SC11 Seattle, WA
Connecting communities through HPC

HPC Archive Solutions Made Simple

Alan Powers, CSC/NOAA

Jason Hick, NERSC

Matt Cary, CSC/NASA

Agenda

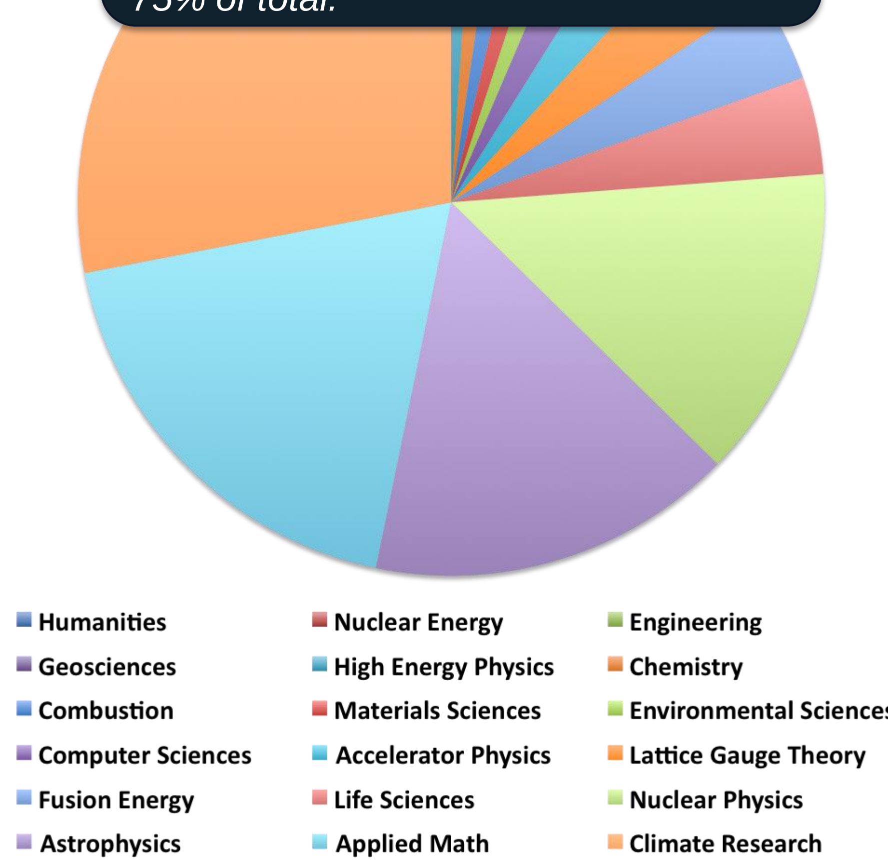
- Facilities — Afternoon Break – 3 to 3:30pm
- Format — Hold questions till the end of a section topic
- Introductions — Jason, Matt, Alan
- Attendees — Show of Hands (SoH) No Archive experience, <5, or 5+
- Attendees — SoH – Archive_Plans, HPSS, SamQFS, DMF
- Attendees — Raise Hands: Provide Name, Company,
  Problem looking to solve? OR Interest in the class?

- Site Background and Archive Configuration
  - Jason Hick – NESRC, jhick@lbl.gov
  - Matt Cary – CSC/NASA, matt.cary@nasa.gov
  - Alan Powers – CSC/NOAA, akpowers@pacbell.net

The Production Facility for the DOE Office of Science

- **Operated by the University of California for the U.S. DOE**
- **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 border soon to be 100Gb
- **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.



NERSC Archive System Statistics

- **HPSS User System** (~12 PB as of 9/30/2011)
  - Single transfers 1GB/sec read/write
  - Aggregate bandwidth 4+GB/sec
  - Average daily IO of 20TB, with peak at 40TB
  - 200TB disk cache
  - 24 9840D, 48 T10KB, 16 T10KC tape drives
  - Largest file: 5.5TB
  - Oldest file: Jan 1976
- **HPSS Backup System** (~13 PB as of 9/30/2011)
  - Single transfers 1GB/sec read/write
  - Aggregate bandwidth 3+GB/sec
  - Average daily IO of 10TB, with peak at 130TB
  - 40TB disk cache
  - 8 9840D and 18 T10KB tape drives
  - Largest file: 3.5TB
  - Oldest file: May 1995

NASA Advanced Supercomputing (NAS)

- History
  - The NAS division was founded in 1984, with the intent to provide high performance computing capabilities to all NASA centers and their collaborators to enable USA aeronautic advancements
  - Today some of the scientific research disciplines are world-wide climate modeling, nanotechnology, astrobiology, and CFD
  - The division's staff includes about 70 civil service employees and about 80 contractor staff
- Mission
  - Develop, demonstrate, and deliver innovative, distributed heterogeneous computing capabilities to enable NASA projects and missions
  - To lead the country in the research, development, and delivery of revolutionary, high-end computing services and technologies, such as applications and algorithms, tools, system software, and hardware to facilitate NASA mission success

Pleiades System Configuration

- SGI Altix ICE+ Blade (Water Cooled)
  - 21 KW/Rack (2100 KW Total)
  - 7 GPM/rack (700 GPM Total)
  - Harpertown, Nehalem, Westmere
  - 112,896 cores, 11,776 nodes
  - 191 TB total memory
  - 1.34 PF peak
  - 9.3 PB (DDN IB RAID/Lustre)
  - Two separate IB fabrics
  - DDR (24p) -> QDR (32p) IB
  - 70+ Miles of IB cables
  - 14 Login nodes (ge, 10ge, IB)
  - 64 nodes with NVIDIA GPU
  - Linux operating system
  - Altair Engineering's PBSpro v9.x batch job scheduler

- Homes File System
  - Three NFS servers
  - NFSv3 via IB/IP
  - 100 TB

Archive Systems

- HSM Software: SGI's DMF
- Archive Silo Capacity 90 Petabytes
- 6 – T950 Spectra Logic Tape Libraries
- 60,000 media slots by 12/20/2011
- Unique data 14.6 PB
- Dual Copy media ~29 PB media
- Lou1 (SEAS)
  - Altix 4700 32 cores, 64 GB memory
  - 280 TB Archive RAID Cache
  - 21 LTO5 - Media Size 1.5 TB
- Lou2 (Science)
  - Altix 4700 80 cores, 160 GB memory
  - 370 TB Archive RAID Cache
  - 28 LTO5 Tape Drives
  - 21 LTO4 Tape Drives

Broad and Diverse CSC's HPC Customer Base

CSC operates HPC systems with aggregate capability with **2+ petaflops,** and manages **25 petabytes** RAID and over **120 petabytes** of archival data across all sites.

Increasing 40TF->1.3PF in 18m

- Consolidating 3 HPC centers
  - GFDL (Princeton), NCEP (Gaithersburg), ESRL (Boulder)
  - Build-out 26,000 sq. ft. computer floor space in NESCC (Fairmont, WV)
  - Dec 2010
    - Deployed a 260 TF system at ORNL/DOE
  - Dec 2011
    - Deploy new HPC system 720 TF at ORNL/DOE
    - Deploy new HPC system 380 TF at NESCC
  - Research Applications (**weather.gov**)
    - Weather, Climate, Hurricanes, Tornado, Severe Weather
    - Forecast (Aviation & Marine), Air Quality, Ocean, Space

NOAA HPC Archive (GFDL) SGI pDMF – 30 PB today

- 155 Million Files, 20 PB -> 120 PB by 2013
- Typically 100-140 TB/day in/out tape drives (66% writes / 33% reads)
- Typically 200-300 TB/day between Archive and Post-Processing Systems via 10ge network
- 8 HA NFS edge servers (16 -10ge), 4 Data Movers, 2 MDS
- 2 PB, 30 GB/s RAID, 20 Filesystems (57-228 TB)
- 7.5 GB/s Tape bandwidth, 52 T10KB, 8 T10KA
- 3 Oracle SL8500 (150 PB) & 3 STK 9310 (~45,000 tape slots)

NOAA HPC Archive (NESCC) IBM's HPSS

- 24 Million Files, 17 PB -> 90 PB by 2014
- 5 VFS servers (10ge & IB), 4 Data Movers, 1 Core, 1 Spare node
- 960 TB, 10 GB/s RAID
- 7.6 GB/s tape bandwidth, 32 T10KC, 30 TS1130
- 1 Oracle SL8500, 10,000 tape slots, 50 PB Archive Capacity
- 3 STK 9310 with TS1130 tape drives, 25,000 tape slots
  - Soon to be retired

Archive Topics

- Why Invest in an Archive Solution?
- Technology
- Terminology
- Common Archive Features
- Architecting an Archive System
- Evaluating an Archive Solution
- Running an Archive System
- Online Resources

Why Invest in an Archive Solution?

- Save Money $$$$
- Most data is static (not changing or being read)
  - Small Daily (Read+Write)/Total_Data (<.02%)
- Future data importance hard to determine
- Linear or exponential data growth
- Data too costly to recreate (Experimental)
- Only copy of the data (Satellite)
- Data requirement preserved for years
- Time consuming to integrate new storage
- Expensive to backup, beyond backup window
- Time consuming to retrieve a file from backups

Why Invest in an Archive Solution?

- Total Data Size
- Age of Data
- Data Reuse
- Data Importance
- Data Management
- Data Growth
- Backup vs Archive Differences
- RAID vs Archive Costs

Backup vs Archive (1 GB Static File)

- **Backups** (Meta Data and Data Blocks)
  - Keep backups for 1 year (Dependent upon System Security Level (Moderate))
  - Daily Incremental, weekly full backups
  - At least 52 copies of a file on tape(s)
  - User wants file: Normally Admin selects backup and reads file and then provides file to user.
- **Archive** (Policy: two copies to tape(s))
  - Meta Data and Data Blocks (DB) are separate
  - Meta Data keep online, 2 data copies on separate tapes
  - User opens file: system automatically reads data from tape
  - Only 52 copies of Meta Data on backup tape

Archive system can be a backup solution, but backup solution can not be an archive solution.

Initial Purchase: RAID vs Archive

- 10 PB data, 60 million files, average file size 125 MB, Cache bandwidth 2.5 GB/s, tape bandwidth 1 GB/s, 7x24 support
- RAID 10PB – using 10ge and pNFS servers
  - Purchase: $, Power, Cooling, Size

    $5-7x, 14x KW, 40x Tons, 1.4x sq ft
- Archive – using 700 TB cache, 10 PB media
  - 8 tape drives, 6700 tape slots.

    $1x, 1x KW, 1x Tons, 1x sq ft

4 Year TCO: RAID vs Archive

- 10 PB data, with 50% data growth, 60 million files, average file size 125 MB, Cache bandwidth 2.5 GB/s, tape bandwidth 1 GB/s. End Year 2, upgrade data cache to 7 GB/s, tape bandwidth to 3 GB/s. Data grows to 50 PB at the end of the fourth year. 7x24 support. Assume new media is purchase twice a year. One copy of the data written to tape.
- Media (PB)
  - Year1:15; Y2:+8 ; Y3:+5,+6; Y4: +8, +8
  - End Year1:15; Y2:23; Y3:34; Y4:50
- Assume staffing cost the same

4 Year TCO: RAID vs Archive

- RAID 10PB-> 50 PB
  - Y1 15 PB @4 GB/s, Y2 +8 PB
  - Y3 +11 PB @7 GB/s, Y4 +16 PB
    - Purchase: $5-7.5x, 30x KW, 50x Tons, 3x *sq ft*
- Archive 10PB->50 PB
  - Y1 700 TB cache @ 3 GB/s, 1 GB/s tape, 15 PB
  - Y3 2 PB @ 7 GB/s, 3 GB/s tape, 50 PB media
  - 8+12 tape drives, 20,000 tape slots.
  - Purchase $1x, 1x KW, 1x Tons, Size 1x sq ft

12 Year TCO: Archive 15x Better than RAID

- "In Search of the Long-Term Archiving Solution – Tape Delivers Significant TCO Advantage over Disk"
  - By Clipper Group in December 23, 2010
- 45% annual growth rate over 12 years starting with 3 PB growing to 86 PB
- RAID and Tape upgrades every three years
- 2 TB disk with RAID6 and LTO5 for tape with 85% utilization for both technologies
- **RAID power cost alone greater than tape solution**

