Identification and Categorisation of Applications and Initial Benchmarks Suite

Alan D. Simpson, Mark Bull, Jon Hill
EPCC, University of Edinburgh, UK
a.simpson@epcc.ed.ac.uk, m.bull@epcc.ed.ac.uk, j.hill@epcc.ed.ac.uk

Abstract

This document contains an analysis of a major survey of HPC systems and applications across the PRACE[1] partners (see [1]). The survey data was used to produce an overall utilisation matrix characterised by scientific area and algorithm. This matrix represents one of the best snapshots of HPC utilisation produced, and could be an invaluable basis to predict the likely utilisation of future European Petaflop/s systems. We discuss a methodology for weighting subset lists of applications to maximise their match with the utilisation matrix. This process is used to guide the selection of applications for inclusion in the representative benchmark suite. By selecting highly used applications from a range of scientific areas and algorithms, we were able to produce a list of no more than 16 applications that are generally representative of European HPC usage and fitted the utilisation matrix well. Drawing on the myriad PRACE expertise in existing and emerging applications areas, we produced a recommended list of nine representative applications, based on the best-fitting list, for use in benchmarking future Petaflop/s systems.

The future of HPC in Europe is also considered. Programming multi-core architectures is seen as a major challenge alongside parallel I/O and techniques for scaling applications to thousands of cores. By comparing large and small systems in the survey it would appear that Particle Physics may well be the main scientific area in a future Petaflop/s system, alongside Materials and Computational Chemistry.

The study detailed in this report is a detailed snapshot of current European HPC, enabling both the choice of a representative benchmark suite and providing insight into what might run on future systems.

[1] The PRACE project receives funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° RI-211528

Table of Contents

List of Figures

List of Tables

Executive Summary

In order to be a success, PRACE needs to understand the software requirements for future Petaflop/s systems. This deliverable identifies they key scientific and technical categories of applications through a survey of most major European HPC systems and the applications that exploit these. It discusses a methodology for identifying a representative subset of applications and recommends appropriate sets of applications for inclusion in a benchmark suite for future Petaflop/s systems.

The bulk of this report is an analysis of a survey of HPC systems and their major applications. We surveyed more than 20 systems representing more than half a Petaflop/s of performance and nearly 70 applications. As well as including most of the largest HPC systems in Europe, this survey also included key HPC systems and applications from all but two of the European countries participating in PRACE.

The analysis of this data gives a snapshot of the current HPC usage across Europe. Particle Physics and Computational Chemistry constitute half of the utilised computational resources. There is a wide spread of job sizes, but there are clear differences between systems, which are not simply related to the size of the machine. The data collected from the specific applications show much variation across sites in Europe, with many codes only being used on a single system.

As part of the survey, the respondents were asked to give their opinion on the future of HPC in Europe. Programming for multi-core architectures was seen as a key challenge. However, as one respondent notes, it may be a lack of personnel that is actually the main challenge.

The data collected from the surveys was collated into a usage matrix characterised by scientific area and type of algorithm. We then selected subsets of the applications and investigated how well they could reproduce, with appropriate weights, the usage matrix. When selecting applications for subsets, we took into account their total usage, the scientific areas and algorithms used, their geographical spread and how scalable they were and as such the usage matrix also acts as a guide on which application areas are of importance. This resulted in a number of subsets which were generally representative of the overall usage and had a good spread of scientific areas and algorithms. It was possible to identify a subset of no more than 16 applications which was in good agreement with the survey data. The applications in these subsets were discussed with all the PRACE partners to produce recommendations for inclusion in the benchmark suite.

The recommended list generated after this discussion is split into two: a core list and a list of possible extension applications. The lists include key applications and spans the full range of scientific areas and type algorithms. Although the best-fit of this data is somewhat poorer than the original list of 16 applications, it represents a more practical view, matching the effort and expertise available within PRACE to the applications.

This list of applications to go into the PRACE benchmarking suite therefore represents the current workload of Tier-1 systems while exploiting the expertise available in the PRACE partnership. A method to provide weights to these codes has been generated that gives the required representation to scientific areas and algorithms.

1 Introduction

The Partnership for Advanced Computing in Europe (PRACE) has the overall objective to prepare for the creation of a persistent pan-European HPC service. PRACE is divided into a number of inter-linked work packages, and one of them (WP6) focuses on the software for petascale systems.

The primary goal of PRACE work package 6 (WP6) is to identify and understand the software libraries, tools, benchmarks and skills required by users to ensure that their applications can use a Petaflop/s system productively and efficiently. WP6 is the largest of the technical PRACE work packages and involves all of the PRACE partners. It is structured into six distinct tasks.

Task 6.1 is responsible for understanding the key existing applications across Europe and identifying likely candidates for future Petaflop/s systems. A key aspect of this work is the establishment of a methodology for categorisation of HPC applications and using this to identify a set of representative applications from those codes currently in use at major European HPC centres. This document discusses the approach that we have taken to understand the existing applications usage across Europe through surveys of major HPC systems and the applications that are used. The data from the surveys have been used to characterise the existing usage and identify representative lists of applications that might be suitable for inclusion in a benchmark suite to evaluate future Petaflop/s systems.

The listed applications will then provide a major focus for future tasks within WP6 which will: optimise and petascale applications (Tasks 6.4 and 6.5); produce a packaged benchmark suite (Task 6.3); investigate the requirements of applications (Task 6.2) and the necessary software libraries for Petaflop/s systems (Task 6.6). The packaged benchmark suite is particularly important for work package 5 which is responsible for the deployment and evaluation of prototype petascale systems in 2008 and 2009.

While the major audience for this report is the other tasks within WP6, the authors believe that the document should also provide valuable information for the entire European HPC and computational research community. This work represents one of the most complete pictures of applications usage of HPC systems ever undertaken and this data should be used to inform future HPC strategy and planning.

1.1 Structure of the Report

The next chapter discusses in more detail the purpose of the survey that we have carried out and the methodology we have employed. Chapter 3 provides an analysis of the data that we have collected about HPC systems across Europe, including the total computational power available, the different architectural classes, the job sizes used, and the breakdown of usage between scientific domains. Chapter 4 analyses the data we have collected on the major HPC applications, including: the ranking of applications by total usage; breakdown of job sizes for each application; and total utilisation by different algorithmic classes, languages and parallelisation techniques. Chapter 5 describes the process used to choose a weighted set of applications which is representative of the usage

on the current Tier-1 systems, in that it is composed of heavily used codes, and adequately reflects the usage distribution across scientific areas and algorithmic patterns. It will also include the results of this process, i.e., a number of recommended weighted sets of applications. Chapter 6 examines possible challenges for the future and tries to look at what could be the workload on a Tier-0 system. Finally, Chapter 7 summarises conclusions of the work. There is also an Annex containing details of the questionnaires used in the survey.

2 Objectives and Methodology

2.1 Objectives

The major objective of this report is a detailed investigation into European HPC applications. As such, we wish to categorise applications usage of the major HPC systems by scientific area and by algorithmic type, and then to identify the major applications in key areas. From these key applications, we will select a modest-sized list of applications that, with appropriate weights, are representative of the overall usage. This list would represent an initial draft of our proposed benchmark suite and the included applications would therefore be a focus for the rest of the activity of WP6. Task 6.3 would then package up the benchmark suite and make it available so that other work packages could analyse the applications performance of PRACE prototype systems and future Petaflop/s systems.

To ensure that the applications in the benchmark suite are chosen for technical, rather than political, reasons, we have undertaken significant surveys of all the PRACE partners covering the major HPC systems and their key applications. We collected 24 system surveys which represents the major systems of each PRACE partner and other large national systems, where possible. The aims of the systems survey were to investigate:

- how much compute power is currently available in Europe;
- the architecture types of the major systems;
- the overall usage for various areas of science;
- the distribution of job sizes;
- the availability of tools;
- what the major future trends in HPC were likely to be;
- which applications were the most important on each system.

Each partner was then asked to complete an applications survey for each application on their system that accounted for more than 5% of the utilisation, and optionally for any other application with was considered to be particularly important for the future. We collected over 100 application surveys representing more than 70 distinct applications. For each application, the information collected, included:

- a brief description of the application and who the authors were;
- the scientific area;
- the implementation techniques used, e.g., language, libraries, algorithm,…;
- utilisation of different job sizes.

The survey questionnaires can be found in Annex A.

The utilisation matrix derived from the surveys consists of 70 categories and is shown in Table 1. These categories are based on ten scientific areas and seven algorithmic

"dwarves". The scientific areas are based on the DEISA [2] benchmark list [3][4][5], largely agree with past work [6] and are:

- Astronomy and cosmology
- Computational chemistry
- Computational engineering
- Computational fluid dynamics
- Condensed matter physics
- Earth and climate science
- Life science
- Particle physics
- Plasma physics
- Other

The dwarves are those algorithm types which constitute classes where membership in a class is defined by similarity in computation and data movement and was first described from Lawrence Berkeley National Laboratory [7]. A dwarf is therefore a grouping of kernels that share both computational and data structure. These dwarves are:

- Dense linear algebra – data is stored in dense matrices or vectors and access is often via unit-level strides. Typical algorithm would be Cholesky decomposition for symmetric systems or Gaussian elimination for non-symmetric systems.
- Sparse linear algebra – data is stored in compressed format as it largely consists of zeros and is therefore accessed via an index-based load. Typical algorithm would be Conjugate Gradient or any of the Krylov methods.
- Spectral methods – data is in frequency domain and requires a transform to convert to spatial/temporal domain. They are typified by, but not restricted to, FFT.
- Particle methods – data consists of discrete particle bodies that interact with each other and/or the "environment".
- Structured grids – Represented by a regular grid. Points on grid are conceptually updated together via equations liking them to other grids. There is high spatial locality. Updates may be in place or between 2 versions of the grid.
- Unstructured grid – data is stored in terms of the locality and connectivity to other data. Points on grid are conceptually updated together, but updates require multiple levels of redirection.
- Map reduce methods – embarrassingly parallel problems, such as Monte Carlo methods, where calculations are independent of each other.

