



ELSEVIER

Parallelism at CERN: real-time and off-line applications in the GP-MIMD2 project

Paolo Calafiura*

CERN, CH-1211 Genève 23, Switzerland

Abstract

A wide range of General Purpose High-Energy Physics applications, ranging from Monte carlo simulation to data acquisition, from interactive data analysis to on-line filtering, have been ported, or developed, and run in parallel on IBM SP-2 and Meiko CS-2 CERN large multi-processor machines.

The ESPRIT project GP-MIMD2 has been a catalyst for the interest in parallel computing at CERN. The project provided the 128 processors Meiko CS-2 system that is now successfully integrated in the CERN computing environment.

The CERN experiment NA48 was involved in the GP-MIMD2 project since the beginning. NA48 physicists run, as part of their day-to-day work, simulation and analysis programs parallelized using the Message Passing Interface MPI. The CS-2 is also a vital component of the experiment Data Acquisition System and will be used to calibrate in real-time the 13 000 channels liquid krypton calorimeter.

1. CERN and the ESPRIT project GP-MIMD2

The ESPRIT project GP-MIMD2 started in March 1993 and terminated at the end of August 1996. Its goal was 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. The project was a follow-up of the earlier GP-MIMD project which developed a transputer-based MIMD machine specially tailored for high performance real-time applications, such as event triggering for High Energy Physics experiments.

CERN, as a leading partner in the project, exploits a 64-node Meiko CS-2 system, provided by the project. Each node consists of a twin-processor board equipped with two 100-MHz UltraSparc processors, 128 Mbyte RAM and 1–4 Gbyte of disk. 400 Gbyte of external disk are attached to the machine.

The nodes are interconnected with a high-performance (40 Mbyte/s) low-latency (less than 10 ms) 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 envi-

ronment (SUN Solaris). MPI, PARMACS and PVM message passing libraries are available for parallel programming.

The other large HPCN platform available at CERN computer centre is the 64-processor IBM SP2. This is equipped with 66.7 MHz power processors and has an internal network providing 40 Mbyte/s bandwidth. The architecture and the performances of the two systems are comparable, but the SP2 is heavily used for generic interactive and serial batch services so that only 8–16 processors are normally available for parallel processing.

2. Communication benchmarks for LHC experiment trigger systems

General-purpose processors connected with high-speed networks can provide the necessary computing power for the trigger systems of the next generation of CERN experiments at the LHC proton–proton collider, due to run around year 2006. The most challenging of these application is probably the “second level” trigger. This trigger must reduce the incoming rate of 100 KHz candidate events by two orders of magnitude, combining information coming from different regions and from different components of the apparatus in the first global view of a candidate.

The EAST-RD11 collaboration has considered the possibility of using commercial HPCN systems for this

* E-mail: paolo.calafiura@cern.ch.

task, in view of their reliability and scalability. An implementation of the ATLAS experiment second-level trigger system [1] on the Meiko CS-2 showed that such system would be already able to reach 50% of the required performance today. Ideally such an implementation should be ported onto multiple potential platforms to measure their relative performance. Today's systems, on the other hand, cannot certainly be considered for the final solution, because of their price/performance ratio. Besides the algorithms to be executed have not been defined in any detail, and the detectors are still being optimized.

Rather than repeating the same complex exercise on every MPP platforms that will become available, it has been proposed [2] to perform a number of measurements on different existing MPP systems, to benchmark their performance in view of the realization of a "second-level trigger farm". The measurement includes seven basic inter-processor communications patterns, as well as three application-specific communication structures, inspired by the implementation on the CS-2.

3. The NA48 experiment and its central data recording system

The networking capabilities of the Meiko CS-2 are well exercised by NA48 Central Data Recording system (CDR) [3]. The idea is to collect, process and store the experimental data at the CERN computer centre using the CS-2 as a data warehouse. Since 1995 the system has been successfully integrated in the NA48 experiment Data Acquisition System. This high precision experiment will study with high accuracy the violation of the symmetry of nature under the combined particle–antiparticle exchange and mirror reflection (CP symmetry)¹. The experiment will run for three years starting in 1997 and is expected to collect a sample of 10 million selected events² extracted from a total of about 120 Tbyte of data.

