High Performance Storage System

Storing Exabytes of Data Across Trillions of Files

Jim A. Gerry
IBM Storage Architect

IEEE MSST 2013
07 May 2013

JGerry@us.ibm.com
713-797-4447

Disclaimer

Forward looking information including schedules and future software reflect current planning that may change and should not be taken as commitments by IBM or the other members of the HPSS Collaboration.

Outline

- HPSS collaboration
- HPSS features
- HPSS 7.4 architecture
- Current HPSS high end examples
- HPSS v8 scalability objectives
- More files, devices, data and clients
- Improving reliability, recoverability & availability
- HPSS user interfaces – what we are doing in the HPSS community
- To exascale and beyond with HPSS...

HPSS Collaboration

- In early 1990s there was no COTS or other scalable archive/HSM on the market or under development meeting projected requirements of Terascale HPC (now Petascale and tomorrow Exascale), and other large scale data intensive storage applications.
- The National Labs had extensive experience with building and deploying mass storage systems, IBM had been doing research related to these systems.
- IEEE Mass Storage Reference Model – modular, scalable, distributable developed in late 1980s, published in 1990 – work of IBM and other vendors, National Labs and other government labs.
- HPSS - a joint DOE Lab/IBM collaboration, utilizing all partner's strengths and experience.
- After twenty years, the collaboration is going strong.

"Since 1992"

