PROCEEDINGS OF SCIENCE

# KEK Central Computer System (KEKCC)

**Go Iwai*, Hiroyuki Matsunaga, Koichi Murakami, Tomoaki Nakamura, Takashi Sasaki, Soh Suzuki and Wataru Takase**
*High Energy Accelerator Research Organization (KEK)*
*E-mail:* go.iwai@kek.jp, hiroyuki.matsunaga@kek.jp, koichi.murakami@kek.jp, tomoaki.nakamura@kek.jp, takashi.sasaki@kek.jp, soh.suzuki@kek.jp, wataru.takase@kek.jp

The High Energy Accelerator Research Organization (KEK) plays a key role in particle physics experiments, as well as supporting the related research communities in Japanese universities. In order to ensure these critical missions, KEK has two large-scale computer systems: the Supercomputer System (KEKSC) and the Central Computer System (KEKCC).
The KEKSC is used mainly for collaborative research studies in theoretical elementary particle and nuclear physics and condensed matter physics, as well as for accelerator simulations. The system is composed of two different systems: Hitachi SR16000 model M1 (System A) and IBM Blue Gene/Q (System B).
The KEKCC caters for the research demands of particle physics, nuclear physics, the photon factory, neutron science, accelerator development, theory computation, etc. In addition this system provides an information infrastructure environment including Web, e-mail and Grid (EMI and iRODS) services and supports the research activities and collaborations of KEK.
As mentioned above the EMI Grid middleware is deployed in the KEKCC for analysing and sharing experimental data over the distributed systems. The system is operated under the Worldwide LHC Computing Grid (WLCG) project. The researchers working on the Belle II, T2K, ILC and Kagra experiments perform their data analysis using the Grid infrastructure to manage large amounts of experimental data.
We would like to share our experiences and challenges related to the security, operation and experiment-specific applications, as well as the requirements for storage and computing resources, of the current KEKCC during the nearly four years of its operation. We focus in particular on the Grid Computing System. In addition we discuss prospects for the next KEKCC system, which will be newly introduced in September 2016.

*International Symposium on Grids and Clouds 2016*
*13-18 March 2016*
*Academia Sinica, Taipei, Taiwan*

*Speaker.





PoS(ISGC 2016)037

## 1. Introduction

The High Energy Accelerator Research Organization (KEK) plays a key role in particle physics experiments, as well as supporting the related research communities in Japanese universities. In order to ensure these critical missions, KEK has two different types of large-scale computer systems: the Supercomputer System (KEKSC) and the Central Computer System (KEKCC).

(a) System A – Hitachi SR16000 model M1 (b) System B – IBM Blue Gene/Q

**Figure 1:** KEK Supercomputer System (KEKSC).

The KEKSC is used for collaborative research studies in theoretical elementary particle and nuclear physics and condensed matter physics, as well as for accelerator simulations. The system is composed of two different systems, System A and System B, respectively shown in Figure 1(a) and Figure 1(b), that each attract a specific type of workload.

System A is suitable for shared memory parallel computation on a single node, as well as inter-node parallel computation with message passing. For parallel computation on the cores of a node, an automatic parallelizing compiler is available. This machine enables high-performance computing in a wide range of sciences.

System B, which is an energy-efficient supercomputer ranked 12th on the June 2012 edition of the Green500 list [1] with a rating of over 2,000 MFlops/Watt, is best suited for massively parallel computation, particularly in ultra-large applications that require super-fast inter-node communications.

The basic specifications of Systems A and B are briefly profiled in Table 1.

| System | Cores per node | Nodes | Cores | Speed (TFlops) | Memory (TB) |
|---|---|---|---|---|---|
| A | 32 | 56 | 1,792 | 54.9 | 14 |
| B | 16 | 6,144 | 98,304 | 1,255 | 96 |

**Table 1:** Basic system specification of System A and System B for the KEKSC

The KEKCC caters for the research demands of particle physics, nuclear physics, the photon factory, neutron science, accelerator development, theory computation and so on.

