HPC applications

- ✓ **Classes of HPC applications**
  - **The 7/16 dwarfs**
- ✓ **HPC in Europe – PRACE, DEISA, GÉANT2**
  - **PRACE survey**
  - **PRACE prototypes**
  - **PRACE training**
- ✓ **GÉANT3 and FEDERICA ?**

Major HPC Application areas

- 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 "Dwarfs" classes
(the Berkeley scheme)

The dwarfs are those algorithm types which constitute classes where:

- Membership in a class is defined by similarity in computation and data movement
- The dwarfs are specified at a high level of abstraction to allow reasoning about their behaviour across a broad range of applications.
- Programs that are members of a particular class can be implemented differently and the underlying numerical methods may change over time, but the claim is that the underlying patterns have persisted through generations of changes and will remain important into the future.

A dwarf is therefore a grouping of kernels that share both computational and data structure (and they use similar numerical libraries)

http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html

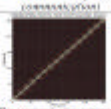
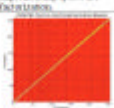
The initial "seven Dwarfs" classes
(the Berkeley scheme)

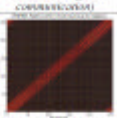
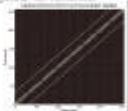
1. **Dense linear algebra** (e.g., BLAS)– data is stored in **dense matrices** or vectors and access is often via unit-level strides. Typical algorithm would be Cholesky decomposition for symmetric systems or Gaussian elimination for non-symmetric systems.
2. **Sparse linear algebra** (e.g., SpMV) – 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.
3. **Spectral methods (e.g., FFT)** – data is in **frequency domain** and requires a transform to convert to spatial/temporal domain. They are typified by, but not restricted to, FFT.
4. **N-body methods** (e.g., Barnes-Hut) – data consists of **discrete particle** bodies that interact with each other and/or the "environment".
5. **Structured grids** (Cactus) – Represented by a **regular grid**. Points on grid are conceptually updated together via equations linking them to other grids. There is high spatial locality. Updates may be in place or between 2 versions of the grid.
6. **Unstructured grid** (e.g., ABAQUS) – 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.
7. **Map reduce methods** (e.g., Monte Carlo)– **embarrassingly parallel** problems, such as Monte Carlo methods, where calculations are independent of each other.

Extensions to the original Seven Dwarfs (the Berkeley scheme)

8. **Combinational Logic** ((e.g., encryption) - Functions that are implemented with logical functions and stored state.
9. **Graph traversal** (e.g., Quicksort) - Visits many nodes in a graph by following successive edges. These applications typically involve many levels of indirection, and a relatively small amount of computation.
10. **Dynamic Programming** - Computes a solution by solving simpler overlapping subproblems. Particularly useful in optimization problems with a large set of feasible solutions.
11. **Backtrack and Branch+Bound** - Finds an optimal solution by recursively dividing the feasible region into subdomains, and then pruning subproblems that are suboptimal.
12. **Construct Graphical Models** - Constructs graphs that represent random variables as nodes and conditional dependencies as edges. Examples include Bayesian networks and Hidden Markov Models.
13. **Finite State Machine** - A system whose behaviour is defined by states, transitions defined by inputs and the current state, and events associated with transitions or states.

The "seven Dwarfs" classes (the Berkeley scheme)

Dwarf	*Description*	*Communication Pattern (Figure axes show processors 1 to 256, with black meaning no communication)*	*NAS Benchmark / Example HW*
1. Dense Linear Algebra (e.g., BLAS [Blackford et al 2002], ScaLAPACK [Blackford et al 1996], or MATLAB [MathWorks 2006])	Data are dense matrices or vectors. (BLAS Level 1 = vector-vector; Level 2 = matrix-vector; and Level 3 = matrix-matrix.) Generally, such applications use unit-stride memory accesses to read data from rows, and strided accesses to read data from columns.	 The communication pattern of MadBench, which makes heavy use of ScaLAPACK for parallel dense linear algebra, is typical of a much broader class of numerical algorithms	Block Tridiagonal / Matrix, Lower Upper Symmetric Gauss-Seidel / Vector computers, Array computers
2. Sparse Linear Algebra (e.g., SpMV, OSKI [OSKI 2006], or SuperLU [Demmel et al 1999])	Data sets include many zero values. Data is usually stored in compressed matrices to reduce the storage and bandwidth requirements to access all of the nonzero values. One example is block compressed sparse row (BCSR). Because of the compressed formats, data is generally accessed with indirect loads and stores.	 SuperLU communication pattern plotted above was the BCSR method for implementing sparse LU factorization.	Conjugate Gradient / Vector computers with gather/scatter
3. Spectral Methods (e.g., FFT [Cooley and Tukey 1965])	Data are in the frequency domain, as opposed to time or spatial domains. Typically, spectral methods use multiple butterfly stages, which combine radix-2 add/subtract operations and a specific pattern of data permutation, with all-to-all communication for some stages and strictly local for others.	 PARATEC: The 3D FFT requires an all-to-all communication to implement a 3D transpose, which requires communication between every link. The diagonal stripe dissipates BLAS-3 dominated linear-algebra step required for orthogonalization.	Fourier Transform / DSPs, Zalluk, PDMP [Zadiuk 2006]

