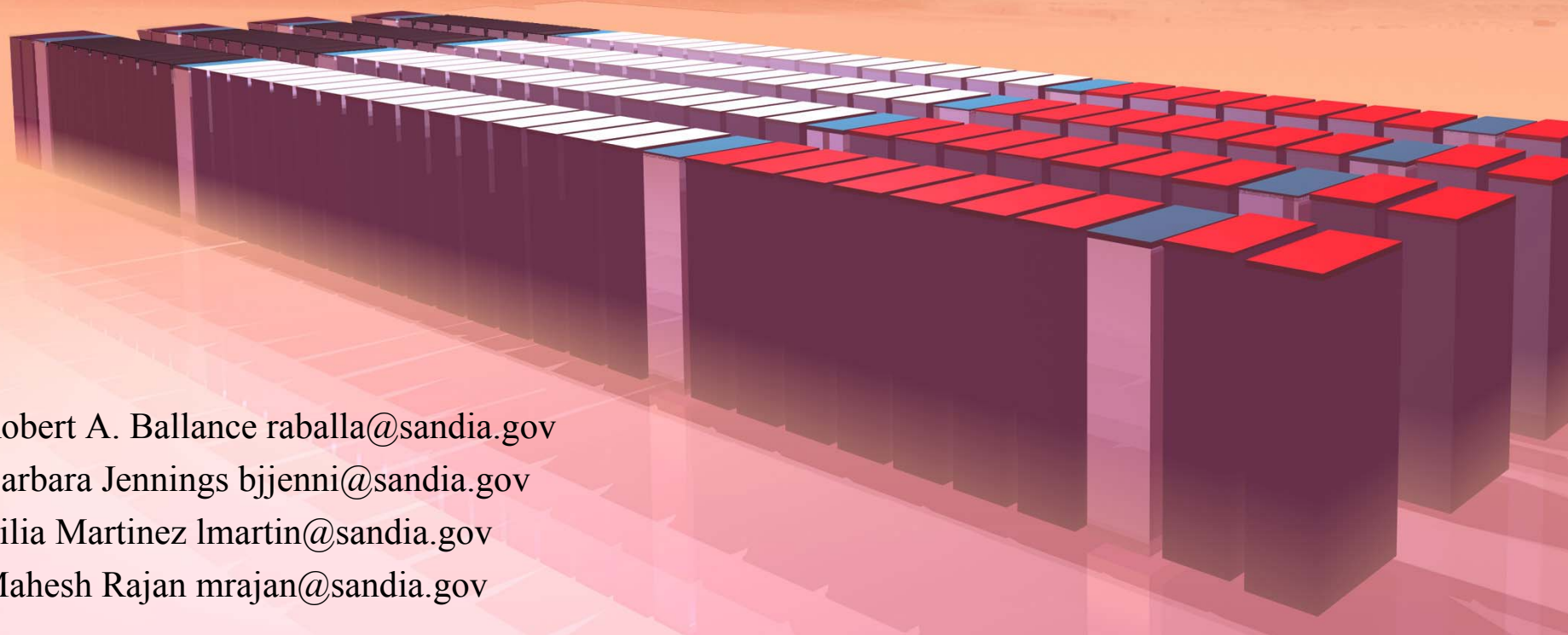
ASCI Alliance Center Site Visit

Sandia National Laboratories
ASCI Computing Systems

September 22-24, 2004

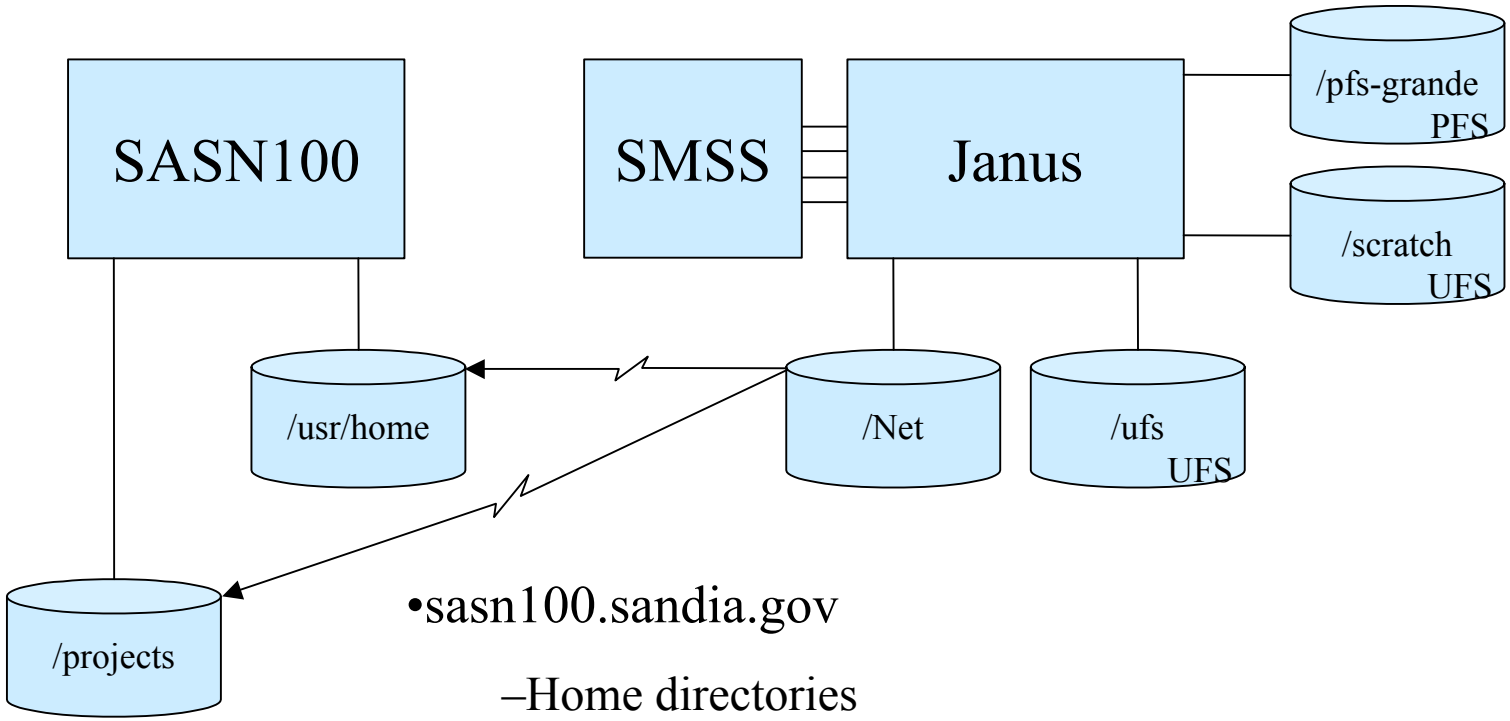


Robert A. Ballance raballa@sandia.gov
Barbara Jennings bjjenni@sandia.gov
Lilia Martinez lmartin@sandia.gov
Mahesh Rajan mrajan@sandia.gov

Unclassified Systems Available for Alliance Partners

- ASCI Red, a.k.a. Janus
- Red Storm, limited availability 2Q05 (calendar year)

Janus System Configuration



- sasn100.sandia.gov
  - –Home directories
  - –Cross compilers
- SMSS (High Performance Storage System, HPSS)
- file transfer between HPSS -> 28 MB

JANUS Operational Configuration

Center

Red/classified

Compute nodes	1166
Service nodes	26
Disk I/O	37
System nodes	1
Network nodes	6

Center switches back every two weeks

Compute Nodes: 2176

Black/unclassified

Compute nodes	1168
Service nodes	26
Disk I/O	36
System nodes	1
Network nodes	6

Can also run virtual nodes doubling the node capacity in batch mode.

The switching schedule can be found at:
https://www.prod.sandia.gov/cgi-bin/cals-SCS/webevent.cgi?cmd=opencal&cal=cal4&

Network

Tom Pratt

Current Lab Networks

- Expected Performance
  - LANL 50 Mbit/sec
  - SNL 32 Mbits/sec
- Dependency
  - Data Movement Apps
  - Latency Tolerance 30-80ms
  - Network Congestion

Proposed Red Storm Network Access

- Ssh, scp to redstorm.sandia.gov
- Likely to require 2-factor authentication
  - Crypto card?
  - Change from Janus, due to new emphasis on security stance
- Load-balancing login node server will forward session to appropriate login node

