U.S. Department of Energy Best Practices Workshop on File Systems & Archives

San Francisco, CA

September 26-27, 2011

Position Paper

doebpw5@nersc.gov.

Yutaka Ishikawa
University of Tokyo
ishikawa@is.s.u-tokyo.ac.jp

ABSTRACT

Two usability issues in storage systems connected with supercomputer are described. One is metada access and the other one comes from open source codes handling file I/O. In the latter one, because the users do not want to improve such codes, they request us to install faster disks such as SDD. According to our experiment, after improving the code, the performance is twice faster. Though twice faster disk was used, the performance was only 10 to 20 % gain. Another topic is related to a distributed shared file system being designed and deployed in Japan.

INTRODUCTION

Information Technology Center at University of Tokyo provides two supercomputer resources, SR11000 and HA8000 cluster, for domestic academic users. SR11000 is six years old machine and will be replaced with this October. HA8000 cluster consists of 952 nodes each of which has two AMD Opteron 8356s (16 cores). Each supercomputer connects with the proprietary parallel file system called HSFS. The total storage size is 1.5 PB.

In addition of computational resource services, we are currently designing and deploying distributed shared storage system, whose total size will be more than 100 PB, accessed by Japanese supercomputers including K computer, as the nation-wide high performance computing infrastructure. This infrastructure is called HPCI (High Performance Computing Infrastructure) supported by Ministry of Education, Culture, Sports, Science, and Technology. As shown in Figure 1, there two storage HUB, West and East. In West HUB, 10 PB storage and 60 PB tape archive will be deployed with K computer. 12 PB storage and 20 PB storage will be deployed in our university. The system is currently under construction and it will be operated from fall in 2012.

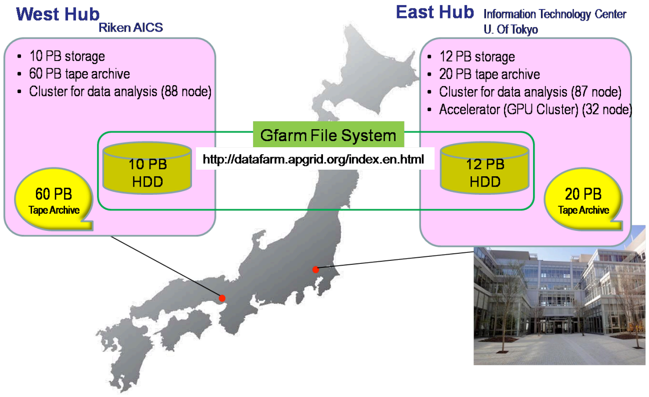
West Hub
Riken AICS
• 10 PB storage
• 60 PB tape archive
• Cluster for data analysis (88 node)
East Hub
Information Technology Center
U. Of Tokyo
• 12 PB storage
• 20 PB tape archive
• Cluster for data analysis (87 node)
• Accelerator (GPU Cluster) (32 node)
Gfarm File System
http://datafarm.apgrid.org/index.en.html
10 PB
HDD
12 PB
HDD
60 PB
Tape Archive
20 PB
Tape Archive

Figure 1 HPCI Storage

Usability Issues in HA8000 cluster

As many others pointed out, the interactive users complained slow meta data accesses in the HSFS parallel file system especially "ls –l" command that involves many meta data accesses to obtain all file statuses. To provide faster response for the interactive users, the vendor has modified its system to handle the interactive users' requests first. This modification with other several changes mitigates this slowness.

Another issue comes from open source codes that are not well programmed for file I/O. The users believe that low performance of such a code results in slow file I/O access, but this is sometimes not true. For example, a bio informatics tool, used by our bio informatics users, consists of two processing modules, genome alignment and data format change. In the genome alignment processing, there are so many critical regions, and the low performance results in those regions. After reducing the number of critical regions, the program is twice faster. In data format change processing, it opens and reads the same file about 1000 times and eventually reads 1 TB data in total. To eliminate this silly code, the program is twice faster. The users have thought if the file system is twice faster, the program would run twice faster. But, though the file system is twice faster using SDD, the performance is only 10 to 20 % improvement.

Usability Issues in HPCI storage system

This workshop may not consider distributed shared storage systems shared by different organizations. But we would like to address this kind of storage systems because data-sharing is important in the data-intensive science. One good example is ILDG (International Lattice Data Grid) where lattice QCD (Quantum chromodynamics) data generated by supercomputers are shared by international organizations. Another example is the climate simulation field. As far as we understand, the research group develops their simulation code, obtains data generated by the simulator, and after the generated data are examined and new information for them is obtained, the data are open to others who are interested in data for other purposes. Thus, the data is eventually shared by others.

The Japan HPCI tends to provide storage resource for not only traditional computational sciences but also data-intensive sciences including life science/drug manufacture, new material/energy creation, global change prediction for disaster prevention/mitigation, manufacturing technology, the origin of matters and the universe. To provide better usability for those users, issues are listed below. All issues arise because many research fields use the storage system, it is not yet predicted that their peek and sustained demands.

- ✓ Prediction of storage capacity
- ✓ Prediction of amount of file transfers

CONCLUSIONS

Dusty code must be replaced with modern code to provide usability for such application users. We have to much pay attention of distributed shared file systems with local file systems.