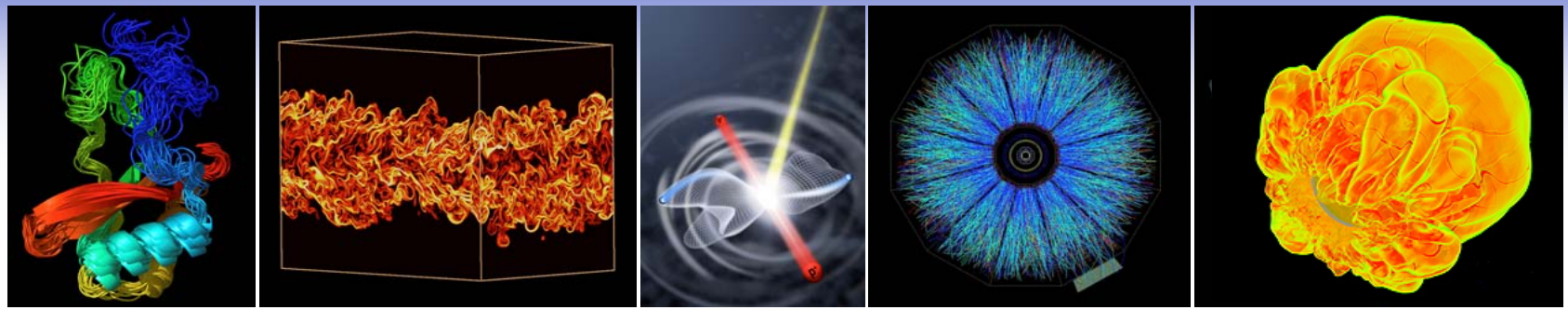


NERSC Global File System Update

Shane Canon
Data Systems Group Leader
Lawrence Berkeley National Laboratory

NERSC All Hands Meeting
September 24, 2008

Berkeley Lab

NERSC

Office of Science
U.S. Department of Energy

NERSC Global File System

- NERSC Global File System (NGF) provides a common global file system for the NERSC systems.
- Currently mounted on all major systems – Bassi, Da Vinci, Franklin (login only), Jacquard, PDSF
- Currently provides Project space
- Targeted for files that need to be shared across a project and/or used on multiple systems.

NGF and GPFS

- NERSC signed a contract with IBM in July for GPFS
- Contract extends through 2014.
- Covers all major NERSC systems through NERSC6 including “non-Leadership” systems such as Bassi and Jacquard.
- Option for NERSC7

Overview of File Systems on NERSC Systems

Today

- Home (Local on each system)
- /scratch (Local on each system)
- /project (global – except Franklin CN)

After next NGF upgrades

- Home (global)
- /scratch (Local on each system)
- /project (global)

Goals for next NGF Upgrade

- Fully connect Franklin (Cray XT4) to NGF
- Improve access to NGF from other systems (Bassi, Jacquard, PDSF)
- Increase Capacity and Bandwidth
- Convert to Global Homes

Franklin Login Nodes

- Franklin has 10 Login Nodes and 6 PBS launch nodes
- Currently mounted over NFS
- Performance is poor but it has allowed access to NGF from Franklin
- Testing with native GPFS client and TCP based mounts
- Intend to use SAN based mount on Login nodes in near future

