RIKEN Accel. Prog. Rep. **37** (2004)

CCJ Operation in 2002-2003

S. Yokkaichi, Y. Goto, H. Hamagaki,[*1] T. Ichihara, O. Jinnouchi, A. Kiyomichi, Y. Watanabe and H. En'yo

CCJ[1], RIKEN Computing Center in Japan for RHIC[2] physics, started operation from June 2000 as the largest off-site computing center for the PHENIX[3] experiment at RHIC. CCJ fills, in PHENIX computing, the three roles of 1) the simulation center, 2) Asian regional center and 3) center of spin physics. Recently DST (Data Summary Tape) production from raw data comes more important than simulation for CCJ role. Other smaller computing centers in PHENIX can handle only relatively small simulation data, however, only CCJ can handle several tens of TB of raw data except RCF (RHIC computing facility)[4] because of our HPSS (High Performance Storage System)[5] and data duplication facility built by us at RCF to transfer the data to CCJ.

Many analysis and simulation projects are being carried out at CCJ, including some PHENIX-official projects. They are shown on the web page: `http://ccjsun.riken.go.jp/ccj/proposals/`. Within the current year, the data from PHENIX Run2 (2001/8 ~ 2002/1) and Run3 (2002/11 ~ 2003/5) were mainly analyzed. The report from each project is described in this volume.

About 50 TB of data were transported from BNL via tape media (42 TB) and network (9.1 TB) within this year. In a typical shipment of tapes, 17 TB of data are transferred in 33 days which includes tape transport by airplane, uploading to CCJ-HPSS and all limited factor of troubles. It means 6 MB/s of transfer rate on average. Via network, the typical transfer rate from BNL is 3~4 MB/s and approximately 12 MB/s was observed as the maximum burst rate retained in 20 minutes. In such large-scale network transfer, multi-stream transfer was mainly used with the *bbftp*[6] command. The data amounts recorded at PHENIX and sent to CCJ are summarized in Table 1. Typical file sizes of the data are also shown.

	Run1 HI	Run2 HI	Run2 p+p	Run3 HI	Run3 p+p
raw data					
recorded (TB)	2.3	20	26	49	39
(transferred to CCJ)	2.3	7	26	-[a]	39
typ. file size(GB)	1.5	1.5	1.5	1.5	1.5
DST(version)	v05	v03	v01/2	v03	v03
produced (TB)	2.6	17.5	-	12.4	4.3
(transferred to CCJ)	2.6	-[a]	8.2	-[a]	-
typ. file size(GB)	0.3-0.6	1.8	0.6	0.5	0.2
μ/nDST (version)	v05	v03	v01/2		
produced (TB)		-	-	-	-
(transferred to CCJ)		4	6.6	-	-
typ. file size(GB)	0.1-0.2	0.3	0.1	-	-

Table 1. PHENIX data amount as of Dec. 2003

The DST hierarchy in PHENIX is as follows. In each stage, the data size is reduced for easier handling. An event reconstruction process writes the DST in 'root'[7] format from the raw data. While the typical file size of raw data is 1.5 GB, DSTs are reduced to 0.6~1.8 GB. The μDST is written in the 'root tree' format and some information included in DST is dropped. Typical μDST size is 0.1~0.3 GB. The nanoDST(nDST)s are subsets of μDST which has only limited events and/or information which is selected with special interest for each group of users. The nDSTs have a size of 10~80 MB typically. In principle, we transfer all DST, μDST and nDST, etc., from RCF to CCJ. For the raw data, all the p+p data for spin physics and a certain fraction of Heavy Ion (HI) data are also transferred. This principle was provisionally confirmed at the CCJ users meeting in Sep. 2003. While raw data and DST are transferred via tape solely by the CCJ administrator, μ/nDST are transferred via network with significant cooperation from users.

The current configuration of CCJ is shown in Fig. 1. We have 176 PC nodes operated using Linux, 166 are calculation nodes and 10 are service nodes. Each calculation node has 1 GB of memory, 2 GB of swap area, 10~31 GB of local work area and dual CPU (Pentium III 700 MHz ~ 1.4 GHz, Pentium 4 2.0 GHz). Red Hat 7.2/kernel 2.4.18 is operated on the calculation nodes, the same as RCF. Upgrading to Red Hat 8 is planned in Jan. 2004. We also have seven SUN servers operated using Solaris, two are service nodes and five are data servers, which are connected to large RAID systems and operated as NFS servers. Service nodes are used for various purposes, for example, login server from WAN, database server for analysis, AFS client to share the PHENIX computing environment, controller of PC cluster, and monitoring of PC nodes and network.

HPSS is used as a mass storage system in CCJ, the same as RCF. Approximately 150 TB/450000 files of data are stored in CCJ-HPSS as of Oct. 2003. Five IBM RS6000/SP servers operated using AIX are used as HPSS core server and data/tape movers.

