Présentation Blue Gene/P CADMOS

Christian Clémençon

christian.clemencon@epfl.ch

24 septembre 2009

Agenda

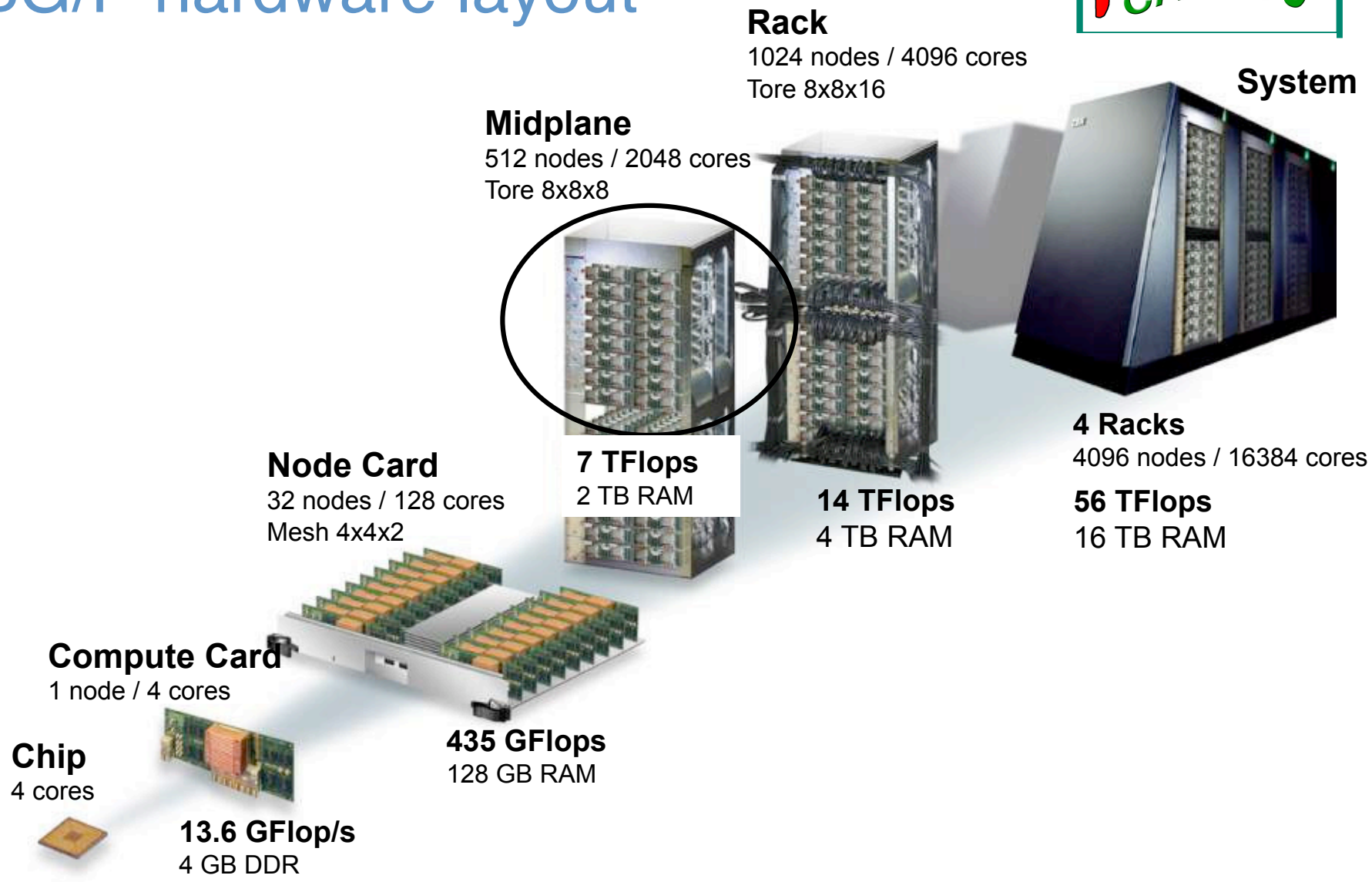
- Caractéristiques principales du Blue Gene/P
- Aperçu du hardware
- Différences avec le Blue Gene/L
- Parallélisme et partage des ressources
- Environnement de développement
- Systèmes de fichiers et sauvegardes
- Accès et utilisation du système
- Conclusion et réponse aux questions

BG/P main characteristics

- IBM Blue Gene/P Massively Parallel Computer
- 4 racks, one row, wired as a 16x16x16 3D torus
- 4096 quad-core nodes, PowerPC 450, 850 MHz
- Energy efficient, water cooled
- 56 Tflops peak, 46 Tflops LINPACK
- 16 TB of memory (4 GB per compute node)
- 1 PB of disk space, GPFS parallel file system
- OS Linux SuSE SLES 10

