System Upgrade of the KEK Central Computing System

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 082038 doi:10.1088/1742-6596/898/8/082038

System Upgrade of the KEK Central Computing System

Koichi Murakami[1], Go Iwai[1], Takashi Sasaki[1], Tomoaki Nakamura[1], Wataru Takase[1]

[1] High Energy Accelerator Research Organization (KEK), 1-1 Oho, Tsukuba, Ibaraki, 305-0801, JAPAN

E-mail : koichi.murakami@kek.jp

Abstract. The KEK central computer system (KEKCC) supports various activities in KEK, such as the Belle / Belle II, J-PARC experiments, etc. The system was totally replaced and launched in September 2016. The computing resources in the new system are much enhanced as recent increase of computing demand. We have 10,000 CPU cores, 13 PB disk storage, and 70 PB maximum capacity of tape system. In this paper, we focus on the design and performance of the new storage system. Our knowledge, experience and challenges can be usefully shared among HEP data centers as a data-intensive computing facility for the next generation of HEP experiments.

1. Introduction

The KEK central computer system (KEKCC) supports various activities in KEK, such as the Belle / Belle II, J-PARC experiments, etc. The system is totally replaced every 4-5 years according to the Japanese procurement process. After the eight-months deployment of the system, the system was put into production in September 2016. The computing resources (CPU and storage) are much enhanced as recent increase of computing demand. We have about 10,000 CPU cores, 13 PB disk storage, and 70 PB maximum capacity of tape system in the new system.

Grid computing can help distribute large amount of data in geographically dispersed sites and share data in an efficient way for world-wide collaborations. But the main data centers of large HEP experiments need to take into serious consideration the aspect of managing huge amount of data. In KEK, the Belle II experiment requires that several hundreds PB data should be stored in the KEK site even if Grid computing technology is fully adopted as a distributed analysis model. The challenge is not only for storage capacity. I/O scalability, usability and power efficiency and so on should also be considered for the design of the system.

In the following, we mainly focus on the design and performance of our new storage system. Our knowledge, experience and challenges can be usefully shared among HEP data centers as a data-intensive computing facility for the next generation of HEP experiments.

2. KEKCC Upgrade

KEKCC is a HPC (High Performance Computing) Linux (Scientific Linux 6) cluster system. It consists of login servers and batch servers connected to large-scale storage systems. The storage system is connected to the computing nodes with high speed InfiniBand interconnect that realizes high I/O throughput. The KEKCC also provides a set of EMI Grid services that are used by experiment

groups (mainly for Belle II). The Grid components like worker nodes and StoRM storage management are fully integrated into the KEKC batch servers and storage system respectively.

The storage system is composed of two types of systems, one is a disk system based on the GPFS (General Purpose File System) parallel file system with 13 PB capacity, and the other is a tape library system with a total capacity up to 70 PB. The HPSS (High Performance Storage System) is used as Hierarchical Storage Management (HSM) for accessing tape data, and data I/O is automatically performed through the GPFS file system by the GHI (GPFS/HPSS Interface). This GPFS/HPSS interface enables to access tape data in the same way as for disk data.

As a job scheduler, IBM Spectrum LSF is used in the batch system. We make continuous efforts to improve job throughputs, monitoring, and optimizing queue parameters. As a result, CPU utilization is kept very high in the system.

The overall system description is shown in figure 1. System comparison with the old system is also summarized in table 1. The system resources of CPU and disk storage system was upgraded by 2.5 times and 1.8 times respectively. The total CPU capacity is counted to 236 kHS06 in the site. The technology in each system component is also upgraded as shown in the table. For example, the InfiniBand interconnect is updated to FDR (56 Gbps) from QDR (40 Gbps), that reinforces the bandwidth between storage system and CPU servers.

