ASC Alliance Center Site Visit

Sandia National Laboratories

September 8-16, 2005

Karen Haskell, Barbara Jennings &
Lilia Martinez

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.

2005 ASC Alliance
SNL Presentation Agenda

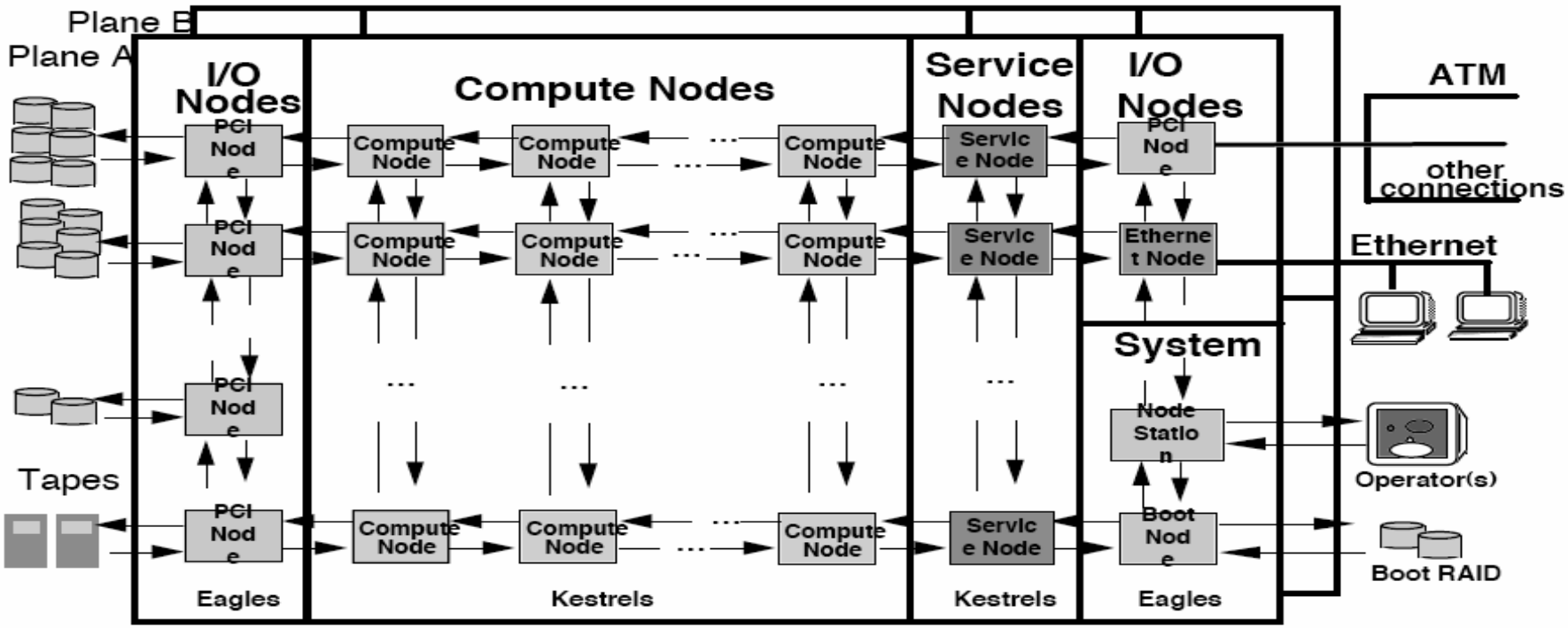
- Unclassified ASC systems at SNL
  - ASCI Red, a.k.a. Janus
  - Red Storm
- Status of Janus
- Red Storm description and status
- ASC Network
- Customer support environment CLIK

Unclassified Systems Available for Alliance Partners

- ASCI Red, a.k.a. Janus will be decommissioned as soon as Red Storm is at full release

-Red Storm, currently released for limited availability

Janus Hardware Configuration



- 2 planes
- Service nodes provide shell access, single-system image

ASCI Red Software

http://www.sandia.gov/ASCI/Red/

Operating Systems

Compute Partition:
Cougar: LWK; Program Loading, Memory Management, Message Passing, Support, Signal and Exit Handling, Run Time Support

Service & I/O Partitions:

'Teraflop Operating System' (OSF Unix); Boot and Configuration Support, System Admin, User Logins, User Commands and Services, Development Tools

Run-Time System

- Job launch – (yod)
- Batch system – NQS

File Systems:

/home/user_name – Home directory; cross mounted on *sasn100* and *janus*

Intel Parallel File System (PFS)

/pfs_grande; /pfs_grande/multi – Higher Performance; larger files

Standard Network and Unix File Systems (UFS)

