

High Performance Storage System

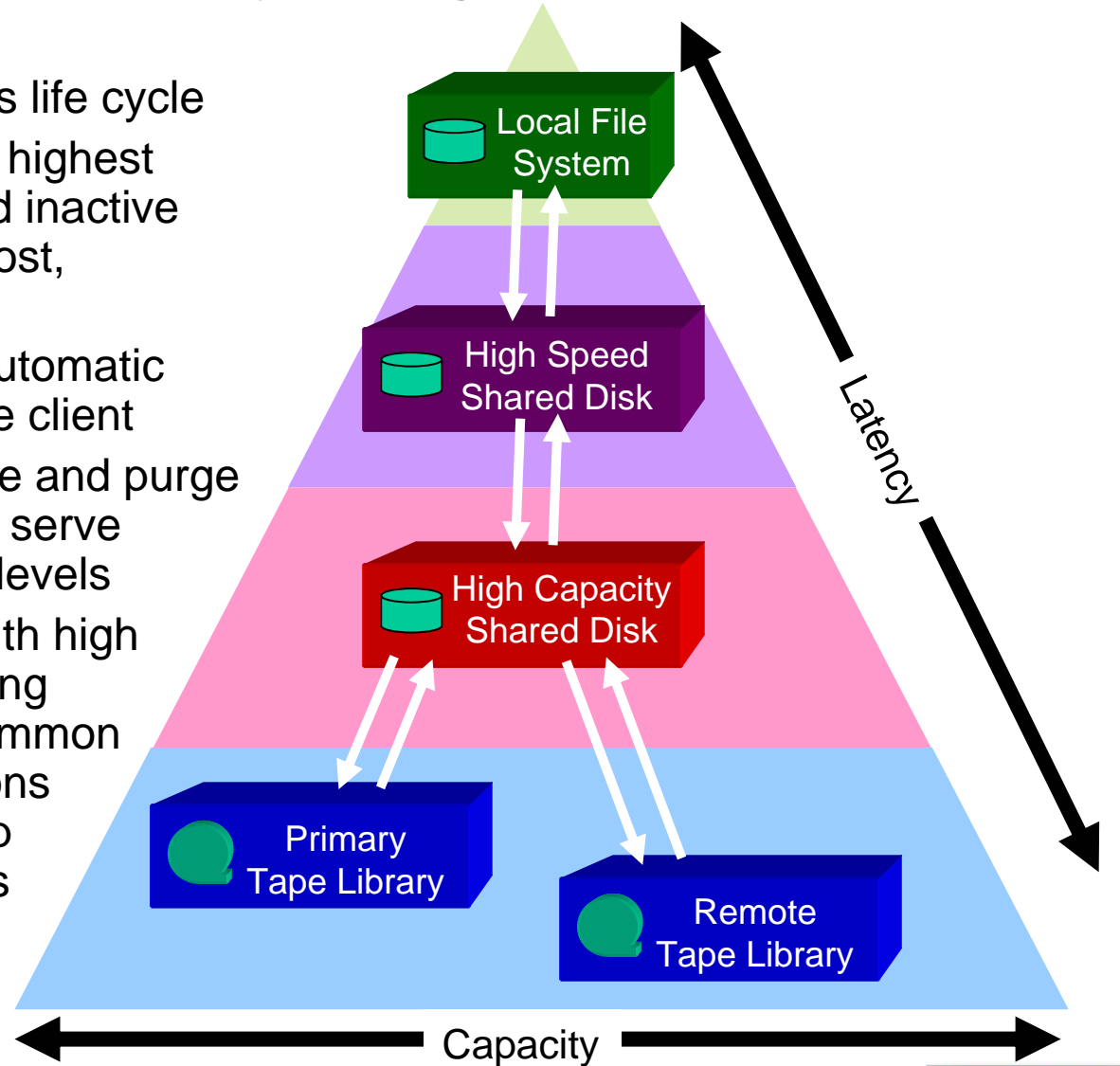


www.hpss-collaboration.org

Hierarchical Storage Management (HSM)

for Data Lifecycle Management

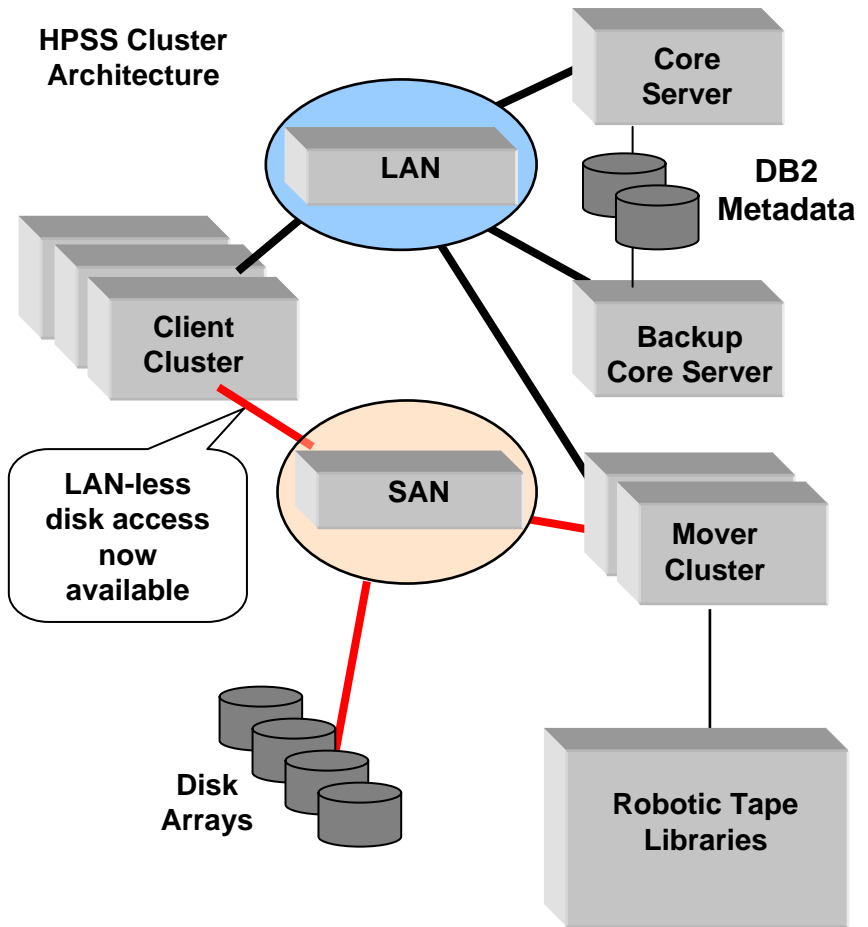
- Manages data over its life cycle
- Keeps active data on highest performing media and inactive data on tape or low cost, high capacity disk
- Migration of data is automatic and transparent to the client
- Asynchronous migrate and purge allows lower levels to serve as backup for higher levels
- HSM is associated with high performance computing and is relatively uncommon in business applications where simpler backup and restore strategies are sufficient