Future releases currently in process

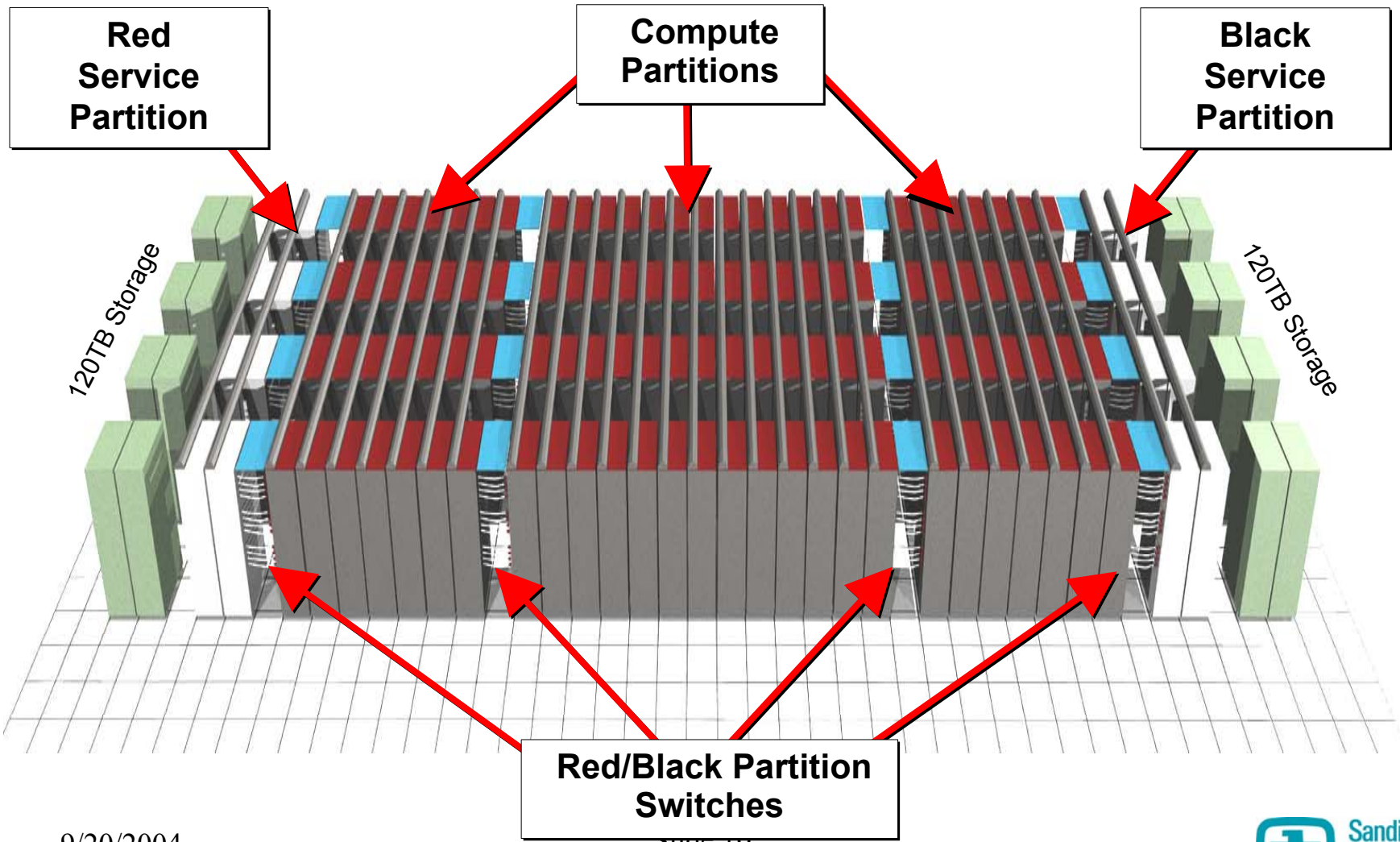
- Newest Supercomputer Red Storm
- Newest Customer support environment CLIK

Project Overview

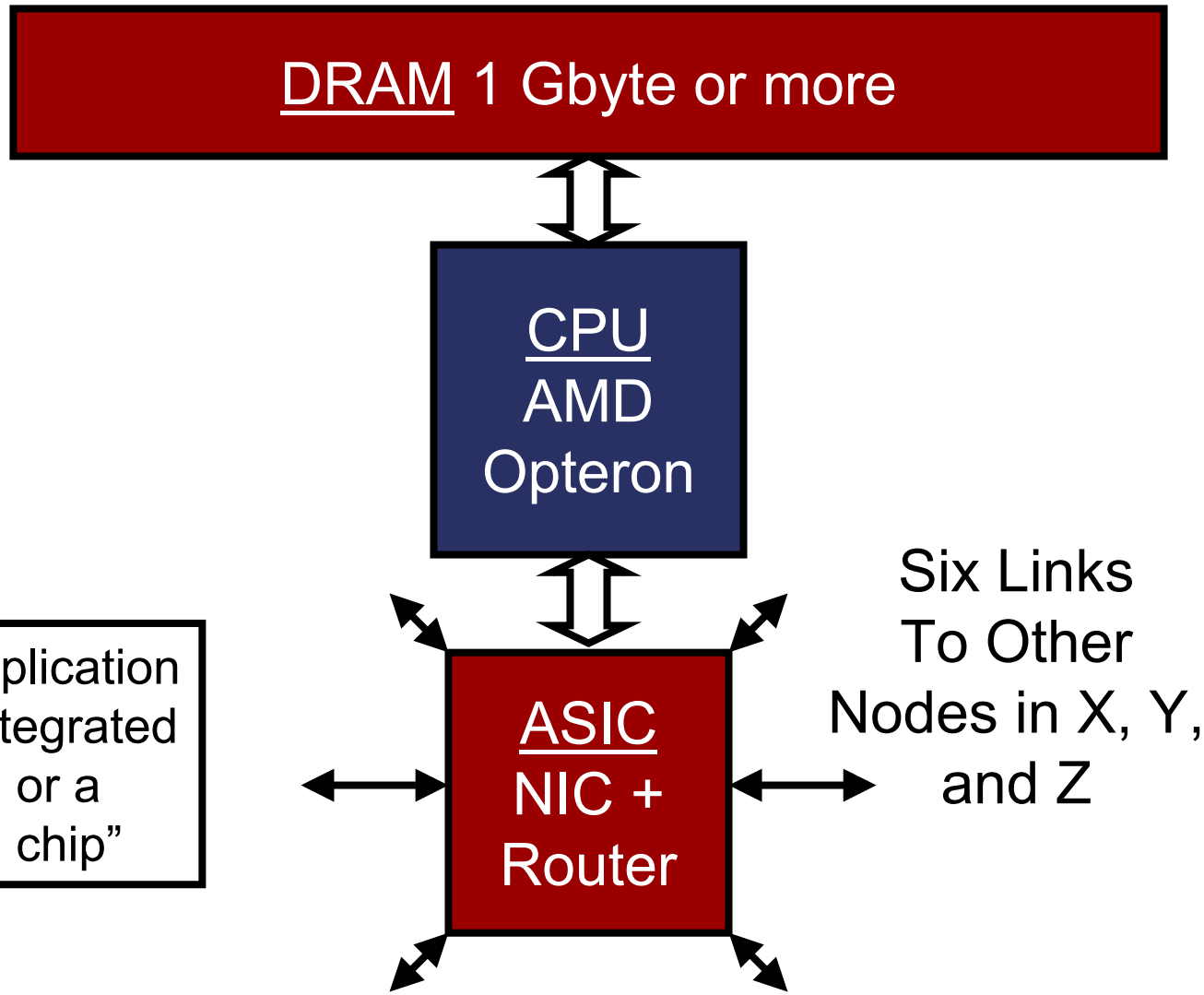
- Red Storm is a nominally 40 TFlops supercomputer that is part of the Advanced Simulation and Computation (ASC) program
- Red Storm is being manufactured by Cray, Inc.
- Initial delivery to Sandia is scheduled for September, 2004
- Limited availability: 2Q05
  - (developers only)

