LANL Computing Systems Update

Harvey Wasserman

Robert Cunningham

HPC Systems Group (CCN-7)

September, 2005

LA-UR-05-6800

LANL Resources Available To Alliance Users

- QSC:
  - Tru64 HP/Compaq system, 256 nodes, 2.5-TF peak
  - Alliances allocated 1/4 of *devq* + ~40% of *smallq* + ~40% of *largeq*, evenly split amonst all 5
  - Usage: Jan: 27.8%; Feb: 24.6%; Mar: 31.2%; Apr 36.1%
    May: 28.5%; Jun: 34.6%; Jul: 9.0%; Aug: 28.1%
  - Total machine usage ranges 65% - 86% over that period.
  - Other users: ADWP (24%); LANL Institutional Computing (33%)
- Likely that if any additional resources are made available they will be Linux+BProc (non-Q) based.
  - Possible trade of some time with Pink (Xeon/Myrinet).
  - Possibly some additional resources on Flash Gordon (Opteron/Myrinet).

Some Critical Directions in LANL HPC

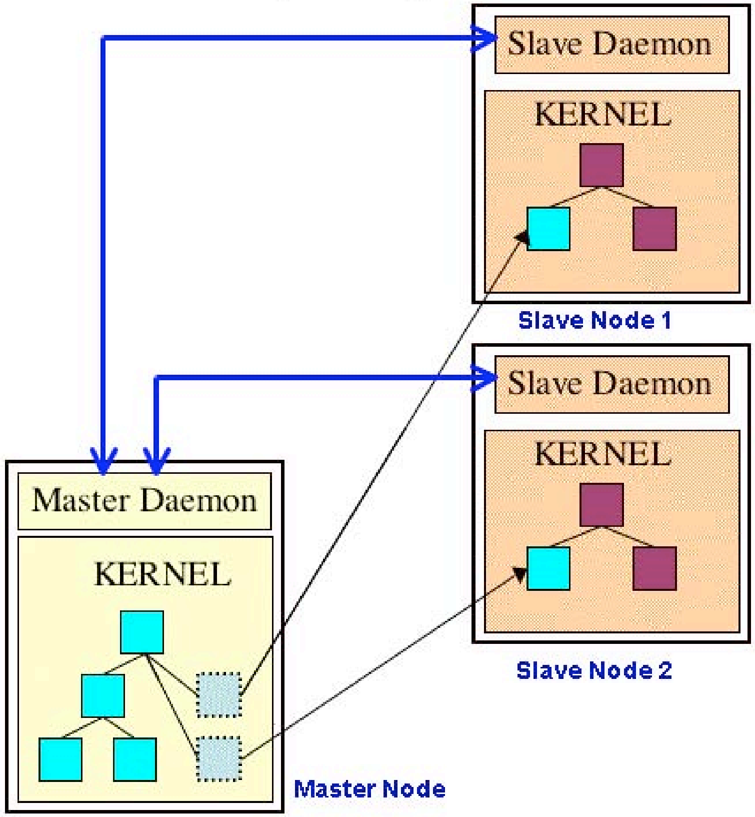
- All users:
  - BProc ("Beowolf distributed Process Space")
  - PaScalBB
  - Capacity computing (Tri-Lab direction)
  - LANL role in system integration
- External users: Turquoise network
  - Designed to enhance collaboration between LANL and external scientific institutions.
  - SSH access via proxy.
  - No export-controlled source or data.
  - Archival storage scheme via TSM (soon). No HPSS.
  - Training highly recommended for new users; issues include file transfer, X, directories, filesystems, etc.

LANL BProc Resources Growing Dramatically

- Flash (Opteron, Myrinet):
  - Currently one segment, 256 slave nodes, one master node, 3-TF
  - Soon three segments, 1@128 nodes, 2@256 slave nodes, one master node each, 6.1 TF total.
- Lightning (Opteron, Myrinet):
  - Currently 5 segments, 255 slave nodes, one master, node each, 11 TF
  - Soon 8 segments, 2048 total nodes, 21 TF total.
- TLC: Turquoise Linux Cluster, 110 Opteron nodes/Myr.
- Grendels, too (Xeon/Myr). Maybe Coyote.
- All this in addition to Pink (Xeon/Myr): 958 nodes

What's All This About BProc?

Process Tree Spanning 3 Machines



- BProc enables a distributed process space across nodes within a cluster.
- Users create processes on the *master node*. The system migrates the processes to the *slave nodes* but they appear as processes running on the master node.
- Stdin, stdout, & stderr are redirected to/from master node.
- R&D100 Award, 2004.Primary goal: High-availability cluster computing environment by making systems easier to build and manage – do more with available resources.

