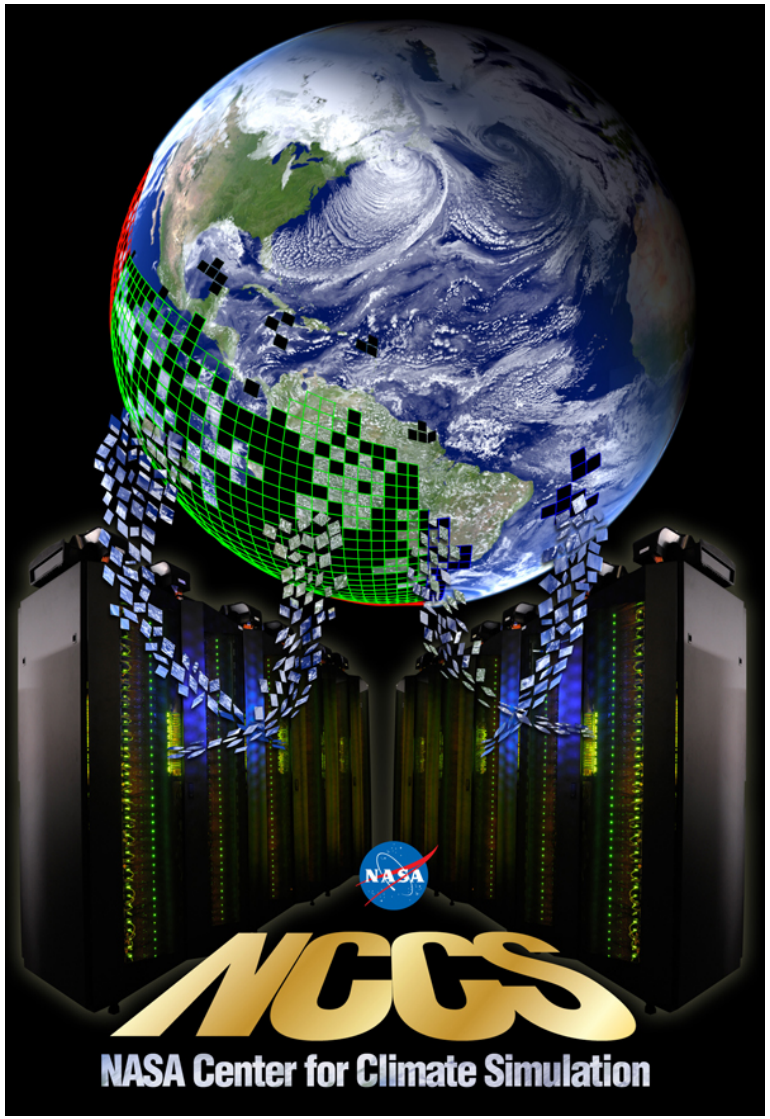
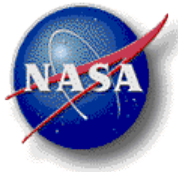


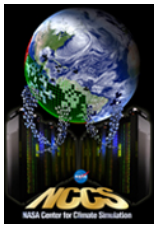
HSM at the NCCS

September 5, 2012

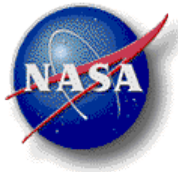
Nicko Acks (CSC)
Nicko.D.Acks@nasa.gov



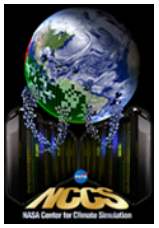
What is an HSM?



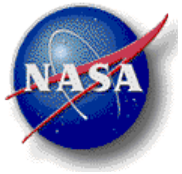
- HSM – Hierarchical Storage Management
- Policy based software that handles the automatic backup/archiving of data with little (or no) user interaction.
- Can make use of a variety of hardware (Tape, cheaper disk, MAID, etc)
- In many environments replaced a human operator that would manually load tape or disk media.
- Many different vendors have their own HSM solutions



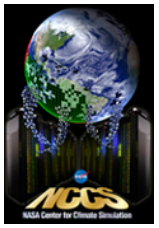
HSM Examples



- EMC DiskXtender for UNIX/Linux – Also formally known as UniTree
- IBM HPSS – the IBM High Performance Storage System is built on top of the GPFS (General Parallel File System) that is based on a collaboration with the US Department of Energy and many other universities and laboratories.
- IBM TSM – Tivoli Storage Manager – formally known as the ADSTAR Distributed Storage Manager
- Oracle/Sun SAM-QFS – based on Sun's Quick File System and StorageTek's Storage and Archive Management software.
- SGI DMF – SGI's Data Migration Facility – based on extensions to the XFS filesystem along with SGI's OpenVault software.

