Contribution à la définition et à la mise en œuvre d'un modèle de programmation à base de composants logiciels pour la programmation des grilles informatiques

Christian Pérez

Projet PARIS
IRISA/INRIA Rennes

Soutenance pour l'habilitation à diriger des recherches

Vendredi 10 novembre 2006

IRISA
institut de recherche en informatique et systèmes aléatoires

PARIS Project-Team

INRIA

Plan de l'exposé

- Un rapide curriculum vitae
- La problématique de recherche
  - Généralités
  - Contexte
  - Les 3 axes de recherche
- Deux exemples de travaux réalisés
  - Un modèle de composants parallèles distribués
  - Un modèle de communication multi paradigme
- Bilan & perspectives

Un rapide curriculum vitae

Points clés : parcours

- 1995
  - Diplôme d'Étude Approfondie d'Informatique de Lyon
    - École Normale Supérieure de Lyon
  - Magistère d'Informatique et de Modélisation
    - École Normale Supérieure de Lyon
- 1995 – 1999 : doctorat au LIP (ENS Lyon)
  - *Compilation des langages à parallélisme de données : gestion de l'équilibrage de charge par un exécutif à base de processus légers*
  - 11/1997 – 08/1998 : scientifique du contingent
    - Établissement Technique Central de l'Armement de la DGA, Arcueil
- 2000 – aujourd'hui : chargé de recherche INRIA
  - Projet PARIS (IRISA)

Points clés : recherche (1/2)

- Bilan succinct (depuis 2000)
  - 3 thèses soutenues, 2 thèses en cours
  - 1 étude post-doctorale
  - 3 ingénieurs experts
  - 4 stages de DEA
  - 2 participations à des jurys de thèse
    - dont 1 fois comme rapporteur
- Publications (depuis 1995)
  - 5 revues internationales, 1 revue nationale
  - 3 chapitres d'ouvrage
  - 32 conférences internationales avec comité de lecture
    - 2 IPDPS, 3 Euro-Par, HICSS, 4 Grid Computing

Points clés : recherche (2/2)

- 1 projet international
  - Projet STAR de l'ambassade de France en Corée, 2003-2005
- 2 projets européens
  - Réseau d'excellence européen CoreGRID, 2004-2008 (2)
  - IST FET Performance Portability of OpenMP, 2001-2005 (2)
- 11 projets nationaux
  - Projet ANR CIGC NUMASIS, 2006-2009 (3)
  - Projet ANR CIGC DISC, 2006-2009 (3)
  - Projet ANR CIGC LEGO, 2006-2009
  - Projets RNTL VTHD et VTHD++, 1999-2004
  - Projet ACI GRID logiciel Grid-RMI, 2001-2003 (1)
  - Projet ACI GRID pluridisciplinaire HydroGrid, 2002-2005 (3)
  - Projet d'animation ACI GRID GRID2, 2001-2004
  - Projet ARC/ACI GRID ALTA, 2004-2005 (1)
  - Projet ARC INRIA Couplage, 2000
  - Projet ARC INRIA RedGrid, 2003-2005 (3)
  - Projet ARC INRIA COA, 2005-2006

(1) Coordinateur du projet
(2) Coordinateur scientifique pour l'INRIA
(3) Coordinateur scientifique pour le projet PARIS

La problématique de recherche

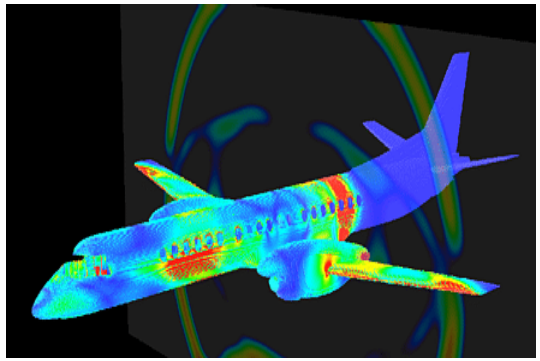
- Généralités
- Contexte
- Les 3 axes de recherche