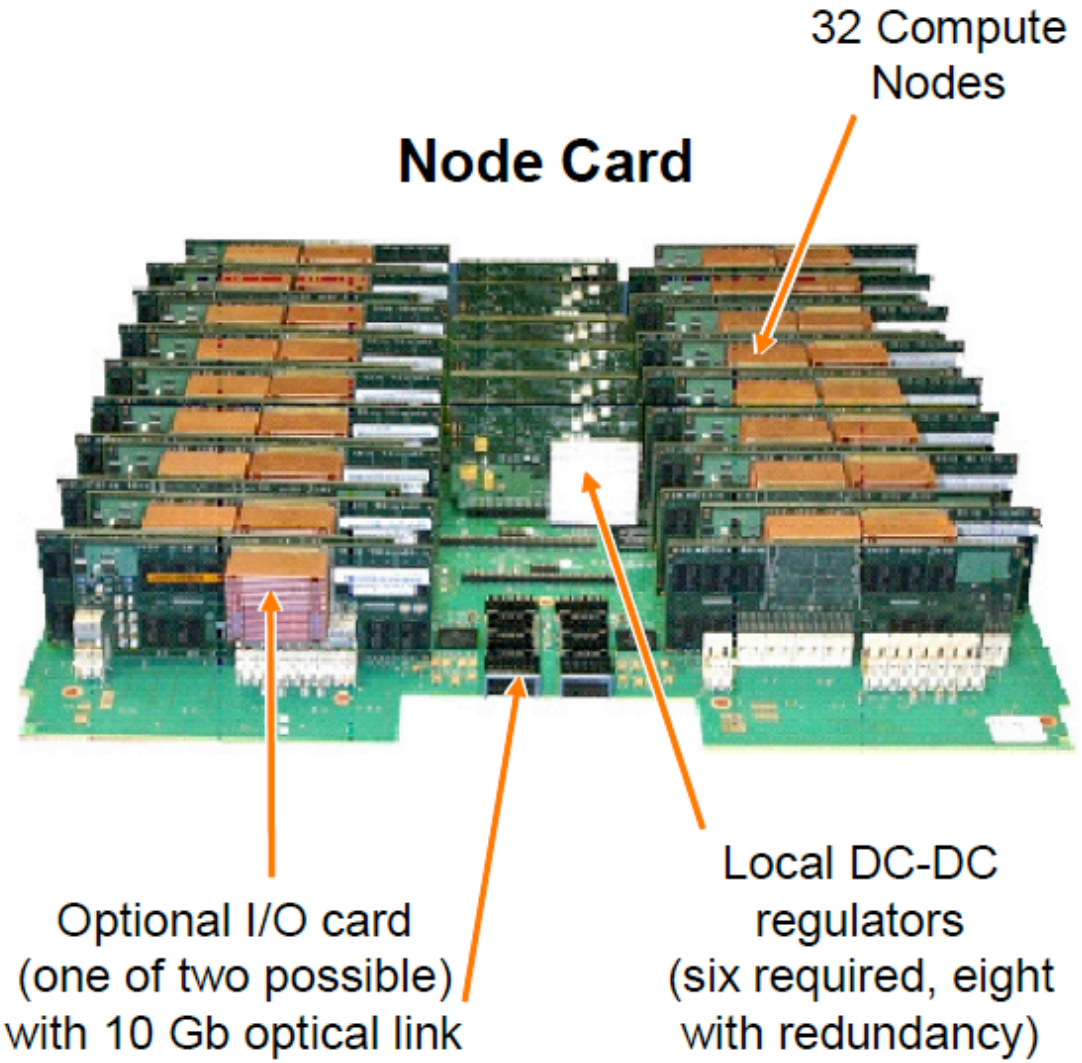
APERÇU DU HARDWARE

BG/P hardware layout

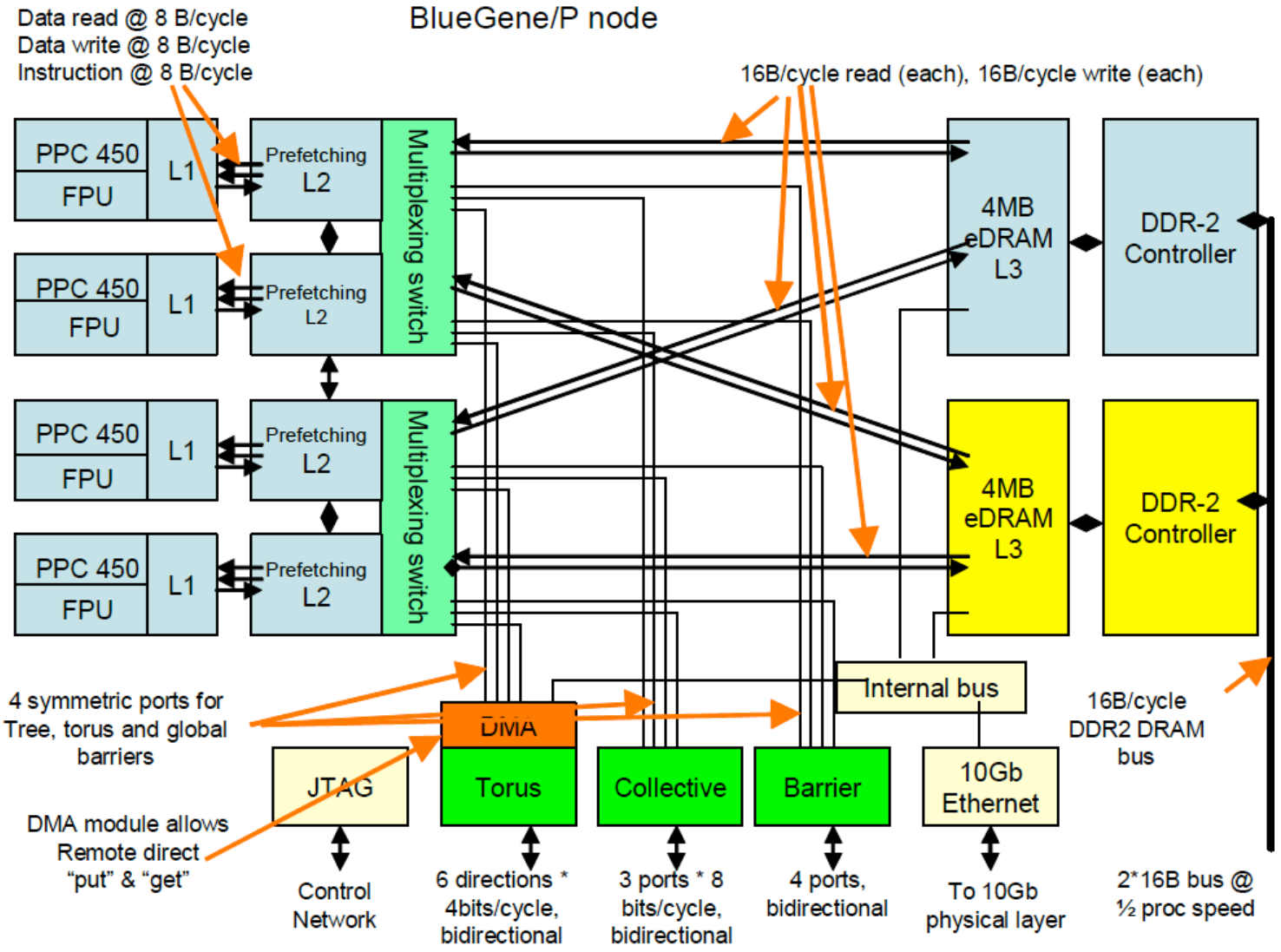