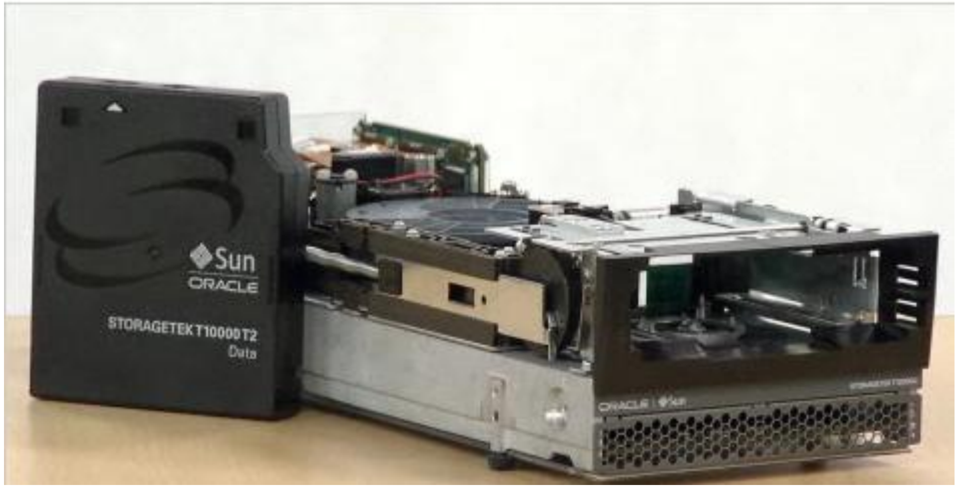
RAID vs Archive

- RAID (+ Strength, - Weakness)
  - + Micro-second to Seconds access to any data
  - + Easily integrates into applications workflow
  - - Takes time to test, integrate new systems and storage
  - - May need to shuffle users' data to new storage
  - - Large power usage (always on, if not being used)
  - - Risk major data loss – (Under Heavy Load – like a RAID set rebuild)
    - **Consumer SATA: Hard Error, 1x10^14 i.e. 12.5 TB, 1-3 days**
    - **Enterprise SATA: Hard Error, 1x10^15 i.e. 125 TB, 10-30 days**
    - **Enterprise SAS : Hard Error, 1x10^16 i.e. 1.25 PB, 70-140 days**

  Note: Disk Hard Error is sector
- Archive
  - + Data older few-24 hours on tape
  - + Keep important and recently used data in disk cache
  - + Low power usage (most data on tape vs disk)
  - + **Hard Error, LTO5 1x10^17, TS1140 & T10000C 1x10^19**
  - - Seconds to minutes to hours to access data
  - - Train staff on new technology
  - - Train users to use archive system in existing workflow

Archive Technology

- Tape Media
- Tape Drives
- Tape Library
- RAID
- MAID
- RAIT



Tape Media

- ½” Media with Barcode Labels (Volume Serial Number – VSN) – AKA – Tape Label & Tape Type
- Cartridge Size – ~4”x~5”x1” – 1984 IBM 3480
- Mount, Load, Seek, Read/Write, Rewind, Dismount
- Cleaning Tape – used to clean tape head
- WORM – Write Once Read Many
- Short (1/5-1/8 less media, ~3x faster to first byte)
- Compression / Encryption
- Shelf Life – 10-30 years
- LTFS Format (New)
  - Makes tape media like removable filesystem
  - “Self Describing” data on tape

Tape Media

- Important Features
  - Entire Tape Reads/Writes
  - Reliability (Bit Error Rate)
  - Lifetime Tape Warranty
  - Pretested Media

Note:

Buy media from same vendor used to qualify the drive

Media needs to be stored in controlled environment

Relative Humidity (Operating, Storage): 20-80% vs 15-50%

Temperature (Operating, Storage): 50-90F vs 59-79F

Dust is a tape killer

Tape Drive

- Open vs Enterprise drive
- Media reuse (TS1100, T10000), increase capacity
- Backward read compatible (1-3 generations)
- Data compression to tape (at full tape speed)
- Next generation drives 2-3 years apart

Roadmap GA	Current Generation		Gen N+1		Gen N+2		Gen N+3	
	TB	MB/s	TB	MB/s	TB	MB/s	TB	MB/s
LTO (FH) 4/1/2010	1.5	140	3.2	210	6.4	315	12.8	472
TS1100 6/3/2011	4	250	8-10	<360	14-20	<540	TBA	TBA
T10000 1/31/2011	5	240	6-10	270-400	12-20	400-600	TBA	TBA

Tape Drive

- Media - $/TB – Likely largest percentage cost in an archive solution, LTO prices decline drastically over time
- Newer tape drives have support the ANSI T10 checksum features, but not yet in OS kernels or device drivers or HBAs
- Short & partitioned media - to decrease (3x) access to first byte
- Important Features

Features	Pricing	Units Shipped (1,000/yr)	Tape Load (s)	Rewind (average) (s)	Tape Unload (s)	Media Bit Error Rate
LTO5 (FH)	$	~100+	12	50	17	1x10^-17
TS1140	$$	~5	15	42	24	1x10^-19
T10000C	$$	~5	13	57	26	1x10^-19

Tape Library

- Also know as silo, tape robot, tape jukebox
- An automatic method (via robot) to mount/dismount media into a tape drive
- Media has Barcode Labels (Volume Serial Number – VSN) and tape type
- Sizes vary from 10's-400,000 tape slots
- Some Vendors – Offer Tape Library Complex
  - Joining Multiple Tape Libraries together via tape mover(s)

Tape Library

- Important Features
  - Lifecycle Management Metrics – Media, Tape Drive, Library
    - This is like Insurance – Never know when something will go extremely bad
    - Easy to track issues between tape drive and media
  - Redundancy/ Reliability
  - Mounts Per Hour
    - Numbers are peak, not measured the same between vendors
  - Density (PB/sq ft)
  - Self Maintenance, SLA
  - Encryption support

Redundant Array of Independent Disk (RAID)

- Storage virtualization layer to read and write data in parallel across set of physical disks in order to increase reliability or performance, or both
- Most RAID controllers support multiple levels (0-6) and nested (01,10, ..), (block-level striping with double distributed parity)
- With RAID6 can lose two disk drives without loss of data
- Archive systems normally use RAID6 devices
- Useful life – 3-5 years

- **Note:** With larger disk capacity rebuild times are an issue

Massive Array of Idle Disk (MAID)

- Used for 1st or 2nd level storage
- Best used for small file access
  - If large percentage of small files (80>%) and
  - Total space of small files (few hundred MB) is small (<20%)
- Emulate a number of tape drives – via a virtual tape interface
- Most of the disks are powered off or in low voltage idle state
- Disk failures extremely low
- **Note:** Not all MAID devices certified/integrate into archive systems.

SC11 Seattle, WA Connecting communities through HPC

Redundant Array of Independent Tape (RAIT)

- RAIT 0 Uses multiple tape drives in writing/reading data in parallel. Increase performance, zero redundancy
- RAIT 5,6 (D+P) Uses multiple tape drives in writing/reading data in parallel with parity blocks spread across tapes. Up to P tape(s) not readable or lost, data file(s) can still be read **(HPSS v8 - In Development)**
- Best used for very large files (100 GB – today) or large file aggregation sets (need to restore whole group of files)
  - Guideline for each file chunk – keep tape drive busy for at least 60 seconds

Note: Not very many sites have large enough files to take advantage of this feature

Terminology

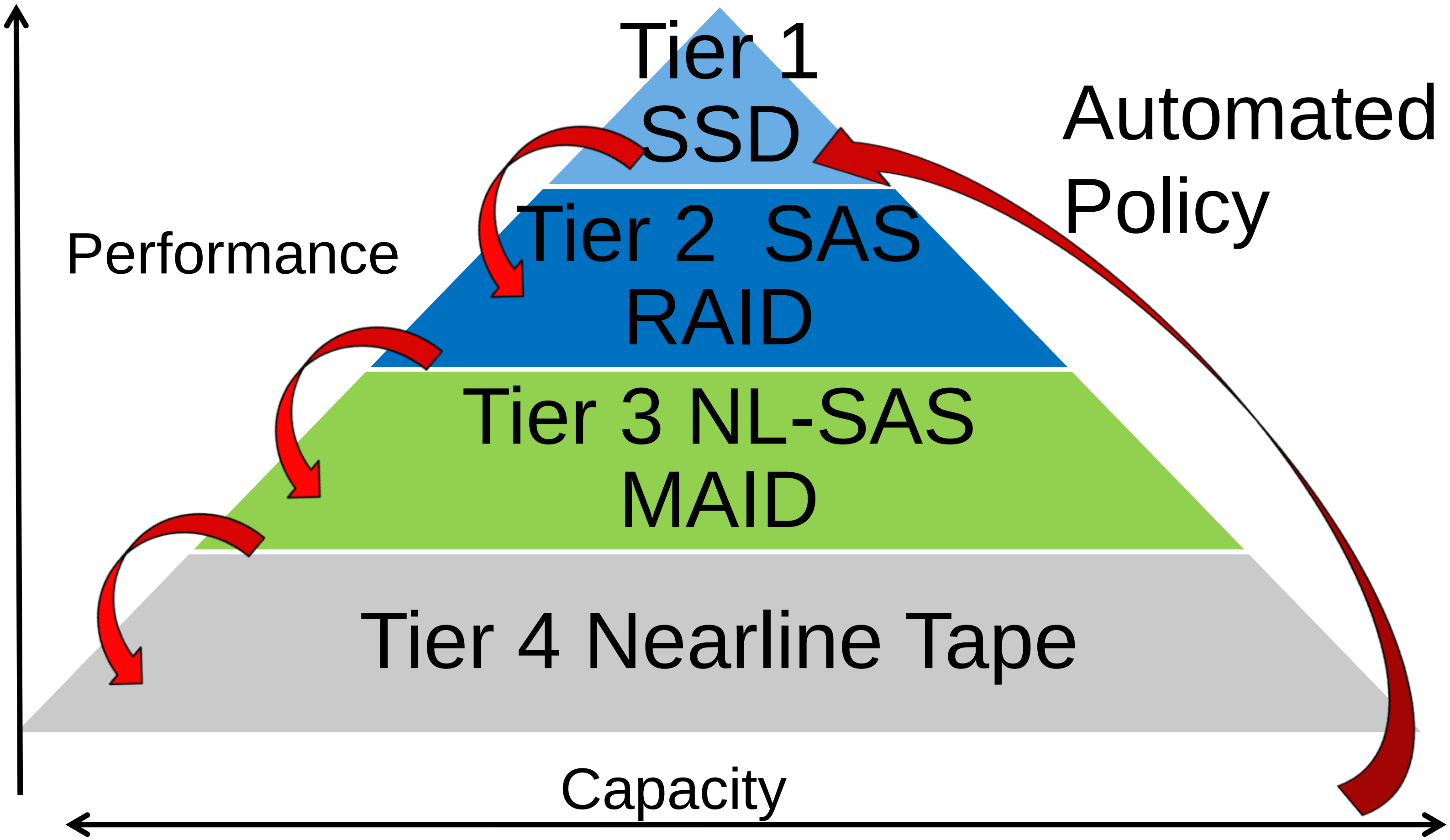
- HSM & Migration Levels
- File Types and States
- Tape Pool, Virtual Tapes
- Tape Compaction and Reclaim
- Archive Cache
- Data Movers, Core Servers, Edge Servers

HSM

- Also known as:
  - Hierarchical Storage Management - HSM
  - Active Archive, Archive System,
  - ILM – Information Lifecycle Management
- Definition:

  Set of policies that **automatically** move data between high-cost to low-cost storage devices.

HSM Migration



File Migration Levels/Concepts

- Both metadata and data on top level (online file – i.e. in Archive Cache)
- Metadata in Archive Cache, data blocks only on lowest level (offline/purge file – i.e. data on tape only)
- Data blocks copying to lower level (i.e. writing to tape)
- Data blocks copying to top level (i.e. reading from tape writing to Archive Cache)
- Data blocks on top and lowest levels (i.e. data on both Archive Cache and Tape)
- Some data blocks on top level (i.e. some data in Archive Cache)
- **Note** – Metadata and directories are left online

File State

- Deleted online files – still exist in backups
- Deleted archive file – meta data exist in database, backups, and data blocks on tape (soft-deleted). Data blocks can be recovered.

Tape Pool, Virtual Tape

- **Tape Pool:** Set of tapes used for certain class of data (user, project, size, age, tag, etc.) for defined set of tape drives.
- **Virtual Tape:** Virtualization layer making a RAID look like a tape drive. MAID devices emulate different tape devices.
- **Tape Compaction:** Process repacks partial empty tapes to a new tape. Once repack complete, data blocks can **NOT** be recovered (hard-deleted)
- **Tape Reclaim:** Process of making empty tapes from Tape Compaction available to use again

Archive Cache

- Large first level storage device, likely a RAID subsystem (SAS, SAS/nearline, or SATA disk)
- Increase reliability use RAID 6
- Archive cache size and bandwidth dependent upon:
  - user access patterns
  - site policy
  - budget
- Archive Cache large enough to ingest incoming peak data load, and write out to tape without overloading cache

Data Movers, Core & Edge Servers

- **Data Movers**: access to local/common filesystem and directed by core server to read/write data from/to a tape drive. (higher level tier) (A process or server)
- **Core Servers:** only servers to have direct access to metadata database. The traffic cops directing data traffic based on site policies.
- **Edge Server:** access to local/common filesystems and NFS exports filesystem to clients. (A process or server)

Questions???

