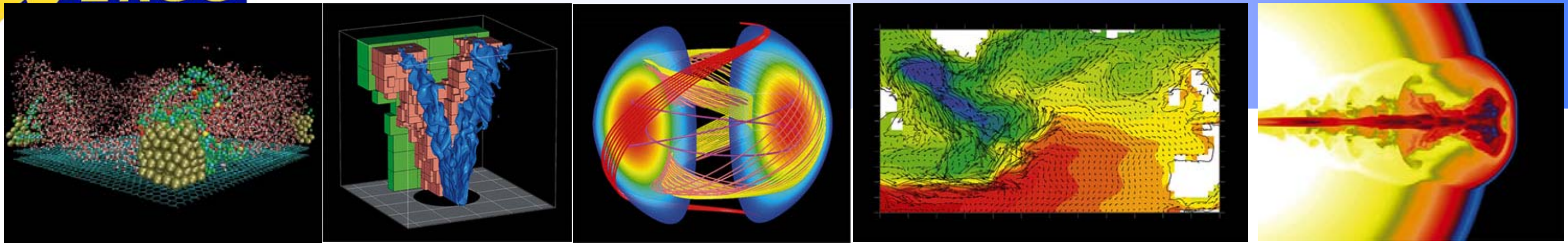


NERSC Experience: Implementation of a Facility Wide Global Filesystem

William T.C. Kramer

kramer@nersc.gov

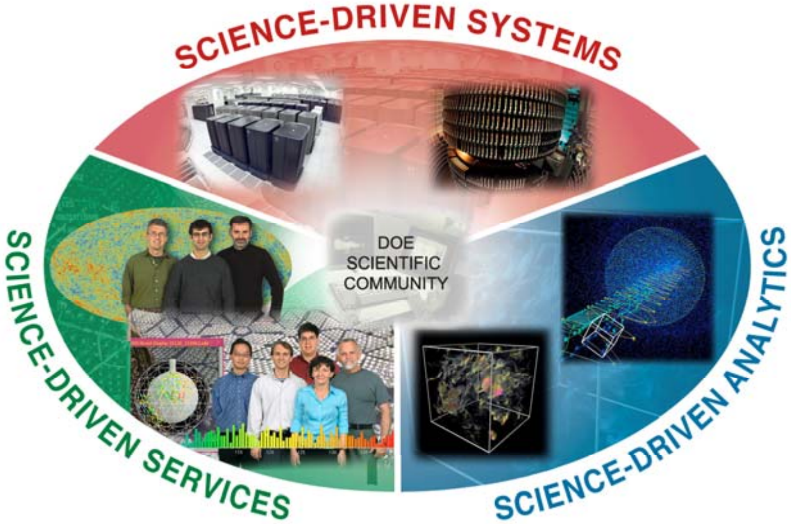
510-486-7577

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**



Two Years Ago
I talked about a project call GUPFS

- **Multi year project to deploy a center-wide shared file system at NERSC**
  - **Purpose to make advanced scientific research using NERSC systems more efficient and productive**
  - **Simplify end user data management by providing a shared disk file system in NERSC production environment**
- **GUPFS - Global, Unified, Parallel Filesystem**
  - **Global/Unified**
    - **A file system shared by major NERSC systems**
    - **Using consolidated storage and providing unified name space**
    - **Automatically sharing user files between systems without replication**
    - **Integration with HPSS and Grid is highly desired**
  - **Parallel**
    - **File system providing performance that is scalable as the number of clients and storage devices increase**
- **Today – we have in it in production for over 1 year.**

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**

Technologies Investigated

- **File Systems**
  - Sistina GFS 4.2, 5.0, 5.1, and 5.2 Beta
  - ADIC StorNext File System 2.0 and 2.2
  - Lustre 0.6 (1.0 Beta 1), 0.9.2, 1.0, 1.0.{1,2,3,4}
  - IBM GPFS for Linux, 1.3 and 2.2
  - Panasas
- **Fabric**
  - FC (1Gb/s and 2Gb/s): Brocade SilkWorm, Qlogic SANbox2, Cisco MDS 9509, SANDial Shadow 14000
  - Ethernet (iSCSI): Cisco SN 5428, Intel & Adaptec iSCSI HBA, Adaptec TOE, Cisco MDS 9509
  - Infiniband (1x and 4x): InfiniCon and Topspin IB to GE/FC bridges (SRP over IB, iSCSI over IB),
  - Inter-connect: Myrinnet 2000 (Rev D)
- **Storage**
  - Traditional Storage: Dot Hill, Silicon Gear, Chaparral
  - New Storage: Yotta Yotta GSX 2400, EMC CX 600, 3PAR, DDN S2A 8500

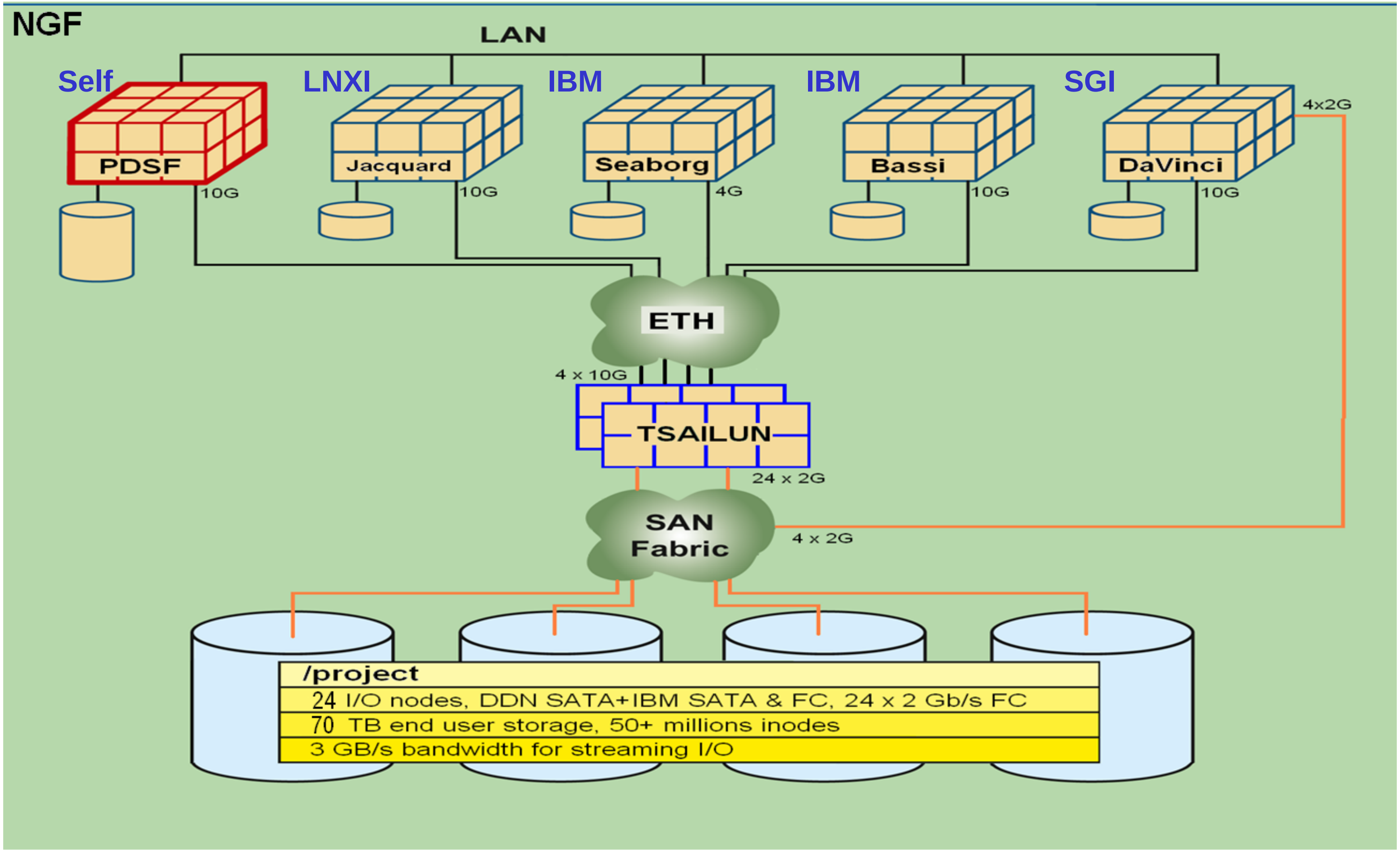
Evolution to Production
Walk and then Run

