

# High Performance Storage System Overview



\* Please review disclosure statement on last slide

# High Performance Storage System



- Hierarchical storage software developed in collaboration with five US Department of Energy Labs since 1992.
- Provides stewardship for 100s of billions of files spanning 100s of petabytes of tapes for the HPC community.
- Licensed and supported by IBM Global Business Services in Houston, Texas.

Highest Performance Computing ↔ Highest Performance Storage



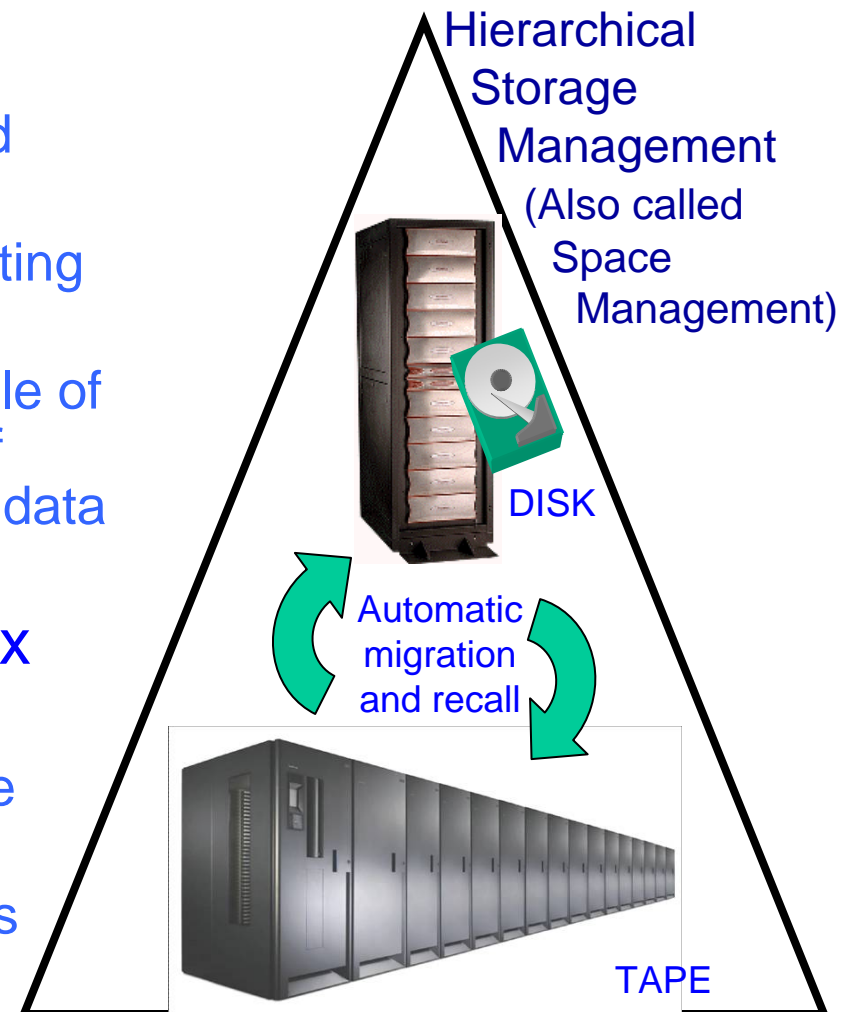
- LANL Roadrunner (top 500 #1)
- ORNL Jaguar (top 500 #2)
- LLNL Blue Gene/L (top 500 #5)
- ANL Blue Gene/P (top 500 #7)
- ECMWF Power 575
- IU Big Red
- LBNL NERSC (top 500 #7)
- LLNL ASC Purple
- NCEP Power 575
- NCSA Blue Water



# High Performance Storage System



- **Disk and tape file repository**
  - Hierarchical storage management (HSM) with automatic migration and recall
  - Highly scalable for high-end computing and storage customers
  - A single instance of HPSS is capable of concurrently accessing hundreds of tapes for extremely high aggregate data transfers.
- **User sees HPSS as a single Unix file system**
  - “Classic” HPSS presents its own file system
  - New HPSS for GPFS extends IBM’s most scalable file system to tape