Archive Common Features

- Scalability
- Tape pools and Storage Classes
- Data movement through hierarchy
- Data replication or redundancy
  - Dual copy
  - RAIT
  - Replication
- Data Integrity
  - Checksums
- Data re-mastering
- Technology refresh capabilities
- Parallel transfer capability
- File handling instructions
- Direct-to-tape operations
- High availability
- System management
  - Logs and monitoring
  - Management consoles/capabilities
- Data reduction
  - Compression, deduplication

Scalability

- Namespace scaling
  - Numbers of files/objects in a given directory (automatic file aggregation)
  - Metadata operations per second performance
  - Parallelization vs. serialization of common metadata operations (file attribute changes, deletion, creation, listings)
- Storage to object scaling
  - Subsystems or multiple file systems
    - Ability to separate the namespace into separate non-shared metadata and storage resources
  - Numbers of devices
    - Scaling to high numbers of disk and tape devices

Pools and Storage Classes

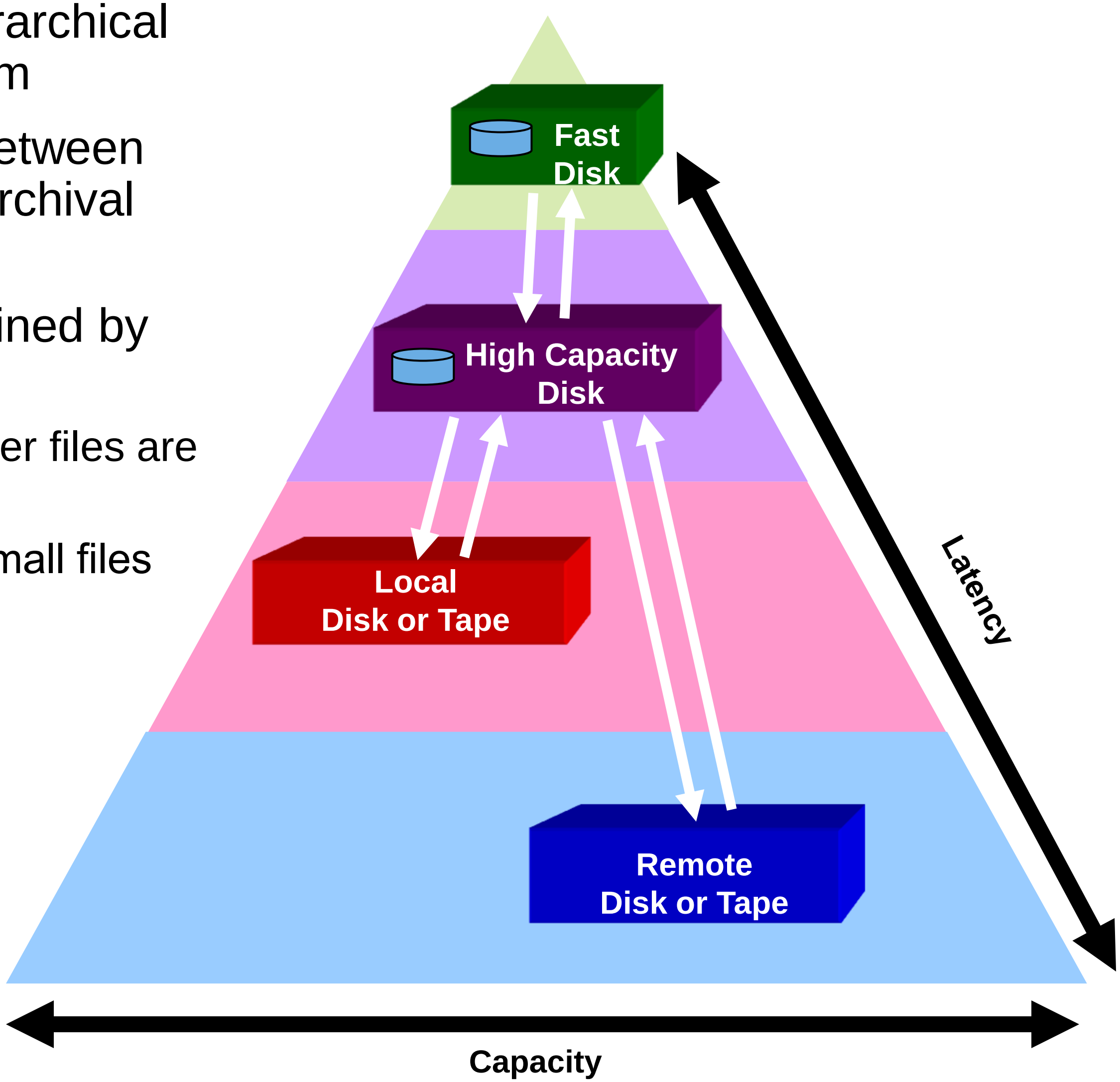
- Methods of dividing files/data in an archive across distinct storage devices (disk and/or tape)
- Typically the method of assigning different kinds of disk (different block sizes, SAS vs. SATA, etc.) to files within the archive
- Many have specific capabilities and limitations around how files move to/from the pools/classes
  - Tivoli Storage Manager (TSM)
    - storage pools = different kinds of storage
    - the primary method of moving data between different kinds of storage, can span servers
  - High Performance Storage System (HPSS)
    - storage classes = different kinds of storage
    - the primary method of moving data up and down the hierarchy within a single HPSS instance
- Send a file to the most appropriate storage devices.

Storage Class 1: FC Disk

Storage Class 2: Tape

Data Movement Through Hierarchy

- Fundamentally define a Hierarchical Storage Management system
- Enable user data to move between storage devices within the archival system
- Migration and purge are defined by policies
  - Temporal (e.g. 10 minutes after files are created, migrate to tape)
  - File size (e.g. don't remove small files from disk)



Dual Copy or Redundancy

- Dual-copy
  - Thought should go into what you want the copy of
  - A copy of the disk file, versus having a copy of the first tape file
  - Or having it in a redundant archive system
  - How coupled or uncoupled are the copies, what is the primary purpose of the copy
- Redundant Array of Independent Tapes (RAIT)
  - HPSS-only, in development, available 2012 timeframe
  - Difficulty has been in compression/EOM/parity block handling
  - Generally to aid in saving media cost over having multiple copies
- Replication
  - Primarily in disk based systems/archives

Data Integrity

- T10 PI standard
  - Device level checksum capabilities
    - HBAs, available today
    - Tape drives: T10KC, TS1140, LTO5 available today
    - Operating systems, not yet.
  - Software
    - HPSS – file level, provided by and back to the client application for use
    - DMF – each data block read/written to/from tape
- End-to-end checksums are desirable
  - Meaning all the way from the data generation source, through all devices and back to the data retrieval client.
    - Any device, NIC, HBA, Disk, Tape that is involved in “writing” the file validates the checksum as data passes through it. This helps administrators identify the problem.
    - The client can then validate the checksum when “reading” the file to ensure it’s what was. This helps the user validate the data.

Data Re-mastering

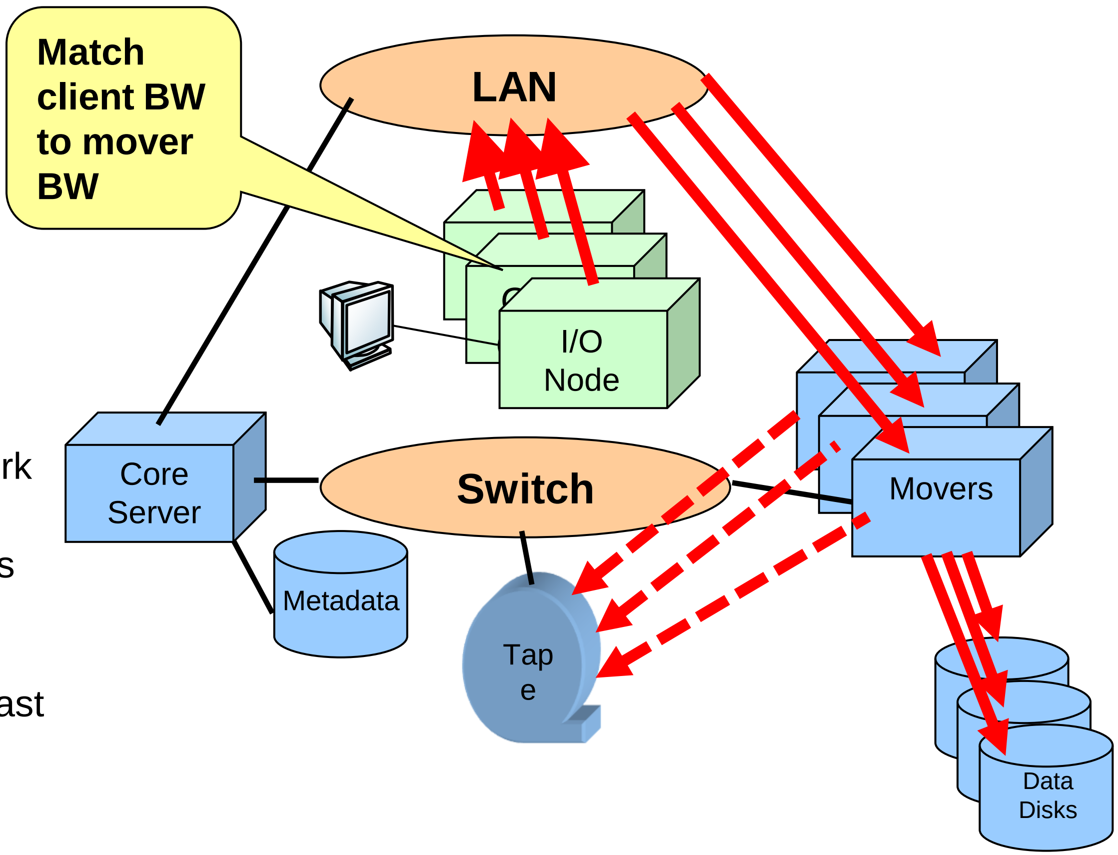
- Ability to copy files or user data from one storage device to another
  - Useful for moving between different storage device types/models
  - Helps reclaim space contiguous space from sparsely populated devices (a.k.a. DMF merge, HPSS repack/reclaim)
  - HPSS recover which copies files from a tape in a lower storage class to upper storage class in the hierarchy
  - Can aid in reading valid data off partially damaged devices
  - DMF move which can move or copy data to different storage pools

Technology Upgrade Capability

- HPSS calls this technology insertion, enabling old (for read) and new (for write) storage devices to exist within the same pool, then use repack to move data from old devices to new devices
- DMF changes an entry in the configuration to direct new data to new storage pool, then use DMF merge to copy data to new storage pool and remove files from old storage pool upon completion
- The idea is to be able to substitute a whole HSM storage level (e.g. tape level 1) with a new storage technology
  - Enables old data to be read from previously used devices
  - All new data is directed to new devices
- Critical to enabling rapid deployment of new technologies

Parallel Transfer Capability

- **Multi-threaded clients**
  - Parallel FTP
  - HSI – Gleicher Enterprises transfer interface
  - PSI – LANL transfer interface
  - GridFTP
- **Multi-node clients**
  - Utilizing multiple client system network interfaces for each transfer
  - Mostly a method to prevent large files from bottlenecking on any one networking interface’s speed
  - Also a method of optimizing time to last byte for a large number of files
- **Striping a file across storage**
  - Utilizing multiple archive data servers and archive devices for a single file (GPFS/TSM at LANL, HPSS)
  - RAIT (HPSS future at NCSA)



Direct-to-tape Operations

- Only for select applications
  - Streaming large data (keeping the tape busy for 1+ minute)
- Archives are capable of not having a disk cache
- Instead allowing the user or application to read/write or put/get directly to/from tape
- Recommended only for large file I/O (streaming, sequential nature of tape), and further:
  - Data you will not read
  - Large data sets that are bigger than the disk cache or with mass ingest (i.e. trying to optimize time to complete ingest)
    - Won't suffer performance loss writing to disk having to compete with migration for long running ingests

High Availability

- No one archive system provides complete HA
- Enable upgradeability and flexibility in not having system down during failures
- Generally aim for HA components within the archive system
  - Storage device HA
    - Server redundancy or failover
    - Device failover
  - Tape libraries
    - Redundant electronics
    - Multiple TCPIP or FC connections
    - Library management server/software
  - Most focus on ensuring the servers don't fail
    - Direct attached vs. SAN attached can improve fault tolerance for tape/disk
  - Tape drives
    - Don't offer HA features
  - One library offers automatic hot-spare of drives
  - Multi-pathing on disk arrays

System Management

- Most have both a command-line and graphical user interface to interact and manage the archive system components
  - Disk devices
  - Tape devices
  - Servers
  - Aid in creating new disk/tape resources
- Logs
  - Transfer logs are unique to archive systems and very useful for operating the system as well as understanding its usage over time
- Monitoring
  - Tape monitoring solutions (e.g. StorSentry, RVA)
  - Most archives are lacking adequate monitoring for the devices