- **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**

NGF Architecture

- **NGF is configured with a GPFS owning cluster**
  - **Separate from all client production systems**
  - **Client system/clusters mount NGF file systems as multicluster remote clients**
  - **NGF owning cluster includes the NGF NSD server, contact nodes, and GPFS manager nodes**
- **Separate NGF owning clusters for each filesystem**
  - **/project has separate owning cluster and servers from global /home**
- **NERSC legacy systems mostly access NGF data via NGF NSD servers over Ethernet**
  - **NERSC-5 service nodes access NGF data via Fibre Channel**

NGF Production Configuration



/project

24 I/O nodes, DDN SATA+IBM SATA & FC, 24 x 2 Gb/s FC

70 TB end user storage, 50+ millions inodes

3 GB/s bandwidth for streaming I/O

Production NERSC Systems using NGF

- **NGF /project is currently mounted on all major NERSC systems (1250+ clients):**
  - **Jacquard, LNXI Opteron System running SLES 9**
    - **First non-IBM hardware to provide with supported GPFS**
      - LNXI has first and second level support responsibilities
  - **Da Vinci, SGI Altix running SLES 9 SP 3 with direct storage access**
  - **PDSF IA32 Linux cluster running Scientific Linux**
  - **Bassi, IBM Power5 running AIX 5.3**
  - **Seaborg, IBM SP running AIX 5.2**

NGF Current Production Configuration

- **Current production (/project) configuration:**
  - **24 I/O Server Nodes, Linux SLES9 SP3, GPFS 2.3 PTF16**
  - **70 TB usable end user storage**
    - **DDN 8500 with SATA drives**
    - **IBM DS4500 SATA drives**
    - **IBM DS4500 FC drives**
  - **50 million inodes**
  - **3+ GB/s bandwidth for streaming I/O**
  - **Storage and servers external to all NERSC systems**
  - **Distributed over 10 Gigabit Ethernet infrastructure**
  - **Single file system instance providing file and data sharing among multiple NERSC systems**
    - **Both large and small files**
    - **Persistent data, not scratch**
    - **Backed up to HPSS**

The /project file system

- **The global** /project **filesystem:**
  - **Access characteristics**
    - **Mountable remotely with R/W access, with *nosuid* and *nodev* mount option**
    - ***root* mapped to *nobody:nobody***
  - **Current usage**
    - **29.7 TB used (39% of capacity)**
    - **3.9 M inodes used (8% of capacity)**
  - **Backed up to HPSS bi-weekly**
  - **Default project quota:**
    - **1 TB**
    - **250,000 inodes**

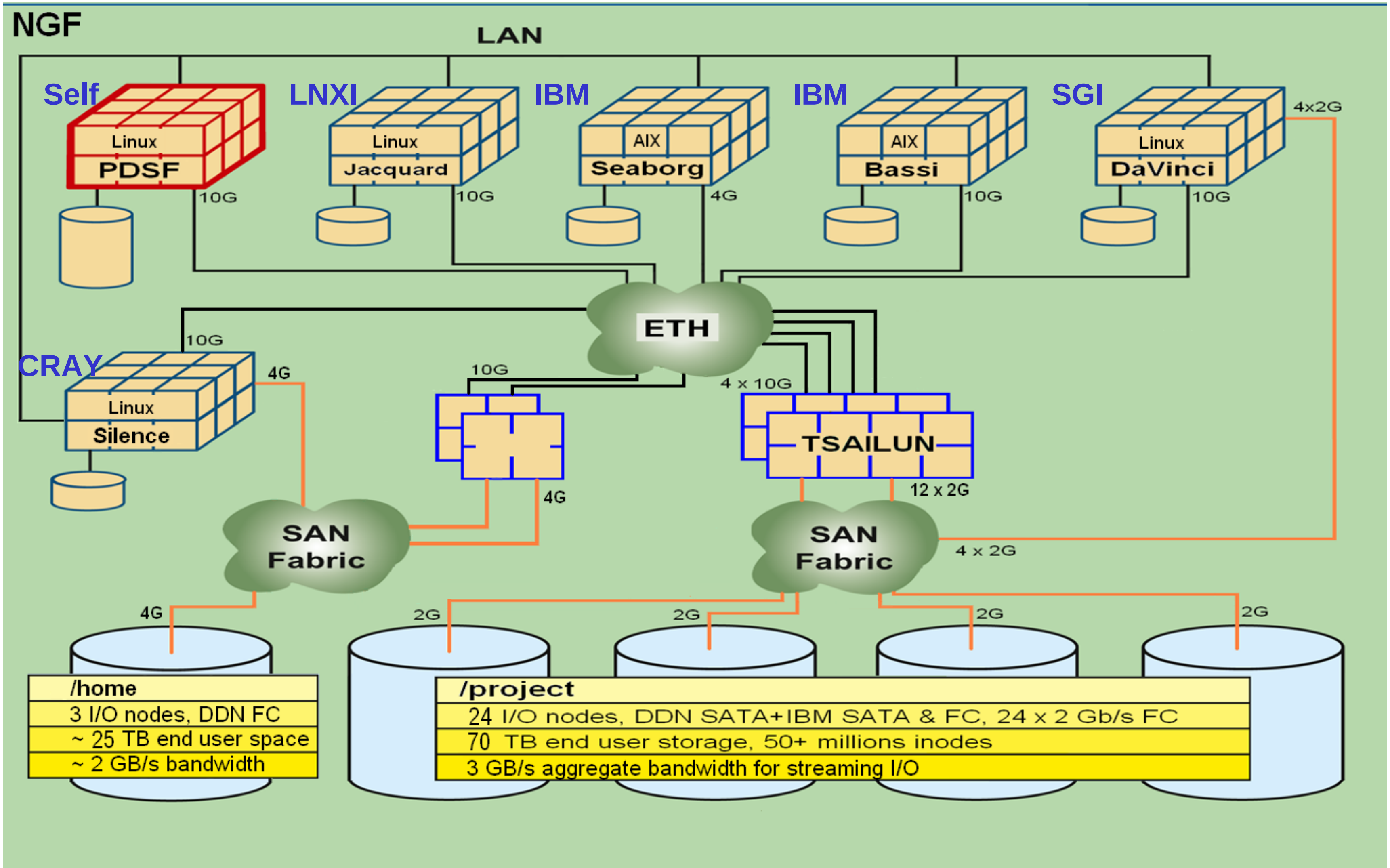
NGF /home Pre-production Configuration

- **Current global** /home **configuration:**
  - **3 I/O Server Nodes, Linux SLES9 SP3, GPFS 2.3 PTF16**
  - **32 TB usable end user storage**
    - **DDN 9550 with 16 Tiers 300 GB FC drives**
    - **26.5 TB currently allocated to home**
  - **50 million inodes**
  - **2.5+ GB/s bandwidth for streaming I/O**
  - **Storage and servers external to all NERSC systems**
  - **Distributed over 10 Gigabit Ethernet infrastructure and via direct Fiber Channel access**
  - **Single file system instance providing file and data sharing among multiple NERSC systems**
    - **Persistent data**
    - **Backed up to HPSS**
- **This is the first implementation of** /home **– it will grow significantly**

The /home file system

- **The global** /home **filesystem:**
  - **Access characteristics**
    - **Mountable remotely with R/W access, with *nosuid* and *nodev* mount option**
    - ***root* mapped to *nobody:nobody***
  - **Currently only accessed by NERSC-5 test system**
    - **Will be accessed by NERSC-5 when it is install**
    - **Will be accessed by legacy systems after NERSC-5 acceptance**
  - **Backed up to HPSS bi-weekly**
  - **Default global home quotas:**
    - **5 GB user quota**
    - **1 TB group quota**
    - **250,000 inodes**

