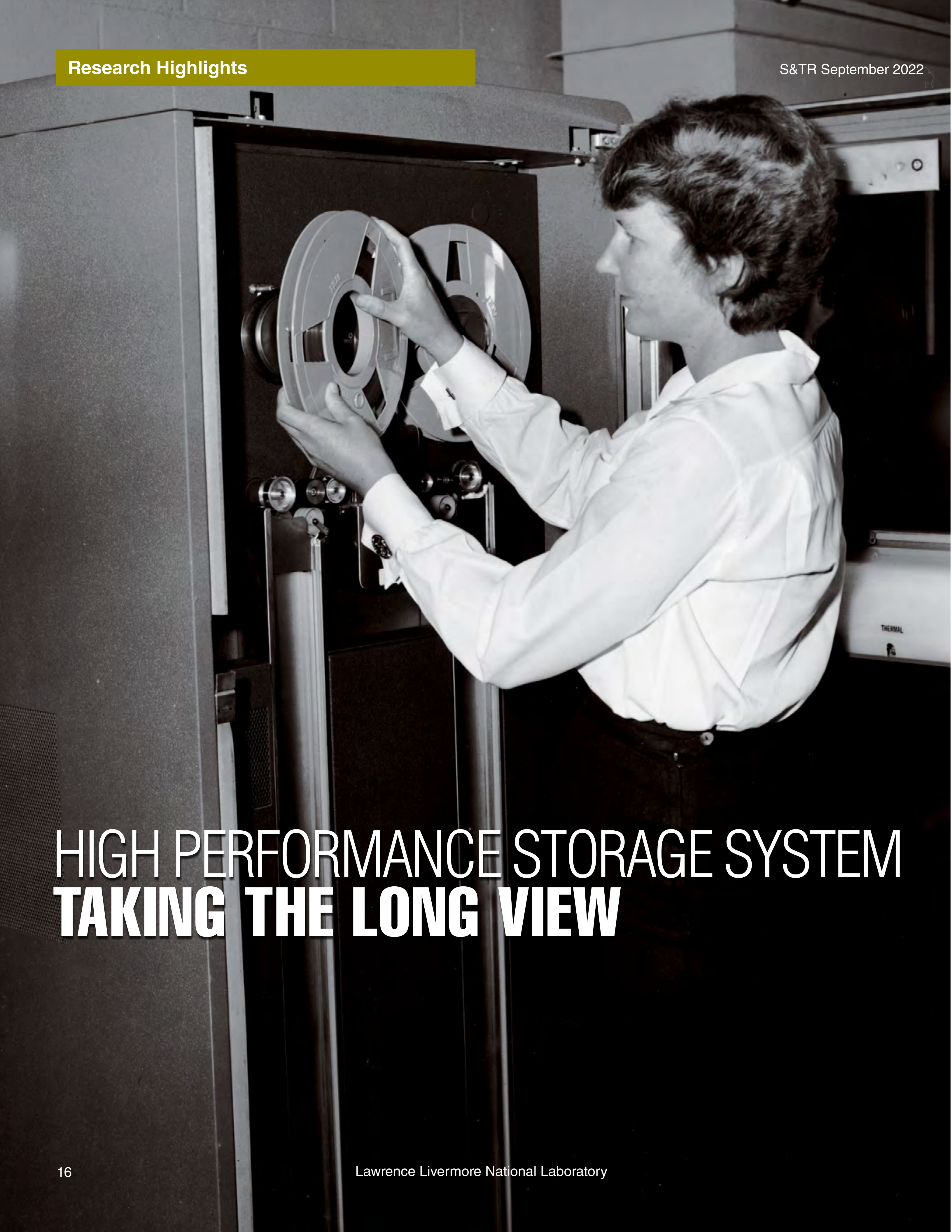


HIGH PERFORMANCE STORAGE SYSTEM
TAKING THE LONG VIEW

THE world of high-performance computing (HPC) is ever advancing with software updates, new hardware, increasingly data-heavy challenges, and the need to access and store immense amounts of data. Today's data dilemmas can be summed up in two simple questions: "Where should all this data go? How should it be managed?" Easy questions to ask, but not so easy to answer.

