

# Archival Storage At LANL Past, Present and Future

Danny Cook

Los Alamos National Laboratory

[dpc@lanl.gov](mailto:dpc@lanl.gov)

Salishan Conference on High Performance Computing

April 24-27 2006

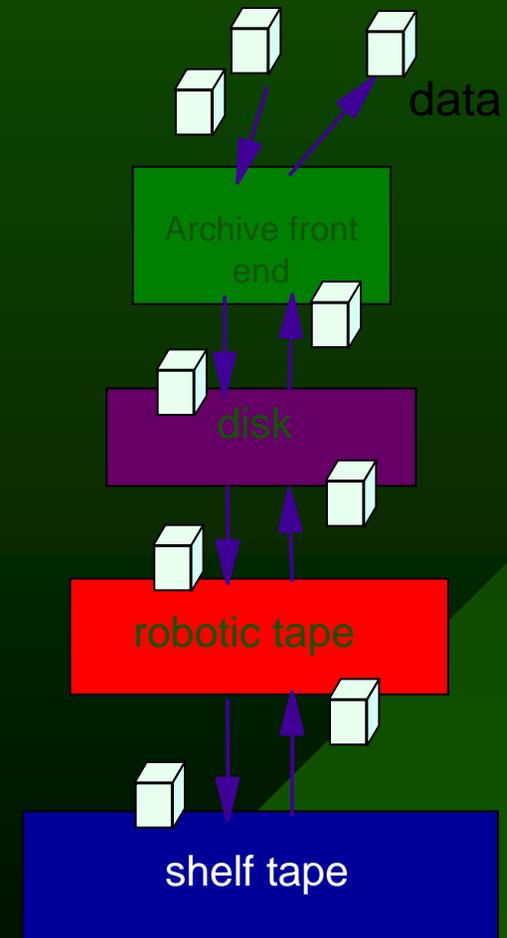
LA-UR-06-0977

# Main points of presentation

- Brief history of archival storage at LANL.
- More detailed history of HPSS with focus on how HPSS has been used at LANL
- Discuss LANL future in archival storage.

# HSM: Hierarchical storage management

- Purposes of HSM:
  - Extend file system space
  - Back up disk files to tape
  - Manage permanent archive
- Can be integrated with file system or usage of an archival utility.
  - Data migrates “down” the hierarchy
  - Migrated files may be asynchronously purged from higher level (e.g. disk) to free up space
- Multiple classes of service in a single name space, for example
  - Disk to tape
  - Tape only
  - High speed disk to low-cost disk to MAID to tape library to shelf tape



# Archives Provide

- Machine/file system agnostic storage solution
- *The most cost effective storage*
- Long-term data stewardship
  - Protection of billions of dollars of data investment
  - *Outlives vendors, machines, operating systems, file systems*
  - Protection from platform disasters (software or hardware)
  - Repack and data rescue tools
  - Multiple copies
- Risk-averse solutions not tied to “latest” changes (OS releases) on platforms
- Scales larger than most file systems - #files, directories, file sizes
- Intelligent resource usage/data placement
  - Classes Of Service,
  - Stage/migrate/purge
- Robotic/atomic mounts of sequential media
- Access to devices that have long inherent delays

# Tape is much less expensive

- In LLNL environment tape is:
  - **6.7** times cheaper to purchase (including drives, robotics, movers and media).
  - **56.7** times cheaper than disk for yearly maintenance
  - **342** times cheaper than disk for electrical power
  - **342** times cheaper for cooling
  - **72** times cheaper net yearly upkeep

# Architectural Tradeoffs

- Removable media reuses I/O infrastructure
  - When disk is full the pipes connecting it go dark.
  - When a tape is full another tape uses its I/O infrastructure
- Tape can be used across multiple generations
  - Our media typically serves 2-3 generations of tape drives
- Removable media allows greater scaling freedom
  - One can more easily buy capacity or performance as needed with removable devices.

*But we can't forget that Disk is random access, fast and easily RAIDed.*

- LANL has been involved in archival storage since at least the late 1970s.
- Systems in use at LANL
  - IBM Photostore
  - Common File System(CFS)
  - High Performance Data System(HPDS)
  - High Performance Storage System(HPSS)

# Common File System

- Development began in 1979
- Deployed in 1981
- MVS/IBM mainframe based
- Primary archival storage system at LANL until 1997.
- All data converted to HPSS in 2001.

# Common File System

- HSM based on disk, IBM 3850 and various tape technologies.
- Network centered, non distributed.
- Highly reliable, robust. Unique features such as User Validation Lists (better than ACLs).
- Limitations in terms of scalability, performance, file size.
- At time of conversion to HPSS, 10 million files, app. 110 TB of data.

# High Performance Data System

- Developed, deployed 1991-1992
- Unix based, distributed model running on workstation class machines.
- Primary user was Connection Machine.
- All data converted to HPSS in 1997.
- Similar work at NSL(NSL Unitree) at LLNL. Some of the first efforts to separate control/data paths.

