



Active Archiving Solution for eResearch Data Sets

Sun StorageTek™ 5800 System with Fedora



The new world of high-performance computing and digital repositories calls for a fresh look at the way digital content is created, managed, and preserved. Digital data is being generated in unprecedented volume, and the challenge is to be able to store, manipulate, and access it in a flexible and cost-effective manner.

Two breakthrough technologies have emerged to address this problem. One is Fedora (Flexible Extensible Digital Object Repository Architecture), which is an open source repository technology now being offered to the world through Fedora Commons, a non-profit organization. The other is the Sun StorageTek™ 5800 system, a storage solution designed from the ground up to help you reduce the cost and complexity of managing large data sets while offering highly flexible data retrieval and delivering unprecedented data integrity.

Taming terabytes of data

Scientific organizations around the world are accumulating data at scales that are a thousand times as large as a decade ago. Previous astronomical data archives were generally measured in tens of gigabytes, whereas today's data archives are often measured in terabytes. For instance, a single data release from the Sloan Digital Sky Survey can measure 30 TB.

All of this data must be archived and maintained with perfect accuracy in case it is needed for later analysis. If, for example, a supernova explosion occurs, it is conceivable that astronomers from all over the world might want to look at all activity in that portion of the sky both historically and on a current day-by-day basis over a period of many years.

Furthermore, there is a need to correlate results across multiple data sets. Coming back to the supernova example, it is likely that different institutions were responsible for different aspects of the data. One astronomical laboratory may have kept ongoing records of wave emissions, another a record of cosmic particle activity, and another of the positions of planetary bodies. Each of these types of data may have been recorded in an independent dataset. However, analysis across all three datasets may be critical to understanding the overall picture.

In addition, flexibility regarding data retrieval and analysis is critically important. Not only is it difficult to predict how the data will be used, but it may also be necessary to integrate today's data sets with future sets that include data types that cannot yet be even conceptualized. As data is recorded in many forms in laboratories around the world, it must be made freely available to researchers so that they can slice and dice it in different ways from how it was originally analyzed. Curation of data pertains not just to its use but also to its future re-use in unanticipated ways.

Finally, the process of both ingesting and accessing data should be as simple as possible. It should be based on a digital object model that allows the encapsulation of metadata information with minimal effort from researchers. The act of archiving should encode as much metadata as reasonably possible to allow easy discovery of objects in the future.

Sun and Fedora eResearch solution

The Sun StorageTek 5800 system is a third-generation, first commercially available fixed content storage system that has been designed to address precisely the problems discussed above. In combination with the Fedora Commons platform, it is an ideal solution for creating, managing, publishing, sharing, and preserving digital content. It addresses the urgent needs of communities of practice such as scholars, educators, museum curators, scientists, and librarians.

Highlights

- Integrated solution with proven success for large-scale repositories at well known institutions such as Johns Hopkins and Purdue University
- Track millions of XML objects in Fedora with capacity to seamlessly scale to hundreds of terabytes with the Sun StorageTek 5800 system
- High data integrity through automatic block and file level checksums and distribution of file fragments across nodes to ensure no single point of failure
- Extreme data protection with redundant design that allows multiple nodes and drives to fail without any resultant data loss

Fedora is an open source storage repository platform that uses a service-oriented architecture to enable the creation of innovative, collaborative information spaces. It is designed for the longevity and integrity of any kind of digital content, and also offers the ability to inter-relate such content from different sources. Fedora was originally created at Cornell University under research grants from NSF and DARPA. This research evolved into a successful open source project jointly managed by Cornell and University of Virginia under a grant from the Mellon Foundation. Currently, development of Fedora continues in the context of the new Fedora Commons non-profit organization with funding from the Gordon and Betty Moore Foundation.

With the help of researchers at Johns Hopkins University, Fedora has recently been integrated with the Sun StorageTek 5800 system, creating a unique combination for eResearch applications. Fedora users now have transparent access to the StorageTek 5800 system through a set of APIs that have been integrated into Fedora as shown in the figure.

