MariCel: The PRACE prototype system at BSC

Gabriele Carteni – Barcelona Supercomputing Center

NSC'09 - PRACE Code Porting Workshop: SNIC Interaction
13-14 October 2009 - Linköping University (Sweden)

CAPACITIES

e-infrastructure

Introduction

- Concurrency and Specialization paradigm
  - Multicore architectures
  - General-Purpose computation on GPU (many-core or hybrid architecture)

- The Cell Broadband Engine Architecture
  - Heterogeneous multicore architecture
  - On chip hardware accelerators

- MariCel, the PRACE prototype system at BSC
  - Prototype anatomy
  - Services provided to the HPC community

Concurrency and Specialization

- The free ride for faster clock frequencies stopped

- Increase concurrency and specialization: the new paradigm
  - Multiple processing units on the same chip @ reduced clock frequency
  - New memory hierarchical model: more cache levels @ different visibility
  - High bandwidth on-chip interconnection bus

- Different design and implementations
  - Homogeneous and heterogeneous multicore architectures
  - General Purpose computation on GPU (GPGPU/Hybrids)

The IBM Cell/B.E. Architecture

- Defined as a **heterogeneous multicore architecture**
- Use a dual-threaded power processor (PPE) running OS and tasks dispatching to specialized co-processors (SPE)
- Each co-processor implements SIMD capabilities with a private/local memory accessible at application level
- RDMA capabilities for accessing central memory from SPEs
- Low power consumption: < **1KW**
- PPC64 Compatible Architecture

The PowerXCell™ 8i Processor

Power Processor Element (PPE)

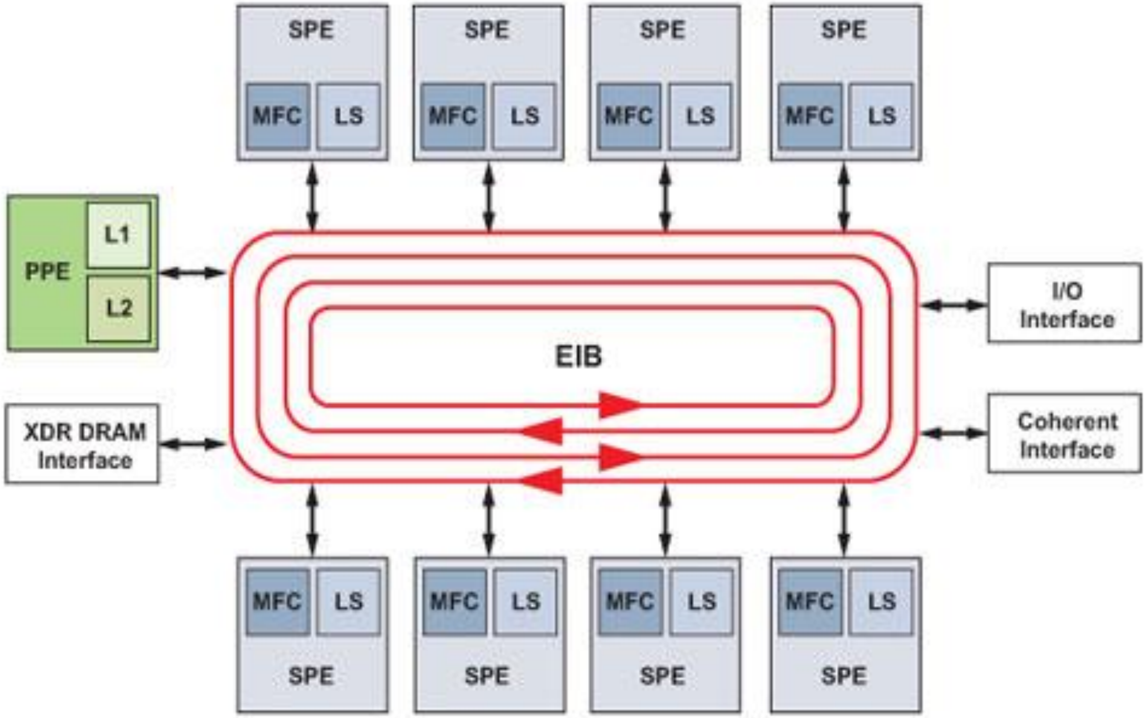
- 32 KB instruction cache, 2-way set associative
- 32 KB data cache, 4-way set associative, write-through
- 512 KB 8-way set-associative store-in L2 cache
- 64 GBps load-and-store bandwidth

