


Systèmes et traitements répartis sur grille

Serge PETITON

serge.petiton@lifl.fr

LIFL M3-ext 232 et Polytech E203

21 septembre 2007

Master 2, 2007-2008

Plan du cours (à ce jour)

- Introduction et premier survol, du vectoriel aux grilles de calcul à grandes échelles(21/9).
- Architectures et calculs parallèles, data parallèles et vectoriels (28/9)
- *An Overview of High Performance Computing and Challenges for the Future* , Jack Dongara, journée Calcul Intensif (28/9, 16h30-17h30, USTL-CERLA)
- Algorithmiques parallèles et réparties pour le calcul global (5/10, 10h15-12-15)

….. Cours de Nouredine Melab …..

- Vers les systèmes et traitements *Petascales* globaux (7/12)

Sommaire Introduction

- Un peu d’histoire en guise d’introduction
- Du calcul vectoriel au GRID
- Calcul global Pair à Pair
- Le Petaflop et les enjeux nationaux et internationaux

Un peu d'histoire en guise d'introduction

Jadis

Calculs *à la main* (mauvaise précision),

calcul parallèle (Richarson en 1924 veut faire calculer 64000 personnes).

Moitié du siècle dernier, vers le Kilofloat

En route vers le calcul flottant,…Fortran,… IBM 360,….

Dès les années 70-80, l'ère du Mégaflop

Machines vectorielles, CDC 203.Cray 1, superscalaires, RISC, VLIW

Milieu des années 80, l'ère des centaines de Mégaflop

Machines parallèles, data parallèles (CM2 : parallèle avec processeurs vectoriels)

Début des années 90, l'ère des Gigaflop

En plus des machines parallèles, Grappes (clusters) et NOW

Milieu des années 90

Couplage de codes entre centres de calcul intensif, GRID

- Organisation virtuelle, contrôle des logins, identifications
- Co-méthodes et méthodes numériques hybrides

Circa 1997, vers les TéraFloat

Calcul global et à grande échelle

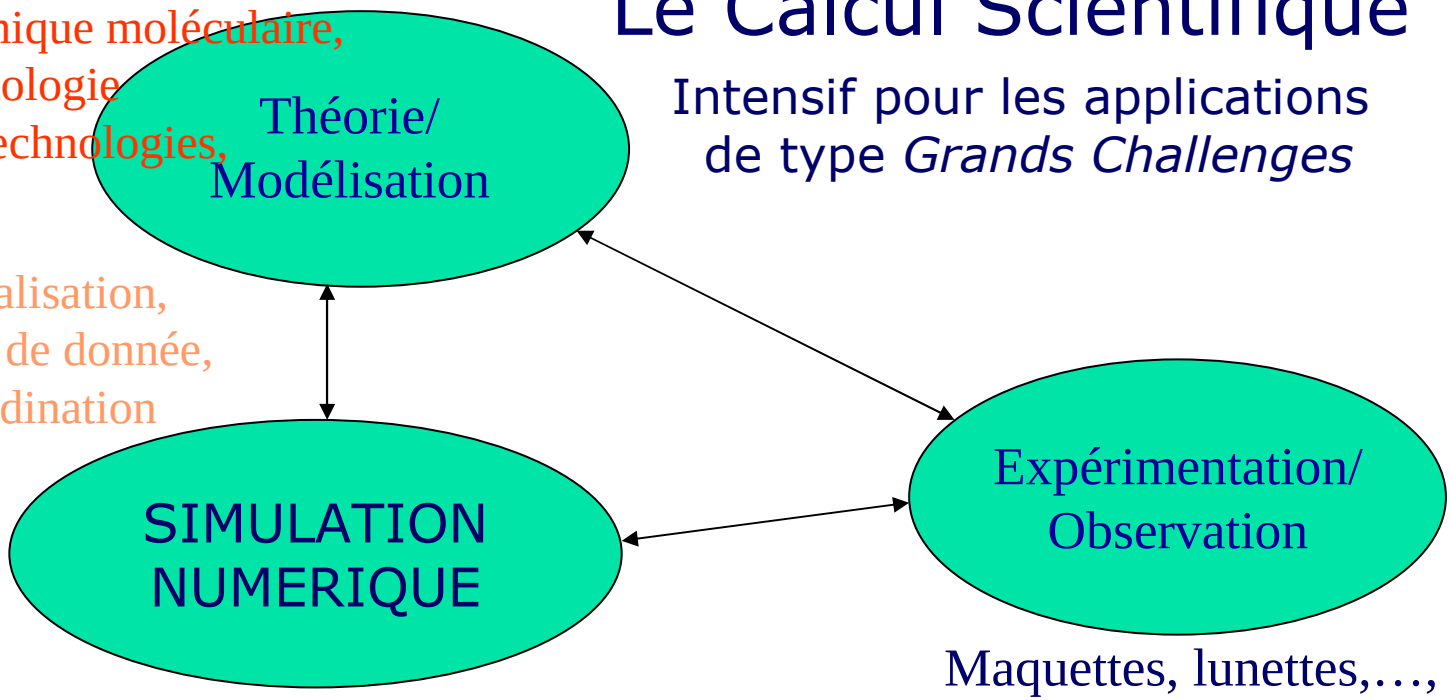
- Grand nombre de participants, calcul global, P2P
- Calculs parallèles et répartis

A venir, le PetaFloat

- Calcul scientifique parallèle complètement répartis et décentralisés sur diverses machines hétérogènes et calcul sur internet…. Mais avec accélérateurs vectoriels, ou autres (Cell, Clearpseed…)

Le Calcul Scientifique

Intensif pour les applications
de type *Grands Challenges*



Dynamique moléculaire,
Climatologie
Nanotechnologies,
etc.…

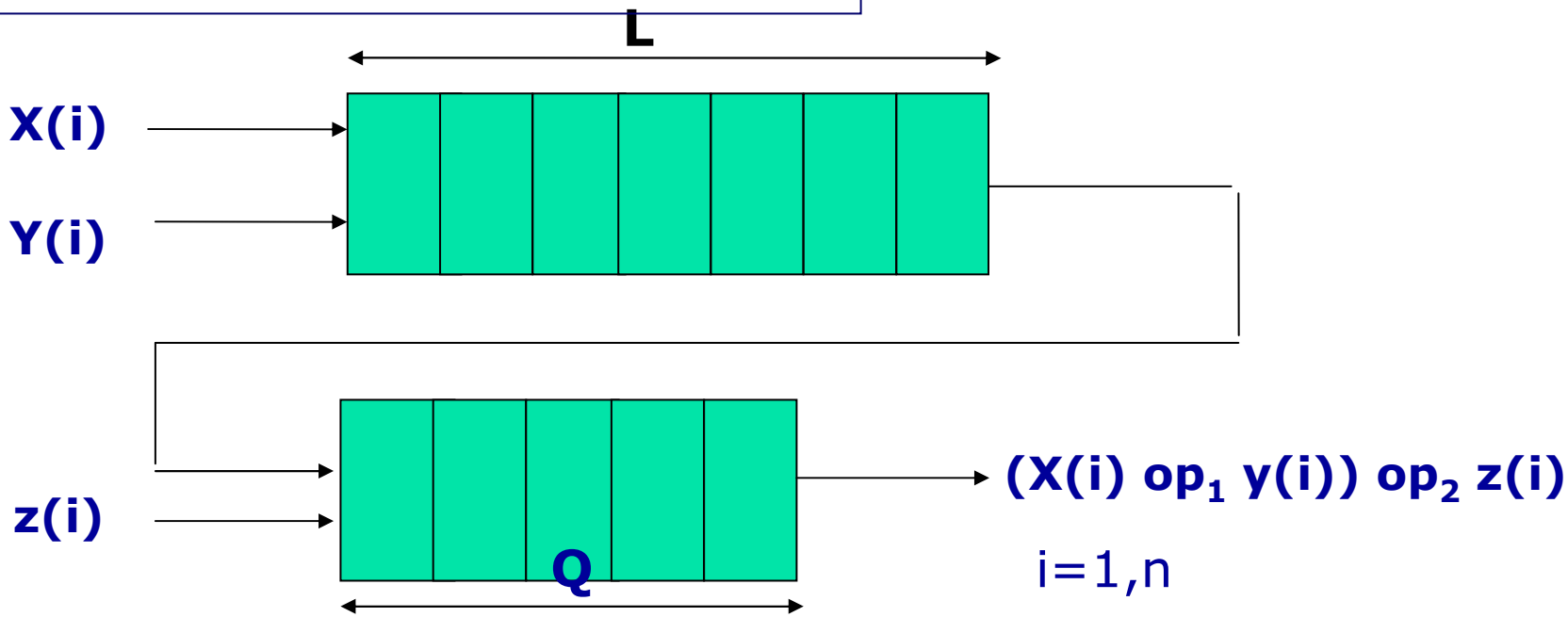
Visualisation,
Fouille de donnée,
coordination

Maquettes, lunettes,…, souffleries,
accélérateurs linéaire, etc.….

