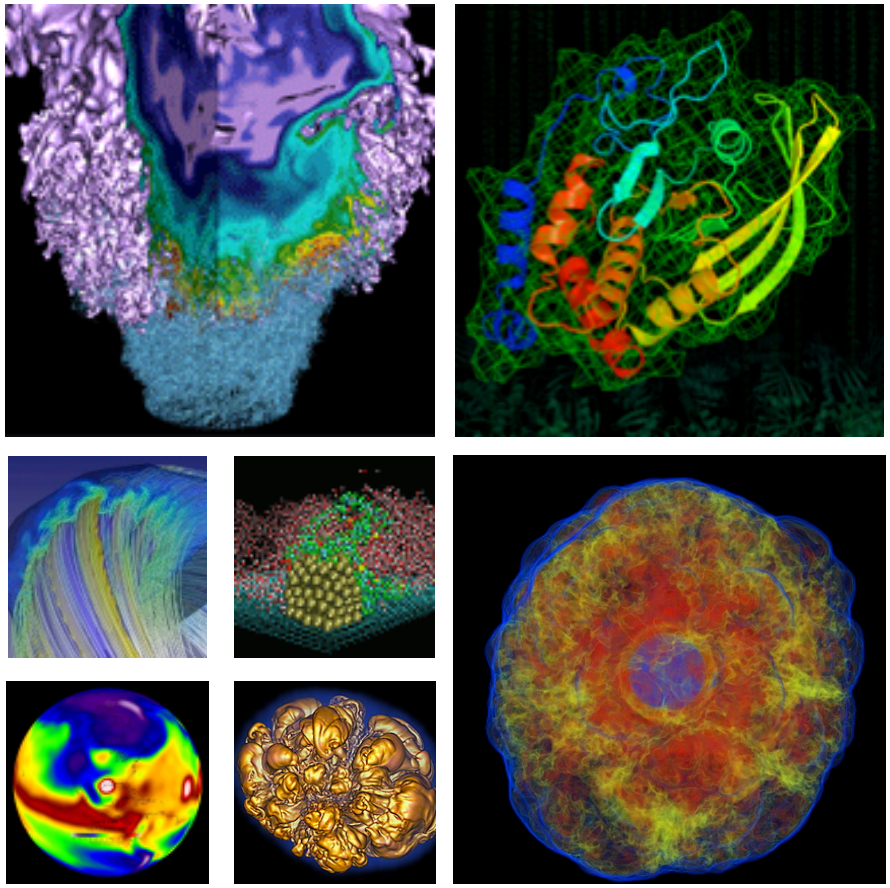
NERSC User Group Storage Update



Damian Hazen
Storage Systems Group

March 24, 2016

U.S. DEPARTMENT OF ENERGY | Office of Science

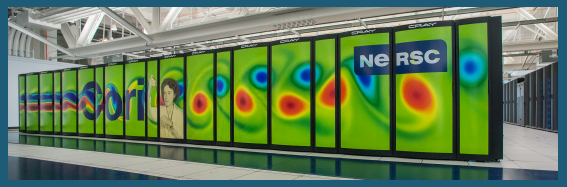
BERKELEY LAB

NERSC Compute and Storage - 2016

Edison: Cray XC-30

5,576 nodes, 133K, 2.4GHz Intel "IvyBridge" Cores, 357TB RAM

Cori: Cray XC-40



Ph1: 1630 nodes, 2.3GHz Intel "Haswell" Cores, 203TB RAM
Ph2: >9300 nodes, >60cores, 16GB HBM, 96GB DDR per node

Data-Intensive Systems
PDSF, JGI,KBASE,Materials

Data Transfer Nodes
Adv. Arch. Testbeds Science Gateways

7.6 PB Local Scratch 163 GB/s

16x FDR IB

28 PB Local Scratch >700 GB/s

2PB "DataWarp" I/O Accl. Layer >1.5 TB/s

32x FDR IB

14x QDR IB

136 GB/s

20 GB/s

5 GB/s

16 GB/s

Ethernet & IB Fabric
Science Friendly Security
Production Monitoring
Power Efficiency
WAN

