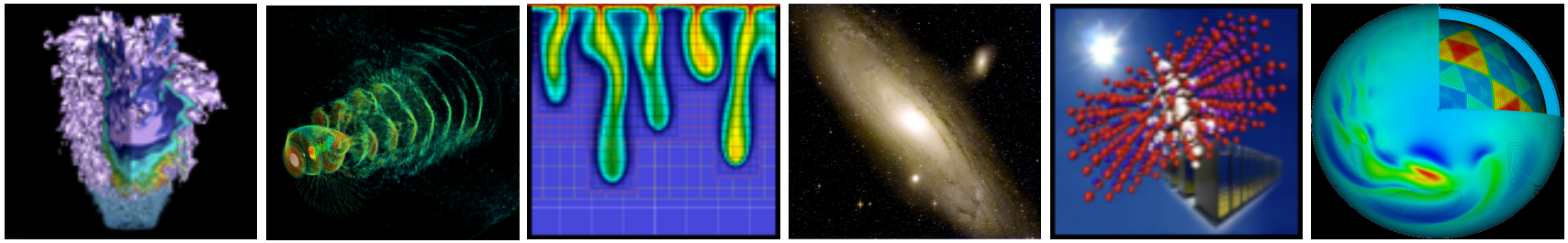


Storage Supporting DOE Science

Jason Hick
jhick@lbl.gov
NERSC LBNL
http://www.nersc.gov/nusers/systems/HPSS/
http://www.nersc.gov/nusers/systems/NGF/

May 12, 2011

U.S. DEPARTMENT OF ENERGY | Office of Science

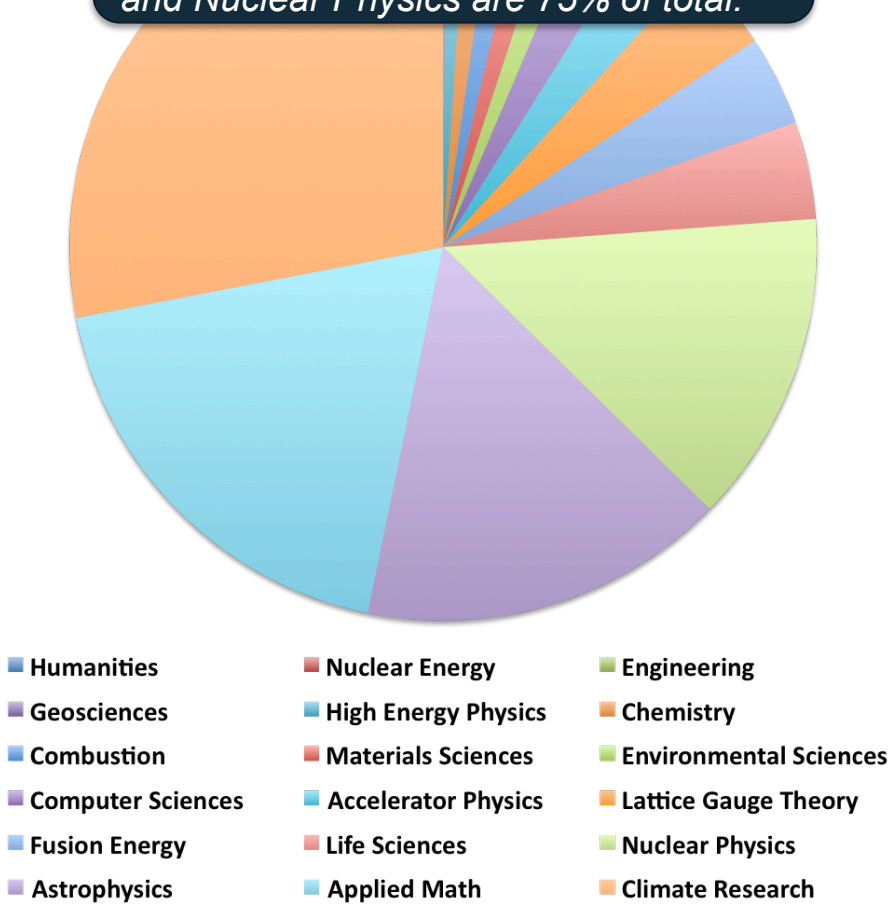
NERSC National Energy Research Scientific Computing Center

BERKELEY LAB Lawrence Berkeley National Laboratory

The Production Facility for DOE Office of Science

- **Operated by UC for the DOE**
- **NERSC serves a large population**
  - Approximately 4000 users, 400 projects, 500 codes
  - Focus on "unique" resources
    - High-end computing systems
    - High-end storage systems
      - Large shared GPFS (a.k.a. NGF)
      - Large archive (a.k.a. HPSS)
    - Interface to high speed networking
      - ESnet soon to be 100Gb (a.k.a. ANI)
- **Our mission is to accelerate the pace of discovery by providing high performance computing, data, and communication services to the DOE Office of Science community.**

2010 storage usage by area of science. Climate, Applied Math, Astrophysics, and Nuclear Physics are 75% of total.



