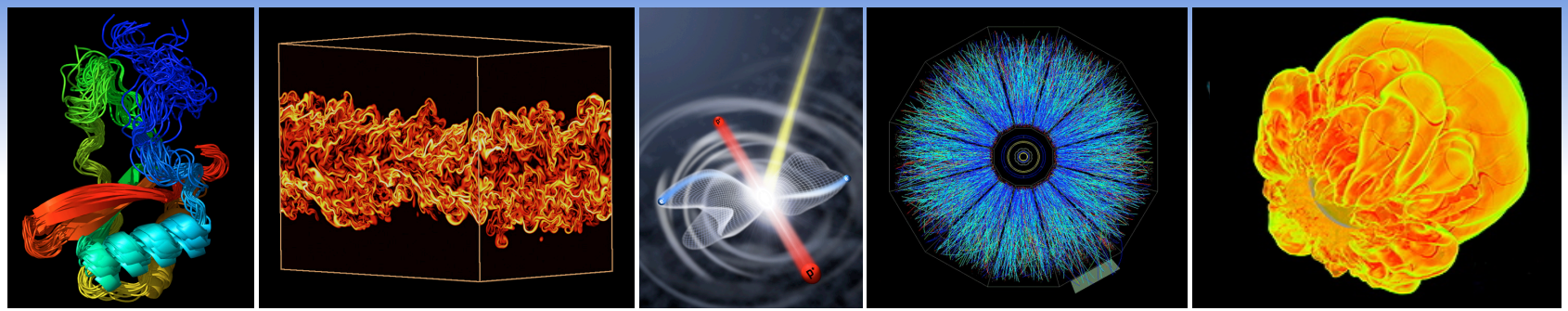


GPFS on a Cray XT

Shane Canon
Data Systems Group Leader
Lawrence Berkeley National Laboratory
CUG 2009 – Atlanta, GA
May 4, 2009

BERKELEY LAB

NERSC

Office of Science
U.S. DEPARTMENT OF ENERGY

NERSC
NATIONAL ENERGY RESEARCH SCIENTIFIC COMPUTING CENTER

Outline

- NERSC Global File System
- GPFS Overview
- Comparison of Lustre and GPFS
- Mounting GPFS on a Cray XT
- DVS
- Future Plans

