DD

CERN - CN 96 - 003

su 9612

CERN-COMPUTING AND NETWORKS DIVISION
CN/96/3
Fevrier 1996

CERN LIBRARIES, GENEVA

CERN-CN-96-003

Remote Data-Recording and Processing for NA 48

F.Gagliardi

Presented at Chp'95 Computing in High Energy Physics Conference,
Rio de Janeiro, September 1995

REMOTE DATA-RECORDING AND PROCESSING FOR NA48

F. GAGLIARDI

CERN, 1211 Geneva 23, Switzerland

Direct fibre data links have been installed to connect the NA48 experiment to the CERN Computer Centre for remote data-recording, 3-level filtering, and online event reconstruction on the Meiko CS-2 and the central CORE services. The Meiko CS-2 is a European parallel and scaleable supercomputer installed at the CERN Computer Centre in the framework of the EC-supported, Esprit Project P7255, GPMIMD2. This computer is a key element of the NA48 data acquisition system because of its parallel and scaleable I/O capability, parallel file system, and large computing capacity (in excess of 1600 CERN computing units). This sophisticated data acquisition system was necessitated by the experiment's high data rate (7 Mbytes/sec initially and 19 Mbytes/sec in 1997) and high volume (40 Tbytes of raw data in the 120 days of physics run foreseen in 1997). By recording the data remotely at the Computer Centre, the experiment will be able to profit from ongoing technological progress in tape recording and tape library technology, which are extremely important in coping with the above-mentioned online data requirements.

1 Introduction

NA48 is a new experiment which studies CP violation by measuring ratios of K^0 decays to the precision level of 10^{-4} . It is the most recent of a generation of experiments (among them NA31) that has developed a very sophisticated detector. This detector has recently started commissioning and it will be in full production by 1997.

The high-precision nature of this experiment requires high statistics both for simulated and collected data.

For the simulation both conventional packages such as NASIM (the NA48 version of Geant) and a special fast event simulation program NMC are used.

NMC requires a fast turnaround to produce several million events overnight. To achieve this a parallel processing approach has been developed based on a computing-farm model¹. A large memory and fast access to disk through a parallel file system are critical elements for simulation speed. This relies in fact on the use of large 'shower libraries' loaded into memory from disks.

2 The NA48 Experiment and its Data Processing Requirements

In order to estimate the required computing resources, we make the following assumptions:

- a trigger rate of 3000 K decays per sec after level 2;
- an average event size of 16 kbytes;
- no level-3 rejection.

The assumptions correspond to a data rate of 7 Mbytes/sec (500 events/sec), averaged over the SPS duty cycle, within the design capacity of the link to the Computer Centre. For 120 days of running at 50% efficiency the total data volume to be stored is 40 Tbytes.

In addition to storing all the raw data, the CS-2 is capable of processing them at these event rates. We estimate the reconstruction time to be 60 ms per event in a single 100 MHz Sparc processor, and therefore 32 dedicated processors are allocated to this task. While the event reconstruction time is based on experience with Monte Carlo simulations, there are choices in strategy and reconstruction algorithms which will only be settled after experience with the running periods of 1995/96. However, we do not expect the reconstruction time to exceed 100 ms unless the trigger purity improves, in which case there would be compensation from a lower average trigger rate.

The calibration of the liquid krypton calorimeter is particularly demanding and will require analysis of Ke3 events where the momenta in the drift chambers can be matched to the energy in the calorimeter for clean electrons. A full calibration of all 13 000 cells will require 24 hours of dedicated running with a modified trigger giving 2000 Ke3 decays per second; this will be complemented by continuous analysis of background Ke3 events recorded during normal data-taking. In both cases the real-time reconstruction pass will provide fast and continuous feedback on the calorimeter status and will ensure that up-to-date calibrations are used for the real-time analysis.

The software will involve inter-linked tasks for single-event reconstruction and for building and solving calibration matrices in different partitions of the calorimeter; this will make efficient use of a parallel architecture such as the CS-2.