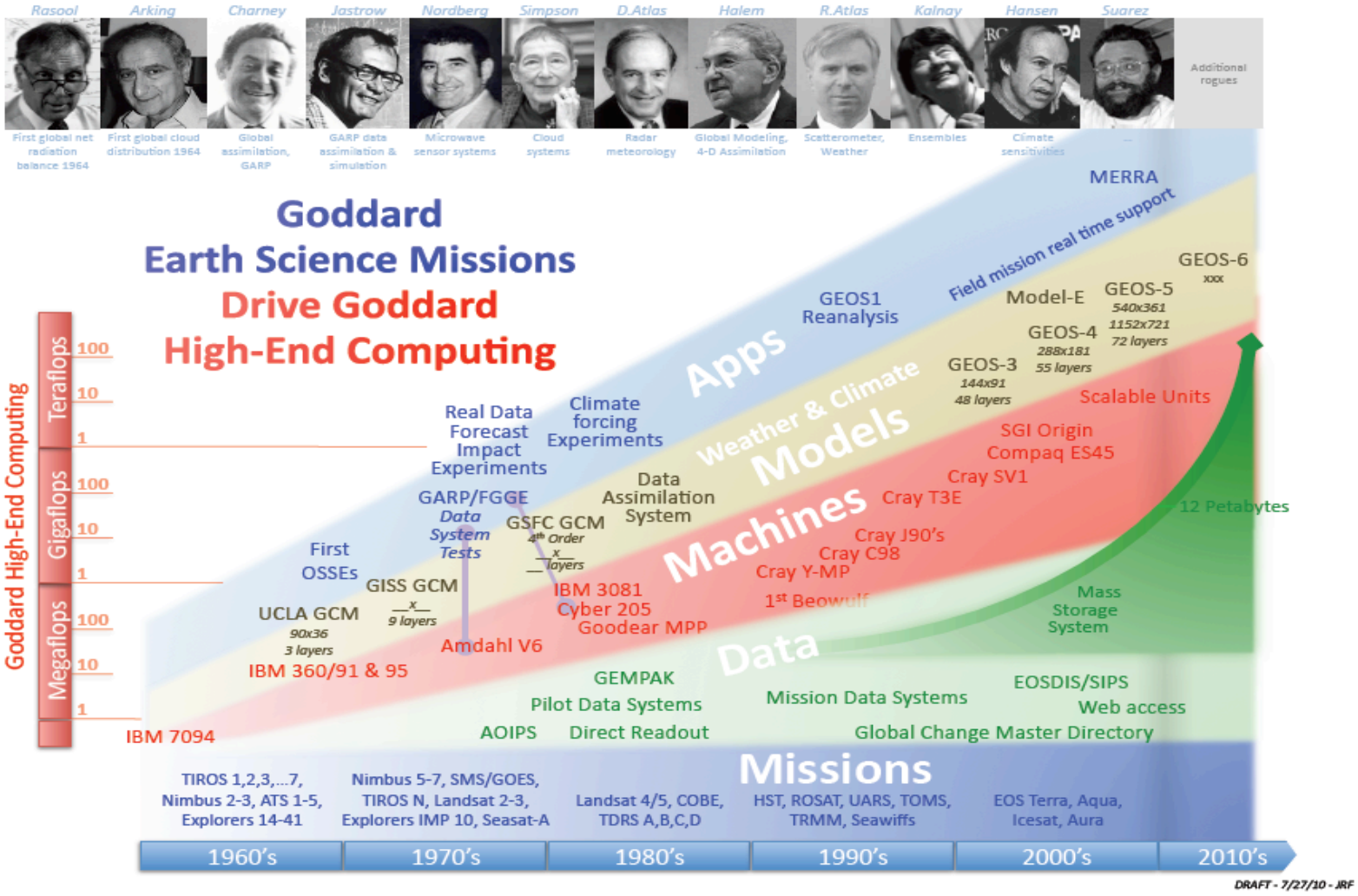


Why use an HSM?

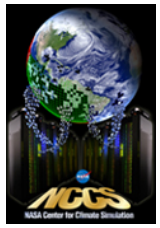


It is all about the Data

Computing History at Goddard



Climate Simulation Data Communities



NASA Scientific Community
- Simulation data consumers
- Advance scientific knowledge
- Direct access to systems
- *Supercomputer* capability required for effective analysis

NASA Modeling Community
- Model development, testing, validation, and execution
- Data creation
- Largest HPC usage
- Requires observational data as input

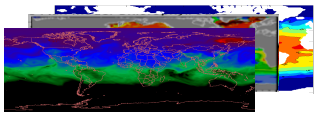
Simulation/Model Data
Observation Data

External Applications Community
- Huge opportunity for impact from climate change data
- Simulation data consumers
- Limited ES data expertise
- Web-based access to systems

External Scientific Community
- Simulation data consumers
- Advance scientific knowledge
- Web-based access to data

Each community has different capabilities and data usage requirements.

NCCS Data Centric Computing Environment



Data Sharing and Publication
- Capability to share data & results
- Supports community-based development
- Data distribution and publishing

Code Development*
- Code repository for collaboration
- Environment for code development and test
- Code porting and optimization support
- Web based tools

User Services*
- Help Desk
- Account/Allocation support
- Computational science support
- User teleconferences
- Training & tutorials

DATA Storage & Management
Global file system enables data access for full range of modeling and analysis activities

Analysis & Visualization*
- Interactive analysis environment
- Software tools for image display
- Easy access to data archive
- Specialized visualization support

Data Transfer*
- Internal high speed interconnects for HPC components
- High-bandwidth to NCCS for GSFC users
- Multi-gigabit network supports on-demand data transfers