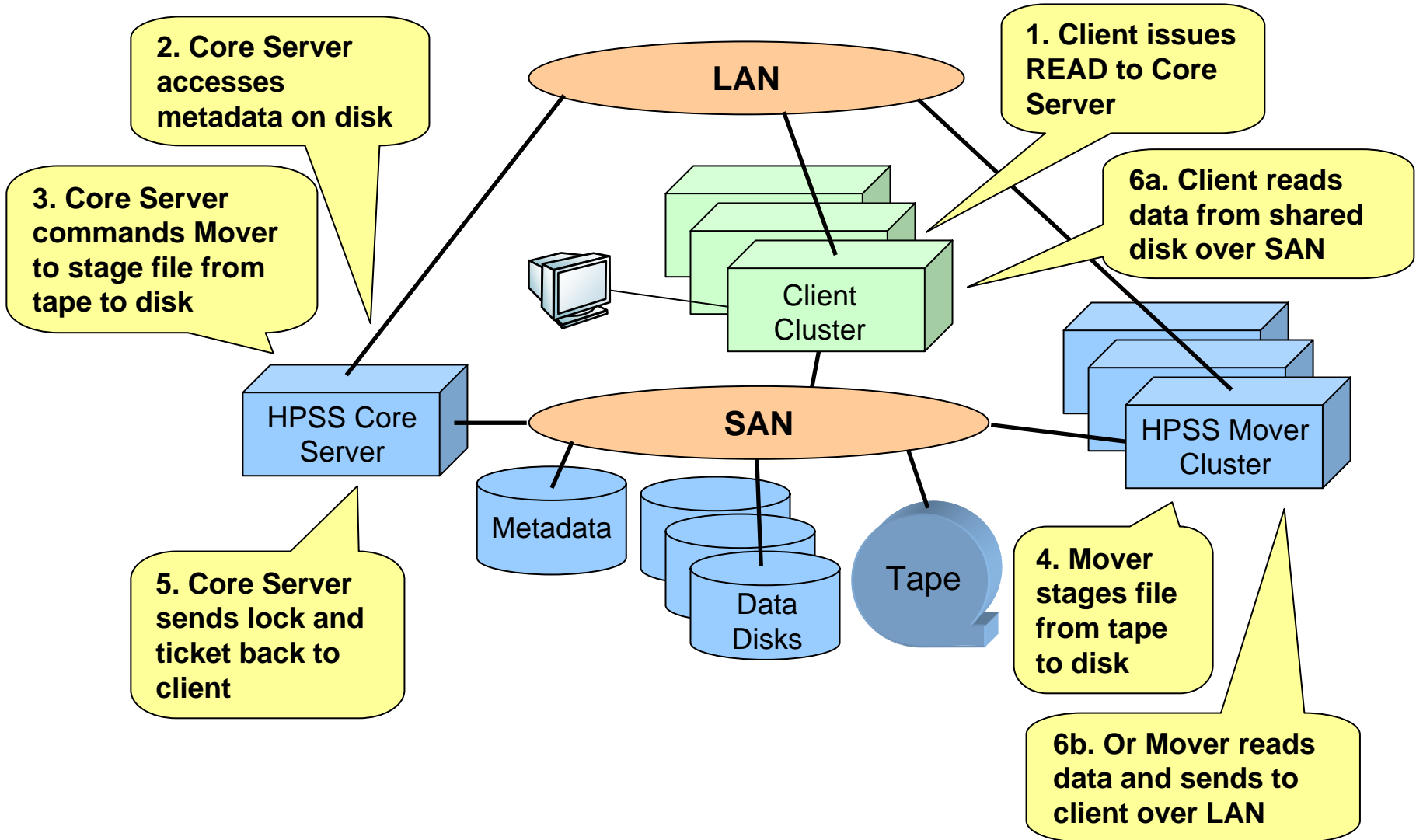
HPSS Cluster Architecture

Cluster storage for cluster supercomputing

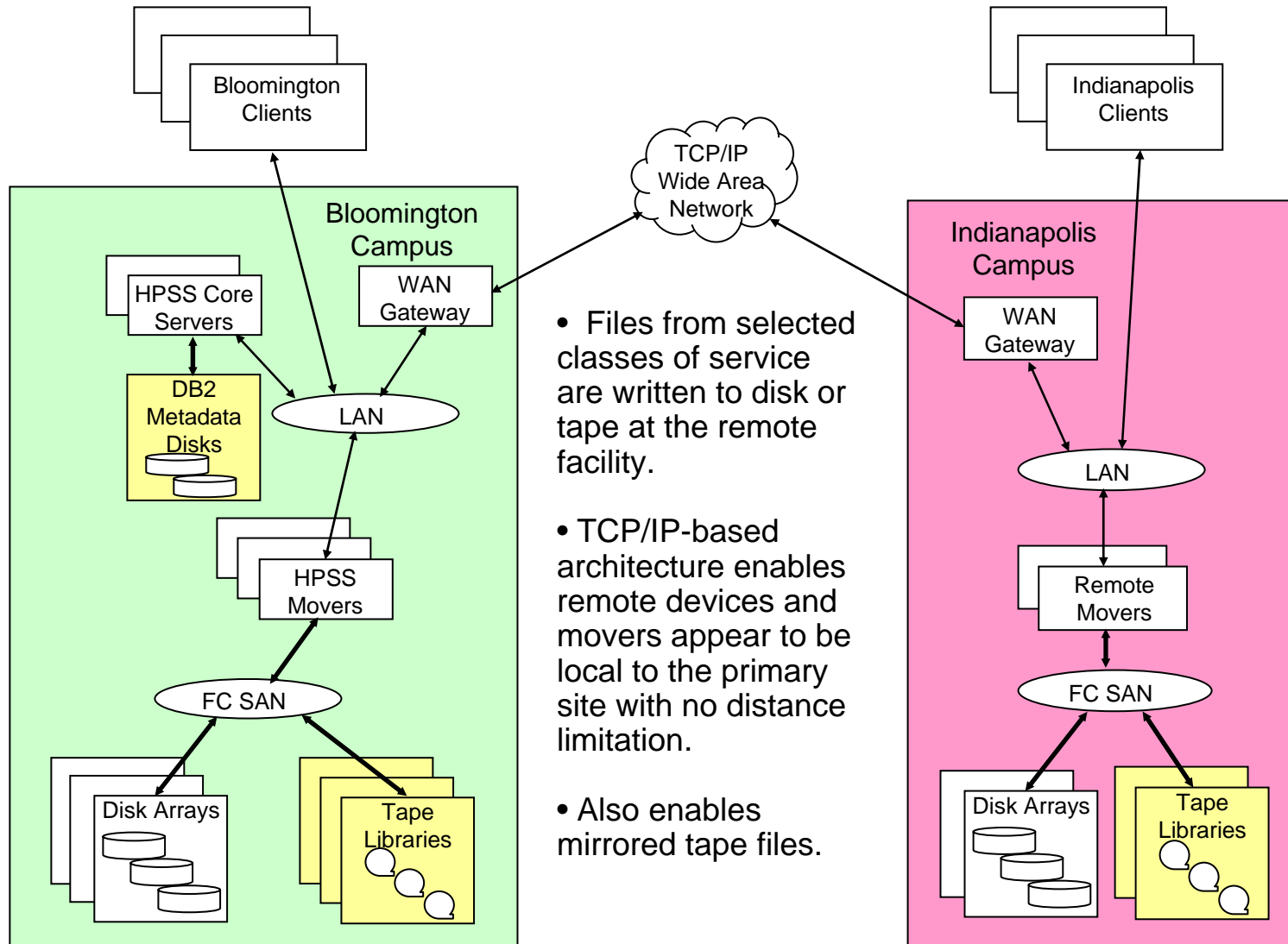
- HPSS is . . .
 - Storage software system suitable for long term retention and rapid staging
 - Very rugged DB2 metadata architecture suitable for medium- and coarse-grain file access
 - Disk and tape, and/or tape only
 - Arguably the most scalable disk- and-tape system anywhere
- Cluster architecture and metadata architecture support horizontal scaling to:
 - 10s of petabytes
 - 100s of millions of files
 - gigabytes per second data rates
 - All in a single system
- Supports technology insertion
 - Add new components, no need to replace
 - Mix and match vendors and models
 - AIX and Linux with client support for IRIX and Solaris



