SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2007-2.2.2.1 - Preparatory phase for 'Computer and Data Treatment' research infrastructures in the 2006 ESFRI Roadmap

PRACE

Partnership for Advanced Computing in Europe

Grant Agreement Number: RI-211528

D7.5.1
Technical Requirement for the first Petaflop/s systems(s) in 2009/2010

Final

Version: 1.0
Author(s): Jonathan Evans, BSC
Date: 26.11.2008

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-211528**	
	Project Title: Partnership for Advanced Computing in Europe	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: D7.5.1	
	Deliverable Nature: DOC_TYPE: Report / Other	
	Deliverable Level: PU *	**Contractual Date of Delivery:** 31 / November / 2008
		Actual Date of Delivery: 30 / November / 2008
	EC Project Officer: Maria Ramalho-Natario	

* - 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: Technical Requirement for the first Petaflop/s systems(s) in 2009/2010**	
	ID: D7.5.1	
	Version: 1.0	**Status:** Final
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2003	
	File(s): D7.5.1.doc	
Authorship	**Written by:**	Jonathan Evans, BSC
	Contributors:	Sergi Girona, BSC Jean-Philippe Nominé, GENCI François Robin, GENCI Norbert Meyer, PSNC Michael Stephan, Juelich Giovanni Erbacci, CINECA
	Reviewed by:	Mark Bull, EPCC Dietmar Erwin, FZJ
	Approved by:	Technical Board

Document Status Sheet

Version	Date	Status	Comments
0.1	30/June/20008	Draft	Define objectives and scope in introduction.
0.2	29/September/2008	Draft	Include comments from WP7 weekly meetings, plus feedback from

			Norbet Meyer, PSNC.
0.3	07/October/2008	Draft	Reorganise with comments from WP7 meeting.
0.4	31/October/2008	Draft	Further refinement after face to face meeting, first architecture values entered.
0.5	9/November/2008	Draft	Add BSC cell based values. Include comments for thin and fat nodes architectures sent by PSNC, TASK (PL) and GENCI/CEA (F).
0.6	11/November/2008	Draft	Final version for WP7 review. Complete introduction, conclusions.
0.7	13/November/2008	Final Draft	Draft for internal review after formatting changes.
0.8	23/November/2008	Draft	Internal review comments.
1.0	26/November/2008	Final version	

Document Keywords and Abstract

Keywords:	PRACE, HPC, Research Infrastructure, Petaflop, Technical Requirement
Abstract:	The PRACE project has the overall objective of preparing for the creation of a persistent pan-European HPC service. Work package 7 within PRACE is titled "Petaflop/s Systems for 2009/2010" and is responsible for providing technical information to the Management Board to facilitate selection of the first Petaflop/s production systems in 2009/2010. Task 7.5 in WP7 is responsible for drafting the technical requirements for a Petaflop/s system which will be used in the procurement process. The key objective of this deliverable is to provide a consistent specification of technical requirements to be used in the procurement process. This is achieved by defining a template of requirements which can then be applied to different system architectures by adding specific sizing values. This approach has the benefit of being expanded as new architectures as become available. This document is the first iteration of the task and will be followed by an update in a year.

Table of Contents

List of Tables

References and Applicable Documents

[1] http://www.prace-project.eu
[2] PRACE FP7-Infrastructures-2007-1 Construction of new infrastructures.
[3] Software Environment Management http://modules.sourceforge.net/
[4] GSI-SSH http://www.globus.org/toolkit/docs/4.0/security/openssh/
[5] Iperf network monitoring http://dast.nlanr.net/projects/Iperf/.
[6] *GridFTP* http://www.globus.org/toolkit/docs/4.0/data/gridftp/.
[7] *Inca: user level grid monitoring* http://inca.sdsc.edu/drupal/
[8] Initial recommendation for the selection of prototypes and first estimates of costs of Petaflop/s class systems, PRACE Deliverable D7.1.1, March 2008
[9] Report on systems compliant with user requirements, PRACE Deliverable D7.2, April 2008
[10] Architectural specifications from user requirements, PRACE Deliverable D8.2.2, June 2008
[11] Final report on application requirements, PRACE Deliverable D6.2.2, September 2008
[12] Preliminary assessment of Petaflop/s systems to be installed in 2009/2010, PRACE Deliverable D7.1.2, November 2008
[13] Report of installation requirements and availability at European sites, PRACE Deliverable D7.3, November 2008
[14] Final assessment of Petaflop/s systems to be installed in 2009/2010, PRACE Deliverable D7.1.3, June 2009
[15] Report on available performance analysis and benchmark tools, PRACE Deliverable D6.3.1, November 2008
[16] Report on evaluation criteria and acceptance tests for procurement, PRACE Deliverable D7.6.3, December 2009
[17] Report on deployment of initial software stack to selected sites for distributed systems management, PRACE Deliverable D4.1.3, December 2008

List of Acronyms and Abbreviations

General terms

ESFRI	European Strategy Forum on Research Infrastructures; created roadmap for pan-European Research Infrastructure.
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing.
PRACE	Partnership for Advanced Computing in Europe; Project Acronym.
RFI	Request for Information
SLA	Service Level Agreement between vendor and IT equipment owner, covering level of support, time to fix, acceptable down time etc.
TCO	Total Cost of Ownership. Includes the costs (personnel, power, cooling, maintenance, ...) in addition to the purchase cost of a system.
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.
WP2	PRACE Work Package 2 - Organisational Concepts of the Research Infrastructure
WP4	PRACE Work Package 4 - Distributed systems management
WP5	PRACE Work Package 5 - Deployment of prototype systems
WP6	PRACE Work Package 6 - Software Enabling for Petaflop/s Systems
WP7	PRACE Work Package 7 - Petaflop/s systems for 2009/2010
WP8	PRACE Work Package 8 - Future Petaflop/s computer technologies beyond 2010

Technical terms

Cell BE	Cell Broadband Engine
FPGA	Field Programmable Gate Array
GPU	Graphical Processing Unit
PFlop/s	Petaflop/s or PFlop/s, i.e. 1000 TF = 10^{15} floating point operations per second

Executive Summary

The PRACE project has the overall objective of preparing for the creation of a persistent pan-European HPC service. Work package 7 within PRACE is titled "Petaflop/s Systems for 2009/2010" and is responsible for providing technical information to the Management Board to facilitate selection of the first Petaflop/s production systems in 2009/2010.

This document is the first deliverable from WP7 task 5 *Drafting of technical requirements* and it provides the management board with the first version of technical requirements to be used in the procurement of the first production Petaflop/s systems in 2009/2010. The key objective of this deliverable is to provide a consistent specification of technical requirements and this has been achieved by defining a template of requirements which can then be applied to different system architectures by adding specific sizing values. This approach has the benefit of being expandable as new architectures become available.

This document is the first iteration of the task and has started to fill out architecture-class specific values and relative weightings to rank the importance of selected requirements. This will be followed by an update in a year. As with Task 7.1 *Survey of technologies, architectures and vendors for Petaflop/s systems to be delivered in 2009/2010* this is a process of initial estimation followed by refinement of system requirements as more information becomes available.

This version contains inputs from a wide range of PRACE tasks and provides a template for ongoing use by the PRACE project. It has started to quantify requirements with architecture-class specific values, particularly on system sizing. Refinement of the technical requirements will be driven by the ongoing work in WP7 to obtain improved proposals from vendors for Petaflop/s systems, from knowledge gained by the efforts to address application Peta-scaling in WP6 and the prototype evaluations in WP5.

1 Introduction

This chapter provides an introduction and sets out the objectives and scope of the document.

1.1 Background and Purpose

The PRACE project has the overall objective of preparing for the creation of a persistent pan-European HPC service. Work package 7 within PRACE is titled "Petaflop/s Systems for 2009/2010" and is responsible for providing technical information to the Management Board to facilitate selection of the first Petaflop/s production systems in 2009/2010.

Task 7.5 in WP7 is responsible for drafting the technical requirements for Petaflop/s system(s) which will be used in the procurement process. This task will be iterated twice.

The first iteration provides technical requirements based on information gathered during the first part of the PRACE project. It provides requirements for major HPC architectures which are already available or considered likely to be available by 2009/2010, without giving preference to a particular architecture. The first Petaflop(s) system(s) are likely to be selected from the current set of architectures. The results are contained in this document.

The second iteration will update these requirements as new information becomes available, particularly the PRACE prototype evaluations including benchmark results. It will also take into account the latest evolution in HPC architectures and will be used for the procurement of the second (multi-) Petaflop/s system(s).

1.2 Objectives

The objectives of this document are:

- To support the PRACE Management Board in selecting the first production Petaflop/s systems by providing a consistent specification of technical requirements to be used in the procurement process,
- To provide specific and measurable technical requirements suitable for a procurement process,
- To provide relative weightings to rank the importance of selected requirements,
- To provide a template which can be expanded with new architectures as they become available.

1.3 Scope

These requirements are drafted here to support the procurement process for future Petaflop/s systems. They cover requirements for different machine architectures, targeted at one or more end user application classes.

