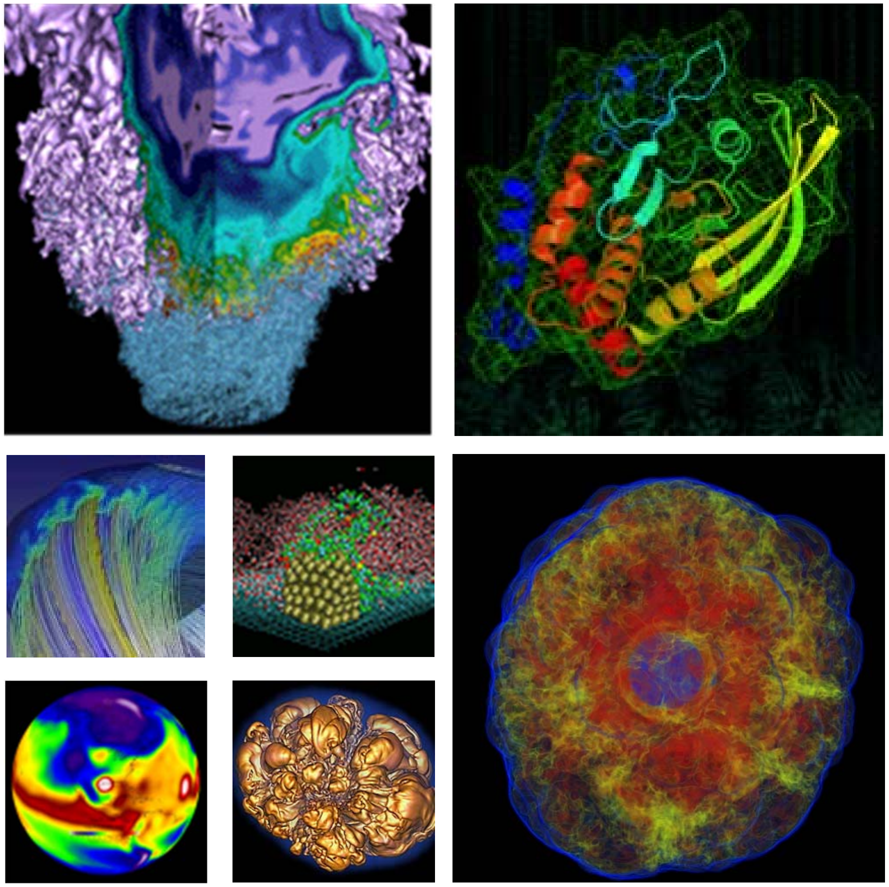
A Storage Outlook for Energy Sciences:

Data Intensive, Throughput and Exascale Computing



Jason Hick

Storage Systems Group

October, 2013

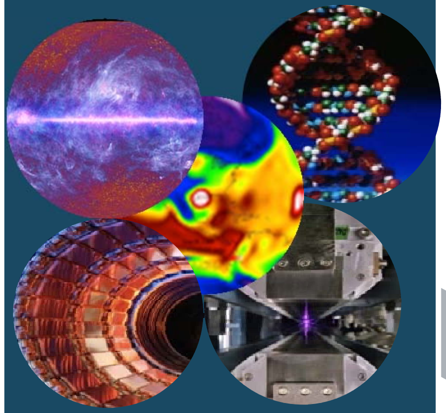
National Energy Research Scientific Computing Center (NERSC)

- **Located at Berkeley Lab**
- **User facility to support 6 DOE Offices of Science:**
  - 5000 users, 700 research projects
  - 48 states; 65% from universities
  - Hundreds of users each day
  - ~1500 publications per year
  - With services for consulting, data analysis and more

Types of Computing at NERSC

Data Intensive

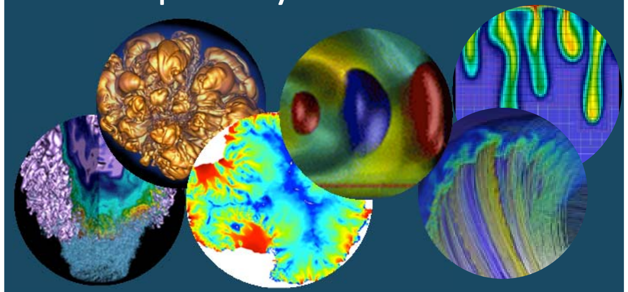
Experiments and Simulations



NERSC ingests, stores and analyzes data from Telescopes, Sequencers, Light sources, Particle Accelerators (LHC), climate, and environment