# The Collaboration



- HPSS Collaboration (copyright holders & developers)

- IBM Global Services in Houston, Texas
- Lawrence Berkeley National Laboratory
- Lawrence Livermore National Laboratory
- Los Alamos National Laboratory
- Oak Ridge National Laboratory
- Sandia National Laboratories



“Since 1992”

- Advantages of Collaborative Development

- Developers are users; they focus on what is needed and what works.
- Source code is available to collaboration members and US stake holders (though HPSS is not open source in the usual sense)
- Customization is allowed (subject to collaboration review).
- *Historically, world wide, all truly high end software has required significant means other than retail sales to spread costs among interested parties, and our unique collaborative method has worked well for HPSS through 7 major releases and with a game-changing 8<sup>th</sup> in development*

# Profile of an HPSS customer



- Already has experience with tape – is not a tape novice
  - Understands the need for tape including acquisition cost benefits, volumetric benefits, long term archive benefits, energy and cooling benefits
  - Has capable staff with tape experience
- Already has experience with tape software
  - Understands difference between file system backup and HSM
  - Has done enough homework to know that one or more easy solutions, and probably their existing solution, won't meet their requirements
- Has one or more difficult requirements, now ***or in the future***
  - Tens to hundreds of tape drives
  - Tens of thousands or more tape cartridges
  - Many small, tape-unfriendly files affecting tape performance
  - Multiple service classes such as single tape, striped tape, mirrored files, remote mirroring, and need to group like file classes on tape
  - Need to keep track of 100s of millions to billions of files in a single name space
  - Need a single solution for file system backup, space management, and archive
  - ***HPSS generally exceeds all others in all of these difficult requirements***

# Sizes of some of the larger HPSS sites

Updated 10/31/09



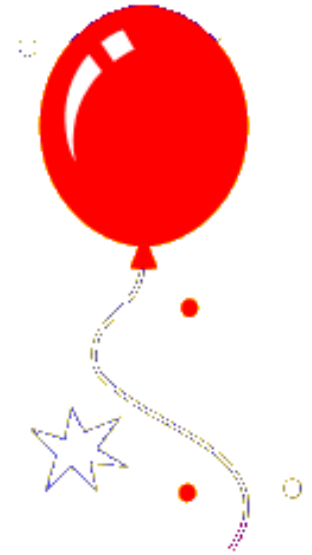
System (Each system shown is a single HPSS instance and namespace)	Petabytes (10 <sup>15</sup> bytes)	Million files	Avg file MB
The European Centre for Medium-Range Weather Forecasts (ECMWF)	15.69	63	238
Lawrence Livermore National Lab (LLNL) Secure Computing Facility (SCF)	15.45	113	131
Los Alamos National Lab (LANL) Secure Computing Facility (SCF)	14.58	126	111
Brookhaven National Lab (BNL)	11.75	67	167
National Centers for Environmental Prediction (NCEP)	10.32	9	1117
Deutsches Klimarechenzentrum GmbH (DKRZ)	9.53	24	378
LLNL Open Computing Facility (OCF)	8.88	117	72
Oak Ridge National Laboratory (ORNL)	8.06	16	474
Commissariat à l'Energie Atomique/Division des Applications Militaires (CEA)	7.27	2	3,233
Institute National de Physique Nucléaire et de Physique des Particules (IN2P3)	7.02	26	258
Lawrence Berkeley Lab (LBL) National Energy Research Scientific Computing Center (NERSC)	5.80	79	70
Stanford Linear Accelerator Center (SLAC)	5.22	7	760
San Diego Supercomputer Center (SDSC)	4.46	58	74
Indiana University (IU)	3.25	25	124
LBL NERSC Backup System	3.09	13	224
LANL Open Computing Facility	2.12	30	68
NASA Langley (LaRC)	1.97	6	304
RIKEN in Japan	1.73	5	346

HPSS follows the convention used by most enterprise disk and tape manufactures and the SNIA that one petabyte = 1000<sup>5</sup> bytes.  
To convert to binary petabytes, where one petabyte = 1024<sup>5</sup> bytes, multiply by 0.888

