John von Neumann Institute for Computing

Application Enabling in DEISA: Petascaling of Plasma Turbulence Codes

Hermann Lederer, Reinhard Tisma, Roman Hatzky,
Alberto Bottino, Frank Jenko

published in

Parallel Computing: Architectures, Algorithms and Applications ,
C. Bischof, M. Bücker, P. Gibbon, G.R. Joubert, T. Lippert, B. Mohr,
F. Peters (Eds.),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **38**, ISBN 978-3-9810843-4-4, pp. 713-720, 2007.
Reprinted in: Advances in Parallel Computing, Volume **15**,
ISSN 0927-5452, ISBN 978-1-58603-796-3 (IOS Press), 2008.

http://www.fz-juelich.de/nic-series/volume38

Application Enabling in DEISA: Petascaling of Plasma Turbulence Codes

Hermann Lederer[1], Reinhard Tisma[1], Roman Hatzky[1], Alberto Bottino[2], and Frank Jenko[2]

[1] Rechenzentrum Garching der Max-Planck-Gesellschaft
Boltzmannstr. 2, D-85748 Garching, Germany
E-mail: {lederer, tisma, hatzky}@rzg.mpg.de

[2] Max Planck Institut für Plasmaphysik
Boltzmannstr. 2, D-85748 Garching, Germany
E-mail: {jenko, bottino}@ipp.mpg.de

The ITER (International Thermonuclear Experimental Reactor) experiment must be accompanied by advanced plasma turbulence simulations. Due to the high demands for compute power and memory, simulations must be capable of using thousands or tens of thousands of processor-cores simultaneously. Highly scalable applications are mandatory.

Through a joint effort of application specialists from DEISA (Distributed European Infrastructure for Supercomputing Applications) and scientists engaged in the theory support for ITER, two important European simulation codes for core turbulence, ORB5 and GENE, have been adapted for portable usage within the heterogeneous DEISA infrastructure.

Moreover, the codes were thoroughly analyzed, bottlenecks were identified and removed, and, most importantly, the scalability of the codes could be significantly enhanced. Through application of the domain cloning concept, the PIC code ORB5 was enabled for high scalability. Efficient usage of ORB5 code could be demonstrated up to 8k processors, both on a Cray XT3 and on an IBM BlueGene/L system. GENE was parallelized through domain decomposition of the five-dimensional problem grid to such a high degree that close to loss-free efficiency on up to 32k processors of an IBM BlueGene/L machine was achieved. Results combined from both strong and weak scaling measurements indicate an even higher scalability potential for GENE. Extrapolations suggest an efficient usage on up to the order of 1M processor-cores of a similarly scalable future HPC architecture is possible, representing a milestone on the way towards realistic core turbulence simulations of future fusion devices.

1 Introduction

ITER[1] is a joint international research and development project that aims to demonstrate the scientific and technical feasibility of fusion power as a viable future energy option offering long-term, safe, and environmentally benign energy to meet the needs of a growing and developing world population. The partners in the project — the ITER Parties — are the European Union (represented by EURATOM), Japan, the People's Republic of China, India, the Republic of Korea, the Russian Federation and the USA. ITER will be constructed in Europe, at Cadarache in the South of France.

EFDA, the European Fusion Development Agreement, is an agreement between European fusion research institutions and the European Commission to strengthen their coordination and collaboration, and to participate in collective activities. EFDA, in supporting ITER, is also favouring a strong theory support that will play an essential role. Large scale simulations require most powerful supercomputers. The DEISA, an EU FP6 project, has

developed and put into operation a grid of the most powerful supercomputing platforms in Europe, but DEISA also provides advanced application support. Application enabling towards petaflops computing of two important European plasma core turbulence codes, ORB5 and GENE, is described here in detail.

2 DEISA

Major European supercomputing centres have combined their competences and resources to jointly deploy and operate the Distributed European Infrastructure for Supercomputing Applications, DEISA[2], to support leading edge capability computing for the European scientific community.

