BES Physical Analyzing Environment Composed of PC Farm

Tianrong Liu, Zepu Mao, Huaimin Liu

Institute of High Energy Physics, Academia Sinica, P.O.Box 918, Beijing 100039, China

June 7, 2000

- Introduction
- Structure of BESPCFARM
- Probation Status and Performance
- Check of BES Reconstruction Codes V102 in Linux and HP-UX
- Existing Problems and Improvements
- Outlay

1 Introduction

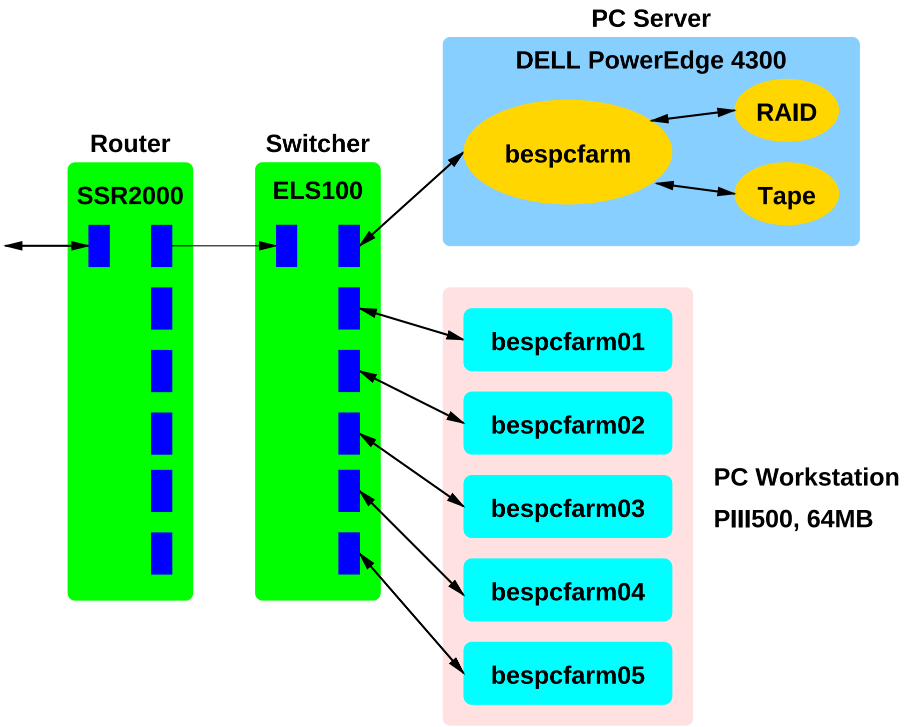
We will collect 50×10^6 j/ψ data within recent two years. Yet the total CPU power is only about 50 SPECINT95 (6×4+2×4.4+17) up to the end of 1998. It is obvious this CPU power is not enough to handle so much j/ψ data. After discussing with experts and leaders of the institute, we determine to increase one HP workstation for data reconstruction in the computer center, and create a physical analyzing platform consisting of PC in hall No.3.
Usually a PC farm is made up of server, workstation, data storage device and network communication system. Server is characterized by good stability, high speed and expanding capability. It is provided with fast network card, Redundant Array of Independent Disks (RAID), multi-CPUs, high-capacity memory and redundant power supply.

2 Structure of BESPCFARM

2.1 Hardware Environment

BES physical analyzing platform is set up on a PC farm called BESPCFARM. As shown in the following figure and table, BESPCFARM is composed of one server with RAID and tape driver and five workstations. They are connected by a 100M switcher.

BESPCFARM Architecture



PC Server	DELL PowerEdge 4300 PIII500 CPU×2, 256MB SDRAM, 17GB SCSI HD
RAID	36GB×6, RAID Mode 5
PC Workstation	PIII500 CPU, 64MB SDRAM, 13GB IDE HD
Switcher	Cabletron SmartStack ELS100-S24TX2M
Tape Driver	EXABYTE EXB-8505ST

2.2 Software Environment

- **Operating System:**
  The operating system of BESPCFARM is Red Hat Linux 6.1 which is obtained from Internet. It will be upgraded to FERMI Lab. Linux 6.1 which has better safety.

- **Configuration of File and User:**
  User home directory, ndst data file, BES library, CERN library and applications are stored on RAID of server. The directory structure is similar to the one on HP workstations. RAID and users are installed onto workstations through Network File System (NFS) and Network Information System (NIS). Each workstation has the same software environment as the server.

