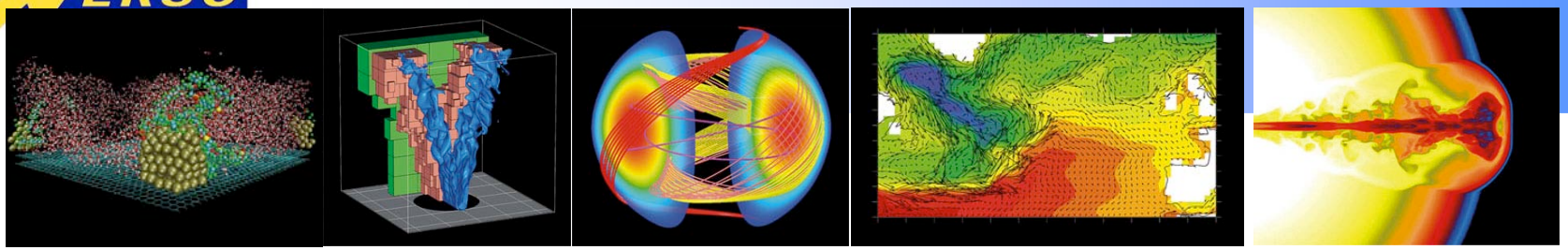


NERSC Experience and Plans for Petascale Data

For the Pestscale Data Storage Workshop – Nov 18, 2006

William T.C. Kramer

kramer@nersc.gov

510-486-7577

National Energy Research Scientific Computing (NERSC) Facility

Ernest Orlando Lawrence

Berkeley National Laboratory

This work was supported by the Director, Office of Science, Division of Mathematical, Information, and Computational Sciences of the U.S. Department of Energy under contract number DE-AC03-76SF00098.

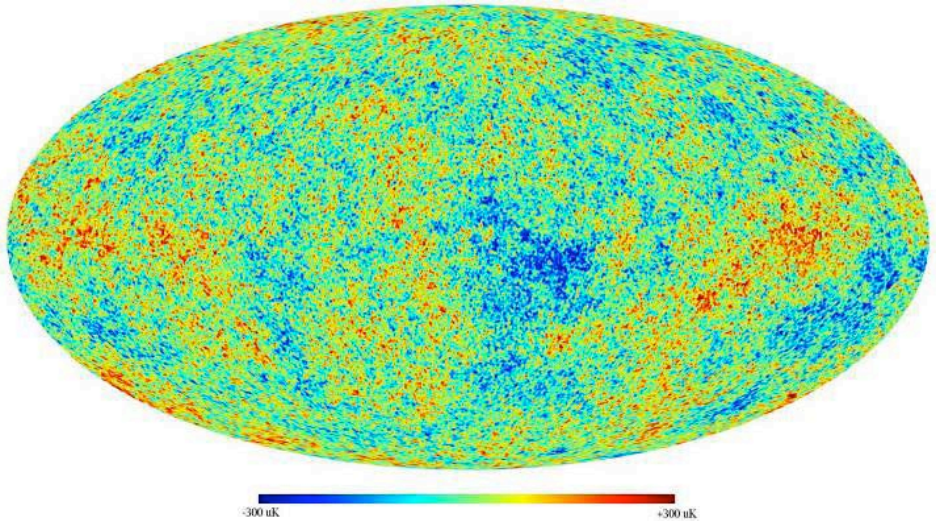
Three Trends to Address

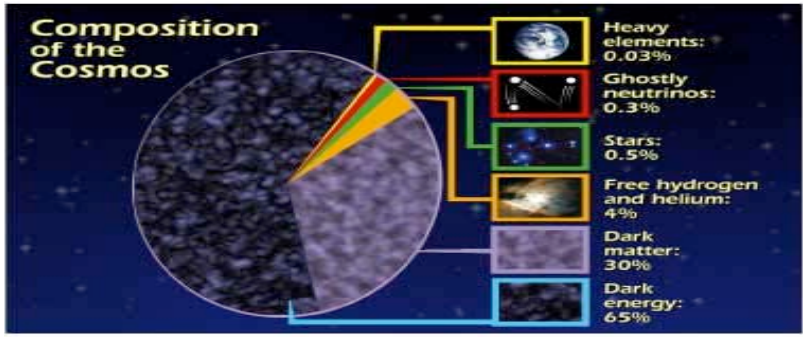
- **The widening gap between application performance and peak performance of high-end computing systems**
- **The recent emergence of large, multidisciplinary computational <u>science teams</u> in the DOE research community**
- **The <u>flood of scientific data</u> from both simulations and experiments, and the convergence of computational simulation with experimental data collection and analysis in complex workflows**

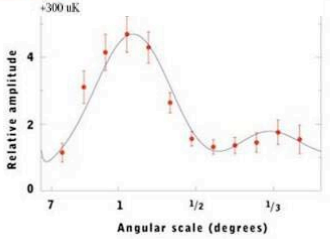
NERSC Storage Vision

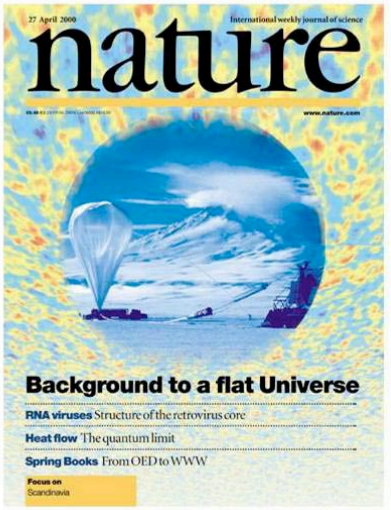
- **Single storage pool, decoupled from individual NERSC computational systems**
  - **Diverse file access - supporting large and small, many and few, permanent and transit**
  - **All systems have access to all storage – require different fabric**
  - **Flexible management of storage resource**
    - **Buy new storage (faster and cheaper) only as needed**
- **High performance, large capacity storage**
  - **Users see same file from all systems**
  - **No need for replication**
  - **Analytics server has access to data as soon as it is created**
  - **Performance near native file system performance**
- **Integration with mass storage**
  - **Provide direct HSM and backups through HPSS without impacting computational systems**
- **Continue to provide archive storage as well as on-line/near-line**
- **Potential geographical distribution**