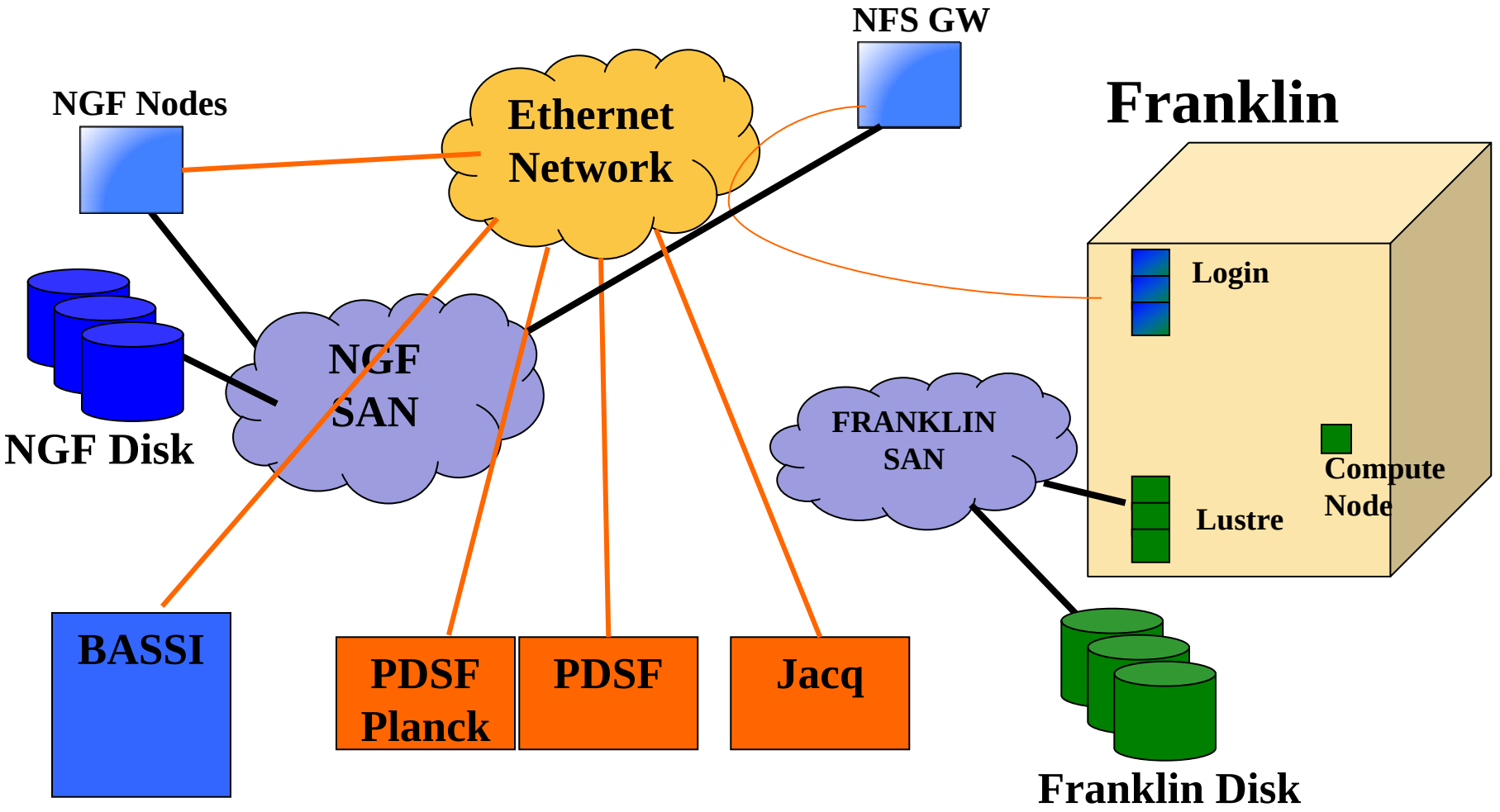
Franklin Compute Nodes

- NERSC will use Cray's DVS to mount NGF file systems on Franklin compute nodes.
- DVS ships IO request to server nodes which have the actual target file system mounted.
- Beginning to test DVS with GPFS now
- Plan to deploy around 20 DVS servers connected via SAN. Initially we expect around 6 GB/s due to the topology, but this could expand to 12 GB/s.