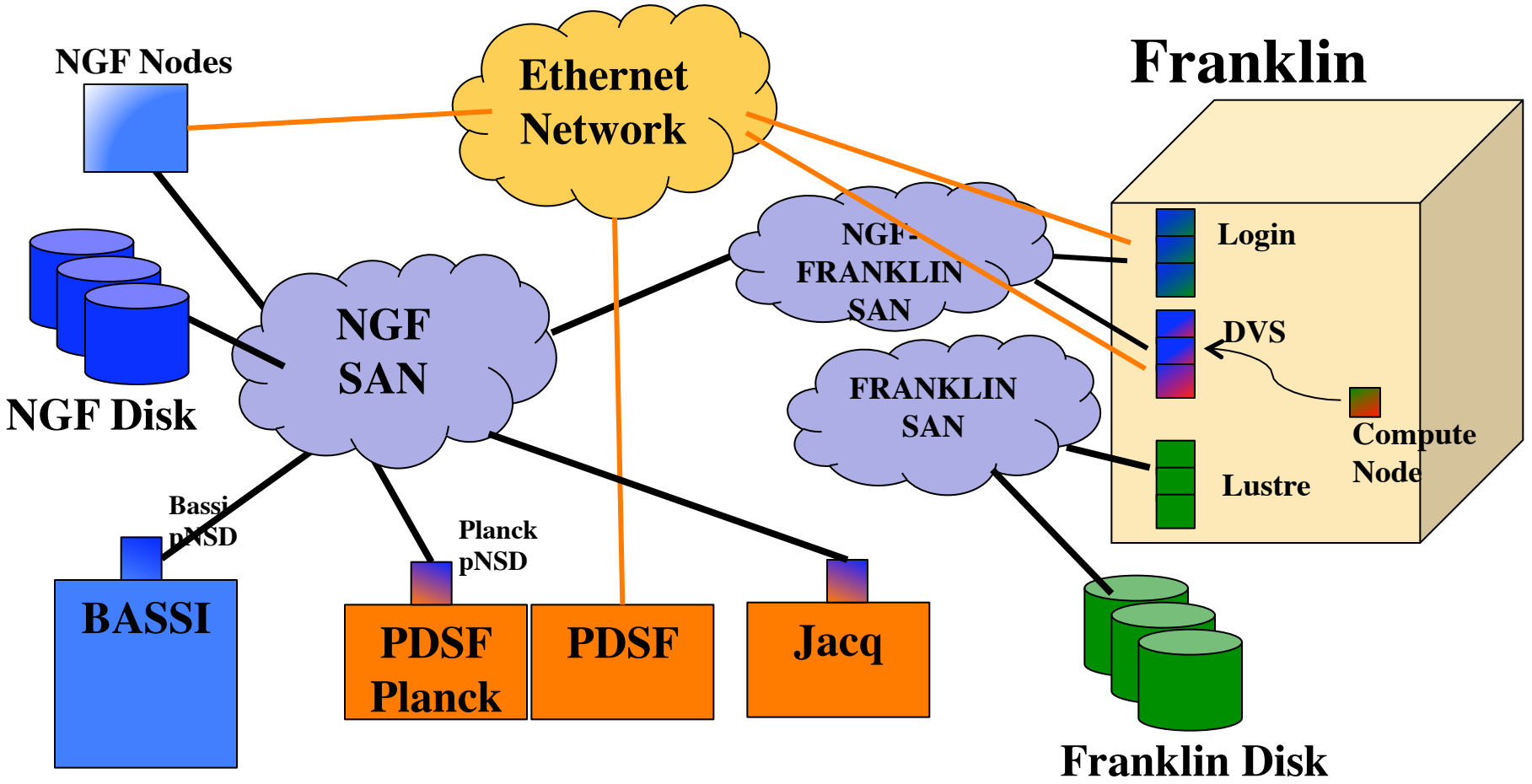
NERSC Global File System

- NERSC Global File System (NGF) provides a common global file system for the NERSC systems.
- In Production since 2005
- Currently mounted on all major systems – IBM SP, Cray XT4, SGI Altix, and commodity clusters
- 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 2008 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

