Architecture Blue Gene/P

Philippe.Wautelet@idris.fr

CNRS - ***IDRIS***

SMAI 2009 - 25 mai 2009

Sommaire

Problématique et approche Blue Gene/P

Problématique

- Puissance électrique dissipée = $frequence^3$
- Vitesse mémoire vive augmente beaucoup plus lentement que puissance crête processeur (*memory wall*)
- Puissance dissipée par cm^2 limitée par le refroidissement

Solution Blue Gene/P

- Si f/2 => consommation d'1 cœur / 8
- Si 4 cœurs/puce => puissance crête x 2 et consommation / 2

=> efficacité énergétique x 4 (FLOP/s/Watt)

Problématique et approche Blue Gene/P

Avantages

Avantages d'une fréquence CPU basse :

- Diminution de écart débits et latences mémoires
- Diminution de écart débits et latences réseaux

=> machine plus équilibrée

- Consommation électrique par cœur faible
- Augmentation de la puissance crête à consommation électrique constante
- Augmentation de la densité (nb de cœurs par rack)
- Refroidissement conventionnel (air)
- Augmentation de la fiabilité (accrue aussi par technologie SoC et tests de validation)

=> puissance de calcul crête élevée pour une consommation électrique donnée, une surface au sol limitée et un nombre de pannes raisonnable

Problématique et approche Blue Gene/P

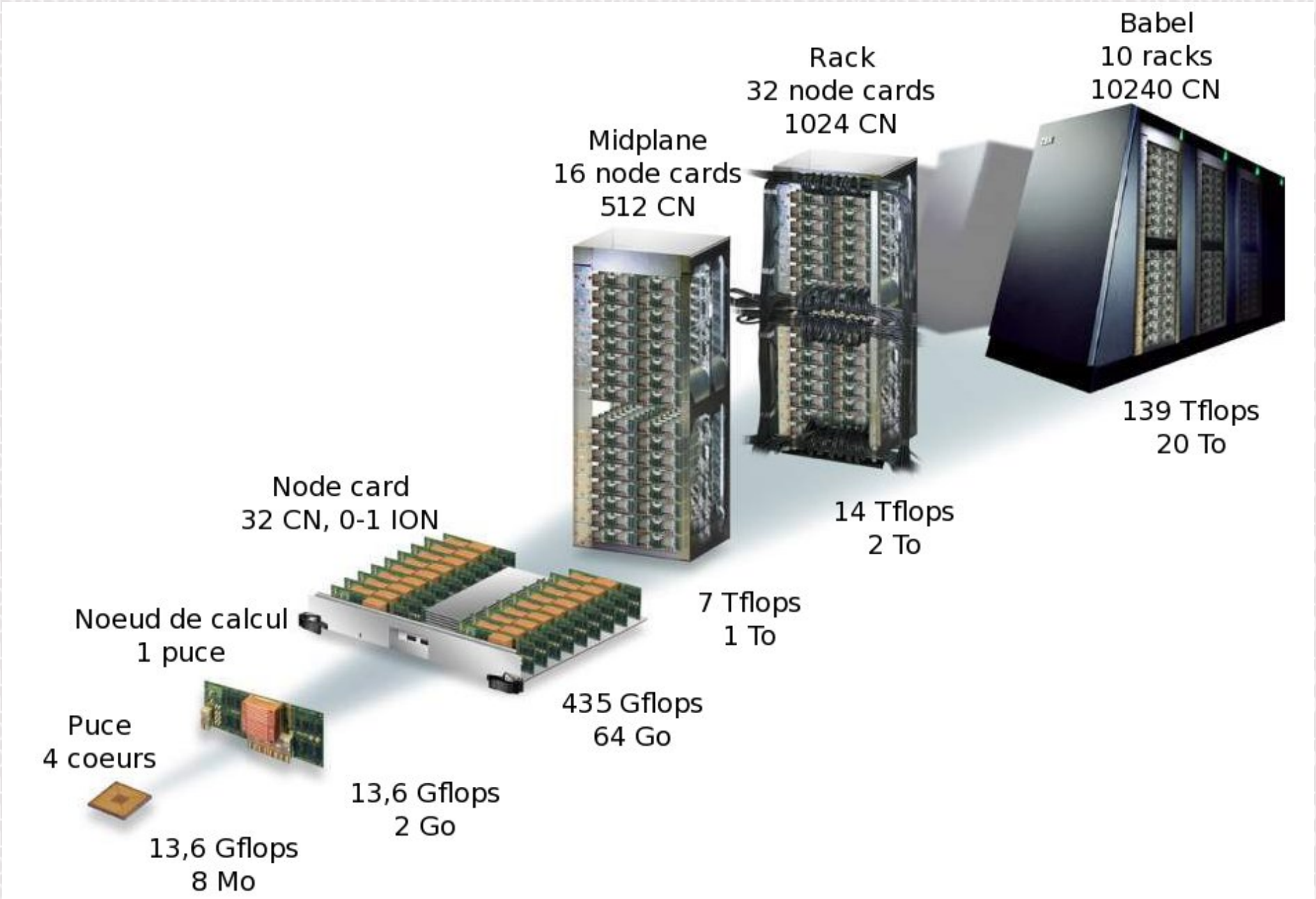
Inconvénients

Inconvénients d'une fréquence CPU basse :

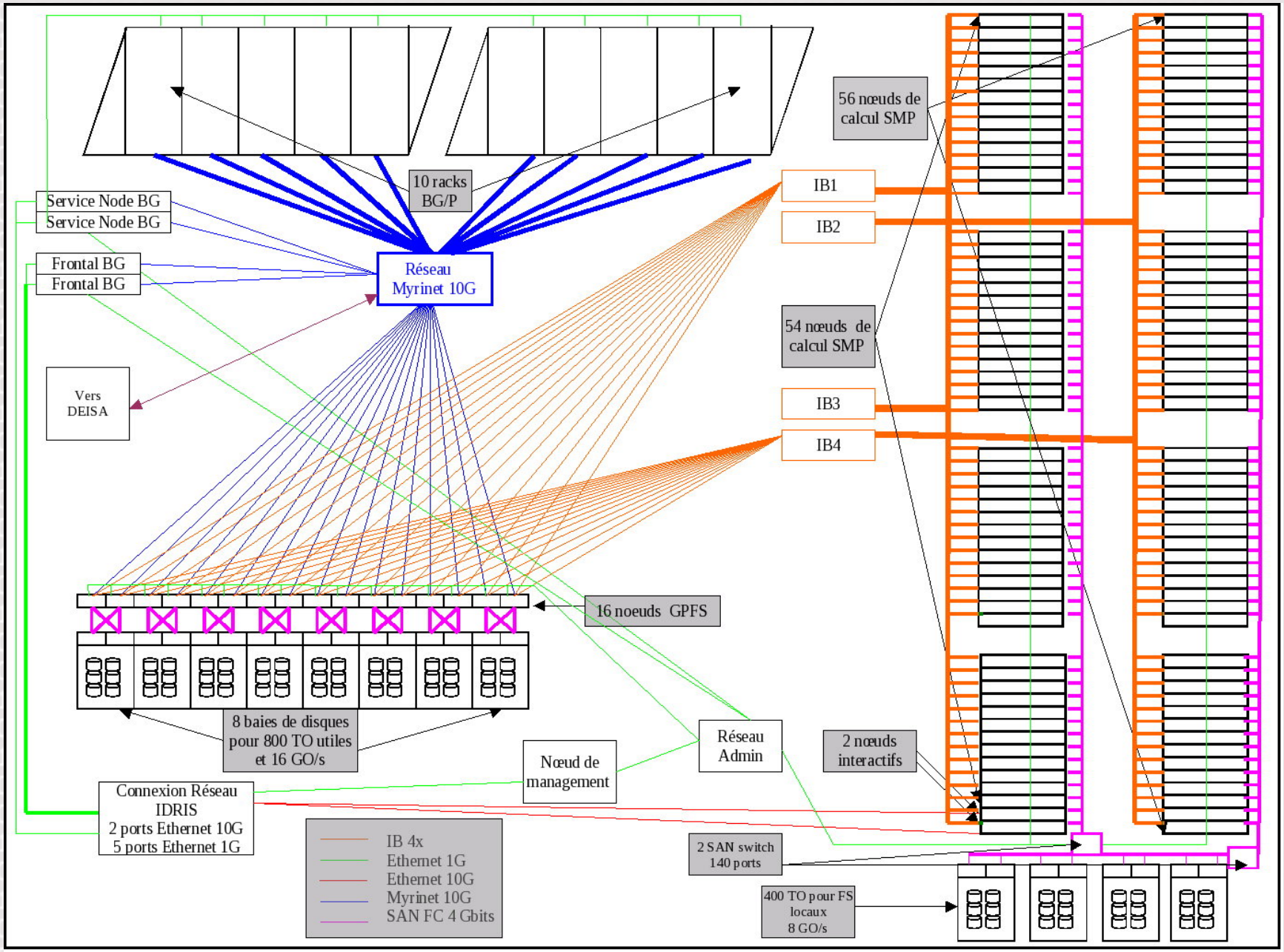
- Chaque cœur est lent
- Grand nombre de cœurs
- Puissance de calcul élevée seulement si grand nombre de cœurs utilisés
- Peu de mémoire par cœur

