SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2011-2.3.5 – Second Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-2IP

PRACE Second Implementation Project

Grant Agreement Number: RI-283493

D7.3
Petascaling and Optimisation for Tier-1 Architectures

Final

Version: 1.0
Author(s): Vegard Eide, NTNU
Walter Lioen, SURFsara
Maciej Szpindler, ICM
Volker Weinberg, LRZ
Date: 24.05.2013

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-283493**	
	Project Title: PRACE Second 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:** 31 / May / 2013
		Actual Date of Delivery: 31 / May / 2013
	EC Project Officer: Leonardo Flores Añover	

* - 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 for Tier-1 Architectures**	
	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:**	Vegard Eide, NTNU Walter Lioen, SURFsara Maciej Szpindler, ICM Volker Weinberg, LRZ
	Contributors:	Nikos Anastopoulos, NTUA Michał Białoskórski, PSNC Michael Browne, ICHEC Gilles Civario, ICHEC Vegard Eide, NTNU Alan Gray, EPCC Nevena Ilieva-Litova, NCSA Michael Lysaght, ICHEC Eoin McHugh, ICHEC Henrik Nagel, NTNU Petri Nikunen, CSC Anders Sjöström, LUNARC Roman Slíva, VSB Filip Staněk, VSB Maciej Szpindler, ICM Tyra Van Olmen, CINES Volker Weinberg, LRZ Niall Wilson, ICHEC

	Reviewed by:	Florian Berberich, FZJ Krasimir Georgiev, NCSA
	Approved by:	MB/TB

Document Status Sheet

Version	**Date**	**Status**	**Comments**
0.1	03/May/2013	Draft	First draft
0.2	10/May/2013	Internal review	
1.0	24/May/2013	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-283493. 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.

Table of Contents

References and Applicable Documents

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

[2] Vegard Eide, Nikos Anastopoulos, and Henrik Nagel, *Best Practice Guide – Generic x86*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Generic-x86.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Generic-x86-HTML.

[3] Roman Slíva and Filip Staněk, *Best Practice Guide – Anselm*, PRACE 2IP-D7.3, May 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Anselm.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Anselm-HTML.

[4] Michał Białoskórski and Maciej Szpindler, *Best Practice Guide – Chimera*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Chimera.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Chimera-HTML.

[5] Alan Gray, Anders Sjöström, Nevena Ilieva-Litova, and Maciej Szpindler (ed.), *Best Practice Guide – GPGPU*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as:

http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-GPGPU.pdf,
http://www.prace-ri.eu/Best-Practice-Guide-GPGPU-HTML.

[6] Maciej Szpindler (ed.), *Best Practice Guide – Hydra*, PRACE 2IP-D7.3, May 2013. Available at the PRACE RI web site [1] as:
http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Hydra.pdf,
http://www.prace-ri.eu/Best-Practice-Guide-Hydra-HTML.

[7] Tyra Van Olmen, *Best Practice Guide – Jade*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as:
http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Jade.pdf,
http://www.prace-ri.eu/Best-Practice-Guide-Jade-HTML.

[8] Maciej Szpindler (ed.), *Best Practice Guide – JUROPA*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as:
http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-JUROPA.pdf,
http://www.prace-ri.eu/Best-Practice-Guide-JUROPA-HTML.

[9] Michael Lysaght, Niall Wilson, Eoin McHugh, Michael Browne, and Gilles Civario, *Best Practice Guide – Stokes*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as:
http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Stokes.pdf,
http://www.prace-ri.eu/Best-Practice-Guide-Stokes-HTML.

[10] Nikos Anastopoulos, Petri Nikunen, and Volker Weinberg, *Best Practice Guide – SuperMUC*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as:
http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-SuperMUC.pdf,
http://www.prace-ri.eu/Best-Practice-Guide-SuperMUC-HTML.

[11] Jacques David, Jeroen Engelberts, Xu Guo, Florian Janetzko, and Walter Lioen, Petascaling and Optimisation Guides for PRACE Systems, PRACE-1IP D7.3, June 2012. Available at the PRACE RI web [1] as: http://www.prace-ri.eu/IMG/pdf/d7.3_1ip.pdf.