Red Storm is a Massively Parallel Processor

System Configuration



Node Architecture



ASIC = Application Specific Integrated Circuit, or a "custom chip"

Comparison of ASCI Red and Red Storm

	ASCI RED	Red Storm
Theoretical Peak (TF)	3.15	41.47
MP-Linpack Performance (TF)	2.379	>20 (est)
Architecture	Distributed Memory MIMD	Distributed Memory MIMD
Number of Compute Node Processors	9,460	10,368
Processor	Intel PH @ 333 Mhz	AMD Opteron @ 2.0 Ghz
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
Interconnect Topology	3-D Mesh (x,y,z) 38 x 32 x 2	3-D Mesh (x,y,z) 27 x 16 x 24

Comparison of ASCI Red and Red Storm

	ASCI Red	Red Storm
Interconnect Performance **MPI Latency** **Bi-Directional Link B/W** **Minimum Bi-section B/W**	**15 µs 1 hop, 20 µs max** **800 MB/s** **51.2 GB/s**	**2.0 µs 1hop, 5 µs max** **6.0 GB/s** **2.3 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 (OSFI)** **VX-Works**	**Catamount (Cougar)** **LINUX** **LINUX**
Red Black Switching	**2260-4940-2260**	**2688-4992-2688**
System Foot Print	**~2500 sq ft**	**~30000 sq ft**
Power Requirement	**850 KW**	**1.7 MW**

Red Storm System Characteristics

- 40TF peak performance
- 108 compute node cabinets, 16 service and I/O node cabinets, and 16 Red/Black switch cabinets
- 10,368 compute processors
- 2 x 256 service and I/O processors (256P for red, 256P for black)
- AMD 64 bit Opteron™ processor
- 10 TB DDR memory
- 240 TB of disk storage(120TB for red, 120TB for black)
- Linux/Catamount Operating Systems
- Approximately 3000 ft^2 including disk systems
- <2.0 megawatts of power and cooling

Red Storm – System Software

Operating Systems

- LINUX on service and I/O nodes
- LWK on compute nodes
- LINUX on RAS monitors

Run-Time System

- Job launch
- Node allocator
- Logarithmic loader
- Batch system – PBS Pro

File Systems

- High performance file system
  - Lustre

User Environment

- PGI Compilers - Fortran, C, C++
- Libraries - MPI, I/O, Math, MPI-2
- Showmesh
- Debugger - TotalView
- Performance Monitor

