CCJ Operation in 2006-2007

S. Yokkaichi, H. En'yo, Y. Goto, H. Hamagaki,[*1] T. Ichihara, S. Kametani, T. Nakamura, Y. Watanabe

1 Overview

The operation of CCJ[1,2], RIKEN Computing Center in Japan for RHIC[3] physics, started in June 2000 as the largest off-site computing center for the PHENIX[4] experiment at RHIC. CCJ was initially planned to perform three roles in PHENIX computing, 1) as the simulation center, 2) as the Asian regional center and 3) as the center of spin physics. Recently, DST (Data Summary Tape) production from raw data has become more important, especially for the p+p data. Out of the many off-site computing facilities of PHENIX, only CCJ can handle the several hundreds of TB of raw data in use of HPSS (High Performance Storage System)[5] for the time being.

A joint operation with RSCC (RIKEN Super Combined Cluster System)[6] was started in Mar. 2004. Most of our computing power is now provided by RSCC. On the other hand, the disk storage and service nodes are still located at the CCJ machine room in RIKEN Main Building and maintained by ourselves. HPSS is shared by CCJ and RSCC, while the CCJ data occupy more than 90% of the data stored in the HPSS.

Many analysis and simulation projects are being carried out at CCJ. They are shown on the web page: `http://ccjsun.riken.go.jp/ccj/proposals/`.

2 Current configuration

In the CCJ machine room, we have approximately 190 PC nodes operated using Linux, 166 nodes are used for calculation and the others are used for various services, e.g., data transfer. Each calculation node has 1 GB of memory, 10–31 GB of local disk and dual CPUs (Pentium III 700 MHz – 1.4 GHz, Pentium 4 2.0 GHz). Scientific Linux (SL) 3.0.5 is operated on the calculation nodes, similarly to at RCF (RHIC Computing Facility)[7], which is the main analysis facility for PHENIX. The upgrade to SL4 is planned in 2008. About 90 nodes are currently operational as calculation nodes and the others are waiting to be retired. Out of the operational nodes, 36 nodes have been augmented with 300 GB of local disk, on which 10 TB of nDST data are located in order to avoid the overhead of the data transfer from the data servers or HPSS. However, a user should take care to submit a job to the node on which the required nDST file is located.

Each RSCC calculation node has 2 GB of memory, 100 GB of local disk space and two Xeon 3.06 GHz processors. Out of the 1024 calculation nodes in RSCC, 128 nodes are dedicated to CCJ usage. They share the PHENIX software environment and can access the CCJ data servers as well as the nodes in the CCJ machine room.

We have two log-in servers. The older server (SUN E450) operated by Solaris 2.6, on which the mail and http servers were also running, was replaced by an HP PC server operated by SL5 in Nov. 2007. Following RCF, we have required the SSH public-key authentication in order to log-in to CCJ since Sept. 1, 2007.

We have a main server (SUN Fire V880) using Solaris 8 for the NIS/DNS/NTP server and the NFS server for the users' home region on the 8 TB SCSI-RAID.

At the end of JFY 2006, two data servers (SUN Fire V40z) and four SATA-RAID systems (9 TB each) were purchased and deployed to replace the systems retired in 2006: a PC data server, two SUN E450 data servers and 24 TB SCSI RAID disks. Each SATA-RAID is served as a single partition using LVM (Logical Volume Manager)[8] with XFS[9].

Including two above servers, we have five V40z data servers, which are operated by SL4 and connected to large RAID systems (67 TB in total). Except for the users' work regions (40 GB is assigned to each user), none of the disks can be accessed by NFS from the calculation nodes to maintain the total I/O throughput. The *rcp* command is used to access the data with a limitation on the number of *rcp* process to avoid congestion in each data server. In the first half of JFY 2007, they tended to hang up being triggered by the heavy I/O load. Particularly the server of the users' work-regions was found to hang up most frequently. We purchased a SUN M4000 operated by Solaris 10 to use as a new NFS server for the users' home- and work-regions, in expectation of the stability. It was delivered at Feb. 2008 along with a 12 TB FC-RAID system.

On the calculation nodes, the batch queuing system LSF[10] is operated. In Mar. 2008, they were upgraded from version 6.0 to 7.0. On the CCJ-dedicated nodes in RSCC, LSF version 6.0 was upgraded to 6.2 in Feb. 2008.

HPSS version 6.2 is used as a mass-storage system at CCJ and RSCC. Approximately 1.2 PB of CCJ data (200,000 files) have been stored as of Dec. 2007. Five IBM p630 servers operated using AIX are used as the HPSS core server and data/tape movers. For CCJ, eight T9940B tape drives (30 MB/s I/O with capacity 200 GB/cartridge), two T10000 drives (120 MB/s I/O with 500 GB/cartridge) and 6,110 tape cartridges are located in two StorageTek PowderHorn 9310 tape robots. A robot can handle approximately 5-6,000

[*1] CNS, University of Tokyo

tapes, and thus the two robots have a capacity of 11,083 tapes and 8,620 tapes are already installed.