In this first version of the document not all requirement values are available because there are other activities ongoing such as benchmarking the PRACE prototypes to provide estimates of performance. An update to the requirement values will be provided in D7.5.2. In addition a specific procurement will need to tailor the requirement values and target them at a specific installation site.

The requirements are designed to leave open the way future procurements are organised through the use of ratios such as memory per compute node. They may either start with a

fixed budget and seek to acquire the best performance for the available budget or seek the lowest price for a fixed performance.

The following technical requirements are within the scope of this document:

1. hardware including systems architecture and sizing
2. benchmark objectives
3. software including operating system, management and programming environment
4. operational requirements including installation constraints
5. maintenance and support requirements
6. training and documentation requirements
7. delivery requirements

1.4 Audience

The intended audience will be both technical (for example HPC researchers, operations staff, HPC vendors) and non technical readers.

1.5 Document Structure

The document is split into 6 chapters plus an appendix:

Chapter 1 provides this introduction.

Chapter 2 explains the requirements gathering process and introduces the requirements format.

Chapter 3, the Guidelines and Constraints for vendor response, provides information on requirement interpretation, benchmarking rules and how total cost of ownership is being calculated.

The Requirements Check List in Chapter 4 defines all possible requirements grouped into related sections.

The Architecture-class Specific Requirement Values in Chapter 5 assigns values to the *hardware including systems architecture and sizing* requirements for each available architecture class, such as MPP or thin node.

Chapter 6 summarises the document and indicates the next steps to be taken.

The Appendix provides supporting information relating to these requirements and covers:

- HPC architecture definitions,
- Performance benchmark descriptions,
- Remaining architecture-class specific requirement values.

2 Technical Requirement Methodology

This chapter explains the approach taken in gathering and presenting the technical requirements. It explains the linkages between different PRACE deliverables, both inside and outside WP7 and then explains how requirements are presented for the different hardware architectures under consideration.

2.1 Dependencies and Relationships

The gathering of requirements for the first Petaflop/s systems has, not surprisingly, strong dependencies with the technical work packages in PRACE WP4, 5, 6 and 7.

Within WP7 the three tasks and their associated deliverables, of which this is one, due at the end of November 2008 are closely related.

Task 7.1 *Survey of technologies, architectures and vendors for Petaflop/s systems to be delivered in 2009/2010* is providing deliverable D7.1.2 [12] which updates the market survey in D7.1.1 [8] providing bounds and ranges for potential Petaflop/s systems in 2009/2010 and feeding into hardware requirements including system sizing. In addition an update of the D7.2 [9] mapping of PRACE applications onto suitable architectures is included which impacts the benchmarking requirements. Finally the total cost of ownership (TCO) methodology is defined, along with cost estimates. More information on TCO is provided in this document in Chapter 3 as this impacts the operational requirements.

Task 7.3 *Installation requirements for Petaflop/s systems* is providing deliverable D7.3 [13] which assesses the capability of the PRACE consortium to host and operate Peta-scale computing facilities based on vendor input of installation parameters and site surveys of existing, and planned, PRACE partner HPC sites. The parameters identified have direct input into the operational requirements and installation constraints.

Task 7.5, which is producing this document, is defining a check list of technical requirements and architecture-class specific values to size and define a future Petaflop system more accurately. This will then feed back into the next iteration of Task 7.1. It will also provide a requirements template for future procurements and the Task 7.6 procurement process template and its deliverables.

Dependencies with other PRACE work packages are examined next.

Work Package 4, Task 4.1 *Distributed management at the tier-0 level* has provided the software requirements for the first implementation of systems management tools, part of the systems software requirements. These will be revised over the coming year and this can be reflected in the updated version of this document.

Work Package 5, concerned with the installation and evaluation of the PRACE prototypes will be a key input to the requirements values for different architectures. Technical assessments with synthetic benchmarks are planned in Tasks 5.2 and 5.3 and with application benchmarks in Task 5.4. The prototype assessments have not started in time for this version of the document but will be included in the next version.

Work Package 6, Task 6.2 *Application requirements capture* has provided input to the check list of technical requirements through deliverable D6.2.2 [11], and D6.3.1 [15] has defined the benchmarking requirements. Further information is found in Annex 7.2.

Work Package 8, Task 8.2 *Multi-Petaflop/s technology* has influenced the check list of technical requirements through deliverable D8.2.2 [10].

Additional information is included from available vendors' system specifications, white papers, and from direct requests sent to vendors mentioned in deliverable D7.1.2 [12].

2.2 Presentation of Requirements

The approach taken is to define a check list containing all possible technical requirements which may apply to a procurement process. As the procurement will be targeting specific architectures which may be optimised for certain classes of applications, the requirement values are listed on a per architecture-class basis in separate chapters. This approach has the advantage of creating a general list of requirements which can be the basis of ongoing procurements for the PRACE HPC service and may be modified as new architectures become available.

Each check list requirement in Chapter 4 includes the following information:

a. A requirement category, which is one of three types,
    1. R (required) is a fixed requirement that a vendor must meet for the Petaflop/s systems,
    2. D (desirable) is a feature which a vendor does not have to meet, but would be advantageous and may be used to differentiate similar vendor bids,
    3. Q (question) is a question to the vendor, where information is needed to evaluate offers,
b. A unique number to allow unambiguous referencing in this and other documents, classified into similar requirement groups, n.m where m provides a unique number within each group and n is,
    1. hardware and system sizing,
    2. benchmarks,
    3. software,
    4. operational,
    5. maintenance and support,
    6. training and documentation,
    7. delivery requirements,
c. A descriptive title,
d. Optional notes to provide a fuller description of the requirement and to help remove ambiguity,
e. A flag indicating if the requirement is scalable with machine size (yes or no).

Some system sizing values, such as memory, are specified in terms of value per calculation node and/or total value for the system. As the number of nodes increases (or decreases) the total value will change and so is scalable with number of nodes. These requirements are flagged as being scalable and are needed to leave open the way future procurements are organised. They may either start with a fixed budget and seek to acquire the best performance for the available money or seek the lowest price for a fixed performance.

Each architecture-class specific requirement value in Chapter 5 includes the following information:

a. A value entry of,

1. "required" to indicate the requirement is valid but no quantitative value is appropriate (for example a statement such as " a capability system is required"),
2. a fixed value indicating no uncertainty, e.g. 5MB,
3. a range providing upper and lower limits, e.g. 5MB-10MB,
4. an upper limit, with no lower bound, e.g. <10ms,
5. a lower limit, with no upper bound, e.g. >5MB,
6. a value of "not available", if information is not available,
7. a value of "not applicable", if the requirement is not applicable to this architecture,
8. a value of "vendor response", where a vendor question is applicable,
9. a blank value indicates no decision has been made about this requirement.

b. A relative weighting (relative within requirement sub-groups, e.g. Hardware - CPU) of,

1. blank - no preference in case of better values,
2. "low" - low priority given to better values for this architecture,
3. "medium" - medium priority given to better values for this architecture,
4. "high" - high priority given to better values for this architecture.

The relative weightings will allow the tender process to compare and contrast additional vendor offerings depending on priorities such as more memory, more storage space, improved SLA (support) or longer warranty. This document does not propose how these weightings will be used in a response evaluation and this is left to task 7.6 *Procurement process template*.

As this is the first iteration of the technical requirements not all requirements may have the final values assigned. The prototype assessments, forming the WP5 deliverables, are a key input into these requirements and no results are available for the first iteration of this document. For this reason only the R (required) rated *hardware requirements including system sizing* are included in the main body of the document as these are the key parameters to quantify the minimum acceptable configuration for a Petaflop/s installation. The remaining requirement values, which are listed in annex 7.3, are less complete and will be expanded in the next version of this document.

3 Guidelines and Constraints for Vendor Response

A procurement process needs to set out ground rules for vendor responses. These rules are intended to help remove ambiguity in the response to requirements and so improve the quality of answers received. This allows responses to be more meaningfully compared and provides transparency in the selection process. The procurement process deliverable D7.6.3 [16] may include this information as part of the vendor selection criteria.

3.1 Requirements Interpretation

Requirement values, unless otherwise stated, are minimum values to be met and the vendor can offer better values. Where a value is identified as global and scalable the actual requirement may depend on the solution offered, typically the number of computing nodes.

Desirable requirements are so categorised to give vendors the option of meeting them or not.

The weightings will allow the tender process to compare and contrast additional vendor offerings depending on priorities such as more memory, more storage space, improved SLA (support) or longer warranty.

Desirable requirements along with their relative weightings will be used in the evaluation process but this document does not propose a response evaluation scheme and this is left to Task 7.6 *Procurement process template*.

3.2 Rules for running benchmarks

Ideally the test machine configuration should be equivalent to the proposed final system but as this may not be possible when procuring leadership class systems the vendor may run benchmarks on the closest existing system and commit to performances on the real system. The vendor should document the discrepancies between systems and explain how they have extrapolated the results.

The test machine software stack should represent a production system. All system services which are running during the benchmarking must be listed.

The test machine size, in terms of processing units, should be at least n% (*value to be defined during procurement*) of the proposed final system.