The Sun StorageTek 5800 has been designed from its very inception to meet the needs of large data repositories such as those created under Fedora. It incorporates clustered servers for both processing and storage, allowing the system to be easily scaled as data sets grow. Early testing units sent to teams building Fedora-based archives and applications at Johns Hopkins University, Purdue University, Alberta Libraries and others, have verified the value of the Sun StorageTek 5800 system for use with Fedora.

Fedora has some core requirements for maximum efficiency, all of which are met by the architecture of the StorageTek 5800 system. In addition to scalability, the required attributes of high data integrity, robustness, reliability, and effective meta-data management have all been designed into the StorageTek 5800 system.

While the StorageTek 5800 system is equally well oriented towards managing data created by High Performance Computing applications such as scientific visualization or of data captured from a particle accelerator, this solution brief is focused on managing research data. Additional

information about how the StorageTek 5800 system can be used with HPC applications can be found in other Sun literature.

eResearch in action

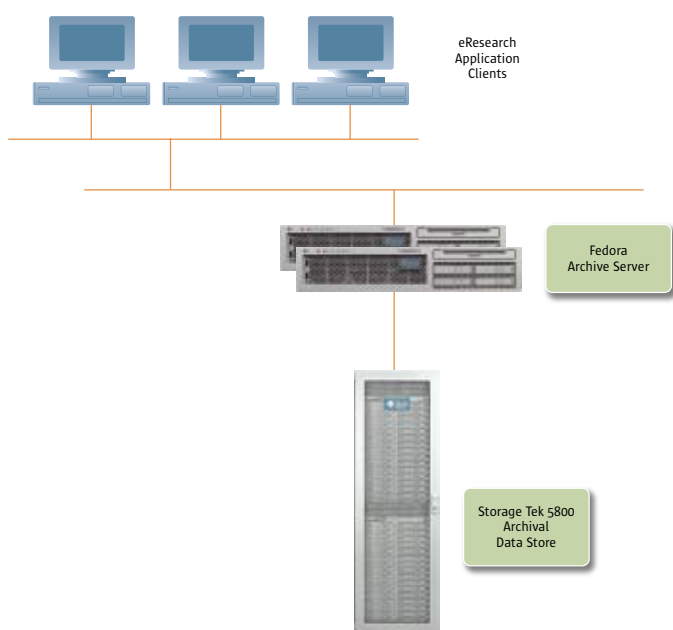
An example of where Fedora is being used in combination with the Sun StorageTek 5800 system is in the National Virtual Observatory Project at the Johns Hopkins University. This project is being funded by the National Science Foundation. Astronomical databases from laboratories and orbital observatories worldwide are being aggregated and made available to today's community of astronomers, students, and scientists.

Another compelling application is at the Max Planck Institute for Meteorology. Researchers at their German Climate Modeling Center are archiving weather data from around the world and making it available to researchers to perform pressure/precipitation/temperature analysis and make predictions on future weather patterns.

Intelligent storage

In coming years, hundreds of millions of objects will be stored in a single archive, taking up petabytes of data storage space. Traditional storage architectures available today such as Storage Area Networks (SANs) are poorly suited to large-scale archival problems since they can be technically and administratively hard to scale. The StorageTek 5800 is object-based storage and, as such, moves space and scaling management down into the storage layer, which reduces the complexity for the administrator.

Primary storage on traditional systems can be more difficult to manage and find the data because there is only one view of it through a file directory structure. As more and more files are added, the directory structure can become extremely complex and hard to navigate. The object based storage on the Sun StorageTek 5800 system with its flat namespace offers multiple views of the data through a simple query interface that gives researchers the flexibility to search for the data they need.



eResearch application clients can use Fedora to transparently access the Sun StorageTek 5800 system data store through the integrated interface.

Searches are based on metadata information that is created when the objects are originally stored. Basic metadata fields such as date and time of creation are generated automatically by the system. The user can also provide additional metadata such as project name, research department name, data type, and source programs when the data set is initially stored. With the built-in, extensible metadata index, this additional data can be queried on by the system itself, alleviating the need for external databases to house and query the data.

