Major HPC Archive Systems

Alan K Powers
HPC CTO
apowers5@csc.com
www.csc.com/hpc

10/23/13

Agenda

NASA visualization system, hyperwall2, created to discover insight into Big Data (Volume, Velocity) for parameter studies, time sequence events and real-time streams from HPC simulations.

A GLOBAL POWERHOUSE IN BUSINESS AND IT TRANSFORMATION

Leading Next-Gen Technology and Business Solutions

WHAT WE DO

Industries

Energy and Natural Resources | Financial Services | Healthcare | Manufacturing | **Public Sector** | Communications and High-Tech | Travel and Transportation | Consumer and Retail

INDUSTRY EXPERTISE

Solutions

Cloud | **Cybersecurity** | **Big Data & Analytics**

Applications Services | Business Process Services and Outsourcing | Consulting | Infrastructure Services | Software and IP

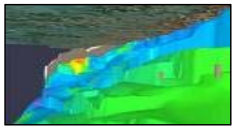
Serving Enterprises and Governments Worldwide*

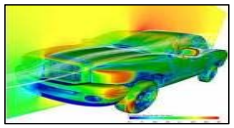
WHO WE WORK WITH

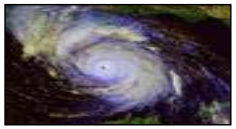
*Representative client list

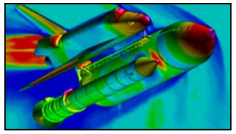
HPC Center of Excellence


Visualization System


Energy & Oil Exploration


Modeling & Simulation


Climate Modeling & Weather Prediction


Space Sciences


Consumer Products Manufacturing


Financial Services

Our Mission

Provide information technology solutions and services that enable clients to achieve their strategic goals within today's and tomorrow's High Performance Computing environments.

Our Objective

Become the preferred, most in-demand customer-focused HPC service provider while maintaining focus on the leading edge of HPC technologies.

HPC CoE Provides Full Range of HPC Capabilities

From Architected Solutions to Outsourcing Services

- Deploying HPC systems for over 20 years
- Managing over $130 million HPC equipment
- Premier HPC clients including NASA, NOAA, P&G
- Industry leader in HPC service delivery –installation, integration, testing and operation of HPC systems.

Applications

Infrastructure Services

Security

Networking

End Users Support

Facilities

Computing

Storage & Data Management

HPC Service Delivery Process

CSC operates HPC systems with aggregate capability across all sites of **5 petaflops**, and manages **42 petabytes** RAID and **180 petabytes** of archival data

