IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Data oriented job submission scheme for the PHENIX user analysis in CCJ

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

2011 J. Phys.: Conf. Ser. 331 072025

(http://iopscience.iop.org/1742-6596/331/7/072025)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 134.160.38.17
The article was downloaded on 24/01/2012 at 10:00

Please note that terms and conditions apply.

Journal of Physics: Conference Series **331** (2011) 072025 doi:10.1088/1742-6596/331/7/072025

Data oriented job submission scheme for the PHENIX user analysis in CCJ

T. Nakamura, H. En'yo, T. Ichihara, Y. Watanabe and S. Yokkaichi
RIKEN Nishina Center for Accelerator-Based Science, Wako, Saitama 351-0198, Japan

Abstract. The RIKEN Computing Center in Japan (CCJ) has been developed to make it possible analyzing huge amount of data corrected by the PHENIX experiment at RHIC. The corrected raw data or reconstructed data are transferred via SINET3 with 10 Gbps bandwidth from Brookheaven National Laboratory (BNL) by using GridFTP. The transferred data are once stored in the hierarchical storage management system (HPSS) prior to the user analysis. Since the size of data grows steadily year by year, concentrations of the access request to data servers become one of the serious bottlenecks. To eliminate this I/O bound problem, 18 calculating nodes with total 180 TB local disks were introduced to store the data a priori. We added some setup in a batch job scheduler (LSF) so that user can specify the requiring data already distributed to the local disks. The locations of data are automatically obtained from a database, and jobs are dispatched to the appropriate node which has the required data. To avoid the multiple access to a local disk from several jobs in a node, techniques of lock file and access control list are employed. As a result, each job can handle a local disk exclusively. Indeed, the total throughput was improved drastically as compared to the preexisting nodes in CCJ, and users can analyze about 150 TB data within 9 hours. We report this successful job submission scheme and the feature of the PC cluster.

1. Introduction

The RIKEN Computing Center in Japan (CCJ) [1] started the operation in time for the beginning of data taking at the PHENIX experiment [2] in 2000. The PHENIX experiment at Relativistic Heavy Ion Collider (RHIC) [3] was designed for two type of physics projects. One of the projects is search for the Quark-gluon plasma state [4] by high energy heavy-ion collisions. Another is to explore the spin structure in proton [5] by using the polarized proton beam. Although CCJ was originally established to provide sufficient computing resources mainly for the spin physics at PHENIX, the resources have been used for the other projects related to Radiation Laboratory [6] in RIKEN Nishina Center [7], for instance, experiments at KEK-PS [8] and the new generation experiments at J-PARC [9]. Thus far CCJ has been providing numerous services as a regional computing center in Japan.

2. System configurations

CCJ has played several important roles in the PHENIX experiment: data reconstruction for the polarized $p+p$ collisions, detector calibrations, simulations and user analysis for all type of data. The collected raw data had been directly transferred from the PHENIX counting room at Brookheaven National Laboratory (BNL) via SINET3 with a 10 Gbps bandwidth maintained by NII [10] by using GridFTP [11]. We receive those data by the four main RAID boxes

Published under licence by IOP Publishing Ltd

parallelly with switching 2 TB buffer area in each RAID boxes. The transferred data are once stored in High Performance Storage System (HPSS) [12] from the RAID boxes before starting the reconstruction process and user analysis. This HPSS is one of the joint projects with the RIKEN Integrated Cluster of Clusters (RICC) [13] managed by RIKEN Advanced Center for Computing and Communication. In December 2008, HPSS was upgraded (version 7.1), and started services using seven IBM p570 servers operated by AIX 5.3 as a core server and disk/tape movers. Twelve LTO-4 drives (120 MB/sec I/O with 800 GB/cartridge) are connected to six tape movers. The associated tape robot (IBM TS3500) can handle the 10,275 LTO tapes in the current configuration, and already stored approximately 1.5 PB of data combining with the data carried from the old archive system. Actually, all systems are located in the CCJ machine room. Figure 1 shows the record of the long range data transfer stored at the HPSS in RIKEN. In 2008 (RHIC-Run8), the transfer rate reached approximately 100 TB/sec as a sustained rate. Owing to the network configurations and archive system, CCJ is the only site that can perform the off-site reconstruction for such a huge amount of PHENIX data.

