Application Performance of Modern Number Crunchers

Gerhard Wellein, Thomas Zeiser, Georg Hager, Regionales Rechenzentrum Erlangen (RRZE),
Martensstraße 1, 91058 Erlangen
Peter Lammers, Hochleistungsrechenzentrum Stuttgart, Allmandring 30, 70550 Stuttgart

Is There Still a Need for Classical HPC Systems or Can We Go Commodity Off-the-Shelf?

In preparation of upcoming procurements in Germany a comprehensive performance evaluation for a selection of applications and kernels has been performed. The main focus of our study was on codes from computational fluid dynamics, which are known to be memory-intensive. In the light of our results we want to comment on the discussion about the need for tailored HPC systems such as SGI Altix or vector computers.

Introduction

Looking at the TOP500 list [1] of recent years, more and more classical HPC systems have been replaced by commodity off-the-shelf (COTS) clusters which do not mainly focus on HPC requirements, but dominate the HPC market due to their (often) moderate price-performance ratio. However, it has also been acknowledged recently that the gap between *sustained* and *peak* performance for scientific applications on COTS platforms is growing continuously [2]. Tailored HPC systems, such as SGI Altix or vector computers, have been designed to meet the numerical requirements of scientific, memory-intensive applications. High *sustained* single processor floating point performance, high memory bandwidth, a balanced interconnection network and a mature software environment (compilers and libraries) are the characteristics for the latter systems. In this report we comment on performance and scalability of tailored HPC systems versus COTS clusters using two applications from computational fluid dynamics (CFD).

Benchmark systems

We have chosen a GBit/Xeon cluster, a Myrinet/Opteron cluster, a SGI Altix 3700 system and a NEC SX6+ based system. Details of the configurations can be found in Table 1.

An important difference between the Intel x86 and AMD x86_64 design is the memory subsystem. While Intel still promotes (also for its new EMT64 architecture) bus based architectures where two or four processors share one path to main memory, AMD uses a separate memory interface for each CPU providing scalable bandwidth within a shared memory node. For this reason, the AMD design is favourable for memory bound applications.

System	GBit/Xeon	Myrinet/Opteron	SGI Altix 3700	NEC SX6+
Basic building block	2-way SMP node with 1 memory path	2-way SMP node with 2 memory paths	4-way SMP node with 2 memory paths	8-way SMP node
CPU	Intel Xeon 2.66 GHz	AMD Opteron 2.0 GHz	Intel Itanium2 1.3 GHz, 3 MB L3	565 MHz
Peak performance per CPU	5.3 GFlop/s	4.0 GFlop/s	5.2 Gflop/s	9 GFlop/s
Memory bandwith per building block	4.3 GByte/s	2x 5.4 GByte/s	2x 6.4 GByte/s	8x 36.0 GByte/s
Interconnect	Cisco 4503 GBit Ethernet switch	Myrinet2000	SGI NUMALink3 2x 1.6 GB/s bidirectional	NEC IXS crossbar 8 GB/s bidirectional
Operating system	Debian Linux 3.0	SuSE SLES 8 Linux	Redhat AS2.1 with SGI Propack 2.4	SUPER-UX
Compiler	Intel ifc 7.1	PGI 5.0	Intel efc 7.1	Native NEC SX

Table 1: Details of platforms and compilers used in the benchmarks.

Application scenarios

Two representative application codes have been chosen, which have been developed at the Institute of Fluid Mechanics (LSTM-Erlangen, Prof. Dr. Durst) and at the University of Erlangen-Nuremberg. Both programs are currently in intense use ranging from single processor runs on Intel Xeon through moderately parallel jobs on SGI Altix (RRZE) up to high end simulations on 512 processors (64 nodes) of the Hitachi SR8000 TFlop/s system at LRZ Munich. The codes have been ported and optim ized by the HPC group of RRZE. Concerning the computational requirements of CFD, usually two scenarios can show up:

Speed-Up: A problem of *fixed size* should be solved as quickly as possible. Time-to-solution is a critical point for applications e.g. from engineering.

