ASCI Alliance Center Site Visit

Sandia National Laboratories
ASCI Computing Systems

Barbara Jennings, Paula McAllister, Lilia Martinez, Thomas Pratt, Mahesh Rajan

Aug. 19 – 27, 2003

Hardware

Mahesh Rajan
Barbara Jennings

JANUS

Compute Nodes 3,342 (1,166 - 2,176)
9298 Pentium II Xeon Processors

Service Nodes 52 (26 / 26)

Disk I/O Nodes 73 (37 / 36)

System Nodes 2 (1 / 1)

Network Nodes Ethernet(1)/ATM(6)
one OC3(17MB) to remote traffic

System RAM 1212 GB Total (256 MB / 512 MB)

Topology Mesh (38 X 32 X 2)

Node Link Bandwidth - Bi-directional 800 MB/s

Cross Section Bandwidth - Bi-directional 51.2 GB/s

Total Number of Pentium II Xeon Core Processors 9298

Processor to Memory Bandwidth 533 MB/s

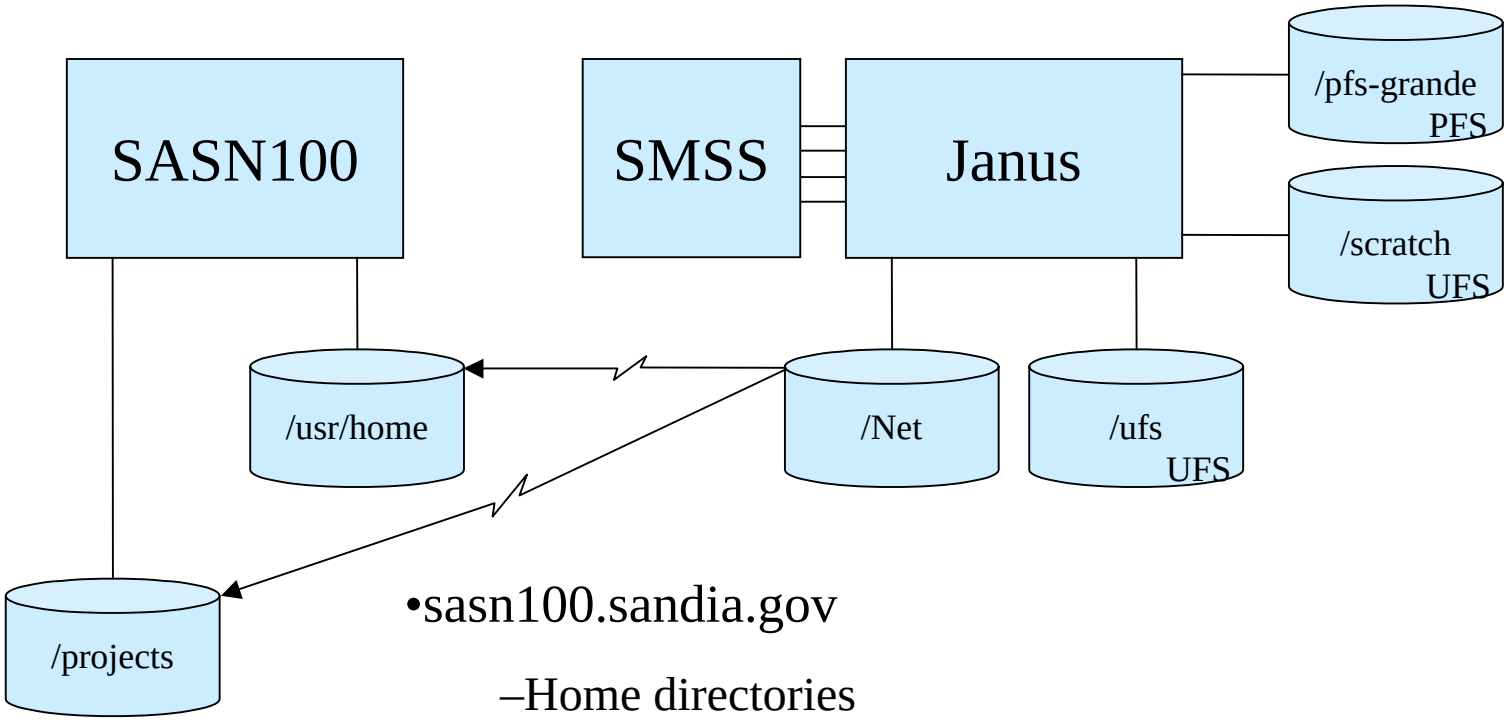
Compute Node Peak Performance 666 MOPs

System Peak Performance 3.1 TOPs

Linpack Performance - Full System 2.38 TOPs (1.6333 TOPs / .581 TOPs)

RAID Disk Storage - Total / per Color 6.25 TB

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&

Janus Support

- CRT for Technical Assistance POC for all other
- FAQ directory
  - Sasn100: /usr/local/FAQ
- Web pages
  - http://www.sandia.gov/ASCI/Red
- http://www.sandia.gov/ASCI/Red/RedFacts.htm
- Email lists *machinename-help@sandia.gov*
- Informational email *machinename-users@sandia.gov*
- Janus schedule from sadn100:
  - $news janus-dedicate
- Also janus-help@sandia.gov

Software

Mahesh Rajan
Barbara Jennings

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*

Compiling and linking

- Cross compile on *sasn100*
  - The cougar f77 compiler is *cif77*
  - The cougar f90 compiler is *cif90*
  - The cougar c++ compiler is *ciCC*
  - The cougar c compiler is *cicc*

They invoke the Portland Group compilers (e.g., *pgf77*) with the cougar switch which sets up the right compile environment for a program that will be run on a cougar compute node.

EXAMPLE: *cif77 foo.f –o foo –lmpi*

(-lmpi links with the mpi library)

Job Launch

Log in to *Janus*: *ssh janus.sandia.gov*

simple run script – call it 'run_script'

```
#!/bin/csh
#snl queue has big proc limit
#snl.day queue for short runs
date
cd $QSUB_WORKDIR
yod -sz 256 a.out < input >& output
```

Key yod options:

-sz size (number of nodes/virtual nodes)

-proc 0 : 1 processor per node (most common & default)

-proc 3 : 2 processor per node(virtual nodes mode)

-proc 1: use second processor for messages

-proc 2: 'asymmetric mode' (some libraries use it)

Submit the job on *janus* prompt with:

>qsub -re -ro -q snl -lT 2:40:00 -lP 256 run_script

Remember: janus compute nodes have no virtual memory and do not support sockets or posix threads. When using –proc 3 the NQS –lp parameter should specify half the number of virtual nodes.