Data Reduction

- Compression
  - Tape drives provide natively, typically 20-50% for scientific data
  - Typically disk solutions don’t provide
  - TSM provides this in the client to save network bandwidth
- Deduplication
  - Only provided on certain disk solutions today

SC11
Seattle, WA
Connecting communities through HPC

Questions
Break – 30m

Architecting Archive Systems

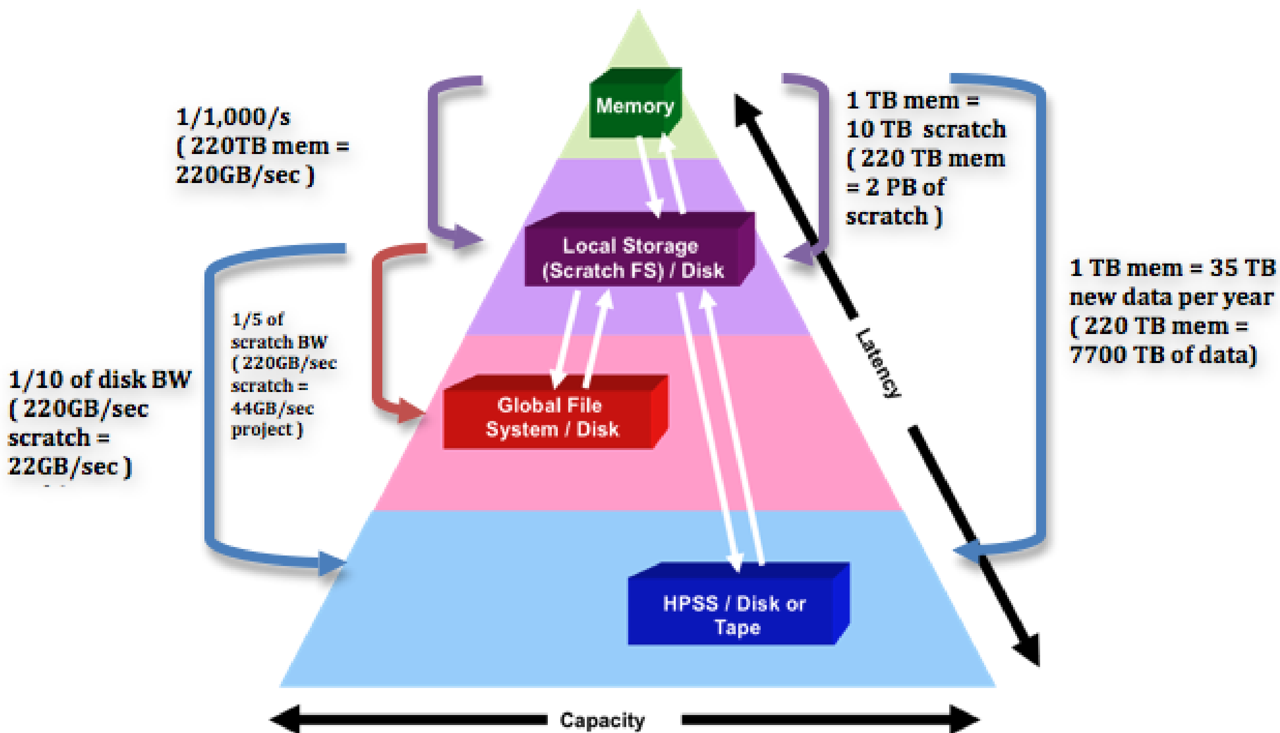
- Guiding principles
- Archive requirements
- Procurement tips
- Building an archive system
  - Tape slot planning
  - Selecting servers
  - Disk cache sizing
  - Tape drive needs
- Environmental considerations
- Challenges

Principles of long-term data stewardship

- Classing data
  - By file size – per file operations vs. bandwidth
  - By access characteristics – frequently needed, frequently updated
  - By protection needs – copy, disaster protection
  - By project – sharing, exporting
- Data protection
  - Backup – on distinct hardware, mirroring and active redundant sites not enough
  - Copies – on distinct hardware
  - Geographic dispersion
  - Re-mastering to new technology every 3-6 years – touching, managing data frequently is good
- Value
  - Cost to maintain
  - Cost to generate
  - Capability to regenerate (actual events, historical)
- Relevance
  - Data format, apps/tools to use it?
  - Uniqueness or commonality
- Hardware and Software solutions
  - Hardware and software must be reliable and dependable (natural and other events must not effect data)
  - Data integrity (data at rest must stand the test of time, 10+ years)
  - Optimize spatial footprint (highest GB/sqft)
  - Solutions must be durable to handle high utilization and allow reuse of assets (1-2 year periods of full mount/rate activity)
  - Components must be cost efficient (lowest $/GB)
  - Solutions should be non-proprietary (survival beyond that of software/hardware technologies and vendors)
- System independence
  - Archive system must survive change in the Center (operating systems, compute systems, file systems, networks)

HPC Archival storage requirements

- These conventions generally apply to HPC sites
- User archive requirements are important to gather, but tend to be quantitatively inaccurate.
  - Actual amount stored is normally a very smooth and predictable growth rate from year to year.
- Gartner states that on average the amount of archived data is growing at 50% per year.
- Look at the amount being generated in file systems per day/month
  - Growth in the archive is normally half the amount of writes per month or year.
- A study at LLNL and validated at NERSC shows that 35TB of data per year is generated for each 1TB of HPC system memory.
- Convention says to target the archive bandwidth at 10% the file system bandwidth.



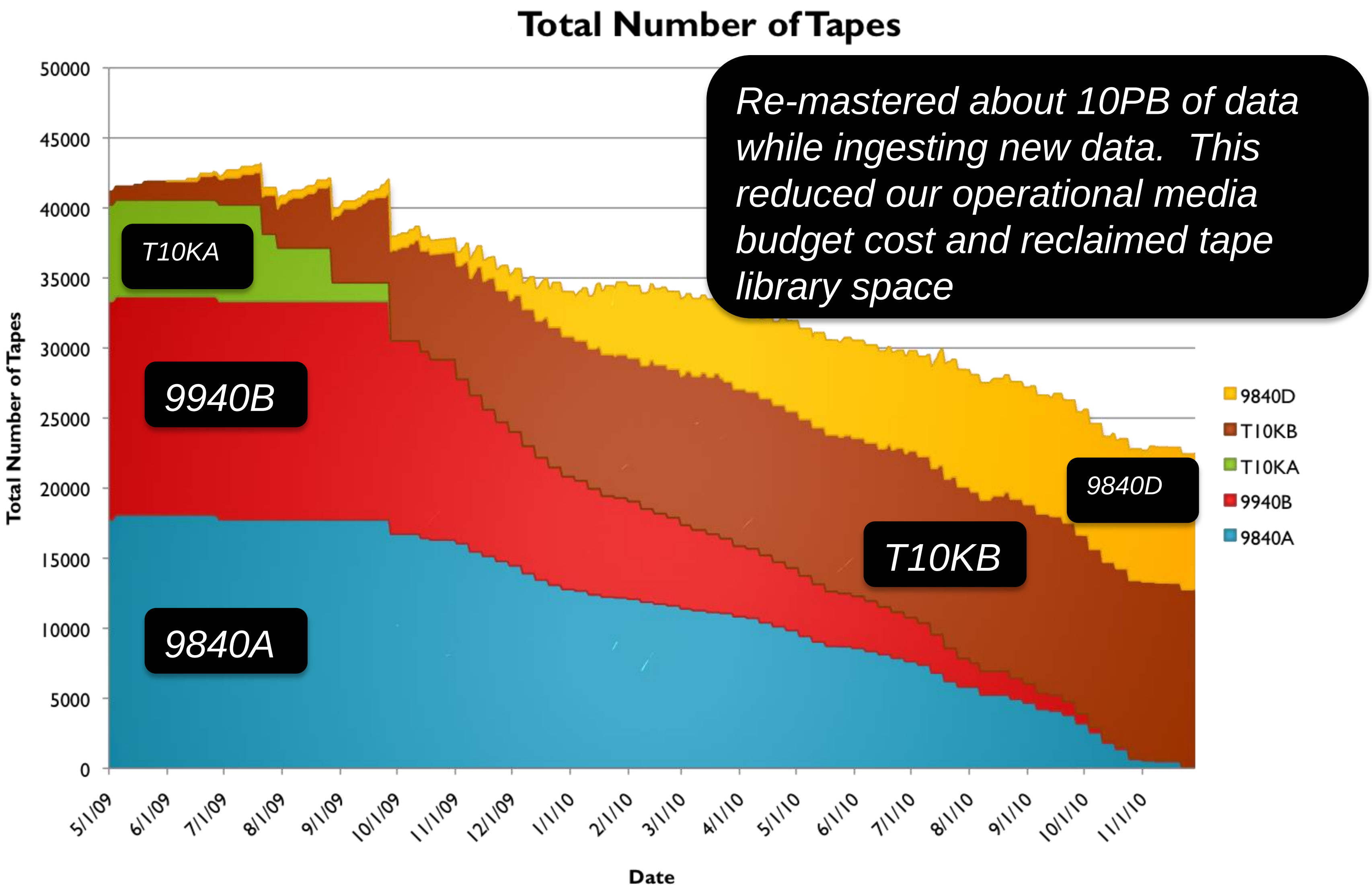
Building an Archive System

- **Determine the number of tape library slots required**
  - Amount of new IO and growth expected and media costs/procurements by type
  - Purchase forecasting for new libraries, tape drives, media
- **Select number and type of servers**
  - COTS vs. enterprise systems
  - Metadata vs. data servers
  - Direct connect vs. SAN
- **Disk cache sizing**
  - Is disk living up to our policy? (e.g. 5 days of peak IO retained on disk, keep small files disk resident forever)
  - Purchase or rebalance disk?
- **Selecting the number and type of tape drives**
  - Concurrent usage for total number of tape drives needed
  - Match bandwidth of disk
  - Generally over provision for failed drives and re-mastering
  - Hot spots within libraries (passthrough activity, tape drive moves)
- **HSM Software considerations**
  - Percent read/writes
  - Manageability of large system
  - Interface features/needs
- **Other considerations**
  - Environmental (fire suppression, contamination/location)
  - Monitoring solutions
  - Disaster recovery

Determining Number of Tape Slots

- Spreadsheets work well
- Typically calculate monthly growth with multiple scenarios (e.g. introducing a new technology or not)
- Plan by month using amount of new data expected
- Account for capacity changes (new capacity tape drives)
- Calculate numbers of tapes needed
- Ensure amount of tape slots available is adequate
- If not, consider purchasing higher capacity drives as mitigation against buying another library
  - Early adoption and aggressive data migration to new technology can reduce number of libraries needed

Understanding a Key Benefit of Tape over Disk



Selecting Servers

- Number of servers
  - Generally about the number of network connections needed to provide necessary bandwidth
  - Also about the number of storage devices a server can support
  - Direct-attached storage devices vs. SAN attached
    - Direct-attached will drive up the number of servers required
- Type of servers
  - Enterprise vs. commodity: predominantly by archive system features in preventing or recovering from failure
  - Each can work, consider balancing investment in-house staff expertise and effort with what a vendor can provide
  - Data movers typically require maximum PCIe slots for tape/disk and network connectivity, modest memory
  - Metadata server typically require large amounts of memory and minimum amount of network connectivity

Disk Cache Sizing

- Establish policies for file retention on disk (NERSC policies):
  - Keep 5 days of peak archive ingest on disk
  - Keep all files <100MB on disk forever
- Budget is the largest factor for disk
  - Balance between scratch file system and archive disk is difficult
  - Some size to around 1/10th of current or intended tape capacity each year
- Performance (bandwidth) for the archive is another factor
  - Disk is still the cheapest method of achieving single file transfer bandwidth
- Ensure you plan/size for write (ingest) and reads from tape if data is copied back to disk in the archive.