# High Performance Storage System

- Next generation HSM. Primary goals were performance and scalability.
- Collaboration formed to focus on a common effort between LLNL and LANL.
- Initial discussion began at Salishan in 1992.
- Collaboration formed between LANL, LLNL, Sandia, Oak Ridge and IBM Federal Systems in 1993.
- Design, development began in late 1993. First production system deployed at SDSC in 1996.
- First LANL production system in 1997.

# High Performance Storage System

- HSM, network based, distributed model
- Separation of data, control paths, parallel IO to tape and disk.
- Centralized metadata manager (initially Transarc SFS, now IBM DB2).
- Security and distributed model based on DCE. Latest version removes DCE. Replaced with GSS model over Kerberos, UNIX and LDAP. RPC HPSS developed.
- Scalability in name space (storage subsystems, DB2).

# High Performance Storage System

- Commercial offering from IBM Federal Systems in Houston.
- Number of sites world wide running HPSS.
- Sites include DOE labs, universities, weather centers, DOD facilities, foreign sites in France, England, Japan, Korea. Currently about 50 sites (classified and unclassified).

# HPSS experience at LANL

- Deployed in 1997.
- Scalability excellent. Has scaled from 3-5 TB growth per month in 1998 to over 300 TB month in 2006. Factor of 100. HPSS ingests in one week what was stored in CFS in 20 years.
- Total storage in secure is at 5.5 PB and app. 55 million files.
- Single file transfer rates have scaled from a max of 60MB/sec in 1998 to 160MB/sec in 2006. With new Titanium drives should reach 600MB/sec. + in 2006. Rates are based on 4 way tape stripes. Larger stripes would give higher rates.
- Currently, aggregate throughput limited by network. New 10GigE switches will relieve this. Prior testing has demonstrated capability to scale to network bandwidth.

# HPSS experience at LANL

- Very reliable, robust. Replacement of SFS with DB2 in HPSS 5.1 in 2004 has enhanced this.
- HPSS 6.2 (available now) will eliminate DCE. Puts HPSS in a strong position to go forward into the future.
- SFS database created problems on 3 occasions. Caused some downtime and a loss of a small number of files(100 or so) on one occasion. Since replaced with DB2 in release HPSS 5.1 in 2004.
- Microcode failure in one STK 9940B drive corrupted a small number of files.

# HPSS Experience at LANL

- STK RAIT project in 1999-2001 did not result in deliverable. Prototype implementation successful, but STK decided not sufficient market.
- STK SN6K with tape mirroring was not deployed due to performance issues.
- Future for RAIT is likely client side software striping.
- Currently, delayed secondary tape copies provide a form of mirroring.
- Have seen some performance issues on small files related to some DB2 locking issues. Resolved in HPSS 6.2

