High Performance Storage System Interfaces

Harry Hulen

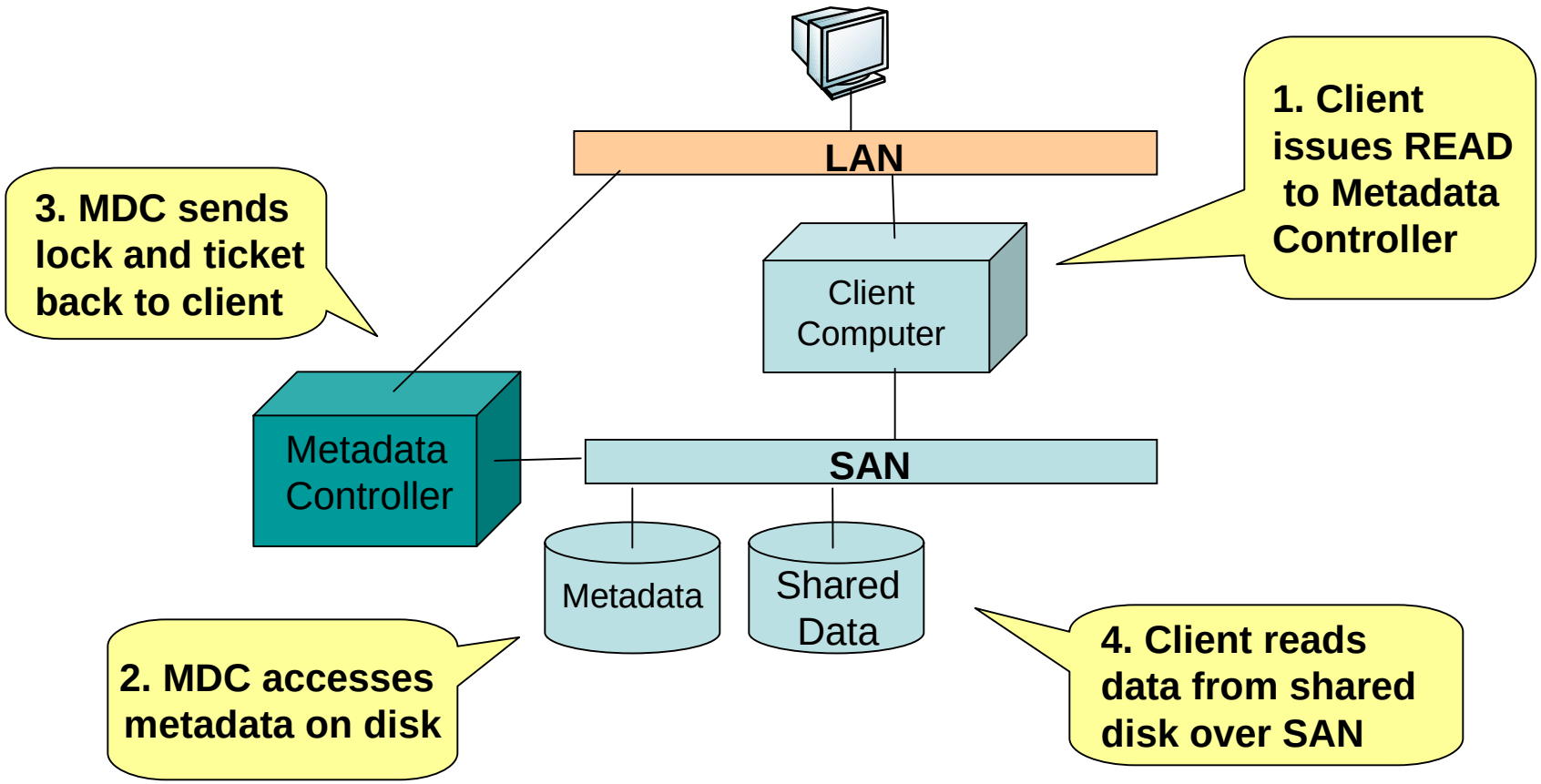
281-488-2473

hulen@us.ibm.com

Outline

- Introduction to HPSS
- HPSS Interfaces
- The HPSS Collaboration

