LANL Computing Environment Update

Robert Cunningham

rtc@lanl.gov

HPC Systems Group (HPC-3)

October, 2006

LANL Resources Available To Alliance Users

- QSC is the workhorse
  - Alliances have 40%,Tru64 cluster, 256 nodes, 2.5 TF peak
  - Usage: Jan: 38.1%; Feb: 19.4%; Mar: 26.7%; Apr: 39.4% May: 33.4%; Jun: 33.0%; Jul: 36.8%; Aug: 43.5%, Sep: 23.8%
  - Past its peak reliability, expensive to maintain
  - Out the door soon (contract expires Dec. 1)
- Additional resources are Linux+BProc based
  - Alliances have 10% of Flash, Opteron/Myrinet, 8.6 TF
  - Possible trade for some time on Coyote (Opteron/IB, 13.5 TF)

LANL HPC Environment Topics

- New batch scheduler on the way: Moab
- Bproc limitations: no more new Bproc clusters
- The schizophrenic computer center: here are resources for you, but you can't use them
  - VPN access
  - Account requests
  - File transfers
- As if VPN wasn't enough trouble for you -- it must now run on government-owned workstation (Sep '06)

VPN Alternatives

- Distribute government-owned computers or disks (or diskless)
- Dial-in, extend yellow network, run from workstations at LANL
- Move to Turquoise network (no VPN requirement)
  - Move 10% of Flash cluster there
  - Swap cycles between Flash and Coyote
- Obtain waiver
  - Remote sysadmin using LANL OCSR
  - Accredited University security
  - Submit to audit/review?
  - Leads to a discussion with your security people

Security Topics for Waiver

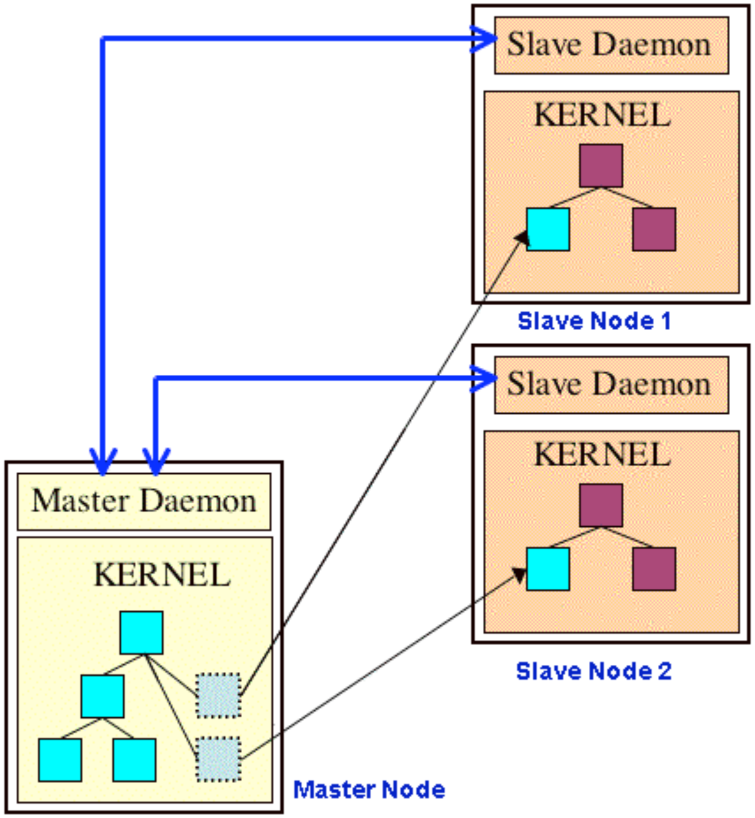
- What kind of security (IPS, IDS)?
- Written security plan?
- How are systems/network configured?
  - Meet LANL standards?
  - Does LANL meet your standards?
  - Network separation/layers/hierarchy?
- What if something happens?
  - Network scanning from LANL?
  - Submit to audit/review?
  - Without explanation, LANL can confiscate a computer!

LANL BProc Resources

- Flash (Opteron, Myrinet):
  - Five segments, 953 compute nodes, 8.6 TF
  - 8GB per node
  - Soon converted to 64-bit addressing (ABI)
  - PaScalBB for I/O infrastructure
- Lightning (Opteron, Myrinet):
  - Currently 13 segments, 7,140 compute nodes, 30.6 TF
- TLC: Turquoise Linux Cluster, 110 Opteron nodes with Myrinet
- Grendels: (Xeon/Myrinet), 126 nodes.
- Older platform: Pink (Xeon/Myr), 958 nodes, 9.2 TF
- Newest cluster: Coyote (Opteron/IB)
  - Five segments, 1275 compute nodes, 13.5 TF
  - 8GB per node

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.