Dwarf	*Description*	*Communication Pattern (Figure axes show processors 1 to 256, with black meaning no communication)*	*NAS Benchmark / Example HW*
4. N-Body Methods (e.g., Barnes-Hut [Barnes and Hut 1986], Fast Multipole Method [Greengard and Rokhlin 1987])	Depends on interactions between many discrete points. Variations include particle-particle methods, where every point depends on all others, leading to an $O(N^2)$ calculation, and hierarchical particle methods, which combine forces or potentials from multiple points to reduce the computational complexity to $O(N \log N)$ or $O(N)$.	 PMEMD's communication pattern is that of a particle mesh Ewald calculation.	(as benchmark) / GRAPE [Tokyo 2006], MD-GRAPE [IBM 2006]
5. Structured Grids (e.g., Cactus [Goodale et al 2003] or Lattice-Boltzmann Magneto-hydrodynamics [LBMHD 2005])	Represented by a regular grid; points on grid are conceptually updated together. It has high spatial locality. Updates may be in place or between 2 versions of the grid. The grid may be subdivided into finer grids in areas of interest ("Adaptive Mesh Refinement"), and the transition between granularities may happen dynamically.	 Communication pattern for Cactus, a PDE solver using 7-point stencil on 3D block-structured grids.	Multi-Grid, Scalar Penta-diagonal / QCDOC [Edinburgh 2006], BlueGene/L
6. Unstructured Grids (e.g., ABAQUS [ABAQUS 2006] or FIDAP [FLUENT 2006])	An irregular grid where data locations are selected, usually by underlying characteristics of the application. Data point location and connectivity of neighboring points must be explicit. The points on the grid are conceptually updated together. Updates typically involve multiple levels of memory indirection, as an update to any point requires first determining a list of neighboring points, and then loading values from those neighboring points.		Unstructured Adaptive / Vector computers with gather/scatter, Tera Multi Threaded Architecture [Berry et al 2006]
7. Monte Carlo (e.g., Quantum Monte Carlo [Aspuru-Guzik et al 2005])	Calculations depend on statistical results of repeated random trials. Considered embarrassingly parallel.	Communication is typically infrequent in Monte Carlo methods.	Embarrassingly Parallel / NSF Terragrid

HPC in Europe

PRACE, DEISA, GEANT2

- **HPC systems, applications, prototypes,training**
- **PRACE objective: Petaflop supercomputers**
  http://www.prace-project.eu
- **DEISA objective: European Virtual Supercomputer**
  http://www.deisa.eu
- **GEANT2 objective: high-bandwidth DEISA interconnect**
  http://www.geant2.net/

PRACE partners

- **Principal Partners**
- France:GENCI – Grand Equipement national pour le Calcul Intensif
- Germany: GCS – GAUSS Centre for Supercomputing
- The Netherlands: NCF – Netherlands Computing Facilities Foundation
- Spain: BSC – Barcelona Supercomputing Center - Centro Nacional de Supercomputación
- UK: EPSRC – Engineering and Physical Sciences Research Council
- **General Partners**
- Austria: GUP – Institut für Graphische und Parallele Datenverarbeitung der Johannes Kepler Universität
- Finland: CSC – The Finnish IT Center for Science
- Greece: GRNET – Greek Research and Technology Network
- Italy: CINECA – Consorzio Interuniversitario
- Norway: UNINETT Sigma AS
- Poland: PSNC – Poznan Supercomputing and Networking Center
- Portugal: UC-LCA – Universidade de Coimbra – Laboratório de Computação Avançada
- Sweden: SNIC – Swedish National Infrastructure for Computing
- Switzerland: ETH Zurich - Swiss Federal Institute of Technology Zurich, CSCS – Swiss National Supercomputing Centre
- **Additional General Partners of the PRACE Initiative**
- Countries that have signed the PRACE Memorandum of Understanding:
- Cyprus: The Computation-based Science and Research Center (CSTRC)
- Ireland: Irish Centre for High-End Computing
- Serbia: The Institute of Physics, Belgrade
- Turkey: National Center for High Performance Computing

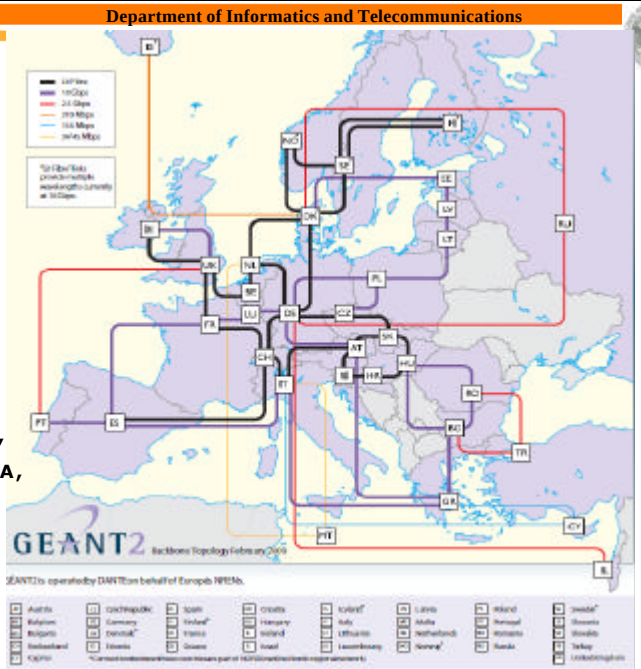
DEISA partners

GÉANT2

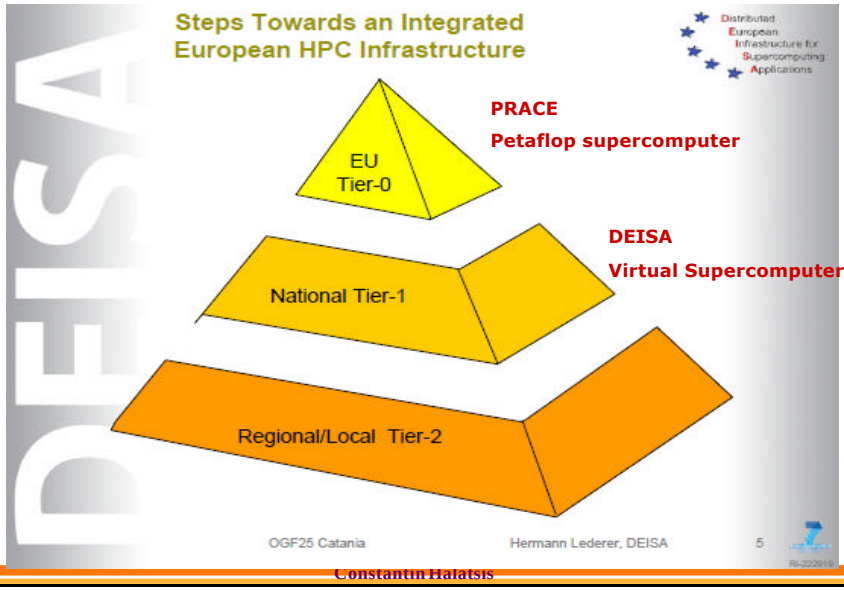
A few Facts & Figures…

- **29 POPs**
- **serving 30 NRENs**
- **~12000 km of fibre**
- **>120 (own) 10G lambdas**
- **22 (leased) 10G lambdas**
- **+ some lower speed links**
- **NREN accesses at up to 10Gbps (+ backup) + P2P**
- **4 x 10G to North America**
- **POP in NY**
- **connections to other R&E networks…**
  Abilene (Internet2), ESnet, CA*net4, SINET, TENET, EUMEDCONNECT, RedCLARA, TEIN2, India