/project — **6 PB DDN SFA12KE**

Sponsored Storage — **1.6 PB DDN SFA12KE**

/home — ***250 TB NetApp 5460***

HPSS — ***100 PB stored, 240 PB capacity, 40 years of community data***

2 x 10 Gb

1 x 100 Gb

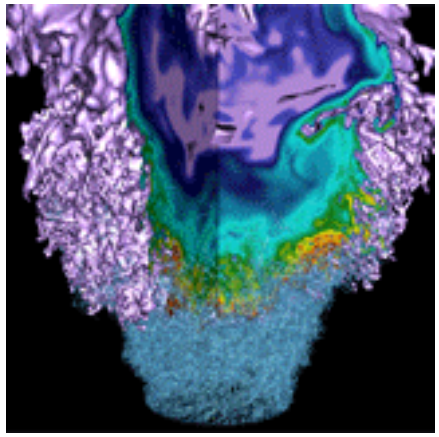
Software Defined Networking

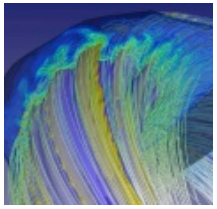
ESnet ENERGY SCIENCES NETWORK

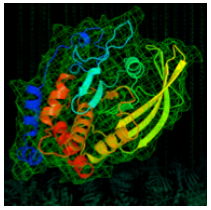
U.S. DEPARTMENT OF ENERGY | Office of Science

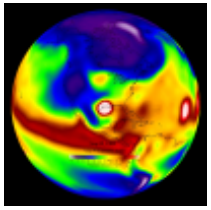
BERKELEY LAB

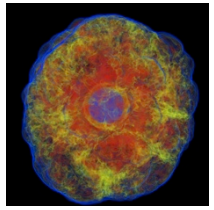
Current Storage Systems

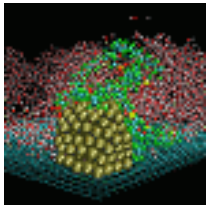












NERSC File Systems

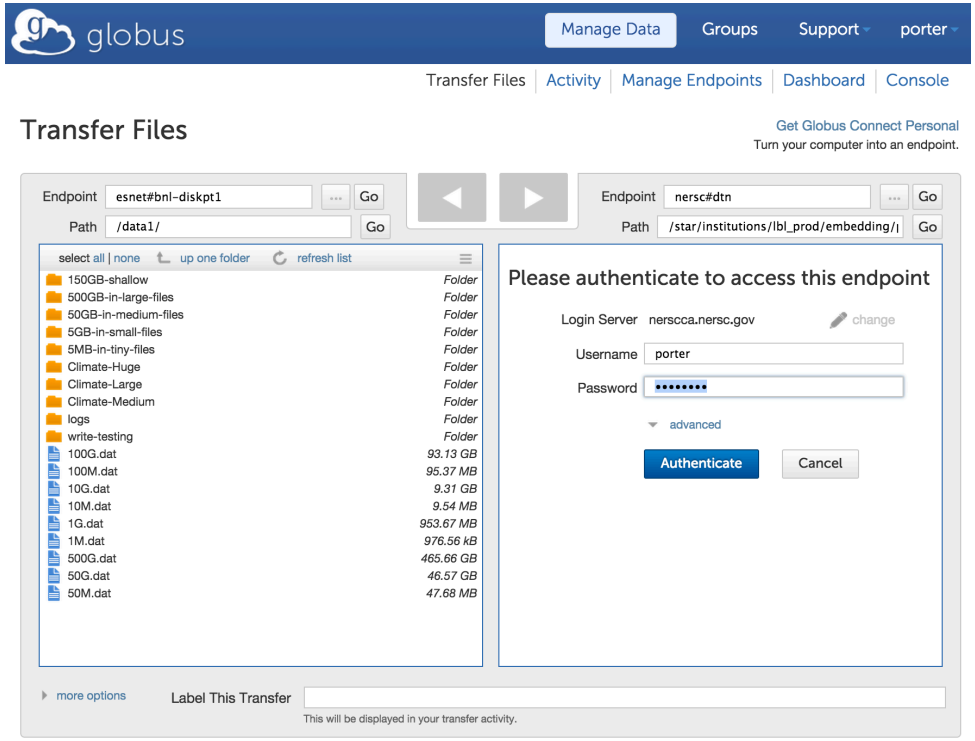
- **Large Compute systems have one or more Lustre scratch file system for local jobs**
  - Edison has 7.6 PB scratch with 163 Gbps aggregate file system bandwidth
  - Cori has 28 PB scratch with > 700 Gbps aggregate file system bandwidth
  - Files are periodically purged from the scratch file systems
- **NERSC has GPFS file systems mounted across all compute platforms**
  - 250 TB for home directories, optimized for small files, and high number of I/O operations
  - 6 PB for project directories. Projects are given 4TB by default. Optimized for larger files, and streaming I/O, and intended for data that is actively being used. There is no purge policy, however inactive projects may be moved to the tape archive
  - 1.6 PB for 'Sponsored File System Storage'. Separate file system, but with the same I/O characteristics as the project file system. This is a 'buy-in' program for projects that need additional space on disk. Program has a fixed price per TB, and a 5 year service agreement
  - DVS is used to project the GPFS file systems onto the Cray compute nodes

