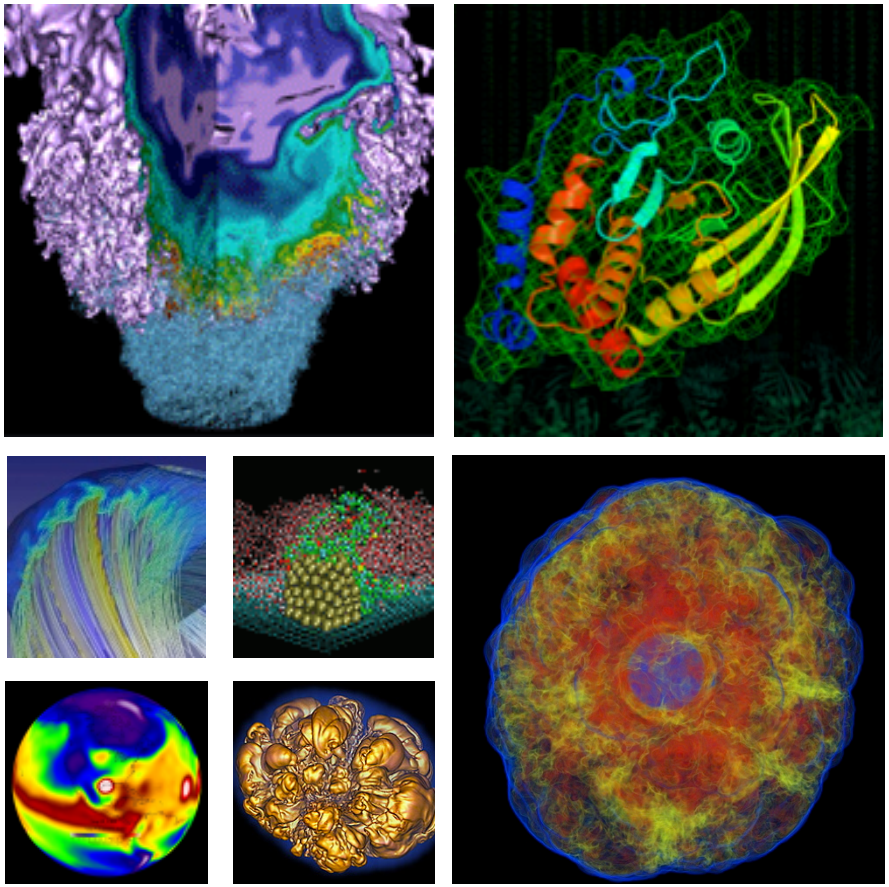
Storage at a Distance



Jason Hick

NERSC Storage Systems Group

What is storage at a distance?

- **Data is not local to the user/resource**
- **Processing and workflow needs are near real-time**
  - Don't want to wait until data transfer is complete
  - Need to see results, make adjustments, and try again
- **Network will become part of the instruments**
  - Telescopes and their data
  - Sequencers and their genome data
  - Light sources and their data
- **Is there an architecture/protocol that is necessary today for successfully providing storage at a distance?**
  - Ethernet vs. IB
  - ROCE vs. RDMA vs. IP

Use case 1: Instruments (beam lines)

- **Shift work (24hr coverage)**
  - Scientists fly in and use the instrument
- **Instrument "shot"**
  - Data ingest, normally locally due to pure bandwidth needs (10's of GB/sec)
- **View the data real-time**
  - Adjust the instrument and redo
- **Post analysis phase**
  - Computationally intensive
  - Potentially I/O intensive

