CCJ operation in 2011

K. Shoji, H. En'yo, T. Ichihara, Y. Ikeda, T. Nakamura*1, 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 at RHIC[3]. Since then, 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 PHENIX, from RHIC Computing Facility (RCF)[4] to CCJ. The transferred data are first stored in High Performance Storage System (HPSS)[5] before starting the analysis. CCJ maintains sufficient computing power for simulation and data analysis by operating a PC cluster running a PHENIX compatible environment.

A joint operation with 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 CCJ, and these projects are listed on the Web page `http://ccjsun.riken.go.jp/ccj/proposals/`. As of December 2011, CCJ has been contributed 29 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 18 PC nodes[a], which were installed in February 2009, and 10 new PC nodes[b], which were added in March 2011; these nodes have been used for the analysis of the PHENIX nDST using the local disks. The details of the data-oriented analysis system on the nodes are presented elsewhere[7]. Numbers of malfunctioned SATA disks in the HP servers (including NFS/AFS servers described in the next section) were 8 out of 190 1-TB disks in Jan–Dec 2011 and 4 out of 120 2-TB disks in Apr–Dec 2011.

We terminated the use of some old nodes, namely, 36 nodes of the IBM server and 18 nodes of the LinuxNetworx server, in March 2011.

The OS on the calculation nodes is Scientific Linux 5.3[8], and the same OS is run on the 20 nodes used by us at RICC. As a batch-queuing system, LSF 7.0.2[9] and Condor 7.4.2[10] were run on the CCJ and RICC nodes, respectively, as of February 2011. Upgrade to LSF 8.0.0 was performed at CCJ in March 2011.

2.2 Data servers

Two data servers (SUN Fire V40 with 10 TB FC-RAID and HP ProLiant DL180 G6 with 20 TB SATA raw disks) are used to manage the RAID disks, which contain the user data and nDST files of PHENIX. The disks are not NFS mounted on the calculation nodes to prevent the performance degradation by the congestion of processes and 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. One of the above data servers, a SUN Fire v40, was replaced in March 2012 with a new data server[c].

The DNS, NIS, NTP, and NFS servers are operated on the server ccjnfs20[d] 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[11]. 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. In Oct 2011–Mar 2012, a controller of the RAID disk connected to ccjnfs20 frequently committed the "link down" error and stopped the operation of CCJ several times. Replacement of the RAID controller and chassis did not solve the problem, and finally, the I/F card was replaced in March 2012.

2.3 HPSS

Since December 2008, the HPSS servers and the tape robot are located in our machine room, although they are owned and operated by the RIKEN IT division. The specifications of this hardware can be found in literature[12]. Version upgrade of HPSS from 7.1 to 7.3 was performed in March 2011. The amount of data and the number of files archived in the HPSS were approximately 1.6 PB and 2 million files, respectively, as of January 2012.

2.4 PHENIX software environment

Two PostgreSQL[13] server nodes are operated for the PHENIX database, whose data size was 56 GB as of January 2012. The data are copied from RCF daily and are made accessible to the users.

In July 2011, one of the two AFS[14] nodes, which copy the PHENIX software environment from RCF

*1 International Center for Elementary Particle Physics, University of Tokyo

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

b) HP ProLiant DL180 G6 with dual Xeon X5650 (2.66 GHz, 6 cores), 24 GB / 20 TB as above, for each node

