THE LO.CO.MO.TION MONTECARLO FARM

E. Leonardi, INFN Roma, Italy

Abstract

The LO.CO.MO.TION project is devoted to the study and realization of a PC-based farm for MonteCarlo production in the L3 experiment at CERN. The farm, located in Rome, Italy, currently consists of 5 machines running the Linux OS. The system has been operational since June 1998 and produced over three millions of fully simulated MC events.

1 INTRODUCTION

1.1 MonteCarlo Production

MonteCarlo (MC) events production is one of the most important as well as resource demanding tasks in the computing structure of a modern HEP experiment: usually a number of events well in excess of a factor 10 with respect to the total experimental data is needed to avoid introducing errors related to low MC statistics.

In the L3 collaboration, one of the 4 LEP experiments at CERN, MC production usually proceeds in 3 well defined steps:

- **Event generation**: phenomenological equations are used to generate a large number of different kinematical configurations with the theoretical probability distribution. Production of a single event usually takes O(10 (s/event)×SpecInt95) and the output event size, consisting only in 4-vectors, is O(1 KB).
- **Event simulation**: events are allowed to evolve and to interact with the experimental apparatus. Each event takes O(600 (s/event)×SpecInt95) and the output, usually in a format identical to that of experimental data, is O(50 KB/event).
- **Event Reconstruction**: data from both the simulation step and the real data acquisition are analysed with the same algorithms in order to reconstruct the original event kinematics. Processing time is here O(60 (s/event)×SpecInt95) and the final data format is O(100 KB/event).

Event simulation is clearly the most computing intensive step and, given the almost complete independence of the task from calibration databases, it can easily be executed wherever CPU power is available.

In L3 requests from physics analysis groups are submitted to the MC production team which takes care of generating the initial kinematics of the needed events. These are then simulated either on a local MC production facility which uses FUNNEL [1] to access unused CPU power on the many graphical workstations of the collaboration or on several MC production systems all around the world. Simulated events are then shipped, either via network or via tape, back to CERN where they are reconstructed and made ready for analysis.

1.2 PC's

Modern PC's, i.e. machines based on Intel or compatible CPU's, offer very high performance, not far from high-end systems on integer computations, for a price tag as low as 50 US$/SpecInt95 for a fully functional system. They have of course a few weak points, namely the floating point performance is usually not nearly as good as the integer performance: a 450 MHz Pentium II CPU boasts a 17.2 SpecInt95 mark[2], quite close to the 18.8 SpecInt95 of a 600 MHz EV5 Alpha processor, but its floating point mark is only 12.7 SpecFP95, compared to 29.2 for the Alpha. Also, high-end workstations and servers usually have more than one PCI bus with speeds up to 528 MB/s while a commodity PC usually has only 1 PCI bus with a meager 132 MB/s bandwidth. Finally, the number of free interrupt lines on a PC is getting lower and lower as more and more new devices become part of a standard configuration.

All in all, PC's look like the perfect machine for running MC production: they give the best price/performance ratio on the market, event when FP performance is taken into account, while low I/O throughput or missing IRQ lines are not an issue for this task.

The only problem is the fact that the most popular operative system running on PC's, Windows in one of its many manifestations, is not much used in the HEP environment (this is changing, though): most, if not all, of the collaborations software is developed on Unix machines so that the porting to a MicroSoft OS is hard and usually beyond the manpower availability of an experiment.

1.3 Linux

A possible solution to this comes from the Linux[3] operative system: this is a Unix-like OS for Intel processors developed by Linus Torvalds in the early nineties. It now works on most modern architectures and supports the majority of the hardware on the market. Also, a huge number of enthusiast fans all around the world guarantee a very high level support.

Thanks to the GNU project[4], compilers for most languages as well as most network services (telnet, NFS, ftp, etc.) are available for free.

The final ingredient to make Linux usable for MC production is the availability of the GEANT package: this is a set of routines and libraries for the geometrical definition of an experimental apparatus and the simulation of all interactions of particles with matter. This package, together with a full set of libraries developed at CERN known as "the CERNLIB package" [5], was privately ported to Linux in 1996 and CERN itself announced its official support on Linux in 1997.

