PRACE
Brief Overview and Background to Accelerator Interest

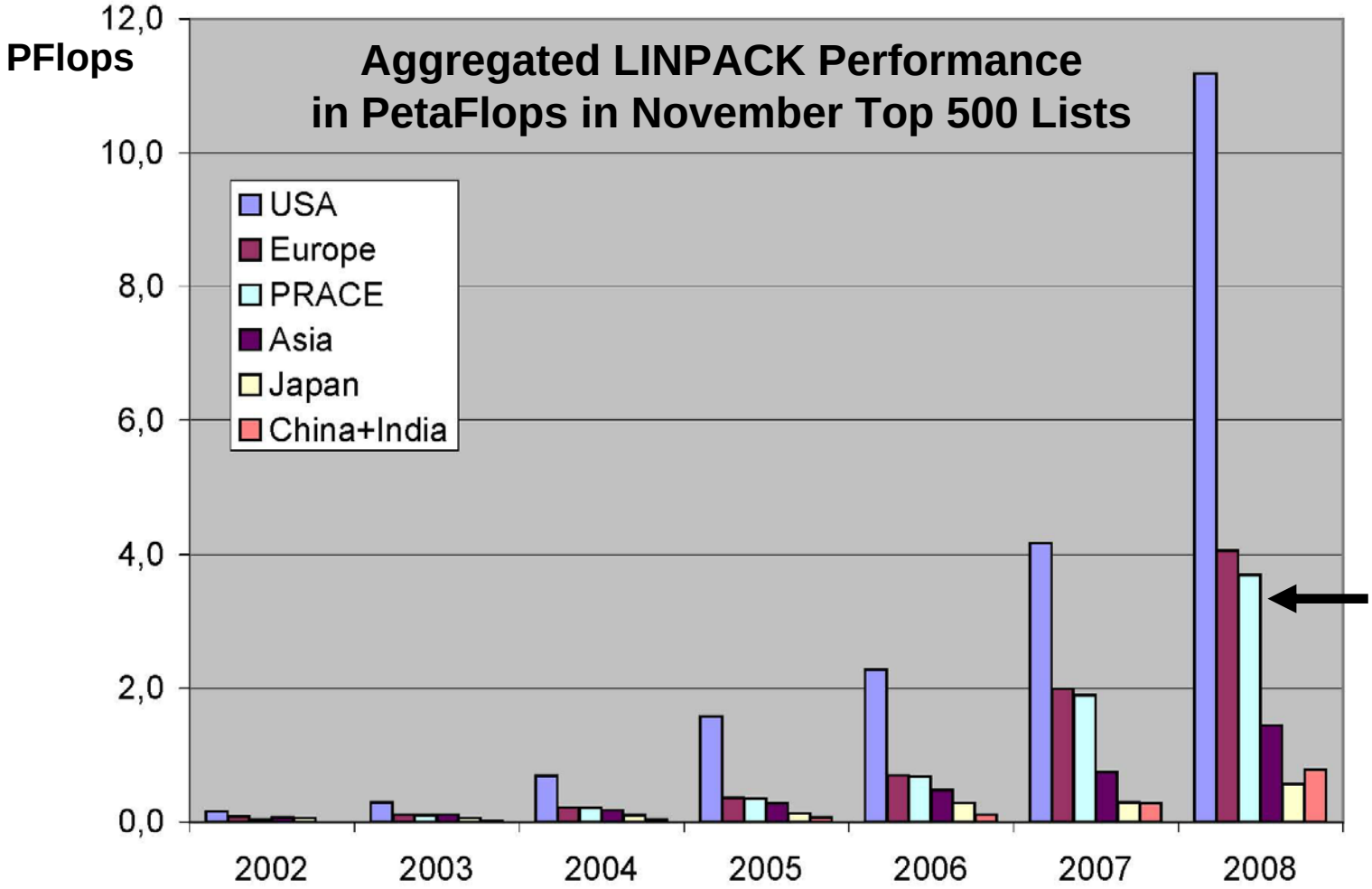
Lennart Johnsson
KTH and University of Houston

PRACE Vision and Mission

- Vision: enable and support European global leadership in public and private research and development.

- Mission: contribute to the advancement of European competitiveness in industry and research through the provisioning of world leading persistent High-End Computing infrastructure

Europe‘s position in HPC through 2008



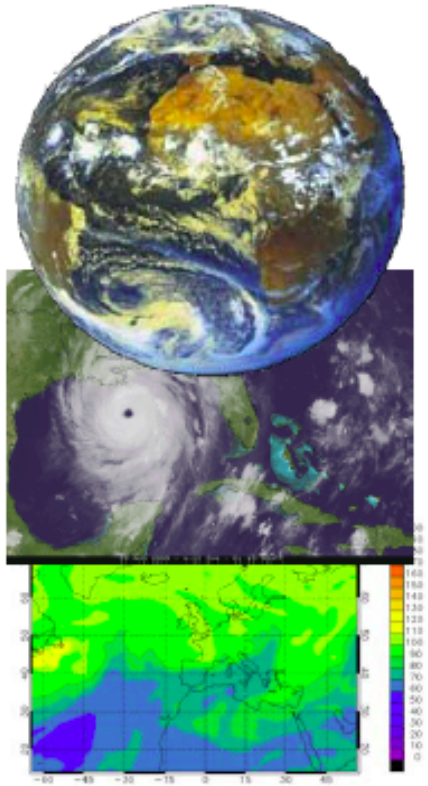
91% of Top500 European HPC-power is within PRACE countries

http://www.prace-project.eu/documents/first-scientific-workshop/eickermann_PRACE%20Infrastructure%20Usage.pdf

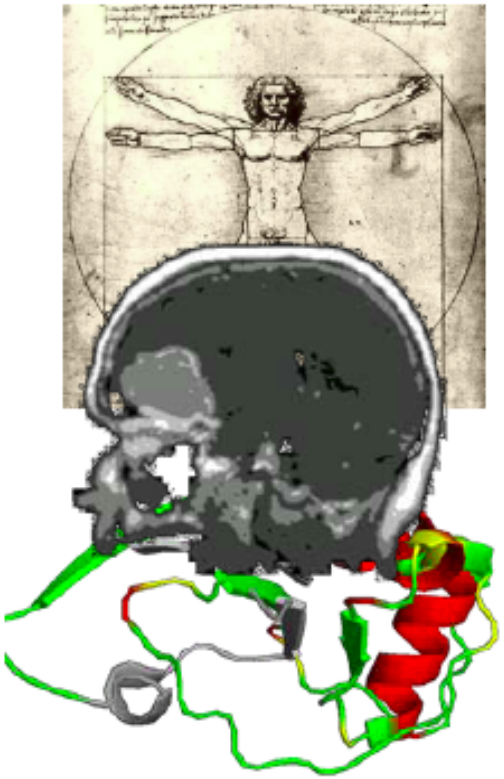
PARTNERSHIP
FOR ADVANCED COMPUTING
IN EUROPE

PRACE

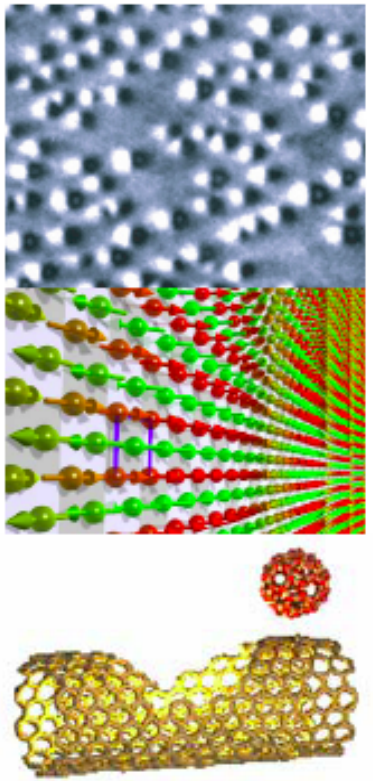
Supercomputing Drives Science through Simulation



