

# The NERSC Global File System (NGF)

Jason Hick

[jhick@lbl.gov](mailto:jhick@lbl.gov)

Storage Systems Group Lead

<http://www.nersc.gov/nusers/systems/NGF/>

CAS2K11 September 11-14, 2011



U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science



National Energy Research  
Scientific Computing Center



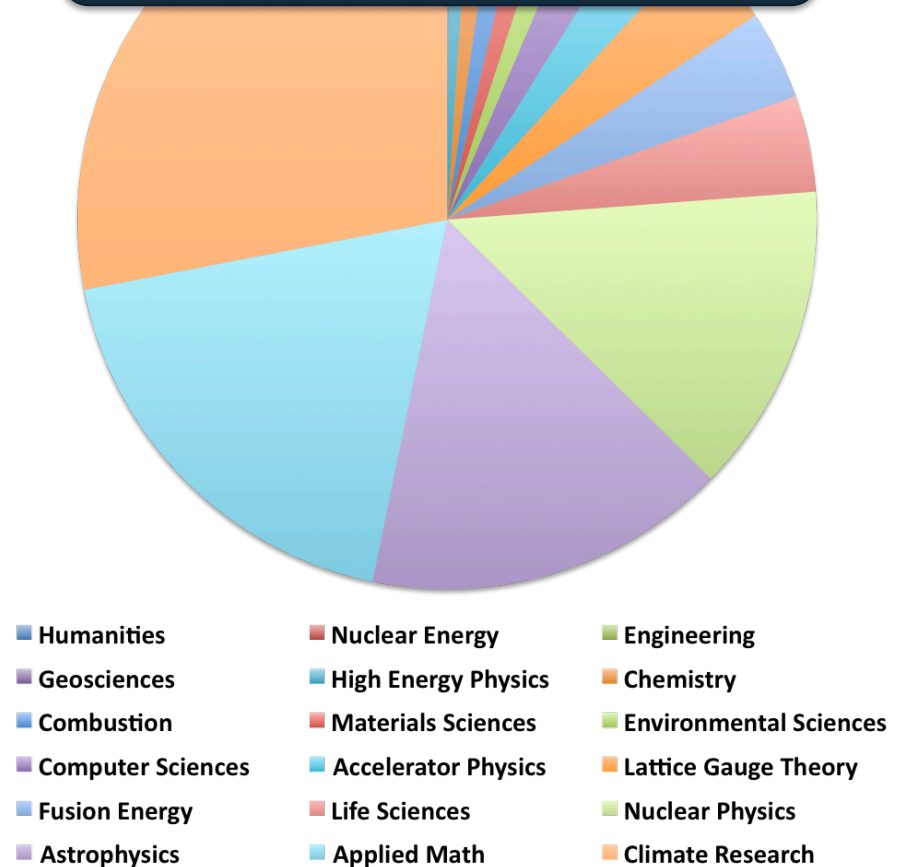
Lawrence Berkeley  
National Laboratory



# The Production Facility for DOE Office of Science

- Operated by University of California for the US Department of Energy
- 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. Advanced Networking Initiative)
- 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.

2011 storage allocation by area of science. Climate, Applied Math, Astrophysics, and Nuclear Physics are 75% of total.





# Servicing Data Needs

- **Storing data:**
  - 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.
- **Analyzing Data:**
  - Large memory systems, Hadoop/Eucalyptus file systems
- **Sharing Data:**
  - Provide Science Gateway Nodes that enable data in /project to be available via your browser (offer authenticated and unauthenticated options).
    - Goals of Science Gateways
      - Expose scientific data on NGF file systems and HPSS archive to larger communities
      - Work with large data remotely
      - Outreach: Broadens scientific impact of computational science: “as easy as online banking”
    - NEWT – NERSC Web Toolkit/API
      - Building blocks for science on the web
      - Write a Science Gateway by using HTML + Javascript
      - Support for: authentication, submission, files access, accounting, viewing queues, user data tables
    - 30+ projects use the NGF -> web gateway
- **Protecting Data:**
  - Frequent (near daily) incremental backups of file system data. Request information on PIBS from Matt Andrews (mnandrews@lbl.gov).



