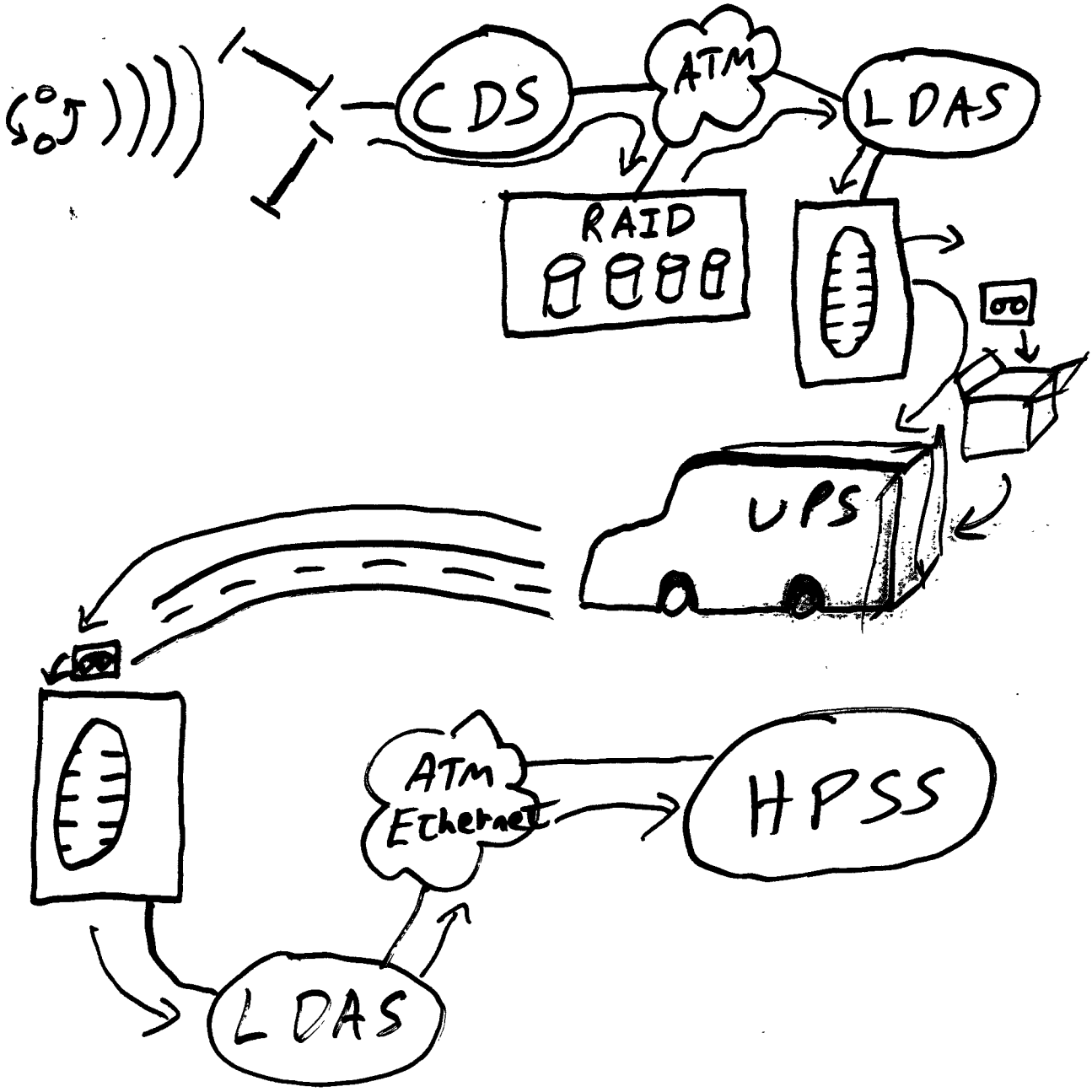
Data Storage Architecture

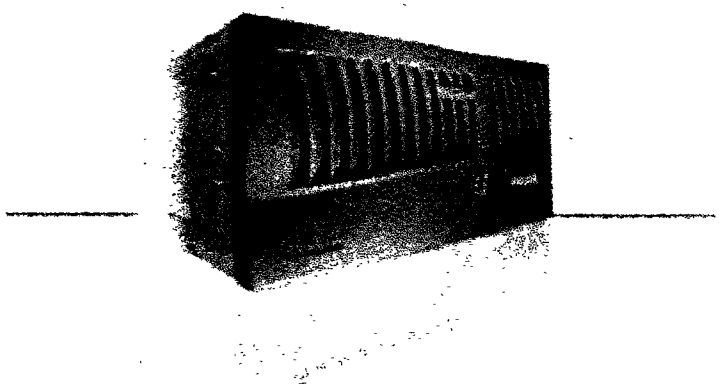
(→ dataflow)