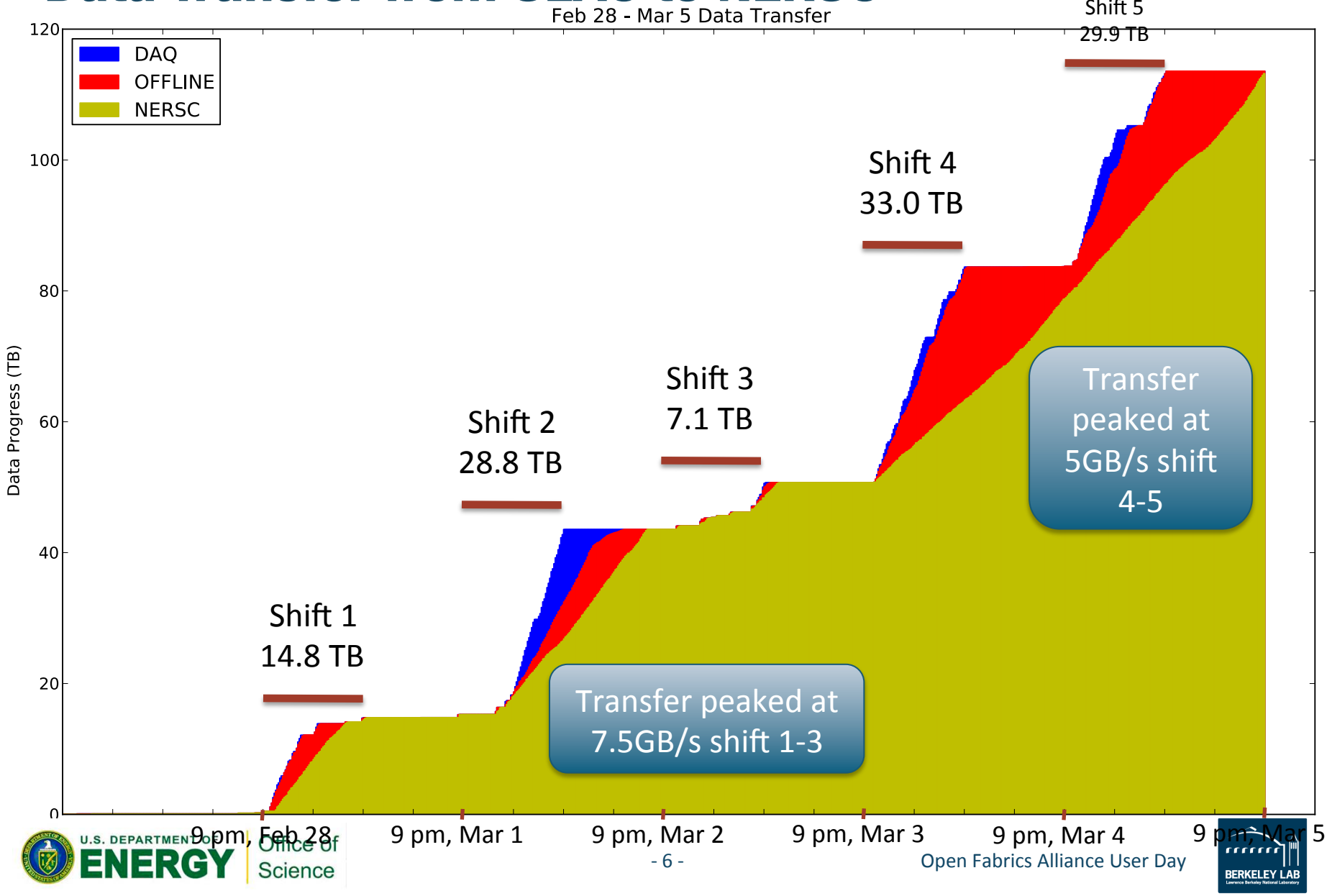
Science DMZ Background (Eli Dart)