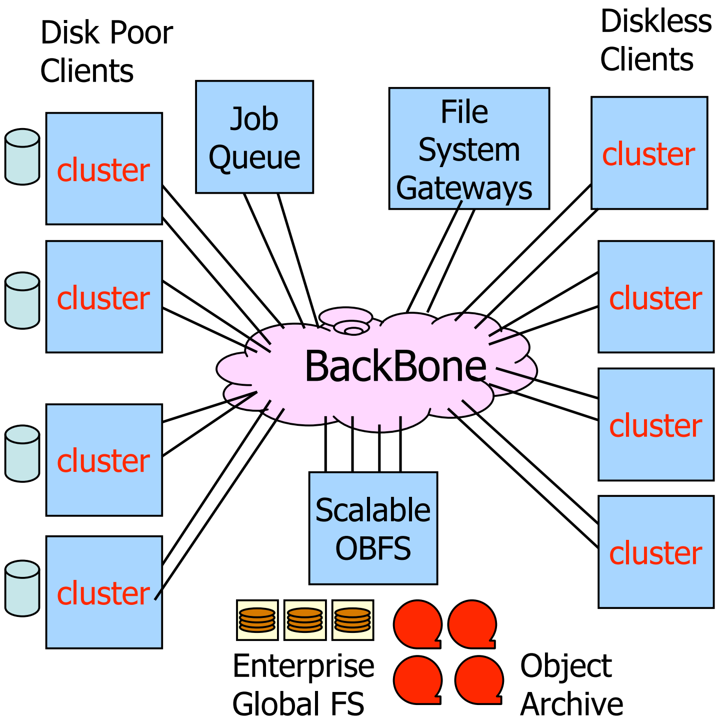
BProc and the User (1 of 2)

- Start with 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. Shell emulation on slave.
- New modulefile naming scheme/usage:
  - Consistency checking between modulefiles; can't load more than one from a given group.

BProc and the User (2 of 2)

- Primary support for LAMPI/OpenMPI
- PGI, PathScale, Intel compilers (others will fall behind).
- 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.
- Most LANL BProc systems currently converting to 64-bit addressing (ABI)
  - 64 bit computing with Fedora Core 3 (2.6.11 kernel), MPI, LSF, Bproc, and Panasas support.

Parallel Scalable Back Bone (PaScalBB)



- Relieve the master node
- Multiple clusters sharing large, global namespace parallel I/O subcluster
  - Examples are Pink/TLC/Coyote, Flash, and Lightning
- 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.

3 LANL Web Sites You Can't Live Without

- `http://computing.lanl.gov` Main documentation

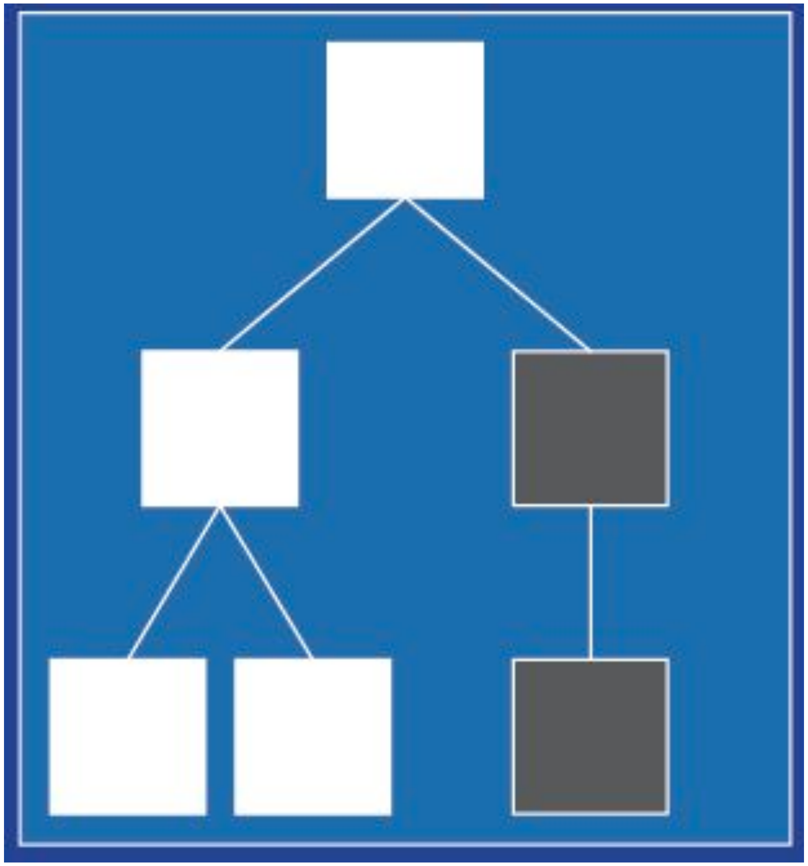
  (or call 505-665-4444 option 3, consult@lanl.gov)
- `http://icnn.lanl.gov` Machine Status
- `http://asci-training.lanl.gov` HPC training

HPC Accounts

Don’t forget Photo Op!