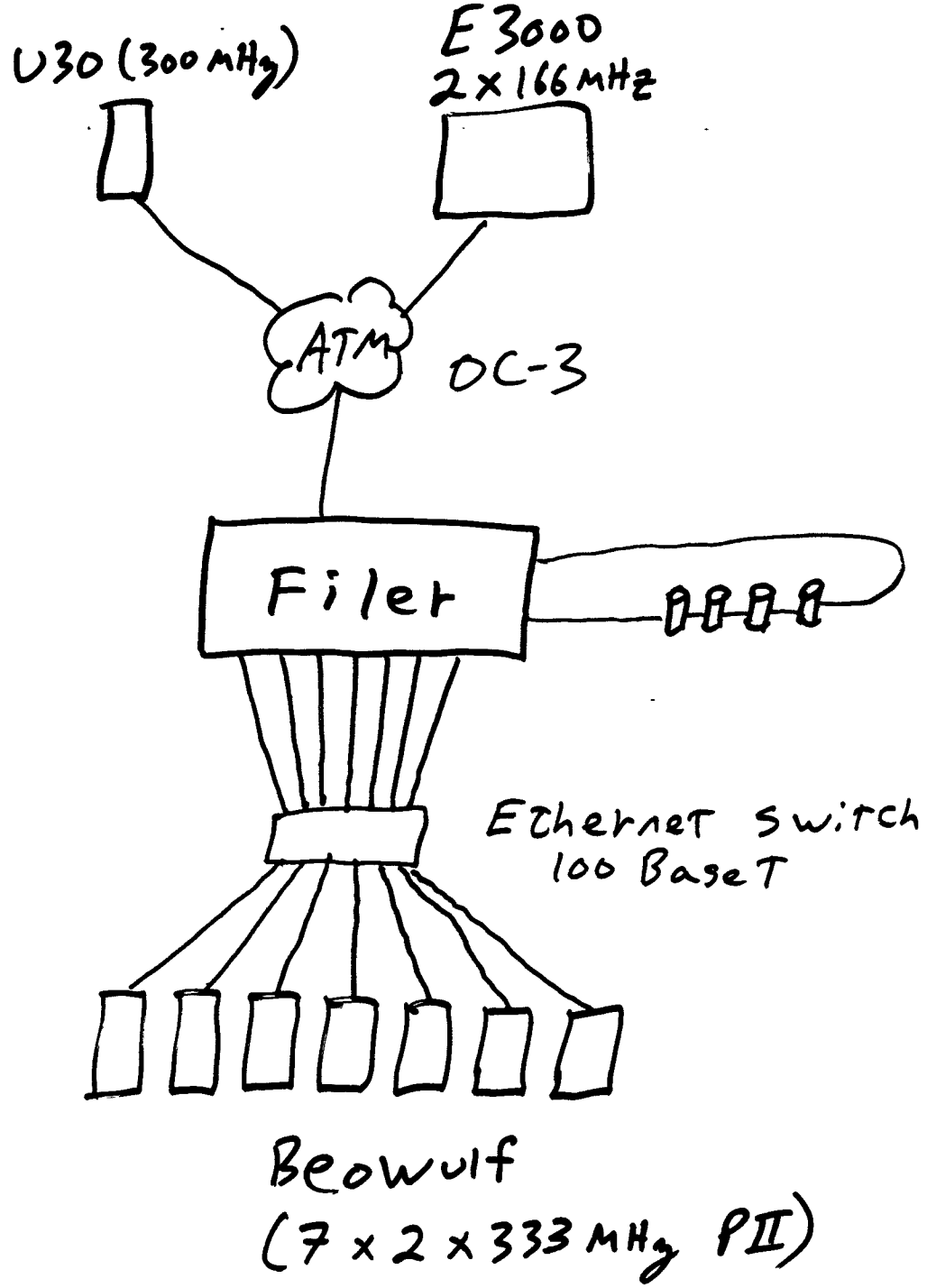
NetApp 760

LIGO is currently testing a direct network attached NFS file server from Network Appliance.

The F760 Filer is built around a 600MHz EV56 Alpha chip (21164) with 1GB of volatile read cache and 32MB NVRAM write cache. The demo unit was configured with 2 quad port fast Ethernet PCI adapters and 1 FORE/ PCA-200E OC-3 ATM adapter. Two shelves of 7x18GB Fibre Channel Seagate/ ST118202FC disk drives where attached to the Filer via a single FC-AL copper loop.



Network Topology (RAID Test)



Performance vs number of clients

Number of hosts (1 client/host)	Read (NFSv3)	
	CPU	kB/s
1	21	11200
2	35	19100
3	45	24000
4	53	27200
5	62	30400
6	66	31700
7	67	31900

This test was run with 1-7 Linux boxes each with its own NIC on the Filer and referencing its own files using dd bs=1024k (NFSv2=8k/NFSv3=32k).

LIGO specific frame file test

Status	CPU	Read (NFSv3)	Write (NFSv3)
		kB/s	kB/s
Nominal	78	23500	8100
Degraded	72	22000	6700
Reconstructing (speed 1)	65	22000	5300
Reconstructing (speed 4)	64	20000	1400

The NFSv3 writer was the U30 (100BT/ES3810/ASX-200BX/Filer) and the 4 NFSv3 readers where Linux boxes using 100BaseT. The files where 1MB frame files and this test made heavy use of the Filer cache.

Performance in degraded mode

Disks (data + parity)	Read (NFSv2)		Read (NFSv3)		Write (NFSv2)	
	CPU	kB/s	CPU	kB/s	CPU	kB/s
11(-1)+1	38	11200	60	20500	71	12200

This test was run with 7 Linux boxes using 7 different 100BasetT NIC's and referencing 7 different files using dd bs=1024k (NFSv2=8k/NFSv3=32k).

Performance while reconstructing

Reconstruct Speed (11+1)	CPU	NFSv3	NFSv2	Disk I/O	
		Read (kB/s)	Write (kB/s)	Read (kB/s)	Write (kB/s)
4	60	13000			2000
1	59	19000			200
9	61	13200			2000
4 (no NFS)	35			41000	3700
4	96		12000		
1	75		12200		
9	96		12000		

This test was run with 7 Linux boxes using 7 different 100BasetT NIC's and referencing 7 different files using dd bs=1024k (NFSv2=8k/NFSv3=32k).

Performance vs number of disks in raidset

Disks (data + parity)	Read (NFSv2)		Read (NFSv3)		Write (NFSv2)	
	CPU	kB/s	CPU	kB/s	CPU	kB/s
1+1	29	11200	28	14400	76	9940
2+1	29	11000	40	19500	78	11600
4+1	28	11200	67	31800	76	12200
8+1	29	11200	67	31900	73	12200
2x(5+1)	Volume with 2 5+1 raidsets identical to single 8+1 raidset					
8+1 (same file)			40	35000		
8+1 (same file/pre-staged in cache)			30	36000		
8+1 (same file/pre-staged in cache plus 1 Sun OC-3 client)			31	38200		

Except where noted this test was run with 7 Linux boxes using 7 different 100BasetT NIC's and referencing 7 different files using dd bs=1024k (NFSv2=8k/NFSv3=32k).

Network Appliance Issues

- Pro:
  - Platfor neutral–LDAS and/or CDS would be free to change compute platforms with no impact on the disk cache.
  - High availability–99.995%.
  - Network attached–not host computer to buy.
  - Low maintenance–no Unix operating System to configure, patch and maintain.
  - Scalability–can grow to a few TByte.
- Netural:
  - Performance–marginal today, expect 2-3X improvement in 6 months.
- Con:
  - Cost–the bare disks are more expensive than build your own.