Synergistic Processing Element (SPE)

- SIMD capability
- SP FP throughput of 25.6 GFLOPS
- DP FP throughput of 12.8 GFLOPS
- 256 KB local store memory
- 128-byte direct memory access (DMA) read and write

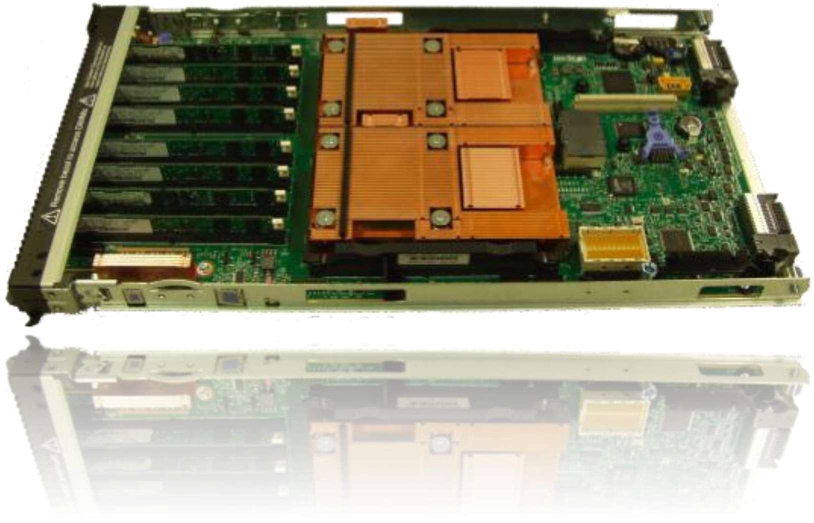
Bus Subsystem (EIB)

- fast on-chip bus connecting all the elements at ~25.6GB/s
- I/O bandwidth: 35GB/s (in) 40GB/s (out)