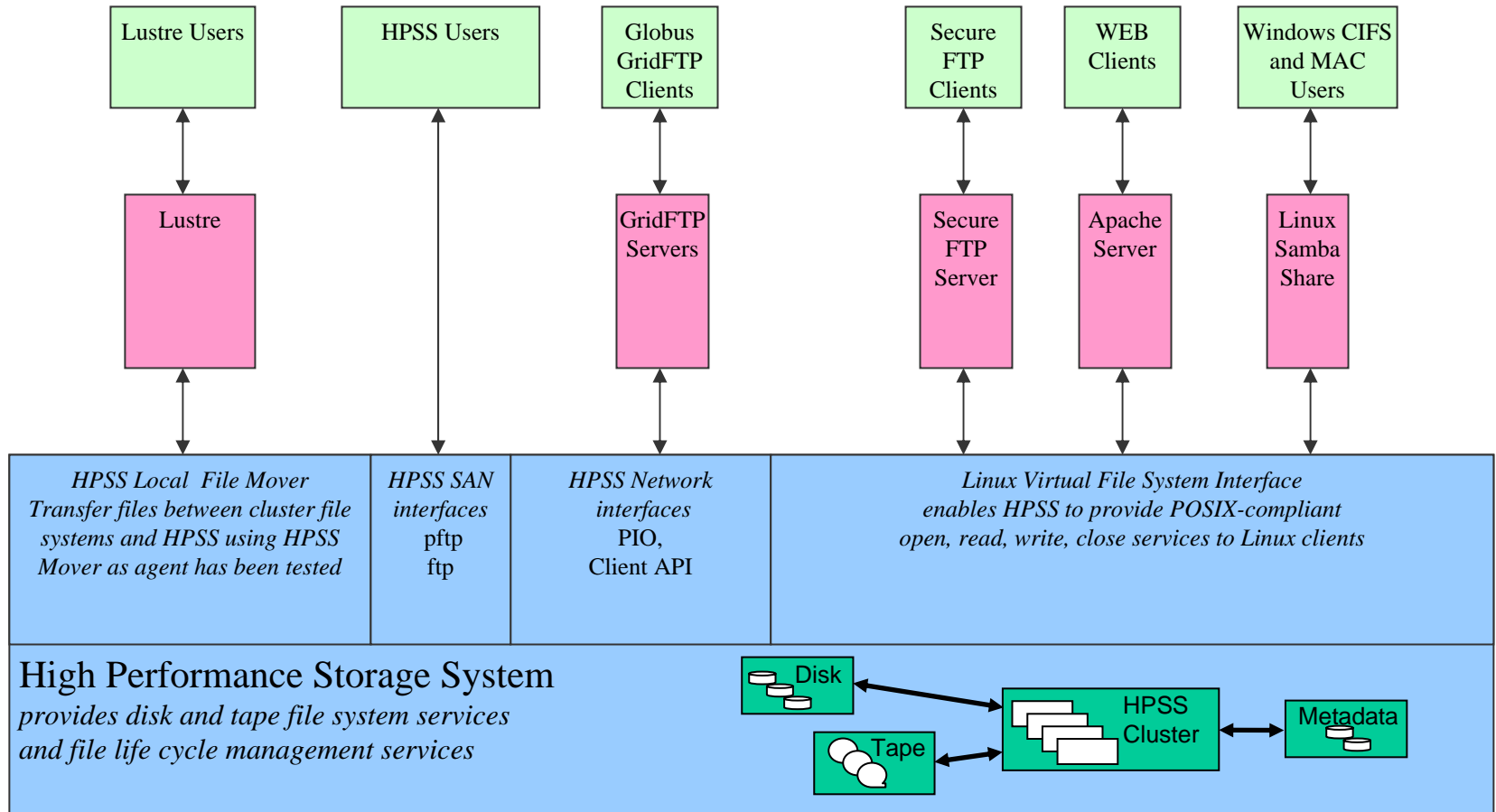
How HPSS Works



IU: Two Sites, One System with Remote Movers



Enterprise HSM Services at IU



Traditional HPSS interfaces

- HPSS Client API
 - A superset of POSIX read-write interfaces
 - General form `hpss_open()`, `hpss_read()`, `hpss_write()`
 - Supports parallel files, parallel servers, parallel clients
 - Supports hints and explicit specification of classes of service
 - The internal basis of all other HPSS interfaces
- HPSS Parallel FTP
 - Semantics similar to ftp
 - Supports parallel files, parallel servers, parallel clients
 - Supports LAN and WAN transfers at high data rates
 - Conventional ftp semantics and capabilities also supported

New HPSS interfaces

- HPSS VFS Interface for Linux
 - Linux applications benefit from a true POSIX standard read/write interface
 - This interface allows many standard commercial programs that do file I/O to use HPSS as file space, essentially turning them into hierarchical disk-tape applications
 - Uses the Linux Virtual File System (VFS) interface
- GridFTP
 - GridFTP is a Grid interface of the Globus Toolkit
 - A downloadable open source GridFTP interface is available for HPSS, developed by the US DOE, Argonne National Lab
 - Now operational at Indiana University

HPSS Interfaces, continued

- Direct SAN access to disks
 - In addition to the original Mover-based disk sharing, disks may now be accessed directly over a SAN
 - SAN access is an option for HPSS Client API and PFTP for AIX and Linux and with the HPSS FS interface described on the next slide
- “Client SAN” access to other file system’s SANs
 - The HPSS Local File Mover feature has been updated and tested to allow HPSS to transfer data between another file system and HPSS over the other files system’s SAN
 - HPSS Movers serve as the transfer agent.
 - This capability works with any cluster file systems offering a Unix/Posix read/write interface, and it has been tested with IBM GPFS, Lustre, and ADIC SNFS

LAN and SAN Access to Data

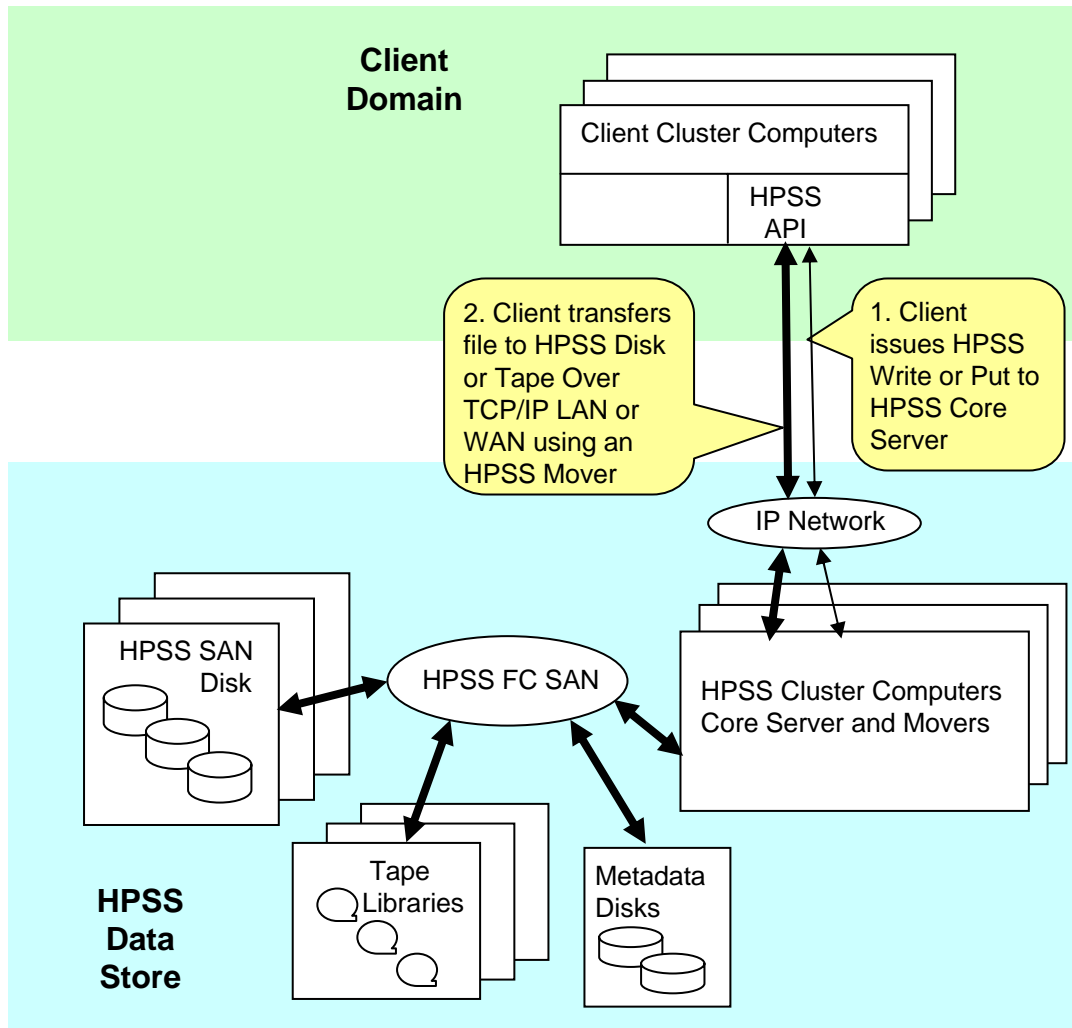
- HPSS supports LAN and SAN access to data
- LAN access means . . .
 - Data is delivered over a LAN, control also over a LAN
 - Block-level transfers simulate (actually precursor to) a SAN
 - Definitely *not* an NFS-type transfer
 - Supports large numbers of clients
 - Uses commodity TCP/IP infrastructure
- SAN access means . . .
 - Data is delivered over a SAN, control over a LAN
 - Fibre channel protocol uses less CPU than TCP/IP
 - Fewer HPSS movers needed
 - Usually limits access to well under a hundred clients
 - Uses more expensive FC HBAs and switches (IP SAN has little or no advantage over HPSS native LAN capability)
- Usually HPSS system engineers recommend LAN access to data for large HPC clusters

