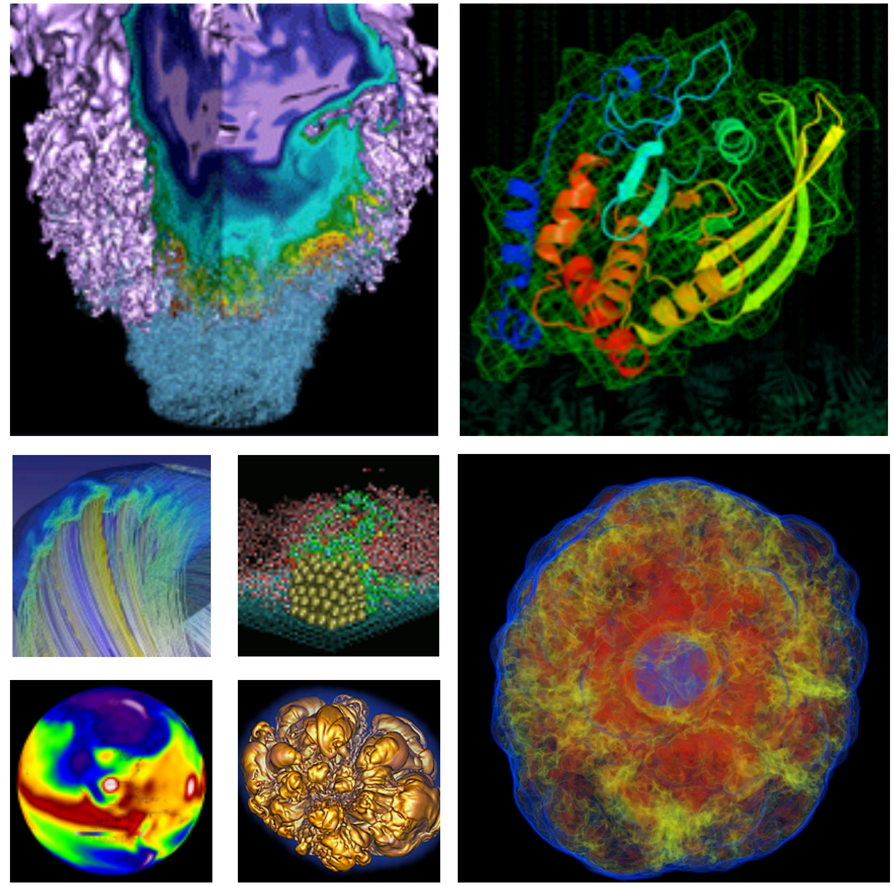
Future Directions and How SPXXL Can Help



NERSC

Jason Hick
Storage Systems Group

May 21, 2015

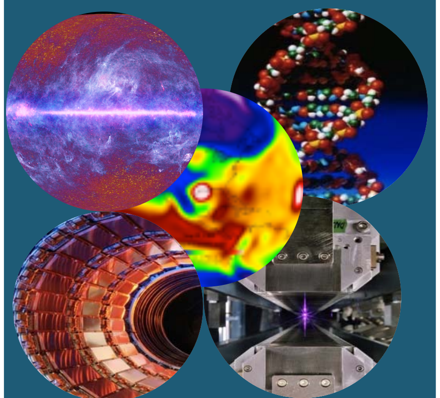
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Storage workloads at a user facility

Data Intensive

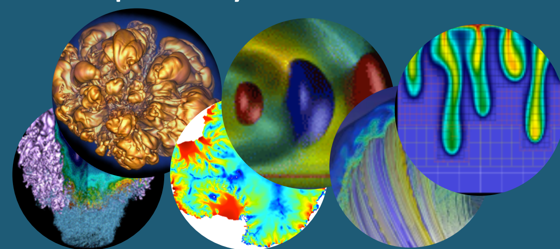
Experiments and Simulations



NERSC ingests, stores and analyzes data from Telescopes, Sequencers, Light sources, Particle Accelerators (LHC), Microscopes, and other scientific instruments