CADMOS BG/P racks

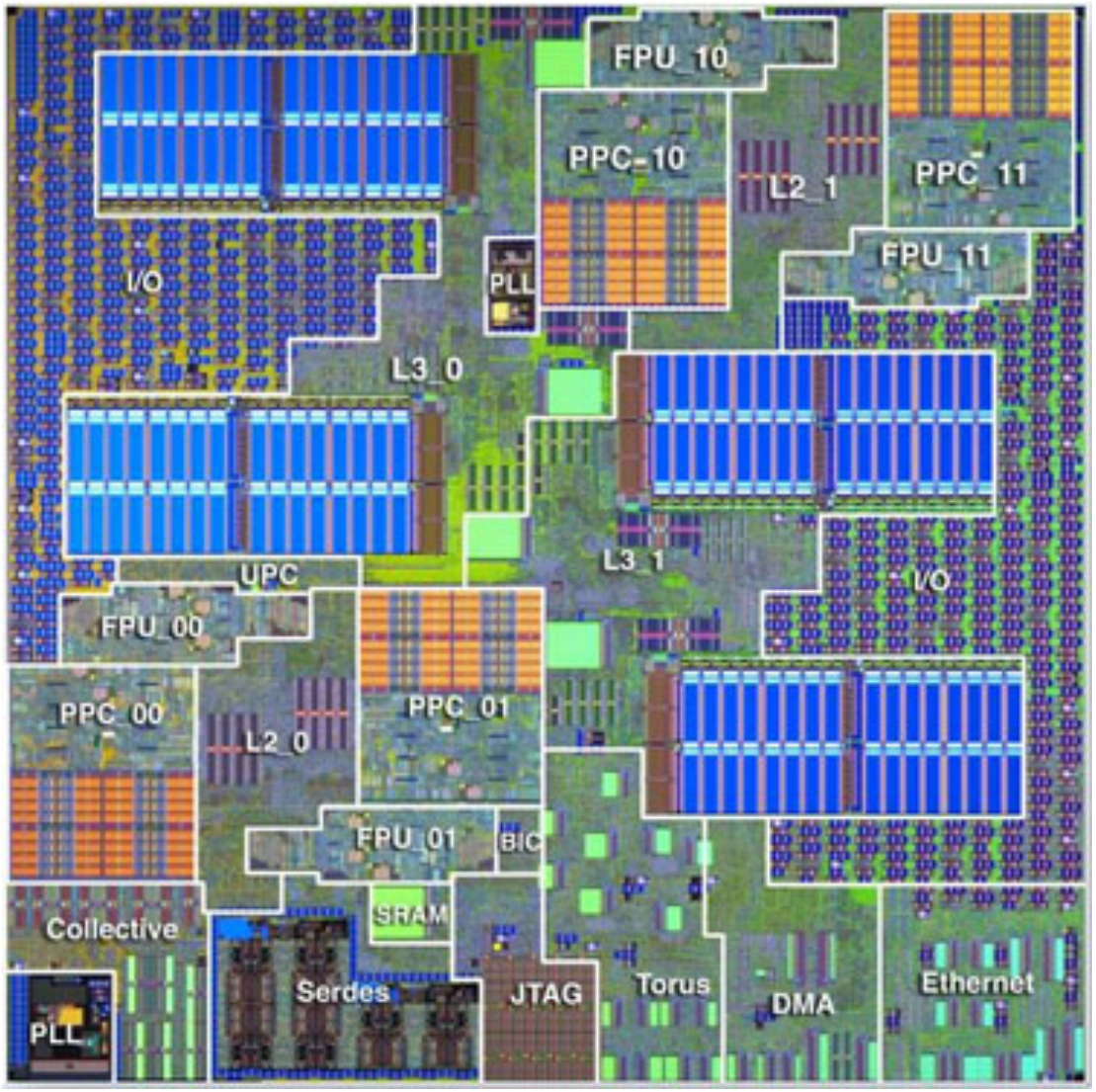
Node Card


32 Compute Nodes
Node Card
Optional I/O card (one of two possible) with 10 Gb optical link
Local DC-DC regulators (six required, eight with redundancy)