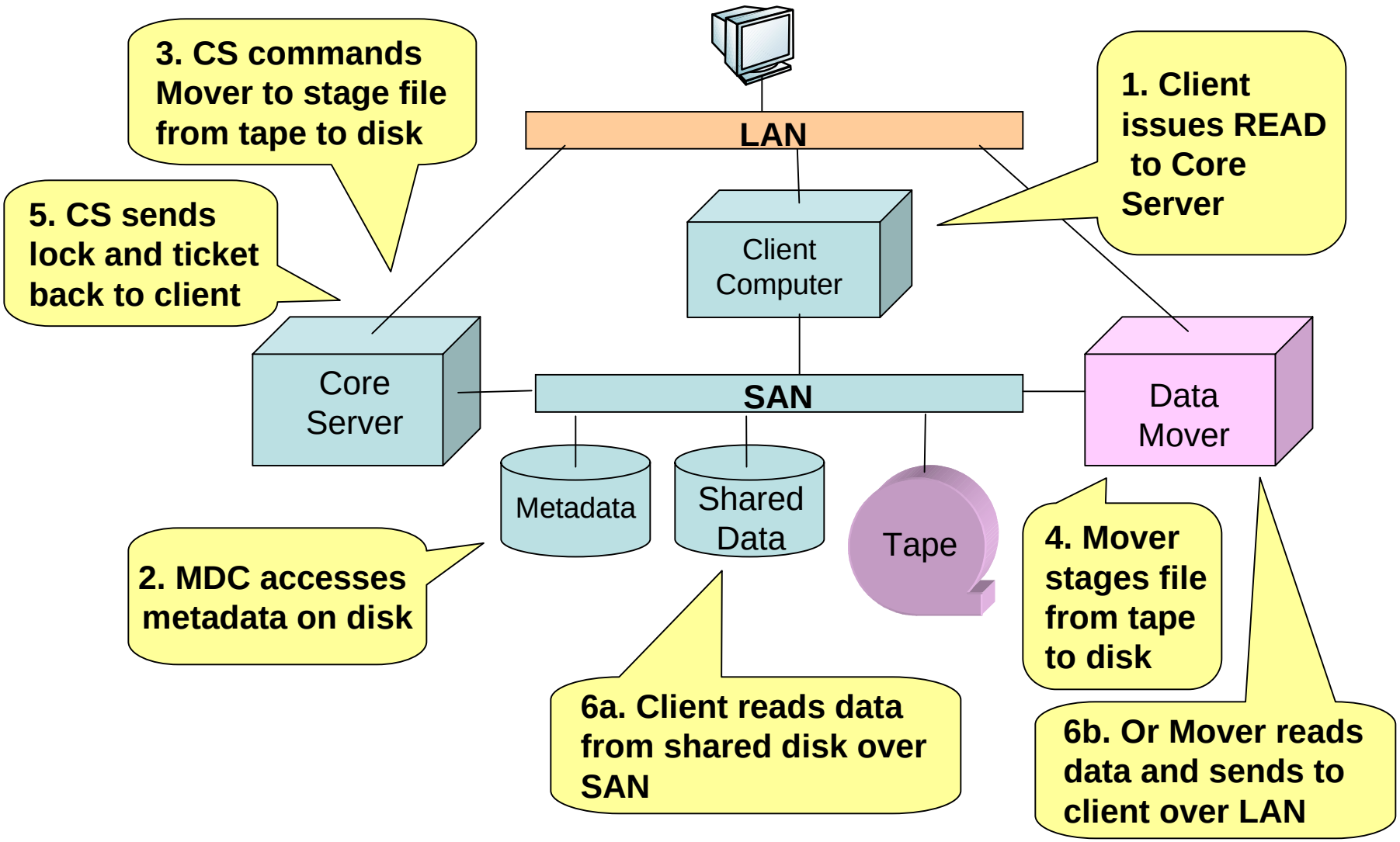
How a SAN File System Works



Examples:

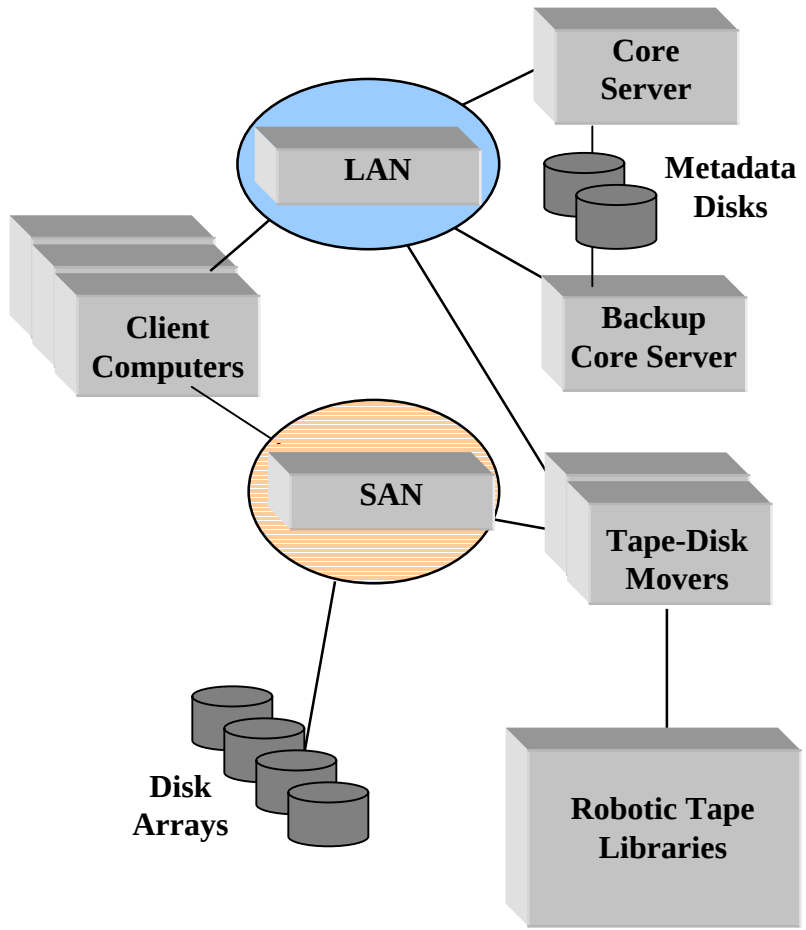
- IBM SAN FS
- ADIC SNFS
- HPSS is similar but adds tape (see next slide)