2.1 Extreme Computing and Applications Enabling in Europe

Application enabling is of key importance for adequate usability of state-of-the-art and next generation supercomputers. A team of leading experts in high performance and Grid computing, the Applications Task Force, provides application support in all areas of science and technology. In 2005 the DEISA Extreme Computing Initiative (DECI) was launched for the support of challenging computational science projects. One of the key tasks is hyperscaling, a unique service not offered by any other European Grid computing project. Hyperscaling aims at application enabling towards efficient use of thousands or tens of thousands of processors for the same problem, a prerequisite for applications to benefit from forthcoming Petaflop scale supercomputers.

2.2 DEISA and the European Plasma Physics Community

DEISA also maintains Joint Research Activities in the major fields of computational sciences. The DEISA Joint Research Activity in Plasma Physics is advised through Principal Investigators like Karl Lackner (former EFDA leader) from Max Planck Institute for Plasma Physics (IPP), Garching, and Laurent Villard from Centre de Recherches en Physique des Plasmas (CRPP), Lausanne. The objective is the enabling and optimization of important European plasma physics simulation codes in DEISA. So far the simulation codes TORB, ORB5, GENE, EUTERPE, and GEM have been enabled and optimized. The enabling of ORB5 and GENE is reported here.

3 Enabling of ORB5 simulation code

The ORB code family uses a particle-in-cell (PIC), time evolution approach, and takes advantage of all the recent techniques of noise reduction and control in PIC simulations. In particular, it uses a statistical optimization technique that increases the accuracy by orders of magnitude. Initiated at CRPP, Lausanne, ORB has been substantially upgraded at IPP, Garching. The ongoing code development is made under a close collaborative effort.

The ORB5 code[345] is able to simulate plasmas of higher complexity which can be used e.g., to simulate effects such as long-living zonal flow structures analogous to those seen in the Jovian atmosphere, or so-called Geodesic Acoustic Modes (GAM). Of course,

a higher complexity of the simulation model has its impact on the complexity of the programming model. Since ORB5 has high relevance for ITER, special effort was given to the ORB5 code to enable it to run with high scalability on existing DEISA and further relevant systems.

3.1 Single Processor Optimization and Portability in DEISA

The code was instrumented to detect bottlenecks and the most CPU time consuming routines were identified. Two bottlenecks were identified: the handling of Monte Carlo particles and the implementation of the FFT (Fast Fourier Transform).

The particle handling was improved by a cache sort algorithm that sorts the Monte Carlo particles relative to their position in the grid cells of the electrostatic potential. Hence, a very high cache reuse of the electrostatic field data could be achieved, significantly improving the Mflop rates of two important routines. The overhead introduced by the sort routine itself was minimized. In addition, a switch was implemented to optionally enlarge a work array for the sorting process: this can speed up the sorting routine by a factor of 3. Usually the resident memory size of the simulation is small enough to take advantage of this new feature.

The bottleneck related to the FFT was due to the inclusion of the source of an own FFT implementation with very poor performance. Therefore a module was written with interfaces to important optimized FFT libraries: FFTW v3.1, IBM ESSL, and Intel MKL. Performance improvements up to a factor of 8 could be measured. All three named FFT library routines have the advantage that they are no longer restricted to vector lengths of powers of two. This results in a much higher degree of flexibility when choosing the grid resolution of the electrostatic potential. With the new FFT module, all DEISA architectures besides the NEC vector system are supported.