=> parallélisme massif, machine non généraliste, gamme d'applications limitée

Architecture Blue Gene/P



Exemple de configuration (IDRIS)



Exemple de configuration (IDRIS)

A l'IDRIS

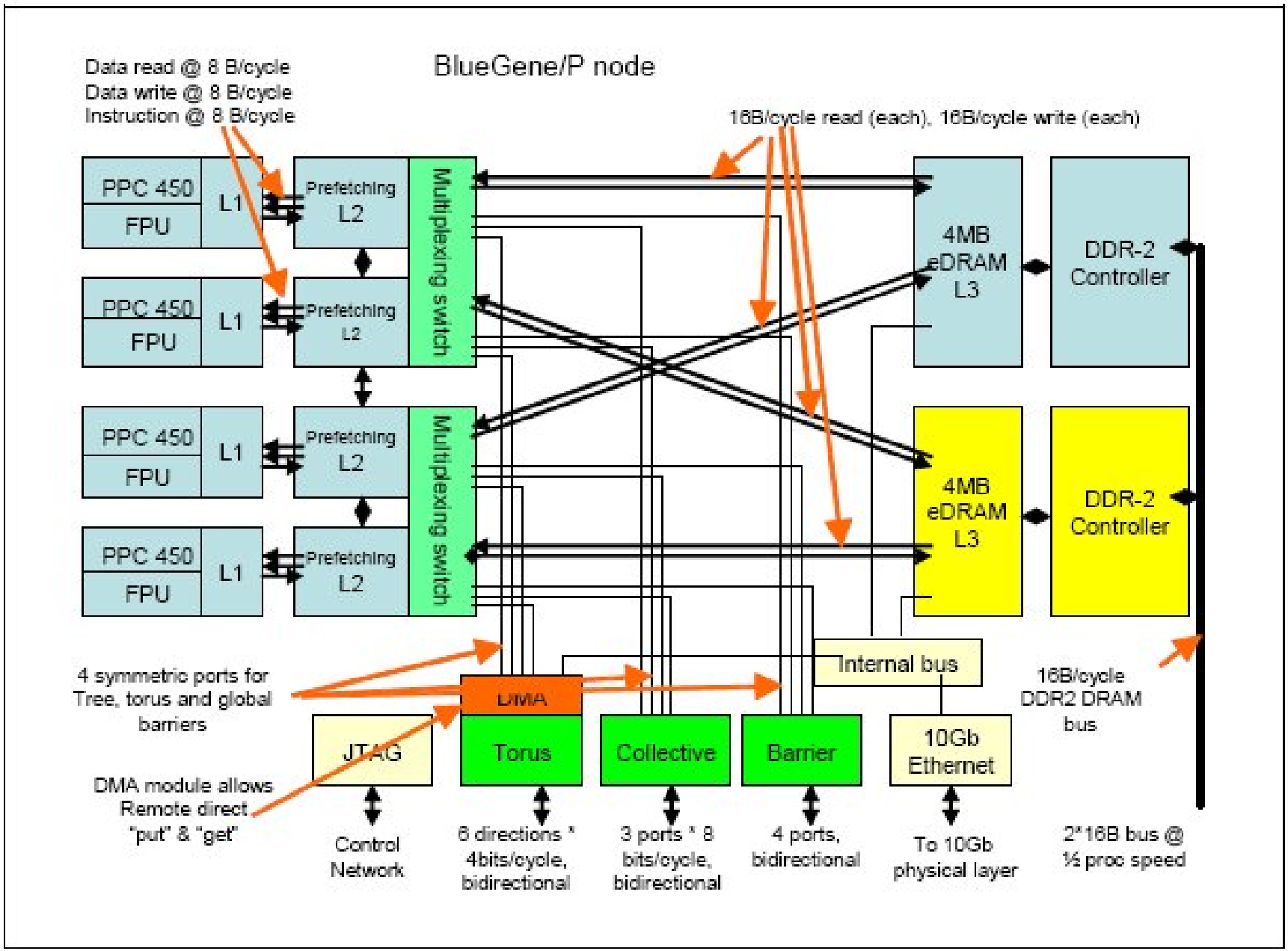
- **Babel : 10 racks Blue Gene/P** :
  - 10.240 nœuds
  - 40.960 cœurs
  - 20 To
  - 139 TFlops
  - 30 kW/rack (300 kW pour les 10 racks)
- **Vargas : 8 racks Power6** :
  - 112 nœuds
  - 3.584 cœurs
  - 18 To
  - 68 TFlops
  - 75 kW/rack (600 kW pour les 8 racks)
- 800 To sur disques partagés BG/P et P6
- 400 To pour Power6
- 950 kW pour la configuration complète

Cœurs/nœuds de calcul

Caractéristiques d'un nœud de calcul

Cœur	PowerPC 450 32-bit
Cœurs par nœud	4
Fréquence horloge	850 MHz
Cache L1 privé par cœur	L1i : 32 ko + L1d : 32 ko
Cache L2 privé par cœur	unité de prefetching
Cache L3 partagé	2 x 4 Mo
Mémoire	2 Go
Puissance crête par nœud	13,6 Gflops
Bande passante mémoire	13,6 Go/s
Consommation électrique	ca 30 W

