 - Support must be provided for all ASCI users regardless of location

*SNL - Sandia National Laboratories

*NNSA - National Nuclear Security Agency

*ASCI - Accelerated Strategic Computing Alliance

Slide 7



How are we going to do it?

- **Determine customer needs**
 - Ask the customer
 - Ask the folks providing support now
- **Look at the success stories**
 - Lawrence Livermore National Laboratory sets a high bar to meet
 - Provide experts on the hot line who rotate between support and lab projects
- **Reward knowledge sharing**
 - Encourage individual user information dissemination
 - Facilitate user cooperation
- **Deliverables**
 - Support via email, telephone, browser
 - Platform specific information: Sample codes to follow
 - Browser accessible Knowledge base
 - Time independent

Slide 8





Five areas chosen for concentrated support.

- **Technical support for use of platforms and applications**
 - Telephone
 - Email
 - Shared Knowledge Base Repository
- **Training**
 - Driver's License
 - Examples of Basic Platform Use
 - On-line
- **Web Accessible Account Requests**
 - Manual process exists but not directly accessible to the user
- **Custom Life Time Support**
 - For developers
 - Includes code tuning and debugging
- **Web Pages Common to Other Resource Sites**
 - Provides a familiar location for resource information

Slide 9



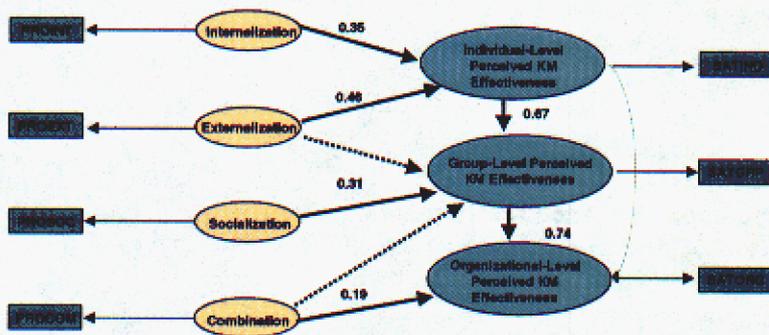
Leverage on what we have and what we know.

- **Support departments are developed over time**
 - Identify the experts
 - Define a path for learning
- **Following the LLNL lead with "experts" on the phone and by email support**
 - Sys Admins have been providing the support - we would like to continue this support
 - Rotating staff between support and regular work allows them to gain knowledge while keeping current in their area
- **Implement ARS tracking system**
 - Using an existing internal system to cut down on learning curve
- **Provide searchable information in the form of the Knowledge Base**
 - Also using an existent internal system - making this available to the users as well
- **Work closely with internal and external support centers**
 - Sharing knowledge with other resource providers
- **Watch Metrics and Stay in touch with the customer**

Slide 10



Emergent Structural Model from Sabherwal and Becerra-Fernandez

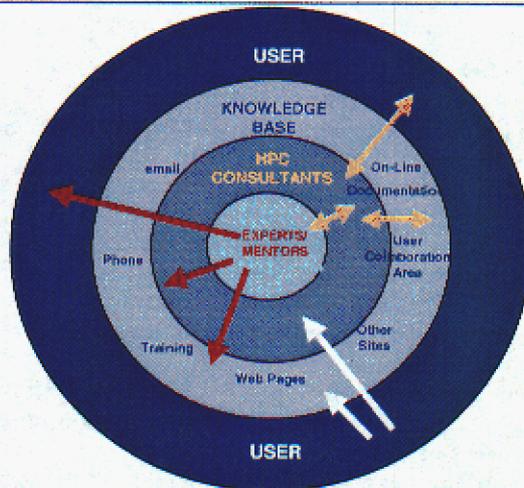


Rajiv Sabherwal and Irma Becerra-Fernandez, (2005). An Empirical Study of the Effect of Knowledge Management Process at Individual, Group, and Organizational Levels. In *Decision Sciences* Volume 34 Number 2 Spring 2005. Printed in the USA, 225-250

Slide 11



Knowledge Culture Support for Sandia National Labs



Slide 12



MTL0165825W

SANDIA NATIONAL