

## Case Study

# Historical Perspective of Supercomputing

Table of Contents

Introduction ..... 2

Modern scalar-processors..... 2

Vector Operations. .... 3

Vector Processors..... 4

Japan enters the supercomputer fray. .... 6

Massively parallel processors. - (MPPs)..... 8

Parallel computers available commercially, in 1990. .... 9

Modern supercomputers – acquire a new taxonomy..... 13

The NEC SX-6 and IBM Power4..... 13

## **Introduction.**

Seymour Cray, the father of supercomputer design, used to say with tongue in cheek that “computer designers are glorified cooling engineers” and that he was a very good plumber”. This is of course an oversimplification, but at the same time very incisive. A computer, as everyone knows, consists of a large number of electronic and mechanical components. In order to obtain maximum speed it is necessary that these components are packed as closely together as possible. Although each individual electronic component works on very low electric voltages and current, the large number (millions to make a computer) consume a substantial amount of electricity. Most of this electric energy is dissipated as heat. The extraction of this heat from close proximity devices has been and indeed still is presenting a great challenge to supercomputer designers.

A modern supercomputer has a large number of independent structures, (such as scalar and vector functional units, central memory and possibly secondary memory, instruction registers, internal channels to connect registers to memory, cache memory and instruction buffers, several central processing units, CPUs, a network or crossbar bus to link CPUs together, an Input Output subsystem including I/O channels and so on) and their ability to interleave during processing provide a substantial amount of parallelism in the total system.

The objective of every computer designer is to produce a computer that performs calculations more cheaply than its rivals. As large computer systems are very complex, the success of their architecture is judged on how well they achieve this objective. This depends to a great extent on the technology used and the ingenuity of the design to deliver a balanced system, and whether the parallelism available in the system architecture can be exploited easily during the solution of large problems. Unbalanced machine architectures pay heavy performance penalties whenever an application highlights a bottleneck in the total system. Hardware engineers always have to make compromises, a trade-off between various system components and their cost. The task of software engineers and application programmers is to get round bottlenecks and to maximize the exploitation of parallel features in the hardware.

## **Modern scalar-processors.**

Modern computers owe their existence to the invention of the silicon transistor in 1961. Early computers needed miles of wiring and made them error prone. Even in the seventies systems programmers and engineers had access to all the wiring diagrams of computers such as the Control Data Corporation’s CDC 7600 and the IBM 370. In fact the patent documentation of the IBM 370 associated with Gene Amdahl, consisted of 12 volumes, together they measured about one meter in thickness. Every “AND”, “OR” logic gate and back panel wiring was described.

The arrival of the silicon chip revolutionized electronic circuit construction and computer architecture. The need for miles of wiring was overnight consigned to history. Circuit logic boards a square foot in area were transformed and packed into a square centimetre chip. Power consumption and hence heat dissipation was also reduced dramatically. More powerful computers could be built with reduced physical size. As silicon chips were needed in bulk their fabrication was mechanized, which produced economies of large scale. Already the trend in electronic components was towards larger scale integration and assembly, smaller size, higher speed, lower energy consumption, and greater reliability. These advances have produced larger memory

capacity and made it possible to increase the complexity of tasks that can be performed by a computer.

One of the earliest computer manufacturers in the U.S.A. concentrating on scientific computing was Control Data Corporation. It was founded in the late 1950s by a small group of computer engineers, included amongst others was Seymour Cray and James Thornton two of the personalities heavily involved in the design of modern scientific computers.

The first modern scalar computer, one with a respectable balance of integer and floating-point processing power, was the CDC 6600 introduced in 1964. The CDC 6600 was the first RISC machine, a supercomputer from 1964 to 1970 and the earliest precursor to the Cray family of supercomputers. It had simple format instructions that were fast to decode. It had a register-to-register format for operations backed by a load/store structure with one level of indexing from a fixed base address. While the average instruction completed in three clock cycles, there were many one and two cycle instructions. It had multiple functional units and provided effective parallelism by the use of an IOP subsystem. The IOP consisted of a number (from as few as four to as many as 16) of small processors.

In 1970 the CDC 7600 and its competitor the IBM 360/195 were vying for first place in the supercomputer stakes. The supremacy of the CDC 7600 was very soon translated into sales statistics (over 80 were sold). After that IBM lost interest in the technical market and concentrated on more lucrative commercial mainframe production.

The CDC 7600 was designed as a successor to the CDC 6600 with a relative four-to-fivefold increase in performance. This machine was still a serial computer without vector pipes and registers. Some parallelism was however possible, as it had independent and partially segmented functional units. It is this attribute of segmented functional units that allowed the important development of vector processing. Other important characteristics of the CDC 7600 were a CPU with, in its time, a very fast (27.5 ns) clock period, a three-level memory arranged in a hierarchy of both speed and size, and three sets of operating registers to move data from the CPU into and out of functional units. So having corrected the floating-point weakness of the CDC 6600 and built using faster technology, it became the workhorse for scientific technical computation from 1970-76.