About three decades ago, Lawrence Livermore and four other Department of Energy (DOE) national laboratories working in HPC collaborated with IBM to tackle this issue. The laboratories and IBM recognized that issues around data storage constantly progress, requiring new and innovative solutions as supercomputing environments evolve. The result of their collaboration was the High Performance Storage System (HPSS), a scalable software solution that addresses the ever-changing data storage challenge facing the worldwide HPC community.

Lawrence Livermore plays a key role in the HPSS collaboration, significantly contributing to the original inspiration for the technology as well as being a co-developer and early adopter. The Laboratory had experience building distributed networked computing, printing, and storage shared environments long before this became commonplace. Engaging with other supercomputing centers that had similar experience, such as NASA and Los Alamos National Laboratory, led to the idea of standardizing how distributed storage could be achieved in the form of the Institute of Electrical and Electronics Engineers (IEEE) Mass Storage Reference model. HPSS was built on that scalable model.

The Laboratory's experience in data storage system development dates back to the 1953 delivery of its first computer system, the UNIVAC-1. Then, computers used vacuum tubes to perform calculations and stored data in mercury tanks and on magnetic tapes. (See *S&TR*, March 2002, pp. 20–26.) Now, supercomputer systems run calculations using thousands of powerful processors and store data on solid-state disks, magnetic disks, in the cloud, and on tape cartridges—high-tech descendants of historic magnetic tape reels.

The Data Conundrum Timeline

Before the advent of 3D simulations, the largest storage management systems at leading HPC sites archived a total of less than 10 terabytes each. (Today, Livermore's HPC archives allow individual files 10 times that size.) Beginning in the late 1980s, data storage became a concern for national laboratories involved in HPC, including Lawrence Livermore. These HPC leaders arrived at hierarchical storage management (HSM) as an archival system that could meet supercomputer performance requirements. HSM automatically moves data between expensive storage media—for instance, solid-state drive arrays—and low-cost media such as hard disk drives and magnetic tape cartridges. The HSM system helps keep costs down by keeping frequently accessed data on fast devices and moving "less used" data to slower devices.

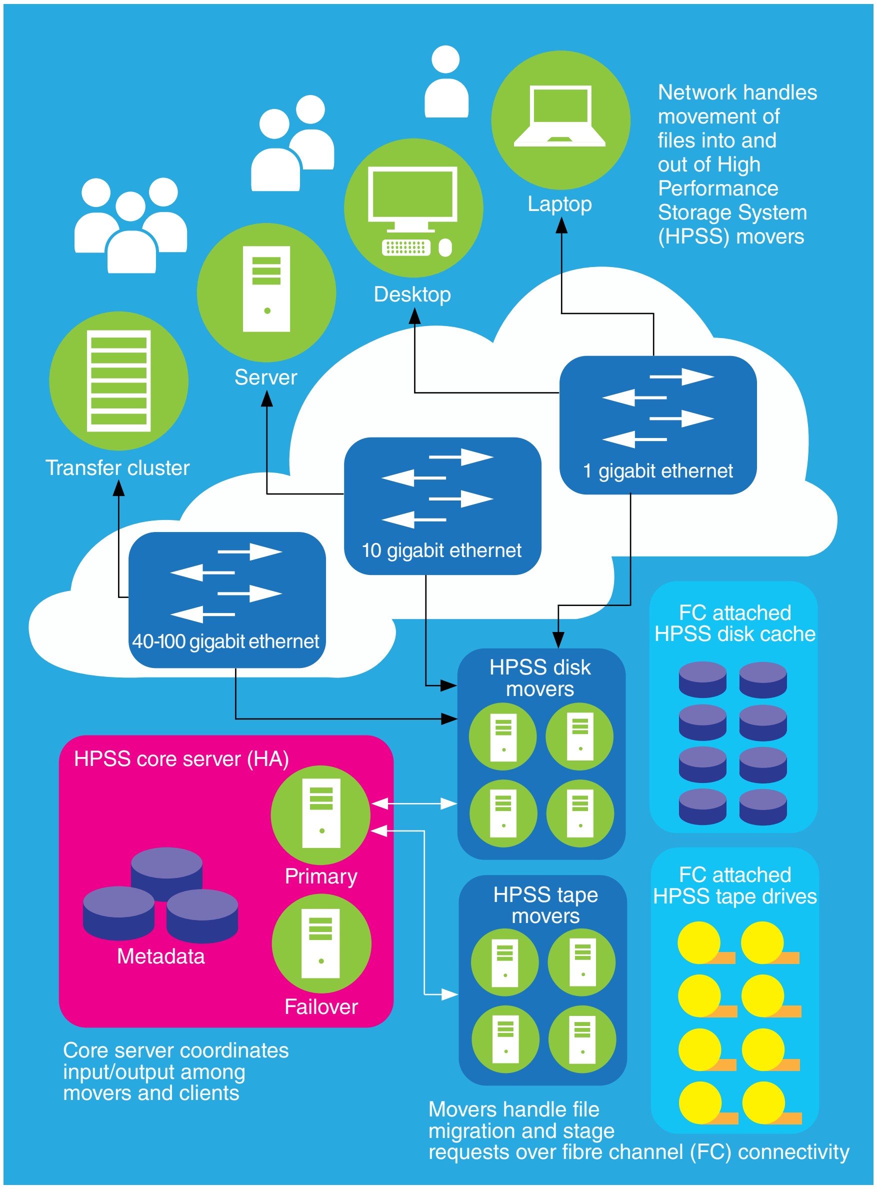
Tape has been one of the go-to media for storing and archiving computer data since the early days of the Laboratory. This photo shows Livermore computer programmer Edna Vienop, who worked on a project calculating the return of Halley's Comet, loading a tape on the IBM 704 in 1959.