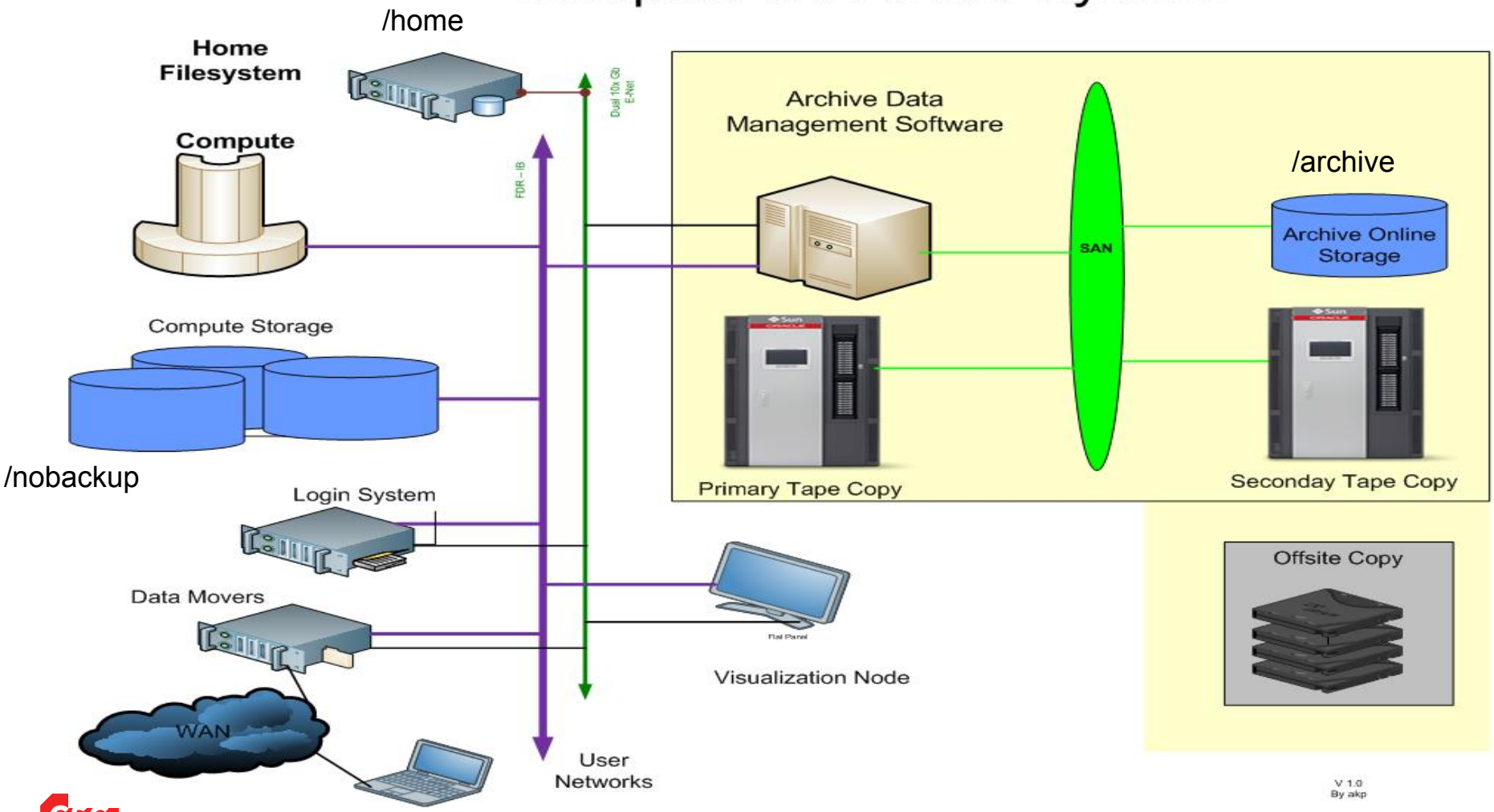
Reasons to Use an Archive Solution

- Daily data access < ~1% (i.e. 99% not accessed)
  - (read+write)/total_tape_data < ~1%
- Future data importance hard to determine
- Linear or exponential data growth
- Data too costly or time consuming to recreate (Experimental or Simulation)
- Only copy of the data (Satellite)
- Data requirement preserved for years (Sport, Movies, Medical)
- Time consuming to integrate new storage
- Time consuming to retrieve file(s) from backups
- Expensive to backup, beyond backup window
  - 4 TB @ 99 MB/s is greater than 11 hours (backup tools slow IO)

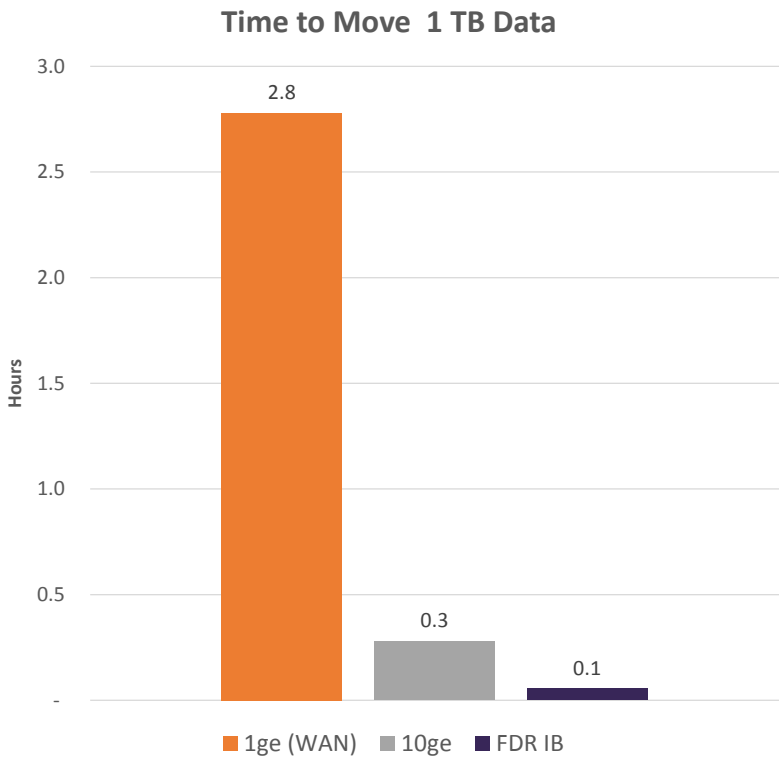
Justifications to Use an Archive Solution?