System Mgmt and Admin

- Accounting
- Red Storm Management System
- RSMS Graphical User Interface

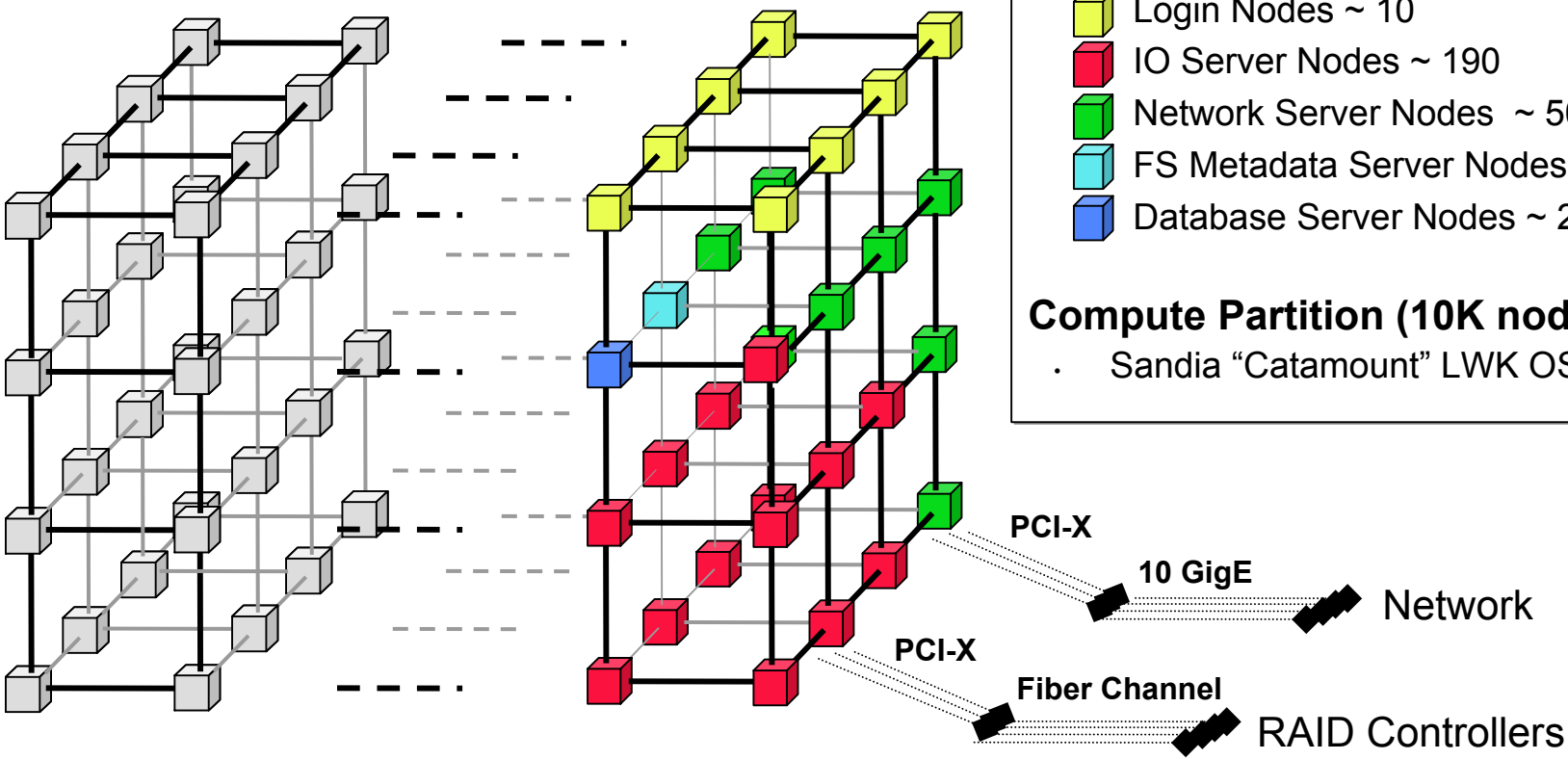
Light Weight Kernel

- Sandia has had very good experiences with LWK
  - Sandia-University of New Mexico Operating System (SUNMOS)
  - Puma
  - Cougar
  - Catamount
- Why?
  - Timing stability
  - Maturity

Breaks in MPP Systems Software

- Unix, Linux, any OS
  - Kernel memory allocation
  - TCP/IP backoff calculations
  - Routing tables
  - Clock synchronization
  - Scheduler
  - Etc., full list unknown, but has been extremely problematic with DOE labs
- Light Weight Kernel
  - None

Red Storm – Partitions, OS, Layout



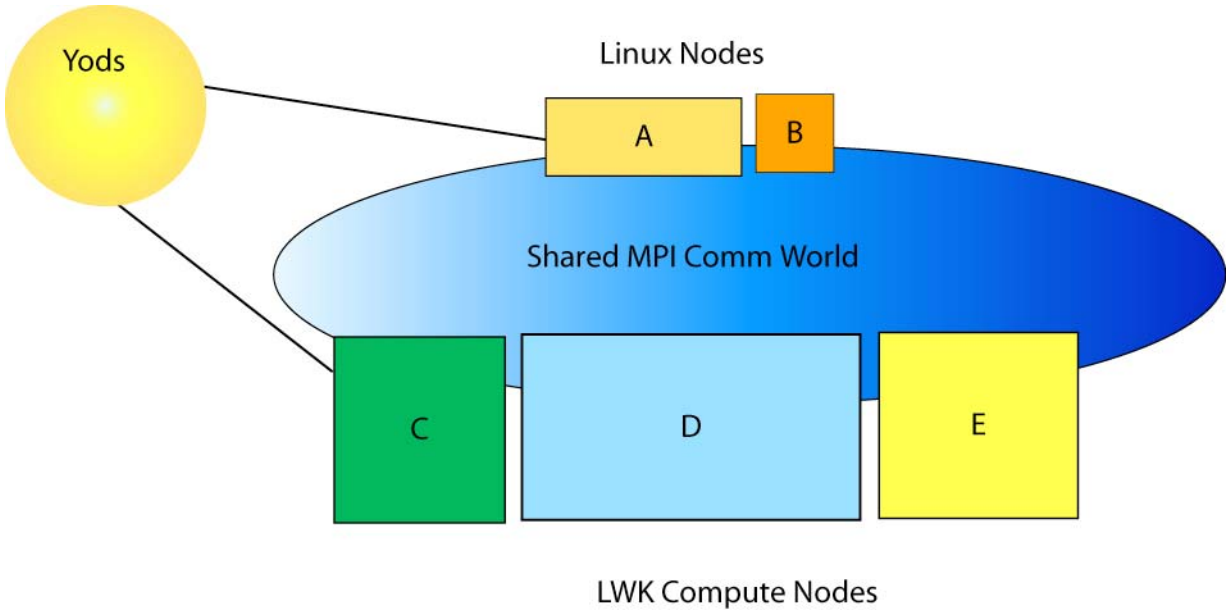
Service Partition (256 nodes)

