

Interactive and Batch Analysis at DESY

K. Künne

*Deutsches Elektronen-Synchrotron (DESY), Notkestr. 85, 22607 Hamburg,
Germany*

Besides the data taking and reconstruction the analysis of the Physics Data is a major part in Physics Computing. Currently a new structure, which should last for the next years (maybe the next millenium), is introduced in the computing for Physics Analysis at DESY, which closely follows a Client-Server-Model, and which is based on the Workgroup-Server-Concept (first introduced at CERN). The talk will give an overview about this new structure, which includes File-Servers, Workgroup-Servers, and Batch-Servers. Special emphasis will be given to the handling of different classes of data inside of such a distributed system. The important point here is to avoid unnecessary network traffic. Other key topics include management of that distributed system and load balancing issues. Also presented in this talk are experiences with that system and future developments.

1 Batch and Interactive Analysis

The first analysis steps are usually a reduction of the amount of data from very huge datasets to smaller samples of selected events. The input datasets for this first step are the reconstruction output datasets. Here at DESY these datasets have a size in the order of several 100 GB. They are kept on tape. In order to process such a big amount of data a lot of processing time (up to one week) is required. That's why this processing is done in batch mode.

After the creation of some event samples these selected events are analyzed in an interactive fashion. This includes the visualization, generation of histograms and so on.

1.1 Requirements

The main requirements for the interactive analysis are a good availability of the system, a short response time, and enough CPU power. Usually the system should be available for work during working hours and the downtimes should be very low. The response to keypresses should be fast enough in order to provide a good editing environment and there should be enough CPU power so that calculations finish in a reasonable time.

For the batch analysis the situation is different because the amount of data which has to be processed is much larger. Therefore a good I/O throughput, lots of disk space, and a fast tape access are required. Because nobody sits in front of the machine and waits until the batch job has finished a certain amount of machine crashes and downtimes are tolerable.

The experience at DESY shows that mainframe-like UNIX systems are not very well suited in order to fulfill the requirements for the interactive analysis. The main problems, which were found, are the availability and the sometimes bad response time. On the other hand it turned out that smaller and specialized machines are

Central Fileservers

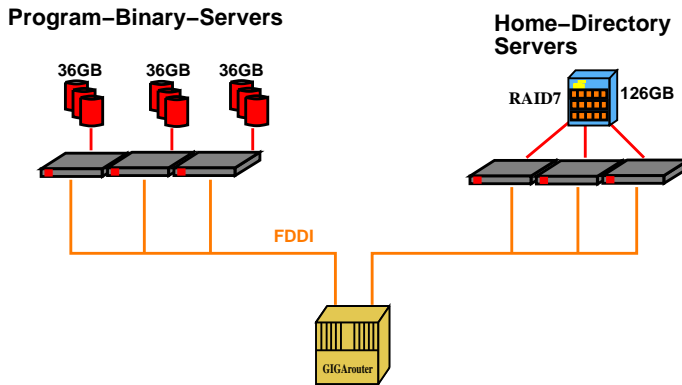


Figure 1: Central Fileservers.

much more stable and they have the advantage that it is possible to assign several machines to the same functionality. That improves the overall availability of such a system by means of redundancy.

2 System Overview

After looking at the experiences of DESY and other big HEP institutes (especially CERN) a decision was made to install a system for the physics analysis, which closely follows a Client-Server model, and which is based on the Work Group Server concept, which was first introduced at CERN ¹.

As the basic network filesystem the AFS filesystem was chosen. The system consists of several parts. One central part are the central AFS filesystems for home directories and application programs (Figure 1). The home directory servers are connected to a RAID array where the home directories are kept. Other parts are the analysis farms of the big experiments H1 and ZEUS (Figure 2 shows the H1 farm). All parts are connected via the network, where the Gigarouter plays a very important role.

The batch servers which were chosen are SMP machines (SGI Challenges) with up to 32 CPU's and up to 2 GB main memory. These machines are connected via HIPPI connections to the Gigarouter. There is up to 450 GB disk space connected to one machine and the machines have a fast access to Ampex and STK robots with a huge amount of tapes.

For the work group servers workstations with up to 2 CPU's and 300 MB memory were chosen. Every server has 8 to 10 GB local disk space and there are usually 10 to 20 users working on one machine. The machines have FDDI and Ethernet interfaces where FDDI is used for the data traffic (from and to the central filesystems) and Ethernet is used for the traffic between the machines and

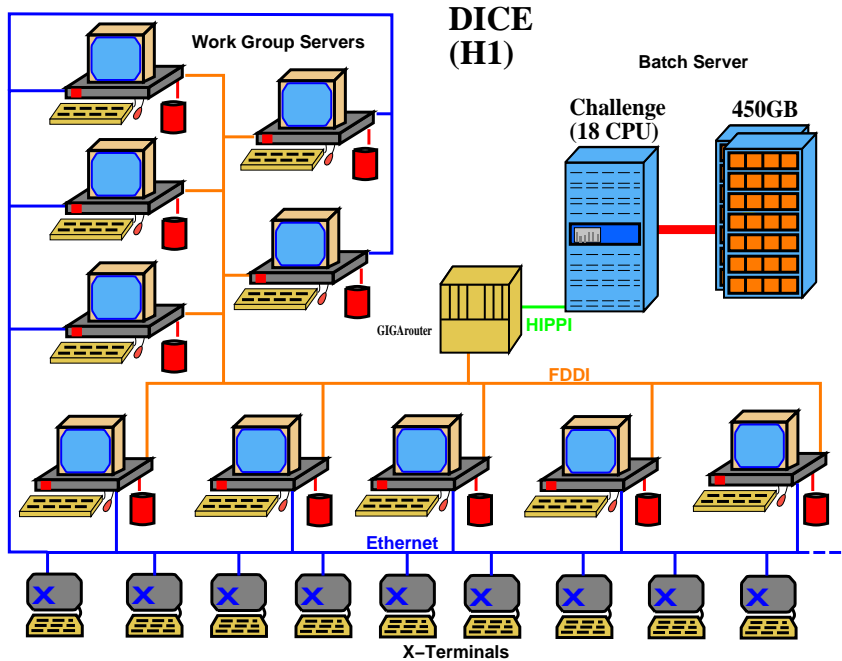


Figure 2: H1 analysis farm.