Large Scale

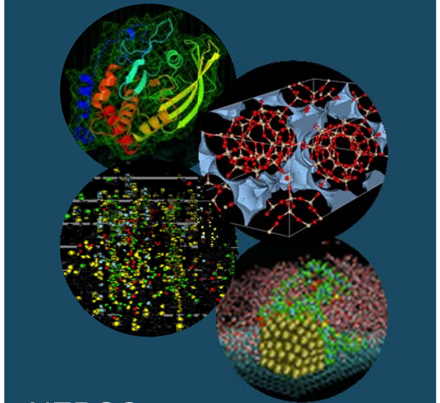
Capability Simulations



Petascale systems run simulations in Physics, Chemistry, Biology, Materials, Environment and Energy at NERSC

High Volume

Job Throughput

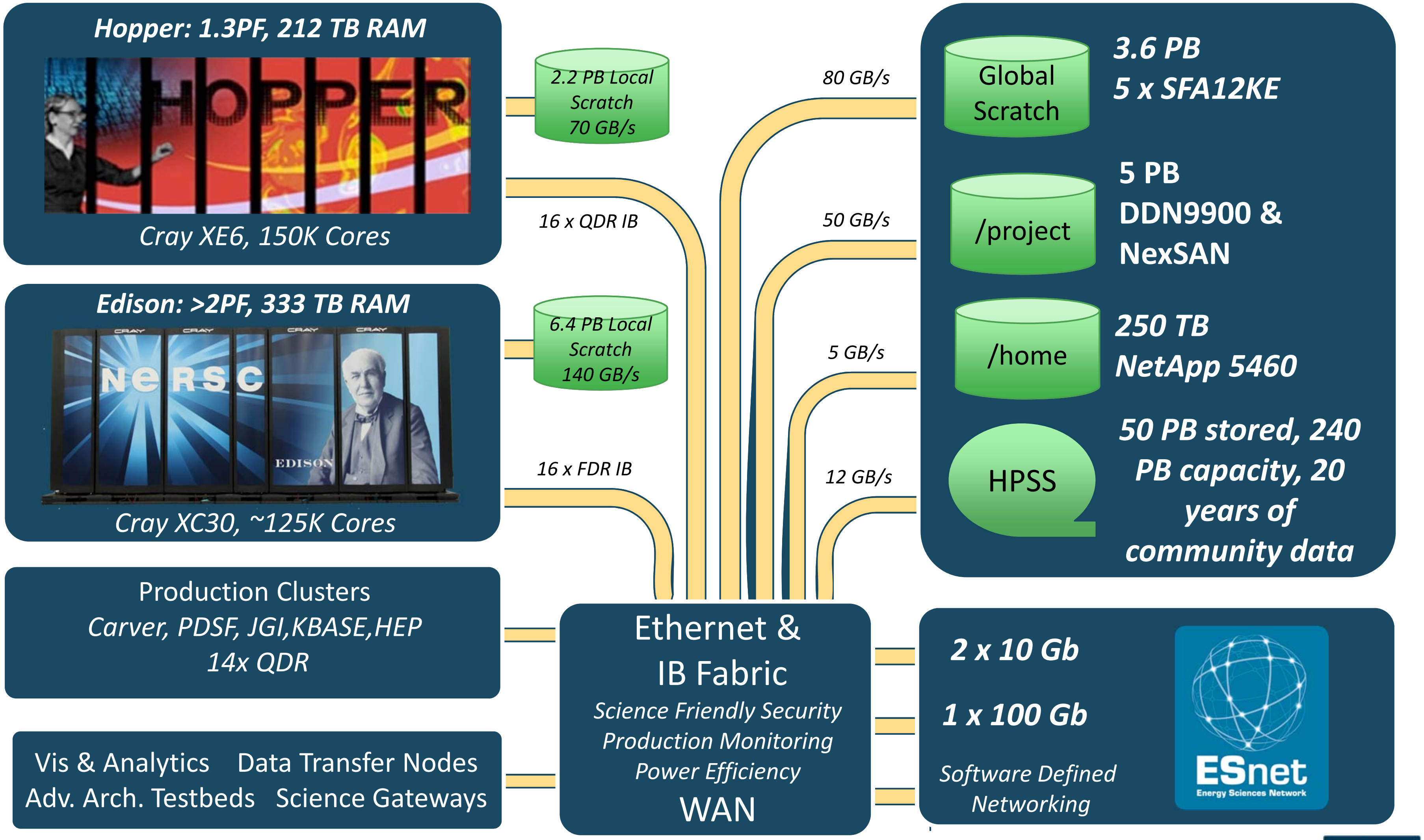


NERSC computer, storage and web systems support complex workflows that run thousands of simulations to screen materials, proteins, structures and more; the results are shared with academics and industry through a web interface

