



# PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

## Prototypes Systems for PRACE

François Robin, GENCI, WP7 leader



## Outline

- Motivation
- Summary of the selection process
- Description of the set of prototypes selected by the Management Board
- Conclusions

## Outline

- Motivation
- Summary of the selection process
- Description of the set of prototypes selected by the Management Board
- Conclusions

## Motivation

- Procurement of the production systems - prototypes are central to the "acquisition pipeline"
  - Architecture evaluation vs user requirements,
  - Preparation of technical requirement, including benchmarks,
  - Risk mitigation, ...
- Integration of different computing centers and systems into a single RI
  - Preparing for operation of the RI and easy user access
  - Strengthening relationship between sites and partners
- Contribution to the European HPC ecosystem and its relationship with international high end initiatives

WP6 : Software enabling for Petaflop/s systems

**WP5 : Deployment of prototype systems**

WP7 : Petaflop/s Systems (for 2009/2010)

WP8 : Future Petaflop/s computer technologies (beyond 2010)

WP4 : Distributed system management

## Intended use of the prototypes

- Systems Management (WP4):
  - Test and deployment of the software for distributed system management
- Deployment (WP5):
  - Technical assessment of the prototype systems under production conditions (e.g. system software, maintenance, system operation)
  - Evaluation and enabling of communication and I/O infrastructure
  - Evaluation and benchmarking of user applications
- Applications (WP6):
  - Preparation of benchmarks
  - Petascaling and optimization of applications
  - Software libraries and programming models
- Petaflop/s Systems for 2009/2010 (WP7):
  - Integration of experience gained from prototypes in technical specifications for production systems

## Outline

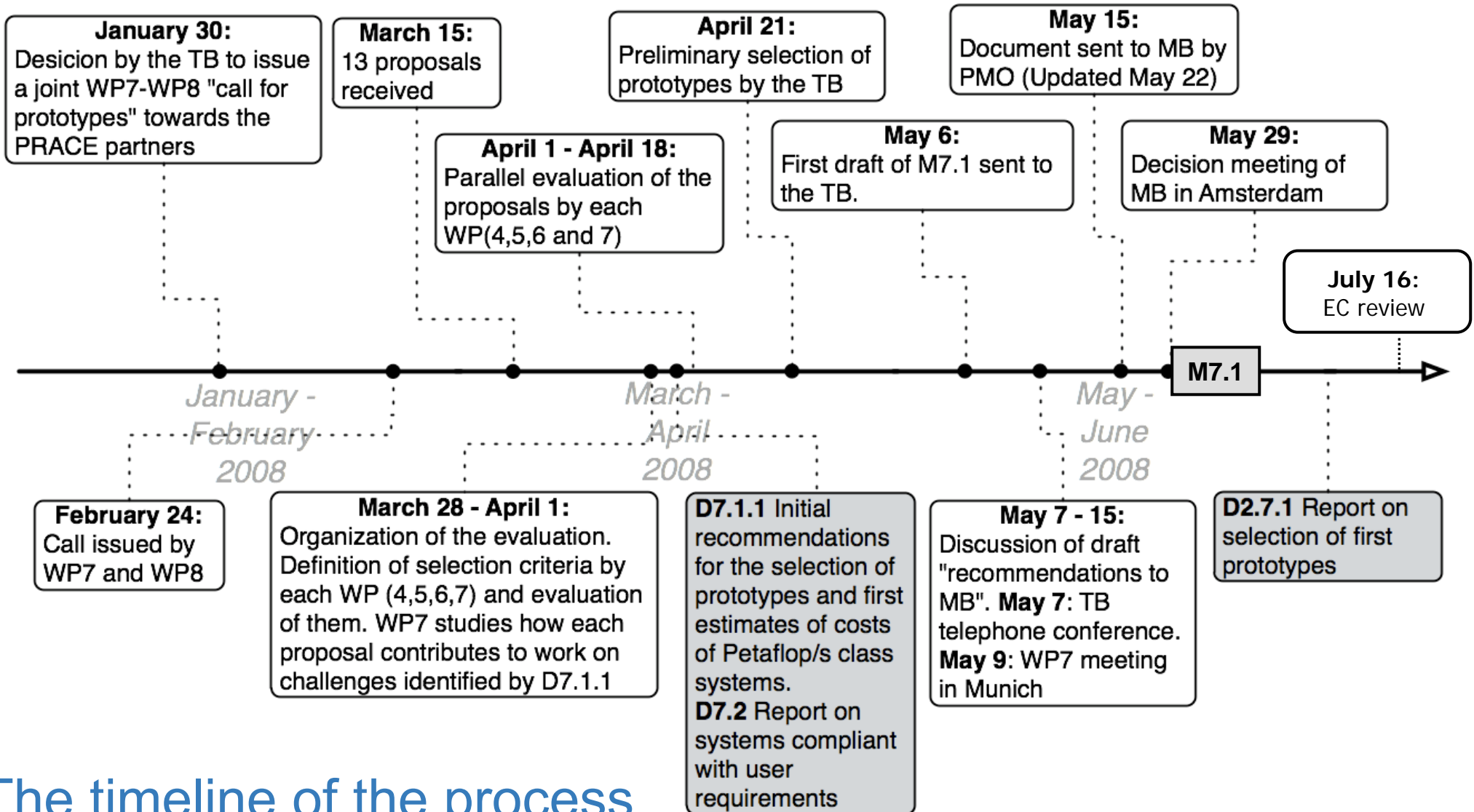
- Motivation
- Summary of the selection process
- Description of the set of prototypes selected by the Management Board
- Conclusions

