The BELLE Computing Facility

Ryosuke Itoh [a]

[a] *Institute for Particle and Nuclear Study, High Energy Accelerator Research Organization (KEK),*
1-1 Oho, Tsukuba, Ibaraki 305, Japan

The BELLE computing facility was constructed for the data analysis at the BELLE experiment. The system consists of a large RISC farm, a large tape and disk storage system with the HSM and many workstations. The detail of the system is presented in this talk.

Key words: large system; RISC farm; HSM

1 Introduction

At the BELLE experiment which is the B-factory experiment at KEK, the typical event rate is expected to be about 200 Hz after the online data reduction with the event size of about 30KB. The maximum data flow is expected to be about 15MB/sec. To process these data without any delay, the required CPU power is estimated to be more than 30000 VAX-equivalent in total where 1 VAX-equivalent is defined as the CPU power with 1 SPECint92 + 1 SPECfp92, while the mass storage is required to be capable of handling data of more than 30TB/year. The BELLE computing facility was designed and constructed to satisfy these requirements using the technologies optimized for the usage at BELLE. The detail of the system is reported in the following sections.

2 Overview

Figure 1 shows the conceptual design of the system. The computing activity at BELLE experiment is divided into 4 categories. One is the online computer farm on which the online event reduction is performed. This is provided as a part of the BELLE data acquisition system. The second is the data server where all the data are concentrated and managed. The third is the offline

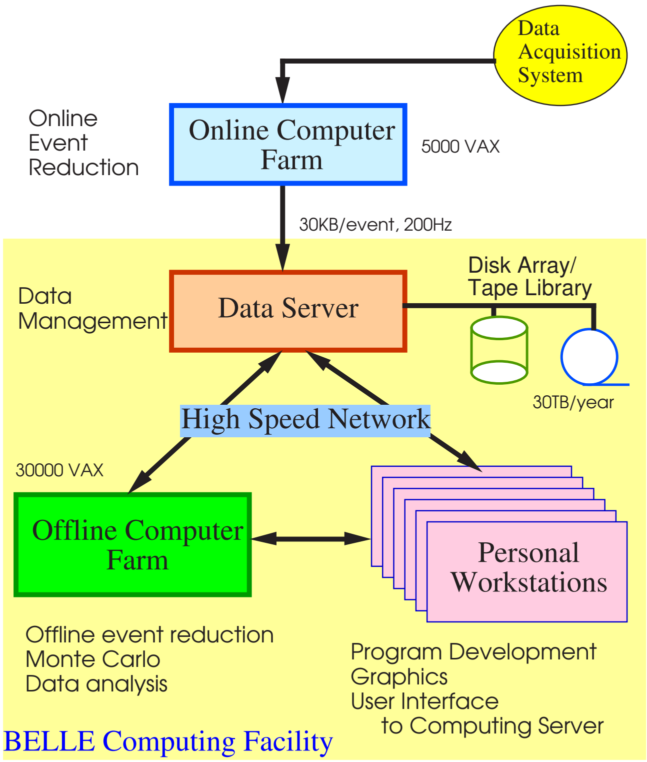

Fig. 1. Conceptual design of computing environment for BELLE

computer farm which provides the main processing power. The fourth is the user workstations on which the program development is done. Among these, the data server, the offline computer farm and the user workstations form the BELLE computing facility.

The BELLE computing facility was obtained by the bidding procedure from a single company. Fujitsu won the bid and the entire system has been started operation from Jan. '97. Fig. 2 shows the configuration of the central system. Fujitsu acted as the system integrator and the total system was constructed using the products from multiple vendors.

3 Offline Computer Farm

According to the detailed study of the data processing at BELLE, the data processing jobs can be categorized into two classes. One is jobs which require a large computing power up to 5000 VAX equivalent. Some of jobs in this class require the high speed I/O (>5MB/sec) with the data server. "Group-shared" jobs like the DST production and the full detector simulation are in this category. The number of these jobs are expected to be limited up to 5-10 at one time. The other is jobs which can be executed on a single workstation. Most of the user analysis jobs are in this category. Sometimes the fast I/O