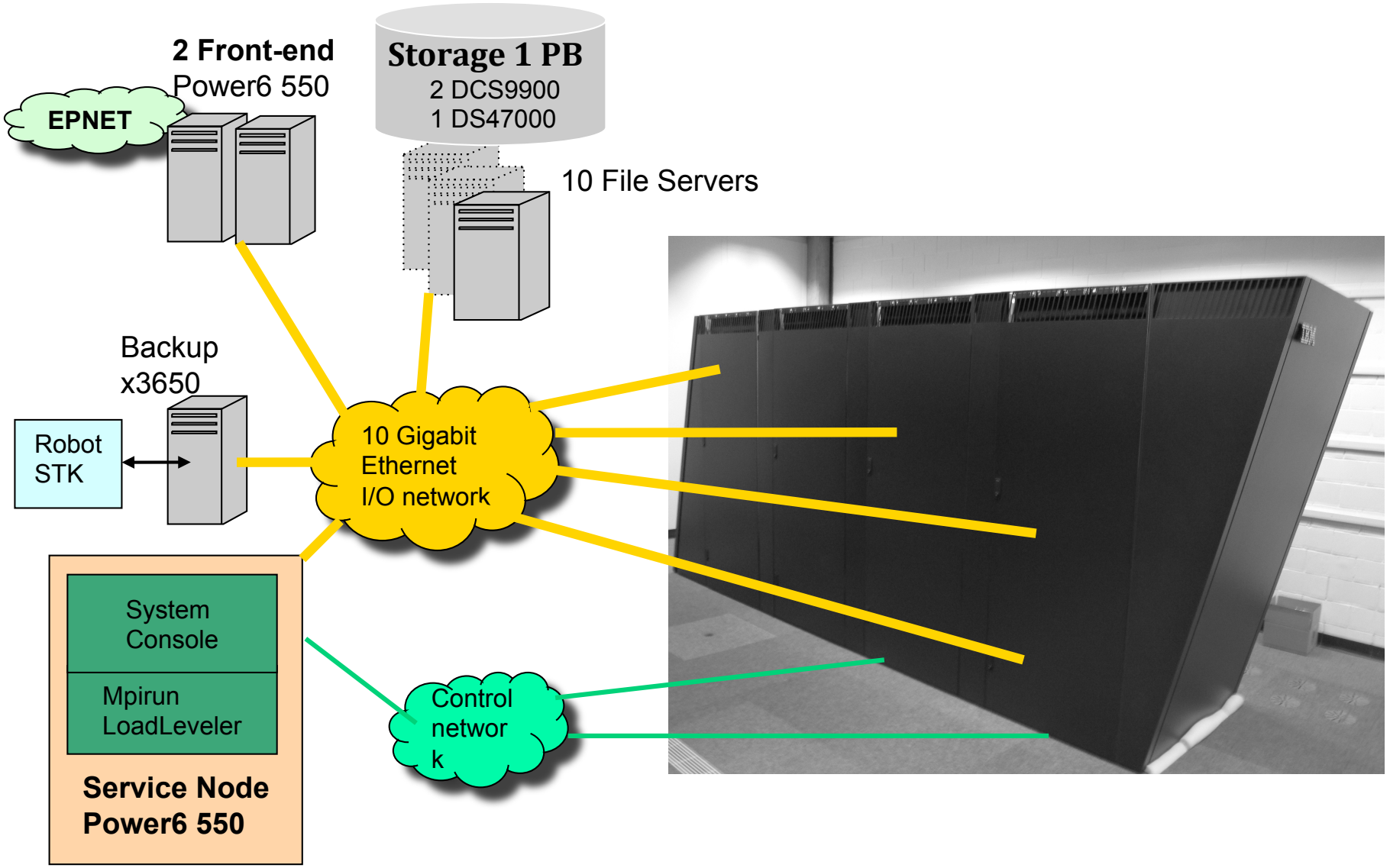
PowerPC 450 ASIC



PowerPC 450 ASIC



CADMOS BG/P system architecture



Front-ends

- Two IBM System-p hosts
  - bg1.epfl.ch (open to the outside world)
  - bg2.epfl.ch (accessible within EPFL network only)
- Configuration
  - 8 Power6 cores 4.2 GHz
  - 32 GB of memory
- Login
  - Via ssh only
    - % ssh username@bg1.epfl.ch

DIFFÉRENCES AVEC LE BLUE GENE/L

Différences entre BG/L et BG/P

	Blue Gene/L	**Blue Gene/P**
Système EPFL	4 racks, 4096 nœuds 8192 cœurs	4 racks, 4096 nœuds 16384 cœurs
Mémoire principale	2 TB	16 TB
Perf. max / LINPACK	23 / 18 TFlops	56 / 47 TFlops
Efficacité énergétique	208.31 MFlops/Watt	371.67 MFlops/Watt
Nœud		
Cœurs	2 x PowerPC 440	4 x PowerPC 450
Fréquence CPU	700 MHz	850 MHz
L3 cache	4 MB	8 MB
Mémoire principale	512 MB - 5.6 GB/s	4 GB - 13.6 GB/s (2 x 16 bytes)
Réseaux		
Point-à-point en tore 3D	2.1 GB/s (6 x 2 x 0.175)	5.1 GB/s (6 x 2 x 0.425) - 3.5 us
Collectif en arbre	2.1 GB/s (3 x 2 x 0.350)	5.1 GB/s (3 x 2 x 0.850) - 2.5 us
E/S système de fichiers	256 x 1 GbE	56 x 10 GbE fibre optique

New hardware features

- Cache coherence
  - In hardware at node level (SMP mode)
- DMA
  - Speedup packet transfer in 3D torus network
  - Overlapping of communications and computations
  - Direct node-to-node memory copy (put – get)
- Shared memory
  - POSIX like

PARALLÉLISME ET PARTAGE DES RESSOURCES

BG/P parallel computer architecture

- MPP (Massive Parallel Processing)
  - Distributed memory architecture
  - 4096 computing nodes
  - Tightly interconnected
  - MPI based applications
  - SPMD parallel jobs, started via mpirun command
- HPC (High-Performance Computing) paradigm
  - Applications make use of the network to share data among multiple MPI tasks

