NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

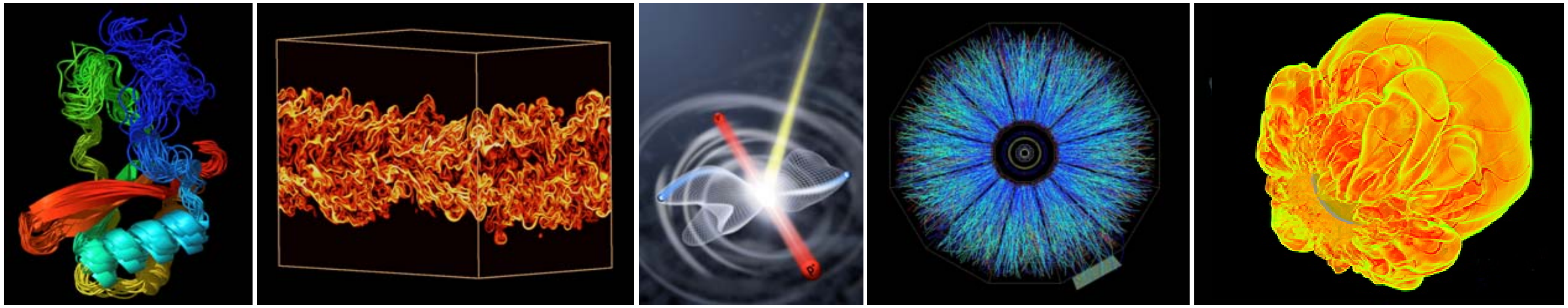
Franklin File Systems & IO

Richard Gerber
NERSC User Services
RAGerber@lbl.gov

NERSC Users Group
Berkeley Lab
Oakland, CA
October 2, 2008

Outline

- File Systems
- System Layout
- Best Practices
- Details
- Reference
  - ▫ www.nersc.gov
  - ▫ www.nersc.gov/nusers/systems/franklin



Franklin File Systems

NUG 2008, Berkeley Lab, NERSC Oakland Scientific Facility

Oct. 2, 2008

What is a File System?

- A special-purpose database for the storage, hierarchical organization, manipulation, navigation, access, and retrieval of data.
  - This is a layer that mediates transactions between the Operating System and the Storage Device.
- A file system deals with “data” and “metadata” (data about the data, e.g. file name, physical location on disk, file size, timestamps)
- We often refer to a “file system name” as the root of a hierarchical directory tree, e.g. “the /home file system.”
  - We can treat this as “one big disk,” but it may actually be a complex collection of disk arrays, IO servers, and networks.

File Systems on Franklin

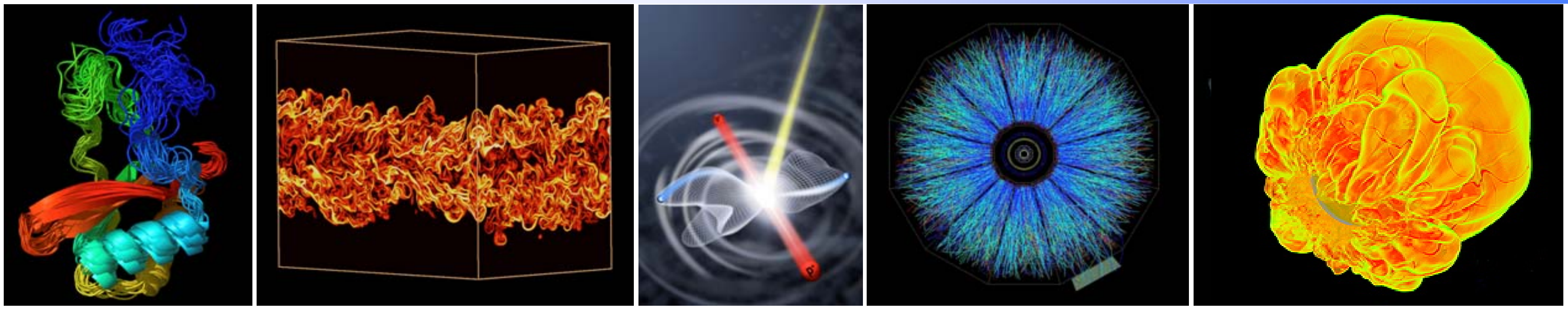
- “Scratch” ($SCRATCH, /scratch)
  - Large temporary high-performance file systems
  - To be used for parallel job IO
  - Not backed up
  - Each user has a unique directory
    - ✓ $SCRATCH (/scratch/scratchdirs/username)
  - Per user quota of 500 GB
  - Purge policy not yet announced, but coming soon
- Home
  - You are in your “home” directory when you log in
  - Permanent storage for source code, binaries, scripts, …
  - Small(ish) quota (15 GB); not intended for data
  - Use $HOME to reference your home directory