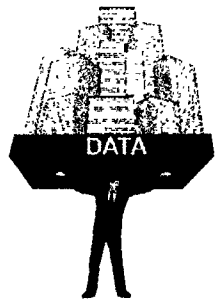
Disk Drives (June 1999)

	IDE (Dave)	SCSI (Fry's)		nStor	NetApp	IBM SSA
	low	low	high	18F	F740	
Single 18GB drive ($)	280	699	1089	1291	2500	
NFS host computer				~30k	0	~30k
~1TB system ($/GB)	16*	39*	61*	105	220	220
I/O (MB/s)				"fast"	55	
* Just the bare drives.						
Albert quote, "need 225 $/GB".						
LDAS doc 3/10/99 has 500GB at 500 $/GB.						
Clarion similar to nStore but twice the cost (has redundant controller and remote admin tools).						

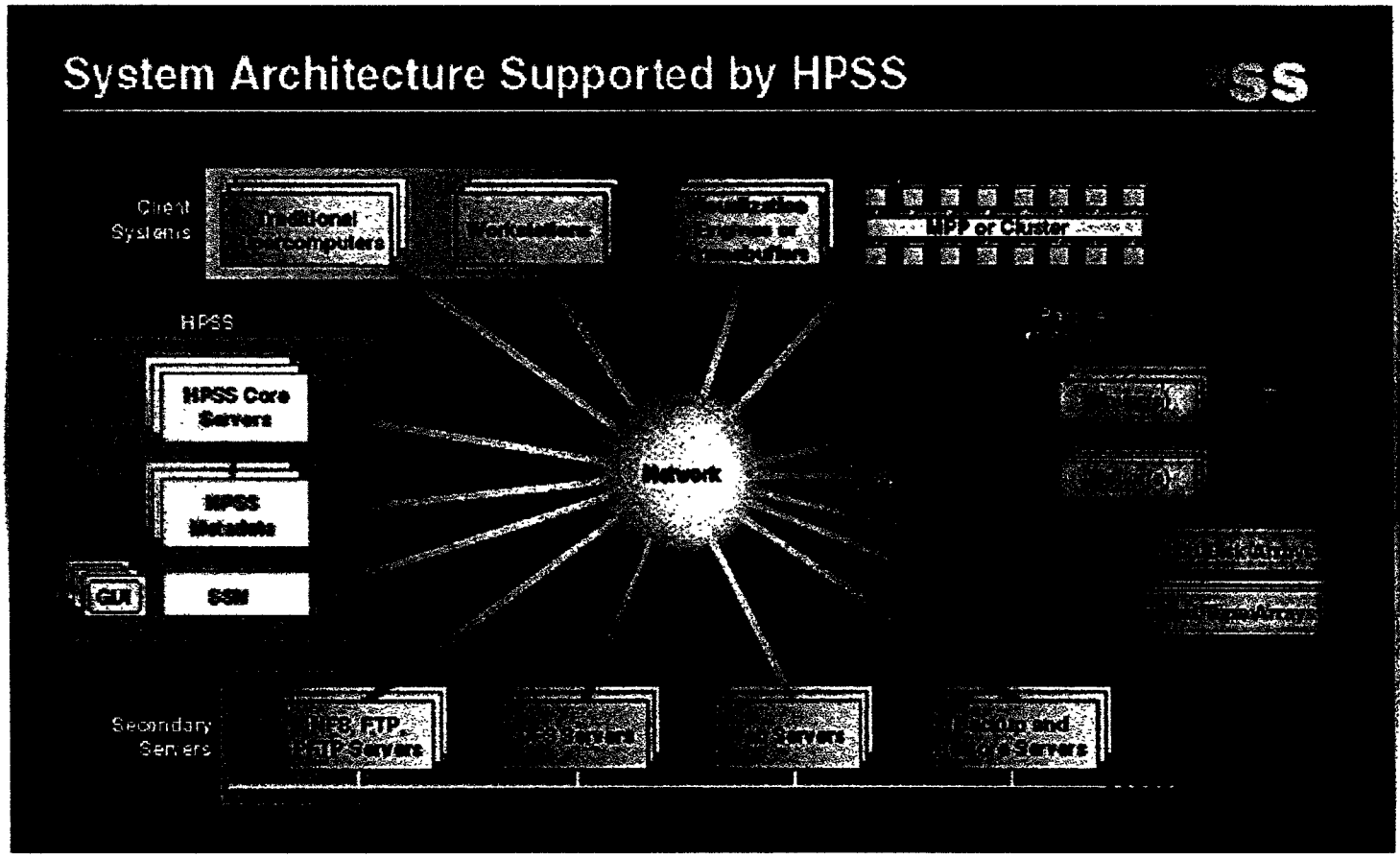
Stuart Anderson
sba@srl.caltech.edu (PGP Public Key)