NGF Configuration Oct 2006

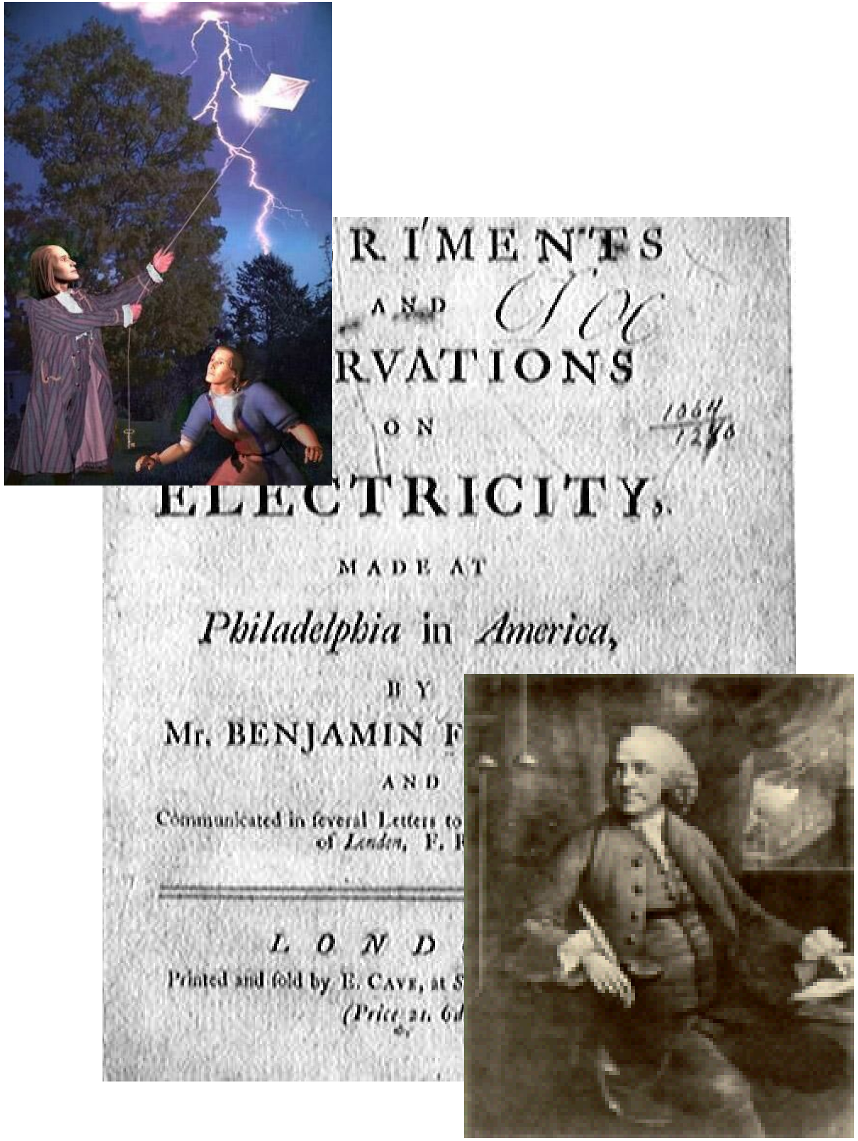


/home
3 I/O nodes, DDN FC
~ 25 TB end user space
~ 2 GB/s bandwidth

/project
24 I/O nodes, DDN SATA+IBM SATA & FC, 24 x 2 Gb/s FC
70 TB end user storage, 50+ millions inodes
3 GB/s aggregate bandwidth for streaming I/O

A Side Track about NERSC-5
franklin.nersc.gov

- **Named after Benjamin Franklin – America's First Scientist**
- **Worked in almost every area of interest to DOE**
  - **electricity, thermal dynamics, energy efficiency, climate and global warming, ocean currents, weather, materials, population growth, medicine and health, and many other areas.**



RIMENTS
AND
RVATIONS
ON
ELECTRICITY,
MADE AT
Philadelphia in *America*,
BY
Mr. BENJAMIN F
AND
Communicated in feveral Letters to
of *London*, F. R
L O N D
Printed and fold by E. CAVE, at S
(Price 2s. 6d

NERSC-5

- **Required all bidder to "integrate" with NGF**
  - **In order for this to work, IBM agreed to make GPFS available to any vendor/bidder**
    - **Binary already runs on LINUX and AIX and prices were set**
      - Vendors could make different support arrangements
    - **If vendor wanted source code, they would have to enter into a different business arrangement**
  - **All bidders agreed to integrate using GPFS**
- **Cray selected in an fully, open Best Value competition.**
- **NERSC-5 will be the largest Cray XT "Hood" system at time of delivery**
  - **Dual Core AMD processors at 2.6 GHz**
    - **This is a "Node" or PE**
  - **9762 Nodes = 19,524 CPUs**
    - **40 are "service node"**
      - All have 4 Gbps Fibre Channel connections
  - **4 GB of memory per node**
  - **New Seastar Interconnect**
    - **A 3D Torus**
    - **50 nanoseconds per hop**

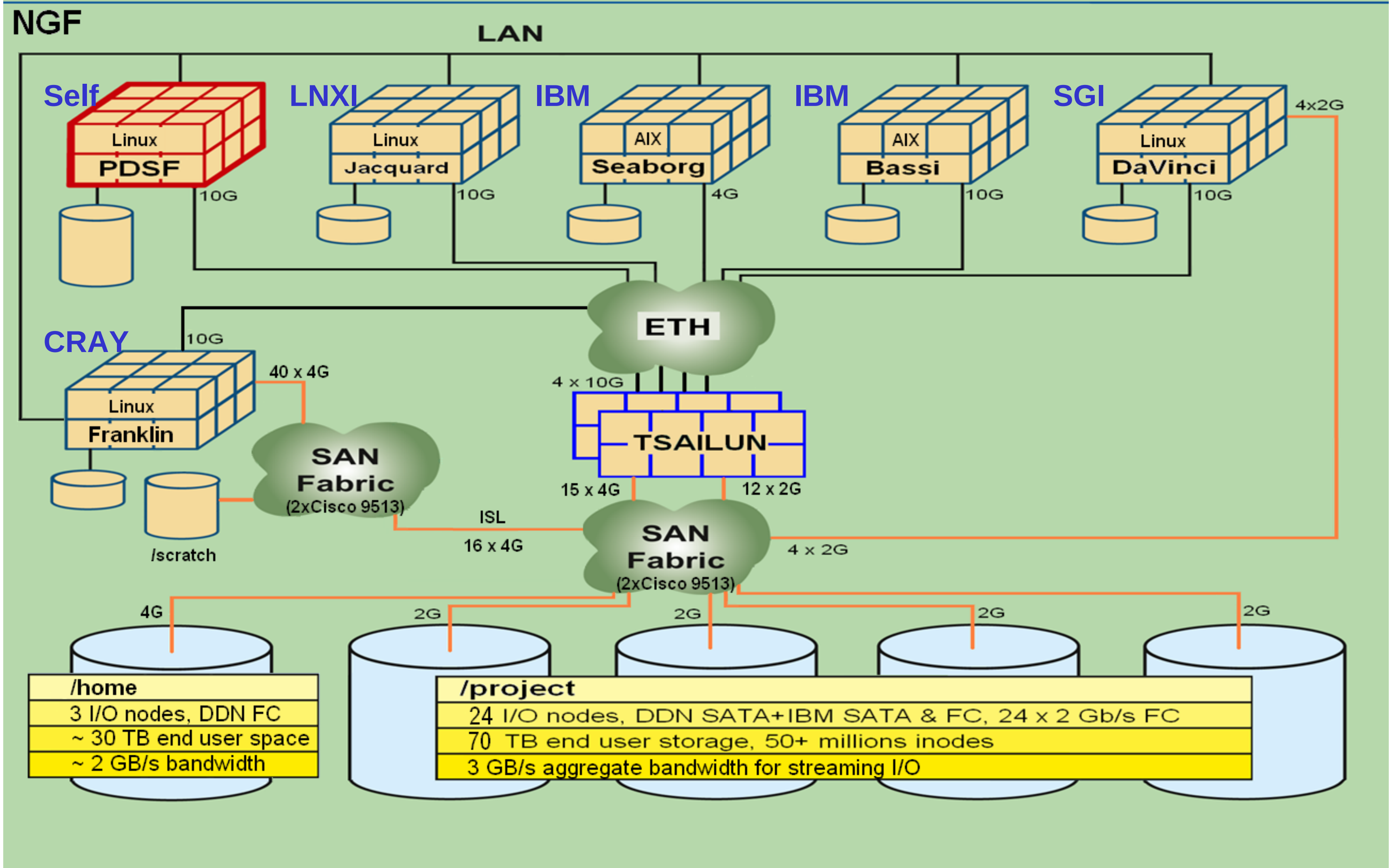
NERSC-5 is almost 10 Times all of NERSC’s Sustained Performance!