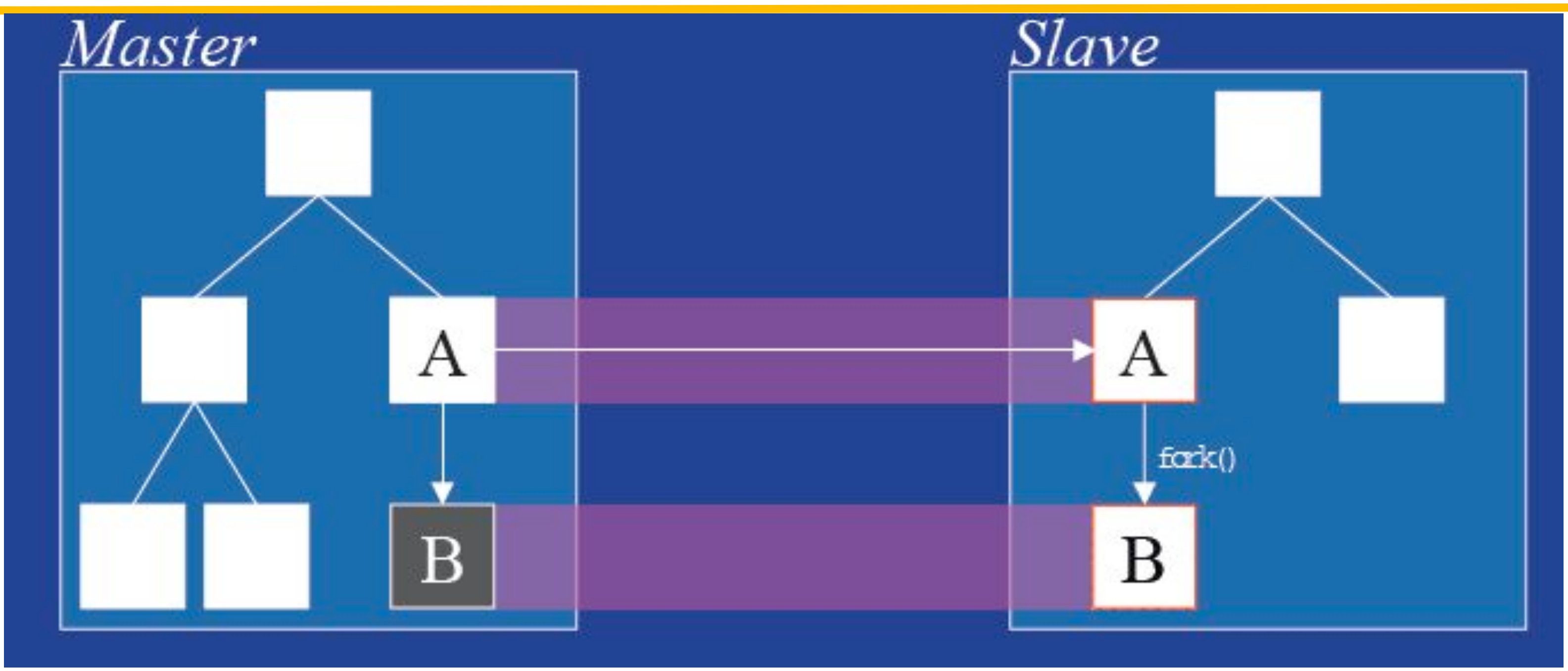
Questions?