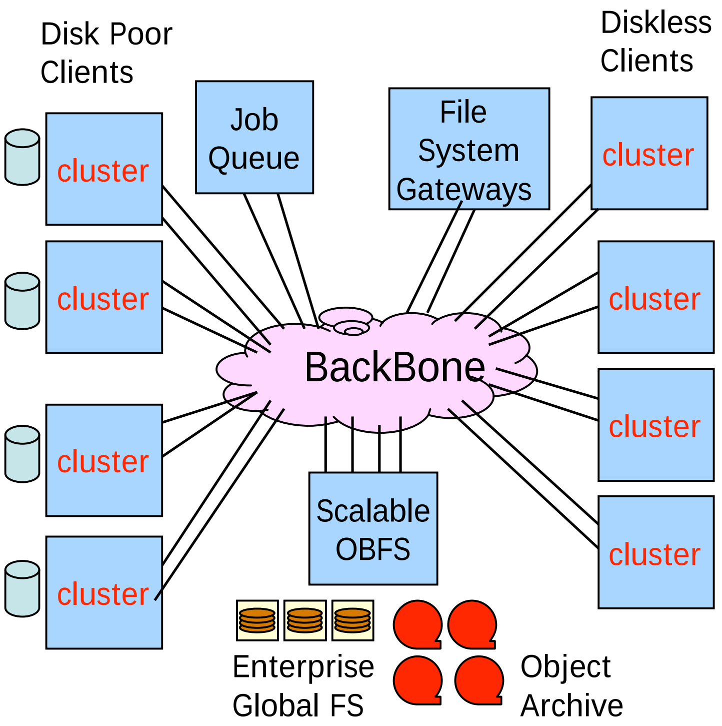
BProc and the User (1 of 2)

- Addition of compile/front-end nodes:
  - Do not **`llogin`** before compiling
- Slight change in how codes are run:
  - `bpsh $NODES a.out.serial`
  - `mpirun -np # a.out.parallel`
- LSF gives you an allocation of slave nodes but your shell is on the master node.
- New modulefile naming scheme/usage:
  - Consistency checking between modulefiles; can't load more than 1 from a given group.

BProc and the User (2 of 2)

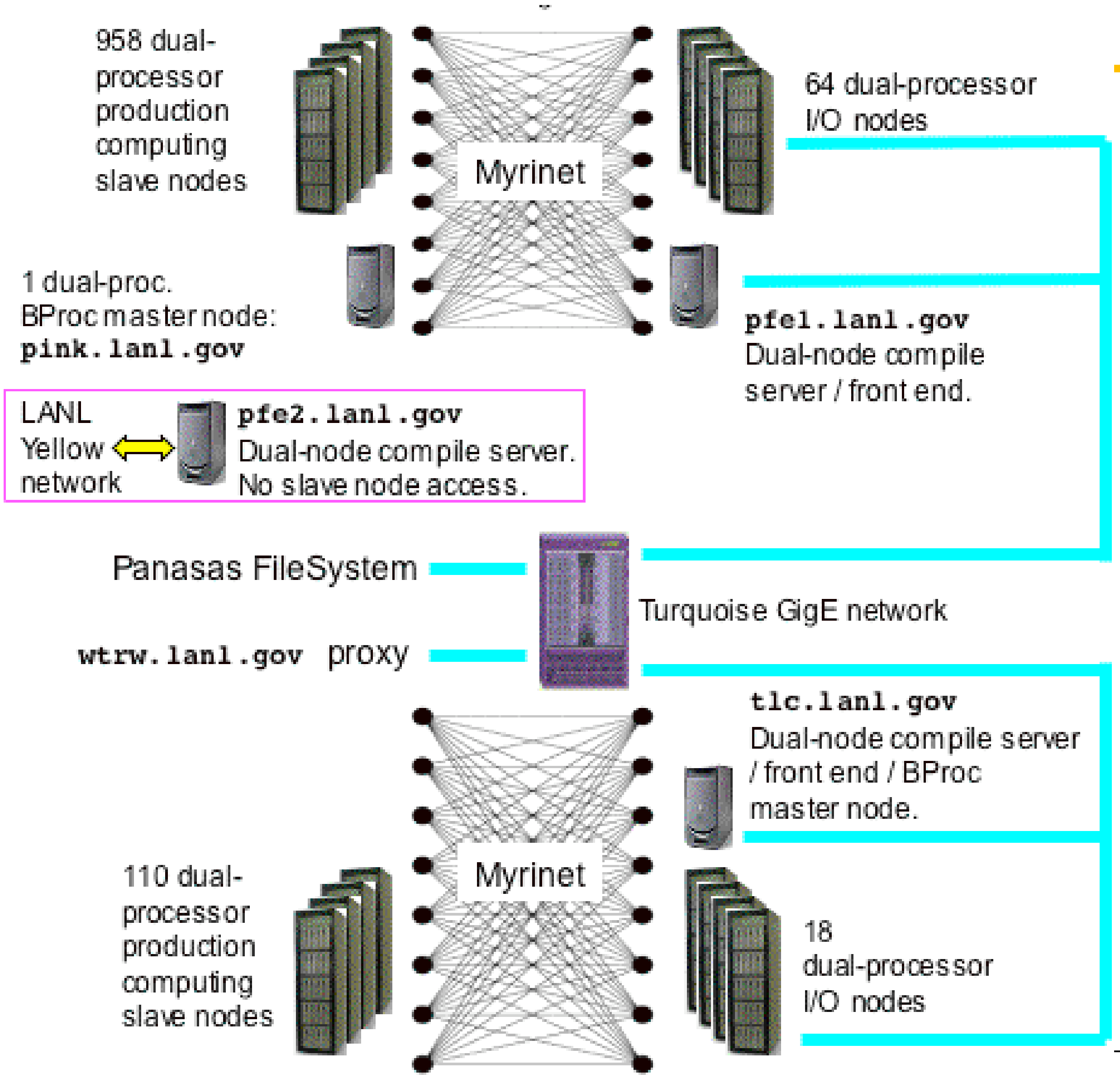
- Primary support for LAMPI; MPICH maybe not at all.
- PGI, PathScale, Intel compilers.
- Some new status commands: `bpps, bpstat, bptop`
  - Must use `llogin` in order to use them.
- TotalView works for serial and parallel; can initiate or attach to running jobs.
- All LANL BProc systems currently operating in 32-bit (Opteron "Legacy") mode.
  - 64 bit computing with Fedora Core 3 (2.6.11 kernel), MPI, LSF, Bproc, and Panasas support someday.