Storage Systems Group: Focused on Data Needs

- **NERSC is a net importer of data.**
  - Since 2007, we receive more bytes than transfer out to other sites. NERSC is a net importer of data.
  - Leadership in inter-site data transfers (Data Transfer Working Group).
- **We present efficient center-wide storage solutions.**
  - NGF aids in minimizing multiple online copies of data, and reduces extraneous data movement between systems.
  - HPSS enables exponential user data growth without exponential impact to the facility.
- **Partnering with ANL and ESnet to advance HPC network capabilities.**
  - Magellan ANL and NERSC cloud research infrastructure.
  - Advanced Networking Initiative with ESnet (100Gb Ethernet).

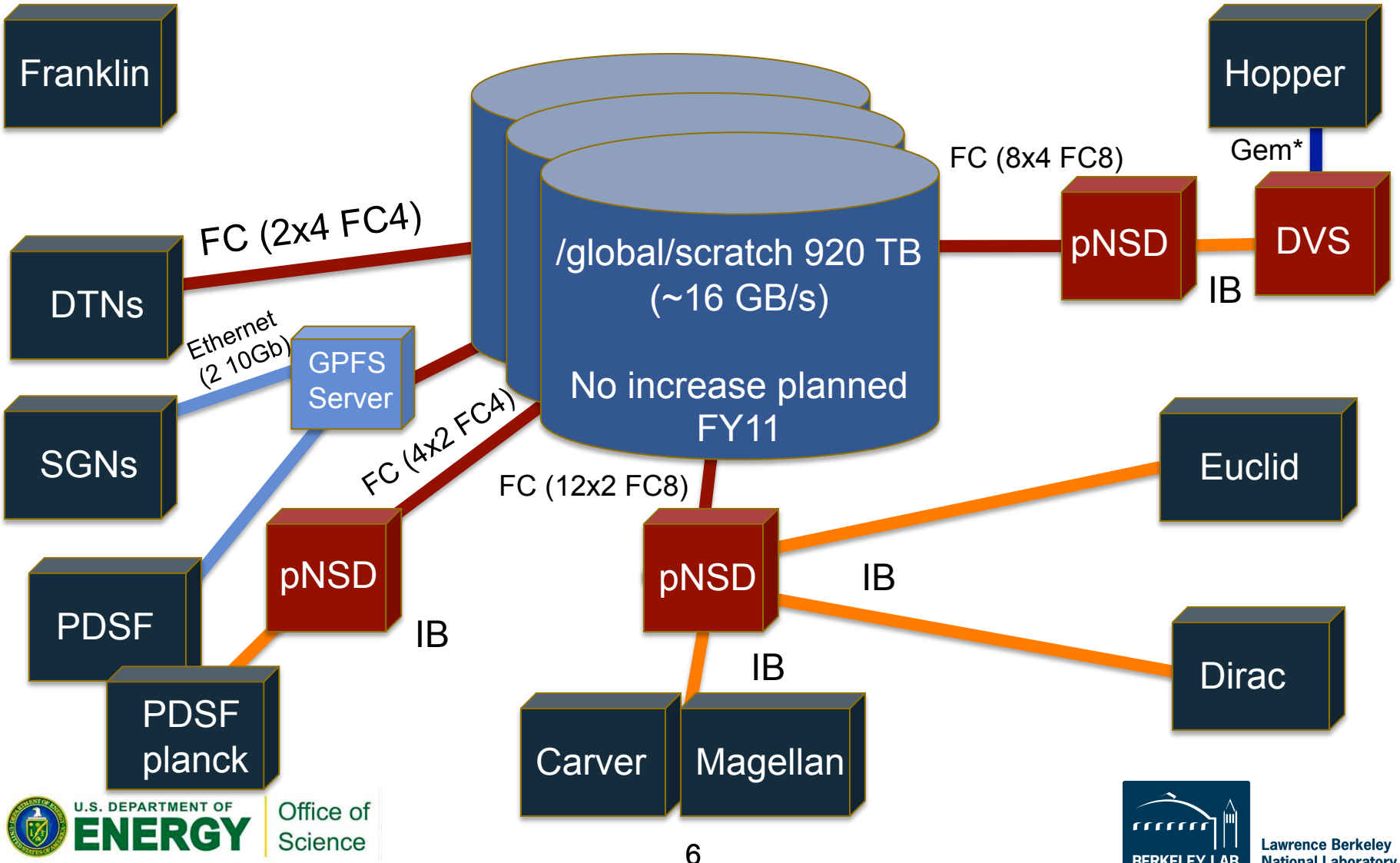
Storage, Who We Are

- Wayne Hurlbert and Nick Balthaser: HPSS system analysts
- Damian Hazen and Mike Welcome: HPSS developers
- Matt Andrews: GPFS backup developer
- Will Baird: Data transfer system analyst
- Rei Lee and Greg Butler: GPFS system analysts

Center-wide File Systems

- **/project is for sharing and long-term residence of data on all NERSC computational systems.**
  - 8% monthly growth, ~100% growth per year
  - Not purged, quota enforced (1TB default per project), backed up daily
  - Serves 200 projects over FC4/8
  - 1.6PB total capacity
  - ~5TB average daily IO
- **/global/homes provides a common login environment for users across systems.**
  - 15% monthly growth
  - Not purged but archived, quota enforced (40GB per user), backed up daily
  - Serves 4000 users, 400 per day over Ethernet
  - 50TB total capacity
  - 100's of GBs average daily IO
- **/global/scratch provides high bandwidth and capacity data across systems.**
  - Purged, quota enforced (40TB per user), not backed up
  - Serves 4000 users over FC8
  - 1PB total capacity