Introduction

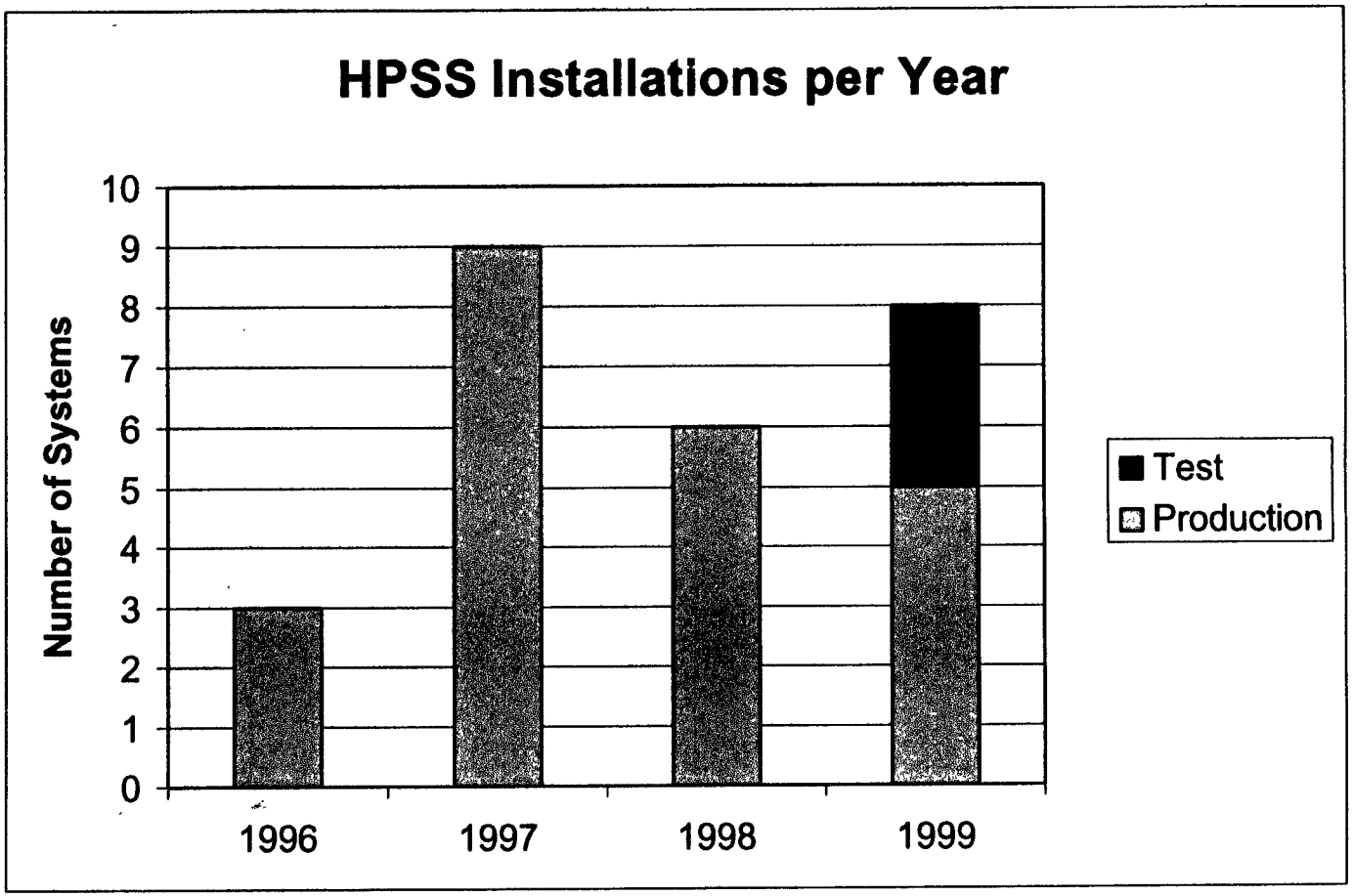
- What is HPSS?
  - Storage software
  - A Hierarchical Storage Manager (HSM)
  - A scalable, parallel storage system
  - A network-centric architecture which supports direct and network attached devices
  - An architecture driven by the IEEE Mass Storage Reference Model
- Why HPSS?
  - Requirements for fast per file and aggregate data transfers
  - Requirements for large data stores
  - Requirements for large single logical name spaces
  - Requirements for distributed, secure access to data



- Availability
  - Released on 7/1/96
  - Available from IBM Worldwide Government Industry as a service offering


System Architecture Supported by HPSS
Client Systems
Traditional Computers
HPSS
HPSS Core Servers
GUI
SSM
Network
Secondary Servers

Number of HPSS Installations



Note: 1999 data is "year to date".

CENTER FOR ADVANCED COMPUTING RESEARCH

Hardware Configuration

- 5 Wide Silver nodes
  - 4 processors / 1 GB memory each
  - 1 node acting as metadata server
  - 4 nodes with both disk and tape movers
- 1 8-way 604 High node for user access
  - Runs FTP daemon
  - "safe" sandbox to contain users
- IBM High Performance Gateway Node
  - 4 SP-Switch
  - 4 HIPPI
  - 2 OC-12 ATM

CALIFORNIA INSTITUTE OF TECHNOLOGY

CENTER FOR ADVANCED COMPUTING RESEARCH

Hardware Configuration II

- Disk Cache
  - 22 18.2 GB SSA drives
  - 32 4.5 GB SSA drives
  - Using SSA Raid adapter with Fast Write Cache
- Tape systems
  - IBM 3494 robot
    - 6 3590 drives
    - Drives ok, robot is a major headache
  - STK 4410 silo
    - 4 Redwoods
    - Silo is great, standard Redwood headaches

