Current Trend of Supercomputer Architecture

Haibei Zhang
Department of Computer Science and Engineering
haibei.zhang@huskymail.uconn.edu

Abstract

As computer technology evolves at an amazingly fast pace, the rate of appearing and disappearing supercomputer systems in the world's top-500 list is high. In each of the biannually released top-500 list, new technologies and systems show up in the list with significant performance improvement and hence replace old systems. This literature survey report explores and compares these top-performance supercomputers from several aspects and analyzes the trend of current supercomputer architecture design.

1. Introduction

In the past several decades, the field of high-performance computing has experienced extensive changes of technologies, vendors, architectures. In spite of these changes the performance of world's fastest supercomputers has been steadily and continuously going up.

At the time of writing, the latest edition (the 22nd edition) of top 500 list of world's fastest supercomputers was released on November 16, 2003 by the University of Mannheim, NERSC/Lawrence Berkeley National Laboratory and the University of Tennessee. The list is available at [1].

In this literature survey report, we explore and compare those top-listed supercomputers from following points of view: interconnection network, cache coherence protocol used, memory consistence model used, latency hiding/reducing technologies, ISA of each processing unit, pipelining facilities of each processing unit and memory hierarchy design.

2. Interconnection Network

Top performance computers exhibit a variety of species. However, All current supercomputers are parallel machines, with a dozen to ten thousand processors. An important distinction is the source and price of the major components. The main issue is that whether to choose a large number of inexpensive components or a small number of expensive, tailor-made components.

Literature shows that the trend is a large number of inexpensive components, or more specifically, cluster of workstations. Number of cluster systems in the top 10 list has grown impressively to seven systems. These systems are built with inexpensive workstations boards. The number of clusters in the full top 500 has grown to 208 systems. In the last edition of top 500 list, this number is 149. The rapidly increasing number shows that cluster is the most popular architecture and is the future trend.

In the 2003 top 10 list, following systems use cluster architecture:

Rank	Cluster / Computer
2	Los Alamos National Laboratory - ASCI Q HP Alpha Server SC45 1.25GHz
3	Virginia Tech – X Apple G5 Dual 2.0GHz
4	NCSA – Tungsten Dell Poweredge 1750 P4 Xeon 3.06GHz
5	Pacific Northwest Laboratory – Mpp2 HP Integrity RX2600 Itanium 2 1.5GHz
6	Los Alamos National Laboratory - Lightening AMD Opteron 2GHz
7	Lawrence Livermore National Laboratory – MCR Linux cluster P4 Xeon 2.4GHz
10	Lawrence Livermore National Laboratory – xSeries cluster P4 Xeon 2.4GHz

Among the above 7 clusters, 6 use popular workstation processors, e.g. Intel Xeon, Intel Itanium , AMD Opteron and Apple G5.

Fast interconnection network has become one key factor that determines the performance of integrated parallel systems and clusters. In the early days of clusters, primitive parallel systems or clusters suffered from high latency caused by lack of bandwidth of the networks, e.g. mostly Ethernet. In the current top 500 list, we found most clusters use following proprietary networks:

1. Infiniband

Among the top 10 systems the Virginia Tech X system uses Infiniband as the primary communication. With switching network architecture, each node connects to the network at 20Gbps full duplex bandwidth. Total of 24 96-port switches are organized in a fat tree topology. According to [2], Infiniband can be used to connect various system components within a system. Using Host Channel Adaptors , the Infiniband fabric can be used for inter-processor networks, attaching I/O subsystems, or to multi-protocol switches like Gigabit Ethernet switches, etc. Infiniband is expected to become relatively inexpensive. The characteristics of Infiniband are rather nice: there are product definitions both for copper and glass fiber connections, switch and router properties are defined and for high bandwidth multiple connections can be employed. Also the way messages are broken up in packets for later reassembling, as well as routing, prioritizing and error handling are all described in the Infiniband standard. This makes Infiniband independent of a particular technology and a good basis to implement a communication library (e.g. Message Passing Interface, MPI) on top of it.