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

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

List of Acronyms and Abbreviations

AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
CINES	Centre Informatique National de l'Enseignement Supérieur (represented in PRACE by GENCI, France)
CPU	Central Processing Unit
CSC	Finnish IT Centre for Science (Finland)
DECI	Distributed European Computing Initiative
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DoW	Description of Work
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)
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HTML	HyperText Markup Language
HX5	(IBM)
IBM	Formerly known as International Business Machines
ICE	(SGI)
ICHEC	Irish Centre for High-End Computing (Ireland)
ICM	Interdyscyplinarne Centrum Modelowania Matematycznego (Poland)
I/O	Input/Output
IP	Implementation Project
IT	Information Technology
IT4I	IT4Innovations (Czech Republic)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
LUNARC	The center for scientific and technical computing at Lund University (Sweden)
MB	Management Board
MCM	Multi-Chip Module
MPI	Message Passing Interface
NCSA	National Centre for Supercomputing Applications (Bulgaria)
NTNU	Norges Teknisk-Naturvitenskapelige Universitet (Norwegian University of Science and Technology, Norway)
NTUA	National Technical University of Athens (Greece)
NUMA	Non-Uniform Memory Access or Architecture
OpenMP	Open Multi-Processing
OS	Operating System
PDF	Portable Document Format
PP	Preparatory Project
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)
RI	Research Infrastructure

RZG	Rechenzentrum Garching (Germany)
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SGI	Silicon Graphics, Inc.
ssh	Secure Shell
SURFsara	Dutch national High Performance Computing & e-Science Support Centre (previously known as SARA)
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
UV	Ultra Violet (SGI)
VSB	Vysoká škola báňská - Technical University of Ostrava (Czech Republic)
WP	Work Package
XML	Extensible Markup Language

Executive Summary

Work Package 7 ‘Scaling Applications for Tier-0 and Tier-1 Users’ provides medium-term petascaling and optimisation support for European HPC applications to ensure that they can make effective use of both Tier-0 and Tier-1 systems. Such applications-enabling projects typically last around six months and provide direct benefits to European researchers.

Through the applications-enabling work, WP7 develops specific expertise on most, if not all, of the architectures which make up the top two tiers of the European HPC system. PRACE-1IP provided best practice guides on the PRACE Tier-0 systems that cover programming techniques, compilers, tools and libraries (cf. [11]). PRACE-2IP supplements these with best practice guides for the other architectures which are important at Tier-1 to allow European researchers to make efficient use of these systems.

PRACE-2IP Task 7.3 is called ‘Best Practice Guides for Tier-1 Architectures’. The main goal of this task is to investigate the efficient use of HPC systems, collect best practices on how to achieve good performance on the systems, and disseminate this knowledge through best practice guides. The target audience are users and support staff who are developing and enabling applications.

Topics for these 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 a generic x86 cluster guide, six system-specific mini-guides for x86 clusters, a generic mini-guide for GPGPU-accelerated clusters, and the SuperMUC guide. The SuperMUC guide admittedly describes a Tier-0 system, however, it started in 2011 with the description of a Tier-1 system: SuperMIG, the SuperMUC migration system.

Finally, because of the total size of the best practice guides, 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], [6], [7], [8], [9], [10]).

1 Introduction

Efficient use of PRACE 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 subtasks, 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 for Tier-1 Architectures’, should present the final version of the best practice guides with a separate chapter for each guide. However, because of the total size of the best practice guides, 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], [6], [7], [8], [9], [10]).

The target audience are users and support staff who are developing and enabling applications.

2 Approach to Best Practice Guides

In the DoW we announced: a generic x86 cluster guide, system-specific mini-guides for x86 clusters, and guides for at least one other architecture.

2.1 Selection of Systems

For the selection of the mini-guides, we started with the list of Tier-1 systems that is being used for the successive DECI calls. Obviously, we tried to match this list with contributors from the corresponding PRACE partners. Furthermore, Hydra and JUROPA have been added based on the available documentation.

Apart from the generic x86 guide and the system-specific mini-guides for x86 clusters, we had to select one other architecture. Here, we selected GPGPU-accelerated clusters.