/scratch – less than 2GB

/ufs/tmp_*

User Environment

PGI Compilers – F77, F90, C, C++

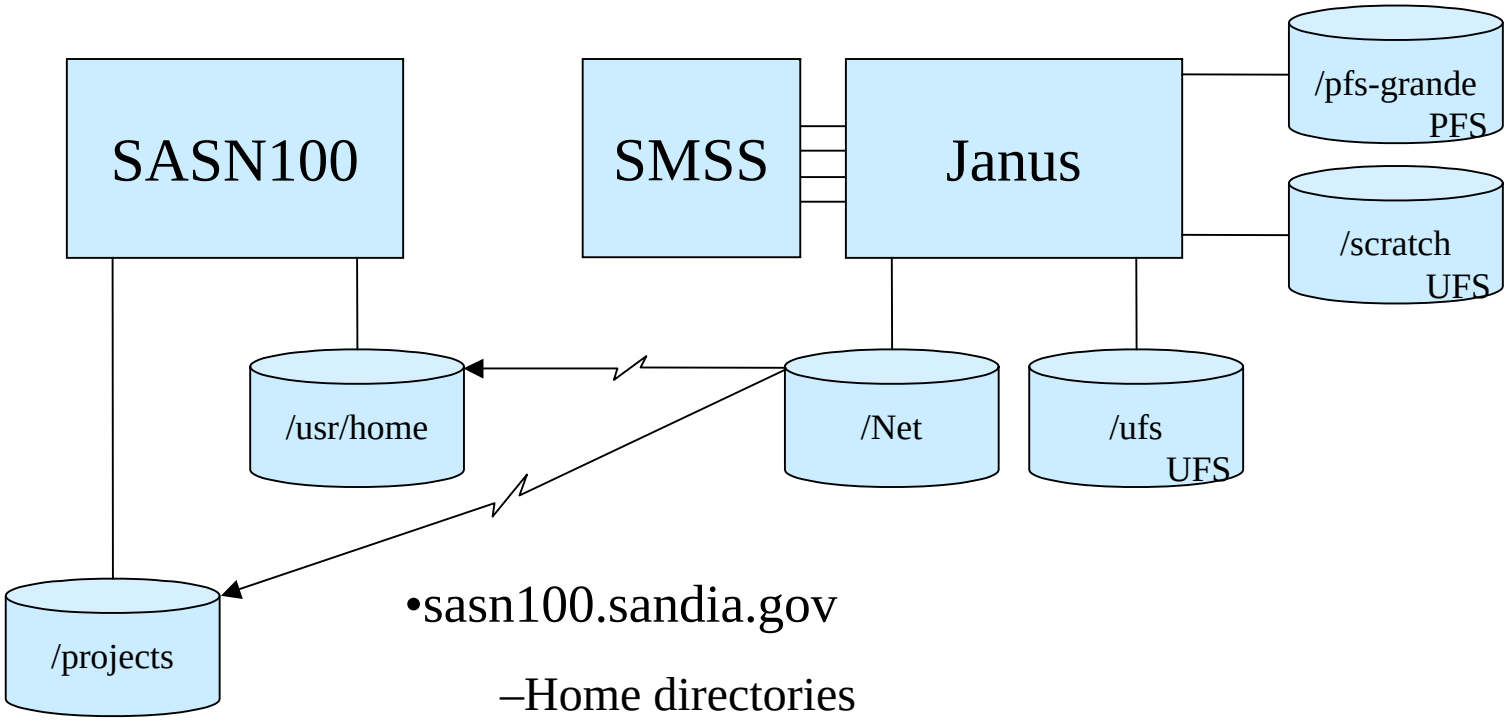
Libraries: *libmpi.a, libc.a, libm.a,libkmath.a,libperfmon.a, libdbmalloc*.

Showmesh

Debugger – *debug, xdebuf*

Performance Monitor – *perfmon*

Janus System Configuration



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

Information for using ASCI Red is available digitally.

