HPSS Capabilities and Interfaces

High Performance Storage System (HPSS) is cluster-based software that provides for stewardship and access of many petabytes of data. When properly provisioned, HPSS is capable of concurrently accessing hundreds of disk arrays and tape drives for extremely high aggregate data transfer rates, thus enabling HPSS to easily meet otherwise unachievable demands of total storage capacity, file sizes, data rates, and number of objects stored.

HPSS has been used successfully for very large digital image libraries, scientific data repositories, university mass storage systems, and weather forecasting systems, as well as defense and national security applications.

Hierarchical Global File System

HPSS is hierarchical file system software designed to manage and access petabytes of data at high data rates. While appearing to the user as a disk file system, HPSS provides policy-driven tiered storage which is an important part of the data life cycle management practice. HPSS moves inactive data to tape based on data life cycle management policy and retrieves it the next time it is referenced.

Distributed, cluster architecture providing horizontal growth

The cluster aspect of HPSS combines the power of multiple computer nodes into a single, integrated storage system. The computers that comprise the HPSS platform may be of different makes and models, yet the storage system appears to its clients as a single storage service with a unified common name space.

Production multi-petabyte capability in a single name space

There are over 15 HPSS systems having one to ten or more petabytes of data in a single namespace. A list of large sites and their reported size in petabytes and number of files is posted on the HPSS web site www.hpss-collaboration.org.

True SAN and virtual SAN capabilities

HPSS provides both SAN access to disks and efficient network access to disks and tape devices using HPSS Mover nodes to create SAN-like access over cost-effective TCP/IP LAN or WAN networks.

HPSS virtualizes disks and tape drives so that they appear to the user to comprise a single POSIX-like file system. This virtualization may be for a single level of tape or a hierarchy of disk and tape. The virtualization may take place over a true SAN or a virtual SAN over a LAN, with most HPC systems opting for the LAN approach.

Metadata engine with IBM DB2

HPSS uses IBM DB2 as its internal metadata repository. There is only one metadata "engine" within an HPSS subsystem (although it maybe fully redundant), and this metadata engine completely characterizes the files whether on single tape, striped tape, mirrored tape, shelf tape, fast disk, high-capacity disk, or any combination of Storage Classes. DB2 is as robust as any commercial software in the world and is the product of a large development staff and a large support staff. DB2 provides its own utilities for copying, mirroring, backup, restore, and consistency checking. DB2 is distributed with HPSS at no additional charge for use as the HPSS metadata engine.

Inodes are not required

There are no inodes, stubs, or other metadata on disk. This is an important part of the reliability and availability of HPSS. When systems with inodes lose a disk, not only is data lost but also the metadata that describes that data is lost. In HPSS, since all metadata is kept in DB2, loss of a disk only causes loss of data, and the metadata is intact. The integrity of the file system structure is thus preserved, and often HPSS can retrieve the lost data from another level of the hierarchy.

Multiple Storage Classes and Classes of Service

HPSS provides multiple pools of storage devices called Storage Classes. As storage devices are added, new Storage Classes can be configured. Storage Classes are organized into Classes of Service. HPSS files reside in a particular HPSS Class of Service which users are able to set or the system can select based on parameters such as file size and performance. A Class of Service is implemented as a storage hierarchy consisting of multiple Storage Classes and rules for migrating data among them. Storage Classes are used to logically group storage media to provide storage for HPSS files.

A Class of Service may be as simple as a single tape, or it may consist of up to five levels of disk and tape. The user can even set up classes of service so that data from an older type of tape is subsequently migrated to a newer type technology. Such a procedure allows migration to new media over time without having to copy all the old media at once.

Striped disks and tapes for higher data rates

HPSS enables disks and tapes to be striped to create files that can be accessed at high data rates through parallel I/O operations. With 16-way striping, single file disk data rates of over two gigabytes per second have been achieved.

Heterogeneous computer support

The full suite of HPSS server software runs on IBM's System p computers with AIX and on System x and System p computers with Red Hat Enterprise Linux. Mover computer nodes may be RHEL, AIX, or IRIX, and client computer nodes may be RHEL, AIX,

IRIX or Solaris. Other platforms may be served with ftp, Secure ftp, gridFTP, NFS, and Samba, as described later in this paper.

IBM provides services to port clients and Movers to other versions of Linux. IBM supports HPSS Collaboration Members' efforts to provide nonrecurring and/or recurring tasks to port and maintain Movers and clients on additional operating systems and versions thereof.

Many tape libraries supported

Enterprise-class tape libraries from IBM, Spectra Logic, Sun StorageTek, and Quantum are supported by HPSS. Furthermore, HPSS supports generic SCSI Tape Libraries that use the SCSI-3 command set. For new SCSI Libraries, HPSS requires documentation of the output from the SCSI "inquiry" command. HPSS best practice is to allow on-site test time for new types of libraries.