CALIFORNIA INSTITUTE OF TECHNOLOGY

Usage and Capacity

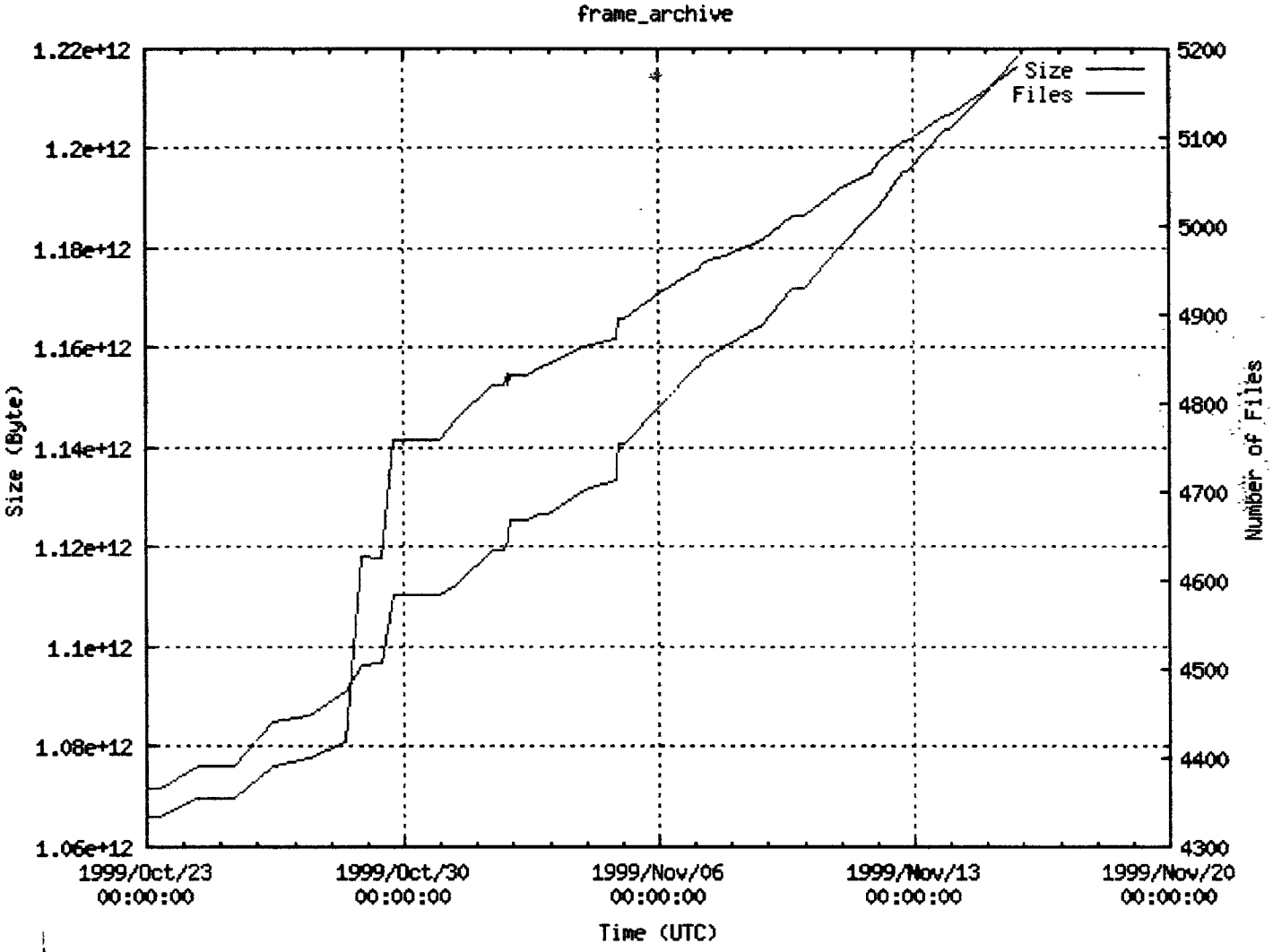
- 9.5 million files, 109 TB
- 500 TB total capacity
- Average of 500 GB/day stored/retrieved
- Peak of 2TB in one day