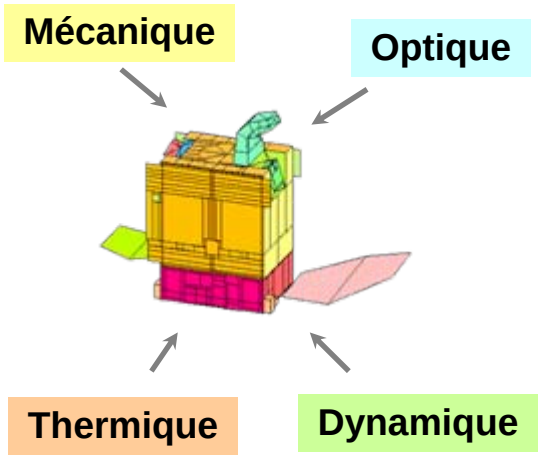
Problématique de recherche

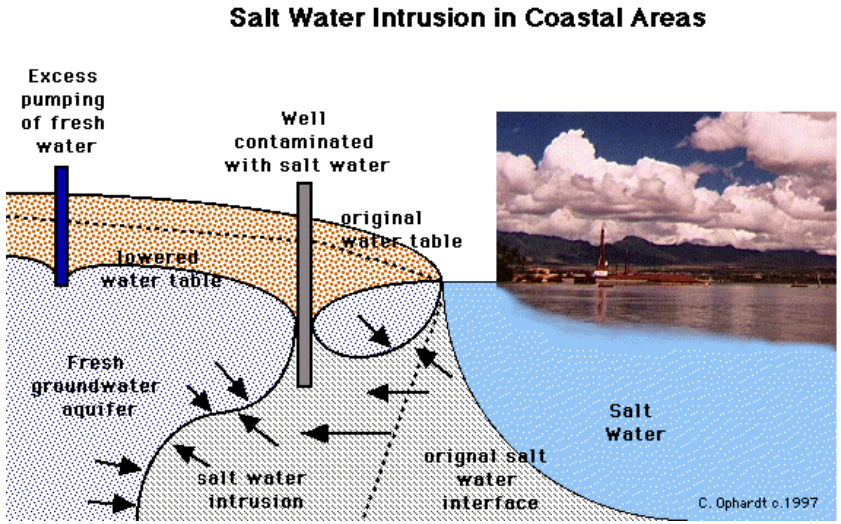
- Le parallélisme: obtenir des applications parallèles
  - Efficaces
  - Portables
  - Simplement
- Démarche
  - Considérer l'ensemble des intervenants
    - Modèle de programmation
    - Modèle d'exécution
    - Support exécutif
    - Système d'exploitation
    - Matériel
  - Confronter l'état de l'art à de plus en plus de types d'applications

Applications cibles

- Simulations numériques
  - Toujours plus précis
  - Toujours plus réaliste
  - Toujours plus complexe
- Besoin de
  - Puissance de calcul
  - Quantité de mémoire

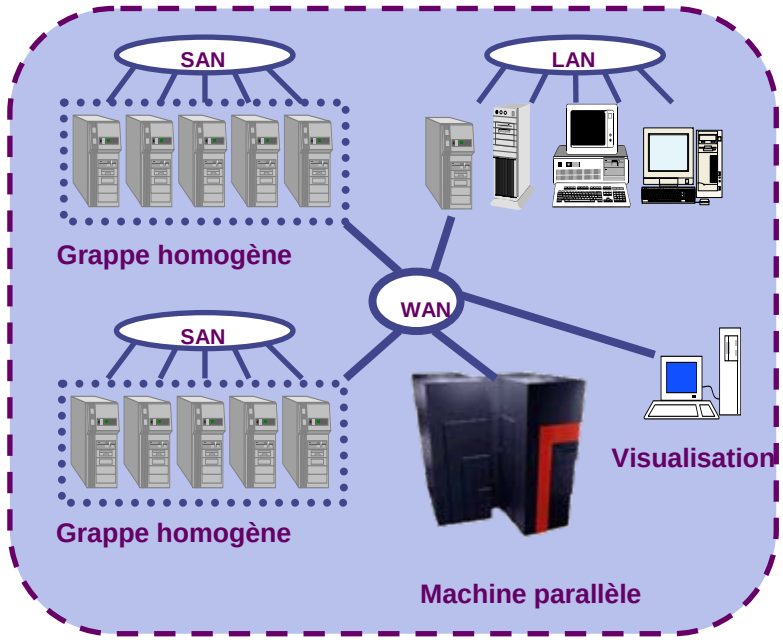






Architectures visées: les grilles informatiques

- Architecture nouvelle
  - Systèmes distribués de machines (potentiellement) parallèles
- Tout type de réseaux
  - WAN : VTHD, Géant, *etc*.
  - LAN : Ethernet
  - SAN : Myrinet, SCI, InfiniBand
- Tout type de ressources de calcul
  - Grappes & machines parallèles
    - Mono/multi-processeur/coeur
- Aspects des systèmes distribués et parallèles
  - Hétérogénéité
  - Performance



- Particularités
  - Force le découplage entre l'application et les ressources
    - Ressources affectées dynamiquement
  - Permet de concevoir de nouveaux types d'applications
    - Puissance de calcul/stockage sans précédent

Mes trois axes de recherches

- Objectif : définir un environnement de programmation pour les grilles informatiques
  - Efficace
  - Indépendant des ressources

- Trois axes de recherche
  - Modèles de programmation
    - Expressivité des modèles
  - Support à l'exécution
    - Liaison entre l'exécutif et le système d'exploitation
  - Modèles de déploiement
    - Liaison entre l'application et les ressources

Axe 1 : modèles de programmation

- Augmenter le niveau d'abstraction des modèles de composants logiciels
  - Modèle d'entités distribuées parallèles
    - Thèse d'André Ribes (2001-2004)
      - Recruté comme ingénieur-chercheur à EDF R&D
    - Support : ACIs GRID RMI & HYDROGRID, ARC RedGrid, RNTL VTHD & VTHD++, ANRs CIGC NUMASIS & DISC
  - Composition via des ports orientés données
    - Thèse d'Hinde Bouziane (2005-)
    - Support : NoE CoreGRID, ANR CIGC LEGO, ACI GRID GRID5000
  - Support du paradigme maître-travailleurs
    - Thèse d'Hinde Bouziane (2005-)
    - Support : NoE CoreGRID, ANR CIGC LEGO, ACI GRID GRID5000
  - Liaison entre composants logiciels et langage d'aspects
    - « Internship » INRIA de Gabriel Lopez (4 mois, 2005)
      - En collaboration avec Jean-Marc Menaud (Projet Obasco, EMN)

Axe 2 : support à l'exécution

- Supporter efficacement les modèles de programmation
  - Transparence d'accès au réseau
    - Thèse d'Alexandre Denis (2000-2003)
      - Chargé de recherche INRIA dans le projet RUNTIME (LaBRI)
    - Support : ACI GRID RMI & HYDROGRID, RNTL VTHD & VTHD++, ARC/ACI GRID ALTA, ANR CIGC LEGO