The CDC 7600 was capable of speeds higher than 10Mflop/s, in particular where instructions could be overlapped. Furthermore, the highly optimised Fortran compiler and the easy-to-use software available enabled this high-performance engine to deliver scientific technical results an order of magnitude faster than its contemporary commercial mainframe computers.

A scalar is a single quantity or value, so a scalar operation is one that works on one pair of operands at a time and produces one result; for example adding two numbers together and storing the result. This is the mode of operation of most general-purpose computers.

## **Vector Operations.**

A vector is a group of values that are accessible under certain machine-dependent conditions. Vector operations involve sets of operands. Using the preceding scalar example, a similar single vector operation would add two sets of operands in pairs, producing a set of results. The additions do not take place simultaneously, as in array processing, but are pipelined. What pipelining means is that as the computer's functional units are segmented, the instruction execution can be

done in steps; each step normally takes one clock period to complete. The pipeline has first to fill up before a result can flow. The time taken to fill the pipe is dependent on the number of segment steps for a particular functional unit. Once the pipeline is full, results are delivered at the rate of one result per clock period. Results may either end up in vector registers or may be stored directly into memory as they are produced. Where vector operations are delivered, is architecture depended and this often has implications on performance.

The delay before the first result is obtained is known as the start-up time and is always equal to functional unit time. Once the first vector operation has been issued the next vector operation cannot be issued until all the results have been obtained. Because a number of pipes exist servicing a number of independent functional units, it is possible to produce results at a rate greater than one result for every clock period. The NEC SX-6 for example, has 16 pipes and produces 32 results per clock period.

### **Vector Processors.**

For the last twenty-six years vector supercomputers and the Fortran language have been synonymous with scientific and engineering computation. They have withstood the test of time, because they are generally available, easy to use, efficient, and provide an evolutionary path protecting users vast computing investment.

Vector supercomputers have come a long way from the early Cray-1 delivered to Los Alamos in 1975. The 8Teraflop/s NEC SX-6 and the 40 Teraflop/s Japanese Earth Simulator manufactured by NEC and dedicated to Earth Systems Science are both vector and massively parallel. The Earth Simulator has 5,120 processors closely connected together. The Earth Simulator shows that success lies in the combination of vector and parallel. The recent LINPACK results put it in a class of its own, delivering 35.6 Teraflop/s over 87% of its peak performance. This machine is so powerful that it exceeds the raw processing power of the 20 fastest American computers combined and far outstrips the previous leader, the IBM-built ASCI White machine and has caused a lot of excitement in the US. The new Japanese Earth Simulator, the world's fastest supercomputer, was financed by the Japanese government and has been installed at the Earth Simulator Research and Development Centre in Yokohama.

For some American computer scientists, the arrival of the Japanese Earth Simulator supercomputer evokes the type of alarm raised by the Soviet Union's Sputnik satellite in 1957.

"In some sense we have a Computenik on our hands," said Jack Dongarra, a University of Tennessee computer scientist who reported the LINPACK achievement. For many years he has maintained the TOP500 list of the world's fastest computers.

The Earth Simulator and the NEC SX series supercomputers use the same technology and their architecture is based on vector processing, a way of using specialized hardware to solve complex calculations that was pioneered by Seymour Cray. As we will see later this concept has generally fallen out of favour in the United States in recent years. The Americans opted for building massively parallel supercomputers by chaining together thousands of off-the-shelf microprocessors. Several United States computer scientists said the Japanese machine reflected differences in style and commitment that suggest that United States research and spending efforts have grown complacent in recent years. "The Japanese clearly have a level of will that we haven't

achieved," said Thomas Sterling, a supercomputer designer at the California Institute of Technology. "These guys are blowing us out of the water, and we need to sit up and take notice."

This accomplishment is also a vivid statement of contrasting scientific and technology priorities in the United States and Japan. The Japanese machine was built to analyse climate change, including global warming, as well as weather and earthquake patterns. By contrast, the United States has predominantly focused its efforts on building powerful computers for simulating weapons, while its efforts have lagged behind in scientific areas like climate modelling.

From its earliest days supercomputing in the USA had an umbilical link to military needs. Supercomputers were and still are utilised as a tool for weapons production, and often the cost for their development comes from the US Department of Defence. The quest for supercomputing, was triggered, by the search for a solution to the hard-silo problem for anti-ballistic missiles, and simulation of nuclear devices. The first vector computer, the Control Data Corporation (CDC) STAR-100 with 100Mflop/s was delivered to Lawrence Livermore in 1973. Its rival Cray-1 was delivered to Los Alamos in 1975. This sequence of events was repeated with the Cray-2, and subsequent MPPs continuing to this day under the auspices of the ASCI programme. In the 1970s and 1980s some 70% of supercomputers were used for simulating weapon designs or other Department of Defence related activities.