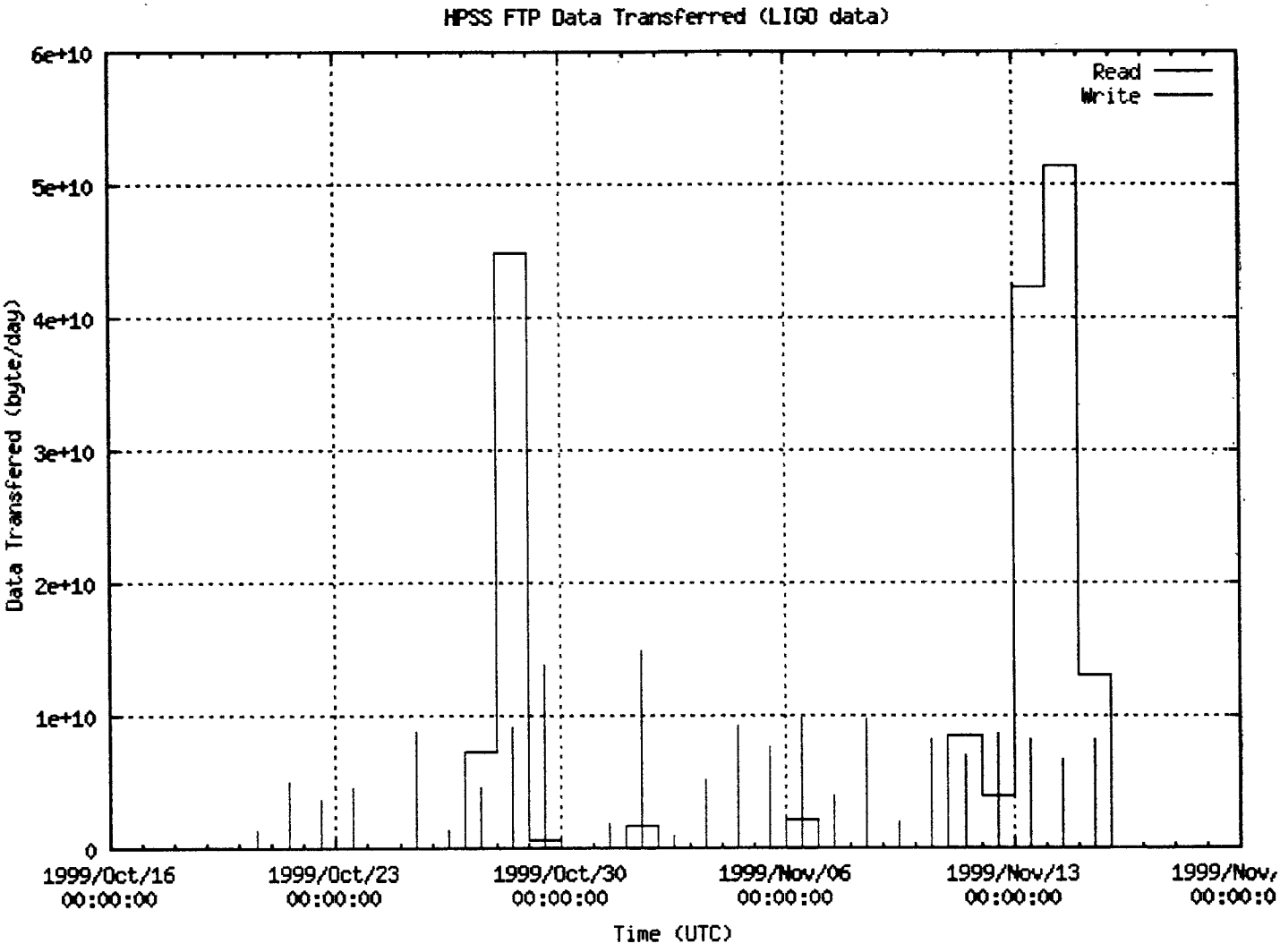
LIGO HPSS Frame Archive

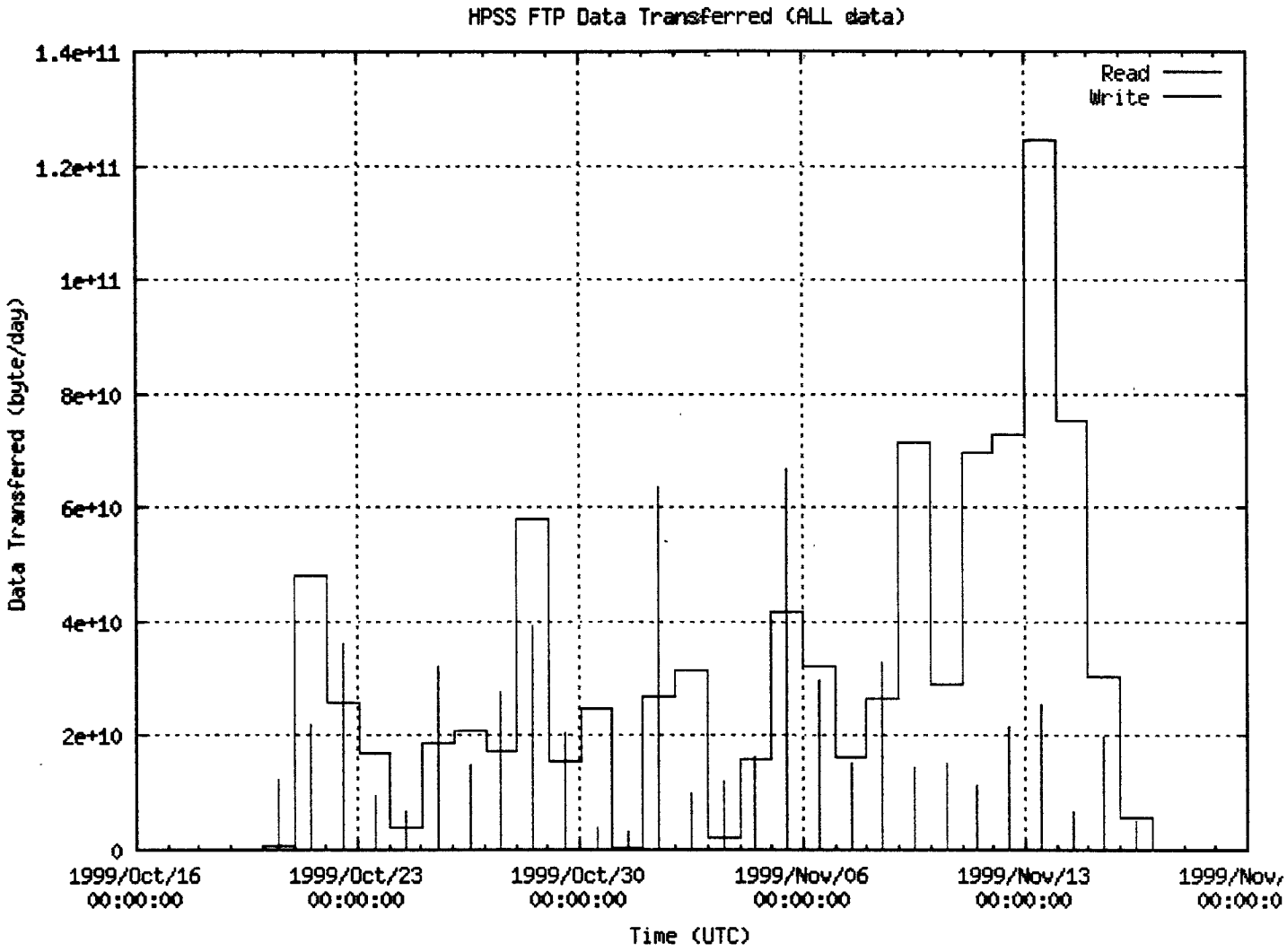
hpss.cacr.caltech.edu:/home/ligo/frame_archive