# Optimizing Inter-Site Data Movement

- **NERSC is a net importer of data.**
  - Since 2007, we receive more bytes than transfer out to other sites.
  - Leadership in inter-site data transfers (Data Transfer Working Group).
  - Have been the an archive site for several valuable scientific projects: SNO project (70TB), 20<sup>th</sup> Century Reanalysis project (140TB)
- **Partnering with ANL and ESnet to advance HPC network capabilities.**
  - Magellan ANL and NERSC cloud research infrastructure.
  - Advanced Networking Initiative with ESnet (100Gb Ethernet).
- **Data Transfer Working Group**
  - Coordination between stake holders (ESnet, ORNL, ANL, LANL and LBNL/NERSC)
- **Data Transfer Nodes**
  - Available to all NERSC users
  - Optimized for WAN transfers and regularly tested
  - Close to the NERSC border
  - GlobusOnline endpoints
  - *hsi* for HPSS-to-HPSS transfers (OLCF and NERSC)
  - *gridFTP* for system-to-system or system-to-HPSS transfers
  - *bbcp* for system-to-system transfers



# Minimizing Intra-site Data Movement

- **Mount file systems on as many clusters as possible**
  - Work with vendor to port client to wide variety of operating systems
  - Today this is easier because primarily systems are variants of Linux
- **Architect file systems for different use cases**
  - Write optimized, large file bandwidth, large quota (/global/scratch)
  - Read optimized, aggregate file bandwidth, long-term storage (/project)
  - Small file optimized, common environment (/global/homes, /global/common)
- **Deliver high bandwidth to each system**
  - Large fiber channel SAN to deliver bandwidth to each cluster using private network storage devices

