CCJ Operation in 2007-2008

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

1 Overview

The operation of CCJ[1,2], RIKEN Computing Center in Japan for the 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 March 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. Renewal of RSCC is on going and new CPUs will be available in August 2009.

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

2.1 PC nodes

In June 2008, 112 PC nodes, which were purchased in 1999–2001, were retired and discarded. After that, we have approximately 80 PC nodes operated using Linux in the CCJ machine room. In these nodes, 54 nodes are used for calculation and the others are used for various services, e.g., data transfer, monitoring of nodes and network performances, database servers, log-in servers, WWW and mail server, and so on. Each calculation node has 1 GB of memory, 2 GB of swap area, 10–31 GB of local disks and dual CPUs (Pentium III 1.4 GHz and Pentium 4 2.0 GHz). Scientific Linux (SL) 4.4 is operated on the calculation nodes similarly to at RCF (RHIC Computing Facility)[7], which is the main analysis facility for PHENIX. At CCJ, upgrade from SL 3.0.5 was performed in April 2008.

Out of the calculation nodes, 36 nodes were augmented with 300 GB of local disk in March 2006, on which 10 TB of nDST (nano-DST) data are located in order to avoid the overhead of the data transfer from the data servers or HPSS. This "Data in local disk" scheme is used by users with applause. These nodes were so old – purchased in March 2002 – that 14 times of system-disk crash occurred in 2008. The second-hand SCSI disks recycled from the retired servers and RAIDs were used for repairing.

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 were dedicated to CCJ usage. In August 2008, the dedicated node was reduced to 64 because the congestion of the entire CPUs in RSCC. The dedicated nodes share the PHENIX software environment and can access the CCJ data servers as well as the nodes in the CCJ machine room. In order to run the PHENIX software, the operating system should be same as PHENIX, namely, SL 4.4 at present. This is why we need the 'dedicated nodes', in other words, we cannot share the CPU with the another RSCC users.

On the calculation nodes, the batch queuing system LSF[8] is operated. Version 7.0 is used in CCJ, and 6.2 is used in the CCJ-dedicated nodes in RSCC.

Six temperature monitors were deployed in September 2008 and monitored by WEB using Cacti[9] graphical tool, which was newly installed in 2008 and is also used to visualize the LSF performance and so on.

In February 2009, following success of the "Data on local disk" scheme, 18 nodes of new PC servers (HP ProLiant DL180 G5) were delivered. Each node has dual CPU (Quad–Core Xeon 2.66 GHz), two 146 GB SAS disks for the system, eight 1 TB SATA disks for the data storage, and 16 GB of memory. I/O bound jobs like the nDST analysis will be performed in the servers.

2.2 Data servers

We had a main server (SUN Fire V880) using Solaris 8 for the NIS/DNS/NTP servers and the NFS server for the users' home region on the 8 TB RAID. This machine was replaced by a new server machine, SUN Enterprise M4000 using Solaris 10 with the 10.5 TB of FC-RAID. The DNS, NIS, and NTP/NFS servers were replaced by the new machine in June 2008, October 2008, and January 2009, respectively. The home region is formatted by VxFS[10] and served by NFS v3. The old server and RAID were discarded in Feb. 2009.

We have five data servers (SUN Fire V40z), which are operated by SL4. They serve five SATA-RAID systems (45 TB) and three FC-RAID systems (26 TB) which are formatted using XFS[11]. In order to keep the total I/O throughput, *rcp* command is used to read the data on these servers instead of NFS.[2]

In 2008, frequency of the hang-up of V40z servers

[*1] CNS, University of Tokyo

was very low unlike the previous year.

2.3 HPSS

HPSS version 6.2 is used as a mass-storage system at CCJ and RSCC. As the first step of the renewal of RSCC, tape robots and HPSS servers were renewed in November 2008. A new robot and servers are located on CCJ machine room, instead of the RIKEN IT Building where the old system was located. Three UPSs (10.5 KVA each), additional power line and additional air-conditioning system for the robot and servers were newly equipped.