- **16.09 TF Sustained System Performance**
  - **Geometric Mean**
  - **Seaborg = .89 TF**
  - **Bassi ~ .8 TF**
- **6.3 TB/s Bi-Section Bandwidth**
  - **7.6 GB/s peak bi-directional bandwidth per link**
- **402 TB of usable disk**
  - **DDN SA 9500 controllers with 32 tiers of 290 GB/10K RPM drives in a 8+1 Raid configuration**
- **4 - 10 GigE connections**
- **32 – 1 GigE connections**
- **56 – 4 Gbps FibreChannel Connections**

The Phasing of NERSC-5

- **Small Test System**
  - **Summer 2006 – user access not planned**
- **Fall of 2006 - Phase 1**
  - **1/3 of compute resources**
  - **~80% of I/O infrastructure**
- **Winter 2007 – Phase 2**
  - **2/3 more compute nodes**
  - **Remaining disks and controllers**
- **Winter 2008 – option to upgrade to at least double the sustained performance**
- **Summer 2008 or earlier – Major software upgrade**
  - **More later**
- **Winter/Spring 2009 – option for a 1 Petaflop/s peak system**
  - **not currently in the NERSC budget**
  - **NERSC currently plans to install NERSC-6 in that timeframe**

NGF Configuration with NERSC-5 During Interim Phase



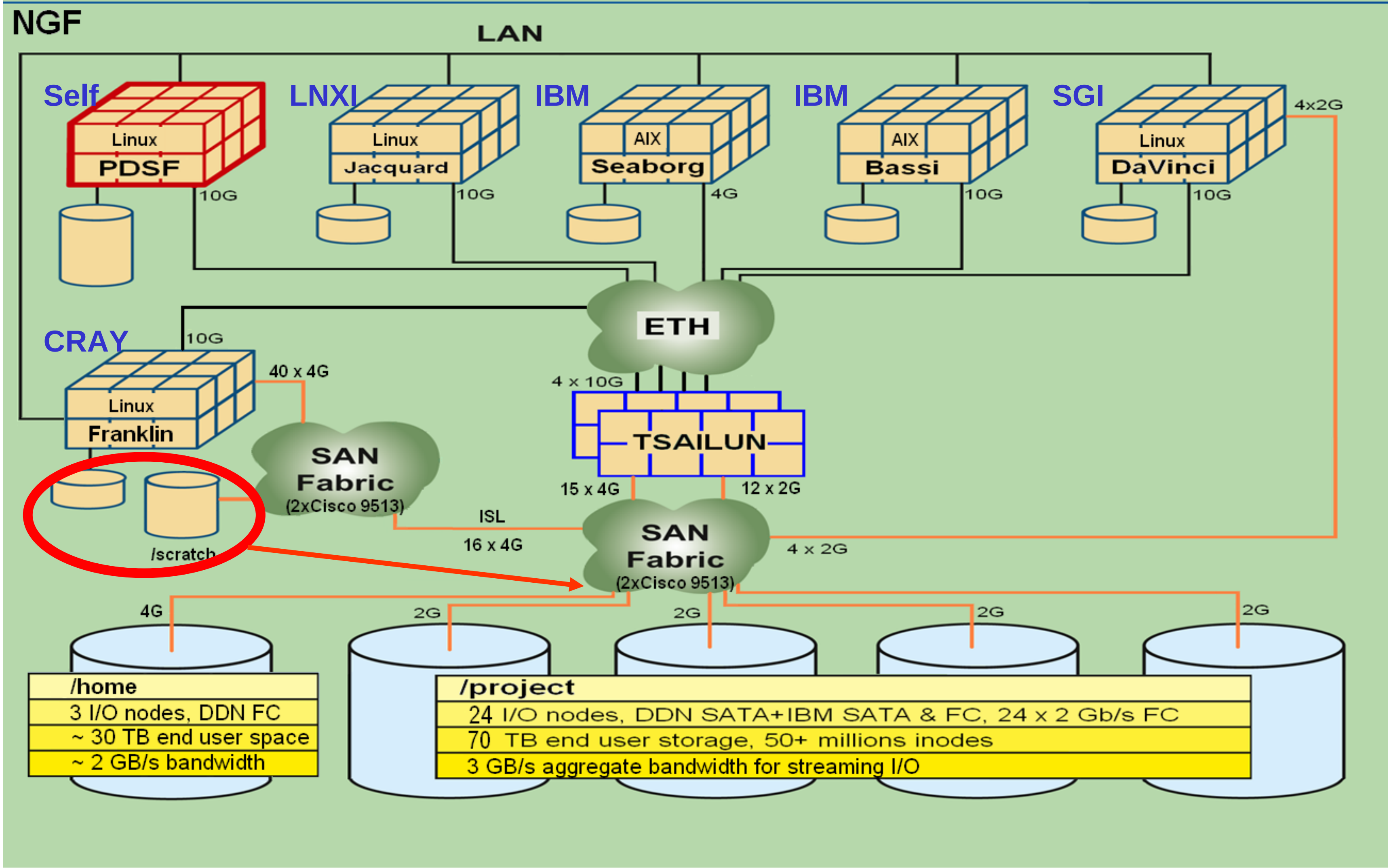
/home
3 I/O nodes, DDN FC
~ 30 TB end user space
~ 2 GB/s bandwidth

/project
24 I/O nodes, DDN SATA+IBM SATA & FC, 24 x 2 Gb/s FC
70 TB end user storage, 50+ millions inodes
3 GB/s aggregate bandwidth for streaming I/O

The Phasing of NERSC-5

- **Initially Franklin will have both GPFS and Lustre running until the "PetaScale I/O Interface" is developed and tested**
  - **Lustre will provide internal scratch file space**
    - **Available to all compute and service nodes**
  - **GPFS will provide all /home and /project file space**
    - **Available only on service nodes**
  - **"Copy-in/Copy-out" mode**
- **Conversion to GPFS-only system in 2008**
  - **"PetaScale I/O Interface" developed as part of the Cray Center of Excellence for System Management and Storage – located at NERSC.**
    - **Provides I/O forwarding to and from compute nodes and service nodes**
      - Much like the T3E implementation
    - **PI/OI will run on Cray's Compute Node Linux**
    - **Designed for performance, scalability and reliability**
  - **Once tested, PI/OI allows compute nodes to have full access to GPFS data.**
- **All Cray local /scratch disk will move into NGF as scratch disk – without perturbing the I/O rates**

NGF Configuration with NERSC-5 Full Implementation



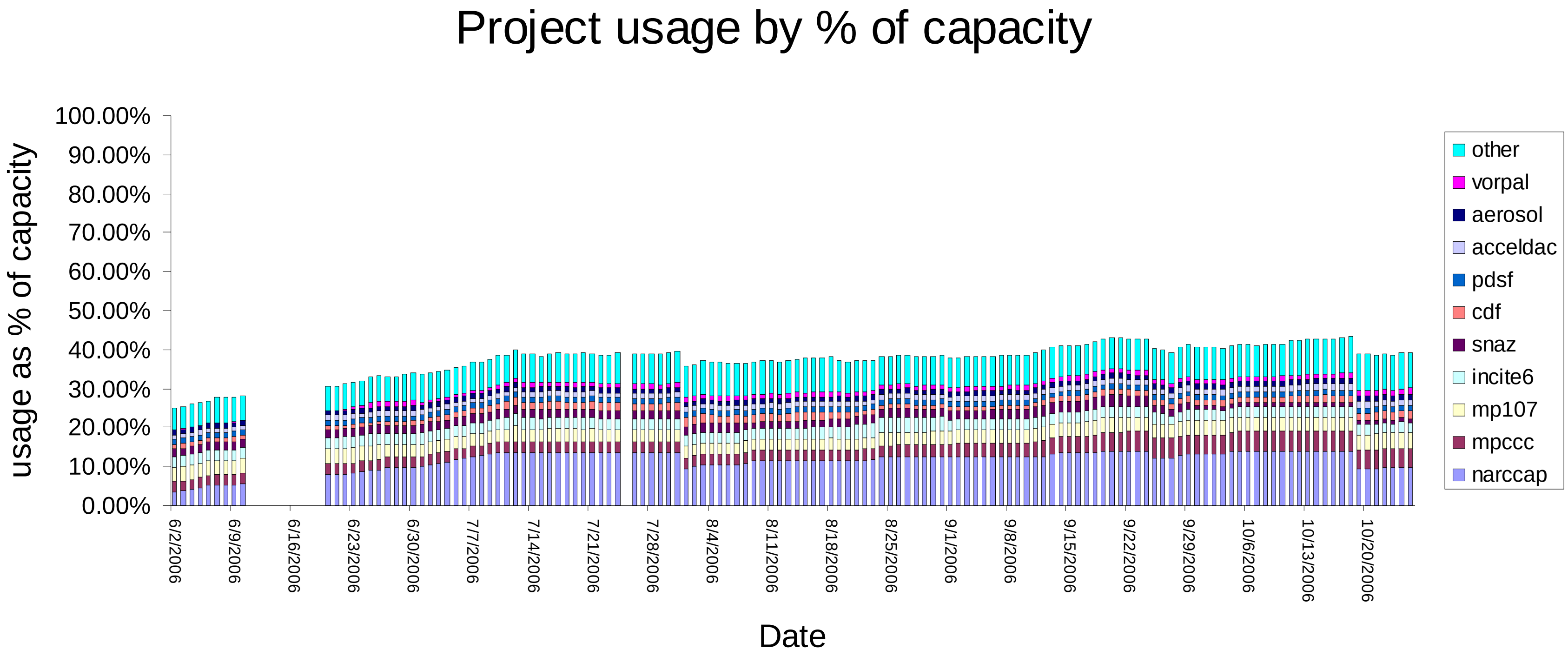
/home
3 I/O nodes, DDN FC
~ 30 TB end user space
~ 2 GB/s bandwidth