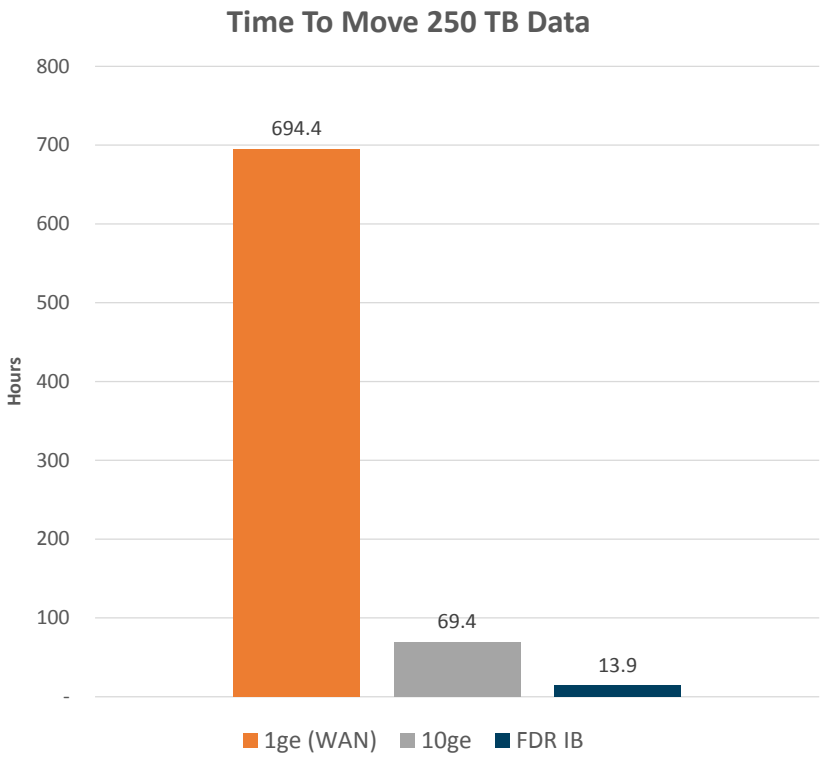
- Reduce TCO – over 5 years 3-6x @ scale: 1+ PB
  - High End RAID : $700/TB @ > 2 GB/s – Moderate Risk
  - NAS: $400/TB @ < 1 GB/s – Moderate Risk
  - DIY: $60/TB @ < 0.2 GB/s – High Risk (HW only)
  - Archive : $100-$200/TB @ > 1 GB/s – Low Risk
  - Cloud : $120-$780/TB/Year < local pipe – Risk?, Security?, SLA?
- Tapes are better at preserving data
  - Better bit error rate on tape drives – 10-1000x better than SAS disks
  - Data not overwritten – append to end of tape
  - Data immediately verified after written to tape by tape drive
- Reduce Backups – monthly full, daily incremental
  - 1 PB Data: Archive 2.3 PB media compared to backups 12.2 PB media
- Reduce Operational Cost
  - Electrical and cooling not needed for tape media
- Industry Paper – disk solution vs archive/tape
  - http://www.clipper.com/research/TCG2010054.pdf (15x savings – TCO 12 years)

Compute & Archive System



Goal: Do the Analysis Where the Data is





Need new tools/methods to copy/verify large amounts of data

NASA Ames/ NAS HPC System

Endeavour
SGI UV2000 – 1536c

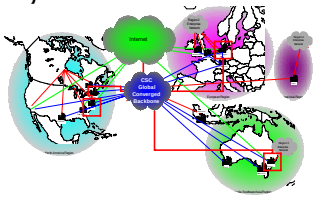
Lustre, QDR IB
DDN, NetApp RAID

NFS IB & Ethernet
Home Filesystems

3 PF, 20 PB RAID, 130 GB/s

1,10 Gb LAN & WAN

Login, Data Transfer,
Batch nodes (PBSpro)



27,800 sq ft Datacenter
6+ MW Power, RUPS
2000+ Tons of Cooling

Pleiades
SGI Altix ICE+
162,496c
417 TB memory
166 Racks
#14 in Top500
as Nov. 2012