La simulation numérique, possible grâce aux puissances de calcul disponibles, change le procédé scientifique

Le calcul scientifique intensif haute performance demande des équipes interdisciplinaires et des moyens très importants

Du calcul vectoriel au GRID



$$t_{seq} = t_{horloge}\ n\ (\ L + Q\ )$$

$$t_{vect} = (L+Q-1)t_{horloge} + n\ t_{horloge}$$

Evaluation des performances

Performances crêtes

Performances soutenues

Performances «TOP500»

Débit asymptotique

N_{max}

$N_{1/2}$

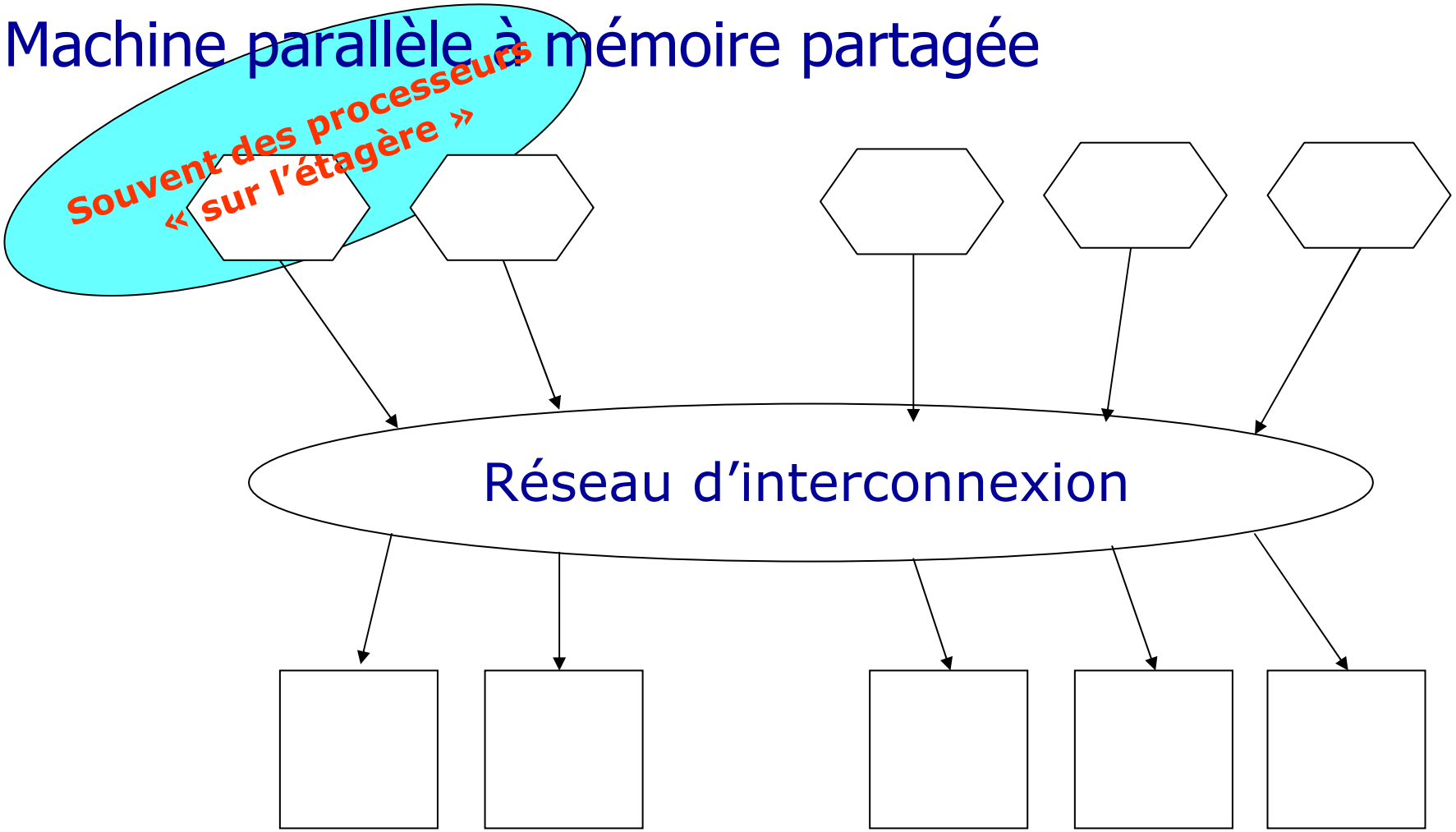
Débit asymptotique =

1/période horloge

D'où la course à la plus petite période horloge.

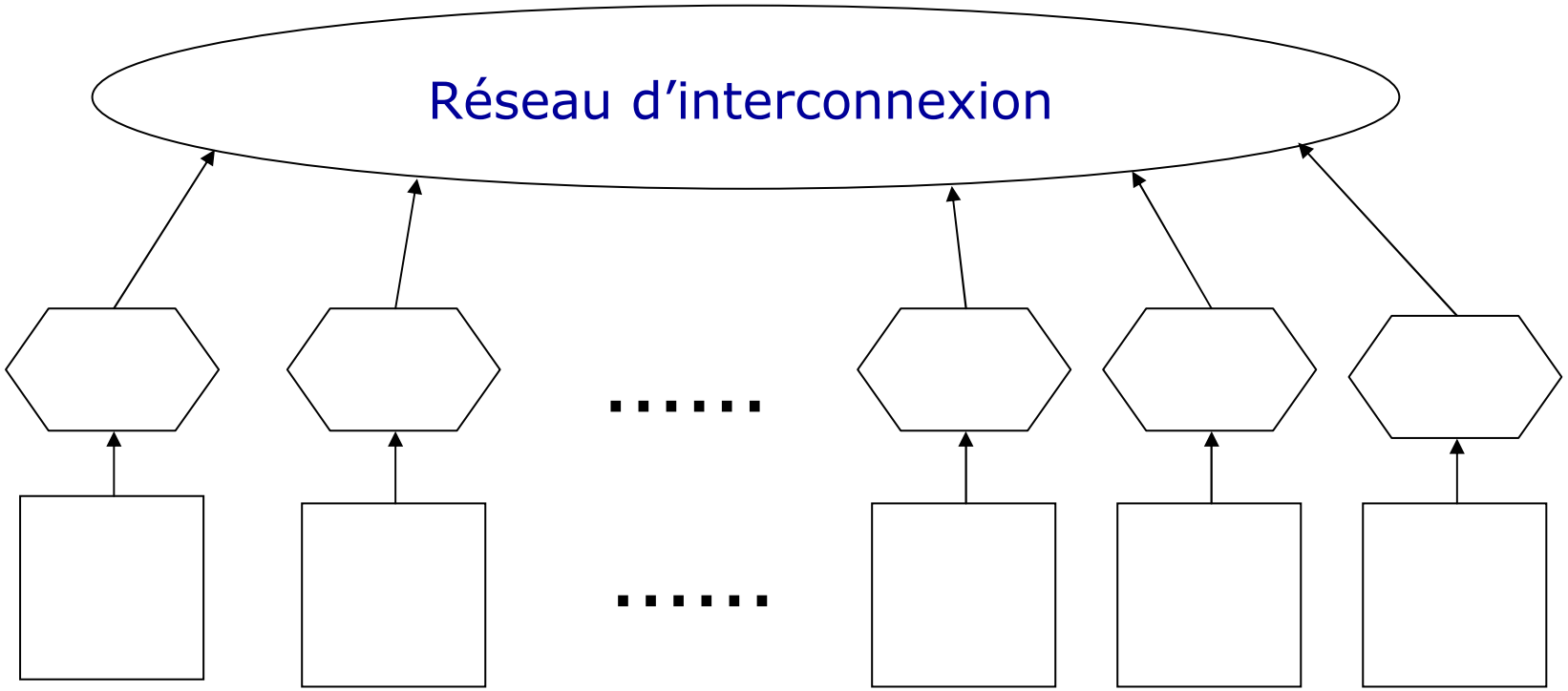
Limites dues au coût, à la technologie et aux évolutions « politiques »

Machine parallèle à mémoire partagée



Uniform Memory Acces (UMA) architectures.
Programmation en MPI ou OpenMP principalement

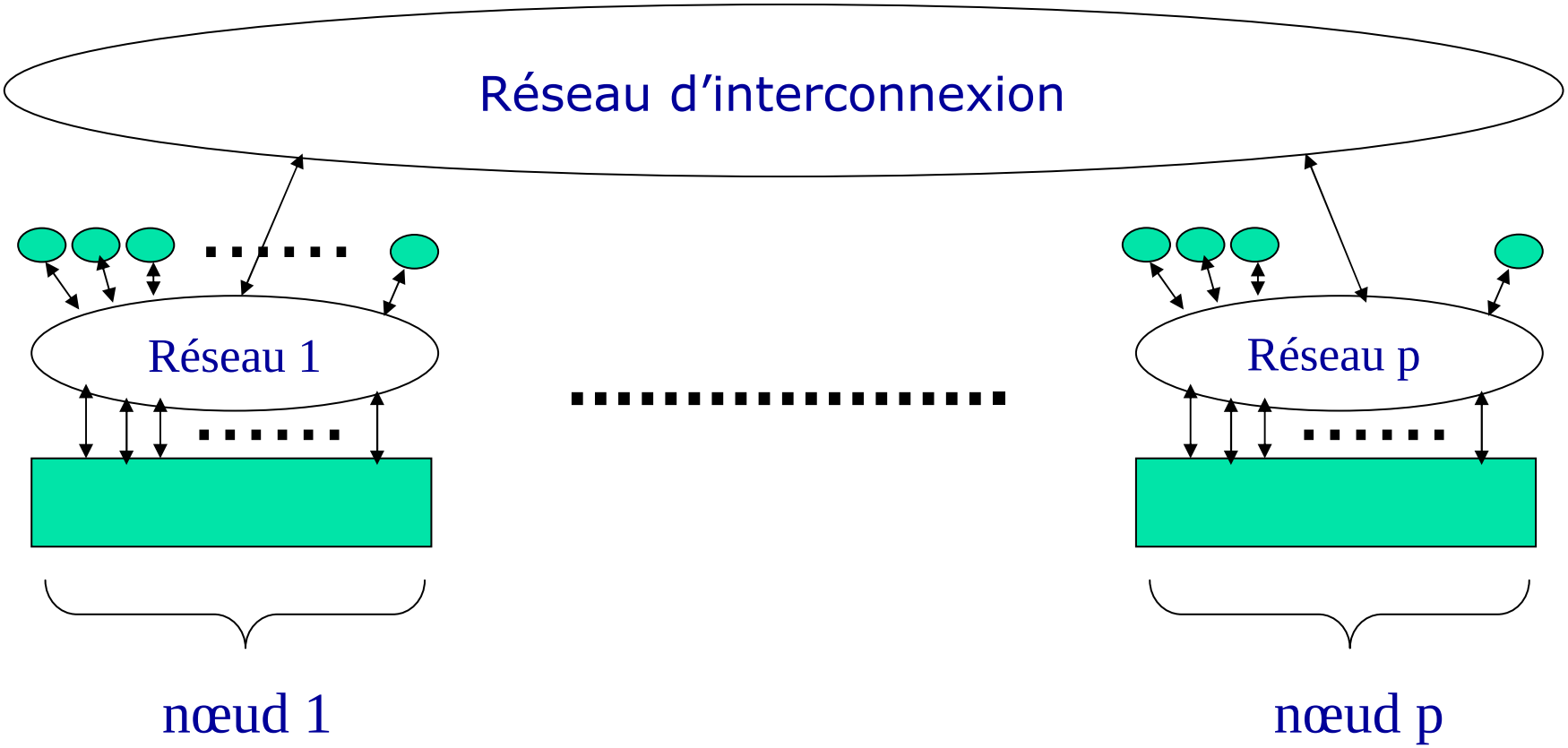
Machine parallèle à mémoires distribuées