/project
24 I/O nodes, DDN SATA+IBM SATA & FC, 24 x 2 Gb/s FC
70 TB end user storage, 50+ millions inodes
3 GB/s aggregate bandwidth for streaming I/O

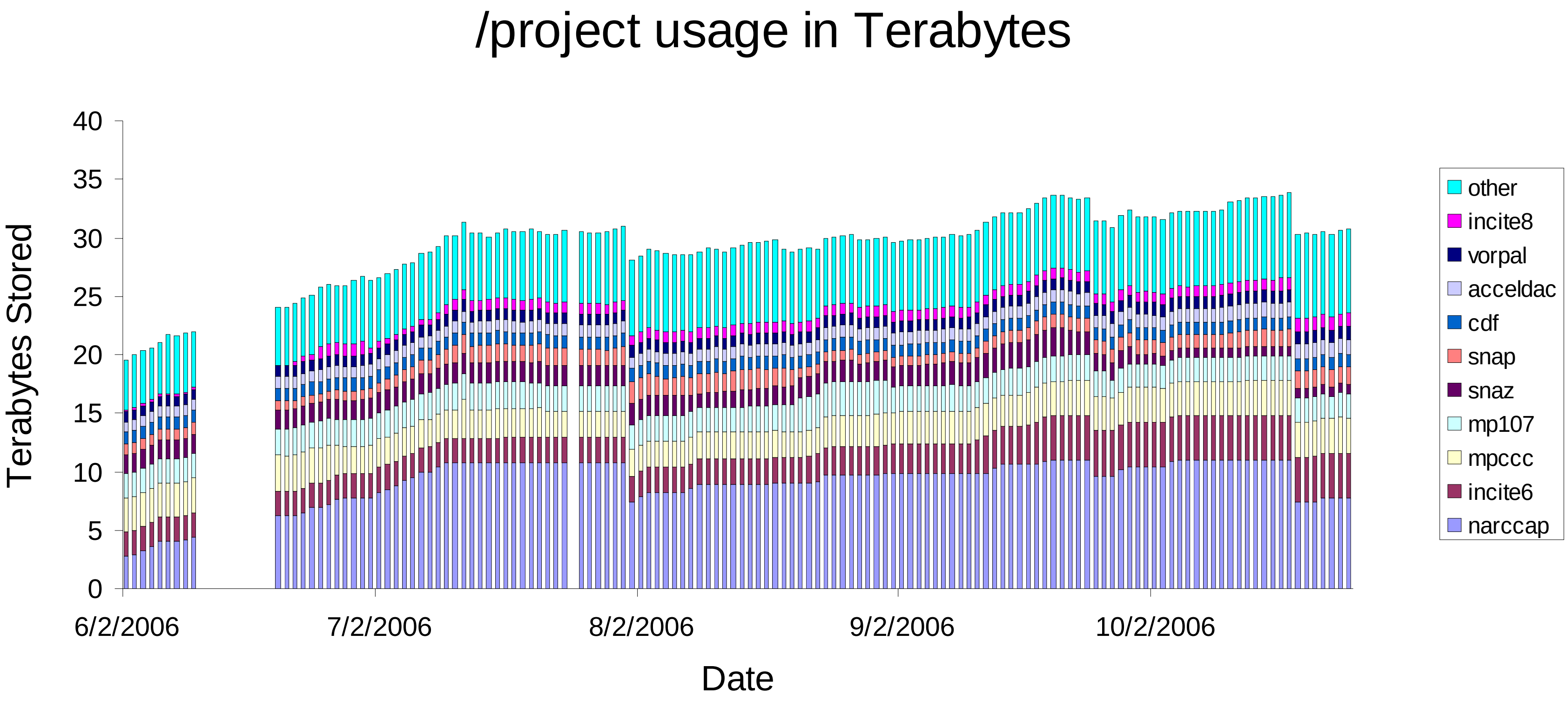
Current NGF Project Information

- **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**
- **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. 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**
    - **Early “Contagious” GPFS problems**
    - **Seaborg outage less impact on users with data in project**

NERSC Multicluster Experiences

General Experince

- **General impression of Multi-cluster is very positive**
- **GPFS multi-cluster is robust and reliable**
  - **Server failover is basically bullet-proof**
  - **Easy to configure and deploy**
- **IBM support has provided good, responsive support for GPFS**
- **GPFS multi-cluster is an excellent network configuration tester**
  - **It expects a well configured network and will let you know about if it is not.**
- **Inter-client communication issues appear to be addressed by GPFS 3.1**
- **Scheduled maintenance is a problem**
  - **Rolling upgrades are you friend**

NGF System-wide Reliability

- **Eight system wide outages in 326 days of production service**
- **System wide MTBF – 41 days**
- **MTTR – 4.75 hours**
- **Last outage 165 days ago**
  - **Outages**
    - **2/10: 09:00 – 21:35: GPFS Upgrade + New Servers (up 70 days)**
    - **2/14: 11:10 – 11:15: Seaborg gateway node crash -> GPFS hung - GPFS bug [PMR69276] (up 4 days)**
    - **3/13: 14:32 – 05:12 FAStT disk failure -> LUN/ARRAY failure (up 27 days)**
    - **4/20: 16:35 – 20:30: Server crashes -> NSD disk down (up 37 days)**
    - **5/10: 08:58 – 10:24: FC switch failures – firmware bug (up 19 days)**
    - **5/11: 03:47 – 09:18: FC switch failures – firmware bug (up 0 day)**
    - **5/11: 10:12 – 11:16: Brocade switch firmware upgrade (up 0 day)**
    - **5/16: 10:30 – 11:53: DDN disk failure -> controller failures - DDN firmware bug. DDN firmware upgrade (up 4 days)**
- **If you treat this as one longer outage**
  - **6 outages in 326 days, MTBF of 54 days and MTTR of 9:24**

Improvements in Reliability

- **Pro-active monitoring**
- **Developing better procedures**
- **Operations staff training and activities**
- **PMRs filed and fixes applied**
- **Replacing old servers in NGF**

GPFS-HPSS Integration

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

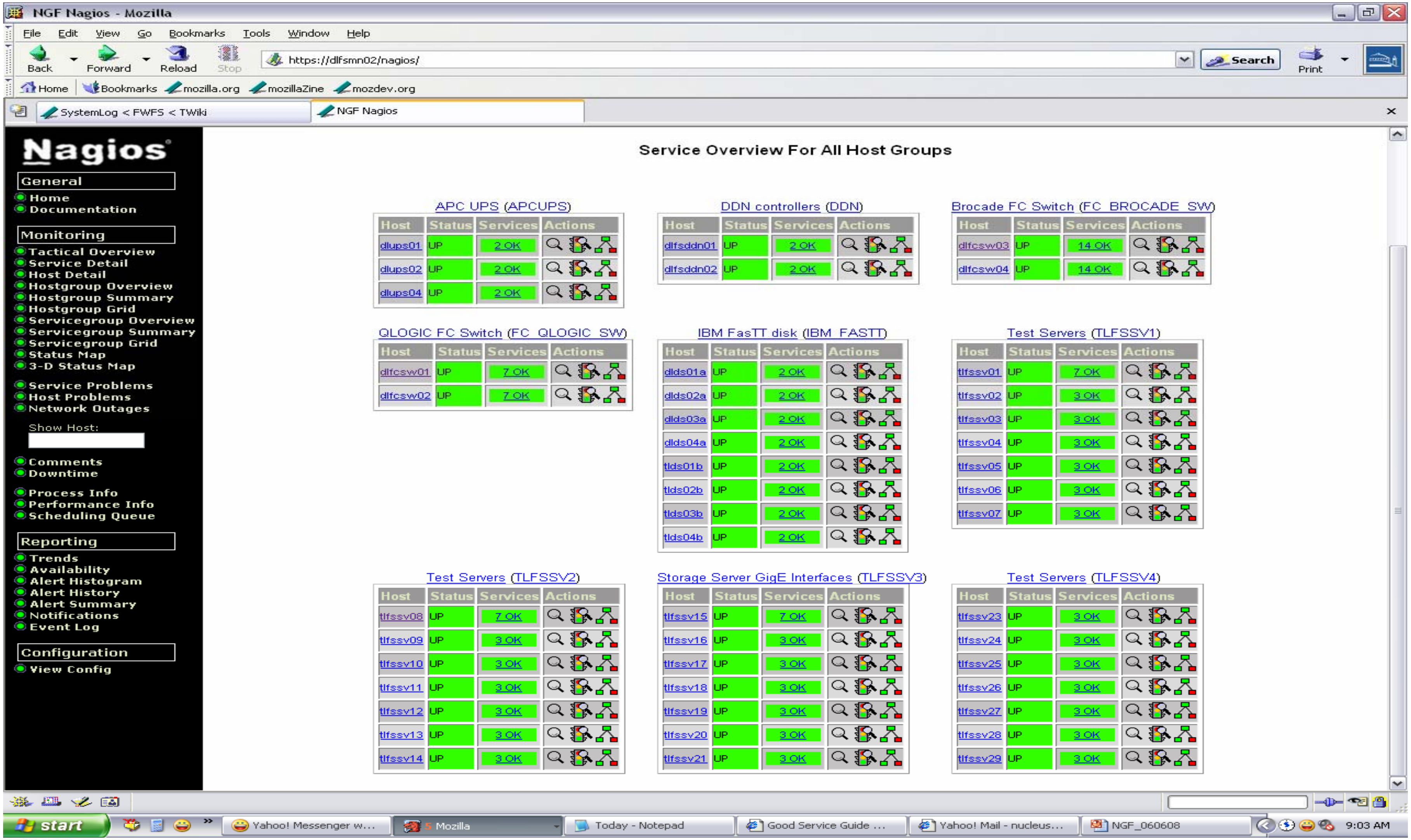
- **Archive mode provides automated archival storage solution for a GPFS file system with minimal impact on file system performance.**
- **Utilizes the policy manager in GPFS 3.2 that enables ILM (information lifetime management) and HSM (hierarchical storage management) functionality.**
  - **HPSS will 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).**