## General principles

- Cover all relevant architectures, with a weighting of their relative importance for user's needs
- Anticipate longer term user user's needs exploiting more advanced architectures
- Cover as much as possible promising hardware and software technologies for Petaflop/s systems in 2009/2010
- Make sure to receive feedback during the project duration from the prototypes

## Remarks

1. Build the best set of prototypes for preparing a timely and seamless deployment of production systems in 2009/2010 - Do not attempt to select the best individual prototypes
2. A fair and open process used was defined by WP7 and the Technical Board and approved by the Management Board
3. A "Call for prototypes" was issued : the PRACE partners were asked to propose prototypes in order to leverage the expertise of the different sites and also to foster a site-vendor relationship network
4. The goal is to evaluate architectures not vendors



## The timeline of the process

## Outline

- Motivation
- Summary of the selection process
- Description of the set of prototypes selected by the Management Board
- Conclusions

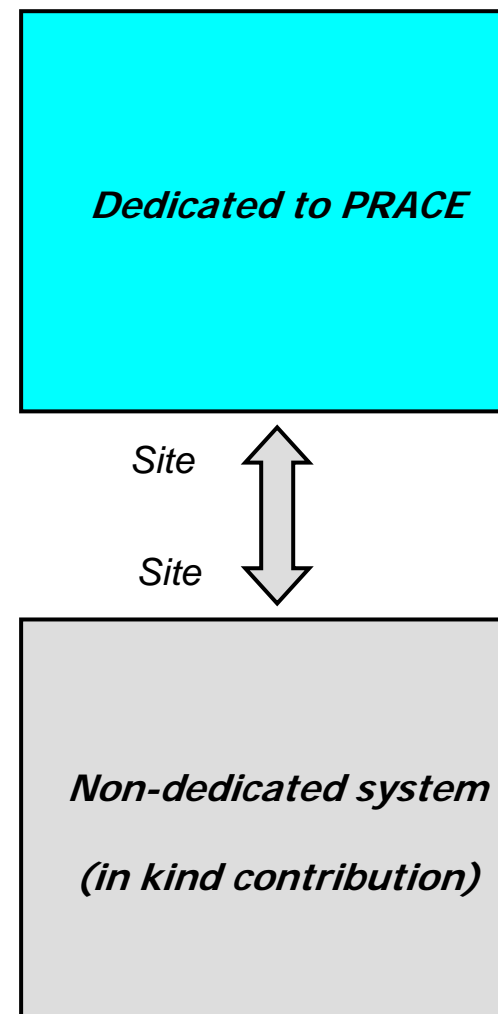


Site	Architecture Vendor/Technology	Point of contact
<b>FZJ</b> Germany	<b>MPP</b> IBM BlueGene/P	Michael Stephan <a href="mailto:m.stephan@fz-juelich.de">m.stephan@fz-juelich.de</a>
<b>CSC-CSCS</b> Finland+Switzerland	<b>MPP</b> Cray XT5/XTn - AMD Opteron	Janne Ignatius <a href="mailto:janne.ignatius@csc.fi">janne.ignatius@csc.fi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