# New HPSS sites

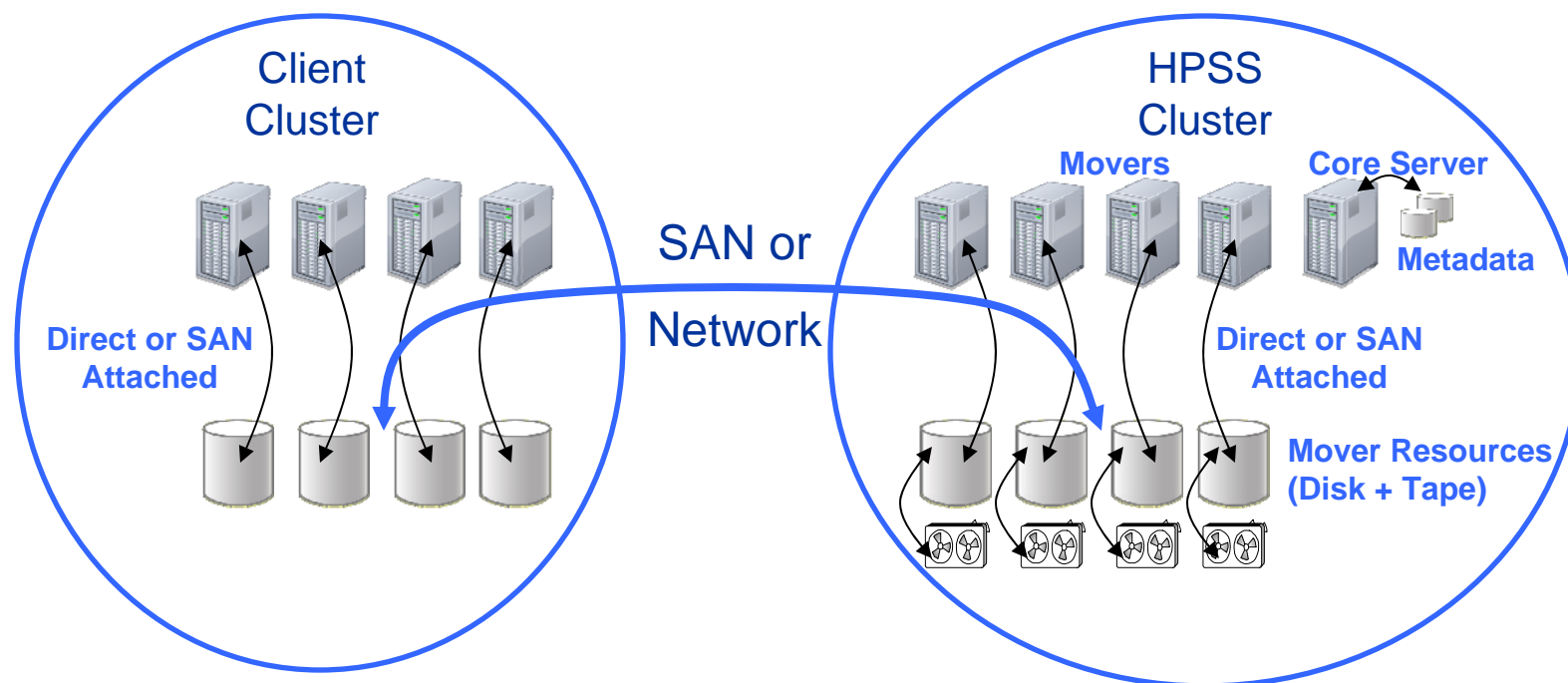


- Recently in production
  - Argonne National Laboratory (ANL)
  - High Performance Computing Center Stuttgart (HLRS)
  - Japan Aerospace Exploration Agency (JAXA)
  - United Kingdom Met Office (UKMO)
- Just starting or developing applications
  - Library of Congress (LoC)
  - National Center for Atmospheric Research (NCAR)
  - Pacific Northwest National Laboratory (PNNL)Northrop Grumman
- A class by itself:  
University of Illinois National Center for Supercomputing Applications (NCSA) and NSF Blue Waters supercomputer project
  - Production in 2011
  - Probably world's largest storage system
  - 10s to 100s of billion files
  - At least 500 tape drives



