

Contribution à la conception de services de partage de données pour les grilles de calcul

Gabriel Antoniu

Equipe-projet PARIS

INRIA - Centre de Recherche de Rennes Bretagne Atlantique

Soutenance pour l'habilitation à diriger les recherches

5 mars 2009



INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE



centre de recherche
RENNES - BRETAGNE ATLANTIQUE

Plan de l'exposé

Introduction et contexte

Problématique de recherche

- Fil conducteur : fournir un accès transparent aux données
- Etapes et actions

Zoom

- Etape 1 : GDS = DSM + P2P
- Etape 2 : Cohérence des données et tolérance aux fautes
- Actions transversales
 - Expérimentation et déploiement à grande échelle
 - Intégration avec des modèles existants

Bilan et perspectives



Contexte applicatif : simulations numériques

Objectifs

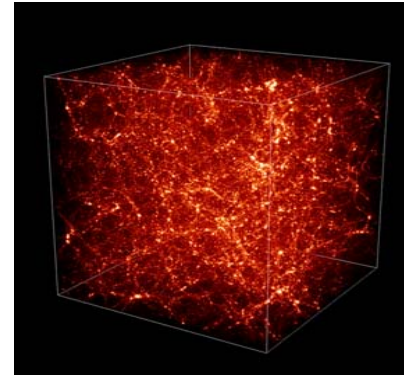
- Plus de précision
- Plus de réalisme

Besoins

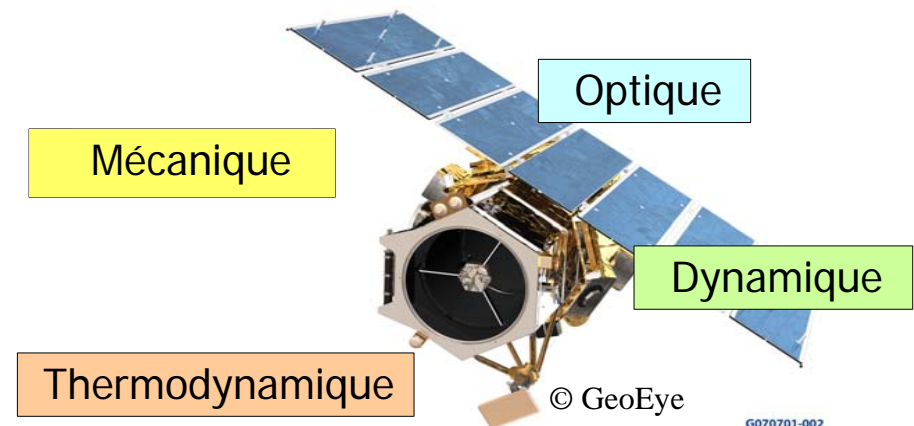
- Puissance de calcul
- Espace de stockage
- Bande passante

Exemples de types d'applications

- Couplage de codes
- Simulations multi-paramétriques



Formation des structures de l'Univers (CEA)



Architectures visées : les grilles de calcul



Analogie avec le réseau de distribution d'électricité

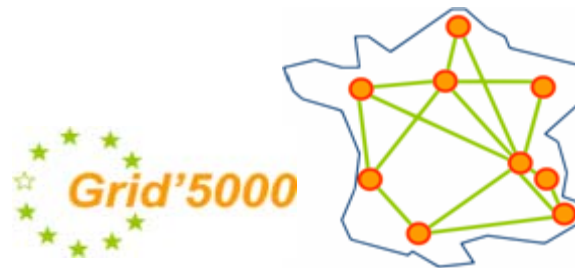
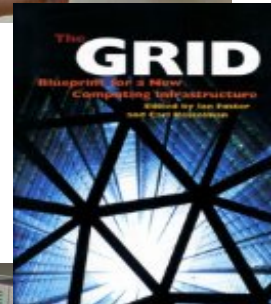
- Simplicité d'utilisation : branchement sur une prise
- **Transparence** : localisation, gestion des ressources

Des ressources virtuellement infinies

- **Agrégation et mutualisation** (calcul, transport et stockage)
- Organisation virtuelle, physiquement répartie

Un cas particulier : les fédérations de grappes

- Topologie **hiérarchique**
- Echelle : 10^3 à 10^4
- Ressources **hétérogènes**
- **Dynamicité** de l'infrastructure



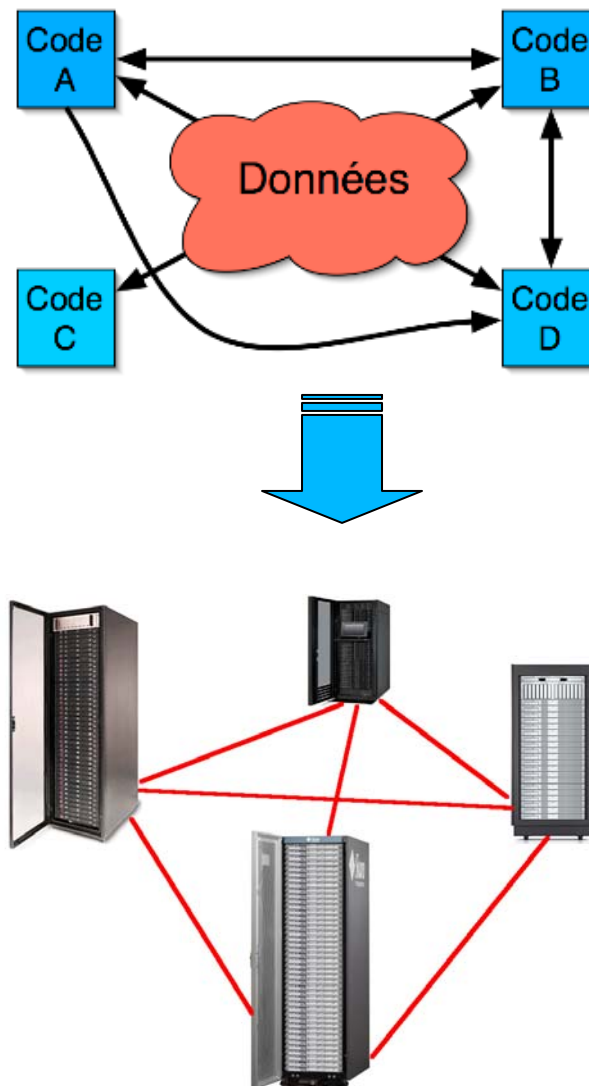
Problématique : la gestion des données

Hypothèses

- Applications et données réparties
- Données partagées

Propriétés visées

- **Transparence**
 - Localisation, transfert
- **Cohérence**
 - Données répliquées sur plusieurs sites
- **Persistence**
 - Dépendance de données entre calculs
- **Tolérance à la volatilité et aux fautes**



Etapes et actions

Transparence : DSM + P2P
Thèse de Mathieu Jan (2003 - 2006)



Etapes et actions

Cohérence et tolérance aux fautes
Thèse de Sébastien Monnet (2003 - 2006)

Transparence : DSM + P2P
Thèse de Mathieu Jan (2003 - 2006)

Etapes et actions

Déploiement
Thèse de Loïc Cudennec (2005 - 2009)

Cohérence et tolérance aux fautes
Thèse de Sébastien Monnet (2003 - 2006)

Transparence : DSM + P2P
Thèse de Mathieu Jan (2003 - 2006)



Etapes et actions

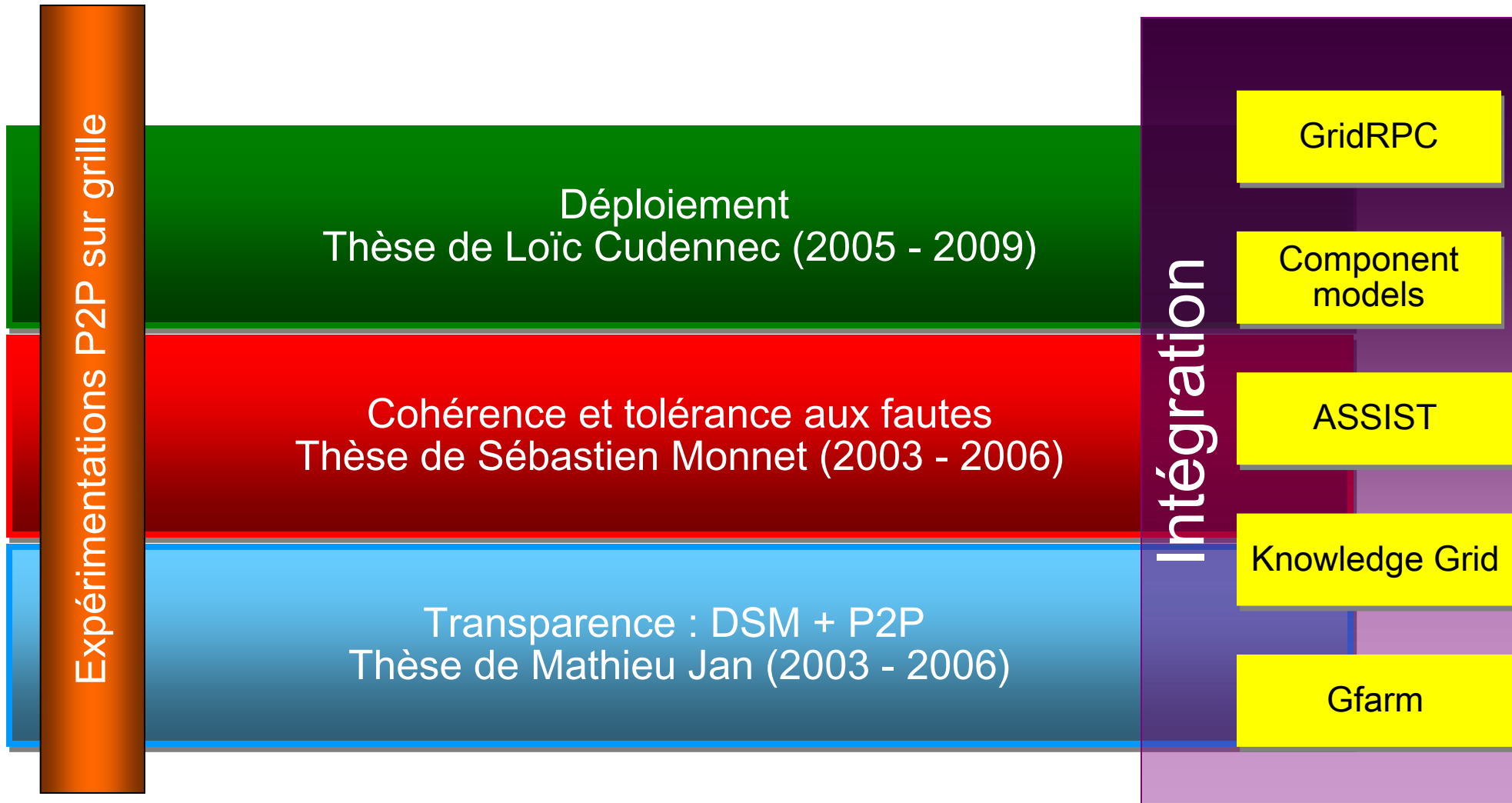
Expérimentations P2P sur grille

Déploiement
Thèse de Loïc Cudennec (2005 - 2009)

Cohérence et tolérance aux fautes
Thèse de Sébastien Monnet (2003 - 2006)

Transparence : DSM + P2P
Thèse de Mathieu Jan (2003 - 2006)

Etapes et actions



Etape 1 : permettre un accès transparent aux données

Jalon : définition d'une architecture de service de partage de données pour grille

Thèse de Mathieu Jan (2003-2006) - INRIA, Région Bretagne

- Ingénieur-chercheur au CEA, LIST, LaSTRE (depuis décembre 2007)

Contribution-clé

- **JuxMem: architecture basée sur l'approche: GDS = DSM + P2P**

Support

- Projets GDS et GdX de l'ACI Masses de Données (2003-2006)
- Projet ACI GRID DataGraal (2002-2004)



2002 : que faisaient les autres ?



Approche majoritaire

- Gestion **explicite** des données par les applications
 - Localisation
 - Transfert
 - Cohérence des copies
 - Pas de tolérance à la volatilité
- ➔ Une complexité qui augmente avec l'échelle



Défi : un partage **transparent** des données, en tolérant la **volatilité**



Première source d'inspiration : systèmes à mémoire virtuellement partagée (DSM)

Propriétés

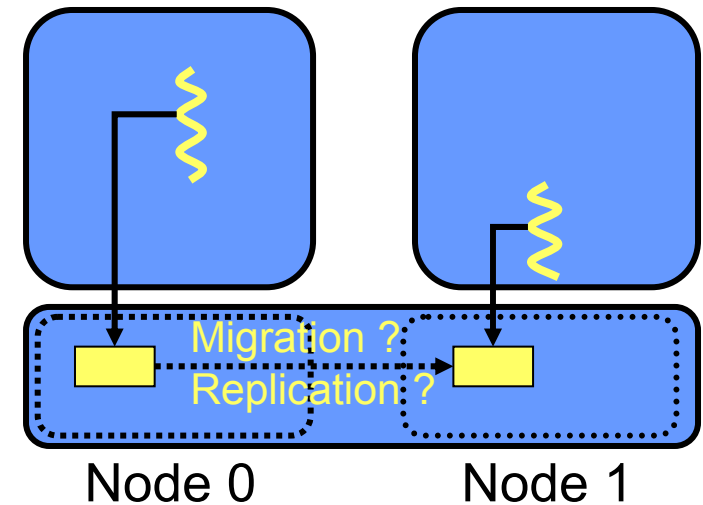
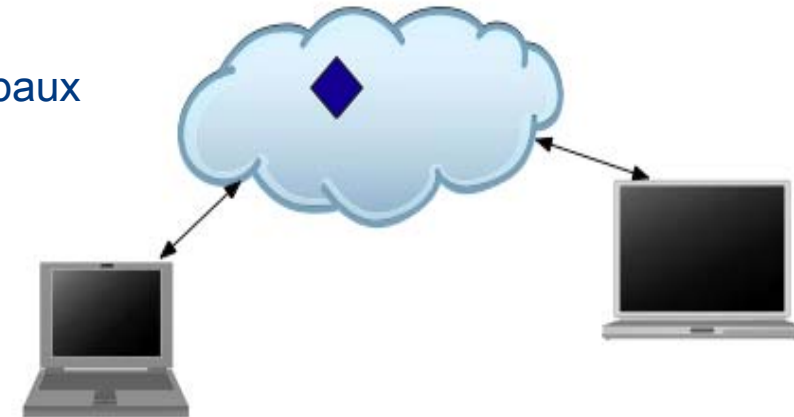
- Accès uniforme aux données via des identifiants globaux
- Localisation et transfert **transparents**
- Modèles et protocoles de **cohérence**