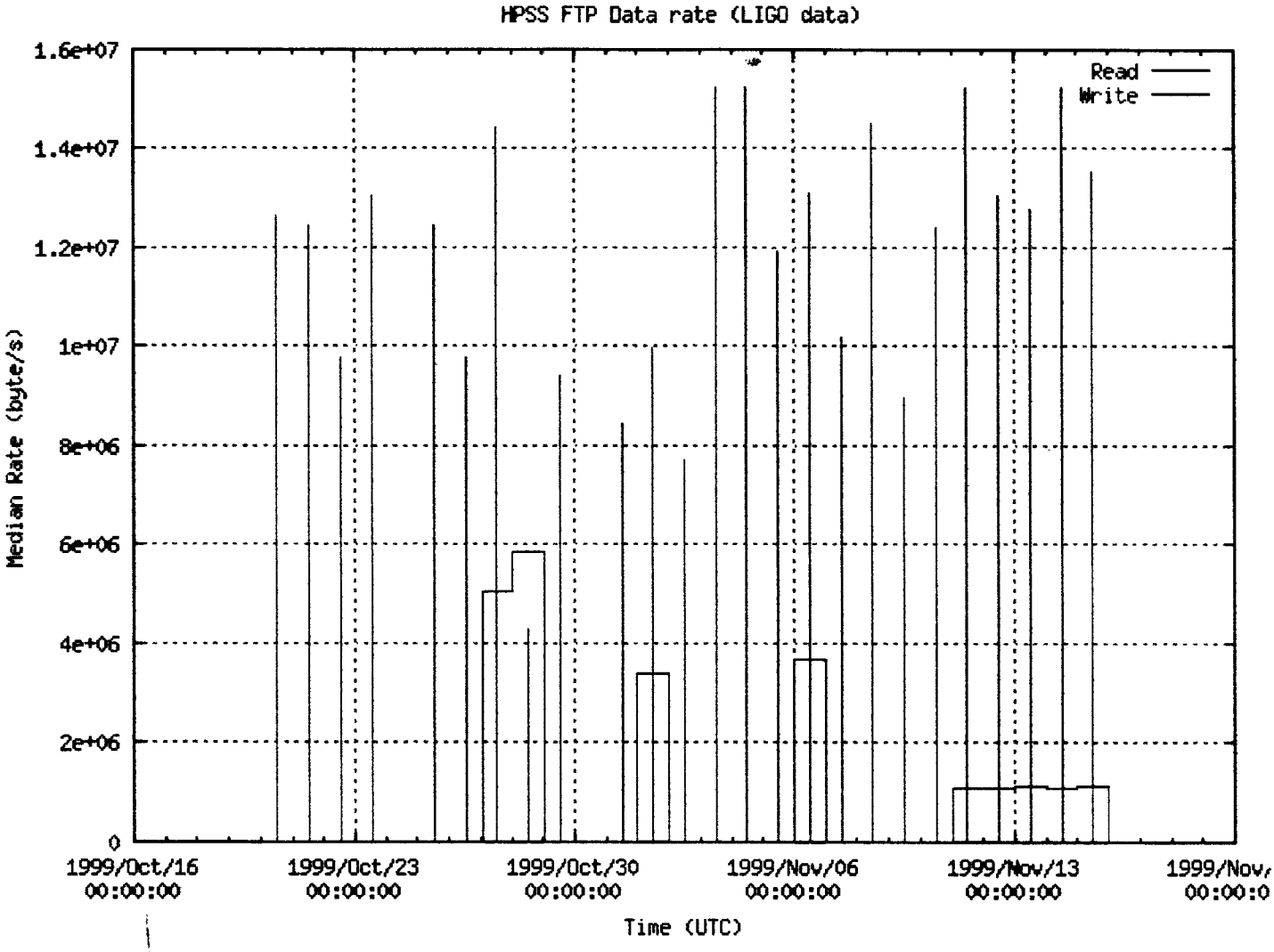
DIRECTORY TREE						ACCOUNT ACCESS							
IFO/	Frame Type/	Dataset	Group	Size (Bytes)	Files	ligo	ligo_lab	lsc_geo	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
40m	frames	1996.hrs2.27	lsc_1	12,799,724,416	1,050	R/W	R	-	-	-	-	R	-
		1996.hrs2.22	lsc_1	6,599,175,556	1,050	R/W	R	-	-	-	-	R	-
		1997.labsuite	ligo	38,127,665,802	323	R/W	R	R	R	R	R	R	R
		1999.oct	lsc_4	363,798,534,243	876	R/W	R	-	-	-	R	-	-
		1999.sum	lsc_3	331,525,724,402	510	R/W	R	-	-	-	R	-	-
		archive	lsc_9	12,334,425,712	25	R/W	R	-	-	-	-	-	-
		total		805,935,820,115	3,792								
	minute-trend-frames	archive	lsc_6	84,174,104	49	R/W	R	-	-	-	-	-	-
	trend-frames	archive	lsc_6	3,939,335,232	200	R/W	R	-	-	-	-	-	-
	Total			809,959,330,451	4,041								
LHO	frames	archive	lsc_2	70,414,672,820	256	R/W	R	-	-	-	-	-	-
	minute-trend-frames	archive	lsc_7	4,498,685,952	225	R/W	R	-	-	-	-	-	-
	trend-frames	archive	lsc_7	127,680,156,160	258	R/W	R	-	-	-	-	-	-
	Total			202,593,514,932	739								
LLO	frames	archive	lsc_5	8,818,455,040	25	R/W	R	-	-	-	-	-	-
	minute-trend-frames	archive	lsc_8	909,567,968	262	R/W	R	-	-	-	-	-	-
	trend-frames	archive	lsc_8	15,441,281,024	182	R/W	R	-	-	-	-	-	-
	Total			25,169,304,032	407								
TOTAL				1,218,491,730,515	5,179								

Updated Mon Nov 15 12:42:58 PST 1999 (R=Read Access / W=Write Access)


frame_archive
Size
Files
1.22e+12
1.2e+12
1.18e+12
1.16e+12
1.14e+12
1.12e+12
1.1e+12
1.08e+12
1.06e+12
Size (Byte)
5200
5100
5000
4900
4800
4700
4600
4500
4400
4300
Number of Files
1999/Oct/23 00:00:00
1999/Oct/30 00:00:00
1999/Nov/06 00:00:00
1999/Nov/13 00:00:00
1999/Nov/20 00:00:00
Time (UTC)


HPSS FTP Data Transferred (LIGO data)
Read
Write
6e+10
5e+10
4e+10
3e+10
2e+10
1e+10
0
Data Transfered (byte/day)
1999/Oct/16 00:00:00
1999/Oct/23 00:00:00
1999/Oct/30 00:00:00
1999/Nov/06 00:00:00
1999/Nov/13 00:00:00
1999/Nov/ 00:00:0
Time (UTC)


HPSS FTP Data Transferred (ALL data)
Read
Write
1.4e+11
1.2e+11
1e+11
8e+10
6e+10
4e+10
2e+10
0
Data Transfered (byte/day)
1999/Oct/16 00:00:00
1999/Oct/23 00:00:00
1999/Oct/30 00:00:00
1999/Nov/06 00:00:00
1999/Nov/13 00:00:00
1999/Nov/
00:00:0
Time (UTC)


HPSS FTP Data rate (LIGO data)
Median Rate (byte/s)
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
Read
Write
1999/Oct/16
00:00:00
1999/Oct/23
00:00:00
1999/Oct/30
00:00:00
1999/Nov/06
00:00:00
1999/Nov/13
00:00:00
1999/Nov/
00:00:0
Time (UTC)

HP55 FTP Transfer Statistics (ALL data)
Number
Size
Median Size (bytes)
Files Transferred (#/day)
Time (UTC)
0
1e+08
2e+08
3e+08
4e+08
5e+08
6e+08
7e+08
0
5000
10000
15000
20000
25000
1999/Oct/16 00:00:00
1999/Oct/23 00:00:00
1999/Oct/30 00:00:00
1999/Nov/06 00:00:00
1999/Nov/13 00:00:00
1999/Nov/20 00:00:00

HP55 FTP Transfer Statistics (LIGO data)
Number
Size
Weight Size (Byte)
Files Transferred (#/Day)
Time (UTC)
0
1e+08
2e+08
3e+08
4e+08
5e+08
6e+08
7e+08
0
50
100
150
200
250
300
350
1999/Oct/16 00:00:00
1999/Oct/23 00:00:00
1999/Oct/30 00:00:00
1999/Nov/06 00:00:00
1999/Nov/13 00:00:00
1999/Nov/20 00:00:00

HP55 FTP Data rate (ALL data)
Read
Write
Median Rate (Byte/s)
0
1e+06
2e+06
3e+06
4e+06
5e+06
6e+06
7e+06
8e+06
9e+06
1e+07
1999/Oct/16 00:00:00
1999/Oct/23 00:00:00
1999/Oct/30 00:00:00
1999/Nov/06 00:00:00
1999/Nov/13 00:00:00
1999/Nov/ 00:00:0
Time (UTC)

LIGO on-site Storage

- StorageTek Robotics (TimberWolf 9740):
  - I/O: few $\times 10$ MB/s.
  - Capacity: $\gtrsim 10$ TB.
  - Footprint: $1.2\,\mathrm{m} \times 1.75\,\mathrm{m}$.
  - Media exchange: 14 cartridge access port.
- Baseline Configuration:
  - 2 "9840" drives with separate disk caches.
    - Create dual-copy master tapes.
    - Playback and master drive spare.
  - Joint tape pool ($\sim$ 1 month @ LHO).