High Bandwidth

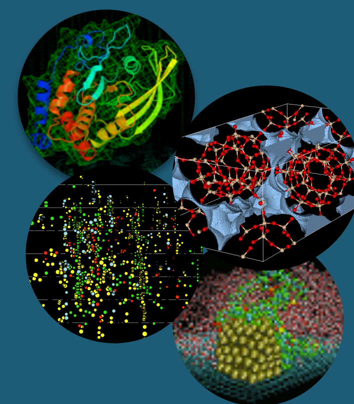
Capability Simulations



Petascale systems run simulations in Physics, Chemistry, Biology, Materials, Environment and Energy typically seeking peak bandwidth

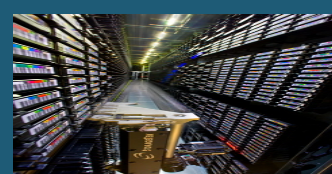
High Demand

Job Throughput



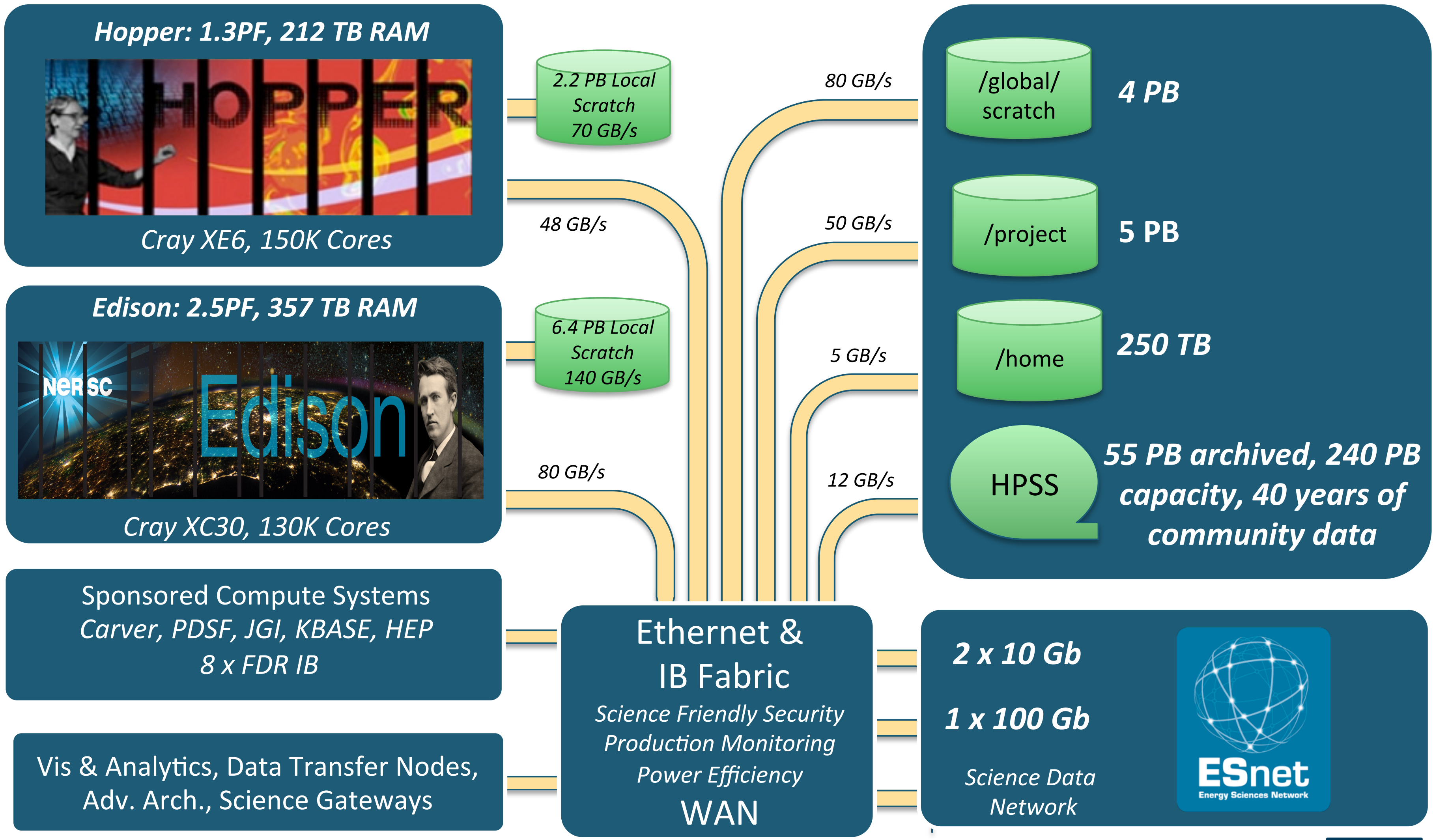
NERSC computer, storage and web systems support complex workflows that generate, index, and analyze data to understand proteins, and materials; the results are shared long-term with academics and industry through web interfaces