Storage Devices Supported

- Tape Drives
 - IBM LTO-3 tape drives on AIX and Linux movers
 - HP LTO-3 Tape Drives* on Linux Movers
 - IBM 3592 and TS1120 tape drives
 - StorageTek (Sun) 9940 and Titanium T10000 Tape Drives
- Tape Libraries
 - IBM 3583*, 3584 and 3494 libraries
 - Sun StorageTek ACSLS-based libraries including SL500 and SL8500*
 - Spectra Logic T Series libraries
 - ADIC Scalar Series libraries
 - Generic SCSI libraries*
- Disk Arrays (both Fibre Channel and SATA)
 - IBM 4200*, 4300, 4500, 4700*, 4800*, and 6800*
 - Engenio disks sold by Sun, SGI, and others*
 - Direct Data Networks (DDN) disk arrays
 - Generic fibre channel disk arrays*
 - Copan Systems MAID device*

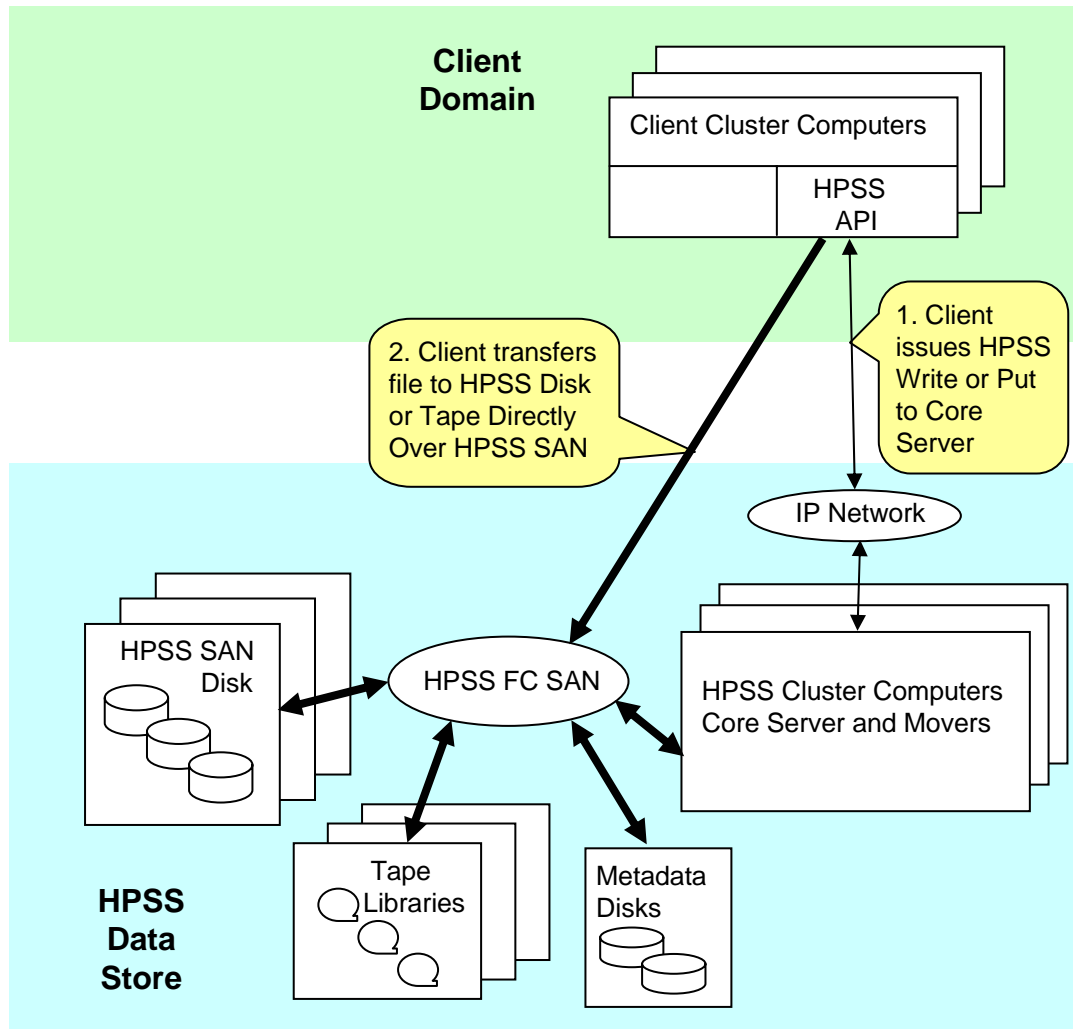
* *We have evaluated and/or tested these devices with HPSS but do not have them permanently in our development lab. These devices can be supported with arrangement to run tests on customer equipment in case diagnostic tests are needed.*

HPSS Write or Put Over TCP/IP Network



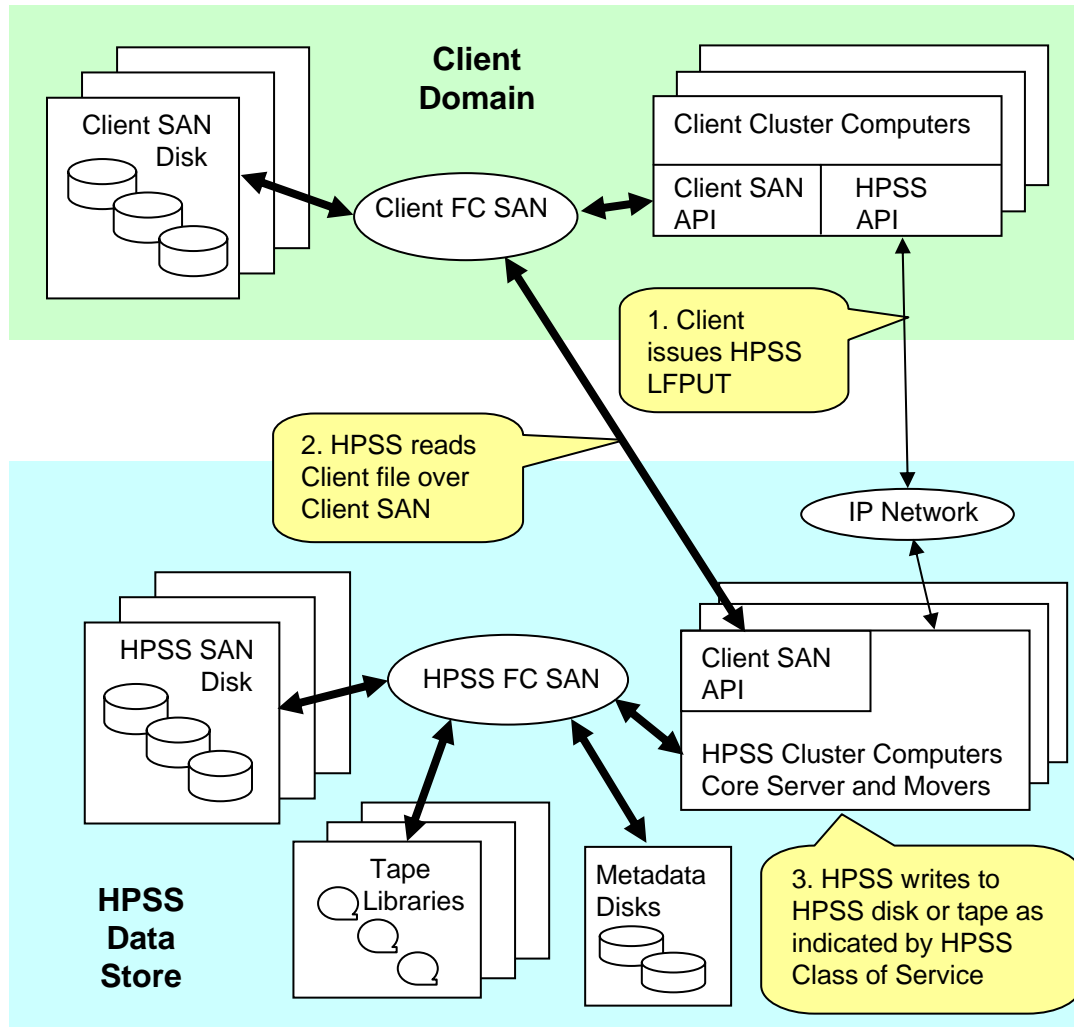
- HPSS Client API
 - Hpsss_write() etc.
 - Optional list form to access discontinuous segments
 - Parallel, gigabyte/s capability
 - Use for performance-critical applications
- HPSS PFTP
 - Parallel FTP
 - FTP-like get-put semantics
 - Parallel, gigabyte/s capability
 - Most-used HPSS interface