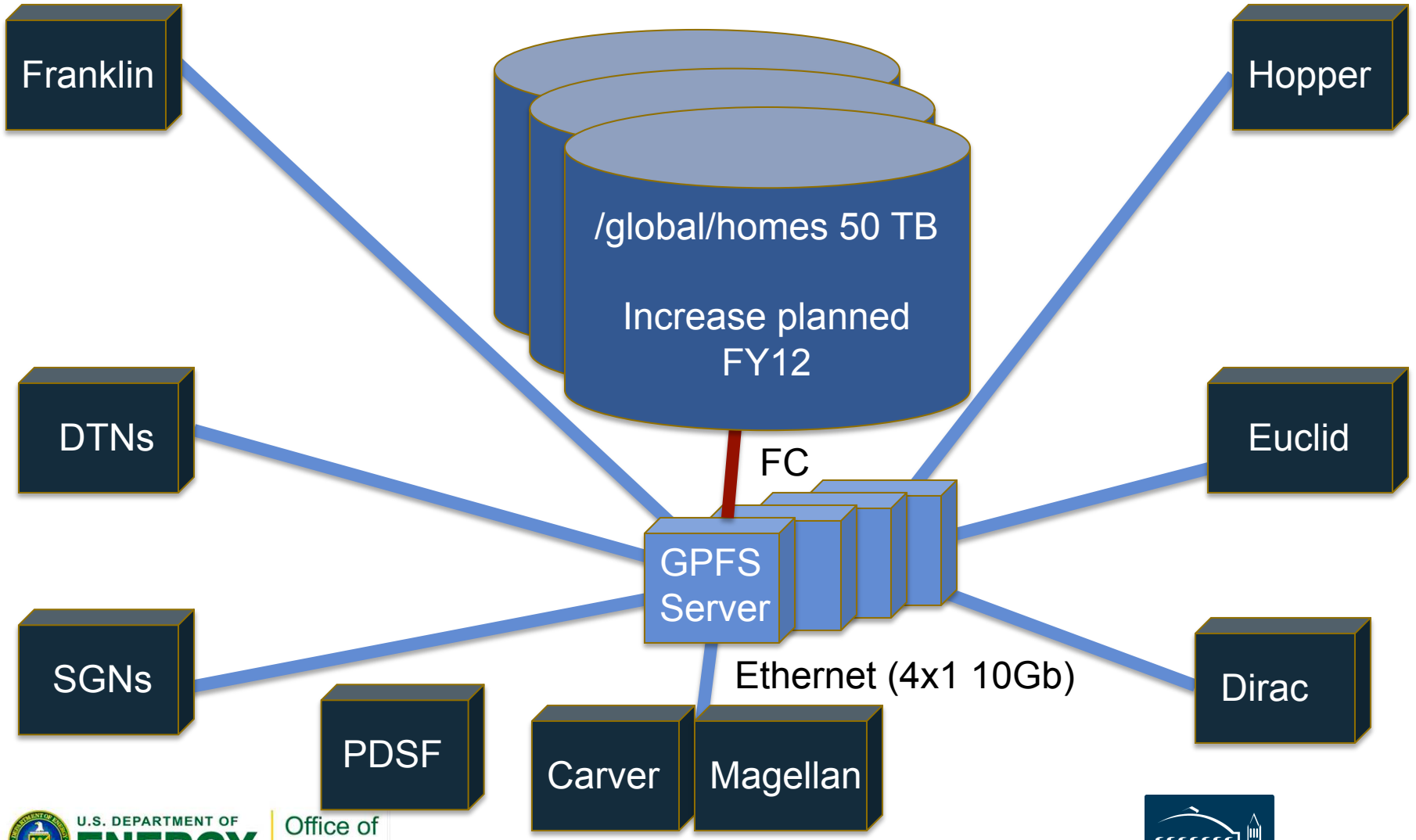
NGF Global Scratch



