A Distributed Network Clustering Linux PCs System and
Study of its Ability
to Reconstruct BES-II Data [1]

Gang Rong Kanglin He Shaoqiang Zhang Dahua Zhang
Jinzhi Bai Shengtian Xue

(BES Monthly' Meeting, IHEP, Beijing, March 14, 2000)

1. Introduction

2. Configuration of the Cluster

3. Environment of Data offline Processing within the Cluster

4. Ability to Reconstruct the BES-II Data

5. Summary

[1]The talk is given at BES Monthly' Meeting at IHEP on March 14, 2000. It was based on many useful discussions with Profs. Z.G. Zhao, Y.N. Guo, Y.C. Zhu, N.D. Qin.
Manny thanks to peolple from data production group of the Department, Mrs J. Su and Dr. G.X Sun from computer center for their kindly help. We would like to thank to Dr. Liaoyuan Dong for his kindly help for installation of SCSI disk.

INTRODUCTION

- Purpose

  Our original purpose to set up a Network Clustering Linux PCs System is for use in study of D physics with exsiting data taken at 4.03 GeV and expecting data to be taken at 3.77 GeV accoriding to D physics research program funded by National Natrural Science Fundation[1] (The proposal of the research program was made in June 1998 and approved by the Committee of the National Natrural Science Fundation in February 1999) and Chisese Academic Special Support Project.

  According to the D physics research program, we will set up a Linux PCs and HP/UX cluster (it will consist of 1 HP/UX and 3 PCs/Linux) which will be used in the study of the D (at 4.03 and 3.77 GeV) physics.

- A Prototype of the Network Clustering Linux PCs System

  In November 1998, a prototype of the Network Clustering Linux PCs System was set up. It consists of two PCs. One of them was set as a server and the other as a slave. The two PCs were connected together through IHEP Key Network. Three Ph.D. students in the D Physics Group finished their thesis research work mainly based on the system.

Since the $\psi^{''}$ D physics is still one of the candidates to be put into effect after collecting J/ψ data at 3.097 GeV (hopefully taking J/ψ data will be finished by the end of May 2001), we should prepare the software enverionment and make the feasibility study of that whether we can use the PCs/Linux to do BES-II data offline reconstruction.

- A Distributed Network Clustering Linux PCs System

  In order to speed up our research program of Charm physics and make best use of some idel CPU (some PCs CPU waste overnight), we set up a Distibuted Network Clustering Linux PCs System. The system consists of 2 Pentium-III, 3 Pentium-II and 1 Celero PCs. Three of the PCs performed a Kernel Network Clustering PCs System which is located in Computer Center. The other three are distributed in Building No. 3 and Building No. 5. We called the distributed three PCs system as Shell Network Clustering PCs System here in the Talk.

  Within the Distributed PCs System, we studied the BES-II data offline reconstruction and physics analysis (See our other talk –"Study of BES-II Data offline Reconstruction on PCs/Linux Platform" given at the same BES Meeting).

  The experince for setting up and operating the system is not only useful for BES-II data offline processing and physics analysis, but also useful for Babar Chinese team's data analysis, which will probably be based on PCs in IHEP, Beijing.

CONFIGRATION OF THE CLUSTER

- Hardware
  - The Kernel Clustering Linux PCs System
    A PIII-500 with 128M RAM and IDE/SCSI Disks has been set to be as a main server for the cluster. A desk printer was connected to the server. Tape drivers can be connected to the server through the SCSI bus.

    Two PII-400s with 128M RAM and IDE disks have been set to be the slaves. One of them was borrowed from MDC Group for temporary use.

    The 3 PCs are connected together though a Soho Basic HUB208b (10Mb/s).

  - The Shell Clustering Linux PCs System
    One PII-400 (Shengtian XUE's PC in Building 3), one PIII-500 (D.H. ZHANG's PC in Building 3) and one Celero-300 in Building No.5 have been set up to be 3 other slaves. The three PCs are with 128M RAM and IDE disks.