Red Storm

- What it is
- Where we are
- Where we're going

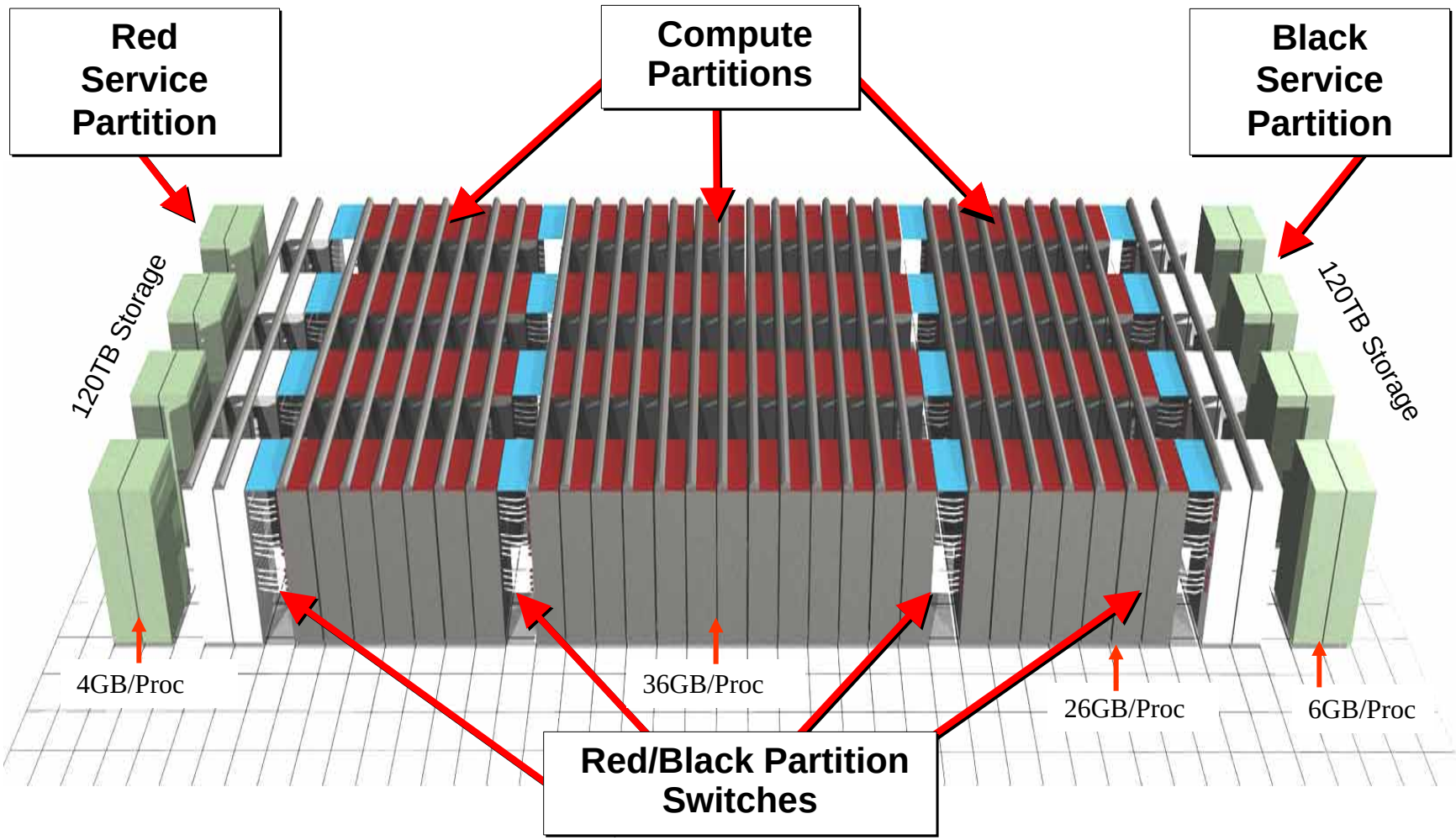
Red Storm Presentations at CUG May 2005

www.cug.org

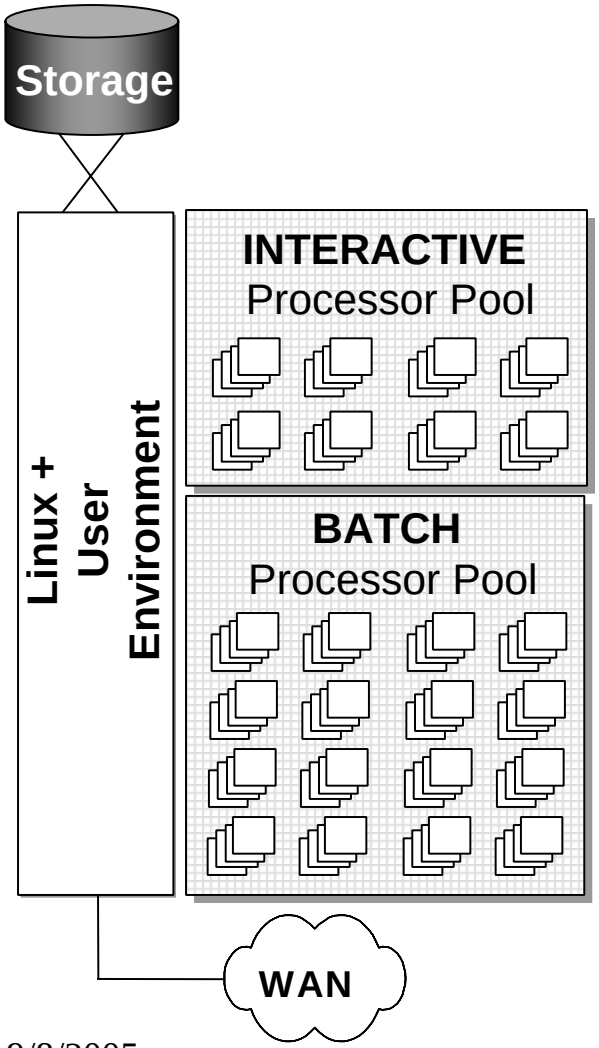
- Robert A. Ballance, Milton Clauser, Barbara J. Jennings, David J. Martinez, John H. Naegle, John P. Noe, and Leonard Stans, *Red Storm Infrastructure at Sandia National Laboratories*
- Suzanne Kelly and Ron Brightwell , *Software Architecture of the Lightweight Kernel, Catamount*
- James Ang, Robert A. Ballance, Lee Ann Fisk, Jeanette R. Johnston, and Kevin T. Pedretti, *Red Storm Capability Computing Queuing Policy*

Red Storm is a Massively Parallel Processor

System Configuration



Red Storm – Usage Model



- Distributed memory Message Passing with MPI
- Global file access and naming
- Global TCP/IP connectivity from service nodes
- Familiar job launch (yod)
- Single system view
- Heterogeneous applications
- Global ability to kill yod jobs
- Programming environment
  - MPI-2, C, Fortran 90
  - Math libraries
  - PGI Compilers
  - TotalView
  - Performance Tools

Comparison of ASCI Red and Red Storm

	ASCI Red	Red Storm
Theoretical Peak (TF)	3.15	41.47
MP-Linpack Performance (TF)	2.379	>30 (est.)
Architecture	Distributed Memory MIMD	Distributed Memory MIMD
# of Compute Node Processors	9,460	10,368
Processor	Intel PH @333Mhz	AMD Opteron @2.0Ghz
Total Memory	1.2 TB	30.95 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	22 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

Red Storm System Software

Operating Systems

- LINUX on service and I/O nodes (SuSE Enterprise Server)
- Catamount LWK on compute nodes
- Linux on RAS monitors

Run-Time System

- Logarithmic job launch (yod)
- Node allocator (CPA)
- 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

Network

- 10x1GigE to login nodes
- 1 GigE to Mgmt Workstations

System Mgmt and Admin

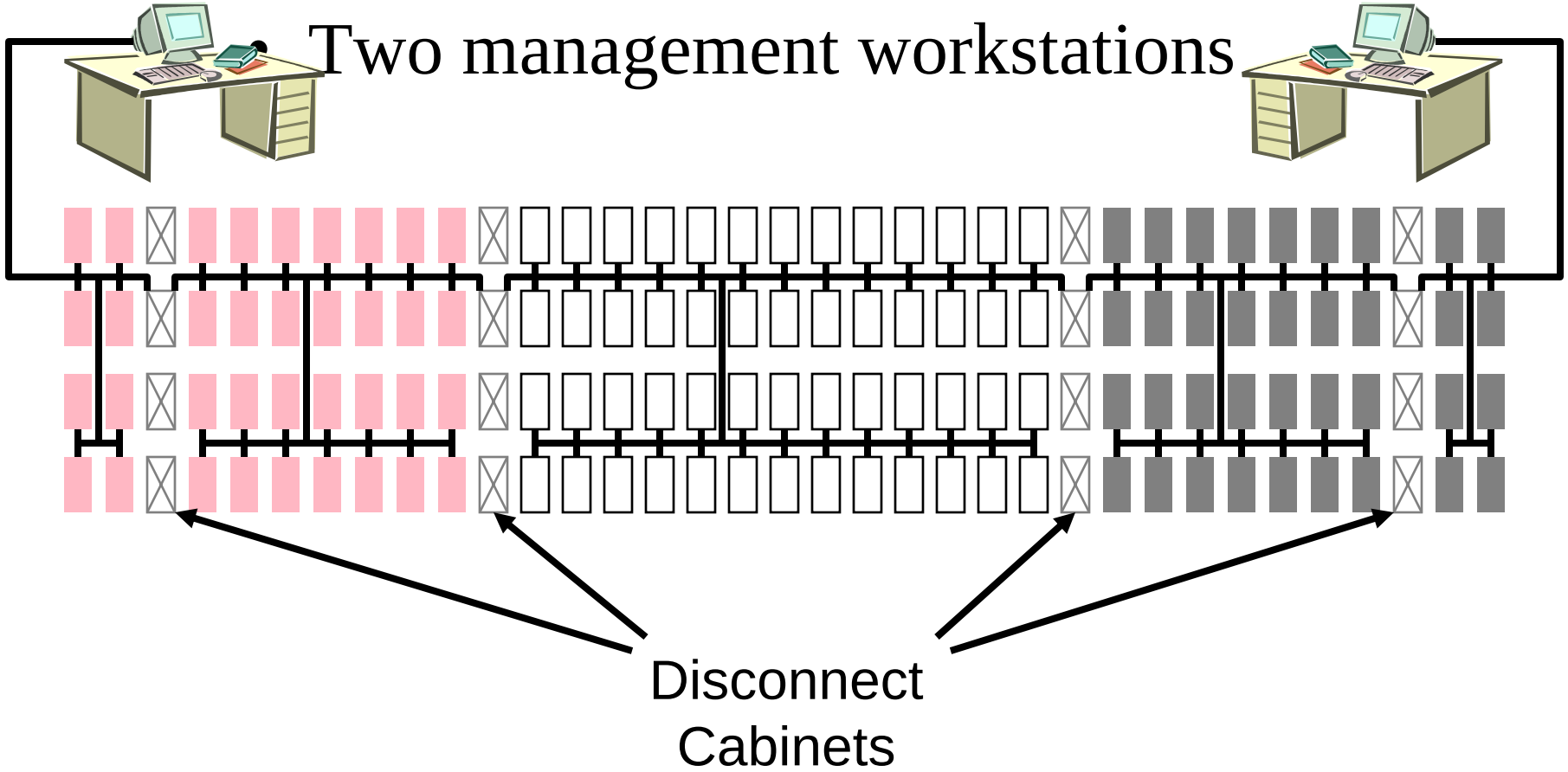
- Accounting
- Red Storm Management System

