

Application Analysis and Porting in the PRACE Project

Peter Michielse
Netherlands National Computing Facilities Foundation (NCF)
The Netherlands
Email: michielse@nwo.nl

1 Introduction

PRACE, the Partnership for Advanced Computing in Europe¹, aims to set up a European HPC ecosystem to facilitate scientific research, with sustainable access to Tier-0 HPC systems, including system management and extensive application support. In order to become successful PRACE will need to understand (among others) the software requirements for future Petaflop/s systems. PRACE has identified the key scientific and technical categories of applications, through a survey of most major European HPC systems and the applications that exploit these, carried out in early 2008. Final goals in this part of the PRACE project are the construction of a benchmark suite, to be used both within the current PRACE project and beyond, when actual Tier-0 systems will be purchased. Other goals include insight in the optimisation and scalability issues with the selected applications, and applicability of synthetic benchmarks and performance analysis tools.

2 Methodology within PRACE

Each benchmark application will be worked on under the responsibility of a so-called Benchmark Code Owner (BCO). The BCO is a person who in most cases belongs to the staff of one of the PRACE partners. The BCO will steer the actual porting, petascaling and optimisation, such that the benchmark code will run on each of the designated hardware architectures for the underlying application. This includes the scheduling of work among the contributing PRACE partners to the benchmark code, and communication with the application owners on all aspects of the application: source code, dataset, output, run scripts, etc. In particular, actual results will first be communicated to the application owner, and through the public status of the deliverable report also to hardware or software vendors, and the rest of the HPC community.

As said, the BCO and his or her coworkers are not only responsible for porting the code to the actual platforms, but also for optimisation and scaling efforts. At this point in time in the PRACE project, porting has been done, and initial proposals and estimates of effort with respect to optimisation and scalability have been formulated by the BCOs.

3 Application Porting to Prototypes

PRACE conducted several surveys among both users of the top national HPC facilities in the PRACE countries, as well as among system administrators of these facilities, in order to establish a representative set of application areas and individual applications. These cover currently the most relevant usage of the national systems in Europe. As a result a list of core applications and a list of possible extensions was created. These are contained in tables 1 and 2. As many applications as possible of the core list should be worked upon in the PRACE project, both to serve in a benchmark suite and to investigate optimisation and scalability aspects.

¹ PRACE has been funded in part by the European Community under INFRA-2007-2.2.2.1 - Preparatory phase for 'Computer and Data Treatment' research infrastructures in the 2006 ESFRI Roadmap under Grant No INFSO-RI-211528. Website: www.prace-project.eu.

| Application name | Application area |
|------------------|---|
| QCD | Particle physics |
| VASP | Computational chemistry, condensed matter physics |
| NAMD | Computational chemistry, life sciences |
| CPMD | Computational chemistry, condensed matter physics |
| Code Saturne | Computational fluid dynamics |
| GADGET | Astronomy and cosmology |
| TORB | Plasma physics |
| ECHAM5 | Atmospheric modelling |
| NEMO | Ocean modelling |

Table 1: The proposed list of core applications.

| Application name | Application area |
|------------------|---|
| AVBP | Computational fluid dynamics |
| CP2K | Computational chemistry, condensed matter physics |
| GROMACS | Computational chemistry |
| HELIUM | Computational physics |
| SMMP | Life sciences |
| TRIPOLI4 | Computational engineering |
| PEPC | Plasma physics |
| RAMSES | Astronomy and cosmology |
| CACTUS | Astronomy and cosmology |
| NS3D | Computational fluid dynamics |

Table 2: Possible extensions to the core list of applications.