The test machine network interfaces should be the same as the proposed final system.

The disk space size should be at least m% (*value to be defined during procurement*) of the proposed final system.

The kernel configuration should remain the same during all benchmark runs.

The following benchmark optimisations are allowed. Separate benchmark runs may be made with one of a or b mandatory and the remaining benchmark runs optional:

a. no modifications in the code and the same compiler and compiler options for all benchmarks,

b. no modifications in the code (except for library changes) and:

    1. dedicated compilers provided by the hardware vendor,

    2. benchmark specific compiler optimisations, with flags generally available to HPC community,

3. code changes to call optimised libraries performing the same algorithms are allowed as long as the libraries used are reported along with version and library provider,
4. library calling sequences and parameter types must be unchanged.

c. modifications to the code, if the results without modification are given and:
   1. code changes are not allowed to alter the algorithm used,
   2. calculations should be run in the same precision as the unmodified version,
   3. code changes should be achievable by the average user,
   4. all changes must be supplied with the results,
   5. knowledge of the output of the benchmark can not be used to skip parts of the code,
   6. optimisations should not require super user privileges.

Where multiple runs are made please provide all the results so that the variance across runs can be determined.

Report the power consumption during the benchmark tests, both total watts and flops / watt if a benchmark gives a flops result. Power consumption should include processing units, I/O units, management servers and network switches.

3.3 Total Cost of Ownership Calculation

The Total Cost of Ownership (TCO) for the system is an important figure which will need to be derived during a procurement process and matched to the available budget. The TCO methodology along with indicative costs is defined in D7.1.2 [12]. Some of the elements which make up the TCO are related to the technical requirements in this document and the list from D7.1.2 [12] is reproduced here to indicate where the TCO calculation has input from these requirements (the relevant requirement section is added in italics):

- Supercomputer including installation – *hardware requirements including system sizing*,
- Related IT equipments needed for the operation of the supercomputer: storage system (including back-up), internal computer centre networks (including connection point for external network connection), including installation – *operational requirements including installation constraints*,
- Maintenance of the supercomputer and related IT equipments and software licences, including vendor support for hardware and software – *maintenance and support requirements*,
- Building (floor space for the IT equipments, the technical facilities, offices for computer centre team) – *installation constraints,*
- Technical facilities including cooling, power supply (transformers, UPS, distribution, …) - *operational requirements* ,
- Maintenance of the building and of the technical facilities – *no dependency,*
- Electricity charge including the cost of the power line and main substation if needed – *operational requirements and installation constraints,*
- The staff including management, computer centre operation, application support, building and technical infrastructure support. Application support may actually be

considered as including development and job submission tools support, code profiling, optimization, porting and scaling. – *no dependency,*

- Training (users and administrators) – *documentation and training requirements,*
- Some (minor) evolutions and upgrades necessary within the 5 years of operation (most likely within the 2 or 3 first years) – *hardware requirements including system sizing.*

4 Requirements Checklist

This chapter contains a check list of the technical requirements, some of which apply generally and some of which are architecture specific. Note that not all requirements will be applicable to a specific procurement. These tables together with specific values for different architectures (Chapter 5) will be the basis for the information to be sent to vendors in the calls for tender.

The numbering system is as follows: Rn.m for a compulsory requirement which must be met, Dn.m for a desirable requirement which is optional but will allow similar proposals to be compared and Qn.m for a request for information from the vendor. See the Section 2.2 for further explanation.

4.1 Hardware Requirements including system sizing

The requirements defined in *Table 1: Hardware requirements including system sizing*, enable the four main parts of the HPC system; CPU, memory, network and disk/IO to be sized. Where requirement values are based on processor cores the term processing unit is used, meaning a core plus accelerators.

The network requirements are presented slightly differently. As a particular architecture in Chapter 5 may have multiple networks for MPI or require a dedicated NFS network for mounting operating systems, the requirements are presented generally for any network. The architecture-class specific requirements will need to be repeated for each network, using the numbering convention n.m.p (where p represents a network id).

Ref	Title	Notes	Scalable
CPU			
R1.1	A capability system is required [2].	This is defined as a system with the ability to run a single MPI application on all calculation nodes requiring fast inter process communication.	No
R1.2	Calculation core bit size (for example 32 or 64).	The calculation processors must support floating point calculations using IEEE-754 representation. Some architectures are designed for 32 bit processors such as IBM Blue Gene.	No

Ref	Title	Notes	Scalable
R1.3	Minimum Peak FLOPS [2] in Petaflops.	Calculated as the theoretical maximum double precision floating point operations per second for the system. A sum of the peak FLOPS for all processing units.	Yes
Q1.4	Vendor to provide information on an upgrade path for processing units after 2 to 3 years of production use.	Vendor to specify what upgrades are available or planned for release in this timeframe.	No
Q1.5	Vendor to provide information on options for cache levels; location, size, associativity.	This is defined as a question for vendors as their are usually very limited user choices for cache sizing.	No
Q1.6	Vendor to specify the number of Simultaneous Multi Threading (SMT) threads supported and the restrictions on how instructions from different threads are scheduled together.	SMT provides CPU efficiency improvements and can help to hide memory latency.	No
Memory			
R1.7	Minimum total memory in gigabytes per calculation node.	Defined as the sum of the memory for each processing unit.	Yes
R1.8	Minimum memory in gigabytes per processing unit.	Needed to set a lower limit on the memory available to an application process.	No
R1.9	Mechanism for error detection and correction in main memory.	This is defined as the use of error correcting codes in memory controllers to automatically reconstruct memory contents using parity bits. As the total amount of memory used by a capability job reaches higher levels failures may become significant.	No
Q1.10	Vendor to provide peak memory bandwidth and minimum latency.	For NUMA architectures specify these values for varying memory hops, local, 1st up to nth.	No
Q1.11	Vendor to provide available memory configurations, including free slots for upgrade and memory unit sizes.	This includes information for memory upgrades during the system lifetime.	No

Ref	Title	Notes	Scalable
Network These will be repeated in the architecture Chapter 5 for each required internal network. The expected networks are; one or more MPI networks, one or more I/O networks, a TCP/IP network for operating system mounting, a management network. The numbering allows a unique number to be assigned to each network requirement, where p will take values 1,2,3,4... depending on the number of networks.			
R1.12.p	All calculation nodes are required to be connected to the network.		No
Q1.13.p	Vendor to specify network technology, for example InfiniBand, MyriNet, Ethernet.		No
Q1.14.p	Vendor to specify network topology.	This defines the physical and virtual connectivity of the computation and other nodes, such as storage servers.	No
R1.15.p	Minimum network bisection bandwidth in gigabits per second.	The bisection bandwidth is the bandwidth across the links which divide the network into 2 equal halves multiplied by the number of nodes. Care should be taken specifying R1.16 and R1.17 not to over-constrain vendor solutions.	Yes
R1.16.p	Minimum network bisection bandwidth in bits per second/ peak performance in flops for calculation node.	Network performance for applications needs to scale with increased calculation rates, hence this ratio. Care should be taken specifying R1.16 and R1.17 not to over-constrain vendor solutions.	No
R1.17.p	Maximum network latency in microseconds between two most distant nodes.		No
Q1.18.p	Vendor to provide information on an upgrade path for network components.		No
External access network			

Ref	Title	Notes	Scalable
R1.19	There should be fast external access to the global file system for user file transfer.		No
R1.20	There should be fast external access to the global file system for data backup and recovery.		No
R1.21	Minimum external network bandwidth in gigabits per second.		No
I/O and Storage – Global			
R1.22	Minimum global disk storage in Petabytes.	This is defined as fast hard drive disk space for use in running applications, as opposed to slower disks or tape for archive storage. This is global to a single tier 0 site.	Yes
R1.23	Minimum number of file system partitions to be supported by the global storage system.		No
Q1.24	Vendor to specify the types of file system supported on the global storage system.		No
R1.25	Minimum size for a global file system partition in Petabytes.		Yes
Q1.26	Maximum global file system size supported in Petabytes.		No
Q1.27	Maximum individual file size supported by the global file system in Terabytes.		No
Q1.28	Maximum number of concurrent clients (computing nodes) which may connect to the global file system.		No
R1.29	Peak bandwidth for global file system reads and writes (updates and creates) in bits per sec / peak performance in flops.		Yes
Q1.30	Minimum file creation and search times in microseconds.		No
Q1.31	A long term upgrade path for global storage is required.	For example, after adding more processing nodes or to increase storage. Vendor to specify what upgrades will be available.	No

Ref	Title	Notes	Scalable
D1.32	Hierarchical/tiered storage management is required.	This provides storage space for running applications using fast disk, with archive date moved to slower disk or tape as defined by site specific policies.	No
D1.33	Percentage of storage space (related to online fast disk) for each level of a hierarchical/tiered storage system.		No
I/O and Storage – Local			
D1.34	Minimum size in Gigabytes of fast local storage for calculation nodes.	This storage is local to a calculation node and not necessarily shared across nodes.	No
D1.35	User space on local storage specified as multiple of calculation node memory.		No
R1.36	Swap space specified as multiple of calculation node memory.	This may be located on local storage or on global storage.	No
D1.37	Minimum bandwidth of local storage reads and writes in Megabits per second.		No
D1.38	Maximum latency of local storage reads and writes in microseconds.		No

Table 1: Hardware requirements including system sizing

4.2 Benchmark Requirements

The application benchmarks are being prepared by PRACE WP6 and further information about these benchmarks is provided in Annex 7.3, with the vendor rules for running benchmarks explained in Section 3.2. The requirements listed here in *Table 2: Benchmark requirements* are a request to run each benchmark so that each architecture-class specific section can choose relevant benchmarks. Latest information on which application benchmarks are available for which architectures is included in Annex 7.2. Also included is the synthetic benchmarks suite which is also being prepared by PRACE WP6.