CMB: Example of one Project ~1% of allocation







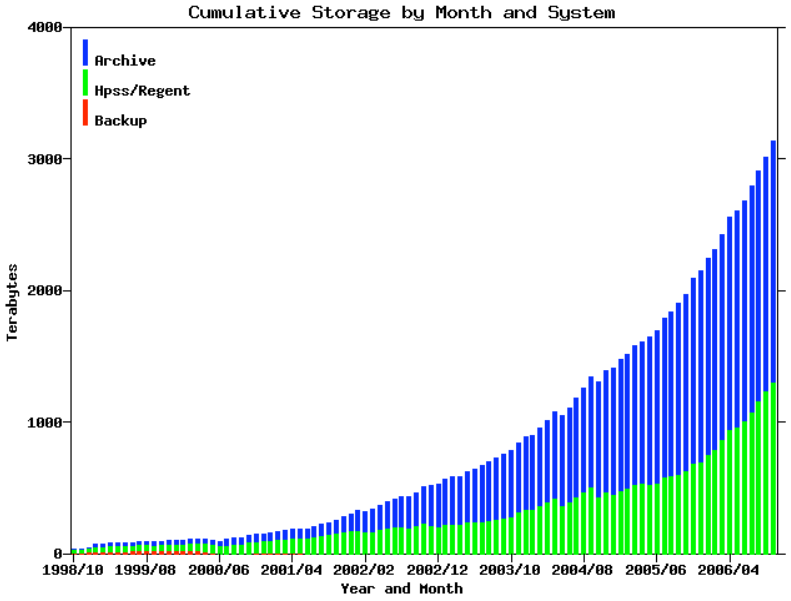


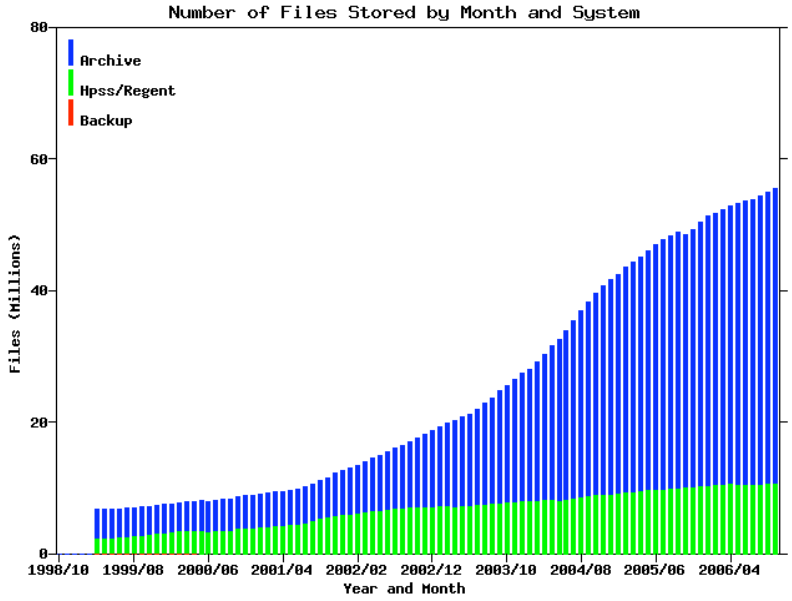
Office of Science
U.S. DEPARTMENT OF ENERGY

CMB: Example of one Project

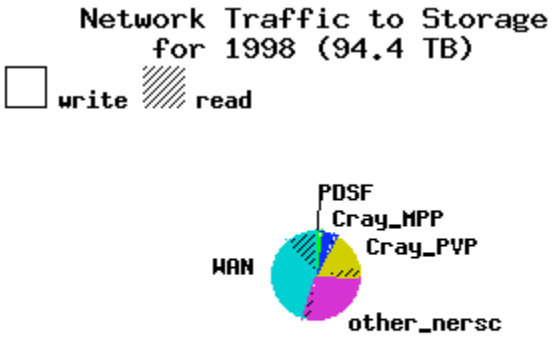
- O(10–100) exaflops of total processing capacity
- O(100) TB of archival file storage for primary data and derived data products.
- O(10) TB of scratch file storage at any one time to support a particular analysis, ideally simultaneously accessible from all of NERSC's machines (
- O(1–10) GB of local tmp file storage on each processor or node to stage intermediate data products and enable out-of-core computations without having to repeat costly I/O subsystem calls.
- Scalable, massively parallel I/O supporting the simultaneous transfer of very large volumes of data across the entire processor set being used; currently much of the Planck-scale CMB data analysis is I/O bound. Stability with respect to the volume of all user traffic over the I/O subsystem is also highly desirable.