Non Uniform Memory Access (NUMA) architectures :

Programmation en MPI principalement, avec langage data parallèle

(Massively Parallel Processors) MPP

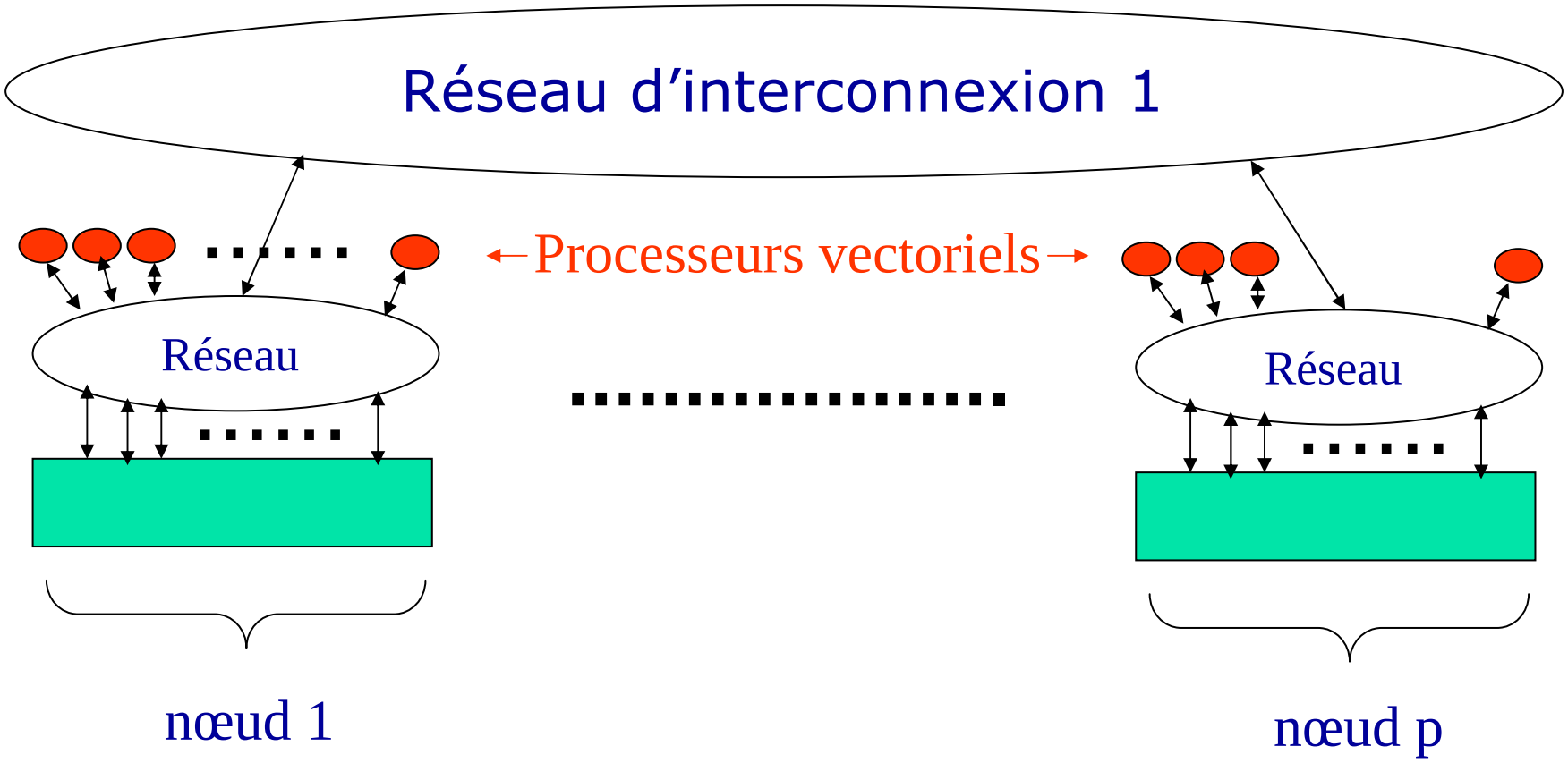


IBM SP4 par exemple, programmation en MPI et/ou OpenMP principalement

MPP/*Earth Simulator*

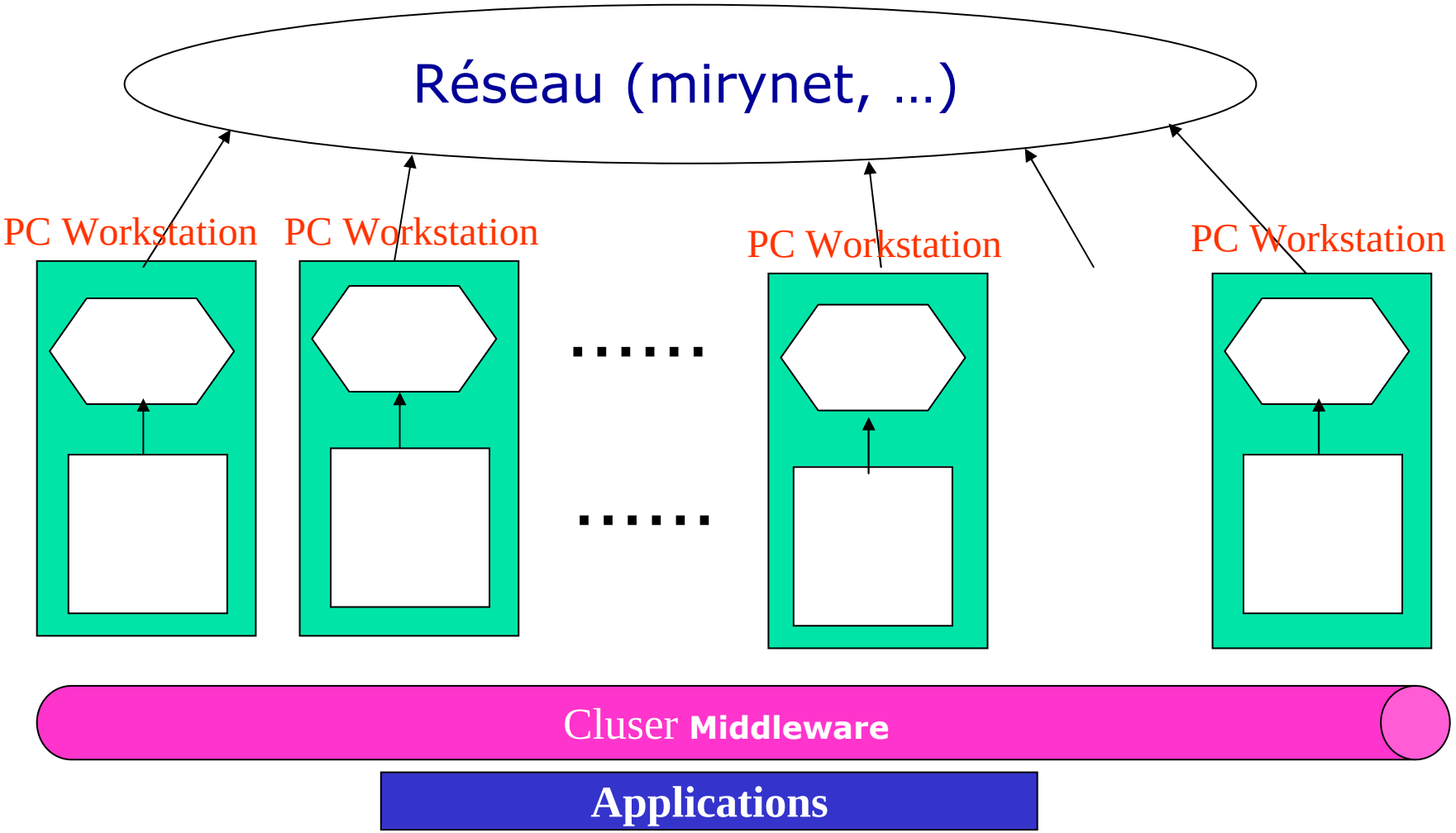
(computenick)

