ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

INFORMATION TECHNOLOGY
DIVISION

The LBNL Perceus Cluster Infrastructure

Next Generation Cluster Provisioning and Management

Gary Jung, SCS Project Manager
http://scs.lbl.gov/

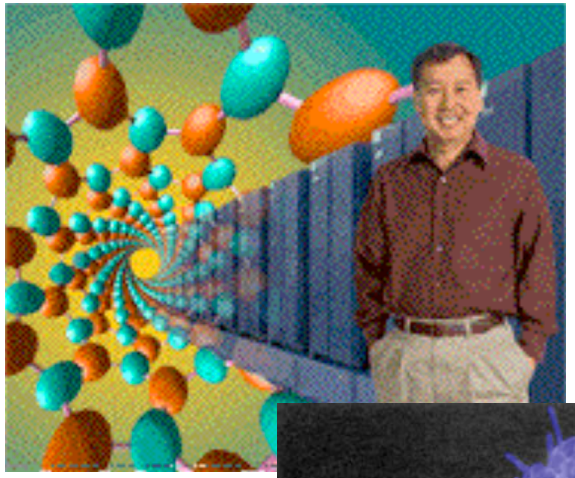
October 10, 2007

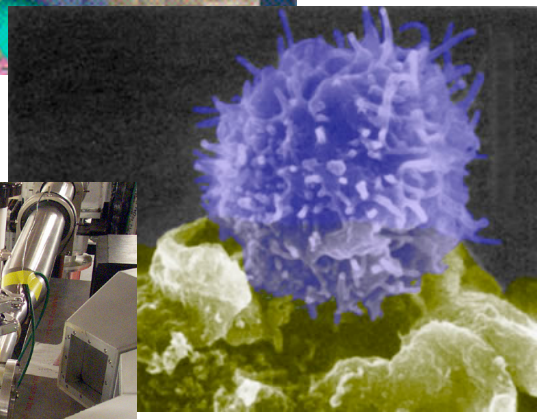
Internet2 Fall Conference

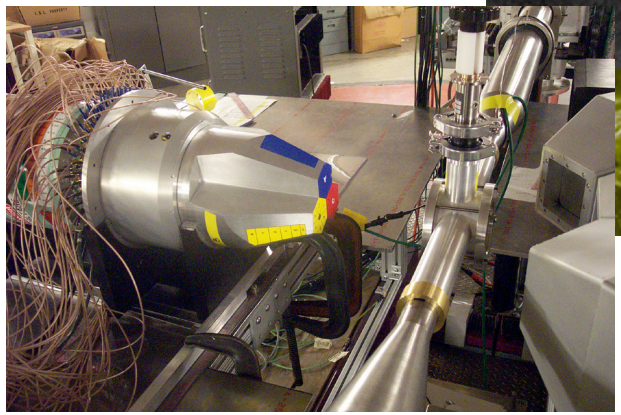
Outline

Perceus Infrastructure

- Introduction
- Cluster Support at LBNL
- Support Methodology (previous)
- The New Infrastructure
- Support Methodology (new)
- What's next?
- Upcoming Challenges







Cluster Support at LBNL

The SCS Program

- Research projects purchase their own Linux clusters
- IT provides comprehensive cluster support
  - Pre-purchase consulting
  - Procurement assistance
  - Cluster integration and installation
  - Ongoing systems administration and cyber security
  - Computer room space with networking and cooling

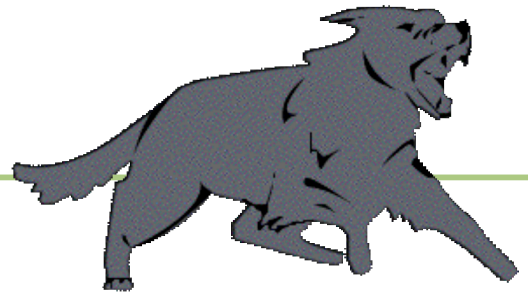
30 Clusters in production (over 2600 processors)

- Huge success! Fifth year of operation
- Examples of recent clusters include:
  - Molecular Foundry 296 -> 432 PE Infiniband (IB) Cluster (Oct 2007)
  - Earth Sciences 388 -> 536 PE IB Cluster (Feb 2007)
  - Department of Homeland Security 80 PE IB Cluster (Dec 2006)