- **BES Library:**
  The current BES version on BESPCFARM is V102. It is based on V102 on HP workstation, and converted into Linux format, then compiled on BESPCFARM to produce the library. To make BES codes suitable for Linux system, during the process of the upgrade of BES V101 to V102, format and algorithm of some codes are modified. With regard to the details of this section, please see the relevant reports of teacher Mao zepu[2].

- **CERN Library and Applications system:**
  The current version of CERN library is 97a. It is downloaded from CERN Website. Other utility and tool packages are also obtained from CERN.

- **Data File Management:**
  The data on BESPCFARM are ndst data now. They are updated automatically from HP workstations at night.

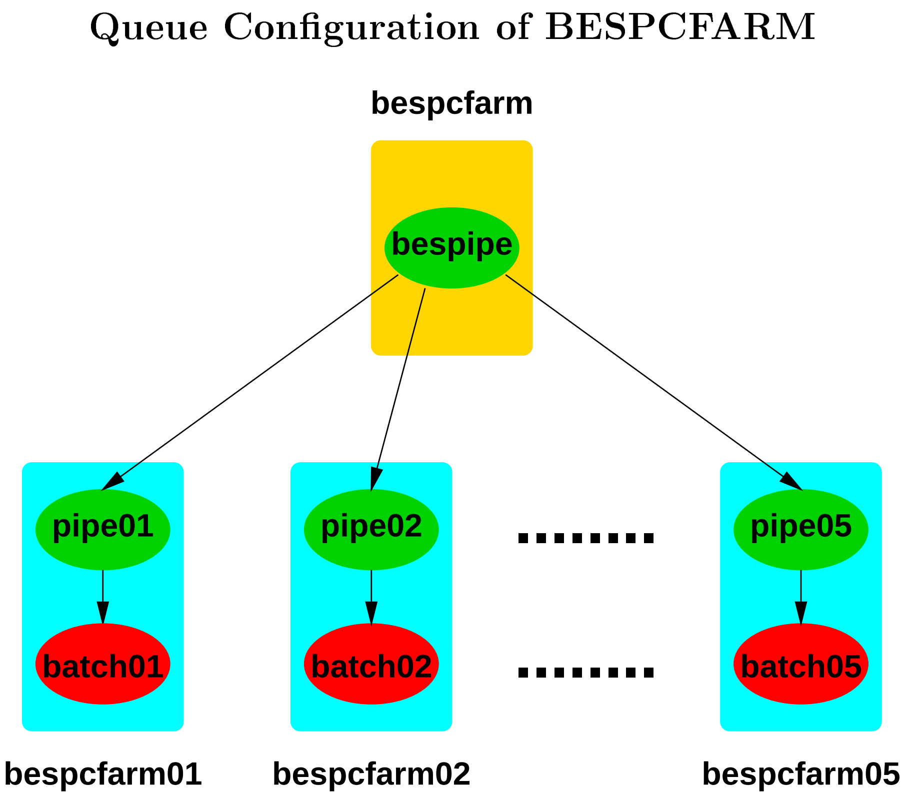
2.3 Network Queueing System (NQS)

2.3.1 Introduction

NQS is the continuing development of Monsanto-NQS, itself descended directly from the original COSMIC NQS, written under contract to NASA by Sterling Software, Inc. NQS incorporates enhancements from Boeing, CERN, and a large number of individual contributors.
NQS allows user to submit batch jobs to queues on local or remote machines, and have the log file returned to the originating machine or another machine. After submitting the job, the user can watch the status of the queue. The user can also affect the job after submitted by holding a queued job from being scheduled, releasing a held job, suspending a running job, resuming job and so on. NQS can only distribute a complete job to one workstation rather than separate it to each workstation.

2.3.2 Working Principle of NQS

NQS use TCP/IP protocol to communicate on network. When it starts, it runs two daemons. One daemon named NQSDAEMON is responsible for handling local jobs and the other named NETDAEMON is for communication on network. There are two main types of queues in NQS: batch and pipe. A batch queue is an execution queue where the job actually runs. A pipe queue provides routing capabilities. When a job is submitted to a pipe queue, it is passed on to another pipe queue for further routing or a batch queue for execution on the same or another machine.
The jobs submitted by users are sent to bespipe queue on server first. Then NQS distributes them to workstations according to the load balance algorithm. Jobs are run on workstations and results are sent back to server.



2.3.3 Detailed Operations

Some useful commands and simple examples are shown below:

- "drunk -v102 -r test": submit a job called test
  Jobs are handled by NQS when using -b or -r options under DRUNK or SOBER.
- "qstat -d": examine the status of jobs submitted by a user.
- "qstat -ad": examine the status of jobs submitted by all users.
- "qdel 16": delete No.16 job waiting in "bespipe" queue on server.
- "qdel -k 12@bespcfarm01": delete No.12 job running on the workstation called bespcfarm01.