Reliability Features of Red Storm

- Light Weight Kernel (LWK) OS 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
- RAS System

Reliability, Availability, and Serviceability

- Separate LAN to all nodes for RAS
- Two management workstations



Catamount: The Light Weight Kernel (LWK) on Red Storm

Why a LWK?

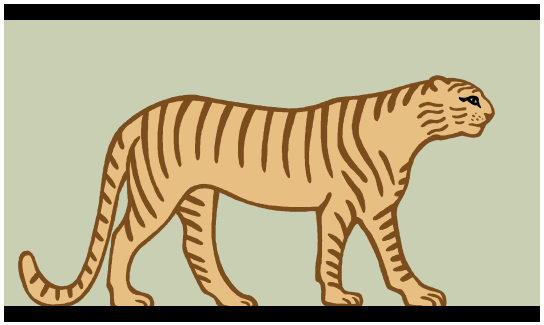
- Purpose of compute nodes is to compute – LWK maximizes the resources devoted to computation
- Every time a processor takes an interrupt, other processors in the parallel computation may have to wait on it
- This may not be an issue for 10's of nodes
- Red Storm has 10,000 CPUs, hence LWK approach important

The Light Weight Kernel (LWK)

- A lightweight compute node OS is a fundamental part of the Sandia architecture
- It is essential for
  - Maximizing CPU resources
    - Reduce OS and runtime system overhead
  - Maximizing memory resources
    - Small memory footprint, large page support
  - Maximizing network resource
    - No virtual memory, physically contiguous address mapping
  - Increasing reliability
    - Small code base, reduced complexity
  - Deterministic performance
    - Repeatability
  - Scalability
    - OS resources must be independent of job size
- Others have realized these benefits
  - nCUBE (Vertex), Cray T3 (UNICOS/mk), IBM BG/L (HPK)
- Similar trend in Linux-based cluster SW – reducing overhead and unused capabilities

History of Catamount

- SUNMOS (Sandia-UNM Operating System) on Paragon
- Puma – Rewrite of SUNMOS as part of a research project
- Cougar – Port of Puma running ASCI Red
- Catamount – Port of Cougar to run on Red Storm



Similarities between Cougar and Catamount

- PGI compilers (C, C++, FORTRAN 77/90)
- No support for threads
- No dynamic process creation (e.g. fork/exec)
- No shared or dynamic libraries (e.g. dlopen)
- No UNIX sockets or IP network stack (portals only)
- Same custom malloc() which rewards large contiguous allocations
- Large (2MB) page support

Differences between Cougar and Catamount

- A cougar node has
  - dual 333MHz Pentium 2 Xeon processors (32-bit)
  - 256MB of memory
- A catamount node has
  - one 2GHz AMD Opteron processor (64-bit)
  - 4,3,0r 2GB of memory
- Cougar uses PGI V3.1; Catamount uses PGI V6.0
- Cougar uses MPICH1 V1.0.12; Catamount uses MPICH2 V0.97 (MPI environment variables will differ)
- Support for TotalView debugger
- Support for PAPI and Vampir
- Support for Lustre parallel and serial file systems
- New processor allocation program, which supports red/black switch horizons

Current and Projected Red Storm Status

Mar 2005 — Red Storm delivery

Oct 2005 — *Initial operations*

Limited availability

Aug 2006 — *General availability*

Upcoming Red Storm Milestones

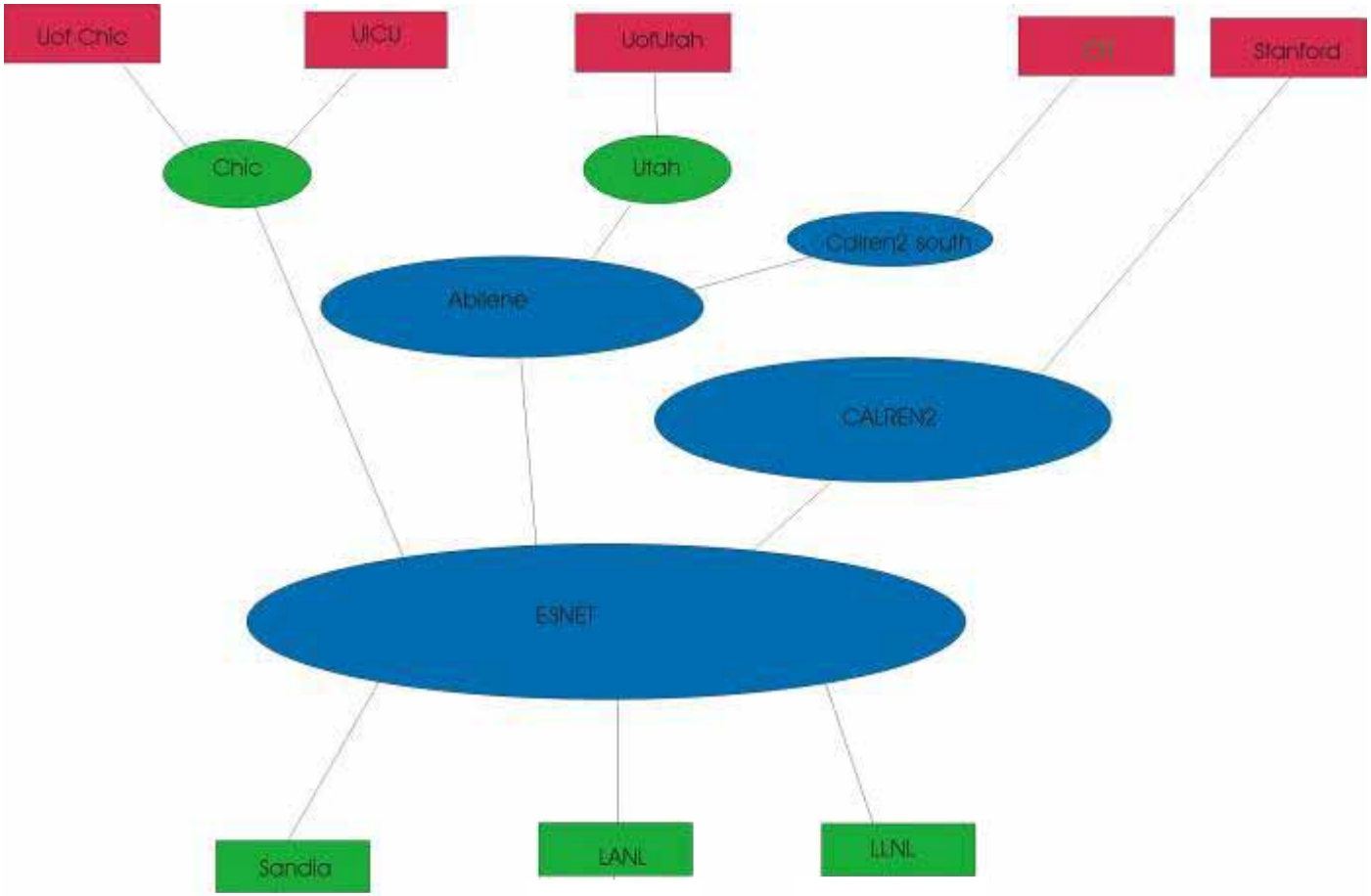
- Complete ASC milestones which will result in a production-ready system for a limited set of users
- Scalability, reliability
- Enlarge file systems
- Looking ahead to next Cray releases (1.2, 1.3)
- Enlarge user base

