SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2012-2.3.1 – Third Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-3IP

PRACE Third Implementation Phase Project

Grant Agreement Number: RI-312763

D7.3.1
Best Practice Guides for New and Emerging Architectures

Final

Version: 1.0
Author(s): Nikos Anastopoulos, GRNET
Maciej Cytowski, ICM
Mark Filipiak, EPCC
Walter Lioen, SURFsara
Philippe Wautelet, IDRIS
Volker Weinberg, LRZ

Date: 19.12.2013

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="6">PRACE Project</td><td colspan="2">Project Ref. №: RI-312763</td></tr>
<tr><td colspan="2">Project Title: PRACE Third Implementation Phase Project</td></tr>
<tr><td colspan="2">Project Web Site: http://www.prace-project.eu</td></tr>
<tr><td colspan="2">Deliverable ID: < D7.3.1></td></tr>
<tr><td colspan="2">Deliverable Nature: <DOC_TYPE: Report></td></tr>
<tr><td>Deliverable Level:
PU *</td><td>Contractual Date of Delivery:
31 / December / 2013

Actual Date of Delivery:
30 / December / 2013</td></tr>
<tr><td colspan="2">EC Project Officer: Leonardo Flores Añover</td></tr>
</table>

* - 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

<table>
<tr><td rowspan="6">Document</td><td colspan="2">Title: Best Practice Guides for New and Emerging Architectures</td></tr>
<tr><td colspan="2">ID: D7.3.1</td></tr>
<tr><td>Version: <1.0 ></td><td>Status: Final</td></tr>
<tr><td colspan="2">Available at: http://www.prace-project.eu</td></tr>
<tr><td colspan="2">Software Tool: Microsoft Word 2007</td></tr>
<tr><td colspan="2">File(s): D7.3.1.docx</td></tr>
<tr><td>Authorship</td><td>Written by:</td><td>Nikos Anastopoulos, GRNET
Maciej Cytowski, ICM
Mark Filipiak, EPCC
Walter Lioen, SURFsara
Philippe Wautelet, IDRIS
Volker Weinberg, LRZ</td></tr>
</table>

	Contributors:	Evangelia Athanasaki, GRNET Lilit Axner, KTH Michaela Barth, KTH Margareta Boiarciuc, IDRIS Eric Boyer, CINES Mikko Byckling, CSC Cédric Cocquebert, CEA Guillaume Colin de Verdière, CEA Guillaume Collet, CEA Maciej Cytowski, ICM Jacques David, CEA Jean-Michel Dupays, IDRIS Bruno Frogé, CEA Silvia Giuliani, CINECA Theodoros Gkountouvas, GRNET Massimiliano Guarrasi, CINECA Xu Guo, EPCC Alexandros Haritatos, ICCS Nevena Ilieva, NCSA Giuseppa Muscianisi, CINECA Konstantinos Nikas, GRNET Hilde Ouvrard, CINES Nikela Papadopoulou, GRNET Jean-Noel Richet, CEA Sami Saarinen, CSC Michael Schliephake, KTH Andrew Turner, EPCC Tyra van Olmen, CINES
	Reviewed by:	Veronica Teodor, FZJ Vít Vondrák, VSB
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	29/November/2013	Draft	First Draft
0.2	05/December/2013	Internal review	
1.0	19/December/2013	Final version	

Document Keywords

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

Table of Contents

References and Applicable Documents

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

[2] Michaela Barth, Mikko Byckling, Nevena Ilieva, Sami Saarinen, Michael Schliephake, and Volker Weinberg (ed.), *Best Practice Guide – Intel Xeon Phi*, produced by PRACE-3IP, December 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Intel-Xeon-Phi.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Intel-Xeon-Phi-HTML.

[3] Philippe Wautelet, Margareta Boiarciuc, Jean-Michel Dupays, Silvia Giuliani, Massimiliano Guarrasi, Giuseppa Muscianisi, and Maciej Cytowski, *Best Practice Guide – Blue Gene/Q*, produced by PRACE-3IP, December 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-BlueGene-Q.pdf http://www.prace-ri.eu/Best-Practice-Guide-Blue-Gene-Q-HTML.