Scale-Up: The problem size is scaled with the number of CPUs/compute nodes used. In this context, basic turbulence research can serve as an example.

Architecture	1 CPU	2 CPUs
Intel Xeon	1.9	3.0
AMD Opteron	2.8	5.7
Intel Itanium2	5.0	7.2
NEC SX6+	38	74

Table 2: Speed-up performance within 2-way nodes given in million lattice site updates per second (MLup/s). On the Itanium2 Altix system the two CPUs chosen for this measurement share one path to the memory.

Scalability of lattice Boltzmann simulations

Owing to the high scientific potential for large scale applications, we have chosen the lattice Boltzmann method (LBM) [3] as a first test case. LBM is a recent method from CFD which is characterised by algorithmic simplicity owing to the explicit nature of the algorithm and equidistant Cartesian grids.

In Figure 1 the parallel performance of both scenarios is presented. A handy performance unit for LBM is million lattice site updates per second (MLups), where 5 MLups ~ 1 GFlop/s holds for our calculations.

Of course the *scale-up* problem is well suited for cluster configurations, where parallel efficiencies of more than 80% on 64 processors can be achieved: A low ratio of communication vs. computation was chosen which remains constant in our application if the problem size is scaled linearly with processor count. Due to higher single processor performance (see Table 2) and almost perfect scalability the SGI Altix system is significantly ahead of the COTS clusters.

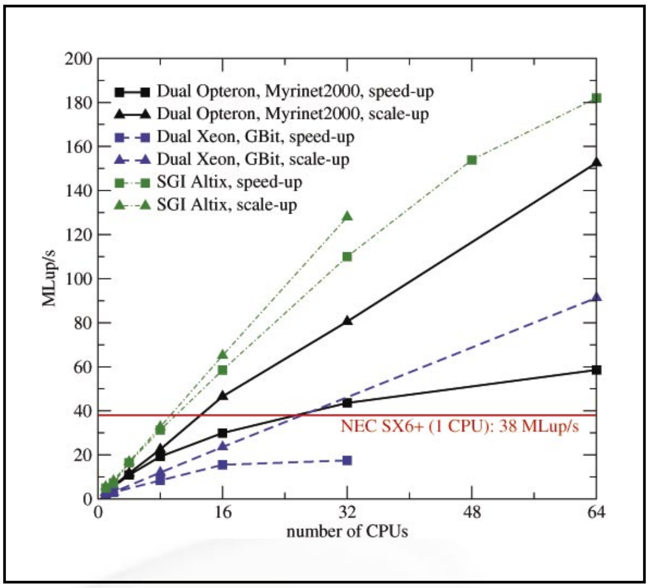

Figure 1: Scalability of LBM for modern clusters and HPC systems. The domain size is 256×129×128 for speed-up and 128^3 per processor for scale-up tests.

The *speed-up* case is a more appropriate test for the balance of parallel computer architectures. With increasing processor count, the computational domain per processor decreases while the communication per processor remains almost constant. Figure 1 clearly demonstrates that the GBit cluster does not scale at all beyond 16 processors, while the Myrinet interconnect does significantly better. Nonetheless, a Myrinet cluster with 24 Opterons is required to achieve the same performance as a single NEC SX6+ vector processor. The SGI Altix scales very well even for the speed-up case sustaining a parallel efficiency of roughly 80% on 64 processors (if using the two processor performance as the base and thus ignoring the bandwidth problem within the 2-way nodes as shown in Table 2). In comparison to the cluster interconnects, the NUMALink3 is in a class of its own, providing roughly 5 times the MPI bandwidth and only 20% of the MPI latency of Myrinet2000. Although the SGI Altix is much better balanced than the COTS clusters, it suffers from limited scalability on node level compared to AMD compute nodes (see Table 2).

Large Eddy Simulations using LESOCC[1]

LESOCC is a finite volume code with the strongly-implicit solver (SIP-solver) according to Stone [3] as the core routine.