HPSS - NERSC Tape Archive

- **HPSS is the archival storage system for long term data retention**
  - Tiered storage system with a disk cache in front of a pool of tapes. The disk cache helps reduce access latencies and improves transfer speeds
  - Provides stable, scalable and high performance system for archiving user data. Contains 40 years of data archived by the scientific community
  - Third largest HPSS in the world in terms of bytes stored with 71 PB as of Jan. 2016
- **Transfers between the archive and file system use a transfer client - there is no direct file system interface**
  - We provide numerous clients: HSI, HTAR, FTP, pFTP, gridFTP, Globus Online, etc.

NERSC Data Transfer Node Services

Data transfer nodes (DTNs) are the resource for moving data between remote centers or instruments, and NERSC



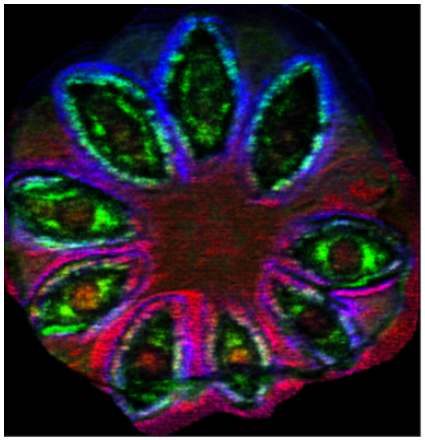
- They are a set of identical systems, with optimized hardware configuration, tuning and network connectivity for data transfers over the Wide Area Network.
- GPFS file systems are mounted on the DTNs, and they serve as Globus endpoints for transfers between the WAN and file systems, AND between the WAN and HPSS.

NERSC Science Gateway Nodes

- Science Gateway Nodes provide a web interface to data housed at NERSC
- Customers can build simple data publishing capabilities, or rich web interfaces and complex portals
- NEWT API provides access to NERSC resources. Backend databases and message queues are also provided
- Three flavors of gateway nodes: bare metal, virtual machine based, or Docker container based