3.2 Scaling of ORB5

The domain cloning concept was implemented in the ORB5 code to optimize scaling and decouple the selectable grid resolution from the number of processors used for the simulation. The relatively new parallelization concept is a combination of the two techniques domain decomposition and particle decomposition. This strategy[6] was applied on ORB simulations[7][8]. With the new ORB5, a simulation with 5×10^8 particles was first tested on the IBM Power4 system at RZG, achieving a speedup of 1.9 from 256 to 512 processors. Next ORB5 was tested on the Cray XT3 system (Jaguar) at ORNL in strong scaling mode up to 8k processor-cores. The Ion Temperature Gradient (ITG) driven simulation was based on a grid of $256 \times 256 \times 256$ cubic B-splines and 512M particles. On 8k processors a parallel efficiency of 70% (relative to 1k processors) was demonstrated for a fixed problem size requiring about 400 GB of main memory. ORB5 was then ported to and tested on IBM BlueGene/L. An Electron Temperature Gradient (ETG) driven simulation with a grid of $256 \times 256 \times 256$ quadratic B-splines and 800M particles was done up to 8k processors on the BlueGene/L system at IBM Watson Research Center. In strong scaling mode, a high parallel efficiency of 88% was achieved on 8k processors (relative to 1k processors, in so-called co-processor mode when the co-processor is used as an off-load engine to help with communication but not with calculations). Tests on higher processor numbers were

not possible due to the principle memory limitations on the BlueGene/L system (0.5 GB per node). This situation will improve with BlueGene/P systems expected soon at DEISA sites. In addition, weak scaling measurements were done for the same ETG simulation (using about 0.8M particles per processor), fully exploiting the available memory by always executing with the largest problem size possible per processor number. After a slight super-linear behaviour on 2k and 4k processors, due to increasing cache reuse, the parallel efficiency approached 1 again on 8k processors. This proves that a 4 TB problem case can be efficiently treated without degradation on 8k BlueGene/L nodes in co-processor mode.

4 Enabling of GENE Simulation Code

GENE[9][10][11][12], is a so-called continuum (or Vlasov) code. All differential operators in phase space are discretized via a combination of spectral and higher-order finite difference methods. For maximum efficiency, GENE uses a coordinate system which is aligned to the equilibrium magnetic field and a reduced (flux-tube) simulation domain. This reduces the computational effort by 2–3 orders of magnitude. Moreover, it can deal with arbitrary toroidal geometry (tokamaks or stellarators) and retains full ion/electron dynamics as well as magnetic field fluctuations. At present, GENE is the only plasma turbulence code in Europe with such capabilities.

GENE version 9 (GENE v9) employed a mixed parallelization model with OpenMP for intra-node communication in an SMP node, and MPI for inter-node communication across SMP nodes. Architectural characteristics of large SMP-based systems were fully exploited. However, for the number of possible MPI tasks, a hard limit of 64 was detected. On large SMP-based systems as IBM p690 with 32 processor-cores per SMP, this was not a limiting factor, since theoretically up to 2048 ($= 32 \times 64$) processor-cores could have been used.

Scalability tests with scientifically relevant problem cases revealed scalability bottlenecks already starting at 256 processors, with a relative speedup of 1.5 from 256 to 512 processors on IBM p690 with the IBM High Performance Switch interconnect. And for small-node based systems with two processor-cores per node, the technical upper limit was usage of 128 processor-cores, with 64 MPI tasks and 2 OpenMP threads per task.

4.1 Improving the scalability of GENE

As described in detail elsewhere[8], the structure and the parallelization scheme of the code was further analyzed. The GENE code has a total of 6 dimensions with the potential for parallelization. The spatial coordinates x and y, originally treated serially, contain significant potential for domain decomposition. A large number of 2-dimensional FFTs are done on the xy planes. If the xy plane is distributed, it must be transposed in order to perform the FFT in the x and y directions. The transposition, however, is communication intensive since it requires an all-to-all communication. A major change to the overall data structure with consequences for many parts of the code was done by the code authors. The new code version GENE v10 was tested on RZG's IBM p690 system. The speedup from 256 to 512 processors is significantly improved from 1.5 to 1.94 with GENE v9. It was assumed that with the described measures, the scalability of the code was pushed towards efficient usability on up to several thousands of processors.