Centerwide Storage



Petascale file systems (NGF), and archival storage (HPSS)

The compute and storage systems 2015

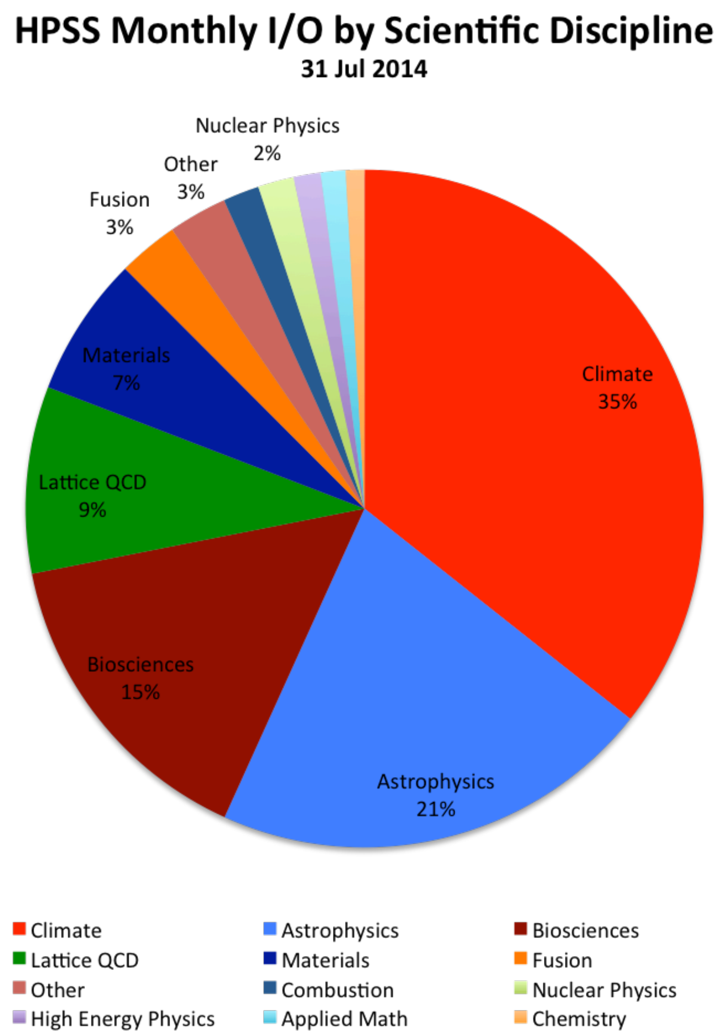


Storage systems and services

- **Parallel file systems (Lustre and GPFS) are primary storage to supercomputers**
  - Total of over 20 PBs of disk available to users
  - Some multi-PB parallel file systems backed up to HPSS (Parallel Incremental Backup System)
    - Has demonstrated it can identify backup candidates from 500M total files and process over 150TBs of backup data in a single day
- **Archival and backup systems (HPSS) are secondary storage for users**
  - 80 PBs of data stored, growing at >1.5PB per month
  - 30% of user IOs are read/retrieve requests from archival storage, so a very active archive
  - Focus on reliability of the system for user data by:
    - Deploying solutions to proactively monitor and maintain health of user data, and environmental parameters necessary for tape
    - Actively migrating/moving data within the system
- **Storage services highly utilized**
  - Inter-facility data transfers
  - Science gateways for sharing data with collaborators

Storage serving the entire range of science

- **Archival data (HPSS) shows a wide range of science data each month**
- **NGF provides at least 1TB project directories to every science repository at NERSC**
- **Users have a diverse set of storage needs that range from**
  - Low latency, small file
  - Long-term data processing
  - Long-term data sharing
  - High availability web hosting
  - Supercomputing job I/O



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

User demand is driving storage

- **DOE Requirements workshops identified a growing gap from 1.6x in 2014-2015 timeframe to 3.5x gap in the 2017 timeframe**
- **We use user/project-level quotas in all our file systems. Recent requests for increase are O(10-100TB)**
- **Our largest requests have come as need arises**
  - Increase in opportunistic data analytics/analysis for science discoveries

