General-Purpose Parallel Computing in a High-Energy Physics Experiment at CERN

J. Apostolakis[1], L. M. Bertolotto[2], C. E. Bruschini[1], P. Calafiura[1*], F. Gagliardi[1], M. Metcalf[1], A. Norton[1], B. Panzer-Steindel[1] K. J. Peach[2]

[1] CERN, CH-1211 Genève 23, Switzerland
[2] Department of Physics and Astronomy, University of Edinburgh, UK

Abstract. The CERN experiment NA48 is actively using a 64-processor Meiko CS-2 machine provided by the ESPRIT project GP-MIMD2, running, as part of their day-to-day work, simulation and analysis programs parallelized in the framework of the project. The CS-2 is also used as a data warehouse for NA48: data coming at high rate from the experiment are processed in real-time and written to DEC DLTs using the Meiko Parallel File System (PFS) as a high-speed I/O buffer.

The ESPRIT project GP-MIMD2 started in March 1993 and will terminate at the end of February 1996. Its goal is to demonstrate the use of a European general-purpose MIMD computer, the Meiko CS-2, for CPU and I/O intensive applications from both the academic and the industrial research communities.

CERN, as a leading partner in the project, exploits a 32-node CS-2 system produced by the British company Meiko Ltd, Bristol. Each node consists of a twin-processor board equipped with two 100-MHz Sparc processors, 128 MBytes RAM and 1-4 GBytes of disk. A substantial upgrade of CERN CS-2 to a 64-node configuration is scheduled for April 1996. The nodes are interconnected with a high-performance (50 Megabytes/s) low-latency (less than 10 microsecond) network developed by Meiko as part of other ESPRIT projects. The machine is connected via Ethernet, FDDI and HIPPI interfaces to the CERN network. The working environment for the end-user is a normal Unix environment(SUN Solaris). MPI, PARMACS and PVM message passing libraries are available for parallel programming.

1 The NA48 Experiment

The computer is used at CERN for advanced High-Energy Physics (HEP) applications. Currently one main user is the NA48 experiment[1]. This high precision experiment will study with high accuracy the violation of the symmetry of nature under the combined particle-anti-particle exchange and mirror reflection (CP symmetry). The violation of CP symmetry is believed to be at the origin of the matter-antimatter asymmetry in the universe. The experiment will compare

* corresponding author: Paolo.Calafiura@cern.ch (phone/fax: 41-22-767-8997/8920)

with permille accuracy the probabilities for long-lived neutral kaons (K^0_L) and short-lived neutral kaons (K^0_S) to decay in two-pion ($\pi\pi$) states[3]. The hearth of the experiment detector is the Liquid Krypton Calorimeter[2]. This calorimeter, at the cutting-edge of High Energy Physics detector technology, will allow to measure with highest accuracy the characteristics of $K^0 \rightarrow \pi^0\pi^0$ decays in an high-intensity, high-background environment. The 13K read-out channels of the calorimeter will produce the bulk of the experiment data at an average rate up to 15 MBytes/sec.

The experiment will run for three years starting in 1997 and is expected to collect a sample of 10 million events[4] producing a total of about 120 TeraBytes of data.

The NA48 experiment is a collaboration of more than 150 physicists coming from 16 European institutions. Half of them are currently using the CS-2 to run their day-to-day parallel applications. These simulation and data-analysis applications were developed originally by about 20 NA48 physicists and parallelized using the message-passing library MPI by the CERN component of GP-MIMD2, in collaboration with the CERN and Edinburgh part of NA48.

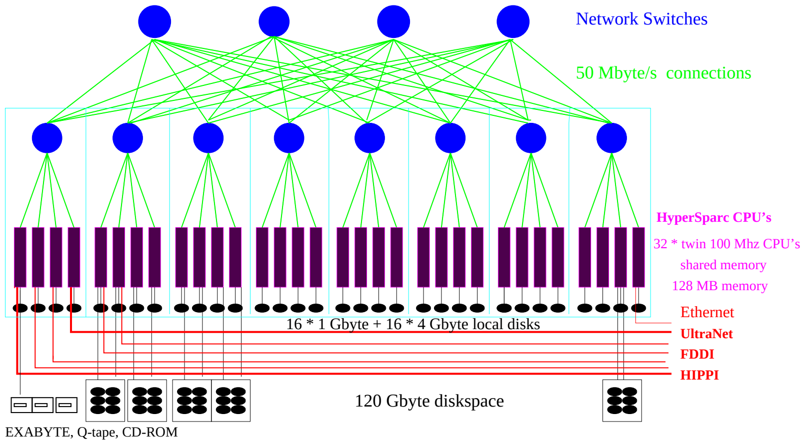

Fig. 1. The Meiko CS-2 in the configuration currently installed at CERN

2 Parallelizing the NA48 Simulation Programs

All the "physics code" that was parallelized is structured as a big loop over the experimental events. Each event is processed by following the track of each