- **GEANT3 Next**



3-Tier architecture



Tier0 / Tier1 Top Layer of the HPC Ecosystem

Distributed European Infrastructure for Supercomputing Applications

T0 : future shared petascale European systems
T1 : leading national supercomputing systems

PRACE
Designing an infrastructure that will enable the operation of shared petascale European systems
Enhancing performance in selected sites and providing wide access to shared systems

In the DEISA-2 environment, scientists will naturally have access to several distributed platforms, and shared systems will have to access remote data repositories. We will be in a situation similar to TeraGrid.

Member states sites

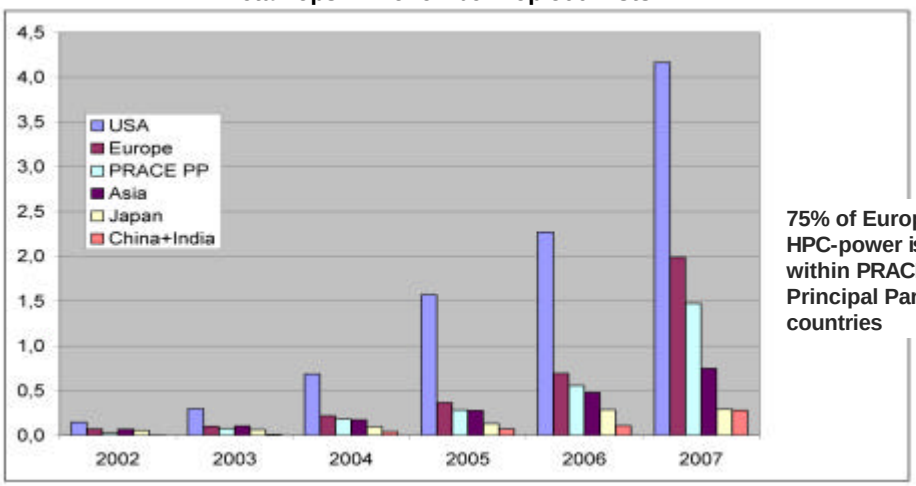
DEISA-2 : strong integration of T0 and T1 systems (automatically provides wide, seamless and efficient access to shared systems and data repositories)

The DEISA-1 services have been tailored for this mode of operation. There is a positive feedback between the two orthogonal lines of action:
- *DEISA is paving the way to the efficient operation of T0 systems.*
- *T0 systems will drive the massive adoption of the DEISA services.*

OGF25 Catania Hermann Lederer, DEISA 7

Europe's current position in HPC

Aggregated LINPACK Performance in PetaFlops in November Top 500 Lists



75% of European HPC-power is within PRACE Principal Partner countries

The PRACE survey utilisation matrix

(distribution of applications into the 7 dwarfs)

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

- **The utilisation matrix is made up of 70 categories; 10 scientific areas and 7 algorithmic "dwarfs". 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.**
- **A dwarf is a grouping of kernels that share both computational and data structure.**

PRACE HPC systems

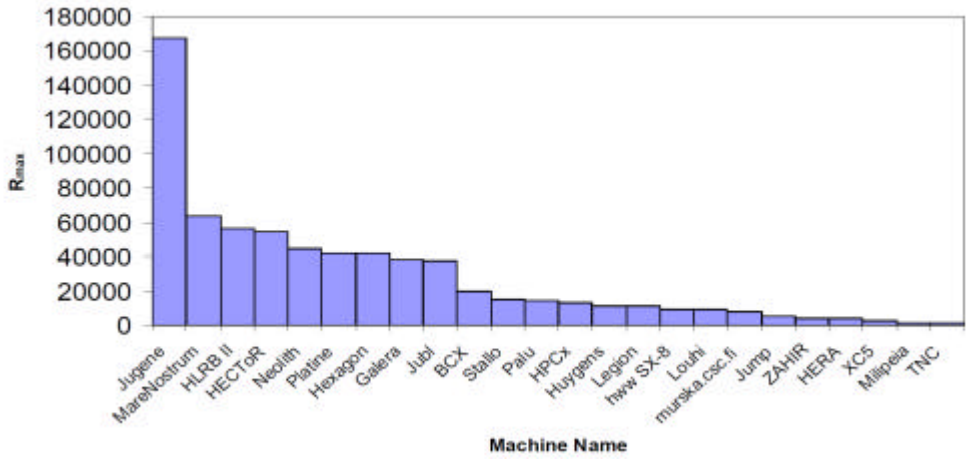
Name	Centre	Manufacturer	Model	Architecture type	R_{peak}	R_{max}	Cores
Jugene	FZJ	IBM	Blue Gene/P	MPP	222822	167300	65536
MareNostrum	BSC	IBM	JS21 cluster	TNC	94208	63830	10240
HLRB II	BADW-LRZ	SGI	Altix 4700	FNC	62259	56520	9728
HECToR	EPSRC	Cray	XT4	MPP	63437	54648	11328
Neolith	SNIC	HP	Cluster 3000 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[7]	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.