Parallel Scalable Back Bone (PaScalBB)



- Multiple clusters sharing large, global namespace parallel I/O subcluster
  - Examples are Pink/TLC, Flash/Gordon, and Lightning/Bolt
- Network is combination of HPC Interconnect + commodity networking bridge
- Panasas
- I/O through a set of fileserver nodes over Myrinet; nodes serve as Myrinet<->GigE routers. Works on Lightning, Pink, & Flash now.

Pink / TLC Configuration



958 dual-processor production computing slave nodes

Myrinet

64 dual-processor I/O nodes

1 dual-proc. BProc master node: **pink.lanl.gov**

pfe1.lanl.gov Dual-node compile server / front end.

LANL Yellow network

pfe2.lanl.gov Dual-node compile server. No slave node access.

Panasas FileSystem

wtrw.lanl.gov proxy

Turquoise GigE network

tlc.lanl.gov Dual-node compile server / front end / BProc master node.

110 dual-processor production computing slave nodes

Myrinet

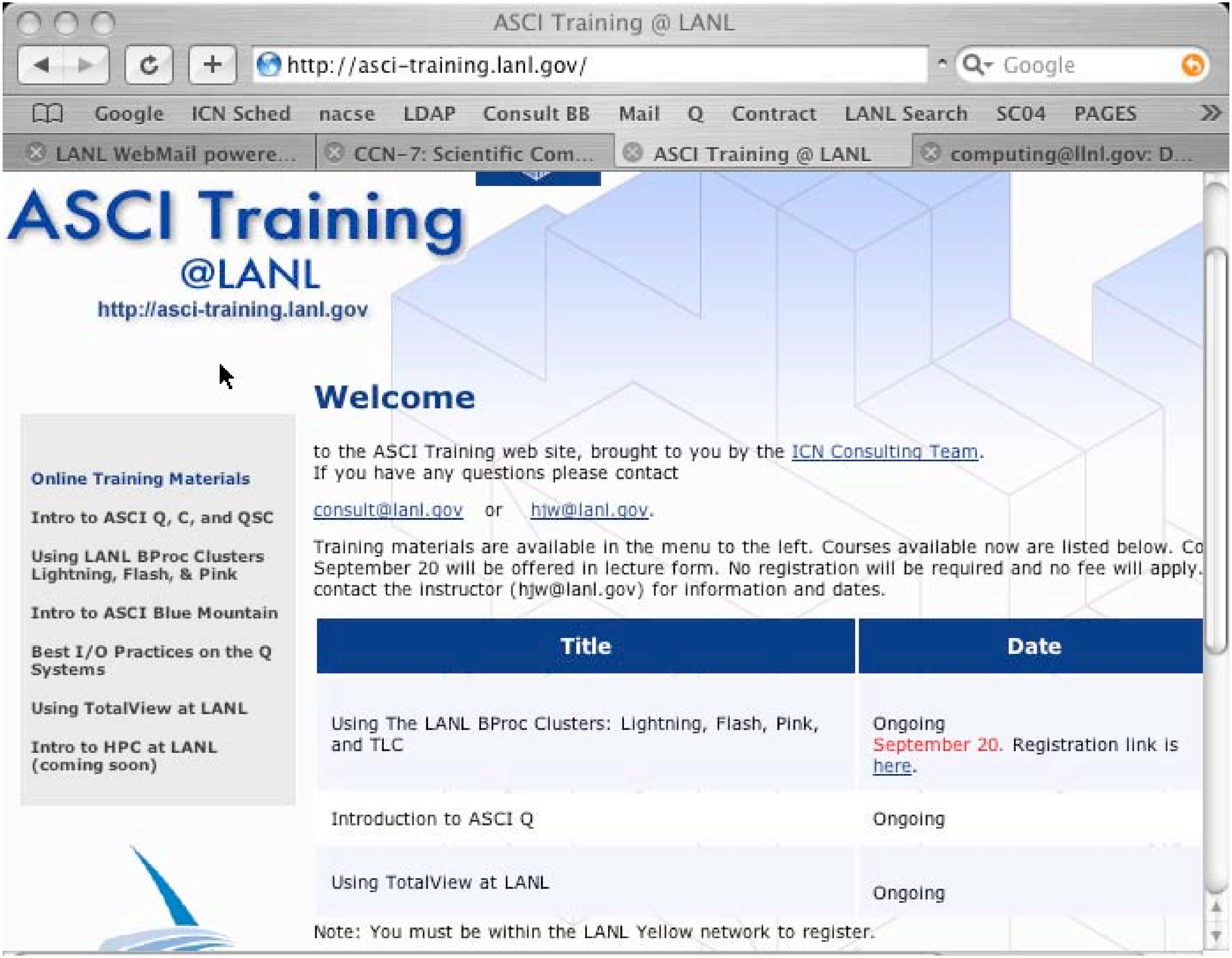
18 dual-processor I/O nodes

Other Recent Changes

- Cheryl Wampler: Was Group Leader of HPC Systems; now Program Manager for Production Computing Systems.
- Steve Shaw: Was Compaq/HP Program manager; now Group Leader, HPC Systems.