NERSC

Petascale Computing, Petabyte Storage, and Expert Scientific Consulting

The compute and storage systems 2013



Hopper: 1.3PF, 212 TB RAM

Cray XE6, 150K Cores

Edison: >2PF, 333 TB RAM

Cray XC30, ~125K Cores

Production Clusters
Carver, PDSF, JGI,KBASE,HEP
14x QDR

Vis & Analytics Data Transfer Nodes
Adv. Arch. Testbeds Science Gateways

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

Global Scratch — ***3.6 PB***
5 x SFA12KE

/project — **5 PB**
DDN9900 &
NexSAN

/home — ***250 TB***
NetApp 5460

HPSS — ***50 PB stored, 240 PB capacity, 20 years of community data***

2 x 10 Gb

1 x 100 Gb

Software Defined Networking

Focusing on storage at our facility

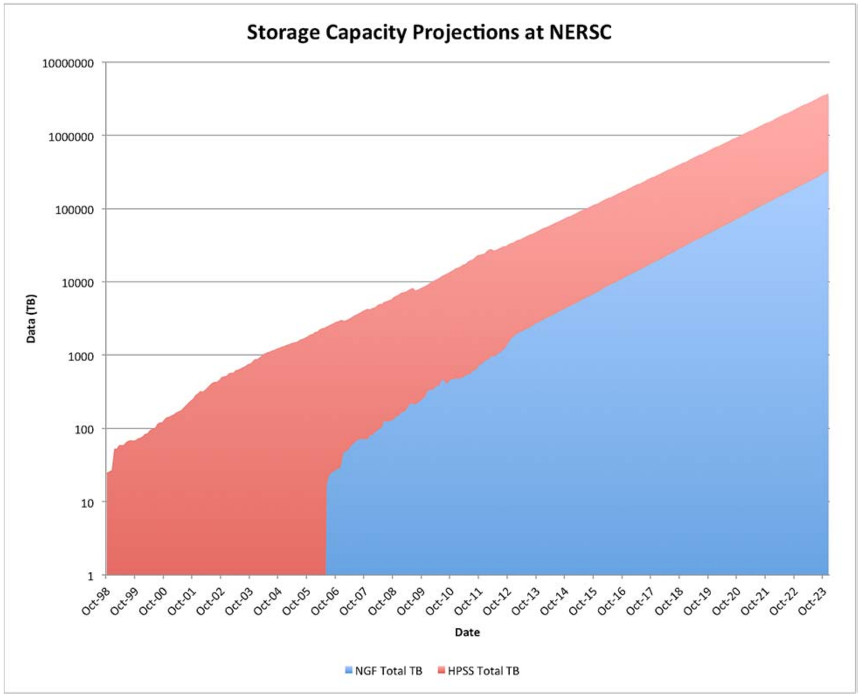
- **Parallel file systems (Lustre and GPFS) are primary storage to supercomputers**
  - Total of about 20 PBs of disk available to users
  - Some multi-PB parallel file systems backed up to HPSS (Parallel Incremental Backup System)
    - Has demonstrated it can process over 100TBs of backup data in a single day currently using direct-to-tape with 12 T10KC tape drives
    - On average, we complete a restore for a user about every other week

- **Archival and backup systems (HPSS) are secondary storage for users**
  - 50 PBs of data stored, growing at >1PB 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
    - Solutions to help us monitor and maintain health of user data
    - Technology choices (SAS disk and Enterprise tape drives)
    - Provides us with five 9's of reliability in practice (http://www.activearchive.com/content/nersc-accelerates-data-access-and-exceeds-reliability-standards-tape-based-active-archive)

Data intensive science challenge

- **Scaling global storage to keep up with demand**
  - Seven large science projects requested increases totaling 2.2PB in October 2013
  - Requests on recent user allocations for FY14 were about 2.5X what we can currently provide
  - Users are projecting PB allocation requests in the near future
    - 1PB in 2017 for NERSC global storage by a single BES project
    - 1PB in 2015 for NERSC global storage by two HEP projects
  - Need a business and technical process to support this
    - Standard charging for storage (disk & tape)
    - Do we continue to scale up each storage system or look
- **Supporting workload for analysis of instrument data**
  - Normal requires bulk data transfer from acquisition storage onto global storage
  - Workload can saturate system and cause issues for shared users
  - Beam line or end station work schedule drives bandwidth needs