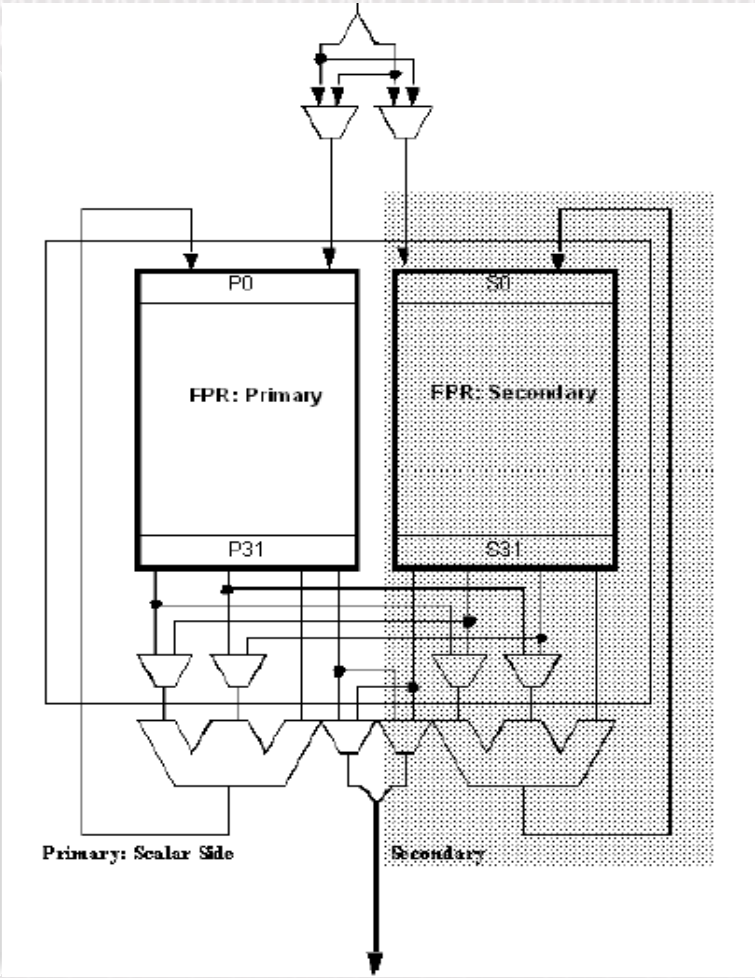
Cœurs/nœuds de calcul



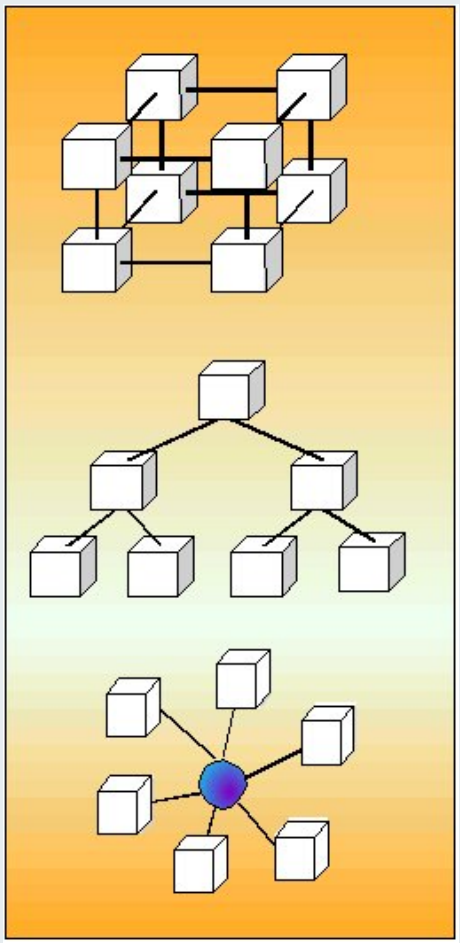
Double FPU

Qu'est-ce que la double FPU ?

- La *double FPU* (Floating Point Unit) est une unité de calcul en réels à virgule flottante double précision (64 bits).
- Cette unité peut réaliser des opérations identiques ou relativement proches en parallèle sur 2 flottants (par exemple : opérations sur des complexes, additions,...).
- Approche de type SIMD.
- Exécute jusqu'à 2 opérations en virgule flottante par cycle. Si FMA (Fused Multiply-Add), 4 opérations/cycle.
- Pas d'exceptions IEEE (si SIMD).

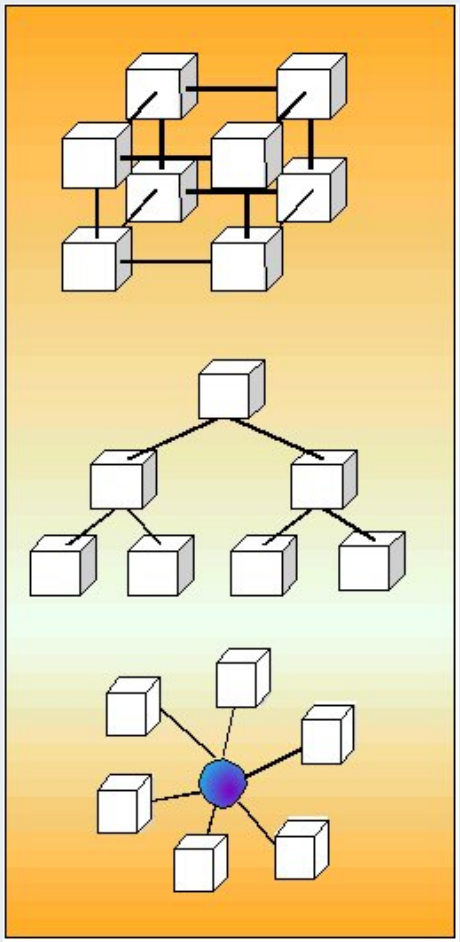


Réseaux



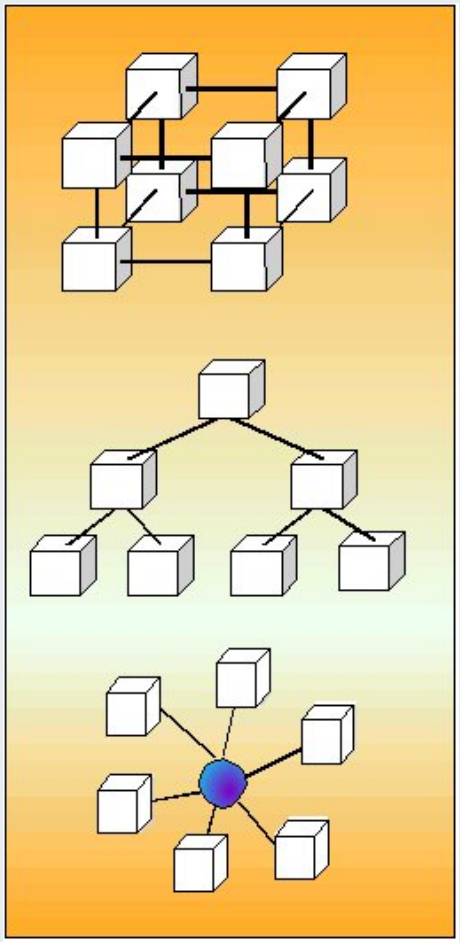
- Tore 3D
  - Connecte chaque nœud à ses 6 voisins (seulement les CN)
  - 6 liens bi-directionnels (12 x 3,4 Gb/s = 5,1 Go/s)
  - Latences MPI : 3 à 10 μs
  - Optimisé pour communications point-à-point et multicast
  - Utilise un moteur DMA => recouvrement calculs/communications
  - Vrai tore seulement si partition = multiple de midplane
- Réseau collectif
- Barrières (Global Interrupt)

