Available online at www.sciencedirect.com
SCIENCE DIRECT®
Nuclear Instruments and Methods in Physics Research A 534 (2004) 53–58
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

Belle computing system

Ichiro Adachi[a,*], Taisuke Hibino[a], Luc Hinz[b], Ryosuke Itoh[a], Nobu Katayama[a], Shohei Nishida[a], Frédéric Ronga[b,1], Toshifumi Tsukamoto[a], Masahiko Yokoyama[a]

[a]*IPNS, KEK, Oho 1-1, Tsukuba, Ibaraki 305-0801, Japan*
[b]*LPHE, EPFL, Lausanne, Dorigny CH-1015, Switzerland*

Available online 30 July 2004

Abstract

We describe the present status of the computing system in the Belle experiment at the KEKB e^+e^- asymmetric-energy collider. So far, we have logged more than 160 fb^{-1} of data, corresponding to the world's largest data sample of 170 M $B\bar{B}$ pairs at the $\Upsilon(4S)$ energy region. A large amount of event data has to be processed to produce an analysis event sample in a timely fashion. In addition, Monte Carlo events have to be created to control systematic errors accurately. This requires stable and efficient usage of computing resources. Here, we review our computing model and then describe how we efficiently proceed DST/MC productions in our system.

PACS: 07.05.Bx

Keywords: Data processing; Pc farm; Belle

1. Introduction

The Belle experiment [1] is the B-factory project at KEK to study CP violation in the B meson system. The KEKB accelerator [2] is an asymmetric energy collider with 8-GeV electrons to 3.5-GeV positrons, operating at $\Upsilon(4S)$ energies. The Belle group has been taking data since June 1999 and has logged more than 160 fb^{-1} of data until December 2003. This means that we have the largest data sample of B-meson pairs around the $\Upsilon(4S)$ energy region in the world. The KEKB reached its design luminosity of $10^{34}\,cm^{-2}\,s^{-1}$ on May 2003. The KEKB performance is still being improved, and we can recently accumulate an integrated luminosity of about 800 $pb^{-1}\,day^{-1}$.

We have to promptly process all of the beam data to provide it for user analyses. To do this, the DST production should be stable and the computing resources must be used in an efficient way. The

*Corresponding author. Tel.: +81-29-864-5320.
E-mail address: ichiro.adachi@kek.jp (I. Adachi).
[1]Present address: IPNS, KEK, Japan.

0168-9002/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2004.07.058

Monte Carlo (MC) data should be generated with statistics large enough to control the experimental systematics. As a result, the data size that we should handle is extremely huge and a mass storage system must be used to avoid network traffic, and data management for entire data sets should be carefully handled without loosing flexibility, for instance, any modification of data distributions.

Provided a large amount of data sample, we have published a variety of physics results related to B-meson decay, which is highlighted by the first observation of large CP violation in B-meson decay [3]. The quick and stable data processing greatly contributed to this remarkable achievement.

In this paper, we describe the details of our computing system after a brief sketch of the Belle software utilities. In the next section, how we proceed DST as well as MC productions will be mentioned, and then, the summary is given.

2. Belle software

2.1. Core utility

In the Belle experiment, a unique software framework, called Belle AnalySis Framework (B.A.S.F.), has been employed for all phases in event processing from online data-taking to offline user analyses. This is a "home-made" core software developed by the Belle group. In this scheme, each program, written in C++, is compiled as a shared object, and it is treated as a module. When one wants to run a program with this framework, the modules defined in its scripts are dynamically loaded.

The data handling is done with the traditional bank system, named as PANTHER, with a zlib compression capability. In this utility, data transfer between different modules is made and data I/O is manipulated. PANTHER is the only software to handle our data at any stage of the data processing.

The typical event data size for raw data is 35 kB, and it is increased up to 60 kB for reconstructed DST data, which contains all the detector information after unpacking, calibration and analysis. For user analyses, a compact data set "mini-DST" is produced, which is approximately 12 kB for one hadronic event.