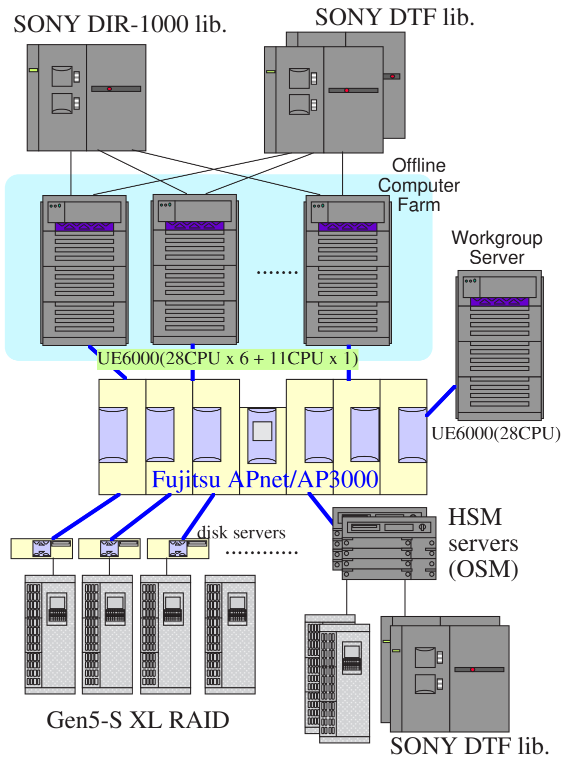

Fig. 2. Central facility of the BELLE computing system

is necessary to access data files, but the transfer rate is not so severe as that required by jobs in the first category (up to a few MB/sec).

Based on these considerations, the offline computer farm was designed to process the jobs in the first category by having 5-10 units of a computer system with the CPU power of more than 5000 VAX-equivalent/unit. However, the CPU power of 5000 VAX-eqv cannot be achieved by a single CPU and the parallel processing is necessary at some level. Since most of jobs are to process many event data, the trivial event-by-event parallel processing is enough. However, the most of users are not experts in the parallel programming, the parallel processing mechanism is required to be hidden from the users. The parallel processing mechanism can be absorbed in the analysis framework and the platform on which the parallel processing nature can be well hidden with the framework is desired[1].

Several possibilities were examined to satisfy these requirements including, 1) SPMD using many workstations, 2) Workstation cluster with event distribution scheme, and 3) SMP-server cluster. After a group-wide discussion, we decided to use a SMP-server cluster system as a offline computer farm.

The realized computer farm consists of 7 units of SUN UltraEnterprise 6000 servers. Each server has 28 UltraSparc-I CPU's running at 167MHz (except one server which has 11 CPU's). The CPU power of one server is more than 7000

VAX-eqv. These servers are connected together by Fujitsu APnet/AP3000 network system which is the network specially designed for Fujitsu's dedicated parallel system AP3000.

All the servers are operated as batch systems and interactive login to the servers by users are not allowed. To control the batch queueing, LSF by Platform Computing is employed and the resource allocation for batch jobs is managed by this system.

The scalability of the SMP server was tested using the realistic applications written on the BELLE analysis framework and confirmed to be ensured up to the expected maximum number of CPU's[1]. The data transfer speed between servers through APnet was also tested using PVM, MPI and TCP/IP sockets, and measured to be more than 50MB/sec for PVM/MPI and more than 20MB/sec for TCP, respectively.

4 Data Server

The data server was designed with four different parts. One is the tape system to store the raw data from the detector (called Tape Library I). Since the maximum data transfer speed from the online computer farm is expected to be about 15MB/sec, and we desired to store the data in a single tape device directly, there was no other choice than the SONY DIR-1000 tape drives together with the library robotics. The library can handle 304 tapes of 80GB/tape and therefore more than 30TB can be stored. In the library, 11 tape drives are installed where 3 drives are connected to the online data transfer system while the remaining drives are connected to the servers of the offline computer farm directly by fast/wide SCSI2.

The second system is the tape library system to store DST's (Tape Library II). This system is used as the final storage to place the DST's and the raw data copied from Tape Library I. A SONY DTF tape library system is used for this system. The library can contain 1790 tapes of 42GB/tape which has the storage capability of more than 75TB. 15 drives are installed in the library and connected to the computer farm servers directly by fast/wide SCSI2.

The third system is a data disk subsystem migrated with a tape library system. This system is intended to place large mini-DST files for the user analysis. As for the disk, 4.2TB of GEN5 S-XL RAIDs are employeed. The disks are connected to SUN Ultra2 workstations which work as disk servers and then connected to APnet. The disk access from the servers of the offline computer farm is made via APnet. These disk system are migrated with SONY DTF tape library with the capability to handle 775 tapes (>30TB in total). 8 tape drives are connected to a HSM server on which SONY PetaServe is operated. SONY PetaServe is based on OSM from Computer Associates and the HSM management is done by this software.