Determining the number and type of Tape Drives

- Considerations for type of tape drive
  - Slots x tape capacity = data archived
  - Small vs. large file handling/speed, per file seek performance
  - Bandwidth of the drive
  - Reliability features
    - Bit error rate, LTO vs. Enterprise
    - T10 PI functionality (checksum)
- Determining number of tape drives
  - Number of concurrent user requests
  - Amount of data to ingest vs. time to move data from disk to tape
  - Service responsiveness from vendor
    - Buying extra drives vs. paying for quick response

Other considerations for number of tape drives

- Concurrent occupancy stats
  - plan to peak for archive
- Utilization stats
  - wear-and-tear on drive
- Numbers of mounts per drive
  - balanced hardware use
- Pass-through port activity
  - drive to data locality
  - Hotspots in library

Chart 21: Archive T9840D Utilization

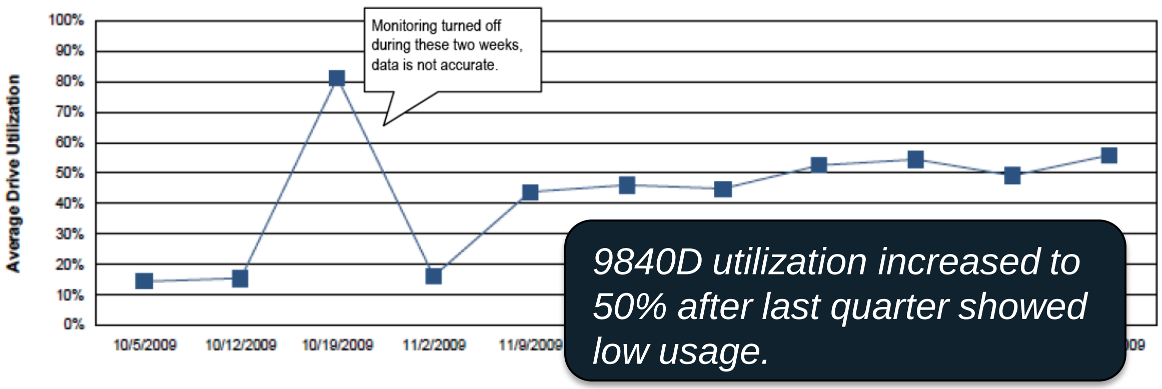


Chart 42: LSM to LSM Movements With More Than 20 Tapes Moved

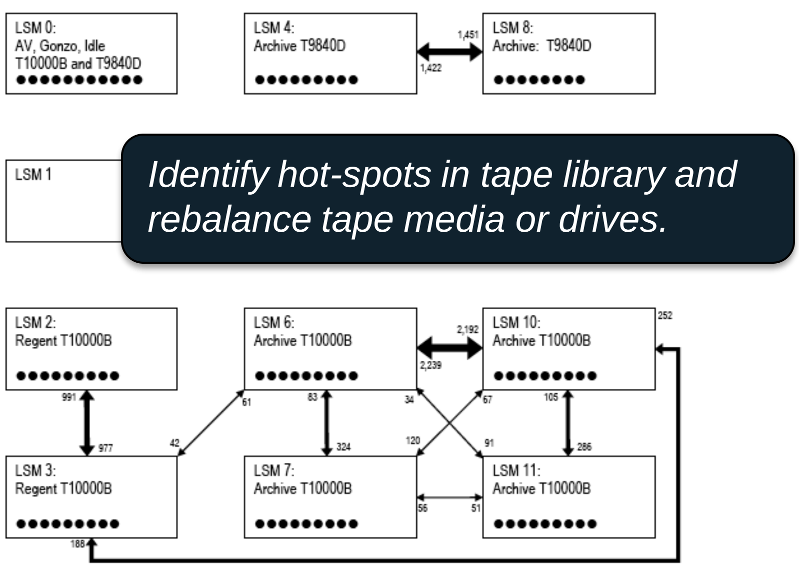
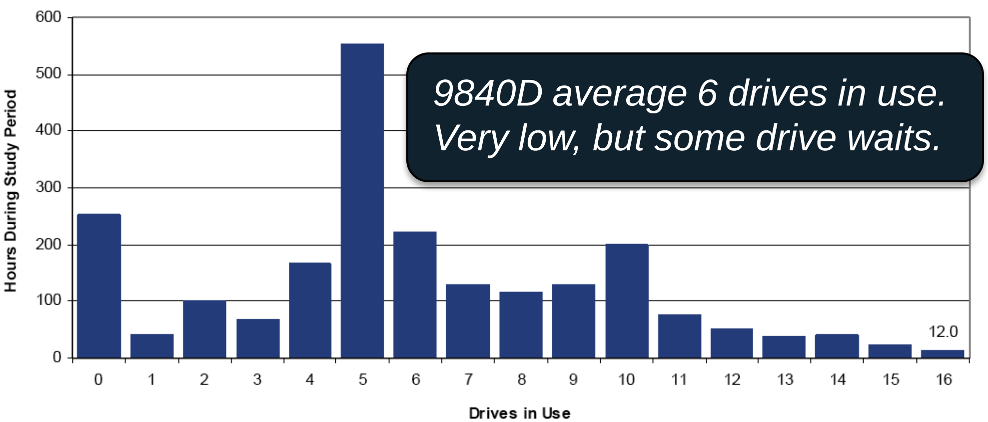


Chart 29: Archive T9840D Simultaneous Drives In Use



SC11 Seattle, WA

HSM Software considerations

- Amount of read/write activity
  - Richness of features in moving data between disk and tape (striping or concurrency to disk/tape)
  - Smarts for retrieving data from tape (scheduling, ordering, performance of reads)
  - Understand which operations are parallel vs. sequential in handling IO (from requests to resources provided)
- Disruption to usage for changes
  - Consider activities that will need to be accomplished during downtimes
- Ability of software to increase or manage hardware
  - Large numbers of disk and tape devices
  - Introduce new disk or tape devices
- Interface needs
  - Parallel or multi-network transfers
  - Data aggregation
  - Data integrity (checksumming, error checking)

Special considerations for active archives

- High frequency read is typically defined as sustaining an average of more than 25% of total IO being read operations.
- Minimum number of tape drives is four
  - One for read, one for write, one for peak, and one for spare
- Monitoring in place to understand peak and average load
  - Gathering usage statistics – per transfer bandwidth, concurrent transfers, commonly requested tapes
  - Determine tape drive utilization, occupancy, concurrency
- Budget tape drives to peak demand
  - We had an average of 16 drives in use a few months ago
  - Our peak demand was 30 drives
- Size the HSM disk cache properly for repeat reads
  - Analyze the cache hit ratio
  - Plan to hold at least 5 days of peak data transfer on disk
  - Study the stage requests (tape to disk)
- Identify and resolve library hotspots
  - Ensures we have cartridge closest to drives for quickest mount time
- Only allow select applications to do direct-to-tape IO
  - Parallel file system backups and restore
  - But gain the benefit of not exhausting a disk cache on massive ingest

Environmental considerations for tape

- Tape is fairly sensitive to airborne particulate because it is removable
  - 0.3 microns is the most damaging to the tape drive heads
  - Most tape drive, library, media specs require ISO Class 8 machine room standards be maintained
- Tape is sensitive to temperature because it is removable
  - Staying in the recommended operating range is key to avoid issues
  - Tape temperature can affect ability to read/write data (mostly about track width)
  - Can be damaged by stiction (tape media surface sticking to tape head) due to temperature issues
  - Tapes are often read/written on different drives (different operating variables)
- Tape is sensitive to humidity changes
  - Constant environment is key
- Best practice is to isolate the tape system from other systems
- Fire suppression is a concern for many with desire to protect the data and prevent further problems with enclosed space (library)

Challenges

- Enabling exponential growth without an exponential budget takes effort
  - Near constant multi-year PB sized data migrations to new technologies
- Be vigilant at handling tape hardware issues
  - Archive is often the last resort for data
- Changes in networking capabilities, compute capabilities, or file system capabilities increase the amount of archived data

SC11
Seattle, WA
Connecting communities through HPC

Questions???

Evaluating an Archive Solution

- Performance
- Acceptance Tests, Benchmarking (at scale)
- Features
- Limitations
- Total Cost of Ownership

Performance

- Tape Library: EMnt/hr
- Tape Drives: Single (ETBw), Scaling, All (OETBw)
- Archive Cache: I/O Bandwidth, Metadata
- Networks: iperf and gridftp: single, all NIC ports
- Exported FS (ex. NFS, GPFS) performance: I/O Bandwidth, Metadata
- Archive Throughput: High Peak Usage – 20m
  - From tape and cache read data and transfer to node
  - From node transfer data to tape and cache

Effective Mounts per Hour (EMnt/Hr)

- **Note** – Vendors Rates – are Peak Numbers
- Vendors – Exchanges per Hour not EMnt/Hr
- EMnt/Hr = 3600/s (Total_Time/Total Mounts)
- For all tape drives in a tape library, simultaneously, Mount, Load, Dismount a random tape, then repeat again with different tape. Track the total time to complete all steps.
- Ex. 1 Tape Library, 32 Tape drives: Select 64 tapes randomly from the silo. Assume 130s to complete above task. (Repeat multiple times)
- EMnt/Hr = 3600/(130/64) = 1772

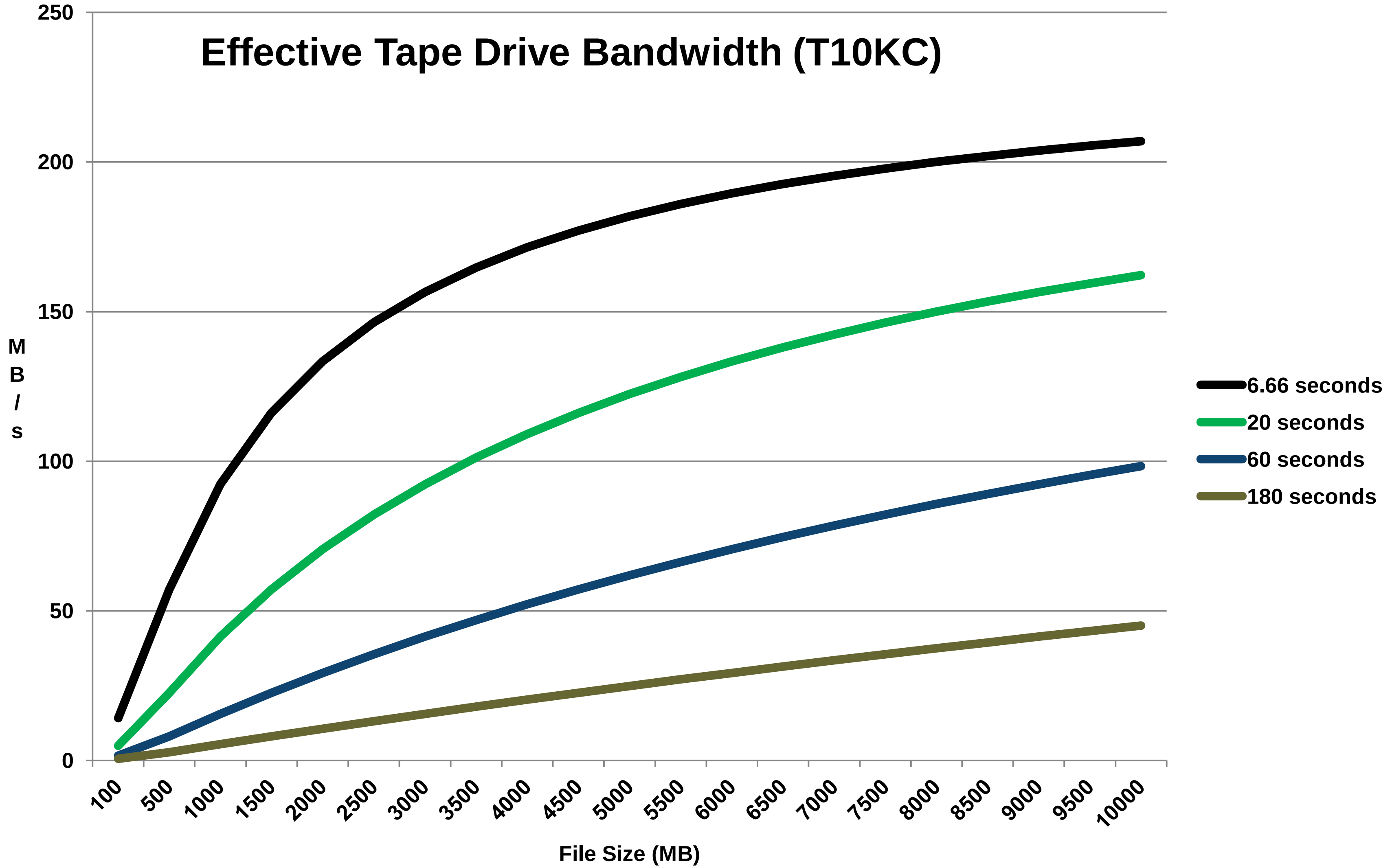
Mounts per Hour has a Huge Impact on Retrieving Nearly All Files (99%) @ GFDL

- Tape Libraries and Drives
  - Mounts Per Hour
    - At GFDL: files <7.5 GB 99% of files, only 46% of total data (average filesize 459 MB)
    - 7.5 GB file – roughly 30 seconds to read using latest tape drive (T10KC, etc.)
    - 70+ GB file needed to achieve overall bandwidth of 200 MB/s with mount/dismount total 60 seconds
  - Effective Tape Drive Bandwidth =
  - Filesize_MBytes / total_time_sec ((mount + seek + dismount) + IO)
  - 70000 / (60 + 291) = 199.4 MB/s

SC11 Seattle, WA Connecting communities through HPC

Overall Effective Tape Drive Bandwidth

- OETBw(7.5g) = ((Total_Data(MB) / (Total_Time)) / #tape_drives
- Write 7.5 Gbyte file on a random tape – 2x number of Tape Drives Times
- For all tape drives in Tape Library, simultaneously, Mount, Load, **Read**, Dismount tape, then repeat with different tape. Track total time to complete all steps.
- Ex. 1 Tape Library, 32 Tape drives: Select 64 tapes from the silo. Assume 190s to complete above task.
- OETBw(7.5g) = 480000/190/32 = 78.94 MB/s



Ex. File_Size / total time (mount + seek + dismount) + IO_Time)) = ETBw

7500 / (60 + 31.25) = 82.19 MB/s

Archive Features

- **Data Integrity, Data Integrity, Data Integrity**
  - How does system check to verify data is the same?
  - How does system check database is correct?
  - How to recover from a major failure?
- Highly Available (not same between vendors)
  - Does client network services drop or blocked until alternate server is started?
  - Does a tape file request re-queue automatically on failure (tape drive or data mover)?