Axe 3 : modèles de déploiement

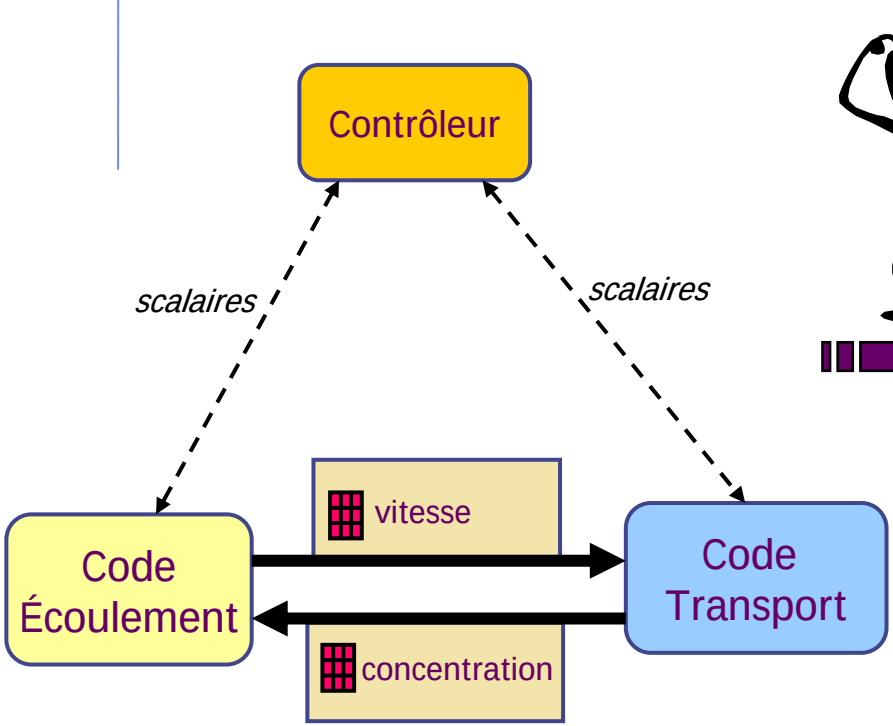
- Liaison entre l'application et les ressources
  - Déploiement automatique d'applications statiques
    - Thèse : Sébastien Lacour (2002-2005)
      - Ingénieur R&D chez Fujitsu Systems Europe (Toulouse)
  - Déploiement automatique d'applications dynamiques
    - Thèse en cours : Boris Daix (2006 -)
      - Financement CIFRE EDF
    - Support : ANR CIGC LEGO & DISC
  - Utilisation du langage gamma pour l'exécution de workflows
    - Post-doctorat de Zolt Nemeth (12 mois, 2004-2005)

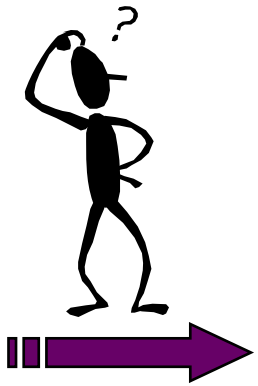
Modèle de programmation

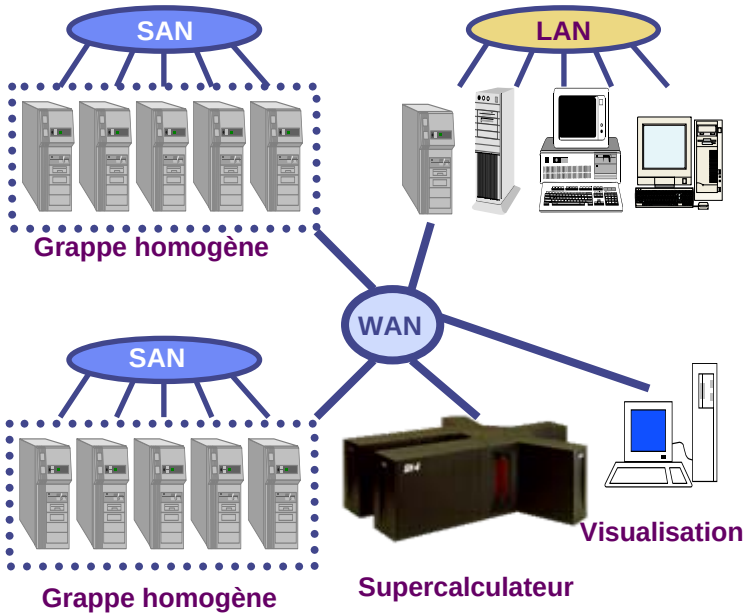
Modèle d’entités distribuées parallèles

Thèse d’André Ribes (2001-2004)

