The New IBM Supercomputer in Jülich – Configuration and First User Experiences

Figure 2: View of the new machine room in Jülich, being built for the IBM supercomputer

Introduction

In August 2002 Forschungszentrum Jülich and IBM signed a contract about the delivery, installation, and maintenance of a new high-performance supercomputer. This system which uses IBM POWER-4 microprocessor technology will be installed in a new machine room at Forschungszentrum Jülich's Central Institute for Applied Mathematics (ZAM). It will replace the two "old" 512 processor Cray T3Es installed in 1996 and 1999 respectively, and will reach a peak performance of 8.1 TFLOPS – a factor of 13 more than ZAM's current Cray T3E-1200 capacity.

Figure 1: IBM p690 node

Configuration

The installation is being performed in three steps: In October 2002 two IBM e-servers p690 with 32 POWER-4 processors (1.3 GHz) each were delivered for tests of the operating system and early application porting. In July 2003 a larger subsystem, consisting of six IBM e-servers p690 with 32 new POWER-4+ processors (1.7 GHz) each was delivered, installed, and offered to scientific computing projects, which have been successfully peer-reviewed by the Resource Allocation Committee of the John von Neumann Institute for Computing (NIC) or the corresponding commission of the Forschungszentrum Jülich. The nodes of the subsystem are connected by Gigabit-Ethernet.
As this may produce communication bottlenecks when parallel programs are executed on more than one node, currently only programs requesting 32 or less processors are run. Together with the installation of the final configuration in December 2003 a high performance switch will be available which enables a node-spanning execution of parallel programs. With 37 IBM e-servers p690 (POWER-4+ processors) this system will reach the theoretical peak performance of 8.1 TFLOPS.
Five of the six nodes currently available are configured as compute nodes and are used for the execution of parallel programs. The sixth node (login node) consists of a login partition with several processors and a partition for data management tasks (I/O, backup etc.).
Batch jobs and interactive parallel programs are both controlled by the IBM proprietary batch system LoadLeveler. The architecture and software of the system supports a variety of parallel programming paradigms: message passing with MPI (including the functions "one-sided communication" and "parallel I/O" defined in MPI 2), one-sided communication with IBM LAPI or Cray SHMEM, multi-threading with OpenMP or Posix Threads or any combination of these paradigms (hybrid programming). Besides the necessary compilers and libraries, which support the different programming models, important software tools are also available, like the parallel TotalView debugger, the performance analysis tools Vampirtrace/Vampir for MPI programs and GuideView for OpenMP programs or several tools from IBM like the callgraph profiler gprof/Xprofiler, the hardware counter profiling tools hpmcount and hpmlib, the MPI profiling tool MP_profiler and the cache simulator Sigma. The offer is complemented by mathematical software, like the Engineering Scientific Subroutine Library (ESSL) or linear algebra libraries (LAPACK etc.) and application software like CPMD (Car-Parrinello molecular dynamics), LS-DYNA or ANSYS (finite elements; restricted use) or GaussianO3 (ab initio chemistry).

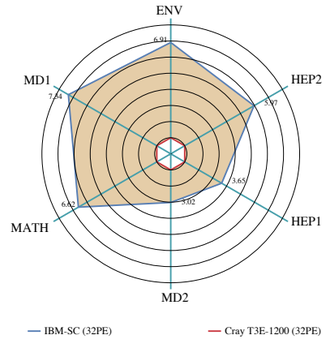
User Experiences

First user experiences show that porting applications from Cray T3E to the new IBM system seems to be no problem at all. Even the data transfer between these two supercomputers including the data conversion was in general straightforward. In most cases the compiler options hot and arch=pwr4 together with the optimization level O3 were sufficient to create well-performing codes. Detailed studies of codes with respect to memory access and the attempt to guide the access to the three cache layers by special environment variables and compiler options often did not improve the performance of the codes much more than the three options given above.
Users report that the performance gain between Cray T3E and IBM p690 is between a factor of 4 and a factor of 10. Measurements of applications from high energy physics (HEP), molecular dynamics (MD), and environmental research (ENV) are compiled in a Kiviat diagram (Figure 3). In this plot we compare runs on a Cray T3E-1200 using 32 processors with corresponding runs on a p690 node (32 processors, 1.7 GHz). On this node large page support was not enabled. Memory affinity, however, was switched on. Of course, up to now we are not able to present any results for codes using more than one node.
It should be mentioned that our experiences very well match with benchmark results of our partners at DESY Zeuthen. They have just recently published a summary of benchmark runs on two different special purpose computers, on a PC cluster, and on a Cray T3E-900, on a Hitachi, and on an IBM p690 system (hep-lat/0309149).

Outlook

We are now looking forward to the upcoming acceptance test of the complete system in November, the installation during the turn of the year in the new machine room and the start of full production at the beginning of February 2004.
For some period, this system will be the largest, made available to the scientific community by the German high-performance computing centres. According to the recommendations of the German Scientific Council, the other two National Supercomputer Centres in Stuttgart and Munich are already preparing the next steps to take the lead in the innovation spiral in high-performance computing in 2005 and 2006.

Figure 3: Kiviat diagram for six applications



• Norbert Attig

Forschungszentrum Jülich, ZAM