



LANL HPSS Review

Hal Armstrong, CCN-7
Storage Team Leader

HPSS User Forum
Santa Fe, NM
May 4-6, 2004



LANL HPSS Open Configuration

- Primary server: IBM p630 4-way
- FTP server: IBM 44P-170
- 10-9940A & 5-9840A STK tape drives
- 534GBs IBM SSA disk
- 4-disk & 11-tape movers (IBM 44P-170s)
- Gig-E data transfer network (18 connections)



Current Open Statistics

- **Archive totals:**
 - ◆ **Users: 2,200**
 - ◆ **Files: 10M**
 - ◆ **TBs: 318**
 - ◆ **Cartridges: 6,200**
- **April totals:**
 - ◆ **Active users: 280**
 - ◆ **Accesses: 342K**
 - ◆ **TBs: 10**

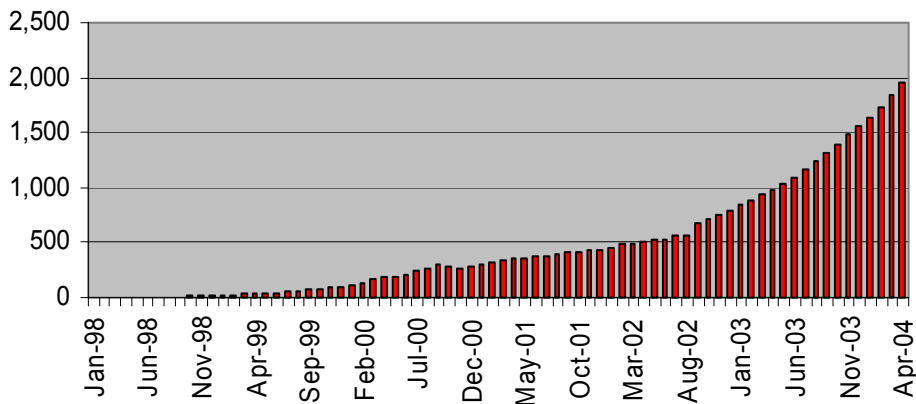
LANL HPSS Secure Configuration

- **Primary server: IBM p270 4-way**
 - ◆ **5.1 server: p650 8-way with 32GB memory**
- **90-9940B, 15-9940A, 12-9840A, STK tape drives**
- **13TBs IBM/SSA & STK/LSI disk**
- **8-disk & 48-tape movers (IBM p170s & p615s)**
- **Gig-E data transfer network (96 connections)**

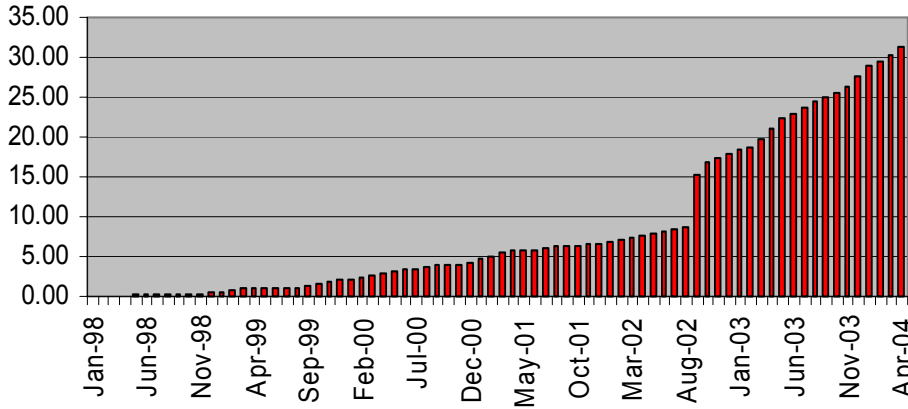
Current Secure Statistics

- **Archive totals:**
 - ◆ **Users: 1,300**
 - ◆ **Files: 31M**
 - ◆ **TBs: 2,000**
 - ◆ **Cartridges: 26,000**
- **April totals:**
 - ◆ **Active users: 340**
 - ◆ **Accesses: 1.3M**
 - ◆ **TBs: 102**

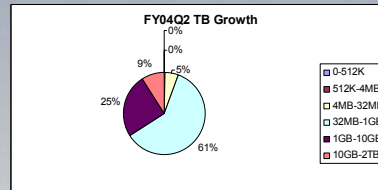
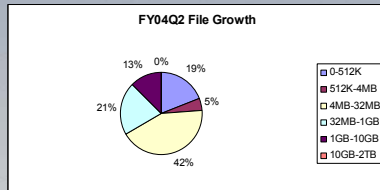
Secure HPSS Terabytes



Secure HPSS Files (Millions)



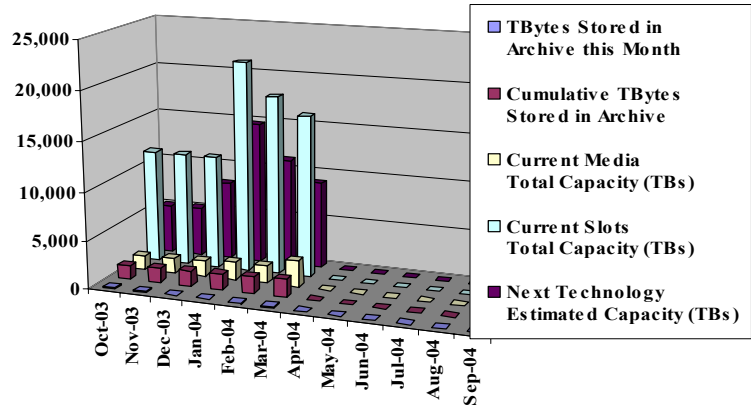
Secure HPSS Growth - Quarterly



• Quarterly growth → 2.7M files : 288TBs

- ◆ Small files (0-4MB) 24% : 0% TBs
- ◆ Medium files (4MB-1GB) 63% : 66% TBs
- ◆ Large files (1GB-2TB) 13% : 34% TBs