- Linux OS
- ***Specialized Linux nodes***
  - Login Nodes ~ 10
  - IO Server Nodes ~ 190
  - Network Server Nodes ~ 50
  - FS Metadata Server Nodes ~ 4
  - Database Server Nodes ~ 2

Compute Partition (10K nodes)

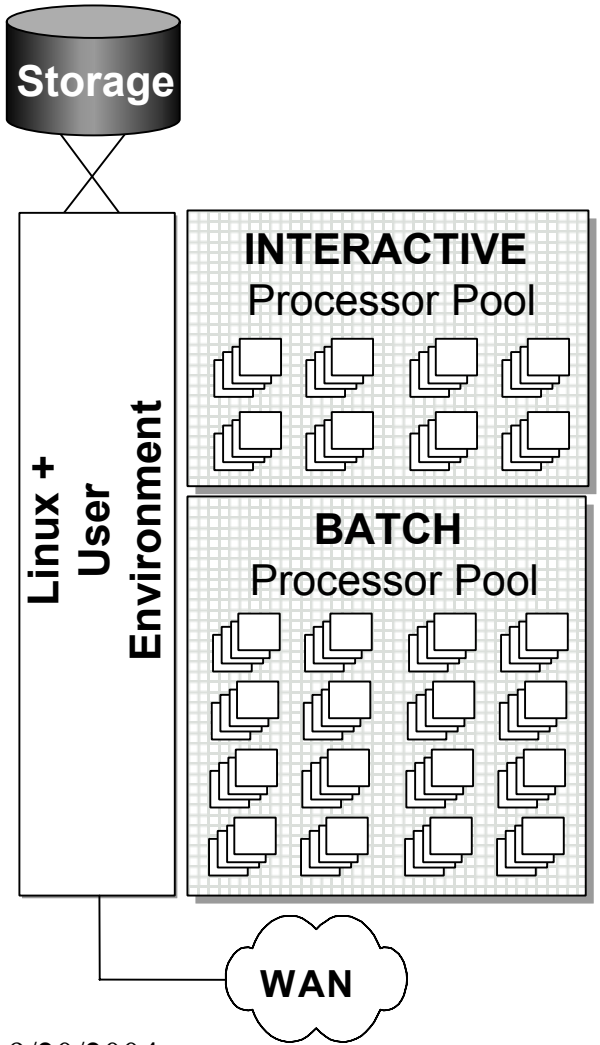
- Sandia "Catamount" LWK OS

Coming: Heterogeneous Apps



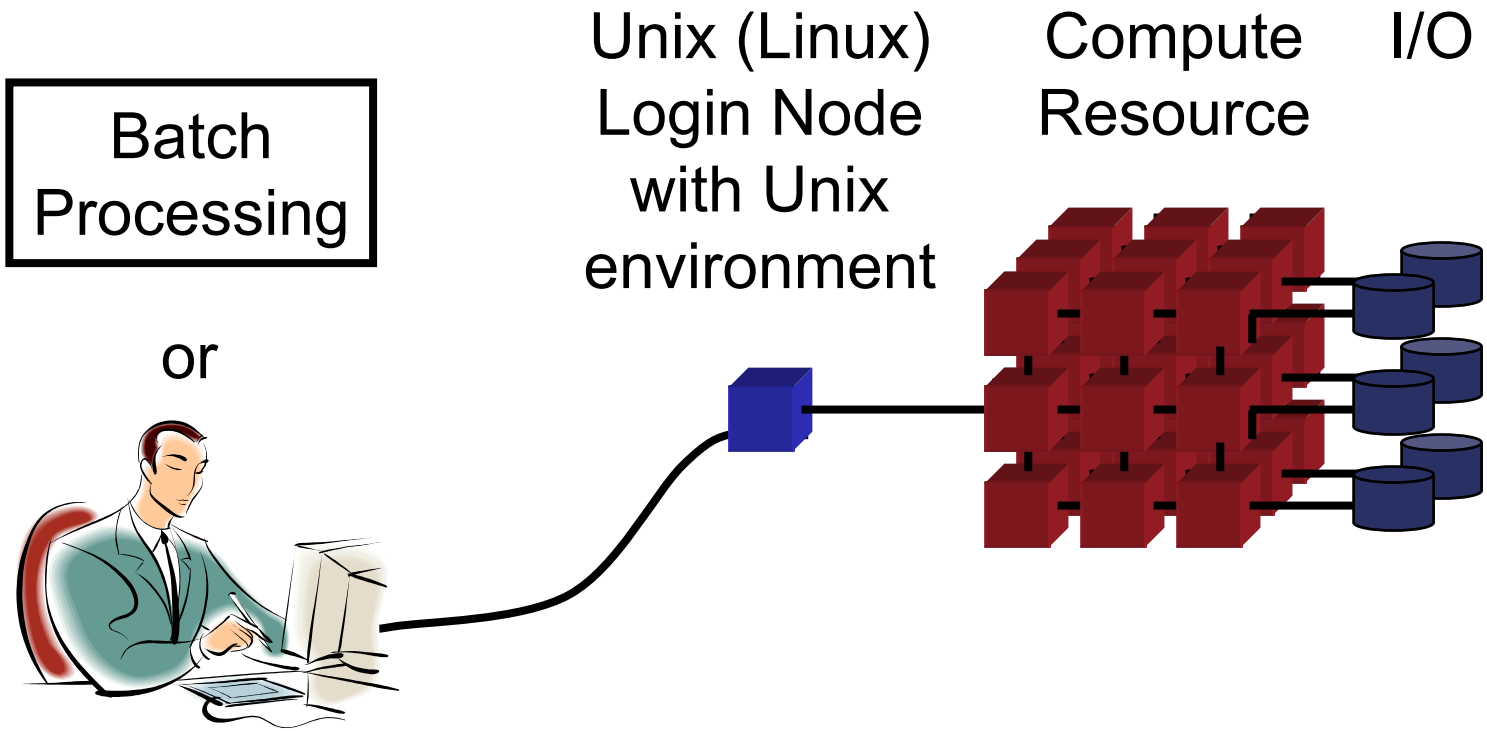
- Single yod controls multiple executables
- App shares MPI_Comm_World
- Can cross node types (Linux, LWK)
- How can you exploit this? (Research opportunity?)

