THE INTRODUCTION TO BES COMPUTING ENVIRONMENT

XU Dong, CHENG Yaodong, CHEN Gang[#], YANG Yi, YANG Dajian
Institute of High Energy Physics, Beijing 100039, China

Abstract

Beijing Electron Spectrometer (BES) is the experiment on Beijing Electron-Positron Collider (BEPC) started in 1980s. The computing facilities were built on VAX and HP-UX and upgraded to PC/Linux between 1995 and 1998. From 2004, BEPC started to be upgraded. The luminosity will increases by 100 times. The new detector BESIII is being built and will be operational in 2007. The paper describes the computing environment for the current BESII experiment and future BESIII experiment.

INTRODUCTION

BES is a detector collecting data in the energy region between 2 to 5 GeV on BEPC. The main task of BES focuses on τ physics. BES experiment started to collect data from 1989. During the period of 1995 to 1998 BES was upgraded first time to BESII. From the May of 2004, BES is being upgraded again to BESIII. The computing environment for BES was initially built on VAX/VMS and HP/UX systems. From middle of 1990s the environment moved to PC/Linux/Cluster system. In this work, the current computing system for BESII is introduced, and proposed system for upcoming BESIII will be discussed.

BESII COMPUTING ENVIRONMENT

BESII Computing Requirements

BES experiment collects data of J/ψ, ψ' and ψ''. BES data consist of raw data, reconstructed data, summary data in DST format, and Monte Carlo simulation data. Table 1 shows the size of real data. A raw event has a size of 2 KB, a reconstructed event has 6 KB, and a DST summary event has 0.6 KB. A physics run has approximately 30000 events. BES experiment collected about 5000 runs till the May of 2004. Monte Carlo simulated data have the same structures. The total real and Monte Carlo data generated by BES and BESII are about 13 TB.

Table 1: Size of BES experiment real data

	Event Size (KB)	1 Run ~3×10^4 events (MB)	~5000 Runs (GB)
Raw data	2	60	300
Rec. data	6	180	900
DST data	0.6	18	90

To manage and analyse the data, a computing system with enough computing and I/O capacities are needed.

BESII Computing Environment

BESII computing environment is running on PC/Linux platform, with free software and physics analysis tools from HEP laboratories such as CERN.

Computing clusters

The layout of the BES computing environment is illustrated in Figure 1. All computing nodes and servers are interconnected through a Cabletron E5 switch and a CISCO switch. There are six file servers used to optimize I/O load balance. These servers are equipped with Pentium III and Pentium IV CPUs and 2 GB memory respectively. There are totally about 7.3 TB disk arrays attached to the file servers. To improve the I/O performance an extra gigabit network adapter is installed on each file server, dedicated to NFS data transferring.

Each computing node is a DELL PC with Pentium IV 2.6GHz or 2.8GHz CPU and 1 GB memory. There are 64 computing nodes. File systems on file servers are mounted to all computing nodes using autofs.

Software platform

The BESII computing system is built mainly on free software platform. RedHat 7.3 is used as operating system. Users' accounts are managed using NIS/YP and AFS. BESII software is written in Fortran and C, common Fortran and C libraries are installed on the system. Also CERNLIB and BESLIB are installed as basic physics analysis tools. To manage batch jobs, OpenPBS is used.

BESII stopped taking data from last May. The detector is being upgraded. BESII computing environment is currently used to analyse the data collected in the previous operation. A new computing facility is being built for the new generation of the experiment.

BESIII COMPUTING ENVIRONMENT

Based on the existing BESII computing environment, following points should be taken into account for the future BESIII offline computing system and software environment: the system should be set up by adopting or referring to the latest technology commonly used in HEP community, both in hardware and software, in order to benefit the collaboration and to have easier exchanges