# HPSS architecture

- HPSS has a clustered architecture, unique among tape-based products
- HPSS scales horizontally almost without limits
- Horizontal scaling is easy by adding cluster components



# HPSS metadata architecture



- Under the hood, HPSS is powered by IBM's DB2 data base engine that scales to your performance needs!
  - DB2 metadata completely characterizes all files whether on disk, single tape, striped tape, mirrored tape, shelf tape, or multi-level hierarchies of disk and tape.
  - DB2 provides its own utilities for copying, mirroring, backup, restore, and verifying.
  - DB2 itself is highly scalable and reliable, bringing those capabilities to HPSS.
  - Enables fast restore after a disruption of service
  - DB2 is bundled with HPSS for this purpose, without extra charge

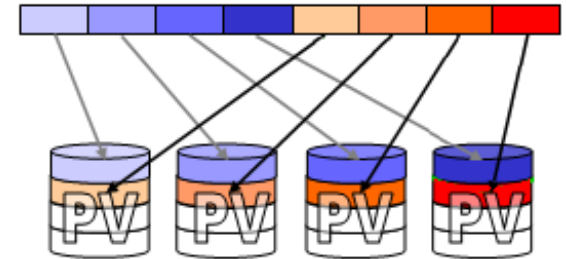


# HPSS tape architecture



- HPSS supports striping to disk and tape.

- At 100 MB/s, it takes almost 3 hours to write a 1 TB file to a single tape.
- Using an 8-way tape stripe, that time is cut to under 25 minutes!
- Software “RAIT” is being developed jointly by IBM and NCSA, to add up to 8+2 reliability to HPSS striping



- HPSS supports small file tape aggregation.

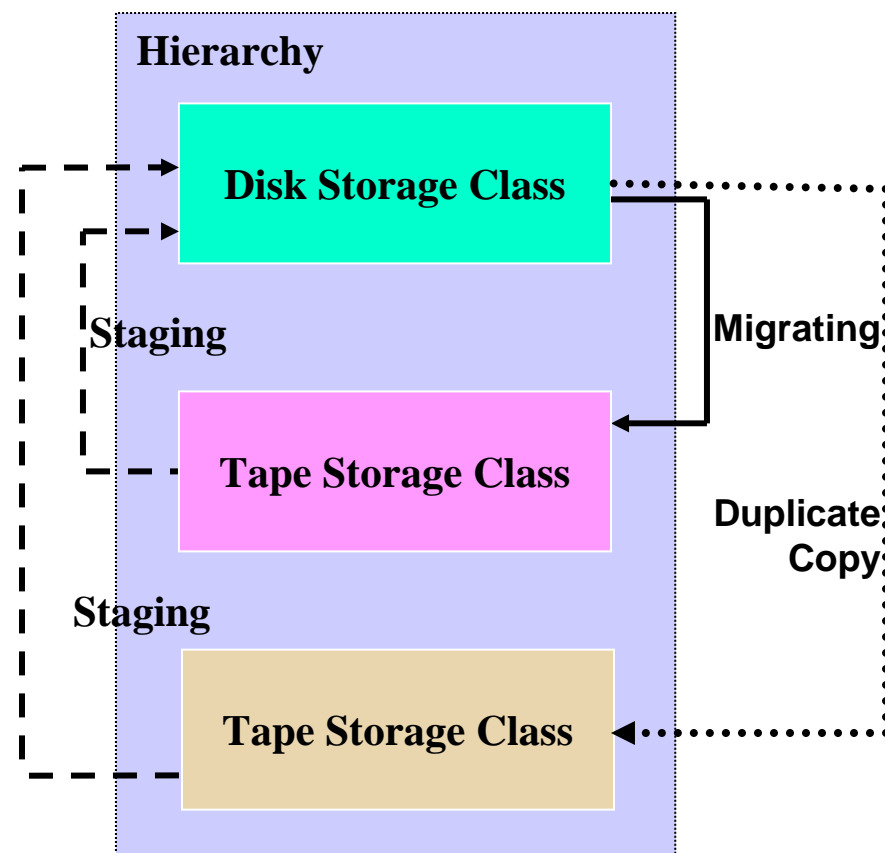
- Tape marks and repositioning for each small file write, prevents tape drives from streaming data to the tape – poor tape performance.
- By aggregating many, many small files to a single object, tape drives are kept moving for longer periods of time -- greatly improving tape drive performance!