With extremely large data sets, quick retrieval of objects can become a major issue. The StorageTek 5800 system helps avoid performance bottlenecks by using intelligent, embedded metadata and virtual file system views that enable parallel execution of the search and retrieval process. With the metadata capabilities of the StorageTek 5800 system, institutions can avoid the added cost and complexity of an external relational database. The metadata features are what Sun calls its second generation object store features. In a future release, Sun plans to extend the value of the StorageTek 5800 system to include its third generation feature, the ability to run data services inside the storage box against the data as it is stored or retrieved, or in a bulk operation. These “storage beans” could, for example, perform transformations against data as it is retrieved (e.g. convert Office95 docs to OpenOffice, convert photo sizes), or encrypt a file's data as it is written.

Flexible data management

Fedora is built with a framework that allows easy extensibility not just to new formats of data but also to new storage platforms and devices. The collaborative work done with the Sun storage team resulted in a Fedora module that allows for seamless integration with the Sun StorageTek 5800.

Another key benefit of the Fedora platform is highlighted in the field of scientific publishing. Fedora enables the creation of collaborative, integrated information spaces where any

information entity can be linked to any other entity. Multiple data sets can be linked to multiple research papers via multiple paths. This allows post-facto analysis such as locating all papers that resulted from a single set of data. Or conversely, one can easily find all data sets referenced by a particular paper, author, or institution. This understanding of linkages can result in new hypotheses and research directions that otherwise would not have been possible.

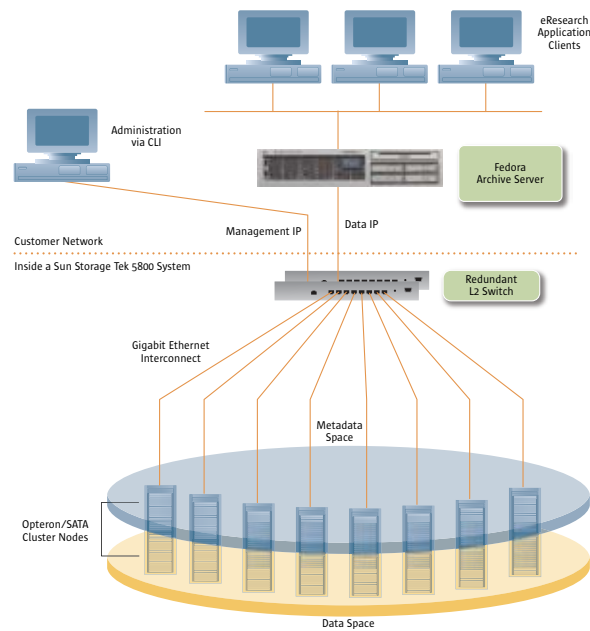
The scalability designed into Fedora also allows a greater depth of understanding to scholars. Often, multi-year or large-population studies result in reams of data, only subsets of which are analyzed in the writing of research papers. Fedora is designed to allow easy access to the raw data underlying these subsets, allowing users the ability to examine and analyze the entire range, regardless of its size.

For example, the Framingham Heart Study, under the direction of the National Heart Institute, has measured the cardiovascular health readings of 13,500 patients over three generations for almost sixty years. The amount

of data generated from this mammoth is nearly impossible to boil down to a level suitable for a single research paper. Researchers have therefore extracted subsets of the data sorted by common characteristics such as gender, age range, diet or exercise as a means to test these. A reader of such a paper archived under Fedora could look beyond the data that was chosen by the authors. They could link to the Framingham Study in its entirety and discern for themselves what advantages or limitations resulted from such a sampling. Because of Fedora's strength in handling multiple media types, in addition to numbers and text, they could also easily view physician charts, CAT-scans, X-rays, and MRIs.

Scalable, reliable infrastructure

The StorageTek 5800 system is the only petabyte-scale object store based on open source technology and architected to become open source. Its scalability results from its modular architecture. A half-cell is composed of 8 nodes, and can be easily upgraded to a full 16-node cell. The system can then be expanded to a multi-cell as data needs increase. With each new cell, more compute



The Sun StorageTek 5800 system is based a scalable architecture that enables parallel execution of the search and retrieval process.

power is also added to the system. The system is thus scalable to hundreds of terabytes.