Longtemps Numéro 1 mondial

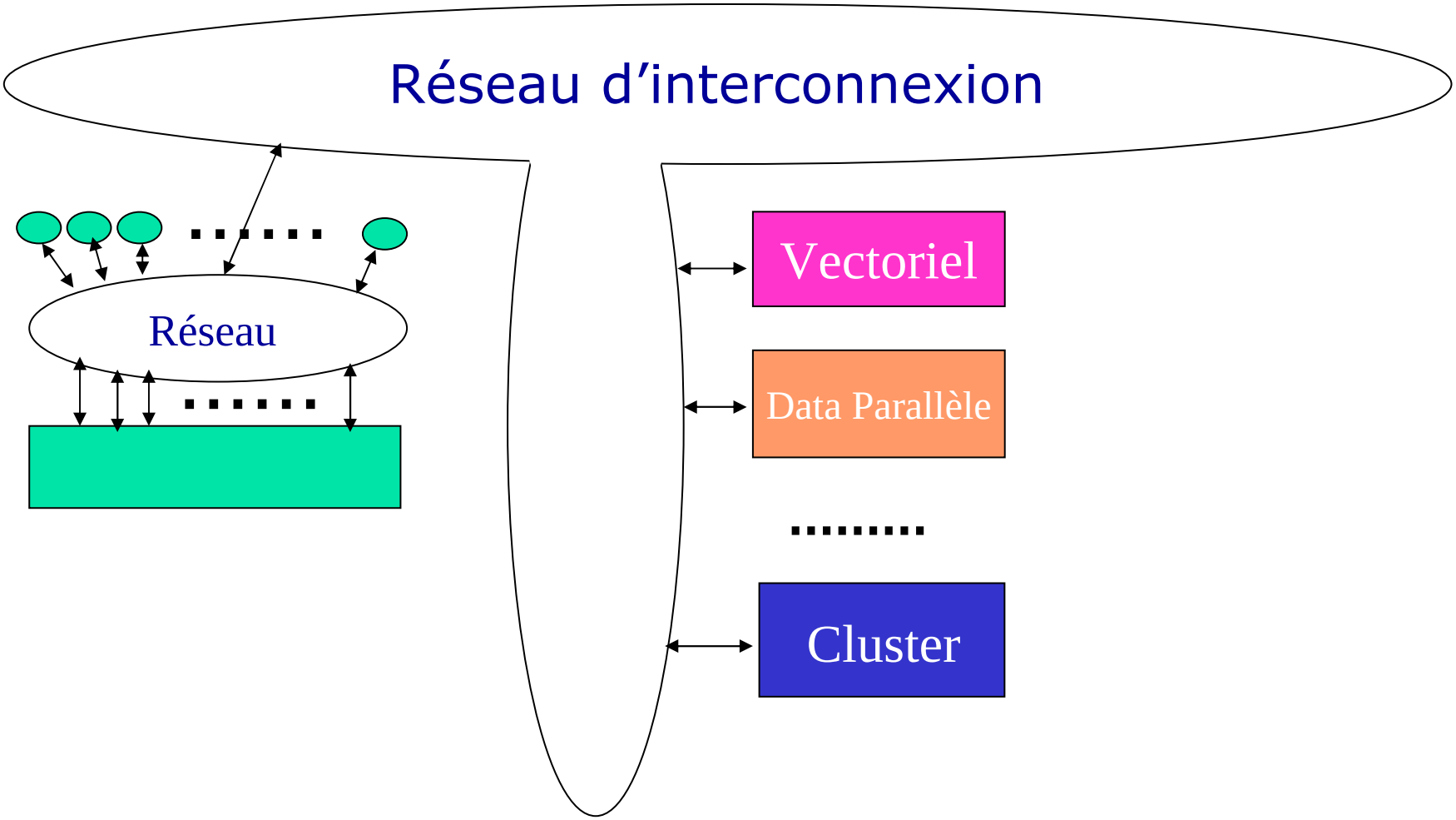


Programmation en MPI et/ou OpenMP ou en HPF

Grappes (Cluster, farm,...)

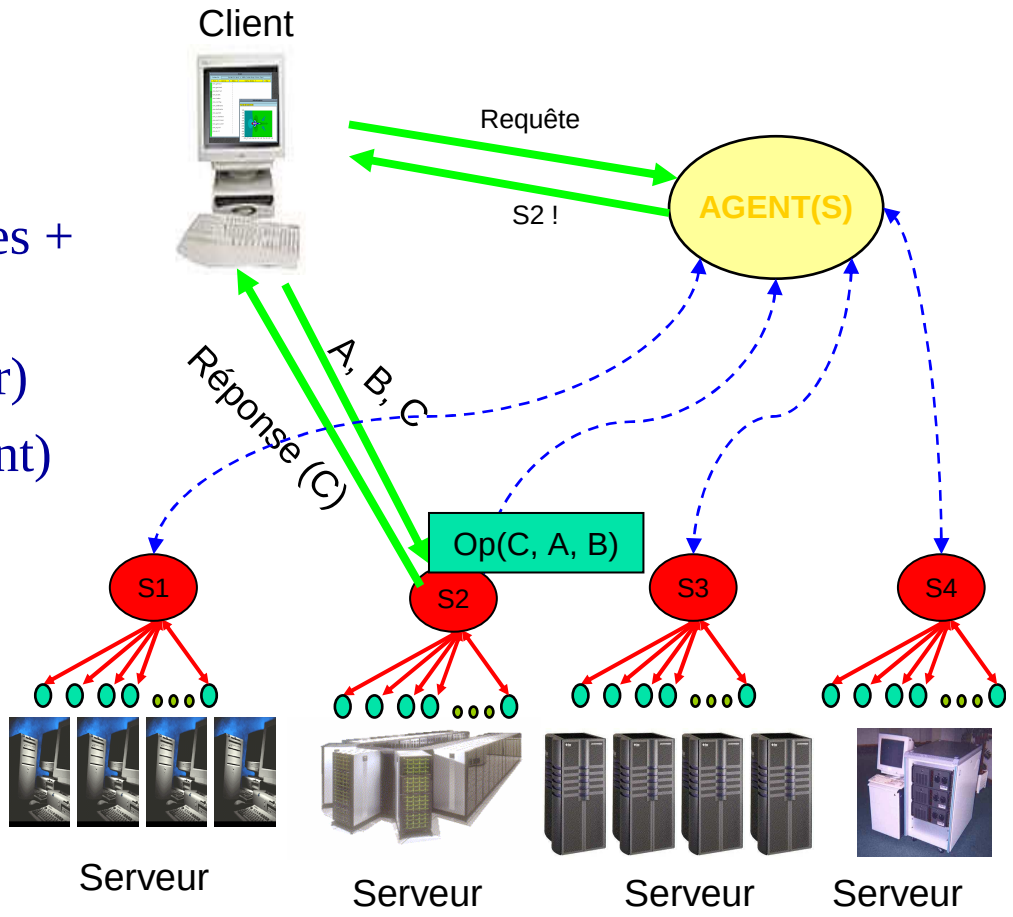


Calculs Intensifs Hétérogènes



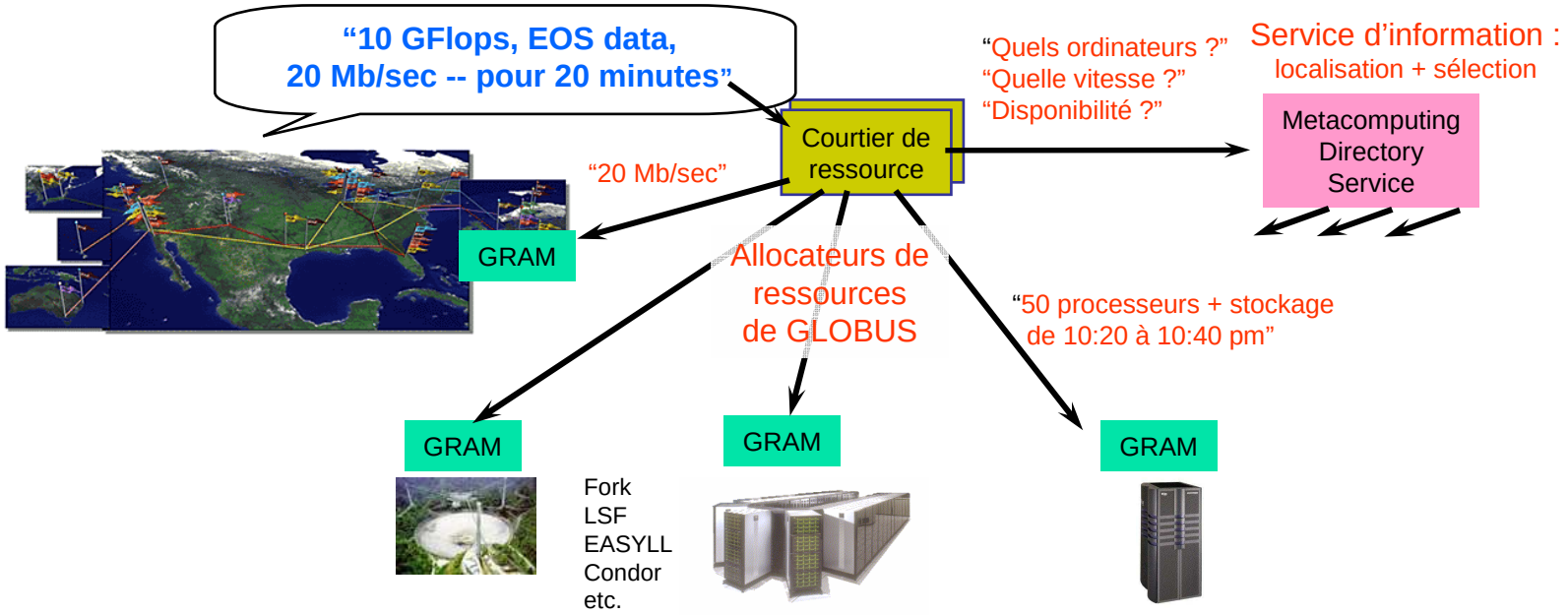
Modèle client/serveur pour les grilles de calcul : le *métacomputing*

- Principe
  - Acheter du service de calcul sur l'Internet
  - Service = applications pré-installées + calculateurs
  - ASP (Application Service Provider)
  - PSE (Problem Solving Environment)
- Exemples
  - Netsolve (Univ. Tennessee)
  - NINF (Univ. Tsukuba)
  - ….