- For very high data rate and very high data volume tape throughput, both striping and file aggregation are required.

# HPSS hierarchical architecture



- HPSS implements hierarchical storage management (HSM) using:
  - Storage classes
  - Storage hierarchies
  - Classes of service
  - Migration and purge policies
  - Automatic staging
- File data are automatically copied to lower levels.
- Typically, file data are automatically staged back to the top level, when accessed by the user.



**Example of a dual copy COS**

# HPSS native interfaces



- **Mountable POSIX-compliant file system:**
  - Linux users can mount HPSS using the Linux virtual file system (VFS) interface.
  - Used when applications needs a file system interface to HPSS for applications like DB2, Apache, SAMBA, etc.
- **Standard FTP and HPSS Parallel FTP**
  - Files are copied using `get` and `put` using standard ftp, or with `pput` and `pget` using HPSS' pftp (parallel FTP).
- **Application Programming Interface (API):**
  - POSIX-like (for example `read( ) ;` becomes `hpss_Read( ) ;`).
  - Supports parallel files, parallel servers, parallel clients.

# HPSS third party interfaces

---



- Interfaces available through HPSS using the Linux virtual file system interface:
  - NFS v3
  - Open SAMBA
  - Apache
  - Secure FTP

# HPSS third party interfaces, continued



- **GridFTP Enabled for HPSS**
  - Extension of the FTP protocol for the grid computing environment.
  - Open source Globus add-on developed by ANL.
- **HSI and HTAR**
  - Two interfaces developed by Gleicher Enterprises.
  - HSI, a widely used Unix-like third-party interface “shell”.
  - HTAR, a stand-alone utility used to aggregate files directly into HPSS.
- **NFS and LUSTRE-HSM bindings**
  - Two interfaces developed by CEA/DAM (France).
  - GANESHA, a multi-usage, large cache NFSv4 server.
  - LUSTRE-HSM bindings in development ([http://arch.lustre.org/index.php?title=HSM\\_Migration](http://arch.lustre.org/index.php?title=HSM_Migration)).
  - Contact: Jacques-Charles LaFoucriere - [jc.lafoucriere@cea.fr](mailto:jc.lafoucriere@cea.fr)