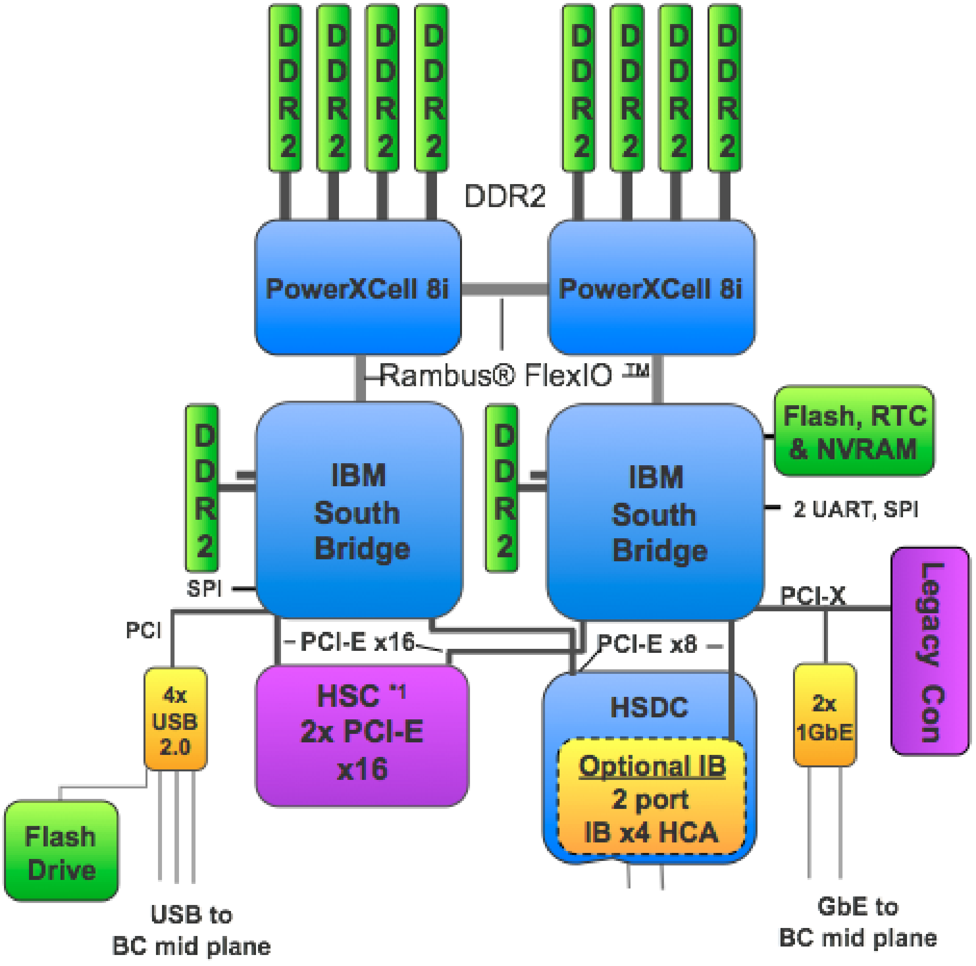


The IBM BladeCenter™ QS22

- Building block for MariCel prototype
- 2 PowerXCell 8i @3.2GHz
- SP/DP peak perf: 460/217 Gflops
- 32GB DDR2 800MHz
- Dual GbE interface
- 4x DDR InfiniBand Host Card Adapter
- SAS expansion card