The integrity of a single data image is maintained no matter how large the system scales. If older technology is removed, the system is designed such that data rebalances seamlessly across the remaining resources.

Regardless of the number of cells, the system has just two IP addresses (for the data and the administrator). There are no RAID sets to manage, no SANs to architect and no file systems to mount. Administration thus becomes a very simple task.

Fedora is similarly scalable since it encapsulates all data as XML objects. As the objects then increase in number, they can be stored on devices ranging from a single small server all the way up to an enterprise-wide multi-server system. In this solution, Fedora maintains a registry of data that is stored across multiple nodes.

The Sun StorageTek 5800 system uses a symmetric, clustered architecture wherein all storage, control, data, and metadata path operations are distributed across the cluster. Each node is independent of all other nodes, and there is complete symmetry in both the hardware and software on each node, providing high reliability and accessibility as well as high performance scaling. The hardware in the Sun StorageTek 5800 system also employs RAID 6 with a dual parity scheme, which is excellent for fault tolerance. No single node in the system contains more than a fragment of a file. This distribution of fragments across nodes prevents loss of data even in the case of multiple node failure.

The system is also self-healing, meaning that it detects failures and moves data away from affected hardware. When failed units are later replaced, Sun software that runs across all nodes automatically detects new replacement hardware and reassembles the data.

The software does automatic data integrity checking and correction, with block and file level checksums validated upon retrieval. The data encoding and reconstruction process means that any two disks on different nodes (or more than two disks on the same node) can be lost without losing data. The system can even sustain multiple, simultaneous drive failures without data loss.

Security is also added at the higher metadata level for an additional layer of protection.

Open solutions

Fedora is open source software and is free. All the code is publicly available on sourceforge.net and is usable in its entirety or as relevant application pieces. The code is tested and robust since it is on its tenth release since 2003. In addition, a large and vibrant user community has produced a wide variety of user interfaces, front ends, middle-ware, applications and utilities.

The StorageTek 5800 system represents the first commercially available fixed content storage system today, which means customers can be assured that they will always be able to get to their data. Since it is based on Open Solaris and Sun's commitment to open interfaces, the StorageTek 5800 system protects against proprietary lock-in. Sun is committed to present open APIs on both the client and the server side. Data is also kept in an open storage format so that future devices that may be based on new advances in technology will be able to read it. More details can be found at opensolaris.org/os/project/honeycomb/.

Educational institutions in particular have adopted Fedora with enthusiasm because of its open source structure, its modularity, and its extensibility. For example, The National Science Digital Library connects end users with over 2.5 million digital resources that promote active, inquiry-based teaching and learning. Fedora was also chosen as the con-

Learn More

To learn more about Fedora and Sun solutions for digital content, visit fedora-commons.org and sun.com/storagetek/.

tent application layer for the Public Library of Science's PLoS ONE project, which is committed to making the world's scientific and medical literature a public resource. There are more than fifty other major digital repositories being developed under Fedora, ranging from the Perseus eHumanities library at Tufts University to the ARROW research collection sponsored by the Australian Department of Education. The entire list can be viewed at fedora.info/wiki/index.php/Fedora_Community_Members.

Breakthrough innovation

Institutional Repositories have been embraced as being critical to managing enterprise-wide digital content and unlocking the value of institutional output. Fedora is a prominent example of an open source software platform that has been widely adopted for the creation of such repositories around the world.

The Sun StorageTek 5800 system is an ideal storage solution for large scale repositories built on Fedora Commons platform. Sun offers the only commercially available petabyte-scale fixed content store designed to be open source and that meets all the required attributes of simplicity, expandability, programmability, and data integrity. Together, Sun and Fedora bring a breakthrough in innovation for digital content. Institutions can now more fully utilize their digital content while reducing the cost and complexity of preserving it throughout its full lifecycle.