Towards Petascale: Application and Benchmarking Activities in the PRACE Project

Dr. Mark Bull EPCC, University of Edinburgh

CAPACITIES

e-infrastructure

Overview

- Introduction to PRACE
- Organisation of the PRACE project
- Applications
- Final benchmark set
- Future architectures
- Future PRACE implementation phase

Introduction to PRACE - 1

- History:
  - First ideas towards European Supercomputing in June 2005
    - Meeting at ISC2005 in Heidelberg: Fr, Ge, NL, Sp, UK
  - Investments in future HPC equipment regarded as beyond the scope and potential of a single country
  - Concluded that “a sustainable HPC infrastructure in Europe, which compares to the US and Japan” would be necessary
    - Tier-0 systems in Europe, comparable to US and Japanese systems
    - How to fund, how to operate and where to install ?
    - What science to facilitate ?
  - But first:
    - Build the science case
    - Seek awareness on European (EC) levels
    - Involve communities and industries

Introduction to PRACE - 2

- Science case:
  - Prepared by HET (HPC in Europe Taskforce):
    - Extensive survey on benefits for science
    - Workshops with scientists
    - Including societal and economic benefits
    - Conducted in 2006EC awareness:
  - High Performance Computing facilities put on the ESFRI roadmap (among 34 other projects)
  - Estimated costs 200-400 M€ on an annual basis (investments and running costs) for one new system each year

Introduction to PRACE - 3

- In 2007:
  - Memorandum of Understanding (MoU) signed:
    - 14 countries: Au, Ch, Fi, Fr, Ge, Gr, It, NL, No, Po, Pt, Sp, Sw, UK
    - To work on the establishment of a European HPC infrastructure
    - To prepare a project proposal to the EC to investigate and to prepare such a European HPC infrastructure
    - To be prepared to match funding to execute the project
    - First known as PACE, later on as PRACE:
      - PaRtnership for Advanced Computing in Europe
  - Project proposal for 20 M€ in total (50% by EC, 50% by 14 countries)
    - 10 M€ by EC in the framework of preparation of the ESFRI projects
  - Project granted and started in January 2008 for 2 years
  - In 2008 and 2009, Ireland, Turkey, Cyprus, Serbia have also joined, more to follow (Bulgaria, Czech)

Organisation of the PRACE project - 1

- 16 partners from 14 countries (Germany with 3 partners: FZJ, HLRS, LRZ)
- Fr: GENCI (CEA, CINES); NL: NCF (with SARA and Univ. Groningen); Sp: BSC; UK: EPSRC (EPCC, STFC); Fi: CSC; CH: CSCS; It: CINECA; ….
- Sustainable Tier-0 infrastructure, preparation phase:
  - "Administrative part": legal issues, peer review, operational and funding models, ownership, seat of PRACE, dissemination, education
  - "Technical part": system software, system management, applications and benchmarking, architecture prototypes, future technology
- Organised in a total of 8 work packages

Organisation of the PRACE project - 2

- Key overall goal:
  - Facilitate European scientific research with top HPC facilities
- In the PRACE project this translates *technically* to:
  - Investigate and improve scalability of applications to petascale and beyond
  - Test applications on petascale equipment
  - Develop a benchmark suite of representative applications for purchasing and testing
  - Investigate future technology for usage in future systems
- This has led to:
  - Applications selection
  - Benchmark definition
  - Prototype systems (mainstream and future)

Prototypes (mainstream technology) - 1

- Hosting partners bought prototypes, matching by PRACE
  - To avoid misunderstanding: PRACE partners have given access to their already existing systems for PRACE usage, and have typically bought a small part in addition to do PRACE-specific work – this is what was matched
- Criteria:
  - Immediate (asap) usage in the PRACE project
  - Coverage of all mainstream architectures:
    - MPP
    - SMP cluster (ThinNode)
    - SMP cluster (FatNode)
    - Hybrid (combination of architecture/processor technology)
  - For both internal and external usage