3 Probation Status and Performance

3.1 Probation Status

BESPCFARM has worked for about two months and distributed about 1000 jobs. There are 38 users and 70GB NDST data from run13500 to run17018 on it. We found some problems in system operation. Some of them have been solved and the others need studying further.

3.2 Performance

The comparison of the CPU power between BESPCFARM and HP workstation has been made roughly. The preliminary results are listed below.

Item	BESPCFARM	HPWS8
Physical Analysis j/$\psi \to \Phi\pi\pi$ (100 runs)	398s	1146s
MC Production $e^+e^- \to j/\psi \to e^+e^-$ (5000 events)	559s	1889s
Data Reconstruction run15254.raw (10000 events)	475s	1496s
Equivalent of CPU (SPECINT95)	51×5	17

4 Check of BES Reconstruction Codes V102 in Linux and HP-UX

The real data run15254 has been reconstructed on HPWS8 and BESPCFARM under BES V102 and corresponding constants. We found there are some differences between the results obtained from the two systems. We can divide these differences into two kinds.
The first one: If the track numbers of a subtype in an event obtained from two machines are different, that is called TRK DIF.
The second one: If the values of the parameters of the subtype obtained from two machines are different, that is called PAR DIF.

4.1 Comparison of Track Number for Each Subtype

The Fig.1 shows TRK DIF.
The first plot is the absolute events number with different track number virus the subtypes, the second plot is the percentage ratio of the events with different track number virus subtypes. From this plot, we can see the maximum difference is about 15% on subtype 4 (TOF). The differences of other subtypes are less than 4%.

4.2 Comparison of Parameters for Subtype 3

The Fig.2 shows PAR DIF for MDC.
The first plot is the absolute PAR DIF track numbers virus the parameters of the subtype 3. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 3. From this plot, we can see the maximum difference is about 80%.

4.3 Comparison of Parameters for Subtype 4

The Fig.3 shows PAR DIF for TOF.
The first plot is the absolute PAR DIF track numbers virus each parameter of subtype 4. The second plot is the percentage ratio of PAR DIF track numbers value virus the parameters of subtype 4. From this plot, we can see the maximum difference is about 25%.

4.4 Comparison of Parameters for Subtype 5

The Fig.4 shows PAR DIF for SHC (BSC and ESC).
The first plot is the absolute PAR DIF track numbers virus the parameters of subtype 5. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 5. From this plot, we can see the differences of all parameters are less than 6%.

4.5 Comparison of Parameters for Subtype 6

The Fig.5 shows PAR DIF for Muon Counter.
The first plot is the absolute PAR DIF track numbers virus the parameters of subtype 6. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 6. From this plot, we can see the difference of parameter 2 is less than 1.5%.

4.6 Comparison of Parameters for Subtype 7

The Fig.6 shows PAR DIF for dE/dX Information.
The first plot is the absolute PAR DIF track numbers virus the parameters of subtype 7. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 7. From this plot, we can see the differences of some parameters are 100%.

4.7 Comparison of Parameters for Subtype 11

The Fig.7 shows PAR DIF for Projected Coordinates of MDC Tracks.
The first plot is the absolute PAR DIF track numbers virus the parameters of subtype 11. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 11. From this plot, we can see the differences of most parameters are from 15% to 60%.

4.8 Comparison of Parameters for Subtype 13

The Fig.8 shows PAR DIF for Vertex Chamber (VC) One-Track Fit.
The first plot is the absolute PAR DIF track numbers virus the parameters of subtype 13. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 13. From this plot, we can see the differences of the parameters is from 1% to 4%.

4.9 Comparison of Parameters for Subtype 15

The Fig.9 shows PAR DIF for Projected Coordinates of VC Tracks.
The first plot is the absolute PAR DIF track numbers virus the parameters of subtype 15. The second plot is the percentage ratio of the PAR DIF track numbers virus the parameters of subtype 15. From this plot, we can see the differences of parameters are all above 25%.

5 Existing Problems and Improvements

- **Blocking of NQS Queue:**
  Sometimes NQS may stop distributing jobs and seems "bespipe" queue is blocked by some job. In this status, the only thing that can be done is to send an "abort" command to "bespipe" queue, then the blocking job is deleted and NQS resume distributing jobs. The current way to solve this problem is writing a program to monitor the running status of NQS. When blocking occurs, the blocking job is resent to "bespipe" queue in a higher priority, and then the "abort" command is sent to "bespipe" queue.