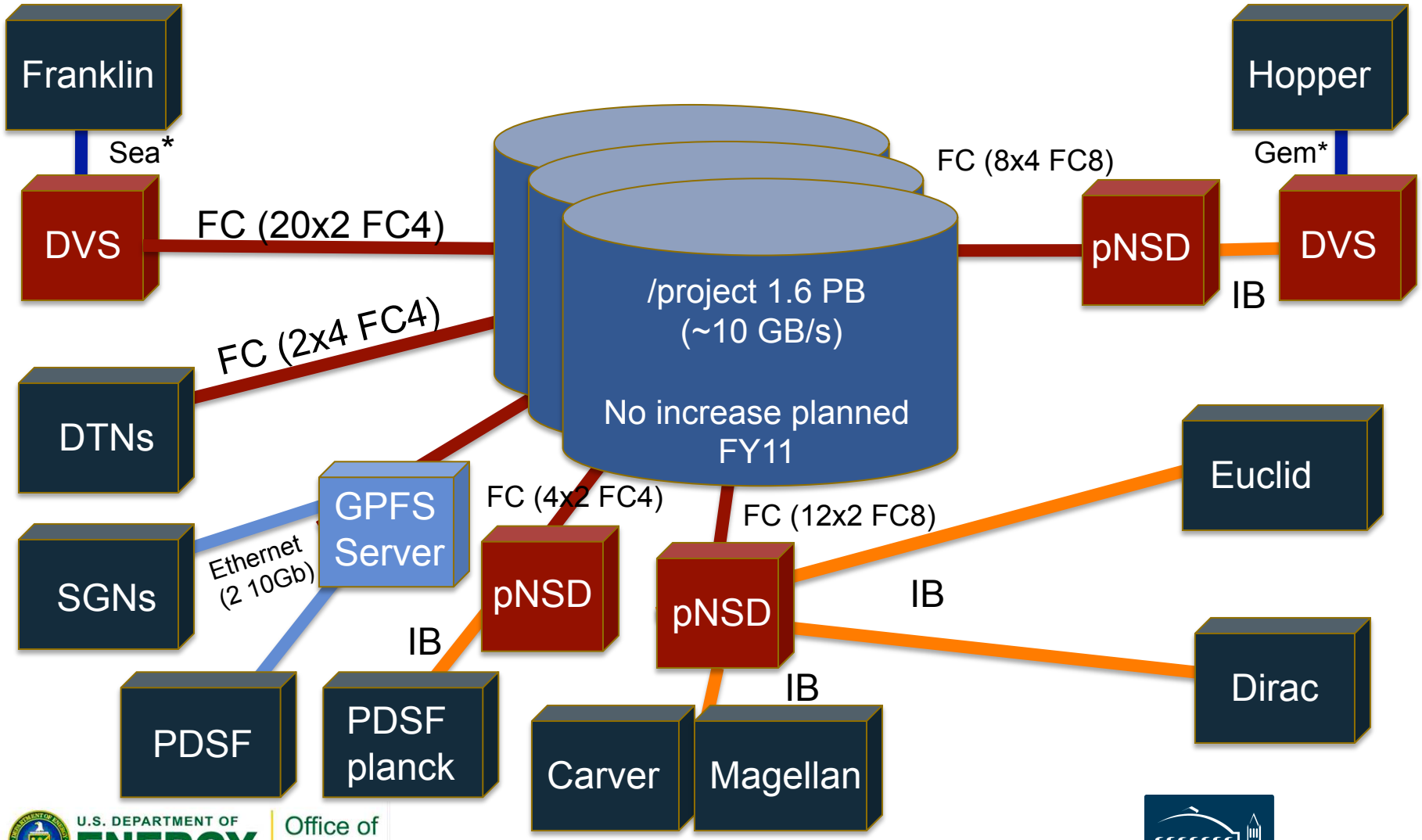
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