Limitations

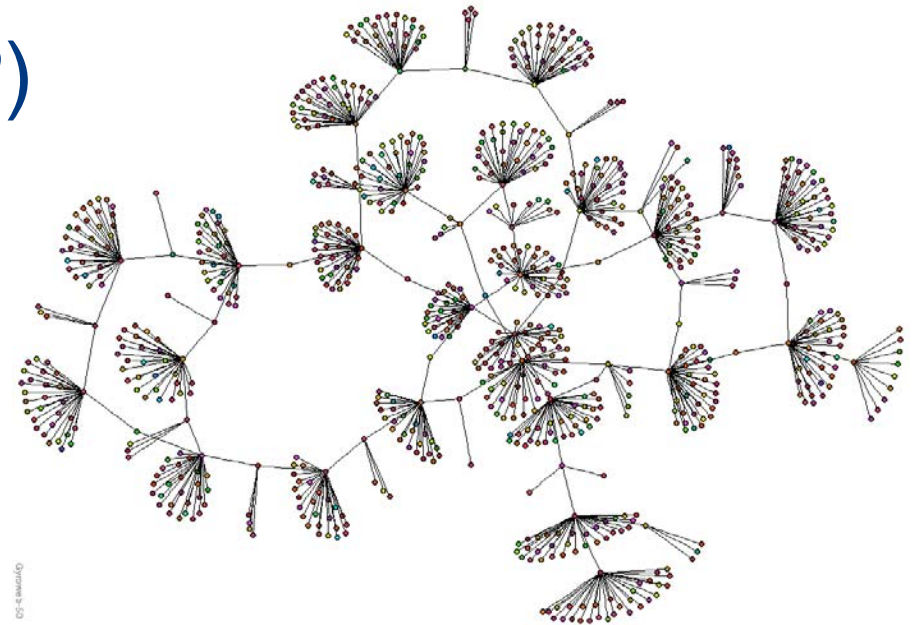
- Petite échelle (grappes),
- Architecture statique
 - Généralement pas de tolérance à la volatilité

Défis pour les grilles

- Intégrer de nouvelles hypothèses !
 - **Passage à l'échelle : grille**
 - **Dynamicité des ressources**
 - **Tolérance aux fautes**



Deuxième source d'inspiration : systèmes pair-à-pair (P2P)



Propriétés

- Excellent passage à l'échelle
- Haute tolérance à la volatilité

Limitations

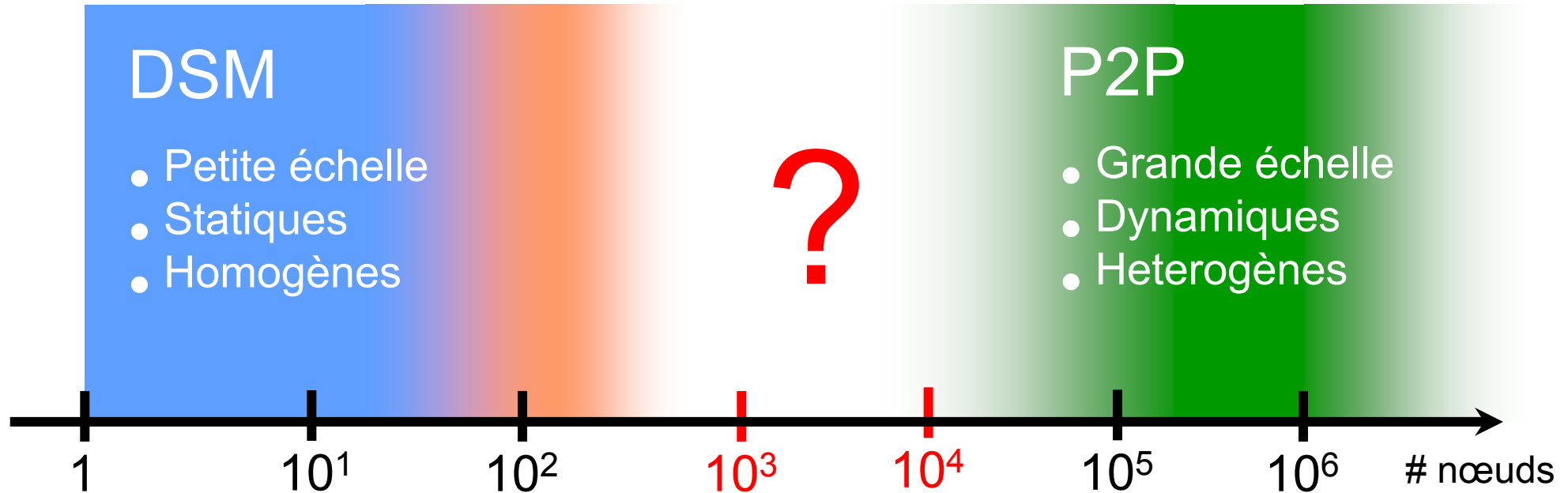
- Données partagées en lecture seule (majoritairement)
- Quelques exceptions: Ivy [MIT], OceanStore [UCB] , Pastis [LIP6, France]

Défi

- Partager des données modifiables















Partage de données : le défi !



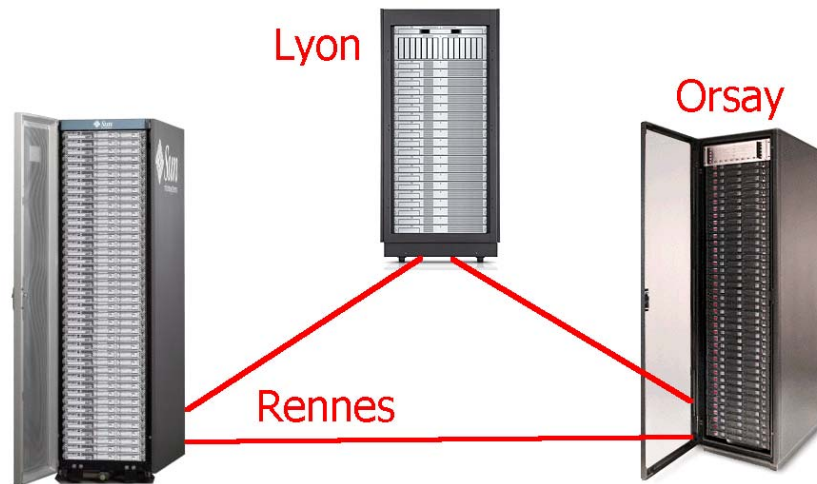
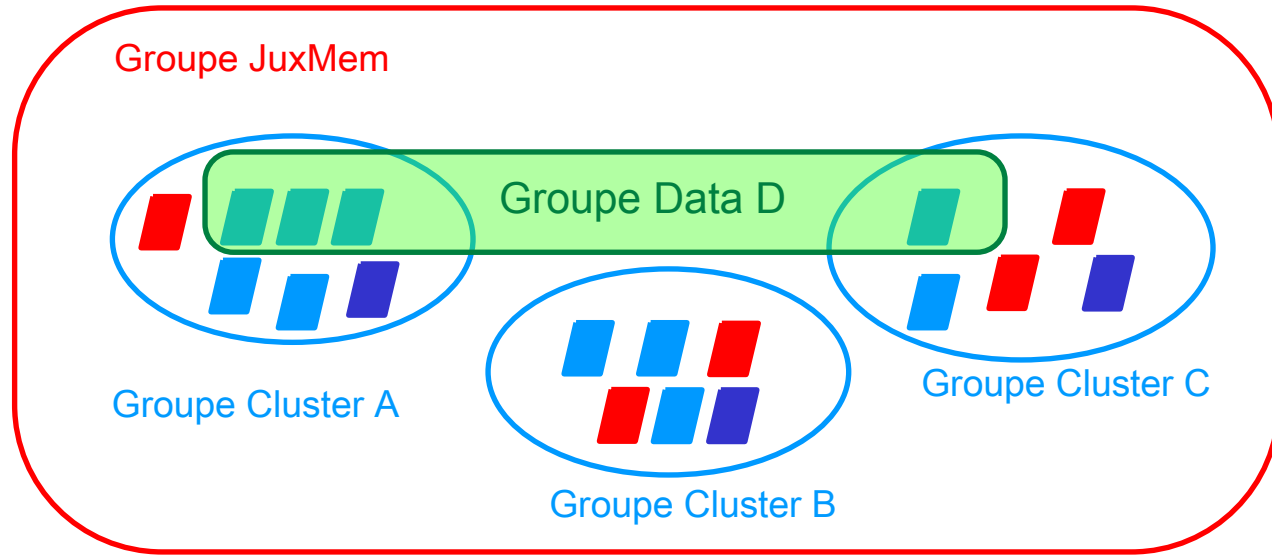
Proposition : service de partage de données pour grille

Grid Data-Sharing Service (GDS)

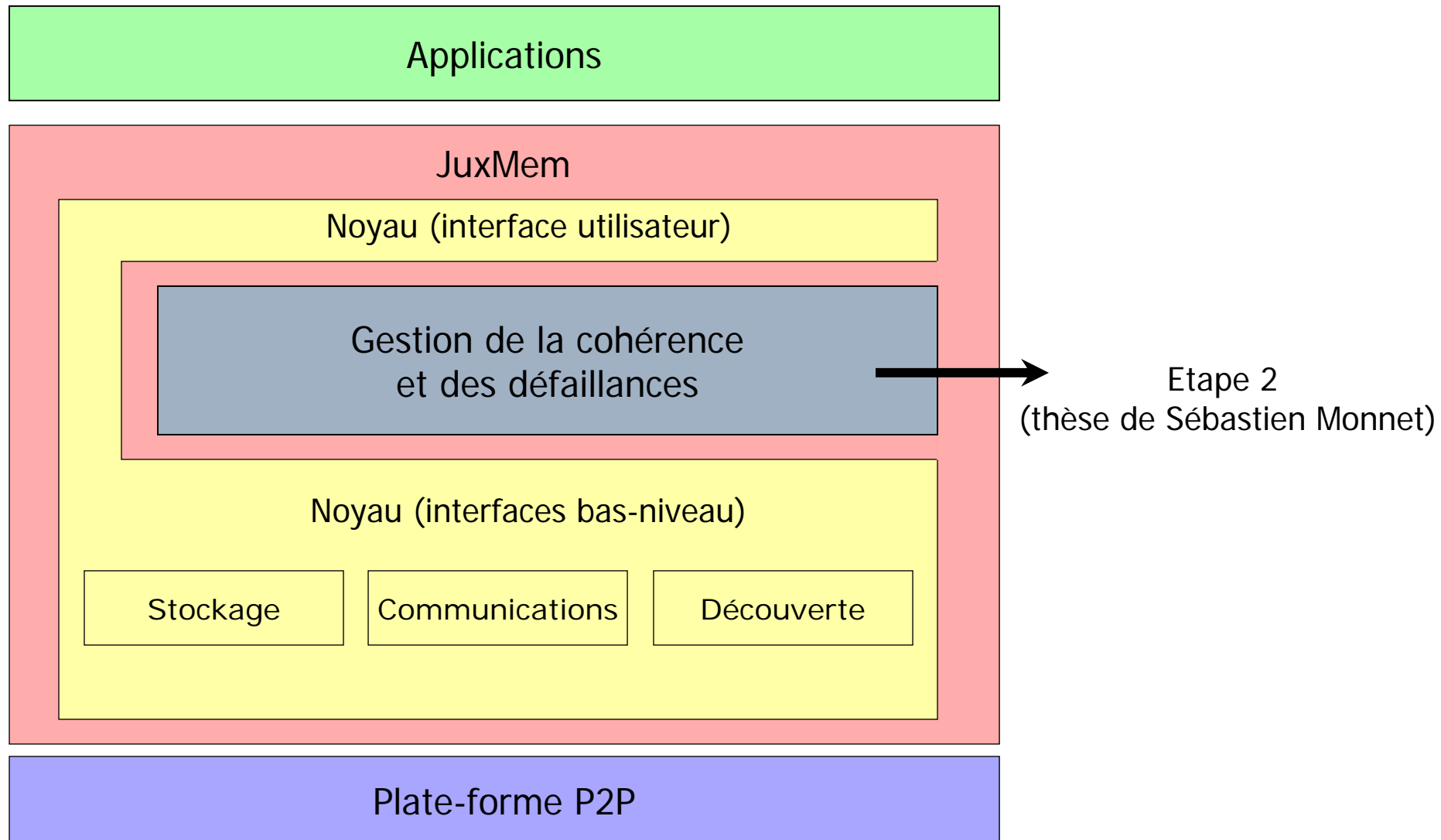
	DSM	GDS	P2P
Plate-forme	Grappe	Grille	Internet
Echelle	10^1 - 10^2	10^3 - 10^4	10^5 - 10^6
Applications	Calcul	Calcul et stockage	Stockage
Transparence			
Cohérence			
Persistance			
Tolérance à la volatilité			



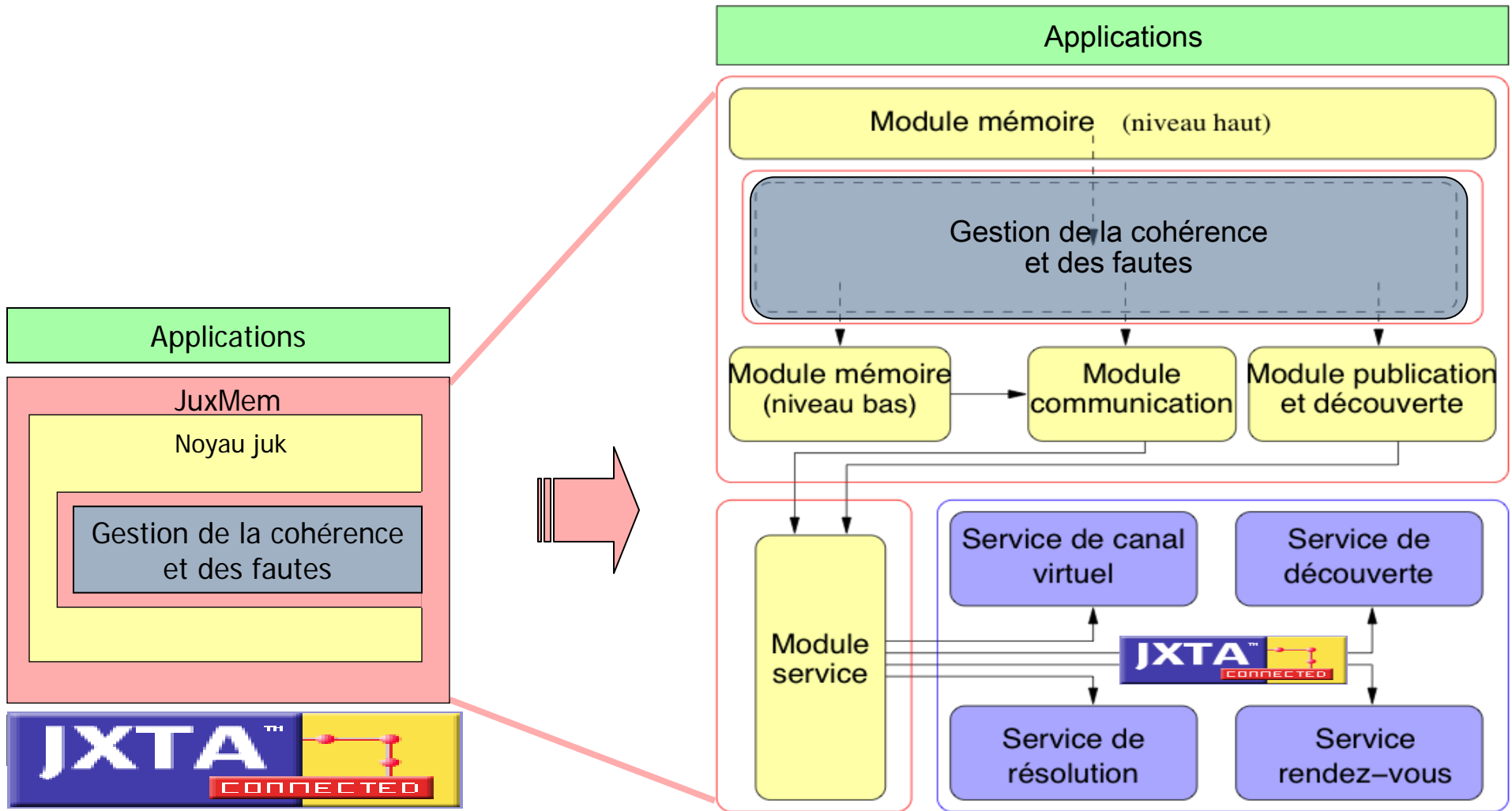
JuxMem : une architecture hiérarchique répartie