Archive Features (Continued)

- MetaData Operations
- Quotas: by inodes – encourage users to group files in tar or cpio file.
- Metrics: Capacity Planning (tape drive usage, data growth by users, projects, system, daily read/write to filesystem and tape drive)
- Important to some sites:
  - Write direct to tape from non-archive mounted fs
  - Partial file read (for large files)
  - Striping large files across multiple tape drives

Archive Limitations - Be Aware:

- of software and systems limitations (max tape drives, NICS, FC ports, pcie buses, size of file, number of files, etc.)
- of the risks of single points of failure
- of disk rebuild times
- if down time needed for any device for BIOS, firmware, software upgrades
- potential hotspots on system
- what logs are used to help isolate issues
- **Vendor numbers are always peak numbers, under real workload peaks are very different.**

Total Cost of Ownership (4 years)

- Software purchase, licenses, maintenance
- Hardware purchase, maintenance
- Upgrade(s) purchases
- Conversions (old to new)
- Staffing
- Installation and Shipping
- Media (Data and Cleaning)
- Facility
  - Space (**Opportunity Cost**), Power, Cooling

$23K/sq ft – Opportunity Cost for Compute Rack

- Inputs:
  - Nodes/Rack - 64
  - Space (with access – sq ft) - 16
  - Cost power ($/KW/hr) - $0.11
  - Price per node hour - $1.20
  - Average Node Utilization - 60%

Opportunity Cost								Date: 10/13/2011	
$/sq ft/Yr	$/node hour	Nodes / Rack	Hours / yr	System Utilization	Rack with Access (sq ft)	KW/Rack	$/KW HR	Power	Cooling
$ 23,043	$ 1.20	64	8760	60%	16	22	0.11	$ 21,199.20	$ 13,779.48

Tape Library 4 year TCO

- Requirement: Support 7x24, 4 hour response
  - CY1 provide 50 PB Tape Library and Media and 7.5 GB/s Tape Bandwidth
  - CY3 add additional 80 PB Tape Library and Media and 10 GB/s of next gen. tape drive
- Evaluate 3 configurations: A-C
- Verify floorspace for any configuration, space cannot be used for another purpose
- Assume system and staff cost stay the same and no data compression
- Inputs: Tape Library, Drive, Maintenance, Media, Cleaning, Install, Shipping

Note: All Pricing are Fictional — **Date** 11/1/2011

Config	4 Year TCO Co	$/ PB	$/ TB	$/ GB
A (T10K)	$ 5,579,832	$ 69,748	$ 69.75	0.07
B (TS1100)	$ 6,684,378	$ 83,555	$ 83.55	0.08
C (LTO)	$ 7,984,510	$ 99,806	$ 99.81	0.10

	Year 1	Year 3
Media (PB)	50	80
Drive Bandwidth (GB/s)	7.5	10

`

Config	Sub Total	Silo Cost	Drive Cost	Maint. Cost 7x24 - 4 yrs	Media Cost	Cleaning Tapes Costs (100)	Install Cost	Shipping Misc Cost	Silo Cells
A - Y1	$ 4,512,232	$ 340,000	$ 768,000	$ 332,400	$ 3,000,000	$ 12,000	$ 19,944	$ 39,888	10,000
A - Y3	$ 1,067,600	$ -	$ 800,000	$ 240,000	$ -	$ 12,000	$ 4,800	$ 10,800	-
B - Y1	$ 5,566,478	$ 750,000	$ 868,000	$ 323,600	$ 3,500,000	$ 10,000	$ 30,742	$ 84,136	12,500
B - Y3	$ 1,117,900	$ -	$ 900,000	$ 180,000	$ -	$ 10,000	$ 14,400	$ 13,500	-
C - Y1	$ 4,496,043	$ 666,667	$ 385,000	$ 357,567	$ 3,000,000	$ 6,000	$ 70,083	$ 10,727	33,333
C - Y3	$ 3,488,466	$ -	$ 342,000	$ 116,280	$ 3,000,000	$ 6,000	$ 20,930	$ 3,256	-

	Drive Type	Number of Drives	Drive Transfer Rate	Drive Cost	Media Count	Media Size (TB)	Media Cost	Media ($/TB)	Silo Cell Cost
A - Y1	T10000C	32	240	$ 24,000	10,000	5	$ 300	$ 60.00	$ 34.00
A - Y3	T10000D	32	**330**	$ 25,000	-	**8**	$ 220	$ 27.50	$ 30.00
B - Y1	TS1140	31	250	$ 28,000	12,500	4	$ 280	$ 70.00	$ 60.00
B - Y3	TS1150	30	**350**	$ 30,000	-	**9**	$ 250	$ 27.78	$ 50.00
C - Y1	LTO5	55	140	$ 7,000	33,333	1.5	$ 90	$ 60.00	$ 20.00
C - Y3	LTO6	38	270	$ 9,000	25,000	3.2	$ 120	$ 37.50	$ 18.00

Goal: CY1 provide 50 PB Silo and Media capacity, and 7.5 GB/s of Tape Bandwidth

Goal: CY3 provide and additional 80 PB Silo and Media capacity, and 10 GB/s of Tape Band

Note: T10000D and TS1150 media size and bandwidth are estimates and will likely change when product is announced.

Assumes - Linear Scaling of Silo Cost, drive, maintenance, and install cost (Likley not be the case, but should be close enough for comparision purposes) Also Assumes Future Tape Drives available near the same time.

Note: Blue Cells are Input Cells

SC11
Seattle, WA
Connecting communities through HPC

Questions???

Running an Archive System

- One copy or two
- System Configuration
- Databases
- Backups
- Tape Pools
- Tape Pool Optimization
- User Issues
- Managing Resource Use
- Moving Forward to New Technology
- Media Destruction and/or Overwriting
- Online Resources

One Copy or Two

One copy

- Lower cost
- Greater potential to lose data
- Consider the total cost to lose/reproduce the data; both system & human.
- Need failure prediction SW (MLM) for single-copy, but this isn't perfect
- Tape recovery is often possible, but not cheap.
- Hardware maintenance means downtime in almost all cases.
- You may have copies of the data in disk caches, so addressing tape failures quickly may allow you to recover some data before it gets flushed from cache.

One Copy or Two

Two copies

- Two copies in the same building protects against media and drive problems, but is not good for disaster recovery.
- Two copies can improve uptime by 1%; even more if you're having persistent HW issues. One copy provides service while the other copy is offline. The disk cache has to handle the writes. Allows for more flexibility for maintenance and upgrades.
- You may have third, fourth, etc, copies of the data in disk caches, so addressing tape failures quickly may allow you to recover data before it gets flushed from cache.
- It's possible to make two copies of only some of the data, but it's not a clean process normally.
- Three tape copies are overkill; high marginal cost for very small marginal benefit. Many of the failures that will ruin two copies will hit three also.

System Configuration

- Just another system in many ways; OS configuration, monitoring, change control, documentation, etc.

- Specialized knowledge is needed and resources need to be allocated for this. Either you're paying a vendor or paying for in-house knowledge, hired or developed. You may need some custom tools, also.

- There may be dependencies or compatibility restrictions between the OS and HSM and related software. Upgrading a single component may be tricky. The information isn't always well-advertised. Vendors will want you to stick with known and tested combinations.

- Upgrade conservatively. It needs to work, period.

Databases

- Keep them clean; don’t ignore anomalies
- Your record retention needs to match your backup retention. (DMF)
- The database should be kept on different hardware (disks, controllers)than the journals.

Backups of the Archive

- Typically a filesystem dump or a database dump of some flavor.
- Backs up metadata and little if any data, since the data is on tape. Takes hours. It's practically impossible to backup the data on an archive.
- Run backups more than once a day if possible

Files "archived" to Mass Storage aren't fully archived until the data is written to tape and the metadata has been backed up. This clashes with user expectations.

- Backup validity tests

  Quick-n-dirty: expected size; "file" command output

  In-depth: Restore files from backup.

- Backup verification scripts have to give positive confirmation of success

A silent backup failure paired with a silent verification failure is a disaster waiting to happen. Our biggest data loss happened this way when RAID hardware failed.

Tape Pools

- Keep them clean – jump on MLM reports of questionable media
  Even two copies aren't safe if errors accumulate
- With two copies how many tapes need to be bad in each pool before the odds are that you lose data? A lot less than is intuitive.
- Tapes are not mirrors; data on a tape will be on several/many tapes in the second pool. A colliding failure of a tape in each pool will cause the loss of a only part of a tape.
- Tapes are written end-to-end, so partially empty tapes have to be read and rewritten to reclaim gaps created by deleted files. Regaining the lost space can take a year or more depending on how long you keep backups and how aggressively you re-write tapes.
- If tapes in both pools are from the same manufacturing tape batch, it increases the chances of a bad batch causing data loss. If the vendor will support it, get tapes from different batches for redundant pools. Ideally, you would have different technology for each copy, but it costs more and may be too much of a headache.

Tape Pool Optimization

- What are your goals?

  Minimize tape mounts when data is retrieved?

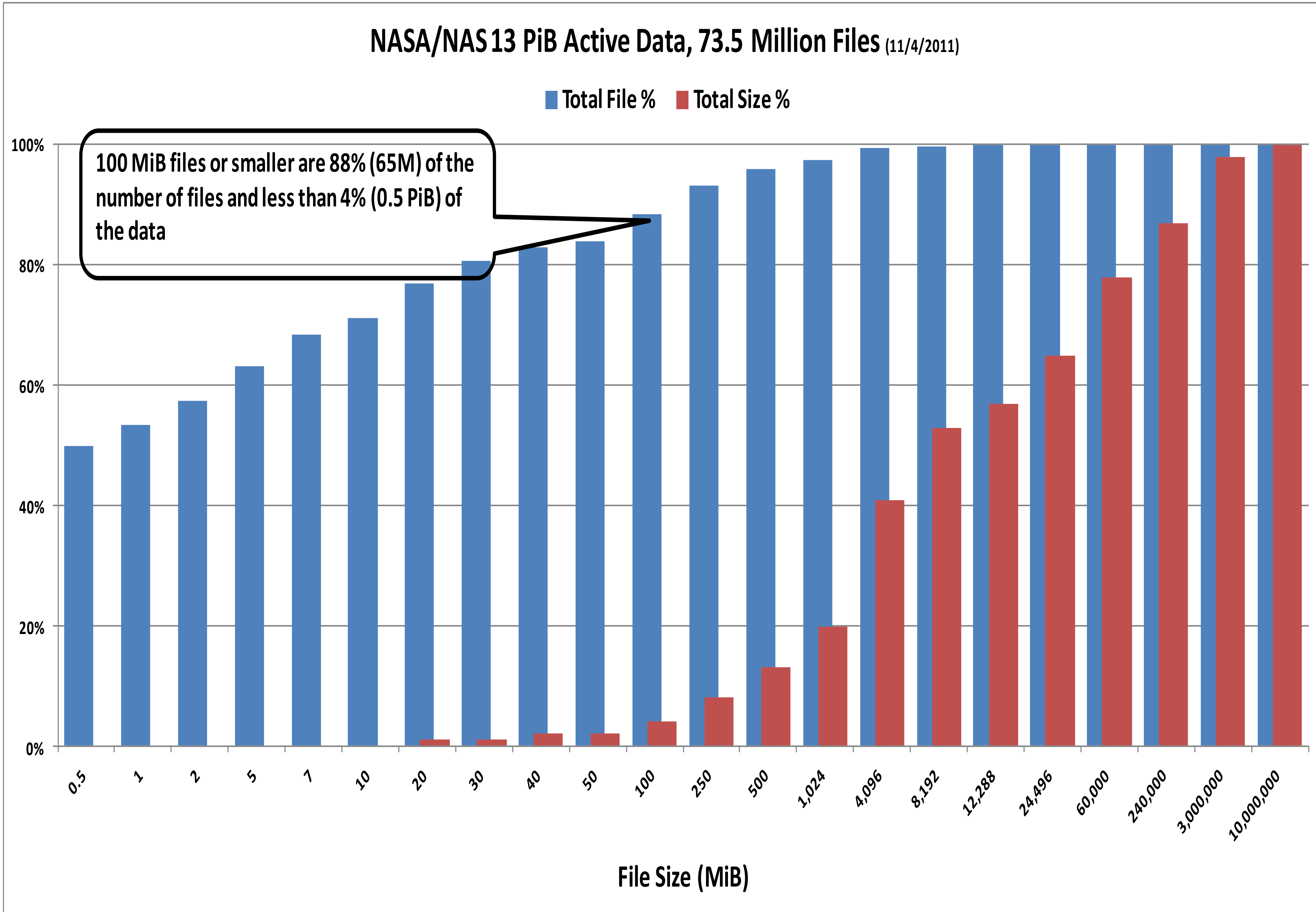
  Maximize recall bandwidth on large files?