Prototypes (mainstream technology) - 2

- Results:
  - MPP: IBM BlueGene/P, FZ Juelich, Germany
  - MPP: Cray XT5, CSC, Finland (in combination with CSCS, Switzerland)
  - SMP-ThinNode: Bull cluster with Intel Nehalem, CEA, France: Sun cluster with Intel Nehalem FZJ, Germany
  - SMP-FatNode: IBM Power6 cluster, SARA/NCF, The Netherlands
  - Hybrid: IBM Power6 + Cell, BSC, Spain
  - Hybrid: NEC vector + Intel Nehalem cluster, HLRS, Germany
- Internal usage:
  - Assessment of the prototypes by running synthetic benchmarks
  - Scalability testing and preparation of application benchmark suite
- External usage:
  - Scientific communities with their own codes through light review mechanism
  - Experiences possibly included in PRACE deliverables

Applications - 1

- Goals in PRACE with respect to applications:
  - Investigate scalability of applications
  - Construct a PRACE Application Benchmark Suite (PABS)
- Process (part 1):
  - Determine application areas
  - Determine codes within application areas, based on usage in Europe
  - Define application/prototype matrix with respect to porting efforts
  - Start testing with this Initial PABS (July 2008-May 2009)

Applications - 2

- Application areas and codes:
  - Particle Physics
    - QCD
  - Computational Chemistry
    - VASP, CPMD, CP2K, NAMD, GROMACS, HELIUM, GPAW, SIESTA
  - Computational Fluid Dynamics
    - SATURNE, AVBP, NS3D, ALYA
  - Astronomy and Cosmology
    - GADGET
  - Earth Sciences
    - NEMO, ECHAM5, BSIT
  - Plasma Physics
    - PEPC, TORB/EUTERPE
  - Computational Engineering
    - TRIPOLI_4

Applications - 3

Benchmark code	Languages	Libraries	Programming Model	IO characteristics
QCD	Fortran 90, C		MPI	no special
VASP	Fortran 90	BLACS, SCALAPACK	MPI (+ pthreads)	no special
NAMD	C++	Charm++, FFTW, TCL	Charm++, MPI, master-slave	no special
CPMD	Fortran 77	BLAS, LAPACK	MPI	
Code_Saturne	Fortran 77, C99, python	BLAS	MPI	read at start, write periodically
GADGET	C 89	FFTW, GSL, HDF5	MPI	
TORB	Fortran 90	PETSC, FFTW	MPI	read at start, write periodically
ECHAM5	Fortran 90	BLAS, LAPACK, NetCDF	MPI/OpenMP	read at start, write periodically
NEMO	Fortran 90	NetCDF	MPI	read at start, write periodically
CP2K	Fortran 95	FFTW, LAPCK, ACML	MPI	checkpoints and output, intense
GROMACS	C, assembler	FFTW, BLAS, LAPACK	MPI	read at start, write periodically, relax
NS3D	Fortran 90	EAS3, Netlib (FFT)	MPI + NEC-microtasking	read at start, write periodically
AVBP	Fortran 90	Hdf5, szip, Metis	MPI	read at start, write periodically
HELIUM	Fortran 90		MPI	read at start, write periodically
TRIPOLI_4	C++		TCP/IP sockets	read at start, write periodically
PEPC	Fortran 90		MPI	read at start, write periodically
GPAW	Python, C	LAPACK, BLAS	MPI	read at start, write at end
ALYA	Fortran 90	Metis	MPI/OpenMP	read at start, write periodically
SIESTA	Fortran 90	Metis, BLAS, SCALAPACK	MPI	read at start, write periodically
BSIT	Fortran 95, C	Compression lib	MPI/OpenMP	read at start, write periodically

Applications - 4

- What kind of work ?
  - Porting of applications to selected prototypes (if not done already)
  - Profiling and optimisation work
    - Recommendations
    - Best practices
  - Investigation and improvement of scalability
    - Recommendations
    - Best practices
  - Prepare for decision-making process on final benchmark suite
- Per application, one person responsible for work on all aspects
  - Responsibility includes porting, optimisation and petascaling
  - Benchmark Code Owner (BCO)
  - May manage a team of co-workers