Program Development tools

Checking Job Status: Use *showmesh*, *qstat* (do *man* for options), *qbyhand*

Debugger: Both command-line and graphical debuggers are available, and are invoked by typing either "*debug*" or "*xdebug*." See Tutorial: http://www.sandia.gov/ASCI/Red/usage/tutorial

Profiling and Instrumentation: See User Guide FAQ

The Intel Pentium II processor higher level counters which can be accessed are:

Name	Definition
mflops	Megaflops
icache	Instruction cache hit rate
dcachemiss	Data cache miss rate
drefs	Data access counters
utilization	Number of data packets sent out over the mesh
contention	Number of outbound data packet clashes

Performance Tips for Janus

- Compiler/Linker options: *–O2*
- To get the best I/O performance:

  Use cread/cwrite from Fortran

  Use read/write from C

  write files to the PFS.

  Make the largest possible read or write requests

  Excellent write up On I/O

  http://www.sandia.gov/ASCI/Red/usage/ioreport.ps

- Branch prediction

  If<condition>{
  No performance hit
  } else {
  Performance hit
  }

- Use the cache (block algorithms)
- Avoid floating point integer conversions where possible
- Avoid divides

ASCI Red - Miscellaneous

- User Account:
  - Contact:
    Univ of Utah, Dav de St. Germain
    UIUC, William A. Dick
    Univ of Chicago, Carrie Eder
    Stanford, Parviz Moin
    Caltech, Michael Aivazis
- ASCI Reps;
  - Barbara Jennings
  - Paula McAllister
- Getting Help:
  - Contact: *janus-help@sandia.gov*
    tflan-help@sandia.gov
- File transfer
  - *sftp, ftp, scp*
- Environment Setup
  - New user should copy; *local.cshrc, local.login, and local.profile* from home directory or from */etc/skel* to their 'dot' files
- Resource Limits/Policy
  - Home directory: 200 Mbytes
  - UFS files: 2 GB
  - PFS files not limited to 2 GB
  - /pfs_multi, /ufs – Temporary – back up to SMSS
  - NQS queues & policy: see FAQ

Network

Tom Pratt

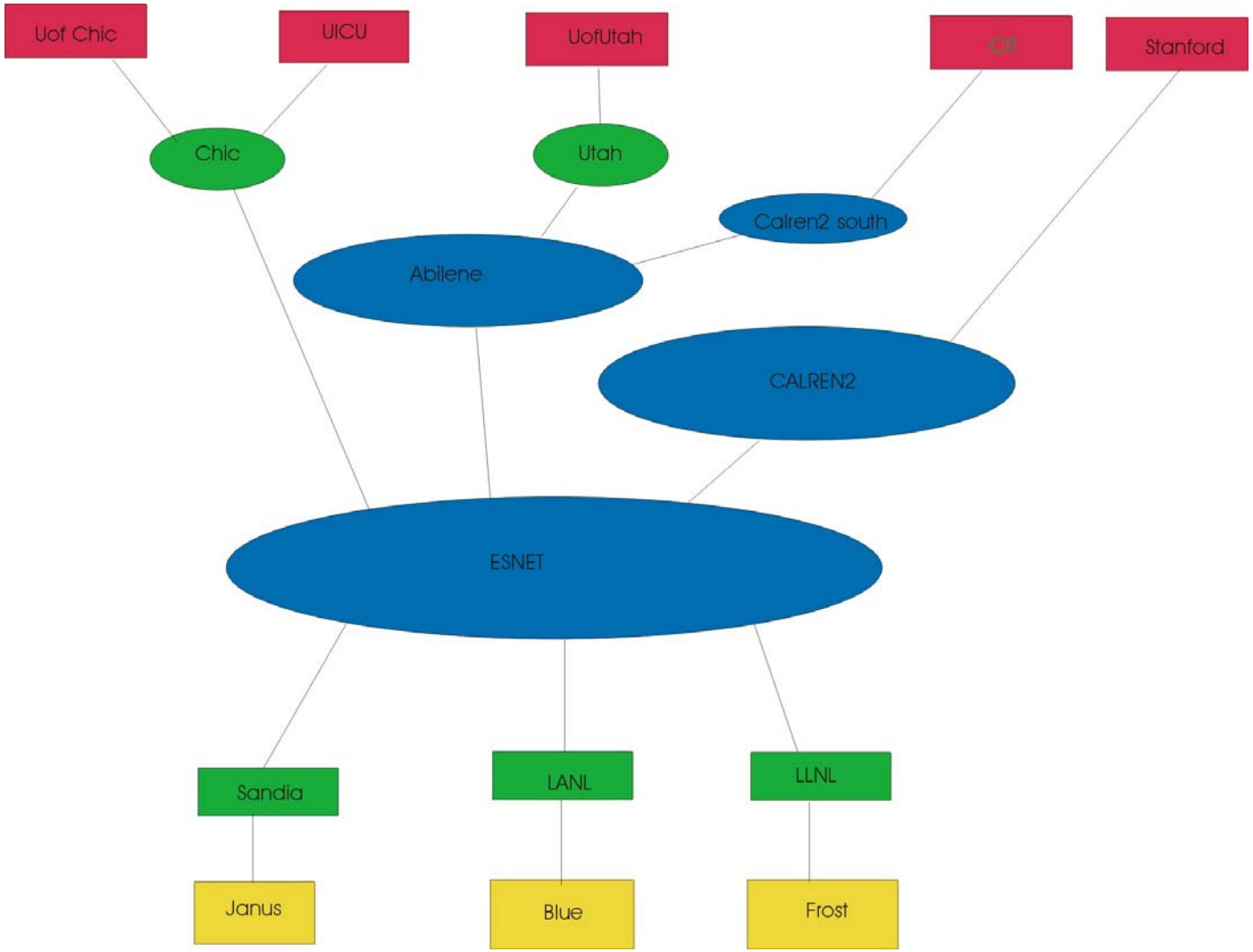
```
[tjpratt@dellux sbin]$ ./traceroute www.utah.edu
traceroute to www.utah.edu (155.99.1.2), 30 hops max, 38 byte packets
 1  ge_router (132.175.244.254)  0.446 ms  0.363 ms  0.339 ms