- Consultants: Jeff Johnson (TL), Sara Hoshizaki, Roger Martz, David Kratzer, Meghan Quist, Robert Derrick. Hal Marshall now "matrixed" to code teams. Rob Cunningham now TL SCR Services.
- New HPC Systems Integration group, Gary Grider, GL
  - Teams: I/O, SSC, Advanced Architectures

3 LANL Web Sites You Can't Live Without

- `http://computing.lanl.gov` Main documentation site
- `http://icnn.lanl.gov` Machine Status
- `http://asci-training.lanl.gov` HPC training classes



ASCI Training

@LANL

http://asci-training.lanl.gov

Online Training Materials

Intro to ASCI Q, C, and QSC

Using LANL BProc Clusters Lightning, Flash, & Pink

Intro to ASCI Blue Mountain

Best I/O Practices on the Q Systems

Using TotalView at LANL

Intro to HPC at LANL (coming soon)

Welcome

to the ASCI Training web site, brought to you by the ICN Consulting Team.
If you have any questions please contact

consult@lanl.gov or hjw@lanl.gov.

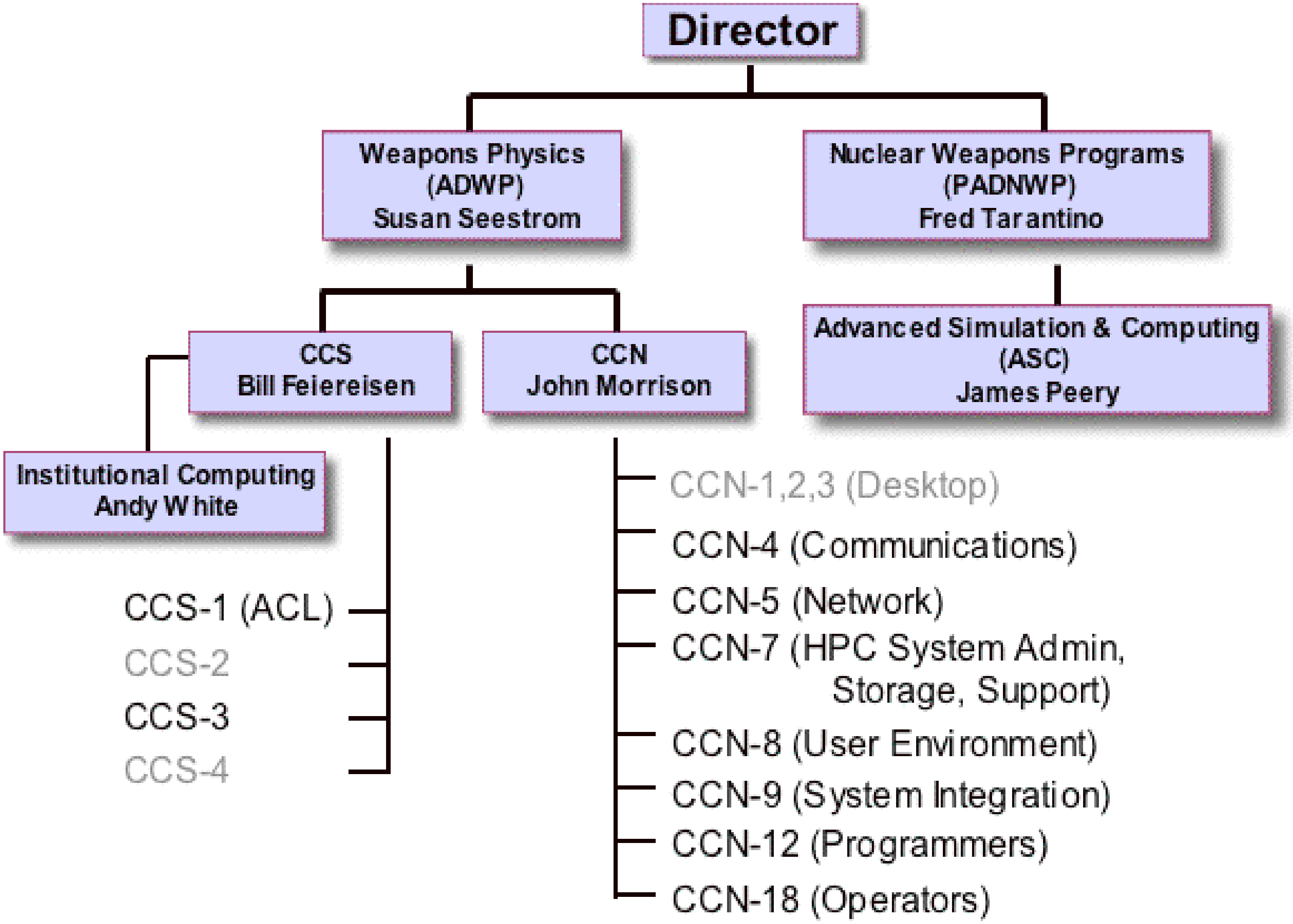
Training materials are available in the menu to the left. Courses available now are listed below. Co
September 20 will be offered in lecture form. No registration will be required and no fee will apply.
contact the instructor (hjw@lanl.gov) for information and dates.