NGF Topology



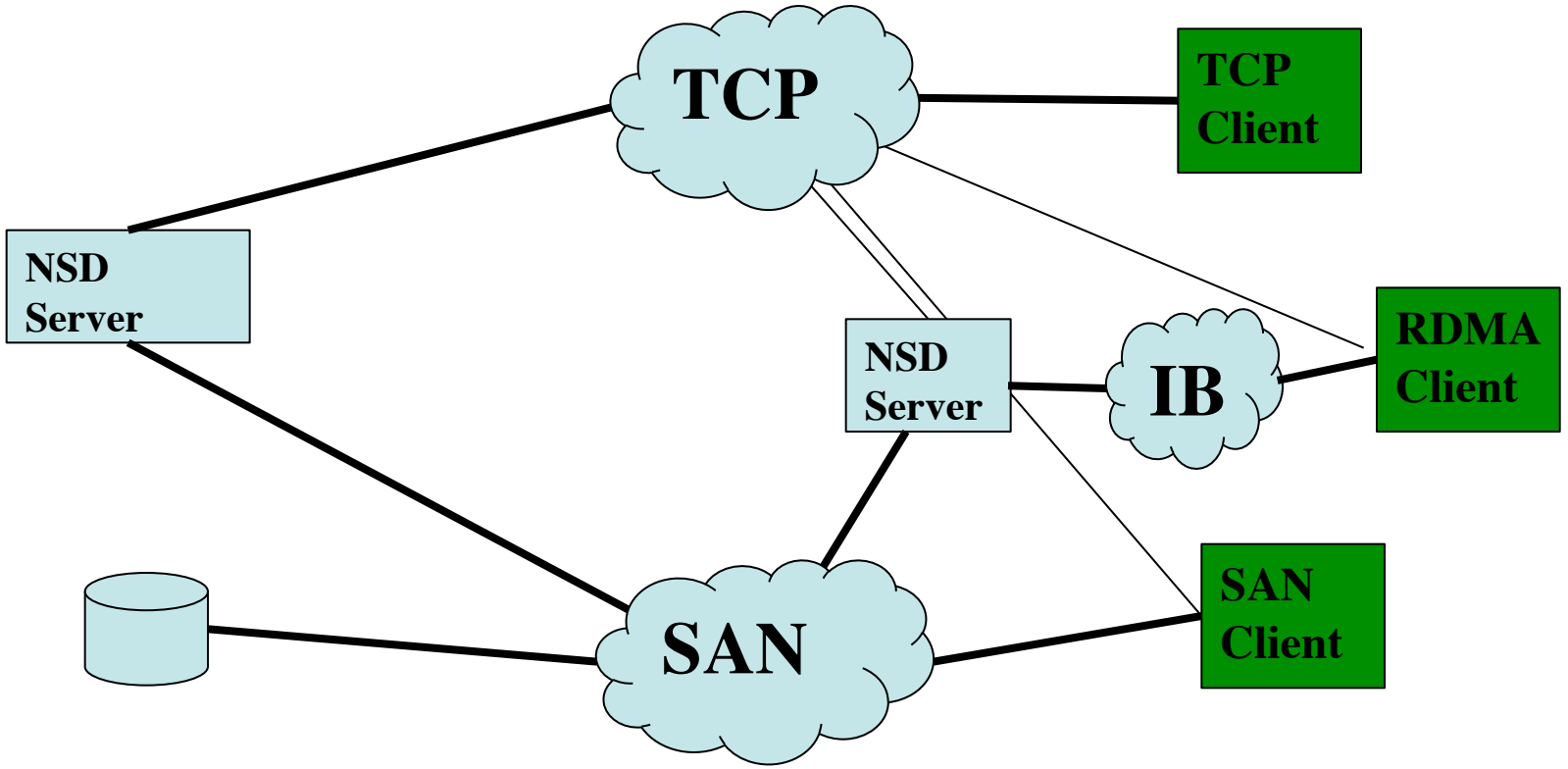
GPFS Overview

- Share disk model
- Distributed lock manager
- Supports SAN mode and Network Shared Disk modes (mixed)
- Primarily TCP/IP but supports RDMA and Federation for low overhead, high bandwidth
- Feature rich and very stable
- Largest deployment: LLNL Purple 120 GB/s, ~1,500 clients