In 1969, the giant General Motors company, signed an option to buy two STAR-100 machines, provided they were delivered by 1971 and that the machine supported ordinary business-type functions. Unfortunately the CDC STAR-100 was not ready for delivery in 1971, and General Motors cancelled the order and bought 5 IBM370/165 computers instead. The STAR-100 became fully operational in 1973. The data processing features, however, survived in its successor machines, the CDC Cyber 203, the Cyber 205 and the ETA-10.

In total, four CDC STAR-100 machines were built. Two of the STAR-100 machines were installed at the Lawrence Livermore Laboratory, one at NASA Langley Research Centre. The fourth one was installed at the CDC Minneapolis Computer Service Bureau.

The logical design of the STAR-100 consisted of one million 64-bit words ferrite core memory with a 1000-ns cycle time, and a central processor unit (CPU). The CPU consisted of a storage access control unit, a data stream unit, two multi-purpose floating-point pipes, and a general-purpose register file. The register file contained 128 64-bit registers, which could also be used as 256 32-bit registers when performing 32-bit half-word arithmetic. The vector pipes were connected to memory through a data stream unit and a storage control unit. There were no vector registers on the STAR-100, and therefore all vector operations were effected from memory to memory. The general registers were used to decode instructions and set up the memory data streams two input and one output, for vector operations. The control of data flow from memory to vector pipes and back was the function of the stream and storage access control units. With two vector pipes active the STAR-100 could either deliver two 64-bit results or four 32-bit results for every 40-ns clock period. Thus it was possible to satisfy the original requirement and deliver 100Mflop/s when working with 32-bit arithmetic.

The slow hardware of the STAR-100 reduced its effectiveness and delivered scalar results some four times slower than the CDC 7600. A long vector start-up of 76 clock periods (3040-ns) made vector operations slower than scalar when dealing with vectors of size less than 100. The STAR-100 was much larger physically than the CDC 7600 or its future competitor the Cray-1. The size problem was an added constraint in the performance characteristics of the machine.

In 1971 a hot debate ensued in the board of CDC to decide which design to pursue for future supercomputer products. The choices were the STAR-100 line or the CDC 7600 line. Seymour Cray favoured the 7600 line but the STAR-100 advocates, Neil Lincoln and Lloyd Thorndyke won the day. Seymour Cray left CDC and with a \$5 million from Los Alamos for the option to have the first machine, set up Cray Research Inc. to pursue his dream.

The Cray-1 architecture was an extension of that of the CDC 7600. The scalar part of the machine with its register-to-register design, instruction stack and independent functional units was retained. However, the peripheral processor units were discarded because of high costs only to be re-introduced in the Cray-XMP and successor Cray products. The design was extended with the addition of vector capabilities and the machine substantially speeded up by improving its various elements. It had one million 64-bit words of fast bipolar memory with a memory bandwidth of 80Mwords/s, and 13 functional units fully segmented. The scalar and vector processing was performed using a 12.5-ns clock period and with chaining the Cray-1 had a theoretical vector peak performance of 160Mflop/s. It was faster than the CDC 7600 in scalar and with a relatively short start up time it was three-to four times faster even with short vectors. Thus, when one compares the CDC 7600 and Cray-1 architecture and their performance characteristics it becomes clear that the Cray-1 delivered the improved performance needed to replace the CDC 7600 as the leading supercomputer in the latter part of the seventies. By the mid eighties Control Data lost its market to Cray Research Inc. and its attempt to revive its fortune in the supercomputing field using its offshoot ETA Systems had fizzled out by 1990.

### **Japan enters the supercomputer fray.**

The 1980's saw a number of new systems from the US, namely the CDC Cyber 205 in 1981, the Cray X-MP in 1982 the Cray X-MP/4, the Cray-2 in 1985, and ETA-10 and Cray Y-MP in 1987.

Japan first entered the supercomputer field with the Hitachi S810-20 in 1983, followed by the Fujitsu FACOM VP-200 in 1984, and the NEC SX-2 in 1985. The NEC SX-2 was the first supercomputer to be announced with a peak performance greater than 1Gflop/s (1.3gigaflop/s). Its designer Tadashi Watanabe has since become a household name in supercomputer circles throughout the world.

The NEC SX-2 used high density LSI logic technology. It employed 1000 gates per chip with a 0.25-ns propagation delay time as logic elements and a 1-kbit bipolar memory with a 3.5-ns access time for vector registers and the 64kbyte cache memory. The logical design of the NEC SX-2 consisted of 32 million 64-bit words MOS static RAM memory. The main memory was arranged in 512 banks allowing a maximum 8 words per 6-ns clock period; i.e. 1.3Gbytes/s. This was achievable not only for contiguous data but also for data spaced with a constant stride. In addition, the SX-2 could have up to 256 million 64-bit words of extended memory that could be used in a hierarchy of memories to hold intermediate results in large-scale scientific computations. The maximum transfer rate of data with main memory was also 1.3Gbytes/s.