[4] Jacques David, Jean-Noel Richet, Eric Boyer, Nikos Anastopoulos, Guillaume Collet, Guillaume Colin de Verdière, Tyra van Olmen, Hilde Ouvrard, Cédric Cocquebert, Bruno Frogé, Alexandros Haritatos, Konstantinos Nikas, Theodoros Gkountouvas, Nikela Papadopoulou, and Evangelia Athanasaki, *Best Practice Guide – Curie*, produced by PRACE-3IP, December 2013. 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.

[5] Maciej Cytowski, *Best Practice Guide – IBM Power 775*, produced by PRACE-3IP, December 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-IBM-Power-775.pdf. http://www.prace-ri.eu/Best-Practice-Guide-IBM-Power-775-HTML.

[6] Andrew Turner, Xu Guo, Lilit Axner, and Mark Filipiak, *Best Practice Guide – Cray XE/XC*, produced by PRACE-3IP, December 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/Best-Practice-Guide-Cray-XE-XC.pdf, http://www.prace-ri.eu/Best-Practice-Guide-Cray-XE-XC-HTML.

[7] 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.

[8] Vegard Eide, Walter Lioen, Maciej Szpindler, and Volker Weinberg, *Petascaling and Optimisation for Tier-1 Architectures*, PRACE-2IP D7.3, May 2013. Available at the PRACE RI web site [1] as: http://www.prace-ri.eu/IMG/pdf/d7.3.pdf.

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

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

List of Acronyms and Abbreviations

AISBL	Association sans but lucrative (legal form of the PRACE RI)
BCS	Bull Coherent Switch
CEA	Commissariat à l'énergie atomique et aux énergies alternatives (France)
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
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)
DARPA	Defense Advanced Research Projects Agency
DoW	Description of Work
DP	Double Precision
EC	European Commission
ECMWF	European Centre for Medium-Range Weather Forecasts (United Kingdom)
EPCC	Edinburgh Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
FZJ	Forschungszentrum Jülich (Germany)
GB	Giga (= 2^{30} ~ 10^{9}) Bytes (= 8 bits), also GByte
GB/s	Giga (= 10^{9}) Bytes (= 8 bits) per second, also GByte/s
GENCI	Grand Equipement National de Calcul Intensif (France)
GNU	GNU's not Unix, a free OS
GPU	Graphic Processing Unit
GRNET	The Greek Research and Technology Network

HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPCS	High Productivity Computing System (a DARPA program)
HTML	HyperText Markup Language
IBM	Formerly known as International Business Machines
ICCS	Institute of Communications and Computer Systems (Greece)
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
KTH	Kungliga Tekniska Högskolan (represented in PRACE by SNIC, Sweden)
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
MB	Management Board
MCM	Multi-Chip Module
MIC	Intel Many Integrated Core Architecture
MPI	Message Passing Interface
NCSA	National Centre for Supercomputing Applications (Bulgaria)
NUMA	Non-Uniform Memory Access or Architecture
OpenMP	Open Multi-Processing
OS	Operating System
PDF	Portable Document Format
POWIEW	HPC Infrastructure for Grand Challenges of Science and Engineering project (Poland)
PP	Preparatory Project
PRACE	Partnership for Advanced Computing in Europe
RI	Research Infrastructure
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SNIC	Swedish National Infrastructure for Computing (Sweden)
ssh	Secure Shell
SURFsara	Dutch national High Performance Computing & e-Science Support Center
svn	Apache Subversion
TB	Technical Board
TFlop/s	Tera (= 10^{12}) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
TGCC	Très Grand Centre de calcul du CEA (France)
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
UEABS	Unified European Applications Benchmark Suite
VASP	The Vienna Ab initio Simulation Package
VSB	Vysoká škola báňská - Technical University of Ostrava (Czech Republic)
WP	Work Package
XML	Extensible Markup Language

Executive Summary

Work Package 7 'Application Enabling and Support' provides applications enabling support for HPC applications codes which are important for European researchers to ensure that these applications can effectively exploit multi-petaflop systems. Applications will be selected either via PRACE Preparatory Access or identified as addressing major socio-economic challenges with the advice of the PRACE Scientific Steering Committee. This applications enabling activity will use the most promising tools, algorithms and standards for optimisation and parallel scaling that have recently been developed through research and experience in PRACE and other projects.

Through the applications-enabling work, WP7 develops specific expertise on most, if not all, of the architectures which make up 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. [7]). PRACE-2IP supplemented 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 (cf. [8]). Finally, PRACE-3IP Task 7.3 called 'Support for European Researchers' supplements these with best practice guides for current and future HPC systems.