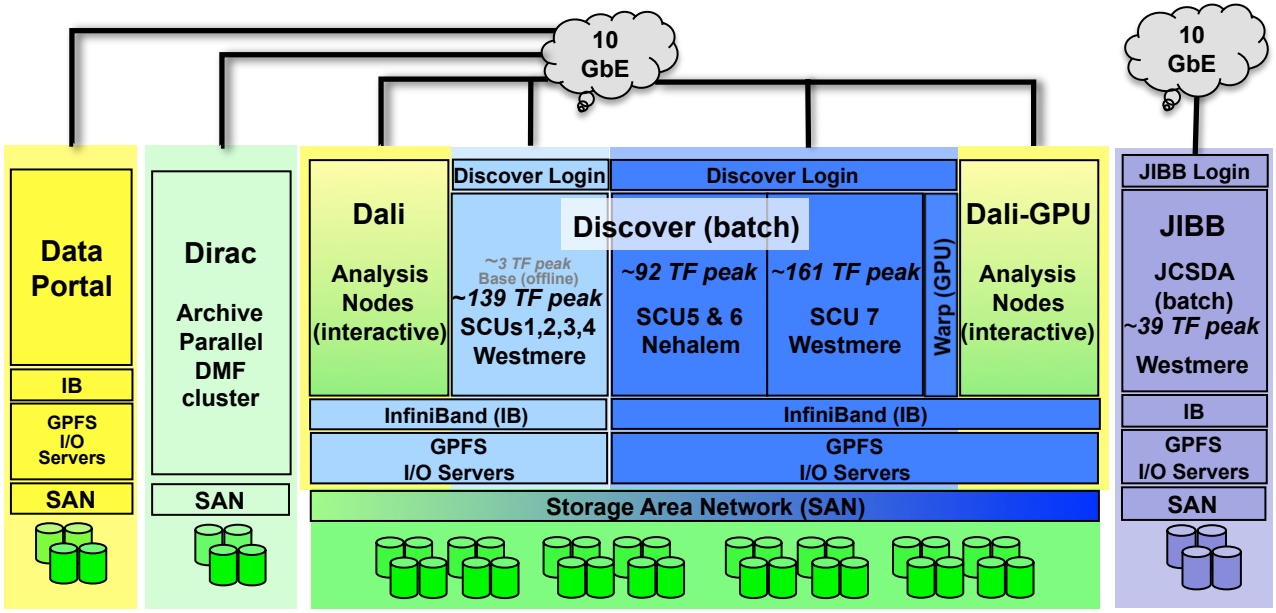
HPC Computing
- Large scale HPC computing
- Comprehensive toolsets for job scheduling and monitoring

Data Archival and Stewardship

*Joint effort with SIVO
*Joint effort with SEN

NCCS HEC Environment

- *Discover/Dali* – High performance computing cluster
- JCSDA (*JIBB*) – NASA/NOAA collaborative high performance computing
- Archive System (*Dirac*) – Mass Storage
- *Dataportal* – Data sharing

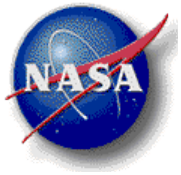


Discover/Dali
- 30,760 cores
- 28,672 GPU streaming cores
- 3,394 nodes
- 400 TFLOPS
- 3.7 PB usable disk storage

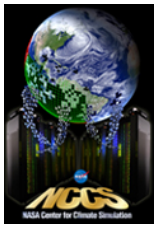
JCSDA/*JIBB*
- 3,456 cores
- 288 nodes
- 39 TFLOPS
- 320 TB usable disk storage

Archive System/Mass Storage/ Dirac
- 16 interrelated servers
- 30 PB archive holdings
- 960 TB usable disk storage

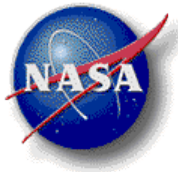
Dataportal
- 16 HP blade servers
- 200 TB usable disk storage



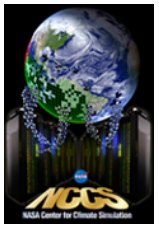
History of the Archive at the NCCS



- Unitree was installed July 1992 (Convex and then later moved to SUN hardware).
  - This is the center's primary HSM.
- DMF was originally implemented at the NCCS for use with the Cray T3E.
  - This was a separate HSM attached for use with the Cray.
- DMF was moved to a succession of SGI based hardware (Origin 2K, Origin 3K and then eventually Altix BX2)
  - This HSM was dedicated for the use of that specific compute platform at the time.
- Unitree was migrated to SAM/QFS in 2003.
- SAM/QFS and DMF were consolidated into a single HSM (DMF) in 2005.