The accelerator physics experiments will generate hundreds of PB of data per year in the near future. For managing such an enormous amount of data within a high throughput I/O, *Hierarchical*


(a) DDN SFA10K: 7 PB of disk storage of which 3 PB is in use for the HSM disk cache. The remaining 4 PB is used directly by the users and experimental groups.

(b) IBM TS3500: 16 PB of maximum capacity in the tape library.

**Figure 2:** KEK Central Computer System (KEKCC) introduced in 2012.

*Storage Management* (HSM) is offered for the storage system of the KEKCC. The HSM in the current KEKCC consists of 3 PB of *General Parallel File System* (GPFS) for the disk staging area and 16 PB of maximum capacity for the tape library as shown in Figure 2(a) and Figure 2(b).

In order to ensure that the research demands of the various experiments are met, the KEKCC consists of several subsystems: the Data Analysis System, the Grid Computing System (EMI [2] and iRODS [3]) and general purpose IT systems, which provide common services such as E-Mail System, a Web System, Certificate Authority and so on. The Grid Computing System is operated under the Worldwide LHC Computing Grid (WLCG) project [4]. The researchers in the Belle II, T2K, ILC and Kagra experiments perform their data analysis using the Grid computing infrastructure to manage large amounts of experimental data.

Table 2 shows a summary of the experiments using the KEKCC. The top four experiments in this table are completed or ongoing projects. These experiments, Belle, and T2K, KOTO and MLF from J-PARC, are using the KEKCC computing resources mainly as a local batch system managed by LSF [5]. The bottom three experiments, Belle II, Kagra and ILC, are future projects. Belle II and ILC are actively using the Grid Computing System, which is a part of the KEKCC. Kagra has just started conducting preparation work for sharing data over the Grid computing infrastructure with other gravitational wave detection experiments, such as LIGO [6] in the US and VIRGO [7] in the EU.

The KEKCC is replaced in its entirety every 4–5 years according to the Japanese government procurement policy for computer systems. We therefore purchase the system through international bidding according to the Agreement on Government Procurement (GPA) of the WTO. The bidding process usually takes at least one and a half years. We purchase a *Service*, which includes leased hardware, as well as service implementation including operational staff. An example of the timeline for a typical bidding process is shown in Table 3.

More detailed hardware specifications of the KEKCC and operational statistics, in particular for the Grid Computing System, during the current KEKCC contract period are described in Section 2. Then the next KEKCC to be introduced in 2016 is outlined in Section 3.

| Experiments | | Start Year | Overview |
|---|---|---|---|
| Belle | | 1999 | Precise measurements for CP violation with an electron-positron asymmetric-energy collider (KEKB). |
| T2K | (J-PARC) | 2009 | Neutrino experiment for measuring neutrino mass and flavour mixing. Neutrinos are shot from Tokai to the detector at the Kamioka mine (a distance of 300 km ). |
| KOTO | (J-PARC) | 2010 | Various experiments for kaon and hadron physics. |
| MLF | (J-PARC) | 2010 | Neutron diffraction, neutron spectroscopy, nano-structure analysis, neutron instruments, muon spectroscopy |
| Belle II | | 2017 | Next generation of Belle experiment with an upgraded accelerator SuperKEKB, which produces higher luminosity of electron-positron collisions than does KEKB. Belle II is aimed to discover new physics beyond the *Standard Model*. |
| Kagra | | 2018 | Gravitational wave detection experiment at Kamioka. |
| ILC | | 2020 or later | More precise measurement of Higgs boson created in electron-positron collisions with long-distance linear accelerator. |

**Table 2:** Experiments using the KEKCC

| Year | Month | Event |
|---|---|---|
| 2015 | Jan | Organizing a committee to determine the system specification. |
| 2015 | Jun | Publishing a draft version of document for system specification. |
| 2015 | Sep | Publishing a final version of document for system specification. |
| 2015 | Dec | Bidding. |
| 2016 | Sep | Launching a new system. |

