CCJ operation in 2012

Y. Ikeda, H. En'yo, T. Ichihara, Y. Watanabe, and S. Yokkaichi

1 Overview

The RIKEN Computing Center in Japan (CCJ)[1] commenced operations in June 2000 as the largest offsite computing center for the PHENIX[2] experiment being conducted at the RHIC[3]. Since then, the CCJ has been providing numerous services as a regional computing center in Asia. We have transferred several hundred TBs of raw data files and nDSTs, which is the term for a type of summary data files at the PHENIX, from the RHIC Computing Facility (RCF)[4] to the CCJ. The transferred data are first stored in a High Performance Storage System (HPSS)[5] before starting the analysis. The CCJ maintains sufficient computing power for simulation and data analysis by operating a PC cluster running a PHENIX-compatible environment.

A joint operation with the RIKEN Integrated Cluster of Clusters (RICC)[6] was launched in July 2009. Twenty PC nodes have been assigned to us for dedicated use, sharing the PHENIX computing environment.

Many analysis and simulation projects are being carried out at the CCJ, and these projects are listed on the web page `http://ccjsun.riken.go.jp/ccj/proposals/`. As of December 2012, CCJ has been contributed 30 published papers and more than 33 doctoral theses.

2 Configuration

2.1 Calculation nodes

In our machine room 258/260 in the RIKEN main building, we have 28 PC nodes[a], and these nodes have been used for the analysis of the PHENIX nDST data. These nodes are operated by the data-oriented analysis scheme that carries out optimization using local disks[7] [8]. The OS on the calculation nodes is Scientific Linux 5.3[9], and the same OS works on the 20 nodes at the RICC. As a batch-queuing system, LSF 8.0.0[10] and Condor 7.4.2[11] were run on the CCJ and RICC nodes, respectively, as of Dec 2012.

Table 2 shows numbers of malfunctioned SATA or SAS disks in the HP servers (including NFS/AFS servers described in the next section).

2.2 Data servers

Two data servers (HP ProLiant DL180 G6 with 20 TB SATA raw disks) are used to manage the RAID of the internal hard disks, which contain the user data and nDST files of PHENIX. One old server (SUN Fire V40 with 10 TB FC-RAID) was terminated in March 2012. The disks are not NFS-mounted on the calculation nodes to prevent performance degradation by the congestion of processes and the network. These disks can be accessed only by using the "rcpx" command, which is the wrapper program of "rcp" developed at CCJ and has an adjustable limit for the number of processes on each server.

The DNS, NIS, NTP, and NFS servers are operated on the server ccjnfs20[b] with a 10-TB FC-RAID, where users' home and work spaces are located. The home and work spaces are formatted with VxFS 5.0[12]. The backup of home spaces on ccjnfs20 is saved to another disk server once a day and to HPSS once a week. The backups on HPSS are stored for 3 weeks.

Table 1. Limitation of number of job slots from LSF queue with cluster node.

	Nodes	Cores	Threads	Jobs
CCJ-hp1	18	144	144	180
CCJ-hp2	10	120	240	200
RICC	19	152	152	144
total	47	416	536	524

Table 2. Malfunctioned HDDs in 2012 and 2011

			Malfunctioned	
type	size	total	2012	2011
SATA	1 TB	192	20	9
	2 TB	120	5	4
SAS	146 GB	38	1	1
	300 GB	24	0	1

2.3 HPSS

Since Dec 2008, the HPSS servers and the tape robot have been located in our machine room, although they are owned and operated by RICC. The specifications of the hardware used can be found in the literature[13]. The amount of data and the number of files archived in the HPSS were approximately 1.7 PB and 2.1 million files, respectively, as of Dec 2012.