---

[3] more precisely, the objective of the experiment is to measure the direct CP violation parameter ε'/ε with a 2×10^{-4} error [1, and ref. therein]

[4] an event, in HEP jargon, is the outcome of the high-energy interaction of particles and nuclei, as observed by the experiment detector

particle of which it is composed through the various sub-detectors. One can classify NA48 events into a dozen categories according to the underlying physical process. The data structures and the algorithms required to process different categories of events have often little in common.

Every month or two, there is a new release of the software that includes bug-fixes and physics enhancements contributed by at least 10 people, typically experts in a particular sub-detector who devote only a small fraction of their time to software development and are not necessarily familiar with parallel processing.

The nature of the problem and the NA48 software community strongly suggested a simple, user-transparent approach to the parallelization: a task-farm distributing (packets of) events to a set of event-workers. The parallel and the (sequential) physics code are thus clearly separated. This simplifies both the maintenance of the parallel versions and hides "obscure" MPI calls from the occasional developer and from the end-user.

The strategy used to parallelize the two simulation programs is described in detail elsewhere[3]. It is worth remembering that the two programs exploit different capabilities of the Meiko CS-2: the first one is a high-accuracy CPU-intensive application (30"/event/CS-2 processor), with relatively low I/O requirements. The second program, called *nmc*, is faster by two orders of magnitude, because it builds complete events using a library of sub-events simulated using the high-accuracy simulation and stored in a 1-GByte data-base with about 500K-entries. With an aggregated network traffic of less than 5 MBytes/s, for a 50 processors run, the network requirements are rather modest. The I/O latency is more critical because to build a complete event we have to extract from the data base up to eight sub-events per event. A low-latency "collision-less" network such as the CS-2's internal one is essential. Also it is important to parallelize the disk-read operations distributing the data-base over several disks. This is done in an elegant and transparent manner, exploiting the Meiko Parallel File System (PFS) capabilities. For a typical *nmc* run the I/O latency accounts for half of the processing time of an event. We can reduce this latency to negligible levels, running two *nmc* processes concurrently in each CS-2 processor, so that there is always one process running on CPU, while the other is waiting for its data base record. We have measured the event throughput of *nmc* when running with one and two processes per processor (Fig. 2). The event throughput with two processes per processor is effectively doubled, confirming that, in our reference *nmc* configuration, I/O latency was accounting for half of the event simulation time and showing that we can hide this latency almost completely. The linear scaling of the throughput with the number of processors confirms that our simple task farm approach does not introduce any significant overhead and that PFS bandwidth scales with the number of dedicated nodes.

3 The Central Data Recording

The NA48 application that exploits most the networking capabilities of the Meiko CS-2 is the Central Data Recording system (CDR)[4]. The idea is to

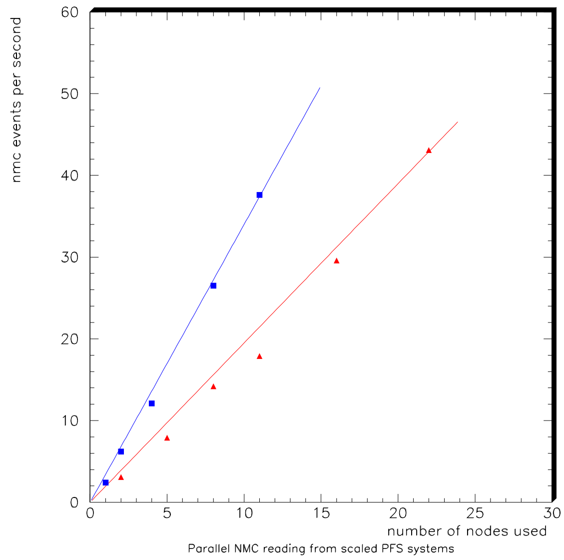

Fig. 2. Parallel *nmc* event throughput reading from scaled PFS, triangles refer to the case of running one *nmc* process per CS-2 processor, boxes to the case of two *nmc* processes per CS-2 processor

collect, process and store the experiment data at the CERN computer centre using the CS-2 as a data warehouse.