**Table 3:** Example of bidding process for the large-scale computer system in Japan

## 2. Resource Scale and Utilisation

As mentioned in Section 1, the KEKCC consists of several subsystems. The Data Analysis System is the most important system in the KEKCC. It provides *Login Servers* and *Calculation Servers*, which comprise 4,080 cores of Intel Xeon X5670 on the IBM iDataPlex. This system also provides the Storage System, which consists of a GPFS parallel file system and *High Performance Storage System* (HPSS) tape library (IBM TS3500 and TS1140×60). The capability of the disk storage of DDN SFA10K is currently 7 PB. 4 PB of GPFS is currently used for users' home directories and shared spaces for experimental groups. In addition to providing disk stor-

age the system provides the HSM with an HPSS for storing and analysing large amounts of data within a high throughput performance environment. The remaining 3 PB of GPFS is used for the GPFS-HPSS-Interface (GHI) as a file staging area for the HSM.

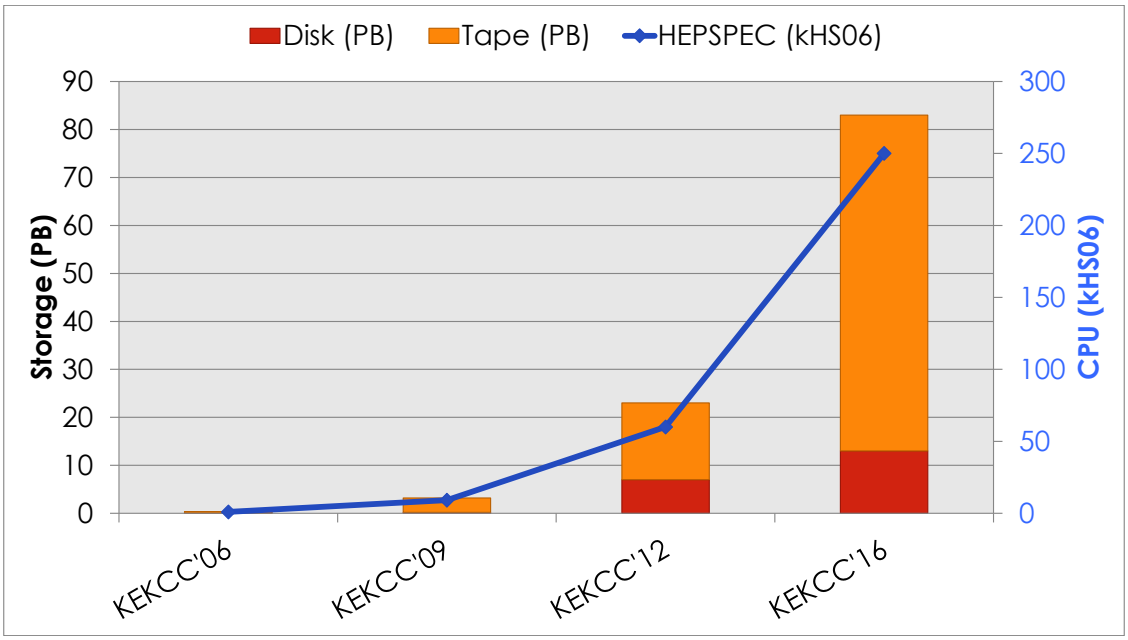
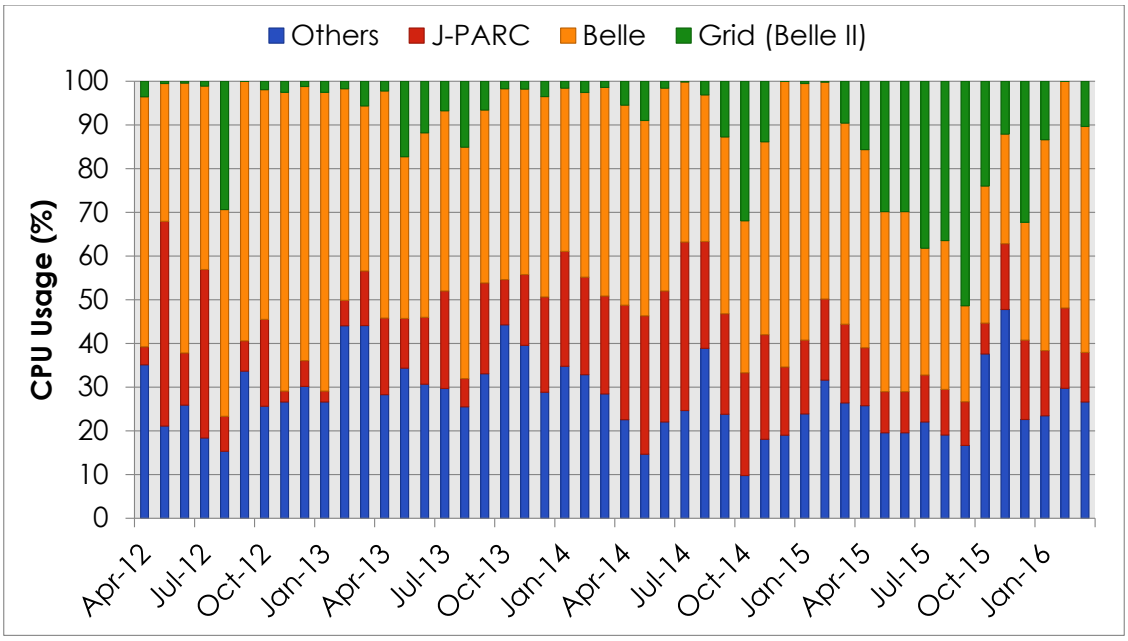