Owing to the complex communication pattern required by the numerical scheme, there is a severe restriction on scalability. Thus we have chosen a typical workload with 12 MPI processes. In Table 3 we show the relative performance of the COTS clusters and the SGI Altix relative to 12 NEC SX6+ vector processors. The vector system is still a factor 5 ahead of SGI Altix, while the COTS clusters only achieve half of Altix performance. Interestingly, performance can also be significantly improved on the Opteron cluster if only one processor per node is used. However, the effect is not as large as on the Xeon. A detailed analysis of the communication pattern on the Altix has identified that a large fraction of runtime is spent in MPI communication with low performance. Whether this is an effect of bad load balance or inefficient communication patterns is currently under investigation. Figure 2 gives a typical example of computations done with LESOCC.

Conclusions

COTS present a highly competitive HPC architecture if the overall system balance is only of minor importance. However if network capability affects the performance of parallel applications, the SGI Altix is the system of choice. With a performance equivalent of 5-8 Itanium2 processors per NEC SX6+ vector processor, SGI Altix has further reduced the gap between cache-based and vector processors.

Concerning our experiences with the SGI Altix, we have found that performance can still vary between identical runs (depending on the load and/or the buffer cache size) and that only minor changes in compiler versions can still have significant performance impact. Thus, we expect a continuous slight performance increase on SGI Altix with improved system parameters and enhanced compiler technology.

Acknowledgements

Part of this work is financially supported by the Competence Network for Technical, Scientific High Performance Computing in Bavaria KONWIHR. We would like to thank CSAR in Manchester for providing access to their computing facilities and helpful support.

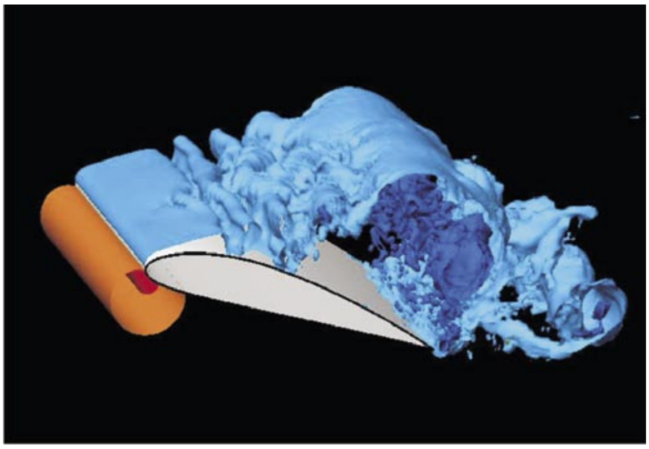

Figure 2: Isosurfaces of constant pressure of the flow around an inclined airfoil using LESOCC (Re=20000). Computations have been performed on the Hitachi SR8000-F1 at LRZ Munich. (Courtesy of N. Jovicic, LSTM - Erlangen)

References

[1] Top 500 list, available at http://www.top500.org (November 2003).
[2] L. Oliker, et al., Evaluation of cache-based superscalar and cacheless vector architectures for scientific computations, in: Proceedings of SC2003, CD-ROM, 2003.
[3] S. Chen, G. D. Doolen, Lattice Boltzmann method for fluid flows, Annu. Rev. Fluid Mech. 30 (1998) 329–364.
[4] H.L. Stone, Iterative solution of implicit approximations of multidimensional partial differential equations, SIAM J. Numerical Analysis, 5 (5), 1968

(Courtesy of N. Jovicic, LSTM-Erlangen)

(Footnotes)
[1] Large Eddy Simulations on Curve Linear Coordinates (LESOCC)

System	12 MPI processes using	Perform. relative to NEC
Xeon/GBit Cluster	6 dual-way nodes	0.06
Xeon/GBit-Cluster	12 dual-way nodes	0.10
Opteron/Myrinet Cluster	6 dual-way nodes	0.15
Opteron/Myrinet Cluster	12 dual-way nodes	0.18
SGI Altix	6 dual-way nodes	0.22
NEC SX6+	2 eight-way nodes	1.00

Table 3: Performance numbers of LESOCC relative to the NEC SX6+ measurements. If 12 dual-way nodes are used in the COTS clusters, only 1 CPU per node is used for computation while the second CPU is idle.