Data come from the detector read-out processors in bursts of 2.5 s, followed by a 13 s interval with no data. During each burst, detector data flow at about 87 Mbyte/s through an HIPPI switch and are stored in the 320 Mbyte 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

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

² An event, in HEP jargon, is the outcome of the high-energy interaction of particles and nuclei, as observed by the experiment detector.

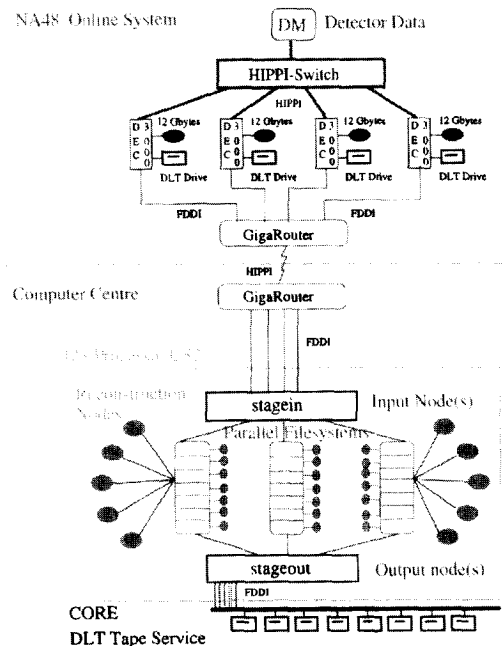


Fig. 1. The central data recording system for NA48 experiment.

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.

NA48 ran in September 1995 with all detector components except the liquid krypton calorimeter. The average data recording rate was 2.5 Mbyte/s (only a fraction of the 19 Mbyte/s foreseen for the final configuration). A second run has been performed a year after in August–September 1996 to commission the complete detector. Data rates were as high as 9 Mbyte/s. The data are sent to the CS-2 multiplexing 3 FDDI links over a single fibre, with a measured bandwidth of 19 Mbyte/s. On the CS-2, the data are written on 8 Parallel File Systems (PFS). Each PFS consists of 4 disks for a total of 16 Gbyte with a transfer bandwidth of 8–10 Mbyte/s. The 32 disks are attached to 12 CS-2 nodes dedicated to the data recording system and to a first real-time analysis of the data quality. The rest of the nodes are split in a login and a parallel partition from where NA48 users could run their own analysis code on the data being taken. The CS-2 128 Gbyte disk space, used as a circular buffer, allowed us to keep the data files on disk for at least 12 h giving a comfortable safety margin to detect and solve problems with the data recording system. 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. This has played a non-negligible role in the success of the two setting-up runs of NA48: in less than two months of data-taking the experiment collected 1.5 Tbyte of raw-data.

The multiple FDDI connection, the CS-2 fast internal network, its CPU power and its disk I/O capability are already adequate to deal with the incoming rate of 19 Mbyte/s of data expected for Summer 1997. As for as data storage is concerned, if the 1997 requirement were to be met using DLT 2000³ drives as in 1996, we would need about 20 of them in parallel. On the other hand, we expect that performance of available storage media will increase to provide a transfer rate of about 5–10 Mbyte/s by 1997. For this reason one of the design criteria of our system is the capacity to integrate and profit from evolving computer technology.

4. Parallelizing high-energy physics simulation programs

Two simulation programs used by NA48 were parallelized in the framework of GP-MIMD2 activities. The strategy used is described in detail elsewhere [6]. In a nutshell, we adopted a task-farm model with a master distributing (packets of) events to a set of event workers. The two programs exploit different capabilities of an HPCN platform: the first one is a high-accuracy GEANT-based application (30/event/CS-2 processor), with modest I/O requirements (typically less than a Kbyte per event). 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 500 K-entries. A low-latency “collision-less” network such as the CS-2’s internal one is essential to extract efficiently up to eight sub-events per event from the data base which is stored in a Meiko Parallel File System (PFS).

