Short description of newly featuring archictectures in the Top 500

Aad van der Steen

November 14, 1995

1.1 The new list of the TOP500 systems

Since the last TOP500 [3] report the list has changed on quite some points. In this section we indicate the systems that have dropped out and we describe the architectures that have come in newly. The new list is given in Table 1.1

The following systems have dropped from the list:

- Cray Computer Corporation: Cray-2, Cray-3.
- Cray Research Inc.: Cray Y-MP, Cray M90, CS6400.
- Hitachi: S820, S3600.
- IBM: ES/9000.
- Maspar: MP2.
- Meiko: CS/i860.
- nCUBE: nCUBE2.
- Silicon Graphics: Challenge.

Twelve systems in all. Of these 7 systems are (multi-)vector processors: all Cray systems, except the CS6400, the Hitachi systems, and the IBM system. The Cray CS6400 and the Silicon Graphics Challenge are shared memory symmetric multi-processor systems. The Maspar MP2 is of the processor-array type. The Meiko CS/i860 and the nCUBE2 are both distributed memory MIMD machines.

For the description of the various types of architectures one is referred to [3] or [2]. In this section we only describe the machines that appear for the first time in Table 1.1 in alphabetical order. For the descriptions of the other systems are given in [3]. Apart from the systems parameter list that is given for each system also the values of R_{peak} and R_{max} as available from [1] are quoted together with the number of nodes associated with these values. The reason that the system parameter list is not sufficient is that the maximal theoretical peak performance is given for the maximal configuration of a system. Such a system may not actually exist.

1.1.1 The Convex Exemplar SPP-1200.

Machine type: RISC-based distributed-memory multi-processor.
Models: SPP-1200.
Operating system: SPP-UX, based on OSF/1 AD microkernel.
Connection structure: Ring.
Compilers: Fortran, C.

Table 1.1: Number of Systems Installed

Table 1.1: *Machines are ordered by frequency of occurrence (N). R_{max} is the cumulative maximum observed performance, R_{peak} is the cumulative theoretical peak performance for each system type. First, medium, and last are the first, median, and last entries of the a system type in the TOP500 list ordered by performance. The system that appear for the first time in the table are marked with a *.*

	N	R_{max}	R_{peak}	first	medium	last
SP2	114	851.3	1265.4	6	269	491
Power Chall	109	362.2	465.5	240	378	475
C90	52	480.2	538.1	60	179	500
T3D	32	666.2	998.2	4	113	388
CM5	23	313.7	685.5	10	157	295
Paragon	21	293.7	431.3	2	174	377
VPP500	19	690.2	872.8	1	89	203
SX3	18	194.5	216.8	39	175	422
Po.Ch.-Array*	16	160.9	250.0	30	144	390
VP2000	14	55.4	70.0	273	286	332
GC/PP	10	49.3	94.7	128	283	478
SPP1000	10	33.1	92.8	354	359	363
S-3800	9	157.7	176.0	27	86	156
iPSC/860	8	20.8	41.0	492	496	499
Triton90*	8	38.8	48.6	191	245	419
SPP1200*	7	25.1	55.7	298	327	485
CM200	5	34.6	70.0	119	183	226
CS	5	20.6	86.4	221	327	484
Paragon SMP*	5	203.8	246.4	3	55	102
C4/XA	3	8.7	9.7	423	424	425
KSR2	3	12.6	16.6	235	269	287
CM2	2	10.4	28.0	216	217	217
SP1	2	9.6	32.0	233	234	234
SX4*	2	76.2	80.0	9	31	52
AlphaServer	1	5.0	7.2	227	227	227
CM5E	1	7.7	20.0	135	135	135
KSR1	1	3.4	5.1	348	348	348
Total	500	4785.4	6903.8			

Mannheim/Tennessee November 10, 1995

System parameters:

Model	SPP-1200
Clock cycle	8.3 ns
Theor. peak performance:	
Per proc. (64-bit)	240 Mflop/s
Maximal (64-bit)	30.7 Gflop/s
Main memory	≤32 GB
Memory/node	≤256 MB
Communication bandwidth:	
aggregate (see remarks)	16 GB/s, 4GB/s
No. of processors	4–128

Performance:
R_{peak} 3.7 Gflop/s
R_{max} 7.7 Gflop/s

Note: The $R_{peak}$and R_{max}values as given above stem from an 32 processor system (4 hypernodes, see below).