Réseaux



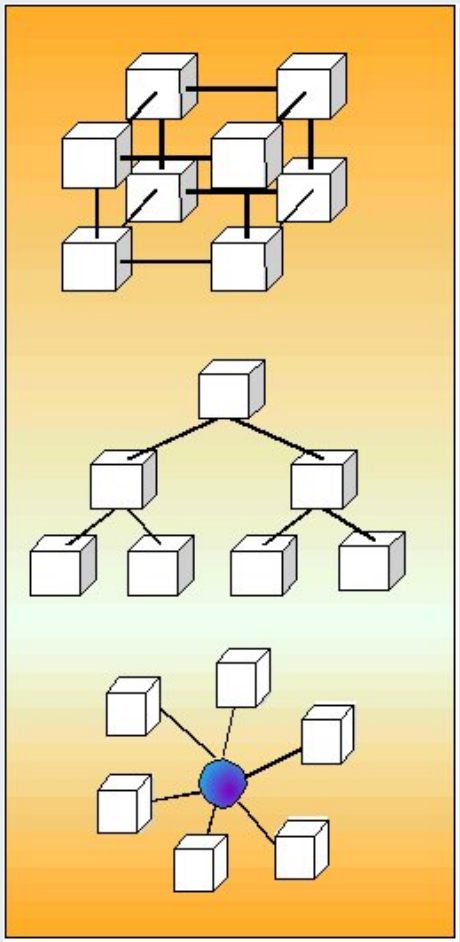
- Tore 3D
- Réseau collectif
  - Connections en arbre (one-to-all) (CN et ION)
  - 3 liens bi-directionnels (6 x 6,8 Gb/s = 5,1 Go/s)
  - Latence MPI : 2 μs (pour la traversée) + 2 μs si suivi par un broadcast
  - Optimisé pour communications collectives (réductions, broadcasts,...)
  - Utilisable seulement si communicateur "rectangulaire"
- Barrières (Global Interrupt)

Réseaux



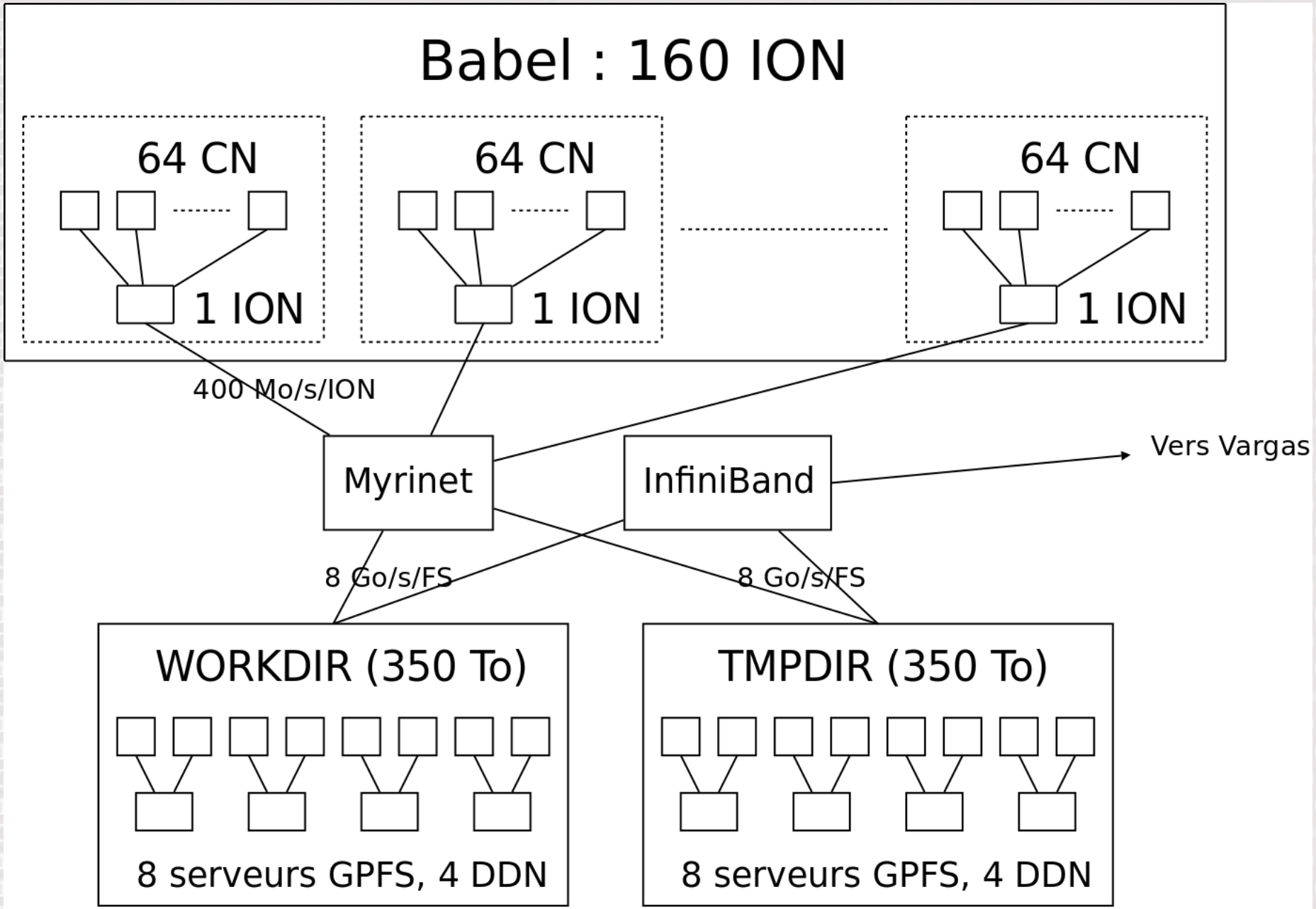
- Tore 3D
- Réseau collectif
- Barrières (Global Interrupt)
  - Réseau à faible latence pour barrières et interruptions
  - Latence MPI : 2,5 μs pour 8 racks
  - Optimisé pour les barrières

Réseaux



- Tore 3D
- Réseau collectif
- Barrières (Global Interrupt)
- 10 Gb Ethernet (utilisé uniquement par ION vers GPFS)
- JTAG (réseau de service)
- Clock

Entrées/sorties (exemple de l'IDRIS)



Entrées/sorties (exemple de l'IDRIS)

Côté clients BG/P (nœuds d'I/O)

- 1 I/O node tous les 64 compute nodes => 16/rack
- les ION gèrent tous les I/O (transparent côté CN)
- ION sont points de sortie des CN
- ION fournissent sockets aux CN
- ION : jusqu'à 400 Mo/s
- Tous les transferts entre les CN et les ION passent par le réseau collectif

Entrées/sorties (exemple de l'IDRIS)

Côté serveurs I/O

- 16 serveurs GPFS (nœuds 4 cœurs Power6 4,2GHz)
- 8 baies DDN (2 serveurs GPFS/baie) à 2 Go/s/baie
- 800 To en GPFS partagés entre BG/P et P6
- 346 To pour le WORKDIR et 346 To pour le TMPDIR (8 Go/s chacun)
- Taille de bloc (WORKDIR et TMPDIR) : 2 Mo
- Plus petite écriture : taille_bloc/32 (2 Mo / 32 = 64 ko)

Système

Frontale

- Linux (Suse)
- seul accès interactif pour utilisateurs
- compilateurs et débogueurs
- soumission des travaux

Nœuds de service

- gestionnaire de jobs
- démon mpirund
- base de données DB2 (comptabilité, problèmes détectés,...)