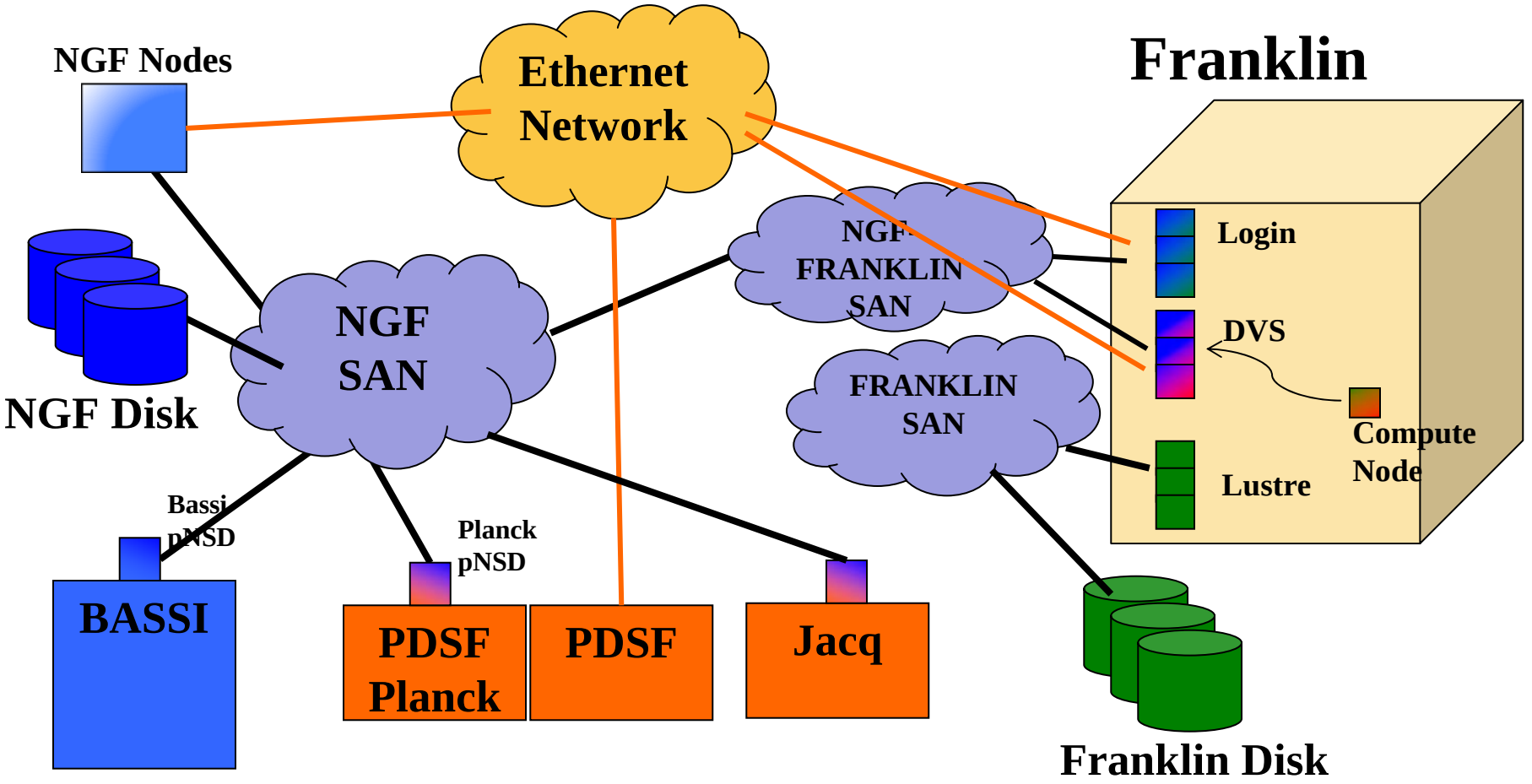
Improving NGF Access

- Improved bandwidth for other systems
  - Bassi
  - Jacquard
  - PDSF/Planck
- Will leverage new capabilities in GPFS to better utilize the native networks in the systems (Federation and IB)

NGF Topology



NGF Topology



Target Bandwidths

	Today	After Upgrades
NGF (Total BW)	~6 GB/s	~10 GB/s
Bassi	1 GB/s	~3 GB/s
Da Vinci	1 GB/s	~2 GB/s
Franklin	<100 MB/s	~6 GB/s
Jacquard	1 GB/s	~2.5 GB/s
PDSF	1 GB/s	1 GB/s
Planck	1 GB/s	~4.5 GB/s

Increasing Capacity

- Will deploy both existing idle storage and new storage to boost capacity
  - Some storage held in reserve while file system options were being evaluated
  - New storage being purchased
- Additional ~75 TB of FibreChannel storage
- Additional ~150 TB of SATA storage
- Should also boost bandwidth to around 10-12 GB/s

Global Homes

- Storage Strategy – Need to evaluate storage options
- Integration with account support – Integration with NIM (mostly complete)
- Review of dot files and other cross system related issues (plan is already in place)
- Transition of existing homes to new system (coordinate with User Services and System Administrators)

Global Homes and "dot" files

- Global Homes will use a system specific sub-directory as the home directory on each system.
- This will allow users to easily customize the environment for each system while still providing a common home directory

```
.../canon/
.../canon/{franklin,bassi,...}/{.login, .cshrc}
.../canon/{franklin,bassi,...}/common -> ../common
.../canon/common/{source, batch scripts, etc}
```

Timeline

Task	Date
Increased capacity	November 2008
Improved Access:	
- Bassi, Planck	December 2008
- Jacquard	November 2008
Franklin Login	November 2008
Franklin Compute	January 2009
Global Homes	January 2009

Transfer Nodes

- NERSC plans to deploy multiple transfer nodes
- Dedicated to transferring data over the WAN
- High-bandwidth access to NGF and HPSS
- Tuned for WAN transfers with documented instructions on how to get good performance from typical end-points (ORNL, ANL, etc)
- Standard transfer tools will be installed and tested (scp, bbcp, globus, possibly SRM)

Filesets

- Project directories will be moved to use GPFS “filesets”
- Allows quotas to be created on filesets (enables directory based quotas)
- Migration policies can be based on file sets (along with other parameters such as age or size).

Other Plans

- Exploring the use of GPFS pools to segment storage. Placing cold files in less expensive storage (slower SATA storage).
- Separating workloads and instituting different policies (i.e. source code and input data, backed up and not backed up)
- Automated performance monitoring

Possibilities

- No current plans to replace Franklin scratch with NGF scratch or GPFS. However, we plan to evaluate this once the planned upgrades are complete.
- Explore Global Scratch – This could start with Jacquard to prove feasibility
- Tighter integration with HPSS

Questions?