Most enterprise tape drives supported

Most current IBM, HP, and Sun StorageTek tape drives and media types are supported, as is Sony AIT. HPSS can support multiple types of tape in one system using Storage Classes and Classes of Service.

HPSS easily accommodates insertion of new tape technology

As new types of digital storage technology are configured into the system, the HPSS Storage Class definition may be updated to the new device and media characteristics. Existing data is accessed normally, but all new media migrations will use the updated definitions and new media. Migration of existing data to the new tape media is accomplished via the Repack utility. The utility is under the control of the administrator and can be automated to run as often as the administrator desires. This process has been used at most HPSS sites in the management of the data lifecycle as customers have upgraded their media technology.

High performance data grid interfaces

For wide-area applications, HPSS includes PFTP, a parallel, multi-threaded, TCP/IP-based service with syntax similar to ftp. PFTP has achieved long-distance file transfers of 200 megabytes per second between Department of Energy national laboratories.

VFS on Linux

Red Hat Enterprise Linux applications benefit from a true POSIX standard read-write interface. This interface allows many standard commercial programs and legacy locally-written programs that include POSIX-type read-write file I/O to use HPSS as a file space, making these applications into hierarchical disk-tape applications.

With this interface, Linux applications may mount HPSS at any point in the directory tree, much as NFS would be mounted. For fine-grain access, large files may then be

copied between the HPSS part of the file tree and a conventional file system using standard commands such as cp (copy).

With this procedure, HPSS can easily be used to back up files from POSIX-compliant Linux file systems such as GPFS, Lustre, and ext3. For application systems other than Red Hat Linux, the HPSS Client API or PFTP may be used, or NFS and SAMBA may be used as described below.

Commercial and Open Source Interfaces Using VFS on Linux

Commercial and open source interfaces may be exported from a computer system employing the Linux VFS interface. Examples are NFS, Samba, Apache, and Secure ftp. HPSS services support the VFS interface to these applications, but the applications themselves are supported by the provider of the interface, whether open source or commercial.

As an example, Samba is in production with HPSS at Indiana University with performance reported in the range of 15 to 20 MB/s. IBM supports the VFS interface to Samba at IU, and Samba is supported through the open source community.

HPSS can support virtually any number of tape libraries in a single system, and the libraries may be of different types and different manufacturers.

Direct SAN access to HPSS disks

In addition to the original Mover-based disk sharing, disks may now be accessed directly over a SAN. This may be done with the HPSS Client API, PFTP, and the Red Hat Enterprise Linux Virtual File System interface. This capability is in addition to the popular HPSS virtual SAN architecture, where block-level data is transferred over less expensive IP networks.

SAN support for disks belonging to other files systems

HPSS can transfer data between a disk-only file system and HPSS over the disk file system's SAN. HPSS Movers serve as the transfer agents using an HPSS interface called Local File Mover (LFM). This is sometimes called a "reverse SAN" approach. The disk file system's POSIX-compliant read-write API must be installed on each HPSS Mover. This capability could be advantageous for a tape-only HPSS installation that is IP-based. By placing a disk file system client on each Mover computer and enabling the Mover computer for SAN access, one store-and-forward "hop" can be eliminated in transferring a file to HPSS tape. In theory, LFM works with any POSIX-compliant cluster file system. It is regularly tested using NFS as the other file system and has been informally tested with GPFS, SNFS, and Lustre clients. Additional system engineering and testing would be recommended to confirm the efficacy of this powerful capability for any site considering its use.

GridFTP

The open source Globus gridFTP is a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks. Software to support gridFTP with HPSS was developed and is offered as open source by Argonne National Laboratory. The first production use has been achieved at Indiana University. GridFTP uses the PIO client interface of HPSS. PIO is supported as a standard HPSS service, and gridFTP is supported through its open source channel.

GPFS HPSS Interface (GHI)

The GPFS/HPSS Transparent HSM (GHI) capability enables HPSS to migrate files from GPFS, IBM's most powerful shared cluster file system. With the release of HPSS 6.2.2, the combination of GPFS and HPSS will provide a virtually infinite high performance file system with robust disaster protection. GHI is an additional feature not included in the standard HPSS offering.

Available and unrestricted data format information

Some file systems and hierarchical storage systems are licensed under rules, or are designed to disallow, physically or logically migrating data out of the system in bulk. The HPSS Collaboration members believe that the ability to efficiently take your data back out of HPSS is fundamental, should you decide to move your data to another file system or hierarchical storage system in the future. This is particularly critical because copying individual files to effect data migration between file systems and storage systems become increasingly painful and in many cases impractical. Utilities are available to list the files stored on each tape, including such information as date and time stored, number of segments, and the location of each file or segment. This information can be used to find and read HPSS files from tape without using HPSS.