 2  sacr3009 (132.175.38.9)  0.642 ms  0.356 ms  0.332 ms
 3  sacr4102 (198.206.223.19)  1.633 ms  0.960 ms  2.107 ms
 4  alb-snla.es.net (134.55.208.125)  2.632 ms  2.262 ms  1.953 ms
 5  elp-alb.es.net (134.55.209.229)  7.059 ms  7.406 ms  9.167 ms
 6  snvcr1-oc48-elpcr1.es.net (134.55.209.217)  32.918 ms  32.804 ms  32.929 ms
 7  snvrt1-ge0-snvcr1.es.net (134.55.209.90)  33.562 ms  33.734 ms  33.433 ms
 8  198.32.11.93 (198.32.11.93)  34.595 ms  34.053 ms  33.901 ms
 9  dnvrng-snvang.abilene.ucaid.edu (198.32.8.2)  58.348 ms  57.621 ms  57.637 ms
10  205.124.237.9 (205.124.237.9)  246.479 ms  196.586 ms  209.824 ms
11  205.124.249.118 (205.124.249.118)  68.359 ms  68.226 ms  68.820 ms
12  crpark-wr1ebc.net.utah.edu (155.99.132.102)  68.715 ms  68.847 ms  68.305 ms
13  r1lib-crpark.net.utah.edu (155.99.132.85)  68.163 ms  68.015 ms  68.030 ms
14  dr1-086lib-202.net.utah.edu (155.97.127.10)  69.156 ms  68.420 ms  68.069 ms
15  www.utah.edu (155.99.1.2)  68.638 ms  68.157 ms  67.489 ms
```

```
traceroute to www.LB-A.stanford.edu (171.64.14.235), 30 hops max, 38 byte packets
 1  ge_router (132.175.244.254)  0.667 ms  0.343 ms  0.340 ms
 2  sacr3009 (132.175.38.9)  0.443 ms  0.383 ms  0.454 ms
 3  sacr4102 (198.206.223.19)  164.205 ms  221.831 ms  2.087 ms
 4  alb-snla.es.net (134.55.208.125)  1.953 ms  1.763 ms  1.963 ms
 5  elp-alb.es.net (134.55.209.229)  7.362 ms  6.136 ms  6.063 ms
 6  snvcr1-oc48-elpcr1.es.net (134.55.209.217)  33.396 ms  33.319 ms  34.252 ms
 7  lbl-snv-oc48.es.net (134.55.209.6)  35.875 ms  35.978 ms  34.773 ms
 8  BERK--esnet.POS.calren2.net (198.32.249.57)  34.091 ms  34.277 ms  34.882 ms
 9  SUNV--BERK.POS.calren2.net (198.32.249.14)  36.369 ms  36.463 ms  36.008 ms
10  STAN--SUNV.POS.calren2.net (198.32.249.74)  35.931 ms  36.039 ms  36.741 ms
11  i2-gateway.Stanford.EDU (171.64.1.214)  36.395 ms  36.278 ms  37.104 ms
12  bbr2-rtr.Stanford.EDU (171.64.1.133)  37.504 ms  37.056 ms  36.860 ms
13  sweet-rtr.Stanford.EDU (171.64.1.166)  38.241 ms  37.434 ms  37.213 ms
14  www5.Stanford.EDU (171.64.14.235)  36.861 ms  37.970 ms  36.343 ms
```