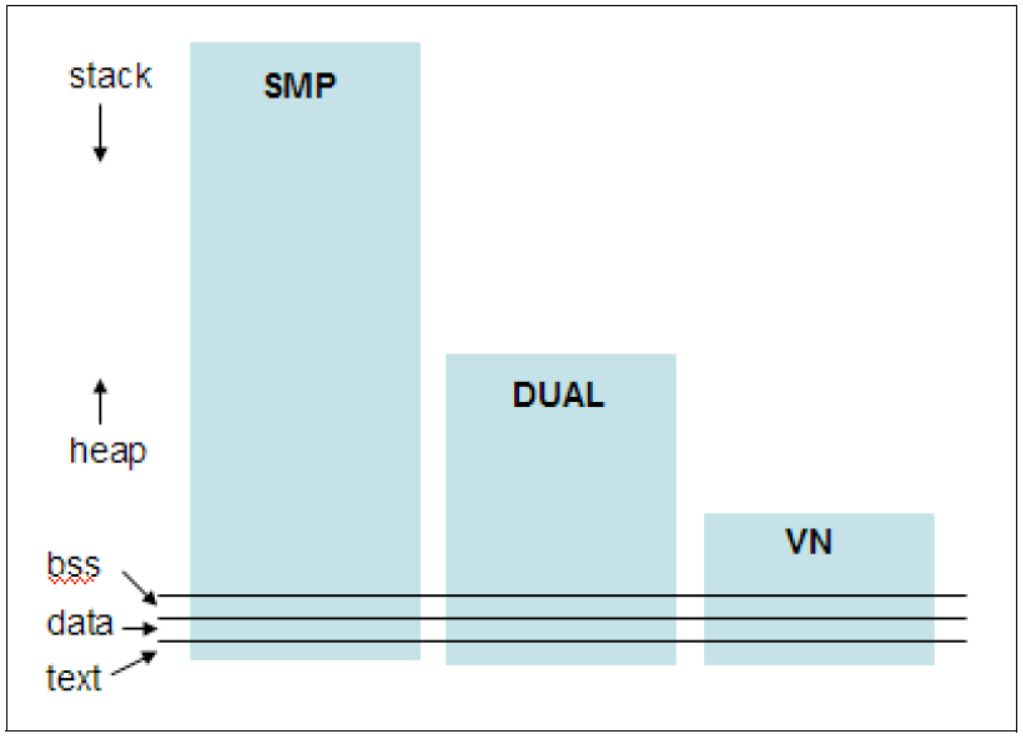
Node level parallelism

- Three mpirun execution modes
  - SMP (Symmetrical MultiProcessing)
    - One MPI task per node, max of 4 threads per task (*1X4*)
    - Each thread is pinned to a physical processor (core)
    - Shared-memory parallelism, via pthread and/or OpenMP
  - VN (Virtual Node)
    - Four MPI tasks per node, one single thread per task (*4x1*)
    - Each MPI task is pinned to a separate core
  - DUAL
    - Two MPI tasks per node, max of 2 threads per task (*2x2*)

Memory considerations

- 4 GB of physical memory per node
  - SMP mode : all memory available to the MPI task running on each node
  - DUAL mode : ½ of the memory available to each MPI task
  - VN mode : ¼ of the memory available to each MPI task
- No virtual paging mechanism (memory swap)
  - Watch for memory leaks !

MPI task address space



- text Application code
- data Initialized static and common variables
- bss Uninitialized static and common variables
- heap allocatable data (brk, malloc system calls)
- stack Automatic data

Computing resources sharing

- Resources are partitioned in physical blocs of nodes
  - Smallest block has 32 nodes
  - Largest block spans the entire system (4096 nodes)
- Blocks are hierarchically organized
- Blocks can’t overlap
- A block is allocated to a single user at a time
- When a user starts a job on a partition, he will keep the block for the entire job duration

Block naming convention

- R0 (entire machine)
  - R00 (I/O rich rack)
    - R00-M0 (test mid-plane)
      - R00-M0-0 (128 nodes)
        - R00-M0-00 (32 nodes)
        - R00-M0-01
        - R00-M0-02
        - R00-M0-03
      - R00-M0-1
      - R00-M0-2
      - R00-M0-3
    - R00-M1
  - R01 (second rack)
    - R01-M0
    - R01-M1
  - R02 (third rack)
    - R02-M0
    - R02-M1
  - R03 (fourth rack)
    - R03-M0
    - R03-M1

ENVIRONNEMENT DE DÉVELOPPEMENT

IBM RedBook


IBM
IBM System Blue Gene Solution: Blue Gene/P Application Development
Understand the Blue Gene/P programming environment
Learn how to run and debug MPI programs
Learn about Bridge and Real-time APIs
Carlos Sosa
Brant Knudson
ibm.com/redbooks
Redbooks