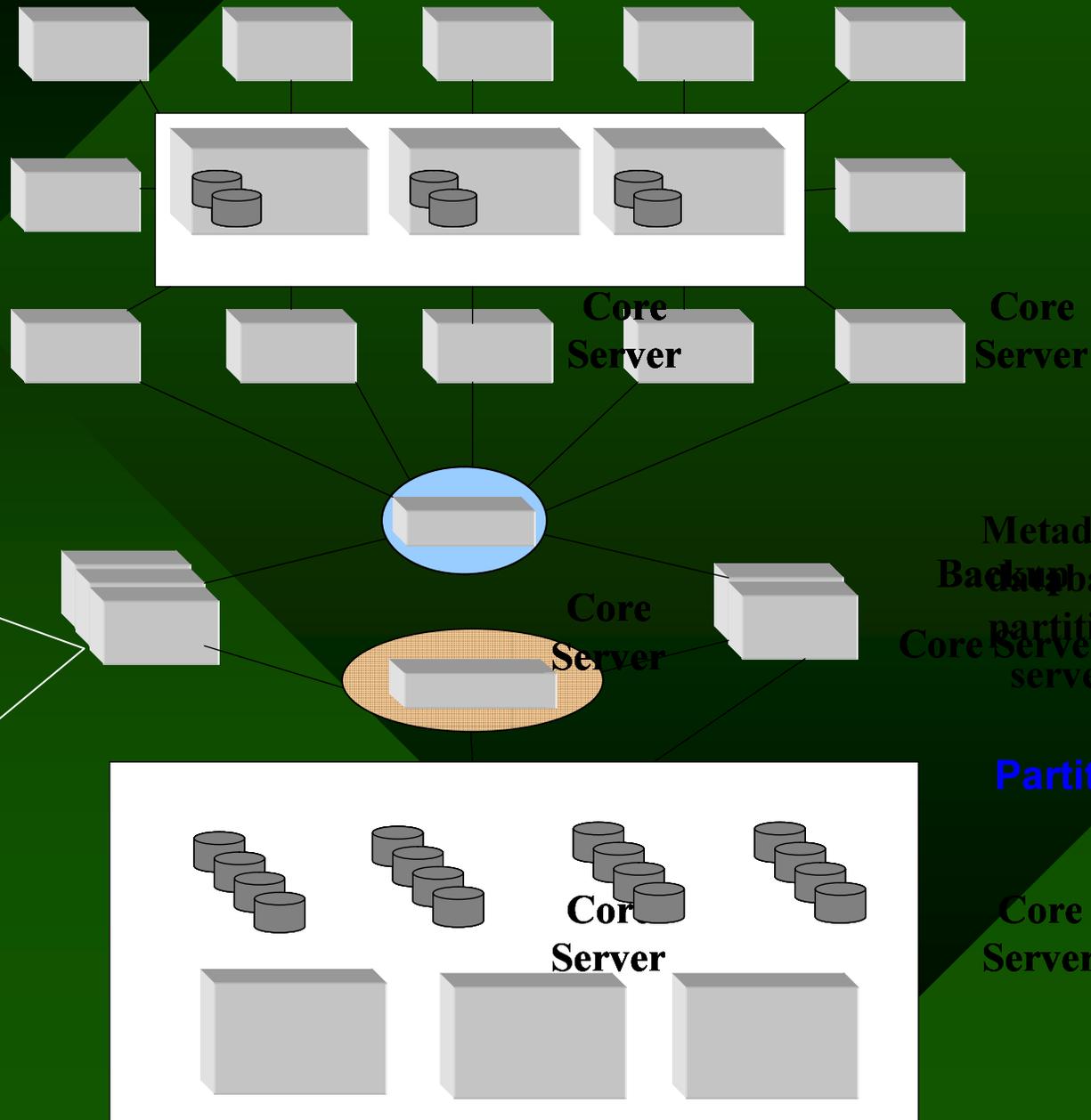
# HPSS Future at LANL

- Complete PSI/HTAR option. Currently in limited production by user. Semi-transparent bundling capabilities. Have demonstrated over 1 million small files/hour.
- Deploy HPSS 6.2
  - Completes infrastructure replacement
  - NFS/VFS Linux
  - GPFS/DMAPI/HPSS (LBL,SDSC)
- HPSS 7.1 development begins. (Summer or fall of 2007).
  - Increase small file performance with a target of a 5x increase. Initial investigations promising.
  - Add tape aggregation to increase migration performance Targeted part of small file performance enhancements.
  - Add client affinity to make the use of Local File Movers more usable.
  - Performance in other areas such as listing operations. Potential 10x improvement in listing performance.

# HPSS Future at LANL

- HPSS 8.1
  - Primary focus is another order of scalability in total throughput. Also oriented towards small files.
  - Multiple core servers to distribute name space load.
  - Use partitioned capabilities of DB2 to distribute and get parallel performance improvements in metadata operations. Believe DB2 provides advantages here.
  - Possible client side caching with lease type lock management.
  - 2009 time frame.
  - Initial discussions begun but at very high level.

# HPSS Architecture



- Clients select Core Servers based on current load
- File metadata is cached at the client and updated periodically, or when invalidated by a Core Server
- Lease based locking is used to provide metadata cache coherency, unix file open semantics and, unix file range locking semantics

# Future challenges for archival

- Petascale computing will present new challenges.
  - Will require small file aggregate archival rates to be in the 1000 files/sec. + range.
  - Amount of data archived will grow by at least an order of magnitude. Has media cost implications. LANL currently spends about \$70,000 month on tape media.
  - Will require advancements in both archival software and hardware.

# Future challenges for archival

- Projections for tape drive capacity and performance.
  - STK currently has the T10K at 120B/sec. and cartridge capacity of 500GB.
  - Tape industry projects capacity doubling every 2 years (16TB by 2015). Speed will ramp up slower. Should approach 1GB/sec around 2015.
  - Sony projects a 32TB helical scan cartridge in the 2011 time frame.
  - Holographic storage now looking possible. Could compete with tape at some point.
  - **Issue: Tape drive capacity may not scale as fast as needed for petascale archival.**

# Future challenges for archival

- Software challenges
  - Small files. Need aggregate of 1000+ inserts/sec. Need tape aggregation. Plans for HPSS to do this.
  - Need a high degree of scalability in aggregate I/O bandwidth. HPSS has this.
  - Reliability/integrity still foremost requirement of an archive.
  - Possible demands for more transparency. Interest in integration with Panasas, Lustre and GPFS. HPSS integrates with GPFS with DMAPI.

# Future challenges for archival

- Alternatives for archival system
  - QSAM/Sam-FS primary competitor to HPSS. Some experience at one existing HPSS site(SDSC). Does not appear to be able to handle the same load as HPSS.
  - SGI DMF not as easily scaled as HPSS. Evaluated approximately 2 years ago.
  - Object based file system(Lustre)/commercial HSM. Current research project at U. Minnesota
  - HPSS

# Some conclusions

- HPSS will be the archival storage system at LANL for at least the next 4-5 years. PSI/HTAR helps significantly with small file problem.
- Modifications to HPSS in the areas of small file performance/tape aggregation/etc. can make it a viable candidate for a much longer future.
- Need to look at Lustre/HSM work at U. Minnesota in more detail. Currently a research project. Need to understand if it is a practical approach that might be an archival alternative at some point in the future.
- Need to take a look at the way archival systems are used. Petascale computing demands may require a more judicious use of archival storage due to storage costs.