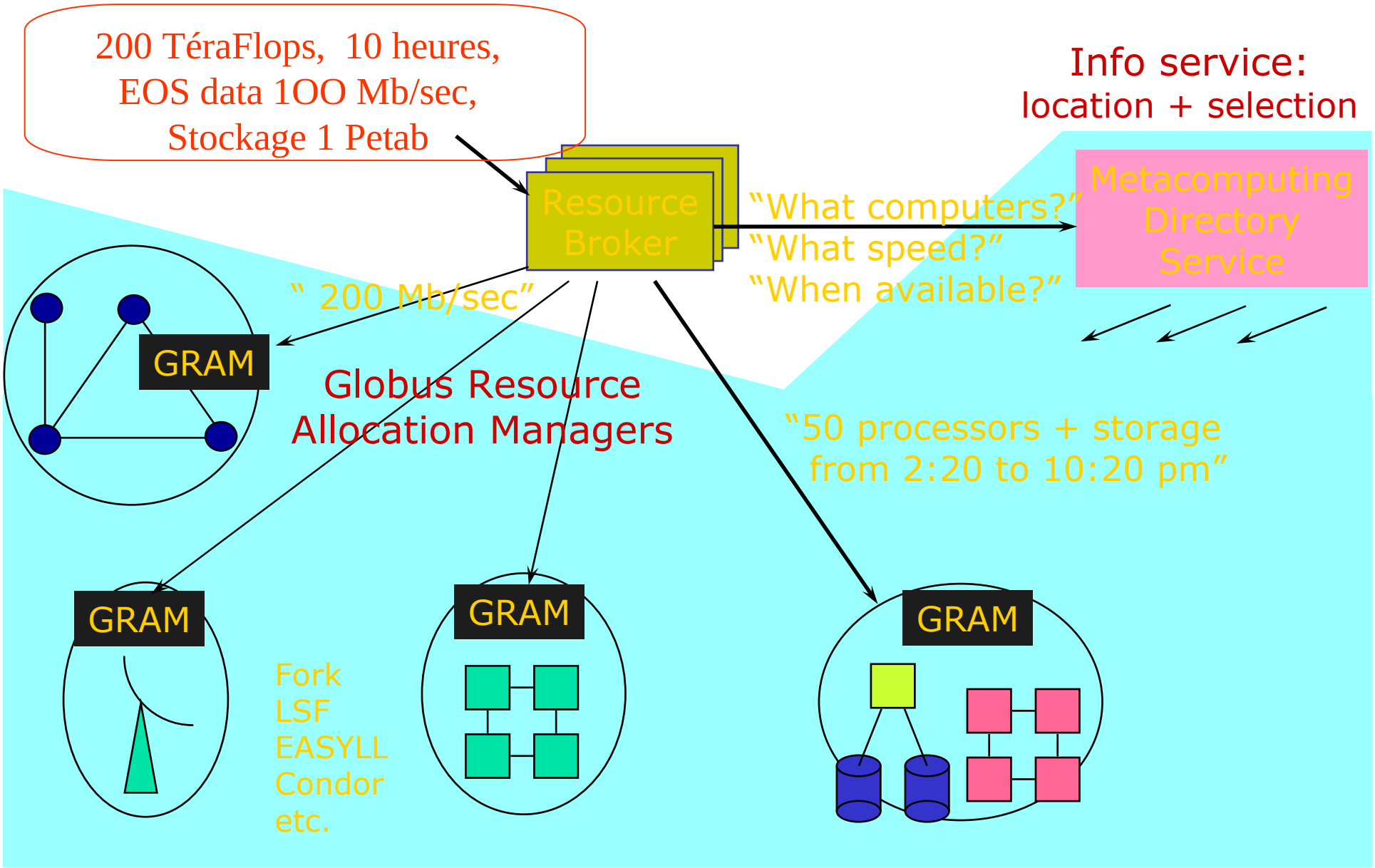


Modèle client/serveur pour les grilles de calcul : le supercalculateur virtuel

- Principe
  - Offrir un supercalculateur parallèle virtuel
  - Faire exécuter ses applications sur des ressources distantes
- Exemples
  - Globus
  - DEISA

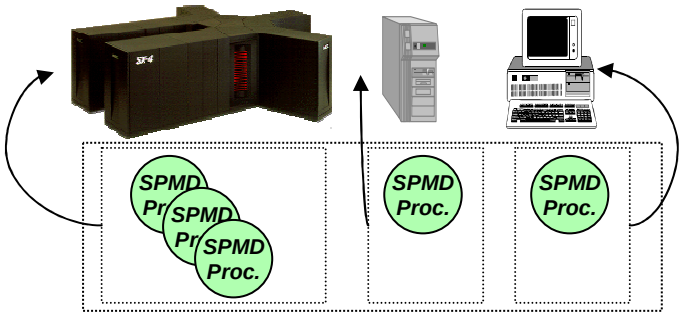


An Example of Globus Services at Work: Resource Management Architecture

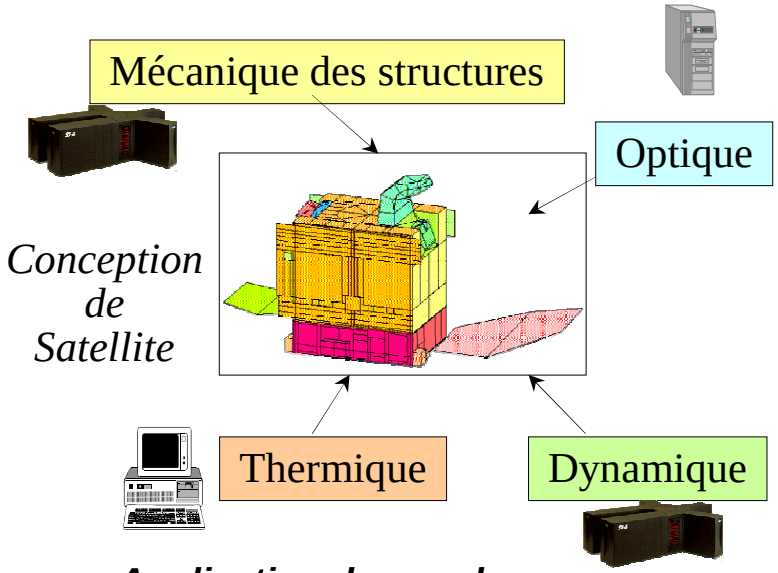


Programmer les grilles de calcul

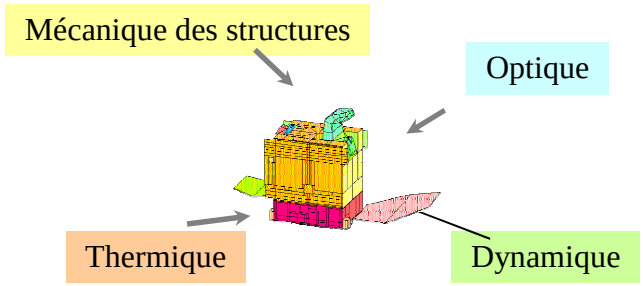
- Un champ applicatif vaste avec des besoins variés…
- Codes parallèles
  - Une grille de calcul est vue comme un calculateur parallèle virtuel (la genèse du Grid)
- Couplages de codes
  - Une application est un assemblage de plusieurs codes de calcul modélisant des physiques différentes



