COMPUTING FOR BELLE

Belle collaboration
Nobu Katayama, Oho 1-1, Tsukuba-shi, 305-0801, Japan

Abstract

The Belle experiment operates at the KEKB accelerator, a high luminosity asymmetric energy e^+e^- machine. KEKB has achieved the world highest luminosity of 1.39 times 10^{34} $cm^{-2}s^{-1}$. Belle accumulates more than one million B meson anti-B meson pairs in one good day. This corresponds to about 1TB of raw data per day. The amount of the raw and processed data accumulated so far exceeds 1.2PB. Belle's computing model has been a traditional one and very successful so far. The computing has been managed by minimal number of people using cost effective solutions. Looking at the future, KEKB /Belle plans to improve the luminosity to a few times 10^{35} $cm^{-2}s^{-1}$, 10–24 times as much as we obtain now. This paper describes Belle's efficient computing operations, struggle to manage large amount of raw and physics data, and plans of Belle computing for Super KEKB/Belle.

STATUS OF BELLE EXPERIMENT

Belle started taking data in 1999 and has nearly been doubling its data size every year. Figure 1 shows the history of its accumulated integrated luminosity.

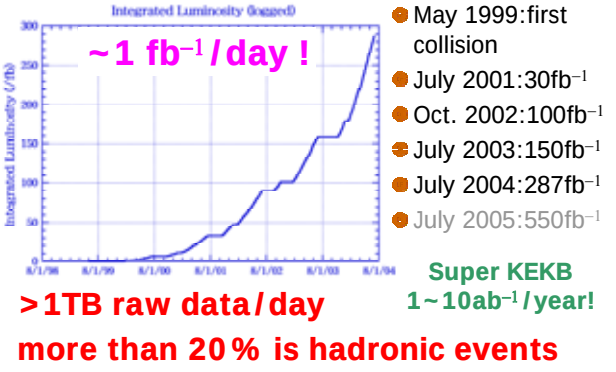

Figure 1: History of Belle integrated luminosity

KEKB now runs with the continuous injection mode.

No need to stop run
Always at ~max. currents and therefore maximum luminosity

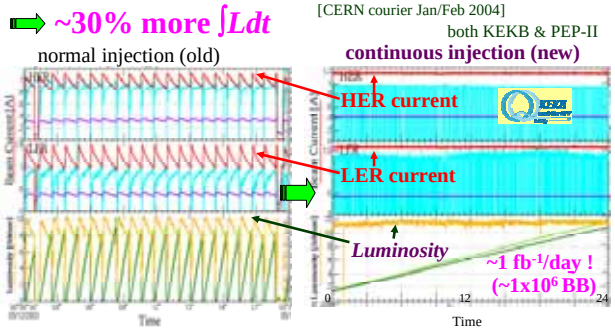

Figure 2: Continuous injection improves int. lum.

As shown in Figure 2, Belle does not stop data taking during beam injections so that the beam currents, thus instantaneous luminosity can be kept at nearly maximum almost all the time. It results in an increase of integrated luminosity by 30%.

BELLE COLLABORATION

The Belle collaboration consists of more than 400 physicists from 58 institutions from 13 countries as shown in Figure 3; Universities and HEP institutions from Russia, China, India, and universities from Japan, Korea, and Taiwan, Australia, US and Europe. KEK dominates in one sense as 30~40 staffs work on Belle exclusively and most of construction and operating costs are paid by KEK. However, universities dominate in another sense as young students from the universities are to stay at KEK, to help operations and to do physics analyses. Shortages of human resources, in particular, in the area of computing and software, are our problem as Belle is running, doing physics analyses and planning the major upgrade.