The CPU consisted of up to 16 arithmetic units arranged in four sets each set of four units used one of the four vector pipes. A scalar pipeline with an additional set of four arithmetic units was also available. All these units were capable of parallel operations. In addition, eight vector load and four vector store pipes connected the main memory to the vector registers and another pipe serviced the scalar registers. A vector mask for compress/expand operations was also available. The operating registers consisted of 128 scalar registers, an 80kbyte vector register file and eight

256-bit vector mask registers. The cache memory was used as a fast memory to hold instructions and operands before execution. The SX-2 architecture allowed for four results per clock period for any one arithmetic operation since each pipeline had its own set of arithmetic units. Given that each of the four pipelined arithmetic units in each set could operate independently, a vector add and a vector multiply could be executed in parallel, producing four addition results and four multiplication results in one clock period hence the peak performance of 1.3Gflop/s.

As I said in my book, "Supercomputers and their Use", in 1985, "The SX-2 with fast scalar and vector arithmetic units, a short start-up time, high memory bandwidth, as well as parallelism in the architecture, had all the right attributes to deliver very high, sustained performance".

In 1984 a number of articles were published in the US press stating that the Japanese have, or are just about to achieve, supremacy over the US in the manufacture of supercomputers. Raul Mendez published two articles in the Society for Industrial and Applied Mathematics (SIAM) News and this triggered intense interest. Mendez stated that: "Supercomputing in Japan was the culmination of the MITI (Japan's Ministry of Trade and Industry) project, which began in 1976. The Japanese view supercomputing as an integral part of their long-term strategy in information technology and have become formidable competitors". Mendez's assertion that the Japanese will gain supremacy in the supercomputer field was immediately refuted by Peter Gregory, then senior Vice President of Cray Research, and Neil Lincoln from ETA Systems in statements reported in High Technology. They both exuded confidence that the US will keep its substantial lead. At the same time they used this assertion as leverage to extract a deluge of new funds from the Department of Defence for research in new supercomputers.

Significantly of the supercomputers promised in the 1985 to 1987 time frame with peak performance greater than 1Gflop/s none of them were from IBM or IBM compatible. Furthermore, the Japanese entry in this market was a major catalyst in broadening supercomputer use into the civilian economy sector.

Supercomputing, became one of the most important research tools by the late eighties. A report from the US President's office in November 1987, re-enforced this view:

"A strong domestic high performance computer industry is essential for maintaining US leadership in critical national security areas and in broad sectors of the civilian economy..."

A study I made for Higher Education Ministry in the UK in May 1989, found that there were about 374 supercomputers installed throughout the world and 17 on order. The USA had 167 Japan 90 while Europe had 92 of which 20 were in the UK, 23 in Germany and 21 in France. The total peak performance of these systems was 23Gflop/s. To put it in perspective, the 500<sup>th</sup> entry in the June 2001 TOP500 list is rated at 96Gflop/s. By year 2000 Germany had 40 systems whilst the UK was stuck at 20.

By 1990 supercomputers were having a profound impact on all branches of science and engineering. The economic pressure of getting a product to market ahead of the competition with a "cost /performance" edge became an active catalyst for the growth of the supercomputer industry.

In the past an improvement of computer power was achieved by speed-up of the electronic circuits. But 1990, the single fast processor supercomputers of the previous 20 years gradually gave way to parallel supercomputers with a relatively small number (4-16) of very fast, tightly connected CPUs. Examples of these machines were the Cray-2, Cray Y-MP, NEC SX3, and the Fujitsu 2000 series.

**Massively parallel processors. - (MPPs).**

Another path, emerging by 1990, which ten years later revolutionized capacity computers, consisted of developments, to build a high performance computer using thousands of relatively slow processors but harnessed to work in parallel. This radical change in computer architecture required an engineering feat to solve the massive connectivity problem and enable the production of a reliable parallel computer from a 1000 to 10000 times faster than today's machines. Once again, in the USA, the R&D phase was and still is partially funded by the Department of Defence in the guise of the ASCI programme started in the mid-nineties.

Certain types of problem lend themselves to massively parallel computation. These problems tend to be "granular" i.e. their small parts (granules) can be processed independently. Although this allows each processor to work on its portion of the problem without having to often exchange information with other processors one major difficulty to be overcome is how to control input/output and communication between processors for the whole system.

Parallel systems are not new. The ILLIAC IV, the Parallel Element Processing Ensemble, PEPE and the ICL DAP are examples from the mid-seventies. These machines failed because technology was not then robust enough to support their ambitious architecture. The industry had to wait for the more mature CMOS technology of the last fifteen years for commercial systems to come on the market.