University of California Lawrence Livermore National Laboratory

Los Alamos NATIONAL LABORATORY

Sandia National Laboratories

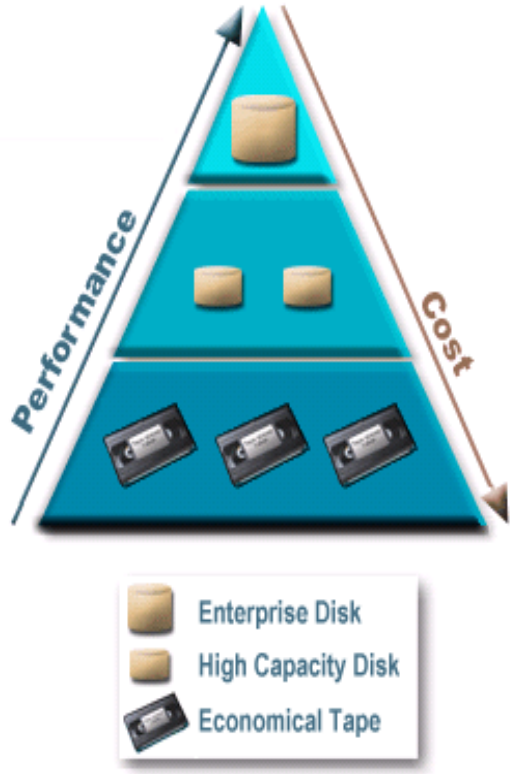
IBM

OAK RIDGE National Laboratory

BERKELEY LAB

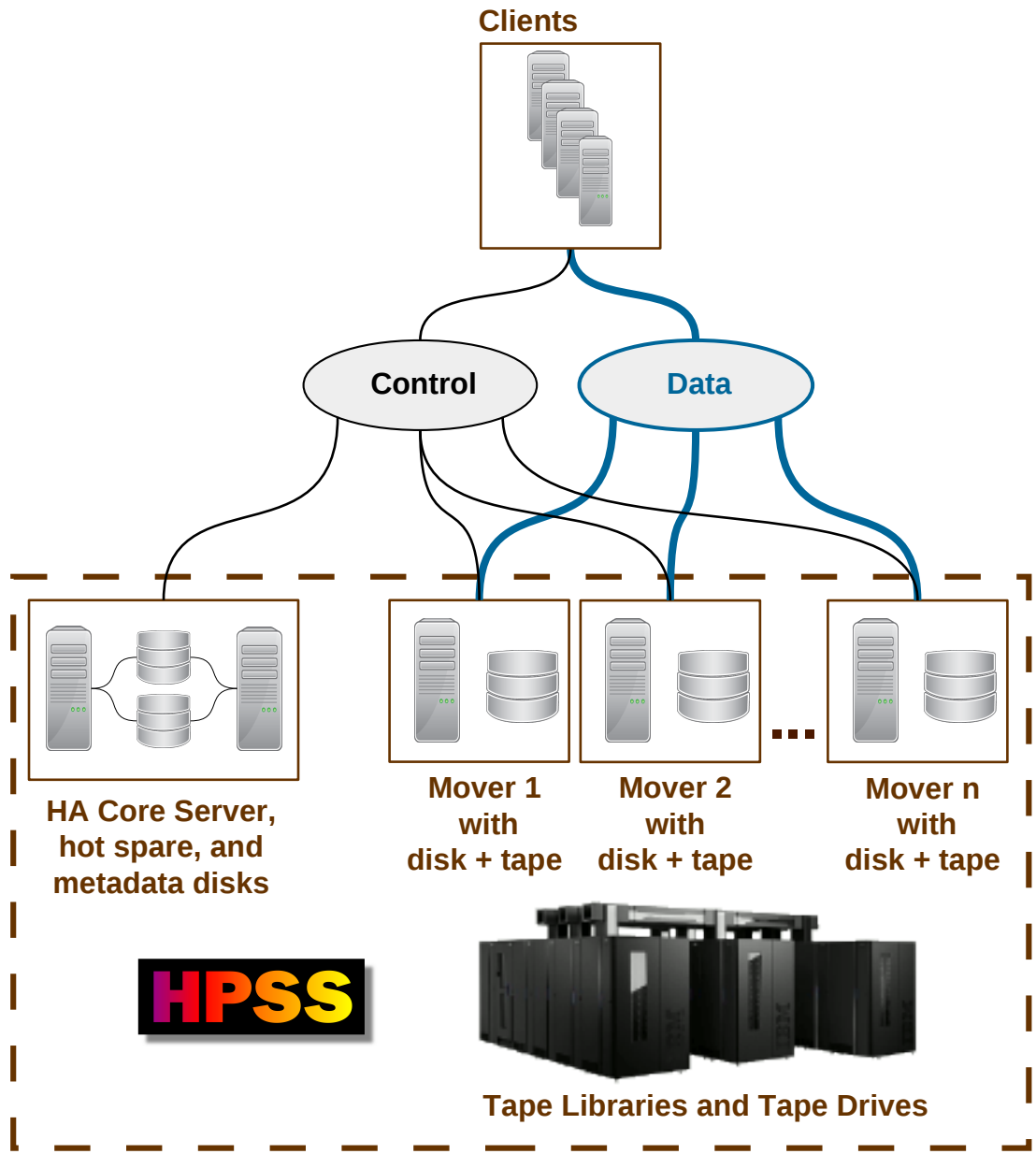
High Performance Storage System Features

- Hierarchical Storage Management (HSM)
  - Active data remains on disk, while dormant data rests on tape
  - Automatic movement of data up and down the hierarchies
- Metadata integrity
  - COTS DBMS, atomic transactions, mirroring, backups
- Data integrity
  - Tape copy (spanning geography), RAIT (RAID for Tape)
  - End-point checksum generation and validation
- Striping of data across storage resources at all levels for large file performance
- File aggregation for improved resource utilization
- HPSS management through graphical and command line interfaces
- Hardware vendor neutral
- Maximize use of COTS and industry standards
- HPSS is an IBM GBS Linux service offering that includes deployment and support services, as well as the HPSS and DB2 software licenses



HPSS 7.4 Architecture

- Single core server
  - Commercial enterprise database for metadata
    - Single partition database
    - File details, and locations
    - Configuration details
  - Manages data movers
  - Manages storage resources
  - Manages client connections
- Separation of control and data
- Distributed data movers and storage resources
  - Scalable capacity
  - Scalable bandwidth
  - Supports IP and SAN data transfers
- High availability using HA software and hot spare computers