It is foreseen to follow the real-time reconstruction pass by a second 'final' pass with full statistics and ultimate minimization of systematic errors and biases. This pass would require similar resources to the real-time pass.

During 1995 and 1996 there will be increasing use of our two complementary Monte Carlo programs NMC and NASIM to simulate and track the performance of the detector before and during installation in the K^0 beam. NMC can provide typically 10^6 events per hour on 15 processors of the CS-2. However, this facility should also offer 'instant availability' so that large runs of 10^6 to 10^7 events can be done with a fast reaction time. By contrast, the detailed studies with NASIM with a typical generation time of 10 s per event can usually be scheduled in advance.

With a fast link to the Computer Centre, a large data-storage capacity, and a powerful parallel processing architecture, it is possible to provide real-time storage and real-time reconstruction for NA48. The experiment can only succeed if all sources of systematic error and bias are continuously controlled to match the statistical error over three years of running at data rates approaching 10 Mbytes/sec.

3 The Computing System

The above physics requirements have imposed the design of a powerful data-acquisition and processing system with high bandwidth from the detector to the data-logging system and to the processing system.

Given also the large error bars in the estimates for the computing requirements, a high level of modularity and scaleability in the computing system is needed.

It was also necessary to provide a software development environment as compatible as possible with general HEP and NA48 programming standards.

When this system was starting to be designed (in early 1993), a new consortium of European supercomputer industries and CERN provided a powerful and scaleable parallel computing platform: the Meiko CS-2.

This offered the right characteristics: familiar FORTRAN development environment under standard SUN/Solaris UNIX system, large distributed memory, powerful parallel file system, high and scaleable I/O (several Mbytes/sec, about 19 Mbytes/sec required in 1997) and processing system (over 1400 CU).

The simulation program was implemented first. Based on a computing-farm approach it could benefit from the large distributed memory of the CS-2 (32 × 128 Mbytes) and fast access to the 'shower libraries' on the disks through the Meiko native parallel file system.

In order to implement an online calibration system, a direct fibre connection between the experiment and the CS-2 at the Computer Centre was installed.

In early 1994 the offline simulation system was successfully implemented and demonstrated. This proved the high level of performance and reliability of the Meiko CS-2.

Therefore a more ambitious system began to be discussed and investigated, while the design of the online calibration system was under way (Fig. 1).

The idea was to try to collect the data directly on the Meiko CS-2 from a series of front-end workstations installed at the experiment and connected via HIPPI to the last stage of the detector read-out system (data merger).

Each workstation is a DEC Alpha 3000 with 320 Mbytes of memory running Digital UNIX and is equipped with a HIPPI-TurboChannel interface. The full system of four workstations is capable of acquiring a full spill of SPS data at about 87 Mbytes/sec from the data merger through a HIPPI switch. Data are immediately written to a local disk and transmitted through FDDI to the memory and then to the parallel file systems of the remote CS-2 at the CERN Computer Centre some 5 km away.

To allow time enough to transfer the spill data through the relatively low bandwidth of the FDDI link, these front-end workstations are used in succession to bring the required capacity of the link down to some 7 Mbytes/sec. When the experiment's requirements grow (about 19 Mbytes/sec are expected in 1997) the number of front-end workstations can be increased as well as the number of FDDI connections to the CS-2. The parallel nature of this machine allows in fact the scaling up of the input bandwidth by simply increasing the number of dedicated I/O nodes and FDDI interfaces.

In the mean time, development of high-performance networking will be monitored and HIPPI or 622 Mbytes/sec ATM could also increase the link performance.

Once on the CS-2, the raw data are immediately logged onto magnetic tapes (DLT 2000 now) using standard CORE system software and the CERN tape robotics systems. The same data are also written to local disc onto the Meiko parallel file system and made available to a partition of 20 processing nodes to run the 'offline'

Figure 1: NA48 online system

reconstruction program 'online' on them. The results of this first-pass analysis are also written onto tapes; once full confidence is gained in the 'offline' algorithms these may also be used to perform 3rd-level filtering of the raw data before writing them to tapes.