2.2. Reconstruction and simulation library

The reconstruction package is built when a major update of programs, which can affect the final physics analysis, is made. Usually, it is built once or twice per year. Once a new library is released, we need to process all the events to produce a consistent data set for analysis.

For MC data, the detector response is simulated based upon the GEANT3 library [4]. Here, background events, calculated from beam data, are overlayed onto MC events. The same reconstruction codes are also applied to MC-simulated events.

The detector calibration constants are stored in the database, for which PostgreSQL [5] is adopted in the Belle experiment. Two database servers are running in our computing system, where one is for public usage and the other for DST/MC productions. The contents of each server are periodically mirrored to the other, and a backup tar file for the contents of the original database server is archived once per week. For linux users, one can start up ones own database server in her/his PC, according to Belle instructions. A user, if necessary, can download original database contents from a KEK B computer for private purpose.

3. Belle computing system

3.1. Computing model overview

Fig. 1 indicates an overview of our computing model for the Belle experiment. As can be seen, it is comprised of four major elements. The first one is the KEK B Computers, which has been operated since February 2001 [6]. This has been a principal system, where data processing as well as analyses have been performed. In addition, we have equipped a 60 TB disk storage area March 2003. In the disk space, more beam and MC data can be kept. As for a network inside Japan, the

Super-SINET link has been available since 2002 [7]. To enrich CPUs for user analyses, PC farms dedicated to analysis jobs were installed in June 2003. These components are interlinked with each other using a fast network of 1–10 Gbps.

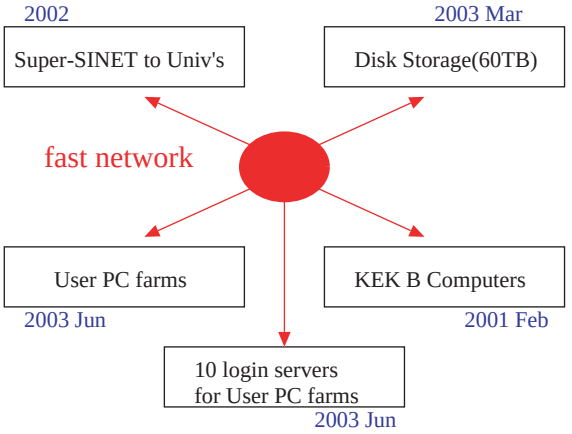

Fig. 1. Computing model overview.

3.2. KEK B computers

The KEK B computers are schematically shown in Fig. 2. This system consists of the three cardinal links. The first one is called the "computing network", which interconnects the 38 Sun servers with the PC farms, which are described below. These Sun hosts are operated as a batch job system controlled by the LSF scheduler. The computing network is also attached to the DTF2 tape robotic device with a 500 TB total volume which is used for raw data and DST data storage. The next network links the nine work group servers of the Sun hosts to the storage devices, which consists of the 8 TB file server and the hierarchy mass storage (HSM) of 120 TB capacity with the 4 TB staging disk. The work group servers are connected via the "user network" to 100 PCs for users. With this network, a user can login from her/his PC to one of the work group servers, from which one can submit a batch job to the 38 Sun compute servers. Table 1 summarizes the CPUs allocated in the KEK B computers.