The combination of dwarves and scientific areas gives the 70 categories used in this study. The categories are not completely orthogonal, but are distinct from each other. To construct the matrix (Table 1) the weighting of each application was calculated by multiplying the percentage of the system utilisation spent in jobs using that application, together with the availability of the system during the survey period and the R_{max} of the system – this is what we term the LINPACK Equivalent Flop/s (or LEF). The

application's weighting was then divided equally over each scientific method and algorithm "dwarf" used by that application on that system. For example, a code that was used in computational chemistry and condensed matter physics, and that used both spectral and map reduce methods, and had a weighting of 12 Tflop/s on that system, would contribute 3 Tflop/s to the four corresponding cells in Table 1. This R_{max} weighted utilisation figure (LEF) does not use the actual number of Tflop/s at which an application performed, but rather a weighting system based on a percentage of the machine's Linpack R_{max} value. Ideally, one would want to measure precisely the time spent in each dwarf, but this was not feasible on the timescales involved.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods
Astronomy and Cosmology	0	0.62	4.91	3.59	5.98	2.99	0
Computational Chemistry	15.35	26.09	1.80	3.45	7.49	0.53	12.98
Computational Engineering	0	0	0.53	0.53	0	0.53	2.8
Computational Fluid Dynamics	0	1.70	7.37	3.05	0.32	3.00	0
Condensed Matter Physics	9.10	15.07	1.62	0.73	1.76	0.28	5.70
Earth and Climate Science	0	2.03	5.83	1.33	0	0.26	0
Life Science	0	4.72	0.94	0.13	0.94	0.28	3.46
Particle Physics	12.50	0	4.59	0.92	0.10	0	89.27
Plasma Physics	0	0	1.33	1.33	3.55	0.42	0.63
Other	0	0	0	0	0	0	0

Table 1. The utilisation matrix based on the survey results.
The utilisation matrix is made up of 70 categories; 10 scientific areas and 7 algorithmic "dwarves". The figure in each cell is an estimate of the number of Tflop/s burned in each category. White boxes are those with no usage. Orange boxes are those with usage greater than zero, but less than 5 Tflop/s usage. Red boxes signify usage greater than 5 Tflop/s.

2.2 Methodology

As indicated above, the PRACE partners collected information on all major systems in Europe to understand the current usage of European HPC. The systems included were any system with a peak performance of more than 10 Tflop/s and in addition, any other significant system at any PRACE partner site, regardless of peak flop rate. For each system data were collected on the system and the top applications running on that system over at least a three month period, the only exceptions to this rule being the two BlueGene/P systems at Jülich and Daresbury which were allowed to submit for shorter time periods. This information was collected using two online survey forms; one for the system and another for applications running on that system.

The two questionnaires shared some key data in order to allow the subsequent analysis to be performed. Both questionnaires asked for:

- Details on the person submitting the form in order for any queries relating to the data submitted to be followed up.
- A unique system identification to be selected from a number of designated choices. It is this data that allows an application survey to be linked to the system that it is run on.
- The period that the survey covered was also asked for both questionnaires.

After the three common sets of data, the system survey then asked the following:

- Generic details of the system: Name, manufacturer, model, processor type, clock rate, memory, configuration of system (cores per chip, chips per node, etc), I/O configuration, cache, interconnect system.
- Performance figures: R_{max}, R_{peak}, availability, utilisation.
- The use of the system: job sizes, scientific areas, number of users.
- System software: scientific libraries, compilers, performance analysis tools, I/O libraries, parallel debugging tools. This information was collected for use in Task 6.6.
- The top applications: applications using > 5% of the available cycles. For each of these an application survey was expected.
- Other information: privacy concerns, future directions of HPC and any other relevant information.

The applications questionnaire asked for the following information:

- Generic information on the application: name, description, authors.
- Scientific areas covered.
- Algorithms used (in the form of the seven dwarves).
- Languages and libraries: languages used, parallelisation techniques, lines of code, libraries required.
- Usage: utilisation percentage of the application on the system in question, job size distribution, parallelisation technique distribution.
- Other information: privacy concerns, other relevant information.

The two online survey forms (see Annex A) were written in PHP and data were stored in both a relational database (MySQL) and as a file, in case of database failure and to keep a record of the information filled in (rather than the database, in which the data were "cleaned" prior to analysis). The data in the database were then checked manually for obvious errors, such as ensuring the forms were correctly processed, spelling errors, acronyms, etc and sanitised. These checked data were then used in the analysis described in later sections. A SQL file containing all data collected is available on the PRACE intranet (BSCW) and is available to all PRACE partners. The data contains some confidential information and is therefore not suitable for public dissemination. In addition, the data collected and presented here was verified and checked independently by a number of PRACE partners, especially CSCS.

The analysis was carried out in Matlab, which has a function for producing least-squares best-fit with non-zero weights. A Matlab function was created that solved the best-fit solution and output the weights and other pertinent variables. An interface to the database was created using PHP, which enabled Matlab to access the current data in the database, rather than relying on manually downloaded data.

3 Analysis of Systems Usage

In this Chapter, we present the results of the systems survey. Table 2 shows the 24 systems for which a questionnaire was completed. For each system the PRACE centre, machine manufacturer and model are given, together with the peak (R_{peak}) and achieved Linpack (R_{max}) flop rates in Gflop/s (as defined by the Top500 list of supercomputers [8]) and the number of cores. The systems represent 14 PRACE partners from 12 countries. The total power of systems is 926 Tflop/s peak, and 675 Tflop/s achieved Linpack, from 169,522 cores.

Name	Centre	Manufacturer	Model	Architecture type	R_{peak}	R_{max}	Cores
Jugene	FZJ	IBM	Blue Gene/P	MPP	222822	167300	65536
MareNostrum	BSC	IBM	JS21 cluster	TNC	94208	63830	10240
HLRB II	BADW-LRZ	SGI	Altix 4700	FNC	62259	56520	9728
HECToR	EPSRC	Cray	XT4	MPP	63437	54648	11328
Neolith	SNIC	HP	Cluster 3000 DL140	TNC	59648	44460	6440
Platine	GENCI	Bull	3045	TNC	49152	42130	7680
Hexagon	SIGMA	Cray	XT4	MPP	51700	42000	5552
Galera	PSNC	Supermicro	X7DBT-INF	TNC	50104	38170	5376
Jubl	FZJ	IBM	Blue Gene/L	MPP	45875	37330	16384
BCX	CINECA	IBM	BladeCenter Cluster LS21	TNC	53248	19910	5120
Stallo	SIGMA	HP	BL460c	TNC	59900	15000	5632
Palu	ETHZ	Cray	XT3	MPP	17306	14220	3328
HPCx	EPSRC	IBM	P575 cluster	FNC	15360	12940	2560
Huygens	NCF	IBM	p575 cluster	FNC	14592	11490	1920
Legion	EPSRC	IBM	Blue Gene/P	MPP	13926	11110	4096
hww SX-8	USTUTT-HLRS	NEC	SX8	VEC	9216	8923	576
Louhi	CSC	Cray	XT4	MPP	10525	8883	2024
murska.csc.fi	CSC	HP	CP400 BL ProLiant SuperCluster	TNC	10649	8200	2176
Jump	FZJ	IBM	p690 cluster	FNC	8921	5568	1312
ZAHIR	GENCI	IBM	p690/p690+/p655 cluster	FNC	6550	3900	1024
HERA	GENCI	IBM	p690/p575 cluster	FNC	3000	3700	384
XC5	CINECA	HP	HS21 cluster	TNC	-	2400	256
Milipeia	UC-LCA	SUN	x4100 cluster	TNC	2200	1600	520
TNC	PSNC	IBM, Sun	ibm e325/sun v40z/x4600 cluster	TNC	1577	1182[2]	330
Totals					926176	675415	169522

Table 2: PRACE partner systems included in survey.
MPP – Massively Parallel Processing, TNC – Thin Node Cluster, FNC – Fat Node Cluster, VEC – Vector.

[2] Estimated value based on 75% of the R_{peak} of this machine

Figure 1 shows the distribution of compute power (as measured by R_{max}) by architecture type. Just under half the compute power comes from MPP systems, just over a third from thin-node clusters, 15% from fat node clusters and 1% from vector systems.

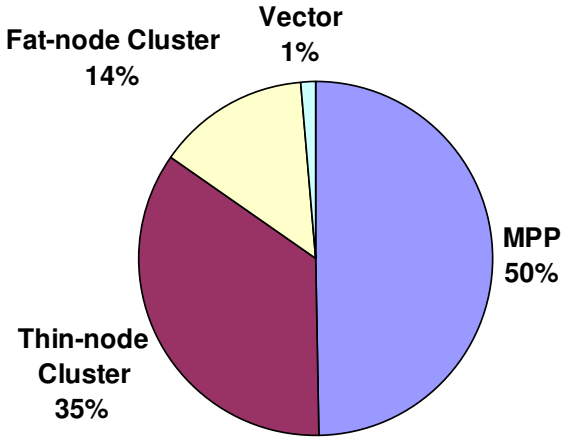

Figure 1: Compute power by architecture type

Figure 2 shows the percentage availability and the percentage utilisation of the available cycles for each of the systems. The mean availability was 97.7%, with a number of systems reporting 100% availability, while the lowest availability was 95%, a figure also reported by a number of systems. Utilisation varies between 20% (Galera) and 95% (Neolith), with a mean of 71.1%. Out of the 674 Linpack-equivalent Tflop/s available, 430 Tflop/s were actually consumed by applications.

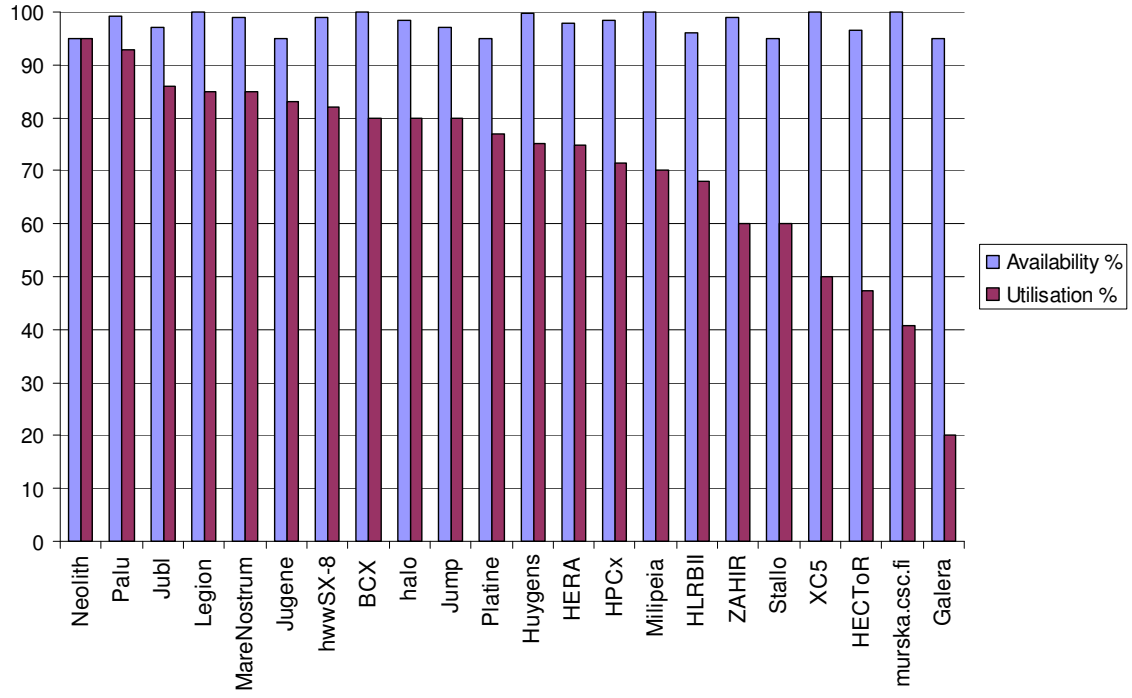

Figure 2: System availability and utilisation

Figure 3 shows the job size distribution of the utilised cycles on each system for five ranges of job size: up to 32 cores, 33-128 cores, 129-512 cores, 513-2048 cores and more than 2048 cores. Note that the distribution is expressed as a percentage of the utilised cycles, not as a percentage of submitted jobs. A wide range of behaviour is observed: some systems run only large jobs, some only small jobs and some a more even spread across the ranges.