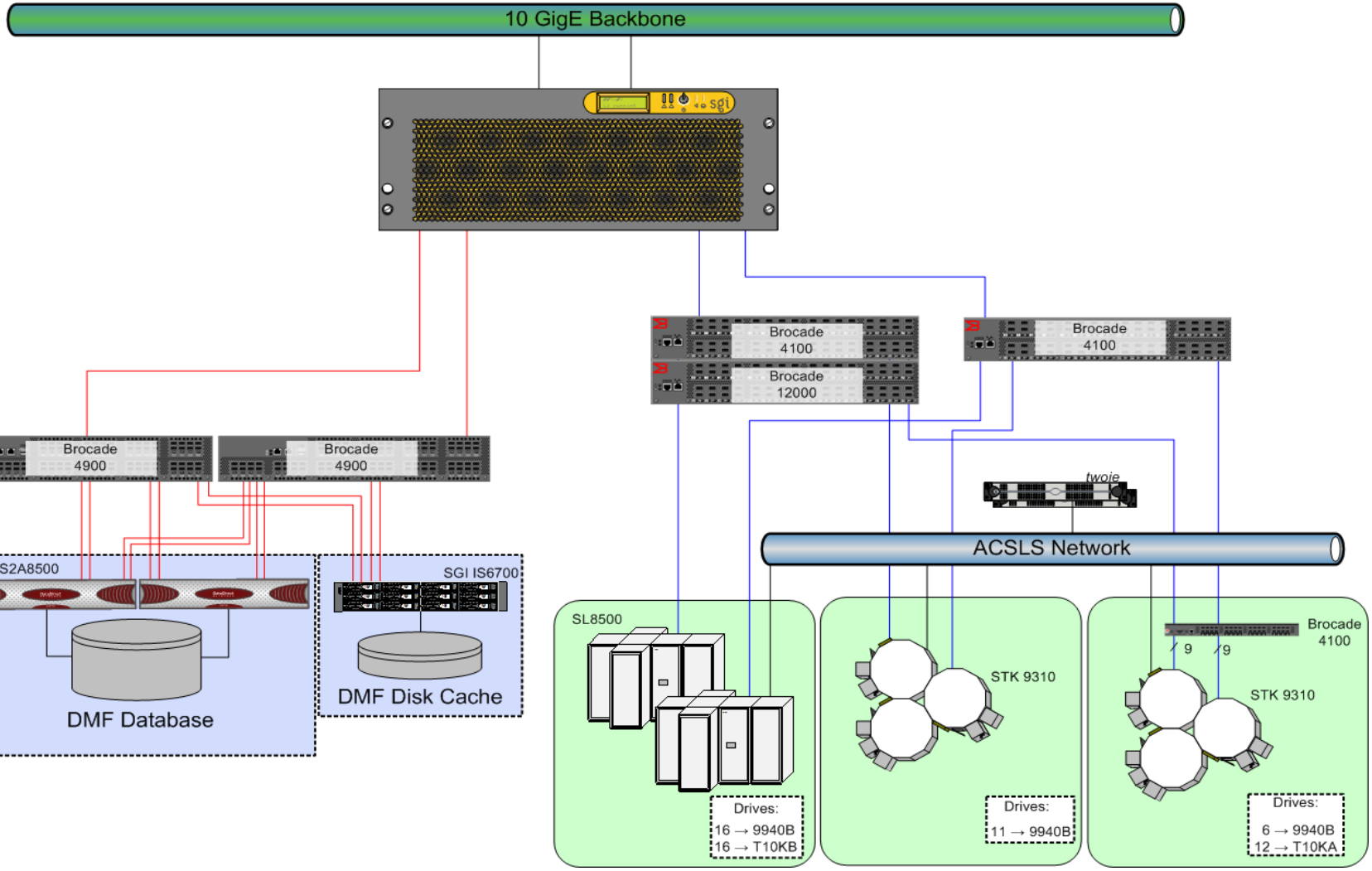


SGI DMF 101

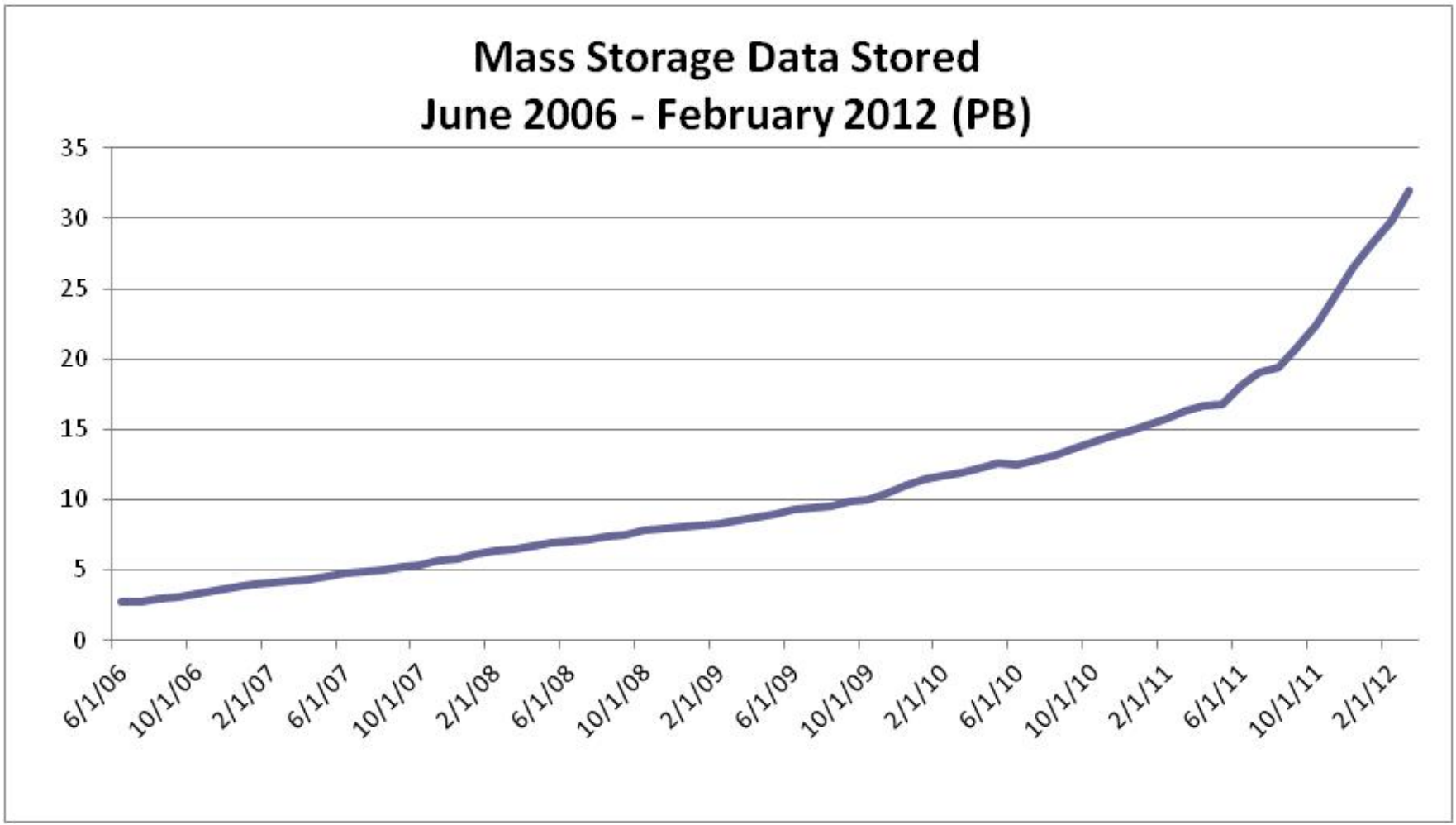


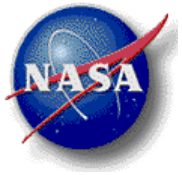
- DMF consists of 3 (or 4) components depending on implementation
  - DMF server software.
  - OpenVault server software (replaced TMF for interfacing with DMF).
  - XFS filesystem (in pDMF the XFS filesystem would be in a CXFS cluster, SGI's cluster filesystems extension to XFS/XVM).
  - DMF distributed client software (not needed, but useful).
- DMF server
  - Manages the files stored on XFS filesystems.
  - Extended attributes in XFS inodes are used to store pointers to database entries that the server manages.
  - Manages a database of all copies of the data stored in the XFS filesystems it manages.
- OpenVault Server
  - Manages access to removable media (MAID, Tape, removable disk, manages tape robot)
  - In a monolithic configuration the DMF server and OpenVault server and client software live together.
  - In pDMF, Parallel Data Movers (pdmo) communicate with the OpenVault server and move data between XFS filesystem and Tape.

Historical DMF at the NCCS

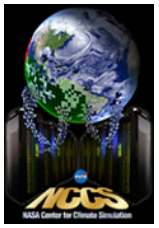


Exponential Data Growth



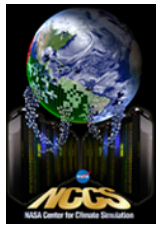