This slide represents the contents of the following slides

## *Site(s)*

- *Specific features*
- *Main contributions to the PRACE project*
- *Availability date (general opening)*



VENDOR

## 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

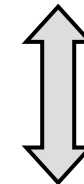
- Specific features:
  - Prototype installed in CSC, joint effort with CSCS
  - Funding goes fully for the dedicated system
  - As additional in-kind contribution access to a larger existing system – similar architecture
- Contribution to the PRACE project:
  - Access to a prototype with MPP architecture 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

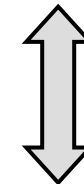
- Specific features:
  - Combination of dedicated test system at CEA and a large production system of same architecture at FZJ
- Contribution to the PRACE project
  - Early access to a prototype of a new product designed by BULL (pre GA machine)
    - High density blade based system with new Intel Nehalem 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

- 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

## 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

- 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

*\* Current estimate, subject to ongoing negotiation*

## Summary

	NCF	CSCS/CSC	BSC	FZJ	CEA/FZJ	HLRS
<b>Dedicated system - makes possible unfriendly tests</b>						
<b>Shared large system - makes possible large runs and assessment under real production</b>						
<b>MPP</b>						
<b>Cluster with thin-nodes</b>						
<b>Cluster with fat nodes</b>						
<b>Advanced (Hybrid)</b>						
<b>Specific Hardware Technologies</b>	Power6	AMD Barcelona and Shanghai	IBM Cell	Blue Gene	Intel NehalemEP	SX8
<b>Specific Software technologies</b>	PERCS Power7 simulator	MPI/OpenMP	Software stack Programming models			
<b>Full featured system with storage and IO</b>						
<b>Connected to the DEISA network</b>					CEA new DEISA associate partner	
<b>Collaboration with vendors</b>	Software technology (IBM)	System reliability, performance, functionality (CRAY)	System design (IBM)		System design (BULL)	System design (NEC)

## Summary of the selected set of prototypes

- A set of full-featured systems of promising Petaflop/s architectures
- Spanning the main aspects of challenges for Petaflop/s systems
  - Including: Petascaling, programmability, reliability
- Assessment of important issues for production systems, including:
  - Distributed systems management
  - Power consumption and cooling
  - I/O and file systems
- "General purpose" systems with advanced architectures and software, allowing Europe to stay on equal footing with ongoing multi-Petaflop/s projects in the USA and in Japan
- Includes most components that are candidate to be the building blocks of 2009/2010 production systems
- Open new technology options that WP8 will explore more systematically
- Enforces numerous collaborations with vendors

## Outline

- Motivation
- Summary of the selection process
- Description of the set of prototypes selected by the Management Board
- Conclusions

## Conclusions: The proposed set of prototypes

- Covers the architectures likely to become Petaflop/s class production systems in 2009/2010 and most hardware/software components that will be the building blocks of such systems
- Includes both general purpose architectures and more advanced architectures and thus forms a broad basis for the decisions about the production systems
- Will contribute to the success of the PRACE project and its recognition in the scientific community
- Since no new architecture or major technology is likely to appear before the end of 2008, the prototype selection has been completed in a single phase

## Conclusions : Next steps

- WP8 prototypes selection will take into account the prototypes already selected in order to maximize the experience gained through them
  - WP8 and WP7 prototypes evaluation will lead to exchange of information and comparisons between prototypes.
  - Feedback towards vendor should be organized to leverage impact of PRACE to the European HPC ecosystem
- A similar process but oriented toward the creation of specifications for procurements (end of 2008)
  - Will take into account aspects related to the operation of Petaflop/s system (including TCO)
- Evaluation of prototypes managed by WP5 (till the end of the PRACE project)
  - Contribution by WP4 and WP6
  - Will produce key inputs for the specifications of productions systems

A blue banner with a grid pattern and a faint map of Europe. On the left, there is a small image of a woman's face. In the center, the text "PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE" is written in white. On the right, the word "PRACE" is written in a stylized font, surrounded by a circle of white stars.