    The 6 PCs are connected through the IHEP Key Network (100MB/s) to form the Distributed Network Clustering Linux PCs System.

- Software
  - Operating System
    The Red-hat 6.1 Linux and Window 98 were used as the operating systems for the PCs Cluster. The main server only run Red-hat 6.1 Linux. the slaves can run the two operating system separately. But when we work on BES-II data offline reconstruction, Monte Carlo simulation, physics analysis and BES/BES-II software R&D jobs, the slaves should run Red-hat 6.1 Linux.
  - System Service, Data, File and Job Manergement
    * NIS (Network Information Service) – YP (Yellow Page)
      The NIS has been set for the PCs' communication each other within the Kernel Network Clustering PCs System.
    * NFS (Network File System)
      Data, Codes and Libaray are shared by the all PCs within the Distributed Network Clustering PCs System by using the NFS.
    * NQS (Network Queuing System)
      All Jobs (Drunk, Sober and others) can be manerged by the NQS.

      The NQS software is not available with the Red-hat 6.1 Linux. We installed the NQS in the PCs within the Cluster. It works well to manerge the jobs.

BES/BES-II SOFTWARE ENVIRONMENT

- BES/BES-II Software

  BES Software of Versions of 8.6, 8.7, 8.8 and BES-II Software of Versions of [100], 101 and 102 are available and runable on PC/Linux platform within the Cluster.

  At present there is no offical Linux runable V102 code, we copied the V102 code from HP/UX BES Software Libarary to the main server, converted the code and built the V102 BES Software there on 14 Feb., 2000. On 15 Feb. the V102 code was runable on the PCs in the System. After fixing a few problems the V102 Drunk jobs for BES-II data offline reconstruction can goes smoothly. Since then we began to reconstruct the BES-II data on the Linux Platform with the PCs in the Cluster.

- CERN Libaray and so on

  CERN Libaray and some other Libaraies are available in the system.

- BES/BES-II data offline processing

  Two data offline processing environments have been set up for BES/BES-II separately within the Cluster. The BES/BES-II Drunk Jobs can be manerged by NQS well.

ABILITY TO RECONSTRUCT THE BES-II DATA

- CPU costs

  Within the Distributed Network Clustering Linux PCs System, we studied the abilities of the PCs to deal with the BES-II data reconstruction. In general It costs 23.6 and 28.7 minutes for PIII-500 and PII-400 to fully reconstruct one run BES-II data and write the data to disks, respectively.

- CPU idel

  Since all slaves PCs read data from IDE or SCSI disks mounted on the main server (PIII-500) through the KNCPCS (Kernel Network Clustering PCs System) or through the KNCPCS + 'IHEP Key Network' during the data offline reconstruction, there are idel CPU during the data reconstruction. In the case of reading the data from IDE disk, the average idel CPU of the PCs located in Building 3 is about 20%, while the average idel CPU of the slave PCs in the KNCPCS is about 5%. The main server has average idel CPU of about 0.3%.

- Average load

  In the case of running one BES-II data offline reconstruction job at each of the PCs, the average load is about 1.1 for the slave PCs, and 2.2 for the main server PC.

- MEM used

  MEM 128M, 125M used, 3M Free.

- Ability to deal with the BES-II data

  Considering the system will deal with some other jobs such as reading/writing the raw data onto tape, selecting Bhabha events for detector calibration, and sorting the raw data and reconstructed data and so on, suppose only 16 hours will be used for data offline reconstruction per day. In this case,

  - The Kernel Network Clustering PCs System

    108 runs BES-II data can be fully reconstructed by Kernel Network Clustering PCs System per day.

  - The Shell Network Clustering PCs System

    73 runs BES-II data can be fully reconstructed by the Shell Network Clustering PCs System per day.

  A total of 181 runs data can be reconstructed per day.

