Distributed Data Sharing with PetaShare for Collaborative Research

PetaShare Team, LSU

The NSF-funded PetaShare project started in August 2006 with the goal of enabling transparent handling of underlying data sharing, archival, and retrieval mechanisms, and making data available to scientists across the state of Louisiana for analysis and visualization on demand. The goal has been to enable scientists to focus on their primary research problems, assured that the underlying infrastructure will manage the low-level data handling issues. The key technologies that are developed as a part of PetaShare include data-aware storage systems and data-aware schedulers, which take the responsibility of managing data resources and scheduling data tasks from the user and performing these tasks transparently. Petashare has two major components – an enhanced version of iRODS to provide a global name space and efficient data access among geographically distributed storage resources, and the Stork data placement scheduler which takes the responsibility of managing data resources and scheduling data tasks from the user and performing these tasks transparently.

PetaShare has been deployed across five state universities and two health sciences centers in Louisiana. These institutions include Louisiana State University (LSU), Tulane University, University of New Orleans, University of Louisiana at Lafayette, Louisiana Tech University, and LSU Health Sciences Centers in New Orleans and Shreveport. PetaShare manages approximately 300 Terabytes of disk storage distributed across these sites as well as 400 Terabytes of tape storage centrally located nearby LSU campus. For connecting all of the participating sites together, PetaShare leverages LONI, which is a statewide 40 Gbps fiber-optic network in Louisiana. It links all major research institutes in Louisiana.

iRODS has been the backbone of Petashare, providing a global name space across all participating institutions. Since its initial implementation, PetaShare has gone through major changes such as integration of MASREP (A Multi-master ASynchronous REPlication) tool, which asynchronously replicates metadata to all other sites to eliminate single point of failure and to provide high availability without sacrificing performance. MASREP even increases the performance because all incoming metadata requests from a respective server are processed within that site. At the front-end, PetaShare provides very light weight interfaces called PetaFs, Petashell, and Pcommands, based on FUSE, Parrot, and icommands technologies respectively. PetaFs is based on FUSE, and the client has recently been ported to Mac OSX which uses MacFUSE. The novel PetaFs and Petashell interfaces enable users to access their remote and distributed data the same way as they access the data on their local disk. While doing so, users need not make any changes to their application, including re-compiling or re-linking, and they also do not need any special privileges on the system to use these interfaces.

PetaShare provides scientists with simple uniform interfaces to store, access, and process heterogeneous distributed data sources. The archived data is well cataloged to enable easy access to the desired files or segments of files, which can then be returned to the requester in a chosen format or resolution. Multiple copies of high priority information can be stored at different physical locations to increase reliability and also enable easier retrieval by scientists in different geographical locations. The data is also indexed to enable easy and efficient access to desired data. The requested data is moved from the source or archival sites to the computation sites for processing as required, and the results are then sent back to the interested parties for further analysis or back to the long-term storage sites for archival.

As of July 2009, PetaShare has been actively used for 25 different research projects by more than 70 senior researchers from 11 different institutions. The supported application areas include: coastal hazard prediction (LSU and SURA), reservoir uncertainty analysis (LSU, ULL, and SUBR), DNA sequencing (Tulane and UNO), high energy physics (LSU, Latech, and DOSAR), X-ray tomography (LSU), numerical relativity (LSU), high speed visualization (LSU and ULL), biomedical data mining (LSUHSC, LSU and Latech), and computational fluid dynamics (LSU). PetaShare has also been an important component in several other state-wide cyberinfrastructure projects in Louisiana such as the NSF-funded CyberTools and HPCOPS, the Louisiana BoR funded LONI Institute, and the DOE-funded UCoMS projects.

Note on PetaFs: A virtual filesystem that allows users to access PetaShare resources as a local filesystem after being mounted on their machines. By using PetaFs, PetaShare resources can be seen in the directory hierarchy of an existing filesystem and accessed in the same way as local storage. PetaFs is based on FUSE (Filesystem in Userspace), a simple interface to export a virtual filesystem to the Linux kernel in userspace. PetaFs translates local I/O calls into remote iRODS calls through intermediate FUSE library calls. In the kernel, FUSE incorporates with the real filesystem and maps these FUSE calls to the actual filesystem calls at one end, and to the remote iRODS calls at the other end. Communication between kernel module and FUSE library is established by specifying a file descriptor obtained from /dev/fuse device file. This file descriptor is passed to the mount system call to match up the file descriptor with the mounted PetaFs virtual filesystem. The advantage of PetaFs is that it makes it possible to work on the data stored in Petashare resources by using standard UNIX commands (ls, cp, etc) and system calls (open, read, write, etc.) as in the real filesystem.