These were exemplified by systems such as, the connection machine CM-2 and the Intel iPSC/2 from the USA, the AMT DAP, the Meiko Computing Surface, and the Parsys DN1000 from Britain and derivatives of the SUPRENUM project (Parsytec) from Germany and France. Most of these systems had their origins in national initiatives funded by DARPA in the USA, and ESPRIT in Europe.

For example, the original Connection Machine (CM-1) and the AMT DAP was Single Instruction Multiple Data (SIMD) fine-grained machine with a mixed topology. It was a grid of bit serial processors arranged into a hypercube with 16 processors on each node. The initial CM-1 announced in 1986, consisted of 65,536 one-bit processors. The CM-1 was a "massively parallel" system and thus differed from its contemporary moderately parallel hyper-cubes such as the Intel iPSC/2 and others mentioned above. The CM-1 machine designed by Dennis Hillis from MIT, and marketed by Thinking Machine Corporation. The CM architecture works well on "embarrassingly parallel" problems. These problems in general handle integer arithmetic and require very little communication between processors. Typical examples are, signal processing, imaging, calculation of prime numbers and Quantum Chromo-dynamics (QCD) theory. Any problem requiring floating point arithmetic was poorly handled and this includes many important classes of problems such as for example fluid dynamics, be it in simulating weather patterns, air flows over aircraft or cars, or the behaviour of fluids in nuclear reactors. Similarly protein folding for drug designs or any other problem using algorithms, which depend on integrating calculated values. The higher the integration of the calculated values, the stronger the connectivity of the underlying hardware must be if it is to deliver high efficiency. As we will see later microprocessor-based distributed memory systems also perform reasonably well with "embarrassingly parallel" codes but poorly on large-scale problems because of their memory limitations and communication weakness.

These systems produced the impetus for an enormous amount of work in parallel software in the late eighties, early nineties, but yet there was a lot to be done, and still is, in this area.

In the field of technology, it became clear that there was still plenty of mileage in silicon: both conventional CMOS and new massively replicated fabrications on wafers. Materials with faster switching properties such as GaAs, were promising to come into common use at the very top end of conventional supercomputers in a few years. By 1990, two strands were set to dominate the pursuit of more computer performance, faster switching circuits and the harnessing of parallelism in newer novel architectures.

### Parallel computers available commercially, in 1990.

In the late 1980's new machines with radical architectures came on the market which had a large number of relatively slow microprocessors harnessed to work in parallel. The more important of these machines are described in my book, "Supercomputers and Their Use".

A number of classifications are used to characterize computers. The Flynn [2] taxonomy distinguishes SISD (single instruction single data), SIMD (single instruction multiple data) and MIMD (multiple instruction multiple data) computers. A second classification is based on type of memory, whether shared, hierarchical or distributed. A third is based on the type of topology the architecture uses, whether the CPUs are tightly coupled, are arranged as a hypercube, a grid array, network or as a hybrid.

The three tables below give a flavour of the main parallel computers, with sixteen or more processors, on the market in 1990.

Company	System	Processing 1-bit Elements	Topology
AMT	510/610 DAP	1-4k	nearest neighbour
Goodyear	MPP	16k	nearest neighbour
Thinking Machine	CM-2	8k-64k	Hypercube

Table 1: SIMD Systems - (Array processor) – 1990

Company	System	Processing Elements	Connection
BBN	Butterfly	2-256	Network
Encore	Multimax	2-20	Bus
Flex	Flex/32	2-32	Bus
Isis (Fr	Isis	8-32	Bus
Myrias(Can)	Myrias	2-1024	Network
Sequent	Symetry	2-30	Bus

Table 2: MIMD Systems with shared memory – 1990.

Company	System	Processing Elements	Connection
Intel	iPSC/2	2-128	Hypercube
Ncube	Ncube-2	64-8192	Hypercube
Meiko	Surface	4-512	nearest neighbour
Parsys	SN1000	8-64	Hypercube/N
Parsytec	Super-cluster	64-512	Hypercube/N
IPS	IP-1	1-4096	Tree
Telmat	TN380	8-256	Hypercube/N

**Table 3: MIMD Systems with distributed memory – 1990.**

A number of other companies were known to be planning systems with sixteen processors or more.

In the supercomputing field it includes Cray Research with the Y-MP/16, Cray Computers with the Cray-3, NEC with the SX-3, Fujitsu, Hitachi, KSR, TERA, SSI (Chen), KEY and PRISMA.

In novel architectures SUPRENUM, MASPAR and TOPOLOGIX were planning to enter the market. It should be noted that the supercomputing field was very fluid; at least 14 companies have stopped producing “supercomputers” from the end of 1988 to 1990.