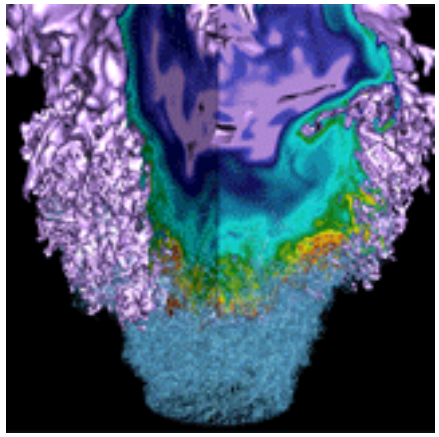
The Materials Project

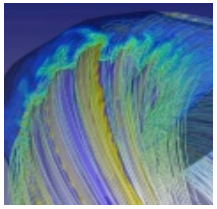
https://materialsproject.org

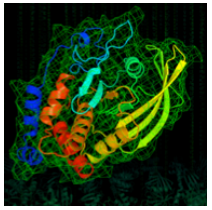


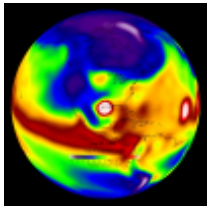
https://openmsi.nersc.gov

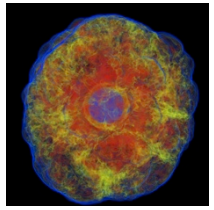
Improvements and changes over the last year

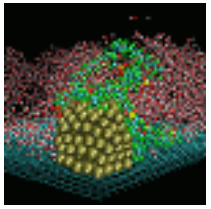








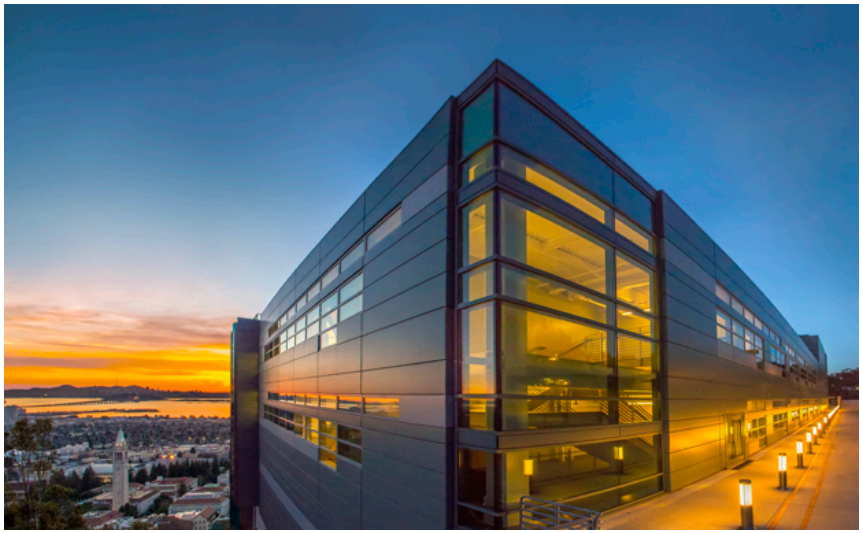




NERSC Global File systems moved to Wang Hall

Challenge and Opportunity

- Move Petabytes of data from the Oakland Scientific Facility to Wang Hall
- Minimize Impact to Users
- Get it done fast!



Implementation

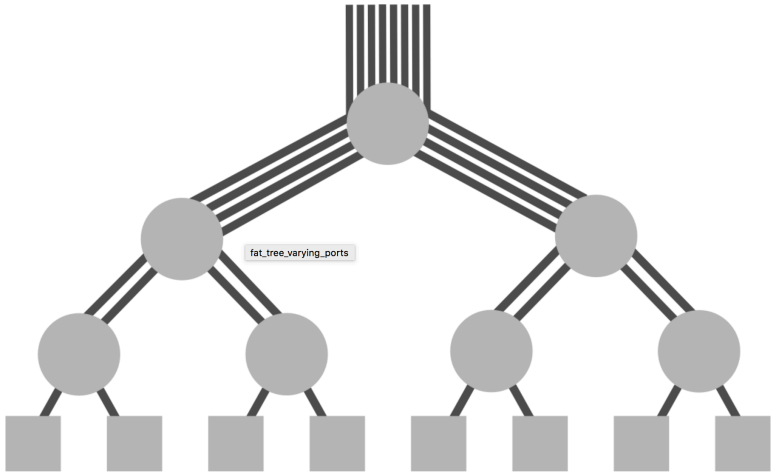
- NERSC and ESnet deployed one of the first 400Gb networks to be put into production by a research and education network
- NERSC architected and deployed 14 advanced Ethernet-to-Infiniband Routers at CRT to route between the WAN link and the internal storage network

Impact and Early Successes

- 400Gb link was used to live-migrate data from OSF to CRT and achieved sustained transfer speeds of 170 Gbps, roughly 1 PB per day (disk-to-disk)
- No Downtime!