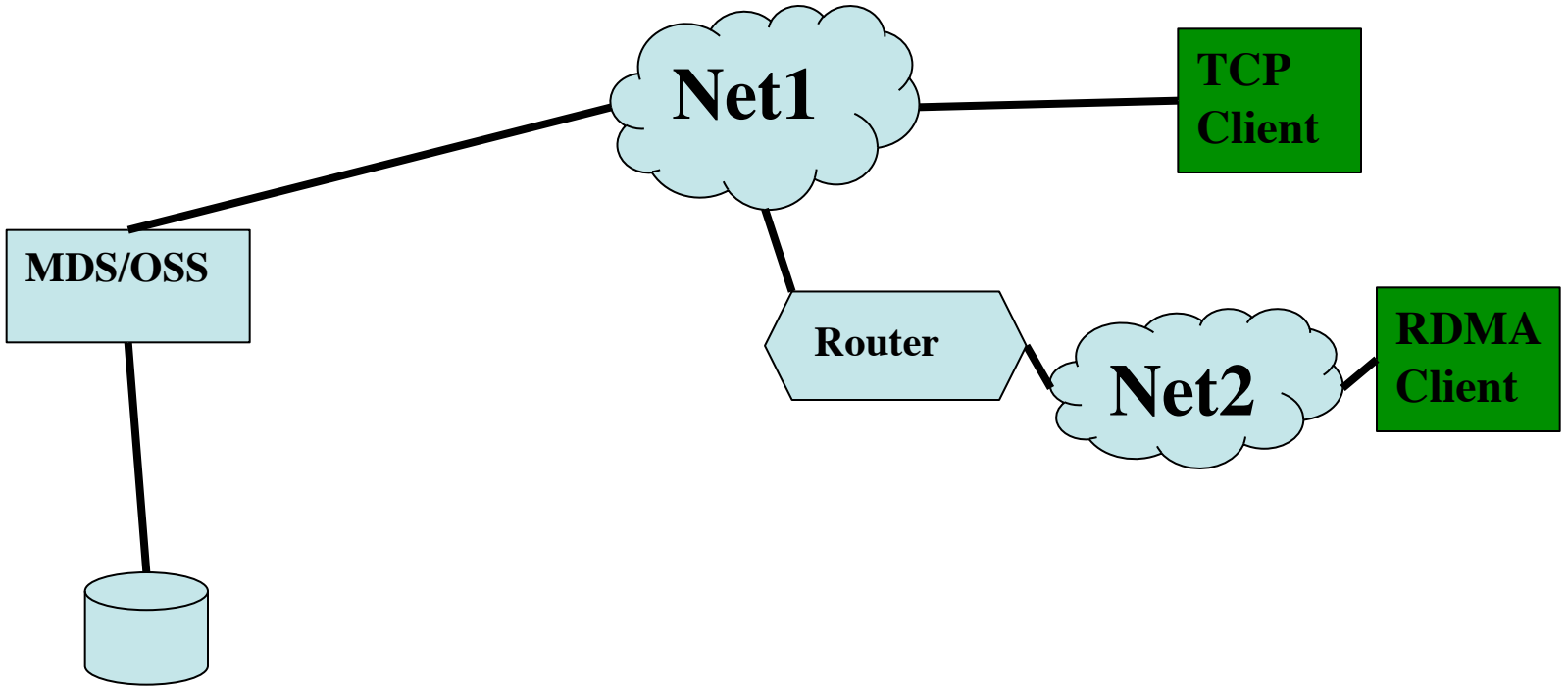
Comparisons – Design and Capability

	GPFS	Lustre
Design		
Storage Model	Shared Disk	Object
Locking	Distributed	Central (OST)
Transport	TCP (w/ RDMA)	LNET (Routable Multi-Network)
Scaling (Demonstrated)		
Clients	1,500	25,000
Bandwidth	120 GB/s	200 GB/s

GPFS Architecture



Lustre Architecture



Comparisons – Features

	GPFS	Lustre
Add Storage	✔	✔
Remove Storage	✔	O
Rebalance	✔	O
Pools	✔	1.8 (May)
Fileset	✔	
Quotas	✔	✔
Disributed Metadata	✔	3.0 (2010/11)
Snapshots	✔	
Failover	✔	O

O-With user/third-party assistance

GPFS on Franklin Interactive Nodes

- Franklin has 10 Login Nodes and 6 PBS launch nodes
- Currently uses native GPFS client and TCP based mounts on login nodes
- Hardware is in place to switch to SAN based mount on Login nodes in near future

GPFS on Cray XT

- Mostly “just worked”
- Install in shared root environment
- Some modifications needed to point to the correct source tree
- Slight modifications to mmremote and mmsdrfsdef utility scripts (to use *ip* command to determine SS IP address)

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.
- DVS has been tested with GPFS at NERSC at scale on Franklin during dedicated test shots
- Targeted for production in June time-frame
- Franklin has 20 DVS servers connected via SAN.

IO Forwarders

IO Forwarder/Function Shipping – Moves IO requests to a proxy server running file system client

Advantages

- Less overhead on clients
- Reduced scale from FS viewpoint
- Potential for data redistribution (realign and aggregate IO request)