Red Storm – Usage Model



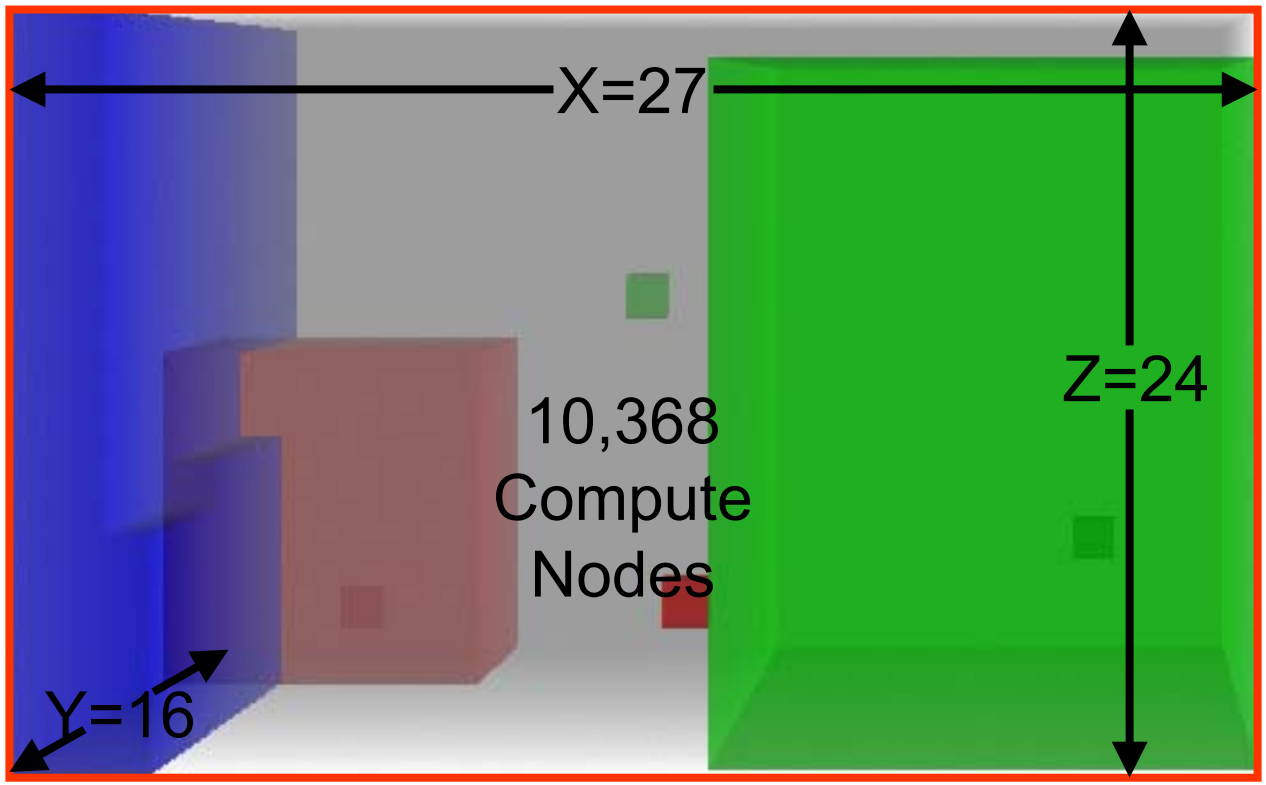
- Distributed memory Message Passing with MPI – single processor per node
- Global file access and naming
- Global TCP/IP connectivity from service nodes
- Familiar job launch (yod)
- Global ability to kill yod jobs
- Programming environment
  - MPI-2, C, Fortran 90
  - Math libraries
  - PGI Compilers
  - TotalView
  - Performance Tool

Usage Model


Batch Processing
or
Unix (Linux) Login Node with Unix environment
Compute Resource
I/O

Space Sharing of Jobs

- Jobs occupy disjoint regions simultaneously
- Example – red, green, and blue jobs:



Reliability

- Reliability for Scientific Applications at Sandia
  - High Mean Time Between Interrupts (MTBI) for hardware and system software
  - High Mean Time Between Errors/Failures (MTBF) that affect users
- What it is not
  - High availability