**Figure 3:** Resource growth history.

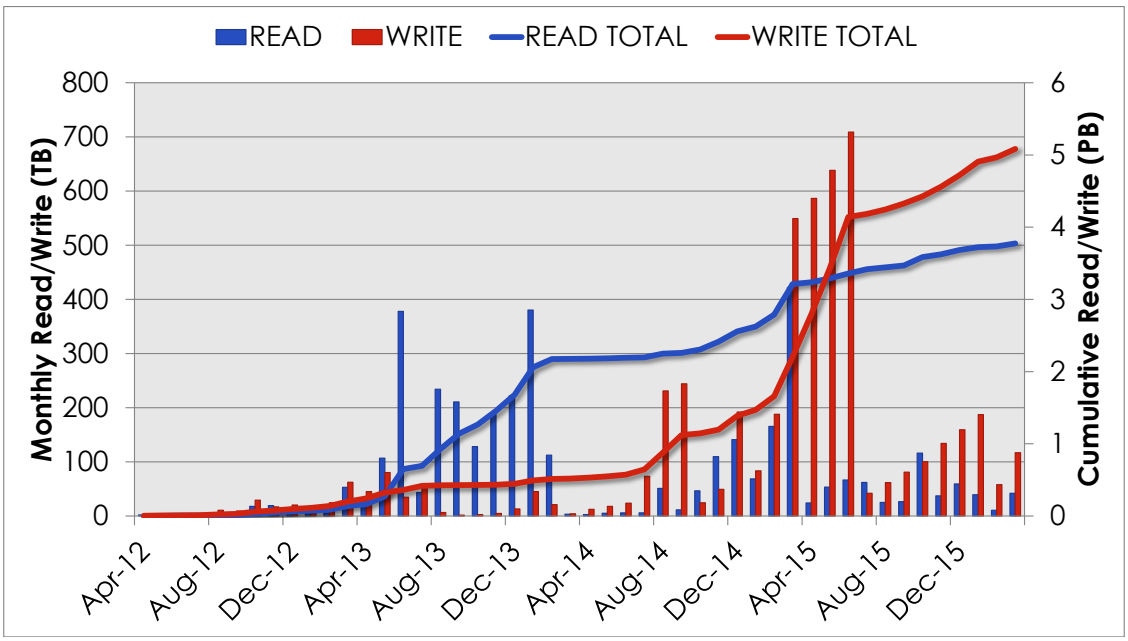
Figure 3 shows the resource growth history for the last three KEKCCs and the expected resource scale in the next system. Bars shaded red and orange respectively show the disk capacity and maximum capacity of the tape library in PB. The blue line shows the total CPU performance of the KEKCC in HEPSPEC [8], which is scaled to the vertical axis on the right-hand side. The current KEKCC, labelled *KEKCC'12*, has 60,000 HEPSPEC of CPU performance, 7 PB of disk storage and 16 PB of maximum capacity for the tape library. 7 PB of data is stored in the tape library as of March 2016. The expected resources for the next KEKCC, labelled *KEKCC'16*, will be 250,000 HEPSPEC of CPU performance, 13 PB of disk storage and 70 PB of maximum capacity for the tape library.