4.2 Porting GENE to the Major Supercomputing Architectures in DEISA

The GENE v10 code originally used Bessel functions from the NAG library. Three new subroutines were written implementing these Bessel functions for GENE. Further library routines used are FFTs. Interfaces were provided or added for the FFT-routines from the optimized libraries IBM ESSL, public domain FFTW, the Math Kernel Library (MKL) from INTEL, and ACML from AMD.

4.3 Petascaling of GENE

Tests of GENE v10 on higher processor numbers were first done on the Cray XT3 system (Jaguar) at ORNL. Up to 4k (single core) processors, a linear speedup was achieved for strong scaling of a fixed grid size, requiring about 300–500 GB of main memory. For tests on higher processor numbers, the BlueGene/L system at IBM Watson Research Center was used. Here a functionally and algorithmically improved code version, GENE v11+, was used, with the same parallelization scheme as in GENE v11. Strong scaling measurements showed close to linear speedup up to 4k processors; on 8k processors, a parallel efficiency of 73% was achieved, revealing some degradation. For further scalability improvements well beyond 8k processors, the second velocity dimension was parallelized in addition. Measurements with the new code version GENE v11+ on BlueGene/L demonstrate the significant improvements achieved, with excellent scaling behaviour well into the range between of 10^4 and 10^5 processor-cores. In strong scaling mode from 1k to 16k processors, a speedup of 14.2 was achieved, corresponding to an efficiency of 89% (Fig. 1). Here, the problem size was $64 \times 32 \times 128$ grid points in the radial, binormal, and parallel direction, 64×32 grid points in $(v_\parallel, \mu)$ velocity space, and two species, corresponding, e.g., to the physical situation described in a recent publication[12]. Complementary strong scaling measurements in so-called virtual node mode, using both processors per node for computation, reveal a comparable good scalability, with only a small performance degradation of approx. 15% (Fig. 2). For weak scaling runs (using a fixed problem size per processor of about 200 MB) an excellent parallel efficiency of 99% from 2k up to 32k processors is demonstrated (Fig. 3). Here, the grid in the 2k case corresponds to the one described above, while for the other cases, the resolution in the parallel direction and in velocity space has been adapted accordingly.

In this context, we would like to note that various processor grid layouts have been measured for each number of processors, and the best performing processor grid layout with the shortest execution time for one time step was selected. The results from the two different machines can, after normalization with the known factor of 1.022, be combined into one extremely long scaling curve, with 15 measuring points covering four orders of magnitude (from 2 to 32k). The result is shown in Fig. 4, after normalization of the absolute values (time per time step) on the result for 2 processors. A measurement for only one processor was not possible, since at least the two species ion and electron had to be considered and these must be treated in parallel.

The results of the strong and of the weak scaling measurements can now be combined to extrapolate the scalability of the code beyond the actually measured maximum number of processor-cores. For strong scaling (Fig. 1) excellent scalability was proven up to 16 times the number of processors used for the base run (1k). The problem size used there ($\approx$0.5 TB), corresponding to the 2k processor measurement in Fig. 4, can now be

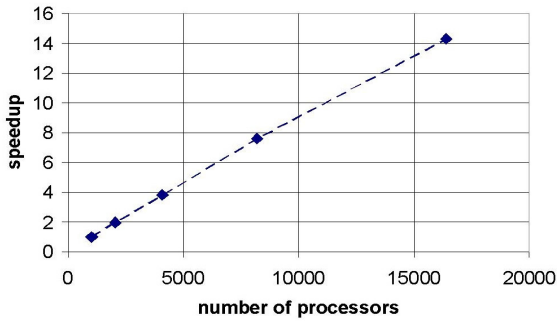

Figure 1. Strong scaling of GENE v11+ on BlueGene/L, normalized to 1k processor result (problem size of ≈300–500 GB; measurements in co-processor mode at IBM Watson Research Center)