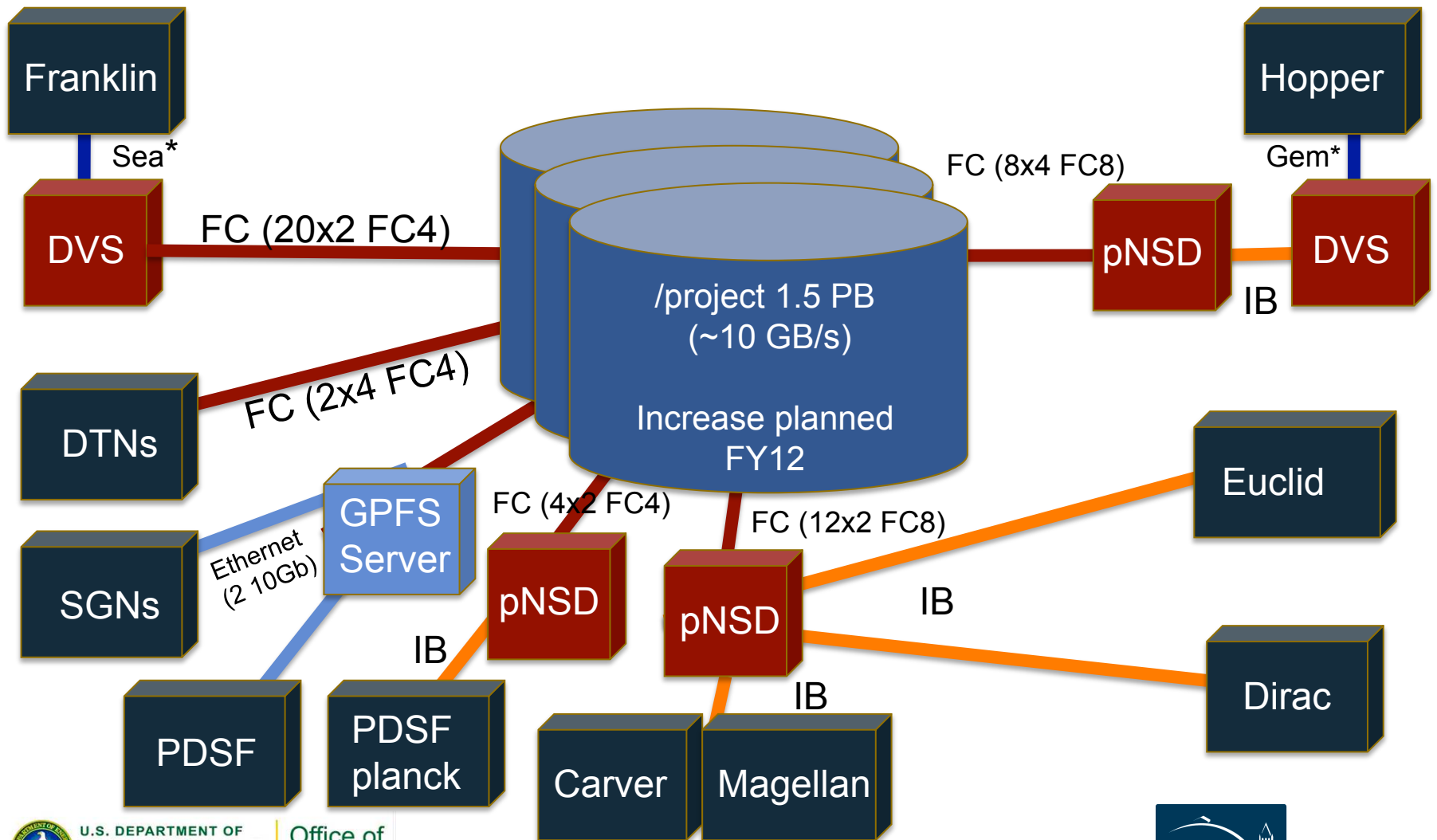


# The NERSC Global File Systems

- **/project is for sharing and long-term residence of data on all NERSC computational systems.**
  - 4% monthly growth, 60% growth per year
  - Not purged, quota enforced (4TB default per project), projects under 5TB backed up daily
  - Serves 200 projects over FC8 primarily, 10Gb ethernet alternatively
  - 1.5PB total capacity, increasing to ~4.4PB total in 2012
  - ~5TB average daily IO
- **/global/homes provides a common login environment for users across systems.**
  - 7% monthly growth, 125% growth per year
  - Not purged but archived, quota enforced (40GB per user), backed up daily
  - Oversubscribed, needs ~150TB of capacity unless we increase quotas
  - Serves 4000 users, 400 per day over 10Gb Ethernet
  - 50TB total capacity, increasing to ~100TB total in 2012
  - 100's of GBs average daily IO
- **/global/common provides a common installed software environment across systems.**
  - 5TB total capacity, increasing to ~10TB total in 2012
  - Provides software packages common across platforms
- **/global/scratch provides high bandwidth and capacity data across systems.**
  - Purged, quota enforced (20TB per user), not backed up
  - Serves 4000 users over FC8 primarily, 10Gb ethernet alternatively
  - 1PB total capacity, looking to increase to ~2.4PB total in 2012
  - Bandwidth is about 16GB/s
  - Aided us in transitioning off Hopper p1 to p2, reconfiguring local scratches, and in inter-site data movement