Topics for these best practice guides include: optimal porting of applications (e.g., choice of numerical libraries and compiler options); architecture specific optimisation and scaling techniques; optimal system environment (e.g., tuneable system parameters, job placement and optimised system libraries); debugging tools, performance analysis tools and programming environment.

This report describes the process which led to the best practice guides. For the best practice guides itself we refer to the online versions on the PRACE RI web site [1] (cf. [2], [3], [4], [5], [6]).

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, D7.3.1 'Best Practice Guides for New and Emerging 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]).

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 best practice guides for current and future HPC systems.

2.1 Selection of Systems

When PRACE-2IP started in 2012, we selected Intel MIC as future architecture. After Intel officially released the Intel Xeon Phi product we changed the name of this best practice guide accordingly. Selecting a new architecture in 2012 was also obvious: there were a couple of Blue Gene/Q installations planned at PRACE Tier-0 and Tier-1 sites.

Furthermore, we created updates of the PRACE-1IP best practice guides for: Cray (adding Cray XC for e.g. the Archer system at EPCC) and Curie (adding so-called extra large nodes).

Finally, based on the PRACE-1IP generic best practice guide on the IBM Power architecture, we created an architecture specific mini-guide for the IBM Power 775.

2.2 Subtasks

Task 7.3 was led by Walter Lioen (SURFsara); the respective subtask leaders were:

B. Volker Weinberg, LRZ, for Intel Xeon Phi
C. Philippe Wautelet, IDRIS, for Blue Gene/Q
D. Nikos Anastopoulos, GRNET, for Curie
E. Maciej Cytowski, ICM, for IBM Power 775
F. Mark Filipiak, EPCC, for Cray XE/XC

Note: the subtask numbering starts with B. since subtask A. addresses the Unified European Applications Benchmark Suite (UEABS) rather than a best practice guide.

2.3 Technology

We built on the experience obtained during the corresponding PRACE-1IP and PRACE-2IP tasks (cf. [7], [8]). 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. [9], [10]) 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.

2.4 Generic Table of Contents

For PRACE-1IP, all best practice guides were created based on the same generic table of contents. For PRACE-2IP we introduced 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. For PRACE-3IP we used the same approach.

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 – Intel Xeon Phi (cf. [2])

This best practice guide provides information about Intel's MIC architecture and programming models for the Intel Xeon Phi coprocessor in order to enable programmers to achieve good performance of their applications.

The guide covers a wide range of topics from the description of the hardware of the Intel Xeon Phi coprocessor through information about the basic programming models as well as information about porting programs up to tools and strategies how to analyze and improve the performance of applications.

3.2 Best Practice Guide – Blue Gene/Q (cf. [3])

This Best Practice Guide is intended to help users to obtain the best productivity from the two PRACE Tier-0 Blue Gene/Q systems FERMI and JUQUEEN. This guide provides information about these supercomputers in order to enable their users to achieve good performance of their applications. The guide covers a wide range of topics from a detailed description of the hardware to information about the basic production environment: how to login, the accounting procedure, information about porting and submitting jobs, and tools and strategies on how to analyse and improve the performance of applications.

The IBM Blue Gene/Q system JUQUEEN is hosted by Forschungszentrum Jülich (FZJ) in Germany. With a peak performance of 5.9 Petaflop/s, JUQUEEN was the first HPC system in Europe to pass the 5 Petaflop/s barrier (quadrillions of calculations per second).

The IBM Blue Gene/Q system FERMI is hosted by CINECA in Italy. It has 10 racks and a peak performance of 2.1 Petaflop/s.

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

This Best Practice Guide is intended to help users to get the best productivity out of the PRACE French Tier-0 Curie system. Curie is a bullx series (designed by Bull for extreme computing) supercomputer, which is owned by GENCI (Grand Equipement National de Calcul Intensif, French representative in PRACE), and hosted and operated by CEA (Commissariat à l'Energie Atomique) in its TGCC (Très Grand Centre de calcul du CEA) in Bruyères-le-Châtel.