For the data reconstruction and individual user analysis, CCJ maintains sufficient computing power by the PC cluster operated by Scientific Linux [14] as well as the same analysis environments with the RHIC and ATLAS Computing Facility (RACF) [15] located at BNL. Both PHENIX offline library, which is provided by some AFS [16] servers in RACF, and several databases on the calibration data are automatically mirrored daily. Then, they are used for the data reconstruction, data analysis and simulation at the PC cluster in CCJ. Approximately 100 TB of disk storage are also available for the users. In February 2009, 18 PC nodes (HP ProLiant DL180 G5) were newly introduced in addition to the preexisting PC nodes as one of the upgrade of CCJ as shown in Fig. 2. Each node has dual CPUs (Intel Xeon E5430 2.66 GHz) and 16 GB memories. These nodes have twelve 3.5 inch bays of HDD for each chassis. Two 146 GB SAS disks are mounted for the operating system and used by RAID1 mode to reduce the down time originating from the troubles on the HDD. 1 TB SATA disks are set for the other ten HDD bays to use as a local data storage in each node. Consequently, the total capacity of the local storage corresponds to 180 TB in 18 PC nodes. This specification is the key feature against the I/O bound type jobs in the data analysis as will be described in the next section. Each node has a 1 Gbps network interface card, and all of the nodes are connected to a network switch mounted on the same rack. This network switch is up linked to the center network switch at CCJ (Catalyst 4900M) via a 10 Gbps connection. All of the analysis jobs are distributed by a batch job scheduler (LSF [17]) for the PC nodes in CCJ. In October 2009, we took over 20 PC nodes from RICC for the exclusive use of the CCJ users. Each node has dual CPUs (Intel Xeon X5570 2.93 GHz) and 12 GB memories. They have completely same environment with the CCJ nodes including user's home area. Condor [18] system is available as a batch job scheduler for this cluster. Thus, a total of 430 CPU cores are presently available as calculation nodes for the CCJ users. CCJ has kept the equivalent scale of number of PC nodes since the start of the CCJ operation in 2000.

The reconstructed data, *i.e.* DST, by CCJ had been transferred back to BNL until 2008. Since enough computing resources were available at RACF, the data reconstruction at CCJ became no longer necessary from 2009. Therefore, the PHENIX raw data were not transferred to CCJ as shown in the bottom inside Fig. 1. However, the size of the PHENIX data are increasing year-by-year by the stable RHIC operation. Additionally, a significant detector upgrade was made in 2010. Installations of silicon vertex detectors, which has about 4.4 million channels, is already completed for the RHIC-Run11. The PHENIX data acquisition system has achieved a performance as 500 - 800 MB/sec to record the raw data in the past run's. However, according to the detector upgrade, it will be also upgraded by factor two or three faster than the last run. It yields much more size of DSTs. Therefore, some provisions must be done to analyze the data under the experimental situations with becoming more and more stable RHIC beam.