Global File System changes

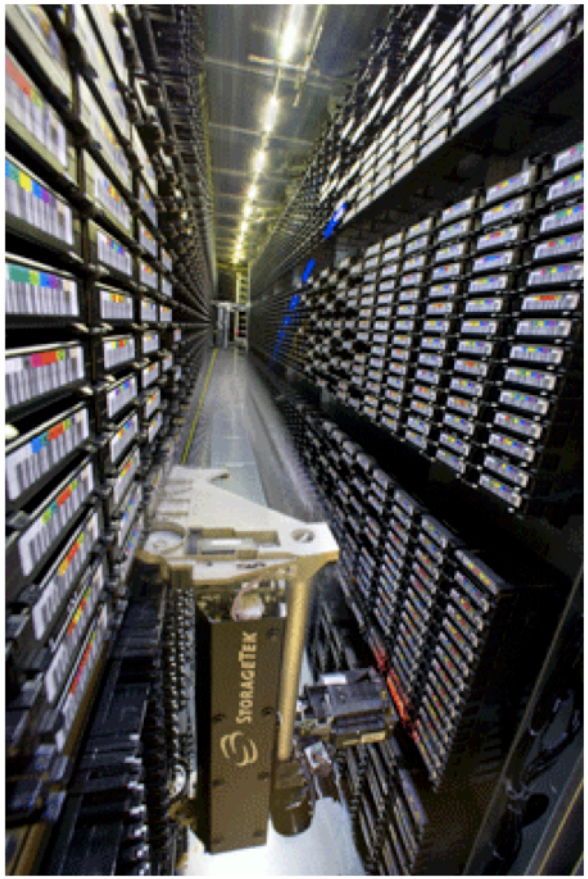
- All GPFS file servers are now connected using FDR Infiniband giving 56 Gbps bandwidth per server. The new storage network topology is a non-blocking fat tree



- Most file system access is over IB. Client access over Ethernet is through the highly tuned Ethernet-to-Infiniband routers used to transfer the file systems to CRT. The gateways act as failover pairs, and have no single point of failure
- The GPFS Global Scratch file system was retired in October 2015. Its functionality will be replaced by mounting the Cori scratch file system on Edison and the DTNs providing much greater capacity.

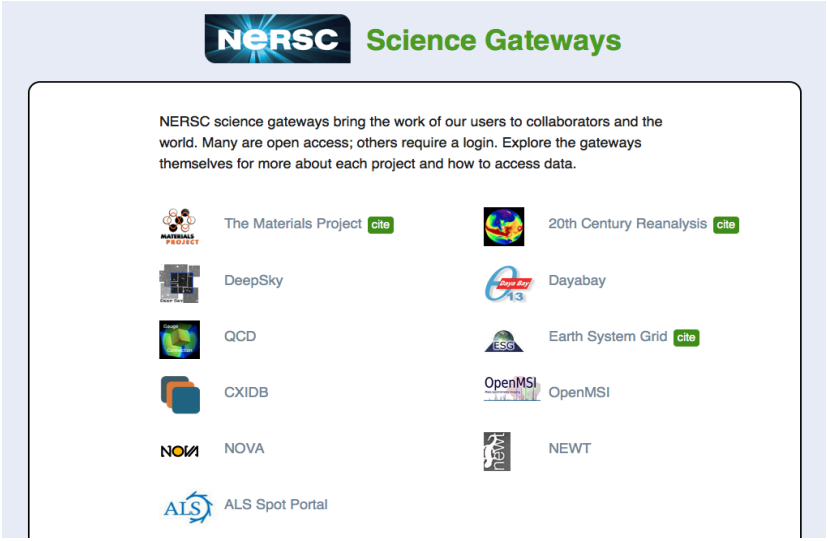
Increased Archive Disk Cache improves HPSS performance for users