Simulations numériques et grilles informatiques



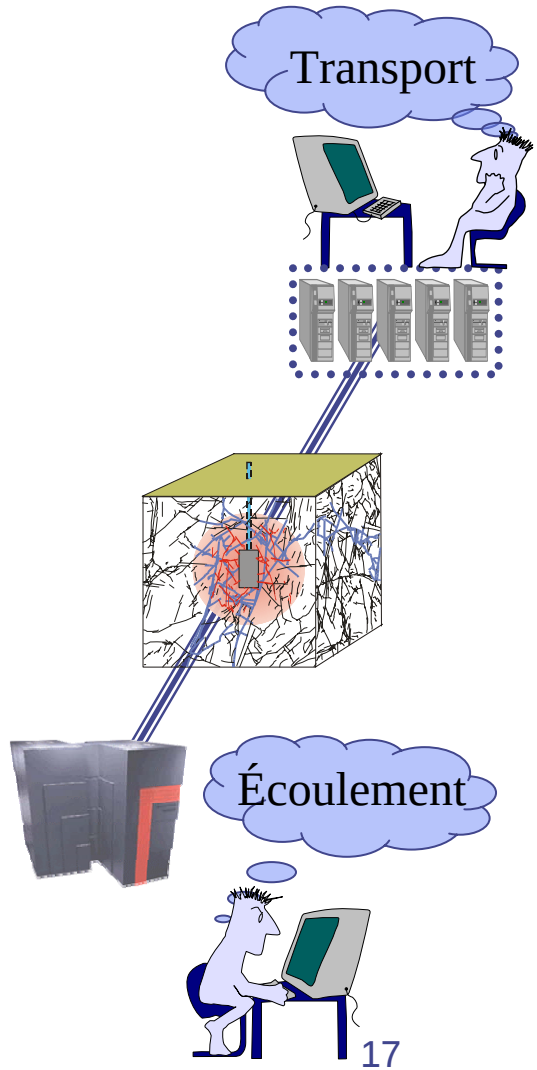




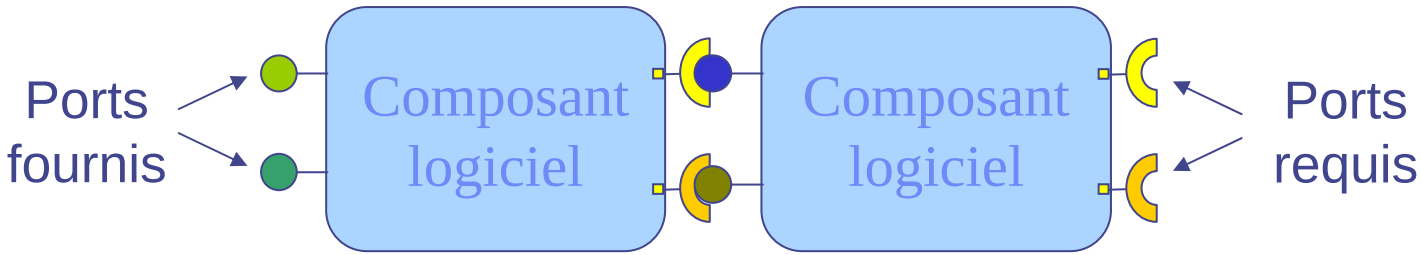
Caractéristiques

- Codes patrimoniaux
  - Multi-langages, multi-OS, multi-processeurs, ...
- Codes parallèles
  - MPI, PVM, OpenMP, ...
- Haute performance
- Déploiement
  - Mono/multi-utilisateurs