Application parallèle

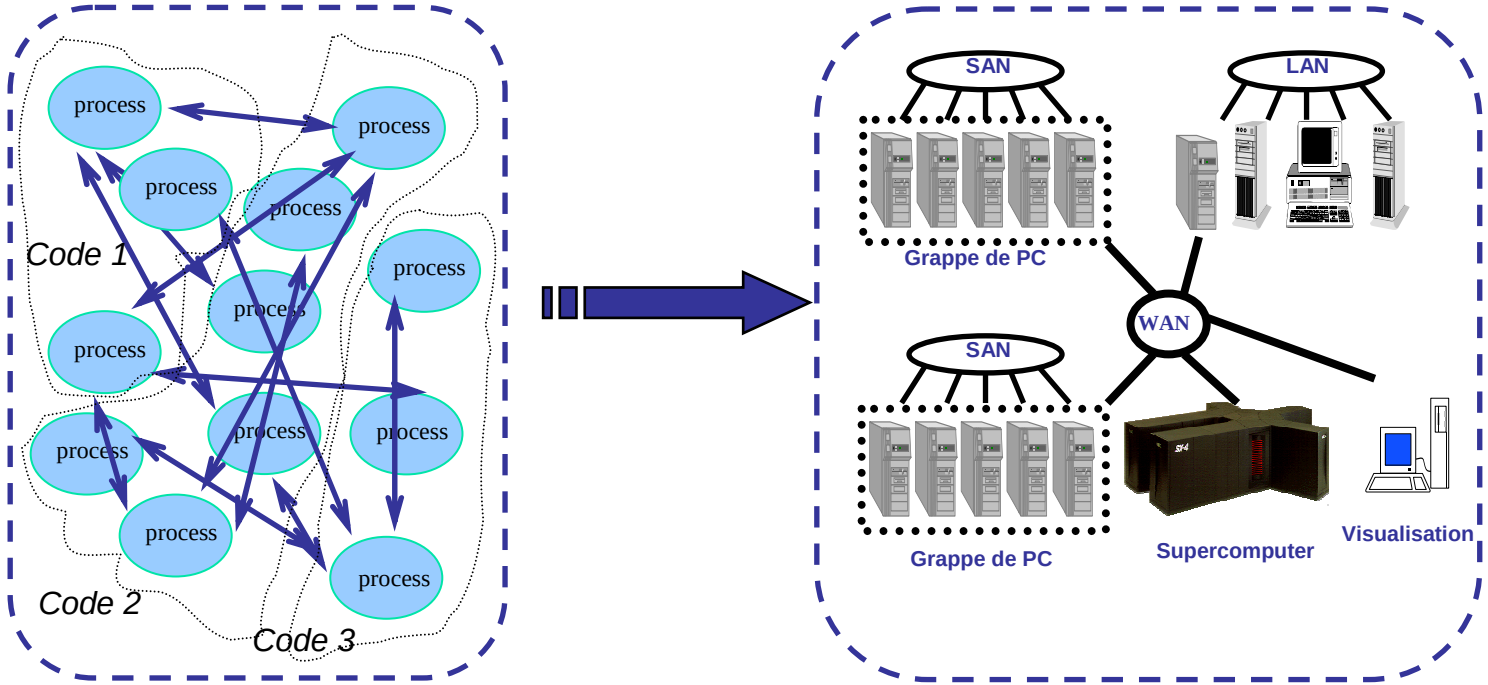


Application de couplage



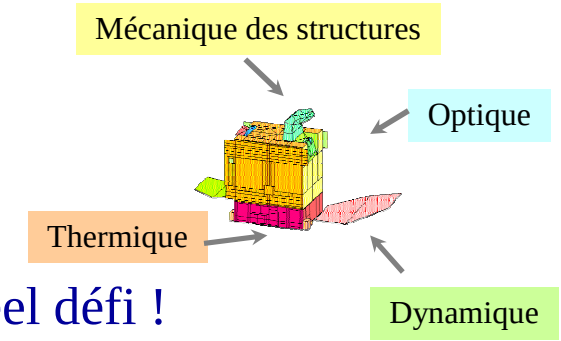
Couplage de codes

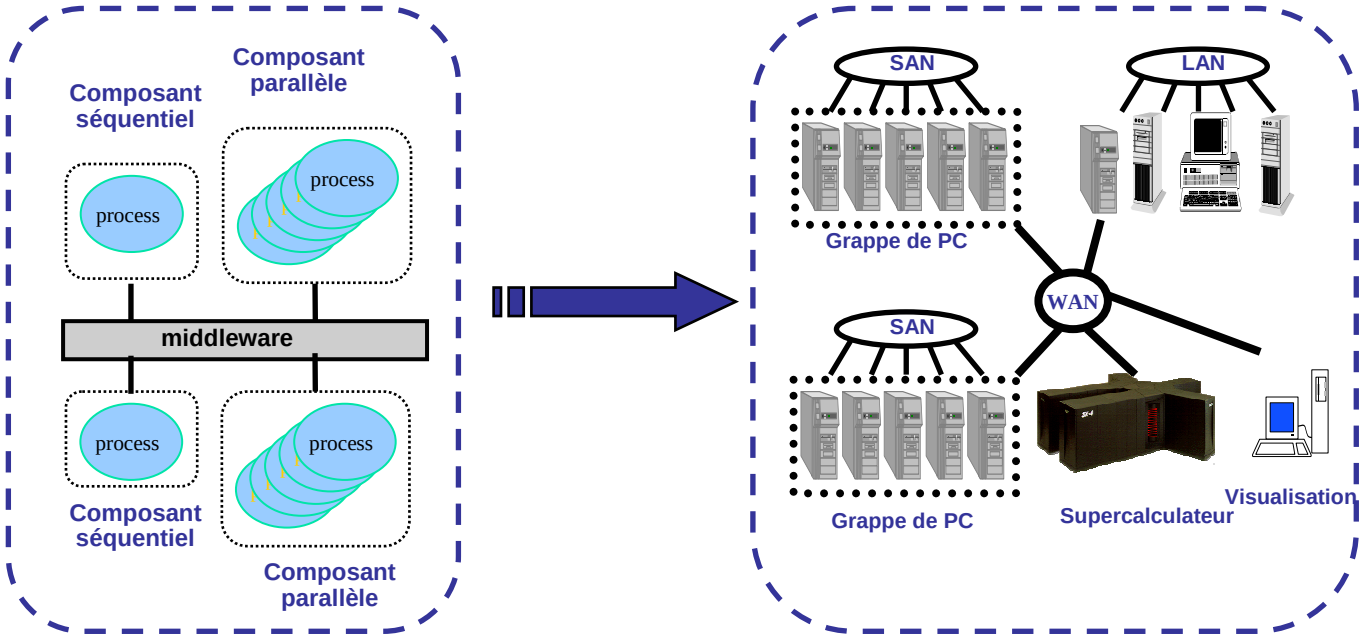
- Utilisation des exécutifs conçus pour la programmation parallèle
  - Une grille de calcul est un calculateur parallèle virtuel, la programmation par échange de messages s'impose, tout repose alors sur le Middleware…



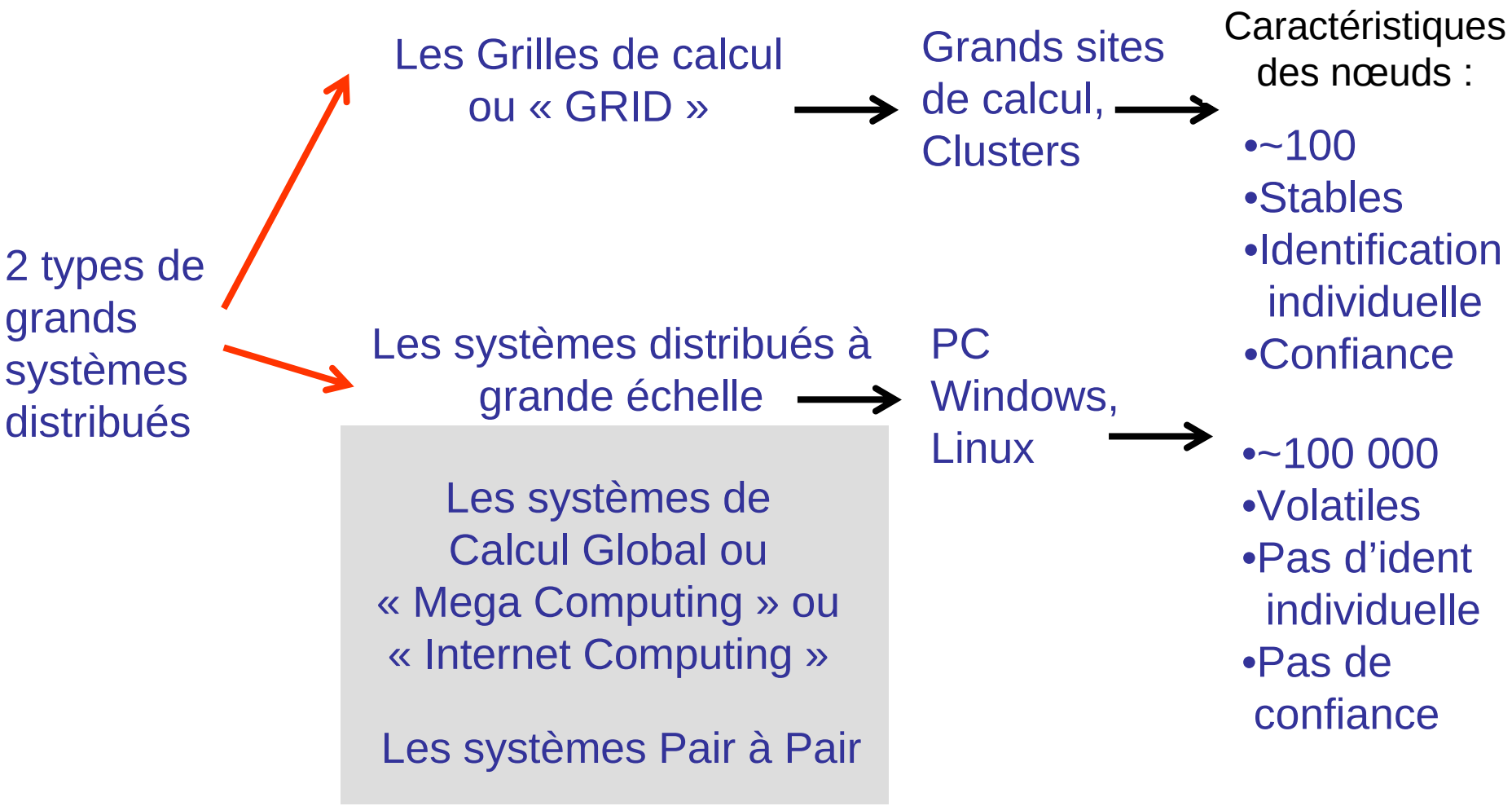
Une approche plus moderne

- Objets distribués/composants
  - Structuration de l'application
  - Encapsulation des codes
- Couplage de codes parallèles
  - Interconnexion des objets/composants ➡ un réel défi !





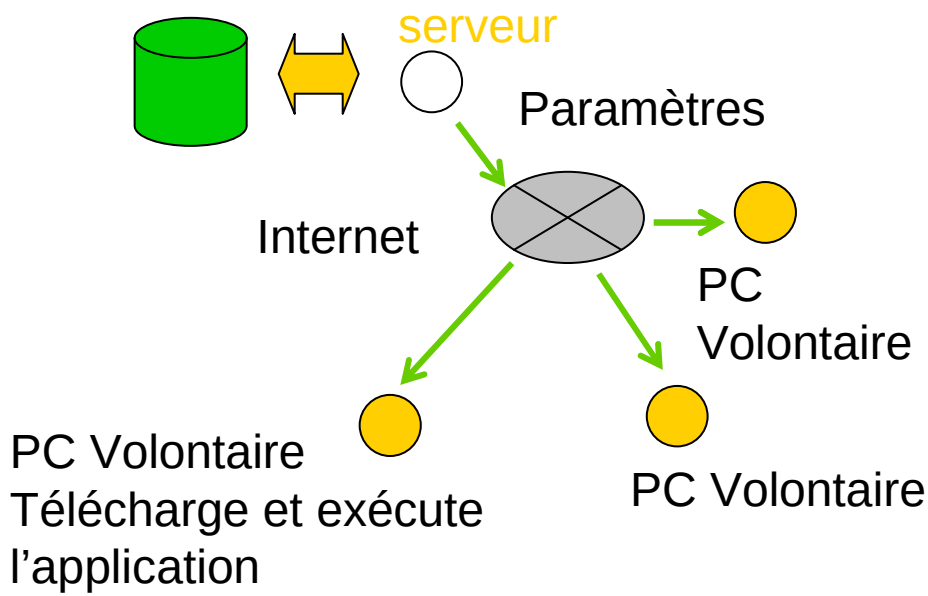
Du calcul global pair à pair



Systèmes de Calcul Global

Calcul Maître-esclave, par vol de cycles sur Internet

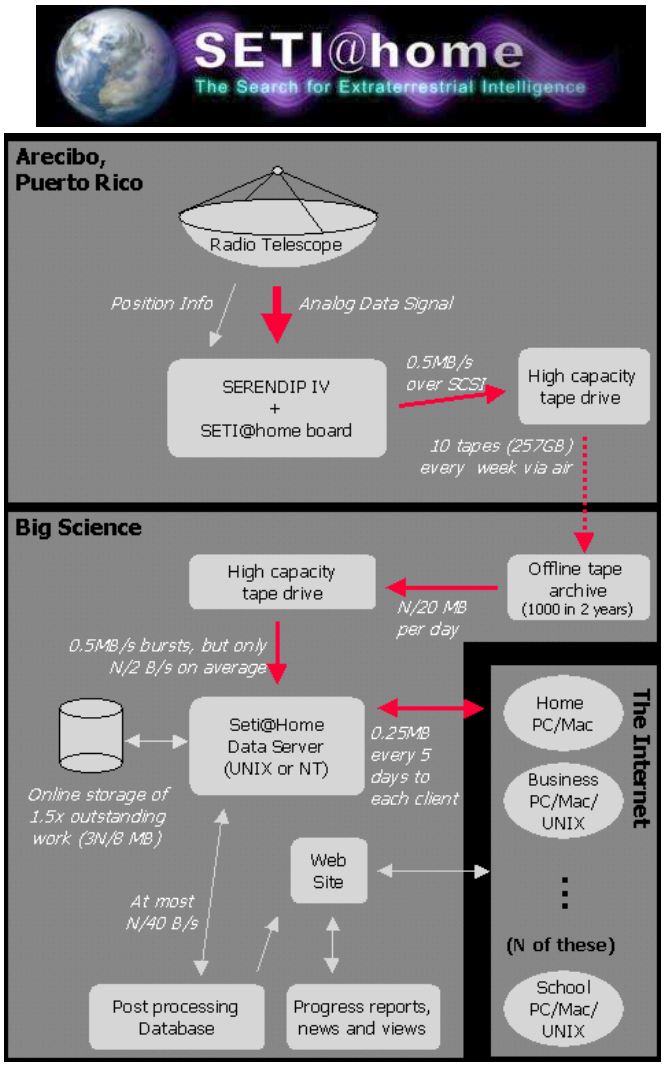
Un serveur centraliser ordonnance des calcul sur des PC volontaires



- Applications dédiées
  - SETI@Home, distributed.net,
  - Décrypthon
- Projet de production
  - Folding@home, Genome@home,
  - Xpulsar@home,Folderol,
  - Exodus, Peer review,
- Plates-formes de recherche
  - Javelin, Bayanihan, JET,
  - Charlotte (based on Java),
  - Ninf (ECL), XtremWeb (LRI),
- Plates-formes commerciales
  - Entropia, Parabon,
  - United Devices,

Large Scale Distributed Computing

- Principle
  - Millions of PCs
  - Cycle stealing
- Examples
  - SETI@HOME
    - Research for Extra Terrestrial I
    - Plusieurs dizaines de Teraflop/s
  - DECRYPTHON
    - Protein Sequence comparison
  - RSA-155
    - Breaking encryption keys



Le calcul parallèle réparti, vers le PetaFloat

Est-il possible de faire du calcul parallèle à gros grain asynchrone sur des plateformes pair à pair hétérogènes de grandes échelles?