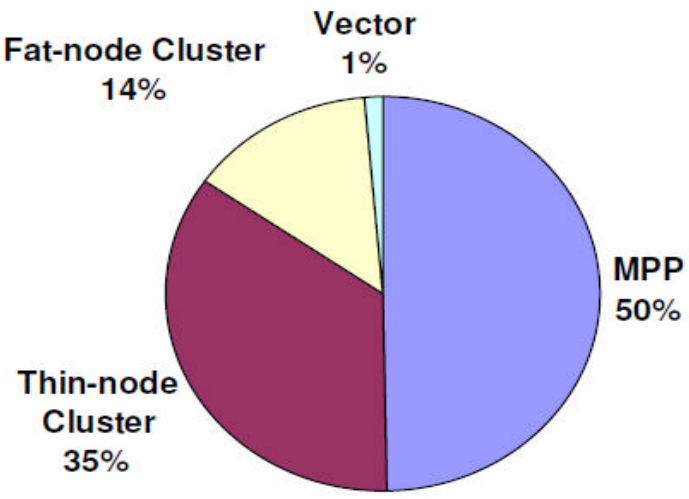
PRACE HPC systems in last Top500 list

Rank	Site	System	Cores	R_{max}	R_{peak}
11	Forschungszentrum Juelich (FZJ) Germany	Blue Gene/P Solution IBM	65536	180	222.82
40	Barcelona Supercomputing Center Spain	BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet IBM	10240	63.83	94.21
44	Leibniz Rechenzentrum Germany	Altix 4700 1.6 GHz SGI	9728	56.52	62.26
46	University of Edinburgh United Kingdom	Cray XT4, 2.8 GHz Cray Inc.	11328	54.65	63.44
55	National Supercomputer Centre (NSC) Sweden	Cluster Platform 3000 DL140 Cluster, Xeon 53xx 2.33GHz Infiniband Hewlett-Packard	6440	47.03	60.02
63	Commissariat a l'Energie Atomique (CEA) France	Novascale 3045, Itanium2 1.6 GHz, Infiniband Bull SA	7680	42.13	49.15
65	University of Bergen Norway	Cray XT4 QuadCore 2.3 GHz Cray Inc.	5550	40.59	51.06
67	Gdansk University of Technology, CI Task Poland	ACTION Cluster Xeon E5345 Infiniband ACTION	5336	38.17	49.73
69	Forschungszentrum Juelich (FZJ) Germany	eServer Blue Gene Solution IBM	16384	37.33	45.88

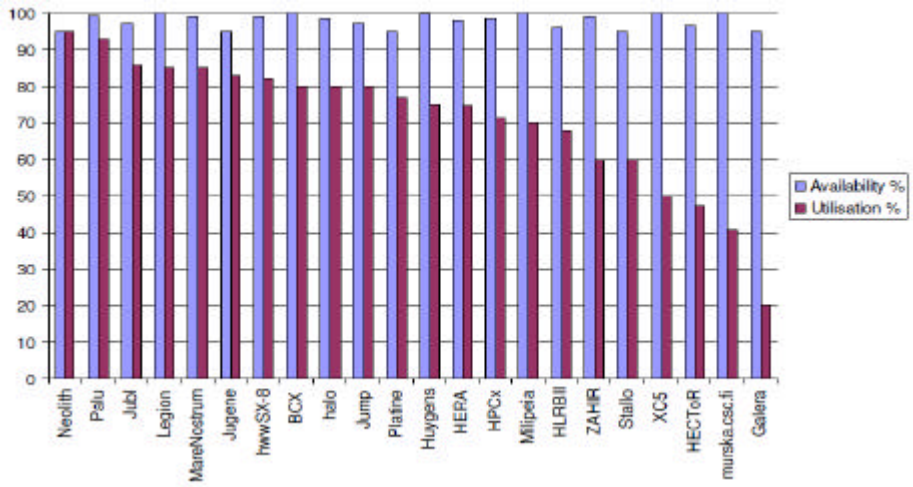
Performance per system (in Tflops)