JuxMem : une architecture en couches



JuxMem : mise en œuvre sur JXTA



Example d'utilisation:

Client :

```
idaA = juxmem_malloc(size, attr, &ptrA) // attr = global_repl, local_repl, consistency_prot  
idaB = juxmem_malloc(size, ..., &ptrB, ...)  
initializeA(&ptrA)  
initializeB(&ptrB)  
idC = remote_multiply(multiply, idA, idB)  
... // wait to be notified
```

Server :

```
local_ptrA = juxmem_mmap(idA, ...)  
juxmem_acquire_read(local_ptrA)  
local_ptrB = juxmem_mmap(idB, ...)  
juxmem_acquire_read(local_ptrB)  
multiply(local_ptrA, local_ptrB, local_ptrC)  
juxmem_release(local_ptrA)  
juxmem_release(local_ptrB)  
idC = juxmem_attach(local_ptrC, ...)□
```

```
ptrC = juxmem_mmap(idC, ...)  
juxmem_acquire_read(ptrC)  
...  
juxmem_release(ptrC)
```

Publications

1. Gabriel Antoniu, Luc Bougé and Mathieu Jan. Peer-to-Peer Distributed Shared Memory? In *Proc. IEEE/ACM 12th Intl. Conf. on Parallel Architectures and Compilation Techniques (PACT 2003)*, Work in Progress Session, Pages 1-6, New Orleans, Louisiana, September 2003.
2. Gabriel Antoniu, Luc Bougé and Mathieu Jan. JuxMem: An Adaptive Supportive Platform for Data Sharing on the Grid. In *Scalable Computing: Practice and Experience*, Vol. 6(3):45-55, September 2005.
3. Gabriel Antoniu, Marin Bertier, Eddy Caron, Frédéric Desprez, Luc Bougé, Mathieu Jan, Sébastien Monnet and Pierre Sens. Future Generation Grids. In Vladimir Getov, Domenico Laforenza and Alexander Reinefeld editors, p. 133-152, **Chapter GDS: An Architecture Proposal for a Grid Data-Sharing Service**, Springer Verlag, 2006.



Bilan

Contributions

- Idée : fournir un **modèle d'accès transparent** aux données à l'échelle d'une grille
- **JuxMem: architecture hybride basée sur l'approche: GDS = DSM + P2P**
- Mise en œuvre sur JXTA: <http://juxmem.gforge.inria.fr> (LGPL)

Qu'avons-nous appris ?

- Faisabilité confirmée
- Performances : besoin d'adapter les protocoles P2P pour une exécution sur grille
 - Améliorer le coût des transferts des données
 - Mettre en place le support pour un déploiement efficace



Etapes et actions

Cohérence et tolérance aux fautes
Thèse de Sébastien Monnet (2003 - 2006)

Transparence
Thèse de Mathieu Jan (2003 - 2006)



Etape 2 : garantir la cohérence des données malgré les défaillances

Jalon : gestion conjointe de la cohérence des données et des fautes

Thèse de Sébastien Monnet (2003-2006) - MESR (GDS, ACI MD)

- Maître de conférence au LIP6, Université Paris 6 (depuis septembre 2007)

Contributions

- **Groupe auto-organisant hiérarchique, tolérant aux fautes**
- Gestion probabiliste de la cohérence des copies au sein des groupes dynamique de grande taille

Support

- Projet GDS de l'ACI Masses de Données (2003-2006)
 - Collaboration avec Pierre Sens, équipe-projet REGAL, LIP6
- Projet bilatéral INRIA-University of Illinois at Urbana Champaign (2005-2006)
 - Collaboration avec Indranil Gupta

Cohérence des données répliquées

Systèmes à mémoire virtuellement partagée (DSM)

- Cohérence forte (*strict, sequential, atomic, causal, ...*)
- Cohérence relâchée (*weak, release, lazy release, entry, scope, ...*)

Systèmes pair-à-pair (P2P)

- Peu étudiée
- Inspirée par les modèles orientés fichiers : *close-to-open, read-your-writes*

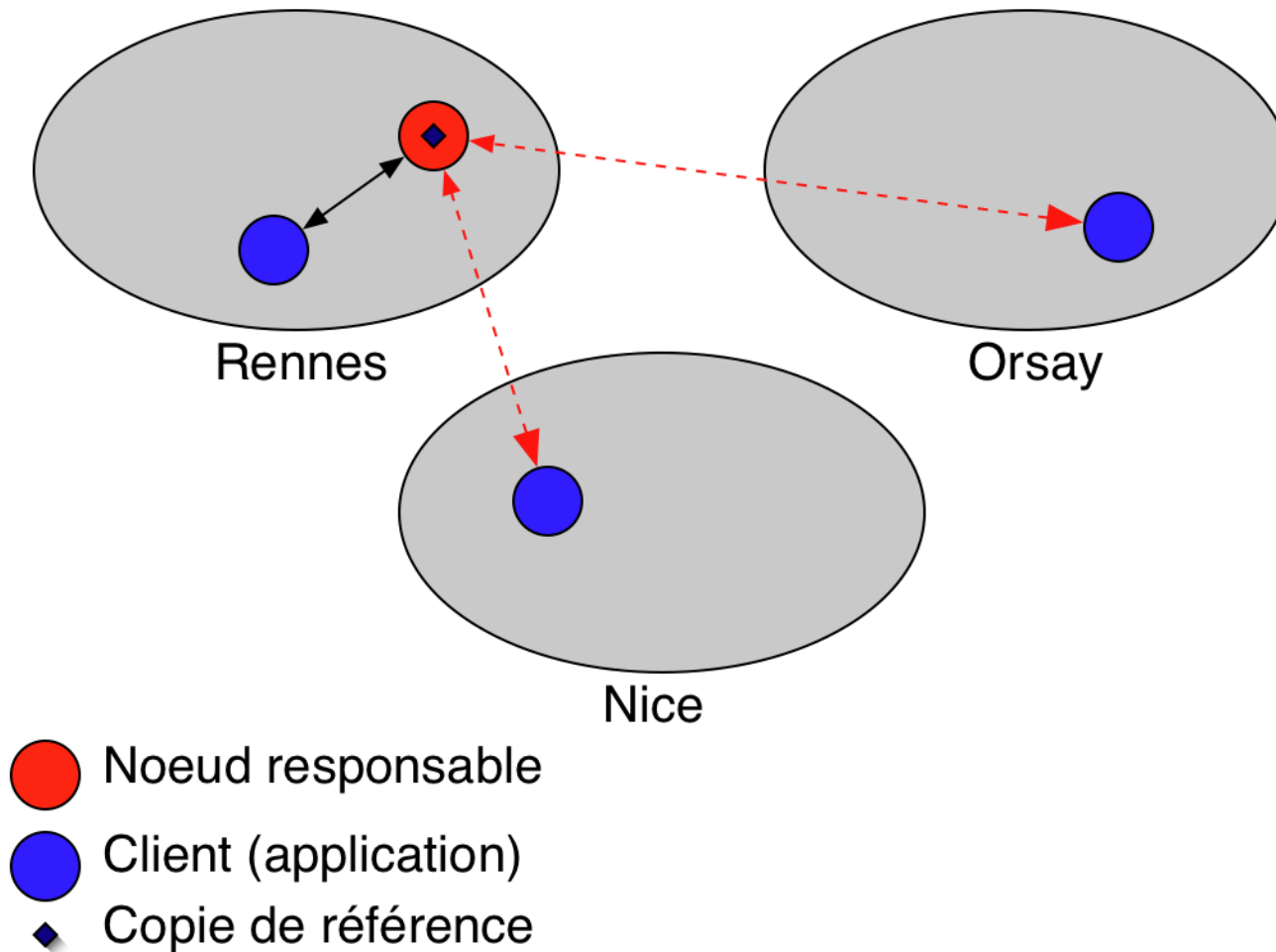
Bases de données

- Cohérence intra-données / inter-données
- Transactions : sérialisation, cohérence à terme (*eventual consistency*)

	DSM	P2P	Bases de données
Modèles de cohérence	Nombreux	Rudimentaires	Spécifiques
Passage à l'échelle	Non	Très bon	Bon
Tolérance aux fautes	Non	Très bonne	Bonne



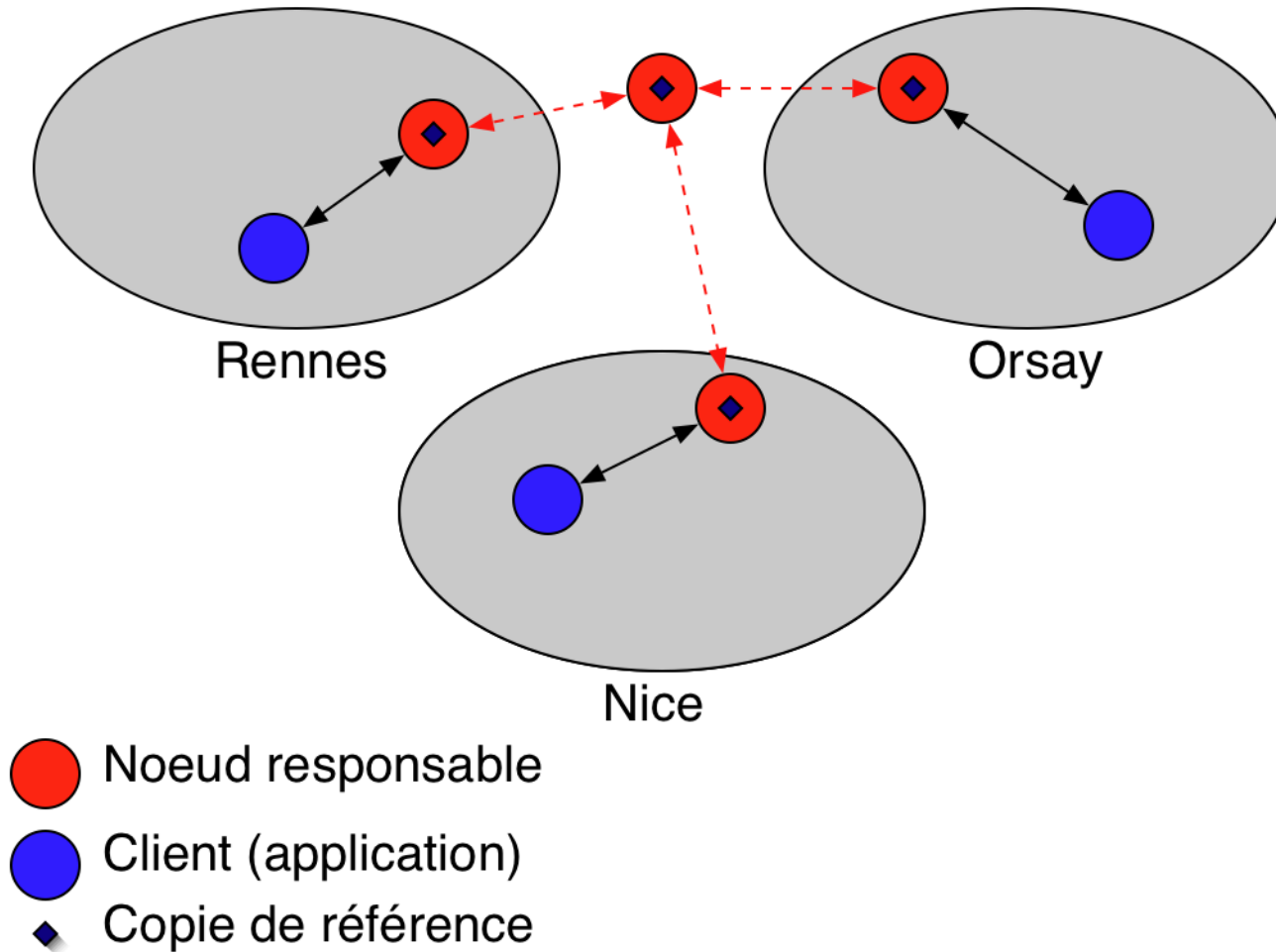
Question 1: comment passer à l'échelle ?



Ratio des latences de l'ordre de **1000** !



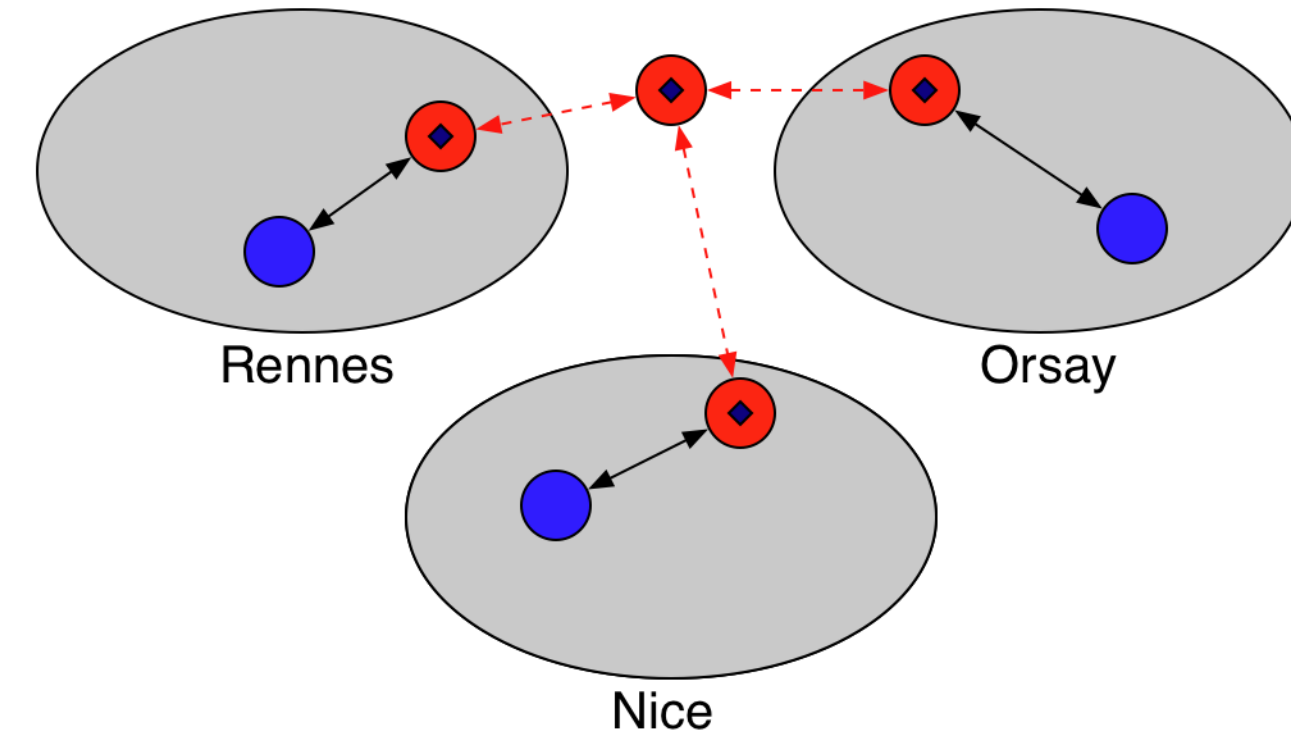
Solution : approche hiérarchique






Basée sur CLRC[LIP6] et H2BRC[IRISA]



Question 2 : comment tolérer les fautes ?