Software development

- Languages
  - C, C++ and Fortran
  - GNU and XL compilers
  - Python
- Libraries
  - MASS (Mathematical intrinsic functions)
  - ESSL (Engineering and Scientific Subroutine Library)
  - Serial & Parallel HDF5
  - LAPACK, ScaLAPACK, FFTW, Global Array, etc

ESSL

- Engineering and Scientific Subroutine Library
  - http://www.ibm.com/systems/p/software/essl.html
- 150 math subroutines tuned for BG/P
  - Linear Algebra Subprograms
  - Matrix Operations
  - Linear Algebraic Equations
  - Eigensystem Analysis
  - Fourier Transforms
  - Random Number Generation
  - Etc …

IBM XL Compilers

- XLF 11.1 / VACPP 9.0
  - /opt/ibmcmp/xlf/11.1/bin
  - /opt/ibmcmp/vacpp/bg/9.0/bin
- Main compiler options
  - -qarch=450, or –qarch=450d, -qtune=450
- Wrappers for compiling MPI programs
  - mpixlc, mpixlC, mpixlf77, mpixlf90, etc
- BG/P specific versions
  - Compiler prefix is bg (bgxlf, bgxlc, bgcc, etc)

IBM XL Compilers (cont'd)

- Thread safe versions
  - bgxlf_r, bgxlc_r, bgxlC_r, etc
  - Support OpenMP
- Libraries
  - static
  - shared & dynamically loaded (new)
- BG specificities
  - dual floating point units (double hummer)
  - SIMD instructions

Programming models

- Pure distributed memory programming
  - MPI
- Hybrid
  - MPI across nodes
  - OpenMP, or pthreads within nodes

BG/P MPI specificities

- MPI-2
  - Derived from MPICH2
- Support for BG/P hardware
  - Point-to-point and multicast (subcommunicator)
    - Torus 3D network
  - Global communication & I/O
    - Collective network, tree topology
  - Barrier
    - Global interrupt network

BG/P MPI specificities (cont'd)

- DMA (New)
  - MPI_Put and MPI_Get
  - Improve point-to-point communications
- MPI-IO
  - Rich set of optimized IO routines
- Can be fine tuned in many different ways

SYSTÈMES DE FICHIERS ET SAUVEGARDES

Storage systems

- Two storage systems
  - data SAN : DataDirect Networks DCS9900
    - 1200 x 1TB SATA disks in two full racks
  - metadata SAN : IBM DS4700 technology
    - 96 x 450 GB FC disks in half a rack
- 10 System-x file servers
  - Each with 2 Xeon quad-core processors
  - 24 GB of memory