The 1990s saw a revolution in supercomputer use, which has fundamentally affected the future of science and technology: scientists and engineers are now able to model “reality” and generate new knowledge. In the same way as the steam engine fuelled the industrial revolution, so scientific and social progress, industrial competitiveness and the understanding and control of the environment are to be governed by the availability of adequate computing power.

Supercomputers are used in many companies to cut R&D costs and shorten the time required to get a product to market. Many manufacturing companies see them as supersavers, and have bought their own systems. More and more Small and Medium Enterprises (SMEs) are using time on university supercomputers – at university of Stuttgart, Germany for example - to solve design problems. The GRID portal infrastructure is likely to accelerate this practice.

The three tables below show examples of systems marketed from 1995 onwards. These include both shared and distributed vector parallel systems as well as distributed MPP systems.

Company	System	Number of CPUs	Memory Size	Memory Bandwidth	Peak Performance	Connection
Cray	T90	2-32			58Gflop/s	Crossbar
Cray	SV1-1	8-32	32GB	9.6GB/s	38.4Gflop/s	Crossbar
Cray	SV1	8-1024	1TB	9.6GB/s	1.2Tflop/s	Network
NEC	SX-4	4-32	4-16GB	to-64GB/s	64Gflop/s	Crossbar
NEC	SX-5	4-512	32GB- 4TB	to-32TB/s	4Tflop/s	IXS Switch
NEC	SX-6	2-1024	16GB- 8TB	to-32TB/s	8Tflop/s	IXS Switch
NEC	ES	5,120	10TB	40TB/s	40Tflop/s	IXS Switch

Table 4: Vector Parallel MIMD Systems with shared memory – 1995=>2002.

Company	System	Number of CPUs	Memory Size	Memory Bandwidth	Peak Performance	Connection
Fujitsu	VP5000	4-128	16GB-2TB	38.4GB/s	1.22Tflop/s	Crossbar
Hitachi	SR8000	4-128	16GB-1TB	1GB/s	1Tflop/s	Crossbar
Hitachi	SR8000F	4-512	16GB-8TB	1GB/s	8Tflop/s	Crossbar

Table 5: D-M MIMD vector multiprocessor Systems – 1995=>2001.

As early as the late 1980s, the proponents of PC based off-the-shelf massively parallel processors (MPPs), and clusters, raised doubts about vector processors survival. Wither vector and Fortran they shouted. Some later, in the guise of the US ASCI programme, made decisions that almost destroyed Cray Research. There were other factors for Cray's difficulties; namely, its dependence on users from the weapons industry that declined after the collapse of the Soviet Union and the end of the Cold War.

Cheap off-the-shelf processors have their attraction. Funding bodies like them because they often give the illusion of cheap capability computing. In reality, up to date, insufficient memory bandwidth and slow inter-connect switches intercede and deliver a mirage. For example, when Intel released the iPSC/2 in the late eighties, offering a 60Mflop/s processor for \$64,000, there was a lot of excitement and a prediction that the Cray workhorse of scientific computing, at that time, has become obsolete overnight. In reality the iPSC/2 could only deliver around 2Mflop/s sustained, about the same as the CDC 6600 twenty-five years earlier and was no match for the Cray Y-MP for large-scale computation. The added difficulty of parallel programming in the late eighties, hammered the last nail in iPSC/2's commercial coffin.

By the mid-nineties a new breed of computers made from off-the-shelf commodity chips arrived on the market. Those funded from the US ASCI programme, consisted of up to several thousand CPUs and grabbed the headlines, but because of communication and memory bandwidth limitations, they often deliver very little of their potential peak performance when solving real problems with irregular data. The reason for this poor return to the user by PC based systems is the relatively small cache supplemented by slow main memory and slow communication

switches. Most civilian installations used smaller systems of 16-256 CPUs or vector parallel processors to solve their problems.

The main problem confronted by all parallel systems is that of connecting CPUs to each other and to the memory. Unfortunately, full interconnection is very costly, growing with order square ( $Op^2$ ) with an increase of  $O(p)$  processors. Various crossbar and network interconnections have been used but even the best ones require  $P\log_2 P$  increase to keep the balance right. This is not available at present in the transistor based MPPs but vendors such as IBM are working hard and promising to partially remedy this deficiency with the Federation switch planned to come on stream in 2003-4.

Company	System	Number of CPUs	Memory Size	Memory Bandwidth	Peak Performance	Connection
Cray	T3E-900	6-2048	2GB-4TB	0.3GB/s	1.843Tflop/s	Torus
Cray	T3E-1200	6-2048	2GB-4TB	0.3GB/s	2.458Tflop/s	Torus
HP	Exem-V2600	16-128	16-128GB	0.96GB/s	0.291Gflop/s	Ring
IBM	SP Power3*	8-2048	8GB-1TB	0.16GB/s	3.07Tflop/s	Crossbar
IBM	P4-690-HPC	16-1024	16GB-16TB	to-64GB/s	5.2Tflop/s	Crossbar

Table 6: D-M MIMD multiprocessor Systems – 1995=>2002.