PARTNERSHIP  
FOR ADVANCED COMPUTING  
IN EUROPE

Some details about the prototypes

## FZJ

- IBM Blue Gene/P (production system)
- Compute configuration:
  - IBM PowerPC 450d processors
  - 16384 nodes
  - 65536 cores
  - 32 TB memory
- Network:
  - Three internal networks (proprietary)
    - 3D Torus + Collective Tree + Global Barrier
  - 10 GigE external for storage
- IO configuration:
  - GPFS
  - 1000TB
- # 6 in Top500 (June 2008)
  - #2 (November 2007)
- Unique features:
  - Best performance / watt ratio ([www.green500.org](http://www.green500.org))
  - Small foot print
  - Extreme scalability ( > 512K cores)
- Existing system can easily be extended
- Fully integrated in production environment

## CSC / CSCS (1/4)

- Hardware
  - Two fully populated cabinets (1440 compute cores) of Cray XT5, initially with AMD Opteron Barcelona quad-core 2.3 GHz processors
  - 1 GB memory per core
  - 3 service/IO/login blades, dedicated disk system
  - All the compute processors will be upgraded to next generation AMD quad-core processor Shanghai in 1H/2009. Early upgrade to the latest technology!
- The prototype is a separate system with its own login nodes and IO system
  - Can be used for dedicated whole machine PRACE runs and unfriendly experiments with a very short notice
  - Allows for studies which could not be easily performed on a production system
    - testing different versions of the operating system, libraries, tools and compilers
    - lowering the network speed to simulate characteristics of future systems

## CSC / CSCS (2/4)

- Early upgrade to the next processor architecture emphasizes the research characteristics of the prototype
  - The latest technology typically includes issues in reliability, performance and functionality of the system. Valuable information for PRACE!
- An excellent environment for testing the hybrid MPI/OpenMP programming model
  - Will be the preferred way to exploit the computing power of the emerging multi-core processors efficiently
  - Strong contribution from CSCS

## CSC / CSCS (3/4)

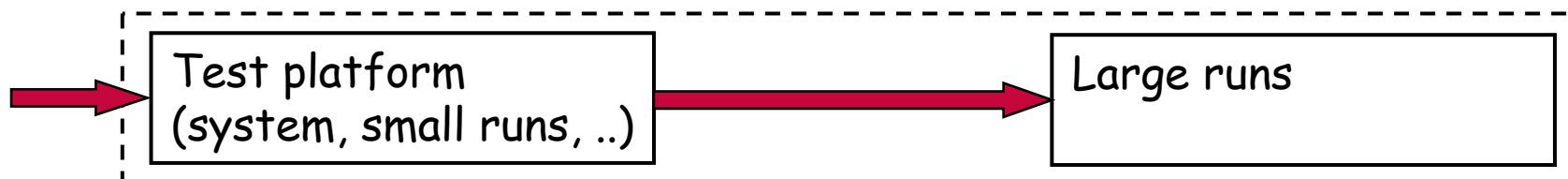
- As additional in-kind contribution access is granted to CSC's Cray XT system 'Louhi' (86.7 TF, 9424 compute cores)
  - Initially up to 15 % on Louhi
  - Special queue arrangements
  - At specific times access to ½ - full machine
  - The Louhi system is fully paid by the state of Finland
- For special testing purposes, it is possible to run an application across combined PRACE prototype + Louhi (100 TF, 11k cores)
  - Requires rebooting of both systems prior to the runs
  - For Petascaling it is essential for the project to have access to as large systems as possible

## CSC / CSCS (4/4)

- Prototype with MPP architecture and fast processors from a leading vendor gives additionally several benefits to PRACE:
  - Access to the highly skilled Cray application team, which will actively support the performance analysis, benchmarking and optimization work carried on within PRACE
  - Cray performance tools are among the very best in the industry and make it easy to pinpoint performance problems
  - Other synergy benefits from Cray/CSCS/CSC (e.g. on-site Cray engineer at CSC)
  - Cray is fully concentrating on HPC. Cray has been estimated to be one of the very few vendors (2?) who will demonstrate this year the Petaflop.