LIGO on-site Storage

- StorageTek Robotics (TimberWolf 9740):
  - I/O: few $\times 10$ MB/s.
  - Capacity: $\gtrsim 10$ TB.
  - Footprint: $1.2\,\mathrm{m} \times 1.75\,\mathrm{m}$.
  - Media exchange: 14 cartridge access port.
- Baseline Configuration:
  - 2 "9840" drives with separate disk caches.
    - Create dual-copy master tapes.
    - Playback and master drive spare.
  - Joint tape pool ($\sim$ 1 month @ LHO).

LIGO off-site Storage

- StorageTek Robotics (4410/Powderhorn):
  - I/O: 50–100 MB/s ($\sim$ 5 drives).
  - Capacity: 2000 slots (1/3 library).
- HPSS Server Hardware:
  - $\sim$ 5 SMP datamover nodes.
  - 1 SMP metadata node.
  - 1 High Performance Gateway Node (HPGN): SP2/HIPPI/Gigabit/ATM.
  - Sun hardware an alternative in next HPSS release (4Q00).

HPSS Issues

- New: $\sim$ 20 production sites world wide.
  - Centralized code support from IBM ($\sim$ \$1M/yr).
  - Development support from DOE labs ($\sim$ 25FTEs).
  - Caltech has been running HPSS for 3 years.
- Stability: 95% uptime (current specification).
  - However, last release order of magnitude more stable.
- Complexity: built on top of other large software packages.
  - DCE (distributed computing).
  - Ensina (robust transactions).
  - Partially alleviated by IBM purchase of Transarc.
- Maintenance: most HPSS sites operate with 2 FTEs.
  - CACR has been successful with 1 FTE.
  - CACR/LIGO MOU for joint 2 FTE HPSS support.
- Licensing: $\sim$ \$100K/yr
  - CACR is currently paying all license fees.
  - CACR will continue to pay license fees under MOU.
- Collaboration: work with CACR.
  - LIGO is actively prototyping LIGO problems on the \$1M CACR production system at no cost.
- Scope: general solution.
  - Designed for a large number of large files.

- Multi-Platform: Sun (SGI)
  - Version 4.2 (4Q00) includes full core server port to Sun hardware.
- Floor Space: CACR facility full.
  - CACR MOU to provide 2000 tape slots.
  - Currently testing with 50GB tapes, however, STK changing to lower density tapes.
- User Interface: ftp, hpss api, nfs, dfs, mpi-io
  - What interface should LDAS use?

STK Tapedrives

dual spindle

	9840	TA42	TA44	TA50
GB	20	20	40	120
MB/s	10	20	20	40
Date	4Q98	2H00	1H02	2003

Single spindle

	TC40	TC42	TC50	Redwood
GB	60	120	360	50
MB/s	10	20	40	11
Date	4Q00	2H01	2003	—

Enhanced Scalability (cont)

Federated Name Space Diagram

Listing of HPSS Sites

as of September 1999

Name of Organization	No.	Phase	Rel.
Maui High Performance Computing Center (MHPCC)	1	P	3.2
Sandia National Laboratory (SNL)	2	P/P	3.2/3.2
California Institute of Technology (Caltech)	1	P	4.1.1
Fermi National Accelerator Laboratory (FNAL)	1	P	4.1.1
Lawrence Livermore National Laboratory (LLNL)	2	P/P	3.2/3.2
University of Washington (UWA)	1	P	3.2
Los Alamos National Laboratory (LANL)	2	P/P	4.1.1 4.1.1
San Diego Supercomputer Center (SDSC)	1	P	4.1.1
Oak Ridge National Laboratory (ORNL)	1	P	4.1.1

HPSS Deployment Phases

SI = System Integration
I = Installation
T = Production Readiness Test
P = Production

Listing of HPSS Sites (Cont.)

as of September, 1999

Name of Orginazation	No.	Phase	Rel.
Lawrence Berkeley Laboratory (NERSC)	3	P/P/P	3.2
NASA Langely Research Center (LaRC)	1	P	3.2
Stanford Linear Accelerator Center (SLAC)	1	P	4.1
European Laboratory for Particle Physics (CERN)	1	P	3.2
CEA Centre de Bruyeres le Chatel (CEA-DAM)	1	P	4.1.1
NOAA National Climatic Data Center (NCDC)	1	P	3.2
University of Maryland (UMCP)	1	P	3.2
Brookhaven National Lab (BNL)	1	P	4.1.1

HPSS Deployment Phases

SI = System Integration

I = Installation

T = Production Readiness Test

P = Production

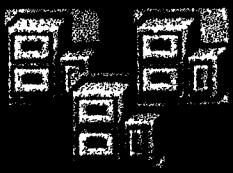
Listing of HPSS Sites (Cont.)

as of September, 1999

Name of Orginazation	No.	Phase	Rel.
Indiana University (IU)	1	P	4.1.1
Institut National de Physique Nucléaire et de Physique des Particules (IN2P3)	1	T	4.1.1
Institute of Physical and Chemical Research (RIKEN)	1	T	4.1.1
Marconi Integrated Systems	1	T	4.1.1
Argonne National Laboratory (ANL)	1	SI	4.1.1

HPSS Deployment Phases	
SI	= System Integration
I	= Installation
T	= Production Readiness Test
P	= Production

Enhanced Scalability (cont)



Storage Subsystem Diagram