5. Real-time reconstruction and the calibration of NA48 liquid krypton calorimeter

Reading back from tape the 40 Tbyte of raw-data produced every year by NA48 should take several weeks and, due to limited tape resources, can only be performed

after the end of data-taking period. To give to physicists access to data during data-taking it is necessary to perform a systematic first-pass reconstruction while data files sit on CDR disk buffer. To achieve the required performance, the reconstruction requires in quasi real-time the calibration constants of the liquid krypton calorimeter response, with a few permil accuracy. To test the system we run the parallel version of *nmc* on 32 CS-2 processors, simulating 300 good calibration events per second, as produced by the experiment. The event simulation workers generated the events and passed the calorimeter data to 16 calibration workers, each one collecting the energies deposited in a 32×32 channels sector of the calorimeter. Each calibration worker performed independently the calibration of its calorimeter sector (1024 channels) using the standard technique called *energy resolution minimization* or *E/p method*. In a 5 h run, 50 million kaon decays and 2.8 millions useful calibration events were generated achieving the required inter calibration accuracy.

6. Parallel interactive data analysis

The Parallel Interactive Analysis Facility (PIAF), a client-server version of PAW originally designed to run on a 5-node HP cluster, was the first distributed-computing CERN library application. This was ported and optimized to the CS-2 [7], using eight nodes as a partition and a dedicated PFS. This has now become a production service and, in fact, so successful has this been that the CS-2 will host the only PIAF service from the end of 1997.

7. Conclusions

The strategy adopted at CERN to parallelize high energy physics applications, in particular off-line applications, has been a pragmatic one. This was also in agreement with the mandate of GP-MIMD2 project to port and run in production real-life applications on a commercial MPP platform.

Given the nature of the applications, and the sociological constraints coming from the large and very active community of users and developers, the typical programming model of choice was message-passing-based event or file farming. This has been considered as the best compromise among efficiency, maintainability and portability. A worldwide accepted, open standard, such as MPI or PVM, appears to be the best possible choice for 10–20 yr long global scientific collaborations like CERN experiments.

High-speed low-latency networking and fast, transparent access to vast amounts of data turned out to be the

³ Quantum/DEC DLT2000 are 10 Gbyte linear recording magnetic tapes capable of 1.2 Mbyte/s transfer rate.

relevant system features for most of the applications considered.

Acknowledgements

I would like to thank J. Apostolakis, L. Bertolotto, C. Bruschini, F. Carminati, F. Gagliardi, R. Hauser, E. McIntosh, M. Metcalf, B. Panzer-Steindel and G. Wirrer for the useful discussions. Special thanks to J. Walsh for his careful proof-reading of the draft.

References

- [1] R. Hauser I. Legrand, in: HPCN '95, Milan, 1995, Lecture Notes in Computer Science No 919, eds. B. Hertzberger and G. Serazzi (Springer, Berlin, 1995).
- [2] R. Bock et al., Benchmarking communication systems for trigger applications, ATLAS DAQ note 48 <http://www.cern.ch/RD11/combench/daq48.ps.Z>.
- [3] J. Apostolakis et al., in: HPCN '96, Brussels, 1996, Lecture Notes in Computer Science No 1067, eds. H. Liddell et al. (Springer, Berlin, 1996).
- [4] G.D. Barr et al., Proposal for a precision measurement of ϵ'/ϵ in CP violating $K^0 \rightarrow 2\pi$ decays, CERN/SPSC/90-92, SPSSC/P253, July 1990.
- [5] W. Bozzoli et al., in: RT '93, Conference on Real-Time Computer Applications in Nuclear, Particle and Plasma Physics, Vancouver, 1993, ed. D. Axen and R. Poutissou, IEEE Trans. Nucl. Sci. NS-41 (1) (1994).
- [6] L.M. Bertolotto et al., in: CHEP '94, San Francisco, 1994, ed. S.C. Loken, LBL 35822.
- [7] T. Hakulinen F. Rademakers, in: HPCN '95 Milan, 1995, Lecture Notes in Computer Science No. 919, eds. B. Hertzberger and G. Serazzi (Springer, Berlin, 1995).