[#] Gang.Chen@ihep.ac.cn

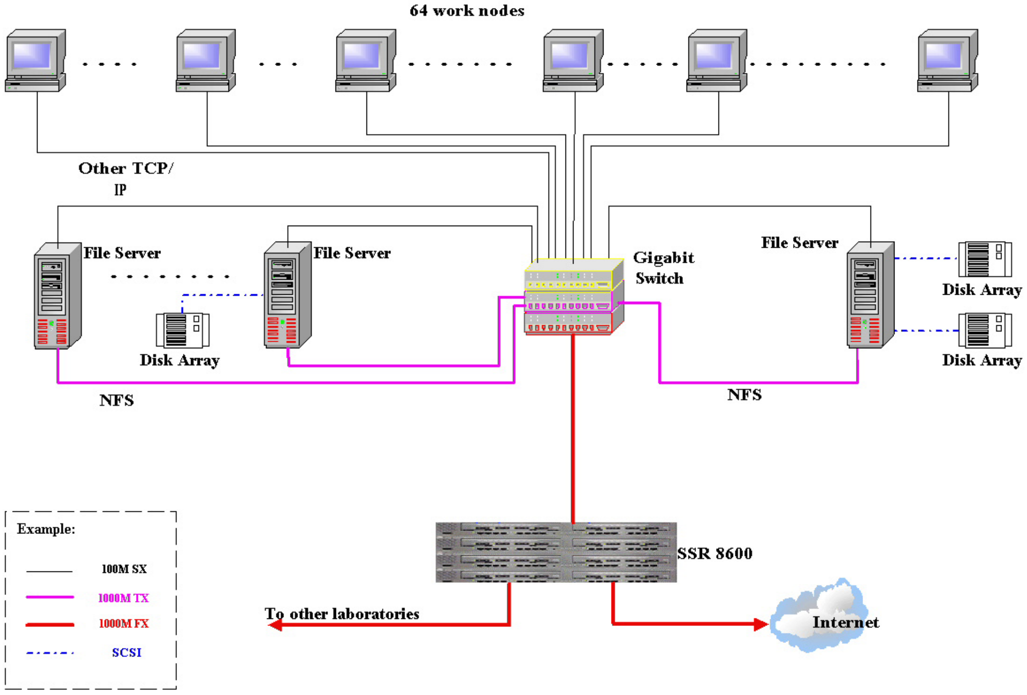

Figure 1: Topology of BESII computing environment.

with other experiments.

The system should support the existing BES software packages and should serve for both experts of the BESII software and new members in the collaboration.

Many of the BESII packages will be modified or re-designed to suit the new computing environment.

BESIII Storage Requirement

The peak luminosity of the BEPCII at the J/ψ resonance will be about 10^{33} cm $^{-2}$ s $^{-1}$. The event rate recorded is estimated to be about 3000 Hz. and the event size is about 12 Kbytes/event for raw data, 24 Kbytes/event for reconstructed data (Rec.) and 2 Kbytes for summary data (DST).

BESIII will take J/ψ data at the beginning of the data taking for one year or more, and then move to ψ' energy region. So the maximum data yields per year is about 1×10^{10} J/ψ data. The total data size in first year is $12\times10^{3}\times1\times10^{10}\approx120\times10^{12}$ bytes.

The total amount of raw data in 5 years is estimated to be about 240×10^{12} bytes, which include 120 Tbytes of J/ψ data and 120 Tbytes of ψ', D and Ds data. Suppose the data reconstruction is repeated four times per year, the total size of the Rec. and DST data will be about 1440 Tbytes and 120 Tbytes respectively. The size of Rec. and DST data from Monte Carlo simulation will be about the same as that of real data.

All raw and reconstructed data, about 3120 Tbytes, can be put on the tape library. A total of 240 Tbytes DST data will be stored on disk array accessible via high-speed network system. Details are listed in table 2.

Table 2: Storage Requirement for BESIII

Data Type	Data Volume (Tbytes)	Storage Media
Raw	240	Tape
Rec.	1440	Tape
DST	120	Disk
M.C. Rec.	1440	Tape
M.C. DST	120	Disk

CPU Requirement

According to the experience of data processing at BESII, required CPU power for data reconstruction is about 20s×MIPS per event. Suppose the total active running time of the computer is about 2×10^{7} seconds per year, and the data reconstruction is repeated three times a year to improve calibration and reconstruction, the required CPU power is about 130000 MIPS. Details are listed in table 3.

Bandwidth for Data Transfer

The bandwidth required for data transfer from the online computing system to the offline data servers should be more than 400 Mbps, which is determined by the product of trigger rate times the event size, i.e.