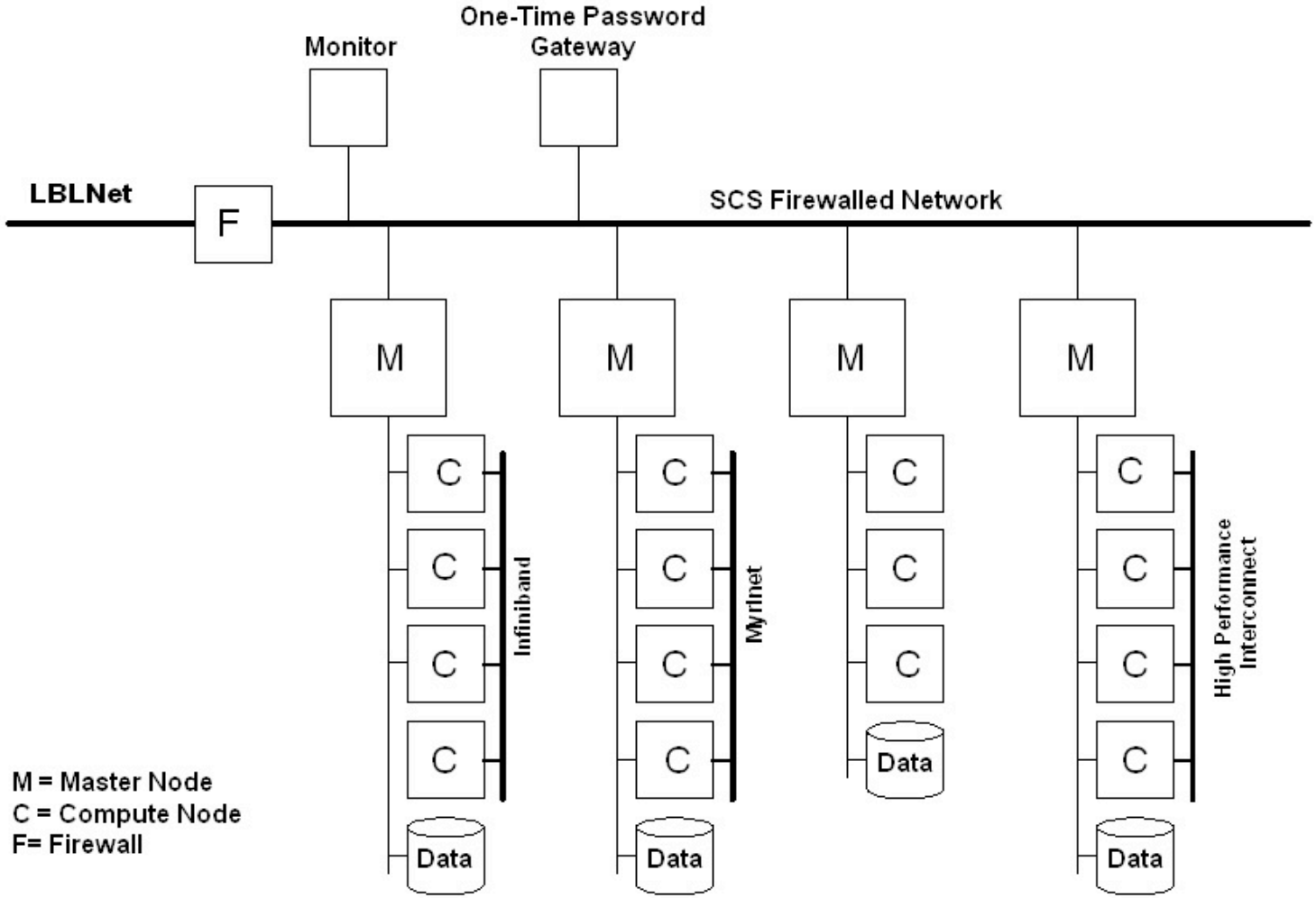
Support Methodology (previous)

Key points

- Standard hardware and software facilitates systems administration
  - Ia32 or x86-64 architecture
  - Choose Linux OS distro, MPI, scheduler, etc…
  - Minimizes skill set requirements for support
  - Expertise transferable between clusters
- Warewulf Cluster toolkit allows us to easily scale up clusters
  - Allows all nodes to be configured from the master node
  - Runs nodes “stateless” (disks not required)
  - Incremental effort to manage more nodes
- Minimal infrastructure to keep recharge costs down
  - Most clusters behind firewall
  - One-time password authentication



SCS Infrastructure (previous)



Support Issues (previous)

What are the issues that we need to solve?