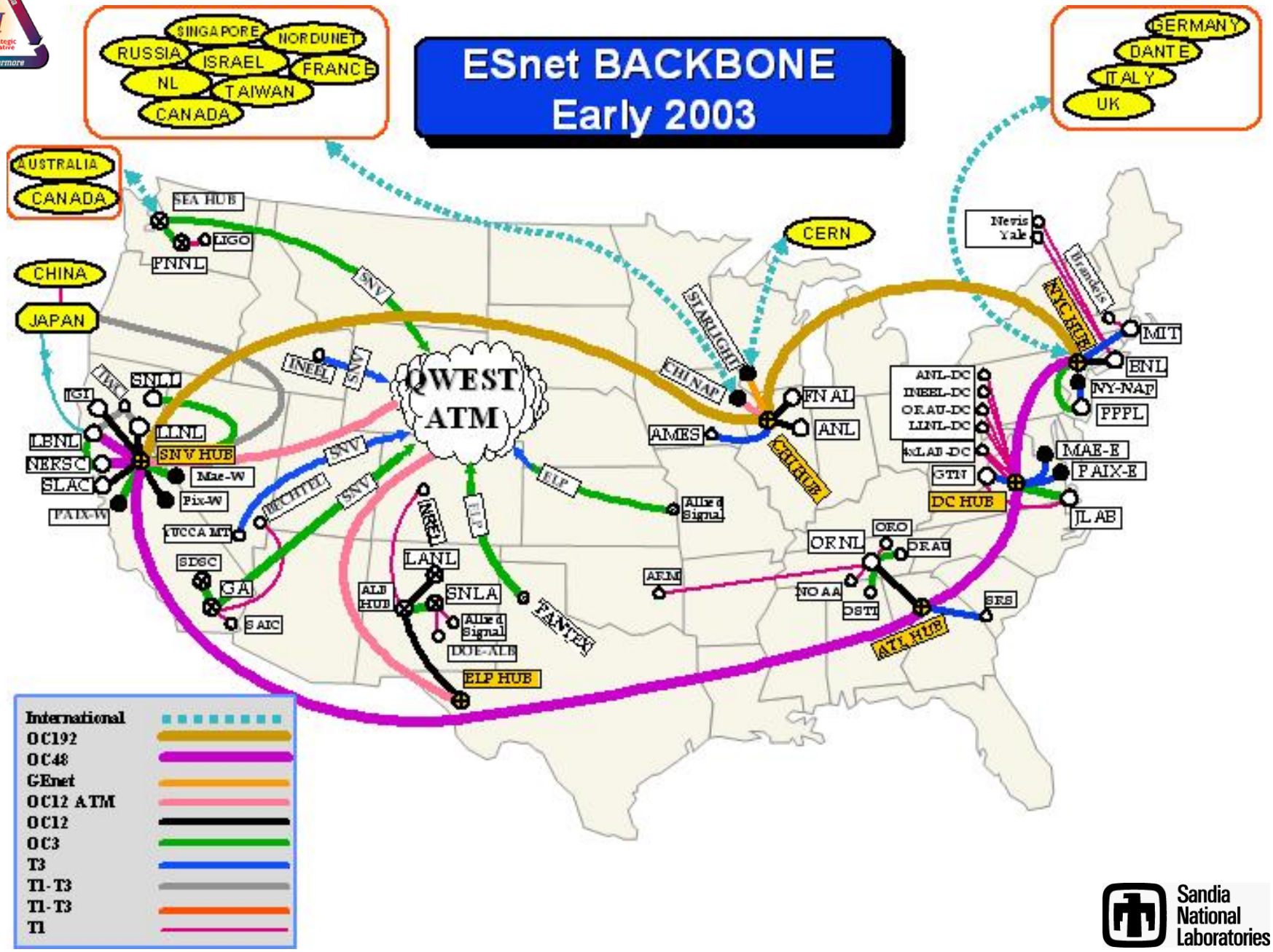
```
traceroute to www.caltech.edu (131.215.220.4), 30 hops max, 38 byte packets
 1  ge_router (132.175.244.254)  0.466 ms  0.343 ms  0.336 ms
 2  sacr3009 (132.175.38.9)  0.416 ms  0.290 ms  0.277 ms
 3  sacr4102 (198.206.223.19)  1.588 ms  1.511 ms  1.165 ms
 4  alb-snla.es.net (134.55.208.125)  2.586 ms  2.561 ms  1.680 ms
 5  elp-alb.es.net (134.55.209.229)  6.881 ms  6.841 ms  7.032 ms
 6  snvcr1-oc48-elpcr1.es.net (134.55.209.217)  33.579 ms  33.452 ms  33.699 ms
 7  snvrt1-ge0-snvcr1.es.net (134.55.209.90)  33.910 ms  33.305 ms  33.419 ms
 8  198.32.11.93 (198.32.11.93)  34.735 ms  33.505 ms  32.991 ms
 9  losang-snvang.abilene.ucaid.edu (198.32.8.94)  40.740 ms  40.582 ms  40.961 ms
10  hpr-lax-gsr1--abilene-LA-10ge.cenic.net (137.164.25.2)  40.816 ms  41.614 ms  40.950 ms
11  dc-lax-12410a--hpr-lax-12410-ge.cenic.net (137.164.22.12)  40.650 ms  41.148 ms  40.671 ms
12  c2-ucla--dc-lax-12410a-pos.cenic.net (137.164.22.57)  45.366 ms  45.199 ms  54.398 ms
13  JPL--UCLA.POS.calren2.net (198.32.248.2)  46.194 ms  45.435 ms  49.206 ms
14  CIT--JPL.POS.calren2.net (198.32.248.6)  47.910 ms  46.234 ms  46.385 ms
15  Caltech-CalREN.caltech.edu (192.41.208.50)  49.410 ms  56.628 ms  49.158 ms
16  Booth-RSM.ilan.caltech.edu (131.215.254.253)  46.694 ms  48.614 ms  45.705 ms
17  booboo.caltech.edu (131.215.220.3)  47.667 ms  47.075 ms  48.278 ms
```

```
[tjpratt@dellux sbin]$ ./traceroute www.uiuc.edu
traceroute to tarantula.cso.uiuc.edu (128.174.5.130), 30 hops max, 38 byte packets
 1  ge_router (132.175.244.254)  0.532 ms  0.354 ms  0.345 ms
 2  sacr3009 (132.175.38.9)  0.386 ms  0.298 ms  0.480 ms
 3  sacr4102 (198.206.223.19)  1.133 ms  1.777 ms  1.298 ms
 4  alb-snla.es.net (134.55.208.125)  1.676 ms  1.989 ms  2.239 ms
 5  elp-alb.es.net (134.55.209.229)  6.368 ms  6.333 ms  6.414 ms
 6  snvcr1-oc48-elpcr1.es.net (134.55.209.217)  33.558 ms  34.117 ms  32.749 ms
 7  chicr1-oc192-snvcr1.es.net (134.55.209.54)  81.442 ms  81.021 ms  82.194 ms
 8  chirt1-ge0-chicr1.es.net (134.55.209.190)  81.287 ms  81.305 ms  81.515 ms
 9  198.125.140.114 (198.125.140.114)  84.345 ms  84.472 ms  85.065 ms
10  uiuc-ibgp1-lnk.gw.uiuc.edu (128.174.0.185)  84.738 ms  85.659 ms  85.052 ms
11  uiuc-fwexit-net.gw.uiuc.edu (128.174.0.196)  85.060 ms  85.226 ms  87.138 ms
12  t-core1b.gw.uiuc.edu (130.126.0.185)  85.868 ms  88.200 ms  86.094 ms
13  tarantula.cso.uiuc.edu (128.174.5.130)  86.058 ms  85.860 ms  85.188 ms
```

```
[tjpratt@dellux sbin]$ ./traceroute www.uchicago.edu
traceroute to krypton.uchicago.edu (128.135.12.53), 30 hops max, 38 byte packets
 1  ge_router (132.175.244.254)  0.475 ms  0.345 ms  0.335 ms
 2  sacr3009 (132.175.38.9)  0.496 ms  0.324 ms  0.311 ms
 3  sacr4102 (198.206.223.19)  1.128 ms  1.468 ms  0.856 ms
 4  alb-snla.es.net (134.55.208.125)  2.410 ms  2.799 ms  1.999 ms
 5  elp-alb.es.net (134.55.209.229)  6.849 ms  6.844 ms  7.155 ms
 6  snvcr1-oc48-elpcr1.es.net (134.55.209.217)  33.596 ms  32.749 ms  32.719 ms
 7  chicr1-oc192-snvcr1.es.net (134.55.209.54)  81.694 ms  80.871 ms  81.301 ms
 8  chirt1-ge0-chicr1.es.net (134.55.209.190)  81.249 ms  81.395 ms  81.677 ms
 9  uofc-peer.es.net (198.125.140.150)  81.747 ms  81.970 ms  82.124 ms
10  msfc-1155-v919.uchicago.edu (128.135.249.129)  83.045 ms  82.843 ms  87.580 ms
11  krypton.uchicago.edu (128.135.12.53)  82.415 ms  81.933 ms  82.788 ms
```

