SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2010-2.3.1 – First Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-1IP

PRACE First Implementation Project

Grant Agreement Number: RI-261557

D7.3
Petascaling and Optimisation Guides for PRACE Systems

Final

Version:	1.0
Author(s):	Jacques David, CEA Jeroen Engelberts, SARA Xu Guo, EPCC Florian Janetzko, FZJ Walter Lioen, SARA
Date:	19.06.2012

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-261557**	
	Project Title: PRACE First Implementation Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D7.3**>	
	Deliverable Nature: <DOC_TYPE: Report>	
	Deliverable Level: PU*	**Contractual Date of Delivery:** 30 / June / 2012
		Actual Date of Delivery: 30 / June / 2012
	EC Project Officer: Thomas Reibe	

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

Document	**Title: <Petascaling and Optimisation Guides for PRACE Systems>**	
	ID: <D7.3>	
	Version: <1.0 >	**Status:** Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2007	
	File(s): D7.3.docx	
Authorship	**Written by:**	Jacques David, CEA Jeroen Engelberts, SARA Xu Guo, EPCC Florian Janetzko, FZJ Walter Lioen, SARA

	Contributors:	Nikos Anastopoulos, ICCS Lilit Axner, KTH Eric Boyer, CINES Mirko Cestari, CINECA Gilles Civario, ICHEC Guillaume Colin de Verdière, CEA Guillaume Collet, CEA Cédric Coquebert, CEA Maciej Cytowski, ICM Jacques David, CEA Jeroen Engelberts, SARA Bruno Frogé, CEA Silvia Giuliani, CINECA Theodoros Gkountouvas, GRNET Xu Guo, EPCC Alexandros Haritatos, ICCS Florian Janetzko, FZJ Stefanie Janetzko, FZJ Walter Lioen, SARA Konstantinos Nikas, GRNET Hilde Ouvrard, CINES Jean-Noel Richet, CEA Jorge Rodriguez, BSC Elda Rossi, CINECA Alexander Schnurpfeil, FZJ Huub Stoffers, SARA Maciej Szpindler, ICM Andrew Turner, EPCC Tyra Van Olmen, CINES David Vicente, BSC Brian Wylie, FZJ
	Reviewed by:	Thomas Eickermann, FZJ Wilhelm Homberg, FZJ
	Approved by:	MB/TB

Document Status Sheet

Version	**Date**	**Status**	**Comments**
0.1	01/June/2012	Draft	First draft
0.2	09/June/2012	Internal review	
1.0	19/June/2012	Final version	

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Best Practice Guide

Disclaimer

This deliverable has been prepared by the responsible Work Package of the Project in accordance with the Consortium Agreement and the Grant Agreement n° RI-261557 . It solely reflects the opinion of the parties to such agreements on a collective basis in the context of the Project and to the extent foreseen in such agreements. Please note that even though all participants to the Project are members of PRACE AISBL, this deliverable has not been approved by the Council of PRACE AISBL and therefore does not emanate from it nor should it be considered to reflect PRACE AISBL's individual opinion.

Copyright notices

Table of Contents

References and Applicable Documents

[1] PRACE RI web site, http://www.prace-ri.eu/

[2] Florian Janetzko, Gilles Civario, Maciej Cytowski, Maciej Szpindler, Stefanie Janetzko, Alexander Schnurpfeil, Brian Wylie, and Huub Stoffers. *Best Practice Guide – JUGENE*. PRACE-1IP D7.3, June 2012. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-JUGENE.pdf, http://www.prace-ri.eu/Best-Practice-Guide-JUGENE-HTML.

[3] Jacques David, Jean-Noel Richet, Eric Boyer, Nikos Anastopoulos, Guillaume Collet, Guillaume Colin de Verdière, Tyra Van Olmen, Hilde Ouvrard, Cédric Coquebert, Bruno Frogé, Alexandros Haritatos, Konstantinos Nikas, and Theodoros Gkountouvas. *Best Practice Guide – Curie*. PRACE-1IP D7.3, June 2012. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Curie.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Curie-HTML.

[4] Andrew Turner, Xu Guo, and Lilit Axner. Best *Practice Guide – Cray XE*. PRACE-1IP D7.3, June 2012. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Cray-XE.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Cray-XE-HTML.

[5] Jeroen Engelberts, Walter Lioen, Huub Stoffers, Mirko Cestari, Silvia Giuliani, Elda Rossi, Jorge Rodriguez, David Vicente, and Guillaume Collet. *Best Practice Guide – IBM Power*. PRACE-1IP D7.3, June 2012. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-IBM-Power.pdf, http://www.prace-ri.eu/Best-Practice-Guide-IBM-Power-HTML.