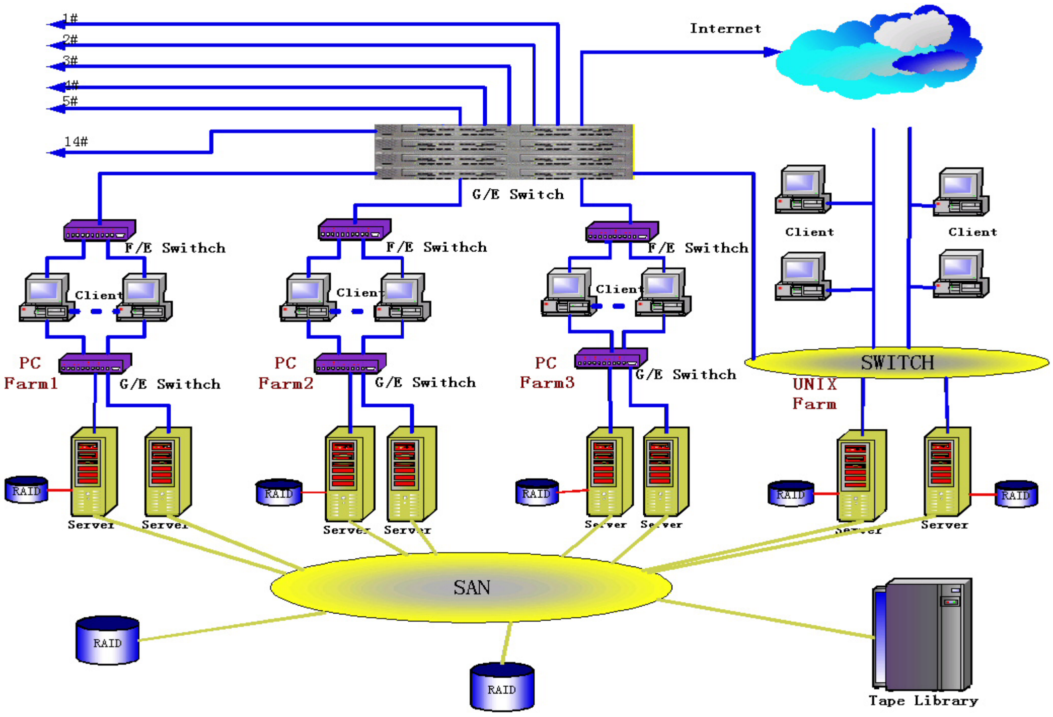

Figure 2: The scheme of the BESIII computing system

Table 3: CPU requirement for BESIII

Job type	Speed/Event (MIPS×s)	Total event (10^{10})	Total CPU (MIPS)
Data Rec.	20	4	40000
MC Sim.	100	1	50000
MC Rec.	20	4	40000
Total			130000

4000×12Kbytes×8. It also requires that the network system should be highly stable and secure to avoid event losses.

The bandwidth required for data transfer from the data servers to the reconstruction (data production) farm depends mainly on the processor speed of selected machines and the number of jobs. It is necessary to create a dedicated BES network environment, and ensure a reasonable efficiency in data transfer.

Computing Environment

The main tasks of the BESIII computing environment are data reconstruction and analysis; data transfer, data storage and communication.

To satisfy these requirements, the system to be built should have high performance, better stability, reliability and flexibility, with a reasonable and acceptable cost. Also the rapid development in both computer hardware and software should be followed closely so that we can benefit from the latest development of technology. Especially a high-speed network is essential for mass storage system, such as tape library and disk array. Figure 2 shows a preliminary scheme of the computing system for the BESIII. The main considerations are the following:

CPU type and architecture: A high performance computing system based on PC/Cluster or PC/Grid technology will be built. The CPU type can be any or all of Intel, AMD or IA64.

Data storage: The BESIII Storage System will adopt the technology of the Disk Array and Tape library with HSM (Hierarchical Storage Management). A SAN (Storage Area Network) could be a candidate to satisfy the requirement of large amount data storage, high access speed and scalability.

Network and I/O control: In order to improve the data access speed and to reduce the interference, a dedicated network based on SAN could be adopted to separate data transfer and normal network traffic. In addition, all servers will have both 100TX and 1000TX network cards, in which 100TX provides traditional TCP/IP services while 1000 TX provides NFS services.