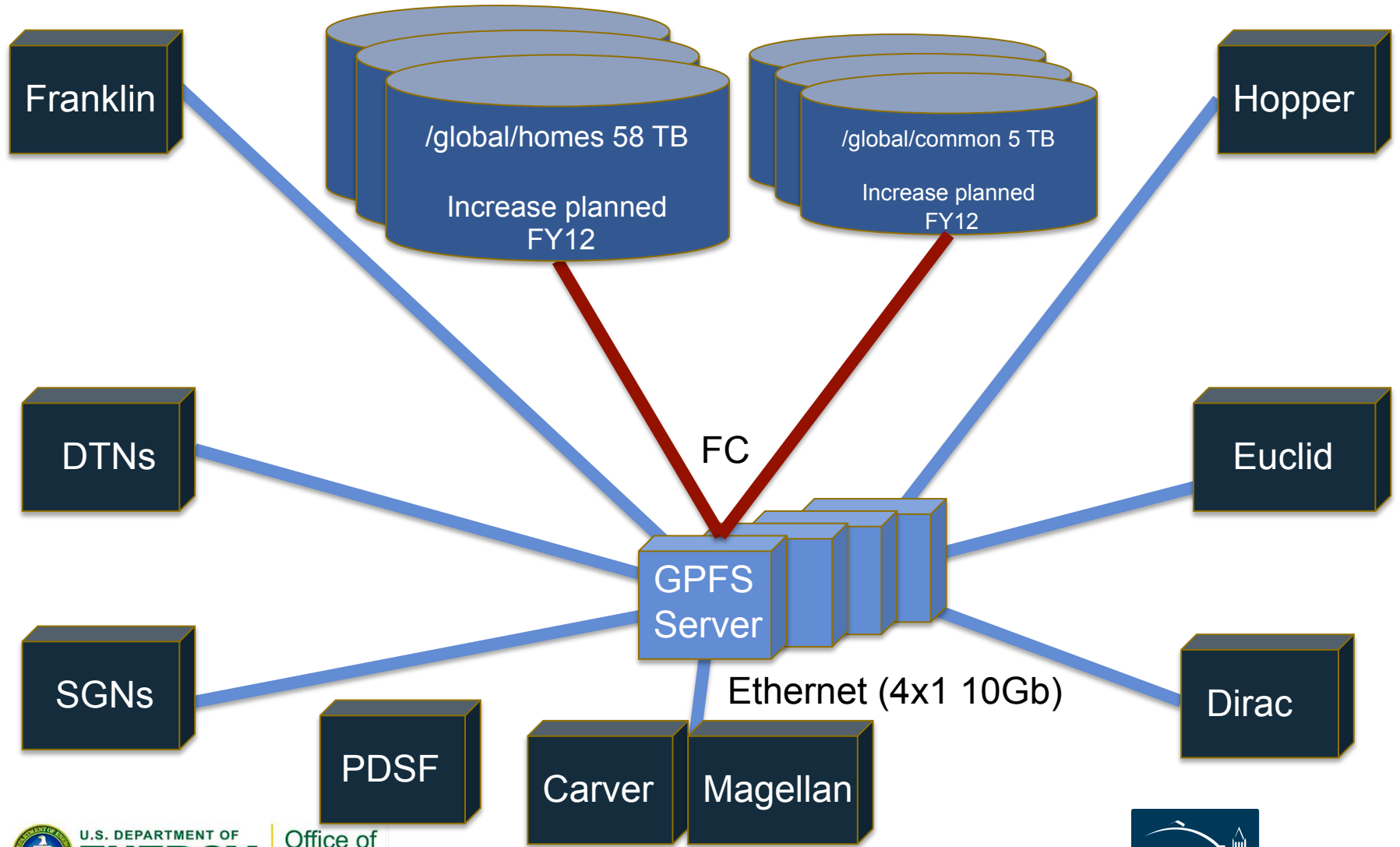


# NGF Project





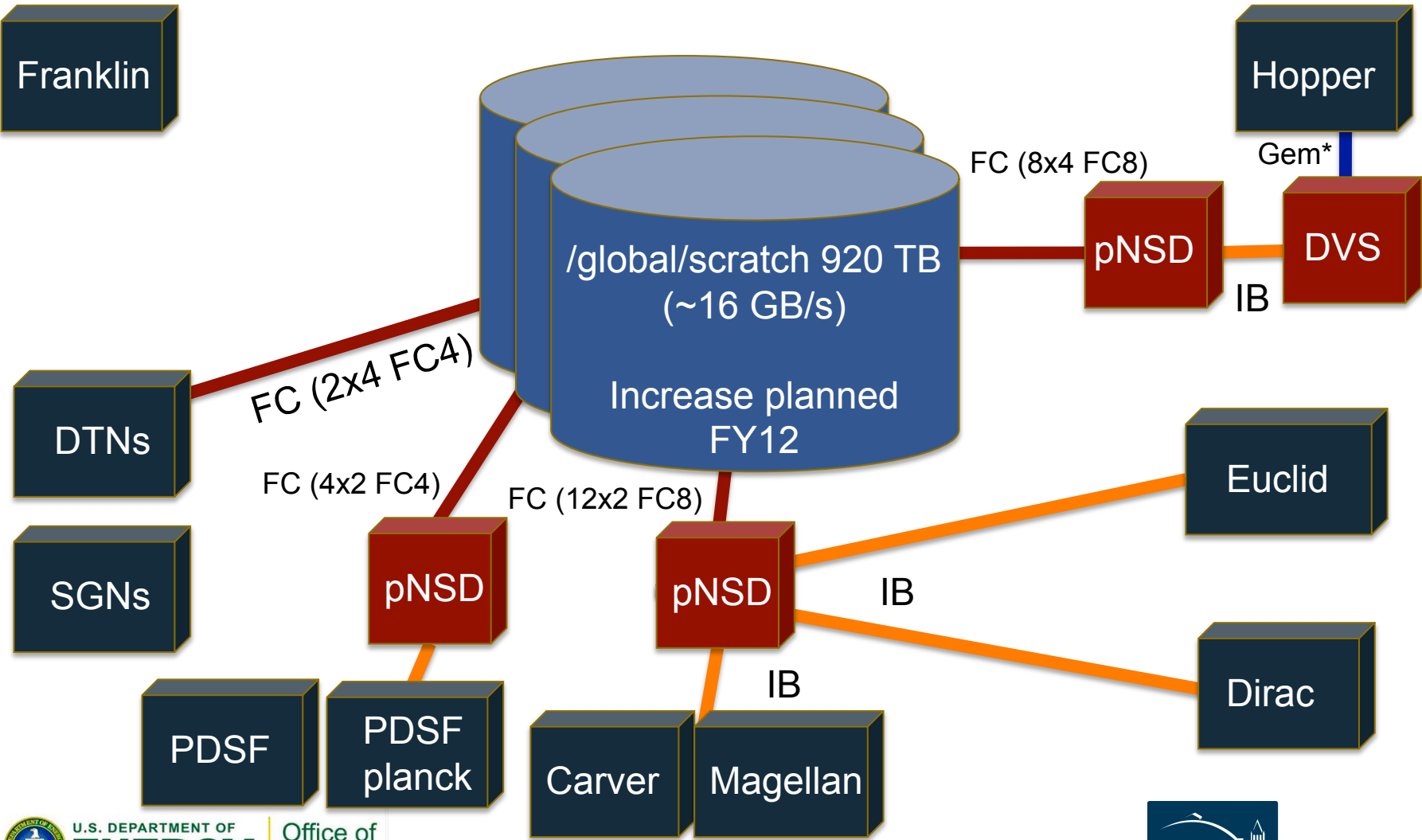
# NGF Global Homes & Common







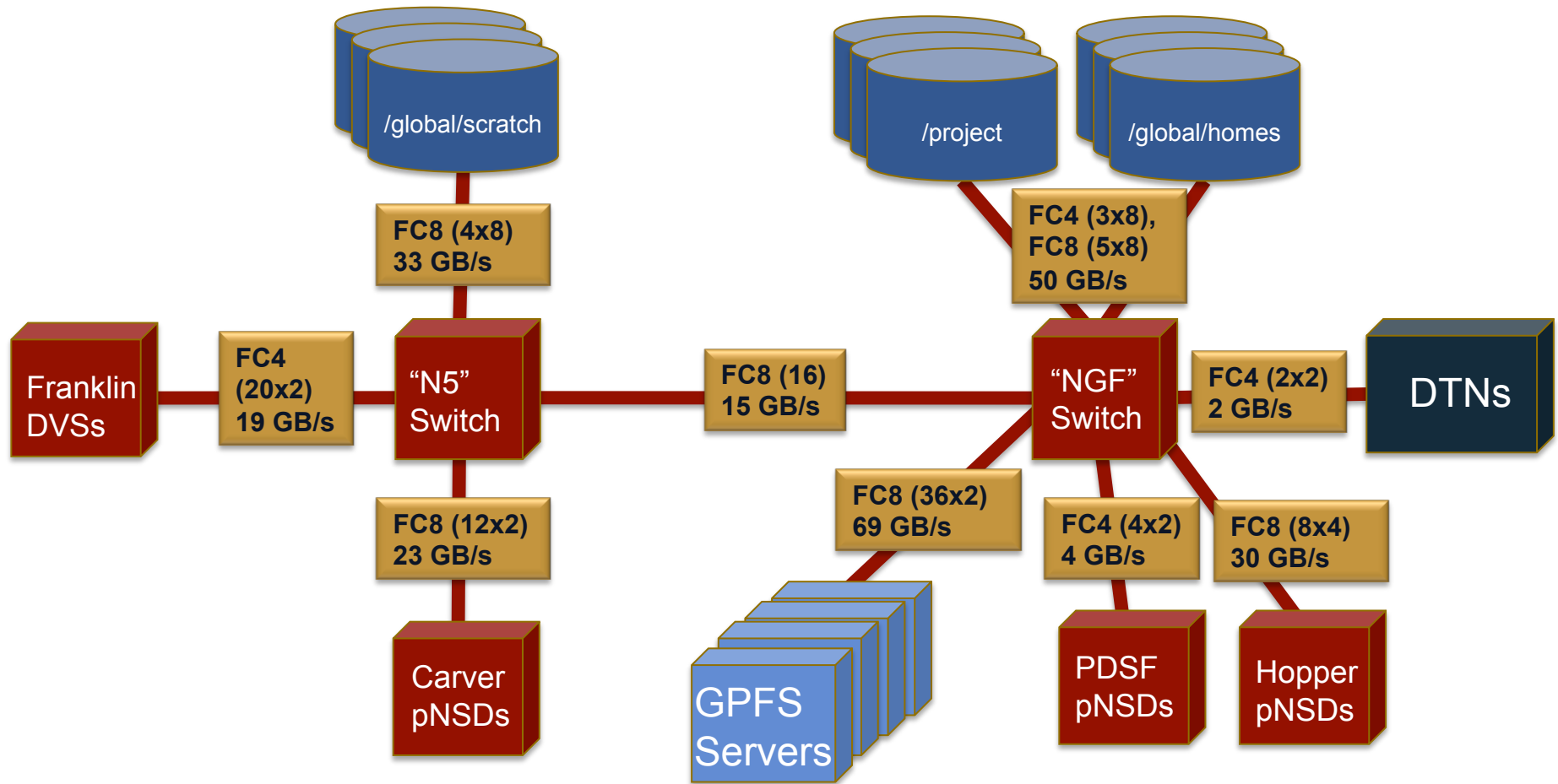
# NGF Global Scratch





## Lesson Learned: Avoid direct-attached FC clients

- **Storage direct-attached via FC to clients**
  - In principle a good idea, in practice it's not
  - Requires a very “broad” fiber channel network (200+ initiators)
  - Leads to de-tuning the FC HBAs (queue depth = 1 or 2) to keep storage arrays from panicking
  - Keeping client FC HBAs tuned correctly is difficult and it only takes one to cause havoc
  - Changes in device configuration (arrays) is difficult for every client (some OS's allow dynamic changes, some don't)
- **FC network is at its final destination**
  - Cisco 9513s are oversubscribed with 24 port 8Gb linecards
