PHENIX Computing Center in Japan

T. Ichihara[1,2], Y. Watanabe[1,2], N. Hayashi[1], S. Sawada[3], S. Yokkaichi[4], A. Taketani[1], Y. Goto[2], H. En'yo[3,2], H. Hamagaki[5]

[1] RIKEN (The Institute of Physical and Chemical Research), Japan,
[2] RIKEN BNL Research Center, USA
[3] High Energy Accelerator Research Organization (KEK), Japan
[4] Department of Physics, Kyoto University, Japan
[5] CNS, University of Tokyo, Japan

Abstract

PHENIX Computing Center in Japan (PHENIX CC-J) is now under construction at the RIKEN Wako campus over a three year period, beginning in April 1999. CC-J is intended as the principle site of computing for PHENIX simulation, a regional PHENIX Asian computing center, and as a center for the analysis of RHIC spin physics. The CC-J will handle the data of about 200 TB/year. The total CPU performance of the CC-J is planned to be 10,000 SPECint95.

Keywords: Gigabit ethernet, Jumbo frame, Linux CPU farm, NFSv3 kernel patch, PBS, HPSS, AFS, RHIC, PHENIX

1 Introduction

The PHENIX [1] experiment at RHIC [2] is scheduled to start early in 2000 at BNL. The spin physics experiment of polarized $p-p$ collision at $\sqrt{s} = 500$ GeV is planned to start in 2001. To produce the physics output promptly, we have been preparing to construct the PHENIX Computing Center in Japan (CC-J)[3]. The purposes of the CC-J are focused to the following three points:

- Detector simulation and theoretical model calculations: These very CPU intensive tasks are invaluable for the extraction and understanding of results derived from PHENIX[1] measurements. The CC-J is intended to handle the bulk of the simulation tasks of PHENIX.
- PHENIX regional computing center in Asia: The CC-J is intended to be a regional center for simulation and analysis of PHENIX physics results. Having regional access to computing and physics data will encourage strong participation from the PHENIX collaborators of China, Korea, India and Japan. The CC-J will emphasize micro Data Summary Tape (DST) production and later physics analysis.
- SPIN physics analysis: Although the workforce for carrying out the SPIN physics program at RHIC PHENIX is centered at the RIKEN BNL Research Center (RBRC), it is very important to keep maintain a strong locus of activity in Japan.

R&D for the CC-J started in 1998 at RBRC and a prototype of the CPU farms and data duplication method were studied. The construction of the CC-J started at the RIKEN Wako campus over a three year period, beginning in April 1999.

2 Overview of the CC-J System

Figure 1 shows the concept of the CC-J system. During the RHIC experiment, data acquisition, raw data recording and monitoring will be performed at the RHIC Computing

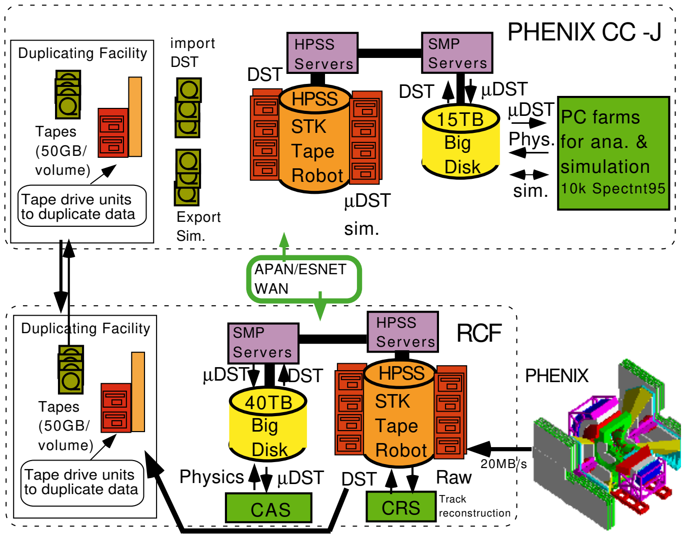

Figure 1: Connept of the CC-J system

Facility (RCF) located at the BNL. After the track reconstruction, the DST will be produced, duplicated and exported to the CC-J. Data mining will be carried out at CC-J as well as RCF from the DST to produce the micro-DST, filtered by the physics motivation. Simulation for the PHENIX experiments will be performed at CC-J and this output will be exported to RCF mainly via SD3 tape cartridge. The following table shows the current and planned size of the CC-J.

	Jan. 2000	Mar. 2001	Mar. 2002
CPU (SPECint95)	1500	5900	10700
Tape Storage (TB)	100	100	100
Disk Storage (TB)	2	10	15
Tape I/O (MB/s)	45	90	112

3 Description of the CC-J System

Figure 2 shows the configuration of the current CC-J system. While the CC-J essentially follows the architecture to the RCF, there are several changes and improvements.

3.1 Data Storage

The following table shows the estimate of the annual data amount for PHENIX experiment at the nominal year and that for CC-J. The Raw data from the PHENIX detector comes at 20MB/s and this yields 290 TB for each year.