Another consideration has been the actual choice of promising architectures, to be assessed in the PRACE project. For the work on applications, this set of architectures (which are production or near-production systems) has been identified by PRACE in May 2008, and deployed as prototype systems to different partner sites (see table 3). Also, for each of the applications, we have selected BCOs who combine knowledge of the particular application, expertise with certain hardware platforms and access to prototype architectures. For most applications, both from the core list as well from the extended list, this has been successful. Contributors to a benchmark code typically qualify if they satisfy at least one, and preferably two or even three of these aspects.

| Architecture type | Actual system | Location |
|---------------------|-----------------------------------|------------------------------|
| MPP-BG | IBM BlueGene/P | FZJ, Germany |
| MPP-Cray | Cray XT5 | CSC, Finland |
| SMP-FatNode-pwr6 | IBM p575 Power6 | NCF/SARA, Netherlands |
| SMP-ThinNode-x86 | Bull – Intel Xeon/Nehalem cluster | FZJ, Germany and CEA, France |
| SMP-ThinNode+Vector | NEC SX-9 + x86 ... | HLRS, Germany |
| SMP-FatNode+Cell | IBM Power6 with Cell | BSC, Spain |

Table 3: Actual prototype architectures in PRACE.

Table 4 shows that all applications from the core list are usable as benchmark codes, on at least 3 target prototype architectures, complemented with 3 applications from the non-core list: CP2K, GROMACS and NS3D. These are the first 12 rows of table 4. SMMP, RAMSES and CACTUS have disappeared from the extended list, as it turned out to be that there was no PRACE partner that could volunteer as BCO. Instead, GPAW (computational chemistry), ALYA (computational mechanics and fluid dynamics), SIESTA (computational chemistry, molecular dynamics) and BSIT (computational geophysics) have joined the application set, mainly to make sure that enough coverage of the SMP-FatNode+Cell platform could be guaranteed. An additional advantage of this is that two other application areas are introduced: computational mechanics and computational geophysics. Each BCO and its contributors have started the work on the benchmark codes and hardware architectures.

Table 4 also shows the current porting status of the applications to the prototype architectures. Green colors denote successful porting, yellow means that porting is in progress, and orange means that porting has not started yet or stopped for the moment because of practical reasons (mostly lack of human resources to do the work).

| Application | MPP-BG | MPP-Cray | SMP-TN-x86 | SMP-FN-pwr6 | SMP-FN+Cell | SMP-TN+vector |
|--------------|--------------|--------------|-------------|--------------|--------------|---------------|
| QCD | Done | Done | | Done | | |
| VASP | Done | | | Done | Stopped | Yet to start |
| NAMD | Done | Done | | Done | Yet to start | |
| CPMD | Done | | | Done | Done | Yet to start |
| Code_Saturne | Done | Done | | Done | Stopped | Done |
| GADGET | Done | | Done | Done | | |
| TORB | Done | | | Done | Yet to start | |
| ECHAM5 | Stopped | Done | In progress | Done | | Yet to start |
| NEMO | Done | Done | | Done | | In progress |
| CP2K | Done | Done | | Done | | |
| GROMACS | Done | Done | | Done | | |
| NS3D | | Yet to start | Done | Yet to start | | Done |
| AVBP | Yet to start | | Done | Done | | |
| HELIUM | In progress | Done | | Done | | |
| TRIPOLI_4 | Yet to start | | Done | | | |
| PEPC | Done | Done | | Done | | |
| GPAW | Done | Done | | Done | | |
| ALYA | | | | | Done | |
| SIESTA | | | | | Done | |
| BSIT | | | | | Done | |

Table 4: Summary on porting efforts for benchmark codes and prototype architectures.

4 Scalability expectations

Apart from porting efforts to the prototype architectures, initial insight in the potential for scaling to petascale systems (and single-CPU optimization) has been obtained. Table 5² contains the scalability potential of each of the benchmark codes, including an estimate on the amount of effort in person months (PM). We have defined scalability to be in the range none via low, medium to high and have assumed one core to deliver a minimum of 10 GFlop/s peak performance. The color codes mean:

- None (red): No speed-up above 2500 cores;
- Low (orange): Speed-up obtained up to 5000 cores;
- Medium (yellow): Speed-up obtained up to 10000 cores;
- High (green): Speed-up obtained for more than 100000 cores.