The statistics of CPU utilisation for the last four years is shown in Figure 4(a). The KEKCC was almost fully utilised by local batch jobs from KEK (blue), J-PARC (red) and Belle (orange), in the first three years. Grid jobs (green) then used 31% of the KEKCC resources on average during the 5th series of Belle II MonteCarlo (MC) campaign, which took place between August and December in 2015. Although Grid jobs come from several Virtual Organisations (VOs), Belle II is dominant in the KEKCC at this moment.

Figure 4(b) shows another four-year statistics for read and write sizes into the HSM via SRM [9] (StoRM [10]). The blue and red bars show the monthly read and write sizes in PB respectively. The cumulative read and write sizes in PB are shown by the blue and red lines, scaled to the vertical axis on the right-hand side. Several Belle II MC campaigns have taken place since the middle of 2014 and a cross-sites Data Challenge took place in 2016 to check the throughput performance. An MC campaign gradually increases the number of production events. This resulted




(a) Monthly CPU usage in the KEKCC




(b) Monthly read and write size into the HSM via SRM (StoRM)

**Figure 4:** Monthly resource utilization statistics in the KEKCC between April 2012 and March 2016.

in a peak of more than 700 TB write size in June 2015. As a result 4 PB of readout and 5 PB of writing was achieved during the four years of operation. It is noted that Figure 4(b) does not show the data size through internal connections.

## 3. New System to be Started in September 2016

The requested resources for the new system are summarised in Table 4. The number of requested resources is 10,800 for CPU cores, 17 PB for disk storage and 65 PB for tape storage. *KEKCC'16* will almost meet these requests.

| Experiments | CPU (cores) | | Disk (PB) | | Tape (PB) | |
|---|---|---|---|---|---|---|
| Belle | 1,000 | | 1.2 | | 3.5 | |
| Belle II | 7,500 | | 9 | | 29 | |
| ILC | 400 | | 0.3 | | 1.5 | |
| CMB | 250 | | 0.5 | | 1 | |
| J-PARC | 1,650 | | 5.9 | | 27 | |
| KOTO | 1,000 | | 5 | | 15 | |
| T2K | 300 | | 0.2 | | 1 | |
| MLF | 50 | | 0.5 | | 8 | |
| Others | 300 | | 0.2 | | 3 | |
| Total Requirements | 10,800 | | 17 | | 65 | |
| KEKCC'16 | 10,000 | (−7%) | 13 | (−24%) | 70 | (+8%) |

**Table 4:** Amount of requested and resources in fact introduced in KEKCC'16

Table 5 shows a summary of the difference between the current KEKCC (KEKCC'12) and the next KEKCC (KEKCC'16).

The number of CPU cores will be 10,000, an increase of a factor of 2.5, in KEKCC'16. The system-wide CPU performance can be estimated at 250,000 HEPSPEC, which is four times higher performance than that of KEKCC'12. Disk capacity will be increased to 13 PB ($\times$1.8) and maximum capacity will be 70 PB ($\times$4.3) for the tape library.