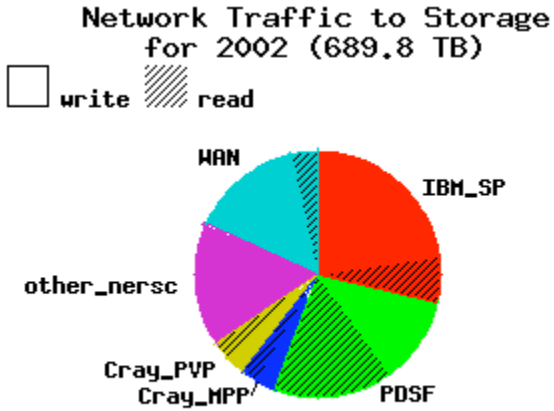
Archive Growth



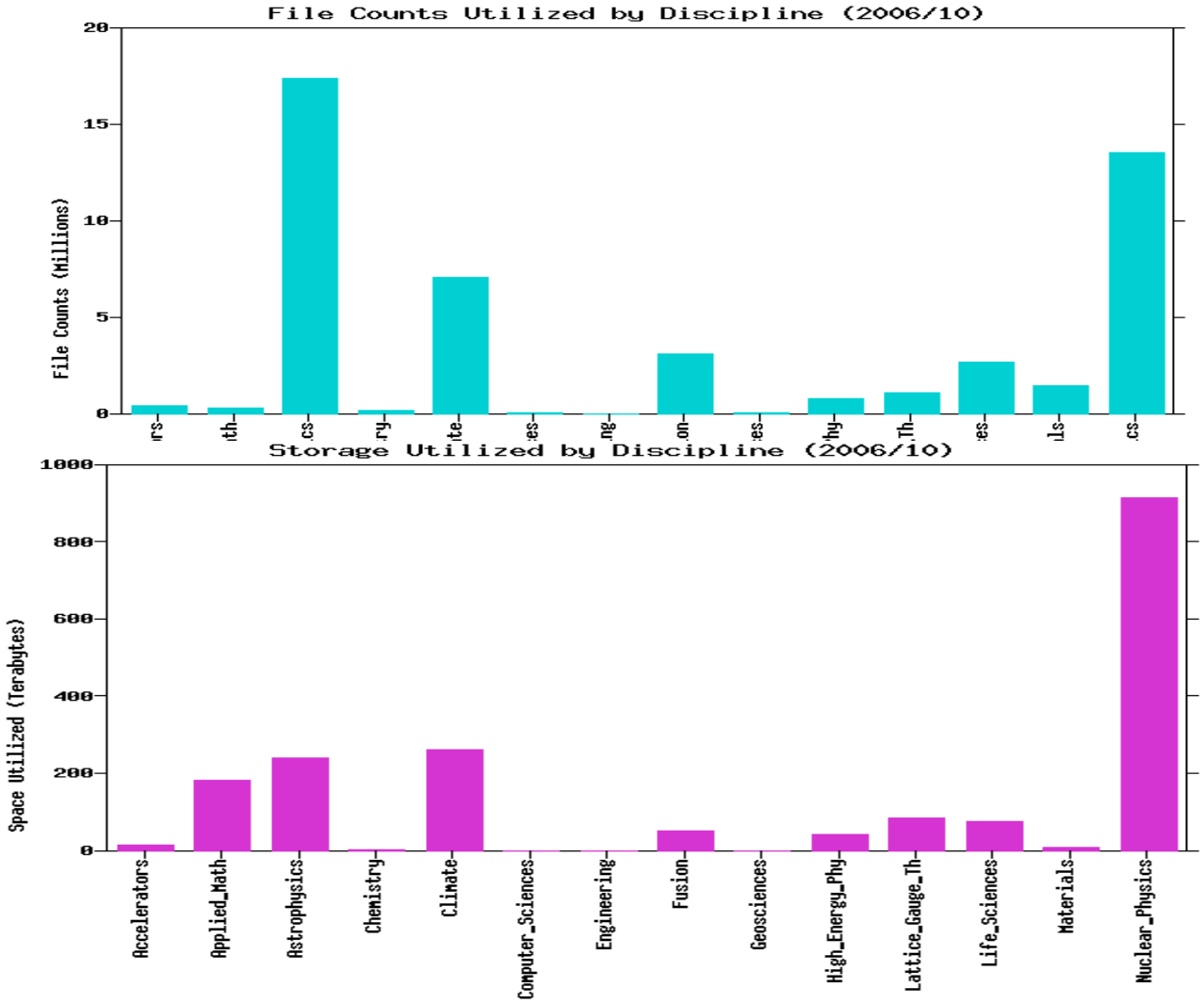


Different Types of Storage



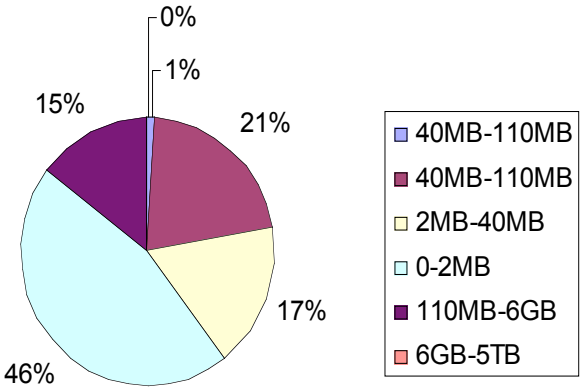


Storage by Discipline

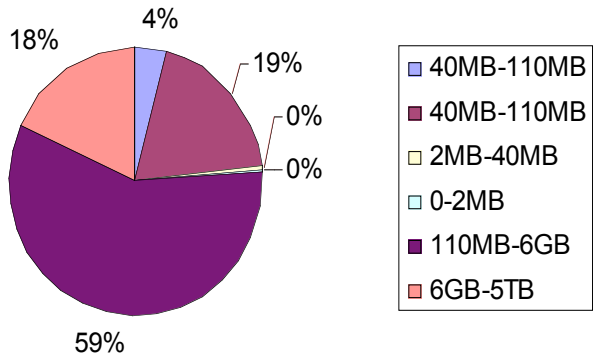


Data and Files

4QFY06 Archive File Growth



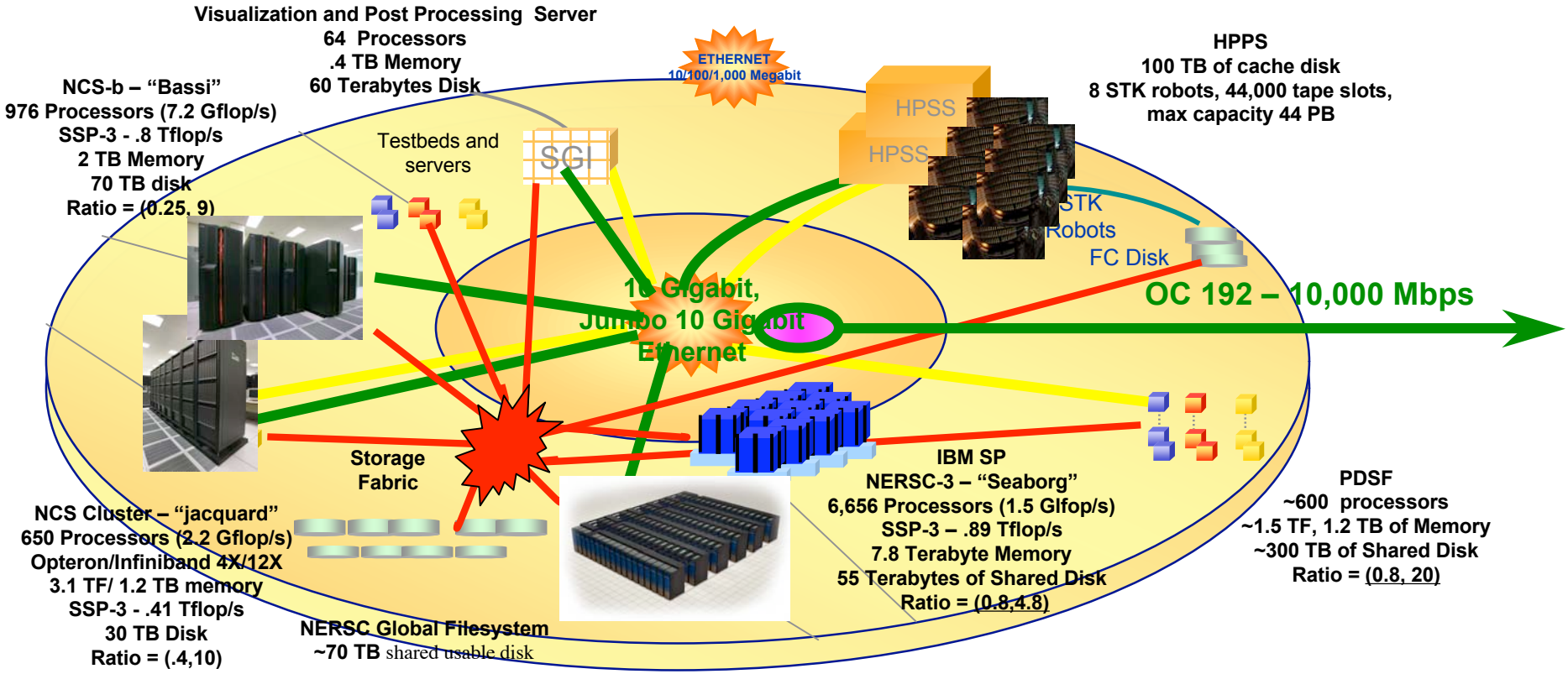
4QFY06 Archive Data Growth



Facility Wide File System – Evolution to Production

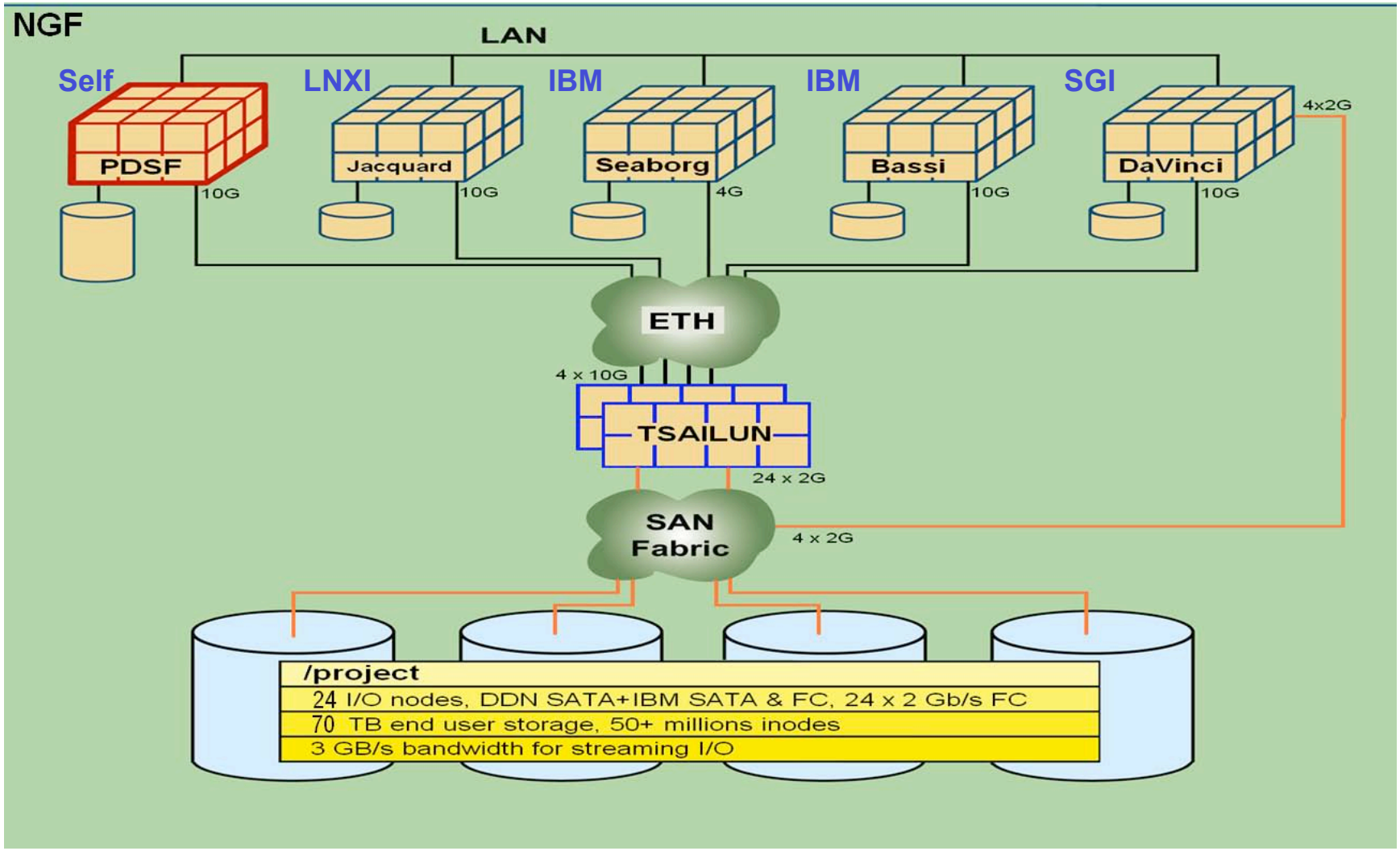
- **The three components – storage devices, connection fabric and filesystem software were sufficiently robust to move to production in summer 2005**
- **I/O performance is a function of hardware first and filesystem software**
  - **Disk heads**
  - **Controllers**
  - **Connections to a host**
- **Decided to provide function first with reasonable performance and then invest in transfer performance**
  - **Metadata performance has to be good to begin with**
- **All systems already had local disk in /home and /scratch**
  - **Need for "project" repositories – so that was the first implementation**
  - **Performance for existing systems limited by system hardware**
- **Started with 5 projects and 20TB of disk – September 2005**