Why pDMF?

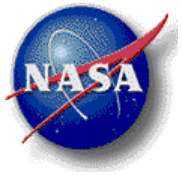


- Reliability
  - Ability to implement High Availability (HA)
  - Decentralized Services
    - Separate Data movement between storage tiers from the DMF server.
    - Separate NFS services from the DMF server.
    - Separate other services (ssh/sftp/scp/ftp) from the DMF server.
- Cost
  - The cost of maintenance on legacy monolithic Origin or Altix based servers was prohibitive.
  - Technology available using X86_64 based hardware was much more reasonable.
- Scalability
  - It is relatively cheap and easy to add more bandwidth by adding more low cost nodes.

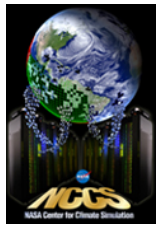
Initial pDMF Cluster Configuration



- Novell SuSE Linux Enterprise Server 11 (SLES11) and SuSE Enterprise Linux High Availability Extension (SLE11 HAE)
- SGI Foundation Software 2.2 and ISSP-2.2
- 2x DMF/CXFS Metadata servers
  - XE500 3u servers
  - 2x Intel Xeon X5670 processors (Hex Core)
  - 72GB of memory each
  - MYRICOM 10Gb ethernet
  - 2x 4-port Qlogic 8Gb FC cards
- 3x Parallel Data Mover (pdmo) nodes
  - Coyote 1u servers
  - 2x Intel Xeon X5570 processors (Quad Core)
  - 12GB of memory each
  - 2x 4-port Qlogic 8Gb FC cards