- Each cluster has its own
  - Software: applications, libraries, compilers, scheduler, etc..
  - Hardware: Interconnect, storage, mgmt infrastructure
  - Not standard enough!
- Provisioning new clusters is time consuming
  - Each new system built from scratch
  - Significant amount of work required to get user applications running first and then performing well with each new cluster
- Major cluster upgrades usually time consuming. Typical process would be to:
  - Partition some part of the cluster
  - Install new software stack on new partition
  - Integrate and test user applications on new stack
  - Migrate compute nodes and users to new environment
  - Overall process usually takes several weeks or months

Support Issues (previous)

- Proliferation of clusters has led to users needing accounts on more than one cluster
- Giving users access to multiple systems is a manual process. No centralized mgmt of user accounts.
  - Requires managing multiple accounts
  - Users have to copy data from one cluster to another to run.
  - Results in replicated user environments (e.g. Redundant data sets, binaries)
  - Minor software differences may require users to recompile
  - Installation of layered software is repeated to a varying degree
- Sharing resources between clusters not possible
  - Idle compute resources can't be easily leveraged
  - As mentioned, software inconsistencies would make for a lot of work for the users
  - Grid might help with sharing, but not with the support effort

How can we improve?

- We've scaled the support effort to manage a standalone cluster very well.
- Now, how do we scale the support effort in another dimension to manage a very large number of clusters?

New Infrastructure

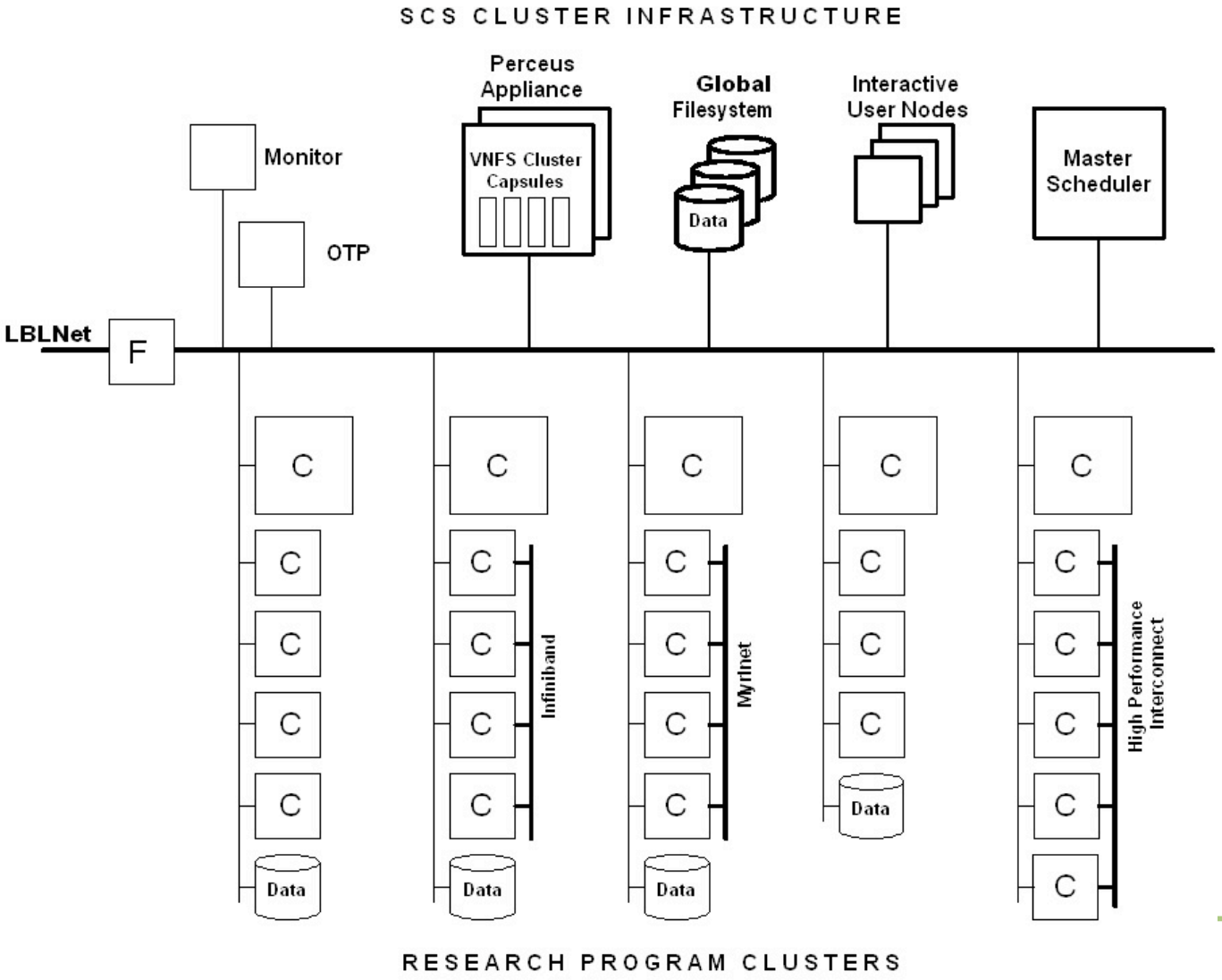
Infiscale Perceus

- Next Generation Warewulf software (http://www.perceus.org)
  - Large scale provisioning of nodes

