



PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

A Survey of HPC Systems and Applications in Europe

Dr Mark Bull, Dr Jon Hill and Dr Alan Simpson
EPCC, University of Edinburgh

email: m.bull@epcc.ed.ac.uk, a.simpson@epcc.ed.ac.uk



Overview

- Background
- Survey design
- Survey results
 - HPC Systems
 - HPC Applications
- Selecting a benchmark suite



Background to the survey

- The PRACE project is working towards the installation of Petaflop/s scale systems in Europe.
- Requirement for a set of benchmark applications to assess performance of systems before and during procurement process
- Benchmark applications should be representative of HPC usage by PRACE partners
- To understand current applications usage, we conducted a survey of PRACE partners' current HPC systems



- We took the opportunity to gather other interesting data as well
- We also devised a method for selecting (and weighting) a set of applications which can be considered representative of the current usage
 - we wanted to do this in a quantifiable way
 - we wanted to avoid political considerations
 - ...but it was not entirely successful!

Survey Design

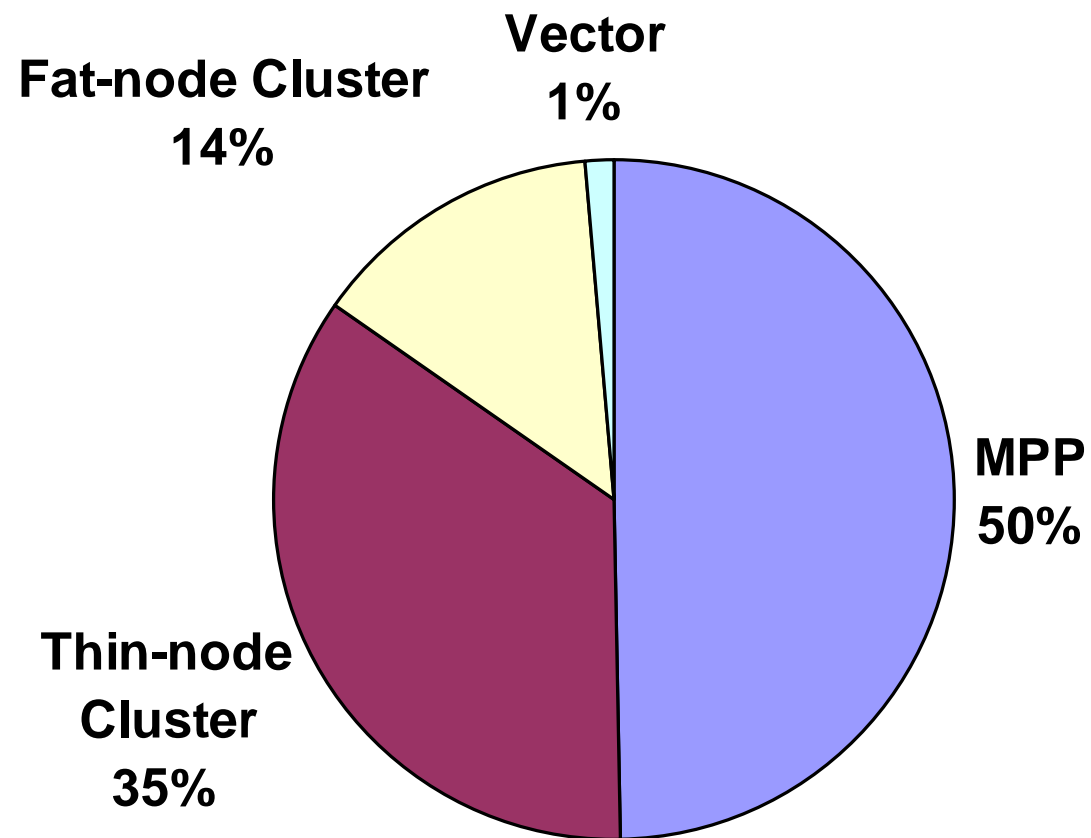
- We asked the PRACE centres to complete:
 - a systems survey for their largest system, and any other system over 10 Tflop/s Linpack
 - an application survey for all applications which consumed more than 5% of the utilised cycles on each system
- We collected data for 24 systems and 69 applications
- Survey was conducted in April 2008
 - data relates to 2007/8



Systems surveyed

System	Centre	Manufacturer	Model	Architecture	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	e325/v40z/x4600 cluster	TNC	1577	1182	330
Totals					926176	675415	169522