To further understand how the systems are utilised, we computed a mean job size for each system, by assuming that all the jobs in each range are on average the midpoint of the range (and that jobs in the >2048 range are assumed to be of size 4096). We then divided this mean job size by the number of cores in the system, to obtain a metric which approximately represents the fraction of the system occupied by the average job. (Note that our assumption about the job sizes over 2048 cores may be quite inaccurate for very large systems: for the two systems with the largest fraction of usage in the >2048 range (Jubl and Jugene), the value was derived exactly from system data logs). This metric is shown in Figure 4. The fraction of the system occupied by the average job varies from just over 1%, to almost 33%. This shows that the way machines are used varies widely: some systems are divided very finely between lots of small jobs, whereas others mostly run a small number of large jobs.

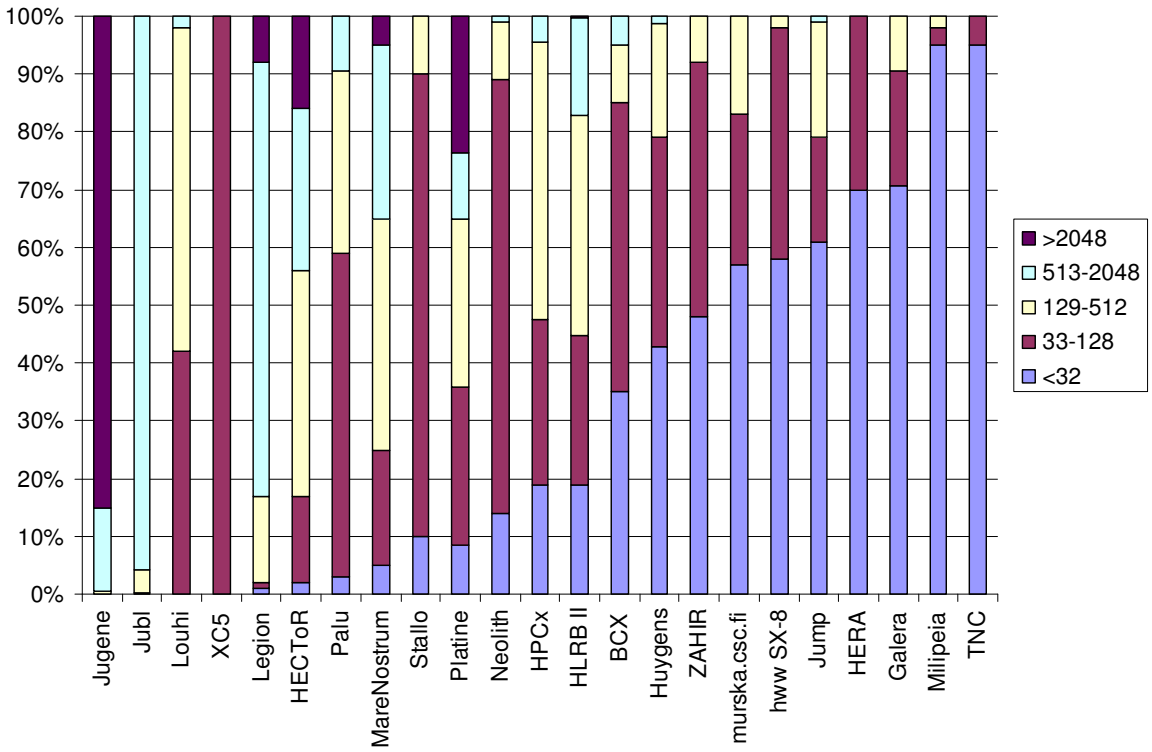

Figure 3: Job size distribution by system

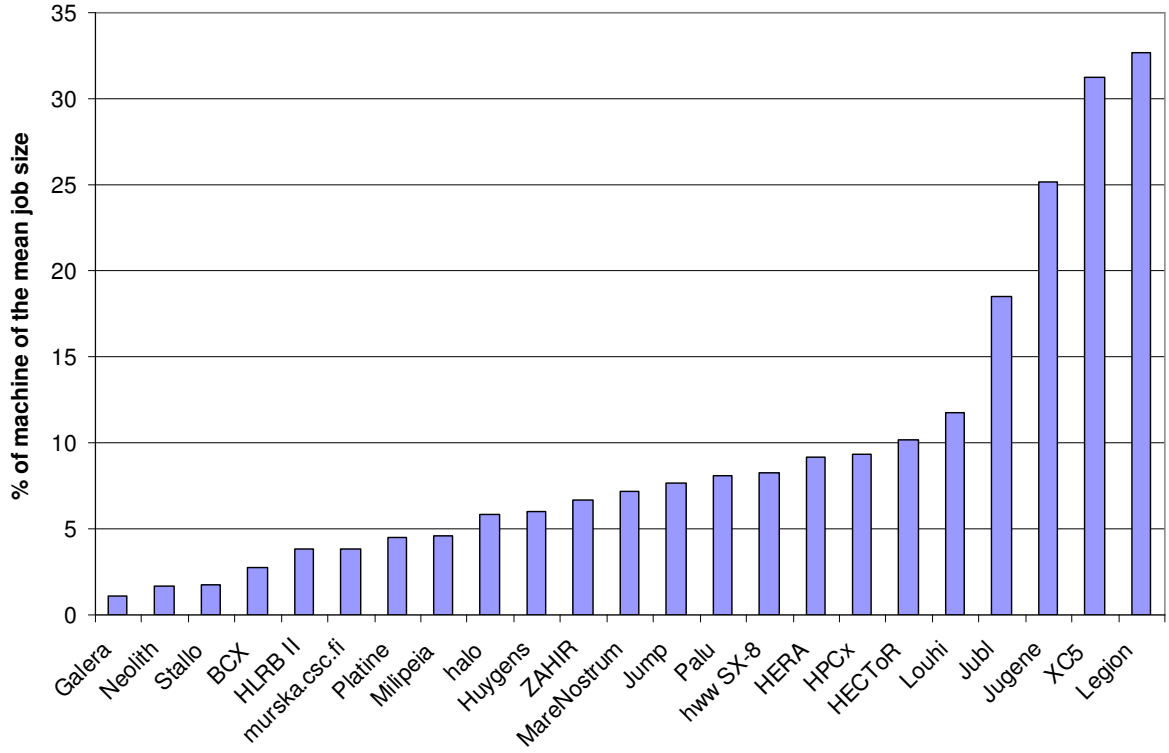

Figure 4: Mean job size as a percentage of system size.
Most data are estimated from the survey (as described in the text). However, data was available on Jugene and Jubl for actual mean job sizes, so exact figures were used for these two systems.

Figure 5 shows the job size distribution aggregated across all systems. More than a quarter of the LEFs were in jobs of more than 2048 processors, though almost all of this comes from a one system, Jugene. The remaining LEFs are divided roughly equally between the other four job size ranges.

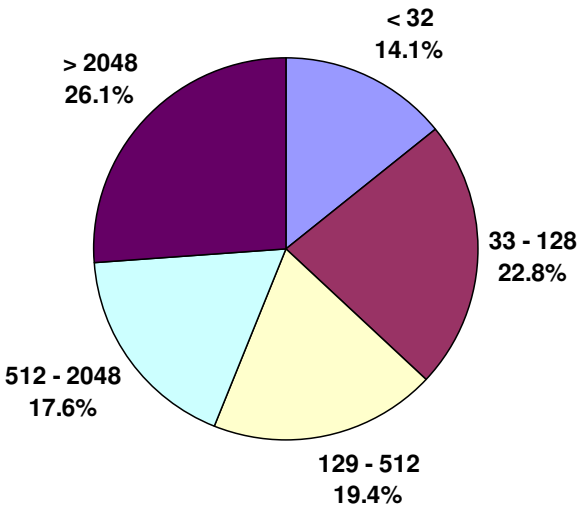

Figure 5: Aggregated distribution of LEFs by job size

Figure 6 shows the distribution of cycles used by scientific area for each system. Only a few systems are dedicated to a small number of scientific areas: most systems have substantial usage from a number of different scientific areas. Condensed Matter Physics and Computational Chemistry show usage across all systems (apart from XC5), but has very high usage on the smaller systems. In contrast Particle Physics has large usage only on the larger machines.