NGF Global Homes



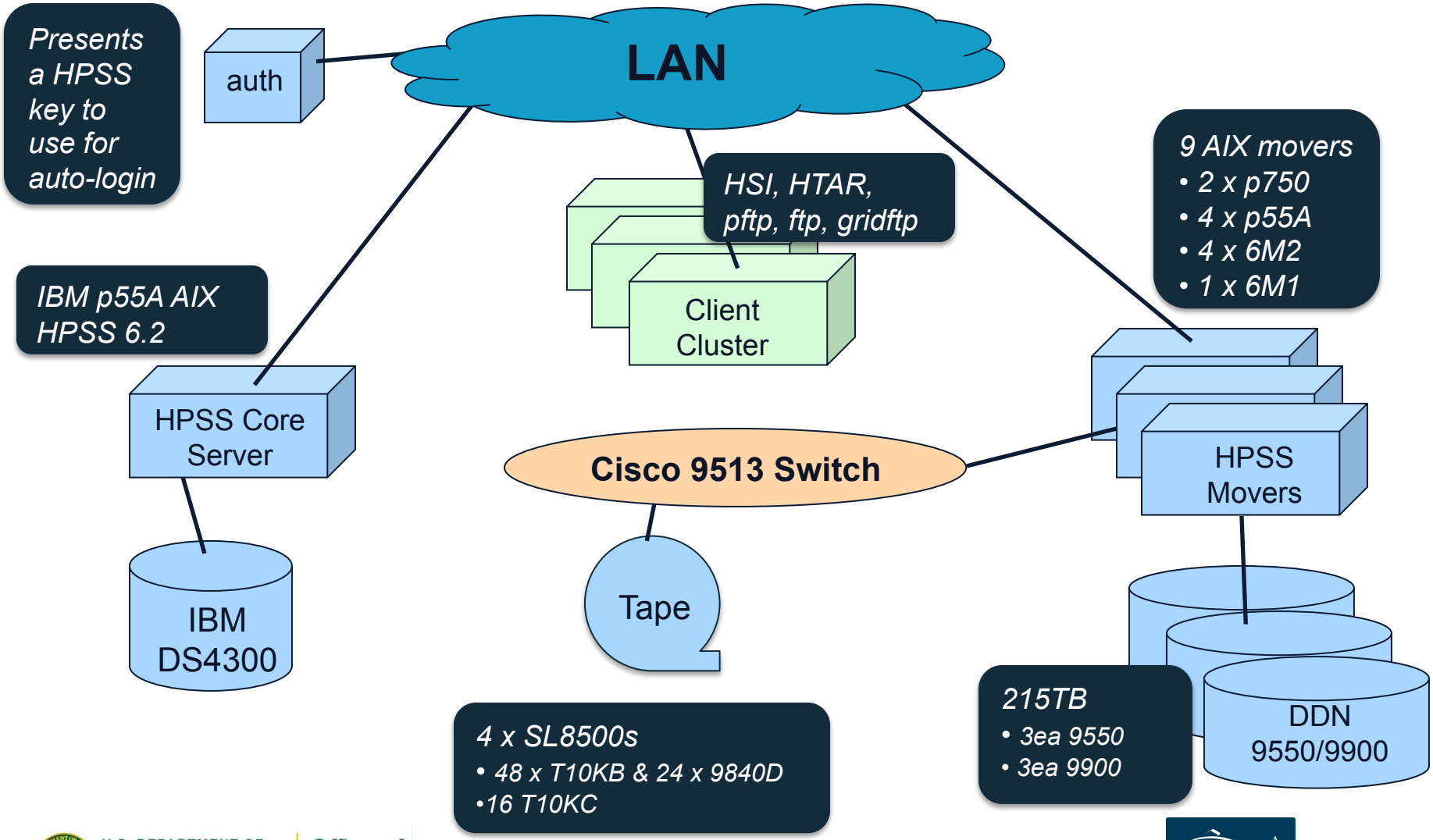
NGF Project


Franklin
Sea*
DVS
FC (20x2 FC4)
FC (2x4 FC4)
DTNs
SGNs
Ethernet
(2 10Gb)
GPFS
Server
PDSF
PDSF
planck
/project 1.6 PB
(~10 GB/s)
No increase planned
FY11
FC (4x2 FC4)
pNSD
IB
FC (8x4 FC8)
pNSD
IB
Hopper
Gem*
DVS
FC (12x2 FC8)
pNSD
IB
IB
Euclid
Dirac
Carver
Magellan