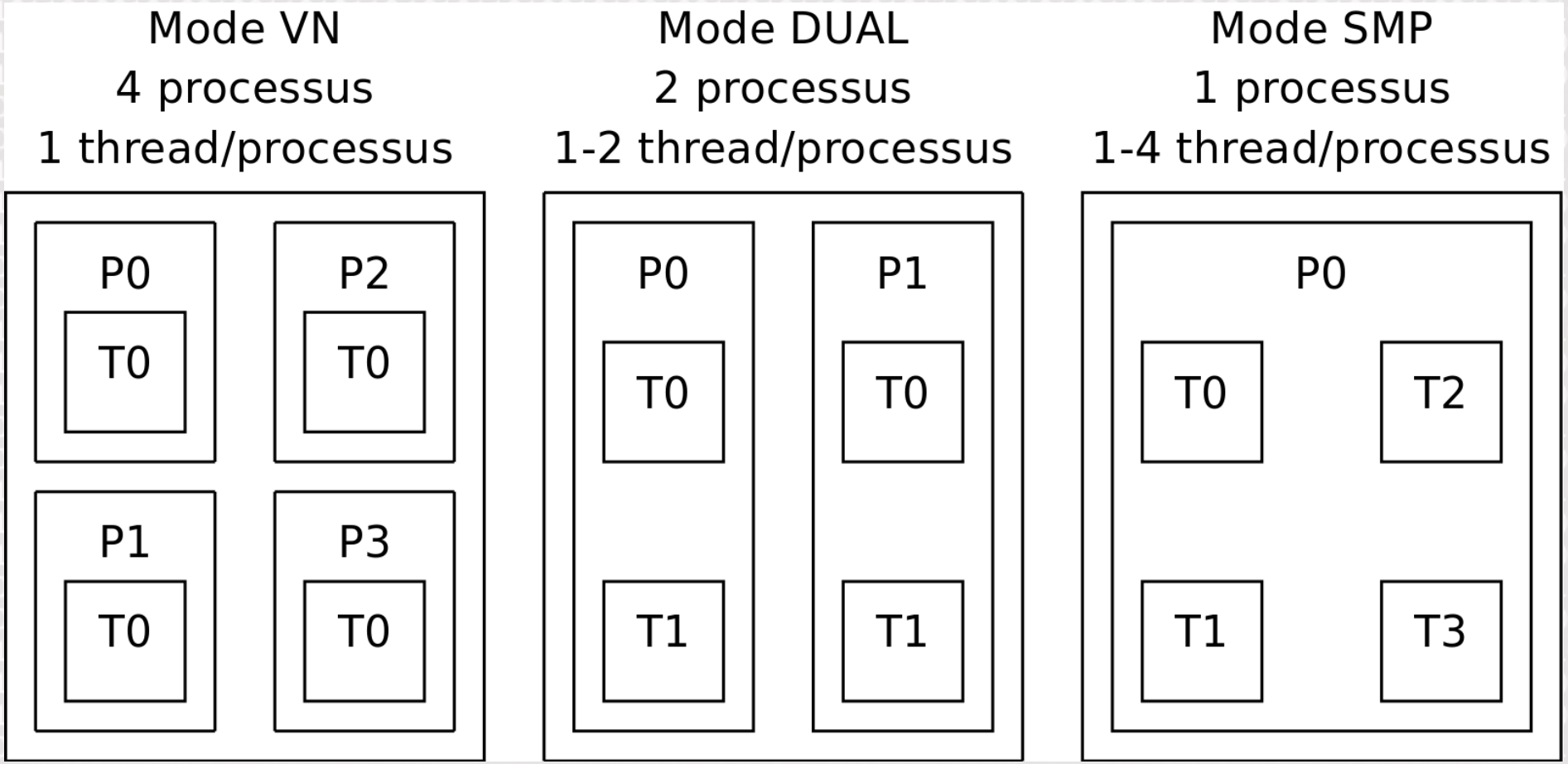
Nœuds d'I/O

- CIOD : démon gérant les I/O
- au démarrage d'un job, reçoit une copie de l'exécutable et le lance sur les CN

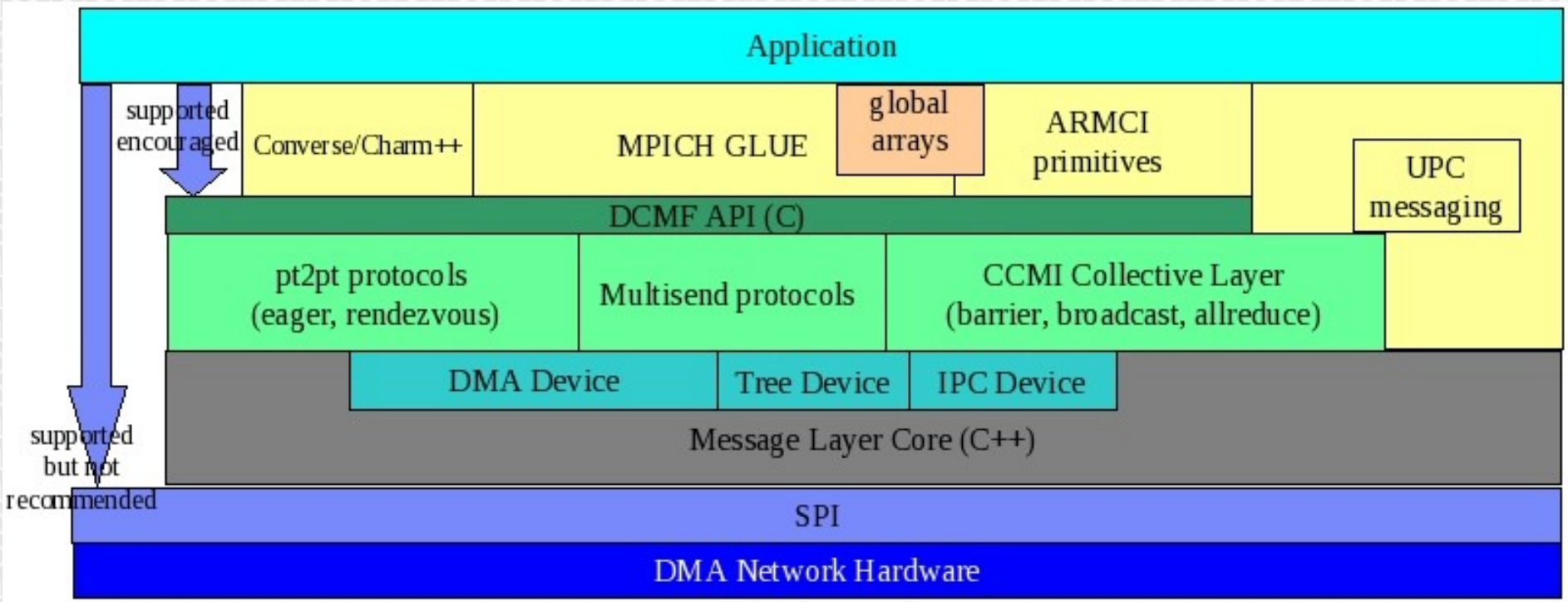
Nœuds de calcul

- CNK (Compute Node Kernel) : OS allégé (pas de démons,...)
- binaire compatible avec linux (linux de l'ION)
- paradigmes de programmation supportés (pas tous implémentés) : MPI, OpenMP, ARMCI, Global arrays, Charm++, UPC
- 1 seul job parallèle simultanément par partition
- 1 seul exécutable par CN => partage des zones text et read-only constant data
- max 1 thread par coeur (assigné statiquement)
- mémoire virtuelle = mémoire physique
- pas de mémoire swap
- support de shared memory (POSIX)
- support complet NPTL (native POSIX thread library ou pthreads)

Modes d'exécution



Infrastructure de communications



Mapping/placement des processus

Définitions

- Chaque processus MPI a un rang dans le communicateur MPI_COMM_WORLD (allant de 0 à Nprocessus-1).
- Chaque processus MPI est placé de façon statique sur un cœur de la machine (pas de déplacements en cours d'exécution).
- Le placement ou mapping correspond à la relation entre le rang du processus et sa position sur la BlueGene/P.

Importance

Plus le nombre de processus est élevé, plus il y a de communications et plus la distance moyenne (le nombre de liens à traverser) entre chaque processus augmente.