X-Terminals.

3 Data Issues

During physics analysis there is a big amount of data to handle. The goals for the data handling are to provide a fast and easy access to the data, not to overload the network by too much data traffic, to ensure that the data can be accessed in a reliable way, and that it's available if it's needed.

In general there are three classes of data to distinguish: very important and small data, experiment data, and medium sized data. The last kind of data is of a somewhat temporary nature but that doesn't necessarily mean that the lifetime is very short, the lifetime can be in the order of several months or longer.

3.1 Important Data

Under important data one has to understand data which is needed for the daily work and which can not easily be reproduced. Examples for this kind of data are configuration files, mail files, reports, text files. Experience at DESY shows that the amount of this data is in the order of 20 to 100 MB per user, the filesize is rather small. It is usually affordable to transfer these small files via the net.

In order to provide a highly available and reliable storage for this class of data it is stored on a RAID array, which is connected to specialized home directory servers.

3.2 *Experiment Data*

The experiment data is the big stuff and includes things like the raw data, reconstructed data, and DST files. Here at DESY this kind of data is in the order of several TB per experiment, the files are usually larger than 100 MB. It is not affordable to transfer such huge files over our “normal” nets, that means Ethernet and FDDI.

The main repository for the experiment data are the tape robots at DESY, large parts of it are also kept on disks which are connected to the batch servers. Because there are fast connections from the batch servers to the tape robots data can easily be exchanged between disks and tapes. Access to the data is usually only possible from the batch servers but there also exists a possibility to access the data on the batch servers disks via NFS, mainly for lookup purposes, not for heavy reading and writing.

3.3 *Medium Sized Data*

Under medium sized data one can understand data which can be regenerated but where the regeneration requires some amount of processing time so that it's usually necessary to keep this kind of data for some time. Examples for this are event samples (the results of batch jobs), selected parts of the DST files (mini-, mikro-DST's) and others. The filesize of these kind of files is in the order of 20 to 100 MB and one physicist here at DESY usually has between 100 MB and 1 GB of such data. Fortunately in general the same files can be shared between a small group of physicists so that the total amount is not the number of physics users multiplied by 500 MB but rather one tenth of it. It is in principle affordable to transfer data of that magnitude over our nets but in order to avoid overloading this should be minimized somehow.

Currently no general solution exists for the storage of this class of data. There are three possibilities. The first solution is to store the data on local disks, which are connected to the work group servers. This solution is what is currently in use at DESY. It is very easy to realize, there is not much preparation needed. The remote access from other work group servers is done via NFS cross mounting of the work group server disks. The disadvantage of this solution is that it establishes dependencies from particular work group servers. The servers do not provide only CPU power and memory any longer but also provide some amount of data. That means, if one of the work group servers fails the data, which is kept on it's local disks, is no longer available. Because the work group servers run a lot of different programs and they are not very specialized the probability of a crash of a work group server is somewhat higher compared to the central file servers. This is also the experience at DESY.

Another solution would be to keep the data on tape as the main repository and to stage it on demand to local disks on the work group servers. The advantage is that one gains independence from the availability of specific work group servers. If one server fails users can go to another server and stage the needed files again to the local disks of that server. A disadvantage is that it's possible that a file is staged twice (or more) to different work group servers. This wastes some amount of disk

space and it also generates more network traffic.

The third solution would be the use of a special data server. This server would be similar to the central file servers but because the amount of data is much higher compared to what is stored at the home directory servers it appears to be necessary to have a tape robot as backing store for the data and only keep a subset on disk. The idea is to have some kind of automatic migration of data between the data servers disks and tapes. The work group servers can access the data via a network filesystem, that could be NFS, AFS or DFS. In order to minimize the network traffic the filesystem should have caching at the client side like AFS or DFS. The data server is still a project but it's foreseen to make investigations into that direction.

4 Administration

The administration of the home directory space is done via group quotas. That means that every major group at DESY gets some amount of disk space and the group administrator can distribute it among the users. In order to do that there exists a tool "group_adm" which makes it very easy to assign quotas to group members and to check the disk usage. This tool is based on tcl/tk and was written by Sergei Kulikov.

The user administration is still based on the old Apollo Domain Registry but there exists a plan to migrate to the DCE registry.

Currently no centralized solution exists for the system administration. There are some pieces but there is no central instance which provides the necessary means to manage the batch and work group servers. For the future it is planned to provide a central solution for the system management.

5 Experiences

Currently there are 153 registered AFS users at DESY but this number is continuously increasing. There are approximately 3 GB home directory space occupied. Unfortunately the batch servers are not yet running with AFS due to delays in the delivery of AFS for this kind of machines. The system is running since 4 months now. Besides one serious crash at the very beginning the central file servers are very stable. One can not say the same about the batch and work group servers. Due to various reasons there were some more problems in that area. But this is exactly what was expected (well, it was not expected that the machines are unstable but there is a higher probability for a crash on a work group or batch server compared to a pure specialized file server).

References

1. C.Jones, "A Strategy for Interactive Services at CERN over the next three years", talk at the HEPiX Conference Pisa, September 1993.
2. K.Künne, "New developments in UNIX Computing at DESY", talk at the HEPiX Conference Saclay, October 1994.