- How manny days needed to reconstruct all data take within a year (about 3500 runs estimated) ?

  It takes about 20 days to fully reconstruct 3500 runs BES-II data with the "Distributed Network Clustering Linux PCs System, or about 35 days with the Kernel Network Clustering Linux PCs System if everything is fine !

- What percentage of the IHEP Key Network Ability is costed by the Distributed Cluster?

  One run raw data is about 60 Mb in size. The averaged real time to reconstruct one run data is about 30 minutes on the PCs in the Shell Network Cluster. These mean that one BES-II data offline reconstruction job running in the Shell Network Cluster PCs needs to transfer data from the Server at 0.033Mb per second. This costs only about 3/10000 of the IHEP Key Network full ability (100Mb/s) or 3/1000 of the Local Aera Network full ability (10Mb/s) for tranfering data.

  On the schedule of the Network Upgrade, the IHEP Key Network full ability to transfer data will be upgraded up to 1000Mb/s and the Local Aera Network full ability will be upgraded to 100Mb/s. So the percentage of the IHEP Key Network full Ability costed by one BES-II data reconstruction job running on the PC in the Distributed Cluster will be only tiny of the full Network ability.

- How about the data reconstruction with HP/UX Cluster?
  - HP/J5000 (hpws9)

    It costs about 36.3 minutes for one CPU to fully reconstruct one run BES-II data. Converting the CPU time into that for two CPUs working together, it costs 18.2 minute to fully reaconstruct one run BES-II data.

    79 runs BES-II data can be fully reconstructed for 24 hours.

    53 runs BES-II data can be fully reconstructed for 16 hours.

  - HP/J2240(hpws8)

    It costs about 58.3 minutes for one CPU to fully reconstruct one run BES-II data. Converting the CPU time into that for two CPUs working together, it costs 29.2 minute to fully reaconstruct one run BES-II data.

    45 runs BES-II data can be fully reconstructed for 24 hours.

    30 runs BES-II data can be fully reconstructed for 16 hours.

  - HP735/125 (hpws0 or hpws1)

    It costs about 101 minutes to fully reconstruct one run BES-II data.

    14 runs BES-II data can be fully reconstructed for 24

Table 1: CPU time costed to fully reconstruct BES-II data.

CPU Type	30000 Events to be reconstructed [minute]	CPU costed per event [minute]	Note
HP/J5000 (hpws9)	$36.3/2 = 18.2$	0.0364	2 CPUs work together
HP/J2240 (hpws8)	$58.3/2 = 29.2$	0.0584	2 CPUs work together
HP735/125 (hpws0,1)	101	0.202	
PIII500	23.6	0.0472	
PII400	28.7	0.0570	
Celero-300	39.5	0.079	

hours.

10 runs BES-II data can be fully reconstructed for 16 hours.

SUMMARY

On the schedule of the D physics research program, we have set up a Distributed Network Clustering PCs System. It consts of 2 PIII-500, 3 PII-400 and 1 Celero-300 microcomputers. One of the PIII-500 is set as the main server to manerge disk files, BES software libaray, CERN libaray and jobs. The others are as slaves. The System can be shared data and files with the HP/UX machines as well.

The system can be divided into two parts. One is the Kernel Cluster and the other is the Shell Cluster. The Kernel Cluster PCs have been runing more than 6 months already (one PII-400 in the Kernel system was added into the Cluster 4 weeks ago). They work fine. The Shell Cluster PCs of the system, distributed in other Buildings, have been added into the system and used to build the Distributed Network Clustering PCs System two weeks ago. They also work fine.

In the system, a parallel jobs processing environment NQS was set up for the BES-II data offline processing and physics analysis. The system can be used in BES/BES-II physics analysis, BES/BES-II software $R\&D$ and data offline processing.

With the Distributed Network Clustering PCs System, we can reconstruct 181 runs BES-II data taken at 3.097 GeV for 16 hours loaded CPUs.