Total Compute power by architecture type

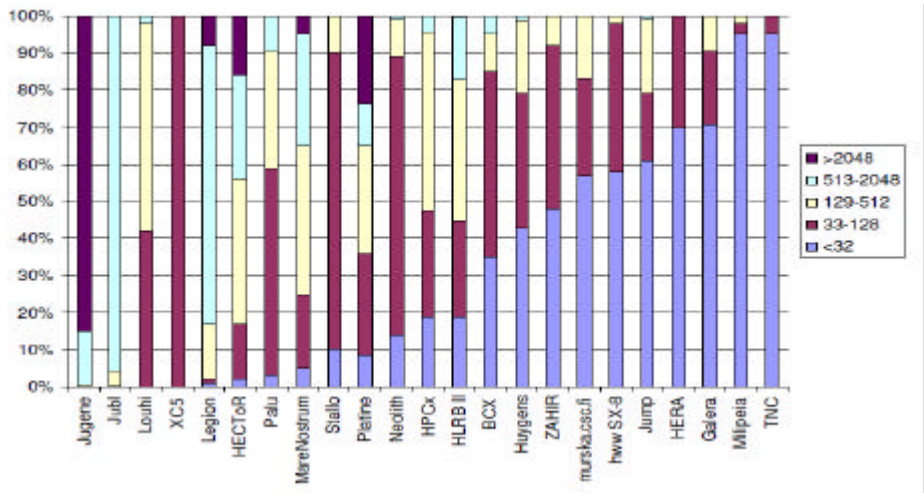


System availability and utilisation

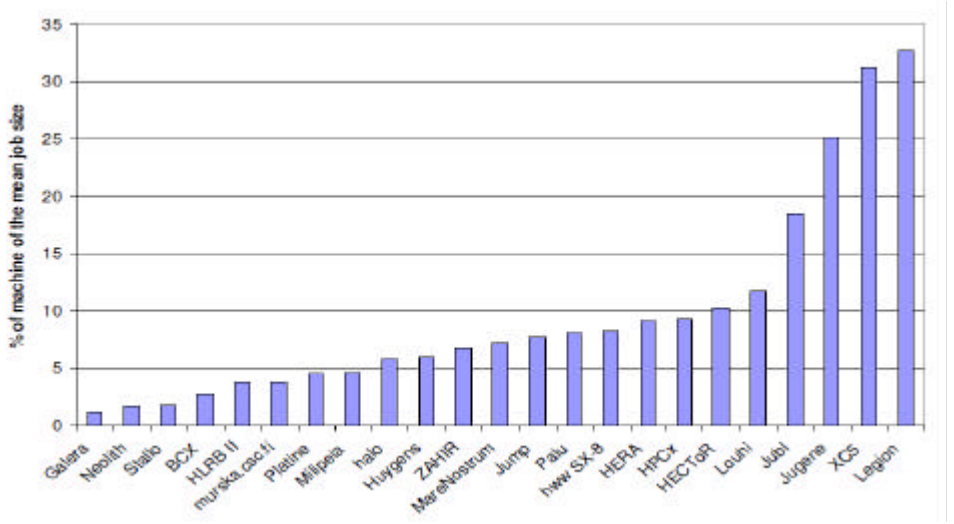


Job size distribution by system

Job size in cores

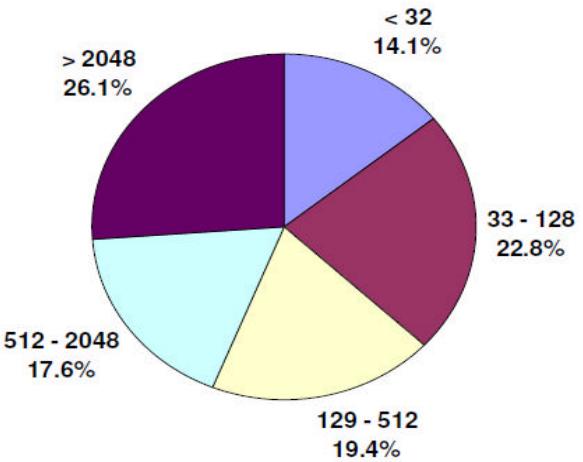


Mean job size as a % of system size

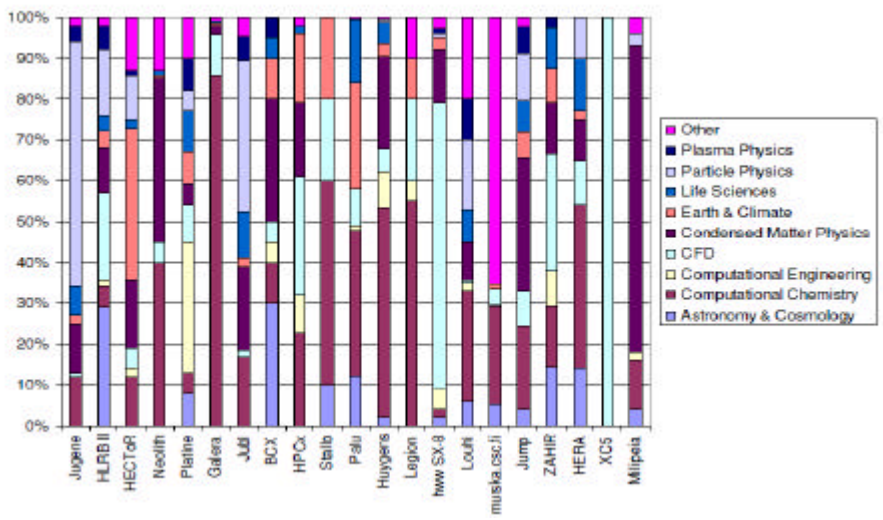


Aggregate distribution of LEFs by job size

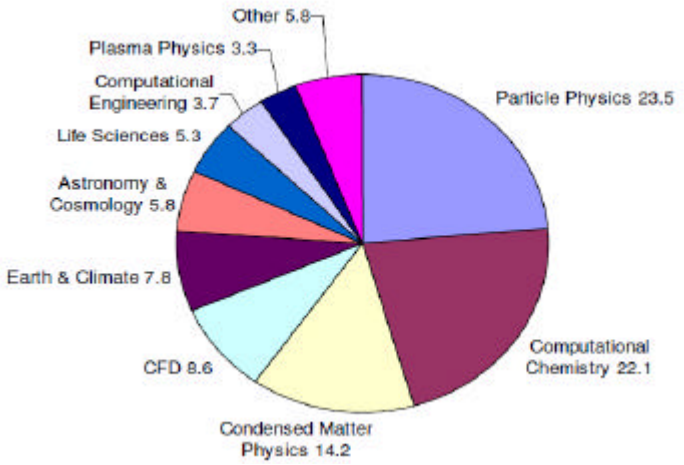
(LEF = LINPACK Equivalent Flop/s)



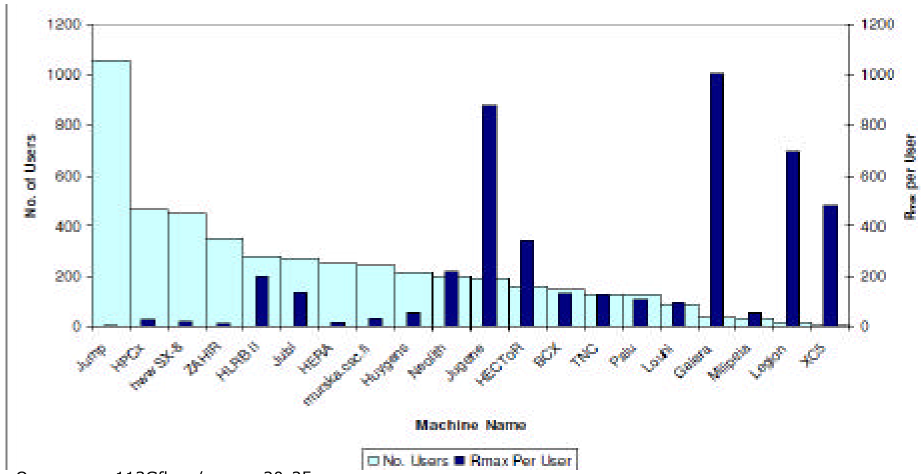
Scientific area distribution by system



Aggregate distribution of LEFs by scientific area



Number of users and R_{max} per user



On average 113Gflops/user = 20-25 cores

Application name

Application Name	LEFs Used (Gflop/s)	Number of systems using this code
overlap and wilson fermions	54923	2
vasp	35766	9
lqcd (twisted mass)	25807	2
lqcd (two flavor)	12393	2
namd	10035	4
dalton	9875	3
cpmd	9688	3
gadget	8412	2
dynamical fermions	7847	1
spintronics	5208	2
materials with strong correlations	4846	2
dl_poly	4779	2
casino	4223	1
quantum-espresso	3982	1
cactus	3708	1
trio_u	3202	1
smmp	3181	2
tfsplano	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

hellium	2248	1
magnum	1598	1
pdkgrav-gasoline	1233	1
crystal	1193	2
su3_ahiggs	1183	1
ukd	1060	1
unified model	992	1
cosmic model	936	1
iamie	884	2
parallel particle mesh (ppm) library	803	1
modelpar	728	1
code stagger code	715	1
pmcm	615	1
ardip	600	5
turbomole	574	1
octopus	480	1
pencil code	435	1
occam	433	1
high resolution computational of local dissipation scales	432	1
climber	422	1
gyaw	368	2
leaflow	353	1
unified cmcp model	295	1
molpro	238	1
metallic layers, electronic and magnetic phenomena	216	1
gaussian	134	1
blast	111	1
iqcs	0	2
siesta	0	2
climer	0	2
alya	0	1
swh	0	1
ccsm	0	1
latflow	0	1
mplot	0	1
espresso	0	1
bats	0	1
cosmos	0	1
vsgr	0	1
code_saturne	0	1
gamess-uk	0	1
fluent	0	1