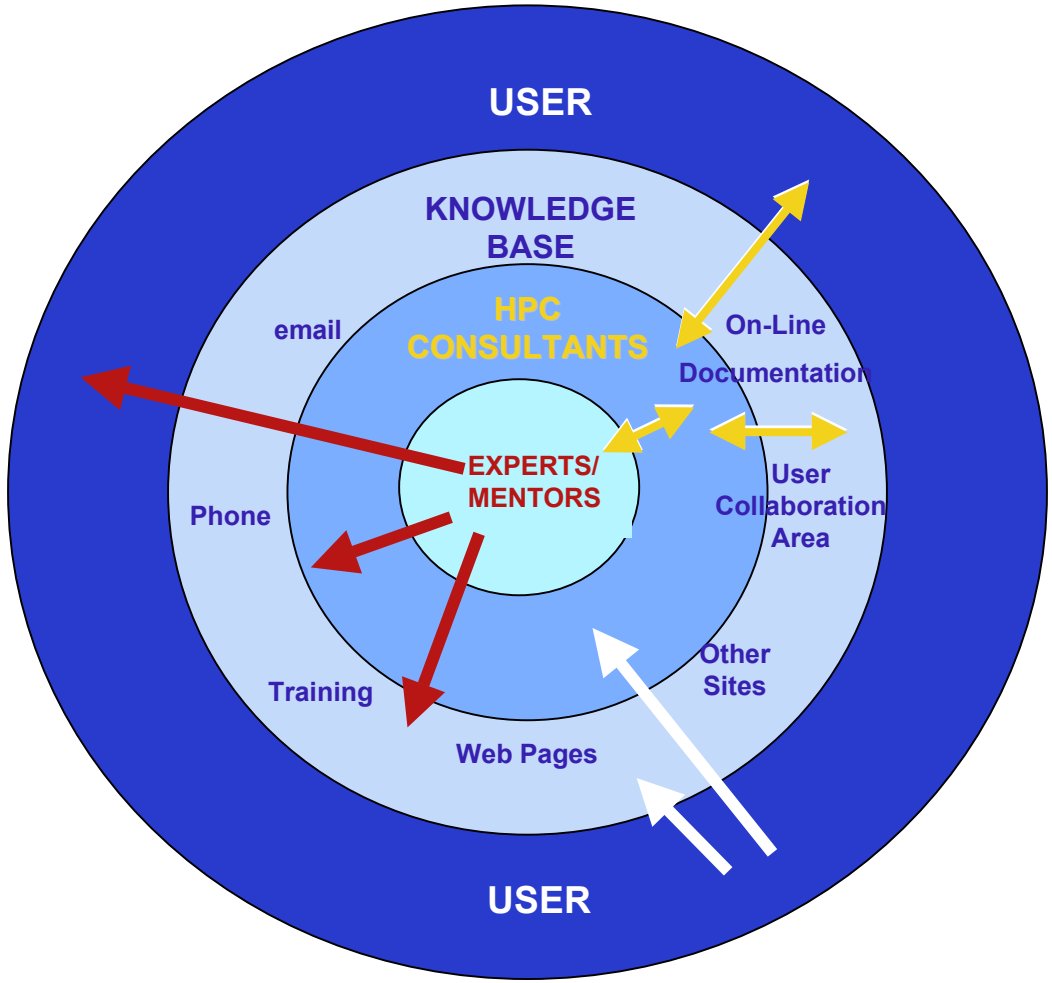
Reliability Features of Red Storm

- Light Weight Kernel (LWK) O. S. on compute partition
  - Less code fails less often
- Monitoring of correctible errors
  - Fix soft errors before they become hard
- Hot swapping of components
  - Overall system keeps running during maintenance
- Redundant power supplies & memories
- RAS System

Sandia Scientific Computing Support (**SSCS**) presents Collaborative Learning Information and Knowledge (**CLIK**)

- Focus
  - High Performance Computing Environment
- Getting Help
  - Experts as consultants
  - Access via phone, web, and email.
- Knowledge Base
  - Web based queries
  - User/Expert Contributions
- Training
  - Web based on-line training

Knowledge Culture Support for HPC Community



HPC

Advanced Simulation & Computing™

High Performance Computing

Sandia National Laboratories

Customer Support Community

Selection...

- CLIK
- Collaboration Area
- Training
- SARAPE
- Code Optimization
- Calendar
- Analyst Information

© Sandia Corporation | Ruth Villegas | Privacy and Security

User Account Contacts

Univ of Utah, Dav de St. Germain

UIUC, William A. Dick

Univ of Chicago, Carrie Eder

Stanford, Parviz Moin

- Deb Michael

Caltech, Michael Aivazis

- Heather Young
- Tracy Sheffer

Remote Account Requests via SARAPE

- Automated, Web Based
- May request accounts to be created or deleted
- Need to know the name and the location of the resource
- Phase V will direct information to Foreign National Request process
- Phase V will allow user to renew accounts
- http://sarape.sandia.gov

Sandia Scientific Computing Support - SSCS

- SSCS-help@sandia.gov
- Email lists *machine-help@sandia.gov*
- Informational email *machine-users@sandia.gov*