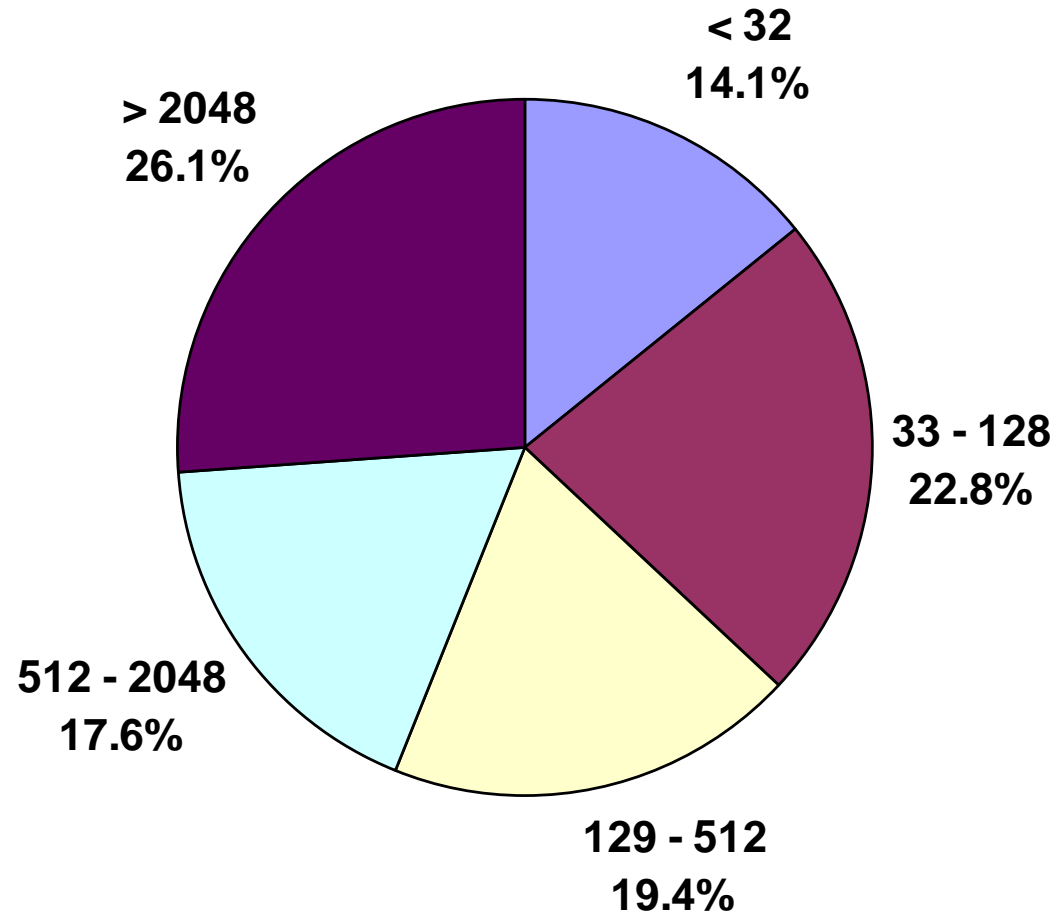
Compute power by architecture type



LEFs

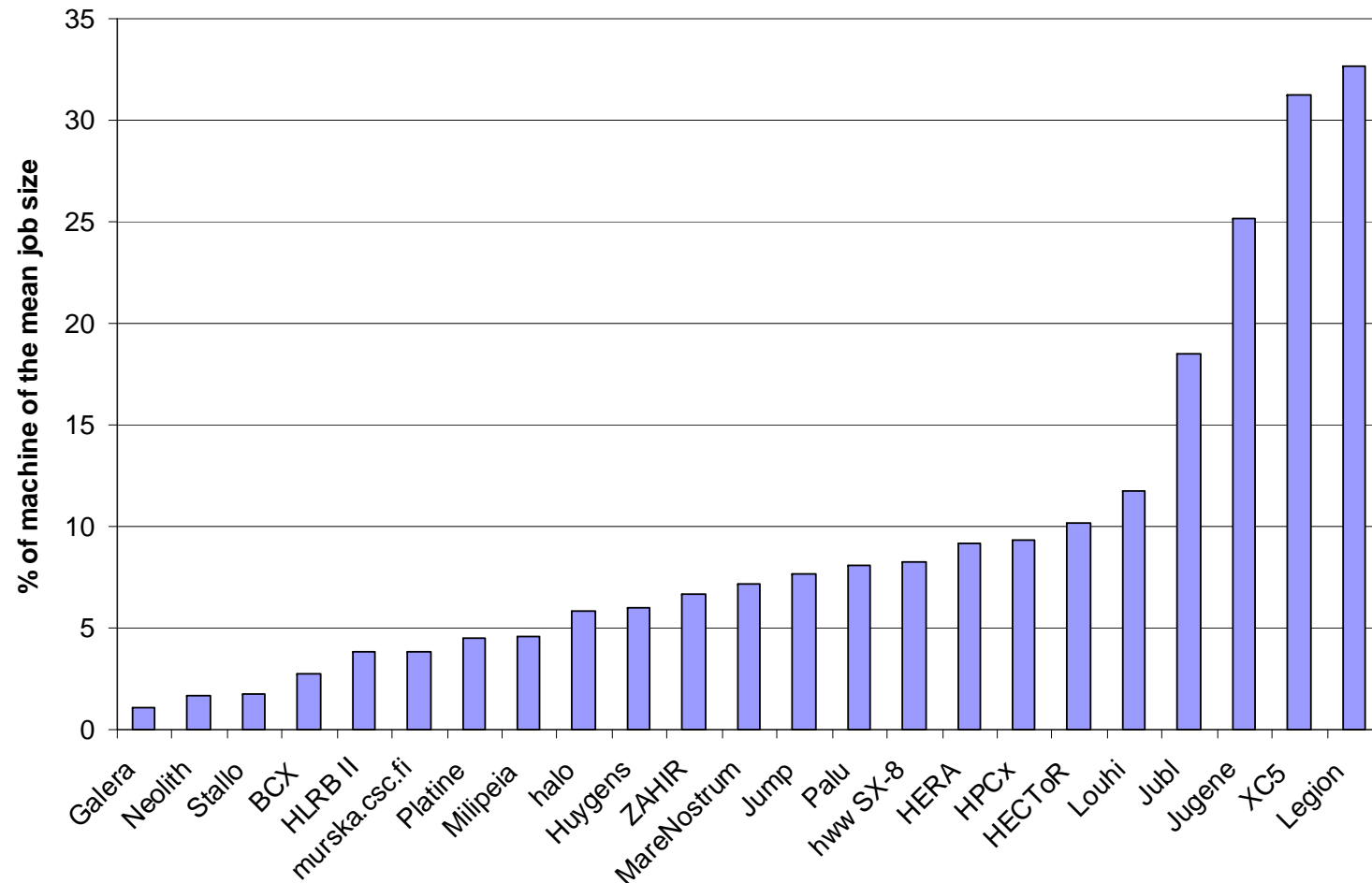
- The measure of computational power and consumed cycles we use is the Linpack Equivalent Flop (LEF).
- A system which has a Linpack R_{\max} of 50 Tflop/s is said to have a power of 50T LEFs
- An application which uses 10% of the time on that system is said to consume 5T LEFs