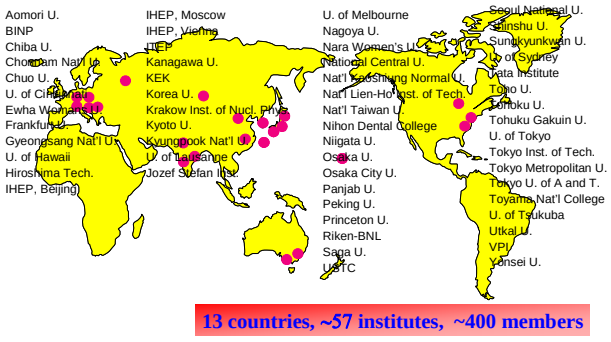

Figure 3: The Belle collaboration

SOFTWARE

Core software

Belle software is supported on two operating systems, Solaris 7 on Sparc CPUs and RedHat 6/7/9 on Intel CPUs. Belle uses the GNU compiler collection of version from 2.95.3 to 3.3. Belle code compiles with SunCC as well. Belle software requires no commercial software. Belle uses QQ from CLEO, EvtGen from CLEO and BaBar, CLHEP, GEANT3 and CERNLIB from CERN, and postgresql. Parts of Belle software (parts of detector simulation code and calibration code) are written in FORTRAN. For I/O, Belle uses a home grown software package called panther and compression using zlib. Panther is the only data format for all stages from DAQ to final user analysis skim files. Index file (pointer to events in the panther data files) is used for final physics analyses.

Belle uses commercial software for hierarchical storage management (HSM) and batch queuing system.

Framework

Belle uses framework called BASF. It has event parallelism on SMP since 1995. It uses UNIX system call, fork as we need to copy Fortran common blocks to sub processes. Event parallelism on multi-compute servers was implemented in 2001. A new, ring buffer based version became available in 2003. In BASF, users' and reconstruction code is dynamically loaded. BASF is the only framework for all processing stages from DAQ to final analysis.

Data access methods

The design of BASF, allowed having an IO package written in C dynamically loaded at run time. However, a single IO package has grown to handle more and more methods of I/Os (disk, tape, etc.) and to deal with special situations. The code has become spaghetti-ball like. This summer we were faced again to extend it to handle new HSM and to test new GRID middleware. We finally rewrote our big IO package into small pieces, making it possible to simply add one derived class for one IO method as shown in Figure 4.

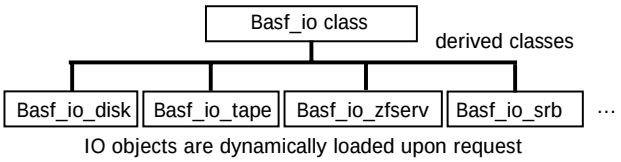

Figure 4: Basf_io class structure

Reconstruction and analysis software

30~40 people have contributed in the last several years in writing Bell reconstruction and analysis software. For many parts of reconstruction software, we only have one package. There is very little competition which is good in one sense and bad in another. We typically identify weak points and ask someone to improve them. The reconstruction software activity is mostly organized within the sub detector groups and motivated by physics. Systematic efforts to improve tracking software exist but it is a very slow process. For example, it took one year to get down tracking systematic error from 2% to less than 1%. When many problems are solved and the reconstruction software is improved significantly, we reprocess all data. Several to tens of people contribute to improve the reconstruction software. We have analysis software packages such as kinematical and vertex fitter, flavor tagging, vertexing, particle identification, event shape, likelihood and Fisher analysis. Belle members tend to use standard packages. We have started a task force (consisting of young Belle members) to improve the core software.

Database

Postgresql is the only database system Belle uses, other than simple UNIX files and directories. We keep one master database server and several replica servers at KEK, many servers at institutions and on personal PCs. The database has about 120,000 records (4.3GB on disk) at present. IP (Interaction point) profile is the largest and most popular data in the database. The management of the database is working quite well although consistency among many replicas is the problem.

Geant4

We have finally started building Geant4 version of the Belle detector simulation program as shown in Figure 5.

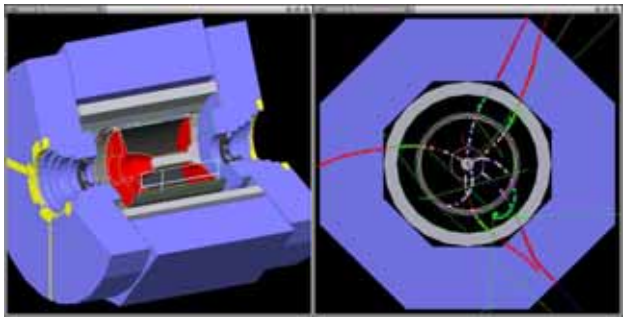

Figure 5: Super Belle Detector in Geant4

Our plan is to build the Super Belle detector (discussed below) simulation code first, then write reconstruction code for Super Belle, increasing number of people who can write Geant4 code in Belle. In one year or so, we hope to write Belle G4 simulation code and compare G4 and G3 versions and real data. Some of the detector code is in F77 and we must rewrite them.

