Computing Center in Japan for RHIC Physics (CC-J)

Takashi Ichihara, Yasushi Watanabe, Naoki Hayashi, Satoshi Yokkaichi, Shinya Sawada[*1], Yuji Goto, Naohito Saito, Hideto En'yo[*2], Hideki Hamagaki[*3] and Masayasu Ishihara

The computing center in Japan (CC-J) for RHIC physics [1] started operation in June 2000 at RIKEN. The CC-J is aimed to be the principle site of computing for PHENIX[2] simulation, a regional PHENIX Asian computing center, and a center for the analysis of RHIC[3] spin physics. Since the PHENIX experiment at RHIC was scheduled to commence in 2000, we started to construct the regional computing center in Japan at the RIKEN Wako campus in April 1999[4]. Prior to the construction, R&D for the CC-J started in 1998 at the RIKEN BNL Research Center (RBRC).

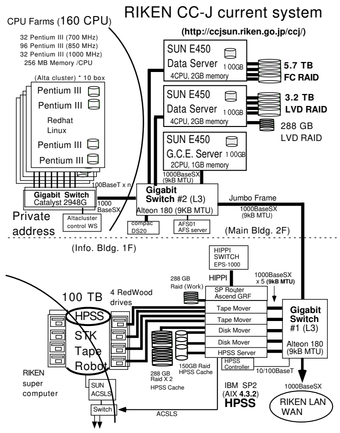
The first PHENIX experiment at RHIC for Au + Au collision was carried out in the summer of 2000. The raw data of approximately 1 TB was transferred from the RHIC Computing Facility (RCF) at the BNL to the CC-J over the Wide Area Network (WAN)[5]. The analyses for the first PHENIX experiment have been carried out successfully at the CC-J. The Ring Imaging Cherenkov (RICH) detector demonstrated the good identifications of the electron[5] . The calibration of the Time-of-Flight (TOF) wall was carried out at the CC-J and revealed clear particle identifications for hadrons[6]. A clear π^0 peak was extracted by the calibration of the Electromagnetic Calorimeter (EMCal) and the background was evaluated by mixed event analysis[7].

The simulation of 100K events for the PHENIX experiment under several conditions was carried out successfully using the PISA code at the CC-J[8]. A fast Monte Carlo code was developed and simulation using the code was carried out at the CC-J to evaluate the acceptance of the EMCal for π^0.[7]

The following table shows the size of the CC-J system. The CC-J is planned to reach its full scale in the three years. At the opening of the CC-J (June 2000), the size of the CC-J was about one-third of the goal, although almost all the planned functions of the CC-J were completed at that time. About 40 user accounts were created.

	1999	2000	2001
CPU (SPECint95)	2400	5900	13000
Tape Storage (TB)	100	100	100
Disk Storage (TB)	2.0	9	15
Tape I/O (MB/s)	45	45	112

Figure 1 shows the configuration of the current CC-J system. The main components of the CC-J consist of (1) a High-Performance Storage System (HPSS), (2)



vspace8.0 cm

Fig. 1. Configuration of the CC-J system.

Data servers and (3) CPU farms. Since the CC-J handles approximately 200 TB of data for year, the HPSS was installed as a hierarchical storage system[9].

Three SUN Enterprise 450 (E450) servers have been installed. One E450 server (ccjsun) is used for login, AFS client and general computing. AFS is a worldwide distributed file system. External users are required to log in to the server using Secure Shell (SSH) to access the CC-J. Other two E450 servers are used for dedicated data servers. A 3.2 TB SCSI Raid5 disk and a 5.7 TB Fiber Channel (FC) Raid5 disk are connected to the servers. A recent benchmark test for 2 GB file access using Bonnie code revealed that the read access of the FC Raid disk (50 MB/s) is approximately four times faster than that of LVD SCSI Raid disks (12 MB/s). The Veritas File System (VxFS), which is a type of logging file system, were adopted to obtain high performance in file operation and fast file system recovery (fsck) after system failure.

These SUN E450 servers, HPSS servers and switching hubs for the Linux CPU farms are connected by the Gigabit Ethernet to ensure fast file transfer and eleminate network traffic jams.