[6] http://www.docbook.org/

[7] http://tldp.org/LDP/LDP-Author-Guide/html/docbook-why.html

List of Acronyms and Abbreviations

AISBL	Association Internationale Sans But Lucratif
BSC	Barcelona Supercomputing Center (Spain)
CEA	Commissariat à l’Energie Atomique (represented in PRACE by GENCI, France)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CINES	Centre Informatique National de l’Enseignement Supérieur (represented in PRACE by GENCI, France)
CMS	Content Management System
CSC	Finnish IT Centre for Science (Finland)
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DoW	Description of Work
DP	Double Precision, usually 64-bit floating point numbers
EC	European Community
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
FZJ	Forschungszentrum Jülich (Germany)
GENCI	Grand Equipement National de Calcul Intensif (France)
GRNET	Greek Research & Technology Network (Greece)
HLRS	Höchstleistungsrechenzentrum Stuttgart (Germany)
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HTML	HyperText Markup Language
IBM	Formerly known as International Business Machines
ICCS	Institute of Communications and Computer Systems (Greece)
ICHEC	Irish Centre for High-End Computing (Ireland)
ICM	Interdyscyplinarne Centrum Modelowania Matematycznego (Poland)
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
I/O	Input/Output
IP	Implementation Project
IT	Information Technology
JSC	Jülich Supercomputing Centre (FZJ, Germany)
JUGENE	Jülich Blue Gene
KTH	Kungliga Tekniska Högskolan (Sweden)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MB	Management Board
MCM	Multi-Chip Module
MPI	Message Passing Interface
NUMA	Non-Uniform Memory Access
OpenMP	Open Multi-Processing
PDF	Portable Document Format
PFlop/s	Peta (= 10^{15}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also PF/s
PP	Preparatory Project
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)

RI	Research Infrastructure
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SNIC	Swedish National Infrastructure for Computing (Sweden)
ssh	Secure Shell
STFC	Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
SVN	Apache Subversion
TB	Technical Board
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
WP	Work Package
XML	Extensible Markup Language

Executive Summary

Work Package 7 'Enabling Petascale Applications: Efficient Use of Tier-0 Systems' (WP7) ensures the effective exploitation of the PRACE Tier-0 systems by increasing scalability and performance of applications.

While the focus in WP7 is primarily on enabling applications for Tier-0 systems, the tasks should also benefit applications performance on Tier-1 systems. Most of the activities had to commence on existing Tier-1 systems while waiting for the installation of the Tier-0 systems.

Task 7.3 is called 'Efficient Use of PRACE Systems'. Efficient use of the PRACE Tier-0 systems requires detailed knowledge of architecture-specific factors influencing performance, including compilers, tools and libraries. The main goal of this task is to investigate such issues, collect best practices on how to achieve good performance on the systems, and disseminate this knowledge to users.

Topics for best practice guides 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.

Task 7.3 covers two Tier-0 systems (JUGENE and Curie) and two common Tier-1 system families (Cray XE and IBM Power) with the potential to evolve in a Tier-0 system.

Finally, since the four best practice guides comprise some 50 – 80 pages each, we decided not to include them as separate chapters in this report but to refer to the online versions on the PRACE RI web site [1] instead (cf. [2], [3], [4], [5]).

1 Introduction

As stated in the DoW, efficient use of the PRACE Tier-0 systems requires detailed knowledge of architecture-specific factors influencing performance, including compilers, tools and libraries. The main goal of this task is to investigate such issues, collect best practices on how to achieve good performance on the systems, and disseminate this knowledge to users.

The purpose of this report is to give a description of the process which led to the best practice guides itself.

In Section 2 we describe: the selection of the systems; the subtask strategy; the technology used for creating the best practice guides; and finally, the generic table of contents.

According to the DoW, this deliverable, D7.3 'Petascaling and Optimisation Guides for PRACE Systems', should present the final version of best practice guides with a separate chapter for each system. However, since the four best practice guides comprise some 50 – 80 pages each, we decided not to include them as separate chapters in this report but to refer to the online versions (both HTML as well as PDF) on the PRACE RI web site [1] instead (cf. [2], [3], [4], [5]).

The intended audience for the best practice guides are the users of the four selected systems.

2 Approach to Best Practice Guides

While the focus in WP7 is primarily on enabling applications for Tier-0 systems, the tasks should also benefit application performance on Tier-1 systems. Most of the activities had to commence on existing Tier-1 systems while waiting for the installation of the Tier-0 systems.

2.1 Selection of Systems

At the start of PRACE-1IP (July 1, 2010), the only readily available Tier-0 system was GCS' Blue Gene/P installed at FZJ: JUGENE (Jülich Blue Gene). The second Tier-0 system would become GENCI's Bull system (Intel based) installed at TGCC/CEA: Curie (the first phase was completed end 2010; the full system became available in October 2011). These two systems were obvious candidates for writing a best practice guide.