Shortly after the announcement of CERN support for the Linux platform, the L3 collaboration ported all of its software, including its MC production program SIL3, to Linux.

2 THE LO.CO.MO.TION PROJECT

2.1 *First Steps*

Within the L3 group in Rome we proposed a small project named Lo.Co.Mo.tion (Low Cost MonteCarlo Production) to investigate the use of Linux-based PC's for MC production in a HEP experiment by creating a small dedicated PC farm. This proposal was submitted to I.N.F.N., the Italian National Institute for Nuclear Physics, which financed it with a total of 18.000.000 Lit (∼11000 US$).

In November 1997 we acquired the first machine: a 200 MHz Pentium MMX PC with 64 MB of ECC RAM and a 1.7 GB hard disk. We used it in the following months to produce some 100000 hadronic events which were then compared to an analogous sample produced on the official MC facility at CERN, this to verify the consistency of the two samples. This test run also showed that MC production speed on this machine was ∼50% faster than a 99 MHz PA7000 machine running HP-UX.

2.2 *The Lo.Co.Mo.tion Farm*

Once the collaboration approved our results, we acquired 4 new machines, three mono-processor and one dual-processor, all based on the 266 MHz Pentium II CPU. One of the machines was connected to a 12 GB SCSI storage system and had two network interface cards in order to act as a front-end for the whole farm. All of the machines were connected to a 8 port Ethernet hub and used NFS to mount the disks of the front-end machine in order to access the MC production code and to store the production results.

This is how we organised the production process:

1. For each new production, composed of several sub-jobs, a job script and an input data file are placed in a conventional directory at CERN.
2. A cron job running each hour on the front-end machine checks this directory and down-loads any new file to the local disks.
3. A daemon running on the front-end machine splits the production into its many sub-jobs and submits each of them to farm machines as soon as they get idle (of course, the dual-processor runs two jobs at a time).
4. Output from the jobs is directly written to the central disks and is moved to an export directory as soon as the jobs end.
5. A cron job running each night on a machine at CERN checks for output and log files in the export directory and transfers them to CERN where they are saved to tape.
6. A cron job runs every morning on the farm checking for successfully transferred files and erasing them from the local disks.

We wrote all of the programs needed for this organisation using the PERL language, with the sole exception of the program which transfers output files to CERN: this is a FTP-like client-server system named fts/ftc and written in REXX by P.A.Marchesini at CERN. All scripts use ssh[6] to implement some degree of security on node-to-node interactions.

2.3 *Remote Monitoring*

Parallel to the development of the farm control system, a web-based monitoring system was implemented. This uses the MRTG (Multi-Router Traffic Grapher) package by T.Oetiker and D.Rand[7] which offers a graphical representation of several farm quantities such as CPU usage, network traffic and disk space occupation, on time scales ranging from one day to one year: a simple look to this graphs is often enough to spot a possible malfunctioning of the system.

Together with MRTG, we developed a CGI script which creates a summary page with information about the current farm status, showing if the daemons are running, which production jobs are currently submitted to the farm and which sub-job is being executed by which machine. All these data are hyper-linked to several log files in order to allow for further investigation should problems arise.

The status of the farm can be monitored by connecting to the Apache web server running on the front-end machine[8].

2.4 *Farm output*

As said before, each sub-job is tuned to create an output file with a size of ∼200 MB[1], so that each sub-job lasts from 12 hours to 3 days and produces from 2000 to 11000 events, depending on their physics content. This gives a daily total of about 2 GB which are transferred via the 8 Mbps link connecting Rome to CERN. Given current line status and occupancy, each nightly transfer takes up to 2 hours.

2.5 *Performance*

Table 1 shows relative performance of the machines. A 99 MHz PA7000-based HP9000/735 is taken as a reference while absolute times are measured on events of an arbitrary type.

Efficiency is the ratio between the CPU time used by the a single job and the real time elapsed. The HP is currently used for interactive work so that efficiency varied during the tests.

[1] This is a traditional size chosen when data were written to 200 MB IBM tapes. Today 20+ GB DLT tapes are used.