File Systems on Franklin

- “Project”
  - Use to share files among group members
  - Not high performance today; will improve
  - Quotas
  - Created by request, /project/projectdirs/proj/ (GPFS)
- /tmp
  - Reserved for system use; DO NOT USE!!
- Archival Storage
  - HPSS mass storage system (archive.nersc.gov)
  - Extremely high capacity
  - Tape storage with disk cache
  - “hsi” is the shell interface utility (ftp-like) to transfer files
  - Can also use ftp

Computational Jobs & File Systems

- **A parallel application (launched with** `aprun`**) can only access $SCRATCH or $HOME**
- **Serial (shell) script commands can access all file systems.**
- **Use** `cd $PBS_O_WORKDIR` **in script to change to submission directory.**
- `STDERR` **and** `STDOUT` **are buffered and returned at job completion.**



System Layout

System IO Architecture

- Should an application scientist or programmer care about these details?
  - Yes! It would be nice not to need to, but performance and perhaps functionality depend on it.
  - You may be able to make simple changes to your code or runtime environment that will greatly improve performance.

Network File Systems

- All disk storage on the XT4 is accessed “externally” as a network file system.
- What is a “network file system?”
  - A file system that supports sharing of files as persistent storage over a network.
  - Network File System (protocol) (NFS)
    - ✓NFS is a standard protocol
    - ✓Widely used and available, but not developed as a standard for high-performance parallel computing
  - Lustre
    - ✓High-performance file systems on the XT4 are Lustre file systems
  - Other examples: AFS, NetWare Core Protocol, Server Message Block (SMB).

XT4 IO Network Fundamentals

- The XT4 has two types of nodes: compute (CNL) and service (login, IO, network; full Linux)
- All nodes are connected by a high-speed Seastar 2 torus network (aka, “The torus”).
- IO service nodes are also connected to large, high-performance disk servers by a fast “Fibre Channel” network.
- Login and batch service nodes are further connected to HPSS and other disk servers via a gigabit ethernet network.

Terminology: Fibre Channel

- Fibre Channel
  - Gigabit network technology primarily used for storage networking. (Franklin is 4 Gb/sec)
  - Fibre Channel Protocol (FCP) is similar to TCP for FC networks
  - Can run over copper or fibre-optic cables.
  - Typically, you have a FC card on a node, similar to a giga-bit ethernet card.

Terminology: Lustre

- **Lustre (derived from “Linux Cluster”)**
- **A clustered, shared file system**
- **Open software, available under GNU GPL**
- **Designed, developed, and maintained by Sun Microsystems, Inc., which acquired it from Cluster File Systems, Inc. in Oct. 2007**
- **Two types of Lustre servers (running on Franklin IO *service* nodes)**
  - **Object Storage Servers (OSS)**
  - **Metadata Servers (MDS)**