Table 1. System comparison between the old and new system

Component	Old System	New System	Upgrade Factor
CPU Server	IBM iDataPlex	Lenovo NextScale	
CPU	Xeon 5670 (2.93 GHz, 6cores)	Xeon E5-2697v3 (2.6GHz, 14 cores)	
CPU cores	4,000	10,024	x 2.5
Interconnect (IB)	QLogic 4xQDR	Mellanox 4xFDR	
Disk Storage	DDN SFA 10K	IBM Elastic Storage System	
HSM Disk Storage	DDN SFA 10K	DDN SFA 12K	
Disk Capacity	7 PB	13 PB	x 1.8
Tape Drive	IBM TS1140 x 60	IBM TS1150 x 54	
Tape Speed	4 TB/vol, 250 MB/s	10 TB/vol, 350 MB/s	
Tape max capacity	16 PB	70 PB	x 4.3
Power	200 kW	250-300 kW	

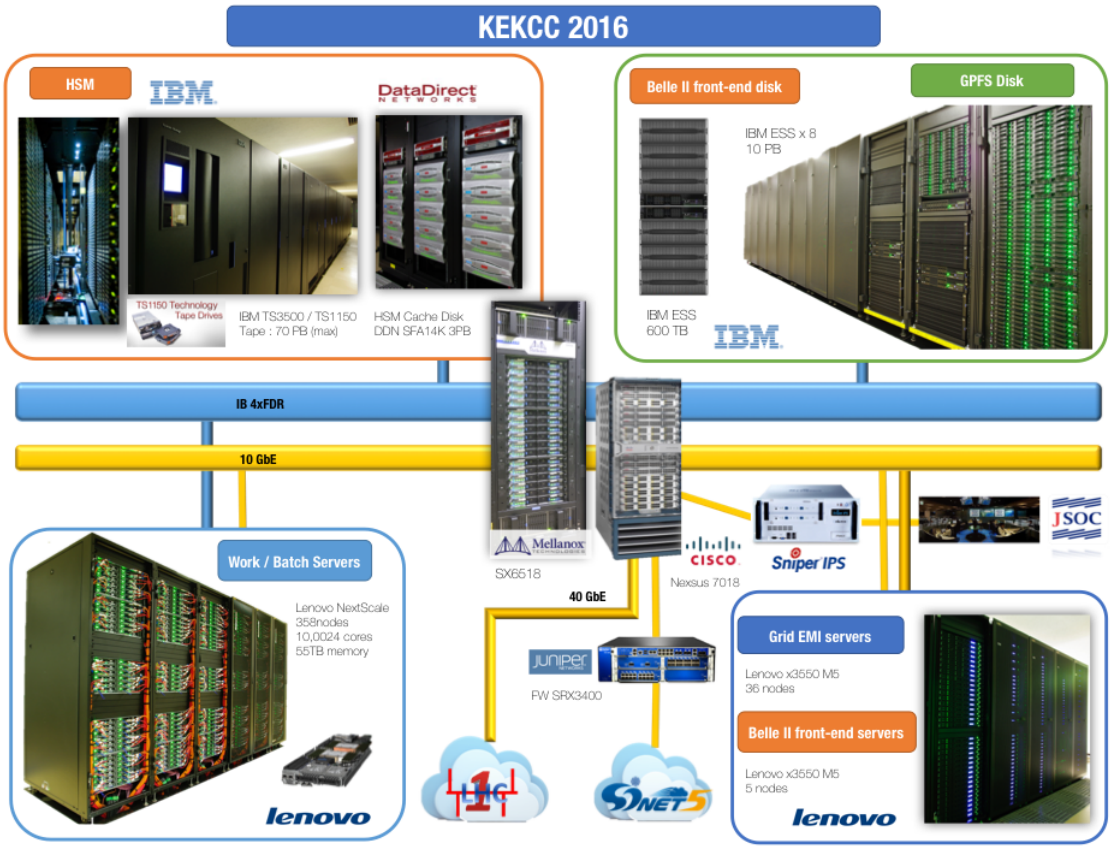

Figure 1. Overview of the new KEKCC system.

3. Storage System

3.1. Requirements on storage system

Before defining system specification, we collected the resource requirements from experimental groups. Each experiment gives the resource expectation (CPU, disk storage, and tape storage) for the next four years, and the total amount of necessary resources are estimated. For tape storage, each experimental group purchases tape medias. In the tape system, the tapes of each group are assigned to a specific storage area and managed according to the policy of class of service. Our data center provides a tape library system that can store the maximum capacity of tapes.

