Large Systems and Specific Solutions
Session F Summary

Fabrizio Gagliardi [a], Akira Ukawa [b], Peter Wegner [c]

[a] *Computer Division, CERN, CH-1211 Geneva, Switzerland*

[b] *Institute of Physics, University of Tsukuba, Tsukuba, Ibaraki, 305, Japan*

[c] *DESY-IfH Zeuthen, Platanenallee 6, D-15738 Zeuthen, Germany*

1 Themes of the session

Many different themes have been developed in this session. However, a common denominator could be identified as "maximize physics output while reducing the overall cost of the computing systems".

Traditionally High Energy Physics (HEP) community has designed and implemented many different dedicated systems to achieve maximum performance in the various computing fields ranging from data acquisition and triggering systems to data analysis, simulations and lattice QCD. With the dramatic evolution of the computing technology in the last several years, a tendency towards the use of more commercial systems often based on commodity components has developed in all of the above fields as confirmed by most of the talks in all the CHEP97 sessions and in this session as well.

The major themes presented and discussed were the following:

- Large parallel systems for QCD
- Parallel systems and their benchmarks
- High performance switches
- Special custom chip developments
- RISC farms
- Software environments
- Management, software distribution tools and operation

We try to summarize the outstanding points of the themes in the following.

Preprint submitted to Elsevier Preprint *27 May 1997*

2 Large parallel systems for QCD

The three major lattice QCD projects around the world were presented, supplementing details to the review talk by R. Tripiccione in the plenary session[1]. A. Gara explained the status of the Columbia QCDSP project[2]. The APEmille, a project under development at INFN, Italy, with contributions from DESY, was covered by P. Vicini and E. Panizzi[3]. A. Ukawa described the Japanese CP-PACS system[4].

The above projects, all at different stages of maturity, are the continuation of a constant effort started in the mid '80s to achieve top computing performance for lattice QCD physics. This requires extremely powerful computing capacity and the use of commercially available supercomputers has traditionally been limited by costing factors. A number of dedicated machines were therefore developed. The aim has been to optimize the three fundamental factors of high performance parallel computing within the specific requirements of lattice QCD, namely, CPU power with high vector capability for large vector data, high network performance with low latency and high throughput, and fast I/O to disk. The three systems presented can be considered as part of the third generation of QCD focused high performance systems.

CP-PACS is already working having been completed in October '96. This system is based on a new RISC processor having vector capability that enhances the PA-RISC 1.1 architecture. The communication is performed via a 3D array of hyper-crossbar switches. The system is capable to deliver over 0.6 TFLOPS. The cost of the project was in the order of 18 MUSD. The system has been developed by Japanese physicists and computer scientists in collaboration with Hitachi Ltd.

QCDSP from Columbia University is well advanced and is planned to start to deliver 0.4 TFLOPS by fall '97. Building of another system with 0.6 TFLOPS is planned at BNL. This system is based on a commodity DSP processor from Texas Instruments. The network is a 4D torus with the ASIC gatearray interconnect switch developed by physicists of the project.

The APEmille associates specially developed hardware to implement a VLIW node architecture with a control network based on commodity PCs running Linux. The control network, a joint development of INFN and DESY, is based on the high speed serial link chip (Autobahn) from Motorola. The cost of the project is estimated in 6 MUSD range and the machine should become available some time in spring '98. The expected power is in the order of 0.5 to 1 TFLOPS.

Two different trends may be identified in these projects. One direction is to pursue simple systems very much specific to QCD applications as exemplified by the QCDSP project. Such systems can be realized at low cost, but generally harder to use. The other direction is toward more flexible systems which are easier to use but more expensive. The CP-PACS system belongs to this direction, while the APEmille is located somewhere in the middle of the two directions.

The fourth generation of lattice QCD systems should provide an order of magnitude more performance and deliver in the order of 10 TFLOPS. Computers of this level of performance will be developed by industry in any case. For instance P. Messina in his plenary talk explained the 10 TFLOPS goal by the end of the decade of the ASCI project. A serious question, however, is whether physicists can have access to these machines even if they become commercially available. In case this will not be the case, funding problems will have to be addressed to develop more QCD focused solutions. It remains to be seen of how much general use these machine will be and how interested industry will be in collaborating with physicists to develop them.

3 Parallel systems

Another class of large parallel computing systems was discussed in this session, which is a class of RISC based MPP systems where a good balance between CPU power, I/O and high scalability are one of the primary goals.

A successful example of this approach was described by E. McIntosh in his talk[5]. The computer, a QSW/Meiko CS-2, is being successfully used by the NA48 experiment at CERN to implement a high performance remote data acquisition and on-line reconstruction system. The most noticeable characteristics of the CS-2 are a commodity processor (SPARC), a commodity operating system (Solaris 2.5) and a very effective parallel file system. The interprocessor communication is a proprietary multistage network capable of 50 MB/sec and with a latency less than 10μsec.

Despite the excellent performance and reliability of these systems, they have had only a limited success in the wider commercial market and most of the installations have been in large defence and research laboratories. Very few customers seem to be prepared to pay a higher cost for the benefit of a proprietary high performance communication infrastructure. The trend is therefore towards the use of more and more commodity components even for the switch fabric.