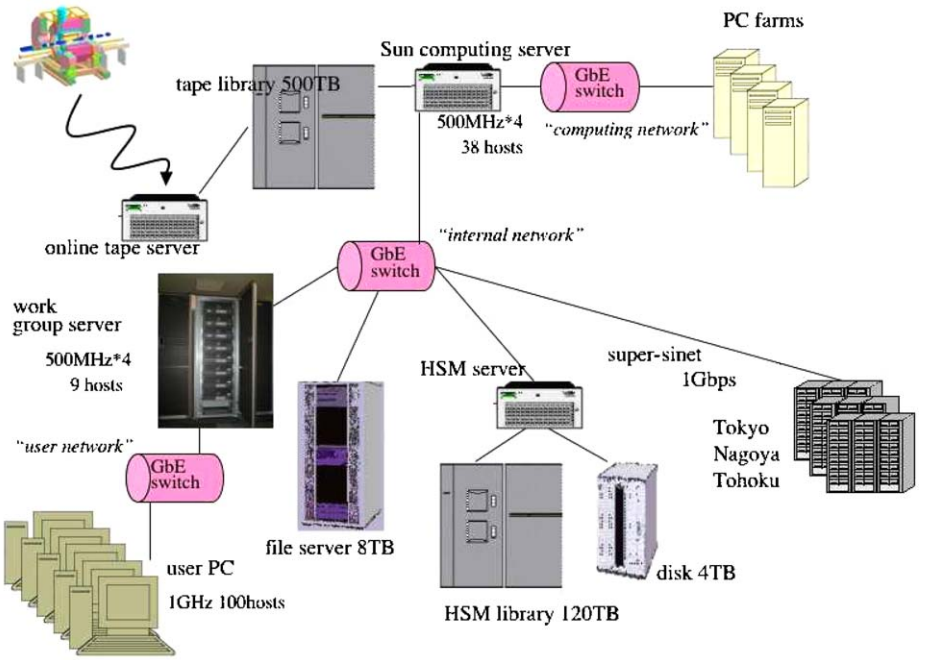

Fig. 2. Belle computing system.

Table 1
Summary of CPUs for the KEK B computers

Host	Processor	Clock	#CPUs	#Nodes
Computing server	Sparc	500 MHz	4	38
Work group serve	Sparc	500 MHz	4	9
User PC	P3	1 GHz	1	100

3.3. PC farm

As beam data have been continuously accumulated, computing needs for event processing are seriously increased. To fulfill this requirement, a bunch of PCs have been installed and connected into this existing network by considering the best usage without making a bottleneck. Table 2 tabulates the PC farm CPUs in our system. As one can see, our PC farms are heterogeneous from various vendors. The best choice at the moment to obtain new PCs usually encouraged us to purchase from a variety of vendors, which effectively reduced the costs for CPUs.

In all of the PCs, linux utilities as well as the Belle library packages have been loaded, and we can use them as clusters of seamless PC systems to process event data.

The primary purpose to add PC farms is that we have to increase the CPU power for DST and MC productions. In 2003, we expanded the usage for our PCs by releasing new PC farms for user analyses, as a possible solution for ever increasing users' demand. The 10 login servers have been arranged with a 6 TB local-disk area, where user histogram files and analysis codes are located. From each login server, a job can be submitted to a user PC farm consisting of 84 PCs with dual Xeon 2.8 GHz CPUs. All PC's are managed by the LSF queuing utility. From user PC farms, beam and MC data samples, which are stored in the 60 TB disk mentioned previously, can be accessed.

3.4. Super-SINET at Belle

A fast 1 Gbps network dedicated to academic activities, Super-SINET [7], has been available between KEK and major Japanese universities. This link enables us to copy a bulk of beam and MC data from KEK to other institutions and vice versa. Moreover, computing resources can be shared as seamless system using Super-SINET. For example, one disk connected to a PC at Nagoya university can be mounted to the KEK computer via NFS as if it were located inside the KEK site. Then, output data can be written onto the disk directly from the KEK computer. It is possible to make efficient and full use of total resources collaboration-wide.

Table 2
Breakdown of the PC farm CPUs

Vendor	Processor	Clock	#CPUs	#Nodes	Total CPU (GHz)
Dell	P3	500 MHz	4	16	32
Dell	P3	550 MHz	4	20	44
Compaq	P3	800 MHz	2	20	32
Compaq	P3	933 MHz	2	20	37
Compaq	Intel Xeon	700 MHz	4	60	168
Fujitsu	P3	1.26 GHz	2	127	320
Compaq	P3	700 MHz	1	40	28
Appro	Athlon	1.67 GHz	2	113	377
NEC	Intel Xeon	2.8 GHz	2	84	470
Total				500	1508

3.5. Data management

A large volume of data is comprised of more than 30 K data files including beam and MC events. Basically, each user has to go through all these files to obtain the final physics results. It is thus very important to notify all of file locations to users for their analyses. These data files are distributed over a bunch of storage disks and sometimes data files have to be moved for various reasons, such as disk failure and so on. From an administrative point of view, it is necessary to maintain flexibility in data management despite any change of the file locations. To solve this problem, we have registered all the attributes of data files, such as locations, data type and number of events into our database, and they serve as "metadata" to access actual beam data. The central database contents for data file information is maintained and updated whenever new data is

available. The web-based interface between database and users is prepared to extract necessary information. In a user's batch job, inquiries to the database are automatically issued, and the user can easily analyze event data without knowing the actual file locations.