COMPUTING

Computing system for Belle at KEK consists of two parts; KEKB rental system operated by KEK computing research center and a system Belle has purchased parts by parts and operates. The first rental system was installed in 1997 for 4 years at a cost of 2.5 billion yen (18 million Euros). The second system was installed in 2001 for five years. Due to a budget reduction the rental period was changed from four years to five years yet keeping the total budget the same (20% reduction of monthly cost!) A new acquisition process has started for the third generation rental system starting 2006.

As there is not enough computing power towards the end of the rental period, Belle has purchased PC farms, disk systems and other hardware to support Belle computing. The KEK Belle group has an yearly operating budget of about 300 million (M) yen (2.2M Euros). We spend 20~50% of the budget on computing, depending on other detector upgrades such as SVD2 and new interaction beam pipe. From time to time, we receive special allocation from the Lab management. So far we received about 200 million yen or so. The computing budgets of our collaborating institutions are quite limited. They range from almost none to 30M yen (220k Euros).

Rental equipments

Figure 6 shows the current rental system (the KEKB computer system). The nine login servers and 40 compute

servers have UltraSparc CPUs with Solaris 2.7 operating system. The UltraSparc compute servers are operated under LSF batch queuing system and are connected to 40 DTF2 tape drives in the tape library. The Solaris/ UltraSparc system is the reference platform of Belle. All Belle members have accounts. Although the login and compute servers are connected to fast disk servers, the disk servers have not enough disk space anymore to keep all Belle hadronic data sample. The system is maintained by Fujitsu SE/CEs under the rental contract.

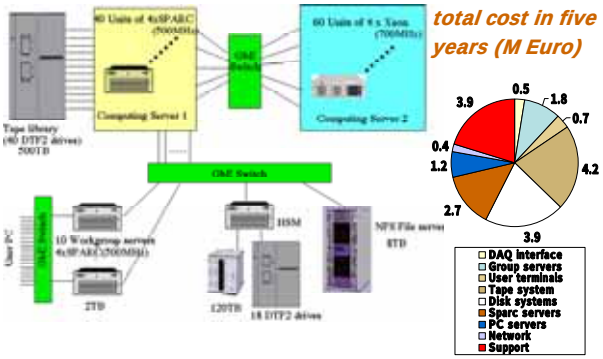

Figure 6: KEKB computer system

PC farms

Figure 7 shows the increase of Belle integrated luminosity and total number of GHz of the PC farms. We are barely catching up with the total luminosity so that reprocessing can be done in a fixed amount of time (three months as we decided). Each year, different vender wins the bid.

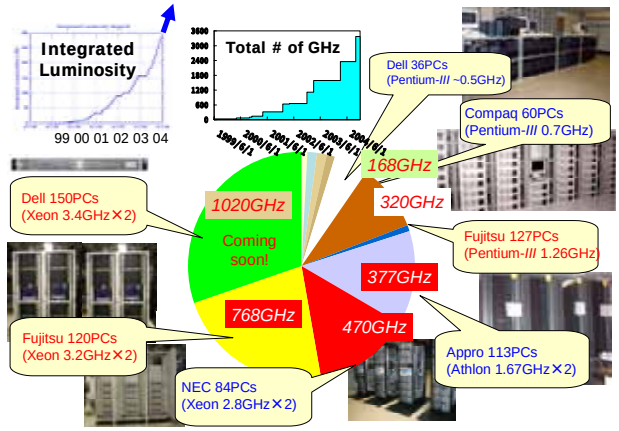

Figure 7: Belle PC farms and its acquisition history

disk servers

As a part of the KEKB computer system, 8TB NFS file servers and 4.5TB staging disks for HSM are mounted on all UltraSparc servers via GbE. About 15TB local data disks on PCs in the farms hold generic MC files. The MC files are used remotely using Belle's own TCP/IP protocol. To keep ever increasing data, we have been purchasing inexpensive IDE RAID disk systems and Linux PC file servers. We started with six 40GB IDE RAID disk system for 200GB disk space. The last few generations of the disk systems are; eight 16 160GB systems totaling about 18TB@100K Euro in 12/2002, eight 16 250GB systems, 28TB@110K Euro in 3/2003, sixteen 16 300GB systems, 56TB@200K Euro in 11/2003 and twenty 16 400GB systems, 96TB@250K Euro in 12/2004 (soon). Disk failures occur quite often and it is painful to maintain this many disk systems.