Title	Date
Using The LANL BProc Clusters: Lightning, Flash, Pink, and TLC	Ongoing September 20. Registration link is here.
Introduction to ASCI Q	Ongoing
Using TotalView at LANL	Ongoing

Note: You must be within the LANL Yellow network to register.

Backup Slides

LANL HPC Organization



Director

Weapons Physics (ADWP) Susan Seestrom

Nuclear Weapons Programs (PADNWP) Fred Tarantino

CCS Bill Feiereisen

CCN John Morrison

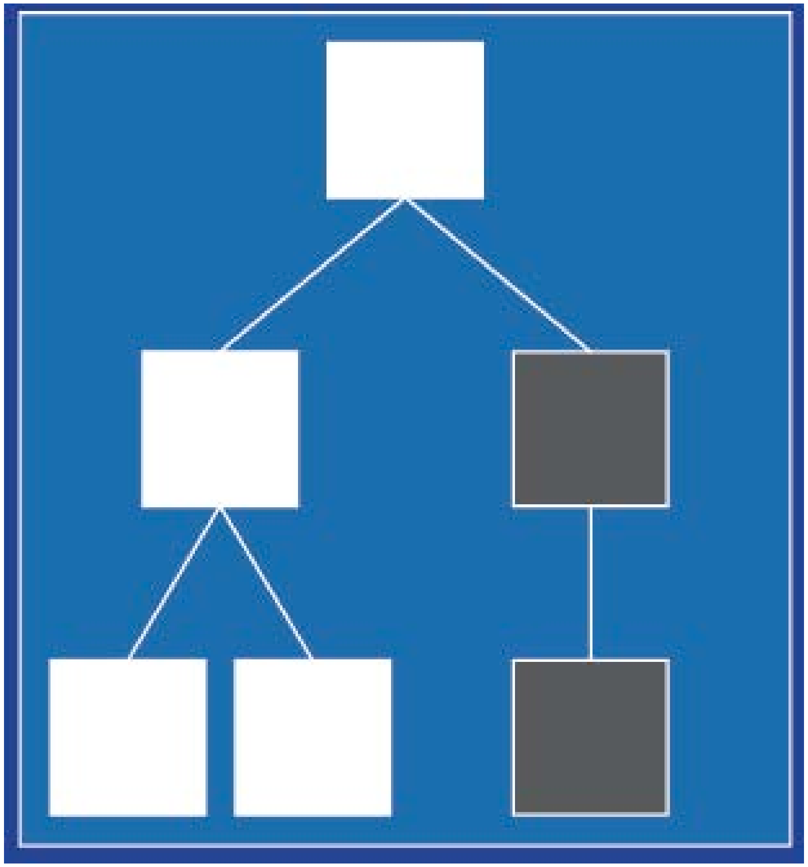
Advanced Simulation & Computing (ASC) James Peery