4. Data processing

4.1. Beam data production

The scheme of the DST production is shown in Fig. 3. In the first step of DST production, one of Sun compute hosts is assigned as a tape server, and two DTF2 tapes; one is for the raw data, and the other for the DST data, are mounted. Then, the raw data are read from the tape and are distributed over the PC nodes. In each PC node, event processing is performed in the B.A.S.F. framework. After the reconstruction is made, event data are sent back to the tape server, where data are written onto the tape as DST.

The next step, the event skimming, is carried out in such a way that DST data are again read from the DST2 tape at the Sun compute server, and an appropriate selection criteria is imposed onto the data. In the case that an event satisfies selection conditions, it is saved onto a disk as a skimmed event. Each event is examined by a set of selections to pick up, for instance, μ-pair events for a detector study or hadronic events for a physics analysis.

In our system, we can process about $1\,\mathrm{fb}^{-1}$ of beam data per day with 40 PC nodes of quad Intel Xeon 700 MHz CPUs each, and it can be increased

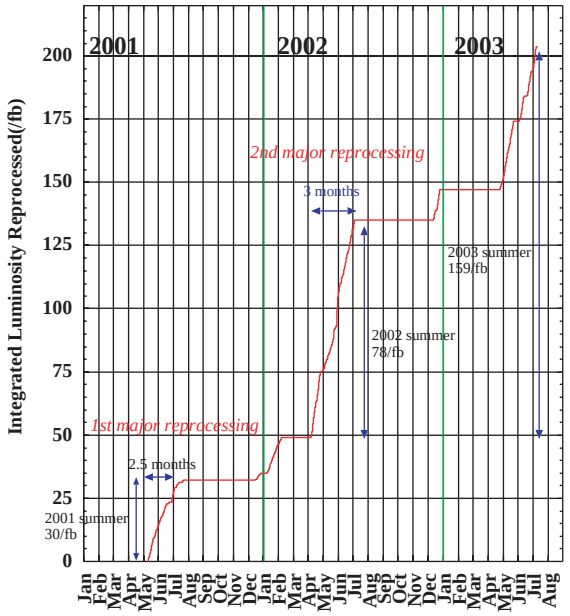

Fig. 4. History of reprocessing from 2001.

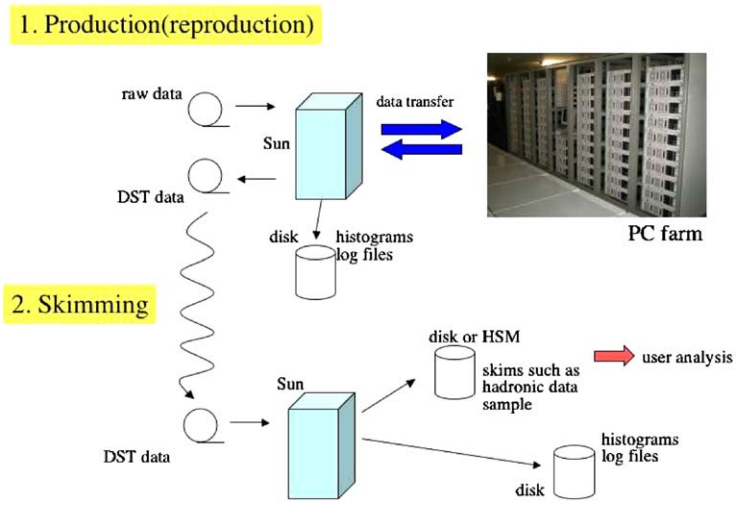

Fig. 3. Schematic drawing of the DST production scheme.