How HPSS Works



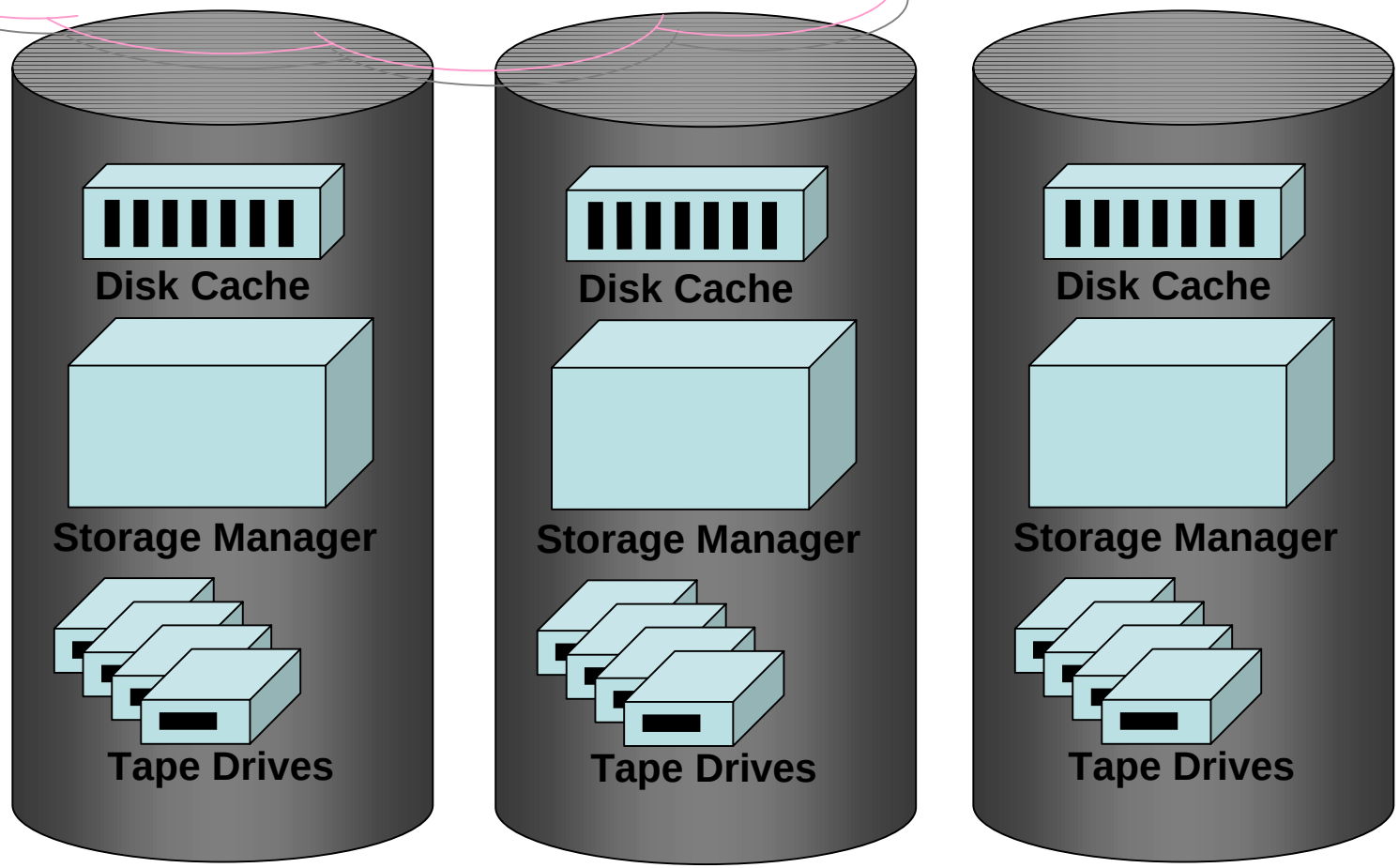
HPSS Architecture

- Hierarchical global file system
- Distributed, cluster architecture provides horizontal growth
- SAN and/or LAN connected
- Metadata engine is IBM DB2
- Multiple storage classes
- Striped disks and tapes for higher data rates
- Multi-petabyte capability in a single name space
- Supports IBM AIX, Linux, Sun Solaris, and some SGI Irix components, mix and match