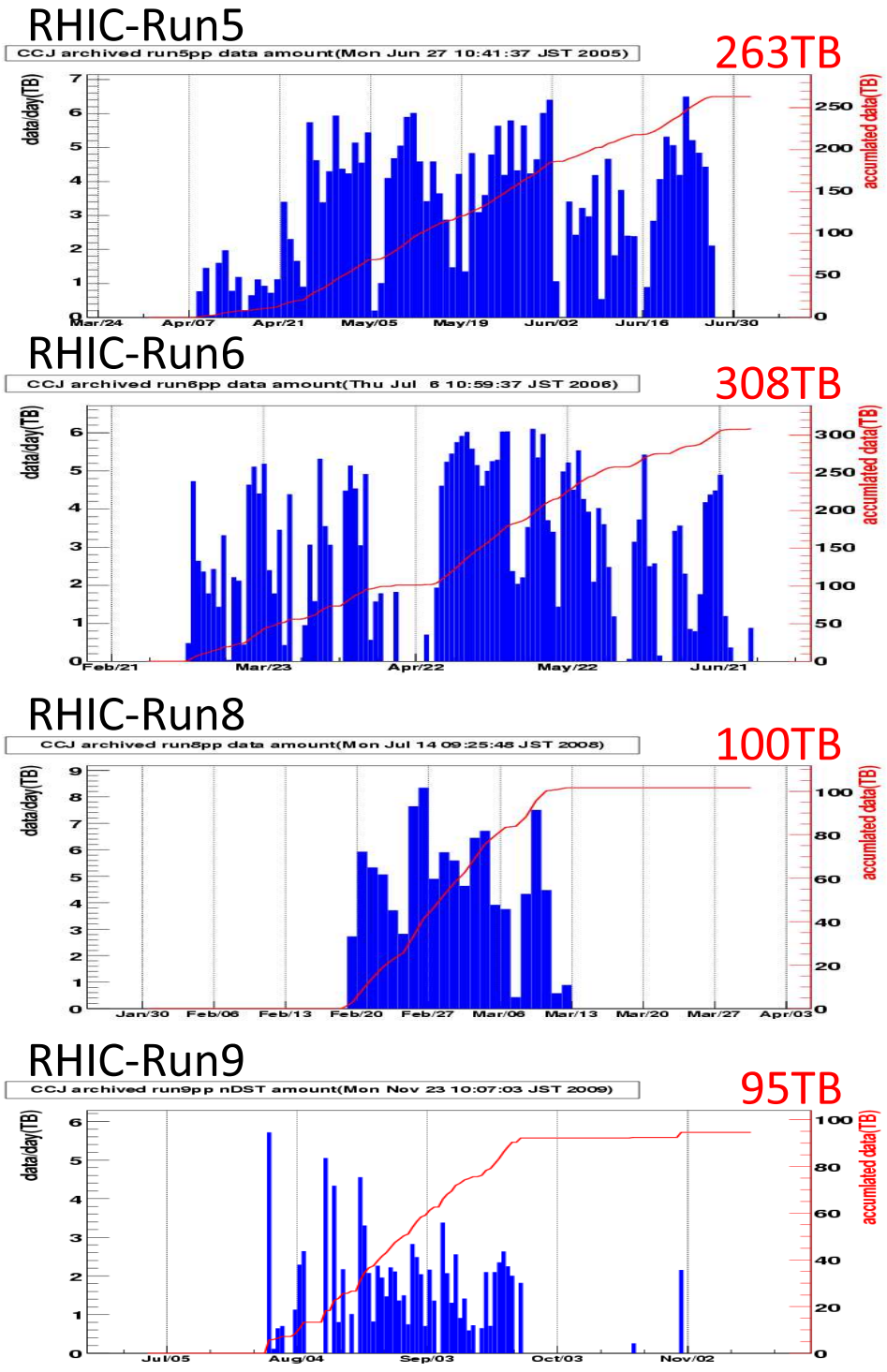

Figure 1. The record of data transfer from BNL to CCJ. Raw data of polarized proton collisions were transferred until 2008 (RHIC-Run8) to be reconstructed. In 2009 (RHIC-Run9), only reconstructed data were transferred as shown in the bottom figure.

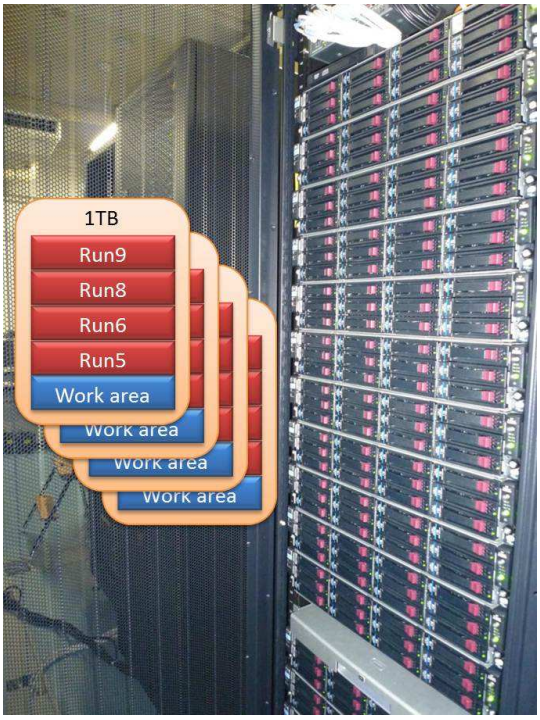

Figure 2. Newly introduced 18 calculation nodes with 12 HDD bays in each 2 U chassis.