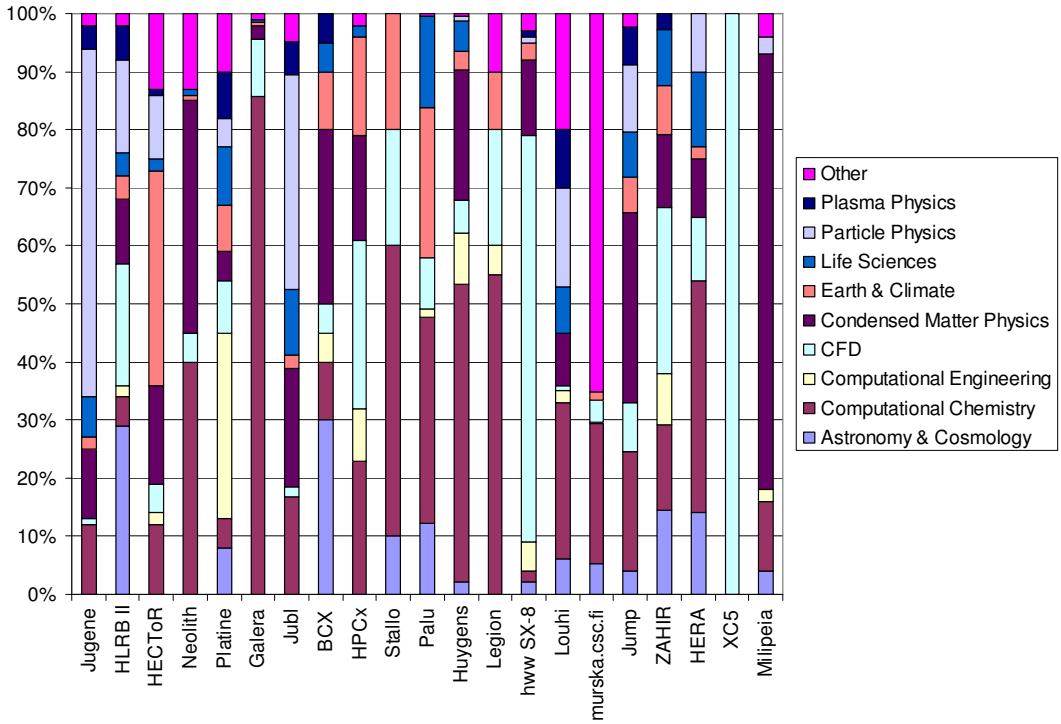

Figure 6: Scientific area distribution by system

Figure 7 shows the aggregated distribution of LEFs used in the different scientific areas across all the systems. Particle physics accounts for nearly one quarter of the LEFs, though these are mainly from one system (Jugene). The next largest areas are Computational Chemistry and Condensed Matter Physics. Together, these account for over one third of the LEFs, and, as we will see in Chapter 4, there is a high degree of overlap between the applications in these two areas. The remaining areas consume between 3.3% and 8.6% of the total LEFs.

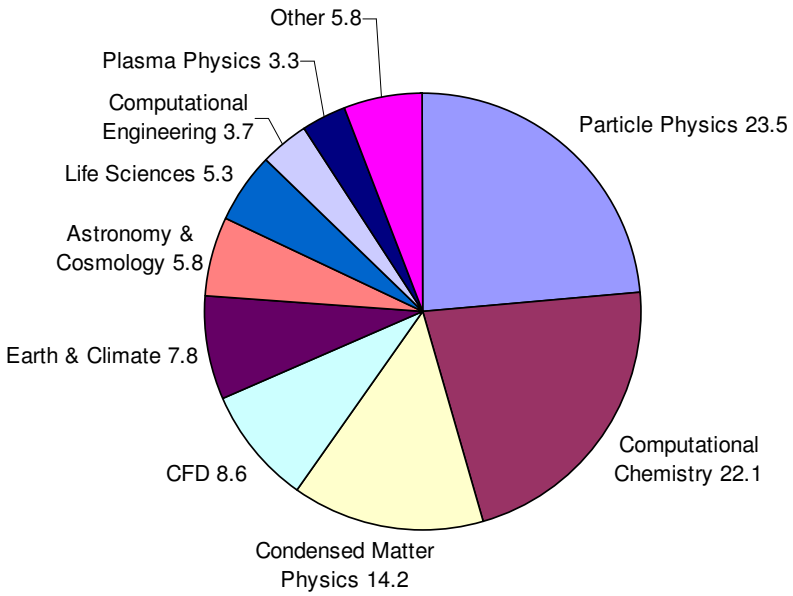

Figure 7: Aggregated distribution of LEFs by scientific area

Twenty out of the 24 systems were able to report the number of users on the system. The number of users, together with the available compute power per user (obtained by dividing the system R_{max} by the number of users), is shown in Figure 8 (note that no R_{max} per user value was available for TNC and the value in Table 2 is estimated). The number of users per system varies widely, from five (XC5) to over a thousand (Jump). The total number of users on these 20 systems was 4733, giving an average of 237 users per system. The R_{max} compute power per user also shows a very wide variation, from 5.3 Gflop/s (Jump) to just over 1 Tflop/s (Galera). On average (over all users) each user has access to 113 R_{max} Gflop/s of compute power, equivalent to around 20-25 fast cores. In practise, the compute power is not evenly shared between users: observations suggest that in many cases a small number of users are responsible for using a high percentage of cycles on a given system.

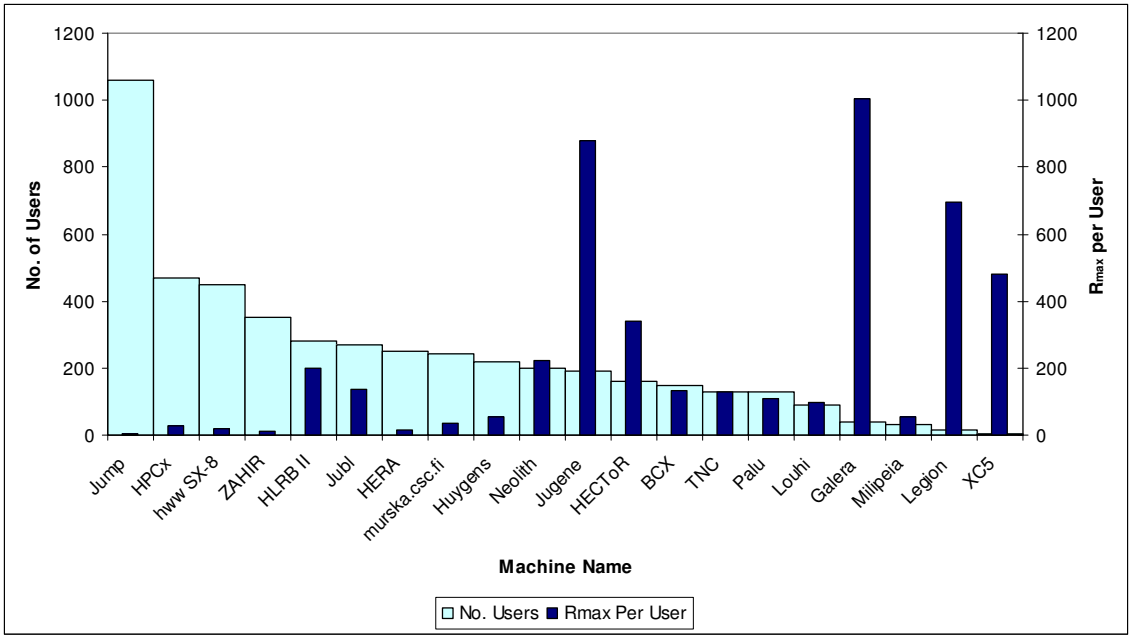

Figure 8: Number of users and R_{max} per user

4 Analysis of Applications Usage

A total of 111 applications surveys were returned across the 24 systems: these come from 69 distinct applications. Table 3 lists these applications, ranked by the number of cycles consumed (as Linpack-equivalent Gflop/s, obtained as the product of the fraction of the machine time used with the R_{max} rating of the system) and the number of systems from which the application was reported. A number of applications survey returns did not include the utilisation figure, and several application surveys were from machines for which the associated systems survey had not been completed: in these cases the number of cycles used is shown as zero. The total usage of these applications in Table 3 is 254 Linpack-equivalent Tflop/s (56%) out of a total of 430 Linpack-equivalent Tflop/s usage reported from the 24 systems.

Of the 69 applications, 47 were reported as being used on one system only, 16 on two systems and only six (VASP, NAMD, DALTON, CPMD, GROMACS and AVPB) on three or more systems. The most widely used code was VASP, which was reported as being used on nine different systems.

Application Name	LEFs Used (Gflop/s)	Number of systems using this code
overlap and wilson fermions	54923	2
vasp	35766	9
lqcd (twisted mass)	25007	2
lqcd (two flavor)	12393	2
namd	10335	4
dalton	9975	3
cpmd	9680	5
gadget	8412	2
dynamical fermions	7947	1
spintronics	5206	2
materials with strong correlations	4846	2
dl_poly	4779	2
casino	4223	1
quantum-espresso	3982	1
cactus	3798	1
trio_u	3202	1
smmp	3181	2
tfs/piano	3092	1
gromacs	2903	3
pepc	2857	2
tripoli4	2802	1
chroma	2745	1
wien2k	2713	1
bam	2713	1
trace	2713	1
bqcd	2713	1
cp2k	2525	1

helium	2249	1
magnum	1398	1
pdkgrav-gasoline	1233	1
crystal	1193	2
su3_ahiggs	1181	1
n3d	1060	1
unified model	992	1
cosmo model	936	1
nemo	884	2
parallel particle mesh (ppm) library	803	1
moldypsi	729	1
oslo stagger code	713	1
parcas	615	1
avbp	601	5
turbomole	574	1
octopus	480	1
pencil code	435	1
occam	433	1
high resolution computational of local dissipation scales	432	1
elmfire	422	1
gpaw	369	2
fenfloss	353	1
unified emep model	285	1
molpro	238	1
metallic layers, electronic and magnetic phenomena	216	1
gaussian	139	1
blast	131	1
iqcs	0	2
siesta	0	2
elmer	0	2
alya	0	1
torb	0	1
ccsm	0	1
hirlam	0	1
mglet	0	1
espresso	0	1
bsit	0	1
coamps	0	1
sage	0	1
code_saturne	0	1
gamess-uk	0	1
fluent	0	1

Table 3: Applications usage

Of the 69 applications, all but two use MPI for parallelisation. The exceptions are Gaussian (OpenMP) and BLAST (sequential). Of the 67 MPI applications, six also have standalone OpenMP versions and three have standalone SHMEM versions. Ten

applications have hybrid MPI/OpenMP implementations, two have hybrid MPI/SHMEM versions and one has a hybrid MPI/Posix threads version. Only one application was reported as using MPI2 single sided communication.
Table 4 shows the usage of different base languages by the 69 applications. Just under half use more than one base language and 16 applications combine Fortran with C and/or C++.

Language	No. of applications
Fortran90	50
C90	22
Fortran77	15
C++	10
C99	7
Python	3
Perl	2
Mathematica	1

Table 4: Base language usage by applications

Figure 9 shows the distribution of LEFs used by the reported applications by scientific area. Comparing this to Figure 7 (which shows the distribution of all cycles on all systems by scientific area) we observe some significant differences: the data from the applications over-represents Particle Physics by almost a factor of two. The over-representation of Particle Physics in the applications data is due to Particle Physics applications being run on the largest systems. Computational Chemistry is also slightly over-represented, while the remaining areas are all consequently underrepresented in the applications data.

Figure 10 shows the distribution of the LEFs used by the reported applications by job size range. Comparing this figure to Figure 5, which shows the same metric for all LEFs used across all the systems, we observe very good agreement.