## CEA – FZJ (1/2)

- BULL - INCA prototype
- Compute configuration:
  - Intel Nehalem-EP processors
  - > 100 dual sockets blades
  - > 800 cores
  - 10 Tflops / > 2 TB of memory
- Network:
  - IB QDR or DDR
- IO configuration:
  - Lustre
  - > 20 TB
- Harpertown based system already installed
- Possibility to extend with Tesla servers for GPU acceleration
- JUROPA2 (production system)
- Compute configuration
  - Intel Nehalem-EP processors
  - 2048+1024 thin nodes
  - 16384+8192
  - 300 Tflops / 72TB of memory
- Network:
  - QsnetIII (or IB)
- IO configuration
  - Lustre
  - 250 TB



*Transparent access / sharing of data*

## CEA – FZJ (2/2) Main features

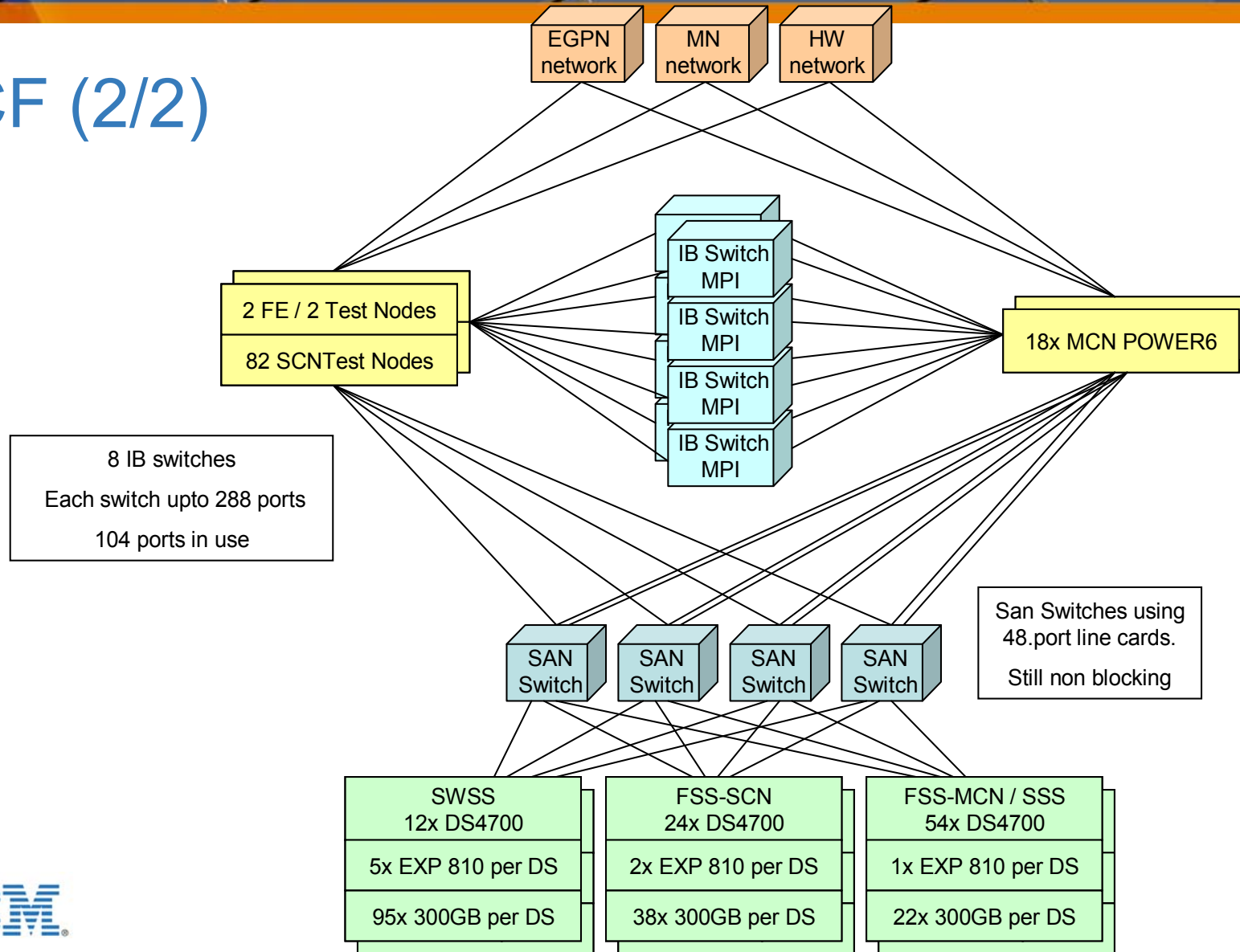
- Based on standard hardware and software technologies
  - Relevant for the largest panel of applications
  - Cost effective technologies for highly productive system
- Advanced architecture :
  - High density relevant for Petascale system
  - No compromise on memory (3GB/core) and interconnect (fastest non proprietary technology)
  - Flexible IO architecture
  - Optimized software stack (MPI library, scalability, management)
- Cooperation with largest HPC engineering team in Europe
  - Cooperation with Bull competence centres
    - Hardware design
    - System software development
    - Application optimization