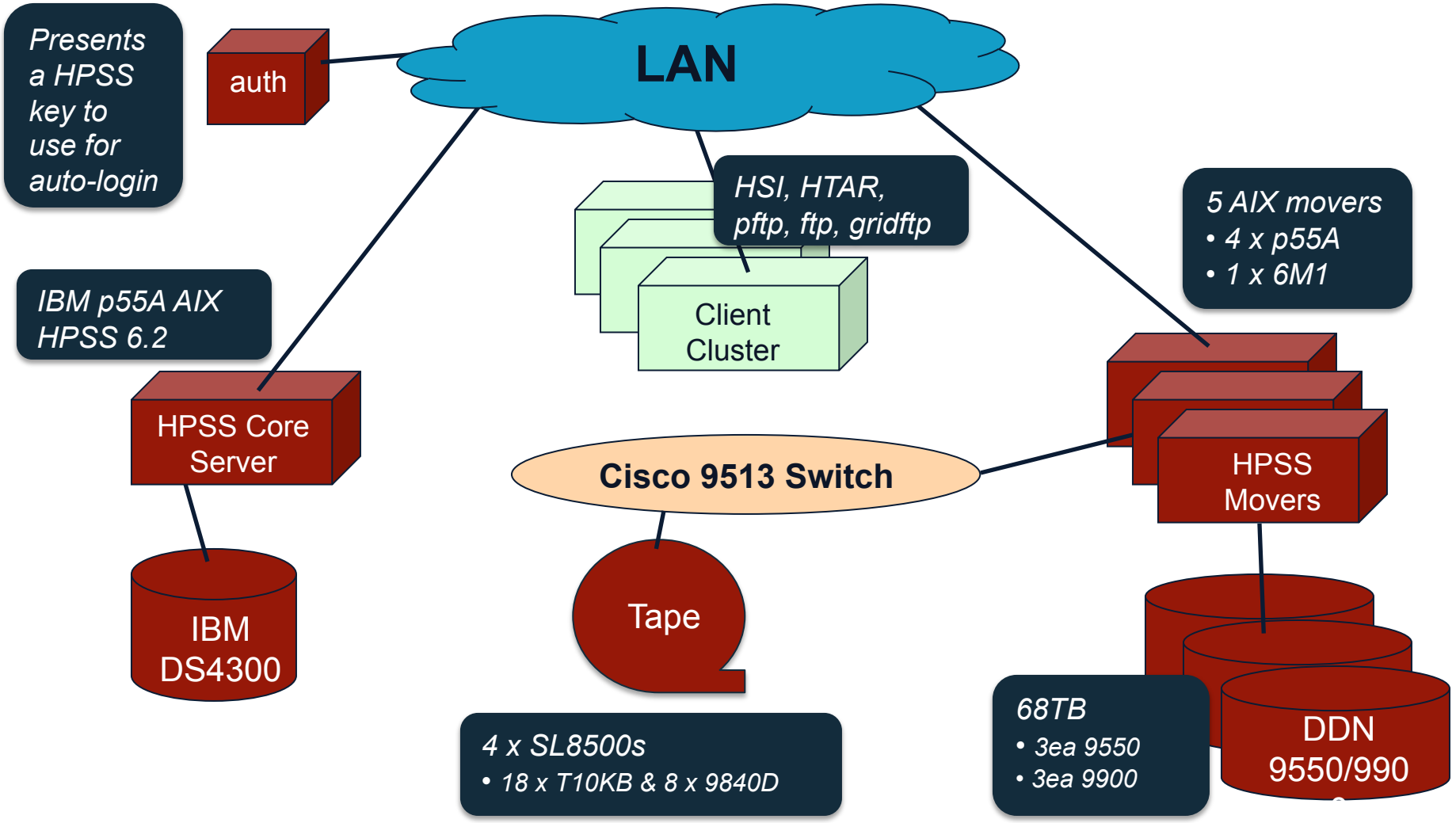
Archival Storage

- **User HPSS**
  - Single transfers 1GB/sec read/write
  - Aggregate bandwidth 4+GB/sec
  - Average daily IO of 20TB, with peak at 40TB
  - 200TB disk cache
  - 24 9840D, 48 T10KB, 16 T10KC tape drives
  - Largest file: 5.5TB
  - Oldest file: Jan 1976
- **Backup HPSS**
  - Single transfers 1GB/sec read/write
  - Aggregrate bandwidth 3+GB/sec
  - Average daily IO of 10TB, with peak at 130TB
  - 40TB disk cache
  - 8 9840D and 18 T10KB tape drives
  - Largest file: 3.5TB
  - Oldest file: May 1995