2007

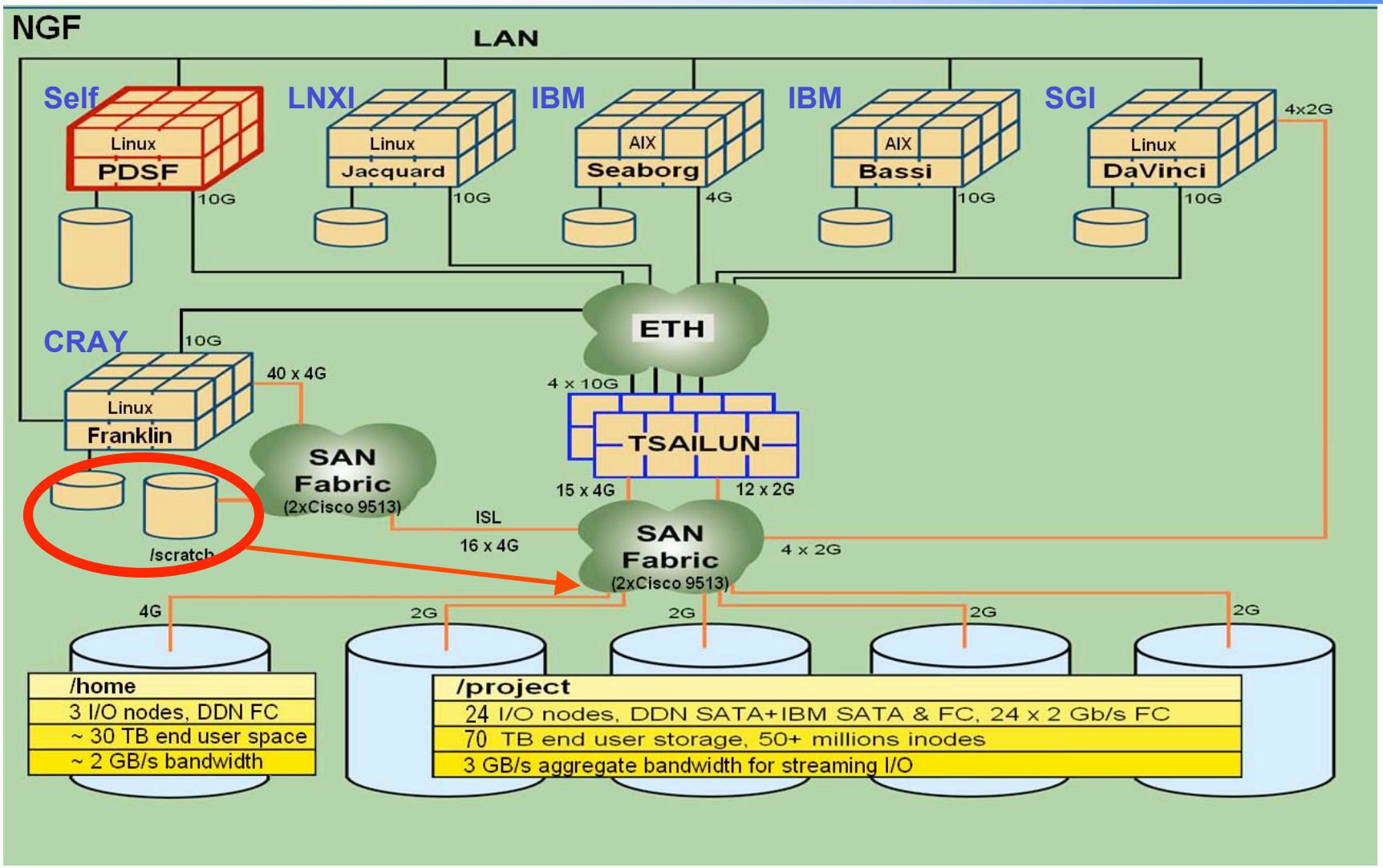


Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

NGF Production Configuration



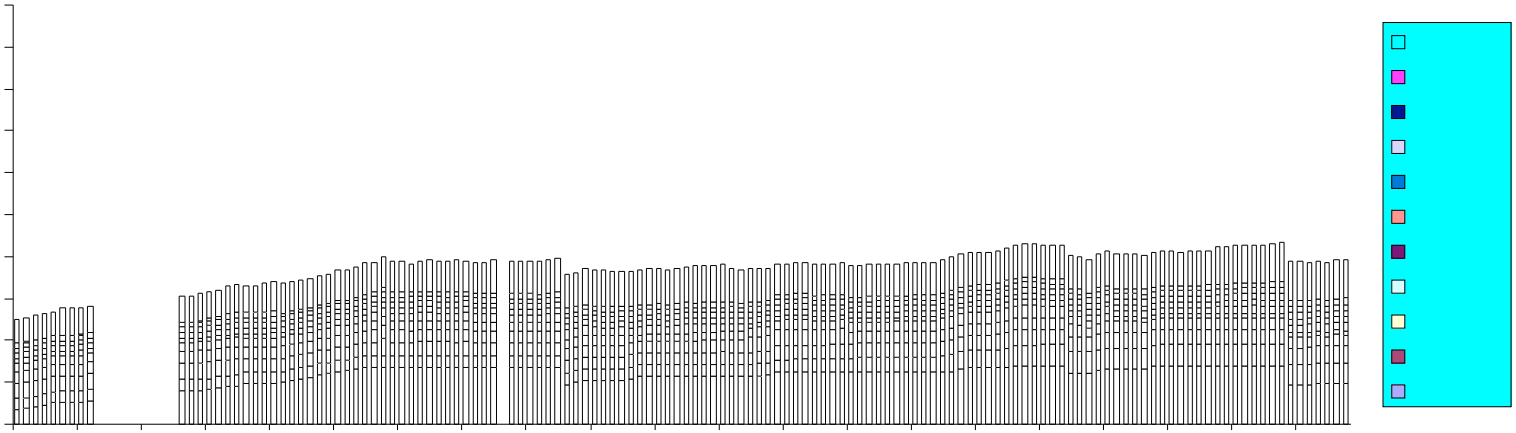
NGF Configuration with NERSC-5 Full Implementation