BProc, the Heart of Clustermatic



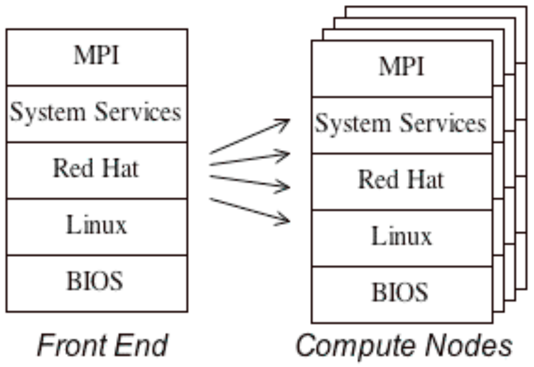
- Bproc = Beowulf Distributed Process Space
- Process Space
  - A pool of process id's
  - A process tree (parent/child relationships)
  - Every instance of a Linux kernel has a process space
- A distributed process space allow parts of a node's process space to exist on another node

Process Creation In BProc

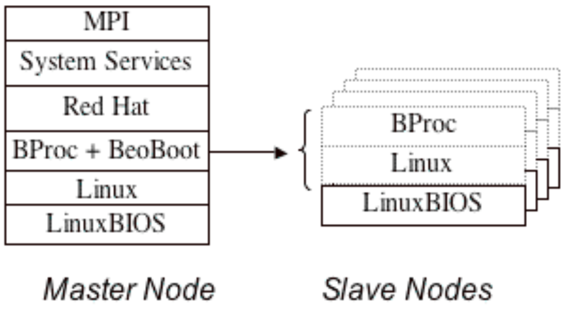


- Process on Master migrates to slave node (1.9s 16MB process on 1024 nodes)
- Process A, on slave, calls fork() to create child process B
- New Place holder for B is created on A (Ghost)
- Not all processes on slave node appear on master space

