High-Performance Storage Systems

Answering the Data Explosion with Massive Scale and Compelling Economics

A White Paper by Jeskell Systems

Executive summary

Data is exploding at an unprecedented rate. Organizations in fields such as scientific research, genomics, aerospace, weather modeling, oil and gas, and national intelligence soon will generate and analyze data volumes in the tens of petabytes and potentially hundreds of petabytes by 2020.

The challenge is how to store, archive, and manage such massive data volumes efficiently and cost-effectively. This white paper details why the High-Performance Storage System (HPSS) is an ideal solution. The paper explains how HPSS works, including its advantages to organizations wrestling with huge data volumes. For example, HPSS is vendor-agnostic, affording organizations great flexibility. It also scales easily to tens and hundreds of petabytes while keeping costs under control.

Finally, the paper summarizes the deep qualifications and value that Jeskell brings to HPSS projects. This includes decades of experience in storage and big data, combined with seasoned consulting and implementation expertise to quickly deliver a complete, multivendor HPSS solution.

With the information shared in this paper, organizations will better understand how HPSS manages massive amounts of data, and they'll gain confidence to deploy HPSS aided by Jeskell's expert guidance.

Introduction

Anyone with a smart phone knows how quick and easy it is to snap pictures that capture nearly every moment of our lives. But after that moment, how often do you go back and look at those images? Yet they remain saved, taking up storage space, as new photos are constantly added.

For five or ten gigabytes of vacation photos, the cost of storage is not huge. But picture an organization conducting massive research projects involving not gigabytes or even terabytes, but many tens of petabytes and beyond. Data is often saved for years for historic preservation, compliance, or long-term trend analysis.

These are common data retention requirements for organizations such as academic and private research laboratories, government entities like the Department of Energy and intelligence agencies, complex manufacturing operations such as aerospace, as well as process industries such as oil and gas or pharmaceuticals. Consider that a single mass spectrometer machine generates three terabytes of data per experiment. Intelligence gathering organizations could generate terabytes of data each month from one video surveillance camera, let alone thousands of global monitoring systems.

Just a couple of decades ago, a large technical-computing environment might have amassed 500 terabytes; today, that figure could easily be 20 petabytes. With IDC projecting the digital universe to reach 40 zettabytes by 2020, individual organizations may be wrestling with hundreds of petabytes, if not exabytes.

How can organizations manage such extraordinary amounts of data securely and cost-effectively while enabling ready access to data that are typically only needed occasionally? For more than two decades, IBM and five U.S. Department of Energy national laboratories have collaborated on a powerful solution: the High-Performance Storage System (HPSS).

The ins and outs of HPSS

HPSS software provides highly scalable hierarchical storage management (HSM), archive, and file system services. Architecturally, it employs disk cache for staging data ingested from or presented to the active technical computing cluster, robotic tape libraries for long-term data retention, and server gateways, called Movers, that manage movement of data between disk cache and tape. (Figure 1)