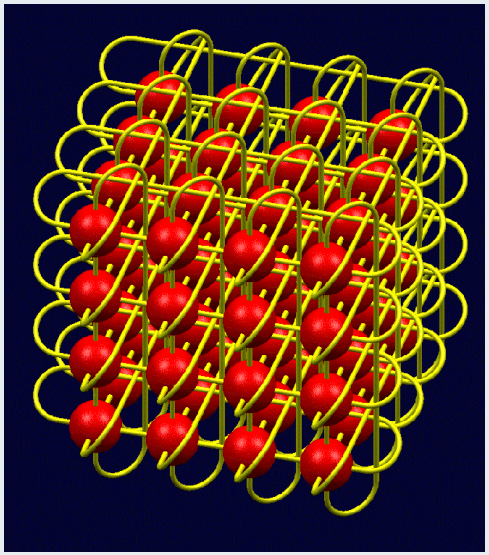
- Or, la latence augmente avec la distance ;
- la contention du réseau augmente si des messages traversent plusieurs liens.

L'impact d'un mauvais placement peut être très élevé.

Mapping/placement des processus

Topologie de la BlueGene/P

- Pour les communications point-à-point et certaines collectives, la topologie réseau est un tore 3D.
- Chaque nœud de calcul est connecté à ses 6 voisins par des liens réseaux bidirectionnels.
- L'idéal est de ne communiquer qu'avec ses voisins directs.



Attention : la topologie n'est un vrai tore 3D que si vous utilisez un multiple de midplanes (512 CN).

MPI

Implémentation

- Normes MPI-1 et MPI-2
- Sauf gestion dynamique de processus
- Implémentation BG/P basée sur MPICH2
- Optimisée pour utiliser les réseaux BG/P
- Moteur DMA => très bon recouvrement calculs/communications
- Fonctions cartésiennes optimisées (MPI_Dims_create, MPI_Cart_create)
- stdin seulement pour le rang 0

Multithreading et OpenMP

Implémentation

- OpenMP (support complet du standard OpenMP 2.5)
- POSIX Threads (basé sur NPTL "Native POSIX Thread Library")
- Maximum un thread par cœur (1 par processus MPI en mode VN, 2 en mode DUAL et 4 en mode SMP)

Approche mixte MPI/multithreads

- Un processus ne peut être décomposé en threads (ou processus légers) que dans un espace mémoire partagé.
- L'utilisation de plusieurs threads n'a un sens qu'en approche mixte sur la BlueGene/P (multithreads à l'intérieur d'un nœud de calcul).

Avantages :

- Moins de communications
- Economies de mémoire
- Gains potentiels d'extensibilité

Bibliothèques à l'IDRIS

Liste bibliothèques

- ARPACK 96
- ESSL 4.4
- FFTW 2.1.5 et 3.1.2
- HDF5 1.8.1 et 1.8.2
- LAPACK 3.1.1
- MASS 4.4
- METIS 4.0.1
- MUMPS 4.7.3
- netCDF 3.6.2
- PARPACK 96
- PETSc 2.3.3 et 3.0.0-p2
- ScaLAPACK 1.8.0

Applicatifs à l'IDRIS

Liste applicatifs

- CP2K
- CPMD 3.13.1 et 3.13.2
- GROMACS 3.3.3 et 4.0.3
- LAMMPS
- METIS 4.0.1
- NAMD 2.6 et 2.7b1
- Vasp 4.6.34

Outils à l'IDRIS

- addr2line
- GNU gprof
- GDB
- FPMPI2, mpiP et MPI Trace
- libbgpidris
- Subversion (SVN)
- TotalView

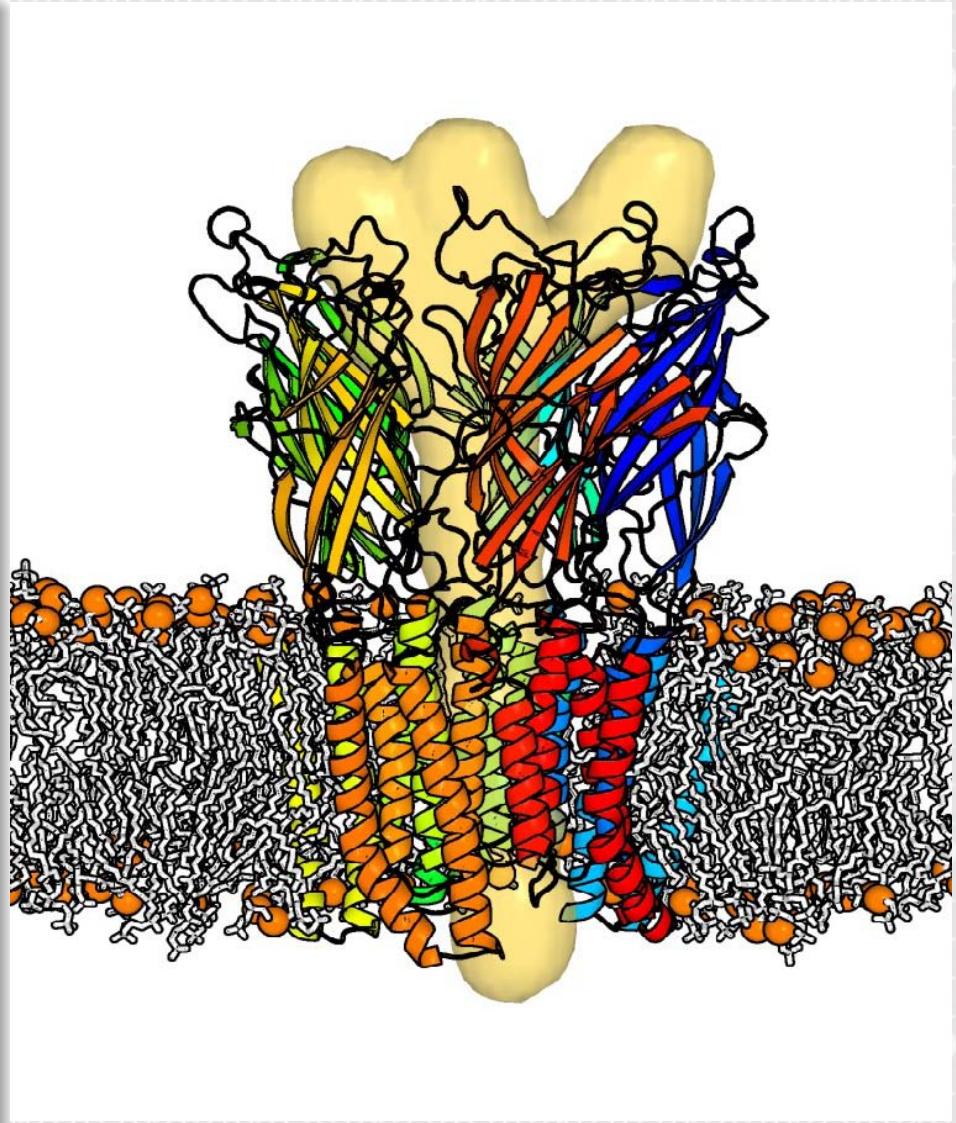
Nico_bacter

Description projet

Simulation d'un analogue bactérien du récepteur nicotinique de l'acétylcholine

- But : compléter étude cristallographique par étude en dynamique moléculaire du récepteur bactérien avec et sans molécules de détergents dans le canal, afin de vérifier la stabilité de l'édifice
- modèle atomique complet incluant les atomes d'hydrogènes
- 20 ns de simulation.
- Conclusion : conformation ouverte stable
- résultats publiés le 05/11/08 dans Nature