Tape libraries

Several tape libraries are in use. The first one is a direct access DTF2 tape library consisting of 40 tape drives (24MB/s each) on 20 UltraSparc servers and 4 drives for data taking (one is used at a time). This library can hold 2500 tapes (500TB). Raw data and reconstructed data (DST) are stored. The data are directly written to tapes using Belle tape IO package (allocate, mount, un-mount, free), proprietary tape management software written by Fujitsu. The second library is the HSM backend. It consists of three DTF2 tape libraries; Each library can hold 200 tapes and has 1.5TB front-end disk system. Files are written and read as if they are under an ordinary UNIX file system. When the capacity of the library becomes full, tapes are moved out of the library and human operators must insert them when the user request to read the files on out-of-library tapes. This operation is quite painful.

Accumulated data

400(700)TB of raw data (DST) have been written since 1999 on 2000(4000) tapes as shown in Figure 8. The amount of mini-DST (hadronic events only, four vectors and errors and PID information) is about 15 TB for 287fb^{-1}. τ and two photon events are about 9TB.

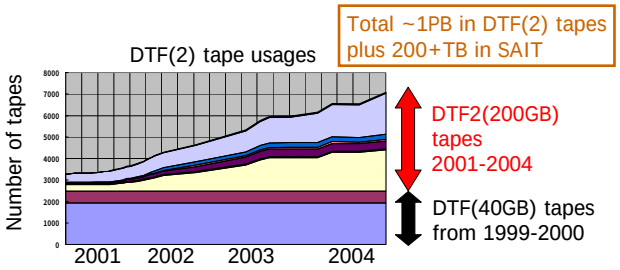

Figure 8: Number of tapes written as a function of time

Mass storage strategy and new HSM system

It was a bad news that the development of next generation DTF drives was canceled by Sony. Sony's new mass storage system will use SAIT drive which utilizes metal tape and helical scan technologies. We decided to test it and purchased a 500TB tape library. Experience with the new HSM system is described in detail[1]. It was installed as the HSM backend of inexpensive IDE RAID systems and PC file servers as described below. We hope to be moving from direct tape access to hierarchical storage system as we have learned that automatic file migration is quite convenient. However, we need a lot of capacity so that we do not need operators to mount tapes.

Batch Queuing System: LSF

Belle has been using LSF since 1999 on Solaris. We started using LSF on PCs since 2003/3. Of ~1400 CPUs

(as of 11/2004), ~1000 CPU are under LSF. CPUs are used for DST production, generic MC generation, calibration, signal MC generation and user physics analyses. DST production uses its own distributed computing (dbasf) and child nodes do not run under LSF. All other jobs share the same CPUs and dispatched from LSF. For users' jobs we use fair share scheduling. LSF is quite effective to make an efficient use of our PC farms which are diverse and different as they have been purchased in several years. We will be trying to evaluate new features in LSF 6.0 such as report and multi-clusters hoping to use it a collaboration wide job management tool.

New rental system

We have started a lengthy process of computer acquisition for the period of 2006–2010. We hope to obtain more than 100,000 specCINT2000_rates of compute power at the beginning, storage of 10PB which is extendable to several tens of PB, fast enough network connection to read and write data at the rate of 2–10GB/s (2 for DST, 10 for physics analysis). The system should be user friendly and should have efficient batch queuing system that can be used collaboration wide. The system should be "Grid capable." In particular, the system should interoperate with systems at remote institutions which are involved in LHC grid computing. We must make a careful balancing of lease and yearly-purchases.

PRODUCTION

On-line production

As we take data, we write raw data directly on DTF2 tape. At the same time we run reconstruction code using 85 dual Athlon 2000+ PC servers [2] as shown in Fig. 9.