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



NGF Nagios - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop https://dlfsmn02/nagios/ Search Print

Home Bookmarks mozilla.org mozillaZine mozdev.org

SystemLog < FWFS < TWiki

NGF Nagios

Service Overview For All Host Groups

APC UPS (APCUPS)

Host	Status	Services	Actions
dlups01	UP	2 OK	🔍
dlups02	UP	2 OK	🔍
dlups04	UP	2 OK	🔍

DDN controllers (DDN)

Host	Status	Services	Actions
dlfsddn01	UP	2 OK	🔍
dlfsddn02	UP	2 OK	🔍

Brocade FC Switch (FC_BROCADE_SW)

Host	Status	Services	Actions
dlfcsw03	UP	14 OK	🔍
dlfcsw04	UP	14 OK	🔍

QLOGIC FC Switch (FC_QLOGIC_SW)

Host	Status	Services	Actions
dlfcsw01	UP	7 OK	🔍
dlfcsw02	UP	7 OK	🔍

IBM FasTT disk (IBM_FASTT)

Host	Status	Services	Actions
dlds01a	UP	2 OK	🔍
dlds02a	UP	2 OK	🔍
dlds03a	UP	2 OK	🔍
dlds04a	UP	2 OK	🔍
tlds01b	UP	2 OK	🔍
tlds02b	UP	2 OK	🔍
tlds03b	UP	2 OK	🔍
tlds04b	UP	2 OK	🔍

Test Servers (TLFSSV1)

Host	Status	Services	Actions
tlfssv01	UP	7 OK	🔍
tlfssv02	UP	3 OK	🔍
tlfssv03	UP	3 OK	🔍
tlfssv04	UP	3 OK	🔍
tlfssv05	UP	3 OK	🔍
tlfssv06	UP	3 OK	🔍
tlfssv07	UP	3 OK	🔍

Test Servers (TLFSSV2)

Host	Status	Services	Actions
tlfssv08	UP	7 OK	🔍
tlfssv09	UP	3 OK	🔍
tlfssv10	UP	3 OK	🔍
tlfssv11	UP	3 OK	🔍
tlfssv12	UP	3 OK	🔍
tlfssv13	UP	3 OK	🔍
tlfssv14	UP	3 OK	🔍

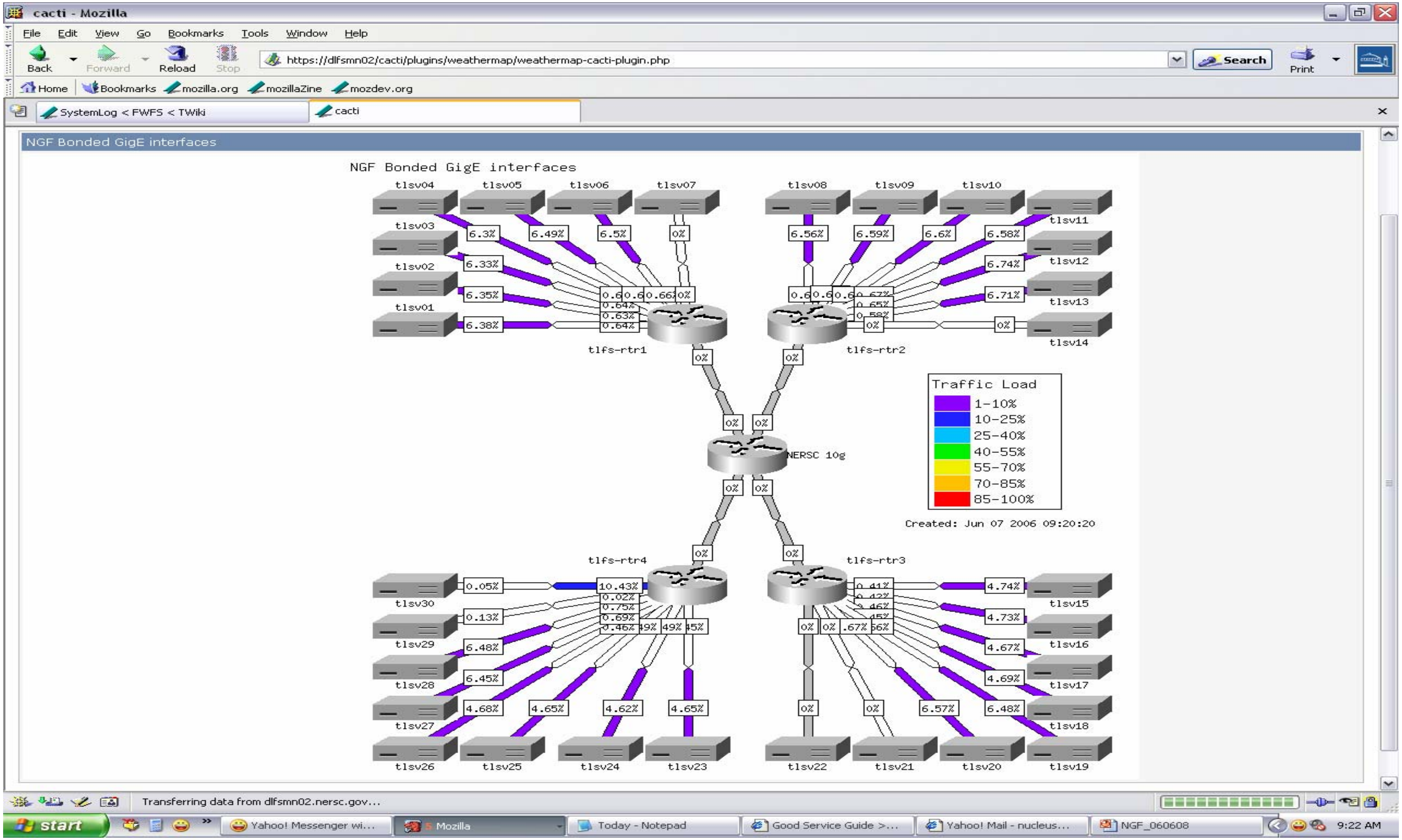
Storage Server GigE Interfaces (TLFSSV3)

Host	Status	Services	Actions
tlfssv15	UP	7 OK	🔍
tlfssv16	UP	3 OK	🔍
tlfssv17	UP	3 OK	🔍
tlfssv18	UP	3 OK	🔍
tlfssv19	UP	3 OK	🔍
tlfssv20	UP	3 OK	🔍
tlfssv21	UP	3 OK	🔍

Test Servers (TLFSSV4)

Host	Status	Services	Actions
tlfssv23	UP	3 OK	🔍
tlfssv24	UP	3 OK	🔍
tlfssv25	UP	3 OK	🔍
tlfssv26	UP	3 OK	🔍
tlfssv27	UP	3 OK	🔍
tlfssv28	UP	3 OK	🔍
tlfssv29	UP	3 OK	🔍

