A Linux PC Farm for Physics Analysis in the ZEUS Experiment

Marek Kowal, Krzysztof Wrona, Tobias Haas, Ingo Martens, Rainer Mankel

ZEUS - DESY, Notkestrasse 85, 22607 Hamburg, Germany
`http://zarah.desy.de/`

Abstract

15 months ago the ZEUS experiment started a still ongoing project to migrate the experiment's reconstruction and batch analysis from the very expensive multiprocessor SGI Challenge computing environment to a PC Farm.

After successful migration of reconstruction software to Linux and first exercises with two batch systems we decided to migrate completely to PCs and provide users with batch facilities.

Right now our computing facilities consist of 47 PCs, two SGI fileservers with 3TB of disk space and tape mass storage (20TB). Old SGI Challenge computers are also used. Two batch systems: NQS and LSF were evaluated.

Special "jobclients" software developed at ZEUS provides users with job submission and retrieval commands from any operating system. PCFarm status is available on line all the time through WWW as well as interface to most of the administrative tasks.

File transfers across network are supported by ZEUS-made RFIO software, which is an ancestor of the CERN RFIO. Access to mass storage facilities is supported by special tpfs filesystem which has been developed at ZEUS.

In this article we give a description of the system, explain software and hardware choices, discuss its performance, scalability and maintenance and also give prices of respective subsystems.

Keywords: Linux, PC farm, batch system, data analysis

1 Introduction

For the past few years the ZEUS experiment was using multiprocessor SGI Challenge XL machines as its main computational platform. Since ZEUS software (as most of the HEP software) is not intrinsically parallel, we do not benefit from parallelism offered by those multiprocessor machines. Therefore the only important issue for us is the processor computing power and the IO rate offered by each subsystem.

On the other side, existing computer resources started to be insufficient in terms of required computational power. This is due to increasing data sample size which ZEUS members process in their analysis jobs.

Since additional processor boards for the SGI machines are a very expensive option, a farm of commodity PCs would be an attractive solution to provide enough computational power to satisfy the needs of the ZEUS users. This requires however that sufficient IO rate as well as an efficient maintenance can be achieved.

Since the SGI machines would be used in parallel for some time a transparent user interface for batch job submission and retrieval is also mandatory.

Table I: PC Hardware

number of PCs	processor type	memory	IDE HDD
17	Pentium Pro 200 MHz	64 MB	2 GB
20	Pentium II 350 MHz	128 MB	6 GB
10	Pentium III 450 MHz	128 MB	8 GB

2 Hardware layout

Currently our PC Farm consists of 47 PCs, 2 SGI fileservers and 3 SGI Challenge XL multiprocessor machines. 17 of the PCs are used for reconstruction purposes and 30 for the batch analysis. In the following we will concentrate on the analysis applications.

2.1 PCs

The PC Farm is consists of 47 Linux PC machines. Table I shows their components. Out of those, two PCs are set up as local NFS fileservers. Their hardware differs from the others' in that they contain more memory and are equipped in 3×8 GB SCSI discs joined together in software RAID 0 array. PC NFS servers are used as local storage areas for tasks like reconstruction or keeping user files submitted for reprocessing in batch system (see also section 4). The files stored there are considered temporary ones and are supposed to be either moved finally to tapes or retrieved by users to their local machines. All PCs are equipped with 100 Mb network cards.

2.2 Fileservers

Total amount of 3TB of disk space is connected to the fileservers, part of them on SCSI interfaces and part on FibreChannel interfaces. The main fileserver - `doener` (4 processors IP27 195MHz, 0.75GB RAM SGI Origin 2000) is equipped with 1TB of SCSI discs and 1TB of FibreChannel discs. The secondary fileserver - `kebap` (4 processors IP19 100MHz, 384MB RAM SGI Challenge DM) is equipped with 1TB of SCSI discs. Filesystems of those machines are exported via NFS for browsing purposes. Actual data transfer is done with the ZEUS version of RFIO software (see section 5).

2.3 Networking hardware

All of the PCs are connected to a Cisco Catalyst 2948G switch (48×100 Mb, 2×1 Gb) and from there via 1Gb interface to the network card of our main fileserver - `doener`. All SGI machines are interconnected via 800 Mb HIPPI interfaces to the HIPPI Giga Router.

3 Batch system

After evaluating two different batch systems, namely NQS and LSF we decided to use LSF on the PC Farm. Its biggest advantages from our point of view are: a much friendlier administrative interface and the possibility to define a load window in which the jobs submitted to the system are executed. The user interface of LSF is completely hidden from the users by the ZEUS `jobclients` software (see section 4). The same applies to NQS installed on the SGI multiprocessor machines. We decided to continue using NQS on SGI machines, as the price of the LSF license per processor is very high and these machines will be phased out at some point in the future.