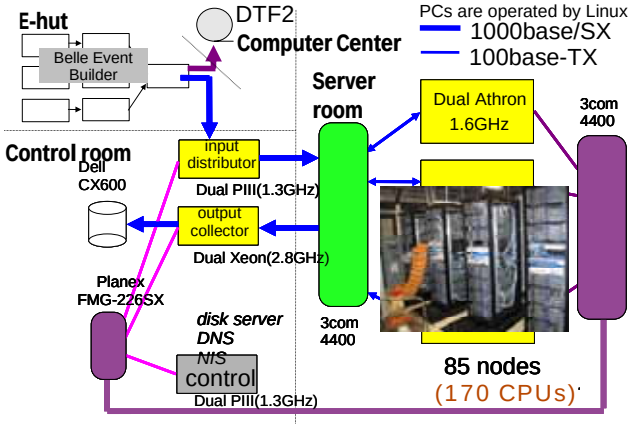

Figure 9: Schematics of online production scheme

Figure 10 summarizes the data characteristics. Using the results of event reconstruction we send feedback to the KEKB accelerator group such as location and size of the collision so that they can tune the beams to maximize instantaneous luminosity, keeping them collide at the center as shown in Figure 11. We also monitor B anti-B production cross section very precisely and change the machine beam energies by 1MeV at a time in order to maximize the number of B anti B pairs produced. The resulting DST are written on temporary disks and skimmed for detector calibration. Once detector calibration is finished, we run the production again and make DST/mini-DST for final physics analysis.

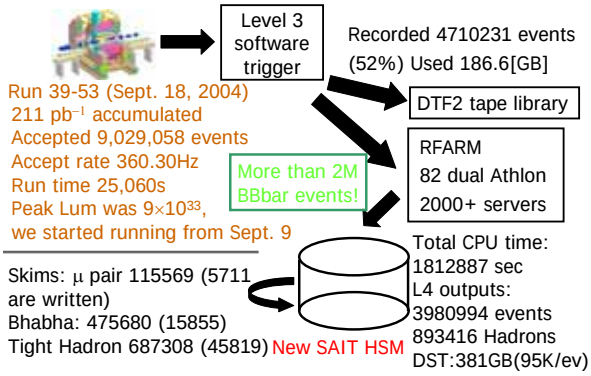

Figure 10: Data size, trigger rates and other numbers for a typical run

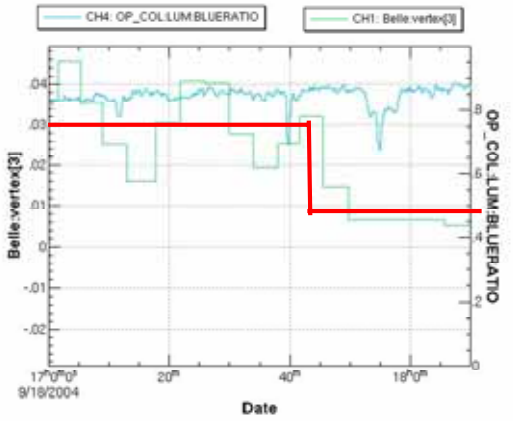

Figure 11: Red line shows the knob that the KEKB accelerator group turned to move the interaction point. Green line shows the actual interaction point measured using the online data processing as a function of time.

DST production

Belle's reprocessing strategy is as follows;

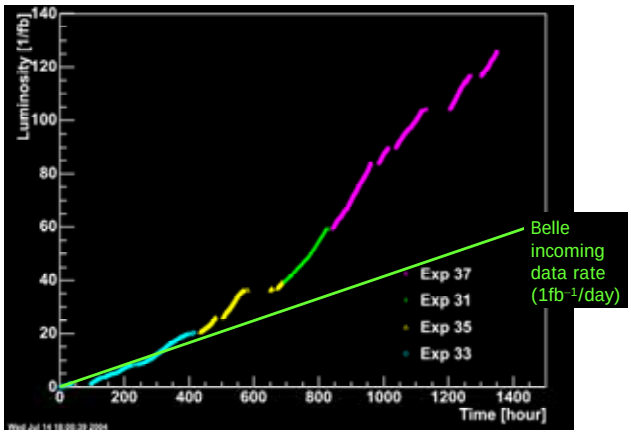

Figure 12: DST production as a function of time (hours)

When we have substantial improvements in reconstruction software and constants, do a reprocessing of all data in three months using all KEK compute servers.

Figure 12 shows the history of the last processing. We have achieved more than 5fb^{-1}/day performance on the best day. The efficiency is between 50% and 70% as we often have to wait for constants to be made and to restart due to bad constants. In 2002, before summer, we had a major software update and reprocessed the entire data set taken before July 2002. In 2003, we reprocessed data we took between October 2002 and June 2003. In 2004 as we upgraded the silicon vertex detector in summer 2003, we reprocessed all data taken since October 2003 and June 2004 in three months (May to July 2004).