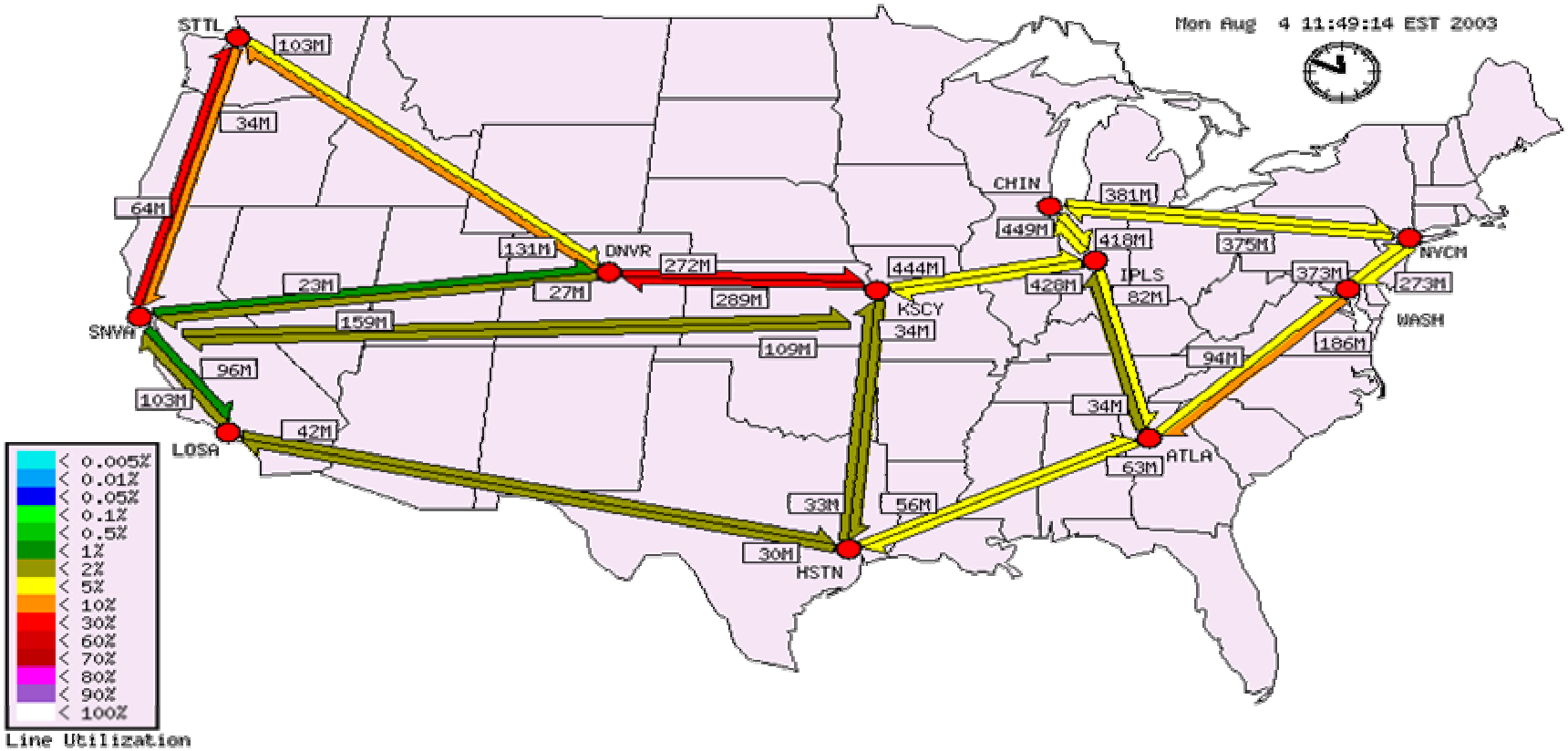

Uof Chic
UICU
UofUtah
CIT
Stanford
Chic
Utah
Calren2 south
Abilene
CALREN2
ESNET
Sandia
LANL
LLNL
Janus
Blue
Frost