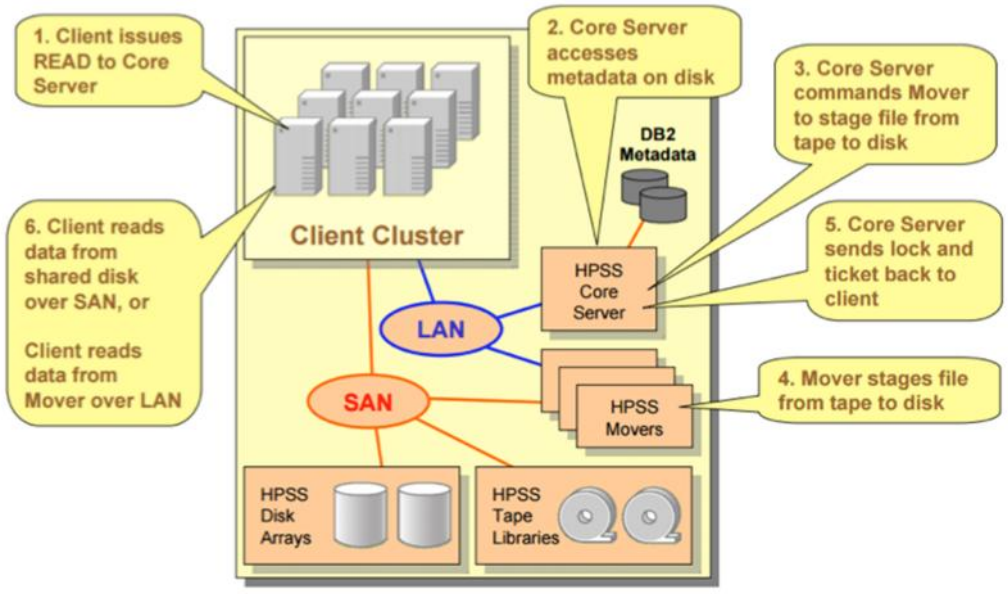

Figure 1 – HPSS Architecture

HPSS manages the data lifecycle by moving inactive data to tape and retrieving it the next time it is referenced. Organizations also can create policies and classes of service to automatically hold data in disk cache for a specified time, and determine when writing of data to tape and purging from disk occurs.

HPSS is ideal for applications that generate massive amounts of data, such as modeling, statistical analysis, computational fluid dynamics (CFD), complex engineering, and computer-aided design. For example, weather forecasting collects data from sensors at sea and in space tracking temperature, moisture, atmospheric pressure, and so on. Systems continually feed current and historical data into models to predict the next big storm or what trajectory a hurricane is likely to take.

Think of the huge amounts of data required in particle collider experiments, design of a jumbo jet, or formulations for new drugs based on genomics and hundreds of clinical trials. It's easy to see the need for a large-scale storage solution like HPSS.

HPSS stands up to massive data growth

For organizations handling massive volumes of infrequently accessed data, HPSS provides a highly efficient, scalable, and cost-effective storage solution.

Flexible hardware and data choices

As a vendor agnostic storage solution, HPSS offers organizations flexibility to choose their preferred hardware. Tape libraries from Oracle (STK), Spectra Logic, or any other manufacturer are available. Disk storage options include Dell/EMC, DDN, NetApp, IBM and other solutions. For servers, companies have an array of choices, including Red Hat Enterprise Linux on Intel, AMD or IBM Power processors. HPSS supports a variety of popular data interfaces including Parallel FTP, the Linux VFS FUSE mount point, Spectrum Scale (GPFS) with GHI for automatic disaster recovery and space management services, Globus gridFTP and OpenStack Swift via HPSS SwiftOnHPSS. HPSS supports an extended POSIX Client API (CLAPI) as its most complete and powerful user interface. With this flexibility, customers avoid long-term vendor lock-in by using common industry standard-based technologies and interfaces.

Unmatched scalability and performance

In addition, HPSS provides unmatched scalability from tens to hundreds of petabytes by simply adding disks and tape libraries. (Figure 2) Yet, regardless of how large the HPSS system grows, it always appears to clients as a single, unified storage service.