➢ ***Composant logiciel !***

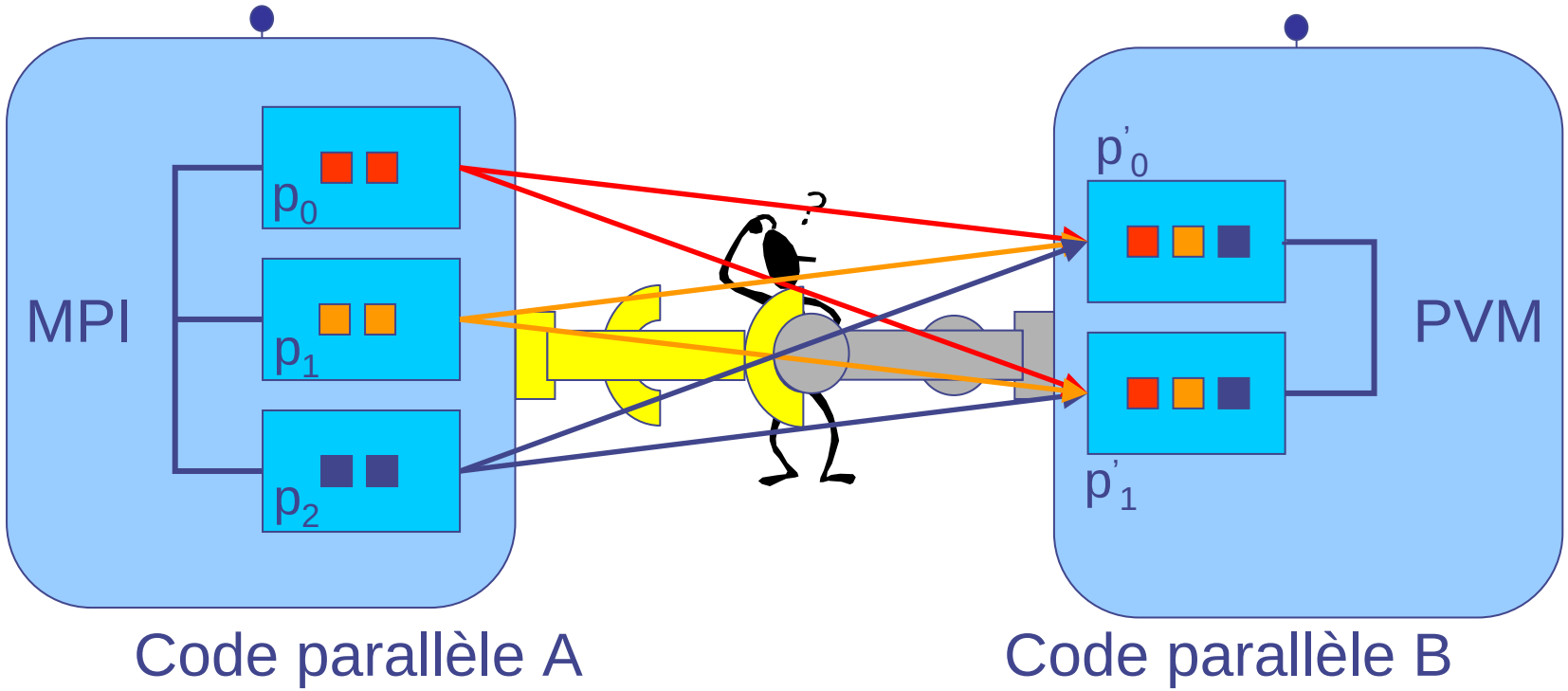


Composants logiciels



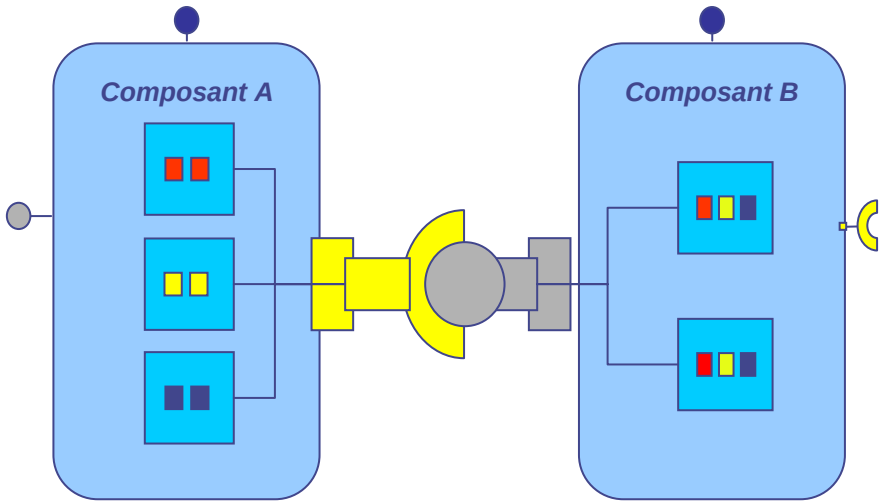
- Modèle boîte noire
- Composition via des ports
  - Appel de méthodes, passage de données, ...
- Modèle d'empaquetage
  - Séparation de la définition de l'implémentation
    - Paquet multi-binaires
- Modèle de déploiement
  - Langage de description d'architecture

Applications de couplage de code



Objectifs

- Proposer une solution qui permette d'agréger les débits
- Ne pas limiter les redistributions de données
- Extension d'un modèle existant
  - Garder la compatibilité avec les composants séquentiels
  - Minimiser l'impact sur les modèles existants
    - Idéalement, sans modification
- Permettre l'obtention de hautes performances



Un modèle d'entité parallèle distribuée

- Motivations
  - Plusieurs types d'entités communicantes
    - Objet, composant, service, ...
  - Plusieurs technologies
    - CORBA, JAVA RMI, EJB, .NET, WS, ...

- Quatre sous-modèles
  - Définition d'une entité parallèle distribuée
  - Connexion et communication
  - Gestion des données distribuées (lors d'une communication)
  - Gestion des exceptions parallèles (lors d'une communication)

Projection du modèle

- Objectif : modèle de programmation concret
  - Objet, composant, service, pair, ...
- Projection directe
  - une entité = un composant
- Projections réalisées:
  - PaCO++ : modèle d'objet CORBA parallèle
  - GridCCM : modèle de composant CORBA parallèle
- Autres projections possibles:
  - Fractal (~GCM), Web Services, JXTA, ...

Example de composants GridCCM

IDL2

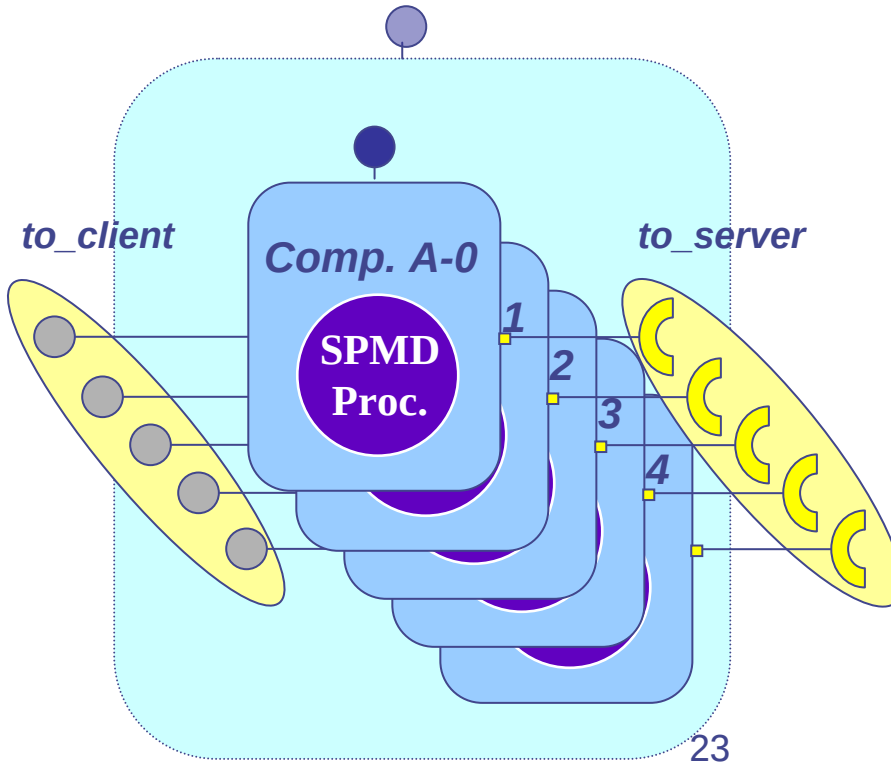
```
interface IExample
 {
   void factorise(in Matrix mat);
 };
```

IDL3

```
component CoPa
 {
   provides IExample   to_client;
   uses     Itfaces2   to_server;
 };
```