ESnet BACKBONE Early 2003



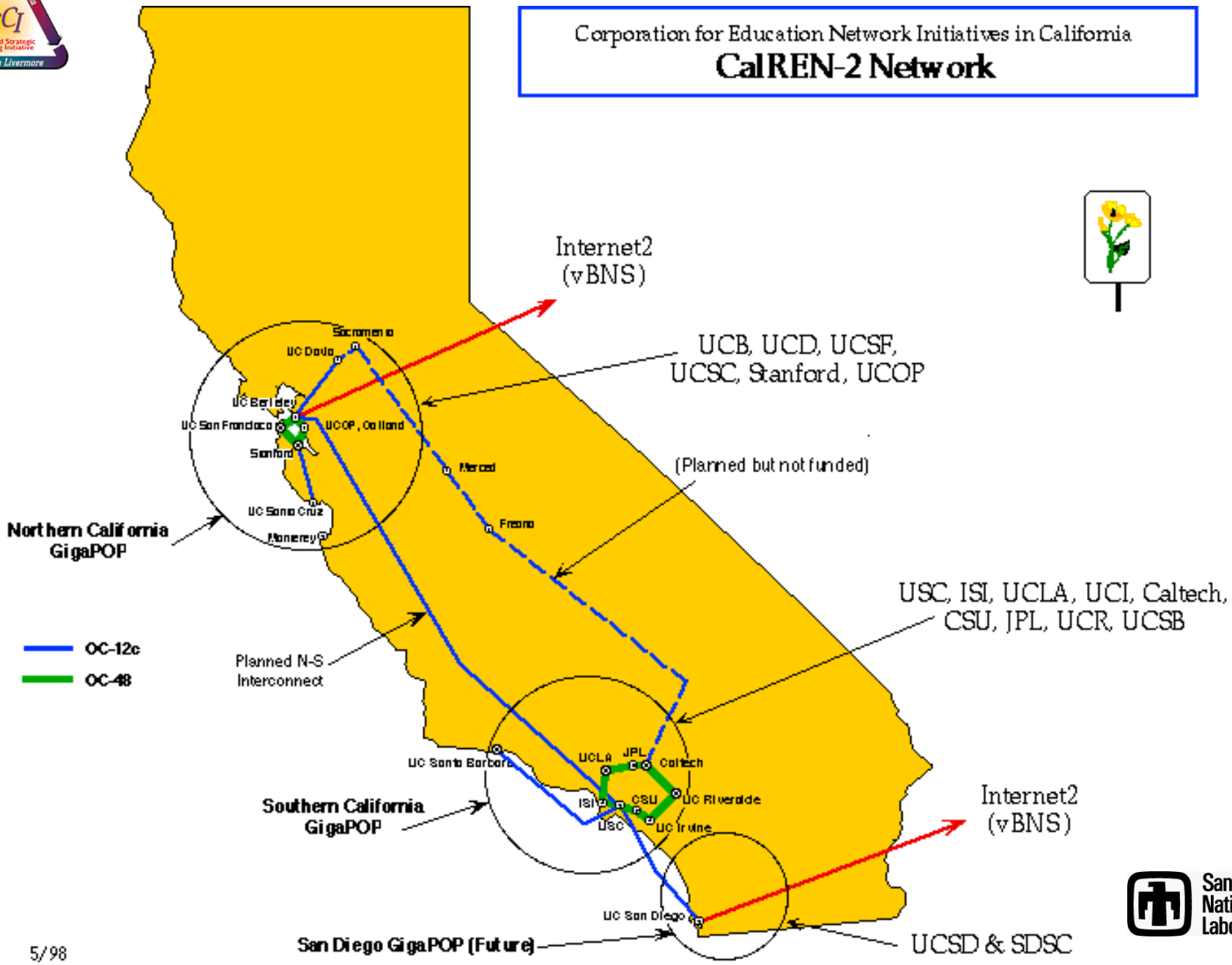
Sandia National Laboratories

INDIANA UNIVERSITY ABILENE NOC WEATHERMAP



Corporation for Education Network Initiatives in California

CalREN-2 Network



5/98

Sandia National Laboratories

Current Lab Networks

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

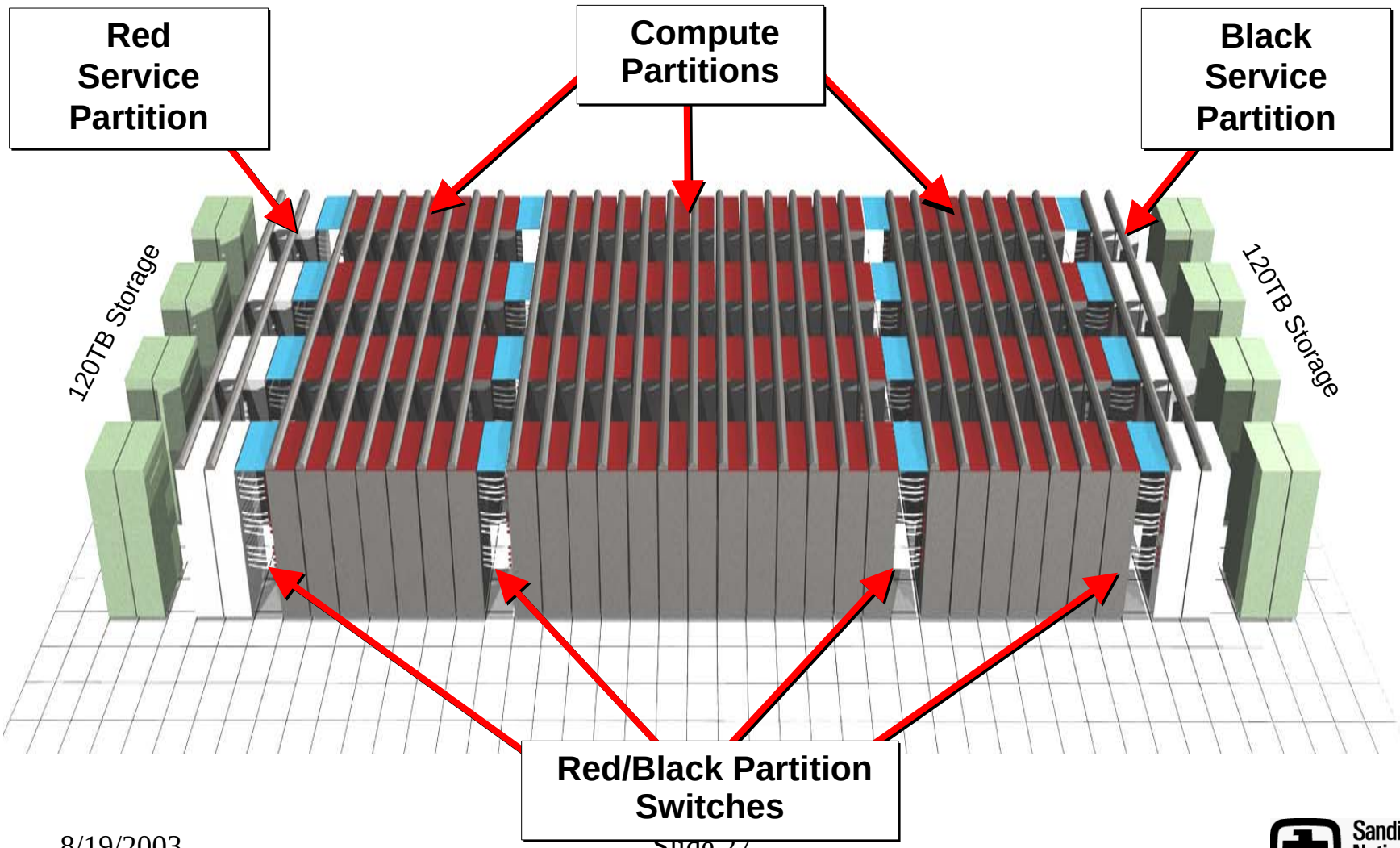
Future

Mahesh Rajan
Barbara Jennings

Futures: ASCI Red Storm Highlights

- $90 million contract to Cray Inc., for a new DOE supercomputer
- Joint Sandia/Cray design focusing on balanced performance of computation, communication, memory and I/O
- Uses AMD 64 Bit Opteron™ processor
- Custom Cray 3d mesh interconnect using Opteron's Hypertransport™
- LWK on compute nodes and full Linux on Service and I/O nodes
- Red/Black partitioning
- Special Red Storm Management System for hardware/software RAS
- Additional Data/Visualization Clusters (Red/Black)
  - Parallel rendering capability for remote visualization
  - 200 TB disk storage
  - 1024 processors each
- Delivery: June 2004 to Sep. 2004; Limited availability to users by Jan. 2005

System Configuration



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 ft² 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

Run-Time System

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

File Systems

- High performance file system
  - Lustre
- Interim file system
  - PVFS - Interim file system