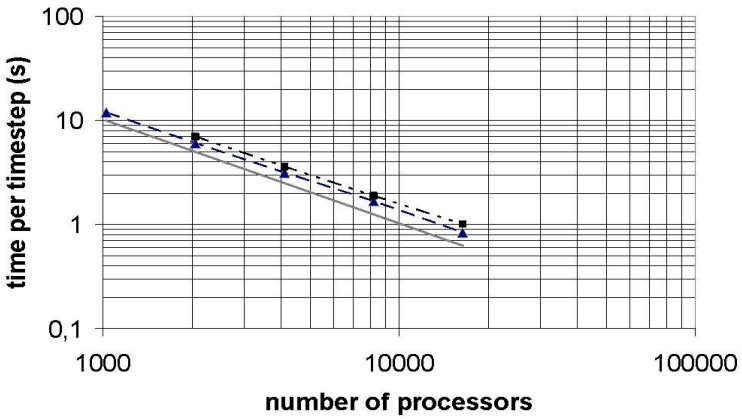

Figure 2. Strong scaling of GENE v11+ on BlueGene/L at IBM Watson Research Center (problem size of ≈300–500 GB; triangles with dashed line: co-processor mode; squares with pointed dashed line: virtual node mode (upper curve); linear scaling: straight line)

stepwise increased (doubled), and the strong scaling curve can be shifted along the weak scaling curve to the right up to the maximum processor number of the weak scaling curve. Through this extrapolation (by increasing the problem size from 0.5 TB to 8 TB), strong scaling of an 8 TB problem size on a scalable architecture is expected to result in good scalability up to $32\text{k} \times 16 = 512\text{k}$ processors.

However, for the BlueGene/L measurements presented so far, only processor grid layouts with y dimension equal to 1 were used, the parallelization of the y dimension has not yet been exploited, since it is most communication intensive. Increasing values of y from 1 to 4 in the grid layout increases the number of MPI threads from 0.5M to 2M for an 8 TB problem size. Scalability tests of y revealed excellent scaling behaviour within an IBM shared memory p690 node (up to 32 processor-cores). Therefore multi-core based archi-

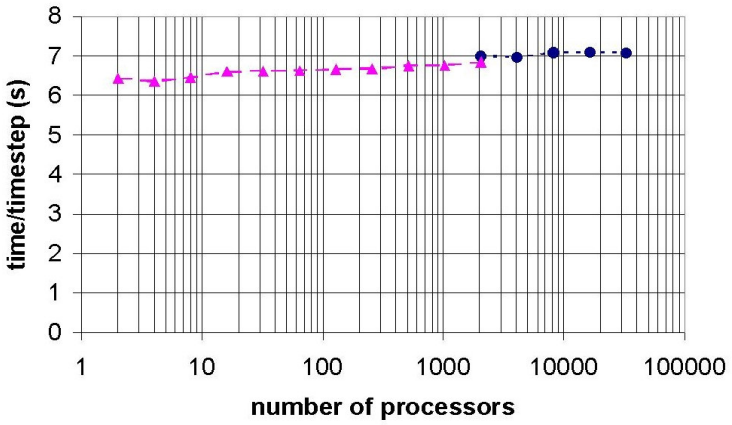

Figure 3. Weak scaling of GENE v11+ on BlueGene/L in virtual node mode (problem size per processor ≈200 MB); circles: measurements at IBM Watson Research Center from 2k to 32k processors (efficiency of 99% with 32k processors when normalizing on the 2k result); triangles: measurements at IBM Rochester Center from 2 to 2k processors; machine speeds differ by 2.2% for the 2k processor results)

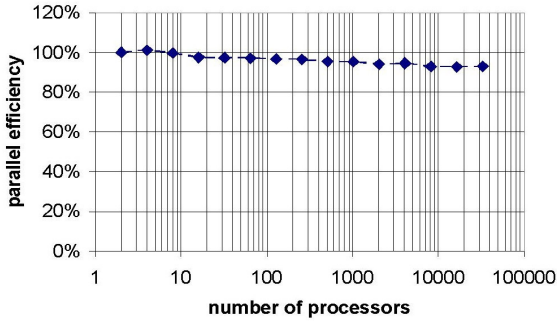

Figure 4. Weak scaling of GENE v11+ on BlueGene/L from 2 to 32k processors over four orders of magnitude (in virtual node mode, problem size per processor ≈200 MB), through combination of the two curves from Fig. 3 and normalization on the 2 processor result. Parallel efficiency on 32k processors, based on the 2 processor run, is 93%

tectures appear especially interesting for the additional exploitation of the y parallelism of GENE.

