# TRENDS IN ARCHITECTURES
# FOR LARGE SCALE SCIENTIFIC COMPUTING

G. Meurant (CEA, Centre d'Etudes de Limeil–Valenton)
and
O. Pironneau (Université Paris 6 and INRIA)

February 1994.

## 1. INTRODUCTION

- *Supercomputers* like the Cray C90 cost around $ 30M and run today at a peak speed between 500 Mflops and 20 Gflops. They can be used by a lot of scientists in a computing center or through a fast network.

- *Massively Parallel Computers* like the Connection Machine CM5, the Intel Paragon or the Cray T3D cost less ($ 1M–10M) but can run very slowly if the application is not parallelizable or up to 100 Gflops when the algorithm is well suited to the architecture and is efficiently coded with a low level language. So far, these machines cannot be really used in a time sharing environment through a network and most of them are dedicated to very few users. However, it seems that this model of computation is one of the way of the future to build cheap, fast and easy to program supercomputers. The key issue for the spreading of massively parallel computers in the industry is the development of software (compilers, programming models, debuggers) for an easy use. It can be that a good compromise of the future will be to build Supercomputers with such a massively parallel machine as a specialized "functional unit" to handle the most parallel parts in algorithms.This is seen today in the cray T3D which has a Y–MP or C90 host.

- *Workstations* cost around ($ 50 000 US) and run at 10–100 Mflops. But they can hardly be used by more than a few people at a time. Workstation clusters compete with low end massively parallel computers. For machines with a real computing power like the IBM SP–1, the prices are around several $M.

An interesting trend is the disappearance of mainframes on the scientific market. Machines like the IBM 3090/600VF with 6 processors was way behind in CPU performance and had no real follower on this market segment.

## 2. MARKET REVIEW

Computers used in scientific computing are made by the USA and Japan, with only a few exceptions in Europe. It is not possible to list all of them because some are still in a development phase but we will briefly review the most important ones.

### US computers.

One should distinguish computers with a "vector" architecture from those with a "parallel" or "massively parallel" architecture although most machines are of some mixed type which makes a classification difficult. On a more rigorous basis the machines can be classified on the type of memory they have (shared or distributed) and the type of control (synchronous–SIMD or asynchronous–MIMD)

When there are more than one processor, the CPU speed is ambiguous because it relies on the parallelism of the computation and on the policy of the computing center (to allow or not some dedicated time with a single user using all the available processors). Manufacturers for their advertisements usually multiply the CPU peak speed of each processor by the maximum number of processors. This gives the absolute peak performance (that is to say the one that can never be reached !). Below we quote this peak performance together with a more reasonable average number which reflects the performance that can be obtained by a good programmer on a well vectorized and/or parallelized algorithm.

The dominant position is still today occupied by Cray Research Inc (CRI) but, today, they are challenged by some of the massively parallel computers manufacturers. Their products today are:

*Cray Y–MP, Cray C90, Cray T3D.*

They have a maximum memory that can go up to 1 Gw for the vector machines and about 10 Gw for the masssively parallel T3D. The number of processors goes from 1 to 16 for the vector machines whose peak performances vary between 300 Mflops for a single processor Y–MP and 16 Gflops for a full C90.

It must be said here that in most supercomputers, the main factor affecting performance is not the maximum speed of the processor but the bandwith between the vector registers or cache and the main memory and also the memory contention that can occur between processors for shared memories and the communication bandwith for parallel machines with distributed memories. From this point of view the Cray machines are well positioned, the C90 having a large memory bandwith and the T3D having probably the fastest network of today parallel computers.

In 1996 CRI plans to release the follow on of the C90 with up to 64 processors which is going to use a new technology and will be much faster than the C90.

The T3D has from 32 to 1024 processors (DEC Alpha) whose individual peak rate is 150 Mflops. The communication network is a 3 dimensional torus. Communication bandwith of 70 Mbytes/s have been observed with a very small latency less than 10 $\mu$s. The original feature of this machine is to have a memory that, although physically distributed, is globally addressable.