The significant change in the Grid Computing System in KEKCC'16 is that many Belle II-critical services, e.g., LFC, SRM, AMGA, FTS and CVMFS Stratum 0, are isolated from the other VOs to provide a more stable operation with no downtime.

For example KEKCC'16 will have two read-write Belle II-dedicated LFCs. In addition to these LFCs, two read-only LFCs with no GSI authentication will be deployed for faster access and larger throughput. It will also have an LFC service for other VOs. This will be shared by several VOs other than Belle II.

As a further example, there will be two different types of GridFTP for Belle II. The first is for raw data transfer from KEK in Japan to PNNL in the US. Belle II generates and stores data into the front-end storage at a data rate up to 2 GB/sec. Then processed data will be transferred to the HSM at a data rate up to 4 GB/sec. There are two dedicated GridFTP servers with four lines of 10 Gbps each for raw data transfer. It is assumed that this configuration will provide sufficient bandwidth until 2020.

For the second type of GridFTP, there is a dedicated server with two lines of 10 Gbps for data analysis, which is for any activities other than raw data transfer. Finally there are also have two GridFTP servers for other VOs. Like the LFCs mentioned above, these services are shared by other

| | Current (KEKCC'12) | New (KEKCC'16) | Upgrade Factor |
|---|---|---|---|
| CPU Server | IBM iDataPlex | Lenovo NextScale | |
| CPU | Xeon 5670 (2.93 GHz, 6 cores/chip) | Xeon E5-2697v3 (2.60 GHz, 14 cores/chip) | |
| Number of CPU Cores | 4,000 | 10,000 | ×2.5 |
| HEPSPEC | 60,000 | 250,000 | ×4.1 |
| InfiniBand | QLogic 4xQDR | Mellanox 4xFDR | |
| Disk Storage | DDN SFA10K | IBM Elastic Storage System (ESS) | |
| HSM Disk Storage | DDN SFA10K | DDN SFA12K | |
| Disk Capacity | 7 PB | 13 PB | ×1.8 |
| Tape Drives | IBM TS1140 ×60 | IBM TS1150 ×54 | |
| Tape Speed | 250 MB/s | 350 MB/s | |
| Tape Max Capacity (PB) | 16 PB | 70 PB | ×4.3 |
| Power Consumption | (actual monitored value) 200 kW | (max estimation) 400 kW | |

**Table 5:** Difference between current system (KEKCC'12) and new system (KEKCC'16)

VOs, except for Belle II.

In addition to Belle II-dedicated services, there will be some generic service instances in KEKCC'16. We will start CVMFS [11] Stratum 0 as well as Stratum 1 in September 2016. Stratum 1 was launched as a pre-production service in September 2015 and shows almost production service quality. We are providing this service for several Japanese as well as other Asian institutes. We will start the Stratum 0 of the domain `kek.jp`. In addition to the CVMFS, we are also starting a Dropbox-like [12] cloud storage service for the users of KEKCC'16.

## 4. Discussion

The KEKCC has constituted a large-scale computing infrastructure for more than 1,400 researchers from the many high energy and nuclear physics experiments including Belle and Belle II. The Belle II *Physics Run* will start in 2017. The KEKCC will therefore focus more energy on the Belle II experiment in the next few generations of the system.

As mentioned in Section 1, the KEKCC is replaced in its entirety every 4–5 years according to the Japanese government procurement policy for computer systems. This policy does not apply only to the KEKCC, but also almost all procurements of large-scale computer systems in Japanese governmental institutes. Therefore, the procurement for the next KEKCC after KEKCC'16 will start in 2019, and then KEKCC'16 will be replaced with the new system in 2020.