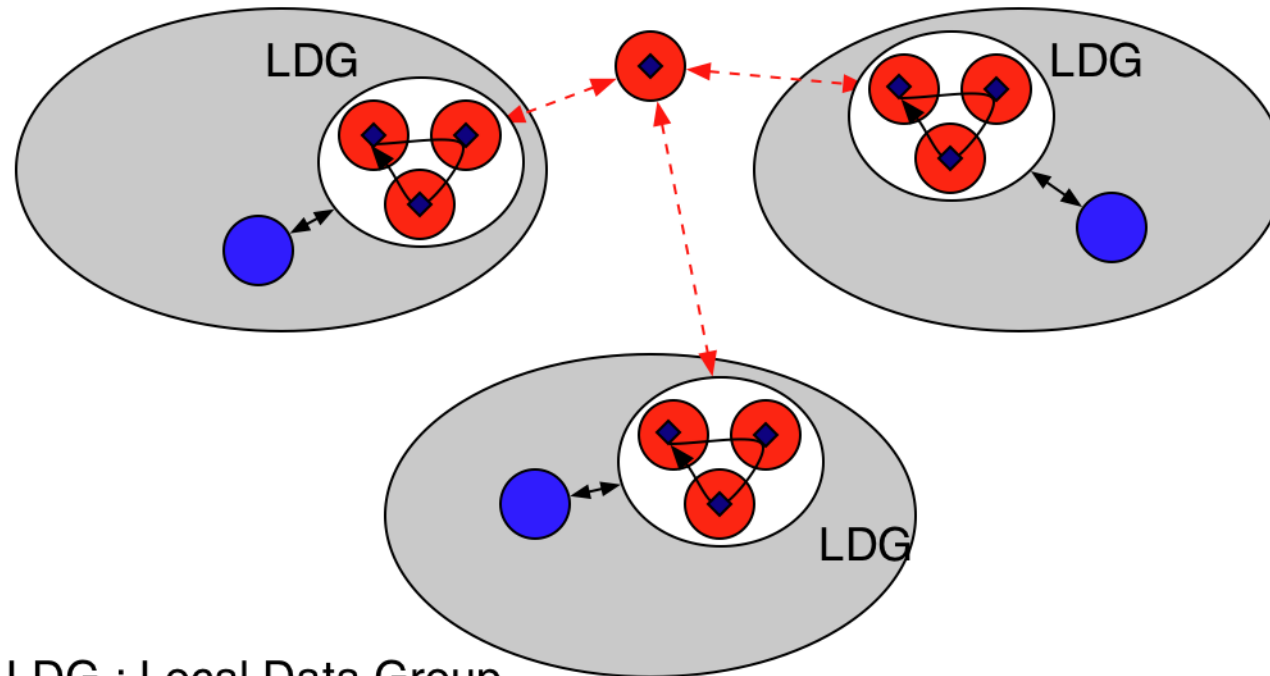


-  Noeud responsable
-  Client (application)
-  Copie de référence

Hypothèses : fautes (crashes) et déconnexions

Problème : disponibilité des données

Solutions : réplication des entités critiques

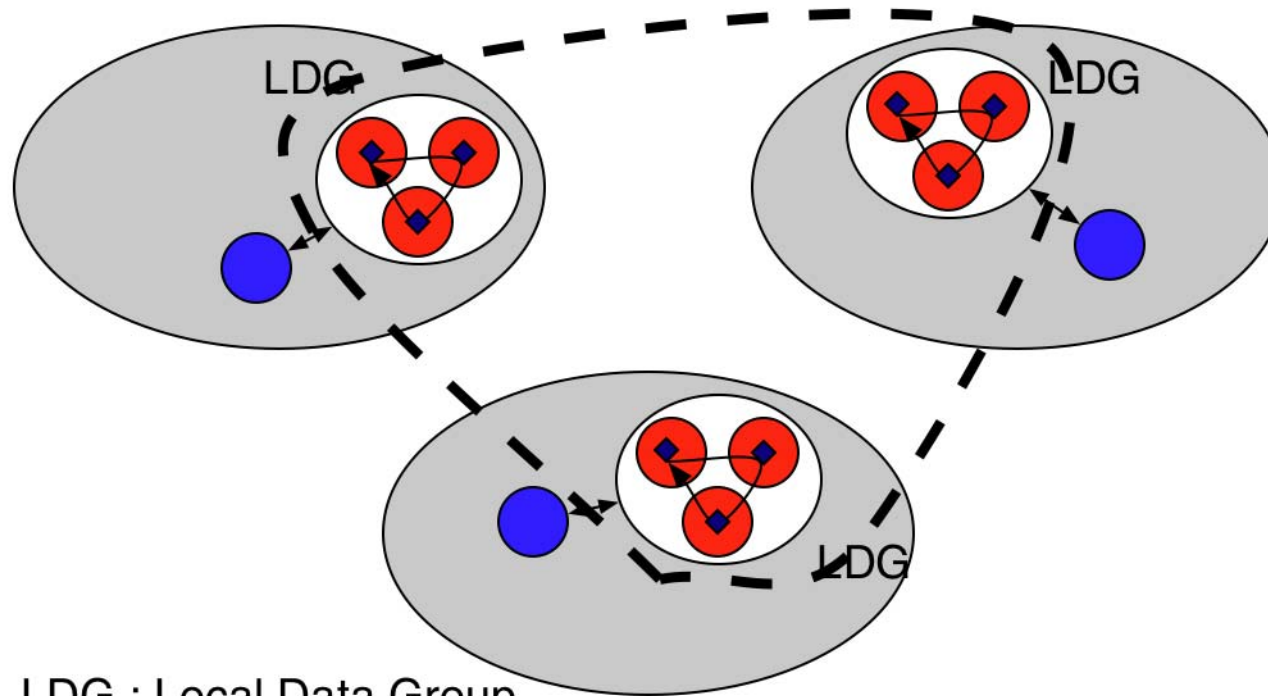


LDG : Local Data Group

Solution utilisée pour fiabiliser de nombreux systèmes

- DARX [LIP6], HORUS [Cornell], LegionFS [U.Va], etc.
- Composition de groupe, diffusion atomique

Couplage des solutions : réplication hiérarchique

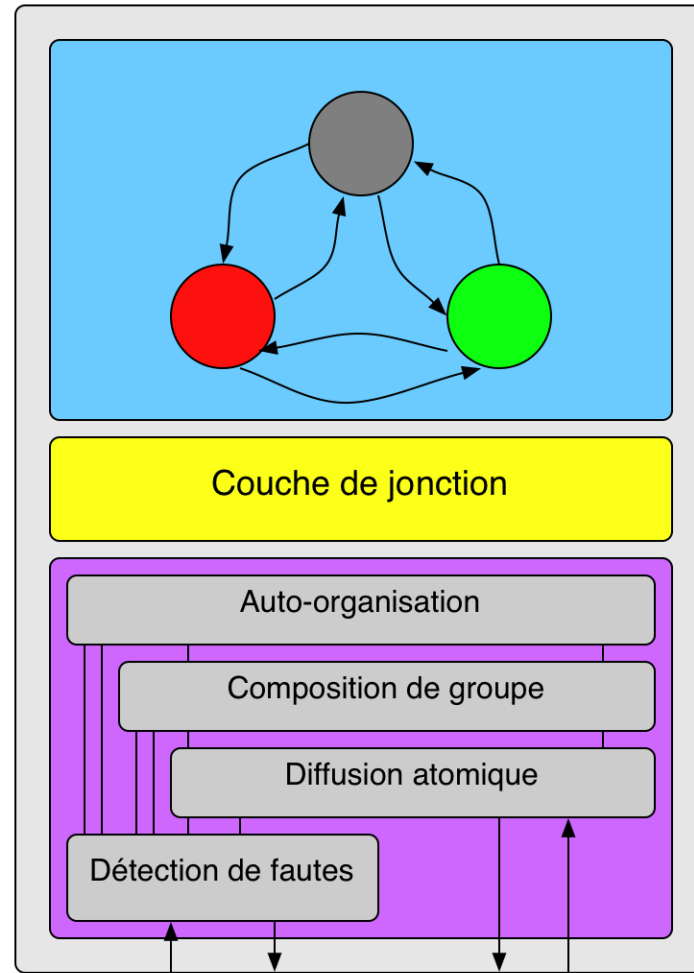


LDG : Local Data Group
GDG : Global Data Group

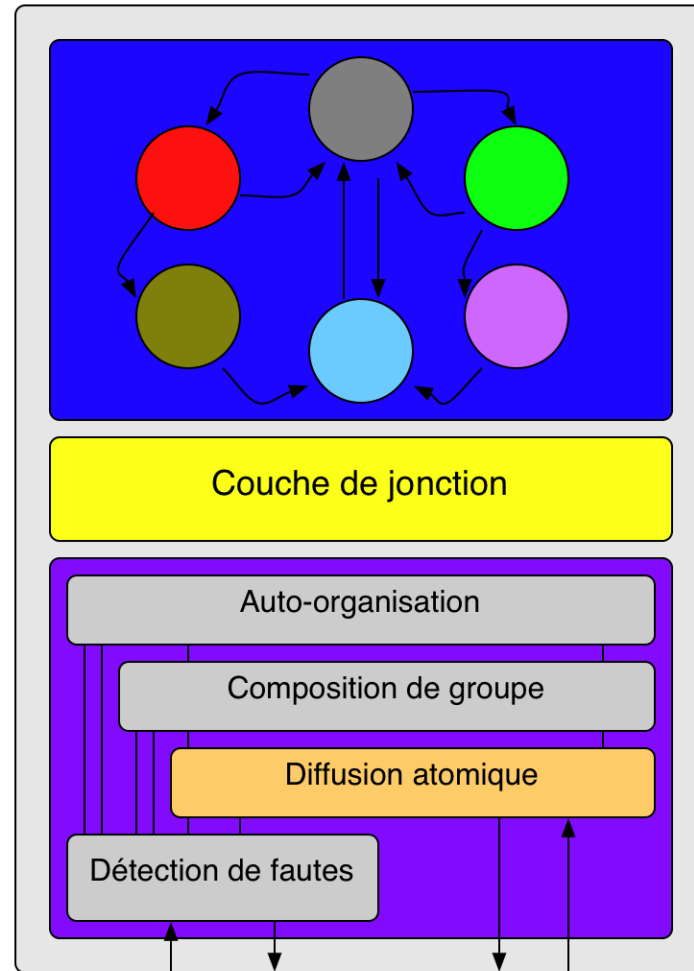
Solution utilisée pour fiabiliser de nombreux systèmes

- DARX [LIP6], HORUS [Cornell], LegionFS [U.Va], etc.
- Composition de groupe, diffusion atomique

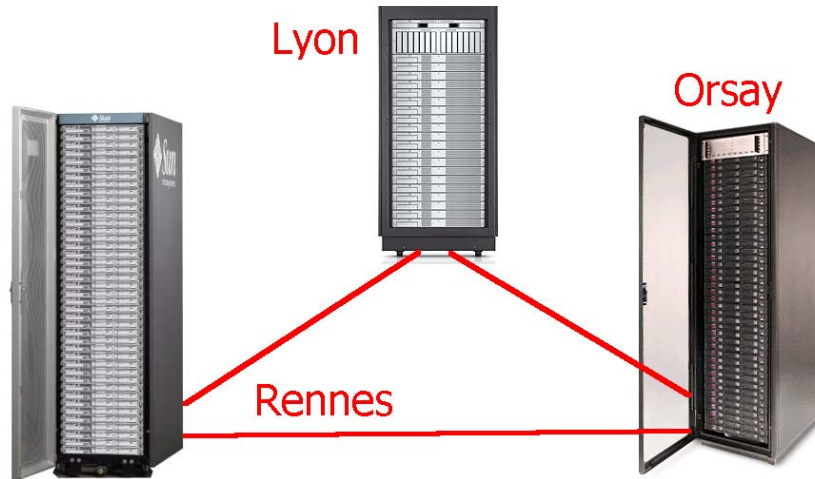
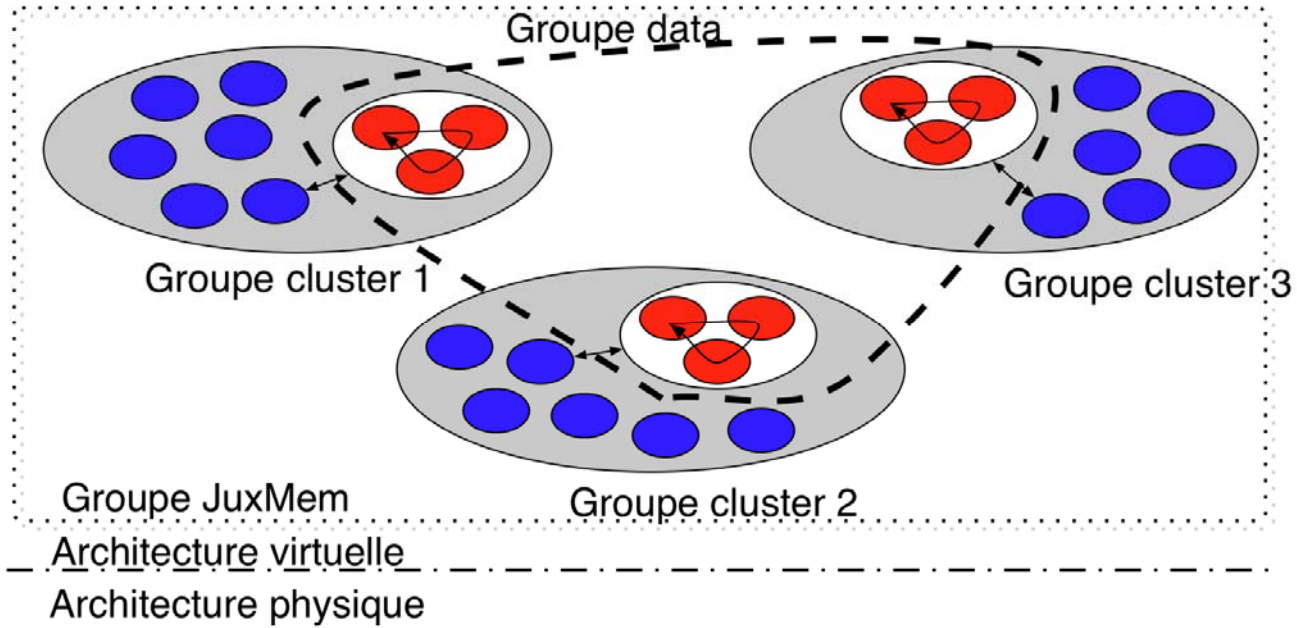
Gestion conjointe de la cohérence et des fautes : Architecture en couches



