

Distributed Data Management with PetaShare

Mehmet Balman, Ibrahim Suslu, Tevfik Kosar

*Center for Computation & Technology, Computer Science Department
Louisiana State University, Baton Rouge, LA 70803, USA
{balman, ihsuslu, kosar}@cct.lsu.edu*

PetaShare¹ is an NSF funded project which aims to solve the distributed data sharing and management problem. Data-aware storage systems, data-aware schedulers, and cross-domain metadata scheme are some of the key technologies being developed in order to prepare an underlying infrastructure for scientist to manage the low-level data handling issues. The initial system which manages 250 Terabytes of disk storage and 400 Terabytes of tape storage has been deployed across the state utilizing the 40Gb/sec LONI² connection at five campuses: Louisiana Tech., University of New Orleans, Tulane University, University of Louisiana at Lafayette, and Louisiana State University. PetaShare enables collaboration between those institutions and helps multidisciplinary research in different application areas such as coastal and environmental modeling, geospatial analysis, bioinformatics, medical imaging, fluid dynamics, petroleum engineering, numerical relativity, and high energy physics.

One important focus in PetaShare research is distributed data handling problem in which data placement jobs should be regarded as first class citizens such that they should be scheduled, monitored, and optimized. Stork³, a batch data placement scheduler, is one of the first efforts for managing data transfers by introducing the scheduling approach. PetaShare brings the idea of data-aware system model which includes data-aware scheduler (Stork), resource allocation and resource selections services, higher lever planners, and workflow managers. Due to the huge data requirements of current scientific application, there has been extra effort to provide efficient data access methods favoring application performance while effectively utilizing the system and network recourses. PetaShare introduces the data management subsystem to be the I/O module in distributed computing systems.

In order to cope with the requirements of today's applications, we require monitoring and tuning tools to dynamically adapt data transfer tasks at the execution time. Optimum values for I/O block size, TCP buffer size, number of parallel streams are some of the examples affecting performance in data placement. Moreover, alternate protocol selection, storage reservation before data transfer, and fault detection mechanisms are some of the crucial properties of an underlying data management system.

The overall infrastructure consists of different storage components such as disk arrays, tape storage, and cache memory in computational nodes. Basically, there are two types of data movement approaches. First, data needs to be prefetched from low level storage layers to the higher levels such that management of data access has to be handled in an efficient manner. Second, data should be migrated between those five contributing institutions; moreover, data should be scheduled and moved between distributed sites. Enstore, SRM, and dcache are some of the used technologies. We concentrate on interaction between the components of distributed I/O subsystem and we design the PetaShare architecture to enhance the overall performance while maintaining a cyberinfrastructure for easy and efficient storage access.

Overall, PetaShare aims to enable data intensive collaborative science across state, by providing additional storage, and infrastructure to access, retrieve and share data. The system includes data-aware storage, data-aware schedulers, and cross-domain metadata. We discuss the overall architecture, data access layers, and organization between PetaShare components. We also explain technologies used to implement the initial system. Furthermore, we give details about the data scheduling approach studied under PetaShare project.

¹ *PetaShare: A Distributed Data Archival, Analysis and Visualization Cyberinfrastructure for Data-intensive Collaborative Research* (www.petashare.org)

² *LONI: Louisiana Optical Network Initiative* (www.loni.org)

³ *Stork: A Batch Scheduler specialized in Data Placement and Data Movement* (www.stork.org)