4 PB of data on the disk storage and 7 PB of data in the tape library should be migrated for the system replacement in 2016. Because of this migration work, we plan two weeks of read-only on

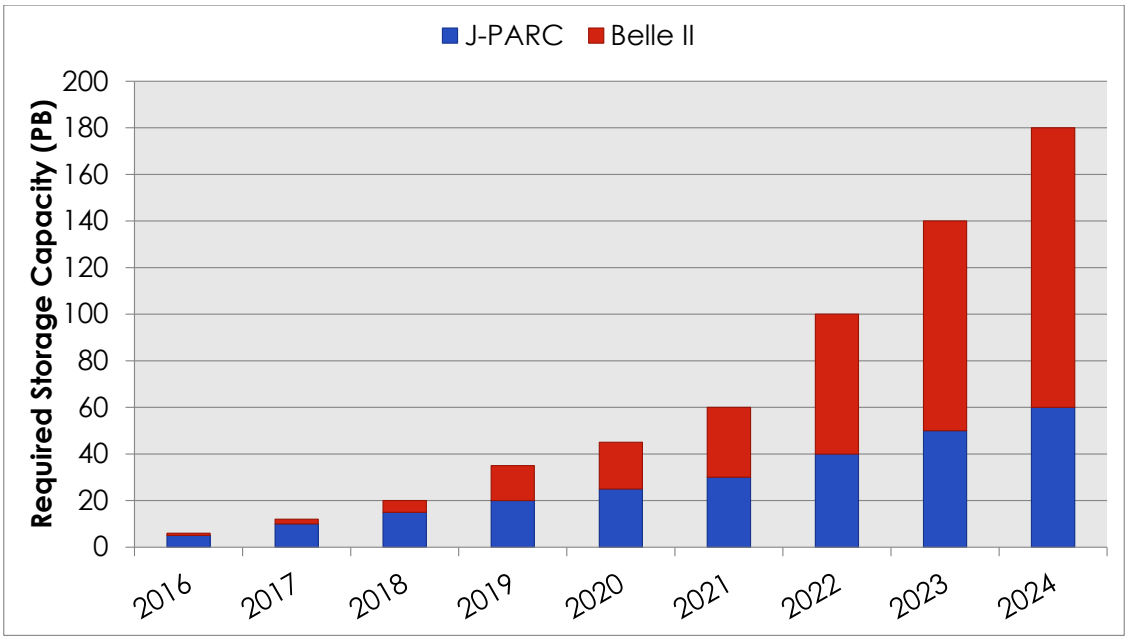



**Figure 5:** Data growth expectation in the near future.

the HSM and a few days of downtime for the disk storage. The new tape library, i.e. IBM TS3500, in KEKCC'16 fortunately has the capability to mount generations of the tape medium, which is currently mounted on the tape library in KEKCC'12. Thus a relatively short duration of downtime will be required to migrate the data on the tape library.

Figure 5 shows the expected data growth until 2024. The blue and red bars respectively show the expected data volume on the disk storage and in the tape library generated by J-PARC and Belle II in PB. It is anticipated that nearly 50 PB of data will be stored by the end of the KEKCC'16 contract. It is suggested that $N \times 10$ PB of data migration will be required in 2020. Thus $240 \times N/D$ GB/sec of file transfer performance for $D$ days downtime has to be achieved. The worst scenario is that an entire data migration will be required because of the incapability of the new tape library. We have to begin to consider data migration without this unacceptable scenario for the next system after KEKCC'16.

*Security* must also be considered. Although we have not detected serious security incidents in the KEKCC during the four years of the current contract period, we have been observing an increasing trend in security issues, e.g., host intrusion, virus infection, data loss and so on, over the whole site of KEK including university groups that we are supporting. Also, the propagation areas and velocity of threats have increased over recent years with the growing popularity of distributed systems, such as Grid and Cloud. It is becoming difficult to detect, identify and trace threats.

On the other hand, in October 2015, Japanese government started *My Number System* that issues a 12-digit ID number to all citizens and residents of Japan individually. There were, however, 200 or more frauds or suspected frauds in two months after launching this system. Thus, many of Japanese people are currently concerned with a keyword of *Security*.

These situation surrounding a keyword of *Security* suggests the necessity of a CSIRT federa-

tion across sites.