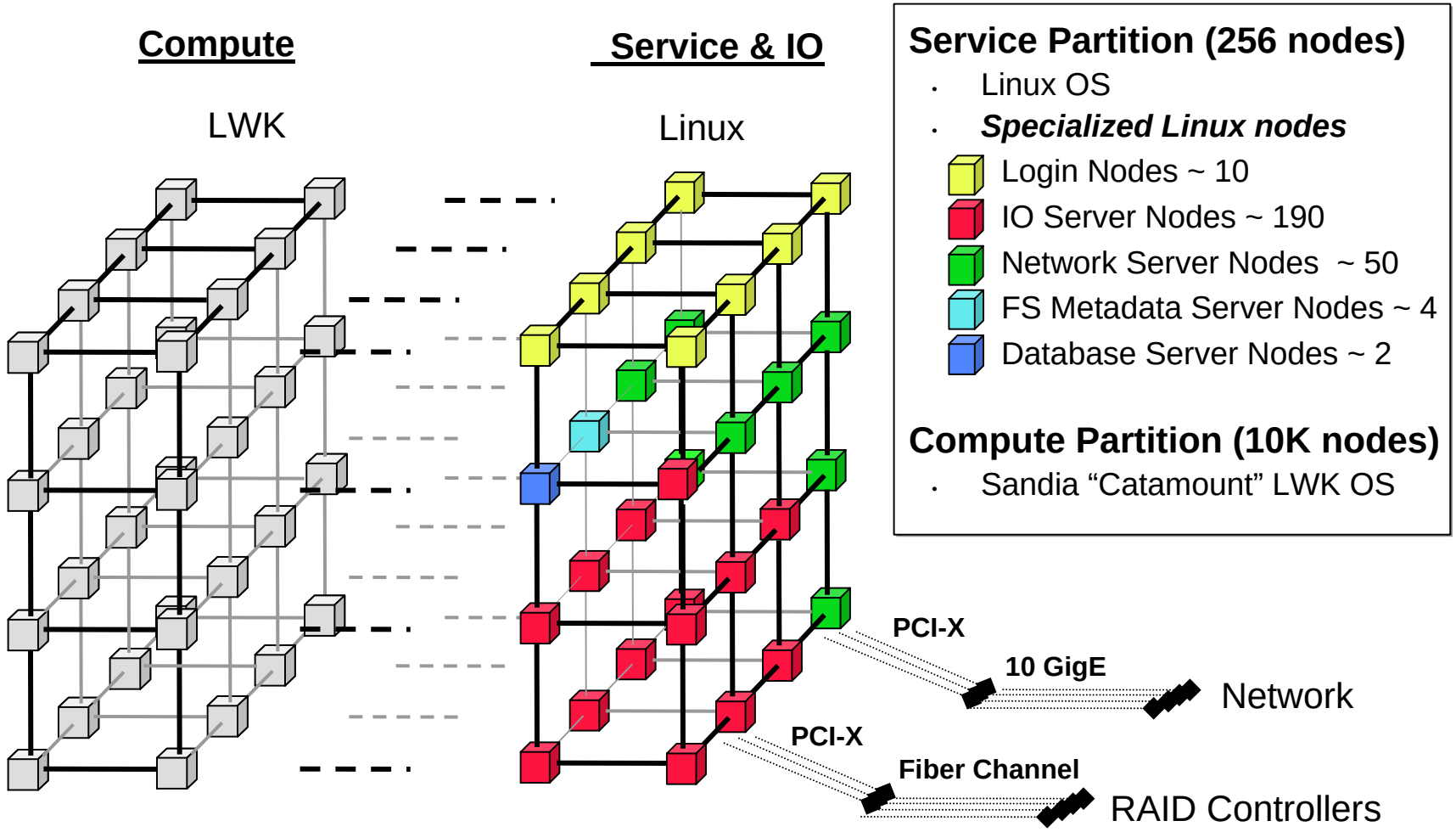
User Environment

- 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

Red Storm – Partitions, OS, Layout



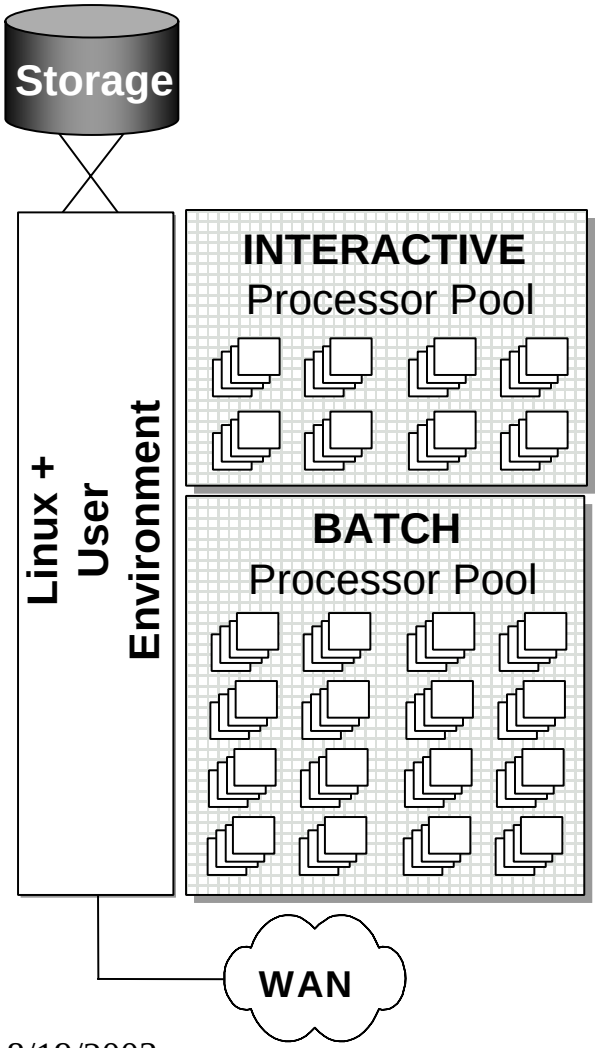
Service Partition (256 nodes)

- 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)