2. Myrinet

Among the 7 clusters in the top 10, two of them use Myrinet as communication. According to [3], Myrinet from Myricom has a bunch of features making itself a nice choice of fast cluster. Besides its higher bandwidth, Myrinet’s the main advantage is that it entirely operates in user space, thus avoiding operating system interference and the delays that come with it. This meant that the latency for small messages was around 10-15µs. Latency and bandwidth compared nicely with the proprietary networks of integrated parallel systems of Convex, IBM, and SGI. Although such a network came at a non-negligible cost, in many cases it proved a valuable alternative to either an Ethernet connected system or an even costlier integrated parallel system. Since then hardware upgrades and software improvements have made Myrinet the network of choice. At Myricom’s web site [4], Myricom provides benchmark measurements of its Myrinet2000 switches , which reaches around 2 Gbps for a Ping-pong experiment with latencies around 10 µs for small messages with the MPI version that Myricom distributes with its networks.

3. Quadrics

Three of the top 10 clusters are implemented with Quadrics network. Quadrics network was originally employed by Compaq AlphaServer SC, yet it is also available for other cluster systems. Like Infiniband and Myrinet, the network consists of two components: the ELAN interface cards, comparable to Infiniband Host Bus Adaptors or Myrinet's Lanai interface cards, and the Elite switch, comparable to an Infiniband switch/router or a Myrinet switch. The topology that is used is a (quaternary) fat tree like in most Infiniband switches. A different design that makes Quadrics switch faster is that it provids two virtual bi-directional channels per link. This means for every link two input/output ports are available with a combined theoretical bandwidth of 400 MB/s in each direction. Quadrics has proved to be a very fast and reliable network and since about two years.

4. Gigabit Ethernet

Although gigabit Ethernet features its wide availability and relatively low cost, its theoretical 125MB/s bandwidth can only satisfy a few applications that are not latency-bound. Therefore it is not an ideal interconnection architecture for current clusters to perform primary computations. Among the 7 clusters in top 10, Virginia Tech’s X uses gigabit Ethernet as its secondary communication to carry Network File Systems(NFS), control, job startup and some IP traffic.

Massively Parallel Processor (MPP) systems used to take a significant part in the top 500 list, but now it is losing its market to clusters. Following MPP systems are found in the current top 10 list:

Rank	Cluster / Computer
1	NEC Earth Simulator
8	Lawrence Livermore National Laboratory – ASCI White
9	NERSC/LBNL - Seaborg

There are totally 165 MPP systems in the November 2003 list, while the number six months ago is 213. 48 MPP systems have vanished from the list in just six months. Just 2 years ago, MPP systems used to be the dominating class of systems in the top 500 with 2/3 of all systems belonging to this class according to [1].

Other historic interconnection architectures such as single processor, SIMD, SMP, have completely vanished from the latest top 500 list. Those systems dominated the top 500 list in the 1990s.

3. Cache Coherence Protocol Used

The term "Cache Coherence" refers to the fact that for all CPUs any variable that is to be used must have a consistent value. Therefore, is must be assured that the caches that provide these variables are also consistent in this respect. There are various ways to ensure that the caches of the CPUs are coherent. One is the snoopy bus protocol in which the caches listen in on transport of variables to any of the CPUs and update their own copies of these variables if they have them. Another way is the directory memory, a special part of memory which enables to keep track of all the copies of variables and of their validness.

A trend can be observed to build systems that have a rather small number of processors that are tightly integrated in a cluster or a SMP node. The processors in such a node are virtually always connected by a 1-stage crossbar while these clusters are connected by a loosely coupled network. For example, among top 10 clusters, IBM SP cluster and linux Beowulfs employs message passing and distributed memory hence avoids cache coherence issue.

The top performance, NEC’s earth simulator, is a shared-memory system. It should be noted that the vector machines are all shared-memory (except for one distributed memory machine), and so manage to maintain their good balance, scalability, and performance despite the negative factors that reduce the performance of the hierarchical-memory shared-memory machines. The vector machines with the best performance characteristics do not employ hierarchical memory, thus greatly simplifying the coherence issue and associated latency penalty. There are two ways to couple the SX-6 frames in a multi-frame configuration: NEC provides a full crossbar, the so-called IXS crossbar to connect the various frames together at a speed of 8 GB/s for point-to-point unidirectional out-of-frame communication (1024 GB/s bi-sectional bandwidth for a maximum configuration). Also a HiPPI interface is available for inter-frame communication at lower cost and speed. When choosing for the IXS crossbar solution, the total multi-frame system is globally addressable, turning the system into a non-uniform memory access (NUMA) system.