Unité de Dynamique Structurale des Macromolécules, Institut Pasteur, Paris



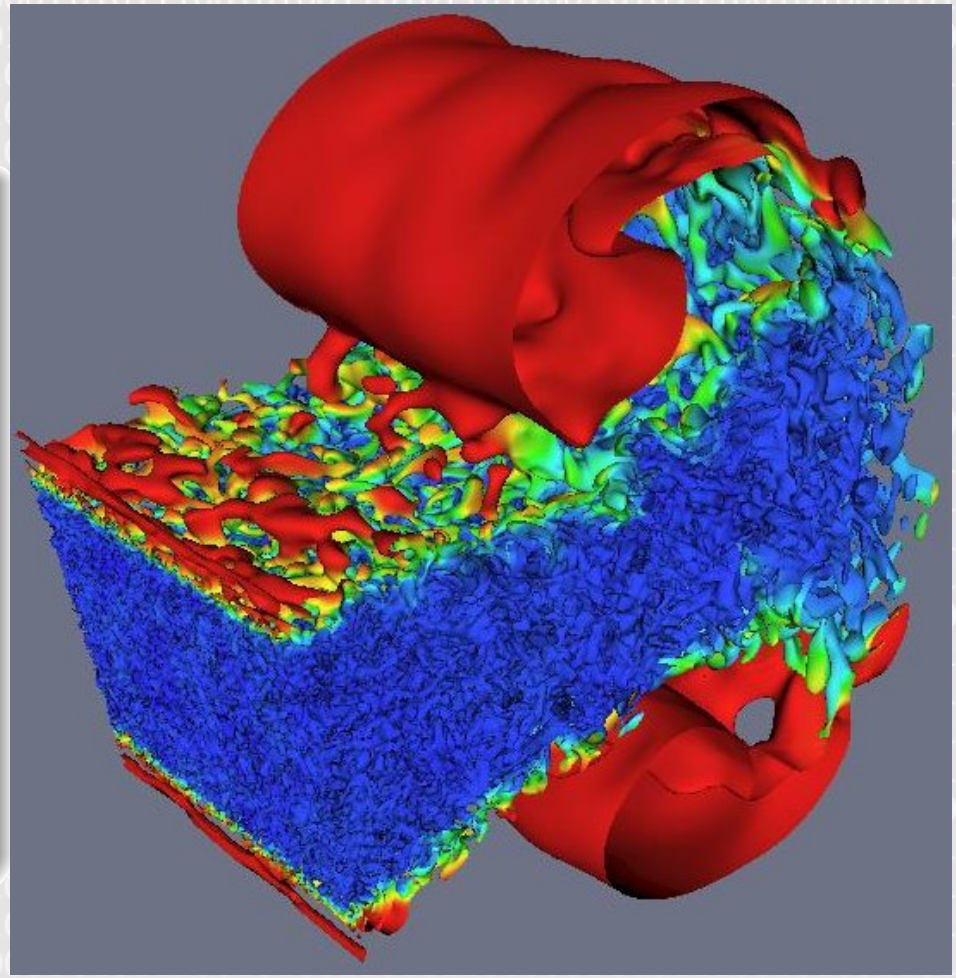
SimCoT

Description projet

Simulation aux grandes échelles et simulation numérique directe de la combustion turbulente

- flamme axisymétrique en moyenne (Re= 2000, 8,5M de noeuds)
- flamme avec une injection de type fente bidimensionnelle en moyenne (Re= 4500, 91 M de noeuds)

CORIA-UMR 6614, Campus du Madrillet 76801 Saint-Etienne-du-Rouvray

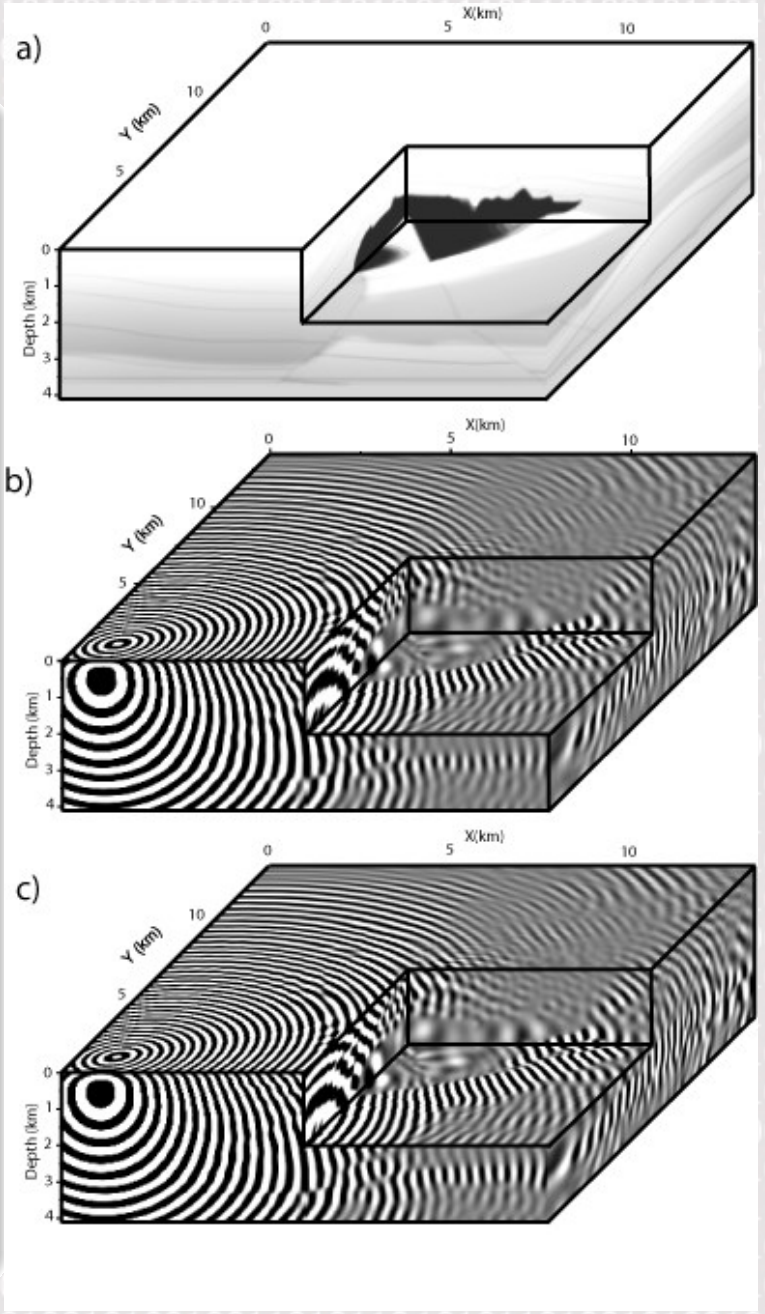


Description projet

Modélisation numérique de la propagation des ondes sismiques en domaine fréquentiel fondée sur un solveur hybride (direct/iterative) massivement parallèle

- imagerie sismique du sous-sol par inversion du champ d'onde complet
- résolution d'un grand système d'équations linéaire creux dont la solution est le champ d'onde monochromatique et le terme de droite est la source sismique
- doit être résolu pour un grand nombre de sources
- méthode de décomposition de domaine avec un solveur hybride direct/itératif et une approche par complément de Schur