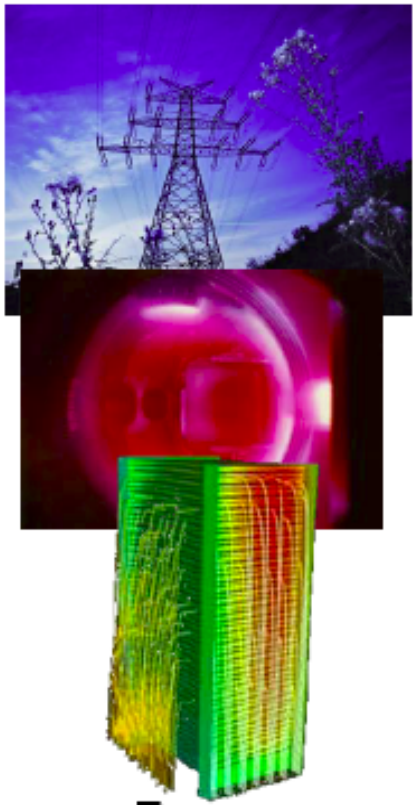
Environment
Weather/ Climatology
Pollution / Ozone Hole



Ageing Society
Medicine
Biology



Materials/ Inf. Tech
Spintronics
Nano-science



Energy
Plasma Physics
Fuel Cells

http://www.prace-project.eu/documents/first-industrial-seminar-presentations/2-%20Bachem%20PRACE%20Industrieseminar%20Amsterdam.pdf

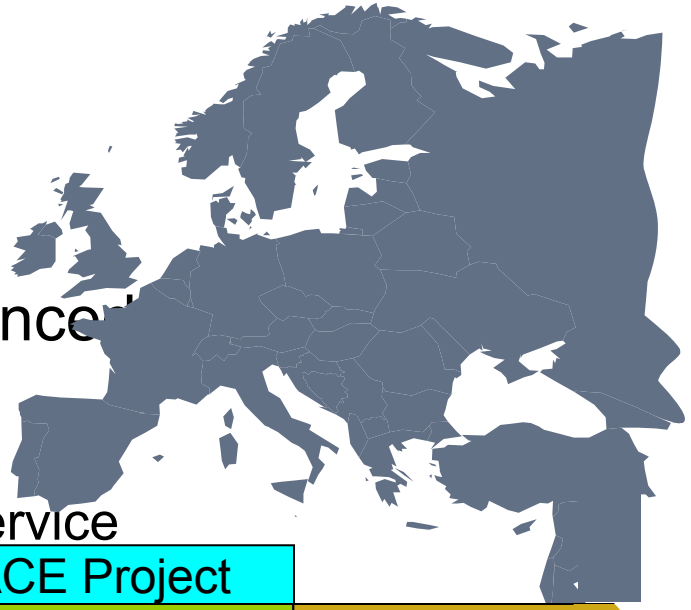
PRACE Partners

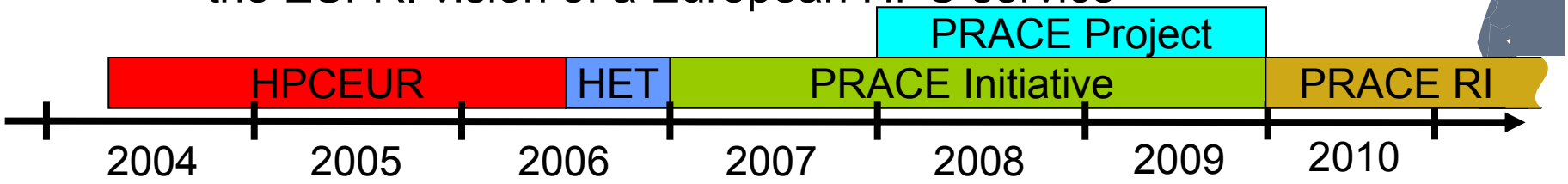
- PRACE Initiative
  - MoU signed by14 countries April 17th 2007:
    - Austria
    - Finland
    - France
    - Germany
    - Greece
    - Italy
    - Netherlands
    - Norway
    - Poland
    - Portugal
    - Spain
    - Sweden
    - Switzerland
    - United Kingdom

Since April 2007 Bulgaria, Cyprus, Czech Republic, Ireland, Serbia and Turkey has joined the PRACE initiative. Additional countries are in the process of joining.

ESFRI – Estimated costs

- Unlike other European Research Infrastructures:
  - Tier-0 resources have to be renewed every 2-3 years
  - Construction cost 200 – 400 Mio. € every 2-3 years
  - Annual running cost 100 – 200 Mio. €
  - Additional effort needed for software
- A truly European challenge – also in terms of funding
- PRACE – The Partnership for Advanced Computing in Europe
  - An Initiative created to implement the ESFRI vision of a European HPC service





(ESFRI = European Strategy Forum for Research Infrastructures)

http://www.prace-project.eu/documents/PRACE_LJ.pdf

PRACE Project Work Packages

- WP1 Management
- WP2 Organizational concept
- WP3 Dissemination, outreach and training
- WP4 Distributed computing
- WP5 Deployment of prototype systems
- WP6 Software enabling for prototype systems
- WP7 Petaflop systems for 2009/2010
- WP8 Future petaflop technologies

PARTNERSHIP
FOR ADVANCED COMPUTING
IN EUROPE

Installed Petaflop Prototypes (WP5/7/8)

IBM BlueGene/P (FZJ)
01-2008 (MPP)

IBM Power6 (SARA)
07-2008

Cray XT5 (CSC)
11-2008

IBM Cell/Power
(BSC)
12-2008

IBM Cell/Power
(BSC)

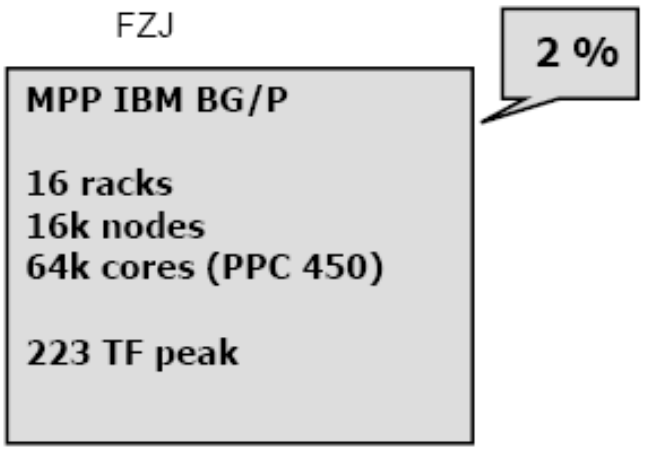
Intel Nehalem/Xeon (CEA/FZJ):
installation date April 2009

http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/AB_Ver%20PRACE_DEISA%20Amsterdam.pdf

Prototypes for 2009/2010 Systems

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



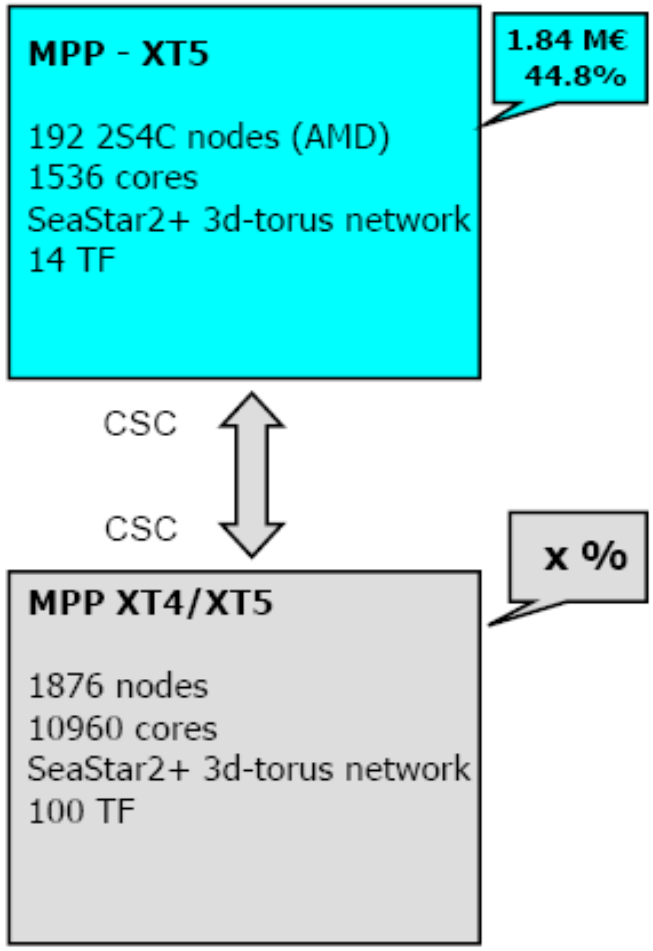
IBM

Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototypes for 2009/2010 Systems