We also asked survey respondents to categorise applications by algorithmic dwarves. In many cases, one application uses more than one dwarf, but it is hard to apportion the LEFs used by an application meaningfully between dwarves. For the purposes of this study, we have simply divided the LEFs used by an application equally between all the dwarves it uses. Making this assumption, the distribution of LEFs used by the algorithmic dwarves is shown in Figure 11. Map reduce methods is the most used dwarf, followed by spectral methods and dense linear algebra.

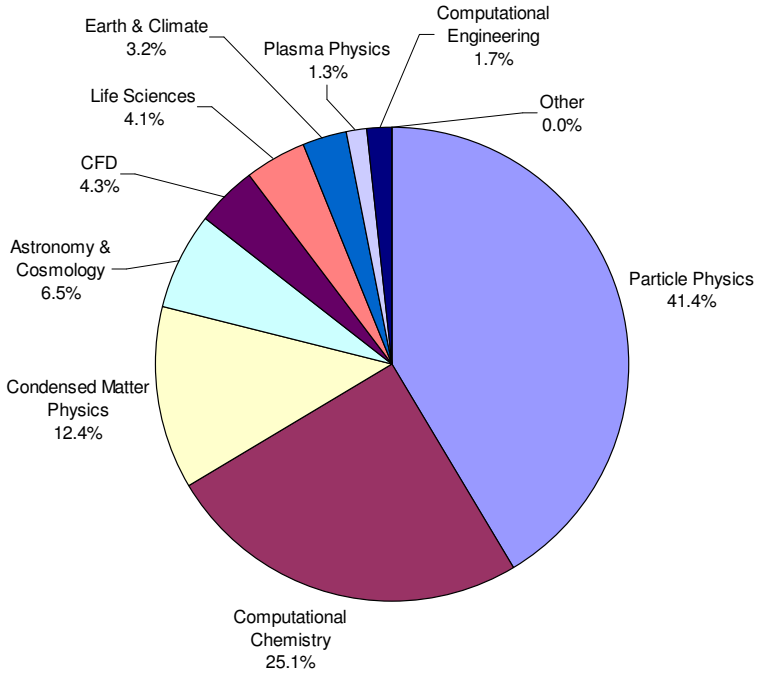

Figure 9: Distribution of applications usage by scientific area

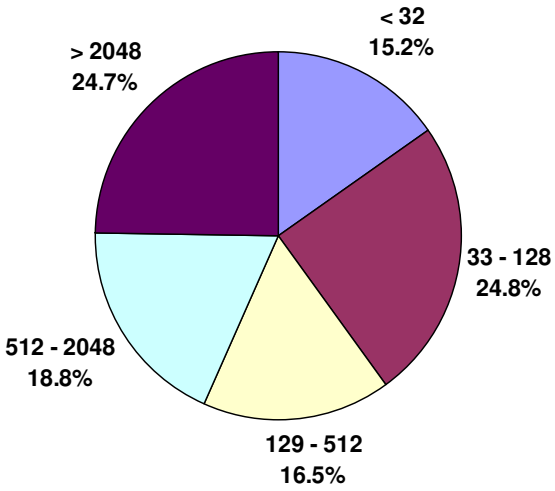

Figure 10: Distribution of applications usage by job size

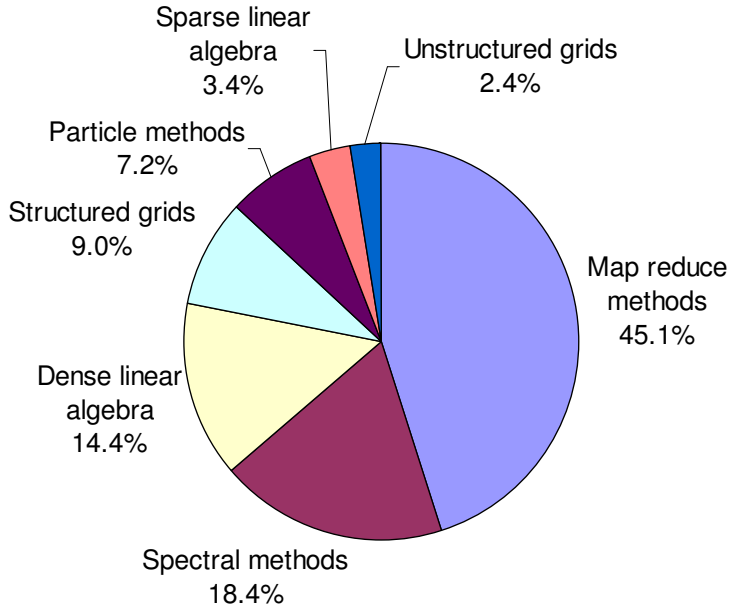

Figure 11: Distribution of applications usage by algorithmic dwarves

The results presented in this chapter show that, apart from the over-representation of Particle Physics applications noted above, the applications reported in the survey (which account for a little over half of the total utilised LEFs) are a good representation of the overall usage of the surveyed systems. The application survey covered a wide range of application areas with 69 distinct applications submitted. Together they form a detailed picture of the usage of current European HPC systems and give some indication to the algorithmic make-up of the HPC ecosystem.

5 Choosing a Representative Subset of Applications

This Chapter describes the process used to choose a weighted set of applications which is representative of the usage on the current Tier-1 systems, in that it is composed of heavily used codes, and adequately reflects the usage distribution across scientific areas, algorithmic patterns (the 7 dwarves), base languages and parallelisation techniques.

The result of this process will be at least one recommended set of weighted applications.

5.1 Method

The usage matrix derived from the surveys (Table 5) is used as a basis of defining the important scientific areas and algorithms. The number in each cell is the amount of LEFs used in that scientific area and algorithm "dwarf" based on the application surveys that were submitted.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods
Astronomy and Cosmology	0.00	0.62	**4.58**	**3.26**	**5.43**	**2.99**	0.00
Computational Chemistry	**15.09**	**24.89**	**1.14**	**2.79**	**7.49**	0.49	**12.98**
Computational Engineering	0.00	0.00	**0.53**	**0.53**	0.00	**0.53**	*2.80*
Computational Fluid Dynamics	0.00	1.70	**7.09**	**1.06**	0.32	**1.01**	0.00
Condensed Matter Physics	**9.02**	**14.33**	**0.96**	0.06	1.76	0.28	**5.70**
Earth and Climate Science	0.00	0.70	**3.31**	0.00	0.00	0.22	0.00
Life Science	0.00	**4.72**	**0.94**	0.13	**0.94**	0.28	*3.46*
Particle Physics	**12.50**	0.00	**4.32**	0.92	0.10	0.00	**89.27**
Plasma Physics	0.00	0.00	0.00	0.00	***2.22***	0.42	*0.63*
Other	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5. Usage matrix from the surveys.
The figures in boxes is the total number of LEFs (Tflop/s) used in that particular scientific area and dwarf. White boxes are those with no usage. Orange boxes are those with usage greater than zero, but less than 5 Tflop/s usage. Red boxes signify usage greater than 5 Tflop/s. Bold figures denote those boxes that have an application representing it in the proposed list (see next section). Italicised figures are those that have an application representing that box in the extended list (see next section).

The application list was then generated by picking the top codes that represented the most heavily used cells in the usage matrix (the red cells in and checking how well those codes (based on their scientific area and algorithms – Table 6) could match the usage matrix using a weighting scheme defined by:

$$\mathbf{Uw} = \mathbf{v}$$

Where U is a two-dimensional matrix containing the functionality of the each application in each category, w is a one-dimensional vector containing the weight for each application and v is the two-dimensional utilisation matrix (Table 5). The fit is the sum of the square of the residuals between the calculated v and the actual v. The aim of this method is to provide a set of weights for any given list of applications that best fits the current utilisation of Tier-1 systems surveyed. The list of applications can then altered by adding and subtracting codes to attempt to reduce the residual. Further modifications were then based on an attempt to spread codes around PRACE centres whilst ensuring the applications were suitable for inclusion in a benchmarking suite, and without unduly affecting the residuals.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods
Astronomy and Cosmology		PDKgrav-GASOLINE	Oslo Stagger Code BAM Cactus PDKgrav-GASOLINE	BAM Cactus	GADGET RAMSES	GADGET RAMSES	
Computational Chemistry	VASP CPMD Dalton GPAW Gaussian Molpro TURBOMOLE Siesta GAMESS-UK Crystal Wien2k CP2K CRYSTAL	SIESTA VASP CPMD NAMD Gromacs ESPResSO GAMESS-UK Crystal DL_POLY Wien2k CP2K	SIESTA NAMD Gromacs GPAW CPMD PPM library DL_POLY	SIESTA GPAW Gaussian Molpro TURBOMOLE CPMD GAMESS-UK HELIUM	NAMD Gromacs Namd Dalton DL_POLY PPM library	Crystal DL_POLY	CPMD DL_POLY CASINO VASP
Computational Engineering			TRIO_U	TRIO_U		TRIO_U	TRIPOLI4
Computational Fluid Dynamics		Pencil Code High resolution computational of local dissipation scales Magnum N3D	AVBP PPM library MGLET N3D TFS/Piano TRACE TRIO_U	ALYA TRIO_U FLUENT N3D Code_Saturne Fenfloss	Pencil Code PPM library	ALYA AVBP Sage TRIO_U FLUENT Code_Saturne fenfloss	
Condensed Matter Physics	Quantum-ESPRESSO VASP GPAW CPMD CP2K	Materials with strong correlations SIESTA VASP CPMD Quantum-ESPRESSO DL_POLY CP2K octopus	SIESTA GPAW CPMD DL_POLY VASP octopus	SIESTA GPAW CPMD	Parcas DL_POLY moldyPSI	DL_POLY	metallic layers, electronic and magnetic phenomena Spintronics DL_POLY
Earth and Climate Science		BSIT Magnum HIRLAM	BSIT NEMO COSMO Model Unified Model OCCAM COAMPS CCSM Unified EMEP model HIRLAM echam5	BSIT HIRLAM	echam5	Sage OCCAM	
Life Science		NAMD Gromacs DL_POLY	NAMD Gromacs DL_POLY PPM library	BLAST	NAMD Gromacs Namd PPM library DL_POLY	DL_POLY	SMMP DL_POLY
Particle Physics	LQCD (Twisted Mass)		PPM library CHROMA bqcd SU3_AHIGGS	CHROMA	Parallel Particle Mesh (PPM) library		Overlap and Wilson Fermions SU3_AHIGGS LQCD (Twisted Mass) LQCD (Two Flavor) CHROMA dynamical fermions
Plasma Physics			TORB gene	TORB	TORB PEPC	Elmfire	PEPC
Other	Elmer IQCS		Elmer	Elmer		Elmer	

Table 6. A list of applications in each of the scientific area/dwarf category. The colours are the same as Table 5 and the order in which the applications appear is of no significance.