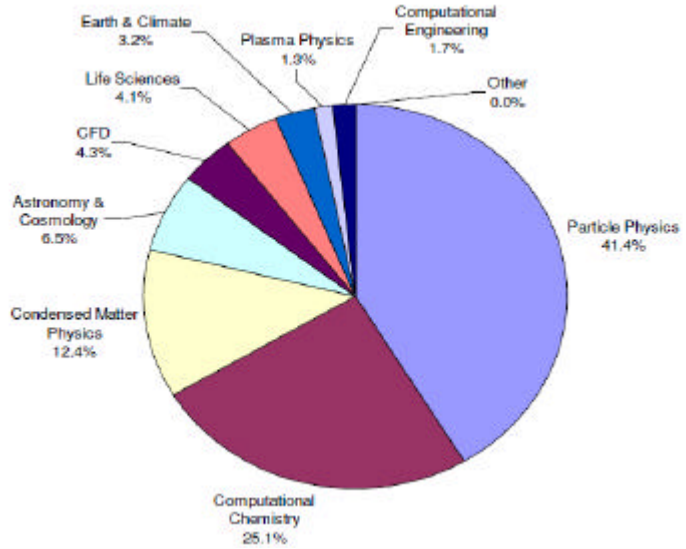
Base language usage by applications

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

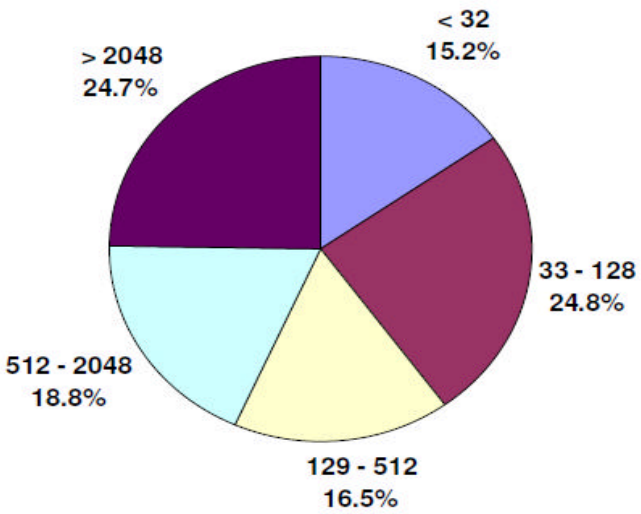
Parallelisation

- 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 stand alone 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.

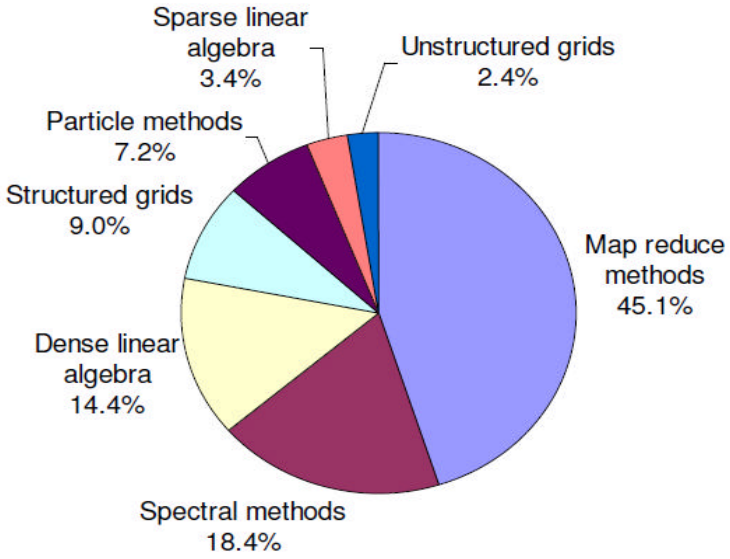
Distribution of applications usage by scientific area



Distribution of applications usage by job size



Distribution of applications usage by algorithmic dwarves



Usage matrix from the surveys

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

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

List of applications in each of the scientific area/dwarf category

Area/Dwarf	Dense linear algebra	Spectral methods	Structured grids	Sparse linear algebra	Particle methods	Unstructured grids	Map reduce methods
Astronomy and Cosmology							
Computational Chemistry							
Computational Engineering							
Computational Fluid Dynamics							
Condensed Matter Physics							
Earth and Climate Science							
Life Science							
Particle Physics							
Plasma Physics							
Other							

Core list of applications to be included in the PRACE benchmarks for the 0-Tier prototypes

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
	Perform 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
nemo	Earth and climate sciences
	Ocean modelling
	http://www.lodyc.jussieu.fr/NEMO/
	http://www.mpimet.mpg.de/en/wissenschaft/modelle/echam/echam5.html

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.

Computational chemistry and condenses Matter Physics are the two main areas

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.

It seems that a future 0 Tier Petaflop/s system might run a large number of map-reduce type jobs, particularly in the area of particle physics.