The IBM BladeCenter™ QS22



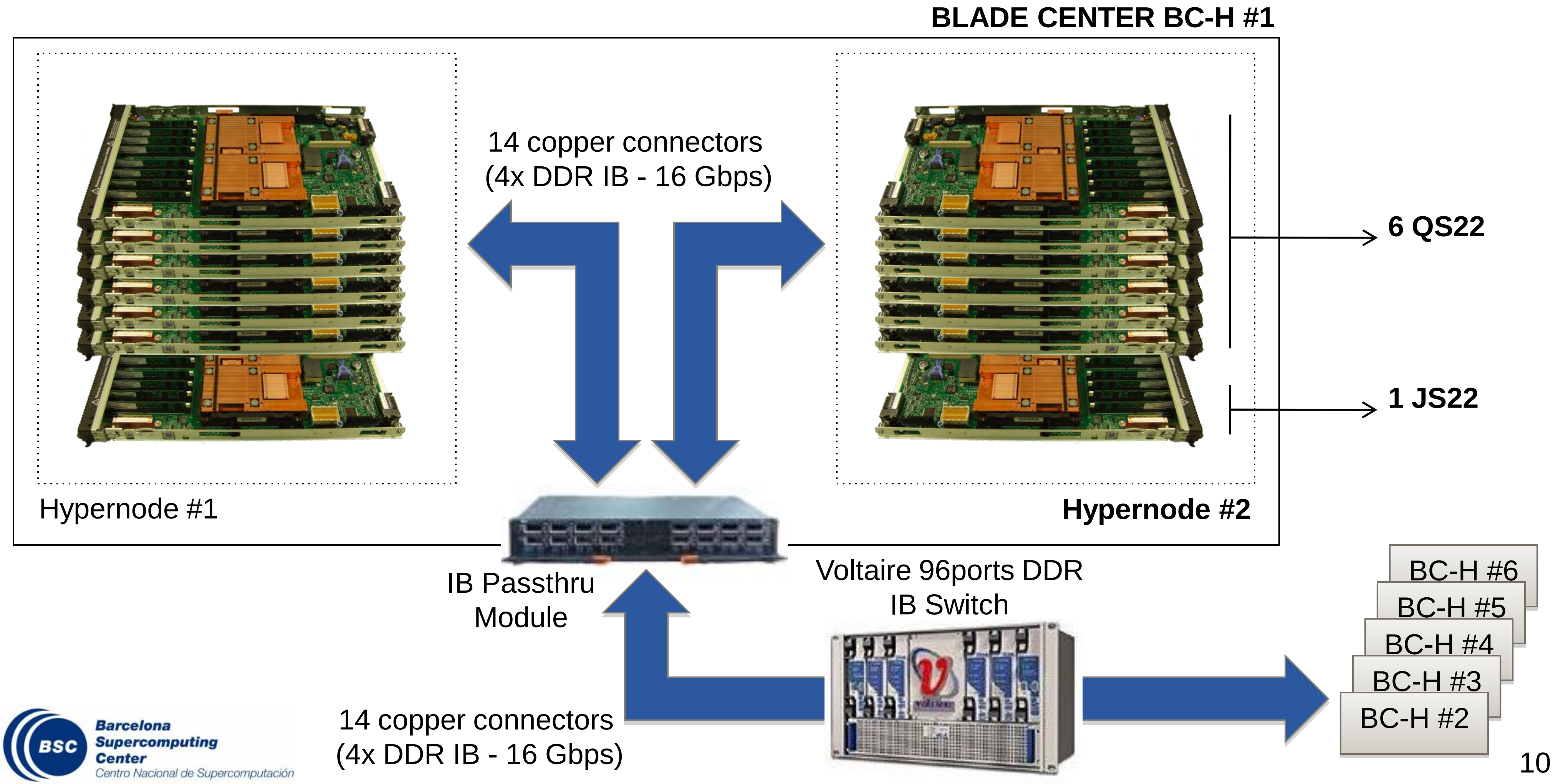
Prace Cell Prototype at BSC: MariCel

- “*mar i cel*” means “*sea and sky*” in Catalan
- Intended for:
  - Testing high performance computing on Cell
  - CPU-intensive application to exploit on-chip parallelism
  - Benchmarking for future petascale systems
  - Taking part as a Tier-0 computing node in the PRACE Research Infrastructure
  - Extending system management techniques to heterogeneous architecture

Prace Cell Prototype at BSC: MariCel

- 6 Blade Center H Chassis
- (12 QS22 + 2 JS22) x 6 BC-H → **72 QS22 + 12 JS22**
- Total of 84 nodes / **1344 cores**
- 960 GB of total memory
- InfiniBand 4x DDR for MPI applications @**16Gbps**
- Peak/Linpack performance: **15.6/10.1 Teraflops**
- 1 TB of GPFS Global Filesystem
- SAS attached disk for fast scratch storage
- Average energy consumption: ~**20kW**
- Energy Efficiency: **500 MFlops / W** (LANL Roadrunner @437)