Network

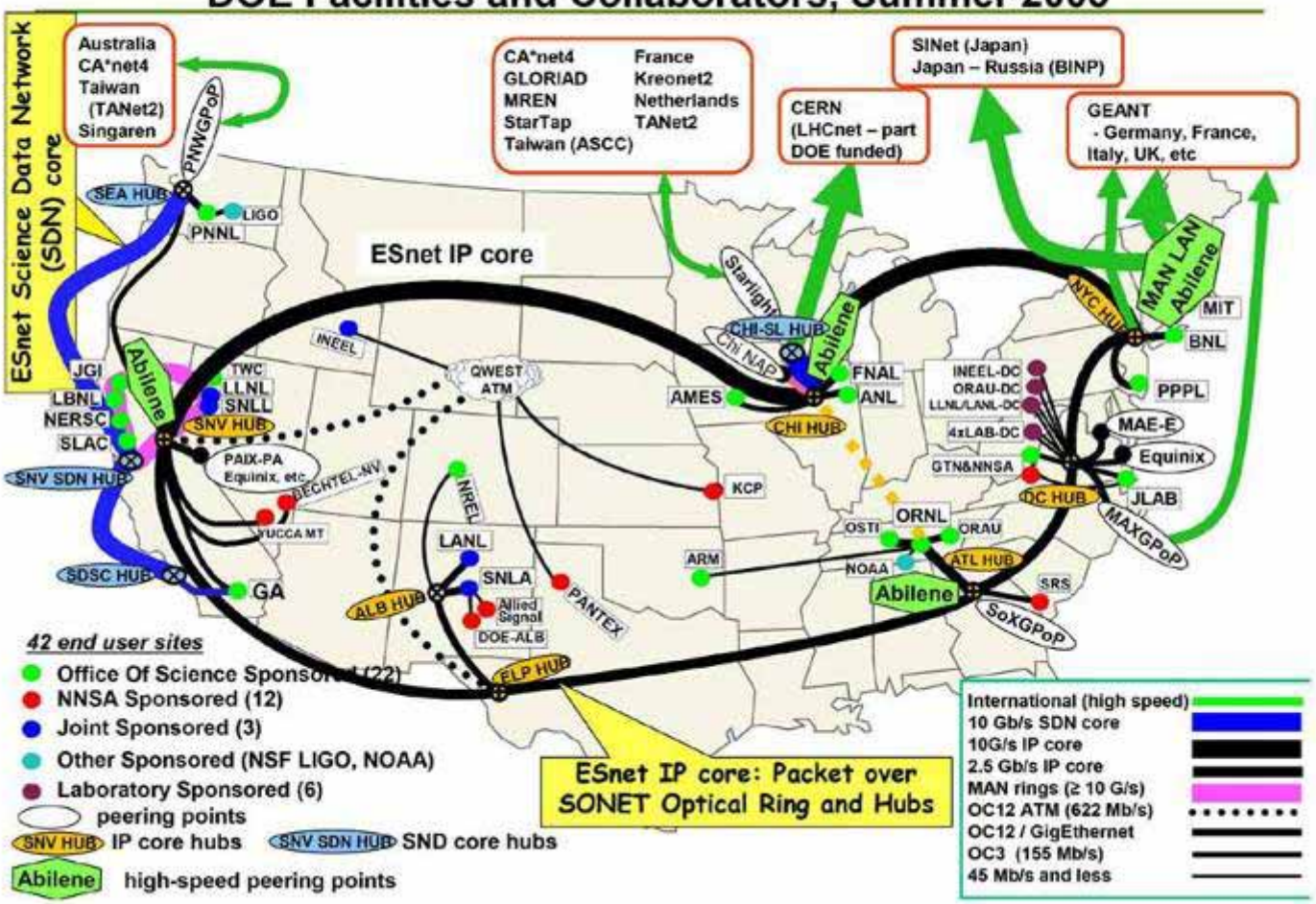
Tom Pratt

Current Lab Networks

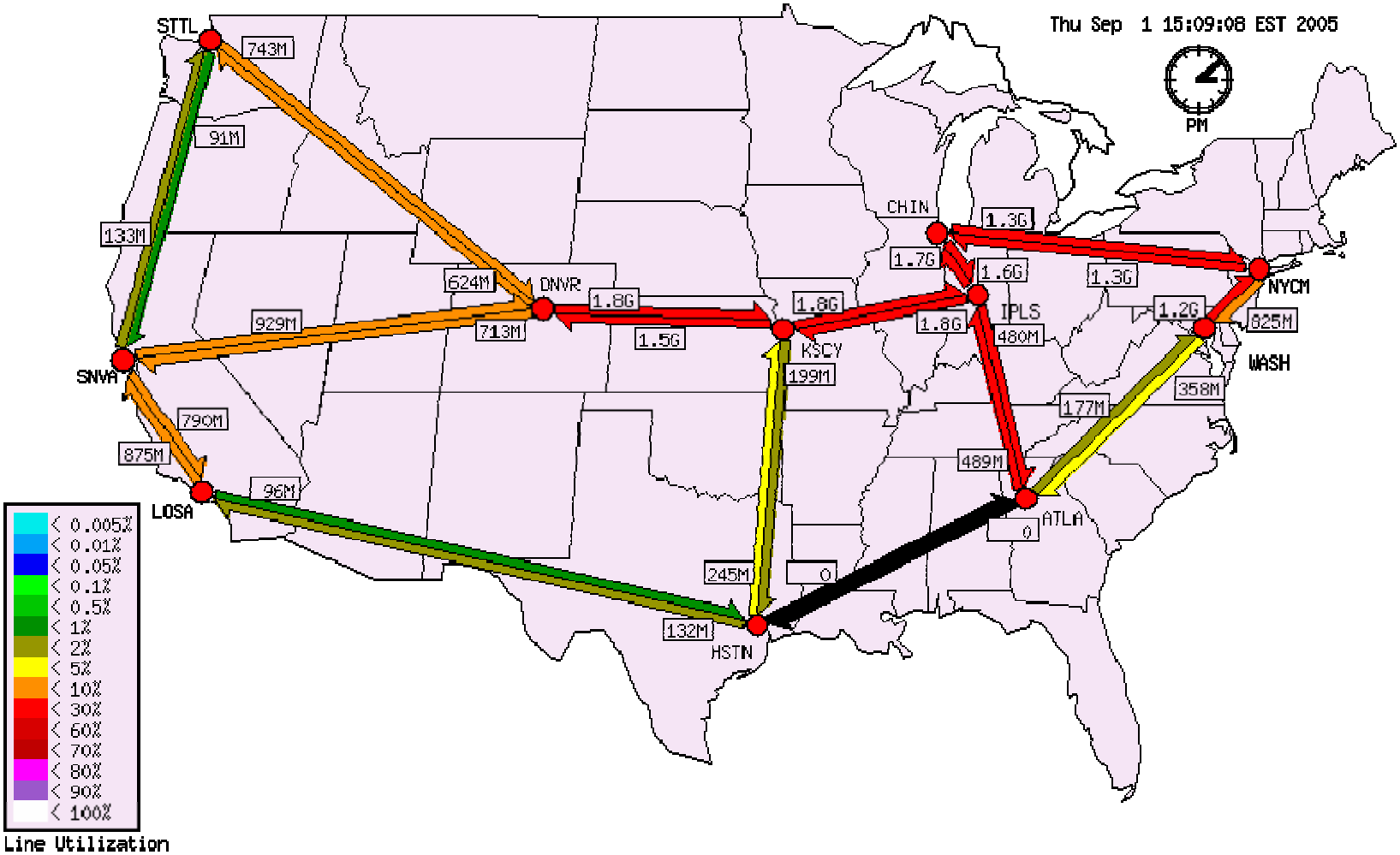
- Expected Performance
  - LANL 50 Mbit/sec
  - SNL 32 Mbits/sec
- Dependency