SAN-Enabled HPSS Write or Put



- Data transferred directly between client and HPSS disk over HPSS SAN
- Control is over TCP/IP network (separation of control and data)
- Supported by HPSS Client API and PFTP
- Currently supported on AIX and Linux
- Used internally to HPSS to move data between disk and tape

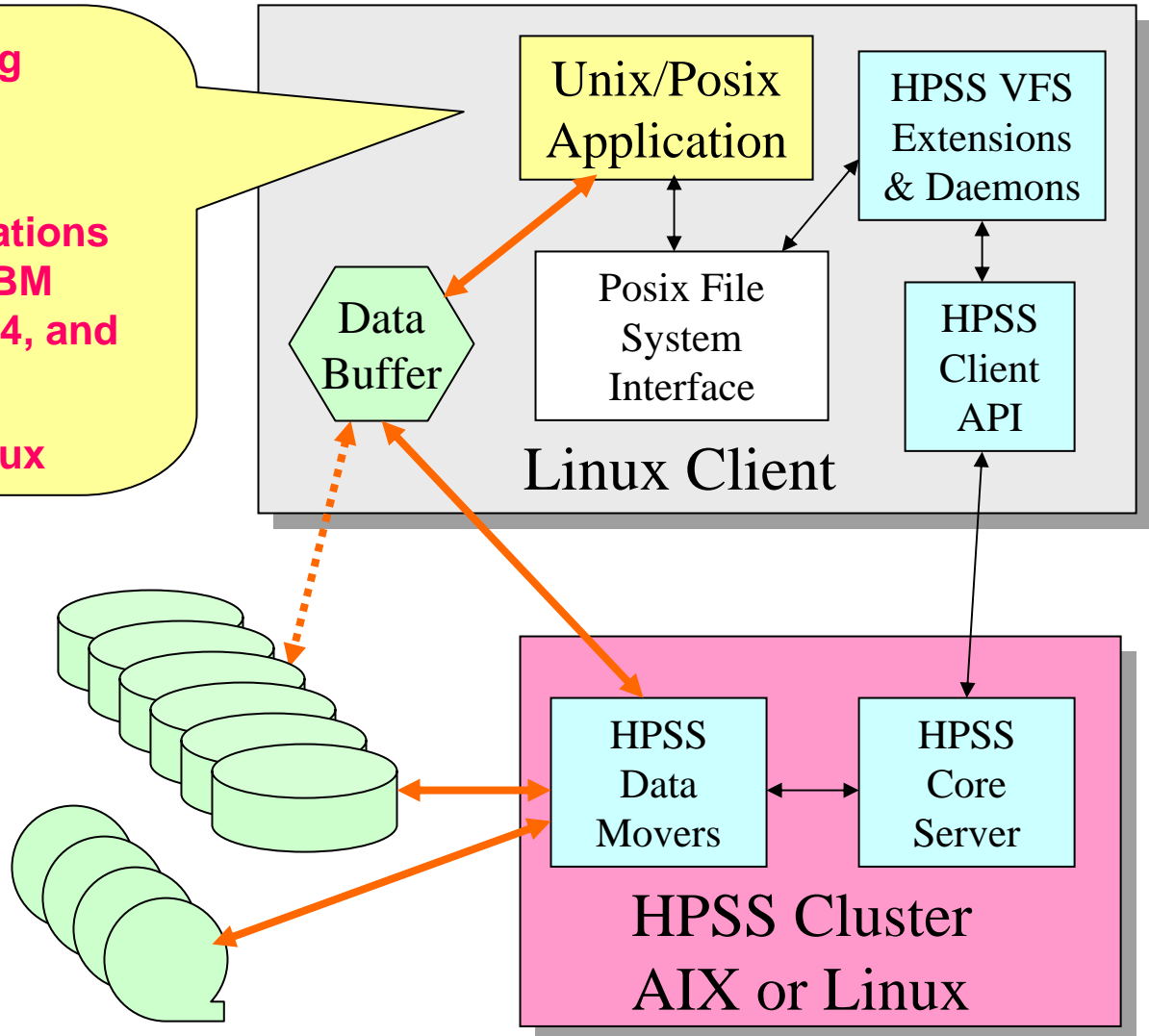
HPSS “Local File Mover” or “Client SAN”



- HPSS accesses data on **client SAN**
- Examples of client SAN: IBM SAN FS, ADIC SNFS, IBM GPFS
- Activated by PFTP LFPUT-LFGET with more options coming
- CPU overhead entirely offloaded to HPSS Movers
- Parallel capability and/or direct tape access via Class of Service options