c) HP ProLiant DL180 G6 with 20 TB SATA raw disks

d) SUN Enterprise M4000 with Solaris 10

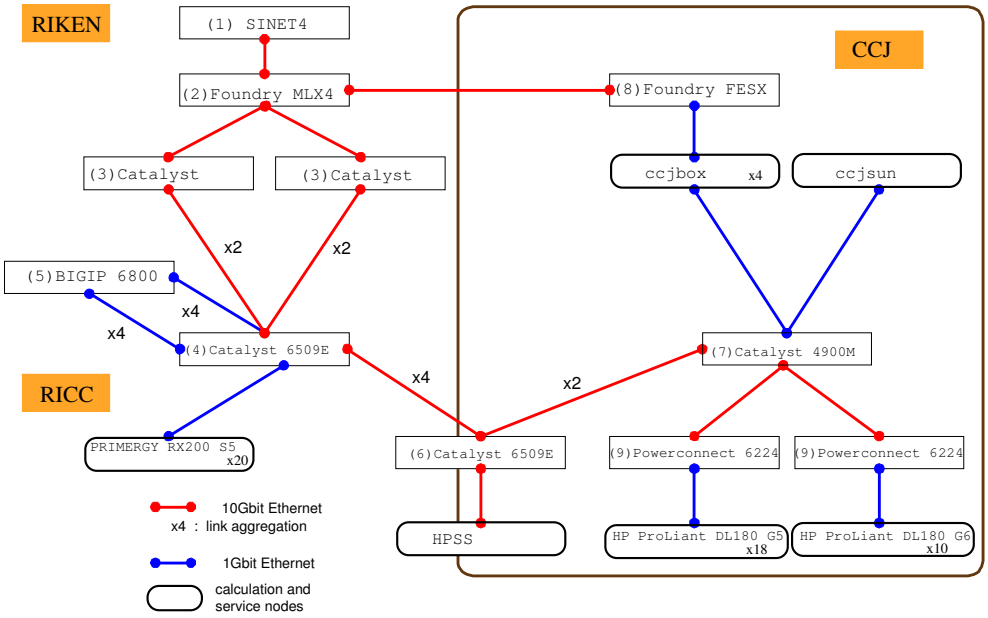

Fig. 1. Schematic view of the network configuration as of January 2012.

and make it accessible to the users, was shutdown and its function was replaced by another one.

2.5 Network configuration

The topology of the network linking CCJ, RICC, and the RIKEN IT division is shown in Fig. 1. We established a link aggregation between Catalyst 6509E and Catalyst 4900M (see 6 and 7 in Fig.1) in December 2011 in order to maintain redundancy, after network suspension in October due to the malfunctioning of the 10GB-LR optical transceiver installed in the 4900M.

2.6 Uninterruptible power-supply system (UPS)

Power consumption of the system, excluding the HPSS, is about 25 kW, and the power is supplied through five UPSs (10.5 kVA each) as of 2011. Two old UPSs were replaced by a new UPS module in March 2012.

3 The earthquake and power cut

In March 11, 2011, "the 2011 off the Pacific coast of Tohoku Earthquake" destroyed a nuclear power plant of Tokyo Electric Power Company (TEPCO). CCJ did not suffer any damage due to the earthquake itself. However, due to the power shotage caused by the disaster, CCJ operation was stopped from the night of March 13 to April 4, although no acutal power outage occured in the Wako Campus.

RIKEN decided to reduce the electric power consumption by 20% of the contracted power at the Wako campus during the summer. But CCJ could continue opperations without any restrictions due to the power-saving measures.

4 Data transfer

Data collected during PHENIX experiment have been transferred from RCF to CCJ using GridFTP[15] through SINET4 (maintained by NII[16]) with a 10 Gbps bandwidth. In 2011, 16 TB of nDSTs of the PHENIX Run-11 were sent from RCF to CCJ, and the data were stored in the HPSS, and also located on local disks on the HP calculation nodes. In 2012, we are expecting additional data transfer from PHENIX Run-11 and Run-12.

References

1) http://ccjsun.riken.jp/ccj/, S. Yokkaichi et al., RIKEN Accel. Prog. Rep. **44**, 228 (2011).
2) http://www.phenix.bnl.gov/
3) http://www.bnl.gov/rhic/
4) http://www.rhic.bnl.gov/RCF/
5) http://www.hpss-collaboration.org/
6) http://ricc.riken.jp/
7) T. Nakamura et al., RIKEN Accel. Prog. Rep. **43**, 167 (2010), J. Phys.: Conf. Ser. **331**, 072025 (2011).
8) http://www.scientificlinux.org/
9) http://platform.com/products/LSF/
10) http://www.cs.wisc.edu/condor/description.html
11) Veritas file system (Symantec Corporation).
12) S. Yokkaichi et al., RIKEN Accel. Prog. Rep. **42**, 223 (2009).
13) http://www.postgresql.org/
14) http://www.openafs.org/
15) http://www.globus.org/grid_software/data/gridftp.php
16) http://www.nii.ac.jp/