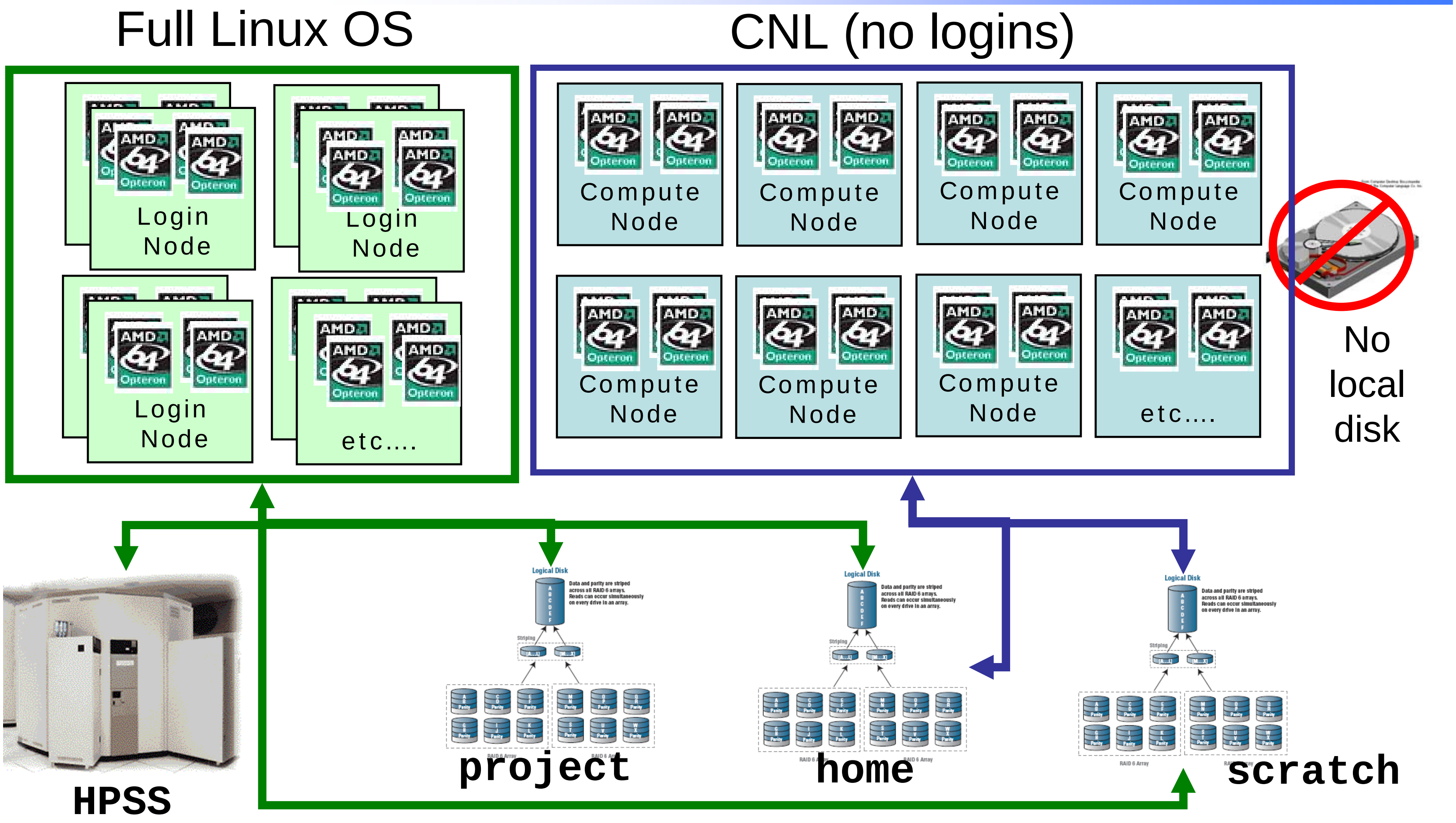
Terminology: Metadata

- **File systems store information about files externally to those files.**
- **Linux uses an inode, which stores information about files and directories** **(size in bytes, device id, user id, group id, mode, timestamps, link info, pointers to disk blocks, …)**
- **Any time a file’s attributes change or info is desired (e.g., `ls –l`) metadata has to be retrieved (from MDS and OSTs) or written.**
- **Metadata operations are IO operations which require time and disk space.**

Types of File Systems on Franklin

- **Lustre**
  - **$SCRATCH is a Lustre file systems.**
  - **$HOME is a Lustre file system.**
  - **A full Lustre client is available for both CNL and Linux, thus Lustre file systems are available from all nodes.**
- **NFS or similar protocol (may be proprietary)**
  - **Project directories (change planned)**
  - **No client or library *support* within CNL, thus no access from compute nodes (change planned).**

Franklin File System Visibility



Full Linux OS

CNL (no logins)

Login Node

Login Node

Login Node

etc....