Another important issue for these machines is the software available. CRI is in good position offering Autotasking which is a way to automatically help the user parallelizing his program at the loop level. An interesting programming model (CRAFT) is available on the T3D merging data parallelism and data and work sharing. Therefore, on this machine two programming models will be available and can be merged together: message passing and CRAFT.

| machine | nb Procs | clock(ns) | memory(Gw) | peak(Gflops) | average(Gflops) |
|---|---|---|---|---|---|
| CRAY 2 | 4 | 4.1 | 0.512 | 1.9 | 0.5 |
| Y–MP | 8 | 6 | 0.256 | 2.66 | 1 |
| C90 | 16 | 4 | 1 | 16 | 6 |
| T3D | 32–1024 | 6.6 | 0.256–10 | 5–150 | 0.5–30 |

One should note that Seymour Cray (Cray Computer Corp.) has not succeeded too much with the Cray 3, but the company is still going on and working on the Cray 4.

There are a few companies competing very strongly on the massively parallel market. May be the most well known is Thinking Machines Corporation (TMC). Their current product is the CM5 ranging from 32 to 1024 processors (a 1024 processors machine is installed at the Los Alamos National Laboratory). The processor is a Sparc engine (32 Mhz) associated with a vector processor allowing a peak rate of 128 Mflops per node. Each node has 32 Mbytes of DRAM local . The data network is a fat tree with a 20 Mbytes/s bandwith. On practical codes a bandwith of 1 to 10 Mbytes/s is observed with a latency of 100 to 300 $\mu$s.

Although the CM5 is an MIMD architecture, it has been tailored to efficiently run the data parallel programming model of the CM2 which was an SIMD computer.

Intel, well known for its microprocessors, has entered the parallel market a few years ago with the iPSC. Nowadays, their product is the Paragon computer ranging from 66 to 1024 processors. The microprocessor that is used is the Intel i860 XPS (50 Mhz) whose peak speed is 75 Mflops.In fact two i860s are associated in each node, the second one handling the messaging operations. Node memory consists of 64 or 128 Mbytes of DRAM. The communication network is a 2 dimensional array. So far, only message passing is available on the Paragon.

Kendall Square Research (KSR) is marketing the KSR1 computer. This is an interesting and new architecture. Each processor, designed by KSR, has a peak speed of 40 Mflops. Processors are connected by a ring (32 processors on one ring) and rings can be interconnected together. The original aspect of this computer is that each local memory is considered as a cache and the data movements between the caches are handled by the system. Therefore, the user can see the memory as being shared although it is physically distributed, leading to a nice style of programming.

However each CPU of Version 1 being rather slow the overall performances are average only. The situation may change with Version 2.

IBM has recently released the SP–1 which is more a cluster of workstations (RS6000) linked by a fast switch than a real massively parallel machine. The follow–on, the SP–2, soon to be on the market will be more interesting.

Convex has also introduced a parallel machine based on HP microprocessors.

One must also consider potential candidates like Tera, the company founded by Burton Smith who was the designer of the Denelcor HEP, a long time ago, which seems to be an innovative architecture.

Finally in the fast changing world of workstations we will list two interesting machines:

*IBM RS 6000* which has several models differing from the processor speed and the cache and memory sizes. On these machines speeds of more than 50 Mflops can be reached.

The *Silicon Graphics* with 16 processors (MIPS based) which has a peak of 50 Mflops per processor and a shared memory. Notice that the largest problem ever solved in CFD had recently been computed at the University Of Minnesota on a cluster of Silicon Graphics machines. The problem was a compressible turbulence analysis in a cube divided in 1024 cells in each dimension. A computational speed of 4.9 Gflops was obtained.

One of the key issues in the design of massively parallel computers and rapid workstations is the increase in microprocessor speed. The contenders in this field during next years are Intel with the Pentium, DEC with the Alpha architecture, HP, the alliance between IBM and Motorola with the Power chip and finally MIPS. It is very hard to tell who will be the winner, but there will be probably only two architectures remaining.

When choosing a computer, it must be kept in mind that many companies have gone into bankruptcy these last years (ETA, SCS, Multiflow, Saxpy, Alliant, SSI, etc…), and there will be more in the near future, so one has to be careful about the future of such companies.