After discussion with several people from BSC, CINECA, EPCC, SARA we additionally selected two Tier-1 architecture / potential future Tier-0 architectures:

- Cray XE
- IBM Power

With time, it became apparent that Hermit (GCS' Cray XE6 installed at HLRS, with a 1 PFlop/s installation step 1 becoming operational in December 2011) would fit the Cray XE family best practice guide. Because of lack of human resources (most notably since HLRS was not involved in T7.3) we were not able to transform the generic Cray XE best practice guide into a Hermit best practice guide.

2.2 Subtask Strategy

Clearly, the natural subtask is at the system level: subtask leaders preferably coming from the hosting partner of the relevant system. The respective subtask leaders were:

A. Florian Janetzko, FZJ, for JUGENE

B. Jacques David, CEA, for Curie

C. Xu Guo, EPCC, for Cray XE

D. Jeroen Engelberts, SARA, for IBM Power

The task was lead by Walter Lioen, SARA.

2.3 Technology

Although all PRACE deliverables are created using Microsoft Word, this did not seem to be the appropriate technology for creating the best practice guides. Having the M9 milestone ‘Electronic versions available on web’ it would have been possible to create cross-referenced PDF documents from Microsoft Word, however, these would not have been true web versions.

It was decided that high quality HTML versions as well as high quality, fully featured PDF versions would be created and made available.

For this an assessment of several technologies was performed, including but not limited to DocBook and the authoring tools used by DEISA and DEISA 2 (some 10 different documents, say).

DEISA was basically using Plone (a CMS) for creating web-based documentation and plone2pdf (having Zopyx support) for creating PDF documents from the web-based documentation (possible downside: not fully featured PDF version). Additional features: collaborative editing environment, roles, possible continuation in PRACE-2IP.

DocBook (cf. [6], [7]) is being used by a lot of open source projects amongst others by the Linux Documentation Project. The key feature is having single (XML) source (which is tracked using svn) and multiple fully cross-referenced output formats: HTML, PDF and more.

The technology assessment resulted in a recommendation to use DocBook after which T7.3 decided unanimously to follow this recommendation.

2.4 Generic Table of Contents

All best practice guides are created based on the same generic table of contents:

1. Introduction
2. System Architecture / Configuration
    1. Processor architecture / MCM architecture (including caches)
    2. Building block architecture (node cards, nodes, drawers, supernodes, racks)
    3. Memory architecture (including NUMA effects)
    4. (Node) Interconnect (including topology, system specific)
    5. I/O subsystem architecture (being system specific and not architecture specific!)
    6. Available File Systems
        1. Home, scratch, long time storage
        2. Performance of file systems
3. System Access
    1. How to reach the system (ssh, portals, file transfer, ...)
4. Production Environment

   1. Module environment
   2. Batch System
   3. Accounting

5. Programming Environment / Basic Porting
   1. Available compilers
      1. Compiler flags
   2. Available (vendor optimised) numerical libraries
   3. Available MPI implementations
   4. OpenMP
      1. Compiler flags
   5. Batch system / job command language
6. Performance analysis
   1. Available Performance Analysis Tools
   2. Hints for interpreting results.
7. Tuning
   1. Advanced / aggressive compiler flags
   2. Single core optimisation
   3. Advanced MPI usage
      1. Tuning / environment variables
      2. Mapping tasks on node topology
      3. Task affinity
      4. Adapter affinity
   4. Advanced OpenMP usage
      1. Tuning / environment variables
      2. Thread affinity
   5. Hybrid programming
      1. Optimal tasks / threads strategy
   6. Memory optimisation
      1. Memory affinity (MPI/OpenMP/Hybrid)
      2. Memory allocation (malloc) tuning
      3. Using huge pages
   7. I/O optimisation (Tuning / scaling of Application I/O)
   8. Advanced job command language (includes defining task topology, affinity, etc.)
   9. Possible kernel parameter tuning (probably less relevant to the ‘average’ user but possibly relevant for large production runs)

8. Debugging
    1. Available Debuggers
    2. Compiler flags

The actual table of contents of the individual guides slightly deviate from this generic one, to best reflect systems specifics.

2.5 Content

For all systems an inventory of the existing documentation was made that could be used as base material for some of the topics mentioned above. Many topics had to be complemented or even written from scratch. Apart from this, experiences learned during the enabling activities in T7.1, T7.2, T7.5, and T7.6 were added. For selected cases, real life experiences have been incorporated as use cases in the best practice guides.

As an internal quality assurance, T7.3 subtask-internal reviews and subtask cross-reviews (every subtask leader did a review of another best practice guide) were performed.

3 Best Practice Guides

The four Best Practice Guides are to be found online:

- Best Practice Guide – JUGENE (cf. [2])
- Best Practice Guide – Curie (cf. [3])
- Best Practice Guide – Cray XE (cf. [4])
- Best Practice Guide – IBM Power (cf. [5])