CSC / CSCS

- Specific features:
  - Prototype installed in CSC, joint effort with CSCS
  - Funding required only for a dedicated system
  - Additional access to a larger existing system, on request – similar architecture
- Contribution to the PRACE project:
  - Access to a prototype with MPP architecture
    - AMD Barcelona, SeaStar2+ torus network
  - Early access to AMD new generation of processors (Shanghai)
  - Additional focus on hybrid MPI/OpenMP parallel programming by CSCS
  - Capability testing on the CSC existing XT system
- Availability December 2008



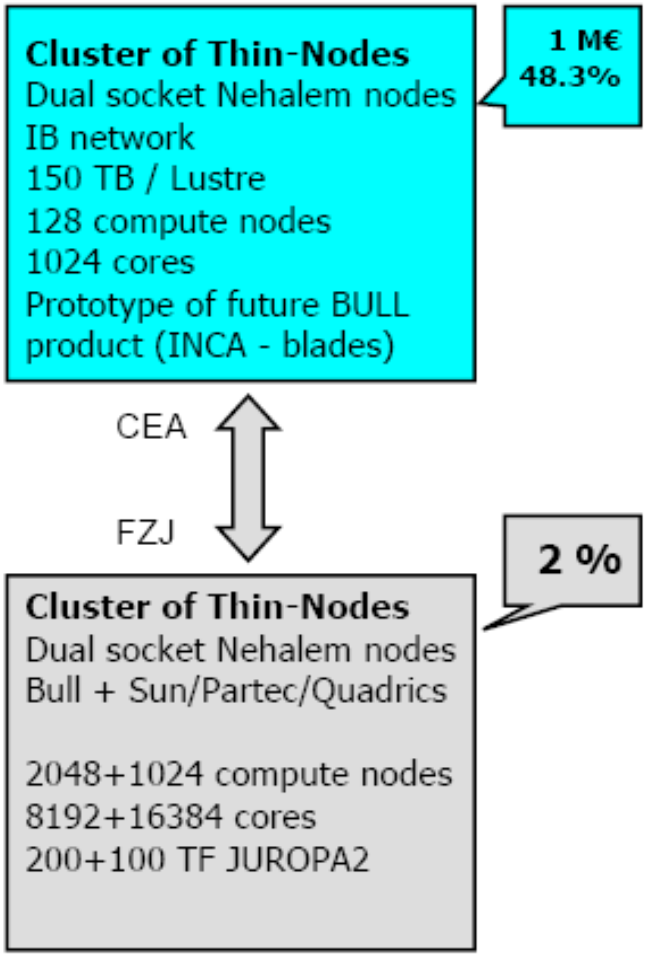
CRAY

Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototypes for 2009/2010 Systems

CEA / FZJ

- Specific features:
  - Prototype distributed over 2 sites
- Contribution to the PRACE project
  - Early access to a prototype of a new product designed by BULL (GA 2H09)
    - 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

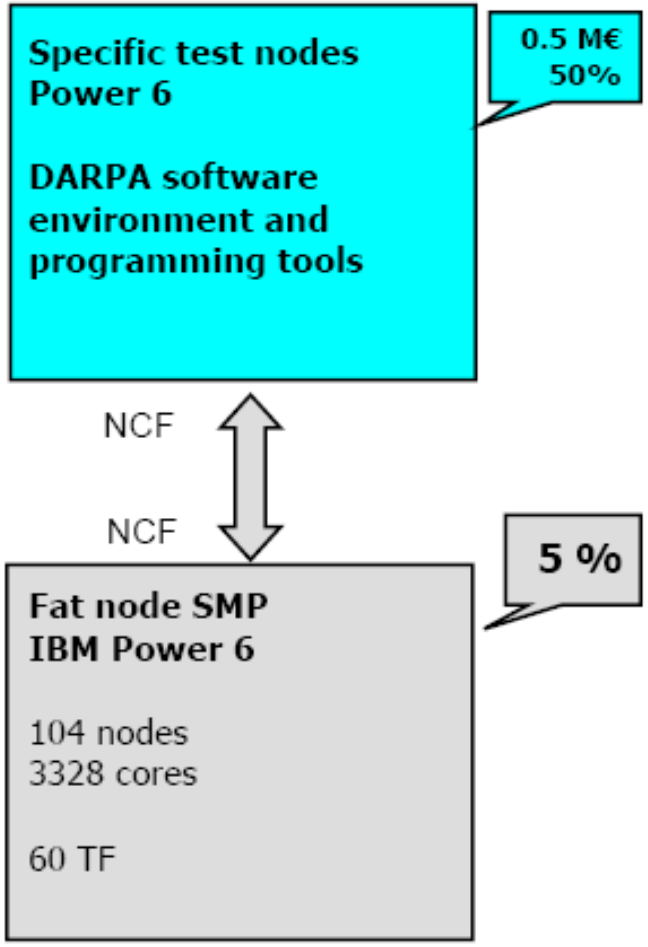


Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototypes for 2009/2010 Systems

NCF

- Specific features:
  - Access to a complete system with new Power 6 processors and IBM Power Cluster fat node architecture (large memory)
- Contribution to the PRACE project:
  - Focus on HPC software from US DARPA research – early access
  - Specific test nodes for aggressive experimentations
  - Larger scale "production" runs on the full system
- Availability October 2008



IBM

Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototypes for 2009/2010 Systems

BSC

Hybrid
IBM Cell + Power6

12 P6 + 72 Cell blades
48 + 1296 CPUs

14 TF

1 M€
48.3%

- Specific features:
  - Dedicated “fine grain” hybrid system
- Contribution to the PRACE project:
  - New Power 6 + Cell processor integration
    - Analog 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

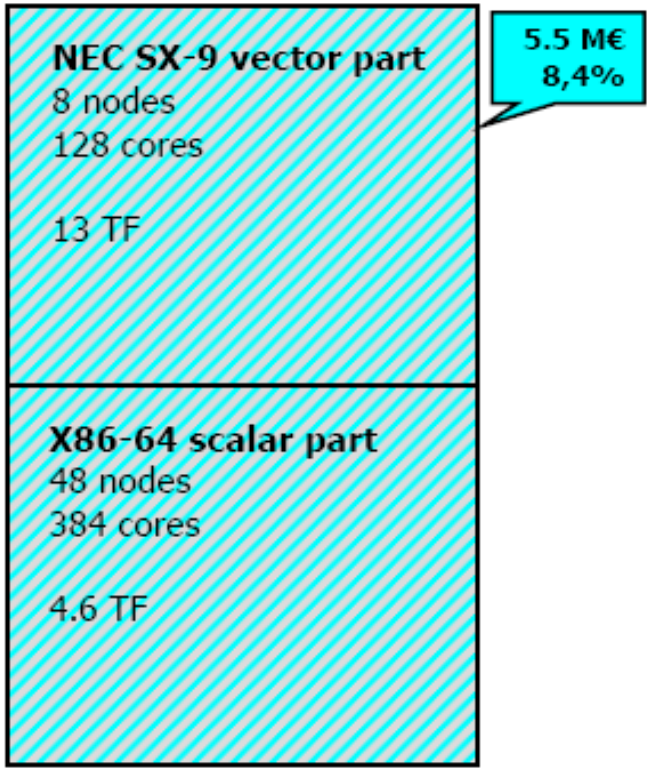
IBM

Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototypes for 2009/2010 Systems