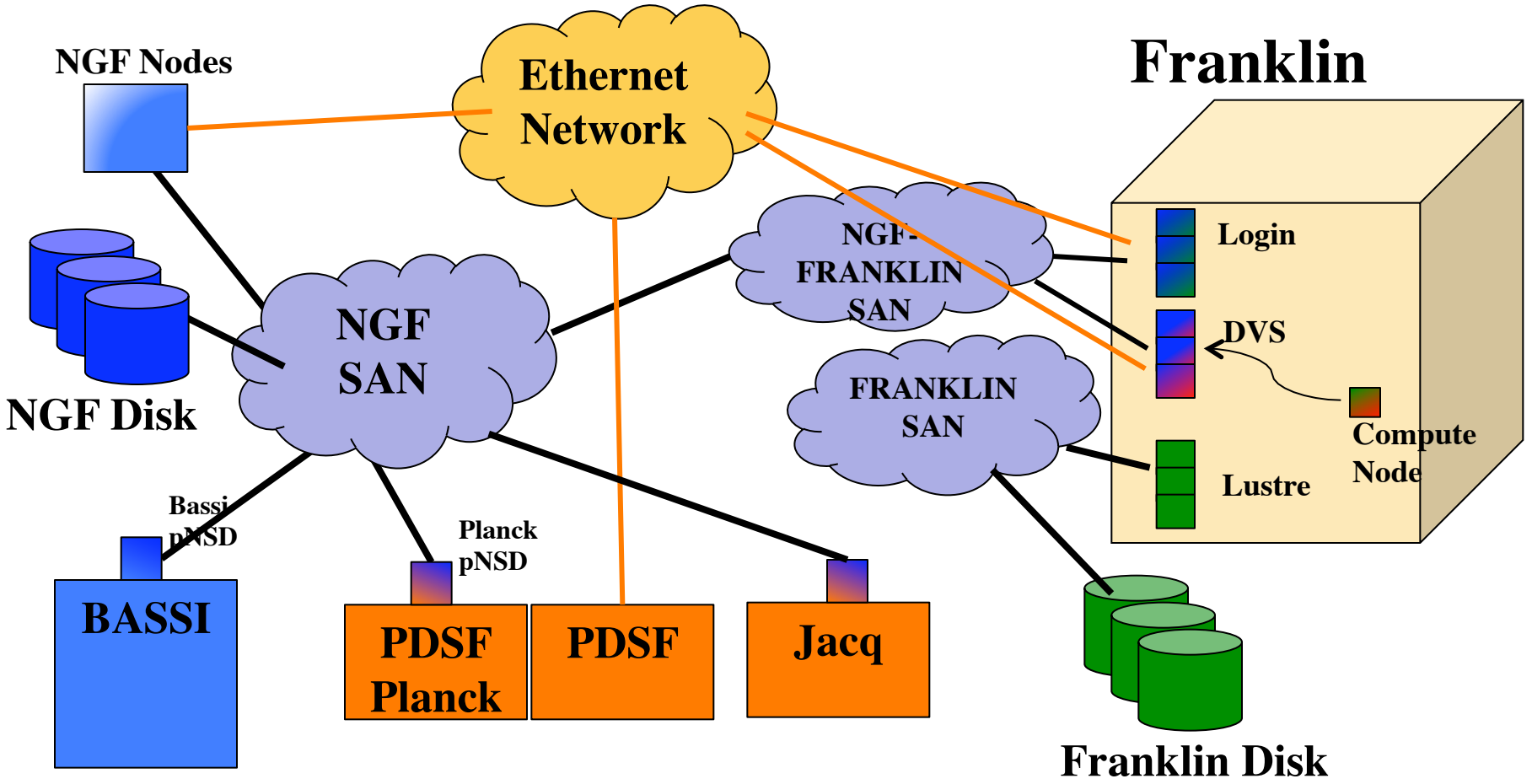
Disadvantages

- Additional latency (for stack)
- Additional SW component (complexity)

Overview of DVS

- Portals based (low overhead)
- Kernel modules (both client and server)
- Support for striping across multiple servers (Future Release-tested at NERSC)
- Tune-ables to adjust behavior
  - "Stripe" width (number of servers)
  - Block size
  - Both mount options and Env. variables

NGF Topology (again)



Future Plans

- 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 smaller Linux cluster to prove feasibility
- Evaluate tighter integration with HPSS (GHI)

Long Term Questions for GPFS

- Scaling to new levels (O(10k) clients)
- Quality of Service in a multi-clustered environment (where the aggregate bandwidth of the systems exceed the disk subsystem)
- Support for other systems, networks and scale
  - pNFS could play a role
  - Other Options
    - Generalized IO forwarding system (DVS)
    - Routing layer with abstraction layer to support new networks (LNET)

Acknowledgements

NERSC

- Matt Andrews
- Will Baird
- Greg Butler
- Rei Lee
- Nick Cardo

Cray

- Terry Malberg
- Dean Roe
- Kitrick Sheets
- Brian Welty

Questions?

Further Information

For Further Information:

Shane Canon

Data System Group Leader

Scanon@lbl.gov