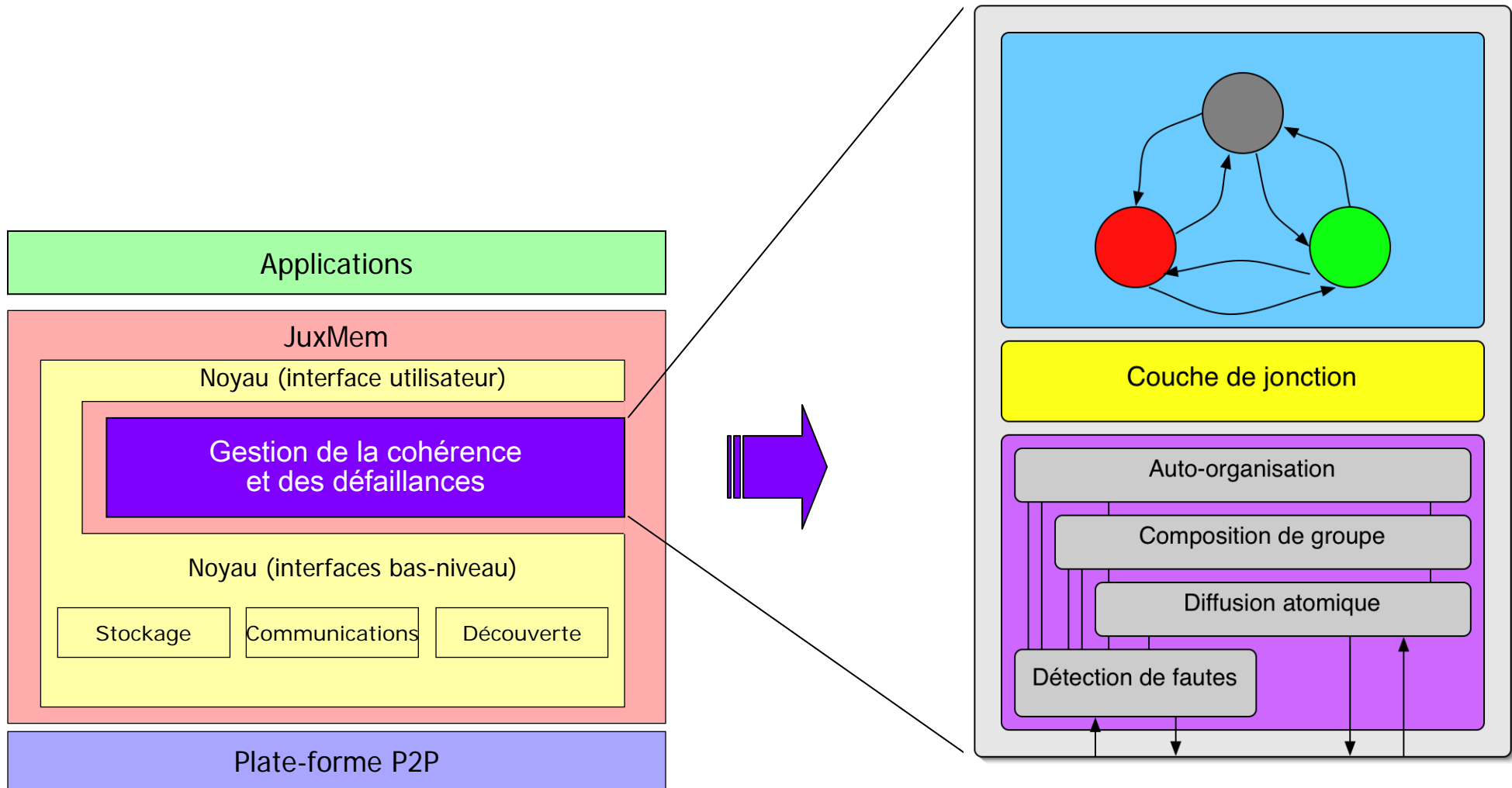
Gestion conjointe de la cohérence et des fautes : Architecture **multi-protocoles**



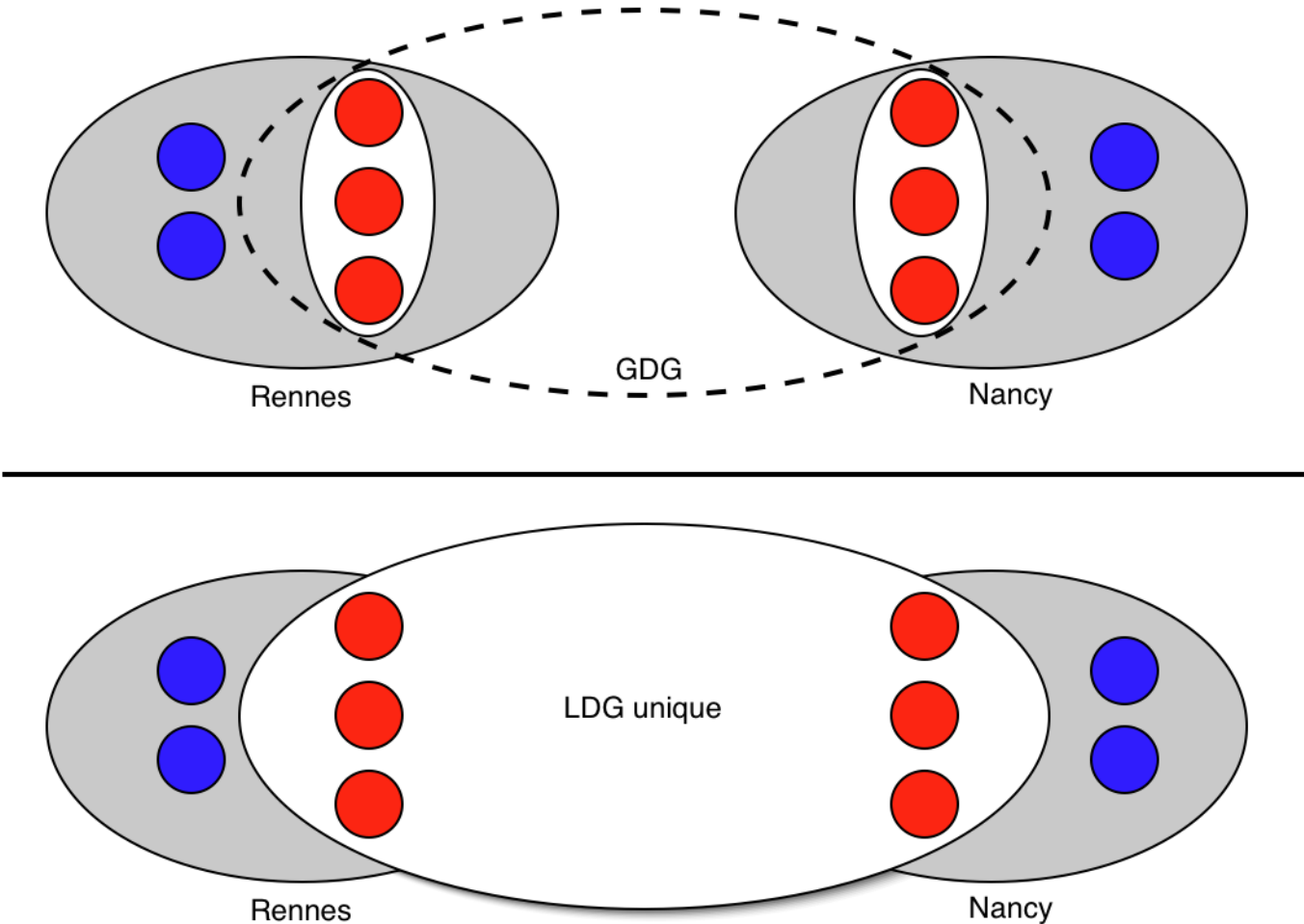
Mise en œuvre dans JuxMem



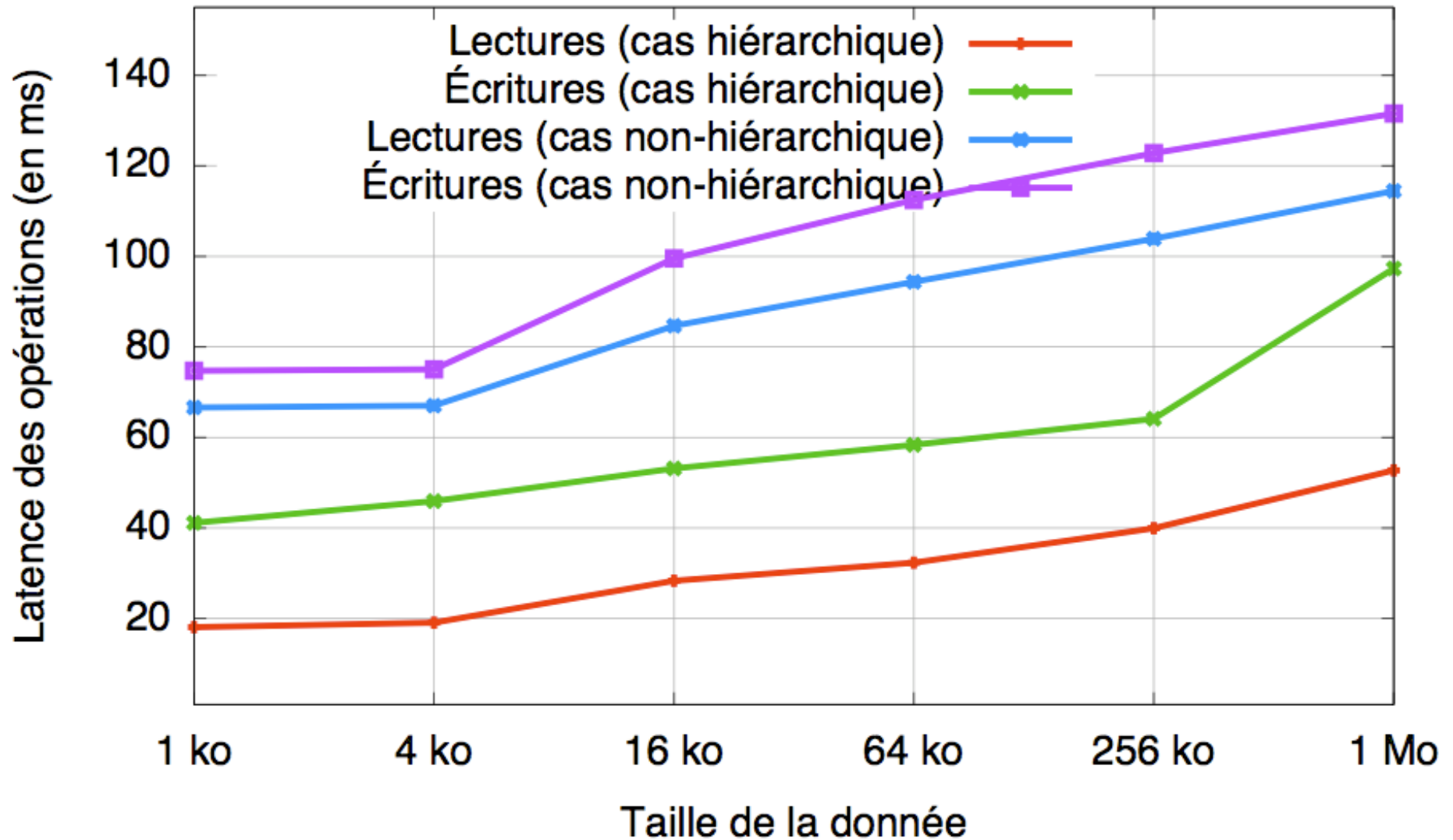
Mise en œuvre dans JuxMem



Gains apportés par l'approche hiérarchique



Gains apportés par l'approche hiérarchique



Publications

1. Gabriel Antoniu, Jean-François Deverge and Sébastien Monnet. How to bring together fault tolerance and data consistency to enable grid data sharing. In *Concurrency and Computation: Practice and Experience*, Vol. 18(13):1705--1723, November 2006.
2. Gabriel Antoniu, Loïc Cudennec and Sébastien Monnet. Extending the entry consistency model to enable efficient visualization for code-coupling grid applications. In *6th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID 2006)*, Pages 552-555, Singapore, May 2006.
3. Gabriel Antoniu, Jean-François Deverge and Sébastien Monnet. Building Fault-Tolerant Consistency Protocols for an Adaptive Grid Data-Sharing Service. In *Proc. ACM Workshop on Adaptive Grid Middleware (AGridM 2004)*, Antibes Juan-les-Pins, France, September 2004.
4. Gabriel Antoniu, Loïc Cudennec and Sébastien Monnet. A practical evaluation of a data consistency protocol for efficient visualization in grid applications. In *International Workshop on High-Performance Data Management in Grid Environment (HPDGrid 2006)*, Vol. 4395:692-706 of *Lecture Notes in Computer Science*, Held in conjunction with *VECPAR'06*, Springer Verlag, Rio de Janeiro, Brazil, July 2006.
5. Jean-François Deverge and Sébastien Monnet. Cohérence et volatilité dans un service de partage de données dans les grilles de calcul. In *Actes des Rencontres francophones du parallélisme (RenPar 16)*, Pages 47--55, Le Croisic, April 2005.
6. Ramsés Morales, Sébastien Monnet, Indranil Gupta and Gabriel Antoniu. MOve: Design and Evaluation of A Malleable Overlay for Group-Based Applications. In *IEEE Transactions on Network and Service Management*, Special Issue on Self-Management, Vol. 4(2):107-116 , 2007.
7. Sébastien Monnet, Ramsés Morales, Gabriel Antoniu and Indranil Gupta. MOve: Design of An Application-Malleable Overlay. In *Symposium on Reliable Distributed Systems 2006 (SRDS 2006)*, Pages 355-364, IEEE Computer Society, Leeds, UK, October 2006.

Bilan

Contributions

- Idée directrice : **gestion conjointe de la cohérence et de la tolérance aux fautes**
- Approche : « gridification », puis couplage d'algorithmes existants
 - Protocoles de cohérence
 - Protocoles de tolérance aux fautes
- Architecture multi-protocoles à deux niveaux

Qu'avons-nous appris ?