  - Data Movement Apps
  - Latency Tolerance 30-80ms
  - Network Congestion

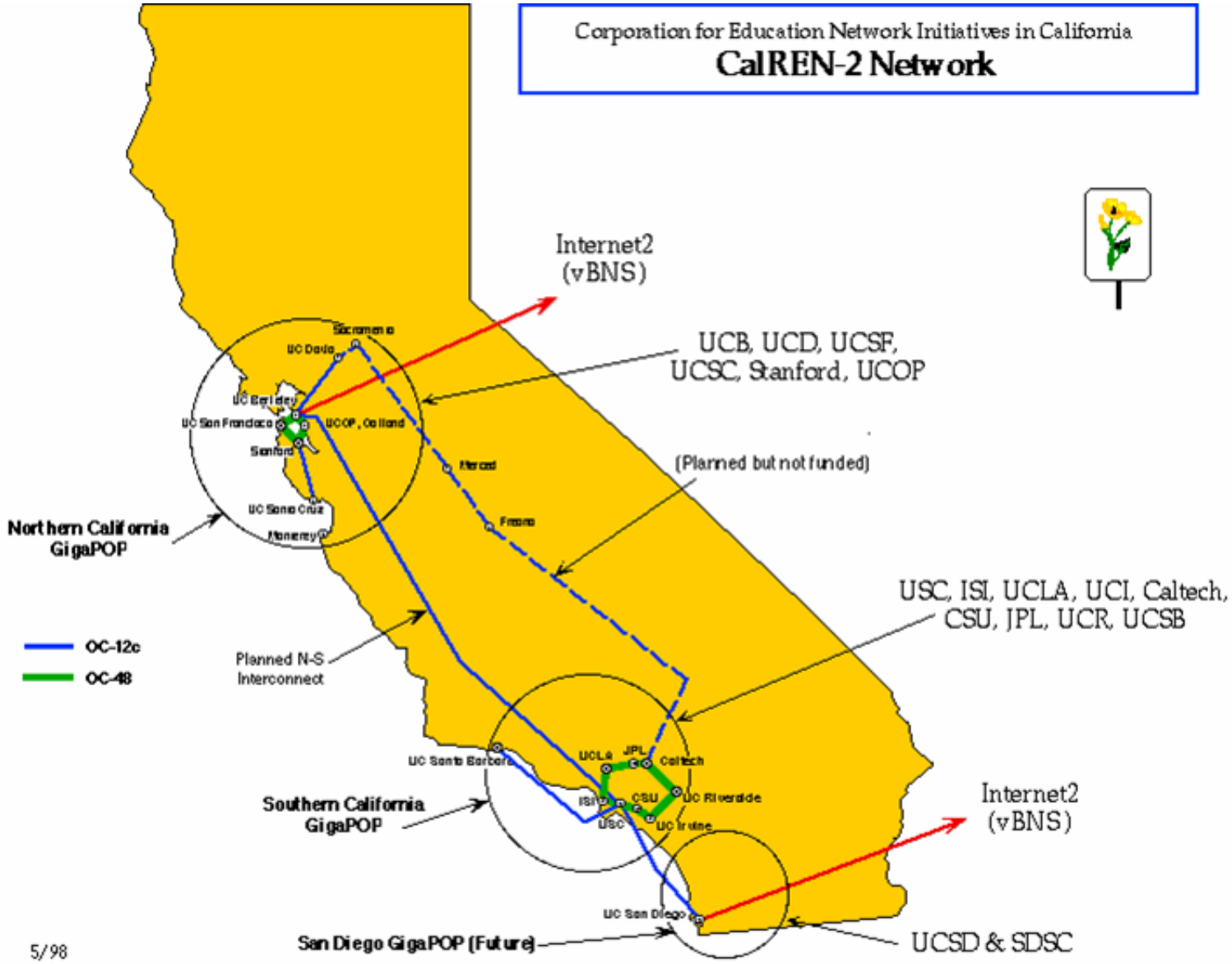

Uof Chic
UICU
UofUtah
OII
Stanford
Chic
Utah
Calren2 south
Abilene
CALREN2
ESNET
Sandia
LANL
LLNL