On the calculation nodes, LSF[8] 4.2 is operated as a batch queueing system. Three general queues, *short/long/bg*, are categorized by CPU time limit, 200 minutes/1 day/7 days, respectively. Special queues for *production* (high priority), *largedisk* (dispatched to the nodes on which large disk/memory space remain) etc. are also set temporally or perpetually for special purposes. The priority of job dispatching in the queues is not 'first-in first-served' but is controlled by the fair-

[*1] CNS, University of Tokyo

[a] transfer is on going

share policy of LSF, taking account of the CPU usage history by each user. Two job slots are assigned to each node because they have dual CPU. In the system, 20∼30 job slots are kept for the *short* queue without being filled by longer jobs to maintain a shorter turnaround time for small-scale jobs. The number of running jobs in each queue is monitored using MRTG[9] in addition to data flow between PCs and data servers or HPSS thus the state of the system is always watched by WEB.

The jobs submitted to CCJ are roughly categorized into two types, CPU bound and I/O bound. Typical CPU-bound jobs are simulation jobs (detector simulation and event generation) and event reconstruction (DST production) jobs. Typical I/O bound jobs are data reduction jobs, i.e., making μDST (or nDST) from DST (μDST) and analysis of μ/nDST, i.e., reading the data and making small histogram files. In CCJ, such I/O bound jobs are limited by the I/O bandwidth of HPSS and/or RAID disks in which the required data are located. Particularly, concurrent NFS access from many PC nodes (typically more than 20) degrades the I/O throughput. In order to avoid such a situation, we use the *rcp* command to access the RAID disk. Jobs running on a PC have to transfer the data between the RAID and PC-local disk by rcp, and only work on the local disk. Actually, many concurrent rcp's also degrade the throughput, hence we limit the number of maximum rcp processes on each data server. By the use of the limited rcp, I/O of RAID is maintained at 30∼50 MB/s in contrast with NFS access which can achieve only 10 MB/s.

Hardware upgrades in this year are as follows. No CPUs were upgraded but several tens of mother-boards were replaced to get rid of the node hanging-up problem. FC RAID disks of 8 TB were newly connected to SUN Fire 880 and the total capacity was increased to approximately 33 TB. In HPSS, tape drives were replaced from four STK 9940A (I/O 10 MB/s, capacity 60 MB/cartridge) to six 9940B (30 MB/s, 200 MB/cartridge). All the data written in the 9940A format have already been rewritten in the 9940B format. Since we have 3000 tape cartridges, the total tape capacity is 600 TB currently. In the data duplication facility at RCF, two Redwood drives were replaced by two 9940B's and the host machine was also replaced, from IBM RS6000 F50 to IBM p630. The connection between RIKEN LAN to WAN was also replaced in June 2003, from IMnet (50 Mbps) to SINET (1 Gbps).

We have encountered the hardware problems described below this year. PC nodes encounter various hardware problems including problems of the fan, disk (itself and SCSI I/F board), memory, CPU and power supply. Approximately 15% of PC nodes (almost belong to older node group) suffered from such problems. Data servers also encounter hardware problems in the RAID controller, I/F PCI card between FC (fibre channel) RAID and server, FC GBIC and FC hub. To solve them, we needed to shutdown at least one data server per month on average. The simple break down of disks in the RAID system is not counted here because such disks are hot-swappable, i.e., they can be replaced without stoppage of service. In spite of these problems, data on the disks have not been lost thus far.

We have proceeded now to a project of integrated operation between CCJ and the new RIKEN supercomputer system. The new system includes 1024 nodes/2048 CPUs of a PC cluster system and we can use dedicated 128 nodes/256 CPUs in the cluster. Our data servers and HPSS can also be accessed by the new cluster nodes the same as for our old PC nodes. HPSS server machines will be replaced by seven IBM p630 servers and we can use eight 9940B tape drives connected to four data/tape movers and one tape robot (STK Powderhorn 9310) dedicatedly. The robot can handle approximately 5000 tape cartridges, thus we can extend the total tape capacity to 1 PB. The integrated system will be available in March 2004.

References

1) `http://ccjsun.riken.go.jp/ccj/`
   Y. Watanabe et al.: RIKEN Accel. Prog. Rep. 36, 262
   T. Ichihara et al.: RIKEN Accel. Prog. Rep. 35, 236
2) `http://www.bnl.gov/rhic`
3) `http://www.phenix.bnl.gov`
4) `http://www.rhic.bnl.gov/RCF/`
5) `http://www.sdsc.edu/hpss/`
6) `http://doc.in2p3.fr/bbftp/`
7) `http://root.cern.ch`
8) `http://www.platform.com/products/LSF/`
9) `http://people.ee.ethz.ch/~oetiker/webtools/mrtg/`

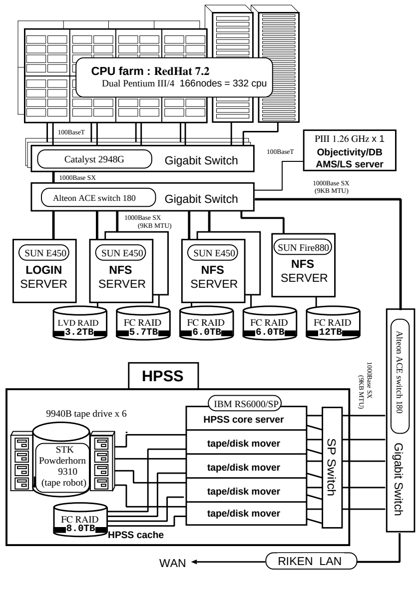

Fig. 1. Current configuration of CCJ