3 PHENIX software environment

Two AFS[11)] clients are operated using OpenAFS on Linux to share the software environment of the PHENIX experiment as analysis and simulation libraries, configuration files and so on. To cope with the instability caused by OpenAFS, one client is experimentally operated as a virtual machine using Xen[12)], in order to reset automatically when it hangs up.

The total size of the PHENIX software copied by AFS daily or weekly from BNL is approximately 300 GB as of Dec. 2007. All the calculation nodes are served the software by NFS, not by AFS. Using a *rsync* server on a PC, the PHENIX software environment are also shared by PHENIX collaborators in Japan. Because the PHENIX analysis libraries are compiled using the shared libraries provided by the operating system on the calculation nodes, the OS used in CCJ should be the same as that used in RCF as mentioned above. It is unrealistic to recompile the software on the other local OS by ourselves, because they contain 14 GB of CVS data, which are being updated daily.

For PHENIX analysis, PostgreSQL is used as the DB engine to store the calibration data. The total amount of data in the PHENIX-DB is 40-50 GB as of Dec. 2007. The data are copied to CCJ by request of users. Three Linux PCs are used to operate the DB in CCJ.

4 WAN and data transfer

In Oct. 2006, a 10Gbps network (Super SINET, which was integrated as SINET3 in Apr. 2007), which is maintained by NII[13)], became available at RIKEN Wako Campus. Then, the bottleneck in the data transfer between CCJ and BNL became approximately 1.5 Gbps of bandwidth (B/W) between the CCJ main switch (Catalyst 4506) and the RIKEN firewall. They were connected by the aggregated two 1000BASE fibers, but the total B/W was found to be limited to 1.5 Gbps due to the firewall performance. To remove the bottleneck, a 10GBASE switch (Foundry FESX424) was newly introduced in CCJ machine room in Nov. 2007 with a 10GBASE line from the switch to the RIKEN main switch.

In the second half of JFY 2007, four PC servers (HP ProLiant DL145 G3) and four SATA-RAID disks (12 TB each) were deployed as 'buffer boxes' for the data transfer between CCJ and BNL. Each node has two buffer areas A and B, each with a 2 TB. The raw data are transferred from BNL using GridFtp[14)] and stored in buffer A. When A is full, B is used to store the data and the transfer to HPSS from A is started, and vice versa.

The buffer boxes are operated by SL5 with a Grid environment (VDT[15)] 1.8.1). Each server has two network I/F. One is connected to the new 10GBASE switch and communicates only with BNL. The limitation is set by the access control list on the RIKEN main switch. The other is connected to the CCJ main switch and communicates with all but BNL. Using the new setup, a 360 MB/sec transfer rate from BNL to CCJ was achieved in Jan. 2008 as shown in Fig.1.

No p+p data were obtained in PHENIX Run-7 (Nov. 2006–May 2007) and thus no raw data were transferred to CCJ in 2007. In Run-8 (Nov. 2007–Mar. 2008), 100 TB of p+p data were transferred using the new data-transfer machines in Feb. and Mar. 2008.

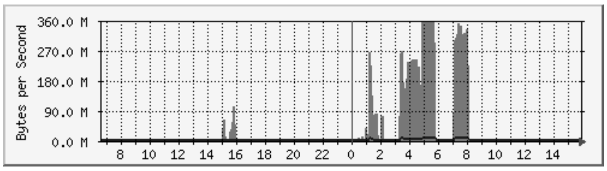

Fig. 1. Data transfer rate between CCJ and BNL achieved in Jan. 2008 with the new 10GBASE switch.

5 Outlook

After five years of operation, RSCC will be renewed in Jan. 2009. We have been taking part in the discussion on the detail of the renewal. Replacement of the tape storage system by disk storage was also discussed but was not adopted.

For CCJ, two big changes are being discussed; 1) a new tape robot for HPSS will be deployed and located in the CCJ machine room, 2) calculation nodes assigned to CCJ will be installed with large disks (the typical size is expected to be ∼1 TB), and the nDST data will be located in the same manner as that of above-mentioned local data disks in a part of CCJ calculation nodes.

References

1) `http://ccjsun.riken.go.jp/ccj/`
2) S. Kametani et al., RIKEN Accel. Prog. Rep. 39, 224 (2006); RIKEN Accel. Prog. Rep. 40, 197 (2007).
3) `http://www.bnl.gov/rhic`
4) `http://www.phenix.bnl.gov`
5) `http://www.hpss-collaboration.org/`
6) `http://rscc.riken.jp`
7) `http://www.rhic.bnl.gov/RCF/`
8) `http://www.lvm.org/`
9) `http://www.xfs.org/`
10) `http://www.platform.com/products/LSF/`
11) `http://www.openafs.org/`
12) `http://www.xen.org/`
13) `http://www.nii.ac.jp/`
14) `http://www.globus.org/grid_software/data/gridftp.php`
15) `http://vdt.cs.wisc.edu/index.html`