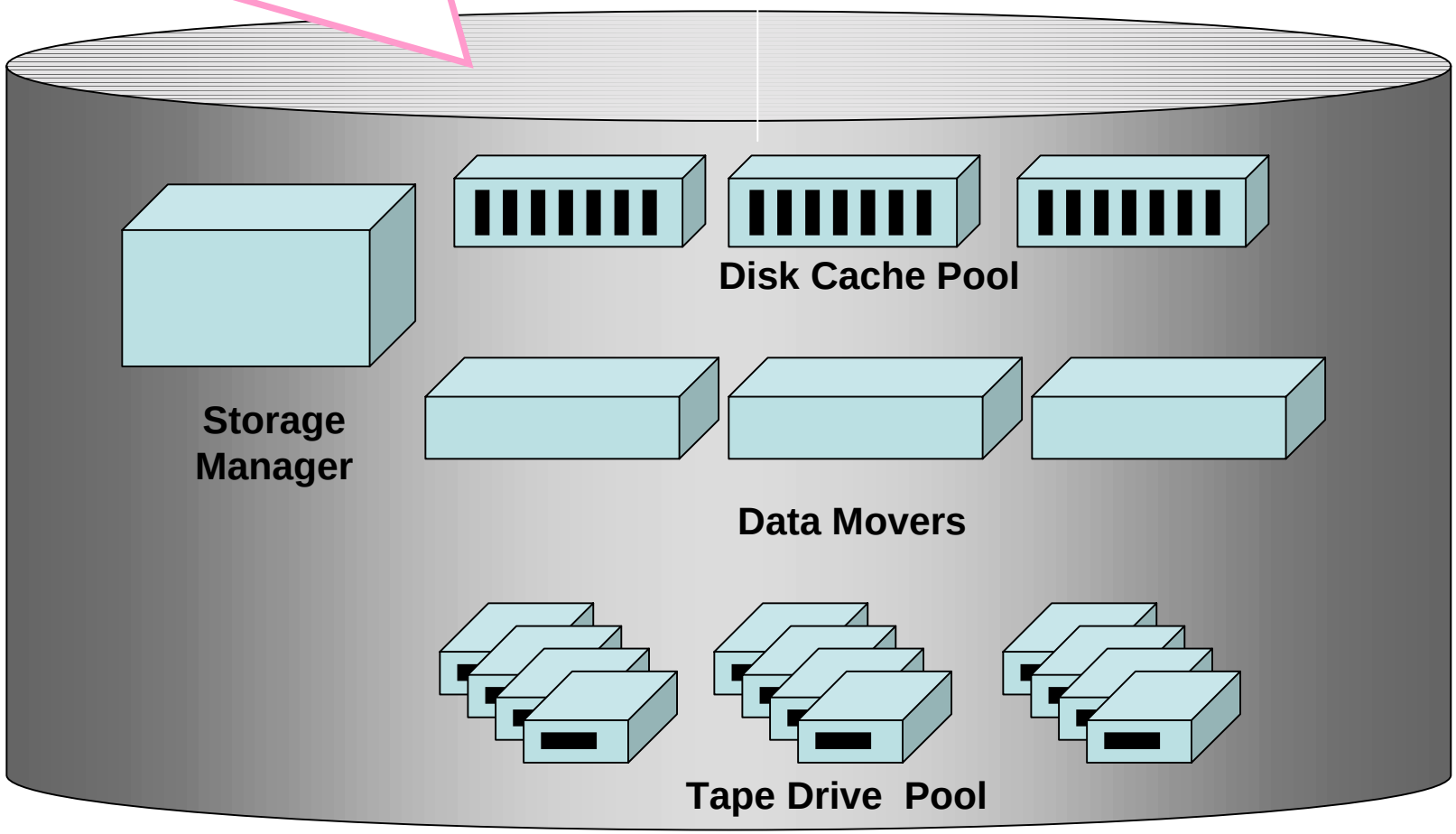
Multiple Storage Systems

- Shared nothing architecture
- *Inefficient* use of resources



HPSS Shared-Resource Archive

A single distributed storage solution with all resources shared



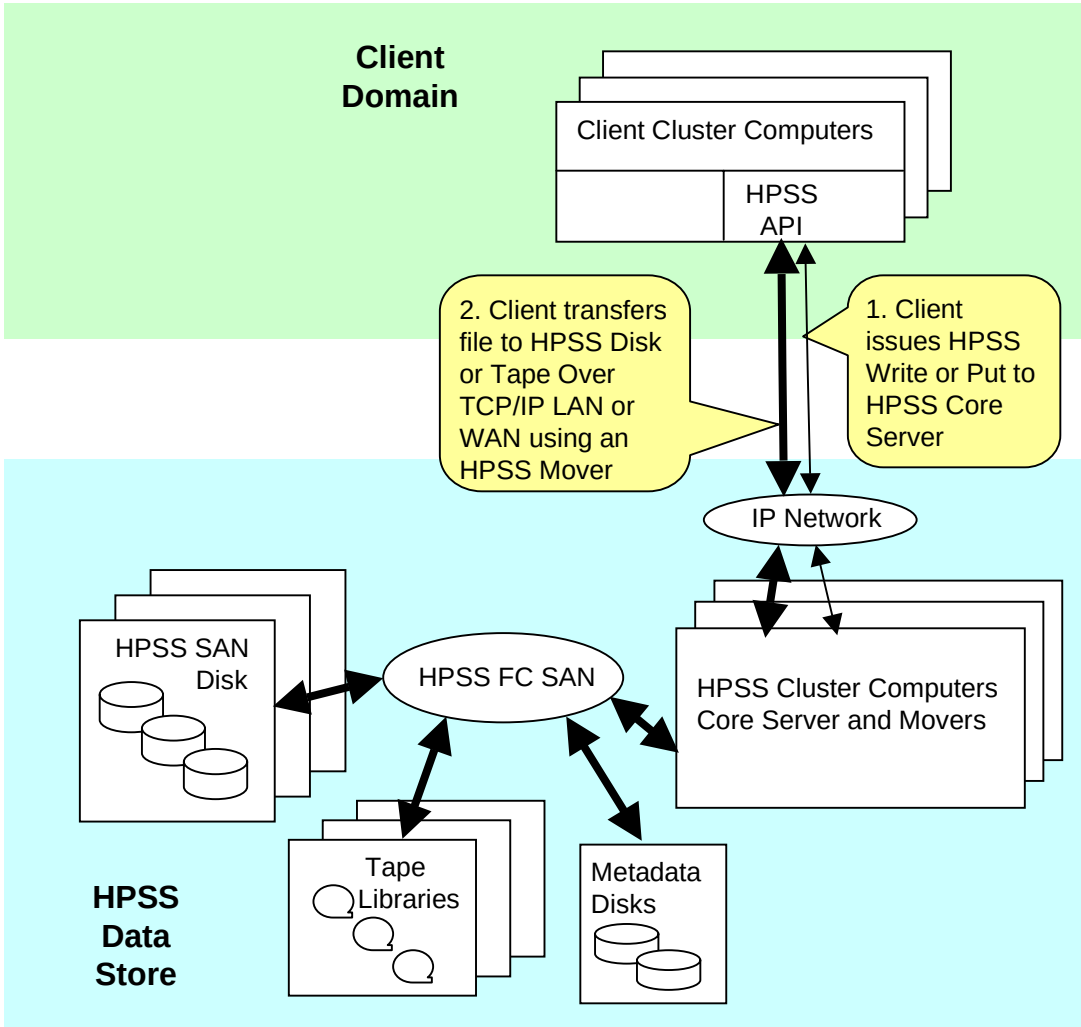
Outline

- Introduction to HPSS
- HPSS Interfaces
- The HPSS Collaboration

HPSS Interface Summary

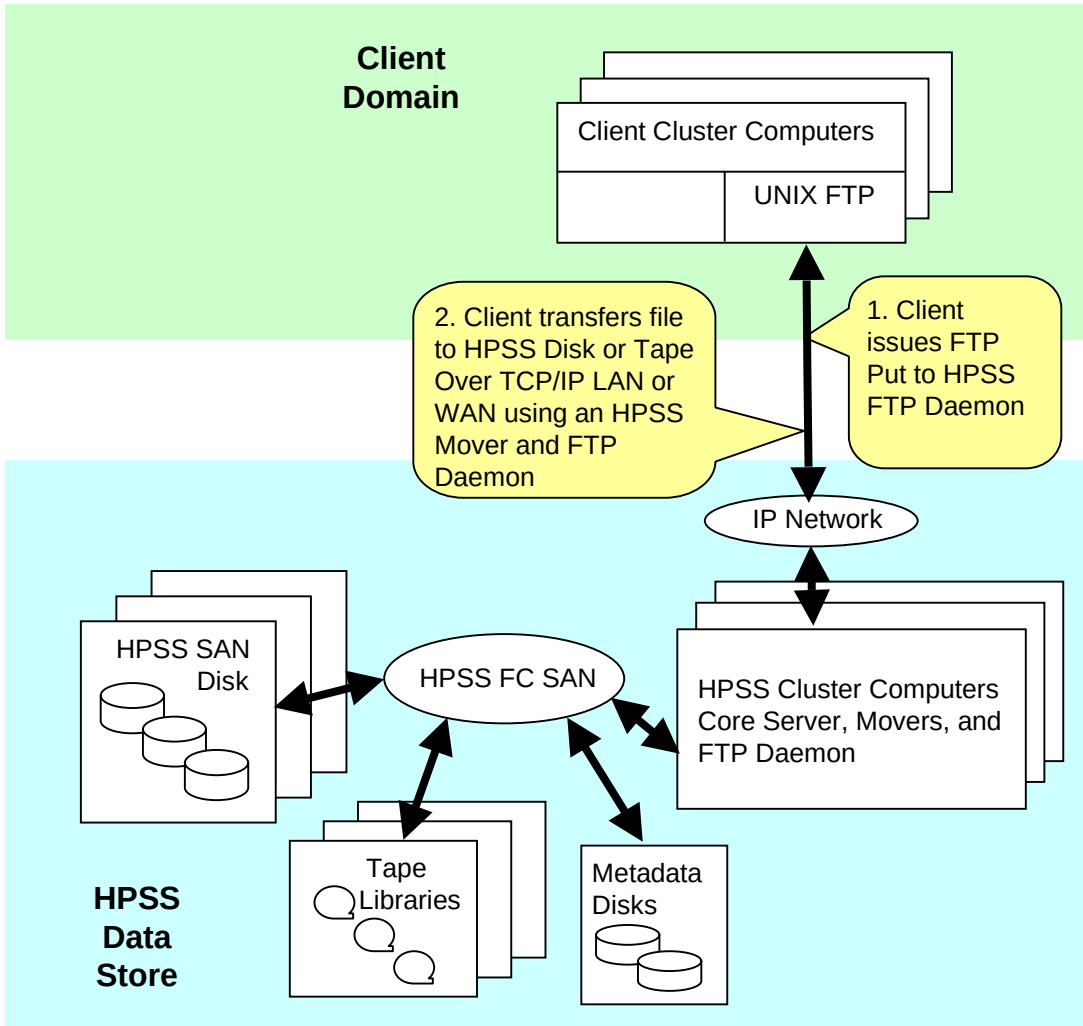
- The following charts illustrate these interfaces, with emphasis on data ingestion (transferring data into HPSS):
  - HPSS Write or Put Over TCP/IP Network
  - FTP Data Flow
  - SAN-Enabled HPSS Write or Put
  - Pull Data from Client SAN to HPSS Disk or Tape