In 1992, predating HPSS, Lawrence Livermore, Los Alamos, Sandia, and Oak Ridge national laboratories along with IBM and other industry partners pooled resources to develop a scalable and efficient data archiving system. These organizations formed the National Storage Laboratory (NSL) Cooperative Research and Development Agreement, setting out to investigate, demonstrate, and commercialize high-performance hardware and software storage technologies that would remove network computing bottlenecks. "The HPC community foresaw a data storage explosion and realized that no single organization had the necessary experience and resources to meet all the challenges involved. This realization fueled the birth of the HPSS collaboration," says Todd Heer, Lawrence Livermore's representative on the HPSS Steering Committee and the committee's co-chair representing DOE. An additional national laboratory, Lawrence Berkeley, joined the other laboratories and IBM in this venture.

The inaugural HPSS-based archive drew heavily on NSL's results and the IEEE's Mass Storage Reference model, which NSL and others helped to create. Spearheaded by Livermore Computing's Dick Watson, who came to Livermore from Stanford Research Institute (SRI), that first HPSS could store billions of objects long before any competing systems, earning it an R&D 100 Award in 1997. (See *S&TR,* October 1997, pp. 18–19.) When the collaboration formed, the terascale era—with computers capable of 10^{12} floating-point operations per second (flops)—was still five years away. The collaboration met the challenge and those that followed by evolving storage architectures to meet the needs of the supercomputing community. In the three decades since, the collaboration has provided more than 10 major releases, with the most current release addressing storage needs for computations at petascale—one quadrillion (10^{15}) flops. Each release focuses on increasing operational efficiencies, performance, and storage capabilities while providing the ever-greater speeds required by its users. Heer notes the rarity of software that thrives for three decades. Yet, HPSS has achieved that milestone and continues to deliver.

The Nuts and Bolts of Now

One of the founding tenets of HPSS was to take magnetic tape—the most economical form of digital media—and scale it up and out. “Using physical storage tapes may seem old-school,” says Heer, “but in actuality, advancements in tape technology have kept tape further than disk from the super paramagnetic limit where it can no longer grow in density.” Heer notes that the tapes used today are quite different from the old eight-tracks in their capacity and capabilities. “Unlike 1970s-era Neil Diamond eight-track audio tapes, today’s digital tapes have upwards of 7,500 tracks spanning their width,” he continues. “The HPSS application must be similarly advanced to handle this technology. We were the first to successfully develop the technology of using redundant arrays of independent tapes—‘RAIT.’ RAIT gives us the ability to write different chunks of a given data set to any number of tape drives simultaneously, providing cost-effective redundancy.” Tape drives have other benefits as well: they save energy, maintain data integrity longer than disks, and are inexpensive compared to cloud storage and hard disk drives.