- Notion de **groupe auto-organisant hiérarchique, tolérant aux fautes**
 - Impact fort de la prise en compte de la hiérarchie sur les performances
- Une **méthode de conception** de protocoles de cohérence adaptés aux grilles
- Utilisation possible dans d'autres contextes



Etapes et actions

Déploiement
Thèse de Loïc Cudennec (2005 - 2009)

Cohérence et tolérance aux fautes
Thèse de Sébastien Monnet (2003 - 2006)

Transparence
Thèse de Mathieu Jan (2003 - 2006)



Etape 3 : simplifier le déploiement sur grille

Jalon : déploiement dynamique et co-déploiement de services

Thèse de Loïc Cudennec (2005-2009) - Sun Microsystems, INRIA, Région Bretagne

- Ingénieur-chercheur au CEA, LIST, LaSTRE (depuis mars 2009)

Contributions

- Un **modèle générique pour le déploiement dynamique et le co-déploiement**
 - Trois propriétés : transparence, versatilité et non-intrusivité
- Une architecture et un prototype, **CoRDAGe** : <http://cordage.gforge.inria.fr/>
- Validation sur Grid'5000, avec JuxMem et Gfarm

Support

- Projet LEGO de l'ANR, CIGC (2006-2009)
- Projet RESPIRE de l'ANR, ARA MDMSA (2006-2009)
- Projet bilatéral PHC Sakura avec AIST/Université de Tsukuba (Japon) et
Projet NEGSTJSPS/CNRS
 - Collaboration avec Osamu Tatebe (équipe Gfarm)

Etapes et actions

Expérimentations P2P sur grille

Déploiement
Thèse de Loïc Cudennec (2005 - 2009)

Cohérence et tolérance aux fautes
Thèse de Sébastien Monnet (2003 - 2006)

Transparence
Thèse de Mathieu Jan (2003 - 2006)

Collaborations industrielle :

Sun Microsystems



JuxMem mis en œuvre sur JXTA

- Travail préliminaire : évaluation et amélioration des performances des protocoles JXTA
- 4 papiers sur l'évaluation de la plate-forme JXTA protocols
 - Euro-Par 2004, GP2PC 2005, HPCC 2005, JavaPDC 2008

Contrat de collaboration de 3 ans avec Sun Microsystems, Santa Clara (2005-2008)

- **Focus : utilisation de JXTA sur Grid'5000**
- Support pour la thèse de Loïc Cudennec
- 3 visites longues de Mathieu Jan et Loïc Cudennec chez Sun (2-3 mois)



Evaluation et optimisation de JXTA

Services de communication de JXTA

- Service point-à-point : communications statiques
- Service de canal virtuel : communication dynamique

Améliorations

- Saturation des liens Giga-Ethernet
- Optimisation des latences

Service de canal virtuel

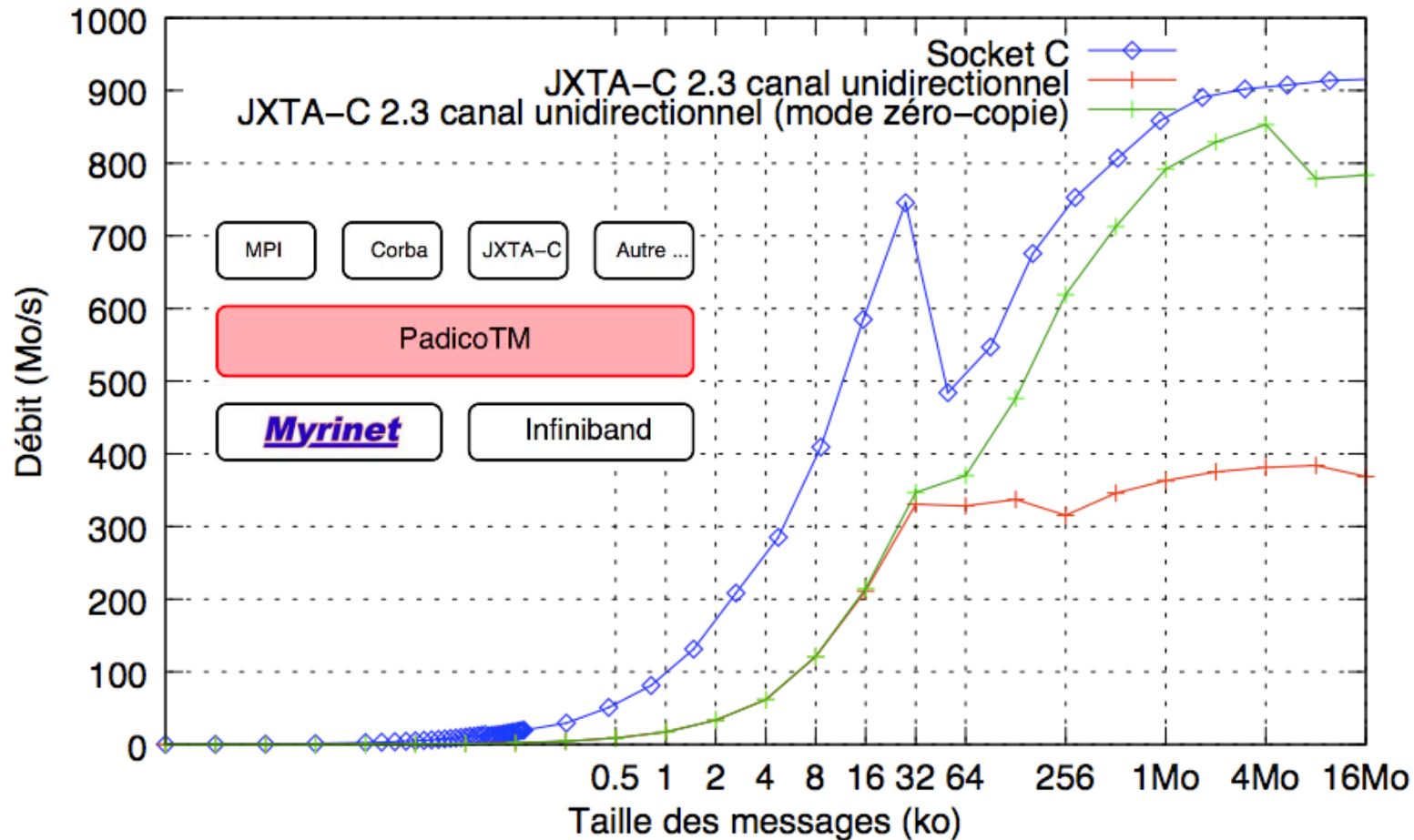
Service point-à-point

Protocole de transport

	Standard	Optimisé
Canal unidirectionnel	294 μs	90 μs
Service point-à-point	149 μs	84 μs
Socket	39 μs	

Evaluation de JXTA-C sur PadicoTM

Service de canal virtuel sur Myrinet (Myri-10G)



- PadicoTM [IRISA/LaBRI]: Plate-forme de communication haute-performance pour grilles
- Mise en œuvre : portage de JXTA-C 2.3

Publications

1. Gabriel Antoniu, Loïc Cudennec, Mike Duigou and Mathieu Jan. Performance scalability of the JXTA P2P framework. In *Proc. 21st IEEE International Parallel & Distributed Processing Symposium (IPDPS 2007)*, Pages 108 (abstract), Long Beach, CA, USA, March 2007.
2. Gabriel Antoniu, Mathieu Jan and David Noblet. A practical example of convergence of P2P and grid computing: an evaluation of JXTA's communication performance on grid networking infrastructures. In *Proc. 3rd Int. Workshop on Java for Parallel and Distributed Computing (JavaPDC'08)*, Pages 104 (abstract), Held in conjunction with IPDPS 2008, Miami, April 2008.
3. Gabriel Antoniu, Luc Bougé, Mathieu Jan and Sébastien Monnet. Large-scale Deployment in P2P Experiments Using the JXTA Distributed Framework. In *Euro-Par 2004: Parallel Processing*, Vol 3149:1038-1047 of Lecture Notes in Computer Science, Springer-Verlag, Pisa, Italy, August 2004.
4. Gabriel Antoniu, Mathieu Jan and David Noblet. Enabling the P2P JXTA Platform for High-Performance Networking Grid Infrastructures. In *Proc. of the first Intl. Conf. on High Performance Computing and Communications (HPCC '05)*, Vol. 3726:429-439 of Lecture Notes in Computer Science, Springer-Verlag, Sorrento, Italy, September 2005.
5. Gabriel Antoniu, Philip Hatcher, Mathieu Jan and David Noblet. Performance Evaluation of JXTA Communication Layers. In *Proc. Workshop on Global and Peer-to-Peer Computing (GP2PC 2005)*, Pages 251-258, Held in conjunction with the 5th IEEE/ACM Int. Symp. on Cluster Computing and the Grid (CCGRID~2005), IEEE TFCC, Cardiff, UK, May 2005. Best presentation award.



Bilan

Résultats

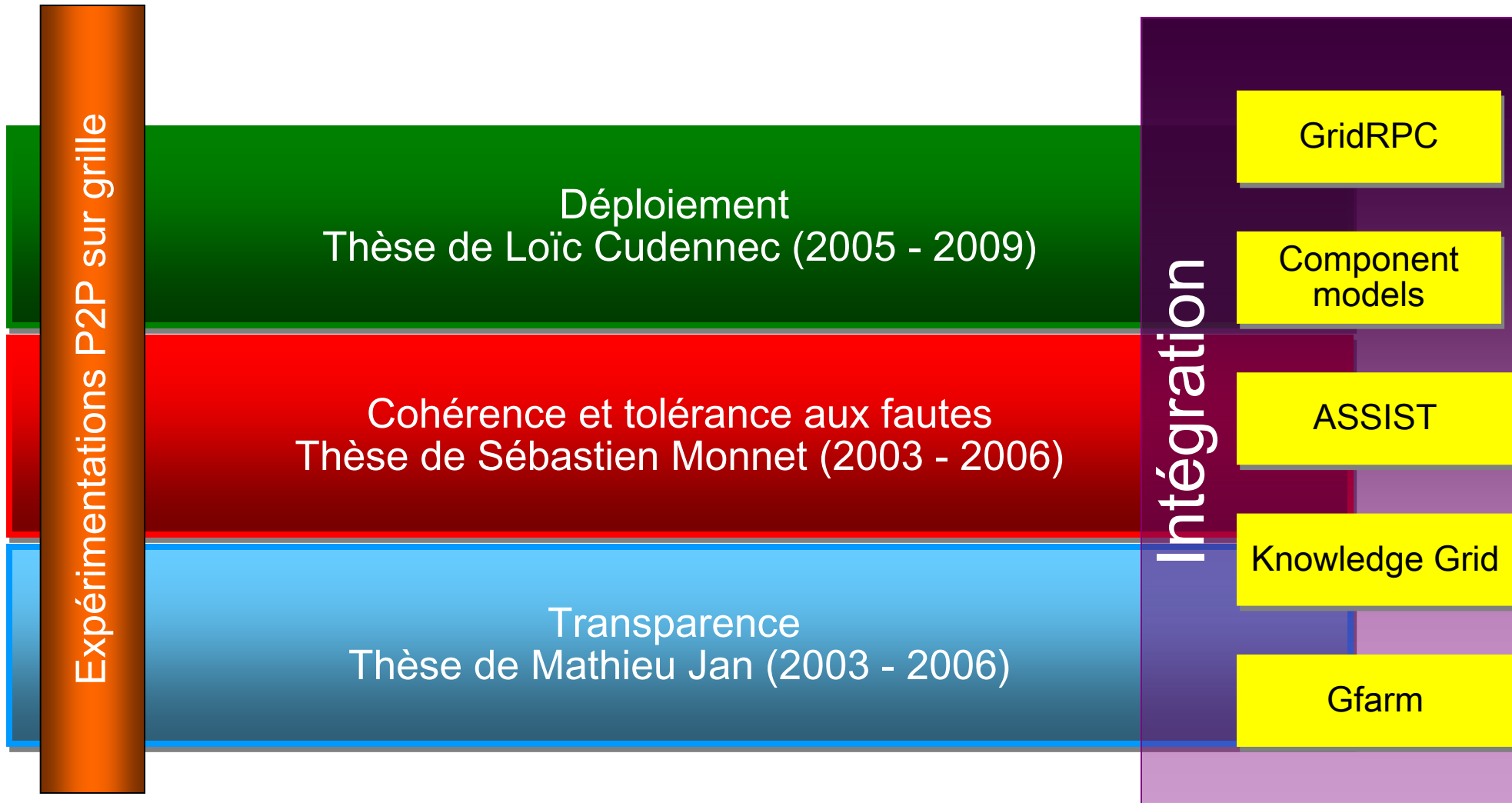
- Contributions intégrées dans JXTA
- **Papier co-publié avec Sun (IPDPS 2007)**
- Les plus grandes expériences connues avec JXTA
 - Déploiement réussi jusqu'à 29 000 pairs
 - Evaluation du protocole de rendez-vous de JXTA: jusqu'à 580 super-pairs sur 9 sites
 - Best Experiment Award (Mathieu Jan, Ecole Grid'5000, mars 2006)
- **Resultats utilisés par Sun pour signer un contrat avec Boeing**

Qu'avons-nous appris ?

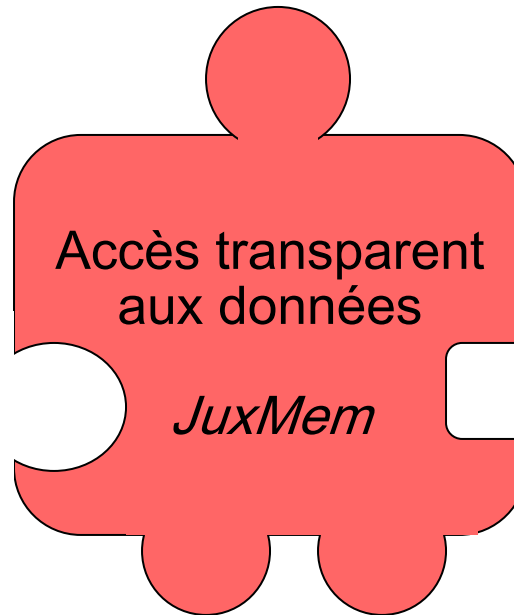
- JXTA « as is » n'est pas adapté aux grilles
- Optimisé, JXTA fournit des performances parfaitement acceptables
 - Saturation de liens haut-débit
 - Fort impact des paramètres de configuration
 - Tuning JVM → jusqu'à 140 %
 - Tuning tampons JXTA → le débit passe de 1 Mo/s à 98 Mo/s sur WAN