4 Job submission software - `jobclients`

Dedicated job submission software was developed at ZEUS in order to facilitate submission of the jobs to the batch system for the users. It defines a set of actions which are important from the user's point of view, namely: submission of job-related files and queueing the job, retrieval of files from the batch system, querying the status of the job, listing the files in a job, killing and purging the job from the batch system and also getting "status" information of the batch system itself. For each of those actions a special "client" binary was prepared. It contacts via TCP/IP protocol a daemon on the batch system side which - in the name of the requesting user - executes the commands in the batch system. Each job is given unique identifier, returned upon submission of the job, which is then used by user for further queries to batch system. Each job is also given its unique directory, where it is executed. This approach has two advantages:

- since each job has its unique remote directory, it is very easy for the user to submit many jobs with different variants of code or data of the same type at the same time, differing - for example - by a value of a single parameter without keeping multiple versions of his/her source directory tree.
- since the job is executed in a remote directory, there is no possibility for user to delete the files after submission but before execution, which is often the case when the users submit jobs in their home directories.

The `jobclients` software is written such that it can be easily customized to any new batch system - thus we can offer the same unique interface to users of both Linux PCs and SGI multiprocessor machines. Jobclients are available for all important UNIX platforms as well as for WindowsNT.

5 The `tpfs` filesystem and RFIO

The `tpfs` filesystem is an NFS server-like daemon, which presents the users with a unique view of all the files both on tapes and fileserver discs and provides them with an automatic staging of requested tape files. It keeps the metadata information on files stored in robots, as the name of the robot, place of the file in the robot's filesystem, file size and creation date. Users can freely browse the `tpfs` directory tree to find the necessary files. By the use of a dynamically loaded library which covers the `open` C library system call it is possible to automatically stage the requested file upon trial to open it. The files are staged to a local disk cache and presented to the user by means of the RFIO protocol. Since the library can be automatically preloaded, one can use it with any program without modification.

In order to facilitate the need for big data transfer across the network which cannot be guaranteed by the NFS filesystem we decided to use the RFIO package. The original package developed at CERN had some important disadvantages so a ZEUS version with the following fixes was prepared:

- The ZEUS-RFIO version is stateless. Therefore we can restart our fileservers without crashing the jobs running on batch machines. Considering the amount of discs connected to the fileservers, some maintenance periods are required and we cannot guarantee that we will not have to shut down the machines once in a while.
- ZEUS-RFIO is available as a dynamically loaded library, which "covers" standard `open, close, read, write` and other file operations on the C library level. Thus it is possible to run **any binary** with RFIO, even if it was compiled when the RFIO package was not available. This means that the same job can be run on local files, over NFS or via RFIO.

6 WWW interface

The status of the batch system and some administrative tasks are available on-line all the time via the WWW interface. One of our machines runs Apache `httpd` daemon which in turn runs CGI scripts querying the necessary information from the batch system. The information that can be found on WWW includes:

- information about the resource availability (amount of free space on discs, uptime of machines etc)
- the load on available machines
- querying running jobs for status information

A graphical interface to JAVA `jobclients` available through WWW is under preparation. More information can be found under the address: `http://zarah.desy.de/`.

7 Scalability, performance and prices

The computing power offered by today's Intel Pentium processors is much better that of IP19-IP25 SGI processors. We found out that the most important issue concerning big IO rate over the network is the installation of an efficient switch. With the use of normal hubs, our jobs had spent most of their time waiting for the connection.

On the other side, the IO rate offered by today's Linux OS in connection with SCSI discs and an almost complete lack of support for FibreChannel interfaces is completely insufficient in terms of required disc IO resources. Therefore `doener` (SGI Origin 2000) is going to remain our main fileserver. Scalability of the Origin architecture is almost unlimited - as much as 50TB of disc space can be connected to it. The same applies to the networking connections between the fileserver and the farm - if needed, several 1Gb connections can be bundled to provide the required bandwidth.

Synchronization of software across different machines is maintained by keeping it in the AFS software directory tree, updated once from one of the farm's machines.

One of the strongest arguments for using the PC farm is its price. The total cost of one PC, together with the storage (rack), LSF license and an 100Mb slot in switch amounts in our case to 3000 DM.

8 Summary and future plans

After more than three years of experience with PC Farms (our first reconstruction farm of 17 PC was built in 1997) we can conclude that they are much easier to administrate and maintain than multiprocessor machines. Few reasons to mention are:

- price per processor is much smaller than that of multiprocessor SGI machine, even if you include costs of racks for storing PCs
- most of the software is free and is available as sources
- failure of a single PC means that only one job fails. In case of multiprocessor machines whole bunch of jobs is crashed

We intend to acquire further PCs and finally remove all SGI Challenge multiprocessor machines (both computational and one fileserver). Once all except one (`doener` fileserver) SGI multiprocessor machines are given up there will be also no need to use HIPPI networking environment anymore.