HLRS

- Specific features
  - Hybrid vector + scalar architecture assessment
    - New NEC SX-9 vector processor
    - Intel x86-64
    - Shared filesystem
  - "Coarse grain" coupled simulations
- Contribution to the PRACE project:
  - Experiment vector components part of future Japan more integrated multi-Petaflop/s systems
  - Specific I/O issues
- Availability March 2009



NEC

Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototypes for 2009/2010 Systems

Summary	NCF	CSCS/CSC	BSC	FZJ	CEA/FZJ	HLRS
Dedicated system - makes possible unfriendly tests						
Shared large system - makes possible large runs and assessment under real production						
MPP						
Cluster with thin-nodes						
Cluster with fat nodes						
Advanced (Hybrid)						
Specific Hardware Technologies	Power6	AMD Barcelona and Shaghai	IBM Cell	Blue Gene	Intel NehalemEP	SX9
Specific Software technologies	PERCS Power7 simulator	MPI/OpenMP CAF UPC				
Full featured system with storage and IO						
Connected to DEISA					CEA new DEISA associate partner	

Courtesy Francois Robin, GENCI, EC Review 2008-07-16

Prototype 2009/2010 Systems Access

- Proposals from academia and industry are eligible, as long as the organisation of the project leader is homed in Europe (or a PRACE initiative country).
- Depending on the system vendor, researchers from some countries cannot be granted access due to export regulations.

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

Prototype 2009/2010 System

- More details on the PRACE prototypes can be found at http://www.prace-project.eu/documents/PRACE-Prototypes.pdf

Technology Prototypes

Technology Prototype

Purpose: Assess technologies and architectures in regards to

- Peak performance/ efficiency
- Programmability
- Energy efficiency
- Density
- Cooling
- Cost

What are the trends and why?

IN HPC speed is never high enough.

Until ≈ 2005 partly solved by increasing clock frequency.

As energy consumption $E = NCV^2f$ (N = # of devices/surface unit
C = capacitance
V = voltage
f = frequency)

Choice was made to increase N, lower V, keep f as low as possible. → More cores on chip with acceptable energy consumption thanks to possible technology shrink:
2005—2006 dual core, 2007—2009 quad-core & 6 core CPUs,
2010—2012 8 to many-core.

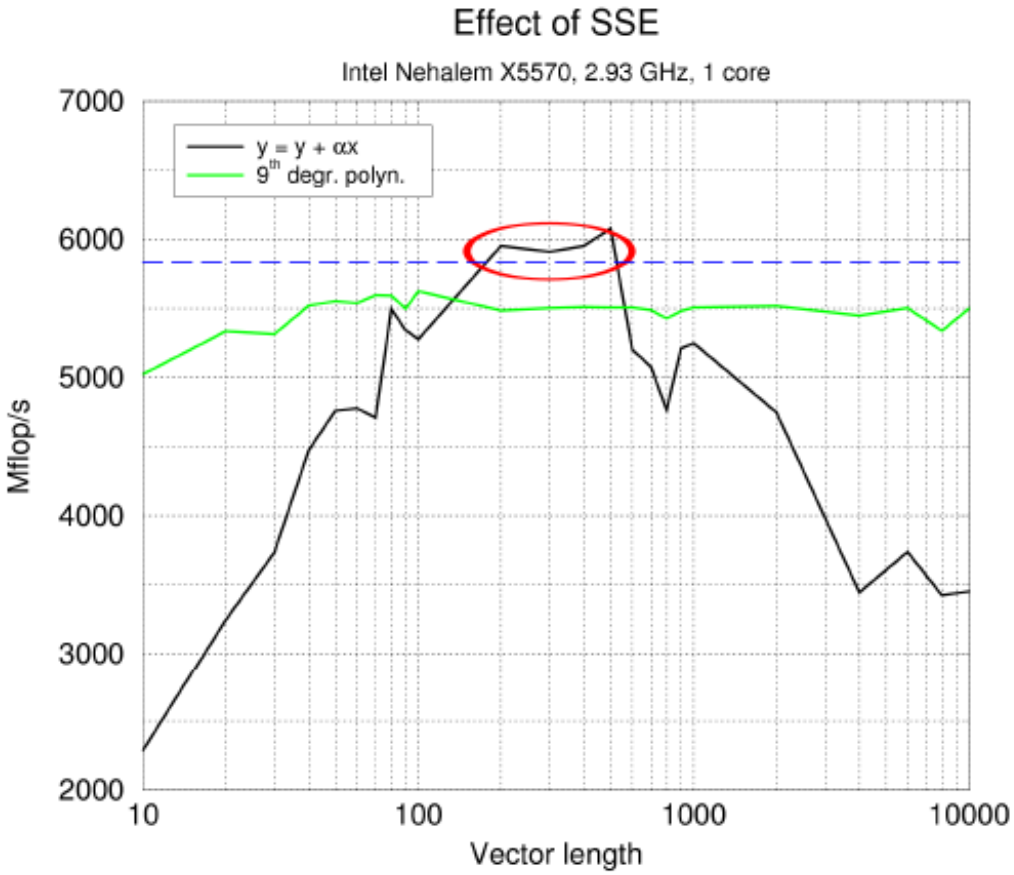
http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/pr_de_talk_vdsteen.pdf

What are the trends and why?

Combatting lack of speed

Vector units SSE*x*, AVX, etc., have been introduced.

Originally for graphics, but also of help for simple regular arithmetic operations:



http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/pr_de_talk_vdsteen.pdf

What are the trends and why?

Combatting lack of speed

External accelerators are a logical extension.

Presently four types are available:

Cell-based systems (IBM)
- BSC, FZJ prototypes

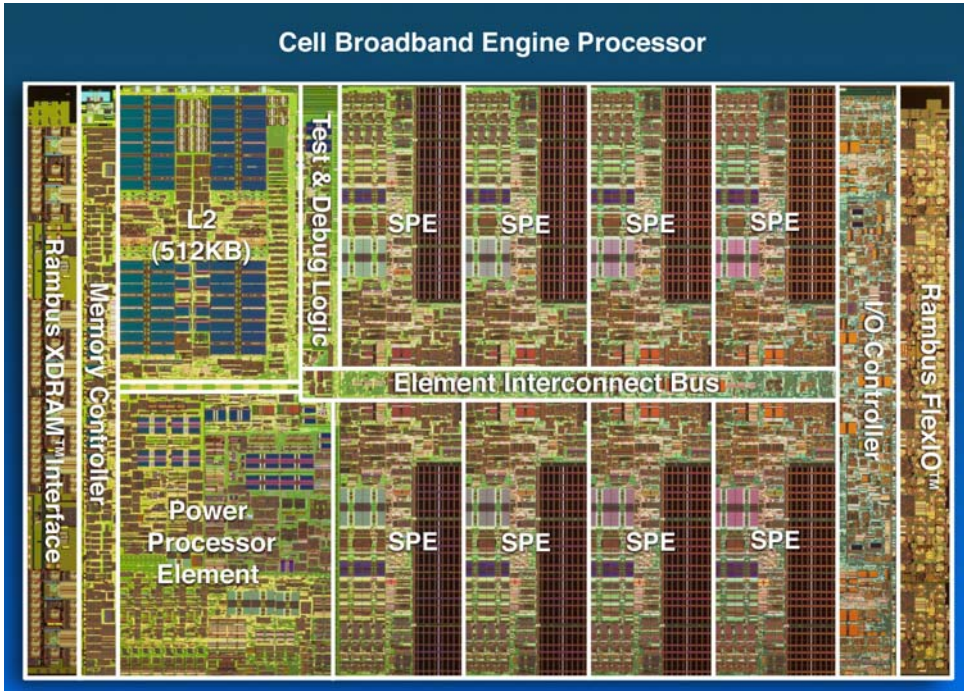
FPGA-based (CPUTech, Mitrion, Convey, Kuberre, SRC,...)
- EPCC prototype