The fourth system is a disk subsystem to place user's home directories. The system is also a set of GEN5 S-XL RAID disks and connected to APnet in the same manner as that for the data disks. The home directory disks are operated using DFS of DCE to provide a versatile security control.

The disk transfer rate was tested by measuring the I/O speed between the data disk and a server of the offline computer farm. The disk was accessed using the NFS and the measured rate was 8.26MB/sec for the disk read, while 8.03MB/sec for the disk write.

5 User Workstations

For the user analysis and the program development for the offline computer farm, all users were expected to have their own workstations. However, the management of a workstation is rather difficult for a modest physicist and some people prefered to use X-terminals connected to one workstation shared by other peoples. So we determined to have workgroup servers for these people together with the workstations.

Since the most of user's analyses are expected to be done on workstations while the computer farm to be used only when a large CPU power is necessary by submitting a batch job using the same executable image of the analysis, the workstations and workgroup servers are required to have the save CPU architecture. 23 units of Sun Ultra 2 workstations are employed for the user workstations while 1 unit of UltraEnterprise 6000 and 2 units of UltraEnterprise 4000 were implemented as workgroup servers.

Fig. 3 shows the topology of workstations and workgroup servers scattered over KEK campus. All the workstations and workgroup servers are connected to the central facility via the campus-wide ATM network at KEK. The user's home directories are all placed in the home directory disks on the data server of the central facility which are mounted by workstations/workgroup servers using DFS. This enables the seemless environment among any workstations and workgroup servers. To reduce the massive access to the mass storage from workstations via network, "replica" mechanism of DFS is utilized to create a disk cache on local disks of workstations / workgroup servers.

To use the offline computer farm from workstations/workgroup servers, the batch submission through the network is done using LSF. DCE is utilized

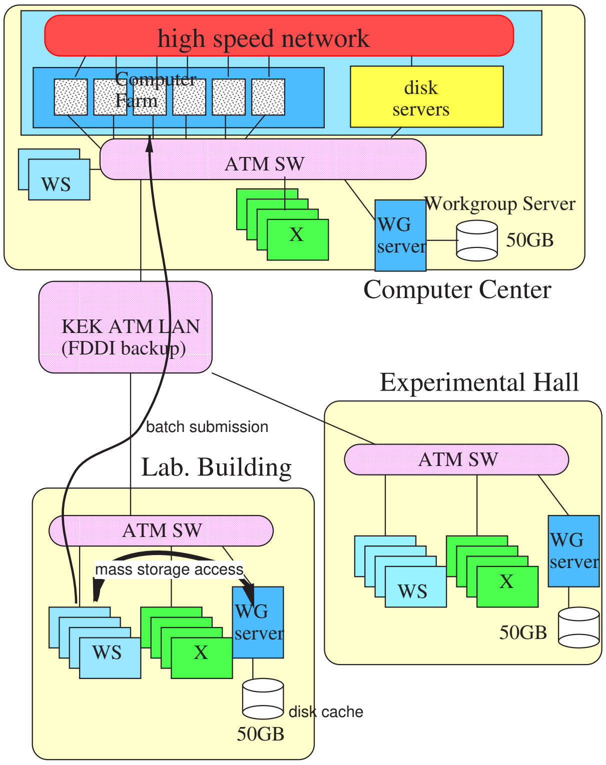

Fig. 3. Workstation/Workgroup server topology at KEK

system wide for the user management to improve the system security.

6 Summary

For the data analysis at the BELLE experiment, the BELLE computing facility was constructed. The system is based on 1) a cluster of SMP servers with a total CPU power of more than 46000 VAX-eqv. as the offline computer farm, 2) SONY tape library system with the storage capability of more than 130TB and a 4.2 TB of GEN5 S-XL RAID disks as the data server, and 3) many workstations and workgroup servers for the program development. The computer farm is operated as a batch system for the workstations/workgroup servers and the batch system is managed by LSF from Platform Computing. A HSM sofware called PetaServe from SONY is employed to migrate disks and tapes. The entire system is managed under DCE environment and DFS is utilized for file management.

Acknowledgement

This presentation was prepared together with N.Katayama, A.Manabe, and

S.Sasaki. I deeply appreciate them for their efforts.

References

[1] R.Itoh, "BASF - Belle AnalysiS Framework", presented at this conference (A244).