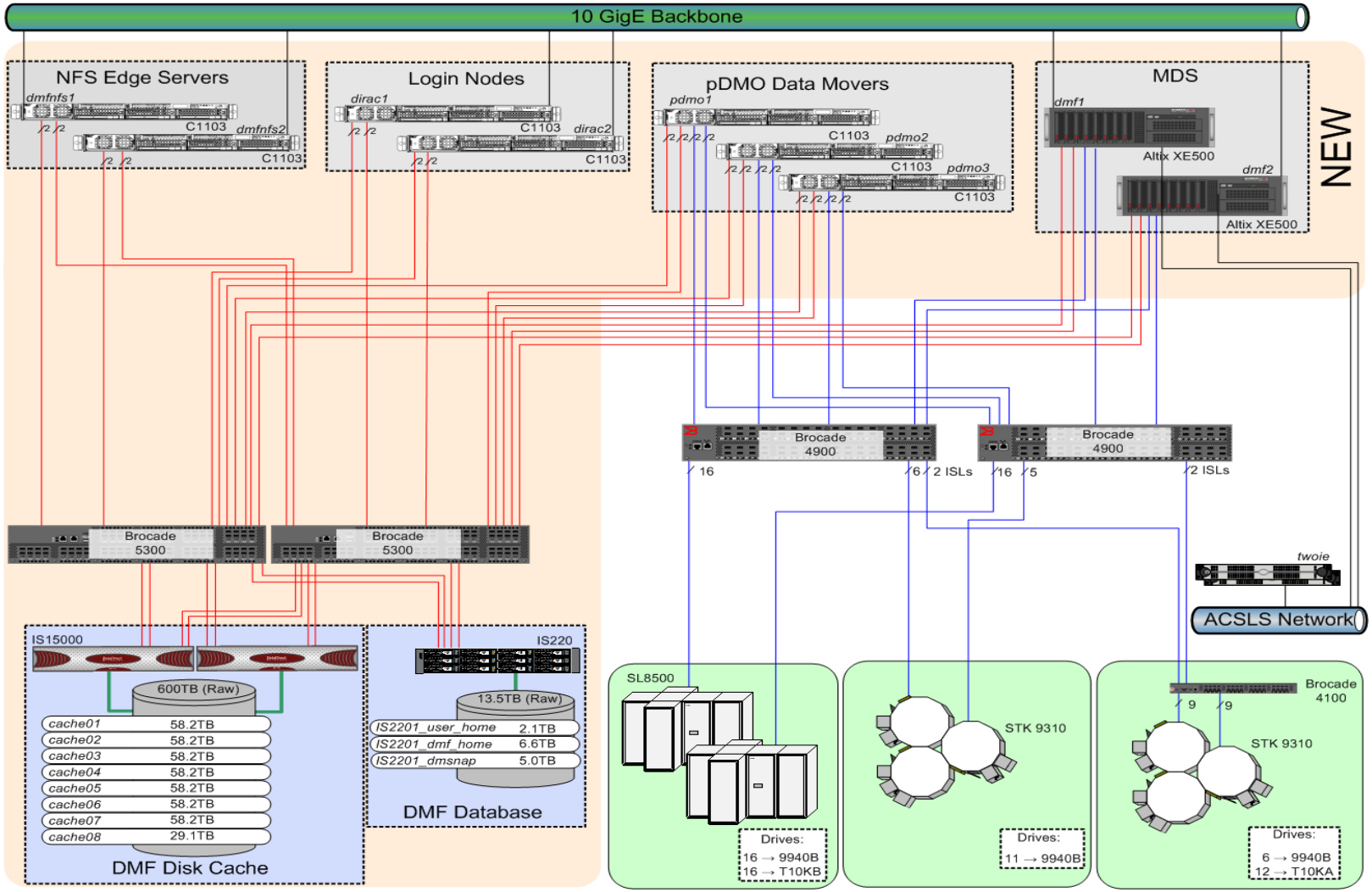


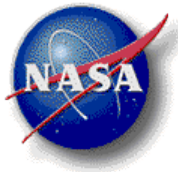
Initial pDMF Cluster Configuration Cont.



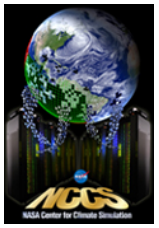
- 2x Login nodes
  - Coyote 1u servers
  - 2x Intel Xeon X5570 processors (Quad Core)
  - 48GB of memory each
  - MYRICOM 10Gb ethernet card
  - Qlogics 4-port 8Gb FC card
- 2x Edge NFS nodes
  - Coyote 1u servers
  - 2x Intel Xeon X5570 processors (Quad Core)
  - 96GB of memory each
  - MYRICOM 10Gb ethernet card
  - Qlogics 4-port 8Gb FC card

Initial pDMF Installation 2010

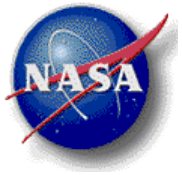




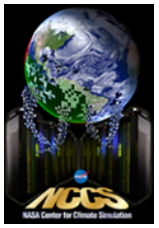
Initial problems



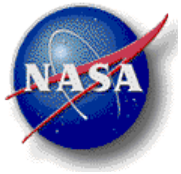
- Problems with HA implementation
  - HA was initially implemented only on DMF/CXFS Metadata server nodes.
  - There was a bit of a learning curve for staff that had never used SLE11 HAE.
  - Getting timeouts correct and dealing with underlying DMF/OpenVault/CXFS problems was complicated by introducing HA.
- Problems with OpenVault
  - This was the first implementation of OpenVault at our site, previously we used TMF exclusively (again a bit of a learning curve).
  - Partition Bit Format issue (Also seen at SARA)
  - We saw issues with long startup times (upwards of 20+ minutes)
- Problems with CXFS
  - Many issues with token deadlocks
- Problems with NFS
  - Edge NFS causes token deadlocks that make using NFS difficult
  - Due to issues with Edge NFS we fell back to using NFS served from the active DMF server. We see NFS read corruption in this configuration.



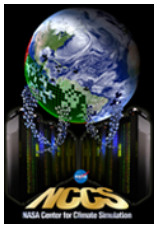
Moving forward



- HA training
  - SGI's class based on SLE11 HAE is highly recommended if you have no experience with Linux HA.
- Multiple ISSP releases and patches to deal with CXFS and OpenVault/DMF issues
  - We started on ISSP 2.2 and SLES11ga
  - Currently running SLES11SP1 and ISSP 2.4. (and many patches in between)
- NFS on DMF servers due to CXFS deadlock issues.