	PHENIX	CC-J
Raw Data	290	0
Calibrated Data	0.1	0.1
Simulated Data	30	30
Data Summary Tape	150	150
micro-Data Summary Tape	45	45
Total	515TB	225TB

To handle such an amount of data, High Performance Storage System (HPSS) [4] was installed early in 1999 at RIKEN Wako campus. The HPSS consists of five nodes of RS/6000-SP server (silver node), a cache disk system of 720 GB, four RedWood SD3 tape

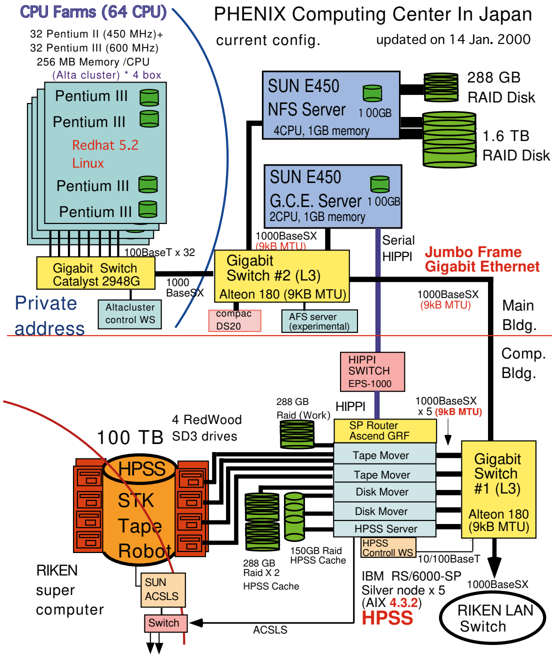

Figure 2: Current configuration of the CCJ System

drives, and a tape robotic storage of 100 TB (half of STK PowderHorn). Each SP node is connected with SP Switch Router and Gigabit Switch. About 50 MB/s transfer rate has been obtained by parallel ftp (pftp) between HPSS clients and HPSS servers over the Gigabit Ethernet with jumbo frame. A dedicated Network File Server of SUN E450 serves the work disk of 1.6 TB and user home directory (290 GB) etc. The preliminary test of exchanging data between CC-J and RCF using the SD3 tape cartridge has been carried out successfully .

3.2 CPU farms

Following table shows the estimate of the CPU requirement for the PHENIX. A large portion of the CPU requirement of CC-J is occupied by the simulation. This estimation shows the CPU requirement of the CC-J is about 11470 SPECint95.

	PHENIX	CC-J
Event Reconstruction	6084	0
Data Analysis	1700	1000
Theoretical Model	800	800
Simulation	7991	8200
Simulation Event Reconst.	1300	1300
Data Analysis of Sim.	170	170
Total	18045	11470

Linux pc's with Intel Pentium II/III are used for the CPU farms. Each node is equipped with dual CPU, 512 MB memory and 14-18 GB local disk. The NFSv3 patch for the Linux latest kernel significantly improved the NFS performance, especially for the NFS write operation. During the simulation production, most of the disk IO are carried out at the local disk. The final result will be transfered to HPSS server using pftp. Hardware-reset and power-control for each Linux pc can be performed via the network through the control pc thus enabling us to maintain the pc farm easily.

To keep the common software environment to the RCF, AFS is used (cell name rhic). To access AFS from Linux stably, the mirroring of the AFS contents to a local disk is carried out daily using *rsync* utility. AFS contents are accessed from Linux pc through the NFS; thereby eliminating instabilities from the usage of AFS on the Linux platforms.

Portable Batch System (PBS) V2.2 are used for the batch queueing. The Multi Router Traffic Grapher (MRTG) is used for monitoring the network traffic, CPU loads and free memory etc.

3.3 WAN environment

RIKEN is connected to BNL via IMnet in Japan at 12Mbps, APAN[5] for Japan-US link, startap and ESnet in USA. Currently APAN has a bandwidth of 73 Mbps but Round Trip Time (RTT) between RIKEN and BNL is about 170 m sec, mainly due to the length of the optical fiber over the pacific ocean. Only 41 KB/s ftp performance between RIKEN and BNL has been obtained using the default TCP parameter. We have measured the ftp transfer rate between RIKEN and BNL for various TCP window-sizes and they are summarized in the following table.

TCP window size	FTP transfer rate (observed)	Theor. limit for 170 ms RTT
8 KB (default)	41 KB/s	47 KB/s
16 KB	87 KB/s	94 KB/s
32 KB	163 KB/s	188 KB/s
64 KB	288 KB/s	376 KB/s
128 KB	458 KB/s	752 KB/s
256 KB	585 KB/s	1500 KB/s
512 KB	641 KB/s	3010 KB/s

It should be noted that a good ftp performance of about 5 Mbps has been obtained over the pacific ocean for large TCP window-size parameter. We expect a certain part of simulated data and DST will be transfered via WAN. Replication of the Objectivity/DB via WAN is going to be tested.

4 Schedule

Stress tests for the entire system have been carried out successfully in May, July 1999 and January 2000. During the test, the physics simulation output have been by-produced. We plan to start the test operation of the CC-J in February 2000 and start the CC-J phase-1 operation in the spring of 2000.

The authors are grateful to Prof. Bruce Gibbard and the staff of the RCF for their earnest cooperation and discussions. They are also indebted to Prof. William A. Zajc, Prof. Barbara Jacak and Dr. Dave Morrison for the coordination of the CC-J in the PHENIX collaboration. Sincere thanks are also expressed to Prof. Ryugo S. Hayano, Prof. Junsei Chiba, and Dr. Naohito Saito for their earnest discussions.

References

1 http://www.rhic.bnl.gov/phenix/
2 http://www.rhic.bnl.gov/
3 http://ccjsun.riken.go.jp/ccj/
4 http://www.sdsc.edu/hpss/
5 http://www.apan.net/