- **Low Efficiency of Network:**
  The running speed of some jobs is limited by the low efficiency of NFS and insufficient speed of network. We are seeking a more efficient way to transfer data to workstations.

- **Classification of Queue:**
  Classifying queue to short, medium and long types can improve the calculating efficiency of workstations under usual circumstances. But it will increase the difficulty of submitting jobs by users.

- **Monitoring of System Status:**
  A software is needed to monitor the status of BESPCFARM and inform the administrator when problems occur.

- **Further Check on Software of the Reconstruction Part:**
  The reconstruction results under Linux and HP-UX should be compared carefully and BES source codes should be modified further to eliminate the differences of the results.

- **Necessity of More Workstations:**
  Based on the operation of BESPCFARM recently, five workstations are not enough to satisfy users. We plan to increase 5-10 PC workstations. So BESPCFARM can run 10-15 jobs simultaneously.

6 Outlay

PC Server	:	~35,000 RMB
RAID	:	~43,000 RMB
PC Workstation	:	~40,000 RMB
Switcher	:	~17,000 RMB

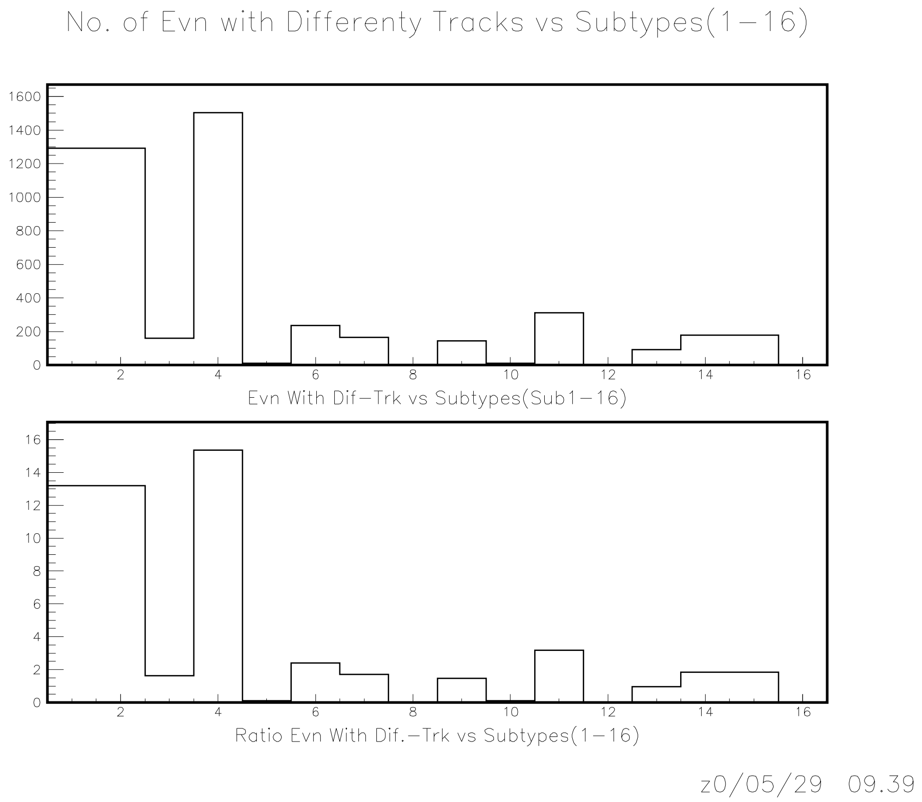


Figure 1: TRK DIF