## NCF (1/2)

- IBM Power6 prototype:
  - Compute configuration
    - IBM Power6 processor (4,7 GHz)
    - 32 cores/node
    - 4-8 GB shared memory/core
    - 104 nodes, 3328 cores
    - 60 TF peak, ~50 TF Linpack
  - Low latency interconnect:
    - QLogic IB, 8-way star topology
  - I/O configuration
    - IBM GPFS file system
    - 4 GB FC SAN, 700 TB net disk storage
  - Operating system: Linux SLES10
  - Very high density (>50kW/rack), water cooled nodes
- PRACE:
  - Access to full production system (5%)
    - Full feature set for all assessments
  - Additional 2-4 nodes for early of testing software from US DARPA/PERCS project assessing specific petascale issues like:
    - Application environment (compilers, libraries, tools, languages)
    - Reliability (system kernels)
    - Job scheduling
    - System management
    - Power7 performance simulation



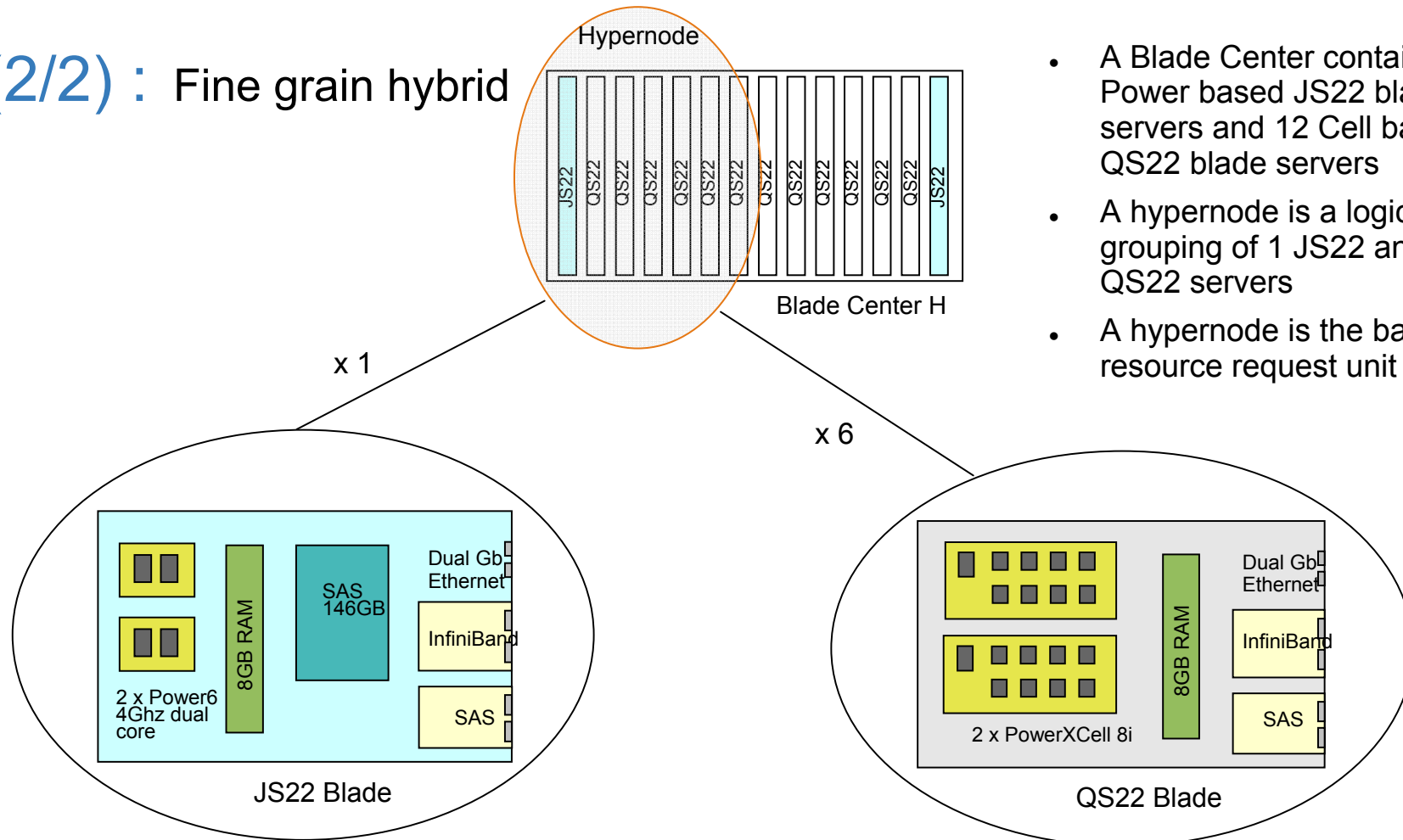
# NCF (2/2)



## BSC (1/2) : Hardware Description

- 12 JS22 Blade servers
  - POWER 6 64 bit 4.0 GHz dual core CPUs
  - 96 Gb RAM total
  - 48 cores
- 72 QS22 Blade servers
  - PowerXCell 8i 3.2 GHz CPUs
  - 864 GB RAM total
  - 1296 cores
- Peak performance
  - 14.4 Tflop/s from 2 rack unit
- Network
  - 4xDDR InfiniBand (16 Gb)
    - MPI, GPFS
  - Gigabit Ethernet
    - Booting, management
    - External services
- I/O configuration
  - JS22 internal disk + external SAS
  - QS22 external SAS
  - GPFS user file space over InfiniBand

# BSC (2/2) : Fine grain hybrid



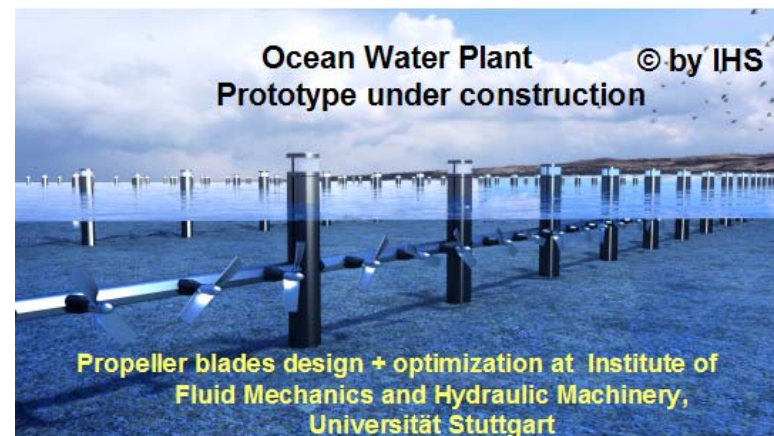
- A Blade Center contains 2 Power based JS22 blade servers and 12 Cell based QS22 blade servers
- A hypernode is a logical grouping of 1 JS22 and 6 QS22 servers
- A hypernode is the basic resource request unit for a job.