Table 1: Performance

Machine	s/event	Index	Eff.
PA7000 99MHz	54.2	1.0	variable
PentiumMMX 200MHz	36.0	1.5	99.9%
PentiumII 266MHz	22.3	2.4	99.9%
2 × PentiumII 266MHz	21.3/2	2.5×2	99.9%

2.6 Dual-processor machines

Dual-processor machines offer a better price/SpecInt95 ratio than single-processor ones as most of the hardware, i.e. case, power supply, hard-disk, NIC, etc., must not be duplicated to accommodate the second CPU.

However, the operative system used on the machine must support some kind of SMP (symmetric multi-processor) protocol in order to efficiently use the second CPU.

Since the appearance of kernel 2.0.x, Linux supports SMP machines[9]. Still the dual Pentium II machine we bought had several problems due to its Tekram P6B40D-A5 motherboard. These problems were solved with release 2.0.36 of the Linux kernel, December 1998, and now this machine runs even more efficiently than single-processors (see table 1).

2.7 Costs

Table 2 summarises the total costs of the farm. Here SI stands for SpecInt95 and the machine tagged “f.e.”, the front-end machine, includes the SCSI storage system, the additional NIC and an extra 64 MB of ECC RAM.

Table 2: Costs

Machine	KLit	US$	SI	$/SI
P.MMX 200MHz	2538	1586	6.4	248
P.II 266MHz (f.e.)	5212	3258	10.8	
P.II 266MHz (2)	5252	3282	21.6	152
P.II 2×266MHz	3760	2350	21.6	109
Hub+Cables	100	63		
Rack	1080	675		
Total	**17942**	**11214**	**60.4**	**186**

Looking to the US$/SpecInt95 ratio one can notice a 40% drop in only 6 months going from the Pentium MMX to the Pentium II machine and a 30% improvement for dual-processor machines. Should one buy a machine now, January 1999, good ratios could be obtained from dual 450 MHz Pentium II, 34.4 SpecInt95 at ∼70 US$/SpecInt95, or, even better, from dual 333 MHz Celerons, 26 SpecInt95 at ∼55 US$/SpecInt95.

2.8 Manpower

Of course, should the set-up and maintenance of the system require a lot of manpower, the TCO of the system would increase. In our case the set-up took about 2 weeks of work for a single person, most of which devoted to the comparison of the results with the official MC production, while, after a first period of fine tuning of the code, maintenance time is now below a couple of hours per month.

2.9 Scalability

All the local factors which could limit the scalability of the system, i.e. shortage of temporary disk space, of ports on the Ethernet hub, of LAN bandwidth, can be easily eliminated with the investment of small amounts of money. The only real bottleneck can come from the WAN connectivity: should the CPU power of the system increase of a factor greater than 3, the time needed to transfer all of the data to CERN would increase to a level which would interfere with the normal activities of the INFN section in Rome.

A temporary answer to this will be given by a bandwidth increase scheduled late this year but the only real solution to the problem would be moving to tape (or CD or DVD) based shipping of the data to CERN, with the associated growth in expenses and manpower.

3 CONCLUSIONS

The realization of the Lo.Co.Mo.tion farm showed that the use of Linux-based PC’s can be a very cost-effective answer to the problem of producing big amounts of MC events for modern HEP experiments.

Future plans for the farm, besides the acquisition of new machines to monitor technological trends in the PC market, include the use of a SQL database system to improve the handling of farm logistics and the creation of a web-based control system which would allow remotization of several tasks, including the possibility of stopping, restarting and rescheduling productions and jobs.

Given the results obtained from this project, a new PC-based farm composed of 19 dual-processor Celeron nodes was recently build in Njimegen (NL) and it is now heavily contributing to the MC production in L3.

4 REFERENCES

[1] `<//www-zeus.desy.de/∼funnel/TOP.html>`

[2] `<//www.spec.org/results.html>`

[3] `<//www.linux.org/>`

[4] `<//org.gnu.de/>`

[5] `<//wwwinfo.cern.ch/asdoc>`

[6] `<//www.ssh.fi/>`

[7] `<//ee-staff.ethz.ch/∼oetiker/webtools/mrtg/mrtg.html>`

[8] `<//pcl3farm.roma1.infn.it/>`

[9] `<//www.irisa.fr/prive/mentre/smp-faq>`