Quelles algorithmiques?

Quels langages?

Quels stockages répartis?

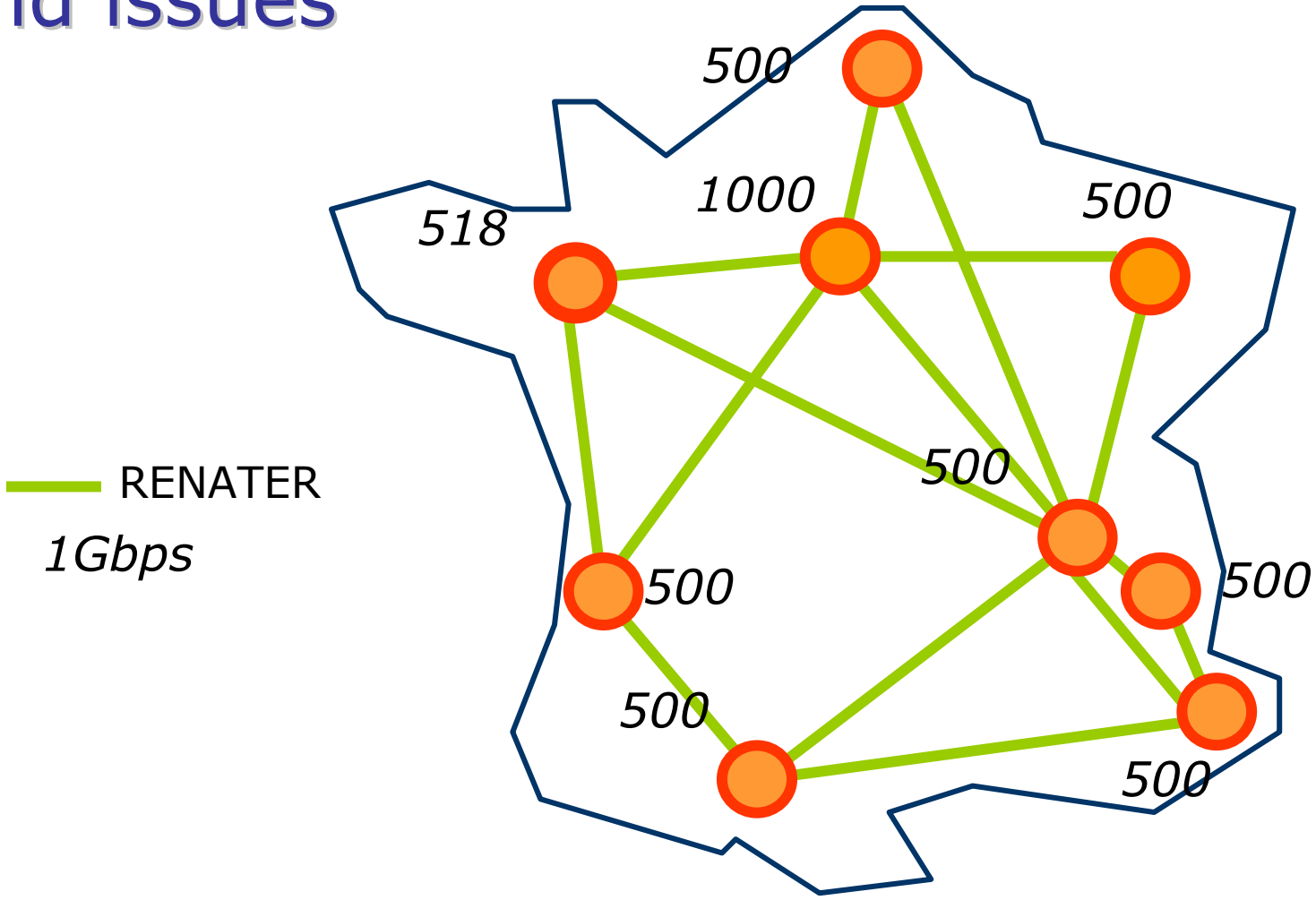
Quelles communications entre pairs : routage, *multicast* tolérant aux pannes?

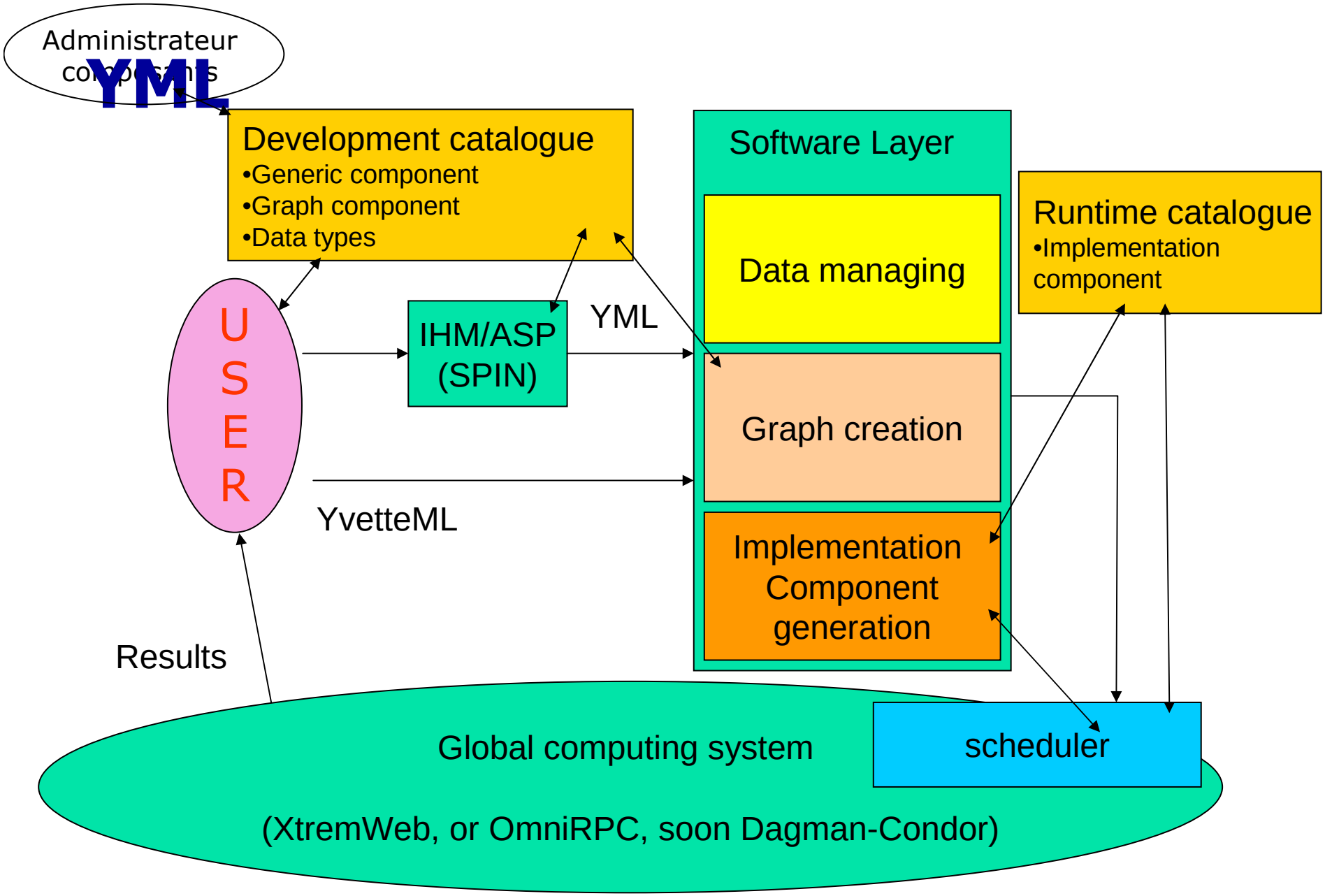
Quels systèmes répartis?