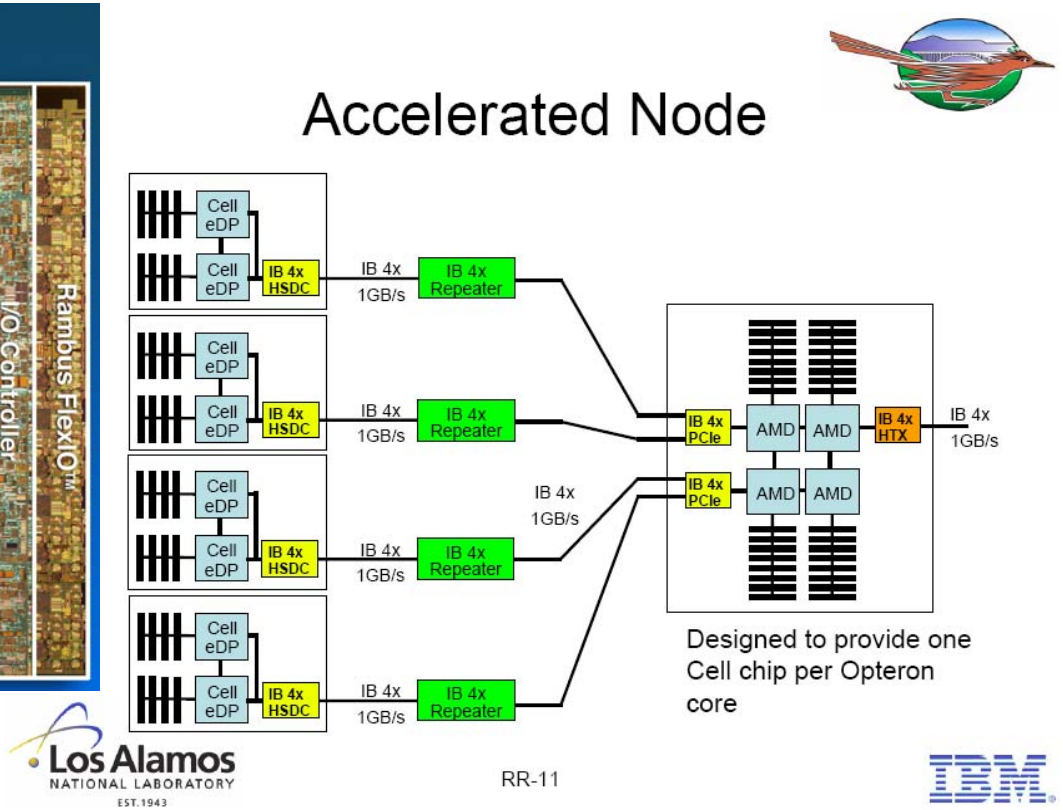
GPU-based (NVIDIA, ATI/AMD)
- GENCI prototype

ClearSpeed CSX700-based (PetaPath)
- CINES, LRZ, NCF prototypes

http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/pr_de_talk_vdsteen.pdf

Cell - Roadrunner





Los Alamos NATIONAL LABORATORY EST.1943

IBM

What are the trends and why?

Combatting lack of speed

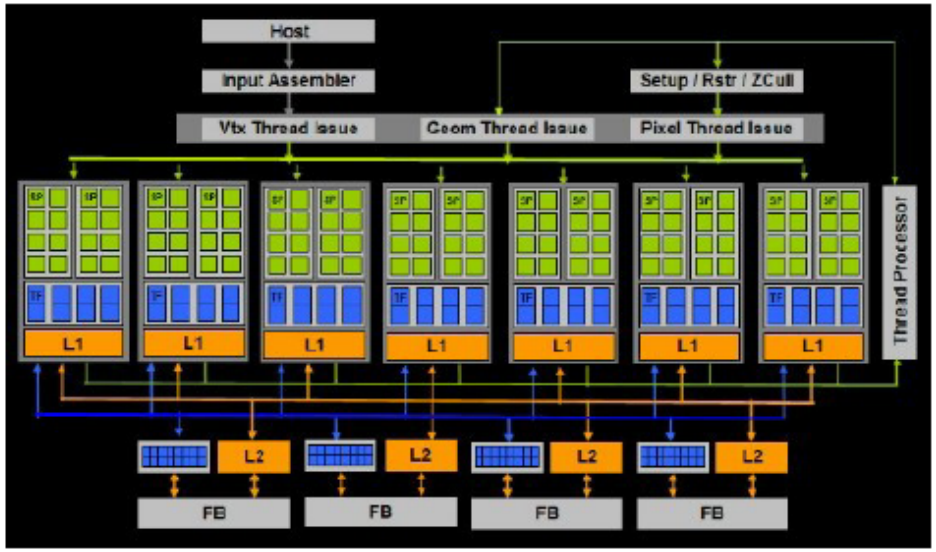
GPU-based accelerators rely on:

Many parallel processor streams

(240/GPU in the Tesla S1070)

At a moderate clock frequency

(1.296—1.44 GHz).



8-byte precision arithmetic is much slower than 4-byte precision (< 10%)

Data transport to/from host is a main issue.

No error correction!

http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/pr_de_talk_vdsteen.pdf

Rank	Computer	Mflop/W	Type	Comment
1-3	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	722.98	Accelerator	PRACE supported prototype
4-5	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, IB	458.33	Accelerator	Moderate size systems, no 26 and 79
6	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, IB	444.25	Accelerator	Large system. No 2
7	GRAPE-DR accelerator Cluster, Infiniband	428.91	FPGA	
8	NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, IB	379.24	GPU	Large system. No 5
9 -20	Blue Gene/P Solution	378.76	Proprietary	
21	Blue Gene/P Solution	366.58	Proprietary	Large system, No 1
22-23	Blue Gene/P Solution	353.98	Proprietary	Large systems, no 4 and no 8.
24	Sun Blade X6440, Opteron 2.5 Ghz, Infiniband QDR	341.42		Top Green500 commodity system
25	iDataPlex, Xeon E55xx QC 2.26 GHz, Infiniband, Windows HPC2008 R2	299.52		
.....	.....	.....		
500	Cluster Platform 4000 BL465c, Opteron DC 2.4GHz, GigEthernet	13.03		

What are the trends and why?

Power issues in HPC

For Multi-Petaflop/s to Exaflop/s systems real disruption in technology is needed. With current type of technology (even with progress factored in) we would have the following situation:

2012	2018
10 Pflop/s	1 Eflop/s
9.3 MW for IT	154 MW for IT
14 MW total	213 MW total
12.3 M€/y	187 M€/y

Assuming: PUE = 1.5, € 0.10/Kwhr.