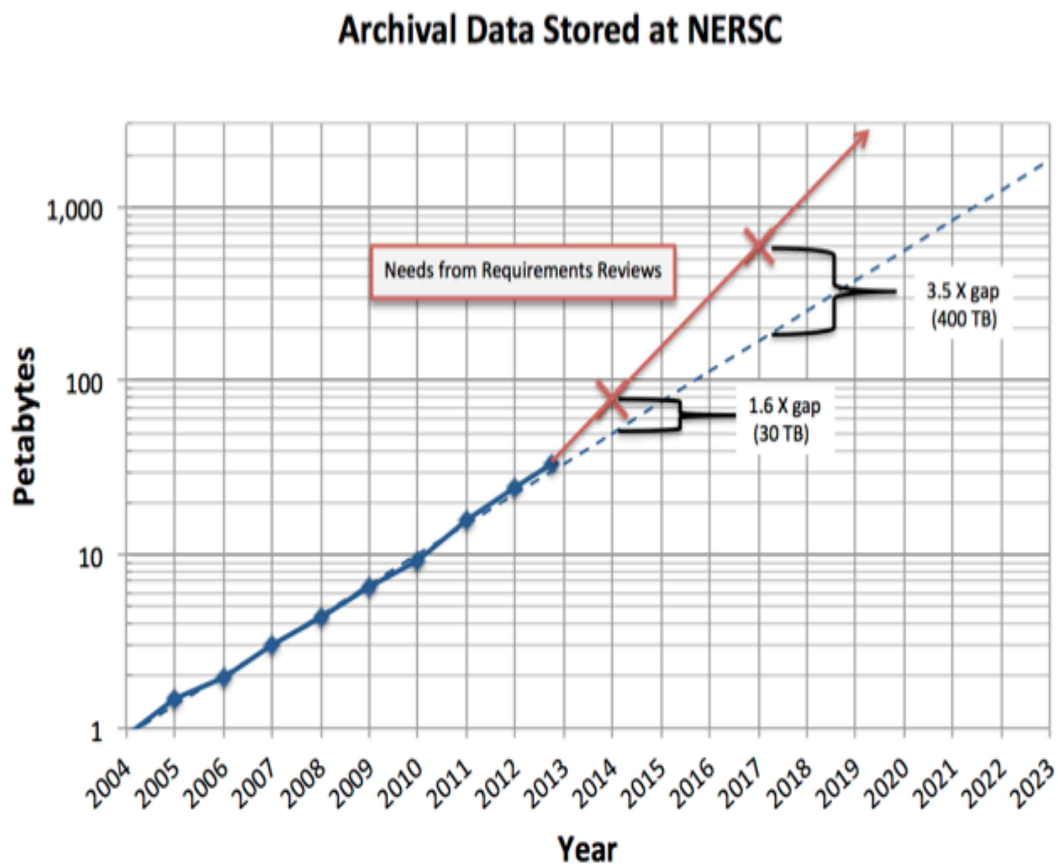
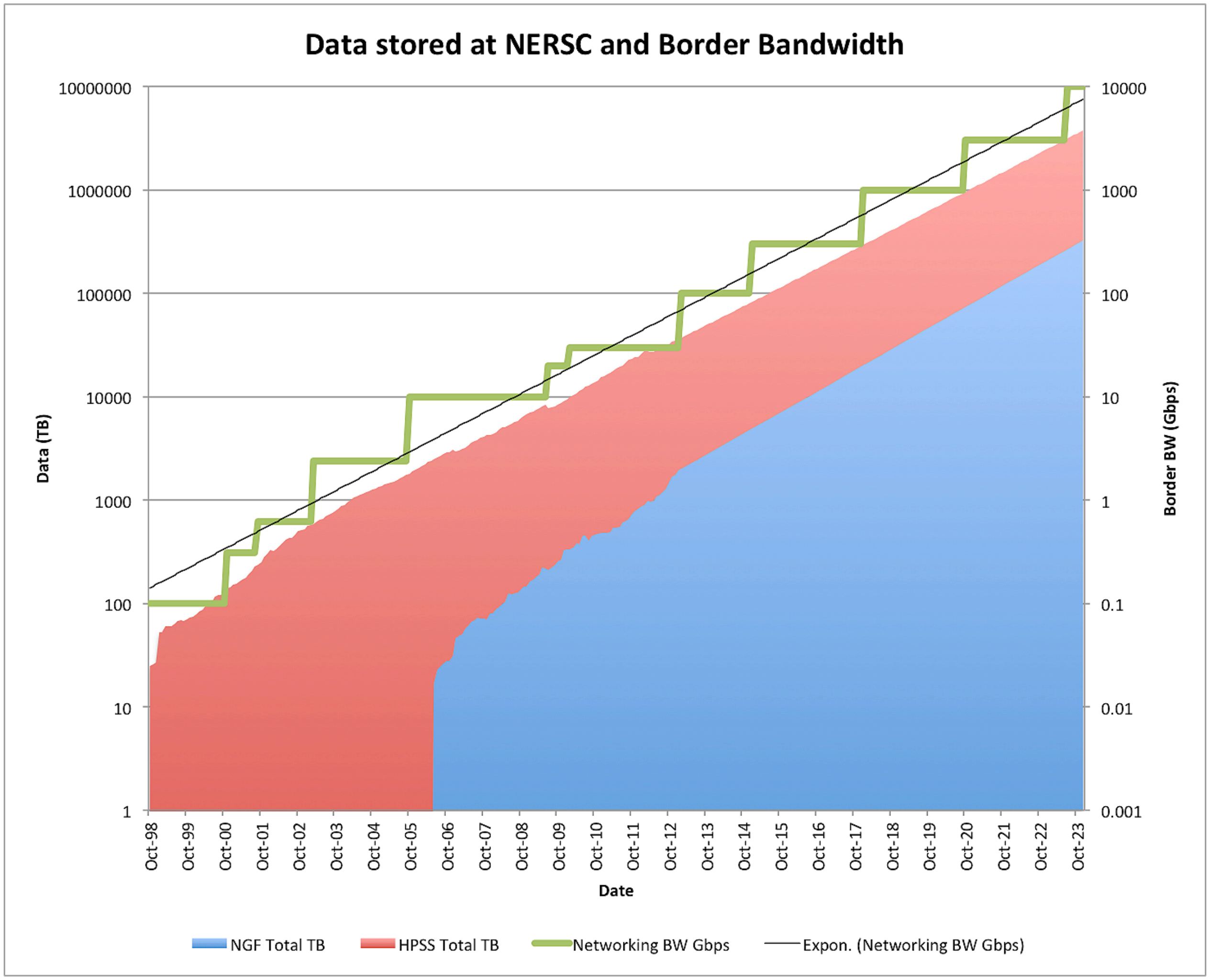


Figure 5. Scientific data stored on the NERSC HPSS archival storage system (solid blue line), and research teams' estimates of their needs for 2014 and 2017. The blue dashed line is a fit to actual usage.