Compute Node

Compute Node

Compute Node

Compute Node

Compute Node

Compute Node

Compute Node

etc....

No local disk

HPSS

project

home

scratch

Logical Disk
Data and parity are striped across all RAID 6 arrays. Reads can occur simultaneously on every drive in an array.
Striping
RAID 6 Array

Application IO

- All IO performed by your job should use the file system designed for HPC applications.
- $HOME at NERSC is not configured for good application IO performance.
- Lustre is currently the only file system that can be used by parallel applications, so we’ll concentrate on it.

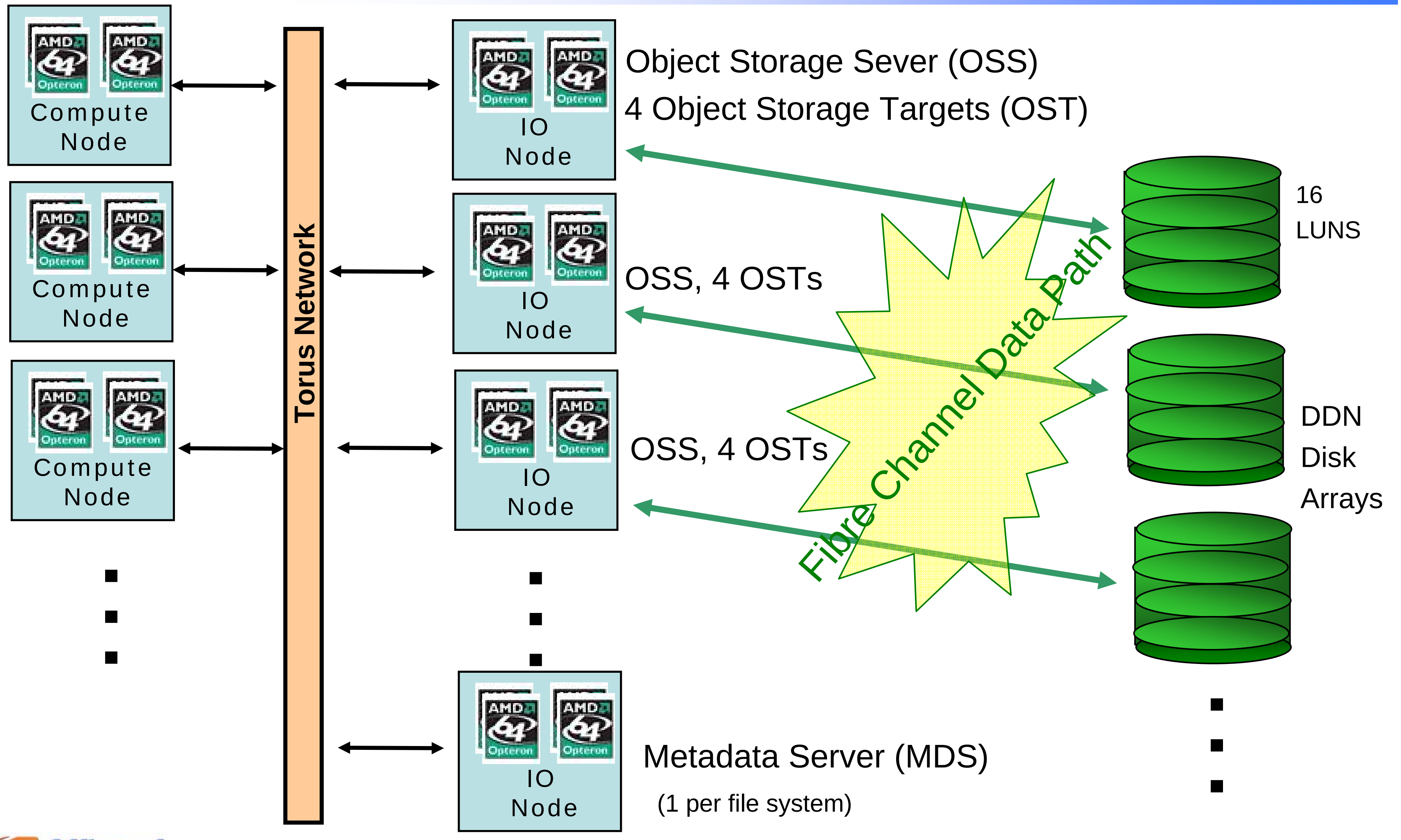
Some XT4 Lustre Terminology

- Object Storage Server (OSS)
  - ▫ Some service nodes are dedicated to IO and serve as OSSs.
  - ▫ 1 OSS == 1 Franklin IO service node
  - ▫ A file system partition (e.g. /scratch) is served by multiple OSSs.
  - ▫ OSSs are connected
    - ✓ To the compute and login nodes via the high-speed torus
    - ✓ To physical disk via a fibre-channel IO network
- Object Storage Target (OST)
  - ▫ Software that presents a single unit of disk to the the OS.
  - ▫ 4 independent OSTs run on each OSS
  - ▫ OSTs are combined into a file system partition that is presented to users
  - ▫ A partition (e.g. /scratch) can be viewed as being built from a number of independent OSTs.
- Metadata Server (MDS)
  - ▫ An IO service node can be configured as an MDS.
  - ▫ The MDS deals with all information about individual files.
  - ▫ One MDS per file system partition.