2.4 PHENIX software environment

Two PostgreSQL[14] server nodes are operated for the PHENIX database, whose data size was 86 GB as of Dec 2012. The data are copied from The RCF everyday and are made accessible to the users. One

[a] HP ProLiant DL180 G5 with dual Xeon E5430 (2.66 GHz, 4 cores), 16 GB memory and 10 TB local SATA data disks for each node, and HP ProLiant DL180 G6 with dual Xeon X5650 (2.66 GHz, 6 cores), 24 GB/20 TB as above, for each node

[b] SUN Enterprise M4000 with Solaris 10

Table 3. DST and raw data files in HPSS on Dec 31, 2012

	DST		Raw data	
Run	size [TB]	cos	size [TB]	cos
1	4	2,3,100	3	3,205
2	24	2,3,4,100	36	1,3,5,205
3	10	2,3,6	46	100,205
4	14	2,3	11	205
5	287	2,3,6,100	292	5,205
6	92	3,6,100	339	11,100
8	22	3	128	12
9	106	3,7	13	
10	32	3	0	
11	142	3	0	
12	3	3	0	
total	736		854	

AFS[15] server node is operated for the PHENIX AFS. The size of the libraries for the PHENIX analysis setup was 2.5TB as of Dec 2012. The libraries are also copied from the RCF by afs everyday.

2.5 Network configuration

The topology of the network linking the CCJ, the RICC, and the RIKEN IT division was not changed in 2012. This topology was shown in the paper entitled "CCJ operation in 2011"[1].

2.6 Uninterruptible power-supply system (UPS)

The power consumption of the CCJ system, excluding the HPSS, is about 25 kW, and the power is supplied through five UPSs (10.5 kVA each) as of Dec 2012, after two old UPSs were replaced by a new UPS module in March 2012. For the HPSS, there is one 7.5-kVA UPS for 100 V and three 10.5-kVA UPSs for 200 V purchased by CCJ. For the latter three, batteries were changed by RICC in Oct 2012.

3 Data transfer from BNL

Data collected during the PHENIX experiment have been transferred from the RCF to the CCJ by grid-FTP[16] through SINET4 (maintained by NII[17]) with a 10 Gbps bandwidth. In 2012, 144 TB of nDSTs of the PHENIX Run-10AuAu (15 TB), Run-11pp (10 TB), Run-11AuAu (116 TB) and Run-12pp (3 TB) were sent from the RCF to the CCJ, and the data were stored in the the HPSS, and 100 TB of this data was moved to local disks on the HP calculation nodes. Files are transferred by grid-FTP at a maximum speed of 280 MB/s.

4 Issue with RAID controller

In the period Oct 2011–May 2012, a controller of the RAID disk connected to ccjnfs20 frequently displayed the "link down" error and stopped the operation of CCJ several times. Replacement of the RAID controller chassis and I/F card did not solve the problem. There was no error recurrence after the firmware of the RAID controller v3.63G.65 was updated to v3.73k.01 in June 11, 2012.

5 Lost data

In Oct 9, 2012, a file system of a user's home directory was damaged due to disk trouble and errors in operation when the ccjnfs20 was booted after a planned power outage. The data were recovered from the backup of Oct 5 (the day before the power outage).

6 Air-conditioners of server room

In Jan 2012, six air-conditioners operated in machine room (another one had malfunctioned). Three of them broke down and the temperature increased up to 28 °C on May (The normal temperature is about 22 °C through a year). Two of broken machines were restored and the temperature returned on June. Another was removed and five have been operational as of Dec 2012.

References

1) S. Yokkaichi et al.: RIKEN Accel. Prog. Rep. **44**, 228 (2011).
2) http://www.phenix.bnl.gov/
3) http://www.bnl.gov/rhic/
4) https://www.racf.bnl.gov/
5) http://www.hpss-collaboration.org/
6) http://accc.riken.jp/ricc/
7) T. Nakamura et al.: RIKEN Accel. Prog. Rep. **43**, 167 (2010)
8) J. Phys.: Conf. Ser. **331**, 072025 (2011).
9) http://www.scientificlinux.org/
10) http://www-03.ibm.com/systems/technicalcomputing/platformcomputing/products/lsf/index.html
11) http://www.cs.wisc.edu/condor/description.html
12) Veritas file system (Symantec Corporation).
13) S. Yokkaichi et al.: RIKEN Accel. Prog. Rep. **42**, 223 (2009).
14) http://www.postgresql.org/
15) http://www.openafs.org/
16) http://www.globus.org/toolkit/docs/latest-stable/gridftp/
17) http://www.nii.ac.jp/