Finally, we selected SuperMUC. The SuperMUC guide admittedly describes a Tier-0 system, however, it started in 2011 with the description of a Tier-1 system: SuperMIG, the SuperMUC migration system.

2.2 Subtasks

The respective subtask leaders were:

A. Vegard Eide, NTNU, for the generic x86 cluster guide
B. Maciej Szpindler, ICM, for system specific mini-guides for x86 clusters and for a generic mini-guide for GPGPU-accelerated clusters
C. Volker Weinberg, LRZ, for SuperMUC

2.3 Technology

We built on the experience obtained during the corresponding PRACE-1IP task (cf. [11]). Although all PRACE deliverables are created using Microsoft Word, this did not seem to be the appropriate technology for creating the best practice guides. It was decided that high quality HTML versions as well as high quality, fully featured PDF versions would be created and made available. To reach this goal, we use DocBook. DocBook (cf. [12], [13]) 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.

A new standardized unified setup has been created for the benefit of the PRACE-2IP (and upcoming PRACE-3IP) best practice guides. The reason for this was the divergence that started during the work on the PRACE-1IP best practice guides. For this we created a much more uniform and self-contained setup (which works on Linux, Windows with Cygwin, and Mac OS); converted/migrated all best practice guides; and created an empty (DocBook) template for new best practice guides.

2.4 Generic Table of Contents

For PRACE-1IP, all best practice guides were created based on the same generic table of contents. We took a similar approach for PRACE-2IP. A fundamental difference is having both a generic x86 guide and many system specific mini-guides. For this we decided to move as much information as possible to the generic x86 guide moving a limited number of system specific items from the generic table of contents to the mini-guides.

We won’t list the generic x86 / mini-guide table of contents but the generic table of contents can be found below.

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 tables 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 other tasks 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 best practice (mini-)guides itself are to be found online.

3.1 Best Practice Guide – Generic x86 (cf. [2])

This guide provides an overview of best practices on using x86 HPC cluster systems. It is not system-specific, i.e. it does not cover information that is targeted for one specific system but instead it focuses on topics that are common to systems based on the x86 architecture. To get details about architectures and configurations for specific systems users are referred to system specific mini-guides, some of which can be found below.

3.2 Best Practice Guide – Anselm (cf. [3])

This mini-guide describes Anselm, a Bull bullx system. It is operated by IT4Innovations and it is hosted at VSB which is the principal partner of IT4Innovations.

3.3 Best Practice Guide – Chimera (cf. [4])

This mini-guide describes Chimera, an SGI Altix UV system, installed at PSNC.

3.4 Best Practice Guide – GPGPU (cf. [5])

This generic mini-guide describes GPGPU-accelerated clusters. *General-purpose* GPU computing is the use of GPUs to do general purpose scientific and engineering computing. The model for GPU computing is to use a CPU and GPU together in a heterogeneous co-processing computing model. The sequential part of the application runs on the CPU and the computationally intensive part is accelerated by the GPU. From the user's perspective, the application just runs faster because it is exploiting the high-performance of the GPU to boost the overall application performance.

3.5 Best Practice Guide – Hydra (cf. [6])

This mini-guide describes Hydra, an IBM iDataPlex system at RZG.

3.6 Best Practice Guide – Jade (cf. [7])

This mini-guide describes Jade, an SGI Altix ICE system at CINES.

3.7 Best Practice Guide – JUROPA (cf. [8])

This mini-guide describes JUROPA, a cluster consisting of Bull NovaScale servers and Sun blade servers. The system was designed by experts from the Jülich Supercomputing Centre and was implemented together with the partner companies: Bull, Sun, Intel, Mellanox and ParTec. JUROPA is installed at FZJ.

3.8 Best Practice Guide – Stokes (cf. [9])

This mini-guide describes Stokes, an SGI Altix ICE system at ICHEC.

3.9 Best Practice Guide – SuperMUC (cf. [10])

This guide describes SuperMUC, an IBM system at LRZ. The main part is an IBM iDataPlex system. The other part is an IBM BladeCenter HX5 system, which originally constituted the SuperMUC migration system named SuperMIG. SuperMUC is a Tier-0 system.