\* Special systems were produced using IBM - RS/6000 SP POWER3 - processors, such as the ASCI White, installed at Lawrence Livermore National Laboratory with 8,192 processors and having a theoretical peak performance just over 12Tflop/s, in 2001. Unfortunately, the connection switch is relatively slow and the system is unbalanced when dealing with irregular data and in that case the user gets a raw deal.

As one of the US Earth Systems scientist told me in May 2000: “In the USA it was as if we were entering, a Grand Prix, and some of us said: don’t give us all that money for the best car, just pay us less and we will buy a commodity off-the-shelf General Motors car and soup it up. Of course we lost”.

The infamous protectionist posture taken by the US government against Japanese vendors with viable vector parallel systems and the near collapse of Cray Research Inc., sent shivers to aircraft companies with a substantial stake in weapons production, who had previously been using Cray vector supercomputers. Their concerns sparked a debate in Washington. This debate is now over. It was resolved that there is a need for high bandwidth systems and money has been made available for Cray Research to develop the SV2 and successor products. The US policy of favouring scalar MPPs in the 1990s gave a great fillip to vendors with PC based systems inside the protected US market, but by year 2000 NEC dominated the aerospace and weather/climate sectors outside the USA.

At present, if you travel in an aircraft or a car, use a fridge or a PC, take medicine, the chances are that you are using a product designed using a supercomputer, but one almost certainly not made in Europe. Today’s supercomputers deliver up to hundreds of gigaflop/s and are approaching the speed for accurate simulations of many tasks. The challenge is to achieve teraflop/s or even Petaflop/s – systems with a thousand-fold better performance by the year 2010.

For example, Teraflop/s are essential for realistic simulation of global climate changes. Today’s weather models are too crude, too large-grained. To extend models to include the earth’s

hydrology (underground water reservoirs as well as cloud) computers require teraflop/s power. As far back as 1990, supercomputer model simulations were exemplified by studies on smog in cities such as Hamburg, London and the Los Angeles basin. The last was used to develop a cost-effective technical basis for key revisions of the US Clean Air Act. The annual cost of environmental control projects exceeds \$100 billion, and a modest cost reduction pays for many supercomputers.

The key to teraflop/s sustained performance is fast CPUs, fast memories and fast interconnect switches. A review of supercomputer architectures from the Cray vector machines to the proposed specialised Blue Gene Petaflop/s machine from IBM shows how the computation and memory bandwidth of Cray and the parallel vector NEC SX series supercomputers are designed to maximise sustained performance made available to the user. This is usually in the order of 40-60%, compared to about 10% for PC based systems at present.

As we entered the twenty first century, two types of supercomputers are definitely parallel and capable to be massively parallel, as the IBM ASCI White and the Japanese Earth Simulator manufactured by NEC have demonstrated. The issue is how to produce a computer that performs calculations more cheaply than its rivals.

### **Modern supercomputers – acquire a new taxonomy.**

At a workshop on climate held in France, last November, Burton Smith from Cray Research Inc., gave an excellent talk titled: “The case for architectural diversity”. He listed the current system providers, Cray, NEC, Hitachi, cluster suppliers (IBM, Compaq, SGI, SUN Microsystems, and so on), do-it-yourself cluster builders and do-it-yourself GRID builders. He then went on to define the two basic types of supercomputers; the clusters and grid systems as (Type T), whose prices are based on transistor costs, and the tightly coupled systems (Type C), represented by NEC and Cray whose prices are based on connection and custom processor costs. The performance of Type C systems is characterised by sparse matrix vector multiply high memory bandwidth and fast interconnection switches.

According to Burton, each system type is adapted to its ecological niche; Type T systems perform well with local data, well-balanced workload, explicit methods and domain decomposition. Type C systems perform well with global access of data, poorly balanced workloads, sparse linear algebra, implicit methods, and adaptive or irregular meshes.

He concluded by saying: There are two principal types of supercomputers, adapted to different niches. These niches are important. Picking the wrong type of supercomputer wastes money because you are paying for unused transistors or connections. Reducing Overhead saves time in Type T systems. Reducing Latency is the preserve of Type C systems. The cost in Type C systems is in wires (connections) not in processors.

### **The NEC SX-6 and IBM Power4.**

Let us briefly look at two representative commercially available systems from the two competing architecture types namely the multi-node parallel vector processor, the NEC SX-6 and the super-scalar HPC version IBM Power 4, P690. Since both systems are scalable as demonstrated by the

IBM ASCI White and the Earth Simulator manufactured by NEC, the comparison below shows 128 nodes. Note that both these systems use similar state-of-the-art technology proprietary chips and not off-the-shelf ones. The difference is that the IBM Power4-P690 is a transistor type T super-scalar whilst the NEC SX-6 is a type C parallel vector system.