Once in full production the experiment is expected to produce a total of 40–50 Tbytes of data per 3-month data-taking.

4 An Example of Use of European Industrial Projects for HEP

The challenges of the computing requirements posed by this experiment together with the image of CERN as a prestigious technology showcase for IT industry attracted some leading European supercomputer firms (Meiko, Parsys and Telmat) to form a consortium with CERN to develop the computing system of this experiment as part of a wider R&D programme. This programme: General Purpose Multiple Processors Multiple Data, version 2 (GPMIMD2) has been approved and generously funded by the European Commission (EC) of the European Union (EU). Without their substantial financial help it would have been very difficult to deploy the necessary computing resources and in particular to hire the young scientists to implement the overall systems.

It has not been easy to reconcile the interests of the EC in supporting the competitiveness of European small- and medium-size enterprises with the research and development interests of CERN and NA48.

The first successful results demonstrate the unique role which CERN and HEP can play as demonstrator of advanced computing technology by offering the flexibility of a research environment together with severe, industrial-quality production requirements.

This project, along with several others of its kind, is an example of a fruitful convergence of interests between industry and research.

5 Conclusions and Plans for the Future

The system has been successfully demonstrated during the end of summer 1995 test runs, including graceful degradation of the performance with the failure of one of the front-end workstations and automatic recovery of the system in case of restart of CS-2 nodes and workstations.

5.1 Data Storage Technology

For the 1997 operation, the bandwidth required is such that it could be handled only by about 20 DLTs in parallel. The installed units will be useful for an emergency use at a fraction of the bandwidth, even if one could expect an improvement of this technology for that time (25Gbytes capacity and 5 Mbytes/sec bandwidth are predicted for 1996–1997).

Moving to higher performance recording systems and automated tape libraries will be considered according to the relevant technology developments.

A hierarchical storage management system (OSM by Legent) is currently being evaluated for possible implementation in the NA48 computing system.

5.2 *Links to the Computer Centre*

The NA48 data rates after 1996 are predicted to be two or three times the above figure of 7 Mbytes/sec. To attain these rates, a solution other than a single FDDI link should then be envisaged. Several FDDI lines could be multiplexed on the same fibre (using HIPPI or ATM as transport protocols) or additional fibres installed (according to financial and technical considerations).

Reference

1. J. Apostolakis et al., "First results from the parallelization of CERN's NA48 simulation program," in: High Performance Computing and Networking Conference, Munich, 1994, v. 1, eds. W. Gentzsch and U. Harms (Springer, Berlin, 1994).
2. D. Asbury et al., 'Remote data recording and processing for NA48', CERN, Cocotime, internal document, v. 3.2, August 16th, 1995.

Acknowledgements

This paper is based on several reports and contributions from the GPMIMD2 project and NA48. Most of the technical work has been done under the guidance and assistance of E. McIntosh, M. Metcalf and A. Norton by a group of EC-supported fellows (J. Apostolakis, L. Bertolotto, C. Bruschini, P. Calafiura and B. Panzer-Steindel). J. Joosten, R. McLaren and B. Segal and their colleagues have contributed to the network part of the project. L. Robertson and CERN senior computing management have supported the project during the different phases. K. Aspola, the GPMIMD2 secretary, has assisted in the preparation of this paper and the associated presentation.

Glossary

CORE

Centrally Operated RISC Environment, p. 1.

CU

CERN Unit of computing, roughly equivalent to the CPU power of VAX 6800, 4 MIPS, p. 3.

Geant

CERN-supported event simulation package, p. 1.

Meiko CS-2

parallel scaleable European supercomputer, p. 1.

NMC

New Monte Carlo, p. 1.

SPARC

SUN standard processor, p. 2.

SPS

CERN Super Proton Synchrotron, fixed-target machine with 2.4 s bursts separated by 12.2 s inter burst periods, p. 3.