Nombreux protocoles, ordonnancements, systèmes répartis à évaluer

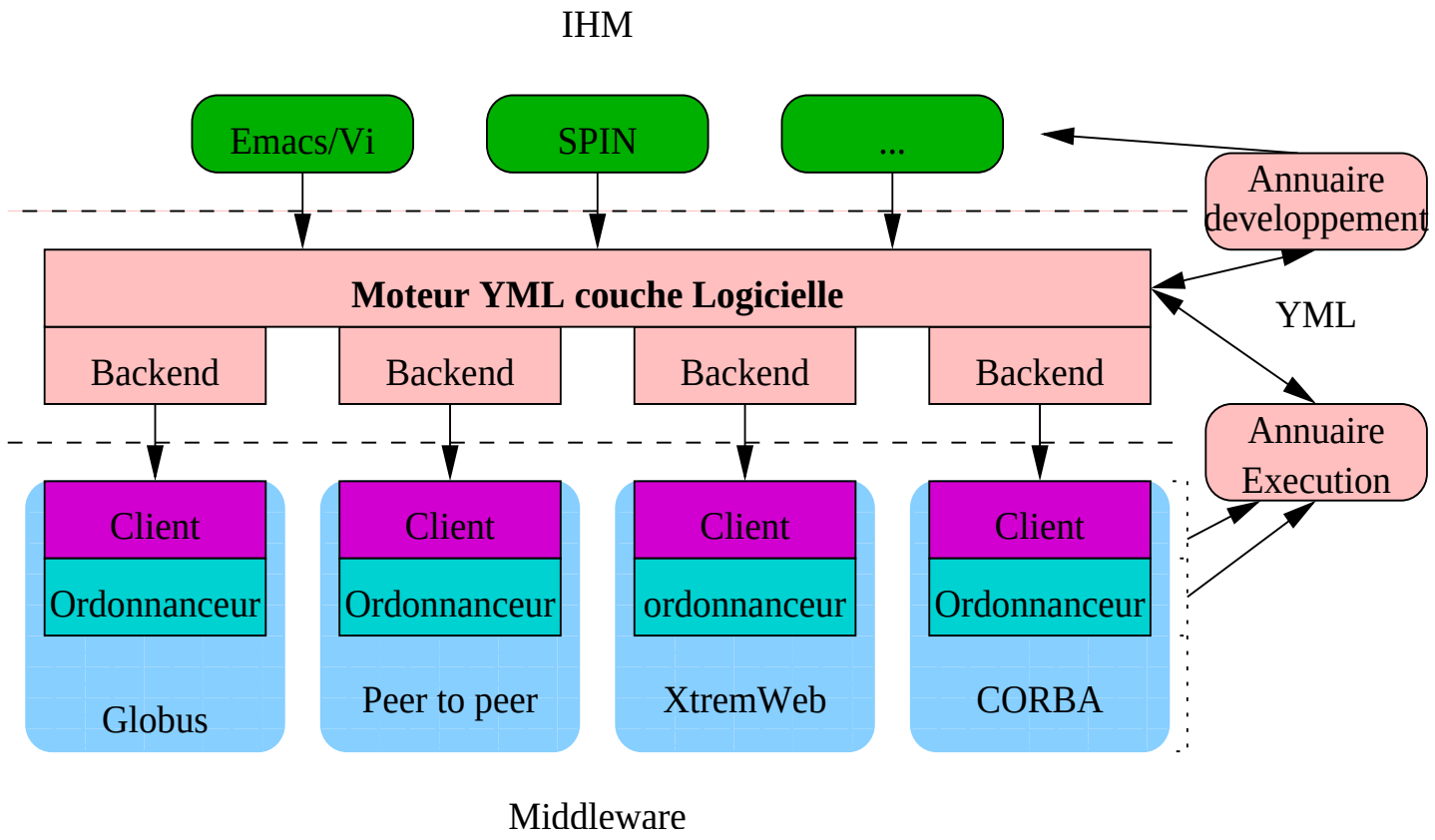
Grid'5000 map (objective)

A large scale Instrument to study Grid issues



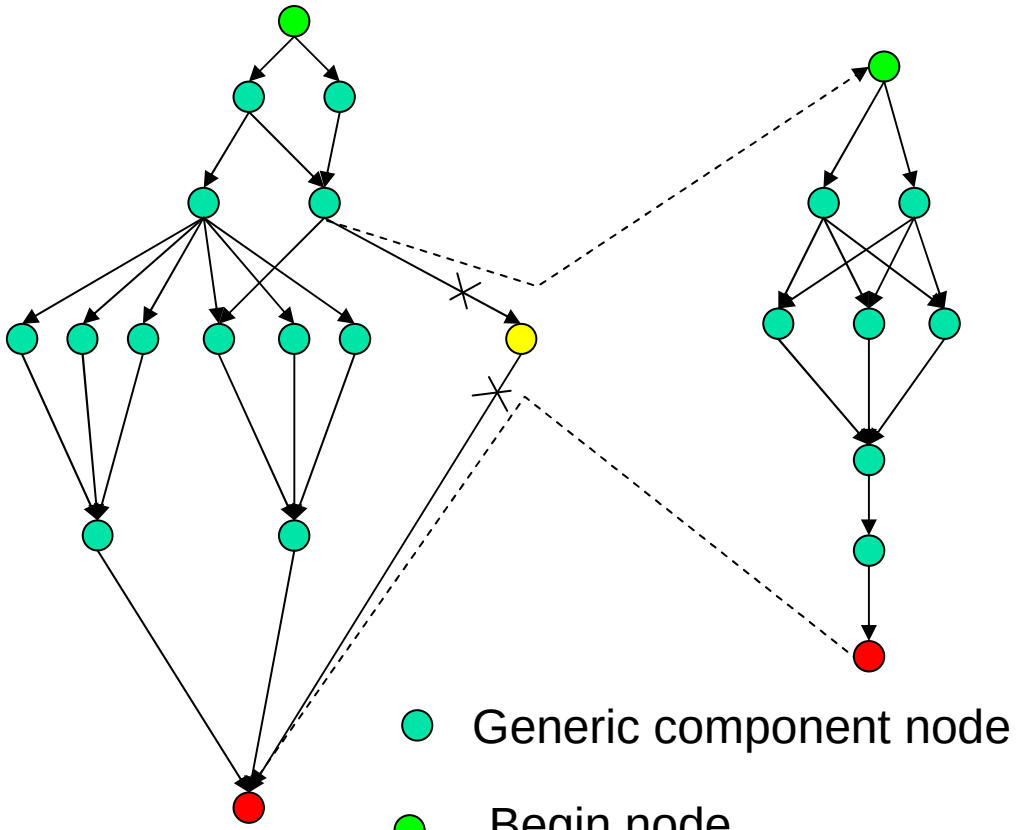

Administrateur
YML
Development catalogue
•Generic component
•Graph component
•Data types
Software Layer
Data managing
Runtime catalogue
•Implementation component
USER
IHM/ASP (SPIN)
YML
Graph creation
YvetteML
Implementation Component generation
Results
Global computing system
scheduler
(XtremWeb, or OmniRPC, soon Dagman-Condor)

YML



Backend pour Xtremweb opérationel

Exemple de graphe de composants/tâches



Generic component node

Begin node

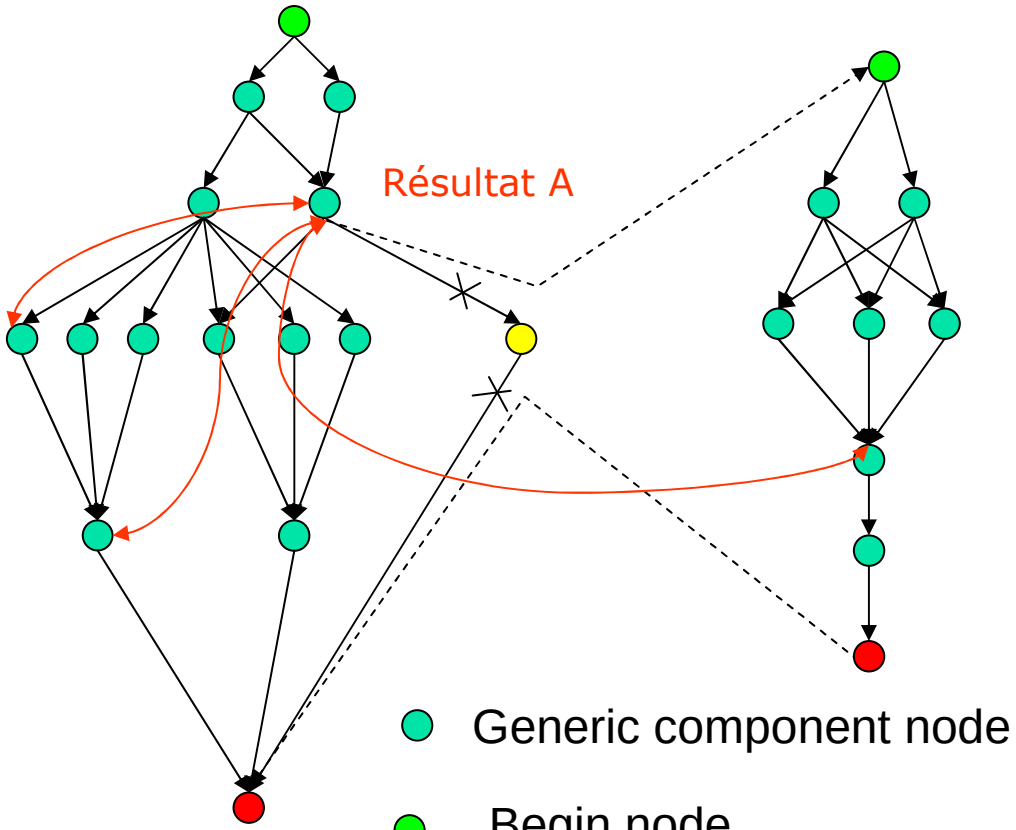
End node

Graph node

Dependence

```
par
  compute tache1(..);
  signal(e1);
//
  compute tache2(..); migrate matrix(..);
  signal(e2);
//
  wait(e1 and e2);
  Par
    compute tache3(..);
    signal(e3);
  //
    compute tache4(..);
    signal(e4);
  //
    compute tache5(..); control robot(..);
    signal(e5); visualize mesh(…) ;
  end par
//
  wait(e3 and e4 and e5);
  compute tache6(..);
  compute tache7(..);
end par
```

Exemple de graphe de composants/tâches



Generic component node

Begin node

End node

Graph node

Dependence

```
par
  compute tache1(..);
  signal(e1);
//
  compute tache2(..); migrate matrix(..);
  signal(e2);
//
  wait(e1 and e2);
  Par
    compute tache3(..);
    signal(e3);
  //
    compute tache4(..);
    signal(e4);
  //
    compute tache5(..); control robot(..);
    signal(e5); visualize mesh(…) ;
  end par
//
  wait(e3 and e4 and e5);
  compute tache6(..);
  compute tache7(..);
end par
```

Du petaflop

What architectures for Petascale computers?

- Larger cluster of clusters?
- Larger massively parallel computers?
- Moore law?
- or new architectures?

Hybrid supercomputer as a solution?

- TSUBAME, Titech, Japan, 10440 AMD + SIMD , TSUBAME2 *Clearspeed* accelerators, 21 TeraOctet Memory,1.1 PetaOctets Disk
- IBM announced a future Petascale computer ‘RoadRunner”: 16000 AMD + IBM Cell (designed for Sony PS3).
- Projet KEISOKU (perhaps : 5000 nodes. 1 vector accelerator and 8 low power (256 cores!) processors on each node. 10 Petaflops.

As a conclusion of this introduction.

- We will « soon » have Petascale computers, interconnected to GRID : supercomputer GRIDs.
- Nevertheless, May we efficiently program those computers and powerfull world-wide GRIDS?
- Many programming, algorithmical, arithmetic and compiling researches!