Data processing cycle is very important for designing the storage system. KEK is the host institute of the Belle II experiment and various J-PARC experiments. We are responsible not only for data store but also data processing. Efficient data processing cycle is the key for data analysis in experiments. We should support from raw data (including MC data) storage to DST production and storage, and storage of physics data used in end user analysis. The raw data of the next generation of experiments are often estimated to become the scale of hundreds PB. From the data center point of view, we should consider the following requirements on the storage system:

- scalability up to hundreds PB
- high availability (for data taking and processing)
- data-intensive processing with high I/O performance
- data portability to Grid service

3.2. Disk storage system

We introduce IBM Elastic Storage Server (ESS) and DDN SFA12K as disk storage system, and adopted GPFS parallel file system to realize high I/O performance. The IBM ESS of 10 PB is used for the disk system, and the DDN SFA 12K of 3PB is located as the cache disk of the HSM system.

The ESS is a storage appliance system. There are two Power8 servers and JBOD disks in each rack. There are no hardware disk controllers in the ESS system. Disk controller and GPFS file server are provided by the software in the ESS. GPFS file system can be configured across multiple ESS nodes. Each ESS node has the maximum throughput 16.8 GB/s. We configured GPFS file systems spanning over 8 racks of ESS, then over 100 GB/s total throughput is realized in the system as a design value.

There are several unique characteristics of the ESS. The first one is that the ESS supports end-to-end data integrity. Actually, a concern about silent data corruption has been reported [1]. Even if we use RAID system, there are risks of data corruption induced by unexpected bit errors in hardware level. The ESS system manages data integrity in every system operation to prevent silent data corruptions in the system. The second one is "Declustered RAID", that can reduce rebuild time by 1/3 comparing to conventional RAID system. In data center, the change of failure disks and rebuilding the disk array are usually performed. In general, rebuilding the disk system is a heavy load task, then performance degradation of the disk system might happen. In the ESS system, GPFS Native RAID uniformly spreads and declusters user data, redundancy information, and spare space across all the disks of a declustered array to reduce the rebuild tasks.

We tried to introduce Local Read-Only Cache (LROC) functionality of the GPFS file system. It enables data cache mechanism at SSD direct-attached storage in batch servers. In practical system operation, many CPU nodes might access small numbers of files in the GPFS system simultaneously. We observed critical performance degradation of the file system in such a case in the previous system. According to system benchmark tests, the readout performance with/without LROC are not so different because remote direct memory (RDMA) access of InfiniBand is fast enough comparing to SSD access speed. Also, the performance of the current system is much better than before. However, LROC could reduce the unpleasant resource concentration, and might help keep the system performance better. Unfortunately, there were bugs in the GPFS software, we had a trouble with LROC, and postponed activating LROC in the system. Now the problem was fixed, so we will plan to activate LROC again and examine its effect.

3.3. Performance of disk storage

System performance test was done for the disk storage system. We measured read / write performance for 500 GiB data files from different servers. The multiplicity of servers differs from 1 to 32, 128, 339, changing thread numbers (1,14,28). The block size of the file system was set to 8 MB. IOR and gpfsperf were used for the measurements, and we obtained consistent results for either software. The result of gpfsperf for sequential and random read / write are shown in figure 2. We obtained about 200 GB/s in sequential read, over 250 GB/s for random read, and 100 GB/s for write respectively.

We also monitor the system in real time. Figure 3 shows a snapshot of the ESS monitoring of throughput in some period. We confirmed that about 180 GB/s reading throughput is achieved in the actual system operation.