  - POSIX VFS Interface
  - NFS Interface via VFS Agent
  - Windows Interface via VFS Agent
- Other possibilities:
  - HPSS Storage Interface (a useful 3rd party suite of easy-to-use interfaces)
  - Reading data from transportable media (”sneaker net”)

HPSS Write or Put Over TCP/IP Network



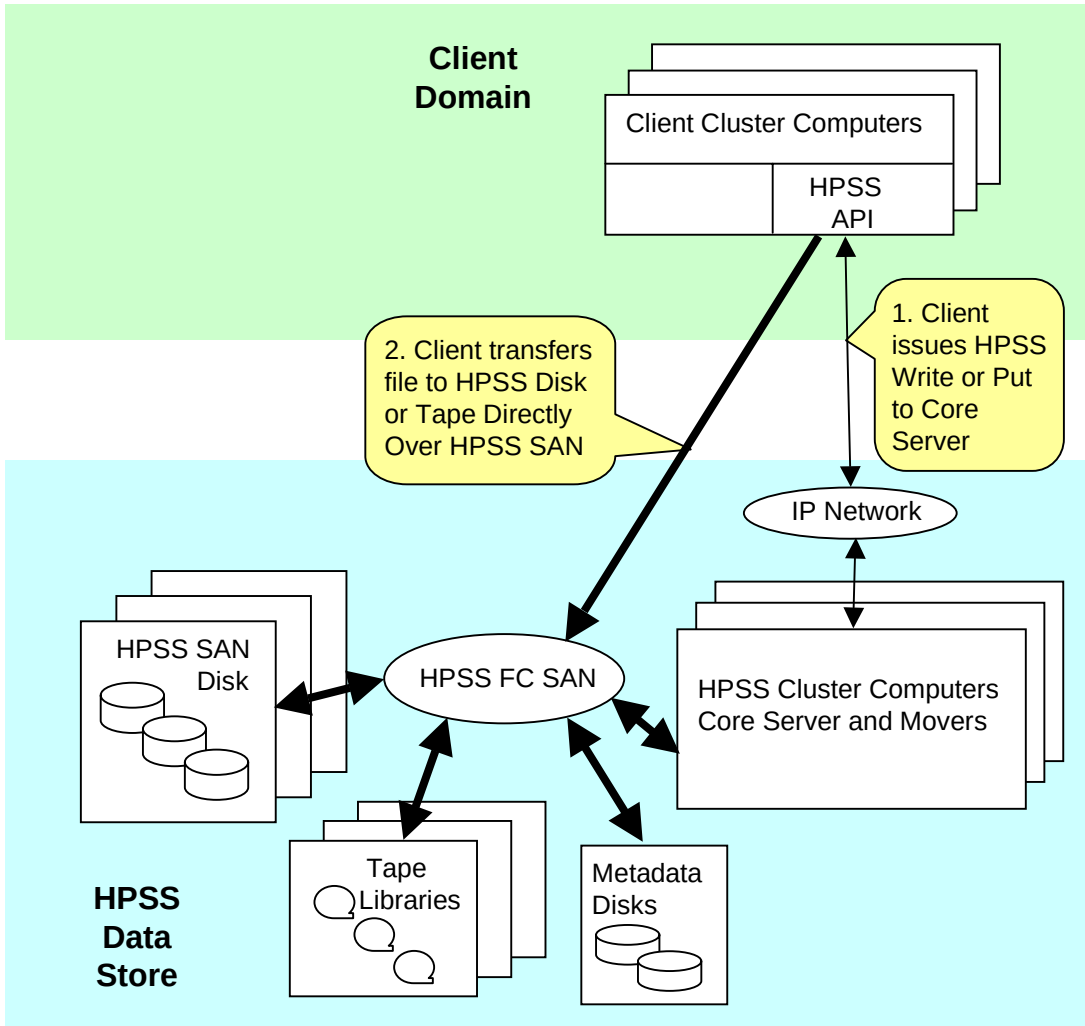
- HPSS Client API
  - Hpss_write( ) etc.
  - Optional list form to access discon-tiguous segments
  - Parallel, gigabyte/s capability
  - Use for performance-critical applications
- HPSS PFTP
  - Parallel FTP
  - FTP-like get-put semantics
  - Parallel, gigabyte/s capability
  - Most-used HPSS interface