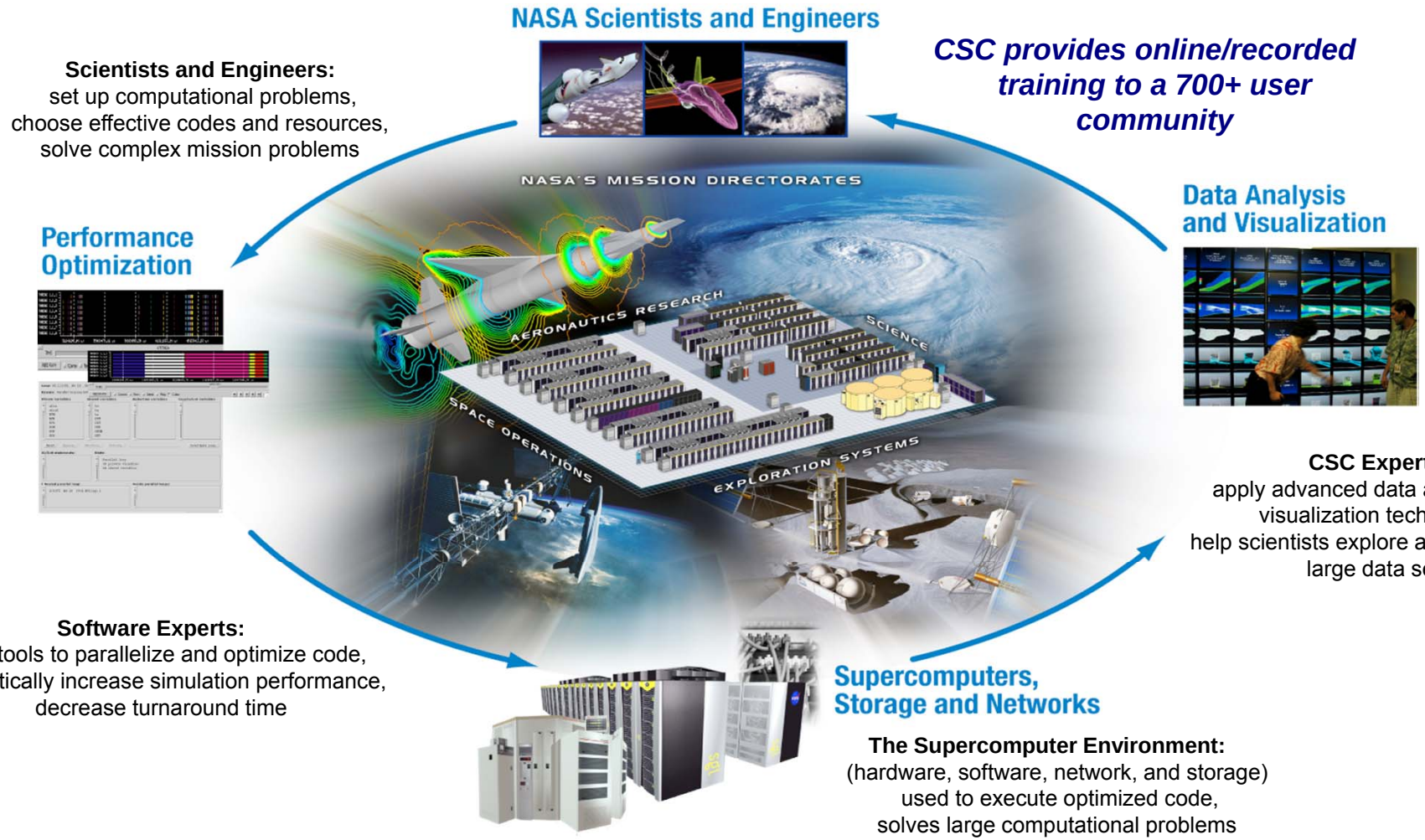
60 PB pDMF
Managed

115 PB Archive
Capacity
(Spectra Logic)

Hyperwall-2 1024c
8x16 20” Displays

All HPC systems have high-speed access to a site-wide filesystem

Integrated Scientific Application Support for Full Life-Cycle High Performance Modeling & Simulation



Scientists and Engineers:
set up computational problems,
choose effective codes and resources,
solve complex mission problems

CSC provides online/recorded training to a 700+ user community

Software Experts:
use tools to parallelize and optimize code,
dramatically increase simulation performance,
decrease turnaround time

CSC Experts:
apply advanced data analysis and visualization techniques,
help scientists explore and understand large data sets

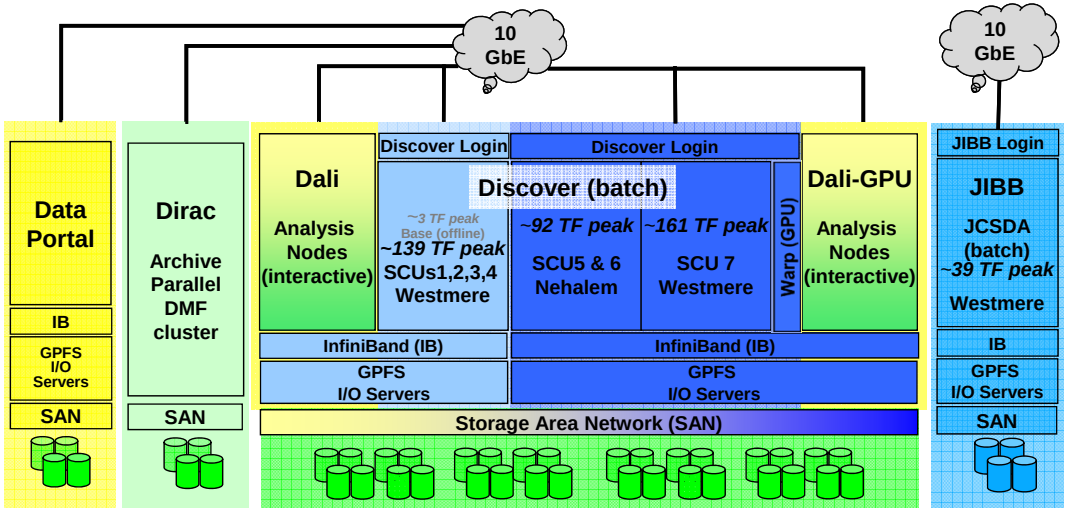
The Supercomputer Environment:
(hardware, software, network, and storage)
used to execute optimized code,
solves large computational problems

NAS Tape Silo & Library Configuration – 62 PB Dual Copy (10/1/2013)