Institutional Computing Andy White

- CCS-1 (ACL)
- CCS-2
- CCS-3
- CCS-4

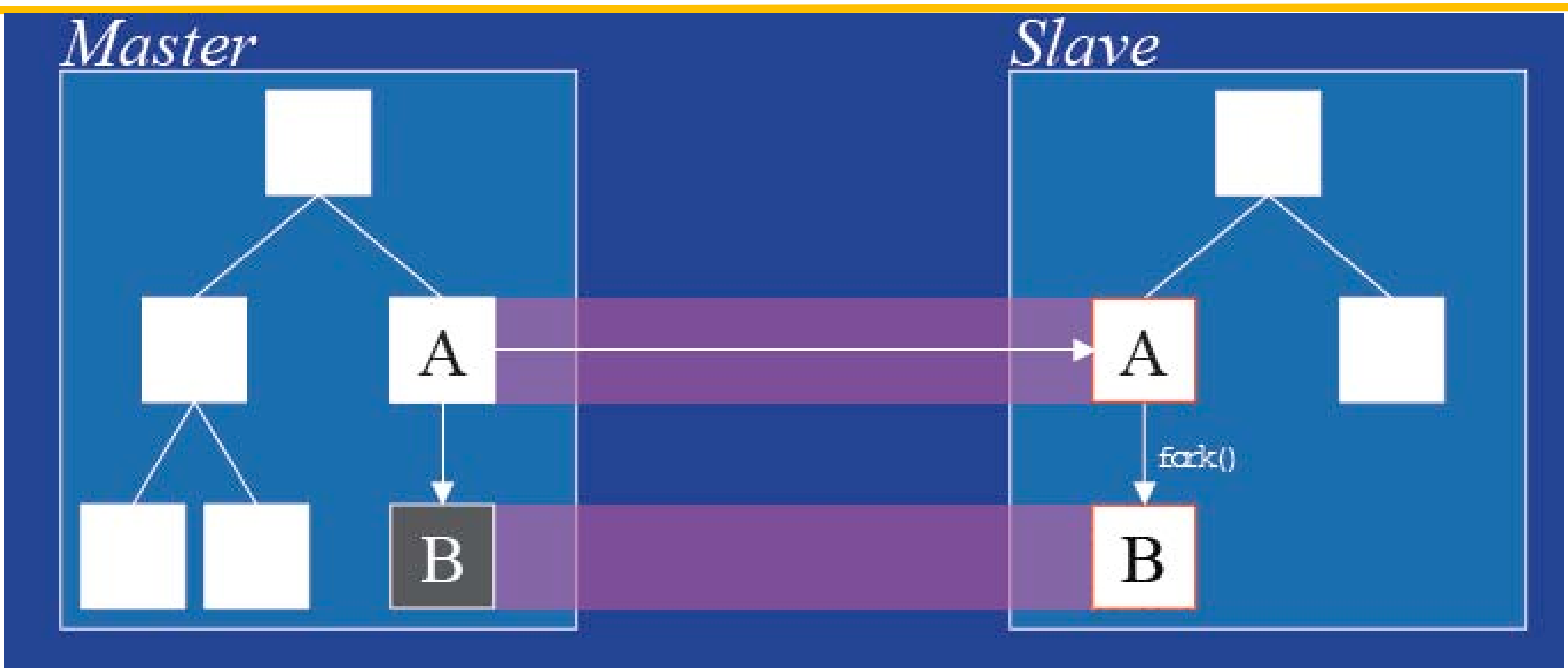
- CCN-1,2,3 (Desktop)
- CCN-4 (Communications)
- CCN-5 (Network)
- CCN-7 (HPC System Admin, Storage, Support)
- CCN-8 (User Environment)
- CCN-9 (System Integration)
- CCN-12 (Programmers)
- CCN-18 (Operators)

BProc, the Heart of Clustermatic



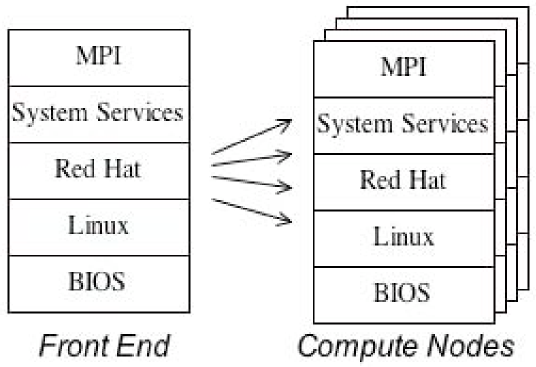
- Bproc = Beowulf Distributed Process Space
- Process Space
  - A pool of process id's
  - A process tree (parent/child relationships)
  - Every instance of a Linux kernel has a process space
- A distributed process space allow parts of a node's process space to exist on another node

Process Creation In BProc

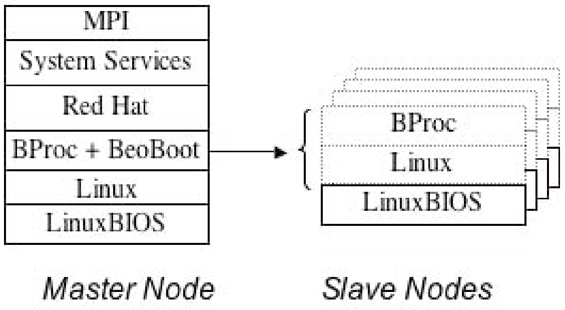


- Process on Master migrates to slave node (1.9s 16MB process on 1024 nodes)
- Process A, on slave, calls fork() to create child process B
- New Place holder for B is created on A (Ghost)
- Not all processes on slave node appear on master space