- Disk cache increased by 10x from 200TB to over 2PB
- Before the increase, files stayed on disk cache for 2.5 days and now stay on 24.5 days (10x improvement)
- Impact for users is enormous, latency to tape is 90 seconds while disk cache is < 1 sec
- Of the files read, 75% are read within 30 days days of writing – disk cache close to optimal capacity

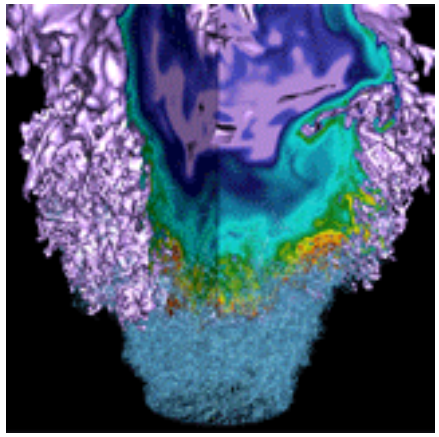


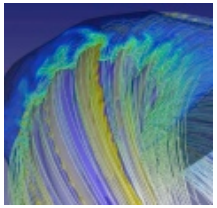
DTN and SGN changes

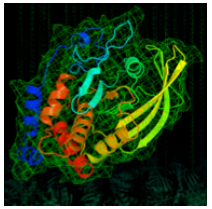
- **Completed deployment of 4 new DTN machines. Each machine is a Globus endpoint for GPFS file systems and HPSS**
- **SGNs are receiving a hardware refresh, and are being migrated to CRT.**

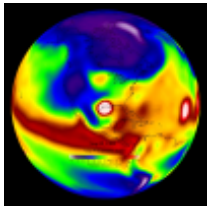


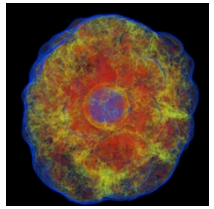
Improvements coming in 2016

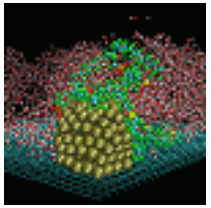












Hardware Refresh for HPSS

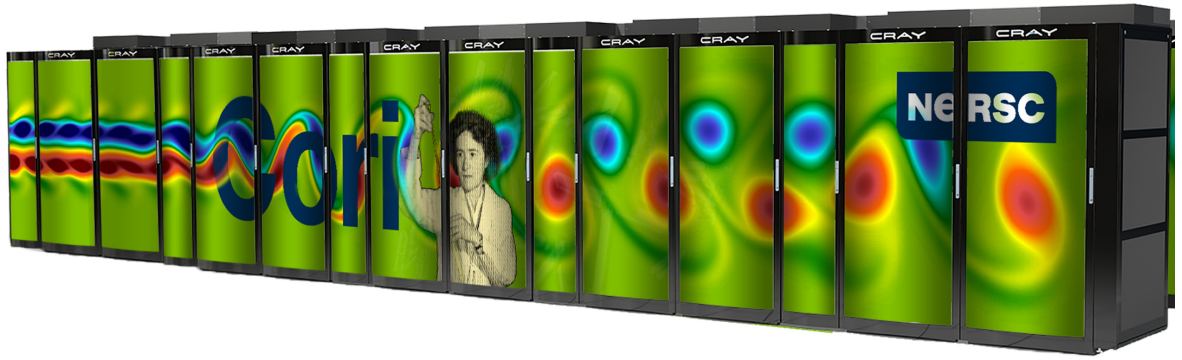
- 8 new data movers doubling, available bandwidth to the disk cache
- New core server will provide a metadata performance boost
- 1+ PB increase in disk cache bringing total to 3PB
- Upgrading tape technology to include 64 STK T10KD tape drives with 8.5TB cartridge capacity and 250MB/s data rate

DTNs and SGNs

- Deploying 8 additional DTNs to increase transfer speeds into and out of the center. Some DTNs may be allocated for specific purposes such as schedulable non-interactive use or specific workflows
- Underlying hardware for the SGN service is being completely re-architected to use containerized services. The design borrows from the cloud service model, and allows us to provide services that are quicker to deploy, more resilient and less impacted by system maintenance