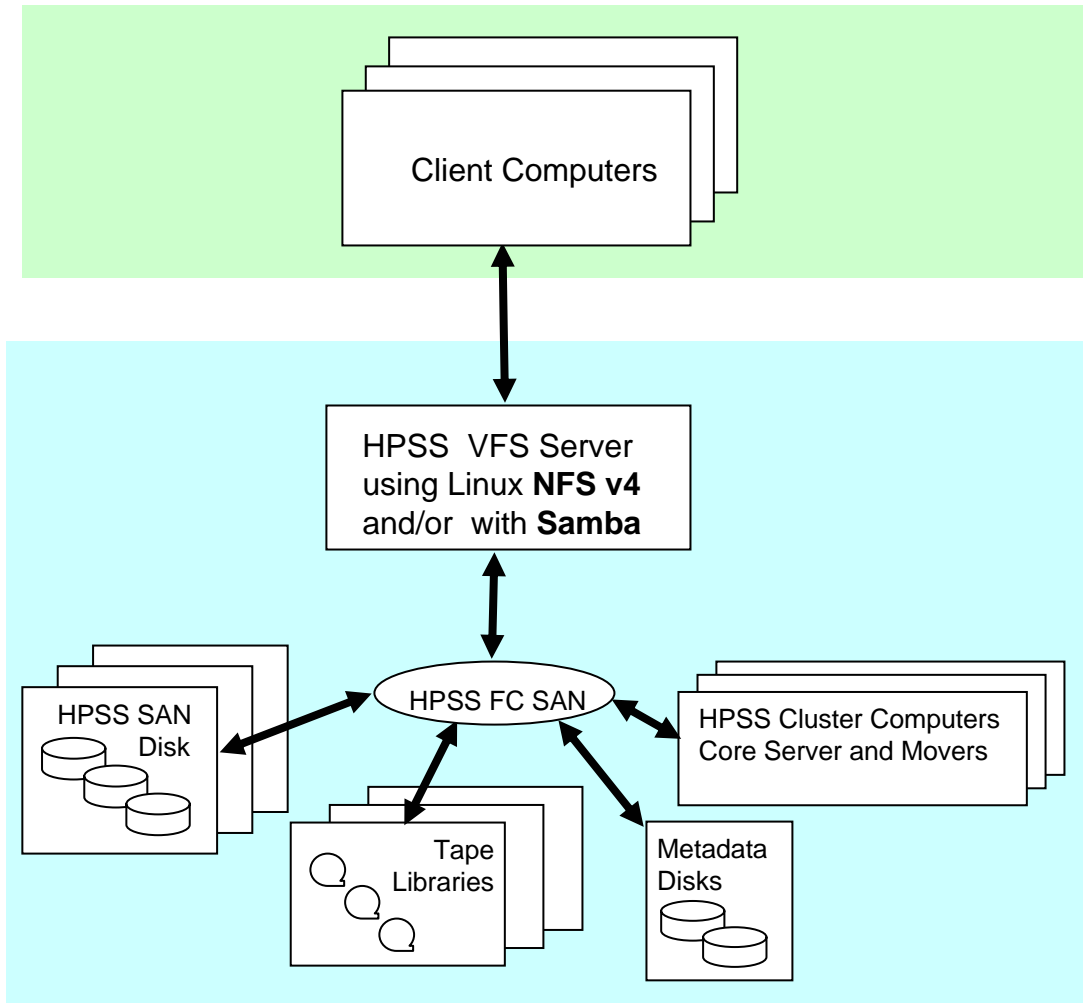
HPSS VFS Interface for Linux

- ✓ HPSS accessed using standard UNIX/Posix semantics
- ✓ Run standard applications on HPSS such as IBM DB2, IBM TSM, NFSv4, and Samba
- ✓ VFS available for Linux



- ↔ Control
- ↔ Data
- ↔ Optional SAN Data Path

Windows Samba Interface

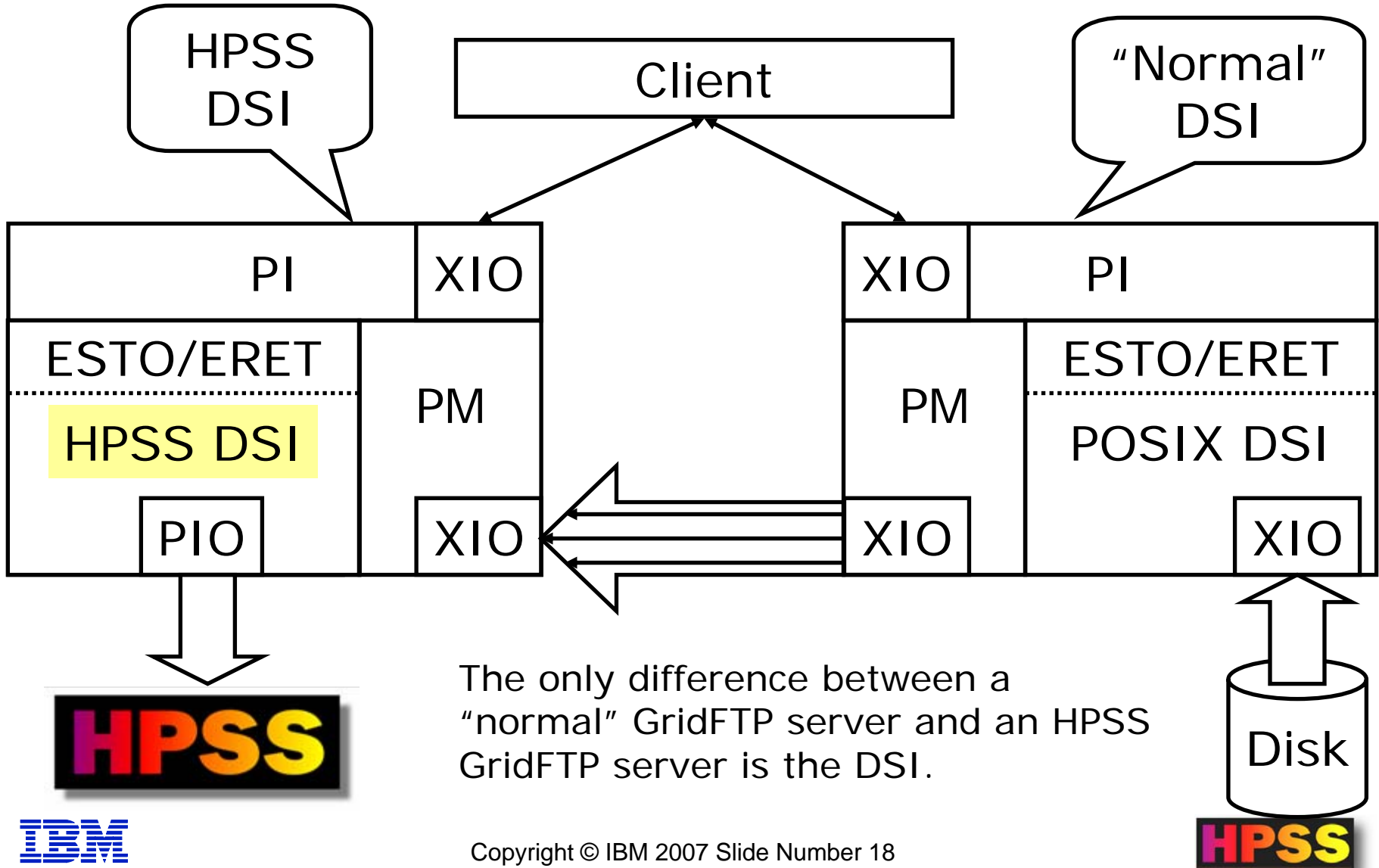


- Samba has been demonstrated to run with the HPSS VFS Interface for Linux
- Makes HPSS storage accessible to Windows
- Operational today at Indiana University

GridFTP is a proposed standard

- GridFTP is based on several existing IETF Requests for Comment (RFCs):
 - RFC 959: File Transfer Protocol
 - RFC 2228: FTP Security Extensions
 - RFC 2389: Feature Negotiation for the File Transfer Protocol
 - Draft: FTP Extensions
 - GridFTP: Protocol Extensions to FTP for the Grid
- Globus has the reference implementation
- There are multiple interoperating implementations
 - Fermi Lab developed DCache GridFTP “door”
 - Univ of Virginia developed GridFTP .NET (Windows)
 - Univ of Wisconsin developed GridFTP lite
 - Argonne Lab developed HPSS GridFTP
- **Indiana University has GridFTP with HPSS in production for TeraGrid**