5 Conclusions

Two important European plasma core turbulence simulation codes, GENE and ORB5, were ported to the major supercomputing architectures in DEISA and beyond, and were optimized and adapted to very high scalability, as a milestone on the way towards realistic core turbulence simulations of future fusion devices. In addition, GENE can be considered a

real physics code able to diagnose the scalability of Petaflops architectures into the range of millions of processor-cores.

Acknowledgements

The authors thank the European Commission for support through contracts FP6-508830 and FP6-031513. We thank IBM for access to the BlueGene/L systems at Watson Research Center and at Rochester Center, and J. Pichlmeier for support on using the systems. We thank Cray Inc. for scalability runs on the Cray XT3 system at ORNL.

References

1. ITER http://www.iter.org
2. Distributed European Infrastructure for Supercomputing Applications (DEISA), http://www.deisa.org
3. T. M. Tran, K. Appert, M. Fivaz, G. Jost, J. Vaclavik and L. Villard, *Energy conservation for a nonlinear simulation code for ion-temperature-gradient-driven (ITG) modes for the theta-pinch*, in: Theory of fusion plasmas: Proc. Joint Varenna-Lausanne International Workshop, 1998, edited by J. W. Connor, E. Sindoni and J. Vaclavik, p. 45, Società Italiana di Fisica, Bologna, (1999).
4. A. Bottino, A. G. Peeters, R. Hatzky, S. Jolliet, B. F. McMillan, R. M. Tran and L. Villard, *Nonlinear low noise particle-in-cell simulations of ETG driven turbulence*, Physics of Plasmas, **14**, Art. No. 010701, (2007).
5. S. Jolliet, A. Bottino, P. Angelino, R. Hatzky, T. M. Tran, B. F. McMillan, O. Sauter, K. Appert, Y. Idomura and L. Villard, *A global collisionless PIC code in magnetic coordinates*, Computer Phys. Comm., **177**, 409–425, (2007).
6. C. C. Kim and S. E. Parker, *Massively parallel three-dimensional toroidal gyrokinetic flux-tube turbulence simulation*, J. Comp. Phys., **161**, 589–604, (2000).
7. R. Hatzky, *Domain cloning for a Particle-in-Cell (PIC) code on a cluster of symmetric-multiprocessor (SMP) computers*, Parallel Comp., **32**, 325–330, (2006).
8. H. Lederer, R. Hatzky, R. Tisma., A. Bottino, and F. Jenko, *Hyperscaling of plasma turbulence simulations in DEISA*, in: Proc. 5th IEEE workshop on Challenges of Large Applications in Distributed Environments (CLADE) 2007, Monterey Bay, pp. 19–26, (ACM Press, New York, 2007).
9. F. Jenko, et al., *Electron temperature gradient driven turbulence*, Physics of Plasmas, **7**, 1904, (2000).
10. F. Jenko and W. Dorland, *Prediction of significant tokamak turbulence at electron gyroradius scales*, Phys. Rev. Lett., **89**, Art. No. 225001, (2002).
11. T. Dannert and F. Jenko, *Gyrokinetic simulation of collisionless trapped electron mode turbulence*. Physics of Plasmas, **12**, Art. No. 072309, (2005).
12. P. Xanthopoulos, F. Merz, T. Görler and F. Jenko, *Nonlinear gyrokinetic simulations of ion-temperature-gradient turbulence for the optimized Wendelstein 7-X stellarator*, Phys. Rev. Lett., **99**, Art. No. 035002, (2007).