FTP Data Flow



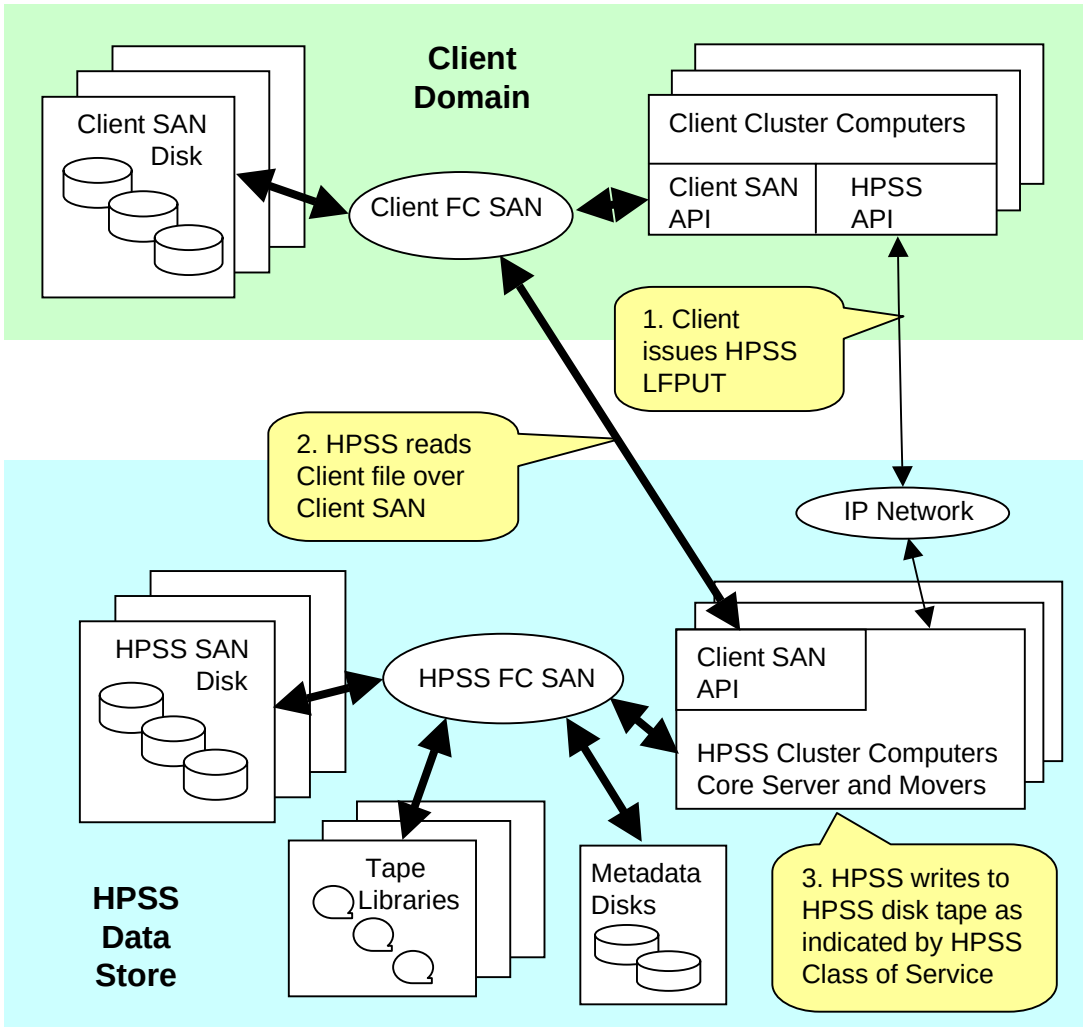
- Uses conventional UNIX, Linux, Windows ftp semantics
- No HPSS API to install
- Most universal HPSS interface
- Performance commensurate with ftp and underlying network
- Not a parallel interface