System software: The BESIII offline computing system will mainly use free software to reduce the cost and to have an easier exchange with other experiments in the world. The main components are the following: RedHat/Linux as the operating system; Oracle or MySQL or PostgreSQL as database system; PBS for the batch system, LSF also be considered; AFS for user management and document management.

Offline Data Analysis System

The main task of the BESIII software system is to convert raw data of the detector responses into physics results. It consists of a main framework, the data reconstruction and calibration package, the Monte Carlo simulation of physics processes and detector responses, the database management and interfaces, various utility packages, and user's physics analysis packages. It should also manage documents, software codes and libraries. The system should take advantages of the Object Oriented technology by using the C++ language, while keeping the possibility to incorporate some of the existing BES Fortran software packages. The system should also take into account practical needs, such as usability, stability and flexibility, and to accommodate conflicting needs between experts and novices.

Framework of BESIII offline software

In order to take advantages of the latest technology and utilize common tools from other HEP experiments in the world, the main framework of the BESIII offline software will be based on the Object-Oriented methodology and C++ language, and take into account the following points: it should support some of the existing BES packages written in Fortran language; it should use as much as possible existing HEP libraries; it should provide a uniform data management, code and library management, and database access.

The BESF as the BESIII software framework, based on the Belle analysis Framework (BASF)[1], has been developed in the summer of 2003. The major packages of the BESF framework are shown in Fig. 3, in which the BesKernel is the core part of the framework that implements the control on data processing. It depends on other four packages: the EventIO package managing event input and output, the UserInterface package providing user friendly interface for running jobs, the Panther[2] package that is an integral data management system and the BesEvent package implementing the interface to the ProxyDict originally developed in the Babar experiment. The ROOT and CERNLIB are the only two external libraries needed by Histogram package.

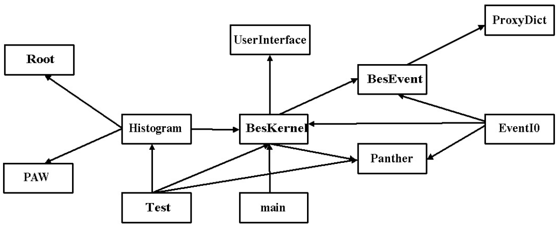

Figure 3: Software packages and dependencies in the BESF.

Calibration and reconstruction

Most of the sub-detectors of the BESIII are different from that of BESII, therefore the calibration and reconstruction codes will mostly be re-written. Whether it is written in C++ or in Fortran, the software system should have a well separated calibration and reconstruction sequence, with a modular structure so that any changes of an intermediate step will not result in modifications of related code in a later stage. If C++ is adopted, some of the objectivity should be compromised, for example, data and operation should be well separated.

Monte Carlo simulation

Most of the event generators of the BESII can be re-used although some modifications may be needed. The simulation of the detector response will be a new package based on the GEANT4 program while a Fortran code based on the GEANT3 program in Fortran will be kept as a backup and for comparison.

The BESIII simulation packages based on Geant4, BOOST, consists of three parts, the event generator, the particle tracking and the detector response. The XML language will be used for the detector description. The raw data format is used for the final output of BOOST. Right now, the hit information from most sub-detectors can be used to test or tune the offline reconstruction program.

Tools and Libraries

Commonly used CERN libraries, both in C++ and in Fortran, will be used extensively. Physics analysis will be based on HBOOK, PAW, PAW++, ROOT, MN_FIT, and so on.

Some of the BESII libraries in Fortran, such as Telesis for kinematical fitting, events vertex fitting and event-kink fitting can be re-used.

The database of the BESIII contains the detector geometry, calibration constants, detector running status and conditions, environment parameters, etc. Some of the tables in the offline database are kept identical with that of the online database while some other tables will only appear in one of the two databases. The database will be managed by Oracle or a free software based on SQL language, such as PostgreSQL and MySQL.

Commercial software packages can also be used, as long as it is well received by the HEP community. For example, the software code will be managed most likely by CVS, RCVS, AFS or DFS and so on.

ACKNOWLEDGEMENTS

The author gratefully acknowledges the support of K. C. Wong Education Foundation, Hong Kong.

REFERENCES

[1] Itoh, R., BASF - BELLE AnalysiS Framework, Talk given at Computing in High-energy Physics (CHEP 97), Berlin, Germany, 7-11 Apr 1997

[2] Shojiro Nagayama, Panther User's guide version 3.0