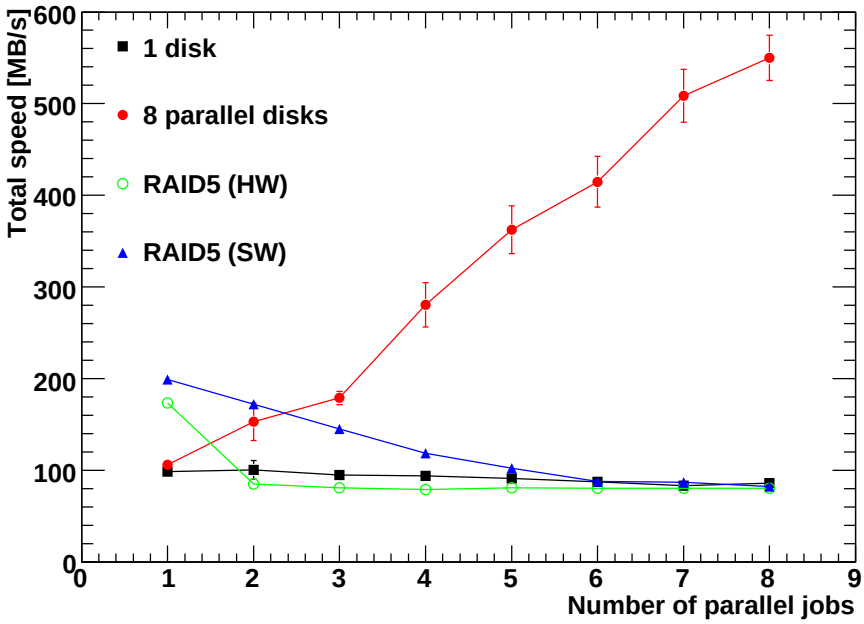

Figure 3. Average total speed for reading 1 GB files in one disk (square), 8 parallel disks (filled circle), hardware RAID5 (open circle), and software RAID5 (triangle) as a function of number of parallel jobs.

3. Development of job submission scheme

The PHENIX data stored once in the HPSS will be transferred to several RAID boxes for the user analysis. Although users can access the data in the RAID boxes through the NFS servers, multiple access from numerous calculation nodes at the same time is not possible because of the decrease in the I/O speed. Therefore, users must transfer the data from the RAID boxes to the calculation node for each batch job. Since the size of PHENIX data is growing steadily, such data transfer becomes a bottleneck in data analysis. This problem is eliminated with the use of the newly introduced calculating nodes, which have large capacity local disks for storing the data a priori (see previous section). However, since these new calculating nodes have multi-core CPUs, which is predominant in the market, data analysis remains an I/O bound type job. Therefore, it is necessary to optimize the composition of the local HDD. We performed a benchmark test to evaluate the I/O performance for the new cluster. Figure 3 shows the average total speed for reading 1 GB files as a function of the number of parallel jobs. Originally, each HDD shows the I/O performance of 100 MB/s. However, the use of a multi-HDD is not advantageous with the RAID configuration, as shown by the open circles and triangles in Figure 3. Since the RAID configuration gives us a single name space, maintaining the data location becomes easy. Nevertheless, we do not chose this configuration to maximize the I/O performance. In October 2009, approximately 96 TB Run 9 $p+p$ data, were transferred from BNL as soon as data reconstruction was completed at RACF. They were stored in local disks along with the previously stored data. Table 1 shows a summary of the dataset in the local disks accessible to users by the batch queuing system.