4. Memory Consistence Model Used

There are mainly two types of memory consistency models: sequential and relax. Sequential memory consistency model requires that each memory operation (either load or a store) is performed before the next one can be issued. If the system uses this model, it also adheres to sequential consistency, although sequential consistency enables some optimizations not possible under strong ordering. Relaxed memory consistency models, such as weak ordering, introduce the notion of synchronization variables; in between the synchronization variables, ordering of memory operations is not imposed, while the synchronization variables themselves have to follow the rules of sequential consistency (ordering is imposed both at acquire and release points). Relaxed does not require sequential consistency. It has some mechanism for enforcing a consistent view of memory at specific times and places in the computation.

Relax memory consistency model is the choice of supercomputers because main memory operations introduce severe latencies. Enforcing strict memory order by waiting on memory operations will significantly waste the CPU resource which clock rate is very high.

5. Pipelining, Latency Hiding/Reducing Technologies and Instruction Set Architecture

Following latency hiding or reducing technologies are used by current processors, most of which are commodity processors used by clusters.

The number of systems in the top 500 list using Intel processors grew in the last six months from 119 to 189, signifying a major shift in this marketplace. With this increase, the Intel processor family is now the most dominant processor used in HPC systems. Intel Itanium 2 uses instruction words of 128 bits that contain three 41-bit instructions and a 5-bit template that aids in steering and decoding the instructions. This is an idea that is inherited from the Very Long Instruction Word (VLIW) machines that have been on the market for some time about ten years ago. The two load/store units fetch two instruction words per cycle so six instructions per cycle are dispatched. The Itanium has also in common with these systems that the scheduling of instructions, unlike in RISC processors, is not done dynamically at run time but rather by the compiler.

Besides Intel’s Itanium, there are remarkable number of clusters in the top 500 list that are built from Intel’s Pentium 4 Xeon. Among all these Pentium 4 Xeon clusters, all of them use dual processor nodes, in which the Pentium 4 Xeon is the processor of choice. The

Intel Pentium 4 has a high clock rate. Its instruction pipeline has as least 20 stages, which doubles the number of stages in that of Pentium III. The allocator dispatches the decoded instructions, “micro operations", to the appropriate queue, one for memory operations, another for integer and floating-point operations. Two integer ALU’s are kept simple in order to be able to run them at twice per clock. In addition there is an ALU for complex integer operations that cannot be executed within one cycle. Both the Pentium 4 Xeon and the Itanium boast of being able to run to instruction concurrently under some circumstances, i.e. the so-called hyper threading technology. Experiments have shown that up to 30% performance improvement can be achieved with a variety of codes.

114 systems among the top 500 are HP’s Superdome series MPP’s using HP’s PA-8700+ as the building blocks Like all advanced RISC processors the PA-8700+ has out-of-order execution, the sequence of instructions being determined by the instruction reorder buffer (IRB) which contains an ALU buffer that drives the computational functional units and a memory buffer that controls the load/store units.

30 clusters in the top 500 list are based on SGI Origin 3000 systems using MIPS processor. As the heart of the SGI Origin 3000 series system, the MIPS R16000 is a typical model of the modern RISC processors that are capable of out-of-order and speculative instruction execution. Like in the HP Alpha processor, there are two independent floating-point units for addition and multiplication and, additionally, two units that perform floating division and square root operations. In MIPS R16000, there are 5 pipelined functional units to be fed: an address calculation unit which is responsible for address calculations and loading/storing of data and instructions, two ALU units for general integer computation and the floating-point add and multiply pipes.

30 clusters in the top 500 list are based on Compaq AlphaServer SC45 series systems. The Alpha EV7x processor used by SC45 clusters features two integer clusters and the two floating-point units. The processor can issue up to 6 instructions simultaneously. The two load/store units draw on a 64 KB instruction and a 64 KB data cache that are both 2-way set-associative. Four instructions can be accepted for processing. Of the 80 integer and 72 floating-point registers, 41 in both register files can hold speculative results. The out-of-order issuing of instructions is supported via an integer queue of length 20 and a floating-point queue with 15 entries. However, as the integer processing clusters do not contain the same functional units, the issuing of integer instructions cannot all be scheduled dynamically. Those instructions that need to execute in a particular unit are scheduled statically.