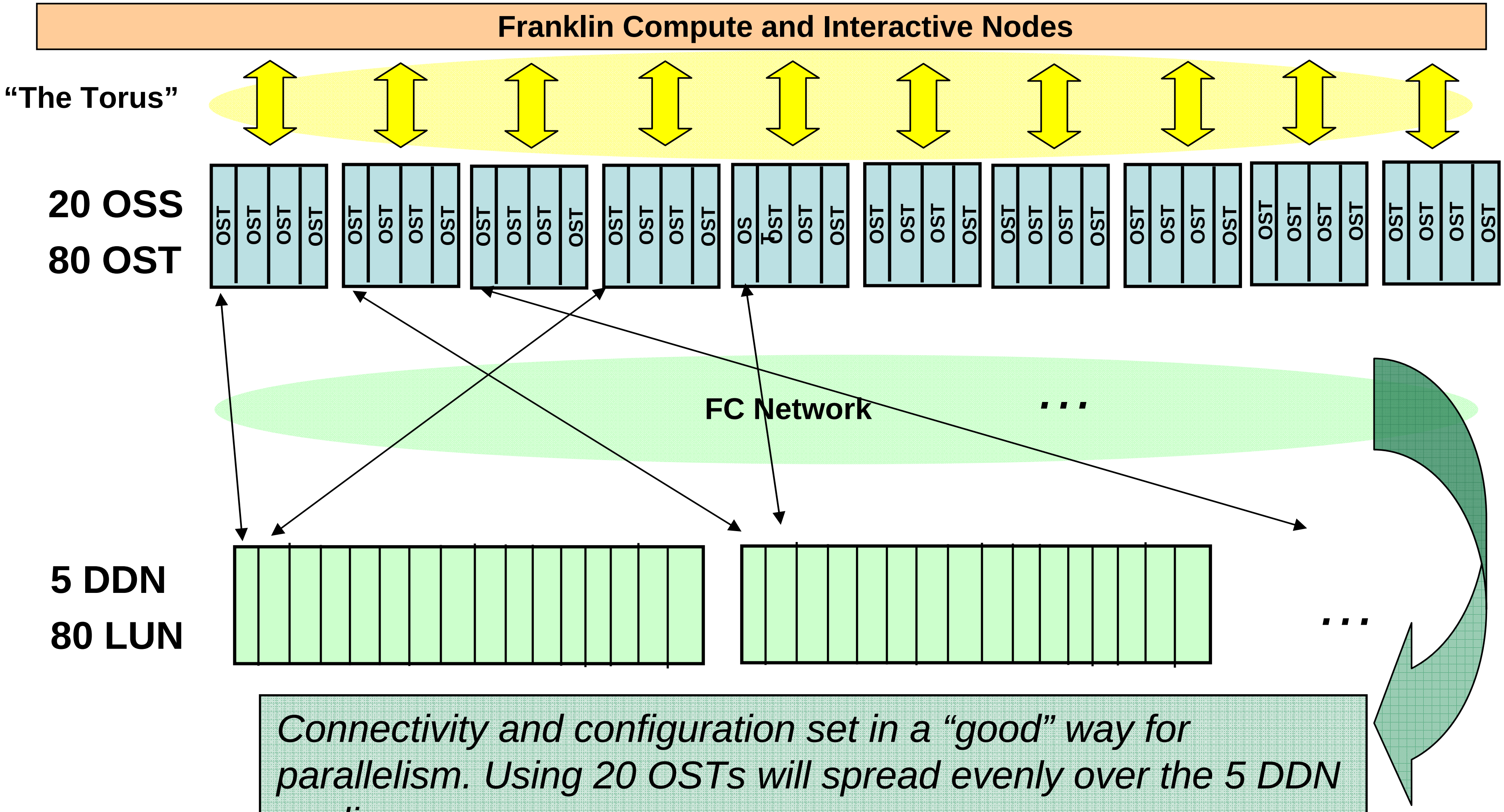
Physical Disks

- Physical disk storage resides on a DDN (Direct Data Networks) Storage Appliance
- The “DDN Disk Arrays” include supporting software and connectivity.
- The DDN server presents collections of hard disks as a Logical Unit Number (LUN) to the file system.