**Japanese Computers**

There is a strong competition for the fastest machine between the US (mainly Cray) and the Japanese manufacturers (particularly Fujitsu and NEC). However, the competition is not so strong in the worldwide market which is still heavily dominated by the US. Few Japanese supercomputers have been sold outside Japan.

The most interesting recent machines have been by *Fujitsu* (VPX/200) and *NEC* (SX3) All these computers operate at Gflops speeds and their memory range from 256 Mw to 1 Gw.

The Fujitsu VPX/200 is a series of computers with different performances. The high end model with a peak performance of 5 Gflops can have 4 scalar units sharing 2 vector units. The memory can have up to 256 Mw and an extended (slower) memory of 4 Gw. The speed is obtained by replication of the functional units.

The NEC SX–3 (known as the SX–X in the US) is a multiprocessor with as much as 4 processors. Each processor will have a 2.9 ns clock cycle sharing 256 Mw of memory. The peak performance is 6 Gflops per processor. Unfortunately the bandwith between the main memory and the vector registers is not large enough and it is likely that the average performance of this machine is much less than the peak.

These computers have also efficient vectorizing compilers which may enable the average users to run their program at say 30% of the peak speed without too much effort.

The Japanese manufacturers have been interested during the last years to put "a vector supercomputer on a chip".

Recently an interesting machine has been introduced by Fujitsu, the VPP500. This a parallel machine with distributed memory (that is globally addressable like in the Cray T3D). The processor has a vector architecture with a peak speed of 1.6 Gflops. There can be from 7 to 222 processors, the last one with a theretical performance of 355 Gflops. Each processor incorporates up to 256 Mbytes of SRAM. The VPP500 is interfaced with a VPX/200. Due to price of the machine, only moderately parallel (10–30 processors) versions can be envisioned.

There are also research projects in Japan on massively parallel machines mainly by NEC.

### European computers

There had been several attempts in Europe to make supercomputers but they have failed so far, except may be the DAP by AMT (UK) which was rather revolutionary for its time. Today, there is only the CS–2 by Meiko (in fact a European consortium called PCI made by Meiko, Telmat and Parsys) that can be considered as an industrial product. This is a massively parallel distributed memory machine having from 8 to 256 processors. The processor is a Sparc engine (cycle time is 20 ns) associated with 2 Japanese vector processors ($\mu$VP).Each node has a peak speed of 200 Mflops. The memory capacity is 32 or 128 Mbytes per node. The CS–2 network is a multi–stage packet switch.

Europe has also some other projects undergoing. Unfortunately, most of these projects are based on US microprocessors as the European industry is particularly weak in this respect.

## 3. NETWORKS

As we said before an important component of any large scientific computing environment is the network that links computers. Either that remote users have to access the supercomputer or that results of computation have to be transferred to graphic workstations. The important things to consider are the physical media and the protocols supported :

–Ethernet linking workstations can go as fast as 10 Mbit/s (average 1Mbit/s) on twisted pairs,

–Hyperchannels (NSC) link supercomputers at 50 Mbit/s

–FDDI (Fiber Distributed Data Interface) is designed to run at 100 Mbit/s

–HIPPI (High Performance Parallel Interface) previously know as HSC is aimed at 800 Mbit/s, for the moment only on short distances and for point to point connections,

–The future is thought to be ATM. Normalization is undergoing. It is expected tthat this technology can be used either in local networks or in wide range fast communications, the so–called information highways.

Therefore, there exist now partial solutions to the communication problems although this aspect of computing is not as well developed as computing machines.

## 4. CONCLUSION

Scientific computing will be striving for larger and larger computers for a long time more. This type of research is not a luxury as in some cases it is the only tool available to scientists.

Computers can be conveniently sorted into three classes: vector supercomputers, massively parallel computers and workstations (clusters). We feel that workstations are needed anyway to handle pre and post processing. Moderately large computing tasks can be computed either on a cluster of workstations or on a supercomputer; the choice is determined by the size of the team and the quality of the network. However very large tasks (hypercomputing) can only be run on a supercomputer and in order not to disturb other users the supercomputer should be partially devoted to hypercomputing. Again it can be done either on site or through a network but if computer images have to be generated today's networks are still too slow. In the near future massively parallel computers combined with powerful scalar processors will make hypercomputing cheaper and will provide the necessary increase in computing power.