² Not all cells in table 5 have been filled yet, as initial analysis after porting is currently work in progress.

Speed-up at a certain number of cores is defined as still improving execution time when comparing the execution time on that number of cores to the execution time on half the number of cores.

From table 5, the following initial observations can be made:

- Within the set of computational chemistry codes (VASP, NAMD, CPMD, CP2K, GROMACS, GPAW) the potential varies from low to high. At first sight, this may seem surprising, as they all cover broadly the same application area, although individual codes may use different approaches. It will make sense to investigate how low scaling codes may benefit from algorithms and implementations used in highly scalable codes;
- The amount of effort estimated to improve scalability to medium or high seems to be reasonable: on average around 4 to 5 person months. This will be carried forward in remaining PRACE work.

| Benchmark code | Expected scalability | Estimated effort | Comments and areas of attention |
|----------------|----------------------|--------------------|--|
| QCD | high | 0-1 person months | |
| VASP | high | | Depends on FFT and BLAS implementations |
| NAMD | medium-high | 8-10 person months | Investigate master-slave (3 pm), investigate shared memory (7 pm) |
| CPMD | high | 2 person months | Well parallelised already, some tuning needed |
| Code_Saturne | medium | 3 person months | Preprocessing stage and IO |
| GADGET | medium-high | 2 person months | Investigate potential OpenMP constructs and MPI implementation |
| TORB | high | 3-5 person months | Adapt code internals (up to now 999 processes is max.) |
| ECHAM5 | low-medium | 2-8 person months | OpenMP optimisation, data output mechanism |
| NEMO | low | 3 person months | Domain decomposition load imbalance, solver implementation, MPI |
| CP2K | low | 5 person months | Load imbalance needs to be solved |
| GROMACS | medium | 8 person months | Optimise communication patterns |
| NS3D | low-medium | 1-6 person months | Very platform dependent - MPI AlltoAll implementation |
| AVBP | medium-high | 2 person months | Focus on MPI implementation (AllReduce area) |
| HELIUM | medium | 3-4 person months | Focus on MPI implementation (synchronisation constructs) |
| TRIPOLI_4 | high | 6 person months | Independent particles, Monte-Carlo approach, IO to be modified |
| PEPC | high | 1 person month | Data structure to be investigated |
| GPAW | medium-high | 3-6 person months | Implement SCALAPACK usage, parallelise over electronic states |
| ALYA | medium-high | 2 person months | Explicit solver ok, implicit solver requires effort, IO to be modified |
| SIESTA | medium | 2-3 person months | Focus on MPI implementation |
| BSIT | high | 1 person month | Embarassingly parallel, need to consider queue management system |

Table 5: Expected scalability potential and estimated effort for benchmark codes.

5 Future Work in PRACE, Relation to IESP and Acknowledgements

As has been mentioned before, porting the applications to the target prototype architectures is work-in-progress. Already a significant part of the sparse matrix has been filled. This work will continue to complete the sparse matrix on applications and prototype architectures.

Another aspect is the fact that already ported applications will enter the stadium of petascaling and optimisation. BCOs will remain responsible for the coordination of optimisation and petascaling aspects.

With respect to the future final benchmark suite for PRACE, there is the issue of usage and licensing of the application codes. This will need to be resolved with the code developers.

With respect to IESP, it seems to make sense to exchange experience and progress on many of the applications, since these are used globally and possibly already improved by US and/or Japanese efforts. Further, alignment of the efforts in PRACE on application scalability with efforts in the USA and Japan, maybe including software developers and hardware vendors, is important.

This white paper is based on the PRACE project's deliverable "Report on available Performance Analysis and Benchmark Tools, Representative Benchmark", dated November 28, 2008. Many people from the project partners have contributed to this public document.