The HPSS DSI enables HPSS GridFTP

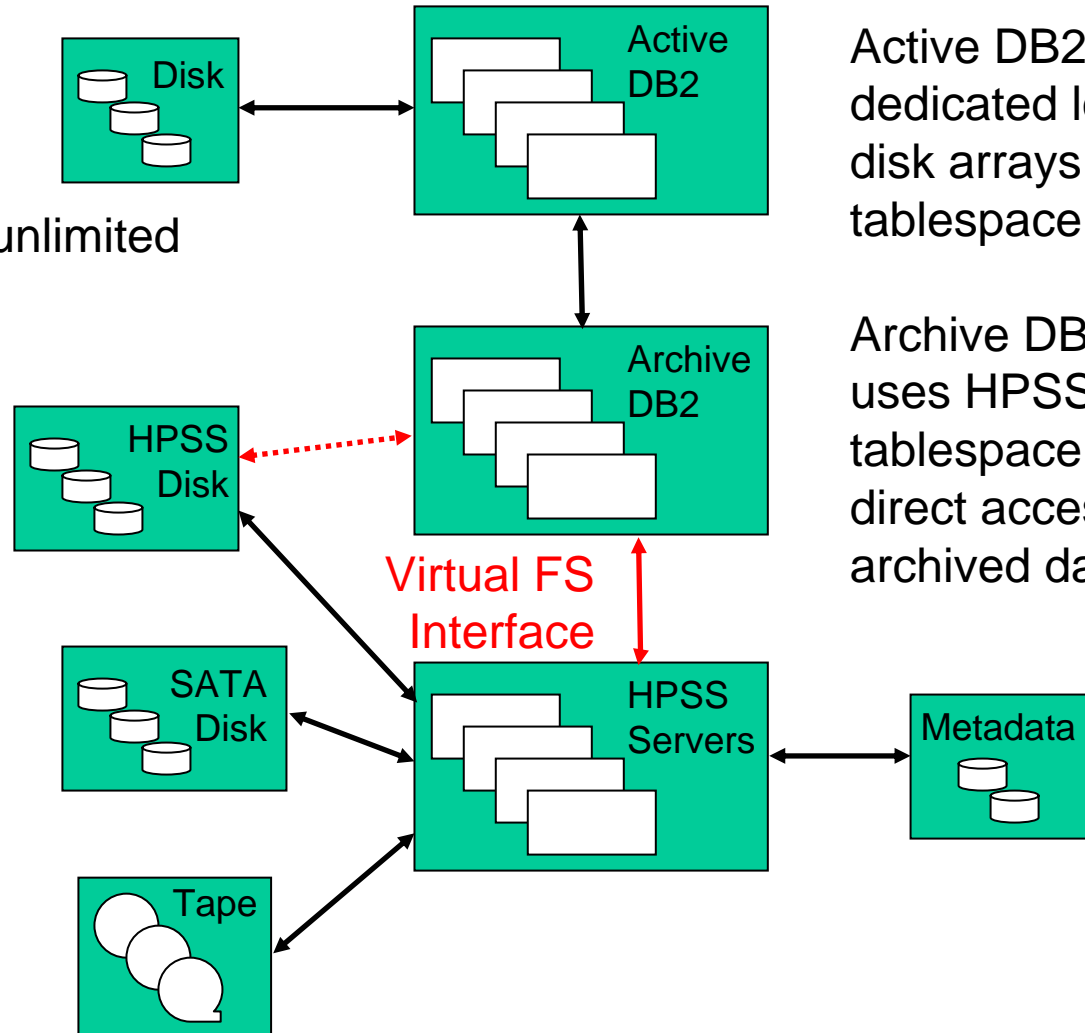


DB2/HPSS Active Tablespace Archive

Recent HPSS developments allow IBM DB2 to run as an HSM

Archiving databases in HPSS allows virtually unlimited accessible data

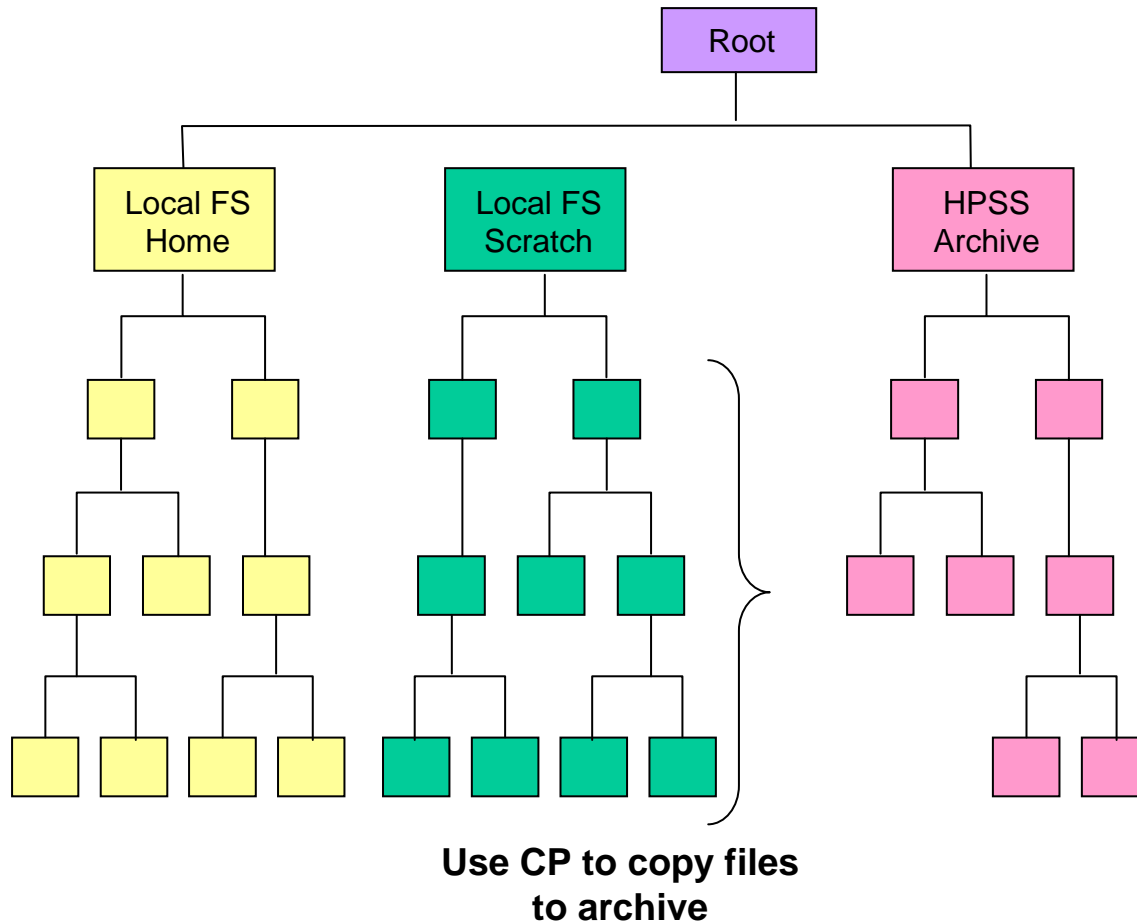
Archive DB2 “sees” Multi-level HPSS storage as virtual disk storage