- 60,000 Slots Spectra Logic T950E
  - 30,000 Slots in each building
- 92 LTO5 IBM Tape Drives
- 30-125 TB/day new data, 2 PB growth last month
- Archive Policy (SGI pDMF software)
  - <262 KB : No Archive
    < 100 MB, 10 GB Zone: Small Files
  - For Medium Files:
    100 MB- 100 GB, 20 GB Zone, Max Chunk Size 50 GB
  - For Large Files:
    > 100 GB, 400 GB Zone, Max Chunk Size 100 GB

NASA Goddard NCCS HPC Environment

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



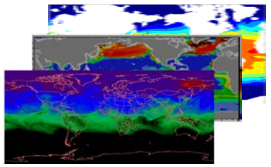
Discover/Dali
- 43,240 Xeon cores
- 28,672 Nvidia GPU cores
- 28,800 Intel Phi (564.3)
- 3,394 nodes
- 619 TFLOPS Xeon
- 3.7 PB usable disk storage

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

Archive System/Mass Storage/Dirac
- 16 node cluster
- 34 PB archive holdings
- 960 TB usable disk storage

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

NASA Center for Climate Simulation (NCCS)
Data Centric Computing Environment



Data Sharing and Publication
- Capability to share data & results
- Support 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

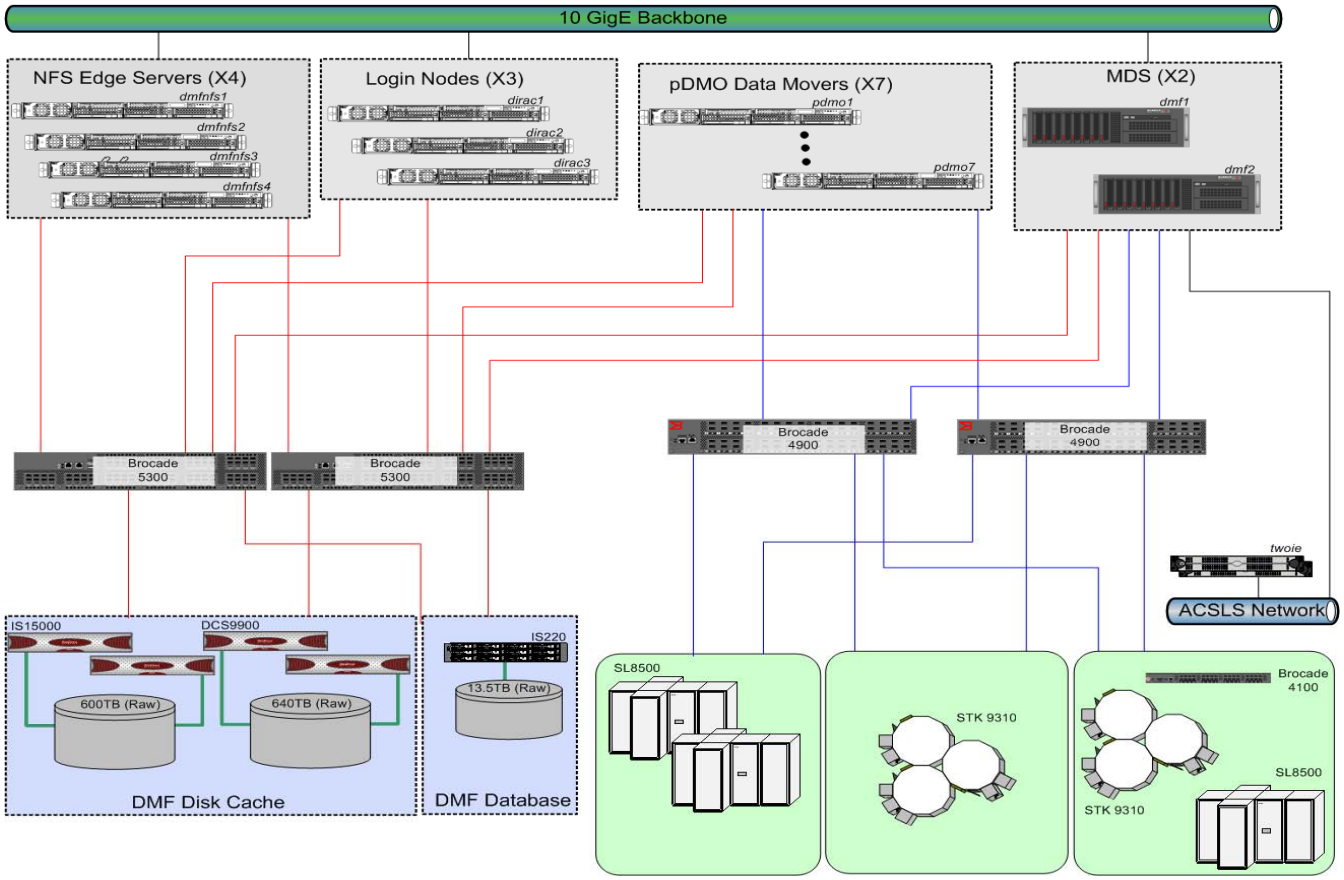
HPC Computing
- Large scale HPC computing
- Comprehensive toolsets for job scheduling and monitoring