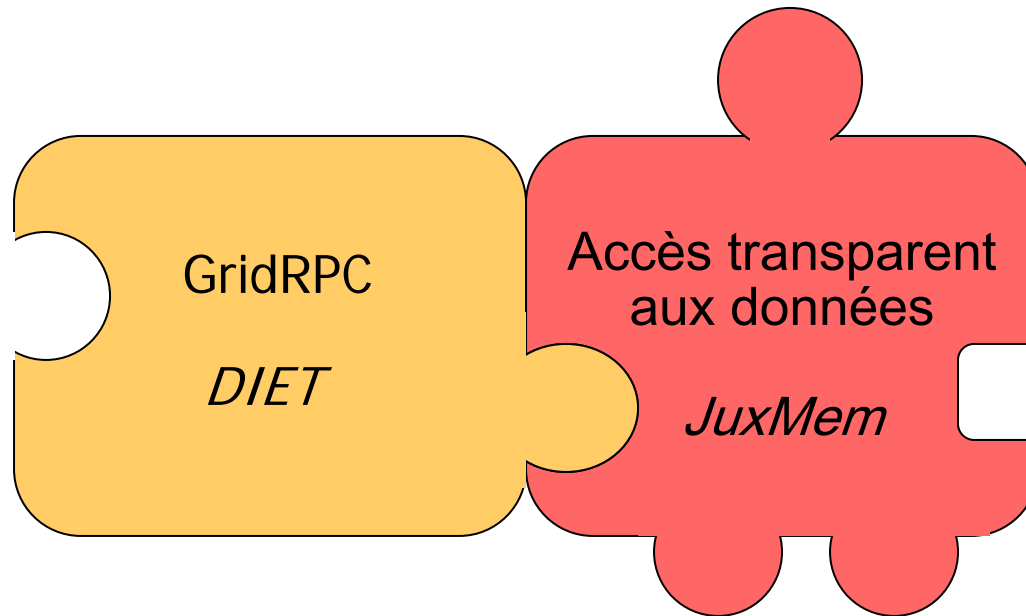
Etapes et actions



Partage transparent de données : GridRPC



Partage transparent de données : GridRPC

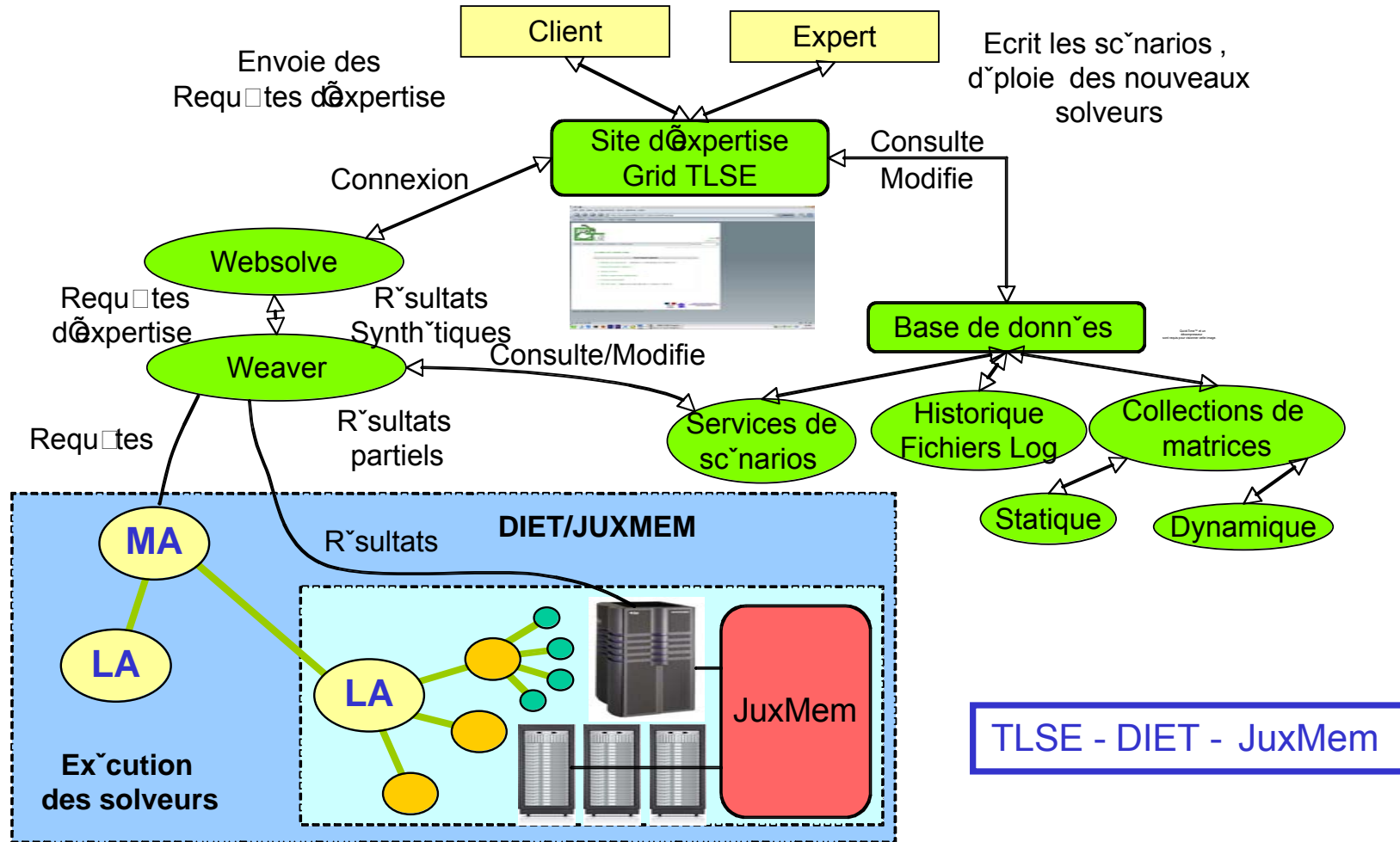


Intégration JuxMem + DIET (GRAAL, LIP, Lyon) : projet GDS de l'ACI MD (2003 - 2006)

Validation avec l'application TLSE (IRIT, Toulouse) : projet LEGO de l'ANR (2006 - 2009)



Partage transparent de données : GridRPC



Intégration JuxMem + DIET (GRAAL, LIP, Lyon) : projet GDS de l'ACI MD (2003 - 2006)

Validation avec l'application TLSE (IRIT, Toulouse) : projet LEGO de l'ANR (2006 - 2009)

Intégration du partage transparent dans le modèle GridRPC (OGF)



Scenario : MUMPS (*MUltifrontal Massively Parallel Solver*)

- $A x = b$
- Metrique: temps total d'exécution de 32 appels GridRPC à MUMPS

Configuration

- E1 : 1 grappe, 1 serveur DIET + 1 fournisseur JuxMem
- E2 : 3 grappes, 1 serveur DIET + 1 fournisseur JuxMem par grappe
- E3 : 3 grappes, 32 serveurs DIET + 8 fournisseurs JuxMem par grappe

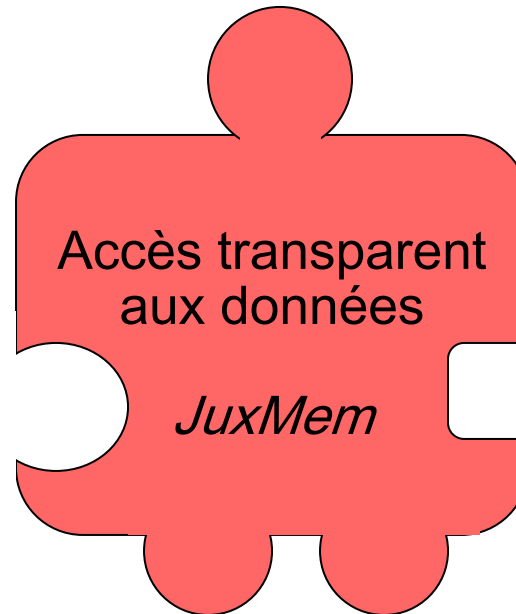


Matrice	A1 (22MB)		A2 (52 MB)	
	Sans JuxMem	Avec JuxMem	Sans JuxMem	Avec JuxMem
E1	36.6	41.3	957	961
E2	92.6	53.7	1420	880
E3	103	103	1358	843

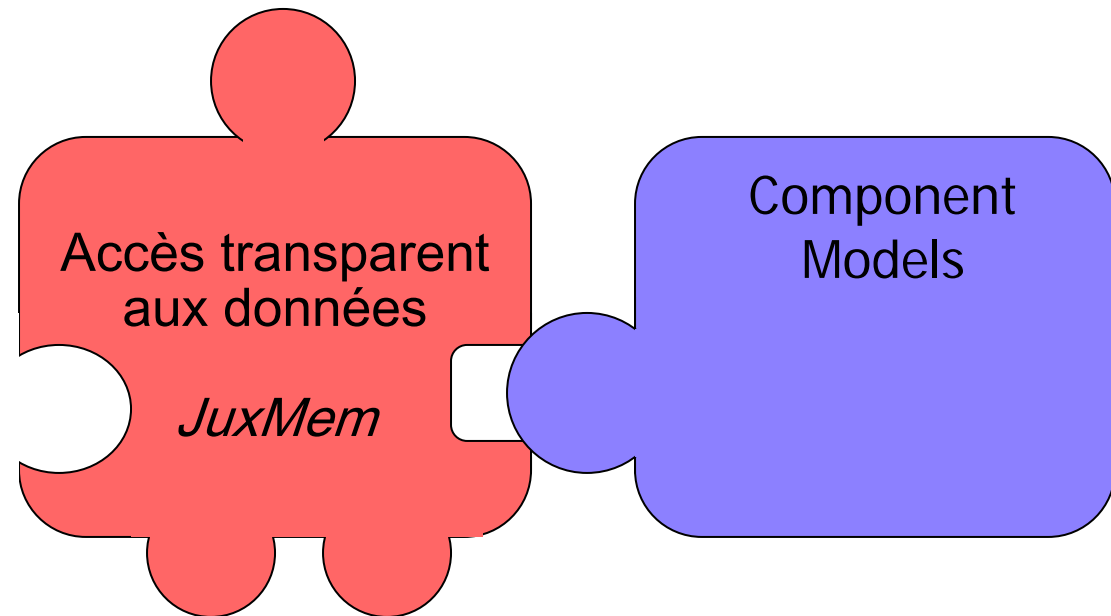
Résultats : gain jusqu'à **40%** en utilisant le stockage persistant de JuxMem (HiPC 2007)



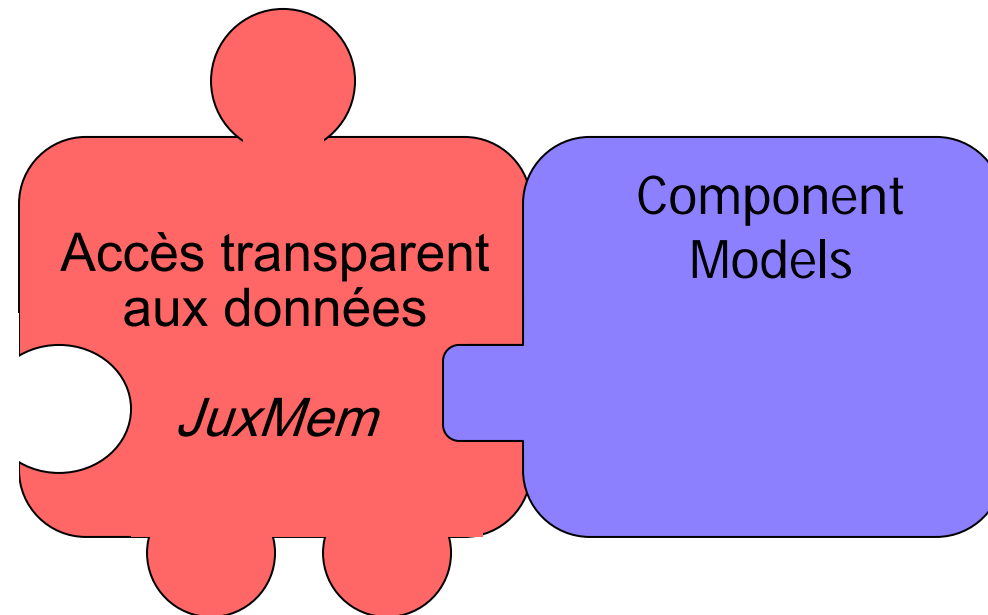
Intégration du partage transparent dans les modèles à composants



Intégration du partage transparent dans les modèles à composants



Intégration du partage transparent dans les modèles à composants



Modèle de composant étendu : port orienté données - projet LEGO de l'ANR (2006 - 2009)

Projection sur CCM et CCA



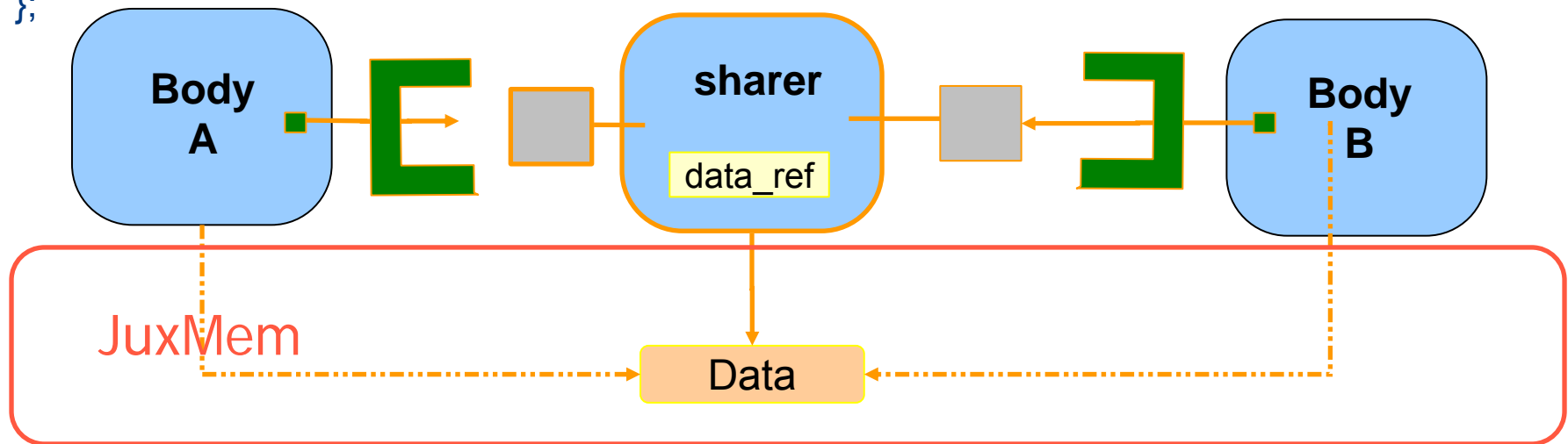
Intégration du partage transparent dans les modèles à composants

Exemple : N-body

```
typedef position float[N][3];
```

```
component sharer {  
    shares position to_bodies;  
};
```

```
component body {  
    accesses position from_sharer;  
};
```



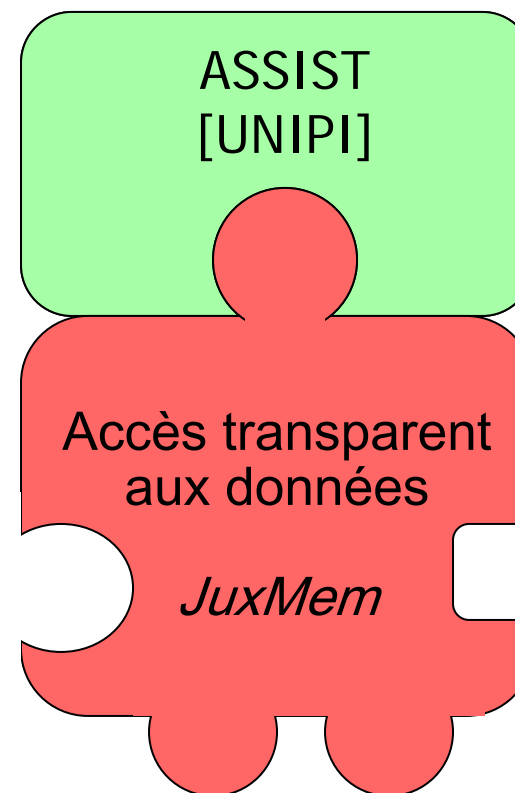
Modèle de composant étendu : port orienté données - projet LEGO de l'ANR (2006 - 2009)

Projection sur CCM et CCA

Intégration avec d'autres systèmes de stockage

Réseau d'Excellence CoreGRID (2004 - 2008)

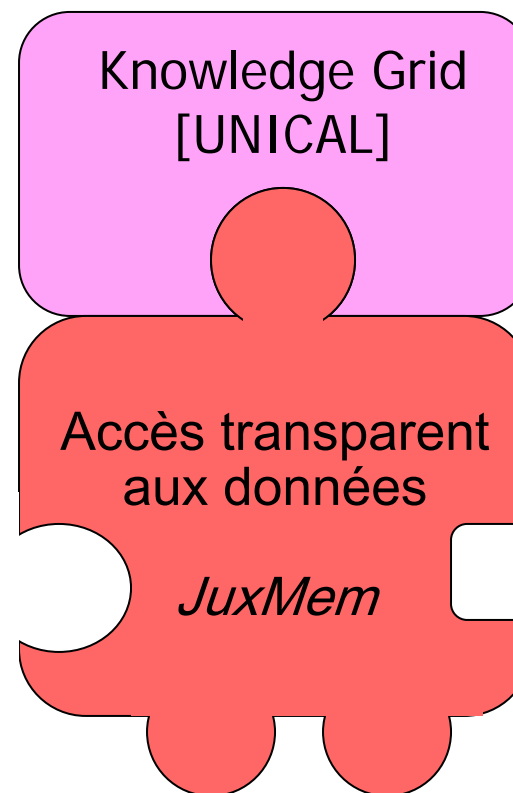
- JuxMem + ASSIST
 - Tolérance aux fautes, partage sur grille



Intégration avec d'autres systèmes de stockage

Réseau d'Excellence CoreGRID (2004 - 2008)

- JuxMem + ASSIST
 - Tolérance aux fautes, partage sur grille
- JuxMem + Knowledge Grid
 - Partage des méta-données