Current HPSS High End Examples

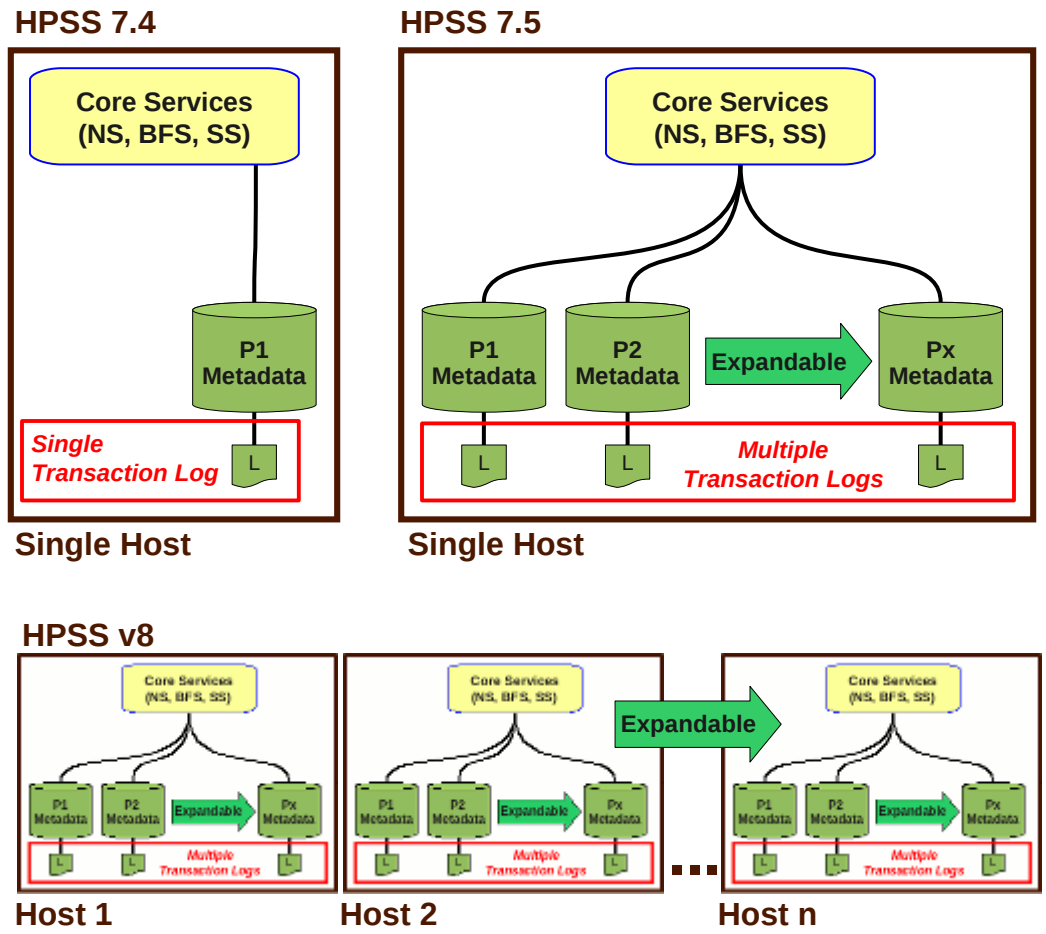
- The largest amount of data stored in a single instance of HPSS is 55 PB.
- The file-count leader in a single instance of HPSS has 384 million files.
- Today's largest HPSS system
  - HPSS capability testing was done with five billion files in the single namespace
  - HPSS DB2 metadata demonstrated over 2,700 file creates per second
  - Configured for 10,000 simultaneous connections
  - 50 data movers using dual bonded 40 gigabit Ethernet
  - Configured with 1-way tape, dual-copy tape, 4+P RAIT and 7+PQ RAIT
  - In the first three weeks of production they sustained 123 terabytes per day
  - Six year projections: 325 petabytes of data across 21.5 billion files
- Single file transfers to tape scales (TS1140 with non-compressible data)
  - 1-way tape @ 234 MiB/s
  - 2-way tape @ 469 MiB/s
  - 3-way tape @ 678 MiB/s
  - 4-way tape @ 915 MiB/s
  - 4+P RAIT @ 700 MiB/s
  - 7-way tape @ 1,555 MiB/s