Data Archival and Stewardship

**Joint effort with SIVO*
**Joint effort with SEN*

NASA/NCCS SGI pDMF

(34 PB - 24 T10KC, 28 T10KB, 27 9940B)



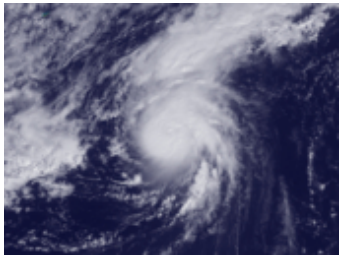
National Oceanic and Atmospheric Administration

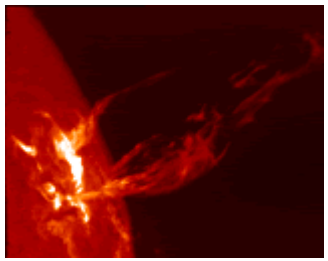
- Scientific agency within US Dept. of Commerce
- Formed in 1970 from three existing agencies
  - US Coast and Geodetic Survey (1807)
  - Weather Bureau (1870)
  - Bureau of Commercial Fisheries (1871)
- 2013 $5.1 billion budget
- **Strategic Vision**
  - Inform society with a comprehensive understanding of role the oceans, coasts, and atmosphere play in the global ecosystem in order to make the best social and economic decisions.
- **Mission**
  - Understand and predict changes in the earth's environment and conserve and manage coastal and marine resources to meet our nation's economic, social, and environmental needs.

Weather Forecast: weather.gov

NOAA has offices in every state

- ~12,000 Employees
- 8 Ships (155'-274' research vessels)
  http://www.noaa.gov/deepwaterhorizon/platforms/ships_index.html
- 5 Planes "Hurricane Hunter"
  http://www.aoc.noaa.gov/aircraft.htm
- Satellites: severe weather, snow, tropical storms, hurricanes, sea level, …
- Saved ~ 33,000 people since 1982 worldwide via emergency beacons (people, ships, planes)
  http://www.noaanews.noaa.gov/stories2012/20121009_sarsataniv.html





Overview of the NOAA HPC Environment

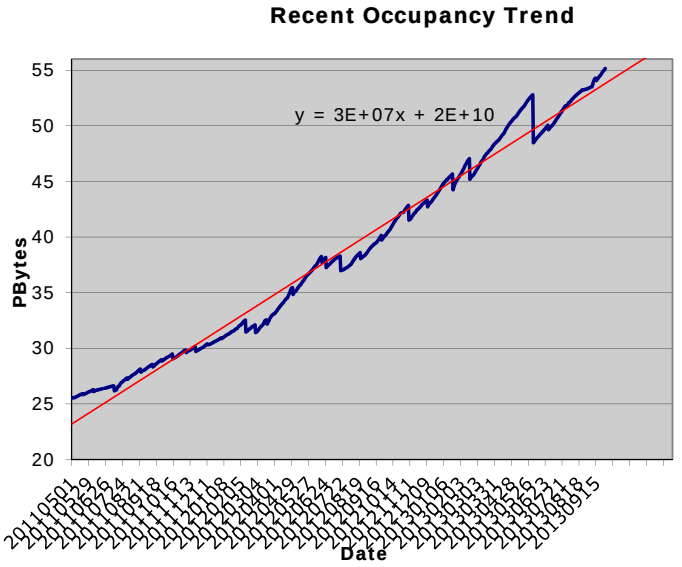
- Distributed sites (GFDL, NCEP, ESRL, NESCC) – one set of NOAA's HPC agency requirements
- Built new facility from the shell/slab – Fairmont, WV (NESCC)
- Collaborate with different Federal agency (DOE/ORNL)
- Distributed CSC staff and customers across time zones
- Implement a Large Scale System with minimal onsite staff
- Coordinate with other NOAA contracts

NOAA Tape Silo & Library Configuration

- GFDL (Princeton, NJ)
  - 55 PB Single Copy (10/1/2013)
  - SGI pDMF – 34 nodes in CXFS cluster
  - 5 PB cache @ 60 GB/s
  - 40,000 Slots Oracle StorageTek SL8500
  - 54 T10000C, 38 T10000B Tape Drives
  - 50-300 TB/day New Data, 2 PB growth last month
  - Currently evaluating T10000D tape drives
- NESCC (Fairmont, WV)
  - 33 PB Single Copy, Growing at 1+ PB/month (10/1/2013),
  - IBM HPSS (2 MDS, 4 data movers, 4 VFS, 1 Spare)
  - 900 TB disk cache @ 10 GB/s
  - 10,000 Slots Oracle StorageTek SL8500
  - 32 T10000C Tape Drives