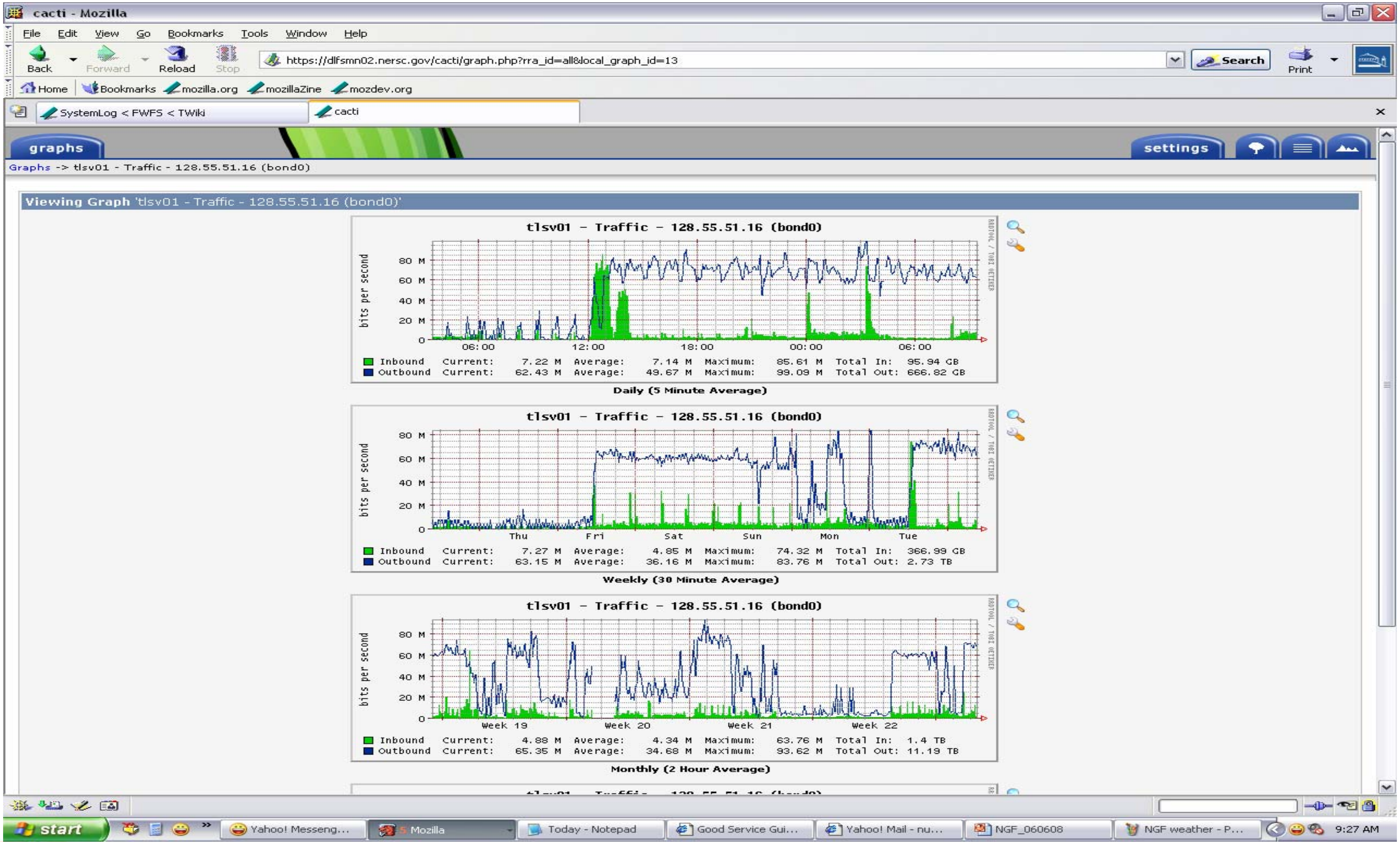
Performance Tracking with Cacti



Created: Jun 07 2006 09:20:20

Transferring data from dlfsmn02.nersc.gov...

NSD Server Network Performance History



https://dlfsmn02.nersc.gov/cacti/graph.php?rra_id=all&local_graph_id=13

Graphs -> tlsv01 - Traffic - 128.55.51.16 (bond0)

Viewing Graph 'tlsv01 - Traffic - 128.55.51.16 (bond0)'

tlsv01 - Traffic - 128.55.51.16 (bond0)

```
■ Inbound   Current:   7.22 M  Average:   7.14 M  Maximum:  85.61 M  Total In:   95.94 GB
■ Outbound  Current:  62.43 M  Average:  49.67 M  Maximum:  99.09 M  Total Out: 666.82 GB
```

Daily (5 Minute Average)

tlsv01 - Traffic - 128.55.51.16 (bond0)

```
■ Inbound   Current:   7.27 M  Average:   4.85 M  Maximum:  74.32 M  Total In:  366.99 GB
■ Outbound  Current:  63.15 M  Average:  36.16 M  Maximum:  83.76 M  Total Out: 2.73 TB
```

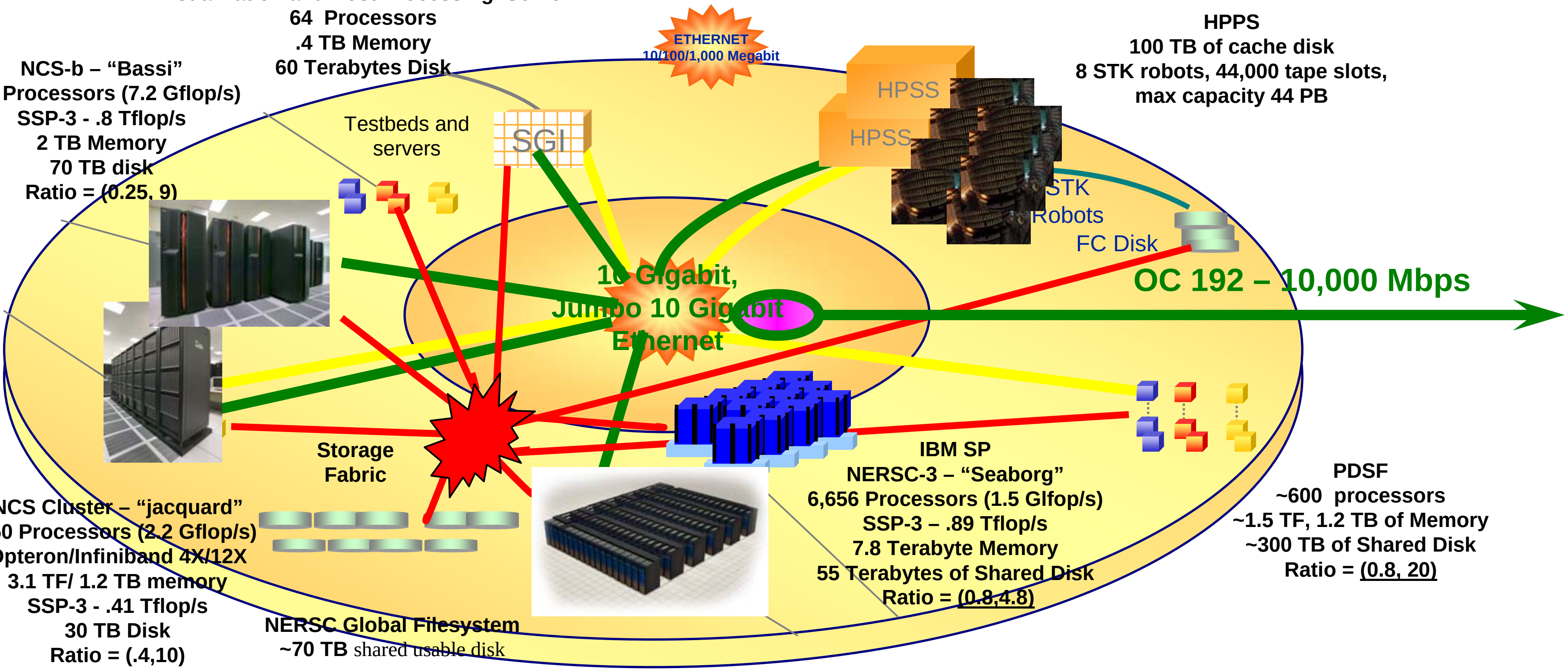
Weekly (30 Minute Average)

tlsv01 - Traffic - 128.55.51.16 (bond0)

```
■ Inbound   Current:   4.88 M  Average:   4.34 M  Maximum:  63.76 M  Total In:  1.4 TB
■ Outbound  Current:  65.35 M  Average:  34.68 M  Maximum:  93.62 M  Total Out: 11.19 TB
```