HPSS v8 Scalability Objectives

Dimension	HPSS Scalability Objectives
Single Namespace Capacity:	
Storage Capacity	1.5 exabytes per year
Namespace Capacity	Over a trillion files
HPSS Clients	Tens of thousands simultaneous connections
HPSS Movers	Over a thousand data movers
Single Namespace Performance:	
File Operations	Over 40,000 file creates per second, with other file operations scaling to match
Daily Throughput	Over four petabytes per day (50 gigabytes per second sustained ingest)
Instantaneous Throughput	Terabytes per second burst
HSM Resource Utilization	Increase parallelism
Storage	Efficient use of tens of thousands of disk and tape
Redundancy, Reliability, and Availability:	
File Metadata	Continue to leverage commercial enterprise database
Data Integrity	End-to-end validation
High Availability	Hardware and metadata redundancy with automatic fail-over
Upgrades	Rolling upgrades of all hardware and software components
Other Dimensions:	
Client Interfaces	FTP (FTP, pFTP, gridFTP), File System (VFS, NFS, pNFS, GPFS, Lustre), Cloud
System Management	Improve the ability to monitor and control thousands of components
Content Management	Scalable user defined attributes that are searchable

HPSS Scalability Plans
More Files, Devices, Data and Clients

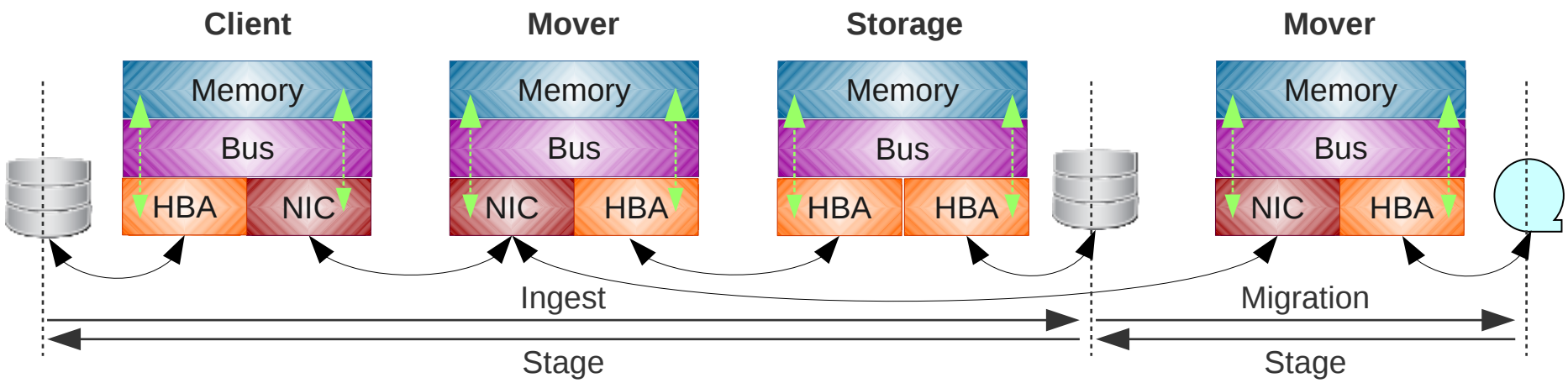
- **Increase parallelism of metadata and HPSS core services**
  - Increase parallelism of the metadata database
    - HPSS v7.5 - take advantage of DB2 partitioning
    - Initial testing demonstrated 4x improvement on a limited test platform over HPSS v7.4
  - Increase parallelism of HPSS core services
    - HPSS v8 - distributes metadata database and HPSS core services across multiple computers
  - Distributed core services, metadata, and movers = virtually unlimited HSM scalability



- **Improve disk-tape I/O management**
  - Smarter use of storage resources
  - Strategies to minimize the impact of device latencies
  - Partner with rule-oriented content managers

HPSS Scalability Plans
Improving Reliability, Recoverability & Availability

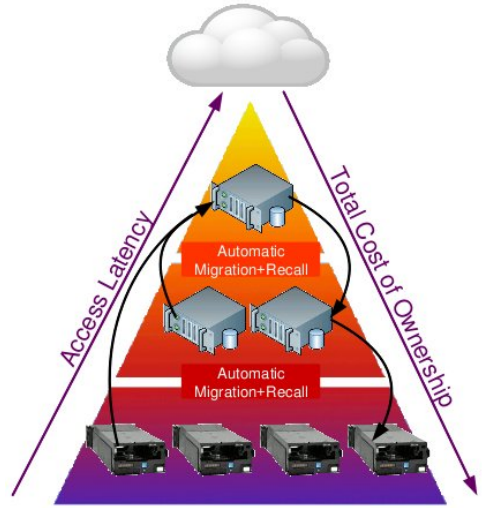
- **Support rolling upgrades**
  - Decouple HPSS from prerequisite software
  - Incremental upgrade without downtime (e.g. OS, DB2, HPSS, etc.)
  - Real time metadata conversion
- **HPSS must be high availability aware**
- **Improve data integrity - multi layered approach**
  - Network paths
  - Data blocks
  - Files