Distribution of LEFs by job size

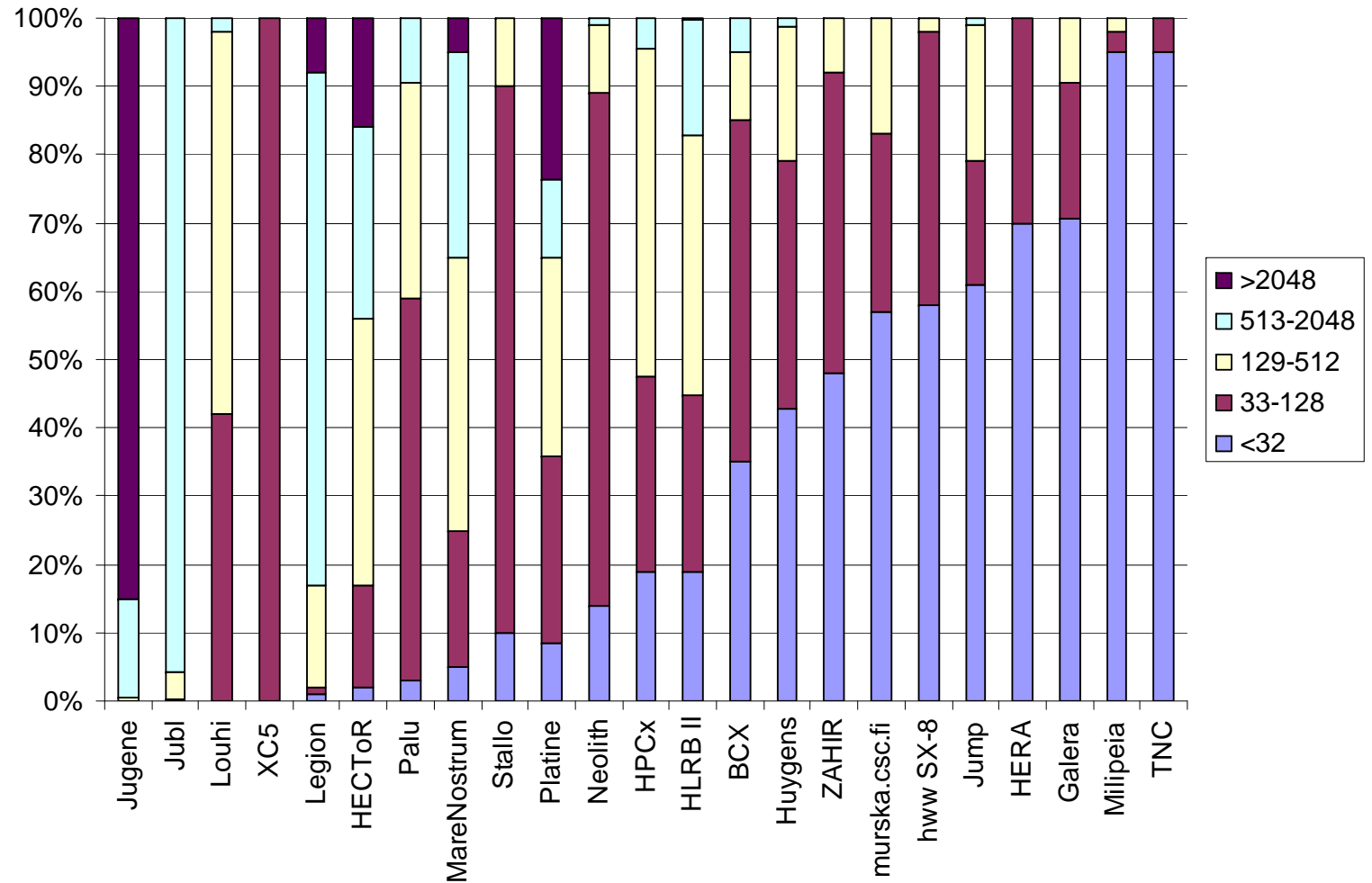




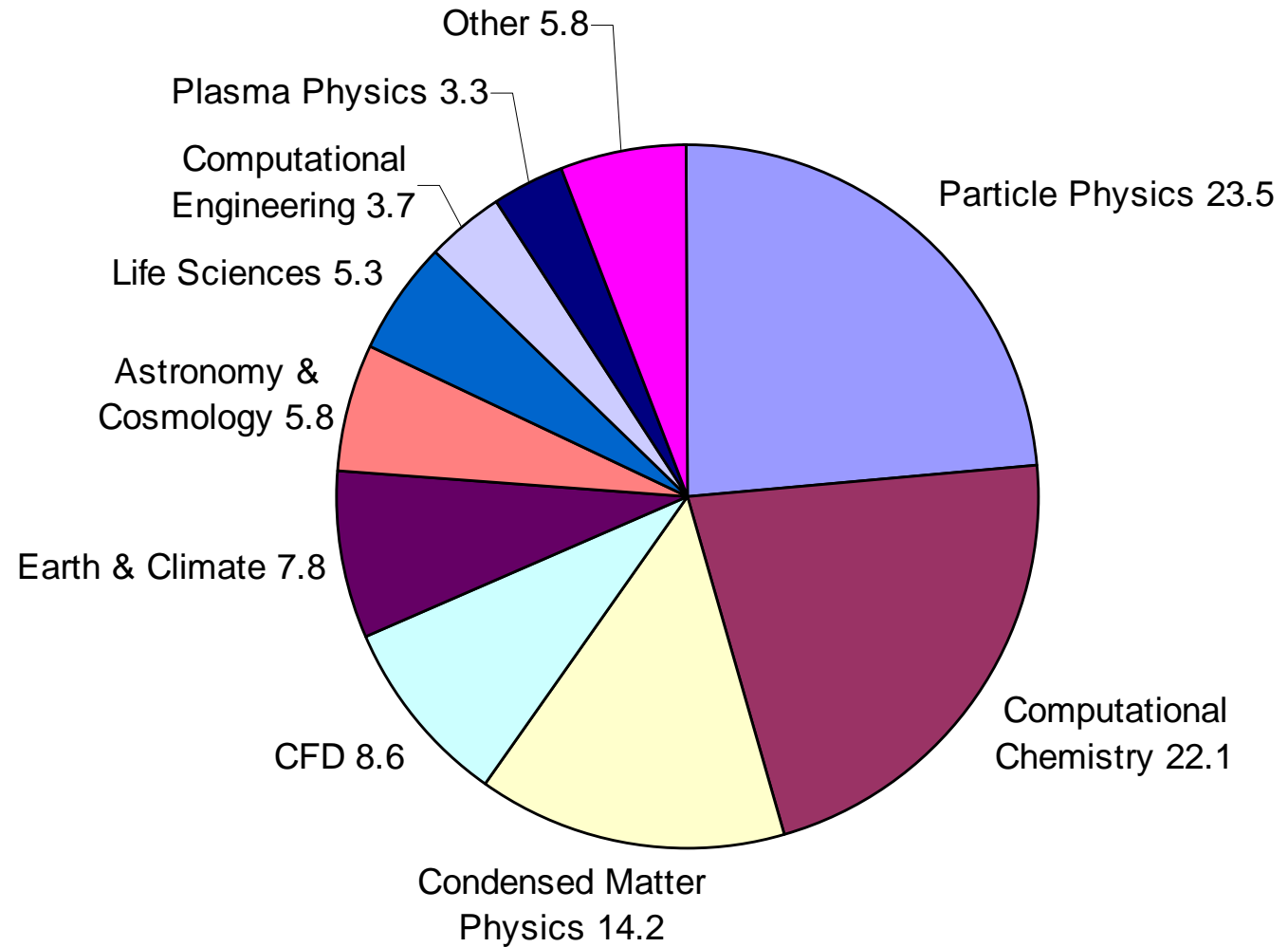
Mean job size as % of machine



Job size distribution by system

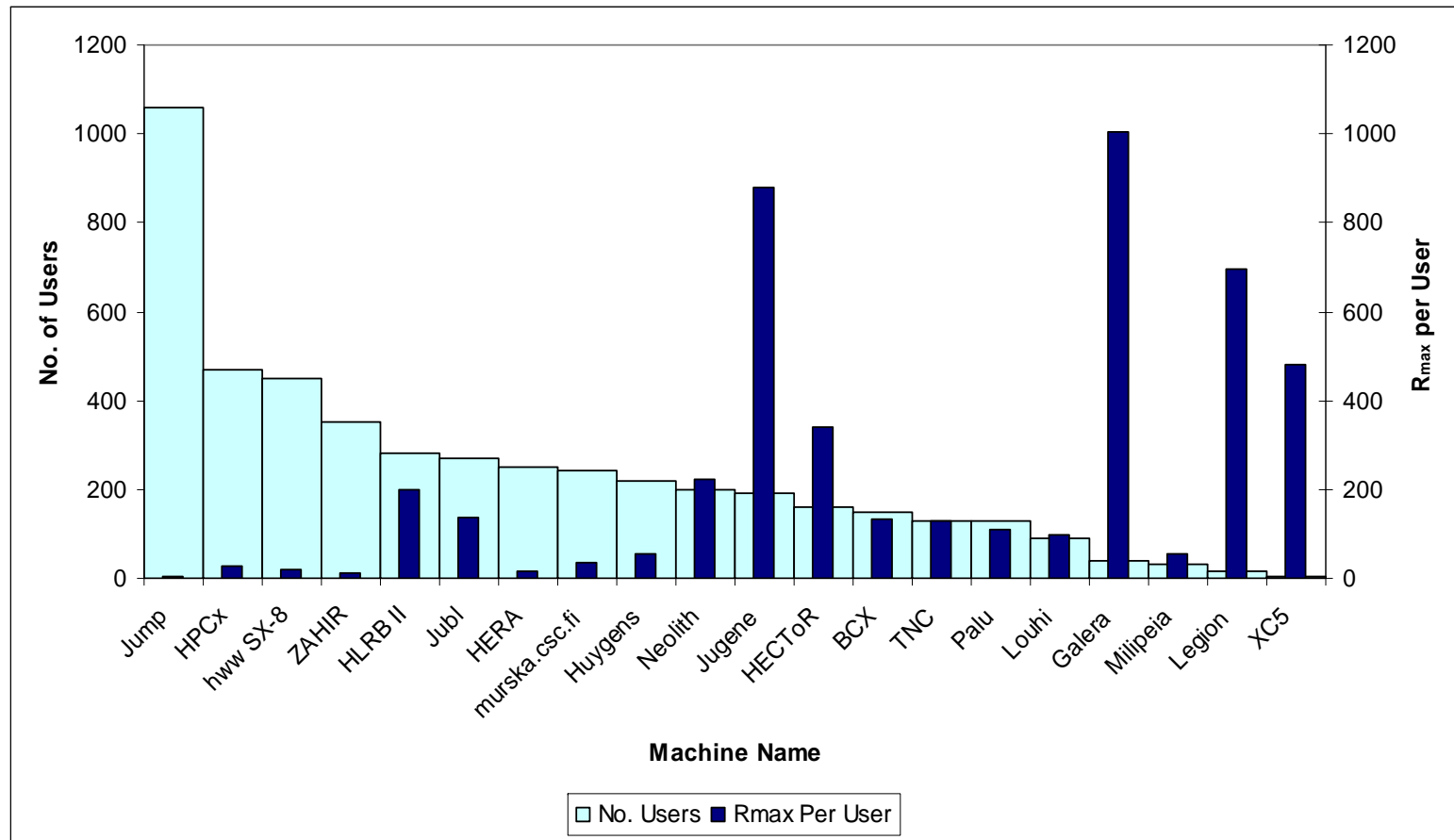


Distribution of LEFs by scientific area