For the new HPSS core server and disk/tape movers, seven IBM p570 servers are operated using AIX 5.3. Twelve (six for CCJ) LTO-4 drives (120 MB/s I/O with 800 GB/cartridge), which are connected six (three for CCJ) tape movers, and 5000 LTO tapes (3000 for CCJ) are installed in the tape robot IBM TS3500, which can handle up to 10275 LTO tapes in current configuration. Since the old tapes cannot be used in the new tape drives, copying the data to the new tapes in the new robot was started in December 2008. Copying approximately 1.34 PB (2.02 M files) of CCJ data stored as of 1 January 2009 will be completed until the end of April 2009.

3 PHENIX software environment

Two AFS[12] 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. The total data size copied by AFS daily or weekly from BNL is approximately 330 GB as of January 2008. All the calculation nodes are served the software by NFS, not by AFS. Using a *rsync* server, the PHENIX software environment are also shared by PHENIX collaborators in Japan.

Three PostgreSQL servers are operated for the calibration database of PHENIX. The data size is approximately 60 GB as of January 2008. The data is copied daily from BNL.

4 WAN, data transfer and DST production

The SINET3 (maintained by NII[13]) connects RIKEN Wako Campus to the Internet with 10 Gbps of bandwidth. Originally, two aggregated 1000BASE lines connected CCJ main switch (Catalyst 4506) and RIKEN Firewall with only 1.5 Gbps of bandwidth. For the data transfer between BNL, dedicated 10GBASE switch (Foundry FESX424) was deployed in CCJ and connected to RIKEN Main switch by a 10GBASE line in November 2007. Also for the transfer, dedicated four PC servers (HP ProLiant DL145 G3) which are operated by SL5 with a Grid environment (VDT[14] 1.8.1) were deployed. Each server has two 1000BASE network I/F, one is connected to the Foundry, and the other is connected to the Catalyst. The data from the BNL are transferred through the former line, and are written to the CCJ-HPSS through the latter line. Using the setup, a 360 MB/sec transfer rate from BNL to CCJ was achieved in the test.

In the PHENIX Run-8 (November 2007–March 2008), approximately 100 TB of p+p data were transferred using the data-transfer machines in February–March 2008. The data were sent from the PHENIX counting house in the 'semi-online' manner, namely, before they stored in RCF-HPSS, to avoid an overhead to read from the tape. As a sustained rate for a day, approximately 100 MB/s (8 TB/day) was achieved using gridFtp[15]. Another 10–20 TB were also sent after the Run period. Thus the amount of Run-8 raw data stored in CCJ-HPSS is 117 TB in total. DST production was conducted mainly in August–October 2008 with the reduced 64 nodes in RSCC and 21 TB of analyzed data (nDST) were produced out of 117 TB of raw data, and sent to RCF.

5 Outlook

In 2009, one of the main task of CCJ is the data transfer of PHENIX run9 p+p raw data, which are expected to be 200-500 TB in the Run started January 2009. The DST production using the raw data is another task. To process such large amount of data, joint operation with the renewed RSCC after August 2009 is important.

Main network switch will be replaced in July 2009 by Catalyst 4900M, which have eight 10GBASE ports. Using this switch, a 10GBASE connection between CCJ and HPSS/renewed RSCC will be established.

The condor batch queuing system[16], which is already used in RCF, should be tested in the new PC servers. When it works well, LSF will be replaced.

References

1) http://ccjsun.riken.go.jp/ccj/
2) S. Kametani et al., RIKEN Accel. Prog. Rep. 40, 197 (2007). S. Yokkaichi et al., RIKEN Accel. Prog. Rep. 41, 159 (2008).
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.platform.com/products/LSF/
9) http://www.cacti.net/
10) Veritas file system, provided by Symantec corporation.
11) http://www.xfs.org/
12) http://www.openafs.org/
13) http://www.nii.ac.jp/
14) http://vdt.cs.wisc.edu/index.html
15) http://www.globus.org/grid_software/data/gridftp.php
16) http://www.cs.wisc.edu/condor/description.html