Ref	Title	Notes	Scalable
R2.1	Run NAMD benchmark and report results.		No
R2.2	Run CPMD benchmark and report results.		No
R2.3	Run VASP benchmark and report results.		No
R2.4	Run GADGET benchmark and report results.		No
R2.5	Run Code_Saturne benchmark and report results.		No
R2.6	Run TORB benchmark and report results.		No
R2.7	Run NEMO benchmark and report results.		No
R2.8	Run ECHAM5 benchmark and report results.		No
R2.9	Run CP2K benchmark and report results.		No
R2.10	Run GROMACS benchmark and report results.		No
R2.11	Run N3D benchmark and report results.		No
R2.12	Run HELIUM benchmark and report results.		No
R2.13	Run GPAW benchmark and report results.		No
R2.14	Run ALYA benchmark and report results.		No
R2.15	Run PEPC benchmark and report results.		No

Ref	Title	Notes	Scalable
R2.16	Run QCD benchmark and report results.		No
R2.17	Run AVBP benchmark and report results.		No
R2.18	Run TRIPOLI_4 benchmark and report results.		No
R2.19	Run SIESTA benchmark and report results.		No
R2.20	Run BSIT benchmark and report results.		No
R2.21	Run Synthetic Benchmarks suite and report results.	Includes HPCC Linpack.	No

Table 2: Benchmark requirements

4.3 Software Requirements including system software and programming environment

The requirements listed in *Table 3: Software requirements* concentrate on the functionality required from the software for the system. Performance requirements such as timings are listed under the Operational Requirements section. Distributed systems management software requirements are related to information which is being assembled for D4.1.3 [17].

Ref	Title	Notes	Scalable
Operating System			
R3.1	The operating system should be UNIX like.	It should be compatible with the X/Open Standard POSIX 1003 (ISO/IEC 9945).	No
Q3.2	Vendor to provide details of supported operating systems.		No
R3.3	Nodes are able to be booted from multiple system images.		No
R3.4	The maximum node CPU usage (%) by the operating system, with no applications running.		No
R3.5	The maximum node memory usage (%) by the operating system, with no applications running.		No
D3.6	Mechanisms to prevent uncoordinated interruption of user processes by O/S tasks to reduce operating system jitter.		No
R3.7	Support for large page sizes	Page sizes much larger than 4 Kbyte.	No
R3.8	Ability to dynamically (on a per process or job basis) alter the number of large pages available on a node, depending on user demands.	No reboot of the system is required to accomplish this task.	No
R3.9	The file systems to be supported on local disk storage.		No
R3.10	Minimum number of open files for each process.	Used to ensure the maximum file descriptor table size can cater with Petaflop scale applications	No
Programming environment - languages, programming models, compilers			

Ref	Title	Notes	Scalable
R3.11	The programming environment should support C, C++ and Fortran.		No
D3.12	The programming environment should support Java.		No
R3.13	Interoperability between Fortran, C, and C++.		No
R3.14	C compiler for compute nodes must at least support a full implementation of the standard ANSI/ISO 9899-1990 („C90“)		No
D3.15	C compiler for compute nodes, supporting ANSI C99 standard.		No
R3.16	C++ Compiler for compute nodes must at least support a full implementation of the standard ANSI/ISO 14882-1998 („C++98“), including the C++ standard library		No
R3.17	Fortran compiler for compute nodes supports a full implementation of the language specifications of Fortran 95 (ANSI X3J3/96-007)		No
D3.18	Fortran compiler for compute nodes supports a full implementation of the language specification of the Fortran 2003 standard.		No
R3.19	A recent version of the GCC is available which supports the system hardware.	Note that some accelerators are delivered with dedicated compilers and are not supported by the GCC directly so this may be limited to CPUs and not accelerators.	No
Q3.20	Vendor to provide details of vendor optimised compilers and libraries including operating system support.		No
R3.21	Compilers support 32 and 64-bit mode.		No
D3.22	Support for PGAS programming model with support for emerging compilers, tools, and libraries.	For example Co-array Fortran, UPC (Unified Parallel C) and vendor-specific constructs for global data addressing such as SHMEM.	No

Ref	Title	Notes	Scalable
R3.23	Ability to run different versions of compilers, linkers, libraries, applications, etc. alternatively or additionally to the standard programming environment		No
Q3.24	Vendor to provide details of supported programming interfaces for any accelerator devices.		No
Programming environment - MPI and OpenMP			
R3.25	MPI library for compute nodes, fully supporting MPI Version 1.2		No
D3.26	MPI library for compute nodes, supporting MPI Version 2.1.	With the exception of dynamic process spawning routines.	No
Q3.27	Vendor to provide details of MPI implementations (version 1 and version 2) supported and level of support (for example which parts of MPI 2 specification supported).	Include details on any MPI optimisations.	No
D3.28	Where compute nodes support threading, the MPI library must implement the highest level of thread safety (MPI_THREAD_MULTIPLE).		No
R3.29	Shared memory applications are able to use POSIX threads.	The implementation should be compatible with the X/Open Standard POSIX 1003 (ISO/IEC 9945).	No
R3.30	If compute nodes have hardware shared memory the Fortran, C and C++ compilers must fully support the OpenMP Version 2.5 standard.		No
Programming environment – tools			
R3.31	The following debugging tools should be usable on the system.	Details to be defined in architecture specific sections.	No
R3.32	The following profiling and optimisation tools should be useable on the system.	Details to be defined in architecture specific sections.	No
R3.33	A parallel debugger for the compute nodes.		No

Ref	Title	Notes	Scalable
R3.34	Sequential performance analysis tool for the compute nodes.		No
R3.35	Parallel performance analysis tool for the compute nodes with profiling capability.		No
D3.36	Parallel performance analysis tools for the compute nodes with MPI tracing and hardware counter capability.		No
Programming environment – libraries			
R3.37	BLAS and PBLAS libraries optimised for the compute nodes		No
R3.38	FFTW version 2&3 libraries optimised for the compute nodes		No
R3.39	LAPACK library for the compute nodes		No
R3.40	ScaLAPACK library for the compute nodes		No
D3.41	LAPACK Version 3.1 library for compute nodes		No
Programming environment - front end/login nodes			
R3.42	A version of the Java SDK ≥ 5.0 is available for the front end nodes.	To run developer tools, software editors, also management tools.	No
R3.43	Front end nodes have access to global file system.		No
D3.44	C, C++, Fortran cross compilers are available.	Cross compilers are applicable for use by developers outside the system or in the case of heterogeneous systems to build code for different target hardware from a single front end node.	No
D3.45	Perl for front end nodes.		No
D3.46	Python for front end nodes.		No
D3.47	Emacs editor for login/front-end nodes.		No
D3.48	Revision control system for login/front end nodes (e.g. CVS, Subversion).		No

Ref	Title	Notes	Scalable
Scheduling, Batch and Resource Management Software			
R3.49	Ability to efficiently manage different workloads of the system.	For example, dynamically grant resources to system tasks depending on a classification scheme decided by the system administrator.	No
R3.50	The resource management software allows global control of nodes, processing units, interconnection networks.		No
R3.51	The maximum time in seconds to start a job on all calculation nodes.		Yes
R3.52	The batch and resource management software is compatible with compatible with all supported parallel programming models.		No
D3.53	The scheduling software supports backfill scheduling.	To avoid underutilisation of reserved nodes.	No
D3.54	The scheduling software provides a means for co-scheduling and/or resource reservation.		No
Q3.55	Vendor to list supported scheduling, batch and resource management software.		No
D3.56	It is possible to drain any compute node after the end of running jobs to block its re-use.		No
Administration Software			
D3.57	The monitoring software should be able to suspend unused nodes and power them up only when required.	As long as the operation of the rest of the system is unaffected.	No
R3.58	Tools to change system parameters without system interruption.		No
R3.59	Software monitors to measure important system characteristics such as; I/O behaviour, disk access behaviour, CPU load, memory load, paging rate and so on.	An easy-to-interpret output is required.	No

Ref	Title	Notes	Scalable
R3.60	Facilities for the on-line detection of hardware errors.	For example faulty memory modules, processors, fans, network links, switches.	No
D3.61	Tools to extract CPU performance information like number of floating point operations, number of integer operations, main memory and cache references, etc. per second from hardware performance counters	Information should be available to the system administrators on a per CPU-core basis without any impact on user codes and without the necessity of any specific changes to those codes.	No
Distributed Systems Management Software			
R3.62	The system can run the modules software (TCL implementation) [3].	This is used to create a standard PRACE environment for grid users.	No
R3.63	The system can run the GSI-SSH software [4].	This software is used by grid users for authentication and encryption.	No
R3.64	System supports job scheduling and monitoring software for grid users.	Initially this will use the local system software.	No
R3.65	System supports resource accounting software for grid users.	Initially this will use the local system software.	No
R3.66	The system can run the lperf network monitoring tool [5].	This is used to measure grid network performance.	No
R3.67	The system can run the GridFTP data transfer tool [6].	Allows users to exchange information between PRACE sites.	No
R3.68	System supports test monitoring and local monitoring software for grid users.	Initially this will use the local system software.	No
R3.69	The system supports the grid monitoring tool inca [7].	This is used for a number of tasks, initially to allow external users monitor the software stack available to them.	No

Table 3: Software requirements

4.4 Operational Requirements including installation constraints

The requirements listed in *Table 4: Operational requirements* relate to the operational use of the system, including reliability and operational management, as well as installation constraints which are linked to information in D7.3 [13].