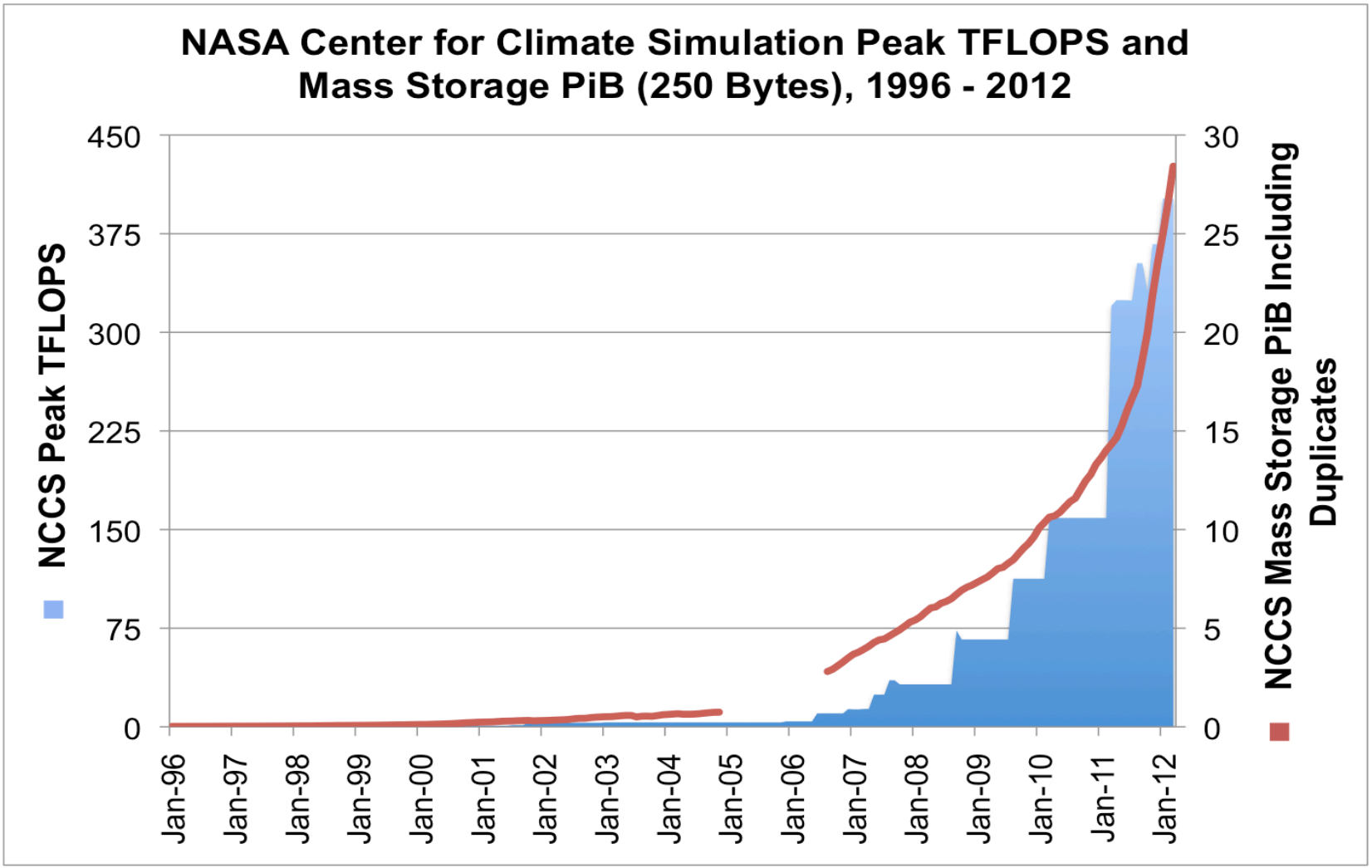


Success!

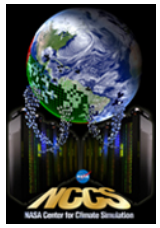


- Late 2011 fixes for CXFS Token problems allow for movement back to using Edge NFS servers.
- Increased stability with HA.
  - This took some time and there were other hurdles introduced by SLES/ISSP upgrades but eventually most of these have been ironed out.
- Victims of our own success.
  - Increased ingest volume (see chart)
  - Hey look! Unlimited disk space! (user training?)
  - SUN STK T10KC rollout in 2011

Success!