UMR Géosciences Azur-CNRS-UNSA-IRD-OCA, ENSEEIHT-IRIT, CERFACS, LGIT



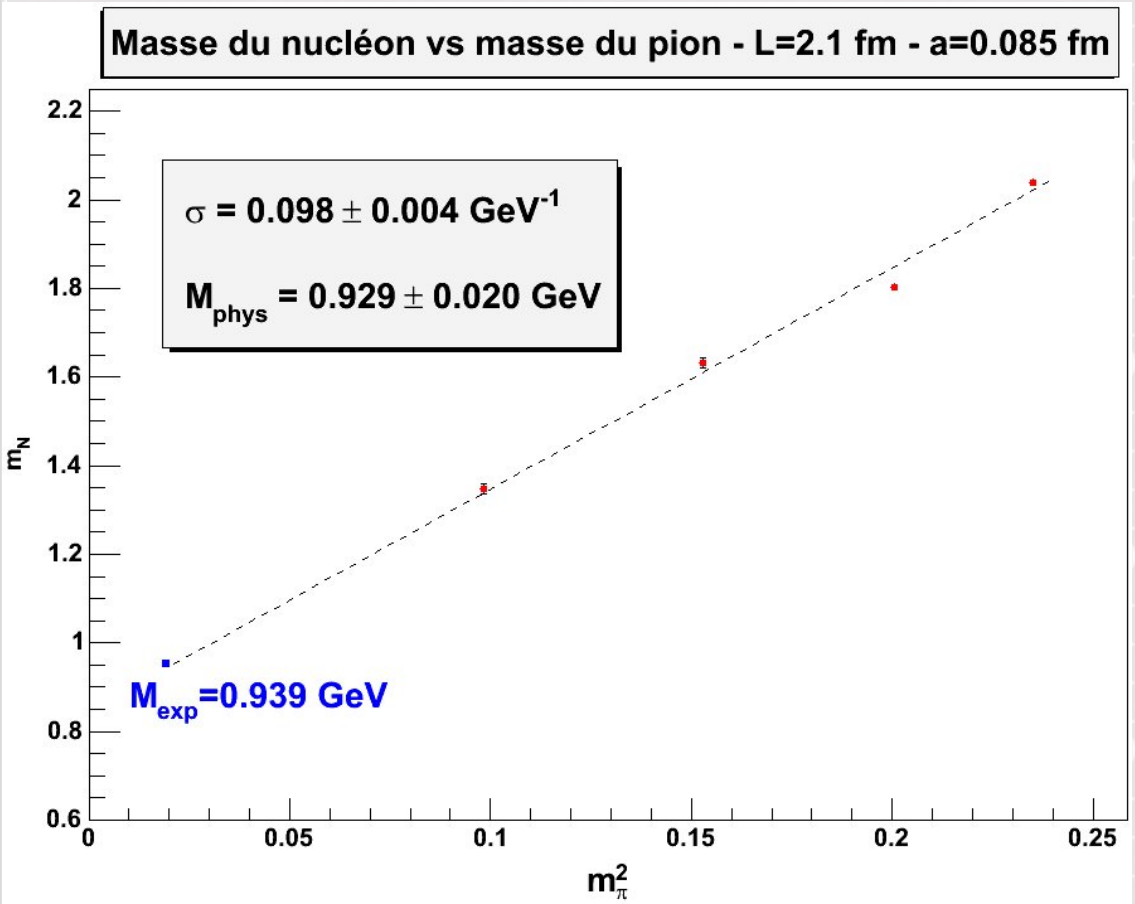
LQCD

Description projet

Calculs de QCD sur réseau en Physique Hadronique

- compréhension des interactions entre particules élémentaires portant une charge de couleur (quarks et gluons)
- expliquer cohésion des noyaux ainsi que la structure des protons et des neutrons
- utilise 16384 processus
- 100 millions d'heures de calcul

Laboratoire de Physique Subatomique et de Cosmologie de Grenoble, Laboratoire de Physique Théorique d'Orsay, Service de Physique Nucléaire de Saclay

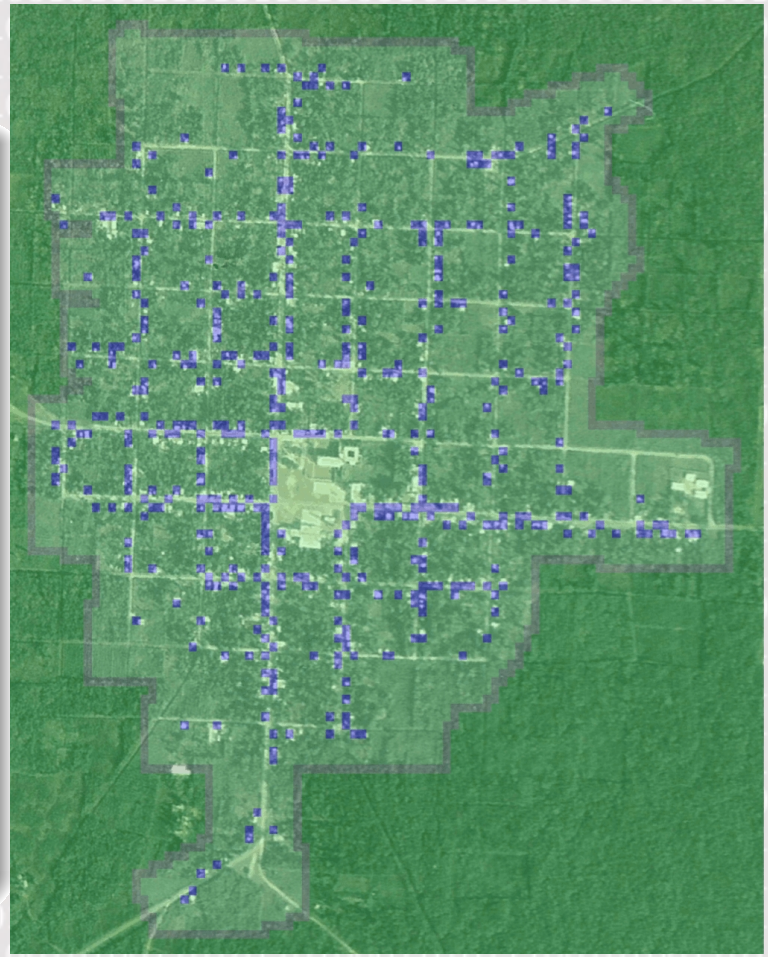


Description projet

Contrôle Spatialisé d'un Vecteur de la maladie de Chagas dans un village réel

- la maladie de Chagas est la maladie vectorielle la plus importante d'Amérique latine
- propagée par des punaises
- But : détermination distribution spatiale des vecteurs et leurs capacités de dispersion

Laboratoire de Mathématiques, Physique et Systèmes, Université de Perpignan et Centre de recherche régional "Dr. Hideyo Noguchi", Université autonome du Yucatan, Mexique



Futur

Blue Gene/Q

La Blue Gene/Q devrait être disponible en 2011. Projet Sequoia (LLNL) :

- 20 PFLOP/s
- 1,6 million de cœurs
- 1,6 Po de mémoire vive
- 18(?) cœurs par puce => approche *many cores*
- 6 MW
- 5 à 10 x plus de FLOP/s/Watt que BG/P

Documentation

Pour en savoir plus

- Le site web de l'IDRIS : http://www.idris.fr (section Support technique -> IBM Blue Gene/P)
- *Blue Gene/P Application Development* : http://www.redbooks.ibm.com/abstracts/sg247287.html
- Documentation des compilateurs IBM pour la Blue Gene/P : http://publib.boulder.ibm.com/infocenter/compbgpl/v9v111/index.jsp
- *Exploiting the Dual Floating Point Units in Blue Gene/L* : http://www-1.ibm.com/support/docview.wss?uid=swg27007511
- *Recommendations for Porting Open Source Software (OSS) to Blue Gene/P* : http://www-03.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP101152
- *Overview of the IBM Blue Gene/P project* : http://www.research.ibm.com/journal/rd/521/team.pdf