up to $2\,fb^{-1}$ per day by adding PC nodes. Fig. 4 shows a history of our beam data processing from 2001. In 2001, we completed the first entire reprocessing with a single version of the reconstruction package, providing $30\,fb^{-1}$ of a data sample for 2000 summer conferences. In 2002, a major update of the software was made in April 2002, and a second reprocessing for $78\,fb^{-1}$ of data using this library was performed from April to June. Last year, we added another $80\,fb^{-1}$ by reprocessing them, amounting to $159\,fb^{-1}$ of the total beam data.

4.2. MC production

We have three types of MC samples; $B^0\bar{B}^0$, B^+B^-, and continuum events. They are produced on a run-by-run basis, where each MC sample file corresponds to each beam data file. After data production for one run beam file is done, run-dependent information, like beam interaction profiles and background hit rates, are calculated immediately and, based upon this information, MC sample data corresponding to the beam run file is created for three types. Beam background hits are overlayed with MC-generated data to mimic actual events as precisely as possible. Fig. 5 shows how we have produced MC samples during the past 2 years. We have produced 2.2 billion events in total by the end of 2003, which is equivalent to 3-times greater statistics for $159\,fb^{-1}$ real beam data.

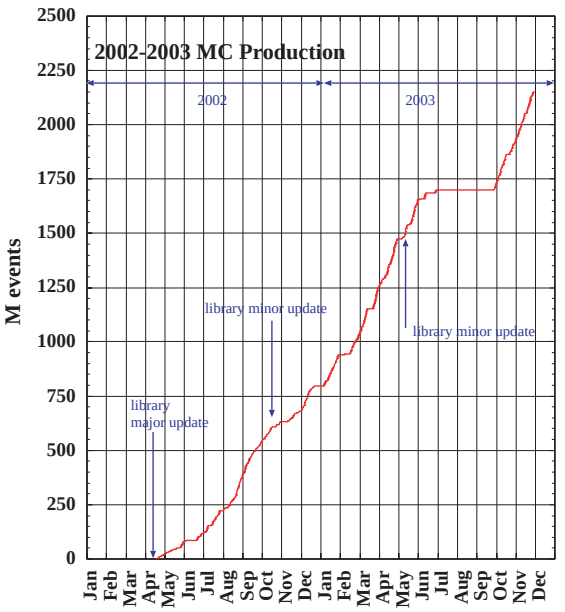

Fig. 5. MC production done in 2002 and 2003.

Another aspect of the MC production is contributions from remote institutes outside of KEK. Approximately one quarter of the MC events have been produced at remote sites, and are transferred to KEK via network.

5. Summary and plan

The Belle computing system has been successfully working, and we have processed a bulk of beam data of more than $250\,fb^{-1}$ so far. For MC data, 2.2 billion events have been produced, which correspond to more than 3-times more statistics than the real data. Those data have been managed via our database, and this scheme allows us to provide data sets both steadily and flexibly.

In the summer of 2003, a new silicon vertex detector was installed. It is expected that this detector will further expand our tracking and vertexing capability. At present, a reconstruction algorithm based upon the new Belle configuration is being developed. After it is completed, beam data in 2003 autumn runs will be processed. To do this, we are planning to double our processing power for beam data by adding more CPUs and storage devices.

References

[1] A. Abashian, et al., Nucl. Instr. and Meth. A 479 (2002) 117.
[2] S. Kurokawa, E. Kikutani, Nucl. Instr. and Meth. A 499 (2003) 1.
[3] K. Abe, et al., Phys. Rev. Lett. 87 (2001) 091802.
[4] R. Brun, et al., Geant3.21 CERN Report No. DD/EE/84-1(1987).
[5] http://www.postgresql.com/.
[6] I. Adachi, et al., Proc. of CHEP03, La Jolla, CA, March 24–28, 2003.
[7] National Institute of Informatics, Japan, http://www.sinet.ad.jp/.