- **The data mobility performance requirements for data intensive science are beyond what can typically be achieved using traditional methods**
  - Default host configurations (TCP, filesystems, NICs)
  - Converged network architectures designed for commodity traffic
  - Conventional security tools and policies
  - Legacy data transfer tools (e.g. SCP)
  - Wait-for-trouble-ticket operational models for network performance
- **The Science DMZ model describes a performance-based approach**
  - Dedicated infrastructure for wide-area data transfer
    - Well-configured data transfer hosts with modern tools
    - Capable network devices
    - High-performance data path which does not traverse commodity LAN
  - Proactive operational models that enable performance
    - Well-deployed test and measurement tools (perfSONAR)
    - Periodic testing to locate issues instead of waiting for users to complain
  - Security posture well-matched to high-performance science applications

NERSC-SLAC-ESnet recent example

- **Needed 125 node Linux system, requested "Data Intensive Pilot Award" for compute/storage in Fall 2012**
- **Ran LCLS instrument in March 2013 where 110TB moved via bbcp between facilities over Bay Area MAN (SLAC-to-NERSC) during multiple shifts**
- **Ultimately a success in moving data, but failed in keeping up with instrument/shifts**
  - Compute system admins to help keep nodes healthy
  - Storage admins to ensure data transfer and storage available
  - ESnet and NERSC network admins to ensure data transfer/ network health

Data Transfer from SLAC to NERSC



Lessons learned – future work

- **Bandwidth guarantees to enable shift planning**
  - Maximal use of current 10Gb, same as we upgrade to 100Gb
  - Can we couple a metroX solution providing RDMA over high latency connections using the optical wavegear?
- **Prefer real-time capability to avoid duplicate storage needs at multiple sites**
  - Data acquisition system
  - Post analysis storage depending on where compute cycles are available
- **Definitely moving from current "push" model to "pull" model**
  - Eliminates need for science team to have expertise/knowledge of both sites, accounts at both, know how to use both sites
  - Instead, will make the data available at SLAC, and consumers at NERSC (or wherever) can pull the data
- **Interested in a publish-subscribe model where data could be streamed from source site to remote consumers**

Use case 2: Highly connected sites

- **Building A or Site A and Building B or Site B**
  - Storage at Building A/Site A
  - Compute at Building B/Site B
- **CRT Building (Berkeley), 6 mile separation, OSF Building (Oakland)**
  - Considering using 100Gb Ethernet wavegear with EDR/FDR IB routers with extra buffers (to handle high latency) to provide 200-400Gb connectivity between buildings
  - This will enable us to provide reasonable bandwidth while minimizing the downtime for the facility
- **MLB.com**
  - Budget for private network at all baseball parks
  - HDTV recordings on numerous cameras
  - Saved locally (disk arrays), then copied via network to MLB.com
  - Video production and archiving
- **What are any organizations/users today doing to highly connect their facilities?**

IB SAN for GPFS currently underway

- **I/O or gateway servers for each compute system**
- **Consolidate cluster of GPFS storage servers (NSDs)**
- **Recently upgraded from QDR to FDR**
- **Using OpenSM as subnet manager**
- **Working through current limitations**
  - Conflict between OpenSM and UFM as SM
  - Interoperability between vendors/drivers
  - Monitoring, standing up UFM in monitoring mode and using custom perl program with perfquery (helps to identify bad cables)
  - GPFS supporting multiple RDMA networks on single NSD

Use case 3: Remote visualization

- **Are there scientific datasets being visualized remotely?**
- **Video conferencing technology**
  - Demand for video meetings is rapidly increasing, resolution is still very low

Summary

- **Current state-of-the-practice for storage @ distance is efficient bulk data movement**
- **Working towards, and scientists desiring remote access to data**
  - Simply mounting file systems at remote sites is challenging (not likely) for security issues it presents
  - But publish-subscribe model is attractive, need to learn more
- **Three use cases for this that I see**
  - Highly connecting 2 or more physical sites
  - Supporting instrument-based facilities and enabling distributed collaborations (and ultimately increasing discovery)
  - Remote visualization
- **Most interested to explore using RDMA with MetroX and 100Gb wavegear**