ESnet High-Speed Physical Connectivity to DOE Facilities and Collaborators, Summer 2005



INDIANA UNIVERSITY ABILENE NOC WEATHERMAP




Corporation for Education Network Initiatives in California
CalREN-2 Network
Internet2
(vBNS)
UCB, UCD, UCSF,
UCSC, Stanford, UCOP
(Planned but not funded)
USC, ISI, UCLA, UCI, Caltech,
CSU, JPL, UCR, UCSB
Internet2
(vBNS)
UCSD & SDSC
Northern California
GigaPOP
Southern California
GigaPOP
San Diego GigaPOP (Future)
Planned N-S
Interconnect
OC-12c
OC-48
Sacramento
UC Davis
UC Berkeley
UC San Francisco
UCOP, Oakland
Stanford
UC Santa Cruz
Monterey
Merced
Fresno
UC Santa Barbara
UCLA
JPL
Caltech
ISI
CSU
UC Riverside
USC
UC Irvine
UC San Diego
5/98

Proposed
Red Storm Network Access

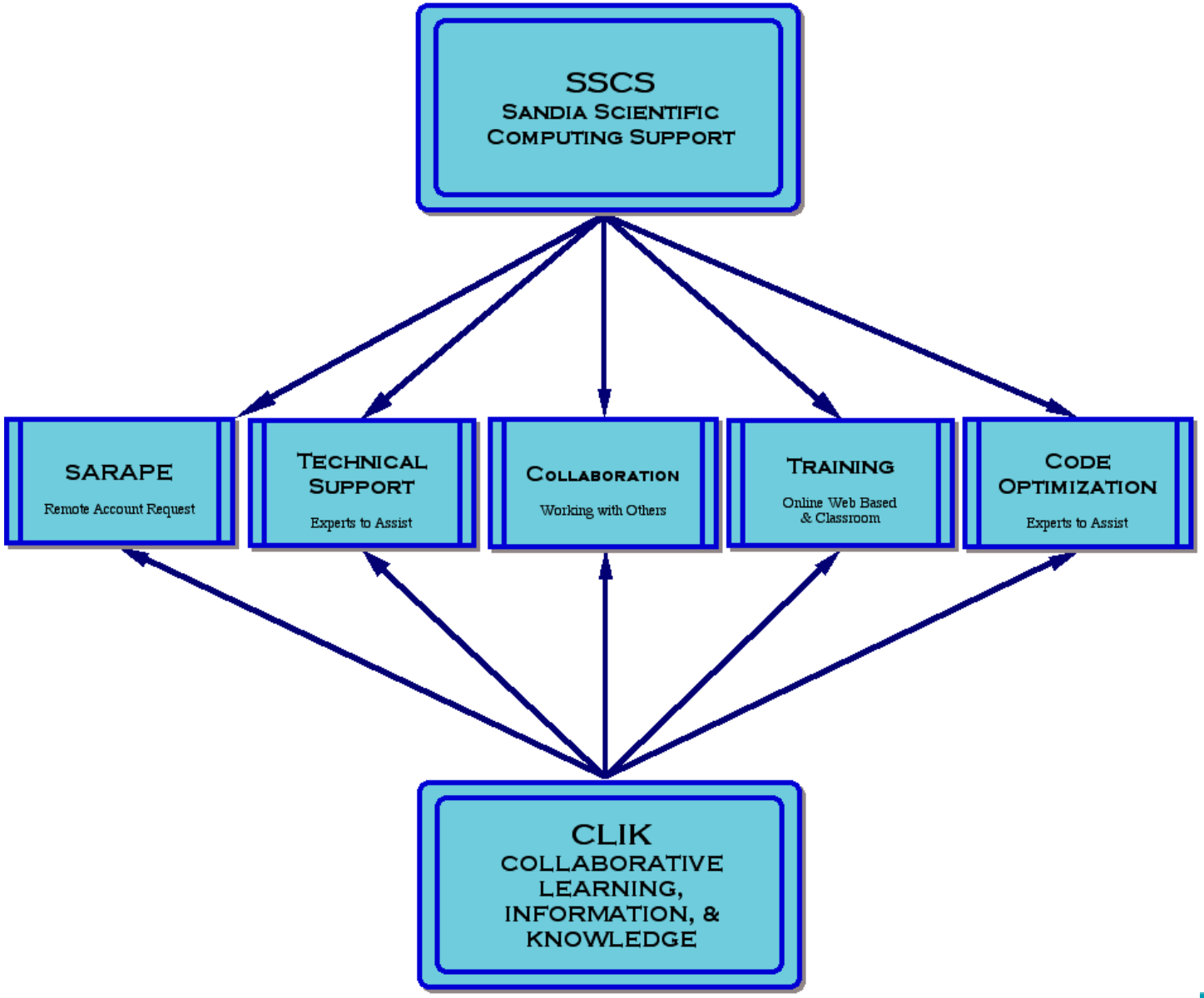
- ssh, scp to srngate.sandia.gov
- Will require 2-factor authentication
  - Crypto card or SecureID
  - Change from Janus, due to new emphasis on security stance
- Load-balancing login node server will forward session to appropriate login node