No. of users and Rmax per user



Parallelisation techniques

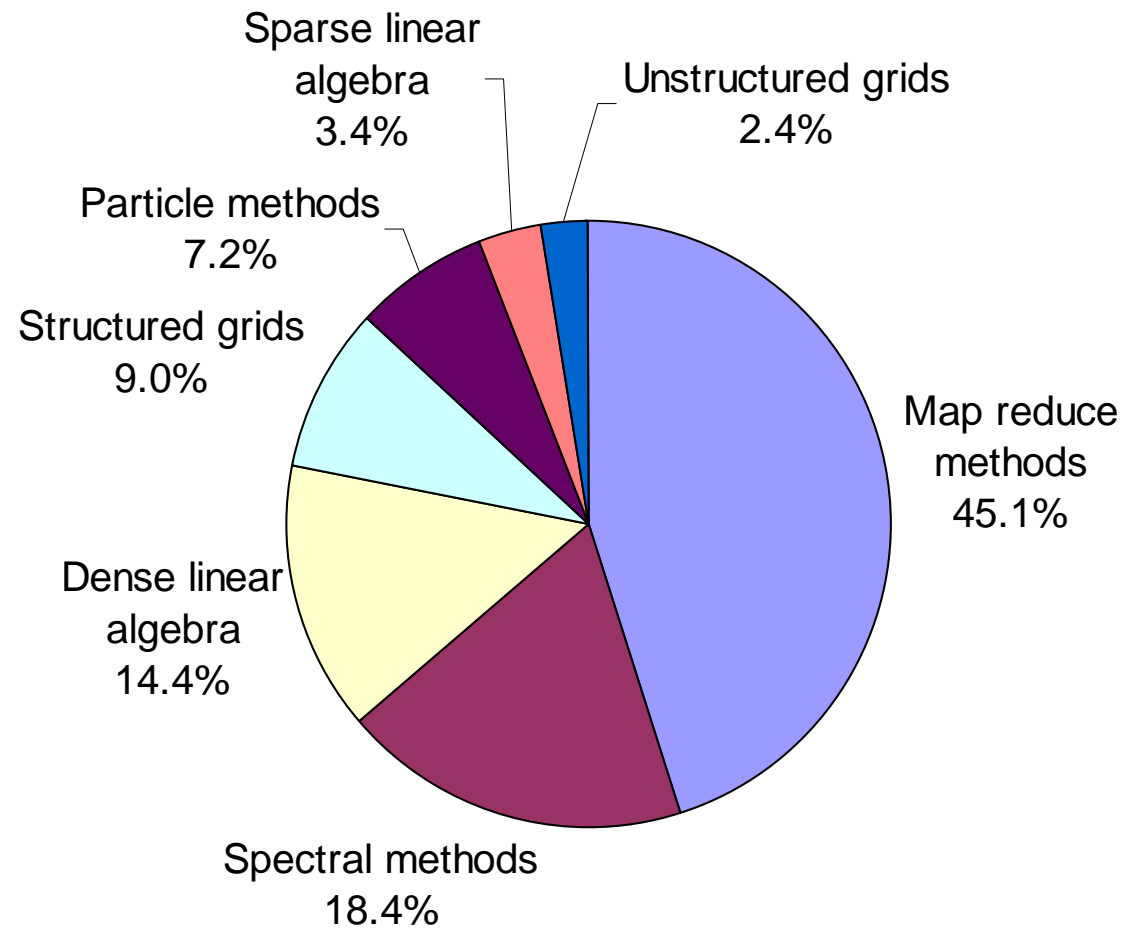
- Of the 69 applications, all but two use MPI for parallelisation
 - exceptions are Gaussian (OpenMP) and BLAST (sequential).
- Of the 67 MPI applications, six also have standalone OpenMP versions and three have standalone SHMEM versions.
- 13 applications have hybrid implementations
 - 10 MPI+OpenMP, 2 MPI+SHMEM, 1 MPI+Posix threads
- Only one application was reported as using MPI2 single sided communication.