Science Appliance vs. a Traditional Cluster



Traditional Cluster Architecture



Science Appliance Architecture

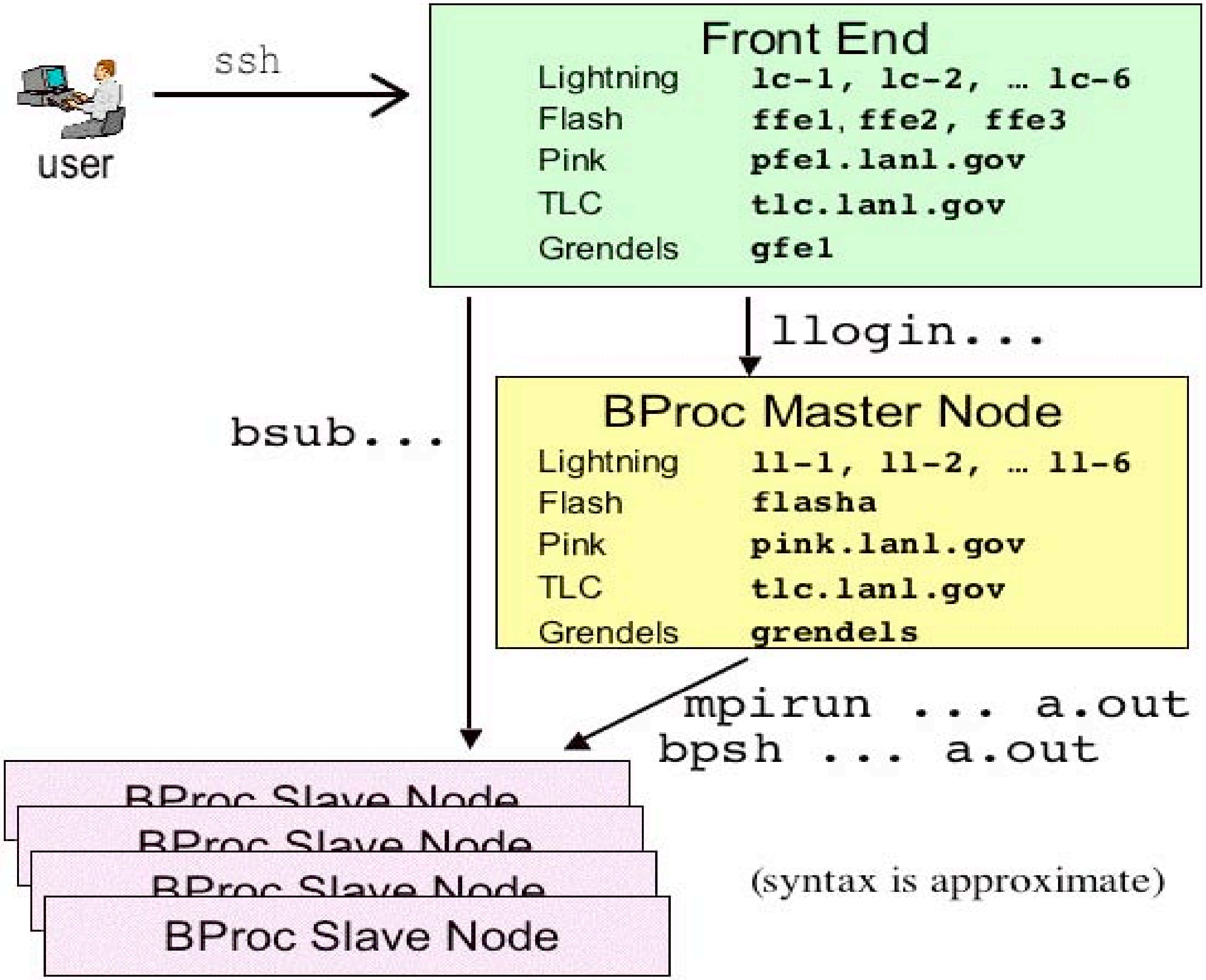
- A traditional cluster is built by replicating a complete workstation's software environment on every node.

- In a Science Appliance, we have master nodes and slave nodes but only the master nodes have a fully-configured system.

- The slave nodes run a minimal software stack consisting of LinuxBIOS, Linux, and BProc.

- No Unix shells running on the slave nodes, no user logins on the slave nodes.

Running Jobs on BProc Systems



user

ssh

Front End

Lightning	**lc-1, lc-2, … lc-6**
Flash	**ffe1, ffe2, ffe3**
Pink	**pfe1.lanl.gov**
TLC	**tlc.lanl.gov**
Grendels	**gfe1**

llogin...