The total required CPU performance for CC-J is about 12 k SPECint95 (120 k SPECint2000). The main part of this is for simulation. We adopted the Linux CPU farms to achieve this requirement. The current CPU farms consists of Linux pc's with 160 units of Intel Pentium III processors at 700-1000 MHz. We adopted 10 boxes of Altacluster as a frame. Figure 2 shows a photograph of the current CPU farms. Each node is equipped with a dual CPU, 512 MB of ECC memory, and a 14-18 GB local disk.

*1 KEK
*2 Kyoto University
*2 CNS, University of Tokyo

Fig. 2. Linux CPU farms

We have developed the *kick-start* installation method for Redhat Linux 6.1 to the farms. The installation of Linux for each node can be carried out with a floppy disk and a common NFS server. It takes approximately 10 minutes to install one node and installation can be carried out simultaneously for several nodes. There is no need to use the keyboard or mouse during the installation. The post-installation procedure is carried out automatically following the kick-start installation.

We are currently using Linux kernel 2.2.14 and 2.2.16. The NFSv3 patch for the Linux kernel significantly improved the NFS performance, particularly for the NFS write operation. During job execution, most of the disk IOs are carried out at the local disk equipped for each node. At the beginning and ending of a job, file transfer will be carried out between the Linux local disk and central data server.

To keep the common software environment to be RCF, AFS is used. To access AFS from the Linux farms stably, *mirroring* of the AFS contents to a local data server is carried out daily using the *rsync* utility. AFS contents are accessed from the Linux farms through the NFS; thereby eliminating instabilities from the usage of AFS on the Linux platforms. The mirrored contents of AFS at the CC-J will be mirrored again at the domestic collaborating institutes using the rsync utility for fast update.

Since the PHENIX collaboration uses C++ for its common software, Objectivity/DB was adopted. Therefore, it is necessary to access the PHENIX Objectivity/DB at the RCF from the CC-J and this method has been studied previously[10].

The portable batch system (PBS), which is free software, was used for the batch queueing in the early stage of the development. Job scheduler software was developed for the PBS. However, since we encountered freezing problems and instability in PBS, we changed the batch queueing system from PBS to the Load Sharing Facility (LSF) in the summer of 2000. LSF is a commercial product. LSF ensures the stable operation of the batch queueing and offers the convenient features of *fairshare* job scheduling and automatic job recovery after node failure.

RIKEN is currently connected to Inter-Ministry research information network (IMnet) at the speed of 15 Mbps and this bandwidth will be updated to 50 Mbps in early 2001. For the international connection, the CC-J joined the Asia-Pacific Advanced Network (APAN) which has a bandwidth of 70 Mbps over the Pacific Ocean. Thanks to the APAN, stable access to the USA and Europe Energy Science Network (ESnet) sites has been achieved. Since the capability for the huge data transfer (200 TB/year) between the CC-J and the RCF is required, a data duplication facility using SD-3 tape cartridges was constructed at the RCF and started operation in late 2000[9]

The CC-J Planning and Coordination Office (PCO) was established. Operation, maintenance and development of the CC-J are carried out under the charge of the PCO. Research plans using the CC-J should be submitted to the CC-J PCO.

The authors are grateful to Professor Bruce Gibbard and the staff of the RCF for their earnest cooperation and discussions. They are also indebted to Professor William A. Zajc, Professor Barbara Jacak and Dr. Dave Morrison for the coordination of the CC-J in the PHENIX collaboration.

References

1) http://ccjsun.riken.go.jp/ccj/
2) http://www.rhic.bnl.gov/phenix/
3) http:/www.rhic.bnl.gov/
4) T. Ichihara et al.: RIKEN Accel. Prog. Report. **33**, 172,(2000).
5) Y. Akiba et al.: in this volume of RIKEN Accel. Prog. Report.
6) T. Chujo et al.: in this volume of RIKEN Accel. Prog. Report.
7) T. Sakaguchi et al.: in this volume of RIKEN Accel. Prog. Report, and *private communications*.
8) N. Hayashi et al.: in this volume of RIKEN Accel. Prog. Report.
9) Y. Watanabe et al.: in this volume of RIKEN Accel. Prog. Report.
10) S. Yokkaichi et al.: in this volume of RIKEN Accel. Prog. Report.