Monthly (2 Hour Average)

2007



Visualization and Post Processing Server
64 Processors
.4 TB Memory
60 Terabytes Disk

NCS-b – "Bassi"
976 Processors (7.2 Gflop/s)
SSP-3 - .8 Tflop/s
2 TB Memory
70 TB disk
Ratio = (0.25, 9)

Testbeds and
servers

SGI

ETHERNET
10/100/1,000 Megabit

HPSS

HPSS

HPPS
100 TB of cache disk
8 STK robots, 44,000 tape slots,
max capacity 44 PB

STK
Robots

FC Disk

10 Gigabit,
Jumbo 10 Gigabit
Ethernet

OC 192 – 10,000 Mbps

Storage
Fabric

IBM SP
NERSC-3 – "Seaborg"
6,656 Processors (1.5 Glfop/s)
SSP-3 – .89 Tflop/s
7.8 Terabyte Memory
55 Terabytes of Shared Disk
Ratio = (0.8,4.8)

PDSF
~600 processors
~1.5 TF, 1.2 TB of Memory
~300 TB of Shared Disk
Ratio = (0.8, 20)

NCS Cluster – "jacquard"
650 Processors (2.2 Gflop/s)
Opteron/Infiniband 4X/12X
3.1 TF/ 1.2 TB memory
SSP-3 - .41 Tflop/s
30 TB Disk
Ratio = (.4,10)

NERSC Global Filesystem
~70 TB shared usable disk

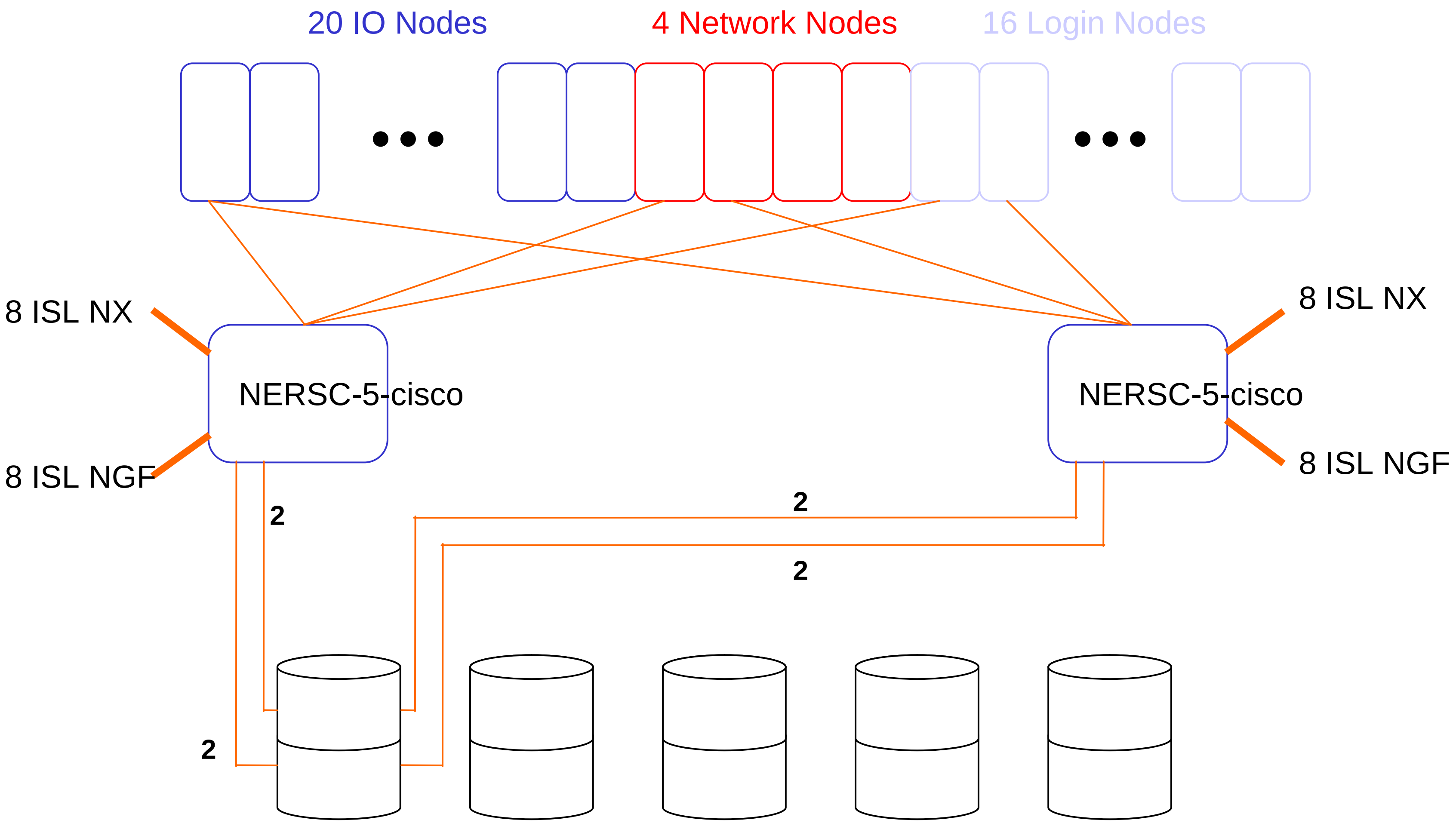
Cray XT
NERSC-5 – "Franklin"
19,584 Processors (5.2 Gflop/s)
SSP-3 ~16.1 Tflop/s
39 TB Memory
300 TB of shared disk
Ratio (.4, 3)

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

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.**

NERSC-5 SAN Fabric Topology



Current Architecture Limitations

- **NGF architectural limitations on legacy system performance**
  - **Majority of NGF access via IP-based GPFS traffic**
    - **Only Da Vinci can access NGF storage directly**
  - **Engenio storage multipath performance deficiencies**
    - **Accessing a L UN via multiple paths results in “flapping” and reduced performance**
      - Requires Engenio/IBM RDAC multipath driver to coordinate
      - RDAC available for Linux and AIX, but unable to test it as all Engenio storage in production
      - Need to test interoperability AIX, and Linux IA32, x86_64, and IA64 systems
    - **Currently, access to a LUN is limited to a single path**
    - **Da Vinci restricted to accessing Engenio storage via IP links**
  - **Limited IP gateway bandwidth between NERSC systems and NGF**
    - **Routing issue on Jacquard**
    - **Bonded Ethernet performance on Seaborg & NGF**
    - **Limited bandwidth into each system**

GPFS PMRs Filed

- **PMR 51072 – GPFS clients unable to communicate to other clients in a remote cluster when its tcpPort is changed.**
  - **A DCR was filed.**
- **PMR 66186 – GPFS crashed during shutdown.**
  - **Fixed in PTF8.**
- **PMR 69276 – GPFS hung when many nodes are failing about the same time.**
  - **APAR IY81318. Fixed in PTF11. Efix available for PTF10.**
- **PMR 69943 – The “ftruncate” bug. User’s make/build failed on Bassi.**
  - **Efix available and the fix was verified on bdev.**
- **PMR 77030 – “untar” performance. User experienced slowness when using /project.**
  - **IBM suggested to tune up the tcp send/receive buffer size but that did not seem to help. IN PROGRESS.**
- **PMR-54400 – File system threading deadlock**
- **PMR-69276 – Failover failed to occur when router died and was repaired**

Problems Encountered

- **Outages**
  - **The current NGF architecture has built-in redundancy to allow NGF to survive from any single hardware failure**
  - **Center-wide NGF outage may occur due to multiple failures or software bugs**
  - **Partial outage (multiple NGF outage within a NERSC system) may also occur due to network failures**
- **Problems**
  - **Server crashes (motherboard faults)**
  - **Disks and Controllers**
  - **Switches**
  - **Software bugs**

NERSC 5 Benchmarks

- Application Benchmarks
  - CAM3 - Climate model, NCAR
  - GAMESS - Computational chemistry, Iowa State, Ames Lab
  - GTC - Fusion, PPPL
  - MADbench - Astrophysics (CMB analysis), LBL
  - Milc - QCD, multi-site collaboration
  - Paratec - Materials science,developed LBL and UC Berkeley
  - PMEMD – Life Science, University of North Carolina-Chapel Hill
- Micro benchmarks test specific system features
  - Processor, Memory, Interconnect, I/O, Networking
- Composite Benchmarks
  - Sustained System Performance Test (SSP), Effective System Performance Test (ESP), Full Configuration Test, Throughput Test and Variability Tests