SAN-Enabled HPSS Write or Put



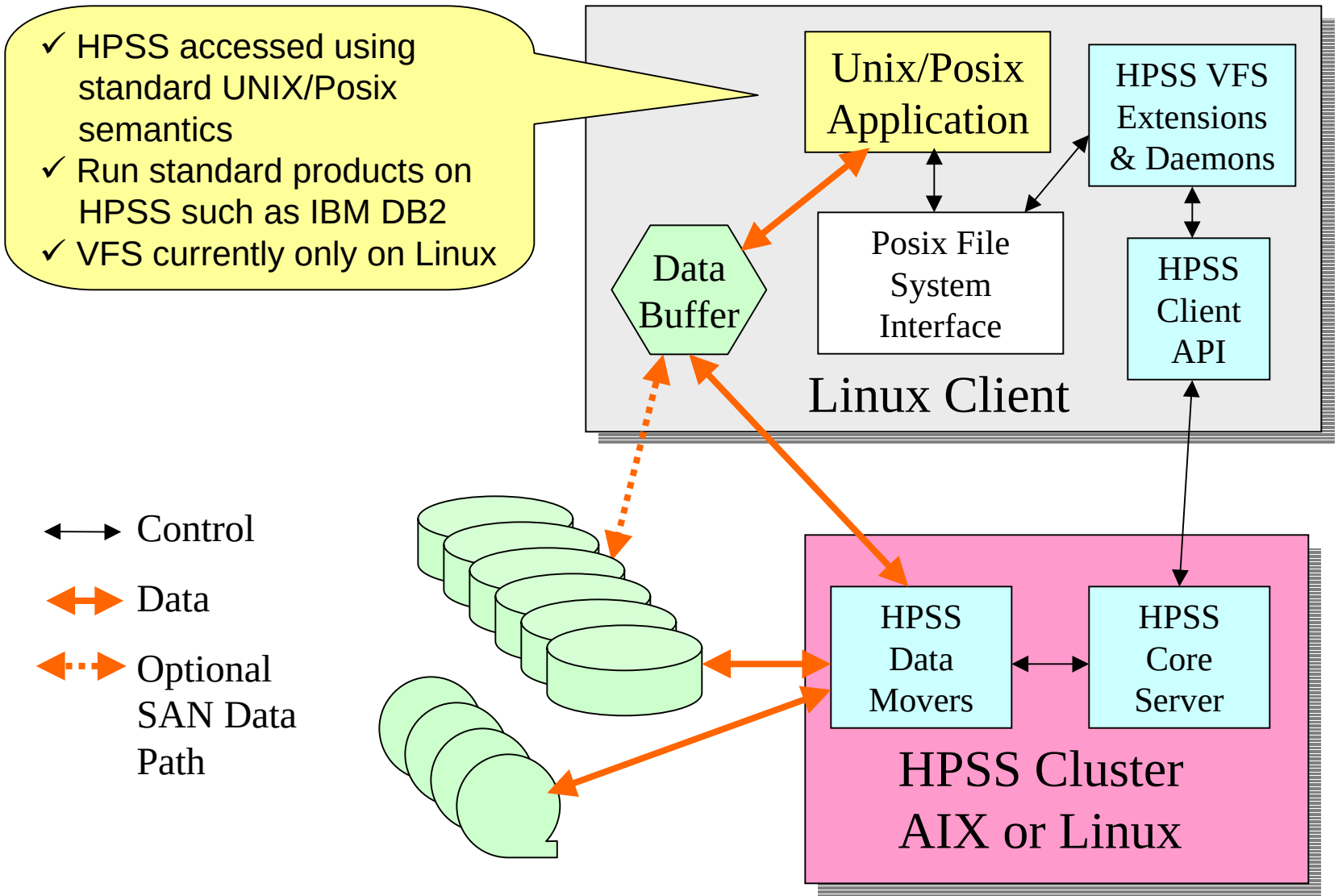
- Data transferred directly between client and HPSS disk over HPSS SAN
- Control is over TCP/IP network (separation of control and data)
- Supported by HPSS Client API and PFTP
- Currently supported on AIX and Linux
- Used internally to HPSS to move data between disk and tape

Pull Data from Client SAN to HPSS Disk or Tape

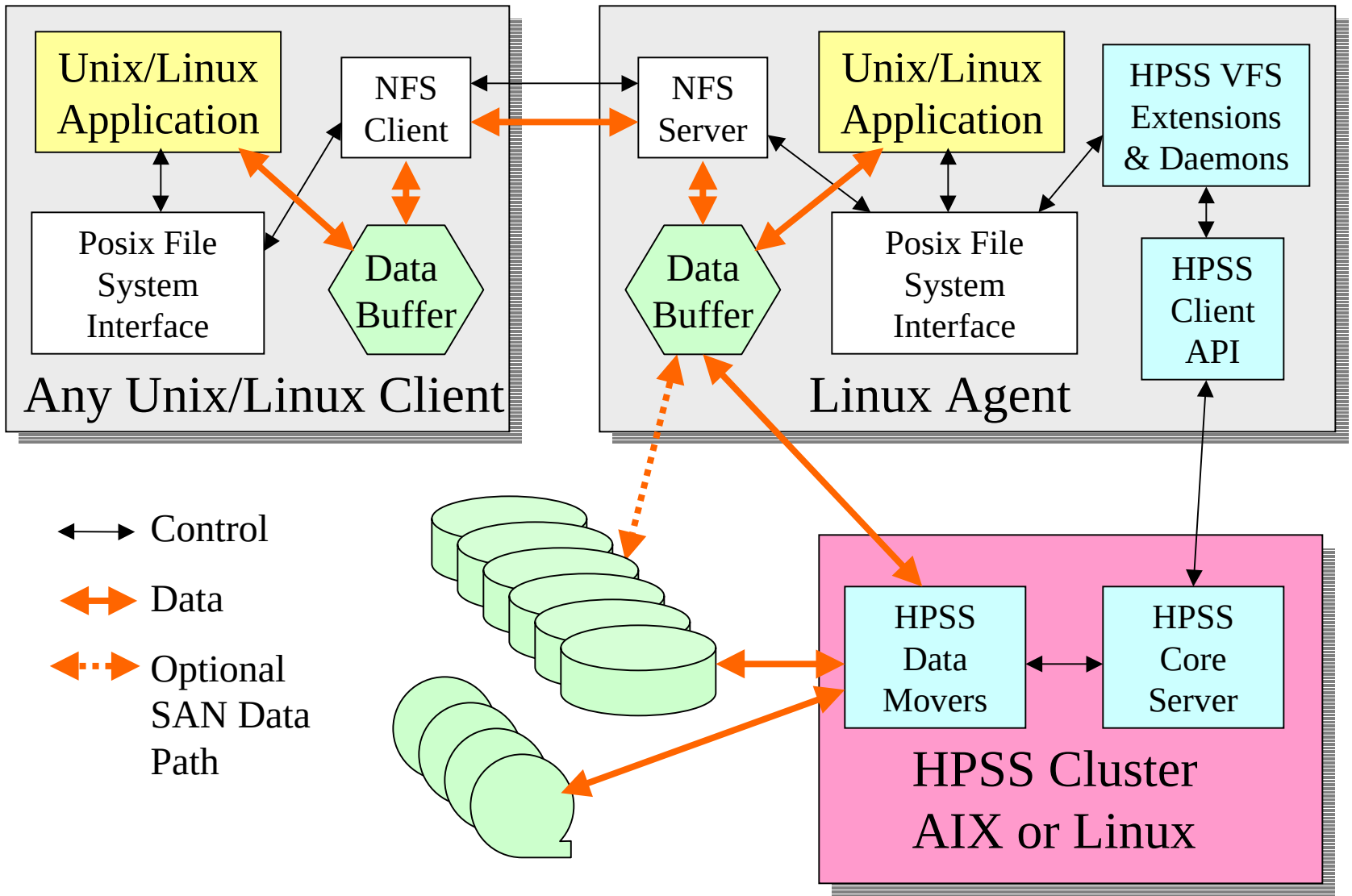


- "Local File Mover"
- HPSS accesses data on **client** SAN
- Examples of client SAN: IBM SAN FS, ADIC SNFS, IBM GPFS
- Activated by PFTP LFPUT-LFGET with more options coming
- CPU overhead entirely offloaded to HPSS Movers
- Parallel capability and/or direct tape access via Class of Service options