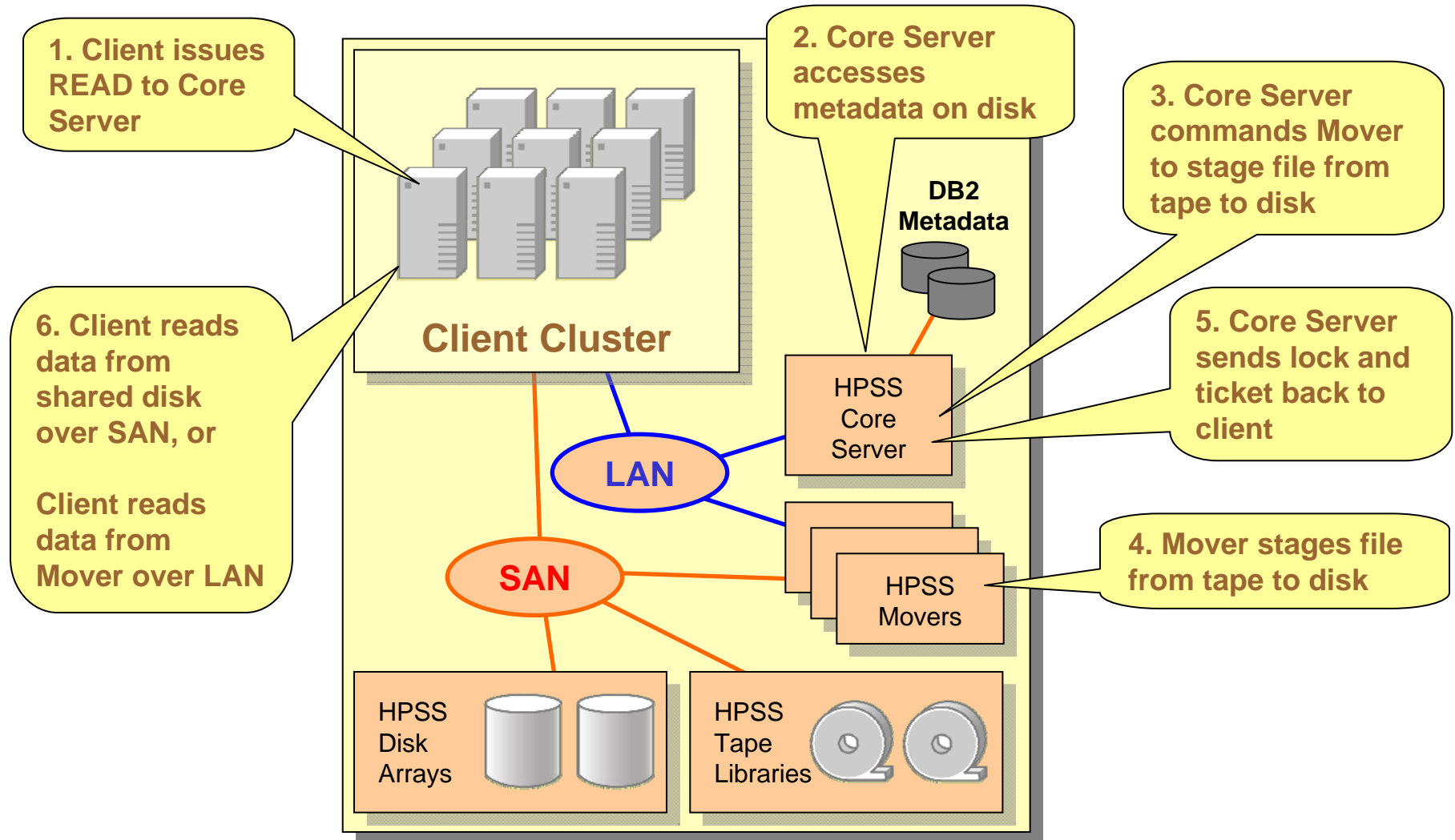
# HPSS Interface to IBM GPFS



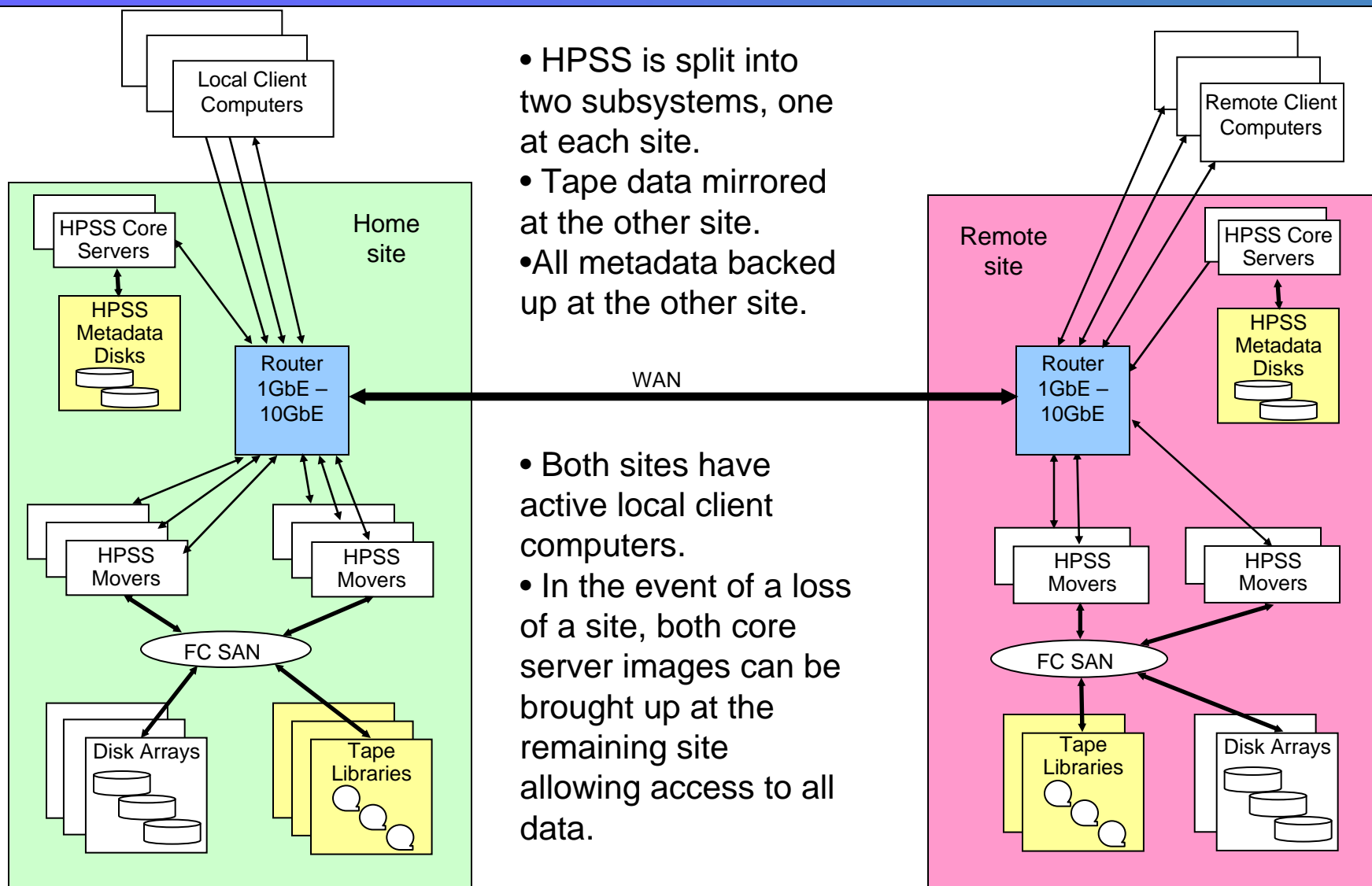
- GPFS is IBM's General Parallel File System
- HPSS can provide a highly scalable GPFS tape pool
- HPSS provides two tightly coupled services for GPFS:
  - Hierarchical space management of GPFS file space
  - Disaster recovery backup of GPFS file systems

# How HPSS works

Example of an hpss\_read



# HPSS at Multiple Sites



- HPSS is split into two subsystems, one at each site.
- Tape data mirrored at the other site.
- All metadata backed up at the other site.

- Both sites have active local client computers.
- In the event of a loss of a site, both core server images can be brought up at the remaining site allowing access to all data.

# HPSS Contacts

---



Jim A. Gerry – [jgerry@us.ibm.com](mailto:jgerry@us.ibm.com)

Harry Hulen – [hulen@us.ibm.com](mailto:hulen@us.ibm.com)

Patrick Schaefer – [pschaef@us.ibm.com](mailto:pschaef@us.ibm.com)

Bob Coyne – [coyne@us.ibm.com](mailto:coyne@us.ibm.com)

# Disclaimer



- Please obtain and read product documentation before deciding to acquire HPSS.
  - Documentation includes the HPSS License Agreement, the Statement of Work, and HPSS and other product manuals.
  - In case of conflict between information herein and product documentation, the documentation shall take precedence.
- Forward looking information including schedules and future product capabilities reflect current planning that may change and should not be taken as commitments by IBM.
- IBM may at its sole discretion discontinue, add, or change HPSS features and function without notice.