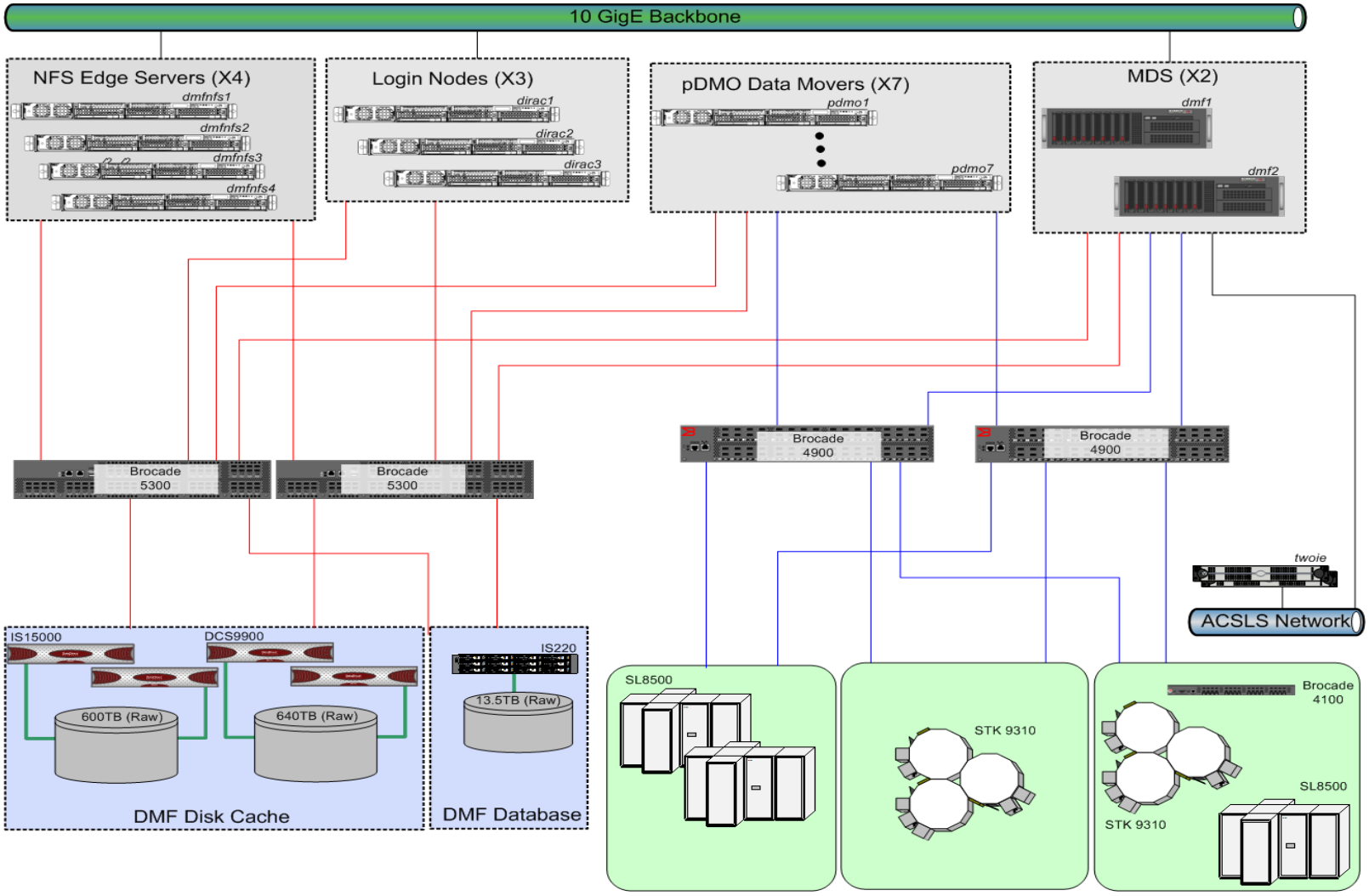

NASA Center for Climate Simulation Peak TFLOPS and Mass Storage PiB (250 Bytes), 1996 - 2012
NCCS Peak TFLOPS
NCCS Mass Storage PiB Including Duplicates
450
375
300
225
150
75
0
30
25
20
15
10
5
0
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Jan-10
Jan-11
Jan-12

pDMF Augmentation 2011

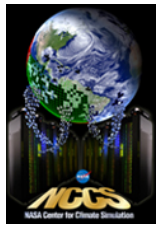


- Sun STK T10KC drives need more bandwidth (feed the beast)
  - 24 additional T10KC drives requires more bandwidth to disk than available with the 3 existing pdmo nodes.
  - Adding 4 more pdmo nodes increased bandwidth to disk and spread the load of driving more tape drives.
- Network bandwith problems.
  - Individual nodes on the compute system (NFS and other transfer methods like bbFTP) was able to saturate 10Gb network to login and NFS server nodes.
  - Adding 2 more edge NFS server nodes and another login node gave a total of 70Gb of available network bandwidth.
- Implement HA for Edge NFS servers
  - Up to this point HA had not been implemented for edge NFS servers.
  - The two additional edge NFS servers allowed for implementation of HA with minimal impact to regular operations.

pDMF Augmentation 2011 Cont.

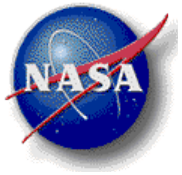


pDMF 2012 and beyond

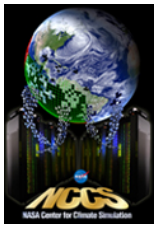


- Increased Compute == Increased Data Growth
- More augmentation to hardware?
- Changes in user behavior?
- Future directions
  - Hybrid GPFS/CXFS nodes?
  - More automated movement of data between DMF and Compute?
  - Integration of alternative data services (iRODS or other cloud based services)?
  - Test environment?

- Upgraded to ISSP-2.5 in June 2012
- Due to replace older DDN S2A9550 disk controllers with SGI IS220 (LSI/NetApp)



Thanks!



Special thanks to the following who helped make this presentation possible:

NASA NCCS Team:

Fred Reitz (CSC)
Ellen Salmon
Tom Schardt
Adina Tarshish
Hoot Thompson (PTP)

SGI Implementation Team:

Jonathan Arbogast
Paul Garn
Zsolt Ferencsy
Ron Kerry