Science Appliance vs. a Traditional Cluster



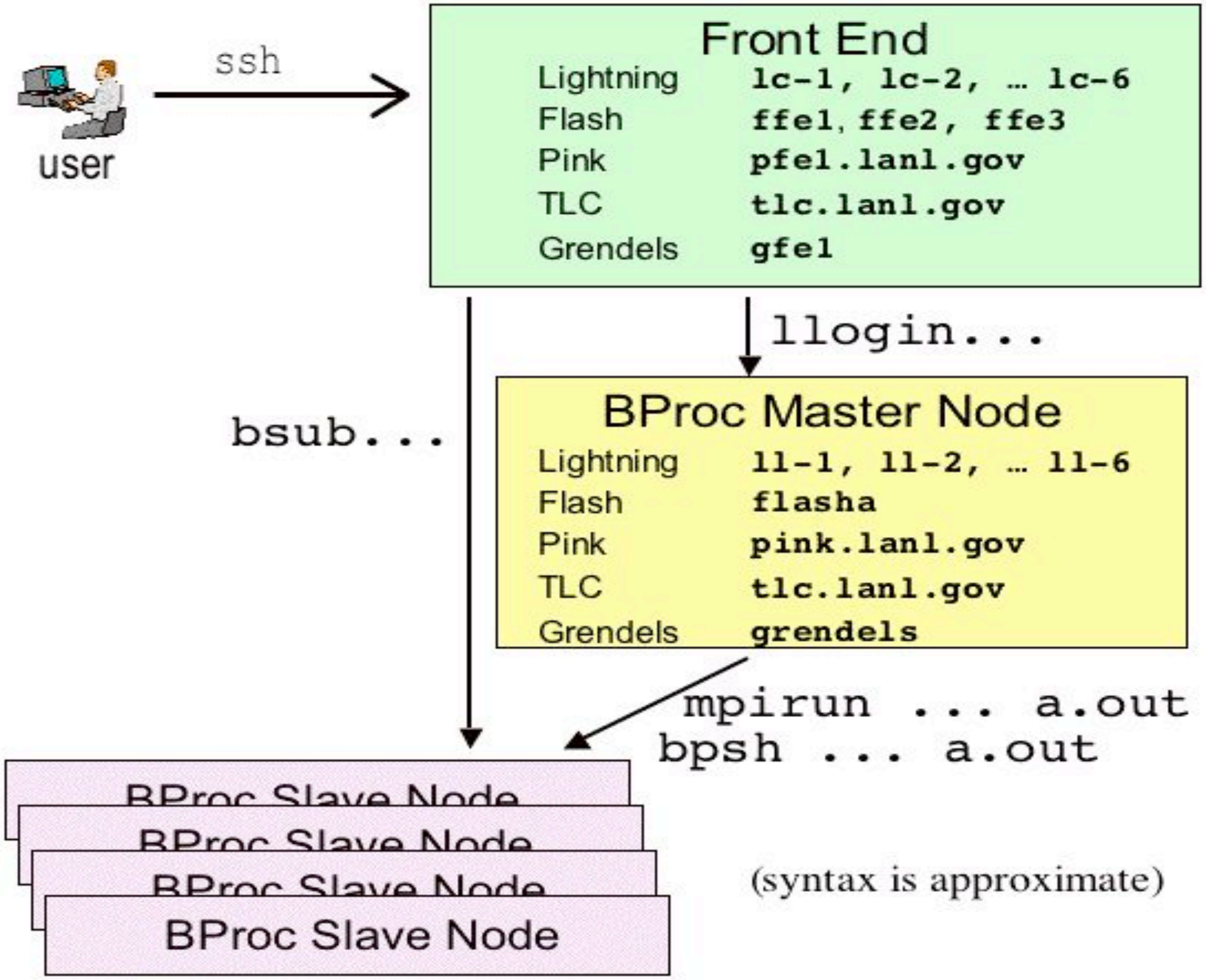
Traditional Cluster Architecture



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...`

`bsub...`

BProc Master Node

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

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

BProc Slave Node

BProc Slave Node

BProc Slave Node

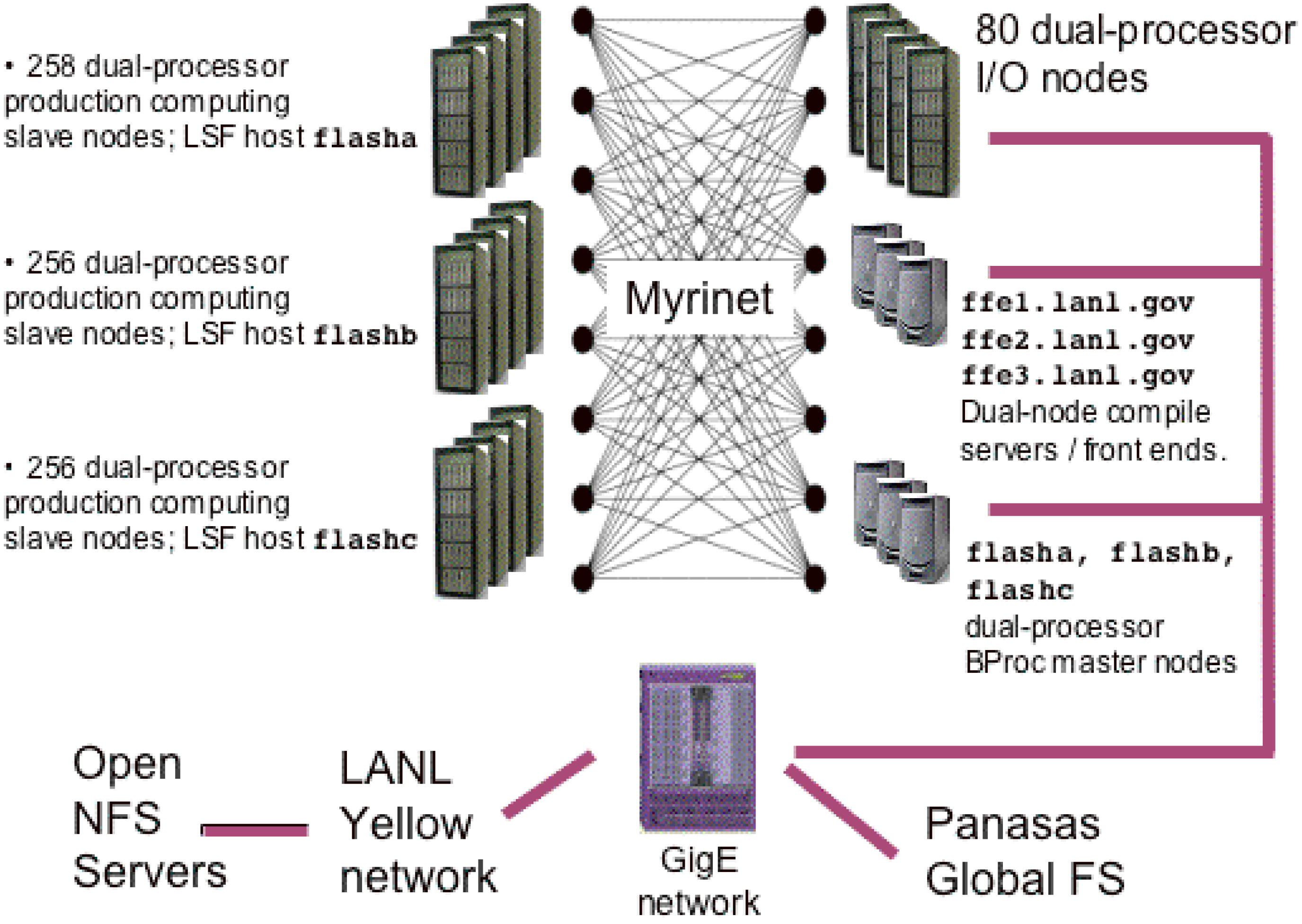
BProc Slave Node

(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