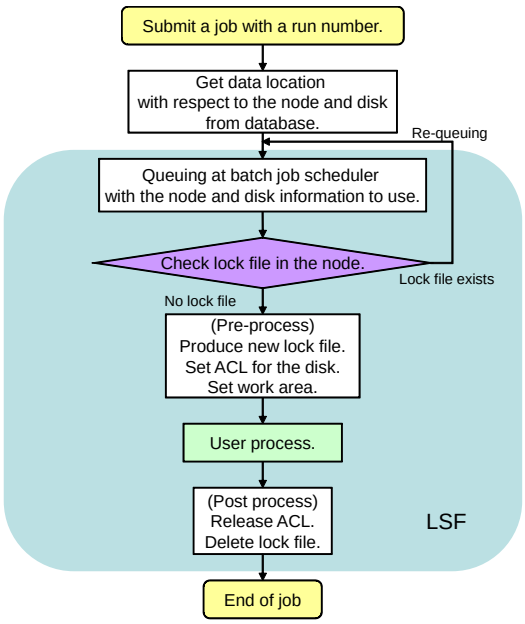

Figure 4. The work-flow of the data-oriented job submission scheme with LSF.

In the calculation nodes, users can process their own analysis code via the batch queuing system (LSF version 7.0 [17]) in the CCJ cluster. We have added some software modules to enable the user specify the DSTs distributed to the local disks during job submission. Figure 4 shows a brief work-flow of the data-oriented job submission scheme. Since all the subsets of the PHENIX data have intrinsic run numbers, the module first obtains the location of the DST by the user-specified run number from a database. Then, the module submits the user jobs to the appropriate node by the LSF. To avoid multiple access for a local disk from several jobs, the

module sets a lock file for exclusive access and grants permission only to the user by the Access Control List (ACL) method. As a result, each job dispatched to a calculating node exclusively handles a local disk. The advantage of this method is that the I/O performance is enhanced, as shown by the filled circles in Fig. 3. Further, a temporary work area for the job is assigned to the same disk with the data location. Therefore, this scheme is effective for eliminating the I/O bound problem even in case of generic jobs as well *e.g.* simulation. Each job are able to save the processing time on the data transfer approximately 10 times as compared to the typical jobs in the preexisting calculating nodes.

4. Summary

CCJ has played a lot of important roles as one of the off-site computing facilities in the PHENIX experiments since 2000. Recently, 18 calculating nodes with 180 TB local disks were introduced for effectively analyzing huge amounts of PHENIX data. A data-oriented batch queuing system was developed as a wrapper of the LSF system to increase the total computing throughput. Indeed, the total throughput was improved by roughly 10 times as compared to that in the existing clusters; CPU power and I/O performance are increased threefold and tenfold, respectively. Thus, users can analyze about 150 TB data within 9 hours. We found this is one of the reasonable and effective solutions to scan the further growing amount of data in large scale experiments with low cost, and it is really practiced. Therefore, we are planning to extend the computing cluster by adding the same type of PC nodes for the upcoming PHENIX data in the future RHIC operations.

Table 1. Summary of DSTs in local disk.

Dataset	DST type	Data amount
polarized $p+p$ 200 GeV (RHIC-Run9)	All type of DST	65.4 TB
polarized $p+p$ 500 GeV (RHIC-Run9)	All type of DST	31.2 TB
polarized $p+p$ 200 GeV (RHIC-Run8)	All type of DST	21.2 TB
polarized $p+p$ 200 GeV (RHIC-Run6)	Except for detector data	14.6 TB
polarized $p+p$ 200 GeV (RHIC-Run5)	Except for detector data	9.9 TB

References

[1] http://ccjsun.riken.jp/ccj/
[2] http://www.phenix.bnl.gov/
[3] http://www.bnl.gov/RHIC/
[4] See, e.g. A. Adare *et al.* [PHENIX Collaboration], Phys. Rev. Lett. **104**, 132301 (2010).
[5] See, e.g. A. Adare *et al.* [PHENIX Collaboration], arXiv:1009.0505 [hep-ex].
[6] http://www.rarf.riken.go.jp/lab/radiation/
[7] http://www.rarf.riken.go.jp/Eng/index.html
[8] http://www-ps.kek.jp/kekps/
[9] http://j-parc.jp/index-e.html
[10] http://www.nii.ac.jp/en/
[11] http://www.globus.org/
[12] http://www.hpss-collaboration.org/
[13] http://accc.riken.jp/ricc_e.html
[14] http://www.scientificlinux.org/
[15] https://www.racf.bnl.gov/
[16] http://www.openafs.org/
[17] http://www.platform.com/
[18] http://www.cs.wisc.edu/condor/