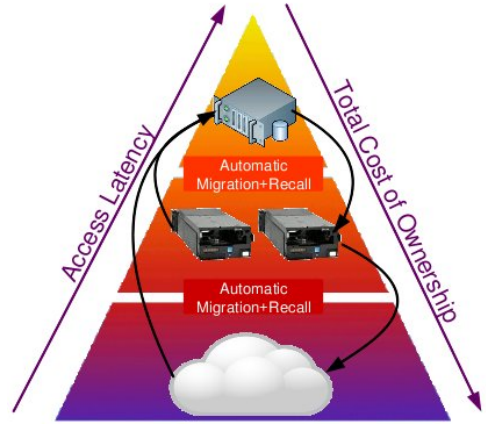


HPSS User Interfaces

What we are doing in the HPSS community

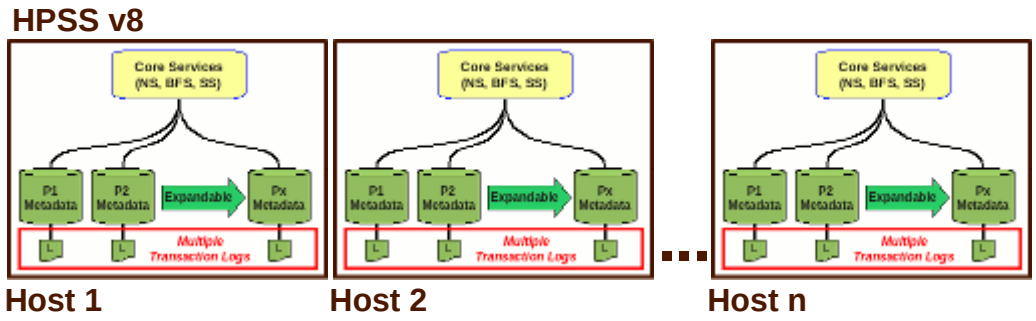
- FTP access methods
  - hpss-collaboration.org for FTP, and PFTP
  - globus.org for GridFTP
- File system access methods
  - hpss-collaboration.org for Virtual File System and GPFS-HPSS Interface (GHI)
  - sourceforge.net/apps/trac/nfs-ganesha for NFSv4
  - lustre.org for copy tool and Robinhood policy engine
- Embracing cloud technologies
  - Integration of HPSS with Mezeo and OpenStack Swift
    - mezeo.com for Mezeo Cloud
    - openstack.org for OpenStack Swift
  - Using a public or private cloud as a level in HPSS storage hierarchy is in work by IBM
  - Browser interfaces
    - globusonline.org for Globus Online
    - HPSS cloud interface service is in work by IBM
- IRODS data management (irods.org)
- Hierarchical Storage Interface and HTAR (gleicher.us)





To exascale and beyond with HPSS...

- **HPSS current architecture has scaled these dimensions by many orders of magnitude over the past 20 years.**
- **By increasing parallelism of HPSS core services and metadata, HPSS will scale these dimensions by several more orders of magnitude.**



- **HPSS will comfortably manage the HSM requirements for the exascale era and beyond.**
- **HPSS will manage the fastest forms of low latency storage, to the most economical higher latency technologies.**

Dimension of Scalability
Single Namespace Capacity:
Storage Capacity
Namespace Capacity
HPSS Clients
HPSS Movers
Single Namespace Performance:
File Operations
Daily Throughput
Instantaneous Throughput
HSM Resource Utilization
Storage
Redundancy, Reliability, and Availability:
File Metadata
Data Integrity
High Availability
Upgrades
Other Dimensions:
Client Interfaces
System Management
Content Management

Many Thanks!