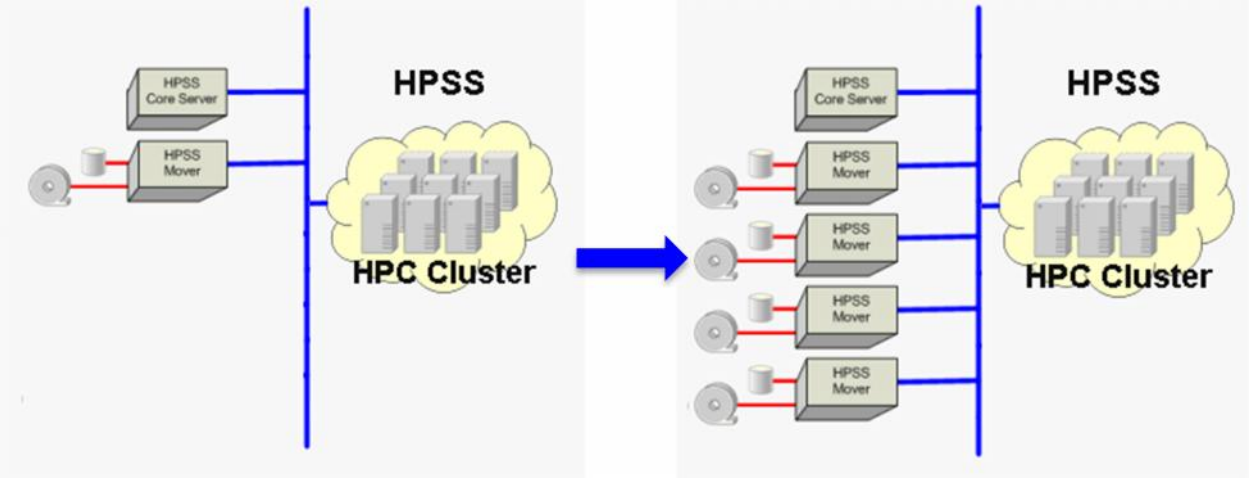

Figure 2 – HPSS Linear Scalability

By distributing data across a high-performance network, HPSS provides both scalability and parallelism. This is important for applications requiring fast access to large files. In addition, HPSS maintains high data transfer rates at scale by eliminating overhead typical of most data transfer processes. For example, the parallel FTP server establishes and controls the transfer sessions, but does not participate in the data transfer itself; the data moves between the parallel FTP client and the HPSS data movers. Additionally, HPSS supports a third party SAN transfer, where HPSS clients can write directly to the HPSS disk devices.

Compelling cost advantages

HPSS also offers compelling cost advantages over storage management systems like Oracle Automatic Storage Management (ASM) and Tivoli Storage Manager (TSM). While these systems charge per terabyte of storage under management, HPSS is a “Software-as-a-Service” (SaaS) offering. The economics for large customers can be extraordinary. While an HPSS customer will experience incremental hardware costs as their implementation grows, there is not an ever-increasing fee associated with capacity, no matter how large the HPSS environment gets.

Organizations also can expect to save on energy-related costs since the bulk of data is stored on tape. Since there are not any moving parts in its inactive state, tape has no power and cooling requirements.

Efficient data handling and protection

Another extremely efficient aspect of HPSS is the ability to perform automatic small file tape aggregation. Instead of writing individual small file to tape, which is very slow, HPSS writes groups of small files to a single tape section. This maximizing tape drive performance and tape cartridge utilization for small files.

Hand-in-hand, for large files, HPSS features tape striping, and tape striping with parity, called Redundant Arrays of Inexpensive Tape (RAIT). The latter ensures data parity, as well as higher performance and reliability by allowing read and write activity to continue even if a tape fails.

For added data protection, HPSS can be implemented across multiple sites. This allows organizations to replicate data from a production HPSS environment to a remote site for preservation of valuable information if a site outage or natural disaster occurs.

For improved tape recalls, HPSS supports tape ordered recalls. Additionally, HPSS supports the enterprise tape drive feature called recommended access ordering (RAO) to further reduce the time needed to recall multiple files from a single tape cartridge – cutting recall times down by 40% to 60%.

HPSS supports a revolutionary end-to-end data integrity strategy that leverages file checksums and T10-logical block protection (LBP) feature of tape drives. Traditional, T10-LBP can only be used to verify that the data written to tape has not changed. The HPSS end-to-end data integrity strategy ensures, that the data being written to tape matches the file checksum, as the data is being written to tape. The files written to tape by HPSS do NOT have to be read back from tape, and verified using an independent checksum calculating process – huge savings in time and hardware resource!

With these exceptional capabilities, HPSS takes researchers, analysts, and engineers far beyond the storage limits of traditional digital backup or archive. It truly addresses the peta-scale requirements of today's most advanced technical computing applications, and is considered “best of breed for tape.”