Border networking needs for storage



- The ethernet network capacity correlates to HPSS capacity at NERSC
- This suggests that NERSC will want to deploy 400Gb Ethernet by 2017 and Terabit networking in 2022.
- Even with Terabit networking, we have a challenge to meet the ~10 Exabyte demand on storage capacity in 2023

Primary storage network needs

- **Centerwide storage network is InifiniBand today and must support a varied workload from multiple compute systems simultaneously**
- **IB topology is soon to be a 2-level fat-tree for maximizing bandwidth between each compute cluster and the file system**
  - Traffic moves predominantly between a given compute system and the file system servers
- **NERSC will require EDR technology in production by 2017 to avoid topology change**

Primary Storage Software/Hardware Needs

- **Rolling upgrades to everything**
  - This supports our primary "evergreen" approach to center-wide file systems that users highly value
- **Generally refresh disk hardware at five years, buy storage increments annually**
  - Require software and methods to support this without downtimes (restriping/rebalancing data, adding/deleting capacity)

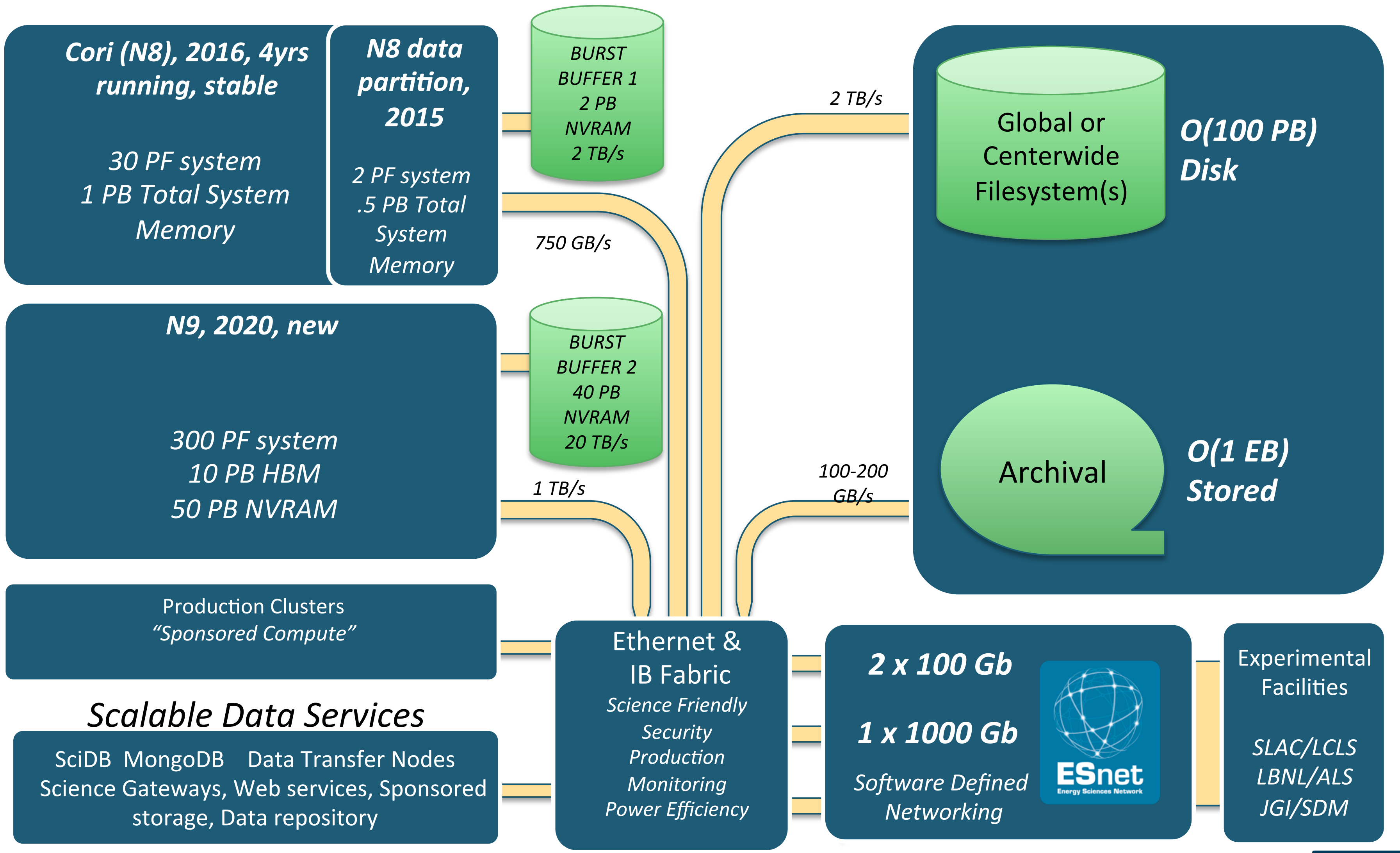
Summary of requirements

- **User requirements:**
  - Keep up with demand for NGF and HPSS resources
    - Develop storage business practices to support scale-out of NGF and HPSS systems
  - Continue providing highest capability for cluster's job I/O
    - Currently, Lustre scratch disk file systems procured through major system vendor
- **Storage networking requirements:**
  - Border bandwidth should be 400Gb ethernet by 2017 & 1Tb ethernet by 2022
  - EDR InfiniBand for centerwide file system by 2017 to maintain server/switch footprint and 2-level network topology
- **Storage hardware/software requirements:**
  - Support an "evergreen" approach

Problems with our storage

- **File and disk systems rarely perform at peak capability**
  - Observed I/O is bursty and non-optimal for storage system
- **Individual file system maintenance is struggling to scale with capacity needs**
  - fsck performance, requiring offline fsck
  - Unhealthy components causing unmounts
  - Lack of advanced health monitoring/determination solution
- **Metadata rates are slow and ultimately don't scale well**
  - Numbers of files per directory
  - Metadata performance can determine whether a file system is usable or not
- **These lead us to segregate and containerize our file systems to serve different workloads**
  - Scratch 1-3, project vs. scratch, homes vs. project

