Operation of Objectivity/DB system at CCJ

Satoshi Yokkaichi, Hideto En'yo, Yuji Goto, Hideki Hamagaki*1, Osamu Jinnouchi, Naoki Hayashi, Takashi Ichihara, Shin'ya Sawada*2 and Yasushi Watanabe

Objectivity/DB [1] is a commercial Object-Oriented Database Management System (ODBMS) and is widely tested and used in the field of high energy/nuclear physics. It has been studied through the CERN RD45 project[2] and the ATLAS experiment. It was used in the Zeus experiment for the tag DB in offline analysis. In the Babar experiment, it was used for online data storage and DST (data summary tapes) storage. However, in some experiments, such as the STAR experiment, gave up to use Objectivity/DB, due to difficulty in maintaining the system.

In the PHENIX experiment, we also use Objectivity/DB as a calibration DB and a run DB in online/offline computing. Single Federation DB (FDB) is being operated on a dedicated Linux machine in the PHENIX counting house. FDB in Objectivity means a set of database files managed under a single namespace. Almost all online programs in the counting house and almost all offline analysis programs on the machines in RCF (RHIC Computing Facility at BNL) access the single FDB.

In RIKEN CCJ[3], we also require the data in Objectivity/DB to run the PHENIX analysis program, particularly for the 'reconstruction' process in the data analysis flow. In this process, the raw data and simulation data are converted to the DST, which has 'reconstructed' particle track information, shower cluster information in calorimeters, Čerenkov ring image in RICH and so on. Then we have to access and read the PHENIX official FDB. Some methods were discussed and we now operate a 'snap-shot' copy method [4] to duplicate the official FDB and deploy it in CCJ. Using the copied FDB, official PHENIX DST production for the data taken in 2000 was performed at CCJ last December and about 40% of the available raw data is processed in CCJ. The remainder was processed at RCF.

For such production, the calibration DB can be used with fixed – 'snap-shot' – data which is copied by hand at the appropriate time. However, during preproduction, calibration works are ongoing and there are frequent updates of the official DB. In this case, frequent updating of the DB in CCJ is also expected. The automatic daily copying is now being tested for two month. Current size of the PHENIX FDB is about 2.4GB. The two main causes of failure of automatic duplicate processes were as follows.

1) Failure of the DB backup process in RCF using built-in command 'oobackup'. This is caused by the remaining DB locking, mainly from some dead or forgotten processes. 2) Network failure between RIKEN and BNL. Instability of WAN broke or delayed the file transfer using the 'scp' command. These two facts convinced us that the 'replication' function of Objectivity/DB, which is a real-time data duplication method via a network which we did not adopt, could cause many problems in the operation of the official FDB in our current network environment.

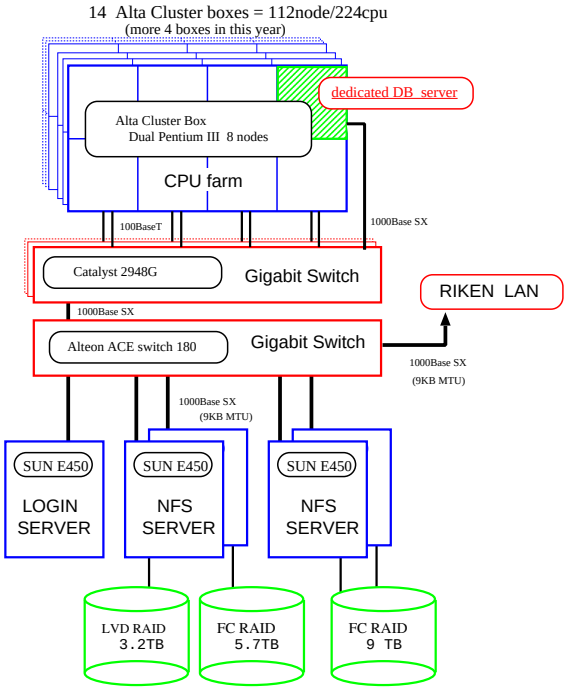

Fig. 1. Current CCJ system configurations.

A dedicated Linux host for DB in CCJ has dual 1GHz Pentium III CPUs, 1GB memory and a 16GB SCSI local disk. 1000Base-SX ethernet connects the DB host machine to the 1000Base-SX/100Base-T ethernet switches. Linux machines as CPU farms on which the reconstruction jobs run are connected with the switches using 100Base-T. In fact the DB host is one of the farm machines but the network interface has been upgraded. SUN E450 as NFS servers with Fibre Channel RAID disks are also connected with the switches using 1000Base-SX. The RAID disks are served to both the DB host and CPU farm hosts. These configurations are shown in Fig.1. In this environment, the network bandwidth was measured using a benchmark program 'netperf'[5]. Over 100MB/s was

*1 Center for Nuclear Study, University of Tokyo
*2 High Energy Accelerator Research Organization (KEK)

achieved using multiple sessions from the DB host to 30 CPU farm nodes, while 12MB/s was achieved from the DB host to typical single CPU farm node. The DISK I/O bandwidth of the DB host was 26MB/s for block read, which was measured using a benchmark program 'bonnie'[6].

We measured the data reading performance of the Objectivity/DB using the PHENIX reconstruction program for 2001 data (Oct.28 version). This program reads about 300MB of calibration data and run-information data from the DB at the initialization stage. Multiple jobs running on multiple CPU farm nodes read the Objectivity/DB data simultaneously (differences in starting times of jobs are less than 60 seconds). Two read methods are tested. One uses the Objectivity Advanced Multi Thread Server (AMS). AMS runs on the DB host , reading the DB data from the local disk and transferring it to the client processes. The other uses ordinary NFS. Client processes read the data from the NFS disk on which the DB data is placed. In both cases, the Objectivity Lock Server also runs on the DB host and processes the read lock of DB.

In the AMS case, about 4MB/s was achieved for a single job, 40MB/s for 10 jobs, 35MB/s for 100 jobs and 28MB/s for 200 jobs. The saturation effect was observed. CPU usage was 95% for over 40 jobs. Local disk I/O does not seem to limit the performance. A cache effect may be effective for such multiple access to the same data at the same time. These were measured using the UNIX command 'netstat -i'. The number of the AMS thread had no clear effect between 8 (default), 4 and 16 with respect to the data output flow.

In the NFS case, about 50MB/s was achieved for 200 jobs, for the DB data placed on the SUN NFS server. This was measured using SNMP (simple network monitoring protocol) at a port of the ethernet switch connected with the NFS server. CPU usage of the NFS server host was about 50% for kernel and 50% was idle.

It seems that the latter is more efficient than the former. However, the NFS method with the Linux disk will be compared with the AMS method on Linux, because the SUN NFS server is more expensive and may not be available in such dedicated way.

We would like to thank our collaborators at PHENIX. Dr M. Purschke advised us regarding a kernel configuration of the DB host machine. Dr T. Chujo and A. Kiyomichi explained the behavior of AMS from their experience of production job management at RCF.

References

1) http://www.objectivity.com/
2) http://wwwinfo.cern.ch/asd/rd45/index.html
3) T. Ichihara et al.: In this volume of RIKEN Accel. Prog. Rep., RIKEN Accel. Prog. Rep. **34**, (2000).
4) S. Yokkaichi et al.: RIKEN Accel. Prog. Rep. **34**, (2000).
5) http://www.netperf.org/netperf/NetperfPage.html
6) http://www.textuality.com/bonnie/