U.S. Department of Energy Best Practices Workshop on

File Systems & Archives

San Francisco, CA

September 26-27, 2011

Position Paper

Dominik Ulmer
CSCS
dulmer@cscs.ch

Stefano C. Gorini
CSCS
gorini@cscs.ch

ABSTRACT / SUMMARY

The Swiss National Supercomputing Center CSCS has introduced a business model which turns data services from a reactively to a pro-actively managed service. A clearly defined center-wide file system hierarchy in conjunction with a set of specialized computers for data analysis allows to optimize storage systems characteristics like bandwidth or latency for different systems and workloads, to plan and manage capacities within the resource allocation process for computing time, and to leverage technical and financial synergies between the different service categories. Storage services are based on Luster and GPFS software with TSM/HSM extensions. A combination of different storage hardware technologies like SATA, SSD, and tape are used for the services depending on the individual requirements.

INTRODUCTION

For many years, data was a side-business to computing for HPC centers. Large-scale storage systems were architected and installed as peripherals of a supercomputing procurement. However, data growth rates exceed performance growth rates of HPC systems and therefore storage systems become an increasingly more significant part of the investment and operational budget of the computing centers. While the Swiss National Supercomputing Centre CSCS recognizes the importance of data for computational sciences, it does not have the intention to turn from a high-performance computing center to a data storage and management center. It is therefore essential to understand the role of data in the workflow of computational scientists using supercomputers and to accordingly architect the data services offered by the center.

SYSTEM BOUNDARIES

The main function of a HPC center is to enable computational scientists to use supercomputers for their research. This involves the preparation and the processing of large data sets, either for preparing input for computing runs or for analyzing data, which may be both, measured data or the result of a computational job. In both cases, one deals with living data for ongoing research projects, i.e. a time span that is well below 10 years. Long-term archiving for documentary purposes is not in the core business of a supercomputing center. A data service at a HPC center in this framework has to address the following topics:

- support of the computational workflow by means of an integrated architecture of computing, storage, and data analysis systems
- a storage hierarchy which is easy to understand by the user and provides a clear basis for the management of technical requirements
- a business model that allows the center to plan investments and operational costs in advance and which is aligned with the business model for providing computational resources.

THE CSCS STORAGE HIERARCHY

CSCS distinguishes three different levels of storage (see **Error! Reference source not found.**):

A) SCRATCH file system

The purpose of the scratch file system is to provide a storage container for running an individual computational job resp. an individual suite of computational tasks. Data remains only temporarily on the file system and must be copied to a different storage level for permanent storage. The file system has no quotas for user or groups. Old files are automatically deleted in order to maintain capacity. The scratch file system is local to an individual computer and its technical characteristics are specified according to the architecture of the system and the expected workload.

B) PROJECT file system

The project file system provides a data management and storage space for an individual computational project. CSCS issues a call for project proposals twice a year. Researchers can request computational and storage resources in their proposals, which are evaluated by an external committee with respect to their scientific quality and impact. The size of the storage request and of the compute cycle request must be justified in the proposal and must be coherent to each other. The project receives a storage quota which is shared between all members of the project team. The project file system is globally mounted on all CSCS user facilities and provides enough bandwidth for efficiently transferring large data sets to and from the scratch file systems. It provides extended user functionalities like snapshots. Data is kept on the project file system for the duration of the computational project (up to 3 years) plus 6 months in order to allow the user transferring the final data to a longer-term storage system or to the storage resource of a successor project.

	Scratch	Project	Store
Size	Large	Very Large	Extreme size
Quota	No	By group	By consortium
Backup	No	Yes	HSM
Data life time	Wiped regularly by system every few weeks	Duration of project + 6 months	As contractually agreed
Locality	Local	Global	Global
Bandwidth	Very high	High	Good (if file on disk)
Current technology	Lustre	GPFS	GPFS
Allocation mechanism	None	Capacity requested and justified in project proposal	Contract; either matching funds or fully paid by customer

Figure 1: Hierarchy of file systems at CSCS

C) STORE file system

Large research projects are often carried out by consortia, which combine many research groups and projects as identified by the CSCS call for proposal process. Research consortia share data between the individual projects and teams and they manage the data sets over a longer timespan.