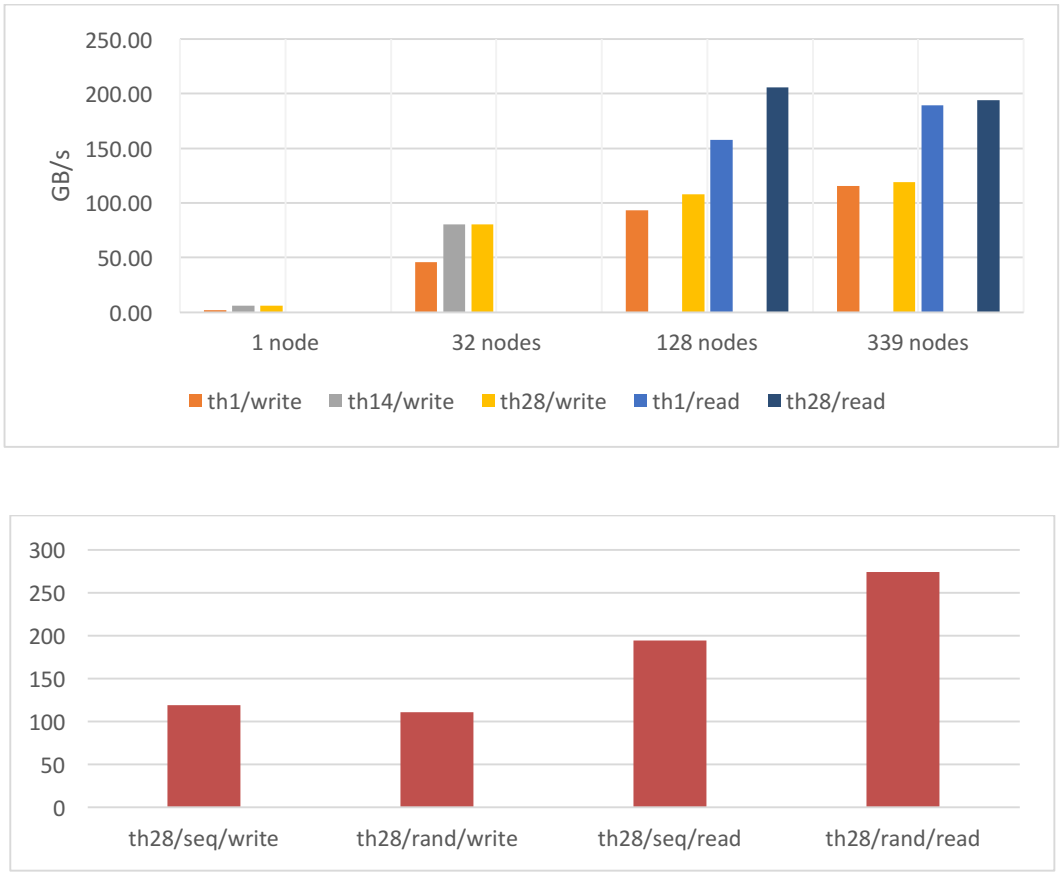

Figure 2. Benchmark results using gpfsperf. We prepared 500 GiB data files, and measure read / write performance from different servers. The multiplicity of servers differs from 1 to 32, 128, 339, changing thread numbers (1,14,28). The upper graph shows the sequential read / write performance. The lower graph shows sequential and random read / write performance.

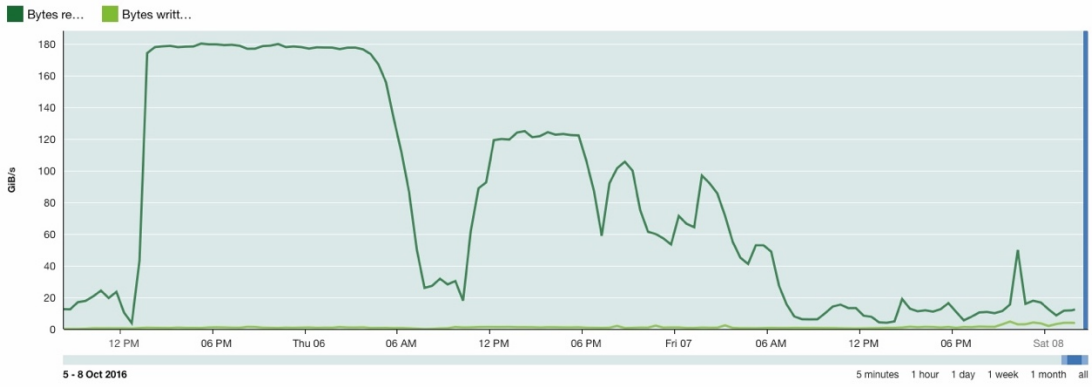

Figure 3. ESS monitoring. We can see the system total throughput (read) of 180 Gb/s as a peak value in real system operation.