The software has been another obstacle to a more widespread use of these MPP systems in large commercial and industrial markets where it has been always difficult to justify the effort to parallelize large legacy codes. HEP problems are, on the other end, intrinsically parallel and parallelism has always been used even if recently this term has become somewhat "deprecated".

An interesting development to approach the problem of performance measurement and bechmarking for a large class of different parallel systems was presented by R. Hauser[6].

The conclusion of E. McIntosh, after having reviewed several issues of distributed versus parallel systems and of many cheap processors versus few expensive ones, were that we should continue to develop parallel applications though some of them may not explicitly refer to a parallel approach.

4 Switches

Several high performance switches are still custom designed. An example is the Macramé discussed by R. Heeley[7]. Autobahn presented by H. Leich[8] is a good example of another trend: to develop, based on industry standard products, a more general purpose system with a generic PCI interface.

5 Special developments

Some other special developments in hardware and software, where real-time response is critical, as for example in trigger and decision systems, were presented. Neural networks, as the one developed for the CMS first level trigger system (presented by R. Nobrega[9]), seem to have become an accepted technology. They provide ease of use and flexibility in systems which are otherwise difficult to simulate and to describe with analytical models. Some times these are implemented with custom chips, as the Neurochip TOTEM presented by I. Lazzizzera[10]. Genetic Algorithms, as described by G. Organtini[11], also fall in this class of techniques.

6 RISC farms

This has become the most common approach for general computing in HEP today. They are used in the on-line data acquisition (i.e. BELLE), in off-line data analysis (IN2P3, FNAL, BELLE etc.) and for simulations. All the most common RISC workstations are used (Digital, HP, IBM, SGI and SUN). RISC farms have replaced the former mainframes in most of HEP computer centres.

The focus is now shifting to farms built with more affordable PCs. Therefore today's RISC farms could soon become "the dinosaurs of tomorrow".

Several presentations in this session covered the subject: IN2P3 farm for H1 (S. Dagoret-Campagne[12]), FNAL farm (S. Wolbers)[13], BELLE farm (R. Itoh[14]), and LAFEX farm (G. Alves, poster[15]).

7 Software

Software remains one of the most important challenge for progress in HEP computing.

The NILE Object Oriented and CORBA compliant system developed for the CLEO experiment (presented by M. Ogg and M. Athanas[16]) stressed the importance of the following issues:

- Code re-usability
- Fast design and ease of maintenance
- Standards and API
- Fault tolerance through object replication

8 Management, distribution and operation

The session was concluded by few talks which described special systems and approaches to minimize the cost of operating, maintaining and supporting HEP computer systems and libraries.

In particular, P. Defert described the ASIS[17] system to distribute HEP code, I. Reguero the SUE Unix management environment at CERN[18] and F. Hemmer reported on an attempt, so far rather successful, to outsource more conventional RISC system management and operations of the CERN central RISC farm to an outside service contractor[19].

9 Conclusions

The large variety of solutions described in this session confirmed a few important points in HEP computing:

- the need for increase of computer power for the next generation of experiments in USA (FNAL run 2, Babar at SLAC, etc.), in Europe (CERN, Desy and LNF) and in Japan continues.
- lattice QCD demands for high performance computing systems also stays as high as ever.
- budgets for research are shrinking almost everywhere.

To reconcile the above conflicting trends much ingenuity will be needed to develop high performance computing solutions with the best price/performance ratios and within reasonable budgets. Some of the approaches were identified as the use of more commodity and standard components, joint developments with industry, and clever use of new software technology such as OO. In any case HEP will not be short of interesting computing challenges in future.

References

[1] Raffaele Tripiccione *Overview on LQCD Computing*

[2] Alan Gara *The .8Tflop project at Columbia*

[3] Piero Vicini *The teraflop supercomputer APEmille: architecture review and project status report*
Emanuele Panizzi *Programming languages and environment for the APEmille parallel processor*

[4] Akira Ukawa *The CP-PACS Parallel Computer - Its Development and Application for Lattice QCD*

[5] Frederick McIntosh *Parallel Processing - Today and Tomorrow*

[6] Reiner Hauser *Abstract Communication Benchmarks in Parallel Systems, for Real-time Applications*

[7] Roger Heeley *Results from the Macrame 1024 Node Switching Network*

[8] Holger Leich *The Application of the AutoBahn Chip for Building a High Bandwidth, Low Latency Network of PC's or Workstations*

[9] Ricardo Nobrega *Neural Networks in HEP triggers*

[10] Ignazio Lazzizzera *The neurochip TOTEM: a case study in HEP*

[11] Giovanni Organtini *A Genetic Algorithm for optimization of selection algorithms in High Energy Physics*

[12] Sylvie Dagoret-Campagne *Reprocessing H1 Data on the IN2P3 computer farm in 1997*

[13] Stephen Wolbers *Next Generation Farms at Fermilab*

[14] Ryosuke Ito *The BELLE Computing Facility*

[15] Alves, Gilvan *The LAFEX Risc Based Farm Solution for CPU Intensive Applications in HEP*

[16] Michael Ogg *Nile Fast-Track: Fault-Tolerant Distributed Processing for CLEO*
Michael Athana *Nile's Distributed Computing Site Architecture*

[17] Philippe Defert *ASIS, a project to manage and to distribute application software in the HEP community*

[18] Ignacio Reguero *SUE: Standard Unix Environment at CERN*

[19] Frederic Hemmer *Managing outsourcing contracts for systems administration*