There are a number of caveats that should be considered when evaluating the suitability of this method to obtain a list of applications for a benchmark suite. The data from the surveys is not perfect. A single code may have been filled in differently by different centres. This is shown clearly using Siesta as an example (Figure 12) Two surveys were

completed and completely different outcomes were produced. A possible explanation may be that one centre is using an older version of *SIESTA* which transform sparse matrices to a dense structure to make use of SCALAPACK and hence spends the majority of time carrying out Dense Linear Algebra. The most recent version of SIESTA carries out the same computations entirely with Spare Linear Algebra. While this was allowed in the survey as a centre may use an application primarily in one scientific area, whereas another centre may use it in a different area, there may be unintentional differences. For heavily-used applications, this may result in noticeable changes to the usage matrix. In addition, the survey is a snapshot – new machines are scheduled for installation in the coming months which will change the data. However, the method of assigning weights will still be valid, only the weights (and possibly applications) may change.

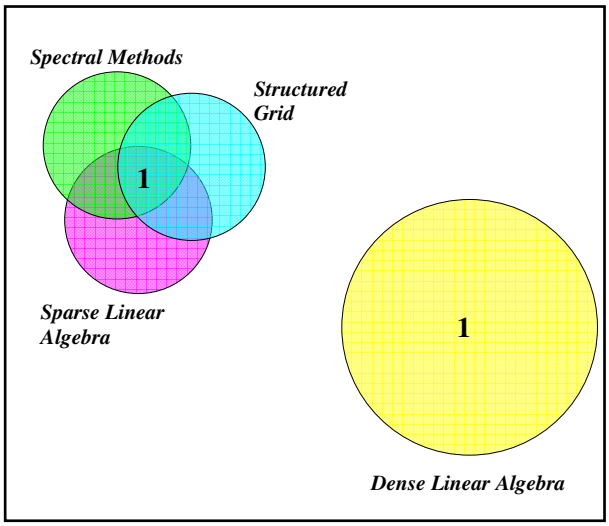

Figure 12. Venn diagram of algorithms in Siesta – a computational chemistry application. Two surveys were completed which show very different algorithms. See text for explanation. Diagram generated by Neil Stringfellow (CSCS).

Secondly, this methodology favours large machines in some respects. While this is not necessarily a disadvantage as a Petaflop/s machine will be, by definition, a large machine, this may not accurately reflect the expected usage on such a machine.

A third caveat is that new machines may not have settled into what may be expected for the final distribution of applications and scientific areas. Unfortunately, the timing of this survey means that FZJ's BlueGene/P is fairly new and the survey had to be completed based on less than the requested three month minimum operating period which may distort that importance of those applications that are currently running on it. However, it is important to note that the applications running on the BlueGene/P are very similar to those on the older BlueGene/L at FZJ, so this may not be a large factor.

The list of applications put into the system is based on human knowledge. Without being able to predict the future, it is impossible to know which scientific areas or applications will be important on a Tier-0 Petaflop/s system. The methodology presented here only assigns weights based on current knowledge. These may, of course, be altered depending on what one thinks might be more or less important on future systems. Of course, what

one person thinks of as an important future scientific area, another may disagree. It is therefore optimal to base the PRACE benchmark suite on data where possible.

Finally, the classification of applications is not perfect. Not all codes fit neatly into a scientific area or algorithmic dwarf. This is apparent from codes such as *Gadget* and *Ramses* (which are used for very different purposes and each contains functionality not available in the other) appear in the same categories. Particle Physics applications are perhaps another example of this possible mis-categorisation as although they use Monte Carlo methods (and as such might be considered Map Reduce Methods) the communication pattern is more indicative of a structured grid algorithm with nearest-neighbour communication. The choice of category in which to place such an application is therefore a difficult one and is not entirely clear. In addition, by categorising codes in this way important differences may be hidden, such as data-access patterns, which will affect performance. Other factors, such as I/O are not accounted for. Ideally, one would wish to:

1. Identify basic computational kernels,
2. Identify basic communication kernels,
3. Identify basic I/O patterns,
4. Measure the use of resources for each of the kernel classes, and,
5. Analyse the applications according to these kernel classes.

However, it would be impossible, in the timeframe available, to cover all of the applications named in the survey in this level of detail. With a total of 70 categories, most applications will fit into one or more of categories used in this study. In addition, with the expert knowledge on the applications available as part of PRACE, it should be possible to ensure applications that have demands not included in the categorisation, such as I/O heavy applications, are included

The issues discussed above can be tested somewhat by examining subsets or modified version of the data, for example removing new, large machines or ignoring the size of the system completely. This has been carried out and is discussed in Section 5.3 in order to address the above caveats. Clearly, this methods has some weaknesses, but there an opportunity to substitute suboptimal applications generated from this methodology. In addition, this method allows weights to be assigned to applications in the benchmark suite to reflects the current HPC usage across Europe, which is the best indicator available to future usage.

5.2 The Application List

The method above was used to create the following list of applications to be included in the PRACE benchmarking suite (Table 7). This list is based on the fit of areas/dwarves and produces a residual of 21.5 Tflop/s for the 13 core applications and 21.1 Tflop/s when the three additional applications are added. A list of applications with a lower error could no doubt be generated (in fact including all applications surveyed would produce a perfect fit), but this list represents a pragmatic and practical view of a possible benchmark list.

Application name	Weight (Tflop/s)	Weight (as %)	Scientific Areas	Dwarves	Sites using code
overlap and wilson fermions	72.3 (72.3)	28.0% (27.2%)	Particle physics	Map reduce methods	FZJ
vasp	61.2 (61.2)	23.7% (23.0%)	Computational chemistry Condensed matter physics	Dense linear algebra Spectral methods Structured grids Map reduce methods	SNIC CSC NCF EPSRC USTUTT-HLRS FZJ BADW-LRZ
namd	27.2 (27.2)	10.5% (10.2%)	Computational chemistry Life sciences	Structured grids Particle methods Spectral methods	CSC SIGMA EPSRC ETHZ
lqcd (twisted mass)	25.0 (25.0)	9.7% (9.4%)	Particle physics	Dense linear algebra Map reduce methods	FZJ
trio_u	13.3 (13.3)	5.2% (5.0%)	Computational engineering Computational fluid dynamics	Structured grids Unstructured grids Sparse linear algebra	GENCI
su3_ahiggs	11.0 (11.0)	4.3% (4.2%)	Particle physics	Structured grids Map reduce methods	CSC
gadget	9.0 (9.0)	3.5% (3.4%)	Astronomy and cosmology	Particle methods Unstructured grids	BADW-LRZ CINECA
cactus	8.5 (8.5)	3.3% (3.2%)	Astronomy and cosmology	Structured grids Sparse Linear algebra	BADW-LRZ
casino	7.5 (7.5)	2.9% (2.8%)	Computational chemistry	Map reduce methods	EPSRC
torb	6.2 (4.9)	2.4% (1.9%)	Plasma physics	Structured grids Sparse linear algebra Particle methods	BSC
helium	6.0 (6.0)	2.3% (2.2%)	Computational chemistry	Sparse linear algebra	EPSRC
nemo	5.8 (5.8)	2.3% (2.2%)	Earth and climate sciences	Structured grids	GENCI EPSRC
spintronics	5.4 (5.4)	2.1% (2.0%)	Condensed matter physics	Map reduce methods	FZJ
Possible extensions to list					
smmp	(3.2)	(1.2%)	Life sciences	Map reduce methods	FZJ
tripoli4	(2.8)	(1.1%)	Computational engineering	Map reduce methods	GENCI
pepc	(2.5)	(1.0%)	Plasma physics	Particle methods Map reduce methods	FZJ

Table 7. The proposed list of applications that are a result of analysing the survey data.

The residuals (Table 8) for the extended set of applications shows that areas of Computational Chemistry (spectral methods and dense linear algebra) are under

represented (positive residual), whereas as Life Sciences and Condensed Matter Physics are over represented (negative values). The high residuals are mainly due to major applications that cover a number of areas and dwarves, such as *VASP* and *NAMD* overlapping with other applications that contribute cycles in the same areas. However, removing these applications and replacing them with similar applications would lead to an increased number of applications in order to cover the scientific areas and algorithms.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods
Astronomy and Cosmology	0.00	0.62	0.66	-0.66	1.50	-1.50	0.00
Computational Chemistry	5.92	12.34	-2.52	0.00	3.16	0.53	0.00
Computational Engineering	0.00	0.00	-1.69	-1.69	0.00	-1.69	0.00
Computational Fluid Dynamics	0.00	1.70	4.51	0.30	0.22	0.25	0.00
Condensed Matter Physics	-7.20	-2.11	-8.95	0.00	1.34	0.00	0.00
Earth and Climate Science	0.00	1.33	0.00	1.33	0.00	0.26	0.00
Life Science	0.00	-4.32	-4.32	0.13	-4.32	0.00	0.00
Particle Physics	0.00	0.00	1.79	-1.79	0.00	0.00	0.00
Plasma Physics	0.00	0.00	-0.32	-0.32	0.64	0.42	-0.64
Other	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 8. Residuals of the 70 categories using the applications list in Table 8.

5.2.1 *Substituting Applications*

One of the primary advantages of this method for PRACE is that applications from the list above can be substituted for applications that cover the same algorithmic or scientific areas. The list presented in Table 7 was refined using the expertise that existed in the PRACE partnership. Proposals implemented were:

- Combining the LQCD applications (*overlap and wilson fermions, LQCD(twisted mass)* and *su3_ahiggs*) into a single LCQD benchmark. These LQCD applications have similar structure and although they consume a large number of LEFs, their use is limited to a small number of systems.
- Adding a weather forecasting code to complement the ocean modelling code *NEMO* in the Earth and Climate Science area. The chosen code was *ECHAM5*.
- Dropping *spintronics* from the list, as it is essentially a harness which uses *VASP* to perform the principal calculations.
- Adding the cosmology application *RAMSES* to the list of possible extensions. The computational methods used by this application differ significantly from those used by *GADGET* and *CACTUS*.

These changes were simple to make and reduced the number of applications, whilst still covering the main scientific areas and dwarves. However, there was further discussion on the CFD and computational chemistry/condensed matter applications. This is primarily due to the difficulty of picking a few representative applications for these two areas – a number of codes exist in each area (see Table 6) and only a limited number could be sensibly included in an application benchmark suite. In order to decide the most

representative applications, PRACE centres were asked to rank the following lists of codes in order of preference.