Languages

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

- 16 applications mix Fortran with C/C

Distribution of LEFs by dwarves



Distribution of LEFs by dwarf and area

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

Choosing a benchmark suite

- Want to choose a set of applications to form a benchmark suite
 - to be used in the procurement process for Petaflop/s systems
- Suggested process: find a set of applications that is a best fit to the area/dwarf table in the sense that it minimises the norm of

$$\|U\mathbf{w} - \mathbf{v}\|$$

where

\mathbf{v} is a linearised vector containing the table entries

U is a matrix describing the area/dwarf combinations satisfied by the applications

\mathbf{w} is vector of weights



- In principle, one could search all possible lists of applications up to a certain length and find the list with the smallest residual
 - in practise, do a manual search
 - we want to include other criteria, such as usage of applications, geographical spread, etc.
- Gives a quantitative measure of how well a benchmark suite represents current usage
- Also gives a weighting for the applications which could be used to weight benchmark results



Problems with this approach

- Classification of codes into dwarves (and to some extent, areas) is somewhat arbitrary
 - some applications use more than one dwarf: we split the LEFs equally between dwarves
- Bias to recently acquired systems
 - high LEFs
 - recently acquired systems may have atypical usage by early users
- Reflects past, rather than future usage



Current status

- We used the above process as a starting point, then swapped some applications to meet some of the concerns
- 12 core applications, plus 8 additional applications
- Core apps: NAMD, CPMD, VASP, QCD, GADGET, Code_Saturne, TORB, NEMO, ECHAM5, CP2K, GROMACS, N3D
- Additional apps: AVBP, HELIUM, TRIPOLI_4, GPAW, ALYA, SIESTA, BSIT, PEPC



- We have undertaken work to port these applications to the PRACE prototype systems, and optimise them for sequential performance and scalability.
- We are currently collecting benchmark data from the prototype systems which have been installed so far.
- Based on this data, we are reviewing the list of applications to ensure that the final benchmark suite contains scalable codes and avoids licensing problems.



Acknowledgements

- The authors would like to acknowledge all those who contributed by filling in survey forms and taking part in subsequent discussions.
- A full report is available from:
<http://www.prace-project.eu/documents/>



PRACE WP6.1 Systems Survey

[Home](#) | [T6.1 Surveys](#)

Please fill in the information below to the best of your knowledge
Note required fields are denoted by a * symbol

About you

Name: *

Institution: *

Email address: *

PRACE partner name: *

Survey period

Start Date (dd-mm-yyyy): *

End Date (dd-mm-yy): *

Machine details

Please select your system from the following list. If your system is not on the list, please email [Jon Hill](#), who will add the system and let you know when this has happened, so you may complete the survey.

Name: *

Hardware architecture

Machine name:

Manufacturer:

Model:

Processor type:



PRACE WP6.1 Applications Survey

[Home](#) | [T6.1 Surveys](#)

Please fill in the information below to the best of your knowledge
Note required fields are denoted by a * symbol

About you

Name: *

Institution: *

Email address: *

PRACE partner name: *

Survey period

Please note dates should be of the form dd/mm/yyyy.