Curie is the second European Tier-0 system, installed in France; it is a Bull system, based on Intel processors. The first phase of the system, a.k.a. the 'fat nodes' (Intel X7560, code-named Nehalem), was completed end 2010; a subsequent 'hybrid' part with Nvidia M2090 GPUs was installed during summer 2011, and the full system, also known as the 'thin nodes' (with more than 80,000 Intel E5-2680 CPUs, code-named Sandy Bridge) became available in October 2011. Late 2012, the 'fat nodes' were converted into 'extra large' nodes, by grouping four of them into a single, 'super-fat' shared-memory node, using the novel Bull Coherent Switch (BCS) architecture.

Within PRACE-3IP, the Curie best practice guide has been updated to reflect the transition from the 'fat nodes' partition to the 'extra large nodes', among other changes. Specifically, we incorporated architectural descriptions and new specifications for the 'extra large' and 'thin' nodes ('System Architecture and Configuration' section). Dependent sections regarding architecture specific issues (e.g. 'Tuning Applications') were updated accordingly. The 'Case Studies' section was largely updated, with new cases and experiments performed on the 'extra large' and 'thin' nodes. Based on the observed behaviour, useful conclusions with respect to the best usage of the machine were drawn.

3.4 Best Practice Guide – IBM Power 775 (cf. [5])

The IBM Power 775 (IH) supercomputing server is a highly scalable system with extreme parallel processing performance and dense, modular packaging. It is based on the IBM Power7 architecture and was designed by IBM within the United States DARPA's HPCS (High Productivity Computing Systems) Program 1. This unique supercomputing environment is currently available in a couple of organisations worldwide, e.g.: ICM, University of Warsaw (Poland), Met Office (United Kingdom), ECMWF (United Kingdom) and Slovak Academy of Sciences (Slovak Republic). The 'Boreasz' system available at ICM,

University of Warsaw is a single cabinet system with 2560 IBM Power7 compute cores and has a peak performance of 78 TFlop/s. The main purpose of the system is research carried out within the POWIEW project which among others includes scientific areas like large-scale cosmological simulations and numerical weather forecasting. 'Boreasz' is also part of the PRACE Tier-1 infrastructure. 'Boreasz' can be very useful for applications based on coarse-grain parallelism where performance depends on high memory bandwidth. A single node has approximately 1 TFlop/s of compute performance, 128 GB of memory and up to 512 GB/s of memory bandwidth. This gives a very good 0.5 Byte/Flop ratio.

'Boreasz' was installed within the 'HPC Infrastructure for Grand Challenges of Science and Engineering' Project and was co-financed by the European Regional Development Fund under the Innovative Economy Operational Programme.

This IBM Power 775 best practice guide was created based on the PRACE-1IP generic best practice guide on the IBM Power architecture. The following list summarizes the main updates and changes:

- Added IBM Power7 processor architecture description,
- Added IBM Power 775 system specific architecture features description (i.e. building blocks, memory architecture, networks and node interconnect, I/O subsystem),
- Added single node level performance descriptions with practical example (VASP performance test),
- Description of 'Boreasz' system access policies,
- Modifications to the section describing the production environment,
- Updates to the sections on programming environment, basic porting and tuning of applications.

3.5 Best Practice Guide – Cray XE/XC (cf. [6])

This best-practice guide is designed to help users get the best productivity out of the PRACE Cray XE and XC (and to a lesser extent XT) systems.

We will focus on providing information that is generally applicable to all Cray XE/XC systems. There are considerable differences in the architecture of Cray XE and XC systems, so these are described separately. Most of the rest of the guide is common to Cray XE and XC systems. Where there are differences between sites (e.g. the login process), we provide the information for each site and links to further information available from the site itself.

Cray XE and XC systems have different node architectures and different interconnection architectures but the programming environment (compilers, batch system) is very similar and the tools (performance analysis, tuning methods, debuggers) are almost identical. Thus, Section 2 'System Architecture and Configuration' has been split into sub-section 2.1 which describes the Cray XE architecture and the new sub-section 2.2 which describes the Cray XC30 architecture. The rest of the new material is integrated with the existing sections of the Guide. The major changes are in:

Section 5 'Batch system/job command language' – The PRACE XC30 systems (Archer and Sisu) use different batch systems from the PRACE XE systems (Hector, Hermit, Lindgren). This section has been updated with batch scripts for the XC30 systems.

Section 7 'Tuning' – Minor changes in several sub-sections for the different architecture of the XC30.

Section 8 'Debugging' – Cray's parallel version of the GNU debugger has recently been updated on both XE and XC systems, so sub-section 8.4 on the GNU debugger has been re-written.