Data come from the detector read-out processors in bursts of 2.5 sec, followed by a 13 sec interval with no data. During each burst, detector data flow at about 87 MBytes/sec through an HIPPI switch and are stored in the 320 MBytes memory of one of the DEC Alpha 3000 workstations of the experiment's Data Acquisition System (DAQ)[5]. Here the data from the burst are written to a disk file, possibly in a compressed format. In a traditional approach, this disk file would be copied immediately onto one of the tape units attached to the DAQ system. Instead we take advantage of a 5 Km-long optical fibre that has recently been laid out between the CERN North experimental Area, where NA48 is located, and the CERN computer centre, to transfer the data in real-time to the CS-2.

3.1 First Experience

NA48 ran in September 1995 with all detector components except the Liquid Krypton Calorimeter. The average data recording rate was 2.5 MBytes/sec (only a fraction of the 19 MBytes/sec foreseen for the final configuration).

The data were sent to the CS-2 on an FDDI link, with a measured bandwidth of 9 MBytes/sec. On the CS-2, the data were written on eight Parallel File Systems. Each PFS consisted of 4 disks for a total of 10-12 GBytes with a transfer bandwidth of 8-10 MBytes/sec. The 32 disks were attached to 12 CS-2

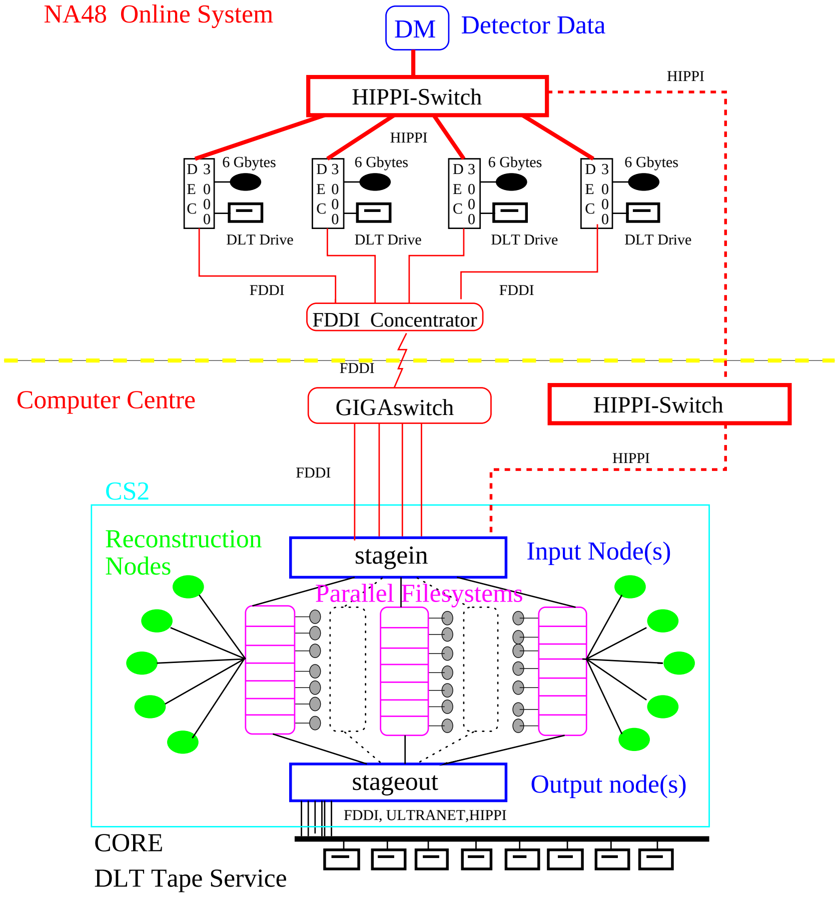

Fig. 3. The Central Data Recording system for NA48 experiment

nodes dedicated to the data recording system and to a first real-time analysis of the data quality. The rest of the nodes were split in a login and a parallel partition from where NA48 users could run their own analysis code on the data being taken.

Although the I/O capability of a single PFS would have been adequate to cope with the incoming data rate, we preferred to use eight of them to spread the load over more nodes and to increase the system resilience to disk or tape-unit failure. The CS-2 90 GBytes disk space, used as a circular buffer, allowed us to keep the data files on disk for at least 12 hours, giving a comfortable safety margin to detect and solve problems with the data recording system. In a month of run we experienced only two noticeable slow-downs in the data processing. The first occurred because the 8 DLT units were fully occupied by one NA48 user who accidentally attempted to read back two thousand files stored on several

DLTs in one go. The second was caused by a CS-2 node serving one of the PFS that hung and did not restart automatically. The operating system, attempting to re-mount that PFS, slowed down considerably every disk operation on the system and in particular the deletion of older files from disk. In both cases the big disk buffer gave us plenty of time to intervene. It is worth stressing that we did not have a single problem with the FDDI connection, and the back-up DLT units attached to NA48 DAQ have never been used.