JS22 runs management tasks, scalar jobs, I/O intensive jobs.

QS22 is optimised for floating point operations with peak performance of 217 Gigaflops per blade.

## HLRS (1/3) : Why does PRACE need the HLRS/NEC Prototype?

- System is unique in the world
- Coupled multi-physics, multi-scale simulations increasingly important
  - Simultaneous simulation of several aspects at once (e.g. Fluid-Structure, Aero-Acoustics, Combustion Procs)
  - Different aspects require different architectures for optimal use → Hybrid systems
- Examples:
  - Optimization of power plants
  - Emission reduction in production
  - Climate ocean-atmosphere
  - Medical apps / surgery planning
- Hybrid systems imposed challenges
  - Alignment of operating systems
  - Integrated scheduling systems
  - Heterogeneous networks
  - Hybrid compilers & analysis tools
- Programming models/ specific benchmarks for future integrated hybrid systems



## HLRS (2/3) : Why does the HLRS/NEC Prototype need PRACE?



- Visibility of PRACE
  - Support development of highly innovative configurations (e.g. Linux for hybrid system, I/O forwarding, hybrid configurations)
  - Better negotiation position
- Knowledge and expertise of PRACE partners
  - Investigation of different approaches for future systems  
→ risk reduction
  - Evaluation of benchmarking applications for future simulation challenges
  - Designing a hybrid system requires knowledge of single systems as well as integration techniques
    - Results of evaluation of single systems by partners help designing future hybrid systems
    - Development of extendable hybrid approach
  - Further refinement of “system of systems” approach

## HLRS (3/3) : Additional aspects of the HLRS prototype

- Energy efficient
  - Applications can use the best suited architecture for their problem, even distinct for distinct parts → No waste of CPU time → no waste of energy
  - Vectorization as programming model is energy efficient
- Industrially relevant applications
  - Coupled multi-physics / multi-scale applications of increasing relevance
  - “System of systems” approach allows coupling of in-house codes with ISV codes
  - Addresses poor availability of ISV codes for high-end systems
- Allow for continuous system update
  - Open architecture allows the addition of new architecture types and to continuously upgrade the system

	1xIA64	1xSX	1xSX + 1xIA64
UNSTRUCT:	2.993,67	<b>7.745,82</b>	3.019,24
STRUCT:	<b>23.887,32</b>	2.870,66	2.869,46
Coupling/Steering:	1.012,37	321,15	554,21
waiting:	0,00	0,00	164,20
Coup calculating time:	1.012,37	321,15	390,02
Total CPU time:	27.893,35	10.937,62	6.278,72
Total elapsed (sec):	27.924,78	10.966,23	3.207,09
Total elapsed (h):	7:45'	3:03'	0:53'
Relative Price	1	1,57	0,58

## About PRACE

- The aim of PRACE is to provide scientists in Europe with unlimited and independent access to fast supercomputers and competent support. PRACE prepares the creation of a persistent pan-European HPC service, consisting several tier-0 centres providing European researchers with access to capability computers and forming the top level of the European HPC ecosystem. PRACE is a project funded in part by the EU's 7th Framework Programme. The following countries collaborate in the PRACE project: Germany, UK, France, Spain, Finland, Greece, Italy, Ireland, The Netherlands, Norway, Austria, Poland, Portugal, Sweden, Switzerland and Turkey. The PRACE project is coordinated by the Gauss Centre for Supercomputing (Germany), which bundles the activities of the three HPC centres in Jülich, Stuttgart, and Garching.
- <http://www.prace-project.eu/>
- The PRACE project receives funding from the EU's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° RI-211528.



PARTNERSHIP  
FOR ADVANCED COMPUTING  
IN EUROPE

END