CCJ operation in 2009-2010

S. Yokkaichi, H. En'yo, Y. Goto, T. Ichihara, Y. Ikeda, T. Nakamura, K. Shoji, Y. Watanabe

1 Overview and current configuration

The CCJ[1–3] (RIKEN Computing Center in Japan for the RHIC[4] physics) started operations in June 2000 as the largest off-site computing center involved with the PHENIX[5] experiment at RHIC. CCJ was initially planned to perform three roles in the PHENIX computing, 1) as a simulation center, 2) as the Asian regional center, and 3) as a center for the study of spin physics. Around 2005, 4) 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 is currently capable of handling several hundred TBs of raw data in use of HPSS (High Performance Storage System)[6]. In 2005, 2006 and 2008, 2-300 TB of raw-data files were sent from RCF (RHIC Computing Facility)[7]to CCJ and analyzed. Recently, the CPU power of RCF has been increased and DST production can be performed at RCF without any help. Thus, we sent only nDST, which is the name of a summary data file at PHENIX, from RCF to CCJ in 2009 and 2010.

A joint operation with RSCC (RIKEN Super Combined Cluster System) was started in March 2004 and completed in June 2009. In July 2009, RICC (RIKEN Integrated Cluster of Clusters)[8] was launched, and the joint operation was continued. Twenty PC nodes have been assigned to us for dedicated use, and the PHENIX computing environment is being shared.

Many analysis and simulation projects are being carried out at CCJ. They are mentioned on the Web page `http://ccjsun.riken.go.jp/ccj/proposals/`. As of June 2010, CCJ has been contributed 23 published papers and more than 30 doctoral theses.

1.1 Calculation nodes

We have 18 PC nodes (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) for the data-oriented analysis nodes; these were installed in February 2009[2] and have been used for the analysis of the PHENIX nDST using the local disks. The details of the data-oriented analysis system are presented in some reports[2,9]. Some old nodes, *i.e.*, 36 nodes of IBM server (with 10 TB local SCSI data disk) and 18 nodes of LinuxNetworx server (with no local data disk), were also present in our machine room 258/260 in the RIKEN main building. New data-oriented analysis nodes (HP ProLiant DL180 G6 with dual Xeon X5650 (2.66GHz 6 cores), 24 GB memory and 20 TB local SATA data disks for each node) have been delivered in March 2011 and replaced the 54 old nodes.

The OS upgrade from SL (Scientific Linux)[10] 4.4 to SL 5.3 was performed in April 2010 for the calculation nodes at CCJ, and the same upgrade was performed in May 2010 for the 20 nodes used by us at RICC. After the upgrade, VMWare ESXi is no longer used on the RICC nodes[2] and the OS is running natively. As a batch-queuing system, LSF 7.0.2[11] and Condor 7.4.2[12] were operated in CCJ and RICC nodes, respectively, as of February 2011. Upgrade to LSF 8.0.0 was performed at CCJ in March 2011. Two old LSF server nodes were discarded.

1.2 Data servers

Five data servers (SUN Fire V40) were used to manage the RAID disks, which contained the user data and nDST files of PHENIX. The disks were not NFS-mounted from 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.

The server ccjnfs11 (with 6.8 TB FC-RAID and 8.9 TB-SATA RAID) was discarded in March 2010. Further, three data servers ccjnfs12 (with 10 TB FC-RAID) and ccjnfs14/15 (with 18 TB SATA-RAID for each) were discarded in March 2011. After the discarding, we have a data server ccjnfs13 (with 10 TB FC-RAID). A new data server ccjnfs16 (HP ProLiant DL180 G6 with 20 TB SATA raw disks) is undergoing tests to replace ccjnfs12.

The DNS, NIS, NTP, and NFS servers are operated on the server ccjnfs20 (SUN Enterprise M4000 with Solaris 10) with a 10 TB FC-RAID where users home and work regions are located. In October 2010, the NFS-write operation from the calculation nodes to the work region was prohibited in order to prevent the performance degradation in the interactive usage of the home region due to the heavy writing operation by user jobs. The home and work regions are formatted VxFS 5.0[13], which has a bug in the quota system. When a large file whose size is more than 2GB is deleted, the quota count is not decreased; thus, the count gets piled up and reaches the limit without the user using up all the allocated space. A superuser should be called to resolve the pile-up. The version upgrade of VxFS will solve this problem.