Storage racks

File systems

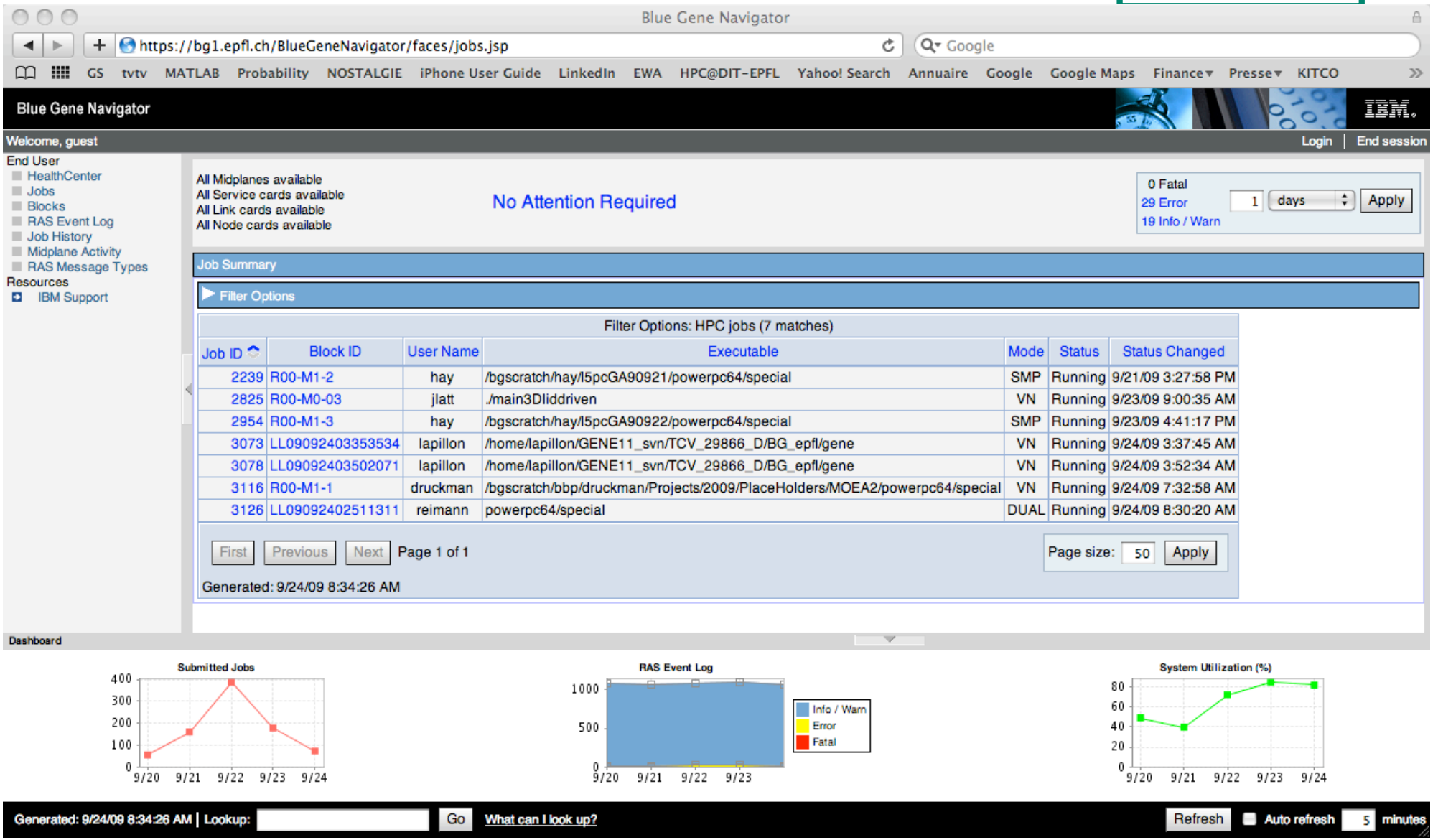
- IBM GPFS (General Parallel File System)
  - /home file system
    - 64 TB, 8 Raid6 arrays, 1 GB/s
    - read-only from compute nodes
    - read-write from front-ends
  - /bgscratch file system
    - 896 TB, 110 Raid6 arrays, 10 GB/s
    - Read-write from compute nodes and front-ends

Data backup

- Only /home is backed-up every day
- NO backups of /bgscratch: copy vital data off the filesystem as soon as possible
- Use archive facility for long-term storage of files
- Data retention policy for /home:
  - Keep 15 last revisions of a file
  - If file is deleted keep 15 last versions of inactive files for the lifespan of the machine

ACCÈS ET UTILISATION DU SYSTÈME

Blue Gene Navigator



Test, debug mode

- Mid-plane R00-M0 reserved
  - 4 x 128 nodes sub-partitions
  - 4 x 32 nodes sub-partitions
- Max 30 minutes per job
  - Don’t keep partition allocated more than required
- Run job interactively

```
% mpirun -partition R00-M0-00
  -mode VN -cwd /bgscratch/clemenco
  -exe hello.bg
```

Batch System : LoadLeveler

- llclass

Class name	Nb of nodes	Max Wall time	Priority
short128	128	4 hours	10
mid	512	12 hours	20
prod	1024, 2048	24 hours	20

- FIFO & Backfill
  - Wall time has to be specified
  - Specify less than the max wall time to benefit from backfilling
  - Use checkpoint/restart to shorten your job wall time

llsubmit <job_cmd_file>

```
#!/bin/sh
# @ comment = "Run single hello example"
# @ job_name = hello
# @ error = $(job_name)_$(jobid).err
# @ output = $(job_name)_$(jobid).out
# @ environment = COPY_ALL;
# @ wall_clock_limit = 00:15:00
# @ notification = complete
# @ notify_user = christian.clemencon@epfl.ch
# @ job_type = bluegene
# @ bg_size = 1024
# @ class = prod
# @ bg_connection = TORUS
# @ queue

CWD=/bgscratch/${USER}
RUNDIR=`pwd`
mpirun -mode VN -cwd $CWD -exe ${RUNDIR}/hello_c
```

CONCLUSION

Conclusion

- System is operational
- You can apply for a test account
  - Application form : http://hpc-dit.epfl.ch
  - Send email to bg-admin@groupes.epfl.ch
- Blue Gene Web site
  - http://bluegene.epfl.ch
- System administrators
  - christian.clemencon@epfl.ch
  - pascal.jermini@epfl.ch

QUESTIONS ?