User HPSS Configuration



Backup HPSS Configuration



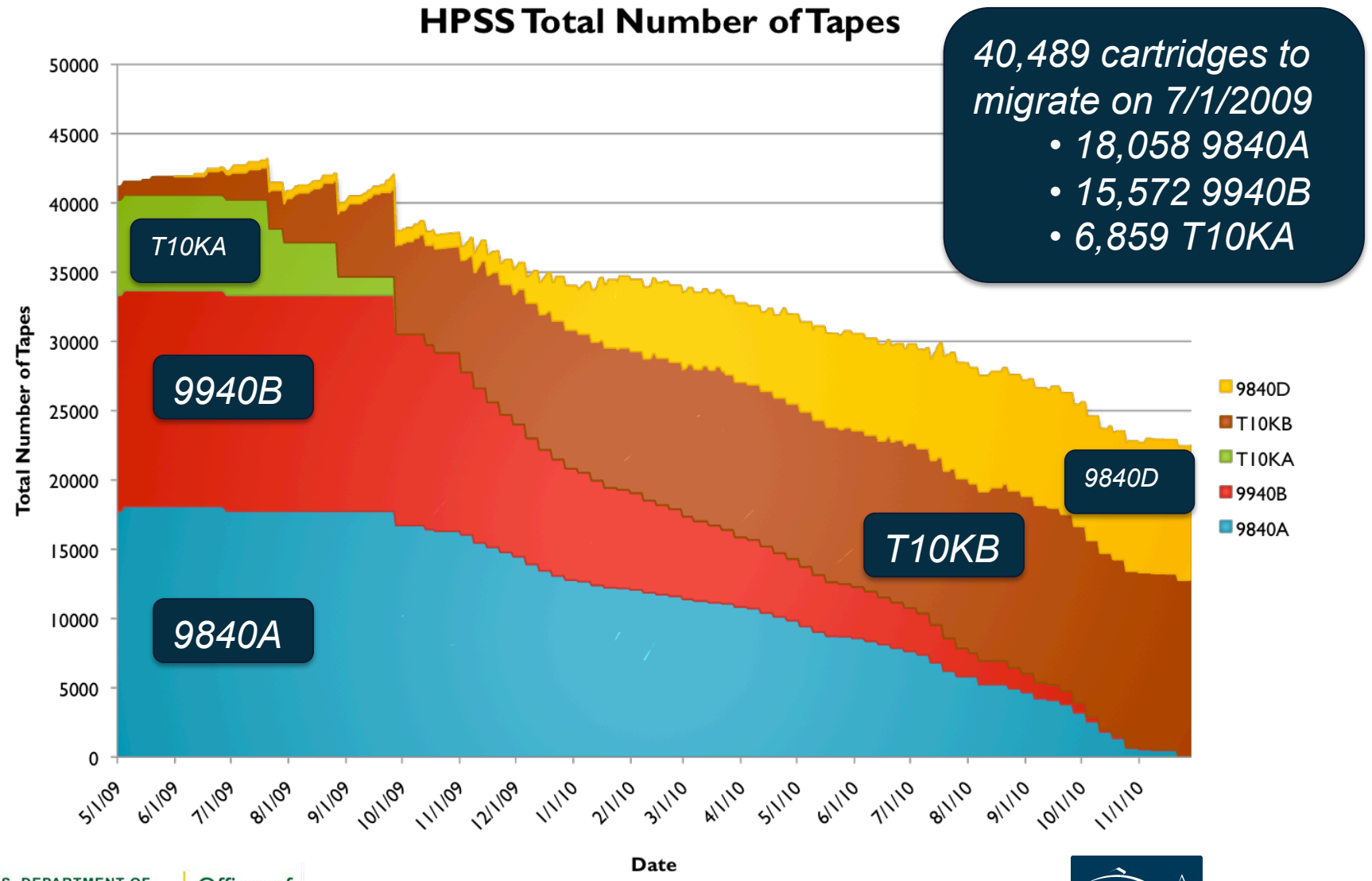
Tape with High Frequency Read

- **Monitoring in place to understand peak and average load**
  - Custom database for gathering usage statistics – per transfer bandwidth, concurrent transfers, commonly requested tapes
  - Crossroads RVA – tape drive utilization, occupancy, concurrency
- **Budget tape drives to peak demand**
  - We had an average of 16 drives in use a few months ago
  - Our peak demand was 30 drives
- **Size the HSM disk cache properly for repeat reads**
  - Analyze the cache hit ratio
  - Plan to hold at least 5 days of peak data transfer on disk
  - Study the stage requests (tape to disk)
- **Identify and resolve library hotspots**
  - Ensures we have cartridge closest to drives for quickest mount time
- **Only allow select applications to do direct-to-tape IO**
  - Our parallel file system backups and restore (10GB files of 600TB)