Applications - 5

- Other assumptions/goals:
  - Each prototype platform has at least 5 to 6 applications
  - Each application is ported and run on at least 3 prototype platforms (if possible)
  - Have proportional distribution of applications to PRACE partners (BCOs) with respect to available manpower (pm's)

Applications - 6

- PRACE aims:
  - To facilitate European scientific research with top HPC facilities
- Hence, needs to cover:
  - Relevant scientific areas – already done
  - Scalability (potential) on Petascale systems
  - Licensing aspects with respect to further usage
  - US/Japan efforts on applications
  - Extension in certain scientific areas
- These aspects (part 2 of the process) have been investigated to define the final PRACE Application Benchmark Suite

Final Benchmark Set

- Final PABS:
  - Particle Physics
    - QCD
  - Computational Chemistry
    - CPMD, CP2K, NAMD, GROMACS, HELIUM, GPAW, Quantum_Espresso, Octopus
  - Computational Fluid Dynamics
    - SATURNE, AVBP, NS3D, ALYA
  - Astronomy and Cosmology
    - GADGET
  - Earth Sciences
    - NEMO, BSIT, SPECFEM3D, WRF
  - Plasma Physics
    - PEPC, TORB/EUTERPE
  - Computational Engineering
    - TRIPOLI_4, ELMER

Current porting status

Application	MPP-BG	MPP-Cray	SMP-TN-x86	SMP-FN-pwr6	SMP-FN+Cell	SMP-TN+vector
QCD	Done	Done	Done	Done		
Quantum_Espresso	Done	Done	Done	Done		Done
NAMD	Done	Done	Done	Done		
CPMD	Done		Done	Done	Done	Done
Code_Saturne	Done	Done		Done		Done
GADGET	Done		Done	Done		
TORB/EUTERPE	Done			Done	In progress	
WRF	Done	Done	Done	Done		
NEMO	Done	Done		Done		Done
CP2K	Done	Done		Done		
GROMACS	Done	Done		Done		
NS3D		Done	In progress	Done		Done
AVBP	Done		Done	Done		
HELIUM	Done	Done	Done	Done		Done
TRIPOLI_4	In progress		Done			
PEPC	Done	Done	Done	Done		
GPAW	Done	Done		Done		
ALYA				Done	Done	
OCTOPUS	Done			Done	Done	
BSIT				Done	Done	
ELMER		Done		Done		
SPECFEM3D		Done				

Future architectures - 1

- Investigation of future Petaflop/s computer technologies beyond 2010:
  - Set up a permanent structure to enable research on future technology and software, with academic and industrial collaborators (STRATOS)
  - STRATOS:
    - PRACE advisory group for **Stra**tegic **T**echnol**o**gie**s**
  - Define promising technologies for mainstream use in the 2011-2012 time frame, and install these as prototype systems
  - Has led to another set of more experimental prototypes (which include software developments)
  - Access to these prototypes to be discussed with respective owners

Future architectures - 2

Sites	Hardware/Software	Porting effort
CEA "*GPU/CAPS*"	1U Tesla Server T1070 (CUDA, CAPS, DDT) Intel Harpertown nodes	"Evaluate GPU accelerators and GPGPU programming models and middleware." (e.g., *pollutant migration code* (ray tracing algorithm) to CUDA and HMPP)
CINES-LRZ "*LRB/CS*"	Hybrid SGI ICE2/UV/Nehalem-EP & Nehalem-EX/ClearSpeed/ Larrabee	**Gadget**,SPECFEM3D_GLOBE, RaXml, Rinf, RandomAccess, ApexMap, Intel MPI BM
CSCS *"UPC/CAF"*	Prototype PGAS language compilers (CAF + UPC for Cray XT systems)	"The applications chosen for this analysis will include some of those already selected as **benchmark codes in WP6**."
EPCC "*FPGA*"	Maxwell – FPGA prototype (VHDL support & consultancy + software licenses (e.g., Mitrion-C))	"We wish to port several of the **PRACE benchmark codes** to the system. The codes will be chosen based on their suitability for execution on such a system."