The SPP-1200 is the successor of the SPP-1000 and structurally there are no differences with this machine. The system can be seen as a logical next step of integration after the Meta Series clusters of Convex. Where the latter is a cluster of workstations that still uses external interconnection channels like Ethernet or FDDI, the SPP-1200 has its communication structure integrated. As in the Meta Series the CPUs used are HP's PA-RISC processors. Up to 8 of those can be placed in what is called a *hypernode* by Convex. A maximal system consists of 16 nodes, i.e., 128

Within each hypernode up to 2GB of memory can be accommodated which can be reached by the local processors via a crossbar with an aggregate bandwidth of 16 GB/s. The hypernodes in turn are connected to each other by a crossbar with an aggregate bandwidth of 4 GB/s. So, the system concept is somewhat hybrid: within a hypernode the machine is effectively a shared-memory system, while between hypernodes it is a distributed memory system. Each node supports local I/O, while external global I/O can be done at an aggregate rate of 4 GB/s.

The Exemplar programming environment complements the SPP-1200 at the software side. This environment includes a message passing programming model (PVM) and a virtual shared memory model which allows the user to have shared-memory view of the system.

1.1.2 The Cray Y-MP T90 (Triton)

Machine type: Shared-memory multi-vectorprocessor.
Models: Y-MP T90.
Operating system: UNICOS (Cray Unix variant).
Compilers: Fortran, C, C++, Pascal, ADA.

System parameters:

Model	Cray Y-MP T90
Clock cycle	2.2 ns
Theor. peak performance:	
Per processor	1.818 Gflop/s
Maximal	58.2Gflop/s
Main memory	≤8 GB
Memory bandwidth:	
Single proc. bandwidth	21.8 GB/s
No. of processors	2–32

Performance:

R_{peak}	19.5 Gflop/s
R_{max}	28.8 Gflop/s

Note: The R_{peak} and R_{max} values as given above stem from a 16 processor T90 (T916).

The T90 is the successor of the the Cray Y-MP C90 and in almost all respects the machines are similar. As in the C90, the number of arithmetic vector pipe sets is four. The performance of a full T90 CPU is slightly less than four-fold that of a maximal C90 system. This is brought about by lowering the clock cycle from 4.1 to 2.2 ns and by doubling the number of CPUs from 16 to 32.

The machines from Cray Research Inc. are at this moment the only ones with a memory bandwidth as seems optimal for vector processors: two operands can be loaded and one result can be stored in one cycle for each pipe set. For the C90 this meant that the relative bandwidth to the CPUs had to be doubled from 24 to 48 bytes/cycle. This has indeed been accomplished and observed results indicate that for the C90 the performance scales up with the clock cycle and the number of functional units.

The Cray Y-MP T90, C90, and M90 systems do not have separate scalar processors but scalar- and vector code have to share the same functional units. Theoretically, the absence of separate scalar processors might impair the throughput speed, however, in practice the drawbacks seem rather limited.

1.1.3 The Intel Paragon XP MP

Machine type: RISC-based distributed-memory multi-processor.
Models: Paragon XP MP
Operating systems: OSF/1, SUNMOS.
Connection structure: 2-D mesh.
Compilers: Fortran 77, C, ADA.

System parameters:

Model	Paragon XP MP
Clock cycle	20 ns
Theor. peak performance:	
Per Proc. (64-bits)	0.075 Gflop/s
Maximal	115.0 Gflop/s
Memory:	
Memory/node	$\leq$128 MB
Maximal	$\leq$192 GB
Communication bandwidth	350 MB/s
No. of processors	64–1536

Performance:

R_{peak} 338.0 Gflop/s
R_{max} 281.1 Gflop/s

Note: R_{peak} and R_{max} are obtained from coupling various XP MP systems up to a number of 6768 processors.

The Paragon MP is a follow-on to the Paragon XP systems that are around since 1992. It follow the trend that is observed in many systems: combination of a number of processors in one node in a shared memory fashion and combining the nodes via a distributed memory model. In the case of the Paragon MP three i860 XPs are combined on one board. When all processors are involved in computation a maximum speed of 225 Mflop/s per node should be attainable. In principle, two processors per node serve as application processors while the third is a message coprocessor. When no communication is going on also this third processor is used for computation.

The memory bandwidth within a node is 400 MB/s while internode communication an aggregate bandwidth of 350 MB/s (175 MB/s in and 175 MB/s out). From release 5.0 on for both the Fortran 77 and C compiler use of the multi processor model is transparant to the user.

The Paragon MP retains compatibility with the Paragon XP/S, XP/E, and the former iPSC/860 systems, an Intel hypercube system preceding them. In particular the the transparent parallel Distributed File System can be used in applications migrated from the iPSC/860. The Paragon has its own parallel file system.

1.1.4 The NEC SX-4

Machine type: Shared-memory multi-vectorprocessor.
Models: SX-4.
Operating system: SUPER-UX (NECs Unix variant).
Compilers: Fortran 77, Fortran 90, HPF, C, C++.