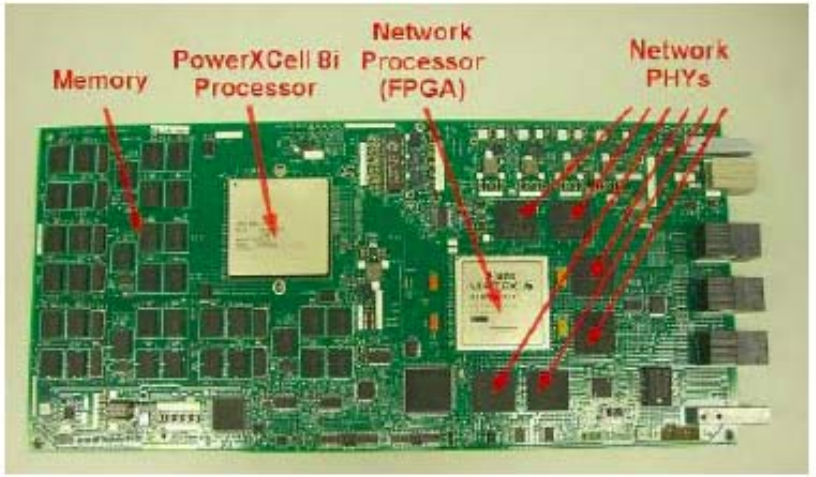
http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/pr_de_talk_vdsteen.pdf

FZJ

eQPACE

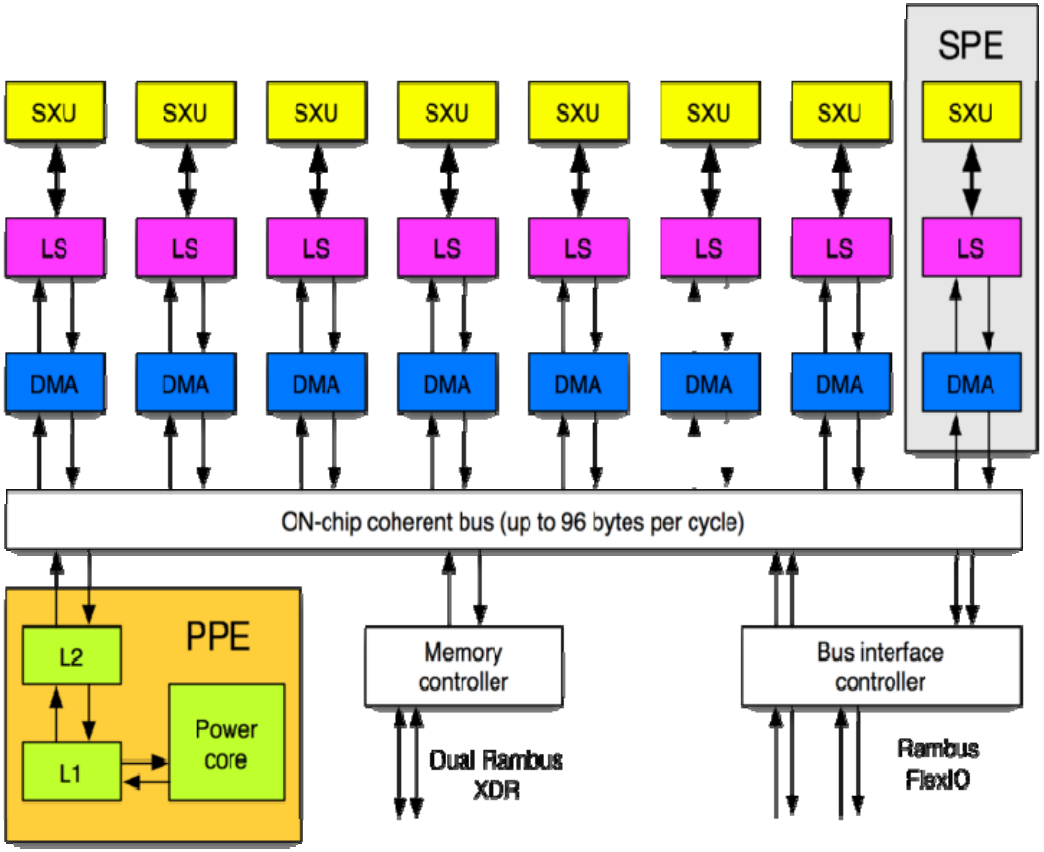
Extend communication capabilities of eQPACE to make it suitable for a wider range of applications. Reach a top position in the Green500 list (FZJ).

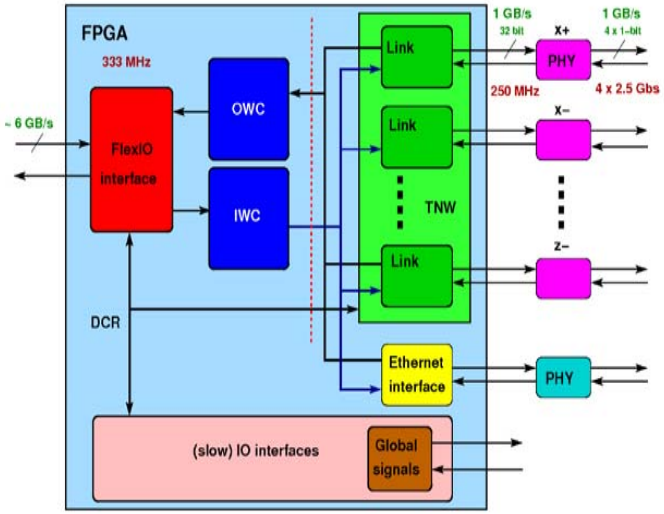
- Hardware: PowerXCell8i processor nodes with custom 3D-torus interconnect.
- Benchmarks:
  HPL, Euroben kernels,
  torus network benchmark,
  applications & iterative solvers.
- Programming environments:
  Cell SDK & CellSs



http://www.prace-project.eu/documents/wp8_prototypes.pdf

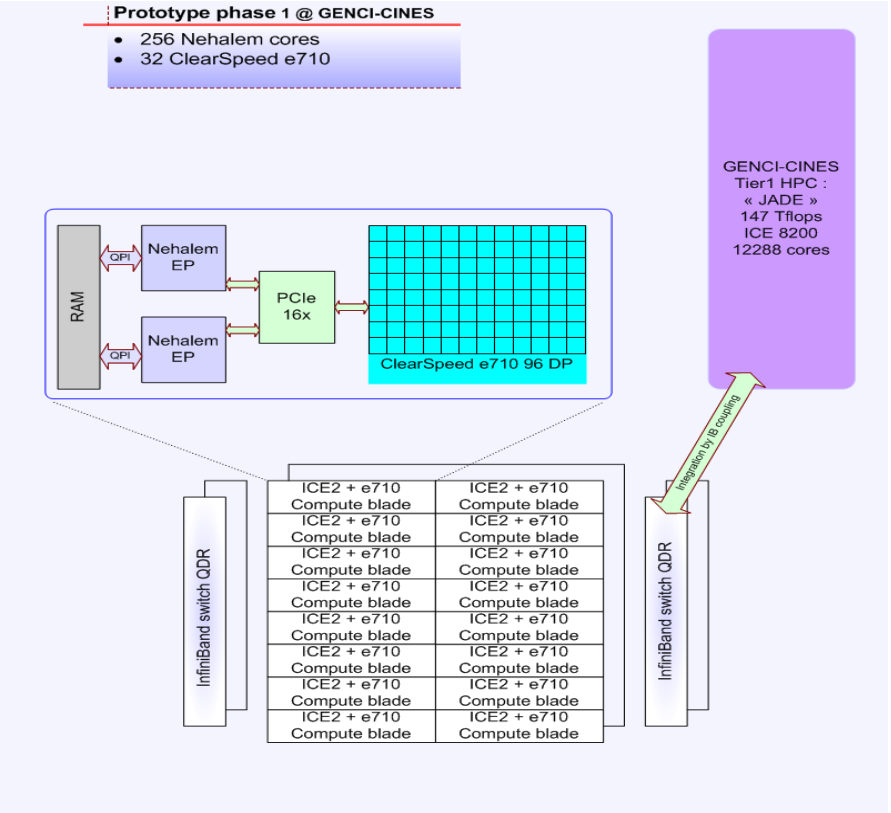
FZJ


SPE
SXU
LS
DMA
ON-chip coherent bus (up to 96 bytes per cycle)
PPE
L2
L1
Power core
Memory controller
Dual Rambus XDR
Bus interface controller
Rambus FlexIO


FPGA
333 MHz
~ 6 GB/s
FlexIO interface
OWC
IWC
DCR
Link
Link
TNW
Link
Ethernet interface
(slow) IO interfaces
Global signals
1 GB/s
32 bit
X+
PHY
250 MHz
4 x 2.5 Gbs
1 GB/s
4 x 1-bit
X−
Z−
PHY