GFDL Metrics: Some of the Most Demanding in the Industry

- 200 million active files, 5 PB cache, 55 PB tape data
- 100-300 TB/day data in/out tape drives
- 100K-1M files per day in/out tape drives
- 500-1200 TB/day of data in/out archive filesystems
- 200-700 TB/day over the 10ge network ports
- ~440 million metadata operations per day to NFS nodes
- ~100 nodes accessing archive filesystems
- ~4000 batch jobs per day used to process data from archive FS
- 14 Analysis nodes directly attached to archive filesystems



2013 Highlight Test Results

- Tape Bandwidth – Goal: 12 GB/s (r/w)
  - 54 T10KC and 36 T10KB tape drives
  - **Results:** 13.17 GB/s, peaks of 14.7 GB/s
- Network Bandwidth – Goal: 18 GB/s (9 bidirectional)
  - 30 PP nodes and 8 NFS servers with 24 10ge ports, gridftp - 4 10 GB files per PP node
  - **Results:** peaks of 21 GB/s
- Tape Load – Goal: 60m, 5 TB read,11 TB write 1mb-100 GB, 54,000 files
  - 42 T10KC and 12 T10KC (54) tape drives in 4 SL8500 Tape Library
  - **Results:** 5 TB in 55m, 11 TB in 45m
- Tape Mounts (ACSLS)
  - 54 T10KC and 36 T10KB (90) tape drives in 4 SL8500 Tape Library
  - **Results:** 3140-3617 mounts/hour.
- IO Bandwidth – Goal: 15 GB/s (all RAID 60 GB/s)
  - 2 RAID Units - disks rebuilding during production workload
  - **Results:** 15 GB/s, peaks 18.8 GB/s

2012 Archive Workload Test (20 m)

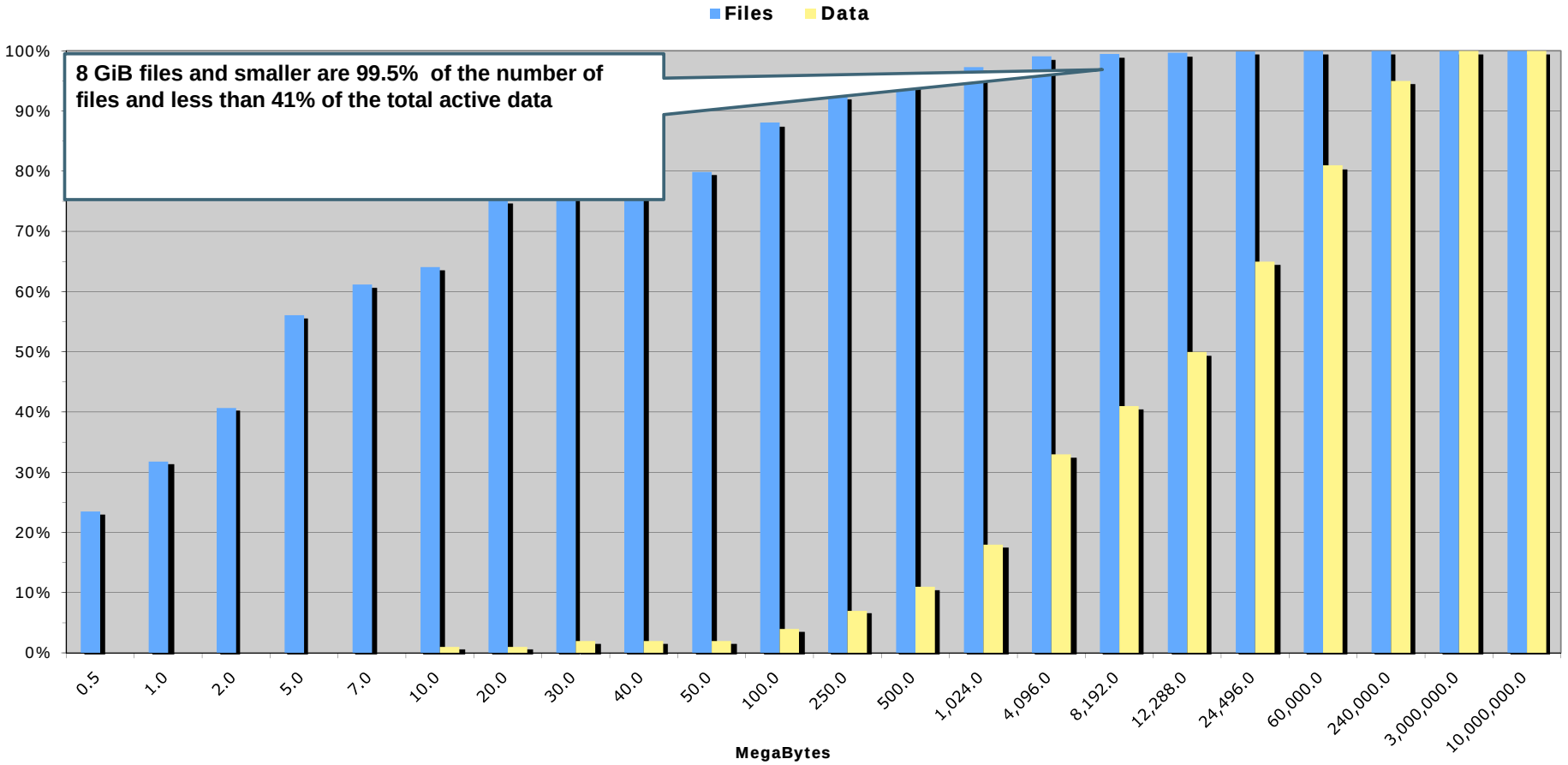
- 52 T10KB & 32 T10KC tape drives, 24 filesystems (57-171 TB), 4 DM, 9 NFS servers (triple bonded 10ge), 36 post-processing (PP) nodes – 10ge
- 504 – 10 GB large files, 1296 – 0.9 GB small files
- Read 252 large files each from single tape and transfer to PP node
- Transfer 252 large files from PP nodes to DMF cache and write to 24 tapes on 24 tape drives
- 648 small files PP <-> Disk Cache Manager