XML

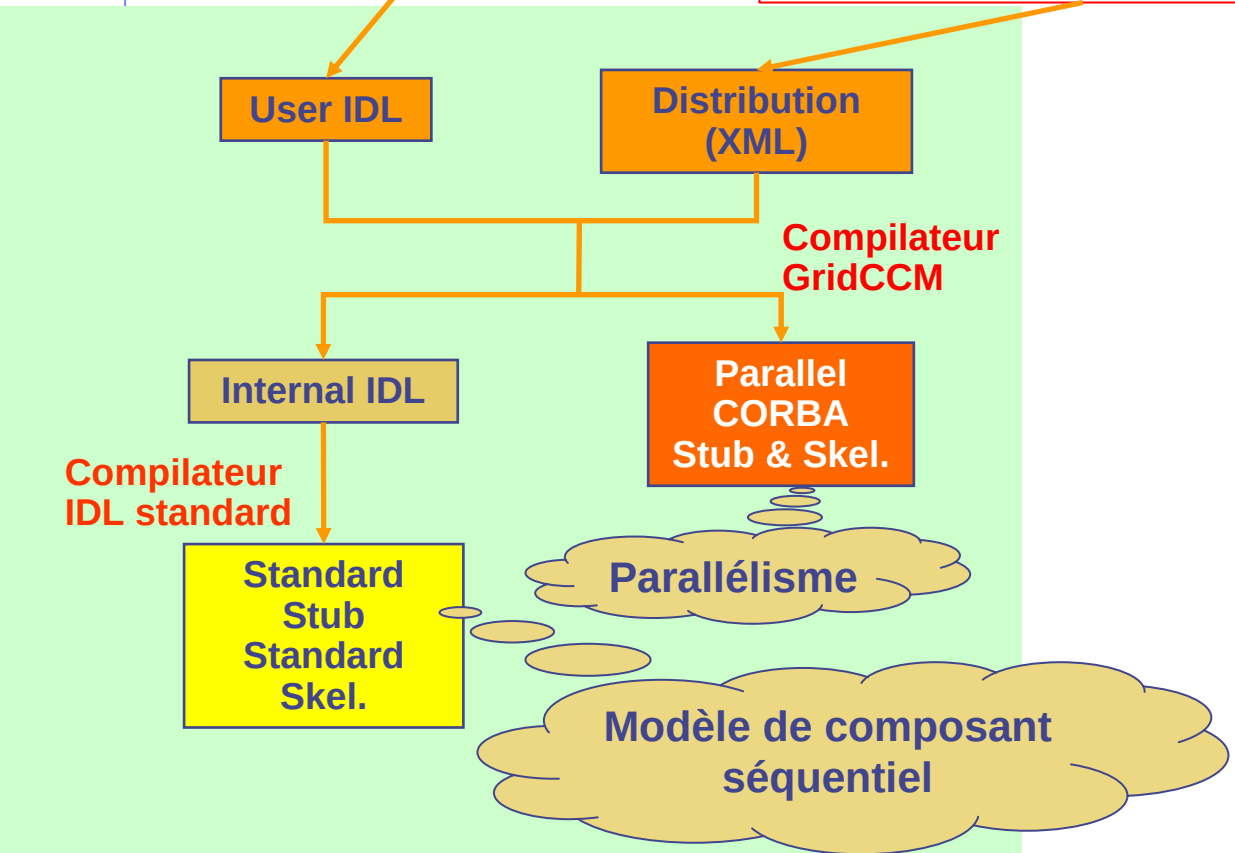
```
Component: CoPa
Port: to_client
Name: IExample.factorise
Type: Parallel
Argument1: Basic_BC[*, bloc]
ReturnArgument: noReduction
```



GridCCM: Génération de code

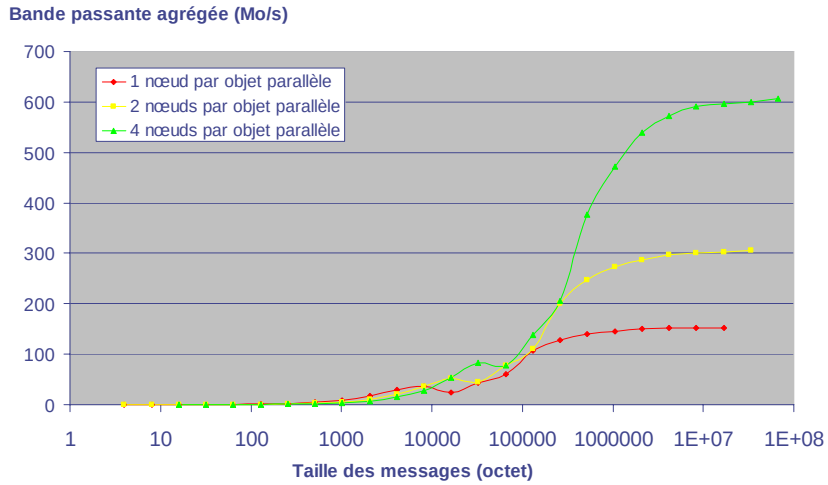
```
Interface Iexample { ... };
component CoPa {
 provides IExample to_client;
 uses     Itfaces2 to_server;
};
```

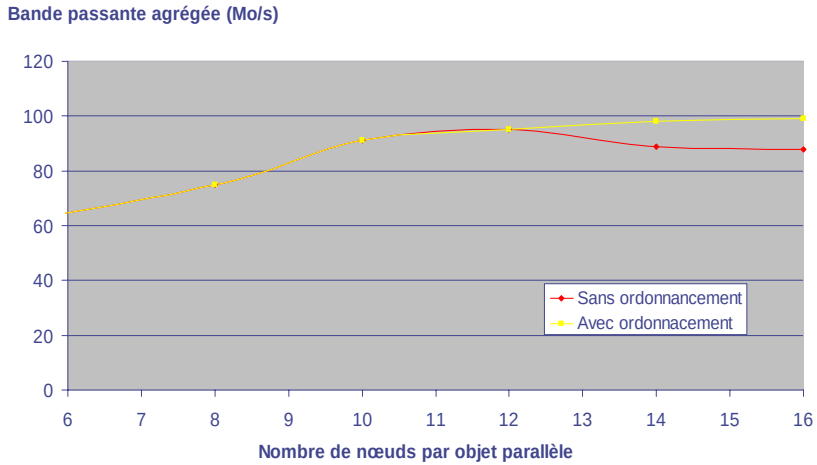
```
Component: CoPa
Port: to_client
Name: IExample.factorise
Type: Parallel
Argument1: BasicBC[bloc]
ReturnArgument: noReduction
```