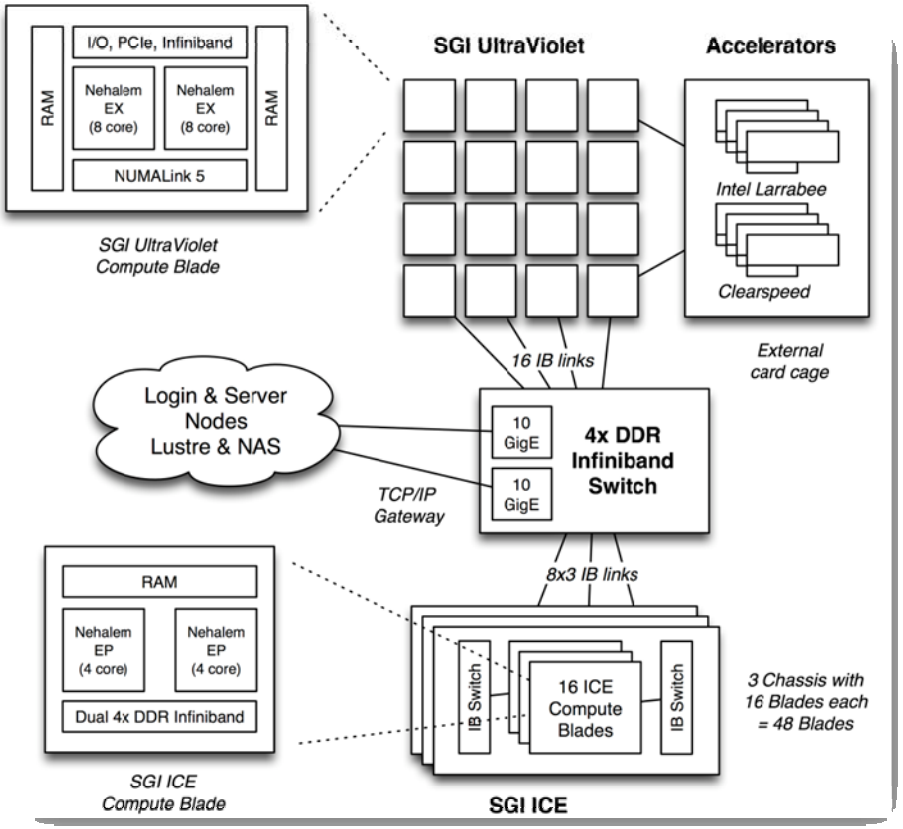
LRZ - CINES

- **SGI hosting platform:**
- 32 Blades XE
  - 64 Processors Intel Nehalem-EP
  - 256 Cores
  - 4 GB per core
- Infiniband QDR
- Estimated peak performance is 3 TFlop/s.
- **ClearSpeed-Petapath accelerators:**
- 32 × e710 cards
  - One per ICE blade
  - 4 GFlop/s (DP) / Watt
- Software development toolkit
- Estimated peak performance is 3 TFlop/s DP.
- The total accumulated peak performance is 6 TFlop/s DP.



BAdW-LRZ

- SGI Altix ICE System with 48 compute blades each containing two Intel Xeon Nehalem-EP 2.53 GHz 4-core processors and 24 GB main memory;
- 4-Socket Nehalem-EX WhiteBox from Intel (α-Prototype: ”Sunrise Ridge“);
- SGI “UltraViolet” prototype with 16 “P1-Level” compute blades each containing two 2.0 GHz Intel Xeon Nehalem-EX 8-core processors and 32 GB main memory;
- 4× DDR Infiniband switch and cables (44 IB ports plus 2 x 10GigE ports);
- SGI UV extension unit with
  - 1 Intel Larrabee accelerator;
  - 4 ClearSpeed-Petapath Advance e710 accelerator boards;
- Two racks with system administration infrastructure (head node);
- SLES10 SP2 operating system and SGI “Tempo” and SGI “ProPack” management software;
- Storage: The hybrid system prototype is connected with the Lustre file systems of the BAdW-LRZ linux cluster (130 TB /scratch and 50 TB /project).



CEA

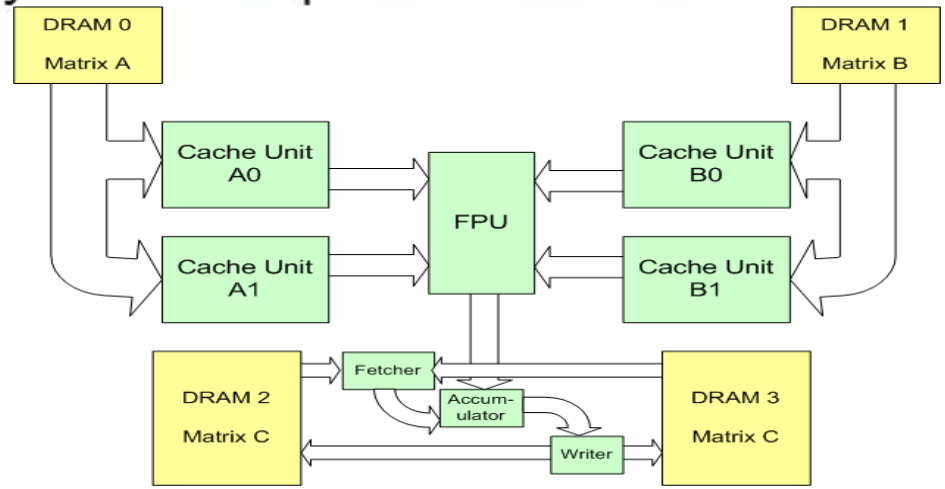
- The prototype is a cluster of 5 nodes. The first node is a login node featuring four Nehalem-EP processors. The last four nodes are BULL R422 servers featuring 2 Haperton CPUs with 16 GB of RAM each. The BULL servers are connected through a PCI-Express 16× to two NVIDIA TESLA servers. Infiniband DDR is used to interconnect the nodes.
- A TESLA server has 2 C1060 graphic boards. A C1060 features 30 multiprocessors at 1,3 GHz and has 4 GB of memory. One multiprocessor has 8 single precision (SP) units and 1 double precision (DP) unit. Therefore, a C1060 can have 240 SP (30 DP) units running in parallel. The peak performance of a C1060 is 78 GFlop/s DP and 624 GFlop/s SP.

EPCC

Maxwell FPGA

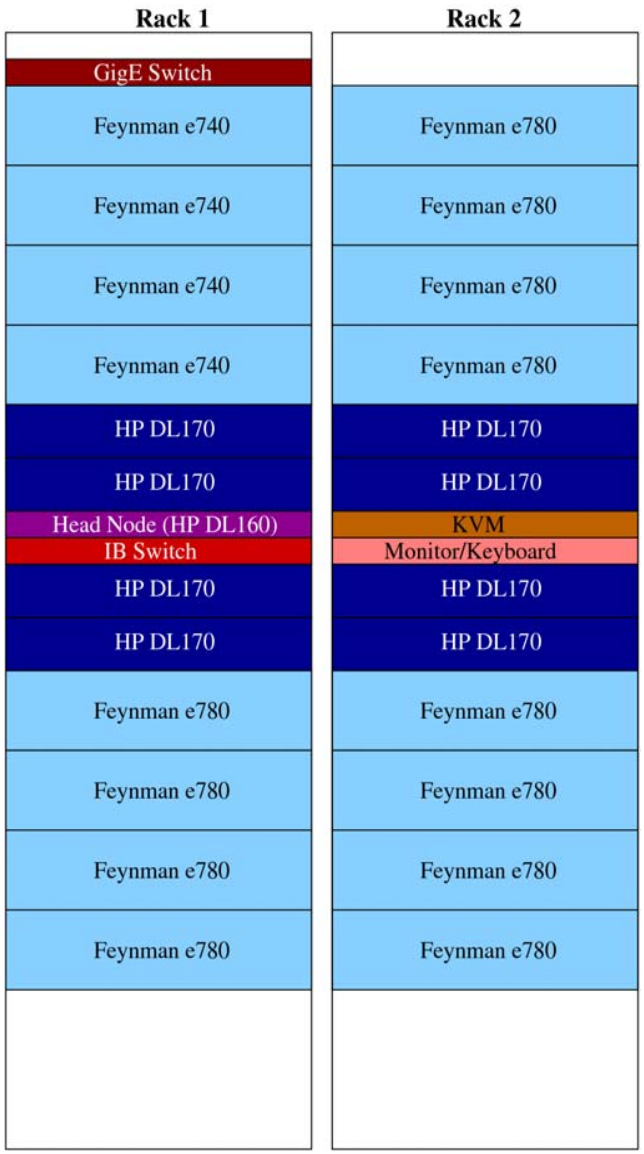
Evaluate the performance and usability of the HARWEST Compiling Environment (EPCC).