- Have tape pools based on file size. We have a Small File pool and a Large File pool and adjust the dividing line as conditions change. We profile the filesystems and determine the split. Others have more than two pools.
- Custom pools – small pools for data that tends to spread out.

  Some users/applications dribble data daily (one user, 3000 tapes)

  When time allows consolidate. One way is to recall a set of files on many tapes and rewrite them to a single tape. Or monitor when a user recalls a lot of tapes and rewrite the tape data at that time.

- Files can be written to tape *and* kept on disk. This reduces tape mounts, but still keeps filesystem dumps small. We keep files < 5M. This will be increased when we get larger disks.



NASA/NAS 13 PiB Active Data, 73.5 Million Files (11/4/2011)
Total File %
Total Size %
100 MiB files or smaller are 88% (65M) of the number of files and less than 4% (0.5 PiB) of the data
100%
80%
60%
40%
20%
0%
0.5
1
2
5
7
10
20
30
40
50
100
250
500
1,024
4,096
8,192
12,288
24,496
60,000
240,000
3,000,000
10,000,000
File Size (MiB)

User Issues

- Monitor user activity for problems/inefficiency; they impact the system and other users (and themselves).
- User education is worth the time. It may seem endless, but you don't notice the problems you solved in advance. Also, a basic understanding of archive can reduce helpdesk requests. If users don't have a working understanding of archive, it may be unusable for them, so they request more scratch space and leave valuable data unbacked-up.
- HSMs can be transparent to users and opaque at the same time. It all just works, but the same operation can take five minutes or five hours based on load with no way for a user to know when they will get their data. We try to provide a look at the request queue. It's crude, but useful. Others are working on this, too.
- Running many instances of parallel file transfer utilities; too many I/O threads slow the disks down. We reduced the default number of streams for parallel file transfer.
- Retrieving too much information from tape at once. Occasionally, more is requested than the disk cache can hold. It thrashes the system and denies service to other users.

User Issues

- Compressing unnecessarily – software compression takes much longer than other operations – 5x or more. Many of our users do it to be good system citizens and don't realize that we get HW compression for free in the tape drives.
- Rsync can be very inefficient for transferring files *out* of the archive, since it will retrieve files inefficiently from tape. Use "rsync --dry-run" to create a list of files to retrieve from tape, then run the rsync.
- Creating unnecessarily large files may disrupt your disk cache and waste resources; a 10TB tar file when retrieved from tape and extracted causes:

  Retrieve data from tape; 10TB tape read + 10TB disk write

  Retrieve data from tar; 10TB disk read

  Extracted files written to tape; 10TB disk read + 10TB tape write
- There are utilities that can split a directory into several/many independent tar files; I've tested only split-tar, which seems to work well.
- Always make table-of-content files (tar –cvv or tar –tv for gtar) to reduce unnecessary retrieval of large tar files.

Managing Resource Use

- Space quotas – Not currently workable in the same way as disk quotas. Quota software may not understand files on tape. Users may not understand compression and charges for multiple copies. The FSF version of du has a –apparent-size option to report space, not disk use.
- For NASA inode counts and metadata are a bigger problem than space. It slows down backups and filesystem scans and bloats databases, so we have inode quotas, but not space quotas. Users consolidate data (tar) to keep the file count down. You should provide help/tools to make this easy and efficient.
- System load can be significant

Rolling forward to new tech every few years means a significant fraction of time is spent with high-volume reads and writes for this activity.

Rewriting tapes to reclaim space can also use significant resources, but this can be trickled.

- Data retention: Focus on the big stuff. Small data isn't worth significant time to clean up. 50GB was an entire tape 10 years ago; 30-100x fits on a tape now.

Typical problems

- Software – this will be the cause of the majority of loss of service. HSM SW focuses on data integrity, but HSMs do a lot of things, so don't expect 99.9% uptime. Expect 97-99% if you work at it.
- Hardware – keep good relationships with your vendors
- Have onsite spares. Avoid waiting for commonly-failed parts to ship

    RAID and system disks, tape drives, possibly even RAID controllers and tape robots.

- Tape failure – be prepared to make a second or third copy. Don't invalidate the "bad copy" until it's replicated.
- Tape drive failure – It needs to be replaced, but what happens when a tape is stuck in the bad drive? Don't let it go out the door with the drive.
- RAID/disk issue – that's what backups are for.
- Disk overload – SATA becomes imbalanced in favor of writes under mixed loads. Can you back up as fast as you can archive in that case?
- What happens when user activity clumps into the perfect storm of data archiving and retrieval plus tape reads and tape writes? Corrupted filesystem? We've had to kill users processes and email apologies.

Typical problems (cont)

- License limits and license expiration can sneak up on you. Some form of warning system is in order if your HSM has licenses.
- Shortage of cleaning tapes. Another sneaky problem. Drives will stop working if they need to be cleaned. Keep extras outside the libraries.
- Fabric issues

8Gb FC can be twitchy - dirty connectors and sub-optimal equipment will cause a variety of hard-to-diagnose fabric errors.

Sub-optimal components (HCAs, SFPs) can fail in non-obvious ways; a problem that seems to be an epidemic of random failures disappears when a single HCA is replace.

Use caps for fibres not in use and plugs for ports not in use.

Clean connectors – there are tools for cleaning fibre ends and for taking pictures of fibre ends. Are they arriving dirty?

Minimize length and the number of patches connections within reason.

Security

- An archive system is a "high-value target", so you need good system security. Our archive is within an "enclave" that requires two-factor authentication. The administration ports are on a private network behind another two-factor bastion.
- Tapes are more vulnerable to "walking away" than heavy disk systems. Normally, the data is considered hard to use if stolen, but could be useful for LTFS tapes or if a particular application's data is written to a custom pool.
- Consider encrypting copies where physical access control is weak.

  We are considering encrypting our remote copy and not the local one.
- Data tapes stolen is data lost for single copy.

Disaster Recovery

- The design depends on the resources available; perfection won't happen.
- As much as possible, test your disaster plans under the same conditions you would use them in the event of a real disaster.
- Over-the-network; for data with long-term value understand that encryption is temporary, i.e. protection on the scale of several years.
- Key management; loss of encryption keys is a potential disaster.
- The ideal case is to maximize geographic separation between the two data copies and have full metadata in both locations.
- If you have two sites or systems, they back each other up.
- At the NAS have an 8Gb FC fabric that extends between two buildings 1k apart. Not ideal, but 14PB is tough to replicate over a WAN. Backups of our meta-data go encrypted to the East Coast.

Moving forward to new technology

- Has to be done every 2-6 years.
- When to roll data forward to new technology –

  trading slots for media

  need for speed and/or capacity
- Write copies to the old and new as long as possible. Switchover to writing data only on the new technology is a dangerous time. We lost data when the new technology failed early on both sides at the same time in a non-obvious way.
- Do lots of testing (weeks), especially for newer technology. Don't be quick to believe the vendor. Some tests we run on new or eval drives/tapes

  Many thousand write / read / checksum

  Write to entire tapes with checks for speed

  As many mount/unmounts as we can manage
- When the old and new tapes are incompatible, this can take considerable effort, on the order of several person-months. If they are compatible, it can be much easier, but isn't free.

Media destruction and/or overwriting

- Old media likely needs to be destroyed.

  We use a vendor from the National Association of Information Destruction (NAID). See www.naidonline.org.

  Get several quotes. Prices varied by 3x in our quotes.

  Usually done as a small fixed cost and then per pound.

  You may want/need certificates of destruction.

  If they are using a giant shredder, you may be able to destroy disks and papers at the same time.

- Can't degauss and reuse anymore. Degaussing erases low-level formatting tracks. Can be used to destroy data, but takes very high magnetic fields for modern tapes. Lots of verification needed.
- Overwriting tapes can be done, but takes a looong time. Most easily done as part of the rolling forward process, but requires more tape drives.
- We were able to find another group in our NASA data security domain and give them $1.5M in tapes.
- You might be able to sell them, but you won't get much.

Questions???

Products

- DiskXtender –
  - http://www.emc.com/products/detail/software/diskxtender-unix-linux.htm
- High Performance Storage System (HPSS)
  - HPSS User Forum - http://hpss-collaboration.org/
  - HPSS Online Docs - http://hpss-collaboration.org/online_doc.shtml
- Parallel Data Migration Facility
  - SGI User Group - http://sgiug.org/
  - DMF Online Docs - http://techpubs.sgi.com/library/tpl/cgi-bin/browse.cgi?cmd=toc&coll=linux&db=bks&pth=/SGI_Admin/DMFx_AG
- Storage Archive Manager (SAM-QFS)
  - SAM-QFS User Group – Jan 10, 2012 at University of Texas, Austin
  - SAM QFS Online Docs - http://wikis.sun.com/display/SAMQFSDocs

Products

- StorNext
  - http://www.quantum.com/products/software/stornext/index.aspx
- StorHouse
  - http://www.filetek.com/products/storhouse
- Qstar HSM Archive Manager
  - http://www.qstar.com/products/qstar-software-products/hsm/
- XenData (Video archive)
  - XenData Digital Archives - http://xendata.com/products09/products.html

Expanded Reading

- FSBP - http://outreach.scidac.gov/fsbp Archive Best Practices
- LTO - http://lto.org Linear Tape-Open
- LTUG – http://ltug.oracle.ioug.org Large Tape User Group
- Hard Error - http://encyclopedia2.thefreedictionary.com/hard+error
- MAID - http://en.wikipedia.org/wiki/Massive_array_of_idle_disks
- PASIG – http://sun-pasig.ning.com – Preservation and Archiving Special Interest Group
- RAID - http://en.wikipedia.org/wiki/RAID
- RAIT - http://www4.clearlake.ibm.com/documents/HPSS_RAIT_Architecture.pdf
- T10000C - http://www.oracle.com/us/products/servers-storage/storage/tape-storage/t10000c-tape-drive-292151.html
- TS1140 - http://www-03.ibm.com/systems/storage/tape/ts1140/

Special Thanks

- DDN – John Josephakis
- IBM – Bill Kneisly and team
- Instrumental – Henry Newman
- SGI - Jennifer Fung and Dale Sutton
- Spectra Logic – Brian Grainger

SC11
Seattle, WA
Connecting communities through HPC

The End

Backup Slides

National Oceanic and Atmospheric Administration

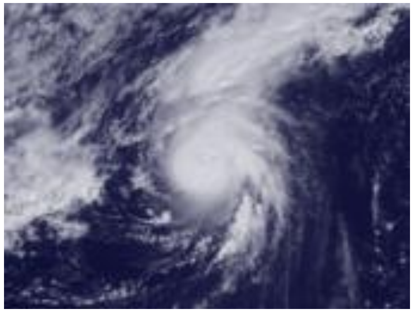
- Scientific agency within US Dept. of Commerce
- Formed in 1970 from three existing agency
  - US Coast and Geodetic Survey (1807)
  - Weather Bureau (1870)
  - Bureau of Commercial Fisheries (1871)
- 2011 ~$5.6 Billion Budget
- Strategic Vision
  - Inform society with a comprehensive understanding 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.