Intégration avec d'autres systèmes de stockage

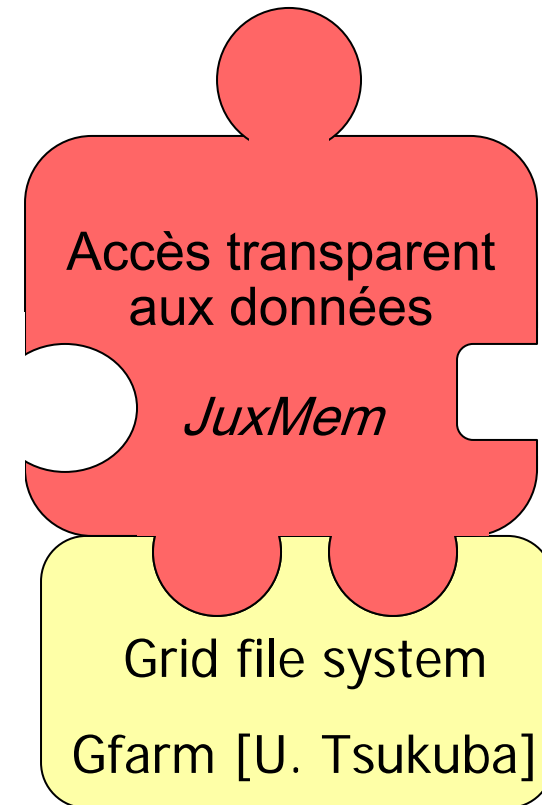
Réseau d'Excellence CoreGRID (2004 - 2008)

- JuxMem + ASSIST
 - Tolérance aux fautes, partage sur grille
- JuxMem + Knowledge Grid
 - Partage des méta-données

PHC Sakura avec l'AIST et Université de Tsukuba (Japon)

Projet NEGST CNRS-JSPS

- JuxMem + Gfarm
- Hiérarchie RAM globale / système de fichier global



Publications

1. Gabriel Antoniu, Hinde Bouziane, Mathieu Jan, Christian Pérez and Thierry Priol. Combining data sharing with the master-worker paradigm in the common component architecture. In *Cluster Computing*, Vol. 10(3):265 - 276, Kluwer Academic Publishers Hingham, MA, USA, 2007.
2. Gabriel Antoniu, Antonio Congiusta, Sébastien Monnet, Domenico Talia and Paolo Trunfio. Grid Middleware and Service Challenges and Solutions. Pages 219-233, Chapter Peer-to-Peer Metadata Management for Knowledge Discovery Applications in Grids, Springer Verlag, 2008.
3. Gabriel Antoniu, Eddy Caron, Frédéric Desprez, Aurélia Fèvre and Mathieu Jan. Towards a Transparent Data Access Model for the GridRPC Paradigm. In *Proc. of the 13th International Conference on High Performance Computing (HIPC 2007)*, Vol. 4873:269-284 of of Lecture Notes in Computer Science, Springer-Verlag, Goa, India, December 2007.
4. Gabriel Antoniu, Loïc Cudennec, Majd Ghareeb and Osamu Tatebe. Building Hierarchical Grid Storage Using the GFarm Global File System and the JuxMem Grid Data-Sharing Service. In Proceedings of the 14th International Euro-Par Conference on Parallel Processing (Euro-Par'08), Vol. 5168:456-465 of Lect. Notes in Comp. Science, Springer-Verlag, Las Palmas de Gran Canaria, Spain, 2008.
5. Gabriel Antoniu, Hinde Bouziane, Landry Breuil, Mathieu Jan and Christian Pérez. Enabling Transparent Data Sharing in Component Models. In *6th IEEE International Symposium on Cluster Computing and the Grid (CCGRID 2006)*, Pages 430-433, Singapore, May 2006.
6. Gabriel Antoniu, Hinde Bouziane, Mathieu Jan, Christian Pérez and Thierry Priol. Combining Data Sharing with the Master-Worker Paradigm in the Common Component Architecture. In *Proc. Joint Workshop on HPC Grid programming Environments and COmponents and Component and Framework Technology in High-Performance and Scientific Computing (HPC-GECO/CompFrame 2006)*, Pages 10-18, Paris, France, June 2006.
7. Gabriel Antoniu, Antonio Congiusta, Sébastien Monnet, Domenico Talia and Paolo Trunfio. Peer-to-Peer Metadata Management for Knowledge Discovery Applications in Grids. In CoreGRID Workshop on Grid Middleware, Held in conjunction with the International Supercomputing Conference (ISC 2007), Dresden, Germany, June 2007.
8. Marco Aldinucci, Marco Danelutto, Gabriel Antoniu and Mathieu Jan. Fault-tolerant data sharing for high-level grid programming: a hierarchical storage achitecture. In Proc. CoreGrid Integration Workshop, Pages 177-188, Krakow, Poland, October 2006.



Bilan

Contribution

- Introduction réussie du modèle d'accès transparent aux données dans des modèles de programmation pour grilles
- Intégration réussie de JuxMem avec d'autres systèmes de stockage

Qu'avons-nous appris ?

- Utilité démontrée du concept de service de partage transparent de données pour grilles
- Faisabilité démontrée
- Intégration dans des architectures hybrides potentiellement utile
 - Meilleures propriétés du système composé
 - Exemple : JuxMem + Gfarm



Bilan global

Contributions-clés

- Architecture de service de partage transparent de données pour grille : **GDS = DSM+P2P**
- Approche pour une gestion conjointe de la cohérence des données et de la tolérance aux fautes
 - **Groupe auto-organisant hiérarchique tolérant aux fautes**
- Intégration du **modèle d'accès transparent aux données** dans des modèles de programmation pour grille
 - GridRPC
 - Modèles à composants
- Mise en œuvre disponible : <http://juxmem.gforge.inria.fr>
 - JuxMem-C : 13 500 lignes de code
 - JuxMem-J2SE : 16700 lignes de code
- Validation expérimentale à grande échelle sur Grid'5000

Contributions annexes

- Evaluation et optimisation des protocoles P2P sur grille
- Un modèle et une architecture pour le déploiement dynamique et le co-déploiement sur grille



Bilan global

Méthodologie

- Travail à la frontière de plusieurs domaines : DSM, P2P, systèmes tolérants aux fautes
- Extension, adaptation et couplage de solutions partielles existantes
- Validation expérimentale à grande échelle (merci Grid'5000 !)
- Collaboration étroite avec de nombreuses équipes
 - Projets nationaux : GDS (coordinateur), GdX, DataGraal, LEGO, RESPIRE
 - Réseau européen d'excellence : CoreGRID
 - Collaborations bilatérales internationales : UIUC, AIST/U. Tsukuba (Japon), U. Calabre (Italie)
 - Collaboration industrielle : Sun Microsystems, Santa Clara, USA

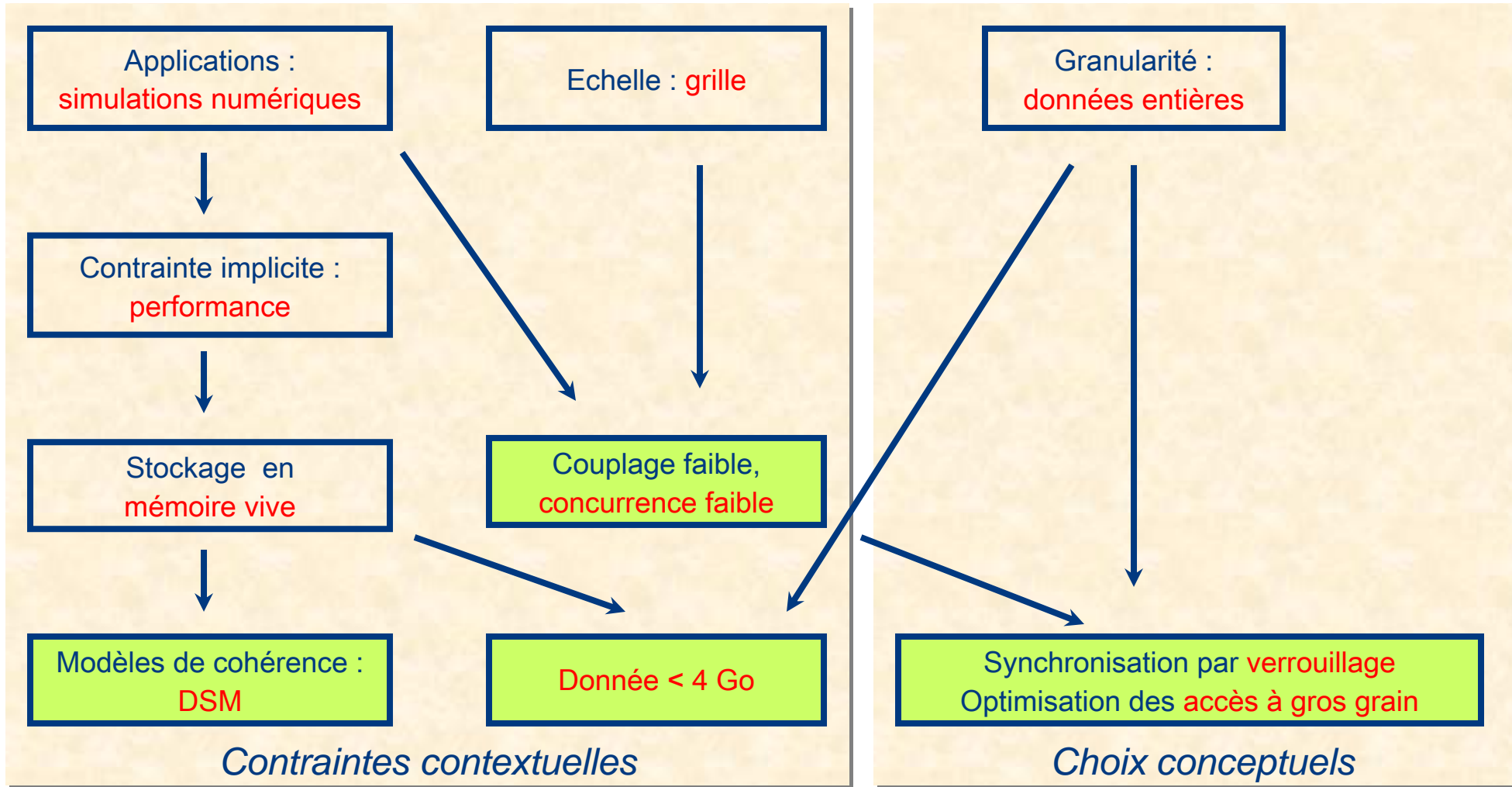
Qu'avons-nous appris ?

- Difficulté : intégration de résultats issus de micro-communautés différentes
- Mise en œuvre sur grille : difficile et laborieuse, nécessite un travail de haute technicité
- Utilité démontrée du concept de service de partage transparent de données pour grilles
- Faisabilité prouvée également



Perspectives

Partage transparent : périmètre du travail accompli



Perspectives

Modèles de cohérence :
DSM

Donnée < 4 Go

Couplage faible,
concurrency faible

Synchronisation par **verrouillage**
Optimisation des **accès à gros grain**

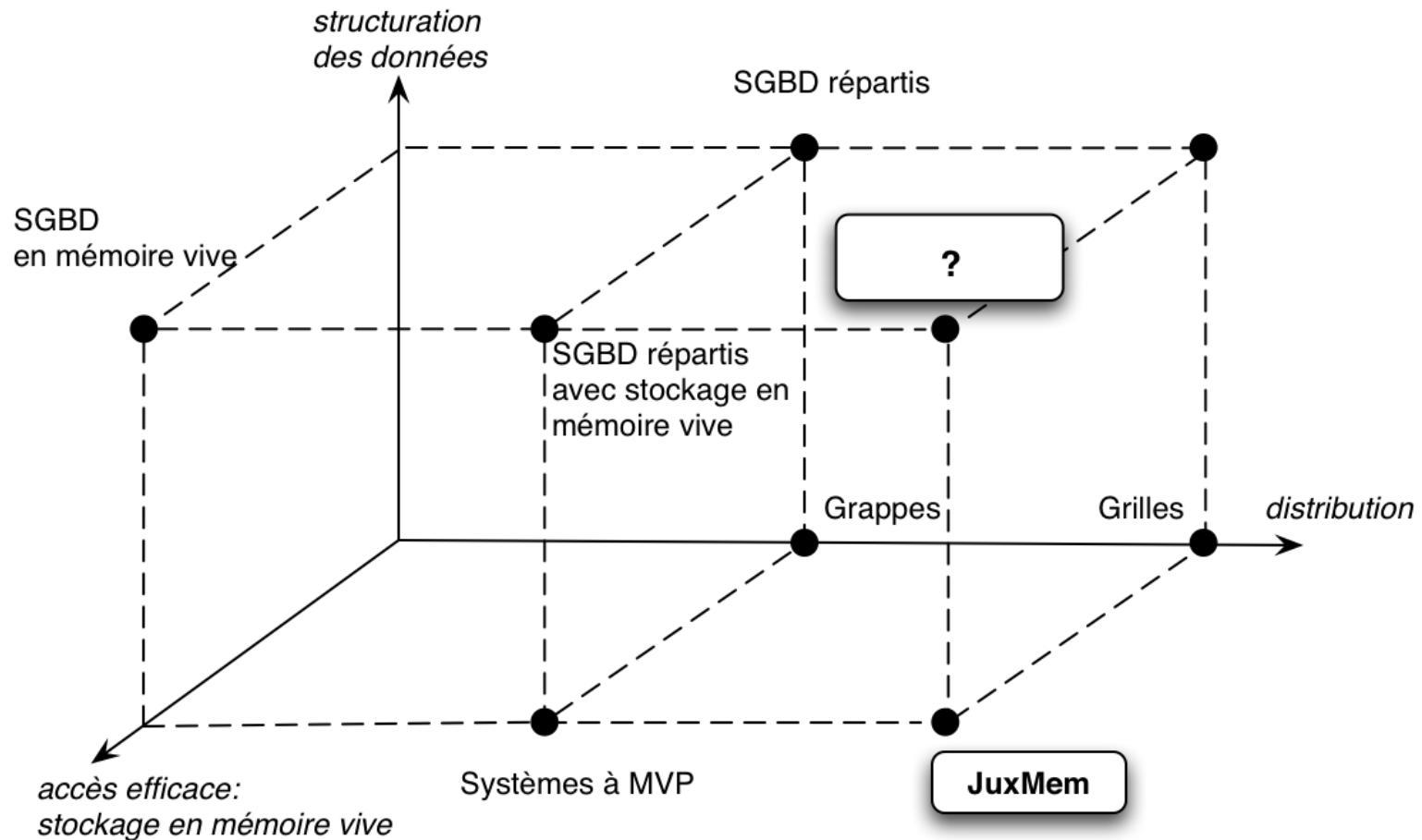
Partage transparent : prochains défis

- **Nouvelles applications**
 - Collaboration à très grande échelle
 - Analyse et traitement de données à grande échelle
 - Applications à base de services répartis
 - *Performance moins critique, cohérence faible*
- **Nouvelles infrastructures d'exécution**
 - Clouds, *desktop grid*
 - Infrastructures mixtes : grille + *desktop grid*
 - *Approche hiérarchique remise en cause ?*
- **Très grandes données** : ~ 1To
 - Fragmentation des données
 - *Gestion efficace des méta-données*
- **Accès hautement concurrents à grain fin** au sein de données géantes
 - *Synchronisation sans verrouillage*

Perspectives

Partage transparent : prochains défis

- Support efficace pour le stockage et le partage de **données structurées**



Remerciements !

Merci à tous ! En particulier :

- A Magda et Alex
- A Mathieu, Sébastien et Loïc (pour ce qui a été fait !)
- A Bogdan, Alexandra et Diana (pour ce qui sera fait !)
- A Luc, Thierry, Christian, aux autres membres de PARIS
- Aux membres du jury
- A l'INRIA et à l'ENS Cachan
- Aux services de l'IRISA