- One Lustre OST maps to one 4TB LUN.

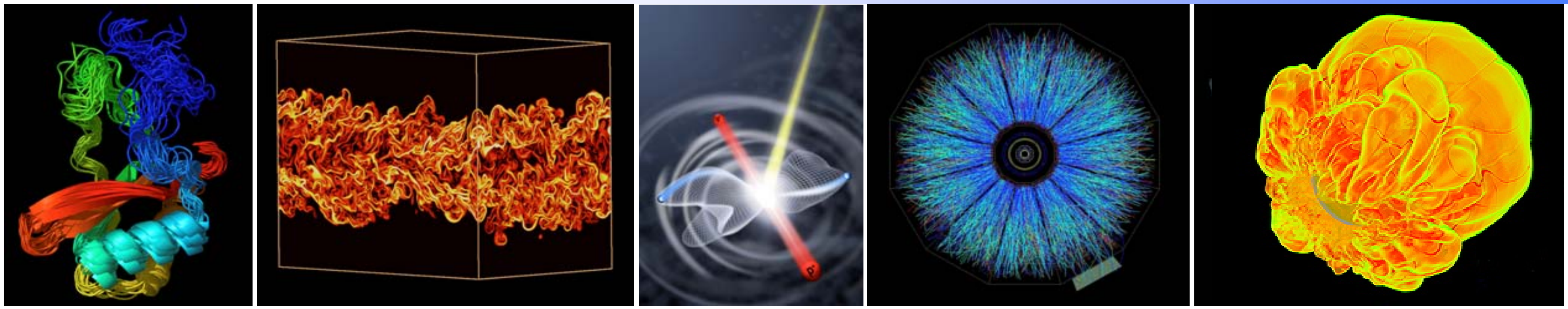
XT4 Lustre Connectivity



Franklin Configuration



Connectivity and configuration set in a "good" way for parallelism. Using 20 OSTs will spread evenly over the 5 DDN appliances.



Next: Application IO and Best Practices