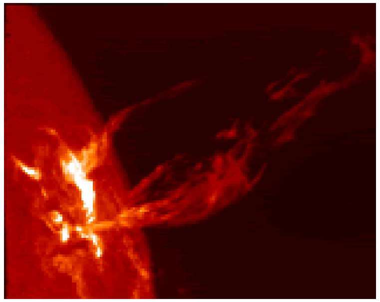
National Oceanic and Atmospheric Administration

- NOAA has number of line and staff offices.
- National Environmental Satellite, Data and Information Services (NESDIS)
- National Marine Fisheries Services (NMFS)
- National Ocean Service (NOS)
- National Weather Service (NWS)
- Office of Oceanic and Atmospheric Research (OAR)
- Office of Program and Planning and Integration (PPI)
- Office of the Federal Coordinator for Meteorology
- Office of Marine and Aviation Operations
- NOAA Central Library

National Oceanic and Atmospheric Administration

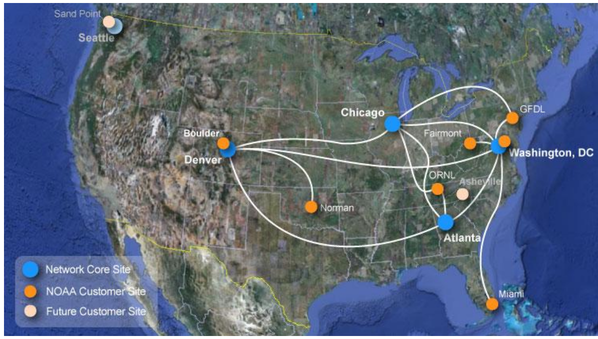
- NOAA has offices in every state
  - Employees ~17,000
  - 8 Ships (155-274’ research vessels)
    http://www.noaa.gov/deepwaterhorizon/platforms/ships_index.html
  - 5 Planes “Hurricane Hunter”
    http://www.noaa.gov/deepwaterhorizon/platforms/aircraft_index.html
  - Satellites: severe weather, snow, tropical storms, hurricanes, sea level, …
    (saved ~ 300 people in 2010 in US)
    http://www.nesdis.noaa.gov/SatInformation.html



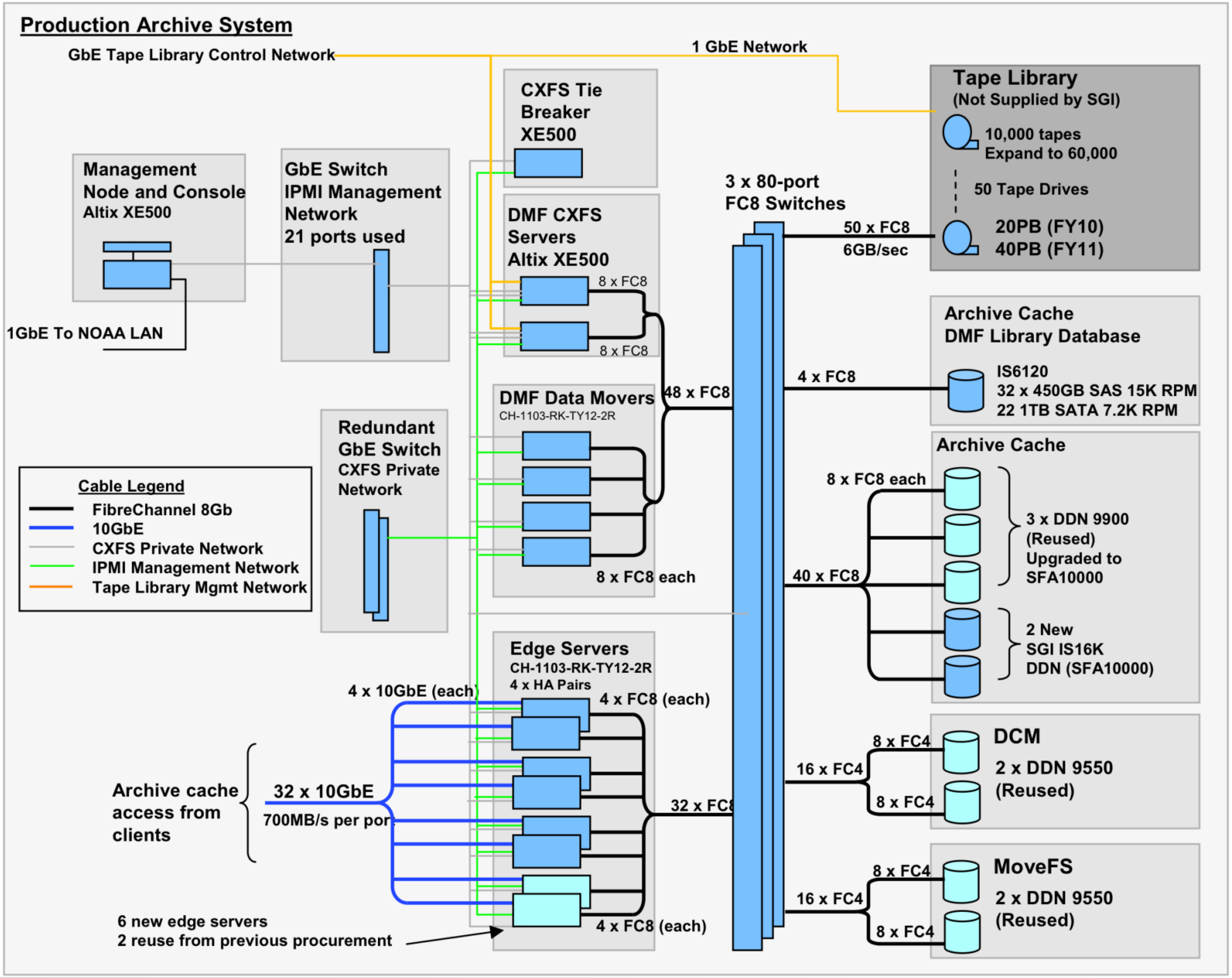


National Oceanic and Atmospheric Administration HPC Program

- Four separate contracts
  - WAN/Nwave
  - Support & Systems
  - Facility
  - ORNL/Compute
- Distributed Sites (one set of agency requirements)
  - GFDL
  - NCEP
  - ESRL
  - NESCC
- New facility building with shell/slab
  - Fairmont, WV - NESCC
- Different Federal Agency affiliations (DOE/ORNL)
- Nation-wide distributed CSC staff and customers
- Limited onsite staff to Implement a Large Scale Compute System

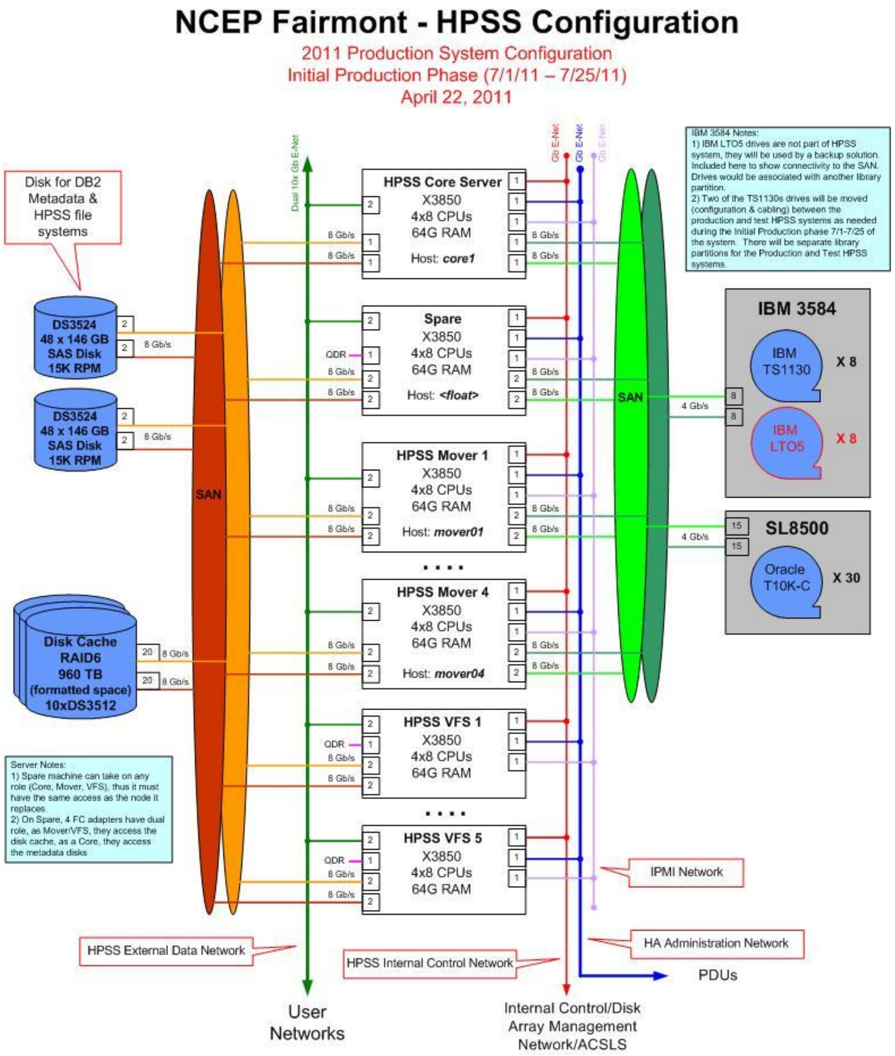


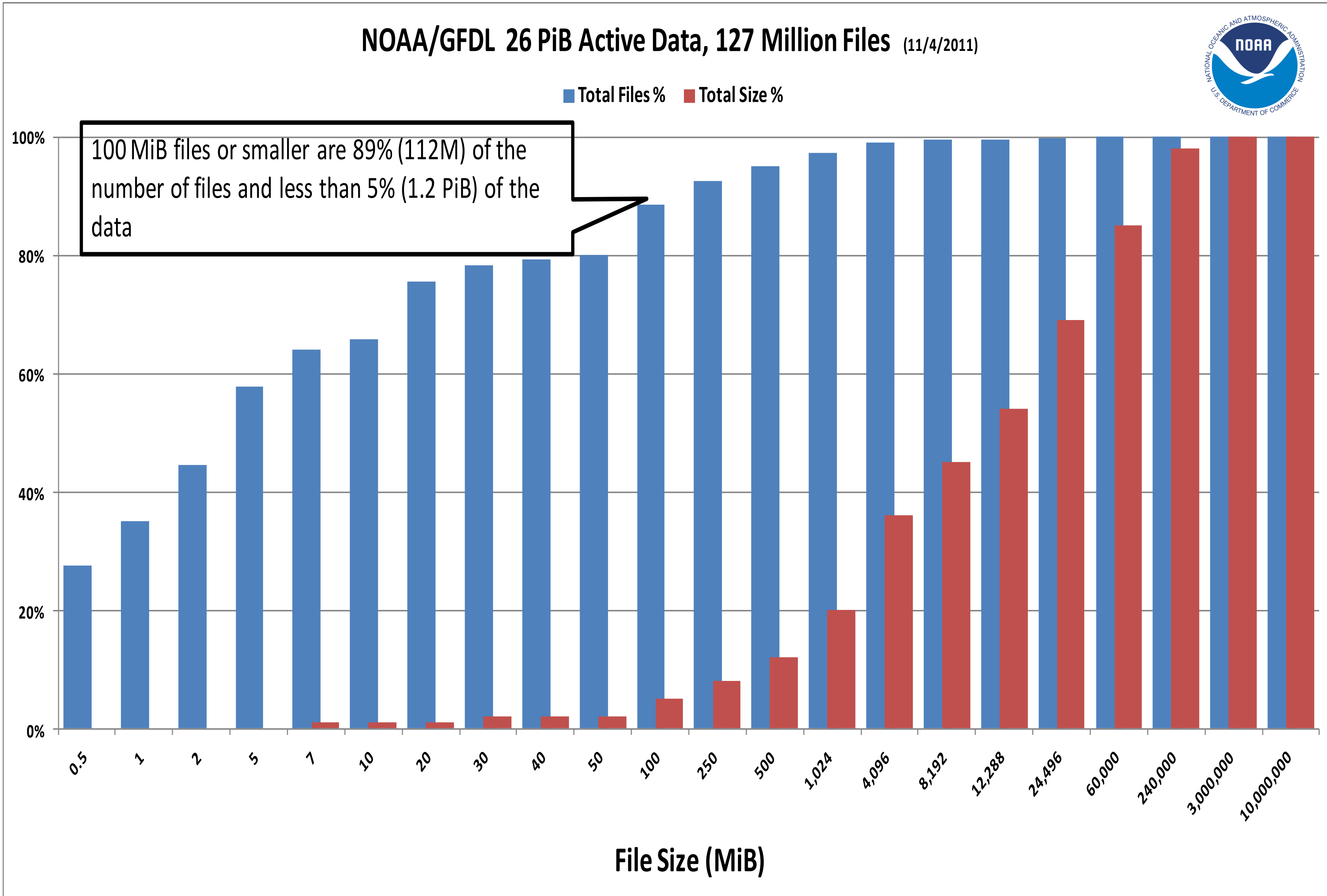
NOAA HPC Archive (GFDL)



SC11 Seattle, WA

NOAA HPC Archive (NESCC)




NOAA/GFDL 26 PiB Active Data, 127 Million Files (11/4/2011)
Total Files %
Total Size %
100 MiB files or smaller are 89% (112M) of the number of files and less than 5% (1.2 PiB) of the data
100%
80%
60%
40%
20%
0%
0.5
1
2
5
7
10
20
30
40
50
100
250
500
1,024
4,096
8,192
12,288
24,496
60,000
240,000
3,000,000
10,000,000
File Size (MiB)