Ref	Title	Notes	Scalable
Users and jobs			
R4.1	Number of concurrent users logged in to the system.		No
R4.2	Number of concurrent batch jobs to be supported by the system.		No
Reliability and Availability			
R4.3	Minimum mean time between interrupts (MTBI) in hours.	This is defined as the mean time between job interrupts, where a job does not complete because of component failure.	No
R4.4	Minimum mean time between failure for system components in hours.	Either one value for mean time across all components or separate times for each component type (calculation node, network and disk components) can be specified.	No
R4.5	Seamless degradation of the system in case of a failure of a compute or I/O node.		No
Q4.6	Vendor to demonstrate how hardware redundancy provides continuity of service to users for different types of component failure.		No
R4.7	Calculation nodes provide temperature monitoring and automated shut down if configurable limits are exceeded.		No
R4.8	Calculation nodes can be stopped and started without interrupting applications running on other calculation nodes.		No
R4.9	Type of hardware redundancy required for global disk storage.	This includes the disk storage hardware topology and disk controllers RAID level.	No

Ref	Title	Notes	Scalable
R4.10	File systems support journaling of meta- and/or user data or some equivalent mechanism		No
R4.11	Ratio of backup storage space to global storage system, including all levels of any hierarchical storage.	Relates to number of full backups which need to be retained.	No
R4.12	Backup of data can be achieved in parallel with and without interrupting running jobs.		No
R4.13	Maximum time in hours to run full global file system backup.		Yes
R4.14	Recovery of data from backup can be achieved without interrupting running jobs.		No
R4.15	Maximum time to recover full global file system in hours.		Yes
R4.16	Means to ensure end-to-end data consistency for global file system.		No
R4.17	Facilities for the on-line detection and correction of media errors in global file system.		No
R4.18	Maximum time to check file system after media errors in global file system in hours.		Yes
R4.19	Hardware redundancy of local disk storage.		No
D4.20	Automated check pointing at application level and restart capabilities.		No
Manageability			
R4.21	All calculation nodes can be started and stopped from a single administration computer		No
R4.22	Software can be installed on all calculation nodes from a single administration computer		No

Ref	Title	Notes	Scalable
R4.23	Time required for an operating system upgrade or complete installation of a new operating system on all nodes in hours.		Yes
R4.24	Maximum time to add a file to the software stack across all nodes in minutes.		Yes
R4.25	Ability to verify correct installation of software changes.		No
R4.26	Maximum time for controlled shut down of system with file system contents preserved in minutes.		Yes
R4.27	Maximum start-up time per node after a controlled shut down in minutes.		No
R4.28	Maximum start-up time for whole system after a controlled shut down in minutes.		Yes
R4.29	Maximum start-up time per node after a forced shut down in minutes.		No
R4.30	Maximum start-up time for whole system after a forced shut down in minutes.		Yes
R4.31	Monitoring the interconnect of the system is possible.	For example, number and size of packets, amount of data sent or received.	No
R4.32	Error tracking and reporting mechanism in cases of operating system errors. (e.g., system dumps) available and officially supported by the vendor		No
D4.33	Modifications of all parts of the operating system (except the kernel) possible at any time, without interrupting the operation, immediately taking effect, and possibility to undo single modifications separately		No

Ref	Title	Notes	Scalable
R4.34	Possibility to carry out all tasks of user administration via scriptable interfaces including setting of new passwords without interactive editing of files by the system administrator		No
R4.35	The ability to monitor power consumption of the system components.		No
System Security			
R4.36	Access control to files on the global file system is managed with Unix groups.		No
R4.37	Minimum number of Unix groups required to be supported.		No
D4.38	Privileges isolation between nodes is required.	Gaining administrative privileges on one node does not automatically mean these privileges are available on other nodes.	No
Power			
R4.39	The maximum power consumed by a calculation node whilst idling in KW.		No
R4.40	The maximum power consumed by a calculation node whilst under full load in KW.		No
R4.41	The maximum power consumed by the whole system in MW.	This relates to system components only, i.e. not cooling power.	Yes
Installation Constraints			
Q4.42	Vendor to specify the system cooling system type.		No
Q4.43	Vendor to specify the system cooling capacity required.		Yes
Q4.44	Vendor to specify the cooling system parameters. Direction of airflow, airflow rate, inlet and outlet temperatures.		No
R4.45	Maximum floor space available to system in m^2.		No

Ref	Title	Notes	Scalable
R4.46	Access constraints for physical access during installation.		No
R4.47	Maximum floor loading available to system in kg/m^2.		No
Q4.48	Vendor to supply any restrictions on cabling distances between components.		No
Q4.49	Vendor to specify electricity supply requirements (current, voltage, phase).		No
R4.50	Maximum heat dissipation per rack in KW.		No

Table 4: Operational requirements

4.5 Maintenance and Support Requirements

Any large HPC cluster is expected to require frequent maintenance and this is particularly important for Petaflop scale machines. The requirements listed in *Table 5: Maintenance and support requirements* address the needs of a high availability Petaflop machine with respect to vendor service level agreements (SLA).

Ref	Title	Notes	Scalable
R5.1	Specify warranty duration in years.		No
R5.2	Percentage availability of calculation nodes during working hours.	Working hours are 8am-6pm, Monday - Friday excluding public holidays.	No
R5.3	Percentage availability of calculation nodes during non working hours.		No
R5.4	Response time to reported problem for redundant hardware (compute nodes) in hours.		No
R5.5	Response time to reported problem for non redundant hardware (network, administration nodes) in hours.		No
R5.6	Support cover hours for redundant hardware (compute nodes).		No
R5.7	Support cover hours for non redundant hardware (network, administration nodes).		No
R5.8	Target repair time for redundant hardware (compute nodes) in hours.		No
R5.9	Target repair time for non redundant hardware (network, administration nodes) in hours.		No
R5.10	Maximum time between response to the problem report and full availability of the entire system in hours.	This is added because repair time of the component may not include re-cabling, software reconfiguration etc. which is required for the system to fully enter production mode.	No

Ref	Title	Notes	Scalable
D5.11	Owner can install third party extensions cards and attach third party network devices without warranty void as long as they are compatible and properly installed. Owner can physically relocate the system without warranty lost as long as it is done according to vendor's procedures included within the documentation.		No
D5.12	Vendor will provide the owner (within the warranty) with access to the problem reporting system where each problem report is identified by a distinct problem id.		No
R5.13	Vendor will provide the owner (within the warranty) with all software upgrades (including operating system and firmware) available for the delivered software and hardware.		No
D5.14	Vendor to provide proactive support by notifying recommended firmware and software updates as they become available.		No

Table 5: Maintenance and support requirements

4.6 Documentation and Training Requirements

Adequate documentation and training are necessary for successful installation and configuration of a large HPC system and these requirements are listed in *Table 6: Documentation and training requirements*.

Ref	Title	Notes	Scalable
Documentation			
R6.1	System is provided with an electronic and optional paper version of a complete list of components.	List includes physical location, model name, serial number and network settings (if applicable).	No
R6.2	Documentation includes system general description, graphical diagrams of all interconnects (including communication and management network) and configuration values (including network settings).		No
R6.3	Documentation describes procedures required for complete disaster recovery.	Physically relocating the system, reconfiguring all components from scratch and reinstalling all provided software.	No
R6.4	Electronic and optional paper documentation for all supplied software.		No
Training			
R6.5	Advanced training at the Owner's location will be provided by the vendor and cover: system and technology introduction, management tools and procedures, system monitoring and optimization, security, applications and programming. The actual system will be used in the training during the labs.		No

Table 6: Documentation and training requirements

4.7 Delivery Requirements

The requirements in *Table 7: Delivery requirements* relate to the procedures for delivery of the system.