1.3 HPSS

Since December 2008, the HPSS servers and the tape robot are located in our machine room, while they are owned and operated by the RIKEN IT division. The

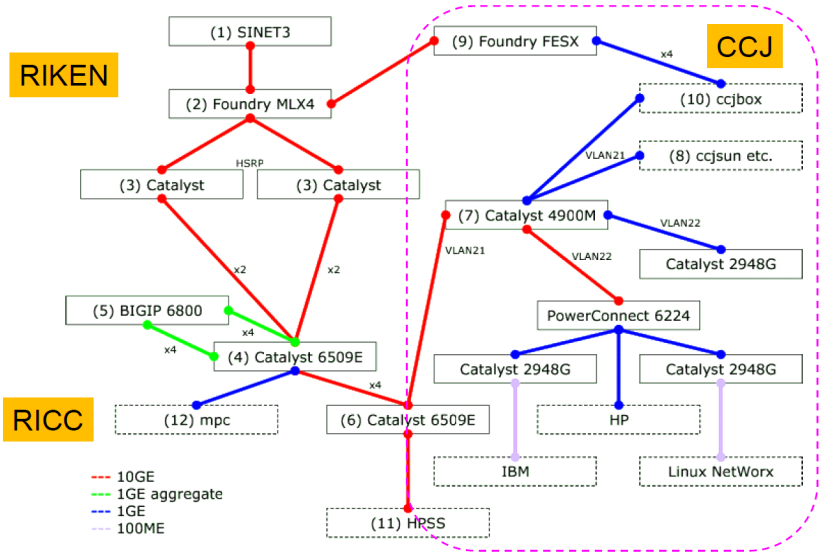

Fig. 1. Schematic view of the network configuration as of June 2010.

specifications of this hardware can be obtained in the literature[3]. The version upgrade of HPSS from version 6.2 to 7.1 was performed in January 2010 and upgrade to 7.3 was performed in March 2011. The amount of data and the number of files archived in the HPSS were approximately 1.5 PB and 2 million files, respectively, as of February 1, 2011.

1.4 PHENIX software environment

Two PostgreSQL[14] server nodes are operated for the PHENIX database, whose data size was 64 GB as of February 2011. The data is copied from RCF daily and is served to the users.

In November 2010, two AFS[15] client nodes operated to copy the PHENIX software environment were combined to one node. Recently, the automatic token-feed system is showing signs of a glitch. It should be fixed or the token should be fed manually and weekly.

1.5 Network configuration

The configuration of the network linking CCJ, RICC and RIKEN IT division has been detailed previously[2]. The topology is shown in in Fig. 1.

2 A big earthquake and a power cut

In March 11, 2011, 'the 2011 off the Pacific coast of Tohoku Earthquake' destroyed a nuclear power plant of TEPCO (Tokyo Electric Power Company). Due to the power limitation caused by the disaster, TEPCO announced to perform the scheduled daily power cut (2-3 hours in a day) from March 14 to April. CCJ was shutdown emergently in the night of March 13, while no actual power outage occurred at RIKEN Wako campus. For salvage of user data and the LSF upgrade, a few servers were temporarily operated in March 18-19 and 28-29. Finally, TEPCO and RIKEN announced in March 29 that the power outage was ended at Wako. Thus, the recovery of CCJ was started and all the system was opened for users in April 4. The recovery of DB and AFS including copying the data took the six days.

CCJ was caused no damage by the earthquake itself.

3 Outlook

In 2011, hundreds TB of nDST from the PHENIX Run-11 will be sent from RCF to CCJ and will be stored on the local disks on the HP calculation nodes. Two old UPS modules, one old data server, and one old login server should be replaced in JFY 2011.

References

1) http://ccjsun.riken.go.jp/ccj/
2) T. Nakamura et al., RIKEN Accel. Prog. Rep. **43**, 167 (2010).
3) S. Yokkaichi et al., RIKEN Accel. Prog. Rep. **42**, 223 (2009).
4) http://www.bnl.gov/rhic/
5) http://www.phenix.bnl.gov/
6) http://www.hpss-collaboration.org/
7) http://www.rhic.bnl.gov/RCF/
8) http://ricc.riken.jp/
9) T. Nakamura et al., Journal of Physics: Conference Series, in press (2011).
10) https://www.scientificlinux.org/
11) http://www.platform.com/products/LSF/
12) http://www.cs.wisc.edu/condor/description.html
13) Veritas file system, provided by Symantec Corporation.
14) http://www.postgresql.org/
15) http://www.openafs.org/