A Building for Red Storm

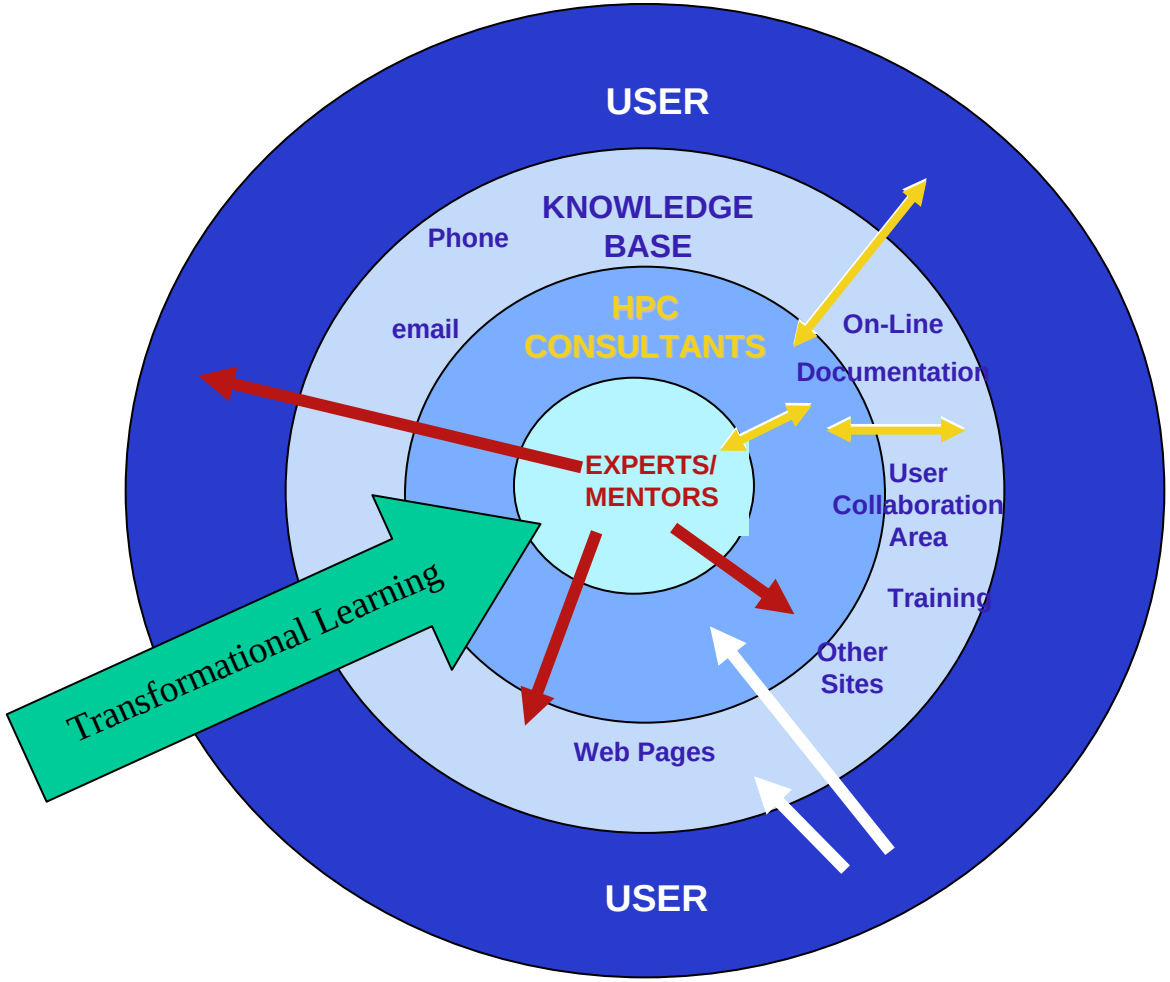
High Performance Computer users needs require unique technical support

We have created a wisdom community to provide support for HPC computing at SNL

- 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


SSCS
Sandia Scientific Computing Support
SARAPE
Remote Account Request
Technical Support
Experts to Assist
Collaboration
Working with Others
Training
Online Web Based & Classroom
Code Optimization
Experts to Assist
CLIK
Collaborative Learning, Information, & Knowledge

Knowledge Culture Support for HPC Community



Unique support is provided to all SNL users by means of a custom tool: http://clik.sandia.gov

CLIK

COLLABORATIVE LEARNING, INFORMATION, & KNOWLEDGE

CLIK Collaborative Learning, Information, and Knowledge

Advanced Simulation & Computing

Sandia National Laboratories

ASC High Performance Computing On-line:

Selection...

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

Find Information in the Knowledge Base Community:

Search CLIK Knowledge Base

Add to the CLIK Knowledge Base:

Knowledge Base Submittal

Need technical assistance?

Technical Assistance Request

Provide Technical Assistance and Track Requests:

TAR Administration

Note: CLIK is for Unclassified Information Only.

© Sandia Corporation | Ruth Villegas | Privacy and Security

Email requests to machine-help generate Technical Assistance Requests (TARs)

Subject: TAR05-Aug-00696 - Submission - ACTION REQUIRED - RE: Problem with redstorm login and net nodes
From: wg-clik-tar@sandia.gov
Reply-To: wg-clik-tar@sandia.gov
Date: 8/30/05 10:35 AM
To: wg-clik-feynman@sandia.gov

wg-clik-feynman:

A technical assistance request (TAR) from Milton J Clauser at Sandia National Laboratories has been submitted.

Please review the request and take action, as appropriate, at
https://cfwebprod.sandia.gov/cfdocs/CLIK/templates/form/tar_submission_rev.cfm?tar_id=1666

------Original Request------
Sent: Tuesday, August 30, 2005 10:35 AM
To: <feynman-help@sandia.gov>
Cc: "Monk, Stephen" <smonk@sandia.gov>, "Haynes, Rena A" <rahayne@sandia.gov>, "Kelly, Suzanne M" <smkelly@sandia.gov>, "Machin, Glenn D" <GMachin@sandia.gov>, "Scott, Randy" <rdscott@sandia.gov>
Subject: RE: Problem with redstorm login and net nodes

(Please respond to Glenn Machin)

Is this also an issue on Red RoSE?

Milt

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
- Questions or assistance with SARAPE call 1(800)417-2634 ext 284-7727 or email sarape-help@sandia.gov
- http://sarape.sandia.gov

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

Sandia Scientific Computing Support - SSCS

URL	EMAIL
• sscs.sandia.gov	SSCS-help@sandia.gov
• CLIK.sandia.gov	Clik-help@sandia.gov
• SARAPE.sandia.gov	sarape-help@sandia.gov
• Email lists	*machine-help*@sandia.gov

In Conclusion ….

Sandia provides an exciting atmosphere to work and an exceptional opportunity to learn in a unique part of the country.

Please let us know if you are interested in obtaining an internship with us.