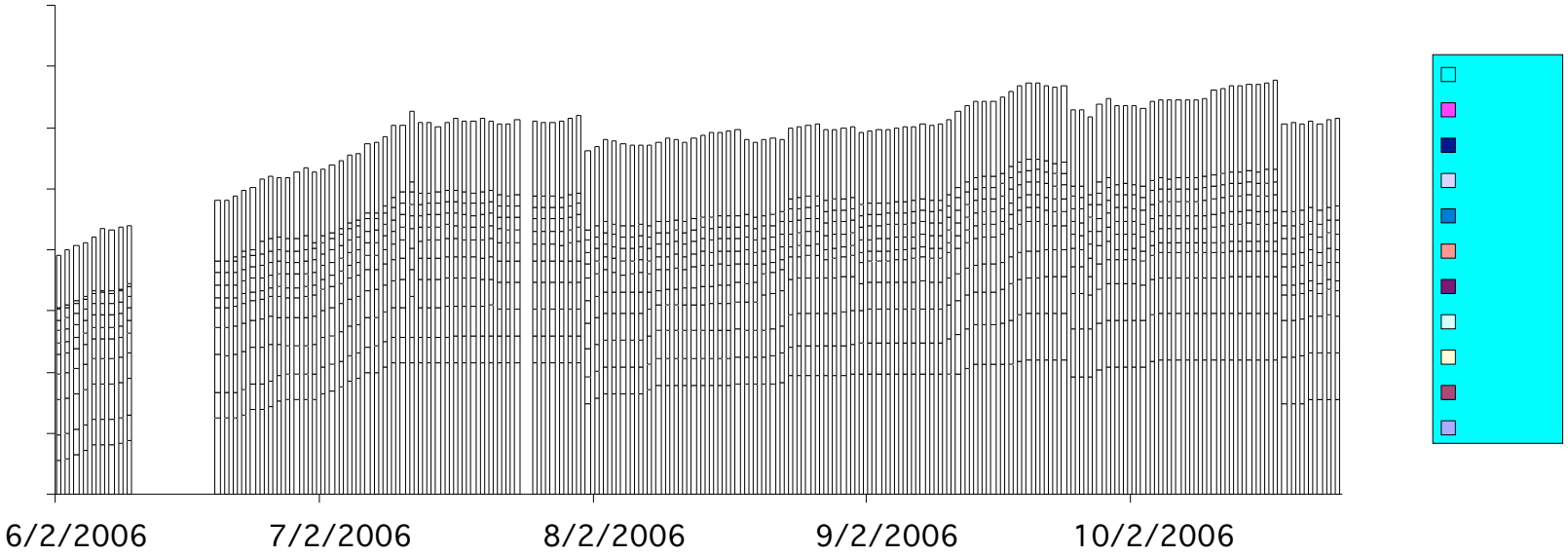
Current NGF Project Information

- **There are 56 projects using /project**
  - **Project directories created by user request**
  - **Utilization ranges between**
    - **0 Bytes (5 projects)**
    - **Multi-TB (4 projects)**
  - **10 projects account for ~75% storage used**
  - **Largest is 10TB quota**
- **When NERSC-5 installed, all users will have their Cray /home data in NGF**
  - **~2,500 users**
  - **~300 projects**

/Project Usage



/Project Capacity Usage



User Feedback

- **User feedback**
  - **Easy of Use**
    - **Unix file groups vs. HPSS repositories**
    - **Quotas**
  - **Performance**
    - **Sufficient for many projects**
  - **Availability and reliability**
    - **Outages of NGF have been noticed – so the good news is someone is using it**
    - **Seaborg outage less impact on users with data in project**
- **System wide MTBF – 41 days**
- **MTTR – 4.75 hours**
- **Last outage >183 days ago**

Integrating HPSS as a GPFS backend

- **Two modes**
  - **Synchronous**
    - **GPFS and HPSS share a name space**
    - **Metadata actions confirmed**
    - **Users DMAPI events**
    - **HPSS metadata slows GPFS down**
    - **Demonstrated at SC 05**
  - **Archive Mode (Asynchronous)**
    - **Operates much like DMF**
    - **Data accessed through the GPFS file system and metadata controlled by GPFS**
    - **File data flows to HPSS using policy (how full the file system is, how old the file, etc.**
    - **Dual residency – means data does not need to be backed up**
      - Need to backup GPFS metadata
    - **Administrators control the flow of data**
    - **Due for demonstration at SC 06**
- **Interface is independent of use in the global file system**

Archive Mode

- **Archive mode provides automated archival storage solution for a GPFS file system with minimal impact on file system performance.**
- **Utilizes the policy manager in GPFS 3.2 that enables ILM (information lifetime management) and HSM (hierarchical storage management) functionality.**
  - **HPSS will appear in GPFS like an external storage pool.**
- **Uses site defined GPFS policy rules to call HPSS provided programs that perform both multi-threaded and multi-noded I/O to HPSS using its client API.**
- **Uses DMAPI I/O events in the GPFS file systems to recall or stage data back from HPSS to the file system for data previously migrated to HPSS.**
  - **Users can explicitly stage data back as well**
- **Leaves metadata for all files (even ones migrated to HPSS) in the GPFS filesystem.**
  - **Will provide a file system backup utility to protect metadata crucial to HPSS data retrieval as well as data that has not migrated.**
- **Status**
  - **Nearing design completion for the archive mode of GPFS-HPSS integration project. Expected design complete is 30 Nov 06.**
  - **Doing proof-of-concept for archive mode for demo at IBM GPFS booth at SC06.**
    - **Demo will include an ability to use the new GPFS 3.2 policy manager in allowing a site to define rules (like SQL and very flexible policy language) to determine when and where data migrates automatically. The demo will show the ability of the GPFS policy manager to move data automatically between a GPFS storage pool and the HPSS external storage pool (a specific HPSS COS).**