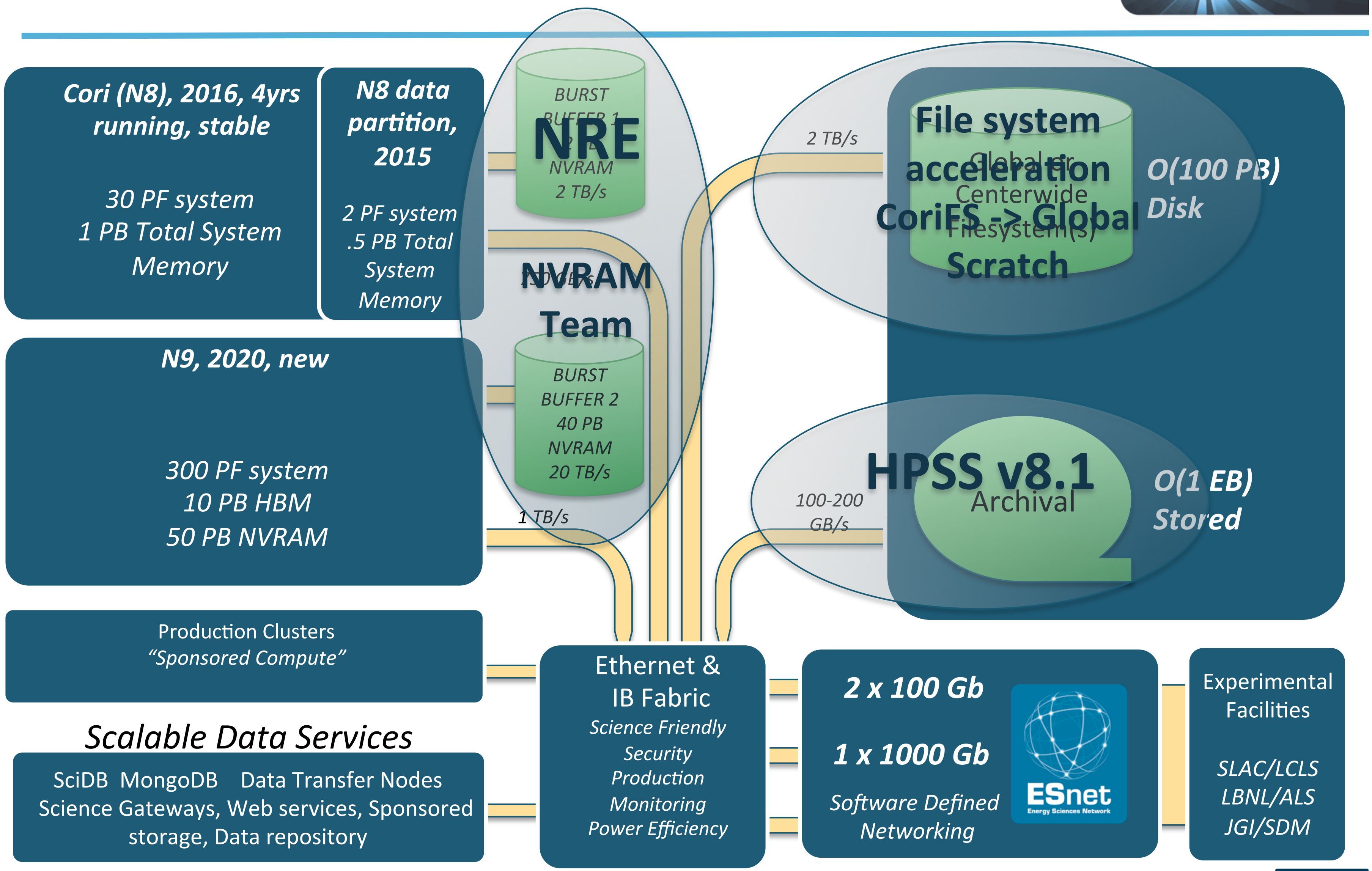
NERSC 2020



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC 2020



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Addressing our problems

- **Supercomputer's primary storage needs the most attention**
  - Bursty and non-optimal I/O suggest flash storage
  - Limits of traditional file systems suggest new software
- **Parallel file systems (PFS) will increasingly serve as center-wide assets**
  - Need software/tools to migrate data to/from non-volatile storage and PFS
  - Look to solutions that allow maximum flexibility for broader multi-vendor infrastructure
- **Need responsive technical engagements for feature development**
  - Storage industry is expanding, embrace it and support feature development using a collaborative approach
  - Require a better process for submitting new features and learning status
  - Consider enabling detailed testing of GPFS at customer sites
    - Better production scale testing of GPFS
    - Aid in feature development and deployment

How SPXXL can help

- **GPFS feature development**
  - From concept & design, and modest involvement in development
- **Support rapid defect resolution for production**
  - Enable GPFS testing at customer sites
- **GPFS maintenance or tool improvements**
  - fsck performance at scale or estimate of completion
  - fsck progress indication
  - Ability to gracefully handle suspended/down NSDs to enable hardware maintenance (avoid dismounting file system)
- **GPFS memory management enhancements**
  - Read caching
  - User space instead of kernel space