An Integrated Storage and Data Management System for a High Energy Physics Experiment

Paolo Calafiura[1]*, Gerhard Wirrer[2]** and Bernd Panzer-Steindel[1]

[1] CERN, IT Division, CH-1211 Geneva 23, Switzerland
[2] Institute of Technical Informatics, Graz University of Technology, A-8020 Graz, Austria

1 Introduction

NA48 is a CERN experiment which studies CP violation by measuring ratios of K^0 decays at a very high level of precision. One of the key elements of the NA48 data acquisition system is a Meiko CS-2, a European, parallel, scalable supercomputer maintained and manufactured by Quadrics Supercomputer World. The CS-2 was made available to CERN in the framework of the ESPRIT project GP-MIMD2 [?]. The experiment's high data rate (7 MBytes/sec initially and up to 20 MBytes/sec in 1997) and high data volume (40 TBytes of raw data in the 120 days of physics data taking foreseen in 1997) require the parallel and scalable I/O facilities, parallel file system and large computing capacity provided by the CS-2 [?]. Storage and data management have to exploit these hardware features, thus hiding the physical location of the data from the user.

In general, the storage system requirements for this class of high-performance data acquisition system can be outlined as follows:

- accept an input data stream that is essentially uninterruptible
- collect multiple Terabytes of data per year
- require (near) real-time, high-performance processing on the input stream

2 Expected File Characteristics and Access Patterns

- The raw data arrive time-ordered in *bursts* every 15 seconds; each burst's data is a single file (size about 100-300 MBytes).
- Files are accessed sequentially (random access within the file is not necessary (or occurs very rarely). No entity smaller than a whole file ist to be managed by the Storage Management System (SMS).
- All files are write-once, delete-never.
- *Online Production:* Multiple streams of selected, reformatted and/or reconstructed data are produced for later analysis. They have to be managed by the SMS together with the raw-data.

* Corresponding author: Paolo.Calafiura@cern.ch (phone/fax: 41-22-767-8997/8920), current address: Istituto Nazionale di Fisica Nucleare, Sez. di Pisa, Italy
** Current address: CERN, ECP Division, CH-1211 Geneva 23, Switzerland

– *Test Analysis:* This s typically run during data-taking, in parallel with online production. Here, the data access pattern is very different and may be purely random.
– *Off-line Production:* After data taking a final production is made, accessing all the data gathered during the physics run in time-order.
– *Off-line Analysis:* Compared to the test analysis, the number of users is greater and longer sequences of consecutive bursts will be read using the results of the off-line production.

3 User View of the Data

The summary output produced by the online or off-line production of NA48 will contain indexing information for accessing the raw-data. We will provide a mechanism, transparent to the user, for retrieving raw-data for selected records by using this index. We shall also enable the user to build second-level indices by selecting records from the summary files. The user filters out the information of interest at the highest possible level and extends his knowledge about these objects by stepping down the hierarchy (user index, summary files, raw-data). Thus, full information has to be retrieved only for a small subset of the data.

4 Storage Management

Due to the requirements imposed by the NA48 experiment, the main goals in designing our storage management system are:

– *Real-time behaviour and guaranteed throughput:* It must be guaranteed that the incoming raw-data and the results of the online reconstruction can be written to disk and tape. Write accesses have priority over any user request.
– *Reliability and fault tolerance*
– *User transparency:* The user has to consider "files" or even abstract "data objects" and not worry about the media and how to access them.

The SMS is intended to be used on the CS-2, but it is easily portable to any other hardware platform on which the required I/O performance is guaranteed. Figure 1 gives a rough overview of the proposed system. The depicted structure is only a logical one. It does not make any assumptions about assigning the processes or their interactions to physical hardware.

There is a clear border-line between the SMS and the rest of the Data Recording System. The only interactions between the two parts are data access to the PFS and user requests to the active instances of the `Storage Server` and the `Directory Server`. All these user requests must be handled through the `Client API`, which provides a clearly defined user interface, independent of the implementation. The `Input Manager` interacts with the SMS like any other user process, but having higher priority then all others. Some other user processes which run in real-time (online reconstruction) are granted a high priority level too, because their output data must also be written within a given amount of time.

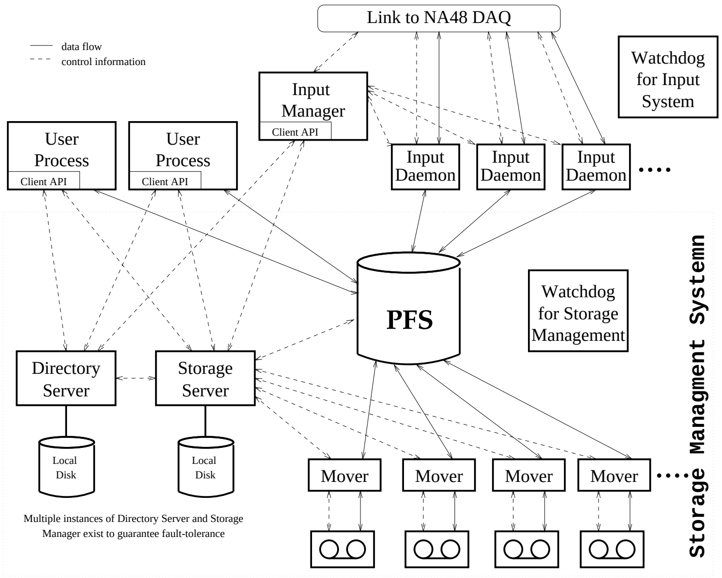

Fig. 1. Logical Structure of the Storage Management System

5 First Experiences and Conclusion

For the 1996 NA48 data-taking period, a prototype of the SMS was implemented. The system performed as expected, allowing us to store 1.4 Terabytes of raw data in 130000 files. Running at nominal intensity, the inflow of raw-data was 6-7 MBytes/sec. Data was written to tape using up to 7 Quantum DLT 2000 tape drives in parallel. The SMS is fully integrated into NA48 off-line applications and used heavily by the physicists doing their analysis (about 100 queries/day).

We try to provide a single system for storage and data management, integrating the recording of the raw-data received from the data-acquisition system of NA48 with the data accesses of end-users. Users should be able to access selected information efficiently via indexing mechanisms, and to build their own customized subsets of the data.