3.4. Hierarchal storage management system (HSM)

The tape system is necessary for storing large amount of data produced in experiments. In terms of cost efficiency and long-term data preservation, tape technology is still important for us. One of disadvantage of tape system is that the way of accessing data is not as convenient as disk system. Hierarchal storage management (HSM) technology is beneficial to hide this demerit and enable seamless data accessing.

For the hardware specification, the tape library is composed of 13 racks of IBM TS3500 library system, that has 70 PB tape media capacity as maximum. There are 54 TS1150 drives and 12 TS1140 drives installed in the library system. The new tape media (JD) has 10 TB per volume, and the data rate is 360 MB/s. In the current tape system, we have about 5,000 tape volumes and 8.5 PB data so far. The latest tape drive supports down to the second latest type of media (JC/JD for TS1150). The old tape medias (JB) are migrated into the new media (JD). JC medias are reformatted to gain more volume (4TB to 7TB) in background operation. DDN SFA12K disk storage system of 3 PB is located as GPFS cache disk. The designed total throughput is 50 GB/s in the disk system.

We have used the HPSS system for over 15 years. Since the previous system, we introduced HPSS/GPFS Interface (GHI), that enables GPFS good performance and scalability for the data in HPSS. Users can access HSM data without minding the tape system existence behind. Once data is staged from the tape system, data can be accessed in the same way as GPFS file system. The file I/O speed is fast and the system is scalable.

We also measured the system performance in terms of tape migration and data staging assuming data processing cycle of the Belle II experiment. HPSS mover I/O performance shows 3 GB/s for concurrent read and write operations. There are 3 tape movers in the system, so the nominal migration speed is 9 GB/s in total. For actual measurements for data migration, we obtained the migration performance of 3.4 GB/s (4GB/file, 24 parallels). It corresponds to over 200 TB / day that meets the requirement of Belle II data taking. We also measured bulk staging from tape to disk assuming a situation of data reprocessing. Data recalling 1GB data file in tape-order shows 1.2 GB/s with 8 parallel processing, that is over 100 TB / day staging. These performance is reasonably good and meets the basic requirements of the Belle II experiment.

4. System Monitoring

We are monitoring the system in different ways. As online monitor, we collect system information using fluentd [2] and processing data with Elasticsearch [3] and visualize in Kibana [3] dashboards for various system resources, in addition to conventional monitoring system like MRTG [4]. The system usage for CPU and memory usage, load average and I/O rate, swap usage, and so on are monitored. When something wrong happens in the system, we can detect the bottleneck in the monitoring system. Monitoring using Elasticsearch and Kibana can be extended to various resource monitoring. We also plan to extend the monitoring system to various system parameters to understand system behavior. Figure 4 shows an example of the Kibana dashboard for the batch system resource monitoring.

Figure 4. Kibara dashboard for the batch system

5. Conclusion

The new KEKCC system was launched in September 2016. The computing resources are enhanced based on the requirements of experiment groups. The storage system is an important component in the system, so we carefully designed the system to make scalable up to hundreds PB. The performance of the disk storage system and tape storage system were measured, and we obtained good performance results that meets the requirements of experiment groups. Our knowledge, experience and challenges can be usefully shared among HEP data centers as a data-intensive computing facility for the next generation of HEP experiments.

References

[1] Bernd Panzer-Steindel, CERN/IT, Draft 1.3 8, April 2007, URL http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797
[2] http://www.fluentd.org
[3] https://www.elastic.io
[4] http://oss.oetiker.ch/mrtg/