Summary

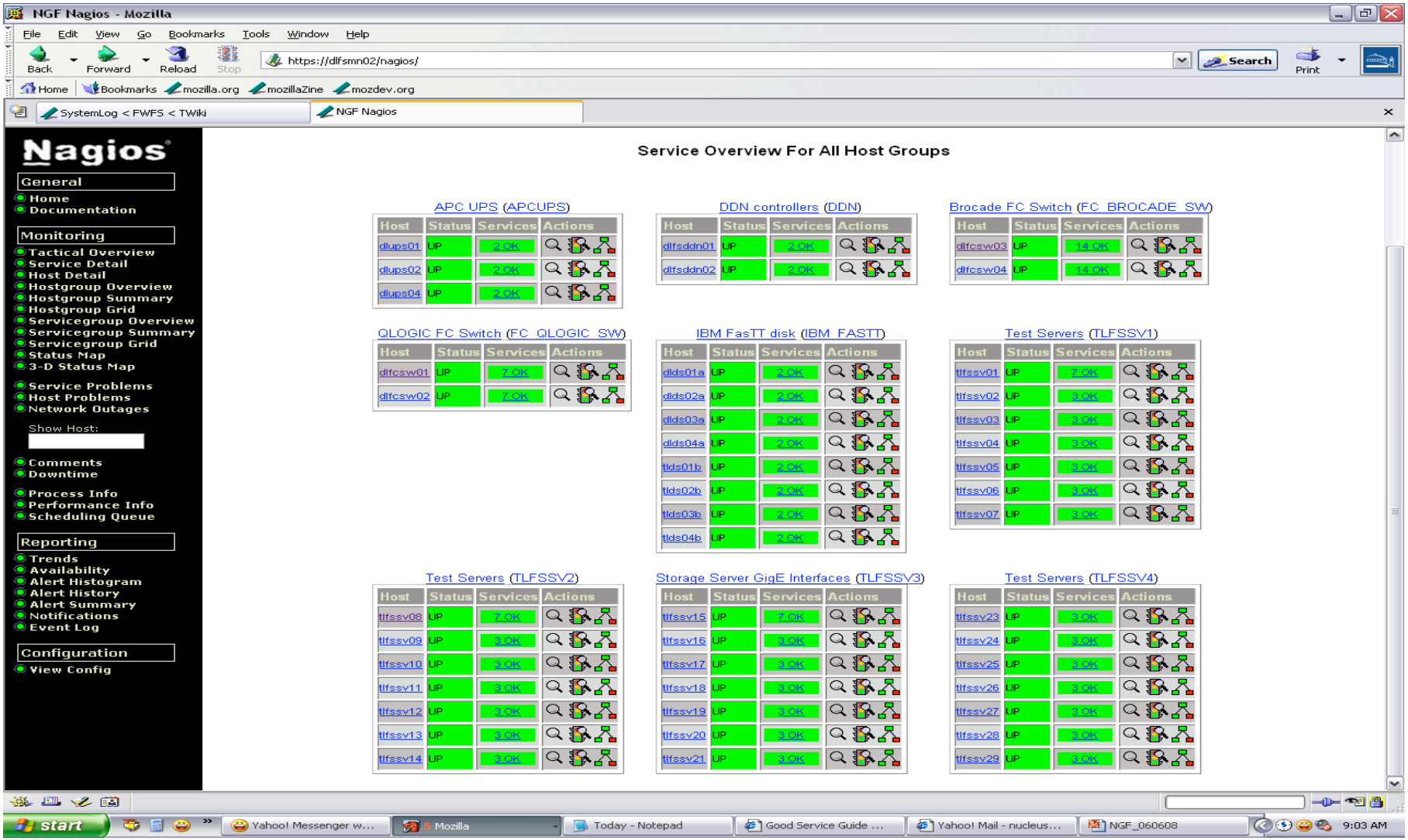
- **Four years ago, NERSC set a goal of a single uniform global file system running at high performance**
- **Two years ago, we understood what needed to be done**
- **Now a global, high performance, production quality filesystem has been realized**
- **We have a pathforward that allows all architectures to participate fully**
- **There are already a huge benefit to a number of users**
- **Two years from now, we expect to report all systems and users are using the global filesystem, many exclusively.**