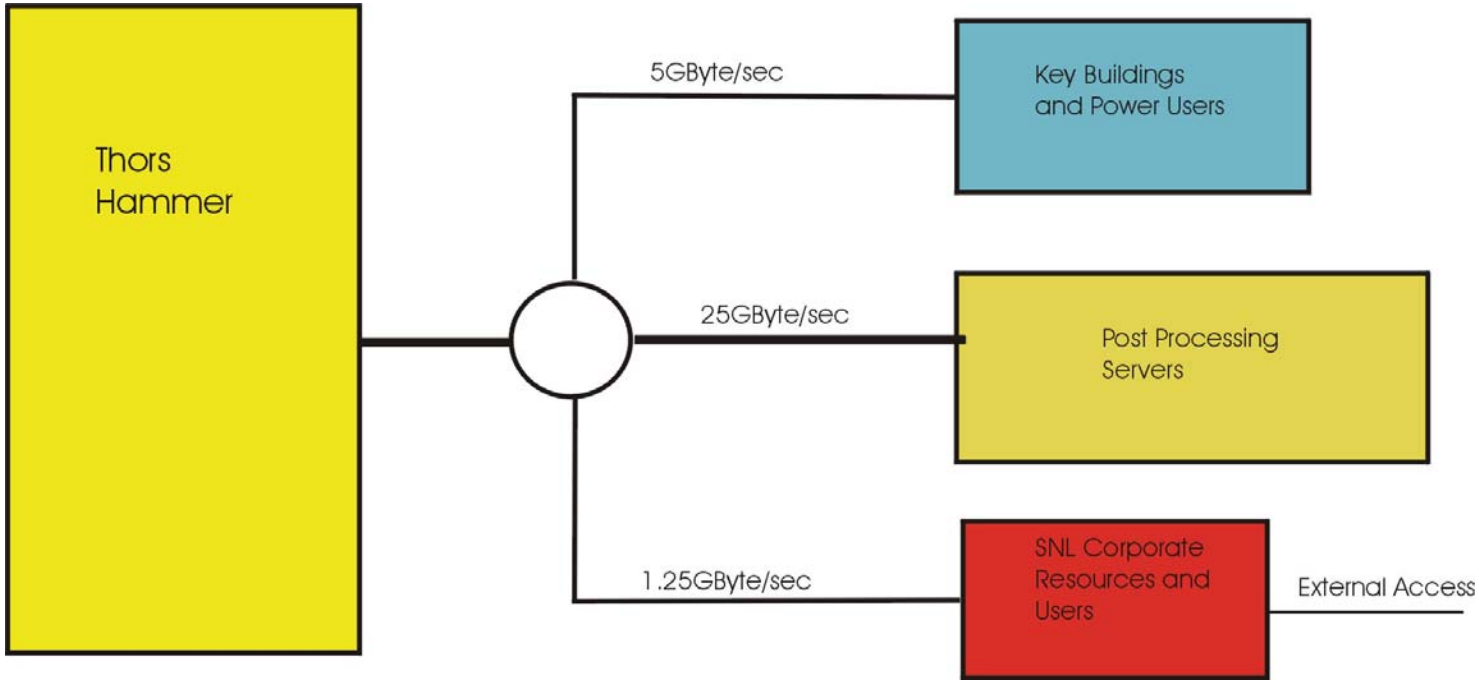
- Sandia "Catamount" LWK OS

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

Red Storm Network



Sandia Scientific Computing Support (SSCS)

Barbara Jennings, Lilia Martinez

- Focus
  - High Performance Computing Environment
- Getting Help
  - Experts on a Hotline
  - Access via phone, web, email and ARS trouble ticket
- Knowledge Database
  - Web based queries by experts and customers
- Web Pages
  - Consistent among the tri-lab community
- Training
  - Web based training

Synchronized Account Request Automated ProcEss (SARAPE)

SARAPE is accessed through Sandia's external WWW and is designed as a front-end to the existing Account Request Process at each of the Labs.

Goal:
Improve the current SecureNet Guest Account Request Form and Process

Driver:
Red Storm will bring additional Guest Account Requests from the NWC

- Phase I – 7/2003 – SNL Computer Account Requests
- Phase II – 8/2003 – SNL Applications/Services Requests
- Phase III – 10/2003 – Tri-Lab/NWC Resources Requests
- Phase IV – 12/2003 – Tri-Lab/NWC Account Delete and Notification Process

SecureNet Classified Guest Account Request Form

Use your browser's PRINT button to print this form.

Section I: GUEST INFORMATION
Guest Name: John DOe **Phone #:** 123-546-6789
Guest's Site Name: SNL/CA **Guest Org:** 9336 **SSN:** 000000000
Guest's Mailing Address:
1500 north
Unclassified E-mail Address: tjp@lall.edu **Mail Stop:** 11
Classified E-mail Address: **Citizenship:** french
Date of Birth: 12/23/56
Guest Site Classified DCE Username: tomflop **Guest Site Classified DCE Cell:** Sandia DCE
Primary Computer: ; **IP Address:** 000.000.00.0

DOE computers are for Official Use Only. Users are responsible for protecting information and passwords they control, avoiding waste, fraud, or abuse of computing resources; using only authorized software; and obeying DOE/Site security policies. Users should not have implicit or explicit expectation of privacy. Use of DOE computers is subject to monitoring and review by authorized personnel. Improper use may result in disciplinary or legal action.
User Signature: _____

HOST ACCESS INFORMATION:
Host Site Name: SNL/NM
Host Contact Name: John brown **Phone #:** 5058569876
Requested Resource(s): ASCI-Red, edison
Requested Username: tomtflop
LANL Resources only:
Z Number: **Project Name/Description:**
Sigma Levels: I don't know

Section II: APPROVALS

I certify that this person requires access to the classified ASCI Site/Resource(s) shown above, has a Q level clearance, and meets the sigma requirements specified above.
Guest Manager Name (print): _____
Signature: _____
Phone #: _____ **Date:** _____

Section III: Guest Processing Agency Authorization

I certify that the above information is true and correct.
GPA Name (print): _____ **Phone #:** _____
GPA Signature: _____ **Date:** _____

(https://sasr089.sandia.gov/cfdocs/prod/sarape/sarape.cfm)

SARAPE

Synchronized Account Request Automated ProcEss

- **You have reached the entry point for initiating a request for a computer account at Sandia National Laboratories. Authorized access is required. If you need assistance, please contact sscs@sandia.gov.**

Initiate Account Request

SARAPE

Sandia Account Request Automated ProcEss

Step 1 of 4. Initiate the account request process by entering your personal information.

*First Name: Rhonda MI: H *Last Name: MacIntyre

*Phone: 555-555-5555 Ext: (Use 555-555-5555 format.)

*E-mail address: rhn@y12.doe.gov

*U.S. Citizen? Yes No

*Social Security Number:

*Reenter to Verify SSN:

*Requestor's site: NWC Subsite (if app.): Y-12

*Request is to: Open a New Account Delete an Existing Account

Notice: DOE computers are for Official Use Only. Users are responsible for protecting information and passwords they control; avoiding waste, fraud, or abuse of computing resources; using only authorized software; and obeying DOE/site security policies. Users should not have implicit or explicit expectations of privacy. Use of DOE computers is subject to monitoring and review by authorized personnel. Improper use may result in disciplinary or legal action.

Initiate Request

Choices:
ASCI Alliance Partners
LANL
LLNL
NWC
SNL/NM
SNL/CA

ASCI Alliance Partners Submenu:
Stanford
Cal Tech
U. Of Chicago
U of Ill Urbana/Champagne
U of Utah

NWC Submenu:
DOE/Albuq
DOE/HQ
Kansas City Plant
Pantex
Savannah River
Y-12