File System Improvements

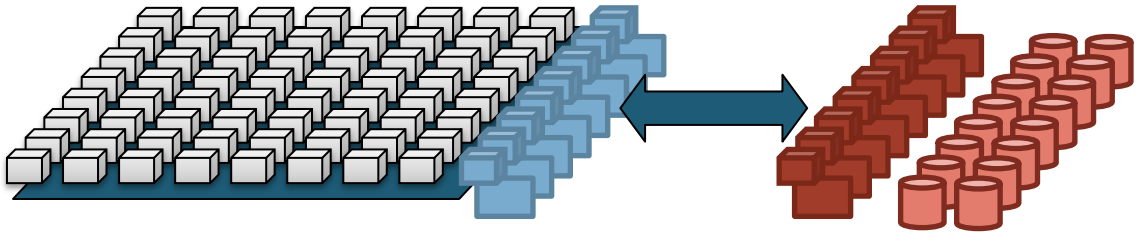
- **Cori's 28 PB scratch file system will be mounted on other NERSC compute resources**
  - First step is to mount on Edison, followed shortly by mounting on the NERSC DTNs
- **Cori NVRAM Flash Burst Buffer as I/O accelerator**
  - 1.5PB, 1.5 TB/sec
  - User can request I/O bandwidth and capacity at job launch time
  - Use cases include, out-of-core simulations, image processing, shared library applications, heavy read/write I/O applications



What will storage at NERSC look like in 2020?

- **This is an open question that NERSC is planning to address in 2016, in concert with the NERSC-9 procurement**
  - The Burst Buffer has added a layer to our storage hierarchy, can other layers be collapsed?
  - As SSDs become more prominent in HPC centers what does that imply for our disk based scratch file systems?
  - What capabilities will enhance data movement between storage tiers and increase user productivity?
- **NERSC is putting together a team to review center storage requirements and make recommendations**

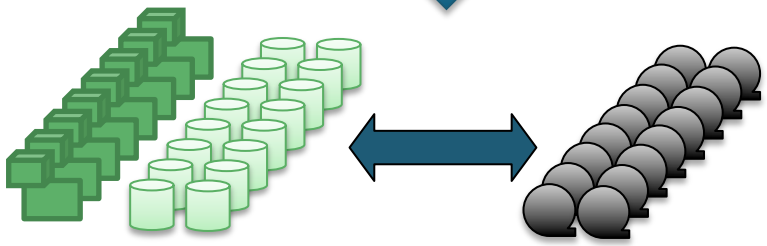
NERSC Storage Hierarchy with Burst Buffer



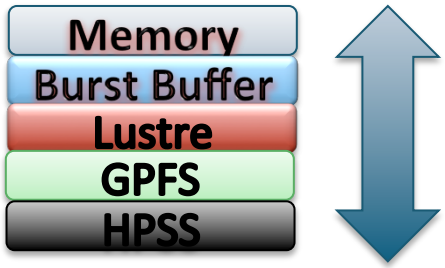
Compute Nodes | HPC Fabric MPI / Portals | IO Nodes Burst Buffer | SAN Fabric OFED | Storage Servers

On cluster

Off cluster



GPFS HPSS



Data movement through the Storage Hierarchy

Data Movement Through The Storage Hierarchy

"We could use better tools for coordinating dataflow between HPSS, project, local scratch, and eventually burst buffer. Currently doing this is a manual pain. – Stephen Bailey, 2015 NERSC User Survey

- **What will storage in the next large system look like? Will almost certainly be tiered.**
  - Flash-> disk-> tape?
  - POSIX Interface? Object interface?
  - Ease of data movement / uniform interface will become more important as storage tiers increase

- **Move towards more automated data transfers**
  - File system interface using GPFS-HPSS integration (GHI)
  - File system interface using Lustre-HPSS integration (Robin Hood)
  - Integration with batch scheduler to stage data from HPSS, and as parameter for determining when jobs are runnable

Thank you