System parameters:

Model	SX-4
Clock cycle	8
Theor. peak performance:	
64-bit precision	1 Tflop/s
Main memory	≤128 GB
Extended memory	≤384 GB
Communication bandwidth	128 GB/s
No. of processors	1–512

Performance:
R_{peak} 64 Gflop/s
R_{max} 60.72 Gflop/s

Note: The R_{peak} and R_{max} values as given above stem from a 32 processor SX-4.

The SX-4 series is the third generation of NECs supercomputers, the first being the SX-2. The architecture is in many respect different from its predecessors although the basic processor resembles the structure of the earlier SX-3 machine. However, the basic processor is implemented in the cheaper CMOS technology instead of ECL which makes them much smaller and cheaper and can be air-cooled. The basic processor in this system will be a 2 Gflop/s vector processor of which up to 32 can be placed in what is called (confusingly) a "node". Up to 16 nodes can be connected to a system of 512 processors maximum with a theoretical peak performance of 1 Tflop/s by an Internode Crossbar Switch (IXS) or HIPPI switch. The maximal model should attain a bisectional bandwidth of 128 GB/s for the IXS switch. The SX-4 shows a similar structure as described for the SPP1200 and the Silicon Graphics Power Challenge Array: A multiple CPU shared-memory architecture within a node and a message passing distributed programming model between nodes. For the message passing MPI will be employed.

Furthermore, Fortran 77, Fortran 90, HPF, and C will be provided with a C++ precompiler in addition. Also profiling and performance analysing tools, ANALYZER/SX and PARALLELIZER/SX will be available.

The range of models that are offered is large: 4 "compact" models with up to four processors are available. Furthermore, within a single-node model 4, 8, 16, and 32 processor are possible. Of the multi-node systems models with 2, 4, 8, and 16 nodes are offered.

1.1.5 Silicon Graphics PowerChallenge Array

Machine type: Shared-memory multi-processor.
Models: PowerChallenge XL Array.
Operating system: IRIX (SGI's Unix variant).
Compilers: Fortran 77, C, C++ , Pascal.

System parameters:

Model	Model XL Array
Clock cycle	11.1 or 13.3 ns
Theor. peak performance:	
Per proc. (64-bit)	0.30 or 0.36 Gflop/s
Maximal (64-bit)	— Gflop/s
Main memory	— GB
Memory bandwidth:	
Proc. to cache/proc.	1.2 GB/s
Main memory/cache	1.2 GB/s
Communication bandwidth	100 MB/s
No. of processors	—

Performance:

R_{max}	26.7 Gflop/s
R_{peak}(64-bit)	46.1 Gflop/s

Note: The value of $R_{max}$is obtained with 128 processors.

The PowerChallenge XL system (see [3]) is used as a computational "node" in the PowerChallenge Array system. Such an XL system can house up to 18 MIPS R8000 RISC processors with, depending on the clock cycle, a peak speed of 300 or 360 Mflop/s. Internally data is transported from the main memory to the CPUs by the so-called POWERpath-2 bus. It is 256 bits wide and has a bandwidth of 1.2 GB/s. This is very fast as busses go but even then the data rates that are needed by the CPUs cannot possibly be fulfilled when no special provisions would exist. These provisions are present in the form of large data and instruction caches for each of the CPUs. The off-chip cache cache for floating-point data is very large: 16 MB. It should reduce the bus traffic as much as possible. All floating-point operations are done by streaming the operands from this large off-chip cache to the floating-point registers.

In the PowerChallenge Array the XL systems are coupled via a HiPPI channel to form a cluster of systems using PVM to communicate between them for the solution of extremely large application problems. SGI provides a "shared-memory" PVM which makes the message passing model homogeneous for the user both within a PowerChallenge XL and between the "nodes". This trend is also to be seen with other vendors (e.g. Convex SPP1200, Fujitsu VPP300, and NEC SX-4).

Parallelisation is done either automatically by the (Fortran or C) compiler or explicitly by the user, mainly through the use of directives. As synchronisation, etc., has to be done via memory the parallelisation overhead is fairly large.

Bibliography

[1] Dongarra J.J., *Performance of various computers using standard linear equations software*, Computer Science Technical Report CS-89-85, Univ. of Tennessee, November 1995.

[2] van der Steen A.J., *Overview of recent supercomputers*, 5th edition, NCF Report, The Hague, January 1995.

[3] Dongarra J.J., H. Meuer, Erich Strohmaier (eds.), *The 1994 TOP500 Report*, Supercomputer, **11**, No. 2/3, June 1995.