Start Date (dd/mm/yyyy): *

End Date (dd/mm/yyyy): *

Machine details

Please select your system from the following list. If your system is not on the list, please email [Jon Hill](#), who will add the system and let you know when this has happened, so you may complete the survey.

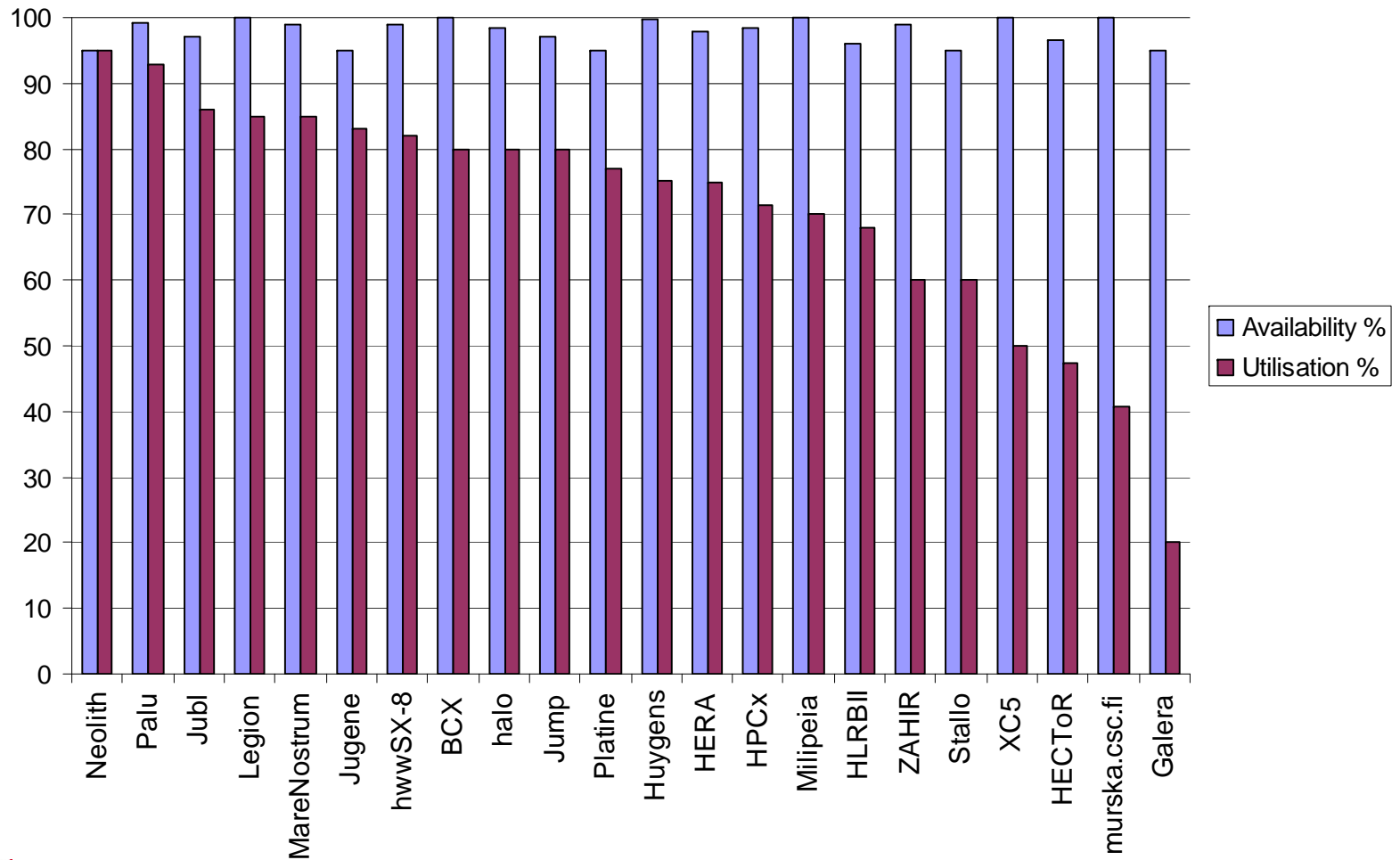
Name: *

About the application

Name:

What is the application's name:

Availability and utilisation



Top 30 applications by usage

Application Name	LEFs Used (Gflop/s)	Number of systems
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