Ref	Title	Notes	Scalable
R7.1	Vendor will provide the owner with complete delivery schedule including starting and ending times of every delivery phases.		No
R7.2	Vendor will adjust to the owner's facility regulations.	Includes safety laws, waste disposal procedures, floor space necessary for packages and system assembly.	No

Table 7: Delivery requirements

5 Architecture-class Specific Requirement Values

This Chapter contains entries relating previously defined requirements to values for representative machines in each architecture class. For this first version of the requirements only the R (required) rated *hardware requirements including system sizing* are listed here as these are the key parameters to quantify the minimum acceptable configuration for a Petaflop/s installation. The remaining requirement values, which are listed in annex 7.3, are less complete and will be revised in the next version of this document.

As discussed in section 2.2 the values are one of either:

1. "required" to indicate the requirement is valid but no value is appropriate,
2. a numeric value or range,
3. a value of "not available", if information is not available,
4. a value of "not applicable", if the requirement is not applicable to this architecture,
5. a value of "vendor response", where a vendor question is applicable.

A blank value indicates no decision has been made about this requirement.

The relative weightings are one of either:

1. blank - no preference for improved values,
2. "low" - low priority given to improved values for this architecture,
3. "medium" - medium priority given to improved values for this architecture,
4. "high" - high priority given to improved values for this architecture.

The architectures considered here are MPP, Thin Node, Fat Node and a cluster of accelerated processors.

Task 7.1 issued a request for information (RFI) to vendors of potential Petaflop systems asking for proposals for a Petaflop system in the 2009/2010 time frame. With no vector systems proposed by vendors it appears that there will be no solutions available and so for this iteration of the document no requirement values have been entered for vector systems. It is hoped that in the next iteration of the document some values can be included here. More information on the background is available in the market survey update in D7.1.2 [12].

As for vector systems, the market survey RFI did not produce any proposals which included accelerated systems (defined in the architecture taxonomy as two different types of compute unit) and there are no suggested values included here. However a second class of accelerated system

was proposed, one a cluster of GPU processors and the other a cluster of Cell processors. In all other respects these could be classified as a cluster of thin nodes but owing to the extra porting work needed for applications they have been separated out.

5.1 Hardware Requirements including system sizing

This table only includes the R rated (required) entries. See annex 7.3 for the remaining entries.

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
CPU					
R1.1	required	required	required	required	high
R1.2	32 bit or 64 bit	64 bit	64 bit	64 bit	high
R1.3	1PFlop	1 PFlop	1 PFlop	1[1] PFlop	high
Memory					
R1.7	2 – 8 GB per node	8 – 32 GB per node	256 -4096 GB per node	8 – 32 GB per node	medium, high (FN)
R1.8	512MB – 1GB per core	1 – 4 GB per core	2 – 8 GB per core	4 - 16 GB per processing unit	medium, high (FN)
R1.9	required	required	required	required	high
Network – message passing					
R1.12.1	required	required	required	required	high
R1.15.1	not available	not available	not available	10 – 20 Gb/s per node	medium
R1.16.1		not available	not available	0.05 - 0.11 bits/flop	medium
R1.17.1	not available	5 µs	5 µs	1.5 – 3.5 µs	medium (MPP, FN), high (TN)
Network – I/O					

[1] This value should be adjusted according to the results of running PRACE benchmarks on accelerated systems.

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R1.12.2	via I/O nodes	required	required	required via I/O nodes	low (MPP), high (TN, FN)
R1.15.2	not available	not available	not available	10Gb * (disk servers * blade centres)	high (MPP) , high (TN, FN)
R1.16.2		not available	not available		medium (TN, FN)
R1.17.2	not available	5 µs	5 µs	not available	low
Network – management					
R1.12.3	required	required	required	required	medium, high (FN)
R1.15.3		not applicable	not applicable	not available	
R1.16.3	not available	not applicable	not applicable	not available	medium
R1.17.3	not available	not applicable	not applicable	not available	low
External Network					
R1.19	required	required	required	required	medium
R1.20	required	required	required	required (separate network)	medium
R1.21	not available	4 Gb/s	N 10 Gbits links, N~10	10Gbs for each network	medium
I/O and Storage – Global					
R1.22	not available	not available	5 PB	1 – 4 PB (25 – 100 times total memory)	medium
R1.23	not available	not available	not available	4	low
R1.25	not available	not available	not available	1 PB	low
R1.29		1.6E-3 bits/flop	not available	not available	medium
I/O and Storage – Local					

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R1.36	not applicable	not available	32 GB	0	low (FN)

Table 8: System sizing values by architecture

6 Conclusions

This document is the first deliverable from WP7 task 5 *Drafting of technical requirements* and it provides the management board with the first version of a consistent specification of technical requirements to be used in the procurement of the first production Petaflop/s systems in 2009/2010. It provides specific and measurable technical requirements suitable for a procurement process, and has started to fill out architecture-class specific values and relative weightings to rank the importance of selected requirements.

The PRACE infrastructure is aiming to have 3-5 tier-0 Petaflop/s systems and these are likely to reflect the different systems architectures that application codes demand. The challenge of providing technical requirements for a variety of system architectures has been addressed in this document by defining a requirement template. This can be thought of as a check list of requirements which can then be applied to a system procurement, most likely targeted at a specific architecture, by including or excluding the check list items and adding sizing values for the architecture in question. This approach has the benefit of being extensible as new architectures become available and provides a template for ongoing use by the PRACE project.

The selection of technical requirements and architecture values was driven by engagement within WP7 in Task 1 (market update and TCO definition), Task 3 (installation requirements) and an update to Task 2 in D7.1.2 [12] (mapping applications to architectural classes). Other PRACE work packages provided input; WP4 (systems management software), WP5 (prototype technical assessment), WP6 (application benchmark requirements). And of course input was gleaned from WP7 member experience.

The document provides technical requirements for the first Petaflop/s system(s) in 2009/2010 with first estimation of system values, concentrating on system sizing. Values have been added for MPP systems, thin nodes, fat nodes and a cluster of Cell processors, categorised at this stage as an accelerated solution. These choices reflect the market survey update provided by D7.1.2 [12]. No accelerated Petaflop/s systems or vector systems were proposed by vendors for the 2009/2010 time frame and so these have not been included. Only certain values have been specified at this time and these will be updated as more information becomes available.

Refinement of the technical requirements will be provided in the next version of this document, D7.5.2 due in November 2009 and will be driven by the ongoing work in WP7 to get improved proposals from vendors for Petaflop/s systems, from knowledge gained by the efforts to address application Peta-scaling in WP6 and the prototype evaluations in WP5. The second version of this document will also take into account the latest evolution of HPC architectures and will be used to inform the procurement of the second Petaflop/s system(s) and, be an updated basis for procurement for the first Petaflop/s system(s) if such systems have not yet been deployed when this version is published.

The next steps in WP7 focus on the procurement process template activities in task 7.6, of which the document is a key part, before returning to the next iteration of task 7.1 with D7.1.3 [14] followed by the next version of this document.

7 Annex

7.1 HPC Architecture Taxonomy

The definition and classification of architectures suitable for Peta-scale machine design is taken from the D7.1.2 [12] architectures definition update.

The architectures of supercomputing systems are mainly distinguished by the processor architecture used (i.e., vector processors or scalar processors) and the number of cores available in one compute node. Due to the ongoing technological improvements in semiconductor technology, the number of processor cores per processor chip will increase in the next years. The number of processor sockets [2] per compute node is used in this document to distinguish between different system architectures using scalar processors. Where as the term 'compute node' (CN) refers to the collection of processing elements controlled by one operating system instance or one Single System Image (SSI).

MPP systems

Consist of a large number of typically single socket compute nodes (CNs). Each has its own dedicated memory. In order to avoid OS jitter, each CN runs a functionally reduced dedicated OS. I/O is handled by separate I/O nodes. The nodes communicate with each other via a high-speed, usually custom, interconnect. Examples are IBM BlueGene L/P or CRAY XT4/5 systems.

Thin Node Clusters

Thin shared memory node cluster systems currently have typically two (max. four) sockets per CN. The processor sockets are interconnected by the processor - or motherboard to form a shared memory node (SMP or ccNUMA). Examples are Bull INCA or SGI ICE systems.

Fat Node Clusters

Fat shared memory node cluster systems basically have about the same number of sockets per processor board as thin node clusters. However the processor and I/O boards are connected by a specialized network to form a big shared memory node (normally ccNUMA) with a large amount of processors, memory, and I/O resources. Examples for fat shared memory node clusters are Bull MESCA, SGI UltraViolet or IBM Power6 575 systems.

Vector Systems

Systems equipped with vector processors. Those are able to calculate mathematical operations on multiple data elements (arrays) simultaneously with very high memory to processor bandwidth. There are currently only two vendors for vector systems: CRAY and NEC. While scalar processors now have more and more vector functionality, and vector processors tend to have caches, there will still be some significant differences between scalar and vector processors in a near future, in terms of CPU performance and memory bandwidth.

[2] A processor socket is a connector on a computer's motherboard that accepts a CPU and forms an electrical interface with it.