  - Can fix with 32 port 8Gb linecards at full rate, but this is a large reinvestment
  - Cisco 16Gb linecards are only intended for ISLs
  - Having increasing trouble matching client-end (IB) speeds with back-end (storage) speeds





## Lesson Learned: Multiple GPFS file systems in one owning cluster

- **To avoid one file system from affecting another**
  - Run file system managers on different NSD servers
  - Seems to eliminate observed ~2 minute delays of file system operations in file system A while file system B has certain commands run (e.g. policy manager, quota commands, ...)
- **Works well otherwise and allows a limited amount of server hardware to maintain a multi-GPFS file system deployment**



# Future Plans

- **Moving to IB based client connections to the file systems**
  - Connect storage directly via FC to GPFS servers
  - Connect GPFS servers to each cluster's IB network, if it doesn't have one, create a small IB network to scale with cluster's proprietary network (Seastar, Gemini, ...)
  - Reduces with eventual goal to eliminate the FC SAN
- **Move from pNSD deployment to NSD deployment**
  - Scalability of server and data network resources
- **NIM Accounting Service & NGF Information integration**
  - Move from administrator-level information to user-level information
  - Administrative backups to user backups
    - Common value in backups is restoring unintentionally deleted files
  - Aid users in knowing their utilization, allocation, and quota information
- **Using the 100Gb WAN for data**
  - Multi-site GPFS?
    - 100Gb WAN makes this feasible, but new GPFS features desired to enable this to work well
  - Inter-site data transfer (bulk data movement)



**Thank you!**

**Questions?**