13 clusters in the top 500 list use AMD processors as their building blocks. Among those there are 5 clusters that use AMD Opteron. The Opteron processor has many features that are also present in modern RISC processors: it supports out-of-order execution, has multiple floating-point units, and can issue up to 9 instructions simultaneously. In fact, the processor core is very similar to that of the Athlon processor.

2 clusters among the top 500 are based on IBM PowerPC 4+ processors, including the Apple G5 processor used by the top 3 Virginia Tech X cluster. The PowerPC 4 core has a separate branch and conditional register unit, 8 execution units in all. The instruction cache is two times larger than the data cache (64 KB direct-mapped vs. 32 KB two-way set associative, respectively) and all execution units have instruction queues associated with them that enables the out-of-order processing of up to 200 instructions in various stages.

6. Memory Hierachy

Intel Itanium 2

Intel Itamium 2’s L1 data and instruction caches are 4-way set associative and rather small: 16 KB each. The L2 cache has been enlarged from the previous Itanium’s 96 KB to 256 KB and it is 8-way set-associative. Moreover, the L3 cache is moved onto the chip and is no less than up to 6 MB. This cache structure greatly improves the bandwidth to the Intel processor core.

Pentium 4 Xeon:

The Pentium 4 Xeon’s L1 cache is quite small by today's standards: 8 KB. This is again to accommodate the high clock speed. With this size of cache it is possible to have a latency of two cycles for the cache. The L2 cache has a size of 256 KB for the Pentium 4 and up to 1 MB for the Xeon processor.

HP PA RISC 8700+

The PA RISC 8700+ features a large L1 cache: 0.75 MB instruction cache and 1.5 MB data cache. Both are 4-way set associative.

SGI MIPS R16000:

Both the integer and the floating-point registers have a physical size of 64 entries, however, 32 of them are accessible by software while the other half is under direct CPU control. The L1 instruction and data caches have a moderate size of 32 KB and are 2-way set-associative. In contrast, the L2 cache can be very large: up to 16 MB.

Compaq Alpha EV 7x

The Alpha EV 7x 's L1 cache includes a 64 KB instruction and a 64 KB data cache that are both 2-way set-associative. The L2 cache size is 1.75MB.

AMD Opteron

The Opteron processor has a 64K data cache, 64K instruction cache, and 1MB of L2 cache.

IBM PowerPC 4+

The PowerPC 4+ features 1.5 MB L2 cache divided over three modules of 0.5 MB each. The L2 cache module are connected to the processors by the Core Interface Unit (CIU) switch, a 2x3 crossbar with a bandwidth of 40 B/cycle per port. The L3 cache is up to 32 MB.

9. Conclusion

The November 2003 top 500 supercomputer list has shown a significant shifting in the supercomputer market. Several architectures has become historic and disappeared from the list. Massively Parallel Processor keeps losing its rank in the fastest computer list. Cluster's rank and share in the list has been steadily rising at a tremendous speed. High performance workstation processing units are used by most clusters. The new Intel Itanium 2 and AMD Opteron, with high clock frequency and several new technologies, have become good alternative choices against most RISC processors available in the market at the moment of writing. With its steep-high rate of growing in the top 500 list, it is predictable that Compaq Alpha, HP PA RISC, SGI MIPS will be faded out from the top performance list in the next few years.

10. References

[1] Top500.org, "Top500 List 11/2003", http://www.top500.org/list/2003/11/

[2] T. Shanley, "Infiniband Network Architecture", Addison-Wesley, Nov. 2002.

[3] N.J. Boden, D. Cohen, R.E. Felderman, A.E. Kulawik, C.L. Seitz, J.N. Seizovic,Wen-King Su, "Myrinet - A Gigabit-per-second Local Area Network", IEEE Micro, 15, No. 1, Jan. 1995, 29-36.

[4] Myricom, http://www.myrinet.com

[5] G. Bell, J. Gray, "What's next in high-performance computing?", Communications of the ACM, February 2002.