Accelerated Systems

Systems based on two different types of compute units: traditional units and specialized units such as Cell BE, GPU, FPGA, ClearSpeed. Accelerated systems do not constitute a homogeneous family, but it is possible to distinguish between Cell based accelerated systems and other accelerators. Examples of Cell based systems are: IBM Roadrunner (a Cell-accelerated blade cluster of commodity x86 Opteron processors) or PRACE WP7 MariCel prototype at BSC (a cluster of Cell blades for computing with a few Power6 service blades). This distinction is not based on specific architectural issues, but instead, for practical reasons to make evident the different activities on accelerators inside PRACE.

These classes cover the essential architectural characteristics of forthcoming Peta-scale machines.

7.2 Performance Benchmarks

Deliverable D6.3.1 [15] has identified a set of representative applications for the scientific community which has considered the following aspects:

- coverage of relevant application areas,
- representative applications within the covered application areas,
- coverage of (the range of) hardware platforms (prototypes) which are relevant for PRACE,
- Peta-scaling opportunities of benchmark codes with relevant datasets,
- optimisation opportunities of benchmark codes.

These applications have been split into a core list and extension list. The core list will be integrated into a benchmark suite, to help improve the benchmarking of Petaflop/s systems. The extension list is included to increase the coverage of application areas and hardware platforms.

The current plan for porting the application benchmarks to hardware platforms is summarised in *Table 9: Summary on porting efforts for benchmark codes and prototype architectures* taken from D6.3.1[15] and this information feeds into the benchmarking requirements in Chapter 4.

Application	MPP-BG	MPP-Cray	SMP-TN-x86	SMP-FN-pwr6	SMP-FN+Cell	SMP-TN+vector
QCD	Done	In progress		Done		
VASP	Done			Done	Yet to start	Yet to start
NAMD	Done	Done		Done	Yet to start	
CPMD	Done			Done	In progress	Yet to start
Code_Saturne	Done	Done		Done	Yet to start	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		
N3D		Yet to start	In progress	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 9: Summary on porting efforts for benchmark codes and prototype architectures

Green colours denote successful porting, yellow means that porting is in progress, and orange means that porting has not started yet. White indicates application is not being ported.

The Peta-scaling of the application benchmarks is ongoing with tasks 6.4 Peta-scaling and 6.5 Optimisation in WP6 and it is expected that more information on the application benchmark suite can be supplied with the next iteration of this document D7.5.2 [12].

There follows a short summary of each of the application benchmarks:

NAMD is a parallel molecular dynamics code for high-performance simulation of large biomolecular systems.

Code author(s): J. C. Phillips and others	
Application areas: Computational Chemistry, Condensed Matter Physics, Life Sciences	
Language: C++	**Estimated lines of code**: 62,000
Parallelisation technique(s): charm++	
URL: http://www.ks.uiuc.edu/Research/namd/	

CPMD is a parallelized plane wave/pseudopotential implementation of Density Functional Theory, particularly designed for ab-initio molecular dynamics.

Code author(s): M. Parrinello, J. Hutter, A. Curioni and others	
Application areas: Computational Chemistry and Condensed Matter Physics	
Language: FORTRAN	**Estimated lines of code**: 40,000
Parallelisation technique(s): MPI and MPI+OpenMP	
URL: http://www.cpmd.org/	

VASP is a package for performing ab-initio quantum-mechanical molecular dynamics (MD) simulations.

Code author(s): Jurgen Hafner, Jurgen Furthmuller	
Application area: Computational Chemistry and Condensed Matter Physics	
Language: FORTRAN90	**Estimated lines of code**: 100,000
Parallelisation technique(s): MPI	
URL: http://cms.mpi.univie.ac.at/vasp/	

GADGET is a freely available code for cosmological N-body/SPH simulations on massively parallel computers with distributed memory.

Code author(s): V. Springel	
Application area: Astronomy and Cosmology	
Language: C	**Estimated lines of code**: 55,000
Parallelisation technique(s): MPI	
URL: http://www.mpa-garching.mpg.de/galform/gadget/index.shtml	

Code_Saturne is a general purpose CFD code, used for nuclear thermalhydraulics, process, coal and gas combustion, aeraulics, etc.

Code author(s): F. Archambeau, N. Méchitoua, M. Sakiz, Y. Fournier	
Application areas: Computational Fluid Dynamics	
Language: C90 and FORTRAN77	**Estimated lines of code**: 400,000
Parallelisation technique(s): MPI	
URL: http://rd.edf.com/code_saturne	

TORB simulates a plasma in a cylindrical θ-pinch configuration. The code solves the coupled system of gyrokinetic equations for the ions, in the electrostatic approximation, and the quasi-neutrality equation, assuming adiabatically responding electrons.

Code author(s): Jürgen Nührenberg, Francisco Castejon, Alejandro Soba	
Application areas: Plasma Physics	
Language: FORTRAN90	**Estimated lines of code**: 80,000
Parallelisation technique(s): MPI	
URL: N/A	

NEMO is a numerical platform for simulating ocean dynamics and biochemistry, and sea-ice.

Code author(s): G. Madec and NEMO team	
Application area: Earth and climate science	
Language: FORTRAN90	**Estimated lines of code**: 100,000
Parallelisation technique(s): MPI, MPI+OpenMP, NEC autotasking	
URL: http://www.lodyc.jussieu.fr/NEMO/	

ECHAM5 is the 5th generation of the ECHAM general circulation weather model. The

model being used for PRACE benchmarking is ECHAM5-HAM which is a combination of the general atmospheric circulation model ECHAM5 and the atmospheric chemistry and aerosol model HAM

Code author(s): L. Kornblueth and E. Roeckner	
Application area: Earth and climate science	
Language: FORTRAN95/2003, some C	**Estimated lines of code**: 100,000
Parallelisation technique(s): MPI + OpenMP	
URL: http://www.mpimet.mpg.de/en/wissenschaft/modelle/echam/echam5.html	

CP2K is a community code to perform atomistic and molecular simulations of solid state, liquid, molecular and biological systems. It consists of several components for classical molecular dynamics, ab-initio density functional theory. etc.

Code author(s): Juerg Hutter, Joost VandeVondele and others	
Application areas: Computational Chemistry and Condensed Matter Physics	
Language: FORTRAN90	**Estimated lines of code**: 500,000
Parallelisation technique(s): MPI	
URL: http://cp2k.berlios.de/	

GROMACS is a versatile package to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is primarily designed for biochemical molecules like proteins and lipids that have a lot of complicated bonded interactions, but since GROMACS is extremely fast at calculating the nonbonded interactions (that usually dominate simulations) many groups are also using it for research on non-biological systems, e.g. polymers.

Code author(s): Erik Lindahl, David van der Spoel, Berk Hess	
Application areas: Computational Chemistry and Life Sciences	
Language: C and FORTRAN77	**Estimated lines of code**: 1,400,000
Parallelisation technique(s): MPI	
URL: http://www.gromacs.org	

N3D solves the incompressible Navier-Stokes equations by Direct Numerical Simulation (DNS).

Code author(s): Ulrich Rist, Markus Kloker	
Application areas: Computational Fluid Dynamics	
Language: FORTRAN90	**Estimated lines of code**: 40000
Parallelisation technique(s): MPI+ NEC microtasking	
URL: none	

HELIUM simulates the behaviour of helium atoms using time-dependent solutions of the full-dimensional Schrödinger equation.

Code author(s): Jonathan Parker	
Application areas: Atomic Physics	
Language: FORTRAN90	**Estimated lines of code**: 14,500
Parallelisation technique(s): MPI	
URL: none	

GPAW is a density-functional theory (DFT) code based on the projector-augmented wave (PAW) method. It uses real-space uniform grids and multigrid methods.

Code author(s): J. J. Mortensen, C. Rostgaard and others	
Application areas: Computational Chemistry and Condensed Matter Physics	
Language: C90 and Python	**Estimated lines of code**: 40,000
Parallelisation technique(s): MPI	
URL: https://wiki.fysik.dtu.dk/gpaw/	

ALYA is a finite element code for Large Eddy Simulation of compressible and incompressible flows.

Code author(s): Guillaume Houzeaux, Mariano Vazquez, Jose M. Cela	
Application areas: Computational Fluid Dynamics	
Language: FORTRAN90	**Estimated lines of code**: 250,000
Parallelisation technique(s): MPI+OpenMP	
URL: none	

PEPC is a parallel tree-code for computation of long-range Coulomb forces. The forces are calculated based on the Barnes-Hut algorithm. The code takes advantage of multipole-groupings of distant particles to reduce the original $O(N^2)$ scaling of the calculation to an O(N log N) scaling.

Code author(s): Paul Gibbon	
Application areas: Plasma Physics	
Language: FORTRAN90	**Estimated lines of code**: 24,500
Parallelisation technique(s): MPI	
URL: http://www.fz-juelich.de/jsc/pepc	

QCD is a particle physics multi-kernel QCD benchmark code.

Code author(s): Hinnerk Stueben, Kari Rummukainen, Bjoern Leder	
Application areas: Particle Physics	
Language: FORTRAN90, C	**Estimated lines of code**: not specified
Parallelisation technique(s): MPI	
URL: not specified	

AVBP is a turbulent Combustion + CFD code.