Generic MC production

The so called generic MC is mainly used for physics background study. With 400GHz Pentium III we can generate 2.5fb^{-1}/day. It takes 80~100GB/fb^{-1} of disk space without raw and GEANT3 hits. When a new release of the library comes out, we try to produce new generic MC sample. For every real data taking run, we try to generate 3 times as many events as in the real data. Detector backgrounds are taken from random trigger events of the run being simulated. If we use all CPUs at KEK we can keep up with the data taking and the generic MC production. However, when we do reprocessing, we use up all KEK CPUs and must ask remote institutions to generate most of MC events. We generated more than 2×10^9 events using QQ event generator. We are switching to EvtGen generator. It requires 100TB of disk space to keep for 3×300 fb^{-1} of generic MC data. Figure 13 shows generic MC production at remote institutions.

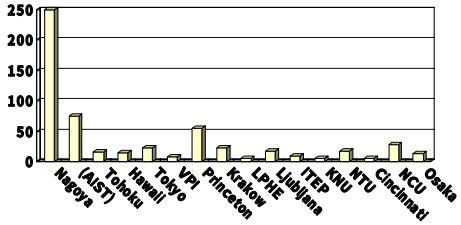

Figure 13: generic MC events generated at remote institutions since between April and July 2004.

Random number generator

The management of pseudo random number can be difficult when we must generate billions of events in many hundred thousands of jobs at many institutions. Each event consumes about one thousand random numbers. Some seeds can have very short periods. Encoding the date and time as the seed may be one solution but there is always a danger that one sequence becomes the same as another. We tested a random generator PCI board which uses thermal noises as the source of random number. The board can generate up to 8M real random numbers in one second. We will have several random number servers so that we can run many event generator jobs in parallel. Figure 14 shows the random number generator built by Toshiba.

Figure 14: PCI random number generator board

NETWORK AND DATA TRANSFER

KEKB computer system has a fast internal network for NFS. To this fast internal network, we have kept adding Belle bought resources; more than 500 compute servers and 24 file servers. We have also connected Super-SINET dedicated 1Gbps lines to four universities to this internal network. We requested that we connect this network to outside KEK for tests of Grid computing.

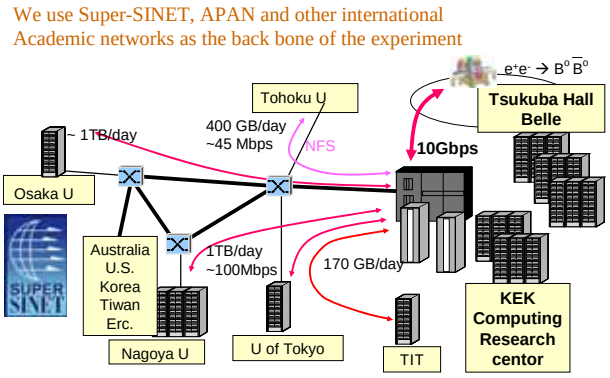

Figure 15: Schematics of Belle's network connection to collaborating institutions

As this network becomes complicated and unmanageable, we added a new Cisco 6509 to separate the above network into three networks. At the KEKB computer, a firewall and login servers make the data transfer miserable (100Mbps max.). DAT tapes are used to copy compressed hadronic event files and MC data generated by outside institutions. As shown in Figure 15, dedicated GbE network to four collaborating institutions as well as 10Gbit backbone network connection to/from KEK were added, thanks to the Super-SINET project. However, we suffer from slow network connection to most of collaborators.

Belle' Grid plans

Belle hasn't made commitment to any Grid technology. Belle@KEK has not thought about it very seriously. Belle@remote institutions might have already been involved in Grid activities, in particular, when, at the same time, they are involved in one of the LHC experiments. As to Grid's parallel/distributed computing aspect, it is nice but we do have our own solution. We use

event (trivial) parallelism and it works fine. However, as we accumulate more data, we may need to have more parallelism (several tens to hundreds to thousands of CPUs) and thus we may need to adopt a standard solution. We have separated stream IO package so that we can connect to any of (Grid) file management packages. We started working with remote institutions in Australia, Japan, Taiwan, and Korea) and other computing centers[4]. We have started testing SRB (Storage Resource Broker by San Diego Univ.) We constructed test-beds connecting Australian institutes and Tohoku university using GSI authentication[5]. We are using gfarm (cluster) at AIST to generate generic MC events with the gfarm software[6].

SUPER KEKB UPGRADES

One of our goals is to establish the CP violation via Kobayashi-Maskawa mechanism. However it is clear that we need something beyond the Standard Model. We may be seeing the first evidence in B decays. Figure 16 shows the roadmap of B factories.