Quelques résultats

- Débit sur SAN
  - Myrinet 2000/GM
  - Agrégation
    - 4 noeuds : 600 MB/s
      - 4x150 MB/s
- Débit sur WAN
  - Réseau VTHD
    - Commutateur à 1 Gbit/s
    - Saturation pour n > 12
  - Débit soutenu
    - Ordonnanceur
      - Algorithme du projet Algorille (LORIA)





Bilan

- Un modèle d'entité parallèle distribuée
  - Applicables à plusieurs types d'entités réparties
    - Et en particulier aux composants logiciels
      - Intégration dans le Grid Component Model du NoE CoreGRID
  - Séparation des différents aspects au travers de quatre sous-modèles
    - ARC RedGrid a permis d'intégrer
      - une bibliothèque de redistribution développée dans le projet Scalaplix (LaBRI)
      - une bibliothèque d'ordonnancement des communications développée dans le projet Algorille (Loria)
  - Transfert technologique chez EDF R&D
    - Recrutement d'André Ribes
    - Évaluation en cours de l'utilisation de PaCO++ dans le modèle de composants Salome

Perspectives

- Un modèle d'entité parallèle distribuée
  - Validation avec d'autres modèles distribués
    - de composants logiciels (GCM, CCA)
    - de systèmes pair-à-pair
  - Validation avec d'autres applications
    - Applications de propagation d'onde électromagnétisme & d'écoulement de fluide (ANR CIGC Discogrid)
    - Applications de sismologie (ANR CIGC Numasis)
  - Négociation de la méthode de redistribution
  - Ordonnancement des communications

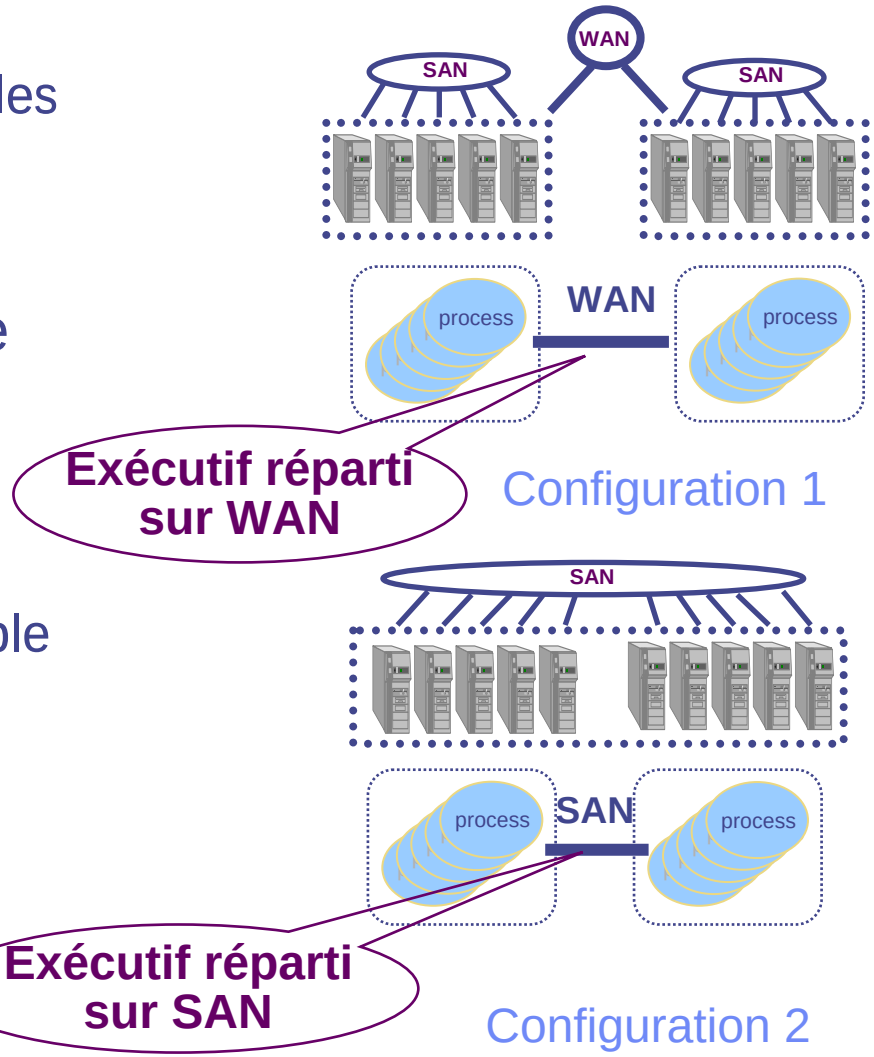
Support à l'exécution

Transparence d'accès au réseau

Thèse d'Alexandre Denis (2000-2003)