CSCS offers the store file system for such consortia. In contrast to the scratch and project file systems, resource on /store is not for free, but requires a financial contribution. Up to a certain limit, academic consortia can get storage space on store on the basis of matching funds. Above the limit and for non-academic consortia, direct investment and operational costs must be fully paid. A consortium must describe its overall research plan and goals, in order to assess the strategic importance of the consortium to science and the HPC center and to define the duration of the contract.

/store is a global file system that can be accessed from all user-accessible computers at CSCS. As it is based on a hierarchical storage management system, which is to a large extent based on tape, bandwidth is lower than to the project file system.

TECHNICAL IMPLEMENTATION

All three storage levels are built with parallel file system technology in order to ensure performance scalability.

The scratch file systems are currently mainly based on Lustre, which allows for optimal read/write performance. Stability is sufficient and enhanced functionalities are not required because of the shared nature of the file system. Both, LSI and DDN storage controllers have been deployed for different implementations, mainly as direct attached scratch. Because of the meta-data performance bottlenecks in the current Lustre architecture, SSDs have been successfully tested for improving meta-data performance, although the fundamental problem of a non-distributed meta-data store can only be eased but not completely resolved with this approach.

The project file system is characterized by the combination of parallel HPC-type file system features with some enterprise storage requirements. It must be able to handle a large number of files with very good meta-data performance and has to offer functionalities like quota, snapshots, and integration with backup software. CSCS uses very similar storage hardware as on the scratch file systems, driven from separate storage servers that are connected to a high-speed Infiniband network backbone. GPFS has been selected as software technology for this file system because of its RAS features but also because superior meta-data performance compared to Lustre.

For the store file system, raw I/O performance is not as important as for the other two file systems. Technical and financial analysis showed that it is easily implemented with the same GPFS technology as /project combined with the TSM/HSM product of IBM. The TSM solution at CSCS also includes a backup and disaster/recovery functionality which enables us in the case of the total loss of the file system to recover all GPFS metadata within a few hours and all critical files within two days. By sharing licenses, infrastructure, and knowhow, operational costs can be kept low.

CSCS would be interested to change from the proprietary GPFS technology to open-source/public domain software. Lustre in its current state does not seem to be a viable option. If Lustre will be developed further in a coherent fashion, with stable funding and a clear roadmap, it could be envisaged to use in the future Lustre as the fundamental file system technology in combination with pNFS for mounting non-HPC clients.

As described above, we consider data analysis systems as an integral part of a data service. CSCS has decided to offer a portfolio of different computer architectures for data analytics: a standard, fat node cluster; a GPU cluster; a large-shared memory system based on the SGI Altix UV architecture; and a massively-multithreaded Cray XMT2 system. Access to these systems is granted within the allocation process for computing time on the main HPC systems.

CONCLUSIONS

By defining a coherent business model for data and storage, the Swiss National Supercomputing was able to simultaneously optimize costs and scientific workflow at the center. For long-term sustainability, however, users additionally have to be educated to rethink their storage needs and patterns by means of in-situ data analysis and rewriting the I/O in their codes.

CSCS considers IBM's GPFS technology to currently be the most advanced solution for highly available and powerful *global* parallel file systems. We use GPFS with its characteristics of an enterprise file system only for the global levels of the storage hierarchy. Thus, the number of required client licenses can be drastically reduced by using a small number of I/O forwarding nodes per system. The bandwidth-hungry local scratch file systems, in which every compute node is a client, are built with Lustre. Solid-state memory technologies have developed into a viable alternative or addition to storage hardware solutions, boosting latency and IOPS-sensitive components of the storage system to new performance levels.