NGF Monitoring

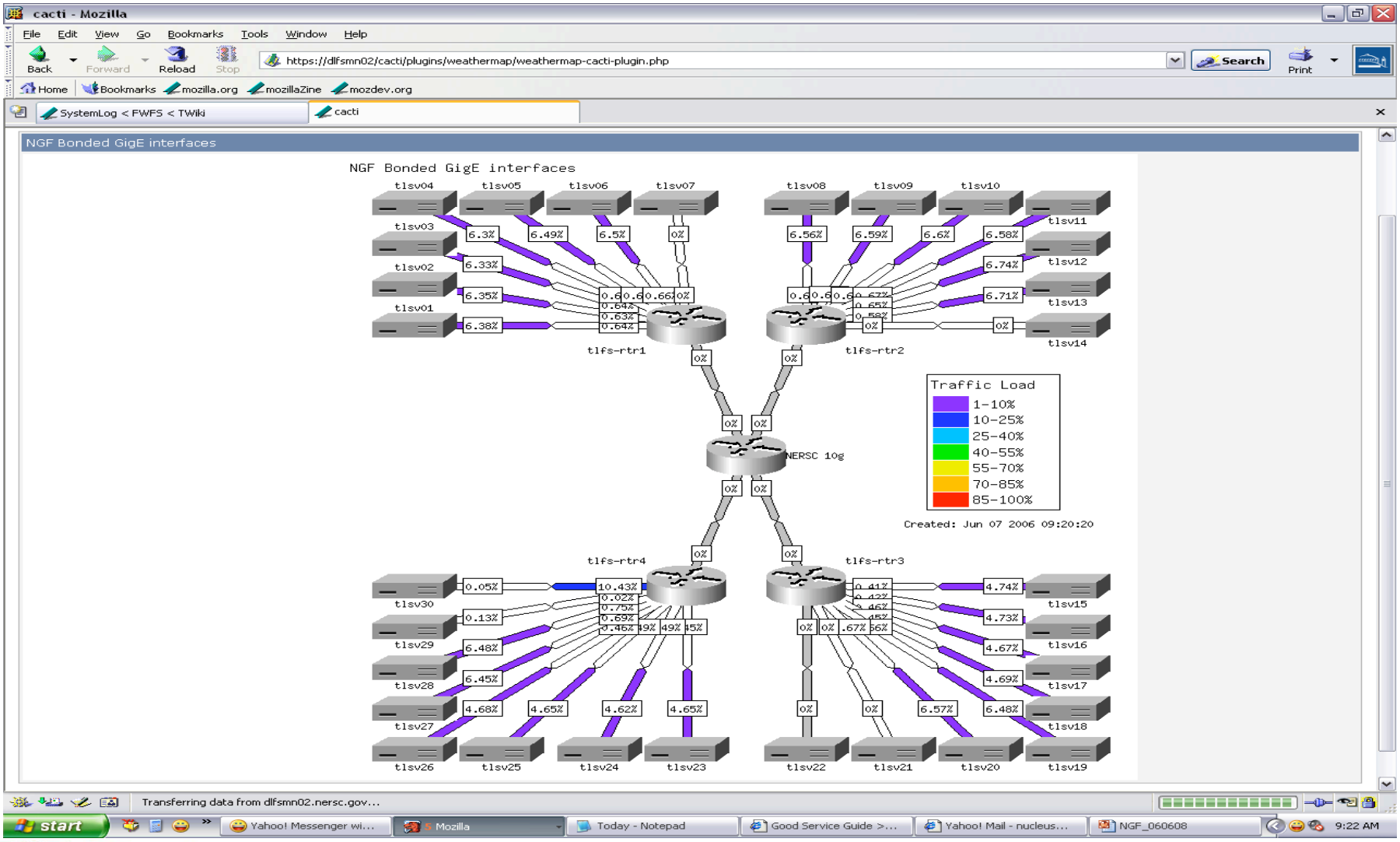
Proactive Monitoring

- **Nagios event detection and notification**
  - **Disk faults and soft failures**
  - **Server crashes**
  - **Nodes/Systems currently being monitored:**
    - **UPS: 3 APC UPS**
    - **FC Switches: 2 Brocade FC switches, 2 Qlogic FC switches**
    - **Storage: 2 DDN controllers, 4 IBM FAStTs**
    - **Servers: 28 NGF servers**
  - **Nagios allows event-driven procedures for Ops**
- **Cacti performance tracking**
  - **NSD servers: disk I/O, network traffic, cpu and memory usage, load average**
  - **FC switches: FC port statistics, fan, temperature**
  - **DDN: FC port statistics (IO/s, MB/s)**

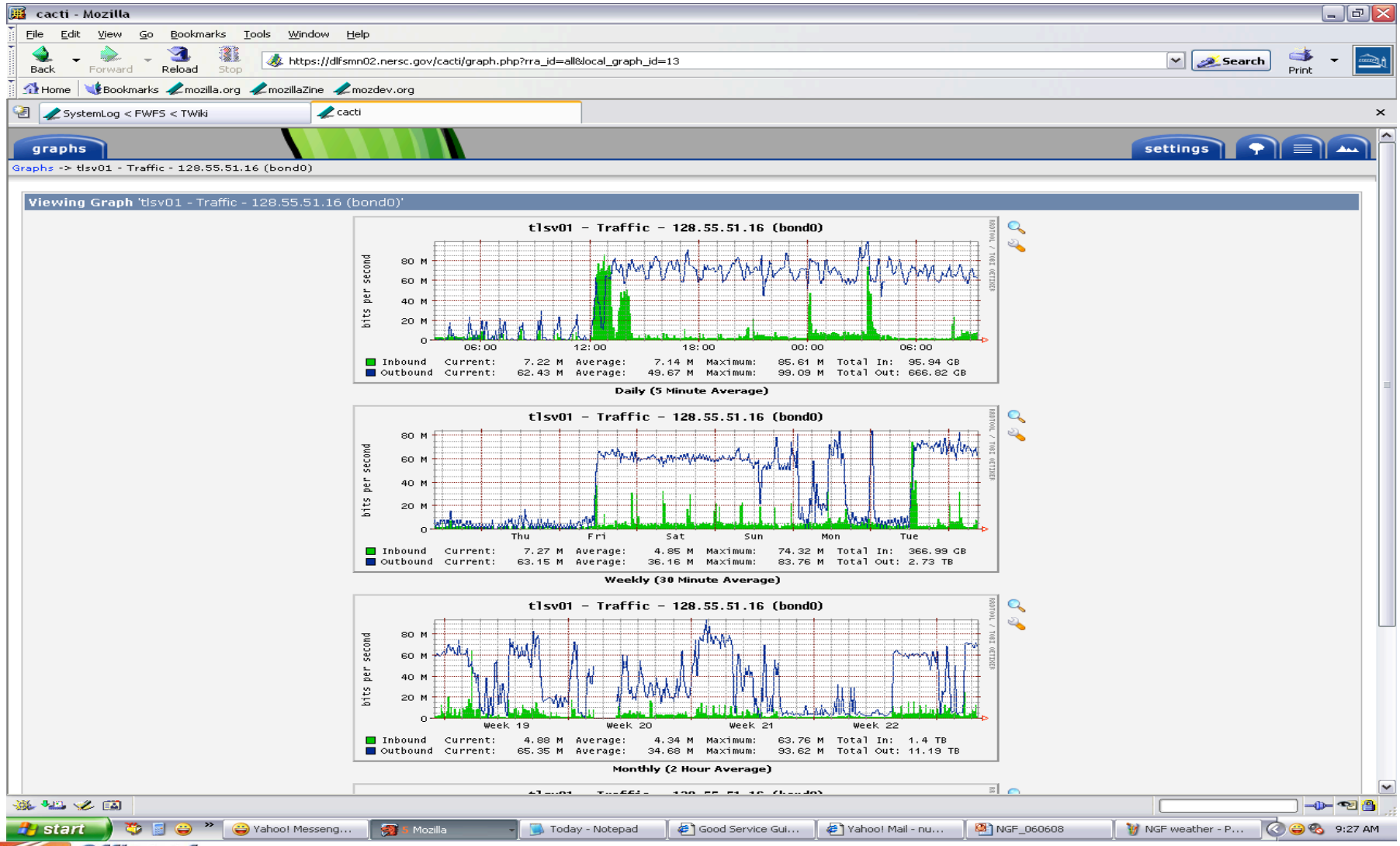
Event Monitoring with Nagios



Performance Tracking with Cacti



NSD Server Network Performance History



NERSC-5 SAN Fabric Topology