Future architectures - 3

Sites	Hardware/Software	Porting effort
FZJ (BSC) "*Cell & FPGA interconnect*"	eQPACE (PowerXCell cluster with special network processor)	Extend FPGA-based interconnect beyond QCD applications.
LRZ "*RapidMind*"	RapidMind (Streaming Processing Programming Paradym) X86, GPGPU, Cell	ApexMap, Multigrid, FZJ (QCD), CINECA (linear algebra kernels involved in solvers for ordinary differential equations), SNIC
NCF "*ClearSpeed*"	ClearSpeed CATS 700 units	Astronomical many-body simulation, Iterative sparse solvers with preconditioning, finite element code, cryomicrotome image analysis
CINECA	I/O Subsystem (SSD, Lustre, pNFS)	-
KTH	Standalone system with AMD Istanbul 6-core CPUs	Exploit novel AMD power control features to maximize energy efficiency

Future PRACE implementation phase

- PRACE project ended at 31-Dec-2009
  - Extension to 31-Jun-2010 to consume unspent budget
- Final deliverables of the project prepare for continuation and/or implementation:
  - Legal aspects like statutes, office seat
  - Procurement schedules
  - Tier-0 access, operation and ownership models
  - European-wide peer review
  - Benchmark set for future procurements

Applications in implementation phase

- Very important part of implementation phase
  - approx 50% of the staff effort devoted to this
- Much more user-focused than the first project
- Main objective is to ensure that applications use time on the PRACE systems as efficiently as possible
  - optimisation
  - scaling
  - choice of system

Applications enabling for capability science

- Ensure effective exploitation of the PRACE Tier-0 systems by petascaling and optimising applications on these systems

- The codes chosen for the applications-enabling work will be selected through either:

1. a competitive process for capability science, or
2. collaborations with applications communities

- Competitive process will be through regular calls for proposals
  - lightweight assessment process.
  - successful applicants will receive ~ 6 months support for optimising/scaling a code or solving a specific science problem.
  - applicants can apply for machine cycles under Preparatory Access scheme
  - will encourage applications from new and emerging science areas, and from all partner countries

Efficient use of Tier-0 systems

- Aim to produce best practice guides to help users make efficient use of the systems
- Topics for best practice guides will include:
  - optimal porting of applications (e.g., choice of numerical libraries and compiler options)
  - architecture-specific optimisation and petascaling techniques
  - optimal system environment (e.g., tuneable system parameters, job placement and optimised system libraries)
  - debuggers, performance analysis tools and programming environment.
- Produce system-specific guides in co-operation with hosting partners
- Collaboration with training programme

Applications Requirements for Tier-0 Systems

- Continue to assess requirements by running application and synthetic benchmarks on the Tier-0 systems
- Assess impact of possible future architectures through performance modelling
- Maintain the benchmark suites and update them based on surveys of actual and potential usage of Tier-0 systems.
  - will include requirements for data storage, libraries, and tools as well

Programming Techniques for High Performance Applications

- Support migration of applications to novel programming techniques

- Investigations and best practice guides covering:
  - analysis of scalable algorithms and libraries to enhance scientific applications
  - optimisation of applications on multi-core/many core systems
  - exploitation of accelerators for real applications
  - porting of applications to novel HPC languages and paradigms

Efficient Handling of Petascale-Class Applications Data

- Investigations into:
  - scalable parallel pre-processing on Tier-0 systems
  - parallel and hierarchical I/O (including applications coupling and data exchange in multi-applications)
  - long-term preservation of applications data
  - upstream approaches to alleviate I/O and post-processing workload, e.g. in situ analysis, data filtering and reduction
  - post-processing and requirements for visualisation tools.

Thank you !

Questions ?

My thanks to Peter Michielse (NCF) for some of the material in this presentation