IBM Blue Gene/L in Jülich: A First Step to Petascale Computing

In early summer 2005, JUBL (JUlich Blue Gene/L), an IBM Blue Gene/L system, was installed at the Research Centre Jülich as the first one of its kind in Germany. At that time, a larger Blue Gene/L system was No. 1 on the Top 500 list. The Blue Gene/L has an innovative architecture and is used by well-known institutions like Lawrence Livermore National Laboratory, Argonne National Laboratory, San Diego Supercomputing Center, and EPF Lausanne for selected applications in high performance computing.

The commitment to this system in Jülich was aimed at the analysis of the architecture of a future leadership-class system which can be extended to very high processor numbers. In comparison to a general purpose supercomputer this system is very competitively priced with noticeably reduced cooling and foot print requirements while maintaining a comparable peak performance. On the other hand only very high scaling applications can exploit the full capacity of the current Blue Gene system meeting at the same time some memory and kernel restrictions. Jülich wanted to give potential users the earliest possible opportunity to determine whether their applications are suited to run efficiently on JUBL. So appropriate applications were selected from a variety of science areas and are now being tested with respect to portability and efficiency.

JUBL consists of one Blue Gene/L rack with 1024 compute nodes, each equipped with two IBM PowerPC 440 processors including additional floating-point units (700 MHz, 2.8 GFlops peak performance) and 512 MB memory. The peak performance of JUBL thus is 5.6 TFlops and the LINPACK performance was measured at 4.7 TFlops. In the Top 500 list of summer 2005, this system is ranked #60.

According to their specific requirements, the applications can use either the coprocessor mode, where one CPU in a node is used exclusively for MPI communication, or the virtual shared node mode, where both processors run independent MPI tasks. In addition to the compute nodes the system has 64 I/O nodes with an external 1 GB/s Ethernet adapter each. There are five different connection networks. For applications, the 3D-torus network and the tree-topology network are important. The latter is dedicated to collective MPI operations. The torus has a maximum latency of 6.4 microseconds and an aggregate bandwidth of 2.1 GByte/s per node. The compute nodes run a reduced Linux kernel without any time-slicing and paging capability whereas the I/O nodes run a complete Linux kernel. The system complex is completed with two larger nodes comprises also one Front-End node for user access and a Service node for system operations. Both run with a standard SuSE-Linux operating system.

The I/O nodes of JUBL have access to an external 2-TByte file system for user data and applications. It is planned that in the near future JUBL can access the HOME file systems of Jülich's IBM p690 supercomputer JUMP (see inSiDE Vol. 2 No. 1) as GPFS client to enable users to access their data from both supercomputers. An additional advantage of this solution is that users get access to the STK data robot systems. The aggregated peak performance of both supercomputers in Jülich is 14.5 TFlops.

Although the system was not available for applications before July, there were about ten serious enquiries in June and July from projects already successfully peer-reviewed and running on JUMP, aiming at capability computing. For first tests, four of them were selected, coming from the research fields lattice quantum chromodynamics (QCD), computational chemistry (Vienna Ab initio Simulation Package, VASP), materials science (Dynamical Mean Field Theory, DMFT), and laser-plasma interaction. All groups have successfully ported and run their codes within a short time, facing only minor problems. It was also possible to establish an early production environment for them, which allows a smooth operation for a handful of users. The applications used JUBL nearly to its full capacity, requesting between 128 and 512 nodes for production runs. Depending on the problem size, the runs were are either performed in coprocessor mode or in virtual node mode.

Comparing preliminary BG/L performance results from these applications with corresponding results obtained on the JUMP system, first estimates can be presented: Considering that a single Blue Gene/L processor has 41 % of the peak performance of a single Power 4+ processor of the JUMP system, the structure optimizations carried out on BG/L with the VASP code reached between 22 % and 43 % of the corresponding JUMP code depending on the chosen structures, the DMFT and the laser-plasma calculations were around 33 %. The best performance – 50 % of a Power4+ processor – was measured with the QCD code.

Looking at the scaling plots of this application the measurements become even more impressive. The major subroutine of the QCD code (Wilson operator) reached a scaling factor of nearly 4 upgrading going from 32 to 128 processors and a factor of nearly 16 upgrading going from 32 to 512 processors. The members of the QCD project at Jülich are very satisfied with these results showing a nearly 100 % scaling with these processor numbers. They expect further performance improvements from already planned code reorganizations and from announced system kernel modifications.

For more information please see
www.fz-juelich.de/zam/ibm-bgl

Figure 1: IBM Blue Gene/L

• Dr. Norbert Attig
• Klaus Wolkersdorfer

Central Institute for Applied Mathematics (ZAM),
Research Centre Jülich

Systems