LANL Secure Archive



Combined Statistics

- **Archive totals:**
 - ◆ Users: 3,500
 - ◆ Files: 40M
 - ◆ TBs: 2,318
 - ◆ Cartridges: 32,200
- **April totals:**
 - ◆ Active users: 620
 - ◆ Accesses: 1.6M
 - ◆ TBs: 112

Archive Storage Performance ASC S&CS Milepost Results

File Size Target	File Size Actual	MB/sec Target	MB/sec Actual	Transfer Type
512KB	538KB	.25	1.1	D → D
4MB	6.5MB	2	8.1	D → D
32MB	32.9MB	16	32.9	D → D
990MB	906.8MB	32	34.4	D → D
1GB	1.2GB	40	52.5	D → T
20GB	92.2GB	150	179.7	D → T
400GB	137GB	600	549.5*	D → T
>20GB	1,062GB	800	1,041*	D → T

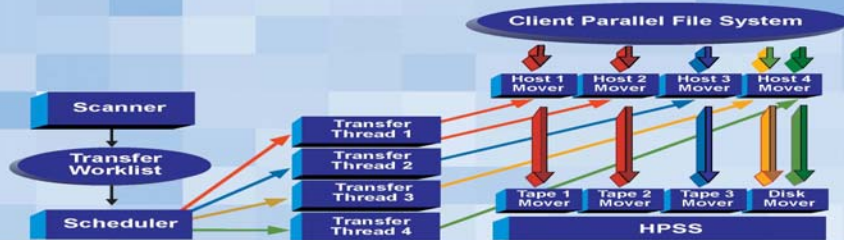
Future Performance

- Aggregate transfers to archive
 - ◆ > 1.5GB/s during 2004
 - ◆ Growing to 10s of GB/s by end of decade
- Single file transfers to archive > 600MB/s
- 2x(+) increase in meta data processing with DB2 and HPSS 5.1
- Metadata processing support for:
 - ◆ Single digit PB archives through 2005
 - ◆ Growing from 10s to over 100 PBs by 2010
 - ◆ Approaching 1 billion files by 2010

LANL's Multi-Node PSI Client

- Parallel Storage Interface (PSI): LANL command level user interface for archiving files
- PSI uses multiple nodes:
 - ◆To maximize bandwidth for any given file
 - ◆To decrease total time to archive for any mix of file sizes
 - ◆To minimize the time to archive for a given job
- PSI processes trees recursively
- PSI archives to disk, a single tape, or striped tape
- A one command example – `psi store file1,...,file4`
 - ◆Transfer to 2-way tape
 - ◆Transfer to 1-way tape
 - ◆Two transfers to archive disk

Archival Parallelism



Storing Files to High Performance Storage System (HPSS)

- 1 - Running PSI
User runs PSI to store a list of files
`psi store file1 ... fileN`
or to store a subtree of files
`psi store -R directory_name`
- 2 - Scanner
The list or sub-tree is scanned and files are added to the work list as they are encountered.
- 3 - Scheduler
 - While work is being placed on the work list, the scheduler dispatches transfer threads to transfer each file on the work list.
 - All optimization is automatic. Work is automatically assigned, using whatever resources (multiple client hosts, network connections, etc) are available to the job.
 - Large files are automatically transferred using multiple hosts.
 - Multiple files are transferred simultaneously.

4 - Example Transfers

Shown are the following simultaneous transfers: one transfer to "2-way" tape using 2 client hosts, one transfer to "1-way" tape using 1 client host, and two transfers to disk using 1 client host.

5 - Recent Test Example

Using 16 nodes of the "C" machine, the command
`psi store file1 file2...file12`
transferred 12 files at a time to 5TK 9940 tape drives: 1 file to "8-way" tape, 5 files to "4-way" tape, and 6 files to "2-way" tape. The peak transfer rate was 1041 MB/sec.

HPSS v5.1 Implementation Plan

- Capture v4.5 metadata statistics – 2hrs.
- Convert the metadata – 10 hrs.
- Tune DB2 & setup backups – 2 hrs.
- Run a verification – 24 hrs.
- Apply LANL local modifications and testing, testing, testing – up to 58 hrs. (12 hours)
- Target implementation date: May 15-18

Problems & Concerns

- HPSS 6.2 deployed by 12-31-2005
- Small file aggregation
 - ◆ MPS Support in 7.1?
 - ◆ Object based mover?
 - ◆ Something else?
- Redundant Array of Independent Tape (RAIT)
 - ◆ HW approaches will have limits
 - ◆ SW approaches let you dial in the width
- Slow WAN transfers (available in 6.1?)

Wish List

- **Small file aggregation**
 - ◆ Object based mover
 - ◆ Faster small file performance
- **RAIT striping – HW or SW**
- **Improved disk allocation algorithm**
- **Dynamic device add/delete**
- **Support for release dates**
- **Grid FTP with multi-node capability**

Questions?

Contact Information

Hal Armstrong
Los Alamos National Laboratory
(505)667-8426
hga@lanl.gov