PRACE prototypes (0 Tier)

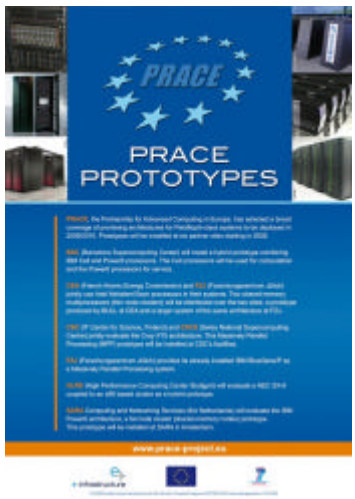
Site	Architecture Vendor/Technology	Point of contact
FZJ Germany	MPP IBM BlueGene/P	Michael Stephan m.stephan@fz-juelich.de
CSC-CSCS Finland+Switzerland	MPP Cray XT5/XTn - AMD Opteron	Janne Ignatius janne.ignatius@csc.fi Peter Kunszt peter.kunszt@cscs.ch
CEA-FZJ France+Germany	SMP-TN Bull et al. Intel Xeon Nehalem	Gilles Wiber gilles.wiber@cea.fr Norbert Eicker n.eicker@fz-juelich.de
NCF Netherlands	SMP-FN IBM Power 6	Axel Berg axel@sara.nl Peter Michielse michielse@nwo.nl
BSC Spain	Hybrid – fine grain IBM Cell + Power6	Sergi Girona sergi.girona@bsc.es
HLRS Germany	Hybrid – coarse grain NEC Vector SX/9 + x86	Stefan Wesner wesner@hlrs.de

Call for access to prototypes

http://www.prace-project.eu/prototype-access

The general conditions for prototype access are:

- Prototypes are intended for testing and not for standard production work.
- As a general rule application support (e.g. code development, optimisation, parallelisation, etc.) during prototype testing is limited.
- Prototype access is granted after technical review by the PRACE centre hosting the prototype.
- The maximum project duration is 3 months.
- Applicants must agree that applicants' names and affiliations, as well as a summary of the project purpose and the results achieved during prototype testing, may be made publicly available in the PRACE website and may be used in PRACE deliverables and other PRACE documents.
- Applicants must agree to summarise the results of the project in an end report to be sent to the PRACE centre hosting the prototype not later than 1 month after project conclusion.
- The project leader is responsible for signing and fulfilling the usage agreements issued by the PRACE centre hosting the prototype.
- Proposals are accessed by the PRACE Prototype Access Committee with representatives from the prototype host centre and representatives from PRACE WP5 and WP6.
- Applicants have the right to reply to the granting decisions. Application replies will be handled by the PRACE Technical Board.



FZJ prototype (Germany)

FZJ

- Specific features:
  - Access to a large existing MPP system, already 1/4 PF with an architecture expandable to 1 PF
- Contribution to the PRACE project:
  - Application scaling, optimization and benchmarking including:
    - Communications
    - I/O
  - Large scale operations on selected applications
  - Assessment of electrical power usage
- Availability July 2008

FZJ

MPP IBM BG/P

16 racks
16k nodes
64k cores (PPC 450)

223 TF peak

CSC/CSCS prototype
(Finland, Switzerland)

CSC / CSCS

- Specific features:
  - Prototype installed in CSC, joint effort with CSCS
  - Funding goes fully for the dedicated system
  - As additional in-kind contribution access to a larger existing system – similar architecture
- Contribution to the PRACE project:
  - Access to a prototype with MPP architecture and fast processors :
    - AMD Opteron, SeaStar2+ 3d-torus network
  - Early access to AMD new generation of processors: all processors (Barcelona) replaced by Shanghai (XTn)
  - Additional focus on hybrid MPI/OpenMP parallel programming by CSCS
  - Capability testing on the CSC XT system
- Availability December 2008

MPP - XT5 (/XTn)
180 compute nodes
3 serv./IO/login blades+disk
1440 compute cores:
AMD Barcelona->Shanghai
SeaStar2+ 3d-torus network
ca. 14 TF

CSC
CSC

MPP XT4/XT5

1684 compute nodes
9424 compute cores
SeaStar2+ 3d-torus network
86.7 TF

CEA/FZJ prototype
(France, Germany)

CEA / FZJ

- Specific features:
  - Combination of dedicated test system at CEA and a large production system of same architecture at FZJ
- Contribution to the PRACE project
  - Early access to a prototype of a new product designed by BULL (pre GA machine)
    - High density blade based system with new Intel Nehalem processors
    - Optimized for HPC
    - Scalable to Petaflop/s
    - Water cooling
  - Scalability testing on the FZJ part of the prototype
- Availability : March 2009

Cluster of Thin-Nodes
Dual socket Nehalem nodes
Quadrics QsNetIII
250 TB / Lustre
128 compute nodes
1024 cores
Prototype of future BULL product

CEA ⇕ FZJ

Cluster of Thin-Nodes
Dual socket Nehalem nodes

2048+1024 compute nodes
16384+8192 compute cores
200+100 TF JUROPA2

NCF prototype (Netherlands)

NCF

- Specific features:
  - Large shared memory (4-8 GB/core) and fast I/O configuration
- Contribution to the PRACE project:
  - Access to the new IBM Power6 processors and IBM Power Cluster fat node architecture
  - Focus on HPC software from US DARPA/PERCS research addressing specific petascale issues– early access
  - Specific test nodes for aggressive experimentations
  - Capability testing and assessments on large production system
  - Very high density (>50kW/rack), water cooled nodes
- Availability October 2008

Specific test nodes
Power 6

DARPA software environment and programming tools

NCF ⇕ NCF

Fat node SMP
IBM Power 6

104 nodes
3328 cores

60 TF

Barcelona SC prototype (Spain)

BSC

- Specific features:
  - Dedicated "fine grain" hybrid system
- Contribution to the PRACE project:
  - New Power 6 + Cell processor integration
    - Comparable to US PF RoadRunner but different CPUs
  - Programming techniques and tools for CPU+accelerators
  - Operation of an hybrid system: queuing system, file system, accounting, system administration, ..
  - Assessment of electrical power usage
- Availability December 2008

Hybrid
IBM Cell + Power6

12 P6 + 72 Cell blades
48 + 1296 CPUs

14 TF

HLRS prototype (Germany)

HLRS

- Specific features
  - Unique "System of Systems" concept
    - Multi-physics / multi-scale apps on optimized hardware
    - Hybrid configuration Vector (SX9) + Scalar (Nehalem)
    - Highly innovative configuration
    - Expandable (e.g. with Cell, GPU, FPGA, …)
    - Shared file system and heterogeneous network
  - Concept enables industry-related applications
- Contribution to the PRACE project:
  - New Programming models and methods
  - Close collaboration with vendor (joint Linux OS porting)
  - Necessary intermediate step towards new hybrid systems (more tightly coupled)
  - Specific I/O and network challenges can be investigated
- Availability March 2009