BProc Master Node

Lightning	**ll-1, ll-2, … ll-6**
Flash	**flasha**
Pink	**pink.lanl.gov**
TLC	**tlc.lanl.gov**
Grendels	**grendels**

bsub...

mpirun ... a.out
bpsh ... a.out

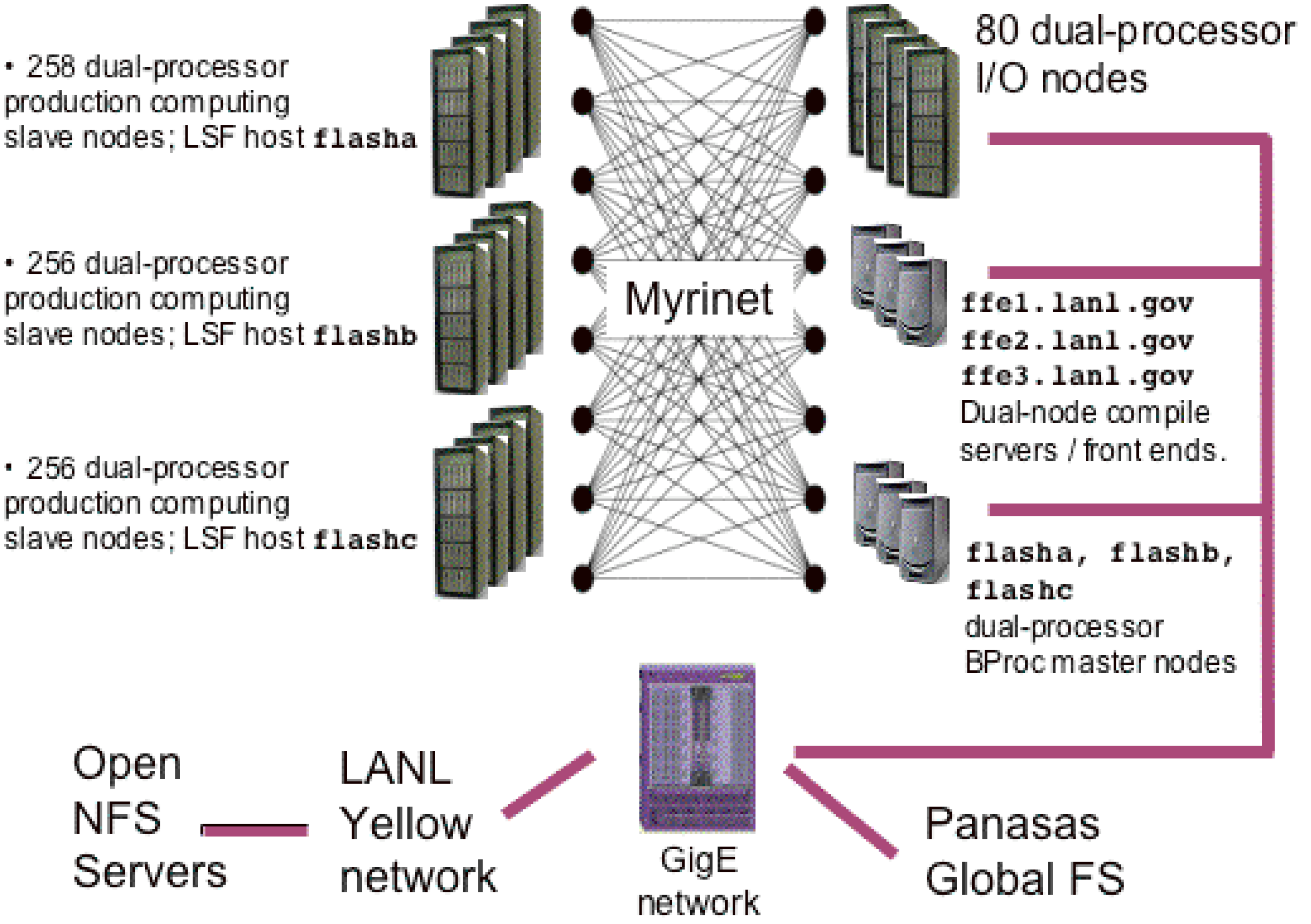
BProc Slave Node
BProc Slave Node
BProc Slave Node
BProc Slave Node

(syntax is approximate)

Debugging on BProc Systems

- **Debugging a Serial Job With TotalView**
  - `llogin`
  - **`module load totalview/`*`version`***
  - `totalview -remote $NODES ./a.out`
  - Dive on the executable name in the "root window." This will bring up the TotalView "process window."
- **Debugging an MPI Job With TotalView**
  - `llogin -n #`
  - **`module load lampi totalview/`*`version`***
  - `totalview mpirun -a -np # ./a.out`

Detailed Flash Configuration-to-Be



- 258 dual-processor production computing slave nodes; LSF host **flasha**
- 256 dual-processor production computing slave nodes; LSF host **flashb**
- 256 dual-processor production computing slave nodes; LSF host **flashc**

Myrinet

80 dual-processor I/O nodes

ffe1.lanl.gov
ffe2.lanl.gov
ffe3.lanl.gov
Dual-node compile servers / front ends.

flasha, flashb, flashc
dual-processor BProc master nodes

Open NFS Servers

LANL Yellow network

GigE network

Panasas Global FS

Abstract

- This talk will be presented to users at the five ASC Alliance university sites (CalTech, Stanford, U.Utah, U.Illinois, U. Chicago) as part of an annual update on hardware and user environment changes in the LANL high-performance computing area. The talk focuses on the new "BProc" Linux clusters and changes in the I/O architecture of several clusters.