Code author(s): Current developer - Gabriel Staffelbach, CERFACS (Toulouse, FRANCE)	
Application areas: Computational Fluid Dynamics	
Language: FORTRAN90	**Estimated lines of code**: 239,578
Parallelisation technique(s): MPI	
URL: not specified	

TRIPOLI_4 is a code for simulating core physics, radiation protection and criticality in nuclear energy using a Monte Carlo method.

Code author(s): CEA Saclay SERMA R&D unit	
Application areas: Atomic physics	
Language: C++	**Estimated lines of code**: 400,000
Parallelisation technique(s): native (TCP/IP sockets)	
URL: http://www.nea.fr/abs/html/nea-1716.html.	

SIESTA is a code for ab initio molecular dynamics simulations of molecules and solids.

Code author(s): Emilio Artacho, Julian Gale, Alberto Garcia, Javier Junquera, Richard M. Martin, Pablo Ordejon, Daniel Sanchez-Portal, Jose M. Soler.	
Application areas: Molecular Dynamics	
Language: Fortran 90	**Estimated lines of code**: More than 105, 000
Parallelisation technique(s): MPI	
URL: http://www.uam.es/departamentos/ciencias/fismateriac/siesta/	

BSIT is a computational geophysics code.

Code author(s): M. Araya, M. Hanzich, F. Rubio, A.C. Lesage	
Application areas: Computational Geophysics	
Language: FORTRAN90 and C	**Estimated lines of code**: 40,000
Parallelisation technique(s): MPI	
URL: none	

7.3 Remaining Architecture-class Specific Requirement Values

These tables contain the remaining architecture-class specific values and have been separated out from the values supplied in 5.1 because at this stage they are less complete and tend to show less variation across architectures. The V (vendor question) rated requirements are not listed here as there is no further architecture-class specific information to add.

The weightings column has not been completed at this stage because many values are not available and because in some cases they reflect site specific procedures and policies and so cannot be generally applied. At the procurement stage these values will be more meaningful to apply.

7.3.1 *Hardware Requirements including system sizing*

The R rated requirements have already been listed in 5.1 and this section contains the remaining D (desirable) requirements.

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
I/O and Storage – Global					
D1.32		not applicable	not available	required	low (FN)
D1.33		not applicable	not available	tape space is 6 – 10 times global disk storage	low (FN)
I/O and Storage – Local					
D1.34	not applicable	146 GB	146 GB	70 – 150 GB	
D1.35	not applicable	8	32 GB	0	
D1.37	not applicable	560 Mb/s	560 Mb/s (read and write)	5 Gbs	
D1.38	not applicable	not available	not available	not available	

Table 10: System sizing values (desirable requirements) by architecture

7.3.2 *Benchmark Requirements*

The requirement to run an application benchmark is dependant on the application being ported to the specific machine being tested and so no entries are offered for R2.1 – R2.20. These can be supplied for a specific procurement. It is expected that the synthetic benchmark suite will be available for most machines in these architecture classes.

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R2.1					
R2.2					
R2.3					
R2.4					
R2.5					
R2.6					
R2.7					
R2.8					
R2.9					
R2.10					
R2.11					
R2.12					
R2.13					
R2.14					
R2.15					
R2.16					
R2.17					

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R2.18					
R2.19					
R2.20					
R2.21	required	required	required	required	

Table 11: Benchmark requirement values by architecture

7.3.3 *Software Requirements including system software and programming environment*

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
Operating System					
R3.1	required (SLES10 + CNK)	required	required (supported by one of the commercial flavours of Linux)	required	
R3.3	required	required	required	required	
R3.4	not available	not available	not available	<5%	
R3.5	not available	not available	10%	<10%	
D3.6		not available	not available	nice to have	
R3.7		not available	required	required	
R3.8		not available	not available	required	
R3.9		not available	not available	not available	
R3.10		not available	not available	5 * number of nodes	
Programming environment - languages, programming models, compilers					

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R3.11	required	required	required	required	
D3.12	required	required	required	required	
R3.13		required	required	required	
R3.14		required	required	required	
D3.15		required	required	required	
R3.16		required	required	required	
R3.17		required	required	required	
D3.18		not available	not available	required	
R3.19	required	required	required	required	
R3.21		required	not available	required	
D3.22			required		
R3.23	required	required	not available	required	
Programming environment - MPI and OpenMP					
R3.25		required	required	required	
D3.26		required	required	required	
D3.28		not available	not available	required	
R3.29	required	required	required	required	
R3.30		required	required	required	
Programming environment – tools					
R3.31	not available	required	debugger with parallel programming support		

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R3.32	not available	thread & MPI profiler/checker	thread & MPI profiler/checker		
R3.33		required	required	required	
R3.34		required	required	required	
R3.35		required	required	required	
D3.36		required	required	required	
Programming environment – libraries					
R3.37		required	required	required	
R3.38		required	required	required	
R3.39		required	required	required	
R3.40		required	required	required	
D3.41		required	required	required	
Programming environment - front end/login nodes					
R3.42		required	required	required	
R3.43	required	required	required	required	
D3.44		required	required	required	
D3.45		required	required	required	
D3.46		required	required	required	
D3.47		required	required		
D3.48		required	required	required	
Scheduling, Batch and Resource Management Software					
R3.49		required	required	required	

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R3.50	required	required	required	required	
R3.51	not available	600s	600s	<300s	
R3.52		required	required	required	
D3.53		required	required	required	
D3.54		not available	required	resource reservation required, not co-scheduling	
D3.56					
Administration Software					
D3.57		required	required	required if doesn't effect rest of system	
R3.58		required	required	required	
R3.59		required	required	required	
R3.60		required	required	required	
D3.61		required	required	required	
Distributed Systems Management Software					
R3.62		required	required	required	
R3.63		required	required	required	
R3.64		required	required	required	
R3.65		required	required	required	
R3.66		required	required	required	
R3.67		required	required	required	

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R3.68		required	required	required	
R3.69		required	required	required	

Table 12: Software requirements by architecture

7.3.4 *Operational Requirements including installation constraint*

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
Users and jobs					
R4.1	not available	1024	5000	100-500	
R4.2	not available	1024	5000	1000 including queued	
Reliability and Availability					
R4.3	not available	not available	not available	100 hours	
R4.4	not available	not available	not available	48 – 72 hours	
R4.5		required	required	required	
R4.7	required	required	required		
R4.8	required	required (node interconnect must remain powered on)	required (node interconnect must remain powered on, does not have to support multi-node applications, such as MPI)		
R4.9		min. RAID5	min. RAID5		
R4.10		required	required		

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R4.11		3	3		
R4.12	required	required	required		
R4.13	not available	not available	not available		
R4.14	required	required	required		
R4.15	not available	not available	not available		
R4.16		not available	not available		
R4.17		required	required		
R4.18		not available	not available		
R4.19		required	required		
D4.20		not available	not available		
Manageability					
R4.21	required	required	required		
R4.22	required	required	required		
R4.23		8h	4h		
R4.24	not available	not available	1min.		
R4.25	required	required	required		
R4.26	not available	10min.	10min.		
R4.27	not available	5min.	10min.		
R4.28	not available	15min.	15min.		
R4.29	not available	10min.	10min.		
R4.30	not available	15min.	15min.		

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R4.31		required	required		
R4.32		required	required		
D4.33		required	required		
R4.34		required	required		
R4.35					
System Security					
R4.36	required	required	required		
R4.37	not available	not available	5000		
D4.38			Required		
Power					
R4.39	not available	not available	not available		
R4.40	not available	not available	not available		
R4.41	not available	not available	not available		
Installation Constraints					
R4.45	not available	not available	not available		
R4.46	not available	not available	not available		
R4.47	not available	not available	not available		
R4.50		8 kW/42U, air cooling 12 kW/42U, liquid cooling The type of cooling depends on site requirement.	12 kW/42U, air cooling or higher for liquid cooling (40 kW); the type of cooling depends on site requirements.		

Table 13: Operational requirements by architecture

7.3.5 *Maintenance and Support Requirement*

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R5.1	not available	3 years	3 years		
R5.2	not available	99%	99%		
R5.3	not available	99%	99%		
R5.4		next business day	next business day		
R5.5		1hr	1hr		
R5.6		9-17 hr	9-17 hr		
R5.7		9-17 hr	9-17 hr		
R5.8		72 hours	72 hours		
R5.9		24 hours	24 hours		
R5.10		not available	not available		
D5.11		required	required		
D5.12		required	required		
R5.13		required	required		
D5.14		required	required		

Table 14: Maintenance and Support requirements by architecture

7.3.6 *Documentation and Training Requirement*

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
Documentation					
R6.1		required	required		
R6.2		required	required		
R6.3		required	required		
R6.4		required	required		
Training					
R6.5		required	required		

Table 15: Documentation and Training requirements by architecture

7.3.7 *Delivery Requirements*

Ref	MPP	Thin Node	Fat Node	Cluster of Accelerated	Weighting
R7.1		required	required		
R7.2		required	required		

Table 16: Delivery requirements by architecture