List 1: Structured CFD

- MGLET
- TRACE
- TFS/Piano
- N3D

List 2: Unstructured CFD

- TRIO_U
- Elmer
- AVBP
- ALYA
- Code Saturne
- OpenFOAM
- FENFLOSS

List 3: Chemistry/Condensed Matter

- NAMD
- VASP
- CASINO
- CPMD
- CP2K
- SIESTA
- DL_POLY
- GROMACS
- GPAW
- Quantum_Espresso

List 1 contains only true structured CFD codes as it was felt that adding a true structured CFD code (*trio_u* uses unstructured grids, which by definition can be structured, but the application is not necessarily using the most efficient structured algorithms) was necessary. Only two centres felt they had the necessary experience to rank the codes in List 1, but *N3D* was considered probably the most suitable application in this area.

In List 2, *Code_Saturne* was chosen as it had the highest number of votes and a PRACE partner willing to co-ordinate effort on this application.

In List 3, *NAMD*, *CPMD* and *VASP* were the three most popular codes, which were included in the main list. The next two most popular were *CP2K* and *GROMACS*: these were added to the list of possible extensions. Therefore, the following changes were made to the list of applications:

- Replacing the CFD application *TRIO_U* with *Code_Saturne*.

- Moving *HELIUM* to the list of possible extensions as, although it was classified as a Computational Chemistry code, it would be more accurately described as Atomic Physics. Furthermore, it is used on only one system, by a small number of users.
- Replacing *CASINO* with *CPMD*, which is more widely used.
- Adding the combustion application *AVBP* to the list of possible extensions, as a complement to *Code_Saturne* in the CFD area.
- Adding the Computational Chemistry/Condensed Matter applications *GROMACS* and *CP2K* to the list of possible extensions.
- Adding *N3D* to the list of possible extensions
- Moving *Cactus* to the extended list

The result of these changes is a core list of nine applications, and a set of ten possible extensions. The core list is shown in Table 9 and the possible extensions in Table 10. Some of the applications in the extended list are direct replacements for those in the core list, e.g. *CP2K* could replace *CPMD*, and either *Ramses* or *Cactus*, could be added, but not both.

Application name	Scientific Areas/Description/URL
QCD benchmark	Particle physics
	This is a synthetic benchmark application designed to include all the key LQCD algorithms
vasp	Computational chemistry, Condensed matter physics
	Performs ab-initio quantum mechanical molecular dynamic simulations.
	http://cms.mpi.univie.ac.at/vasp/
namd	Computational chemistry, Life sciences
	Molecular dynamics code aimed mostly at simulating biomolecules
	http://www.ks.uiuc.edu/Research/namd/
cpmd	Computational chemistry, Condensed matter physics
	Density function calculations with molecular dynamics
	http://www.cpmd.org
Code_Saturne	Computational fluid dynamics
	General purpose CFD code, used for nuclear thermalhydraulics, process, gas combustion
	http://rd.edf.com/code_saturne
gadget	Astronomy and cosmology
	Cosmological N-Body simulations
	http://www.mpa-garching.mpg.de/~volker/gadget/index.html
torb	Plasma physics Solves gyrokenetic equations using a “particle in cell” method
echam5	Atmospheric modelling Earth and climate sciences http://www.mpimet.mpg.de/en/wissenschaft/modelle/echam/echam5.html

Application name	Scientific Areas/Description/URL
nemo	Earth and climate sciences
	Ocean modelling
	http://www.lodyc.jussieu.fr/NEMO/

Table 9. The proposed core list of applications.

Application name	Scientific Areas/Description/URL
avbp	Computational fluid
	dynamics Massively parallel CFD code that solves laminar and turbulent compressible reacting flows
	http://www.cerfacs.fr/cfd/avbp_code.php
cp2k	Computational chemistry
	Condensed matter physics Performs atomic and molecular simulations of solid state, liquid, molecular and biological systems.
	http://cp2k.berlios.de/
gromacs	Computational chemistry
	Life sciences Molecular dynamics package, primarily designed for biomolecules.
	http://www.gromacs.org/
helium	Other
	Computational Atomic, Molecular, and Optical Physics code that simulates the Helium atom
smmp	Life sciences
	Protein folding and interactions in silicio
tripoli4	Computational engineering
	General purpose radiation transport code using Monte Carlo methods to simulate neutron and proton behaviour in three dimensions.
pepc	Plasma physics
	Laser-plasma interactions
	http://www.fz-juelich.de/jsc/pepc
ramses	Astronomy and cosmology
	Built to simulate formation of large-scale structures and galaxy formation
	http://irfu.cea.fr/Projets/COAST/ramses.htm
cactus	Astronomy and cosmology
	Simulates the evolution of black holes using finite-difference techniques
	www.cactuscode.org
N3D	Computational fluid dynamics
	Incompressible Navier-Stokes equation direct numerical simulation.

Table 10. Possible extensions to the core list of applications

5.3 Weighting Schemes

To test the robustness of the methodology and attempt to arrive at a set of weights for the applications, various subsets of the data were used to generate a set of weights for the core list of applications (Table 9). The datasets used were:

1. All data.
2. R_{max} set to one for all machines – this removes the size of the machines from the utilisation matrix meaning only percentage utilisation is considered ("Flat" data)
3. Using PRACE Principal Partner systems only
4. Using PRACE General Partner systems only
5. Removing, large, new machines (HECToR and Jugene – see Table 2)
6. The codes were placed into their scientific areas and the weight corresponds to the percentage weight of that scientific area obtained from the systems surveys (see Figure 7). For multiple applications in the same scientific area, the percentage was split equally.

Together, the four additional datasets address many of the issues of the original methodology (see Section 5.1). In order to carry out this comparison, entries in the database had to be made the QCD Benchmark. For the purpose of this exercise, the QCD Benchmark was classified as particle physics using map-reduce methods, structured grids, spectral methods, dense linear algebra, and sparse linear algebra. In addition, *nemo* and *echam5* were placed in the list together, as according to the classification and utilisation matrix (Table 6 and Table 5) one of these applications would have zero weight, despite the differences in algorithms and scientific area, which the classification presented here does not capture.

The weights derived from each dataset (Table 11) show remarkable similarity. The main differences are between the computational chemistry applications (*NAMD*, *VASP* and *CPMD*) and the QCD Benchmark. This reflects the usage patterns of these applications – QCD is utilised more on large machines and hence gains a greater weight when large machines are included (Principal Partners, and All data), but reduces when the size of the machine is not considered or large machines are removed. The computational chemistry applications are relatively more popular (in terms of percentage utilisation) on the smaller systems.

Code	All	"Flat" data	Principal Partners	General Partners	No New machines	From System
namd	11.3%	18.5%	7.7%	24.4%	20.4%	13.6%
cpmd	1.6%	3.2%	0.9%	4.2%	0.0%	13.6%
vasp	26.4%	37.9%	22.7%	39.9%	37.8%	13.6%
qcdbenchmark	49.4%	19.2%	62.4%	2.5%	17.7%	26.5%
gadget	3.9%	4.3%	3.2%	6.3%	8.9%	3.3%
Code_Saturne	2.9%	4.0%	1.3%	8.4%	4.9%	3.4%
torb	2.2%	6.7%	1.2%	5.8%	5.1%	8.8%
nemo/echam5	11.3%	18.5%	7.7%	24.4%	20.4%	13.6%

Table 11. Weights for the core list of codes using a variety of datasets based on the survey data. See text for more explanation.

Calculating normalised residuals across all five methods shows little change in the residuals, which are poorer than the list of applications shown in Table 7 as expected.

5.3.1 *Proposed Weights*

The above analysis shows that giving equal weights to applications in the benchmark suite is unjustified as giving, say, *Code_Saturne* and *VASP* the same weight would not agree with any of the above analyses. It is therefore proposed that a weighting scheme should be implemented in due course, once the benchmarking suite is complete.

Rather than use the (perhaps spuriously accurate) actual weights in Table 7 or Table 11, we propose using a "banded weighting" system. We propose five bands: 25%, 10%, 5%, 3% and 1%. Applications can then be placed into the appropriate band. The number of applications in the final two bands would depend on whether the primary or extended list was used.

5.4 Comparison to DEISA

Given that a European benchmarking suite already exists, the DEISA benchmark suite, it is perhaps interesting to compare the list proposed with this. All the codes in the DEISA suite were in the survey, so the above analysis can be repeated using the list of applications in the DEISA benchmark suite (Table 12).

Code	Weighting (TFLOP/s)
su3_ahiggs	93.77
cpmd	58.26
dl_poly	19.71
namd	14.54
ramses	8.96
nemo	5.36
fenfloss	4.99
pepc	4.18
gene	1.33
iqcs	0
quantum-espresso	0
bqcd	0
echam5	0
gadget	0

Table 12. Applications from the DEISA list.
The weighting is derived from the analysis of these applications using the data collected in the surveys.

The weighting of the codes show that some of the codes are unnecessary using this classification: *gadget* is identical to *ramses*, *quantum-espresso* shares the same characteristics as *CPMD*, *BQCD* is covered by *su3_ahiggs*, and *echam5* is identical to *nemo*. In reality this is not the case as *nemo* and *echam5* do not exactly replicate each

other, neither do *gadget* and *ramses*. In addition, *IQCS* (a Quantum Computer Simulator) has no weight as there was no recorded usage in the survey for "Other" scientific areas (in which *IQCS* falls in this categorisation).

Examining the residuals (Table 13), it is clear that the DEISA list has high residuals in computational chemistry and condensed matter physics. This is partly due to the applications in one area being used in the other. It is therefore difficult to improve the fit without the addition of applications to "balance" the existing applications. Life science is also under-represented. Clearly, the DEISA list was not created to replicate the current European HPC usage (as it was created three years ago), the weighting of the applications was not considered and as such cannot be expected to replicate the recently collected data. The above is not a criticism of the DEISA work, but a comparison of the PRACE proposed application list and the DEISA application list and establishes that using the DEISA benchmarking suite unmodified would not fulfil the requirements of the PRACE project.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods
Astronomy and Cosmology	0.00	0.62	4.92	3.59	1.50	-1.50	0.00
Computational Chemistry	8.88	15.88	-8.41	-3.02	3.75	-0.78	5.19
Computational Engineering	0.00	0.00	0.53	0.53	0.00	0.53	2.80
Computational Fluid Dynamics	0.00	1.70	6.73	0.03	0.22	-0.03	0.00
Condensed Matter Physics	-4.25	-0.47	-7.31	-6.47	0.03	-1.31	4.11
Earth and Climate Science	0.00	1.33	0.00	1.33	0.00	0.26	0.00
Life Science	0.00	-3.74	-3.74	0.13	-3.74	-1.31	1.87
Particle Physics	12.50	0.00	-42.39	0.92	0.00	0.00	42.39
Plasma Physics	0.00	0.00	0.00	1.33	1.46	0.42	-1.46
Other	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 13. Residuals for the 70 categories using the DEISA benchmark suite.