## 5. Conclusion

In this paper, the KEK Central Computer System (KEKCC) was described, in particular regarding Grid computing. The current KEKCC has been in operation since April 2012 and will be decommissioned in August 2016. An entirely replaced new KEKCC system will come into service in September 2016.

As discussed in Section 4, the KEKCC has constituted a large-scale computing infrastructure for many HENP experiments, which generate vast amounts of data. Therefore we have been relying on the HSM to provide a cost-effective storage with high I/O performance.

In the new KEKCC, several Belle II-dedicated services will be provided for more stable operation of the Belle II-mission-critical services. In addition to deploying these dedicated services, we are upgrading the UPS as well as the cooling system in the new KEKCC facility to attain operation without downtime.

In the future, at the time when the entire system is replaced, data migration will present some challenges. We are required to perform feasibility studies and so on prior to the actual migration work.

## References

[1] W. Feng and K. W. Cameron, *The Green500 List: Encouraging Sustainable Supercomputing*, *Computer* **40** (2007) 50–55.

[2] "European Middleware Initiative (EMI)." `http://www.eu-emi.eu/`.

[3] A. Rajasekar, R. Moore, C. Hou, C. A. Lee, R. Marciano, A. de Torcy, M. Wan, W. Schroeder, S. Chen, L. Gilbert, et al., *iRODS Primer: Integrated Rule-Oriented Data System*, *Synthesis Lectures on Information Concepts, Retrieval, and Services* **2** (2010) 1–143.

[4] "Worldwide LHC Computing Grid (WLCG)." `http://wlcg.web.cern.ch/`.

[5] S. Zhou, *LSF Load Sharing in Large Heterogeneo us Distributed Systems*, in *Workshop on Cluster Computing*, 1992.

[6] A. Abramovici, W. E. Althouse, R. W. P. Drever, Y. Gürsel, S. Kawamura, F. J. Raab, D. Shoemaker, L. Sievers, R. E. Spero, K. S. Thorne, R. E. Vogt, R. Weiss, S. E. Whitcomb, and M. E. Zucker, *LIGO: The Laser Interferometer Gravitational-Wave Observatory*, *Science* **256** (1992), no. 5055 325–333.

[7] C. Bradaschia, R. D. Fabbro, A. D. Virgilio, A. Giazotto, H. Kautzky, V. Montelatici, D. Passuello, A. Brillet, O. Cregut, P. Hello, C. Man, P. Manh, A. Marraud, D. Shoemaker, J. Vinet, F. Barone, L. D. Fiore, L. Milano, G. Russo, J. Aguirregabiria, H. Bel, J. Duruisseau, G. L. Denmat, P. Tourrenc, M. Capozzi, M. Longo, M. Lops, I. Pinto, G. Rotoli, T. Damour, S. Bonazzola, J. Marck, Y. Gourghoulon, L. Holloway, F. Fuligni, V. Iafolla, and G. Natale, *The VIRGO Project: A Wide Band Antenna for Gravitational Wave Detection*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **289** (1990) 518–525.

[8] M. Michelotto, M. Alef, A. Iribarren, H. Meinhard, P. Wegner, M. Bly, G. Benelli, F. Brasolin, H. Degaudenzi, A. D. Salvo, I. Gable, A. Hirstius, and P. Hristov, *A Comparison of HEP code with SPEC benchmarks on multi-core worker nodes*, *Journal of Physics: Conference Series* **219** (2010) 052009.

[9] "Storage Resource Management (SRM) Working Group." http://sdm.lbl.gov/srm-wg/.

[10] R. Zappi, L. Magnoni, F. Donno, and A. Ghiselli, *StoRM: Grid Middleware for Disk Resource Management*, in *CHEP'04*, pp. 1238–1241, 2005.

[11] P. Buncic, C. A. Sanchez, J. Blomer, L. Franco, A. Harutyunian, P. Mato, and Y. Yao, *CernVM – a virtual software appliance for LHC applications*, *Journal of Physics: Conference Series* **219** (2010) 042003.

[12] "Dropbox." https://www.dropbox.com/.