NEC SX-9 vector part
4-8 nodes*
64-128 cores*

6.5-13 TF

X86-64 scalar part
Dual Socket Intel Nehalem
64-512 nodes*
512-4096 cores*

6.1-50 TF

Second set of PRACE prototypes - 1

- **CINECA** (HPC consortium of 36 universities, Italy) will address *metadata performance of I/O subsystem* solutions for petascale machines on an HP XC4 prototype and study NFS and pNFS over RDMA. The usage of SSD based technology to accelerate metadata performance will be tested with Lustre, NFS and pNFS and compared with traditional disks technology based solutions.
- **EPSRC-EPCC** (Edinburgh Parallel Computing Centre, UK) within the *FPGA* High Performance Computing Alliance (FHPCA) will evaluate porting efforts and the ratio of performance versus power consumption of PRACE benchmarks on their *"Maxwell" FPGA prototype supercomputer.*

Second set of PRACE prototypes - 2

- **ETHZ-CSCS** (Swiss National Supercomputing Centre, Switzerland) will study new parallel programming *paradigms like the Partitioned Global Address Space (PGAS)* programming approaches (Co-Array Fortran, UPC) and the upcoming DARPA High Productivity Computer System (HPCS) language (like Cray's Chapel) on a 3328 cores Cray XT3 system.
- **FZJ** (Forschungszentrum Jülich, Germany) will provide a power efficient *special-purpose architecture called eQPACE for lattice Quantum Chromodynamics (QCD*). This 25.6 TFlop/s peak performance prototype is based on *IBM PowerXCell 8i* processors and a custom 3d-torus interconnect implemented within FPGAs supporting presently only nearest-neighbor communication. One of the main goals will be to extend the concept to general all-to-all communication.

Second set of PRACE prototypes - 3

- **GENCI-CEA** (Commissariat à l'Energie Atomique, France) offers a hybrid system composed by *nVIDIA Tesla S1070* coupled with BULL Novascale R425 systems. The purpose of this prototype will be to evaluate different programming environments like CUDA, HMPP (from CAPS Entreprise), OpenCL and the GPU aware version of Allinea's DDT debugger.

- **BAdW-LRZ** (Leibniz Supercomputing Centre, Germany) will assess the productivity of the new *data stream parallel programming language RapidMind* on x86 multicore systems and multiple accelerators (nVIDIA and AMD/ATI GPUs, IBM Cell and Intel Larrabee).

Second set of PRACE prototypes - 4

- **BAdW-LRZ** and **GENCI-CINES** (Centre Informatique National de l'Enseignement Supérieur, France) will install a joint prototype in Garching and Montpellier based on SGI thin nodes (ICE system) and fat nodes (UltraViolet) coupled with *Clearspeed/PetaPath (e710 boards) and Intel Larrabee GPUs*. The two partners are planning to *evaluate novel hardware* (Intel Nehalem-EP/EX processors, NUMAlink5 and 4X QDR Infiniband networks, Clearspeed/PetaPath accelerators and Intel Larrabee manycore GPUs) as well as software components (*Lustre filesystem*) on synthetic benchmarks and real applications.

- **NCF-SARA** (Stichting Academisch Rekencentrum Amsterdam, Netherlands) will install a compact hybrid system composed by 12 standard Intel compute nodes coupled with 12 *ClearSpeed/PetaPath* CATS 700 units primarily dedicated to the *evaluation of large-scale applications* (astrophysics, iterative solvers, geophysics and image analysis). Comparisons with GPU based versions of some applications are planed as well as collaboration with the joint BAdW-LRZ / GENCI-CINES prototype.

PRACE training
immediate requirements

- **Training in mixed-mode (hybrid) programming**
  - Two-thirds of respondents indicated that they had no competency in hybrid programming techniques
- **Increased awareness of Partitioned Global Address Space PGAS) languages**
  - 90% of respondents are unfamiliar with this approach to parallel programming
- **Information on latest developments in HPC and parallel programming**
  - 80% of respondents consider themselves inadequately informed
- **High quality training material on visualization**
  - The vast majority of respondents rated the training they had received in visualization tools and techniques as 'poor'
- **Promotion of the use of numerical libraries**
  - Proficiency and awareness of numerical libraries as rated as 'low' by most users

PRACE training
short-term requirements-1

- **Formal training courses in modern Fortran programming**
  - Fortran was considered the most important traditional language
  - Courses should cover the more sophisticated constructs and promote modern software engineering principles
- **Training material on code optimization, debugging and code testing**
- **Training material on parallel I/O**
  - Two-thirds of respondents are using parallel I/O but the vast majority have little or no knowledge in this field

PRACE training
short-term requirements-2

- **Training in multi-core programming**
  - Just one-fifth of respondents had received training in programming for multi-core architectures
  - The vast majority of respondents consider they have no proficiency in topics such as multi-core cache optimization, and multi-core memory and bandwidth management
- **Training in standard Unix/Linux and HPC system skills**
  - 70% of respondents would benefit from training material in areas such as Unix environments, batch systems, and development tools including version control systems

PRACE training
long-term requirements

- **Increased awareness of developments in next generation programming languages**
  - 90% of respondents are unfamiliar with next-generation HPC programming languages such as Chapel, X10 and Fortress
- **Training in scripting languages**
  - There is an increased uptake of scripting languages (particularly for providing the glue within workflows)
  - 60% of respondents considered the need for training in scripting languages as either somewhat or very important
- **Training in distributed computing and Grid tools**
  - The emergence of Grid technology as a fundamental component of a distributed computing infrastructure requires that users be familiar with Grid tools to enable exploitation of the PRACE HPC Research Infrastructure

PRACE training
general requirements

- **Face-to-face training was considered the most useful channel for delivering training**
  - Training must be given by experts in the HPC field, a factor that would influence users' willingness to attend
- **Existing training material must be improved**
  - 90% of respondents consider that training material is inadequate
  - No training course was designated as "excellent" by any user
  - 50% of users aren't adequately served by their local HPC centre
    - PRACE should work with local centres to ensure training needs and expectations are met for all users
  - 95% of respondents consider that they would benefit from a pan-European centralized training repository
    - Material must be world-class quality and regularly updated