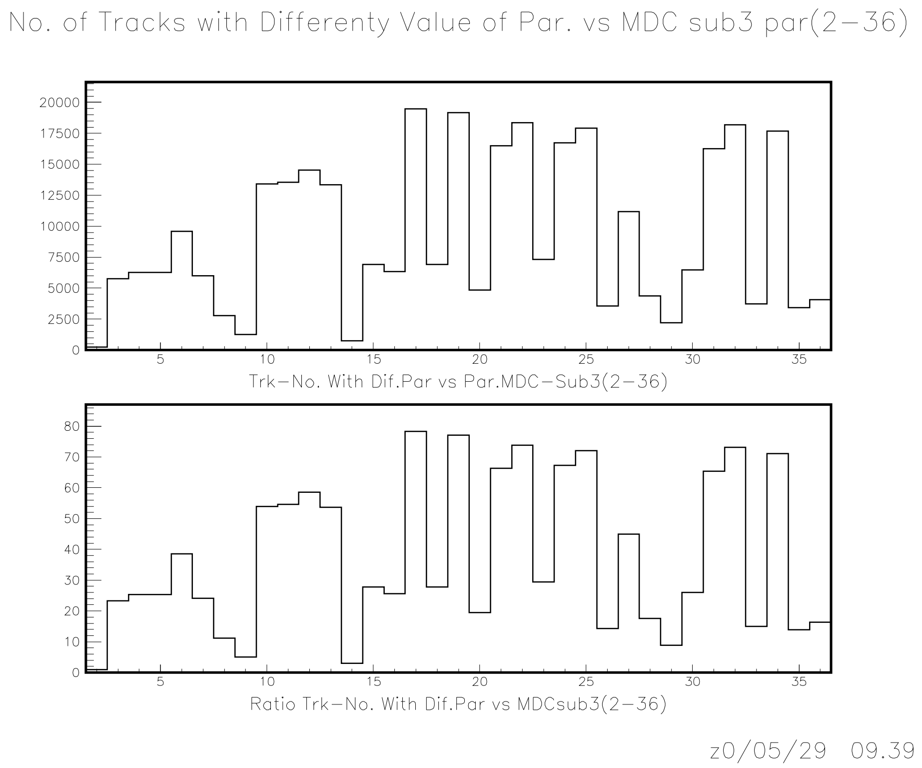


Figure 2: PAR DIF for MDC

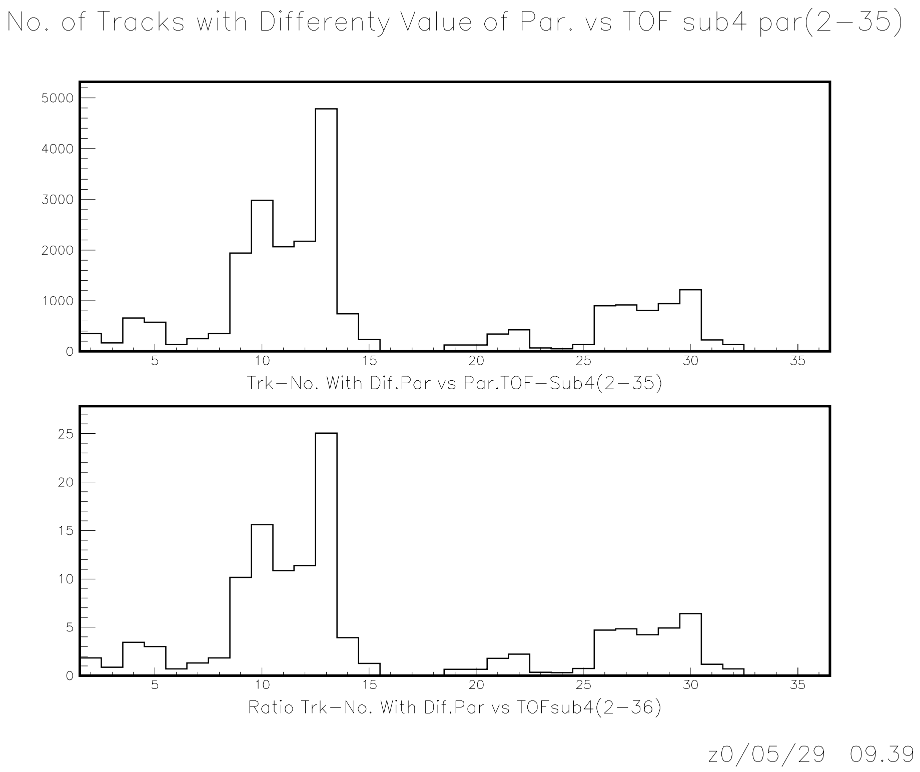


Figure 3: PAR DIF for TOF

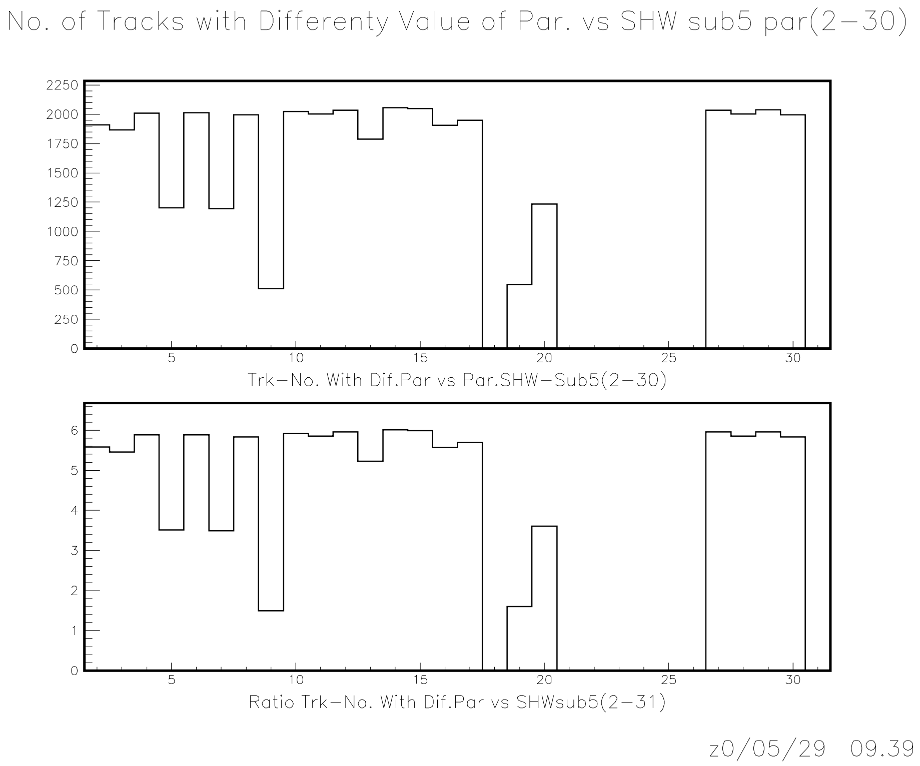


Figure 4: PAR DIF for SHC

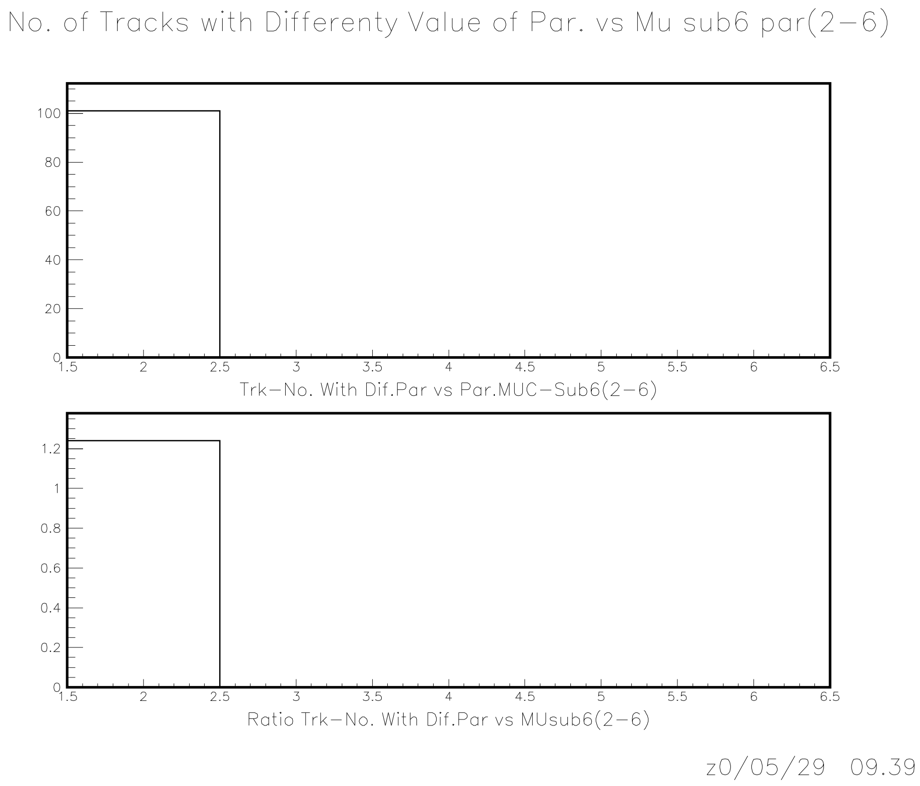


Figure 5: PAR DIF for Muon Counter

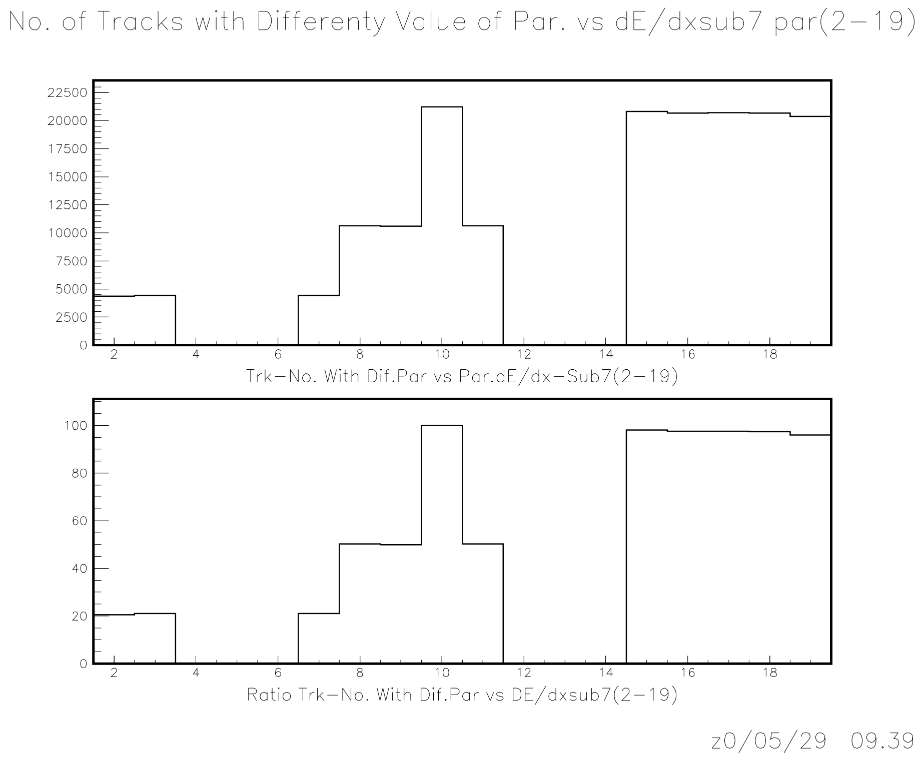


Figure 6: PAR DIF for dE/dx

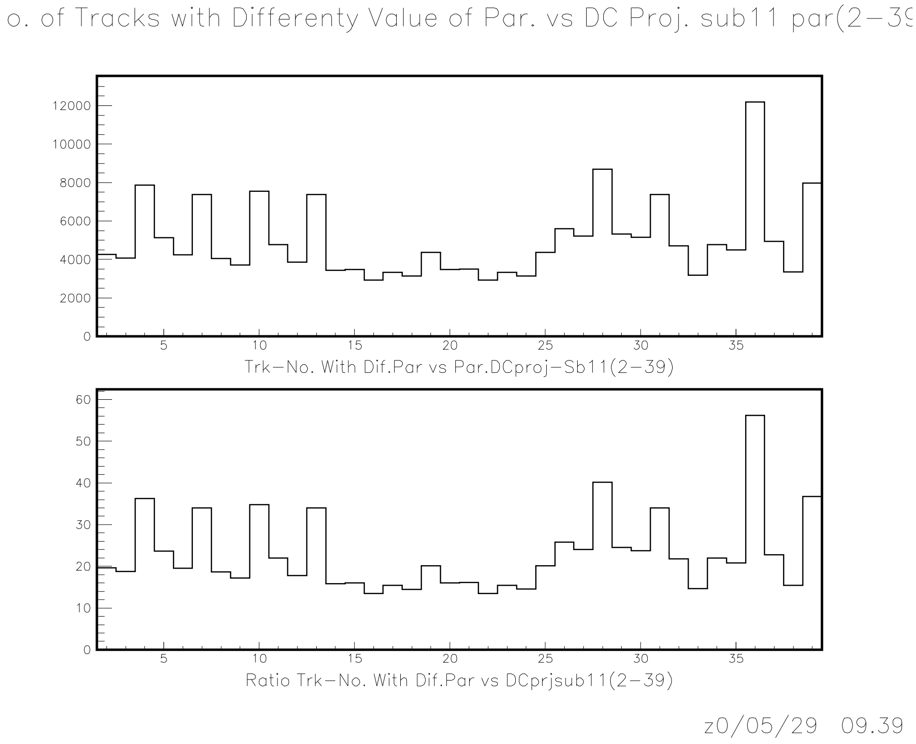