Storage Challenges & Goals

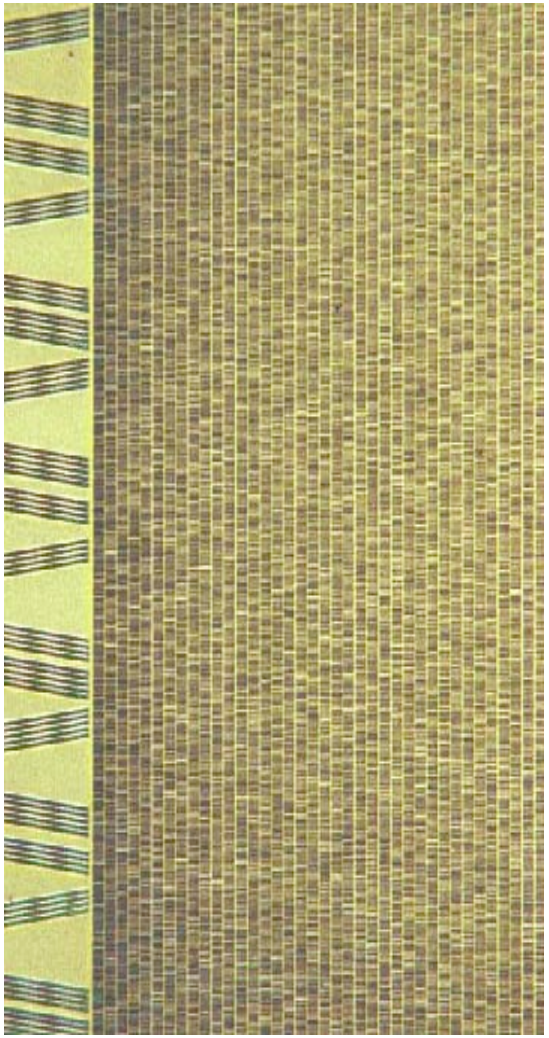
- **Move storage policies from the group/sysadmins to users**
  - Consistent with the success of turning quotas into user allocations!
  - GPFS File System backups, significant effort to scale to PB, and are only intended for disaster recovery.
  - These probably have more value to the user, and they can help determine critical data. They can help reduce the scale of the solution.
  - Provide tools to aid the user in data management (archiving their file system data with one-click, backing up their data with one-click).
- **Balancing growth across disk and tape**
  - File system growth should not exceed archive growth, otherwise we have two problems (backups and resource imbalance)
- **Ensuring our storage network can meet the demands of Exascale computing/storage.**
  - 40Gb/100Gb Ethernet demonstrations, Cloud storage for DOE?
  - Fiber Channel deployment improvements, retracting our SAN
  - InfiniBand testing as center-wide storage network
  - Consider SAS connections for storage to servers

Migrating Data from Old to New



Actual Reliability of Data on Tape

- **We read all data on 40,489 tapes**
  - 6,859 T10KA (up to 2yrs old)
  - 15,572 9940B (up to 8yrs old)
  - 18,058 9840A (up to 12yrs old)
- **We found 36 tapes that had some data that couldn't be read.**
  - 24 9840A, 8 9940B, 4 T10KA
  - One of those had 558 files and couldn't be mounted.
  - Two others had 136 and 43 files that couldn't read, the remainder had less than 6 with most having only 1.
- **0.0009% error rate for tape cartridges or 99.9991% with 100% readable data.**
- **But wait! It's not the whole cartridge, the unreadable data was contained in 850 files (84.6 M total) representing 3.0 TB of data (8,056 TB total).**
- **0.00001% error rate for files or 99.99999% of files with 100% readable data.**
- **Unreadable data is normally in one or two blocks of data (250-500MB of data) with remainder of file readable, but we don't recover partial files unless user requests.**



Our Experience Shows Enterprise Tape is Reliable

The data migration (7/09 – 12/10) involved reading 22,065,763m of tape, the distance of flying San Francisco to Tokyo to Paris to Nova Scotia.

Unreadable data resided in at least one block of 850 files. These files represent 178m of tape, approximately the length of two Boeing 777 jets (70m) or half the length of most cruise ships (350m).

National Energy Research Scientific Computing Center

U.S. DEPARTMENT OF ENERGY | Office of Science

Lawrence Berkeley National Laboratory