6 Outlook Towards the Future

The above work surveys the current usage of Tier-1 systems across Europe as it is impossible to collect data on a system yet to be built. However, as part of the survey we asked what the future may hold for HPC in Europe – which are the emerging technologies and scientific areas that may be used or part of the future Tier-0 system. In addition, one can also look at the differences across existing centres in order to glean some insight into what the workload of a Tier-0 system may look like.

6.1 Principal and General PRACE partners

The PRACE project consists of two types of partners; Principal and General Partners. This gives an opportunity to assess the effect that the large machines have on the survey as, in general, the large machines reside at Principal Partner sites. There are some exceptions, of course, but this split is pre-defined by the PRACE project and was felt to be adequate for this analysis.

Using the system surveys results, job sizes show a major difference with the vast majority of jobs at General Partner sites (83%) consisting of less than 128 cores (capacity computing). In contrast, more than 55% of jobs at the Principal Partner sites consist of more than 512 cores (capability computing). This difference is largely due to Jugene (the BlueGene/P machine) at FZJ. In terms of scientific areas there are only a few minor differences. In General Partner sites Computational Chemistry is used for 46% of the cycles, compared to 14% in Principal Partner sites. However, Principal Partners use 31% of their cycles in Particle Physics, compared to just 1% at General Partner sites. Other than this exchange of time used in Particle Physics for time in Computational Chemistry, the other scientific areas remain largely consistent.

The differences in scientific areas are also reflected in the applications that are used on the two sets of systems. *VASP* appears in both sets and is the most heavily used application at General Partner sites, compared to third (although the top chemistry code) at Principal Partner sites. Apart from the inclusion of a number of LQCD applications (mainly from FZJ) there is little difference between the two sets of applications.

Examining the usage matrices produced using the applications surveys for both sets of systems highlights the differences in scientific area, but also some differences in the algorithms used. The main difference is that heavy usage of map reduce methods at the principal partner sites (Table 14 and Table 15), which burn over 100 Tflop/s using this algorithm type. This may be due to the inherent scalability of these methods and hence their suitability to running on very large machines. Although a large percentage of these are in Particle Physics, there is also a significant increase in Computational Chemistry and Condensed Matter Physics.

Based on these results, one may expect that a future Tier-0 Petaflop/s system might run a large number of map-reduce type jobs, particularly in the area of particle physics. However, this is heavily influenced by the BlueGene/P system, which is a relatively new machine (and as such may not have settled into a steady state of usage patterns), but is the nearest system Europe has to a Petaflop/s system.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods	Total (%)
Astronomy and Cosmology	0	0.62	1.33	0	0	2.99	0	**9.3%**
Computational Chemistry	10.99	5.98	0.72	0.54	7.07	0	0	**47.5%**
Computational Engineering	0	0	0	0	0	0	0	**0.0%**
Computational Fluid Dynamics	0	0.92	0.1	0	0.32	0	0	**2.5%**
Condensed Matter Physics	6.8	6.98	0.3	0.06	1.34	0	0	**29.1%**
Earth and Climate Science	0	0.7	1.72	0	0	0	0	**4.5%**
Life Science	0	0.55	0.65	0.13	0.65	0	0	**3.7%**
Particle Physics	0	0	0.69	0	0.1	0	0.59	**2.6%**
Plasma Physics	0	0	0	0	0	0.42	0	**0.8%**
Other	0	0	0	0	0	0	0	**0.0%**
	33.4%	**29.6%**	**10.3%**	**1.4%**	**17.8%**	**6.4%**	**1.1%**	

Table 14. Usage matrix based on the General PRACE Partner sites.
Percentages at the right-hand side and along the bottom show the percentage of LEFs used in that scientific area or dwarf. Note that these figures are based on the usage of applications (utilisation multiplied by the R_{max} of the system the application is run on), not on the results of the system surveys.

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods	Total (%)
Astronomy and Cosmology	0	0	3.59	3.59	5.98	0	0	**6.4%**
Computational Chemistry	4.36	19.44	0.42	2.25	0.42	0.53	12.98	**19.7%**
Computational Engineering	0	0	0.53	0.53	0	0.53	2.8	**2.1%**
Computational Fluid Dynamics	0	0.79	7.27	1.06	0	1.52	0	**5.2%**
Condensed Matter Physics	2.3	7.42	0.66	0	0.42	0.28	5.7	**8.2%**
Earth and Climate Science	0	0	2.32	0	0	0.26	0	**1.3%**
Life Science	0	4.17	0.28	0	0.28	0.28	3.46	**4.1%**
Particle Physics	12.5	0	3.9	0.92	0	0	88.68	**51.6%**
Plasma Physics	0	0	0	0	2.22	0	0.63	**1.4%**
Other	0	0	0	0	0	0	0	**0.0%**
Total (%)	**9.3%**	**15.5%**	**9.2%**	**4.1%**	**4.5%**	**1.7%**	**55.7%**	

Table 15. Usage matrix based on the Principal PRACE Partner sites.
Percentages at the right-hand side and along the bottom show the percentage of LEFs used in that scientific area or dwarf. Note that these figures are based on the usage of applications (utilisation multiplied by the R_{max} of the system the application is run on), not on the results of the system surveys.

6.2 Future Trends

As part of the survey, we asked respondents of the systems survey to write about three "future directions" of HPC. There were clear themes across all responses, which were:

1. Adapting to multi-core hardware
2. Parallelisation of I/O
3. Scalability of codes as machines gain more and more cores
4. Hardware acceleration

The first of these relates to the current trend of processors to increase the number of cores on a chip to increase performance, rather than simply increase the clock frequency. This trend can be seen across a range of different architectures, in particular those built from commodity components, but also more bespoke HPC architectures, such as BlueGene/P. In order to fully utilise the extra cores a mixed-mode parallelism (e.g. MPI and OpenMP or MPI and pThreads) is thought to be necessary. This will, of course, increase the complexity of applications.

Parallel I/O is seen as important as the performance of disk-based data access is not increasing as quickly as processor performance. In order for I/O not to be a bottleneck in performance, parallelisation of the I/O is seen as the obvious way forward. Tools such as HDF5 and NetCDF help in this respect.

Clearly, if new Petaflop/s systems have hundreds of thousands of cores, then applications must scale to thousands of cores in order to benefit from the increased computing power available. However, this is not straightforward to achieve and often requires re-writing of key algorithms and kernels. While future tasks in WP6 will investigate this scaling, the number of applications involved will be limited. As one respondent noted, the major future challenge may not be due to hardware or software, but simply having enough personnel able to carry out the necessary changes in applications.

Hardware acceleration means the building of heterogeneous systems. In such a system, a conventional system will have hardware accelerators, for example FPGAs, Cell Processors or GPUs, available for applications to use. The idea is that the key kernels can be re-written for these accelerators. Most accelerators required a lot of work in order to maximise the performance, especially FPGAs, although tools are being developed to reduce this problem.

Other issues raised were the availability of performance analysis tools, the development of new languages and compilers (perhaps related to the hardware accelerator point), the handling of large amounts of data and the introduction of new scientific areas to HPC. Clearly, the HPC community has a lot of work to do in order to use a Petaflop/s system effectively.

7 Conclusions and Further Work

The data presented here represent a snapshot of current HPC usage in Europe. The data is not perfect and there are some issues that could be considered if repeating this exercise. Nevertheless, this data provides a detailed view of current usage and some insights into what might be run on a future Petaflop/s systems. Of course, the main purpose of collecting this data was to provide information on which applications should be considered for the PRACE benchmarking suite. By generating lists of applications and checking their fit to the usage data, for some set of weights, the resulting applications give a good representation of the current HPC usage, both in terms of scientific area and type algorithms. The list generated to best fit the data (whilst keeping the list of applications reasonable) was then modified based on expert knowledge within PRACE. We fully recognise that the BlueGene/P at Jülich exerts an influence on the final results, but as this system is the closest that Europe has to a Petaflop/s system it was prudent to include it in this survey, despite the reservations on the length of time it has been operating.

The data collected was then analysed in a number of ways to examine what changes occur to the weighted list of applications when the data is modified. Five additional datasets were used and the major differences between them occurred in the areas of Particle Physics and Computational Chemistry/Condensed Matter Physics. The former has a heavier weighting when large machines (in particular Jugene) are included and the latter have a higher weighting when the larger systems are excluded. This difference is particularly apparent when examining the different usage patterns in the Principal and General PRACE Partner systems.

In addition, the DEISA benchmarking suite was also examined using the utilisation information collected. Unmodified, this suite is not suitable for use within PRACE, but the proposed benchmarking application list and the DEISA list share significant overlap.

Finally, the PRACE partners were asked about the challenges that face European HPC in the next few years. Programming multi-core architectures, parallel I/O and the scalability of applications were seen as key challenges.

The final application lists (core and extensions/replacements) is representative of the current HPC usage and includes some of the most popular applications currently being used in Europe, and therefore forms a suitable basis for the PRACE benchmarking suite.

Tasks 6.3, 6.4, and 6.5 will use this document to inform them of a choice of applications that would make a benchmarking suite. This may be different to those lists proposed above due to practical, licence or other reasons. In addition, these tasks will carry out performance analysis, petascaling and optimisation on the list of applications.

Tasks 6.6 will use information from this survey to inform the software and libraries that will be required on the future Petaflop/s systems.

A follow-on document will analyse the hardware requirements of the benchmarking applications and will look at the users' requirements of future systems. This will complete the picture presented here and in previous WP6 deliverables to present a

rounded view, from the hardware, applications, HPC centres and users' point of view, of what any future Petaflop/s system should look like.

8 References

[1] http://www.prace-project.eu

[2] http://www.deisa.eu

[3] Identification of an Initial Set of Application Codes and Low Level Benchmarks. *Deliverable ID: eDEISA-D-eSA4-B1*

[4] Release of Version 1 of DEISA Benchmarks. *Deliverable ID: eDEISA-D-eSA4-B2*

[5] Benchmarking Activity Status Report. *Deliverable ID: eDEISA-D-eSA4-B3*

[6] HET, 2006. The scientific case for European Super Computing Infrastructure. Available from http://www.hpcineuropetaskforce.eu/deliverables

[7] Asanovic, et. al. 2006. The Landscape of Parallel Computing Research: A View from Berkeley. Technical Report No. UCB/EECS-2006-183. Available from http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html

[8] Top 500 Supercomputer Sites http://www.top500.org/