What Features are Important for an Archive System?

- Important Features
  - Data Integrity, Data Integrity, Data Integrity
  - Scale-Out - Systems, Storage, Tape Drives, Software, Silos
- How the system is architected is as important as Scale-Out
  - How many tape drives, network interfaces, front-ends, data movers, tape libraries, etc?
  - How big should the data cache be? How many tape pools? How many copies of the data should be created?
- Obvious feature is to minimize tape mounts
  - Keep the “right” files in cache
  - Optimize Tape Mounts. i.e. pull all files off a tape you need in one tape mount

Most important throughput feature – Tape Mount performance

NOAA/GFDL: 194 Million Files, 44 PiB Active Unique Data (20130813)



Mounts Per hour has Huge Impact on Retrieving Nearly All (99%) Files

- Silo and Tape Drives
  - Mounts Per Hour
    - At GFDL, files <8 GB 99.5% of total files only 41% of the total data (average file size 225 MB – 8/30/2013)
    - 7.5 GB file – roughly 30 seconds to read using latest tape drive (T10KC, etc)
    - Need over 70 GB file to achieve overall bandwidth of 200 MB/s if mount/dismount total 60 seconds
  - Effective Tape Drive Bandwidth =
    - MBytes / total time -> (mount + seek + dismount) + IO

      7,500 / (180 + 31.25) = 35.5 MB/s ; 225 / (180 + 0.94) = 1.24 MB/s

      7,500 / (60 + 31.25) = 82.19 MB/s ; 225 / (60 + 0.94) = 3.69 MB/s

      7,500 / (20 + 31.25) = 146.44 MB/s ; 225 / (20 + 0.94) = 10.74 MB/s

      70,000 / (60 + 291) = 199.05 MB/s

Areas Where Archive Software can Improve

- Current
  - Reduce Recall Storms
    - Prioritize recalls (Part of DMF user wish list – now in latest release of pDMF)
    - Share recall request across all copies of media (round-robin) with tier priority
    - Develop a process to look for user data spread across 1000s of tapes, then repack them to few tapes.
    - Fair share of tape drive usage between users (user a - 100K files, and user b – 100 file)
  - Manage all the nodes like a cluster (single source to quickly make changes, install, test OS upgrades)
  - Capacity planning & bottleneck isolation: Create graphical views to aggregate all the resources (r/w bandwidth rates for Filesystems, tape and network)
- Future
  - Federate archive repositories (be resilient across whole site outages)
  - Help manage all the data at a site, instead of just the archive system
  - Integrate Archive Software to move data to/from Cloud storage & LTFS

Main Points to Discuss with a Friend or Peer

- CSC 20 years in HPC, Customers – NASA, NOAA, PG
- Compute – 5 PF, Archival Data - 180 PB
- Archive systems are cost effective (3-6x over 4 years) and better in managing most data repositories compared to total RAID solution
- Analyze the data in place (reduces copies, save time)
- Tape library mount rates have major effect on overall IO throughput
  - Migrating all data to next generation tape drive
  - Retrieving 1000's small files from different tapes
  - Writing lots of new data to tape in a day

Questions

Future Questions

Contact
Alan Powers
HPC CTO
akpowers@csc.com
408-800-7340

Effective Mounts per Hour (EMnt/Hr)

- Exchanges per hour different from EMnt/Hr
- EMnt/Hr = 3600s / (Total_Time / Total_Mounts)
- For all tape drives in a tape library, simultaneously:
  1. mount, load, dismount a random tape
  2. repeat again with different tape
  3. track the total time to complete all steps
- Ex. 1 tape library, 32 tape drives: Select 64 tapes randomly from silo. Assume 130s to complete above task. (Repeat multiple times.)
- EMnt/Hr = 3600/(130/64) = 1772