Active DB2 uses dedicated local disk arrays for tablespace

Archive DB2 uses HPSS for tablespace allowing direct access to archived data

HPSS can be mounted as a directory subtree using the HPSS VFS Interface



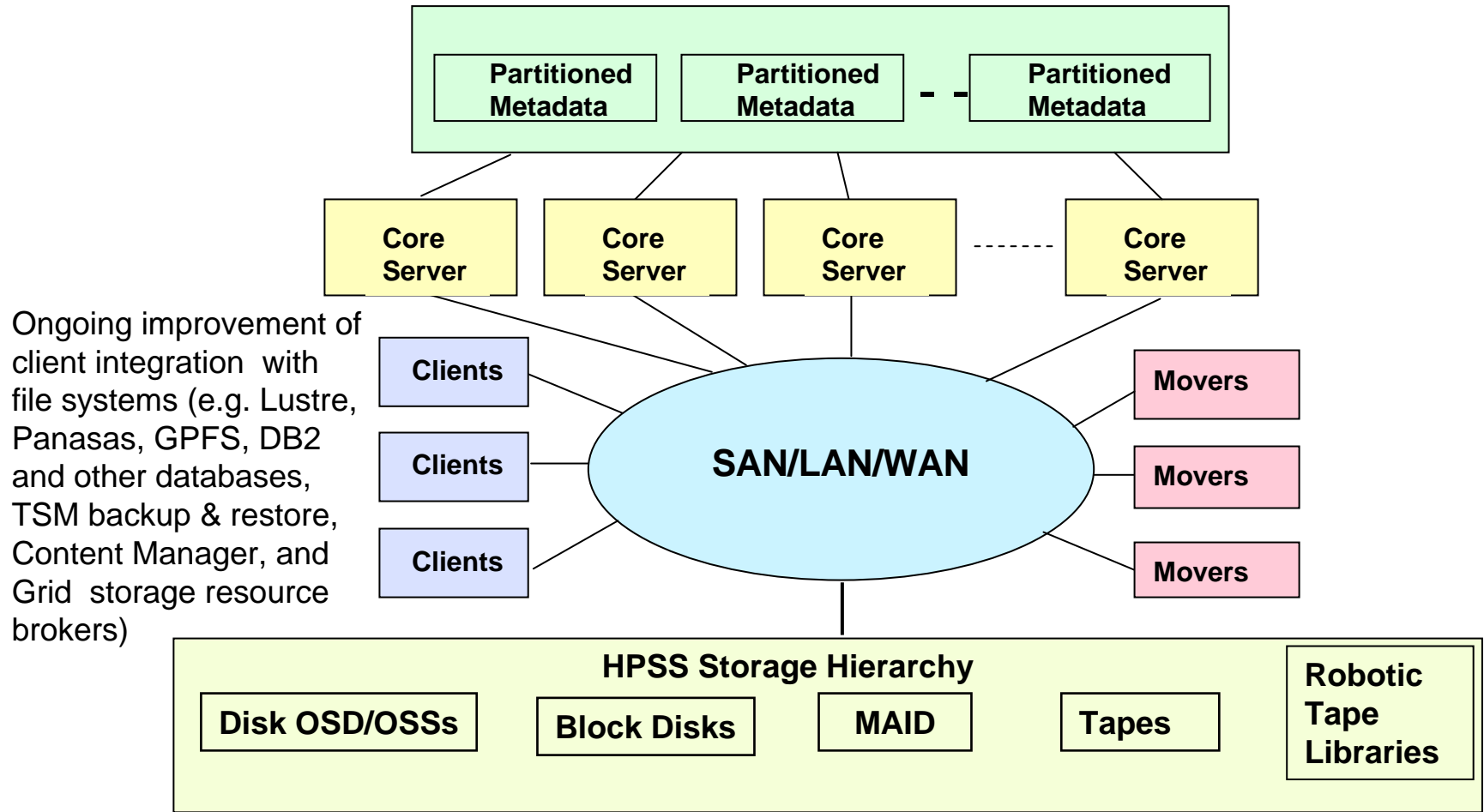
- Applications can use standard POSIX reads and writes
- Clients can use the cp shell command to copy files between an active subtree managed by GPFS and an archive subtree managed by HPSS
- *Currently available on Linux*

Coming: GPFS/HPSS Interface (GHI)

- GPFS/HPSS Interface is a collaborative project to develop synergy between IBM's General Parallel File System and HPSS
- Goals
 - Extend GPFS pool concept to tape and other long term storage
 - Use the GPFS rule-based ILM to centralize ILM administration
 - Provide backup for GPFS that makes effective use of ongoing file migration to tape under ILM control
 - Unequaled scalability
- Participants
 - HPSS Collaboration member NERSC/Lawrence Berkeley Lab
 - IBM Research, Almaden Lab
 - IBM GPFS Product Development in Poughkeepsie NY
 - IBM HPSS Development and Support in Houston TX
- Status
 - Ongoing work demonstrated at SC06 conference November 2006
 - Expect first release general availability by end of 2007

HPSS in 2009 – 2011

from Dick Watson, LLNL, co-chair of HPSS Executive Committee



Ongoing improvement of client integration with file systems (e.g. Lustre, Panasas, GPFS, DB2 and other databases, TSM backup & restore, Content Manager, and Grid storage resource brokers)

HPSS Collaborative Development

“Since 1992”

- Owner-Developers (copyright holders)
 - Lawrence Livermore National Laboratory
 - Los Alamos National Laboratory
 - Lawrence Berkeley National Laboratory
 - Sandia National Laboratories
 - Oak Ridge National Laboratory
 - IBM
- Other Stake Holders and Collaborators
 - Argonne National Laboratory (Grid)
 - San Diego Supercomputer Center (Grid)
 - CEA DAM (France)
 - KEK Lab (Japan)
 - NCEP (US weather service)
 - ECMWF (European weather service)
 - Indiana University (Grid)
 - DDN (high speed disk storage)
 - ADIC (Tape libraries)
 - Spectra Logic (tape libraries)
 - Gleicher Enterprises
 - All our customers!
- Role of IBM Global Services in Houston, Texas
 - Project leadership and management
 - Commercial licensing, distribution, and service
 - Quality assurance and testing (SEI CMMI Level 3)
 - Access to IBM technology (including DB2, GPFS)
- Advantages of Collaborative Development
 - Developers are users: focus on what is needed and what works
 - Software is open and source code is available to collaboration members and US/NATO stake holders

How HPSS is Sold and Supported

- **HPSS is an IBM Service Offering**
 - When you purchase conventional proprietary software, you pay an upfront license fee and then pay for maintenance
 - When you purchase HPSS, which is jointly developed by five U.S. Department of Energy laboratories and IBM (see “*HPSS Collaborative Development*” slide), there is an end-user license agreement document but no license fee, and you purchase a subscription for the services you need from IBM
 - HPSS subscriptions include Basic, Standard, or Premium for most customers and Full Service for customers with very complex requirements (see “*HPSS Offerings*” slide)
 - Charges vary depending on options and level of service, but significantly there are no capacity charges (see “*HPSS Price Examples*” slide)
- **HPSS Source Is Available**
 - All source code developed by the HPSS Collaboration is available without charge to U.S. customers
 - Requires a non-disclosure provision in the HPSS license or a separate non-disclosure agreement
 - HPSS is bundled with some proprietary software including IBM DB2, for which source code is not available
- **HPSS is COTS (Commercial Off-The-Shelf, a government contracting term)**
 - GSA-equivalent terms and conditions are standard for U.S. contracts
 - HPSS has always met commerciality requirements of government agencies of the U.S. and other countries, educational institutions, and government-funded research organizations
 - Listed on the NASA SEWP GWAC (Government-Wide Acquisition Contract) -- www.sewp.nasa.gov
- **HPSS is sold and supported by IBM Global Business Services**
 - Marketed by reputation and referral
 - Proactive support model with high customer satisfaction
 - 12 successful years as a commercial offering

Size of some of the larger HPSS sites

System (Each system shown is a single HPSS instance and namespace)	Petabytes* (10 ¹⁵ bytes)	Million files	Avg file MB	As Of Date
Los Alamos National Lab (LANL) Secure Computing Facility (SCF)	11.0	85.0	124	9/7/2007
The European Centre for Medium-Range Weather Forecasts (ECMWF)	8.5	37.0	219	9/7/2007
Lawrence Livermore National Lab (LLNL) Secure Computing Facility (SCF)	7.0	54.5	122	9/6/2007
Brookhaven National Lab (BNL)	6.5	36.5	168	9/7/2007
LLNL Open Computing Facility (OCF)	4.8	50.3	90	9/6/2007
San Diego Supercomputer Center (SDSC)	4.6	43.5	100	9/4/2007
Stanford Linear Accelerator Center (SLAC)	3.6	4.2	824	8/30/2007
National Centers for Environmental Prediction (NCEP)	3.5	4.1	807	9/7/2007
Commissariat à l'Energie Atomique/Division des Applications Militaires (CEA)	3.0	1.3	2219	4/16/2007
Institute National de Physique Nucléaire et de Physique des Particules (IN2P3)	2.8	20.3	132	9/7/2007
Lawrence Berkeley Lab (LBL) National Energy Research Scientific Computing Center (NERSC)	2.5	50.1	48	9/4/2007
Oak Ridge National Laboratory (ORNL)	1.7	8.9	178	9/10/2007
LBL NERSC Backup System	1.6	11.6	130	9/4/2007
RIKEN in Japan	1.3	2.8	449	8/30/2007
LANL Open Computing Facility	1.1	18.5	58	9/7/2007
Indiana University (IU)	1.0	12.4	71	9/7/2007
National Climatic Data Center (NCDC)	0.9	38.6	23	8/30/2007
NASA Langely	0.7	3.9	165	9/7/2007

* HPSS follows the convention used by most enterprise disk and tape manufacturers and the SNIA that one petabyte = 1000⁵ bytes.
To convert to binary petabytes, where one petabyte = 1024⁵ bytes, multiply by 0.888