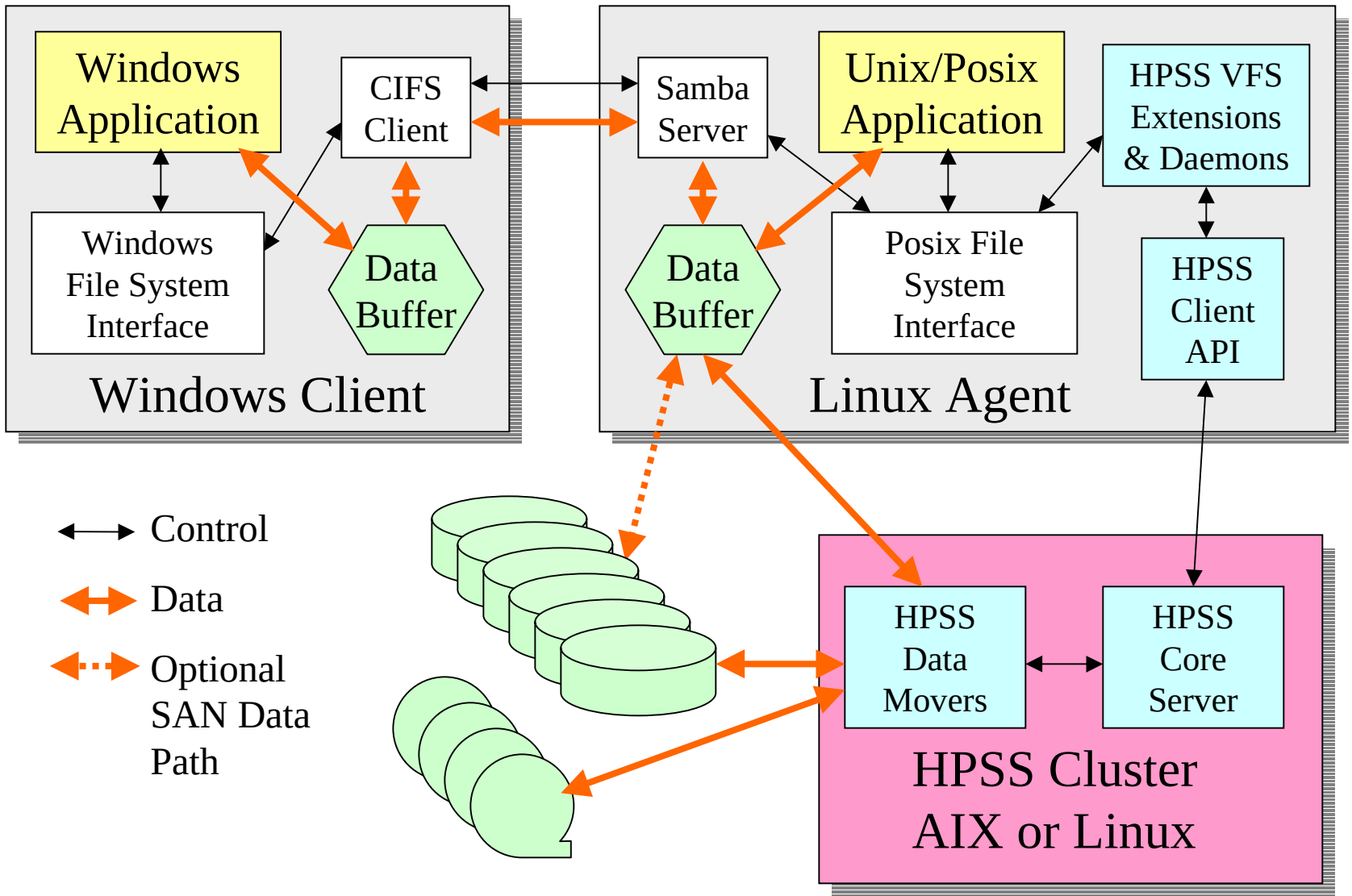
POSIX VFS Interface



NFS Interface via VFS Agent



Windows Interface via VFS Agent



Outline

- Introduction to HPSS
- HPSS Interfaces
- The HPSS Collaboration

The HPSS Collaboration

- U.S Department of Energy Laboratories are Co-Developers
  - Lawrence Livermore National Lab. - Sandia National Laboratories
  - Los Alamos National Laboratory - Oak Ridge National Laboratory
  - Lawrence Berkeley National Lab.
- IBM Global Services in Houston, Texas
  - Access to IBM technology (DB2, for example)
  - Project management
  - Quality assurance and testing (SEI CMM Level 3)
  - Commercial sales and service
- Advantages of Collaborative Development
  - Developers are users: focus on what is needed and what works
  - Keeps focus on the high end: the largest data stores
  - A limited “open source” model for collaboration members and users
- “Since 1992”

Some Large HPSS Sites

- 2+ PB: Brookhaven National Laboratory (BNL)
- 1+ PB: Commissariat à l'Energie Atomique/Direction des Applications Militaires (CEA/DAM) Compute Center in France
- 1 PB: The European Centre for Medium-Range Weather Forecasts (ECMWF) in England
- 1.1 PB: Lawrence Livermore National Laboratory (LLNL) open system
- 2+ PB: Los Alamos National Laboratory (LANL)
- 1+ PB: National Energy Research Scientific Computing Center (NERSC)
- 1 PB: San Diego Supercomputer Center (SDSC)
- 1.4 PB: Stanford Linear Accelerator Center (SLAC)