Figure 7: PAR DIF for Projected Coordinates of MDC Tracks

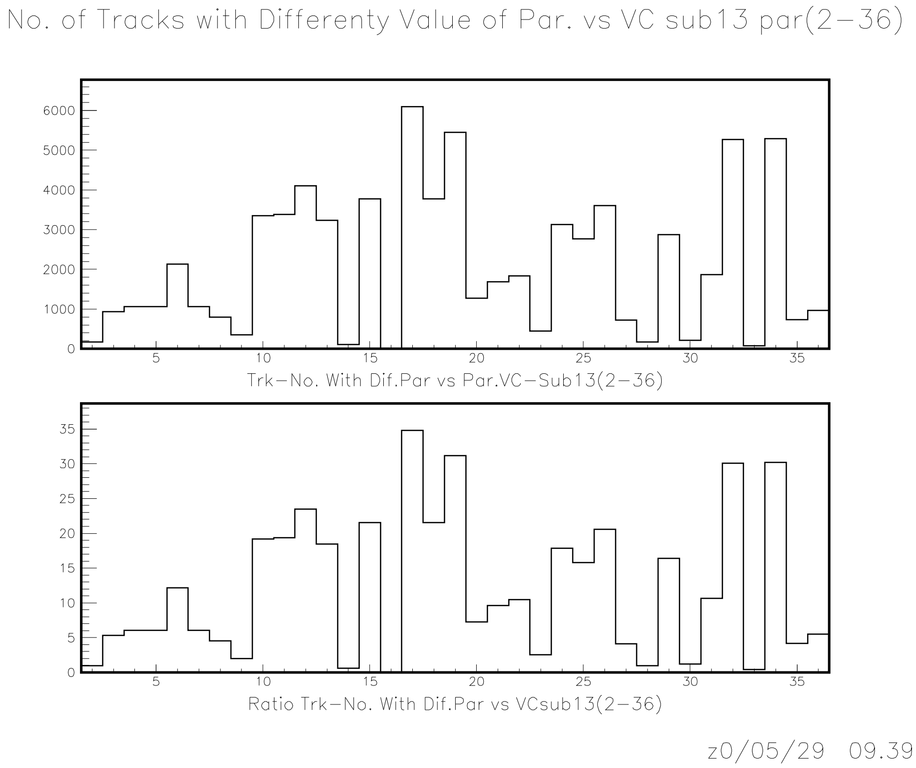


Figure 8: PAR DIF for Vertex Chamber

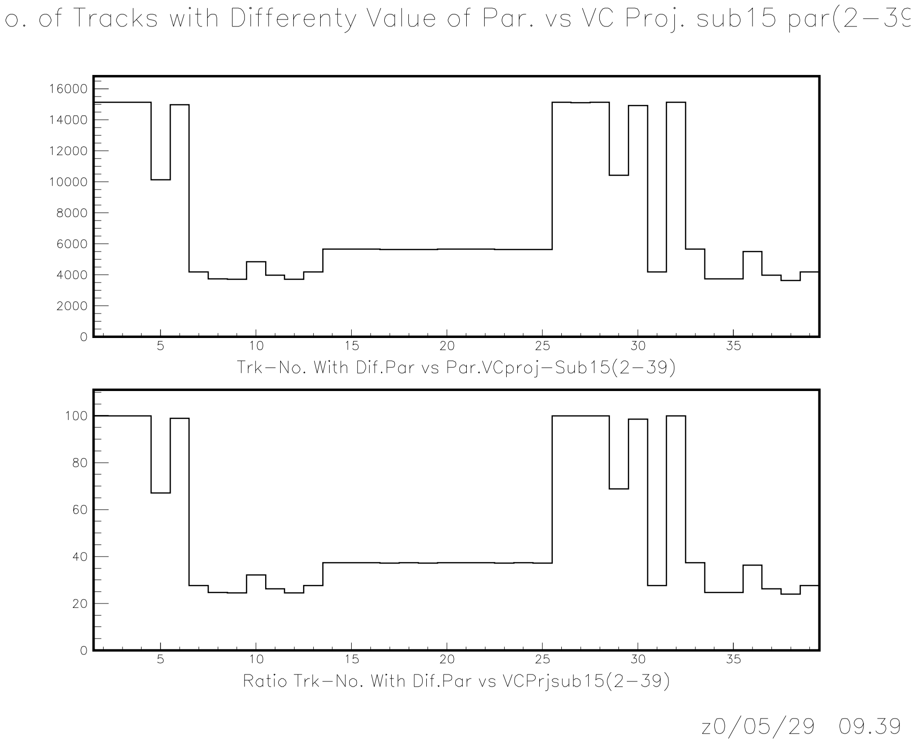


Figure 9: PAR DIF for Projected Coordinates of VC Tracks

7 References

[1] The improvement of the BES offline analysis envoronment, inner report 1998. 10.

[2] The Working Summary of Offline Calibration Group of 1999/11–2000/6 reported in this meeting.