Addressing large scale capability simulations

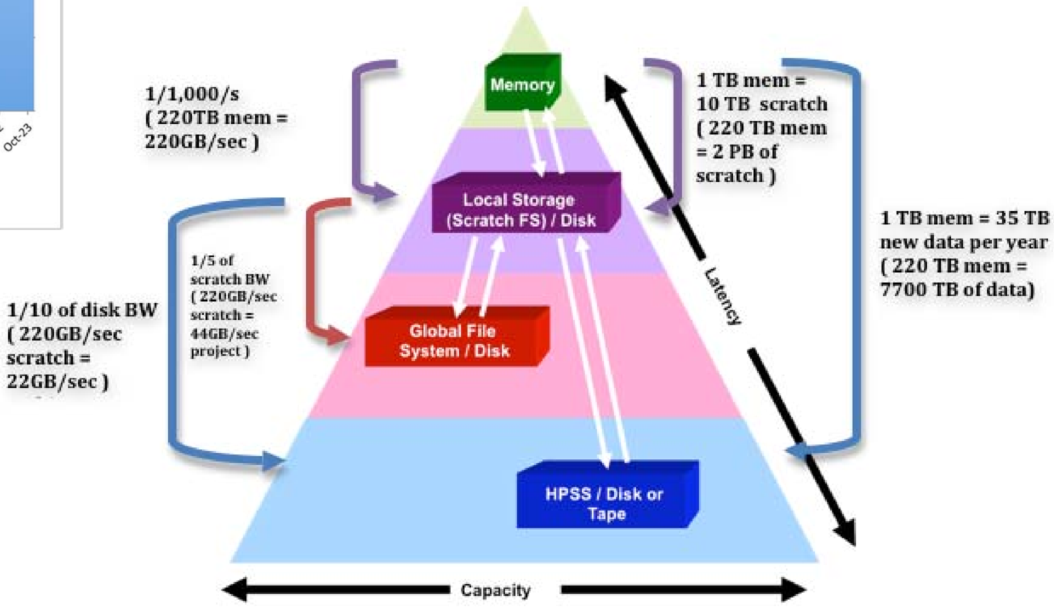


Capacity determined by looking at utilization over time to forecast need for disk (blue) and tape (red)

Capacity increases expected for disk and tape technology support a feasible system supporting exascale computing

Bandwidth desired typically determined by capacity of system memory (top of storage hierarchy)

Bandwidth demand will be the primary challenge for storage systems supporting exascale computing



1/1,000/s
(220TB mem =
220GB/sec)

1 TB mem =
10 TB scratch
(220 TB mem
= 2 PB of
scratch)

1/5 of
scratch BW
(220GB/sec
scratch =
44GB/sec
project)

1/10 of disk BW
(220GB/sec
scratch =
22GB/sec)

1 TB mem = 35 TB
new data per year
(220 TB mem =
7700 TB of data)

Specific data system needs for exascale

- **Terabit networking speeds for archival storage by 2020**
  - To enable data ingest/growth demand projected
- **Continued improvements in software design and performance to support extreme scale**
  - File system design
  - Archive system design
- **Successfully integrate new storage technology between memory and disk**
  - NAND Flash or NVRAM (Probe, memristor, …)
  - To enable bandwidth of primary storage to keep up with Exascale systems (total memory & nodes)

Supporting throughput computing

- **Experiment steering and quality of service**
  - Reservations and scheduling optimizations
- **Data ingest and processing**
  - Complicated workflows are sensitive to transitory events
  - Hangs better than failures (even if fails and recovers)
  - Determining how best to stage data to storage required (between memory and file system, and file system and HPSS)
- **Data management is challenging**
- **Ultimately bandwidth isn't as important as capacity**
  - Steps in workflow are typically very short
- **For HPSS/tape systems, the challenge is high usability and robust automation**
  - They want automated migration and staging between storage systems

Summary

- **NERSC currently manages 3 distinct types of computing**
  - Data intensive (Big Data) predominantly requiring extreme scaling of data resources and special workload considerations
  - Large capability (simulations) for which future bandwidth demands will require integration of a new tier of storage
  - High volume (high throughput) requiring workflow and usability improvements for data system software
- **Capacity, reliability and bandwidth of tape and disk devices are key factors to enabling continued use in high scale storage systems**
  - NERSC requires doubling of tape and disk drive capacity every two years
  - Media reuse, compression, and reliability of tape are critical differentiators for using tape in archive systems
  - Reliability of tape must continue to lead other storage technologies
- **File systems and HSM/archival systems will continue to use tape, disk, and integrate a new bandwidth-oriented storage tier (SSD/NVRAM)**
  - SSD/NVRAM will handle ingest/bandwidth requirements
  - Storage software will change to make use of new storage tier and may cause application/workflow change

National Energy Research Scientific Computing Center