Jeskell as your trusted HPSS advisor

Implementing the hardware infrastructure for HPSS can be challenging since there are often multiple vendor integrations required. Many organizations also have unique technical and business requirements for extending HPSS into existing systems such as high-performance computing (HPC) farms or specialized processes such as CFD modeling and testing.

Jeskell has helped enterprises, academia, research laboratories, and government agencies address their high-performance storage requirements for more than two decades. Each HPSS implementation is tailored based on individual business needs and budget. In addition, Jeskell is the only company outside of IBM authorized to deliver this powerful solution.

Consultative, multivendor approach

Organizations choose Jeskell because we offer a holistic, multivendor approach. We offer each customer a variety of choices for disk cache, core HPSS servers and data movers, and tape libraries to create the best mix of technologies that best suit your use case. And we bring the integration and implementation expertise to put together a complete HPSS solution quickly, correctly, and cost-effectively.

We take a consultative approach to assess your data challenges and determine your precise storage needs. Based on our assessment, we will recommend the optimal HPSS configuration for your organization, applying all the standards and best practices established by IBM and the DoE. We also have the expertise to go beyond IBM's base HPSS solution and address your special requirements.

Comprehensive design and implementation

Jeskell offers complete design services to develop an HPSS architecture based on the requirements and parameters established during initial consultations. The design will address capacity and performance needs by ensuring the solution is configured with the correct number of data movers, disk systems, and tape libraries. We also take into consideration any existing tape libraries and augment those only as necessary. This is where Jeskell's decades of experience in storage and HSM prove especially valuable.

We have strong relationships with all the major equipment vendors, allowing us to execute purchase arrangements quickly and efficiently. Once components are on-site, Jeskell provides complete installation, integration, and testing services to get the HPSS solution up and running effectively. This includes establishing interfaces to existing systems and linkages to all data sources.

Importantly, Jeskell works collaboratively with IBM, which is responsible for installing the HPSS software and configuring any data replication specified for the solution. In this way, we ensure that the solution adheres to all IBM and Department of Energy testing requirements and operating procedures.

Our focus: HPSS deployments with ease

Throughout the project, Jeskell is focused on making your transition to HPSS as easy and non-disruptive as possible. We deliver a complete multivendor solution with comprehensive technical services and support, all managed under a single contract at a competitive monthly fee. And as your data storage and archiving demands grow, Jeskell is your one call to expand the HPSS environment as needed.

Conclusion

With advancing analytic capabilities, exponential gains in supercomputing capability, and a rapidly expanding Internet of Things, the data explosion witnessed in recent years will only accelerate. Today, large research, analysis and technical computing environments may have 10, 20, even 50 petabytes of data to store and manage. Those same organizations will likely grow to hundreds of petabytes within the decade. And it will not be long before we are talking about exabytes.

Traditional storage and hierarchical management systems simply cannot keep up with such data growth in an efficient and cost-effective way. As we have seen, HPSS can scale virtually without limit to support organizations with massive data requirements well past 2020. In addition, HPSS can handle this growth with only incremental hardware costs and no increase in service fees.

Jeskell offers decades of experience to help your organization adopt HPSS with flexible, vendor-agnostic component options and a full portfolio of design, implementation, and support services. We remain the only IT solution provider entrusted by IBM to deliver HPSS solutions. Our thorough, consultative approach ensures organizations receive a solution tailored to their unique technical and business requirements. And our agility means fast deployment, as well as prompt attention and responsive support throughout the project.

In conclusion, HPSS is the industry's best-of-breed, large-scale solution for tape. With Jeskell as a partner, organizations across academia, government, and industry can adopt HPSS with ease, and be confident they are getting the optimal configuration at a highly competitive price.

Please contact Greg Lefelar to set up an appointment:

Greg Lefelar, Vice President of Sales

Jeskell Systems, LLC / An IBM Premier Business Partner

(301) 776-3400 x902 | glefelar@jeskell.com