LBNL Perceus Cluster Infrastructure

- Flatten network to connect everything together
- Use Infiscale Perceus software to allow sharing of certain de-facto resources
- Perceus Appliance is essentially a “SuperMaster” node
- Single Master job scheduler system
- Global home filesystem
- Everything else becomes a diskless cluster node
  - Interactive nodes
  - Compute nodes
  - Storage nodes

New Infrastructure



Support Methodology (new)

Perceus Appliance (SuperMaster)

- Almost everything is booted from the Perceus Appliance
- Two master nodes provide high availability
- Cluster configurations stored in Perceus Berkeley DB (BDB) database on shared filesystem
- VNFS capsules
  - Supports the use and sharing of multiple different cluster images
  - Provides versioning of cluster images to facilitate upgrades and testing
- Facilitates provisioning new clusters. Instead of building each new cluster and software stack from scratch, we do the following:
  - Users test their applications on our test cluster nodes
  - Purchase and install new cluster hardware
  - Configure and test. New systems uses existing shared software stack, home directories, and user accounts already setup.
  - Saves several days effort. As a result, new systems are provisioned much faster

Support Methodology (new)

Global Filesystem

- Shared global filesystem can minimize the copying of data between systems
- Users compile once on Interactive node and can submit jobs to clusters accessible to them from one place
- Software installed on global filesystem can be shared by all
  - Minimizes the need to install software on every cluster
  - Uniform path to software facilitates running of code across all
  - Sharing compilers is more cost effective

User Environment Modules

- Allows different versions and builds of software to coexist on the clusters
- Users can dynamically load specific version of compiler, MPI, and application software at run time
- Facilitates finding the “right” combination of software to build older applications
- Allows us to upgrade software without breaking applications

Support Methodology (new)

Master Scheduler

- Single master scheduler to setup to schedule for all clusters.
- Allows users to submit jobs to any accessible clusters
- Features in OpenMPI will allow us to run jobs spanning multiple clusters across different fabrics.
- Facilitates monitoring and accounting of usage

Shared Interactive Nodes

- Users login to interactive nodes; compile their programs and submit jobs to the scheduler from these nodes
- Interactive Nodes includes support for graphical interactive sessions
- More Interactive Nodes can be added as needed

Dedicated Resources

- Clusters can be still dedicated as before
- Users see what they need
  - Default to their own cluster (but can have access to others)
  - Always see their “global” home directory
  - User still sees a cluster. We see groups of nodes

What's next?

Shared Laboratory Research Cluster (LRC)

- Integrating a large shared resource makes more sense now
- Users can use either the LRC or their own cluster
- The goal will be better utilization of resources to meet demand

Cycle Sharing

- Opportunity to make use of spare cycles
- Now more a a political, instead of technical, challenge

UC Berkeley

- Separate institution also managed by the University of California
- Recent high profile joint projects now encourages close collaboration
- Proximity to LBNL facilitates the use of SCS services
  - One cluster in production; 3 more pending
  - Perceus will be used to setup similar infrastructure
  - Future goal may be to connect both infrastructures together and run clusters over WAN

Upcoming Challenges

Global Filesystem

- Currently used for home directories only
- Can see that this is a good path
- Need to select appropriate technology to scale up
  - HPC parallel filesystems usually complex, require client app, and $$
  - Conventional enterprise technology will work for now
- Can it run over WAN to UCB?

Scheduler and Accounting

- Much more complex than scheduling for small groups
- Need method for tracking user allocations in cpu hours
- Many solutions; not clear which is the best

Security and Networking

- Clusters are more tightly integrated. Need to evaluate current practices
- What are the security implications of extending our Perceus Infrastructure to the UC Berkeley campus
- How do we technically do it?

Conclusion

Summary

- Next evolutionary step in providing cluster support
- Simplifies provisioning new clusters
- Reduces support effort and keeps our technical support staff sane
- Transparent to the users for the most part, but becomes a great help when they start using more than one cluster
- Leveraging of multiple resources will benefit researchers