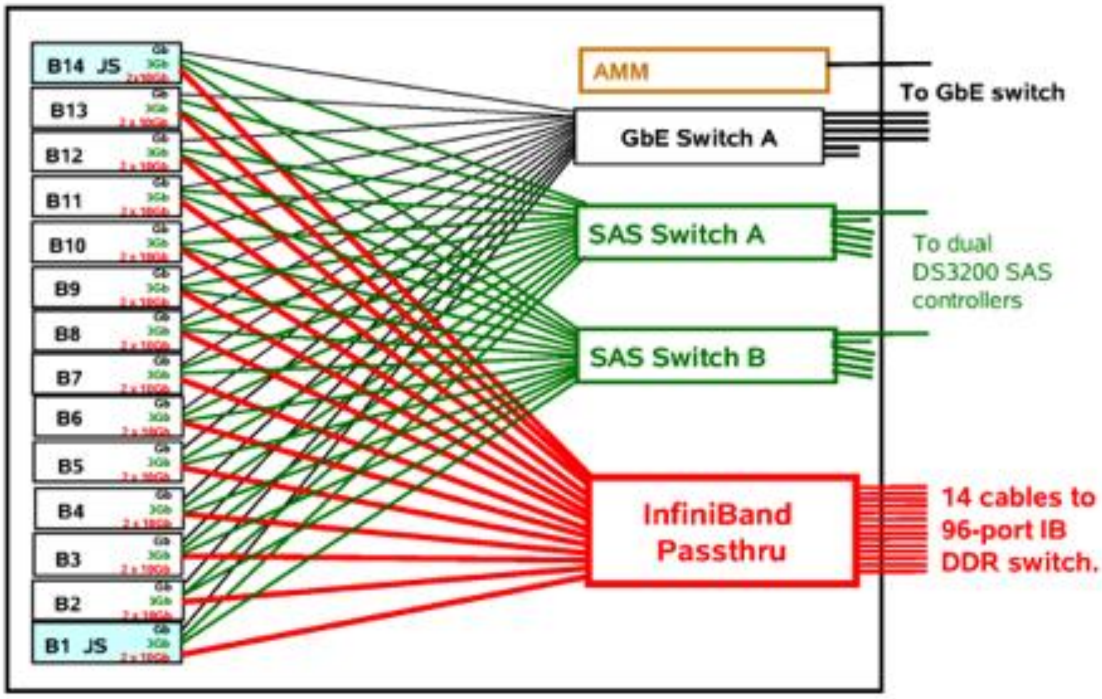
MariCel: InfiniBand network



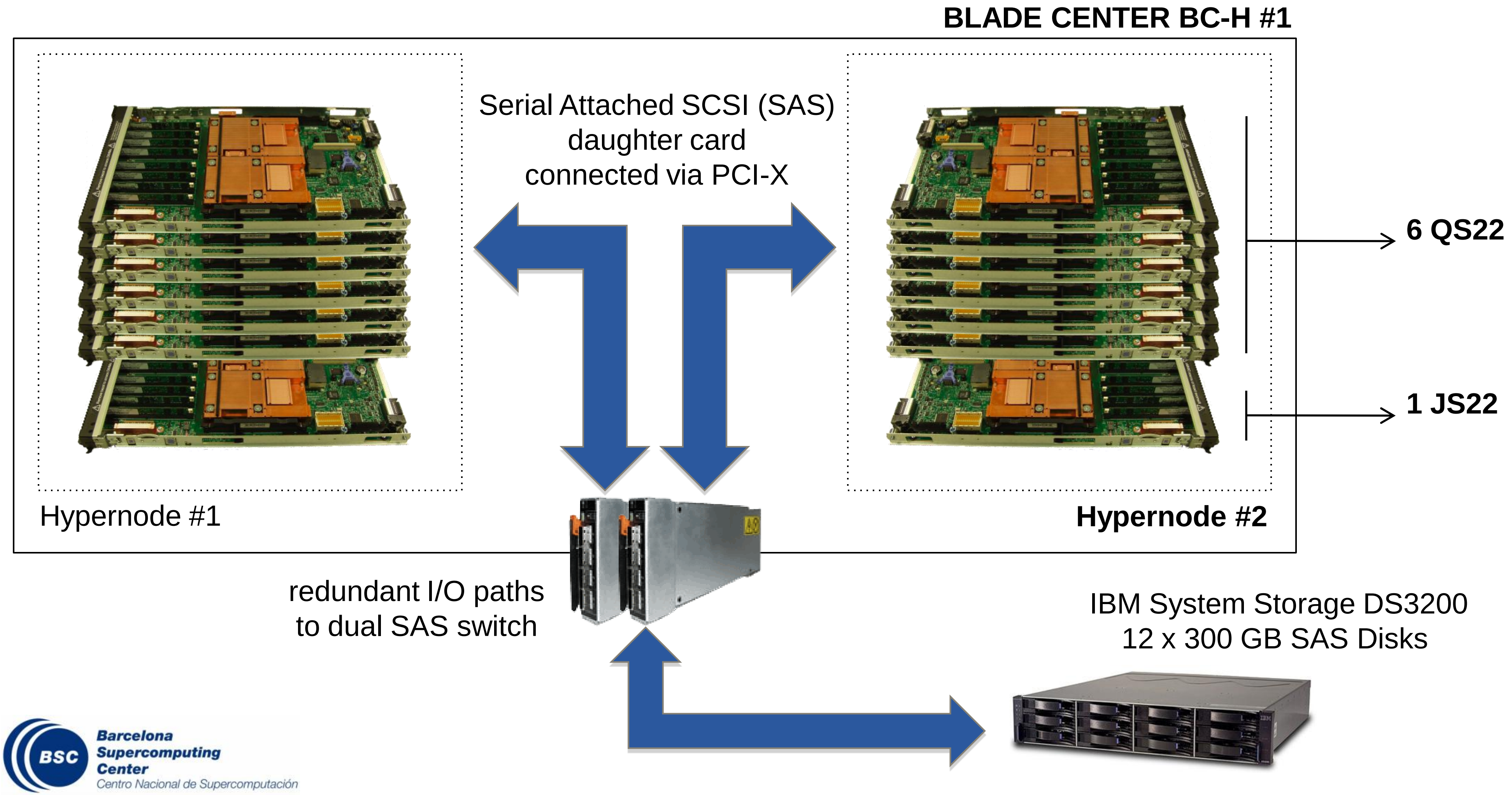
MariCel: BC-H networks

B#n

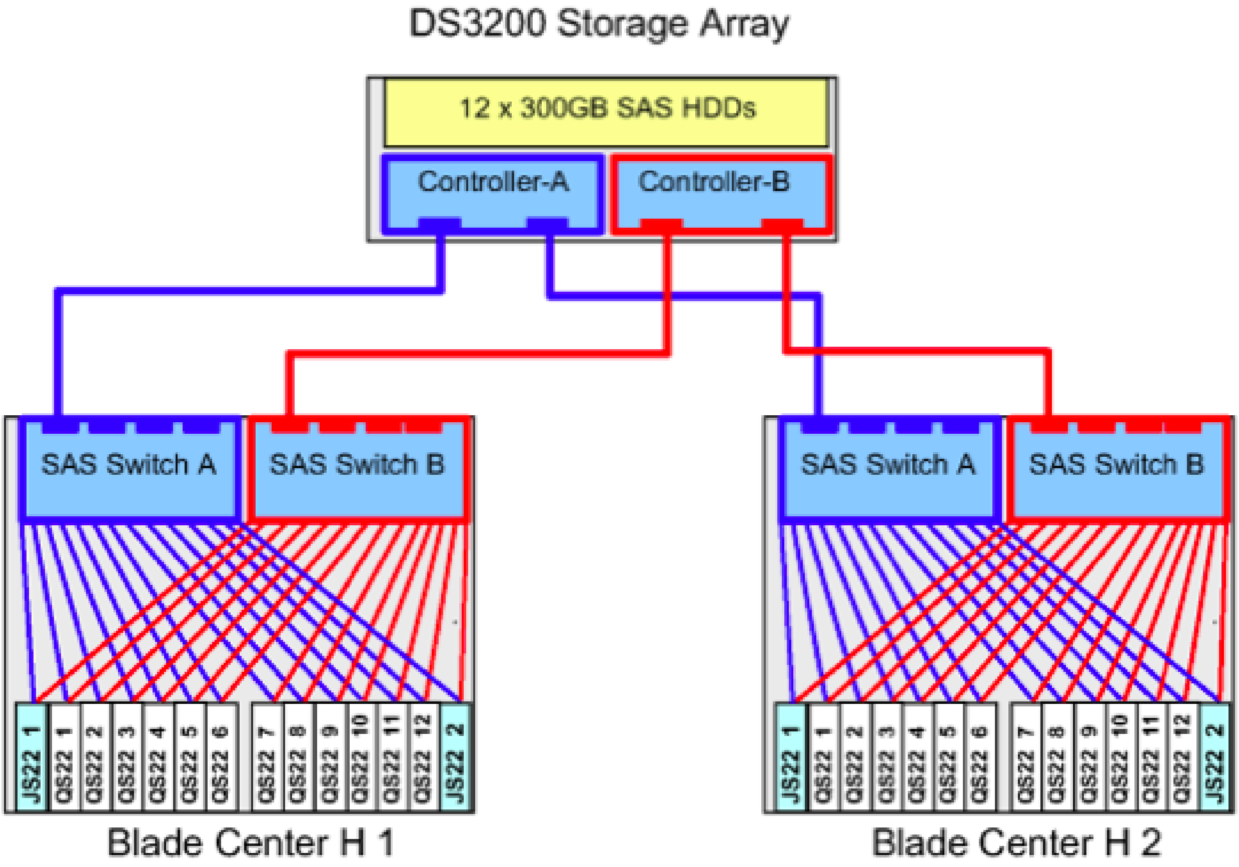
1 Gbps
3 Gbps
16 Gbps



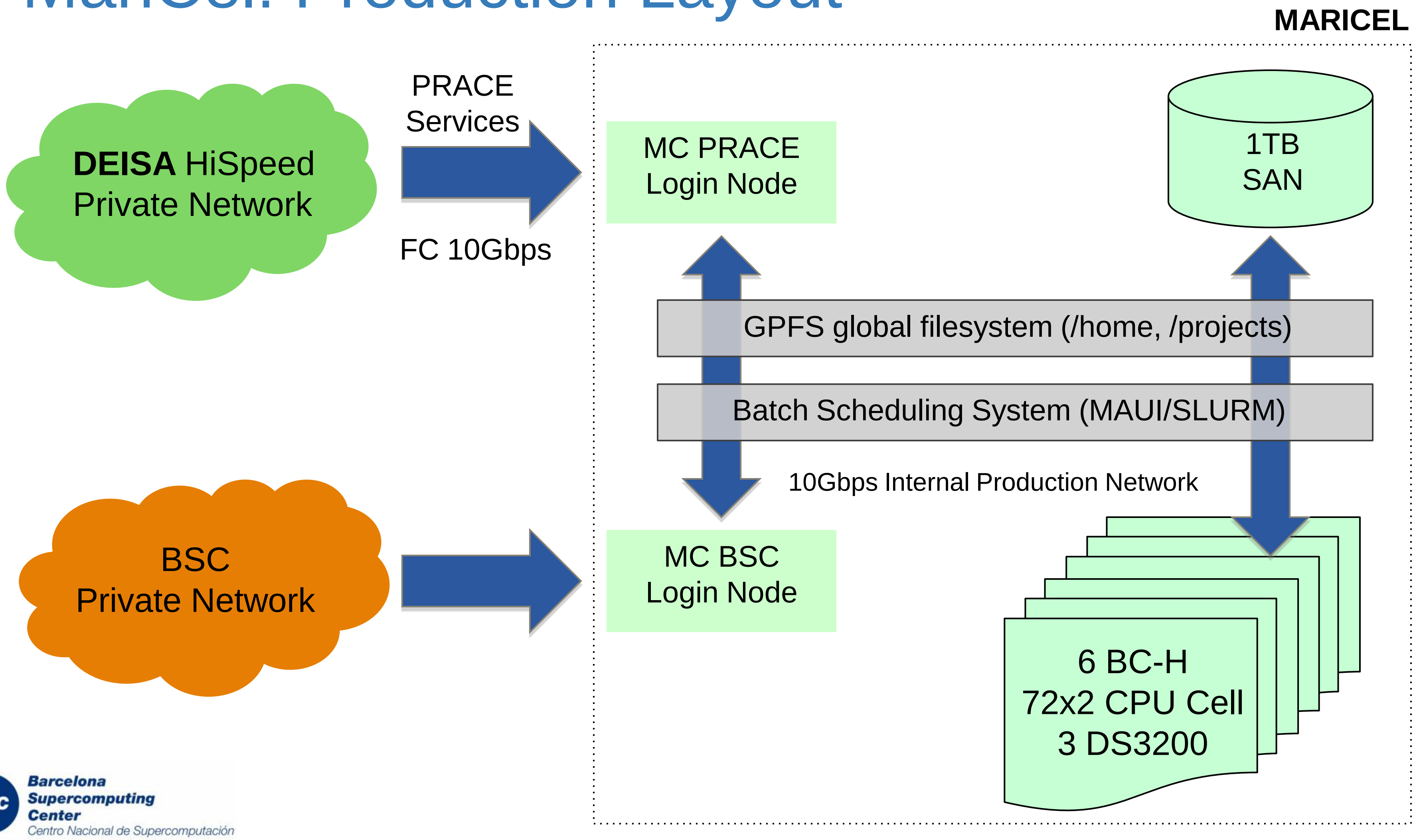
MariCel: SAS attached storage


BLADE CENTER BC-H #1
Serial Attached SCSI (SAS) daughter card connected via PCI-X
6 QS22
1 JS22
Hypernode #1
Hypernode #2
redundant I/O paths to dual SAS switch
IBM System Storage DS3200
12 x 300 GB SAS Disks

MariCel: SAS attached storage



MariCel: Production Layout


MARICEL
DEISA HiSpeed Private Network
PRACE Services
FC 10Gbps
MC PRACE Login Node
1TB SAN
GPFS global filesystem (/home, /projects)