Network handles movement of files into and out of High Performance Storage System (HPSS) movers
Laptop
Desktop
Server
Transfer cluster
1 gigabit ethernet
10 gigabit ethernet
40-100 gigabit ethernet
FC attached HPSS disk cache
HPSS disk movers
HPSS core server (HA)
Primary
Metadata
Failover
HPSS tape movers
FC attached HPSS tape drives
Core server coordinates input/output among movers and clients
Movers handle file migration and stage requests over fibre channel (FC) connectivity

“Clients,” which can be anything from laptops to supercomputers, have their own disk-based file systems suitable for short-term data storage. When a user selects files for archiving, copies of the files are transferred by disk movers to the first tier of the archive—the fast access disk cache. Files stored in the cache can be instantaneously retrieved. If a file stored on tape is requested, a robot grabs the relevant tape, mounts it on a nearby drive, and the data is transferred up through the disk cache and out to the user.

Today, approximately 40 sites around the globe including Japan’s High Energy Accelerator Research Organization, Germany’s Max Planck Computing and Data Facility, the National Oceanic and Atmospheric Administration, Indiana University, and Australia’s University of Tasmania use HPSS. Roughly 20 organizations have contributed to the HPSS effort over the years, including the core team of five U.S. national laboratories and IBM. Heer says, “The value of the feedback mechanism from some 40 sites cannot be overstated. These scaled-out production sites—a few even bigger than Lawrence Livermore’s archive—report back on bugs and features yielding a battle-hardened codebase worthy of keeping DOE’s most important data long into the future. DOE benefits significantly from this model compared to a model in which only national laboratories ran the software. At the same time, DOE keeps its HPC national interests at the fore by co-developing the codebase.”

From the start, HPSS was designed to be an HSM application to meet the high volume and speed requirements of the world’s fastest supercomputers while providing the security to protect data at multiple sensitivity levels. The application is hardware-agnostic, allowing a smooth transition to new devices and systems as the industry evolves and permitting users to choose from the best vendor technologies available. Worldwide, HPSS serves a total of more than 4.5 exabytes (4.5 quintillion bytes) of production data.

Pictured (left to right) are Lawrence Livermore High Performance Storage System team members Todd Heer, Debbie Morford, and Geoff Cleary in front of a Laboratory tape storage system. (Photo by Garry McLeod.)

Accelerating into the Future

At Lawrence Livermore, where more than 50 terabytes of data are produced daily for the archive, HPSS runs on a scalable and lightning-fast, clustered architecture. The Laboratory's HPSS features multiple storage tiers including a fast access disk cache in front of a tape-based archive. Livermore's team leverages economies of scale to deploy an archive disk cache that enables files placed in the archive to be immediately retrievable from a disk cache within a year (on average) before needing to be read from tape.

The Laboratory has five Spectra Logic TFinity tape libraries, including the world's largest, capable of holding half an exabyte of data. Livermore's HPC center boasts a total of 1 exabyte of on-premise storage capacity. "We have archive data dating back to the late 1960s that—along with everything stored since—must be saved in perpetuity," says Heer. "HPSS makes it possible for us to do so, knowing that no matter what new storage technologies are available in the future, the data will remain accessible. HPSS continues to play a key part of a total HPC center storage solution, lifting the burden of the more expensive storage closer to the computer."

The collaboration keeps a steady eye on the challenges ahead as it celebrates impressive accomplishments to date. The amount of data that user sites need to archive continues to accelerate. For example, archiving rates are expected to increase by more than 500 petabytes per year for DOE HPC sites as a whole. Meanwhile, the demands of the exascale computing era bring significant speed requirements to HPSS. "Parallelization of the data retrieval process—that is, increasing the number of tapes we can simultaneously pull data from and the speed at which we can spin the drives and thus access the data—is helping HPSS match the data bandwidth of these systems," says Heer. "Looking forward, the team is also working to incorporate more open-source software, exploring integration of cloud-native technologies, and increasing data discoverability so users can better understand data trends and patterns." In these ways and more, the HPSS collaboration continues to create ever-evolving, software-defined, scalable datastore systems, honoring its primary mission of long-term data stewardship for government, academic, and commercial organizations worldwide.

— Ann Parker

Key Words: archive, data storage, data retrieval, disk cache, exascale, high-performance computing (HPC), High Performance Storage System (HPSS) collaboration, IBM, magnetic tape libraries, Los Alamos National Laboratory, National Storage Laboratory (NSL), Oak Ridge National Laboratory, Sandia National Laboratories, Spectra Logic TFinity, tape drive.

For further information contact Todd Heer (925) 423-9850 (heer2@llnl.gov).