This was a première for CERN. Traditionally, experiments have written data using tape units located at the experiment site. Every few days tapes were physically transferred to the computer centre tape-vault from where they could be read and analysed.

Making the best use of the computing resources available at the computer centre, the Central Data recording concept gives to the average NA48 physicist the possibility to run his analysis programs from the familiar CS-2 working environment on the data taken a few minutes earlier. Many more people than in the past can therefore contribute to checking the quality of the data that are being taken. This has played a non-negligible role in the success of the September 1995 setting-up run of NA48: in a month of data-taking the experiment collected 600 GBytes of physics data that are now used to perform detailed tests on the detector and several analyses on the rates of occurrence of three interesting particle disintegration processes.

3.2 Evolution towards the Full System

This summer there will be a new run of NA48, mainly devoted to commissioning of the Liquid Krypton Calorimeter with data rates up to 7 MBytes/sec. We will test the final system that in Summer 1997 will be used to store 40 TeraBytes of data flowing at an average rate of about 19 MBytes/sec.

The CS-2 fast internal network, its CPU power and its disk I/O capability are already adequate. However we need to increase both the link throughput and the tape I/O speed.

The FDDI link could be replaced by an HIPPI connection that has already been tested at CERN[6] to provide more than 80 MBytes/sec transfer rate. On the other hand, native HIPPI is a low-level, point-to-point unidirectional link and currently, there is no implementation of a high-level protocol (such as FTP) for Digital Unix. For this reason we favour the option of installing several FDDI links. These could be multiplexed over a single fibre running either HIPPI or ATM, or more fibres could be installed. Technical and financial considerations will dictate the final choice.

As far as Data Storage is concerned, if the 1997 requirement were to be met using DLT 2000[5] drives as in 1995, we would need about 20 of them in parallel. On the other hand, we expect that storage performances will increase to provide a transfer rate of about 5-10 MBytes/sec by 1997. For this reason one

[5] Quantum/DEC DLT2000 are 10GBytes linear recording magnetic tapes capable of 1.2 MBytes/sec transfer rate

of the design criteria of our system is the capacity to integrate and profit from evolving computer technology. In particular the software model must maintain transparent, reliable and efficient access to data at different storage levels. A user-interface to access the data stored on permanent media and on disk in a reliable and transparent fashion is also essential, especially since we want to protect innocent users from the risk of misusing the system during data-taking.

4 Conclusion

The EC GP-MIMD2 project is continuing its collaboration with NA48 to enhance the performance of its parallel applications and of the Central Data Recording system, to be able to cope with the vast amount of data expected during the next few years: if NA48's schedule is maintained, by the end of 1997 NA48 physicists will have collected tens of TeraBytes of data that will be stored, retrieved, analysed and simulated using the CERN CS-2 computer.

References

1. G. D. Barr et al, *Proposal for a Precision Measurement of ε'/ε in CP violating* $K^0 \rightarrow 2\pi$ *decays*, CERN/SPSC/90-22, SPSSC/P253, July 1990
2. G. D. Barr et al, *Performance of an electromagnetic liquid krypton calorimeter based on a a ribbon electrode tower structure* , Nucl. Instr. and Meth. **A370**, 413-424, (1996) `http://www.cern.ch/NA48/Welcome/papers/Overview.html#95`
3. L. M. Bertolotto et al, *Feasibility studies for a High Energy Physics MC program on Massive Parallel Platform*, in: CHEP '94, Computing in High Energy Physics Conference, San Francisco, 1994, ed. S. C. Loken
   `http://www.cern.ch/GPMIMD2/papers/laura.ps`
   J. Apostolakis et al, *First Results from the Parallelization of CERN's NA48 Simulation Program*, in: HPCN '94 High Performance Computing and Networking Conference, Munich 1994,v. 1 ed. W. Gentzsch and U. Harms (Springer, Berlin, 1994) `http://www.cern.ch/GPMIMD2/papers/hpcn94.txt`
4. F. Gagliardi, *Remote Data-Recording and Processing for NA48*, in: CHEP '95, Computing in High Energy Physics Conference, Rio de Janeiro, 1995, (World Sc. Singapore, 1995) `http://www.cern.ch/GPMIMD2/papers/CHEP95paper.html`
5. W. Bozzoli et al, *Data transfer and distribution at 70 MBytes/s*, in: RT '93, Conference on Real-Time Computer Applications in Nuclear, Particle and Plasma Physics, Vancouver, 1993, ed. D. Axen and R. Poutissou, (IEEE TR NS, 41(1), Feb. 1994)
6. A. Van Praag, *Testing a long distance serial HIPPI link for NA48, unpublished*, `http://www.cern.ch/HSI/hippi/longdist/shtest1.htm`