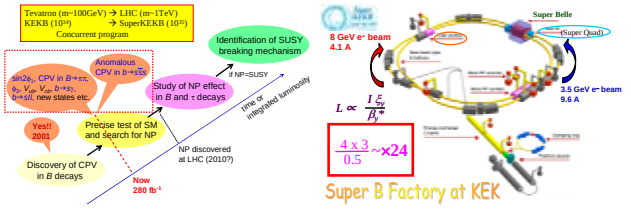

Figure 16 and 17: Road map of Super B factories and schematic view of KEKB upgrade plans

To see such evidence, we need many more B meson decays. As shown in Figure 17, the super KEKB upgrade will increase the luminosity by a factor of 10–24, making the instantaneous luminosity to $1–5\times10^{35}\mathrm{cm}^{-2}\mathrm{s}^{-1}$.

With this much luminosity, minimum of DAQ event rate of 5 kHz (B anti-B pair production at a rate of 100–500Hz) and event size of 100KB/event are expected, making the data rate of 500MB/s. This corresponds to 10^{15} bytes/year. 800 4GHz CPUs are required to catch up with the data taking. For reprocessing and generic MC production, more than 2000 4GHz 4CPU PC servers and 10+PB storage space are necessary. The mini-DST will exceed 300TB–1PB MDST/year. Figure 17 shows the Super Belle detector plan.

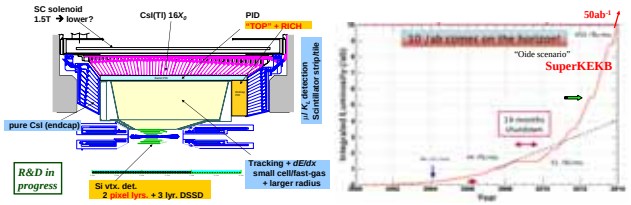

Figure 18 and 19: Super Belle detector upgrade plan and projected integrated luminosity

With Super KEKB and Super Belle upgrades, the integrated luminosity is predicted to be as follows, shown in Figure 19. With $5\mathrm{ab}^{-1}$, if there is an effect from super symmetry in "$\sin2\phi_1$" measured in $B\to\phi K_s$, we should see the effect at more than 6 sigma significance, assuming that the central value remains the same as the 2003 world average as shown in Figures 20. The letter of intent has been written[7] in which we discuss physics cases, detector and accelerator upgrades in detail.

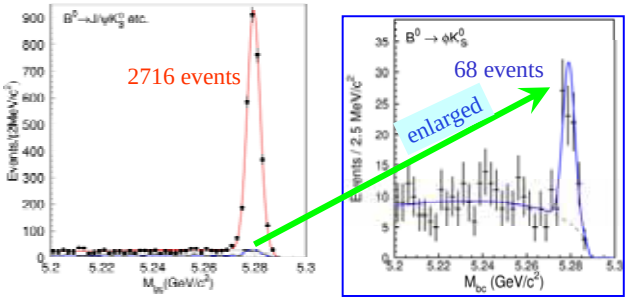

Figure 20: With 100 times more data, ϕKs signal is as strong as J/ΨKs now

CONCLUSIONS

Belle has accumulated 275M B anti-B pairs by July, 2004; data are fully processed and everyone is enjoying doing physics (tight competition with BaBar). A lot of computing resource has been added to the KEKB computer system to handle floods of data. The management team remains small (about 5FTEs, two of the KEK staffs involved are group leasers of sub detectors!) and less than 5 SE/CEs. In particular, the PC farms and the new HSM are managed by a few people. We look forward to Grid solutions at KEK and at the remote Belle collaborating institutions so that the management of computer resources of the Belle collaboration can be done globally with minimum number of people involved.

REFERENCES

[1] N. Katayama et. al., The new compact HSM system (contribution to this conference).

[2] R. Itoh et. al., The new online DST production system at Belle (contribution to this conference).

[3] F. Ronga et. al., Belle's offline DST production (contribution to this conference).

[4] G. Molony et. al., Belle's Grid efforts (contribution to this conference).

[5] Y. Iida et. al., SRB at Belle/KEK (contribution to this conference)

[6] O. Tatebe et. al., gfarm (contribution to this conference)

[7] Super KEKB Letter of Intent (http://belle.kek.jp/superb/loi)