Batch Scheduling System (MAUI/SLURM)
10Gbps Internal Production Network
BSC Private Network
MC BSC Login Node
6 BC-H
72x2 CPU Cell
3 DS3200

Prace Cell Prototype at BSC: Services

- System Software
  - RedHat Enterprise Linux 5.3 OS
  - IBM Virtual I/O System on AIX 5.3 for virtualized services
  - OpenMPI 1.3.2 / OFED 1.4.1
  - MAUI/SLURM Scheduling System and Resources Allocation Manager
  - Diskless Image Management
  - Ganglia Monitoring
  - Distributed infrastructure Services (PRACE)
  - UNICORE 6
  - Globus services (GridFTP+GlobusRFT, GSI-SSH)
  - PRACE Accounting Management
  - INCA services monitoring tool
  - MODULE for PRACE User Production Common Environment (testing)

Prace Cell Prototype at BSC: MariCel

- MariCel is one of the 6 selected T0 PRACE prototype
- Synthetic benchmark evaluation (PRACE-WP5)
  - Floating Point rate of execution of double precision problems
  - MPI performance
  - Sustainable memory
  - Memory latency and bandwidth
  - Network Communication Capacity (Infiniband)
  - I/O read/write access rate on shared file system and fast local storage
  - Benchmark on real applications (PRACE-WP5)
  - Ready for a permanent pan-European HPC service and infrastructure for PRACE users (PRACE WP4)

Programming On Cell: BSC and IBM make it easy

- BSC Cell Superscalar framework (CellSs)
  - http://www.bsc.es/cellsuperscalar
  - Source to source compiler and a runtime library
  - Source code annotation to define directives
  - Automatic exploitation of functional parallelism to use SPEs at execution time
  - Simple and flexible programming model
  - Locality-aware task scheduling support

- IBM SDK for multicore acceleration v.3.1
  - http://www.bsc.es/projects/deepcomputing/linuxoncell
  - Optimized compilers (XL Compilers for C, C++, Fortran)
  - Rich set of scientific libraries
  - Support to GNU compilers (ppu-gcc, spu-gcc)
  - Performance and Debugging Tools
  - Cell Simulator
  - Integration to the Eclipse Development Environment

Q&A

Special notices and Copyrights:
This presentation contains images and contents protected by © Copyright IBM Corporation