- Hardware: FPGA prototype "Maxwell" (32 FPGAs) from both Alpha Data Ltd and Nallatech Ltd using Virtex-4 FPGAs supplied by Xilinx Corp.
- Benchmarks:
  4 Euroben kernels
- Languages:
  - VHDL
  - HCE



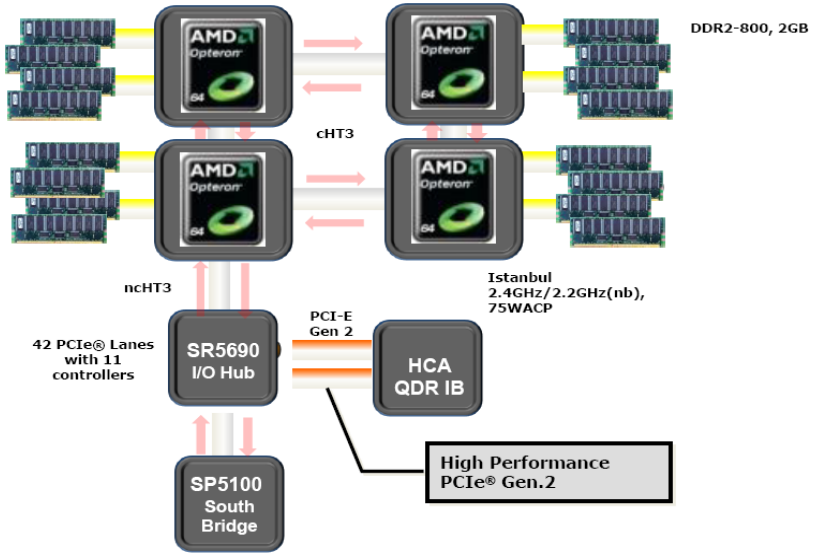
NCF

- "Feynman e740/e780" contain 4 or 8 ClearSpeed CSX700 processors for a total of 114. The dark blue HP DL170 host nodes each connect to two Feynman boxes by PCI Express Gen. 2 16× (8 GB/s).
- The HP host nodes are connected by 4× DDR Infiniband in order to employ them, and the attached accelerator boxes.
- Internally a Feynman e780 connects to its 8 processors through a PLX 8648 PCIe switch. The bandwidth to each individual processor is 1 GB/s. The e740 boxes contain only 4 CSX700 processors but the bandwidth to/from the individual processors is doubled to 2 GB/s. This should help for bandwidth-hungry applications. The peak performances are 768 GFlop/s and 384 GFlop/s in 8-Byte precision for a Feynman e780 and e740, respectively. The energy efficiency is very high: ≈ 4 GFlop/s/Watt.
- The HP DL170h G6 host nodes contain 2 Intel X5550 Nehalem processors at 2,67 GHz and 24 GB of DDR3 memory/node. Furthermore, each host node contains dual disks of 250 and 500 GB, for system and user data, respectively. Access to the system is provided through the HP DL160 G6 head node. The system is divided in a development part (1 DL170 + 2 Feynman e740s) and a production part (7 DL170s + 2 Feynman e740s + 12 Feynman e780s).



SNIC-KTH

- 4-socket blade with 6-core 2.1 GHz AMD Istanbul HE CPUs, 32GB/node
- 10-blade in a chassis with 36-port QDR IB switch
- 180 nodes, 4320 cores, full bisection IB interconnect
- 2TF/chassis, 12 TF/rack



First RI call 2010

Merit based access as determined by evaluation conducted by PRACE Scientific Committee comprised of leading European scientists in disciplines in need of High-End Computing

Outreach and Education

Industry seminars:
1st Seminar Sept. 3, 2008 Amsterdam, The Netherlands

Summer & winter schools:
Stockholm, Athens

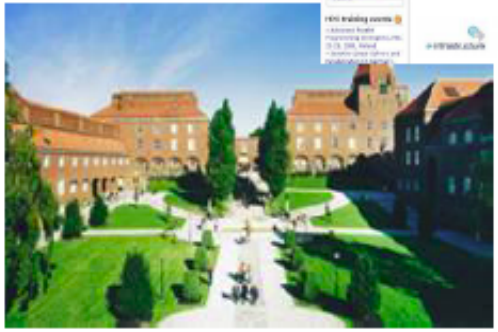
PRACE booths:
ISC, ICT, SC,

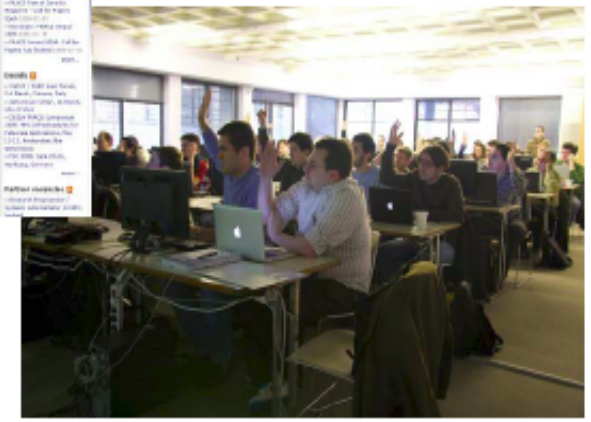
PRACE booth at SC08

1st Industry seminar, 3.9.2008

PRACE website


ICT 2008, PRACE-Booth


PRACE Summer School Stockholm


PRACE Winter Schoolat the OTE academy, Athens 26-29.8.2009

http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/AB_Ver%20PRACE_DEISA%20Amsterdam.pdf

High-End Computing Software Survey

European Large Scale Applications characteristics (Linpack Equivalent Flops (LEFs) used in TF units)

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
CFD	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

http://www.prace-project.eu/documents/PRACE_LJ.pdf

Mapping Applications to Architectures

- Identified affinities and priorities
- Based on the application analysis - expressed in a condensed, qualitative way
  - Need for different "general purpose" systems
  - There are promising emerging architectures
- Will be more quantitative after benchmark runs on prototypes

Code	**MPP** (i.e. BlueGene L/P or CRAY XT4/5)	**Thin node clusters** (i.e. Bull INCA or SGI ICE)	**Fat node clusters** (i.e. Bull MESCA, SGI UltraViolet or IBM Power6)	**Vector systems** (NEC SX8-9, Cray X2	**Accelerated systems** (i.e. scalar or vector + GPU, FPGA or Clearspeed).	**Accelerated systems – Cell based** (i.e. Roadrunner, Mariceli)
NAMD				E	E	E
CPMD						E
VASP						E
QCD				E	E	E
GADGET						
Code Saturne				E	E	E
TORB						E
NEMO						
ECHAM5						
CP2K	E					
GROMACS					E	E
N3D		E	E			
AVBP	E					
HELIUM						
TRIPOLI 4	E		E			
GPAW						
ALYA						E
SIESTA						E
BSIT						E
PEPC				E	E	E

Table 4 : application mapping to Petaflop/s systems archite

http://www.prace-project.eu/documents/PRACE-Sample-Presentation-2009.pdf

Acknowledgements

Workshop Sponsors

www.prace-project.eu

http://www.prace-project.eu/documents/deisa-prace-symposium-presentations/AB_Ver%20PRACE_DEISA%20Amsterdam.pdf