Data storage and Network in offline Analysis of ALEPH

Ronald Hagelberg
European Laboratory for Particle Physics (CERN) CH-1211 Geneva 23, Switzerland

ALEPH collaboration

Presented at the Conference on Computing in High Energ Physics 1992
21-25 September 1992 in Annecy, France

In large HEP Experiments, CPU is more and more distributed with the increasing use of
workstations. A typical analysis accesses several Gigabytes of data. For computer platforms
linked by fast networks, redundancy of data should be avoided to simplify management.
Taking the ALEPH experiment as an example, we show the approach taken for networking
and data storage. ALEPH at CERN uses a CRAY, an IBM mainframe, a centrally-operated
UNIX cluster "SHIFT", and a locally managed UNIX and VAX cluster for data analysis.
Efficient and transparent access to the data is ensured for workstation users by a fiber-optics
based FDDI link between several computer platforms.

The LEP experiments are now in the forth year of data taking and the volume of data is
increasing continuously. On the example of ALEPH, one of these experiments, we explain how
the data are accessible for analysis from the different computer platforms at CERN.

1 Data

Data recorded by the experiment are written onto disk. When 200 Mbytes are accumulated the
run is ended and the raw data are copied to 3480 cartridges at the pit and processed by a Vaxstation
farm running the standard reconstruction program. The output is called the Production Output
Tape (POT). In the next step all events not useful for most analyses are suppressed and a Data
Summary Tape (DST) is written. The Mini-DST files contain the same events as DST's, but in
a compressed form and only information needed for further analysis is kept.

Raw-data are generally mounted only 5 times for detector studies, monitoring or very
special searches e.g. fractional charge particles. Analyses which need the POT's are rare.
Sometimes the DST is needed but the aim is to have the information on Mini's complete enough
for the majority of analysis work.

Reducing the data volume greatly simplifies the problems of accessing and storing the
data. Table I illustrates the storage requirement for ALEPH in the first 4 years.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>12</td>
<td>90</td>
<td>180</td>
<td>145</td>
<td>427</td>
</tr>
<tr>
<td>POT</td>
<td>5</td>
<td>25</td>
<td>55</td>
<td>60</td>
<td>145</td>
</tr>
<tr>
<td>DST</td>
<td>2</td>
<td>11</td>
<td>24</td>
<td>34</td>
<td>71</td>
</tr>
<tr>
<td>Mini</td>
<td>-</td>
<td>1.9</td>
<td>3</td>
<td>3.7</td>
<td>8.6</td>
</tr>
<tr>
<td>Nano</td>
<td>-</td>
<td>0.16</td>
<td>0.3</td>
<td>-</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Table I: Volume of data in Gbytes, only real data no Monte Carlo.
The data were recorded in a standard data format, EPIO [1], readable and decodable on all platforms to give freedom in the choice of the computer [2]. The blocks are 32040 bytes long. To avoid at the beginning all problems with multiple files each cartridge contains until now only one single file of up to 200 Mbytes [3]. Event directories [4] with 4 words per event namely run, event and record number with classification bits are generated for all POT's, DST's and Minis's. These 0.8 Gbytes of Event directories allow fast selection of events and reading of the data from disk in random access mode.

2 Distributed CPU's and data storage

From the start we aimed at writing our code in standard FORTRAN allowing us to take benefit of all CPU's available for ALEPH at CERN. This implies to have access to the data on all these platforms. In addition to the centrally operated mainframes CERNVM and Cray, the Unix system "SHIFT" [5] is used. Aleph has its own VAXstation cluster [6] and its own cluster of DECstations [7]. We have several platforms with different operating system all offering interesting amount of CPU. Figure 1 shows the network linking the different computers and the disk space available to ALEPH.

All Mini's and most recent DST's are permanently on disk on ALWS and CERNVM. The rest of the DST's and the POT's is on cartridges accessible by a robot. Staging these cartridges allows an optimal usage of disk space and reading the data in random access mode.
3 Network Access

Accessing the data, opening and closing the files is done in FORTRAN inside the analysis programs with the help of datacards prepared by the ALEPH physics bookkeeping package SCANBOOK. To hide network access from the user by software is a large effort.

On SHIFT the data are accessed from FORTRAN by issuing a system call staging the data onto a pool of disks if they are not yet on disk. A symbolic link in the directory /aleph/data is created by each staging process giving the pathname for the datafile. For staging the Cray or a Sun computer is used. The Cray is the only unix system with access to the cartridge robot. The transfer rate to the SHIFT disks for a tape read on the Cray is 0.65 Mbyte/sec. For the Sun the transfer rate was measured to be 0.30 Mbyte/sec.

On Decstations access to data on ALWS and SHIFT is possible. In the directory /aleph/data symbolic links giving the path to the data on ALWS are maintained. Similarly in the same directory symbolic links to the files on the SHIFT data disk pool will be created. If the data don't exist on disk, a staging command is issued to the Cray respectively the Sun.

On ALWS we can read the data from Shift disks. Unix files have no record format. We have to change the record attributes of the shift files on the VAX-cluster to be able to read it there.

A Remote File Access System (RFIO) provides an efficient way of accessing files stored on SHIFT disk servers from compute or tape servers.

Table II gives the transfer rates measured [8].

<table>
<thead>
<tr>
<th>compute server</th>
<th>network</th>
<th>data server</th>
<th>transfer rate in Mbyte/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>dxal10</td>
<td>local</td>
<td>dxal10</td>
<td>1.35</td>
</tr>
<tr>
<td>dxal01</td>
<td>FDDI</td>
<td>dxal10</td>
<td>0.88</td>
</tr>
<tr>
<td>dxal01</td>
<td>ethernet-same segment</td>
<td>dxal10</td>
<td>0.6</td>
</tr>
<tr>
<td>dxal03</td>
<td>ethernet-one bridge</td>
<td>dxal10</td>
<td>0.55</td>
</tr>
<tr>
<td>dxal05</td>
<td>ethernet-several bridges</td>
<td>dxal10</td>
<td>0.3</td>
</tr>
<tr>
<td>dxal10</td>
<td>ethernet-Multinet</td>
<td>alws</td>
<td>0.14</td>
</tr>
<tr>
<td>dxal10</td>
<td>ethernet with RFIO</td>
<td>alws</td>
<td>0.36</td>
</tr>
<tr>
<td>dxal03</td>
<td>ethernet with RFIO</td>
<td>alws</td>
<td>0.36</td>
</tr>
<tr>
<td>dxal03</td>
<td>local with RFIO</td>
<td>dxal03</td>
<td>1.1</td>
</tr>
<tr>
<td>workstation</td>
<td>FDDI with RFIO</td>
<td>SHIFT</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table II: Transfer speeds measured
4 Conclusion

The main ideas to get good network access to tenth of Gbytes of data for ALEPH offline analysis were: - reduce data volume as much as possible
- standard data format readable on all platforms
- event directories with classification bits for random access mode
- data on robot cartridges staged to disk pool
- ULTRANET and FDDI to improve transfer speed
- remote file input output (RFIO) optimizing remote file access
- on UNIX systems symbolic links are used extensively
- large effort to hide network access from the user by software

Acknowledgements

The realization of the data access for ALEPH was possible only thanks to fruitful collaboration between members of ALEPH and the Computing and Network Division at Cern.

References