The NEC SX-6 is a massively parallel vector system that comes in multiple nodes. An SX-6 node consists of up to 8 vector processors. The SX-6 CPU has four vector pipes (one each for add, multiply, logical and divide) which all are eight-way parallel. For the calculation of peak performance, NEC considers only the Add and Multiply pipes. Since both are eight-way parallel, they are capable of 16 results per clock. With a clock of 500 MHz, this results in an 8Gflop/s per CPU peak performance, that is 64Gflop/s per node. It is usually said: “that vendors guarantee that peak-performance will not be exceeded”. This is not true for the SX-6: If a programmer is able to use all three of the arithmetic pipe-sets (add, multiply and divide), the code will achieve more than 8Gflop/s. NEC staff call this a “supersonic” loop. The SX-6 uses 0.15micron lithography technology delivers one processor on a chip, has ultra fast register file and cache memory. The processors are coupled to a uniform shared main memory of 64Gbytes capacity and a memory bandwidth of 32Gbytes/s from each CPU, i.e. 256Gbytes/s per node. Multiple Node Models scale from 8 to 1024 CPUs (1 to 128 Nodes), delivering up to 8Teraflop/s peak performance, with a maximum share-distributed memory capacity of 8TBytes and maximum memory bandwidth of 32TBytes/s. Its inter-node crossbar switch has a peak data transfer rate of 1TBytes/s. The I/O also has a peak transfer rate of 1TBytes/s. The ultra fast memory bandwidth and inter-node crossbar switch ensures that a substantial portion of the peak performance is available to the user application.

The IBM Power4-P690 HPC option is massively parallel super-scalar system that comes in multiple nodes. A P690 HPC node consists of up to 16 IBM P690 Power4 processors. A Power4 node produces 4 floating-point results per clock, i.e. a theoretical peak performance of 5.2Gflop/s. The P690 uses 0.18micron lithography technology delivers one processor on a chip, has dedicated fast level-2 memory and level-3 cache memory. The processors are coupled to the level-3 cache of up to 16MBytes and onwards to a distributed DDR main memory of 16GBytes capacity and a memory bandwidth (between L3-cache and memory) of 12GBytes/s. For 1024 CPUs it delivers a peak of 5.2TFlop/s and a total memory of up to 16TBytes. The Achilles heal of the IBM system is likely to be the slow 12GByte/s memory bandwidth and the 1GBytes/s per port (500MBytes/s per direction) inter-node Colony switch. The next generation switch is expected to have a bandwidth of 4GBytes/s.

To put it in perspective, the memory bandwidth of a single NEC SX-6 processor (64Gbytes/s) is greater than a whole IBM Power4 Node or that of a Compaq Alpha based Node for that matter. While the Cray and NEC SX systems are well balanced, the ratio of floating point to memory bandwidth in the P4 is 0.025 and even at the L-2 to L-3 cache it is only 0.33. Added the fact that the P4 has a distributed memory of 16Gbytes then one can see the bottlenecks magnifying when dealing with large applications. Amdahl's Law implies that small inefficiencies have very large effects on degrading performance. With regular data the pre-fetch cache mechanism in the P4 compensates somewhat, but the lack of scatter/gather and fast bandwidth in the IBM architecture is likely to have severe degradation on sustained performance when handling irregular data in large-scale problems.

This is presumably why in the recent ECMWF procurement IBM contracted to deliver a hardware system 35 times more powerful than the 200Gflop/s sustained performance requirement to ensure conformance to the RFI. This is a lot of unused transistors and this simple fact speaks volumes about the unsuitability of Type T systems to meteorological capability problems.

In conclusion, vector massively parallel supercomputers such as the NEC SX series have a balanced architecture and tend to give better value for money when measured across the whole spectrum of applications. Transistor based MPPs with relatively small memory and slow inter-connect switches perform poorly on large-scale applications with irregular data such as sparse matrix calculations. For example, using Jacobi solvers the SX-6 is about 12 times faster than the IBM 690 Power 4 and this increases to about 18 times faster in measured sustainable memory bandwidth in HPC (TRIAD).

Although the supercomputer market is relatively small, about \$5Billion a year, it is overlaid with strategic importance spanning technological and scientific advances as well as national security imperatives.

It was alleged that for the ECMWF procurement, IBM heavily discounted the Power 4 by several times its book price to get the order. Of course, like its cousin peak performance, book price is theoretical and not reflecting reality. Nevertheless, if this is even half true it provides an insight about “dumping” when trading in world markets and also shows how fierce is the competition in this market.

Copyright: Christopher Lazou, HiPerCom Consultants, Ltd., UK.

Email: [Chris@lazou.demon.co.uk](mailto:Chris@lazou.demon.co.uk)

June 2002.