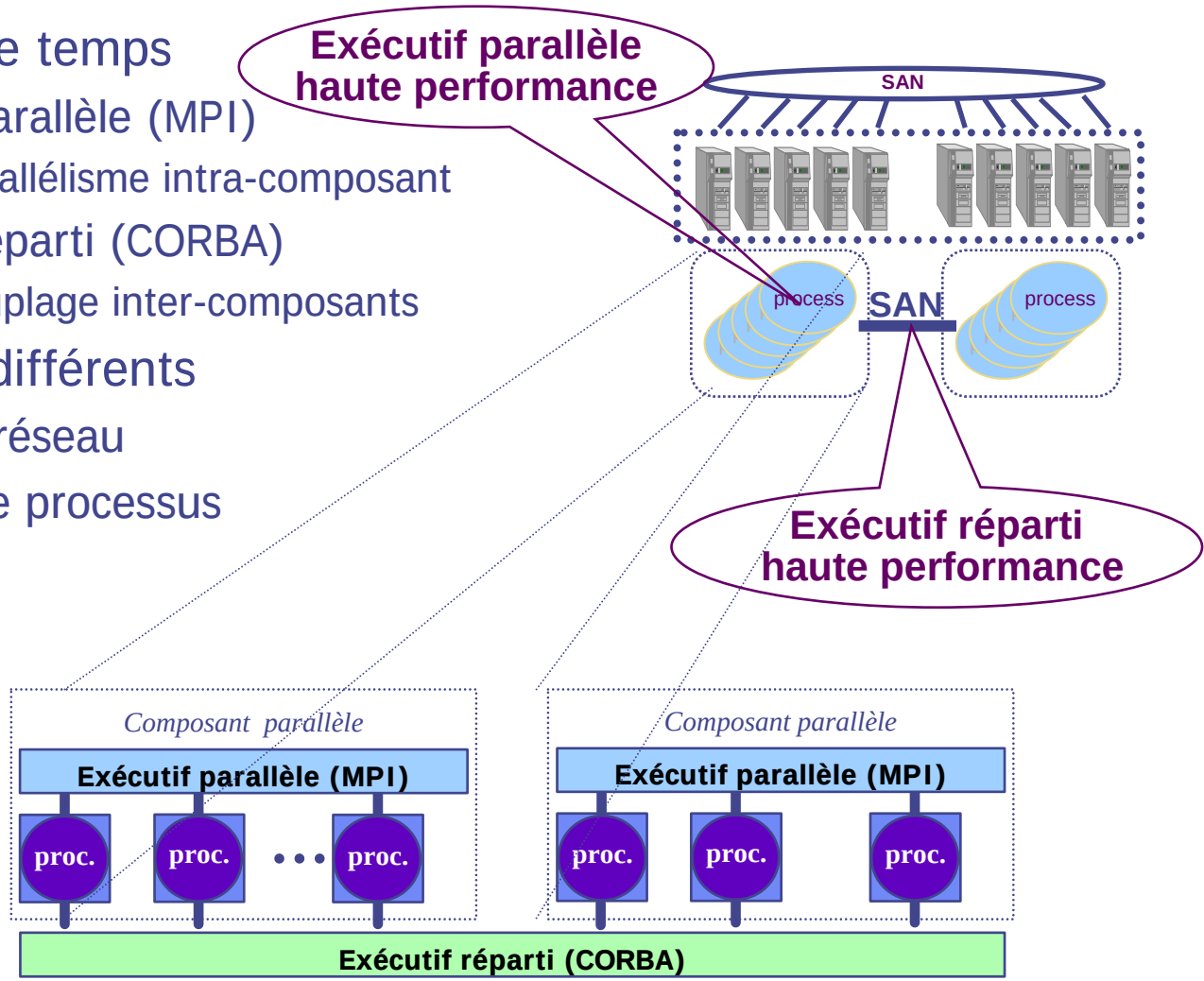
Transparence d'accès aux ressources

- Déploiement
  - Ressources allouées variables
  - Déployer le même code
- Exploitation transparente de différents réseaux
  - Sans modifier l'application
  - Sans modifier l'exécutif
  - Le plus efficacement possible
- La plate-forme s'adapte au réseau disponible



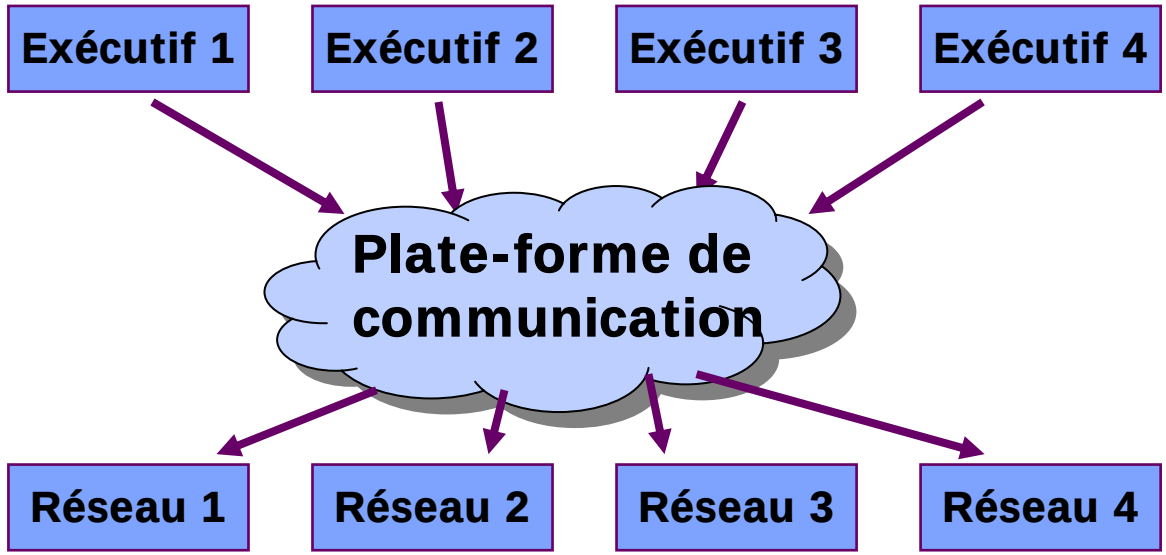
Concurrence d'accès aux ressources

- Utiliser en même temps
  - Un exécutif parallèle (MPI)
    - Pour le parallélisme intra-composant
  - Un exécutif réparti (CORBA)
    - Pour le couplage inter-composants
- Deux exécutifs différents
  - Sur le même réseau
  - Dans le même processus

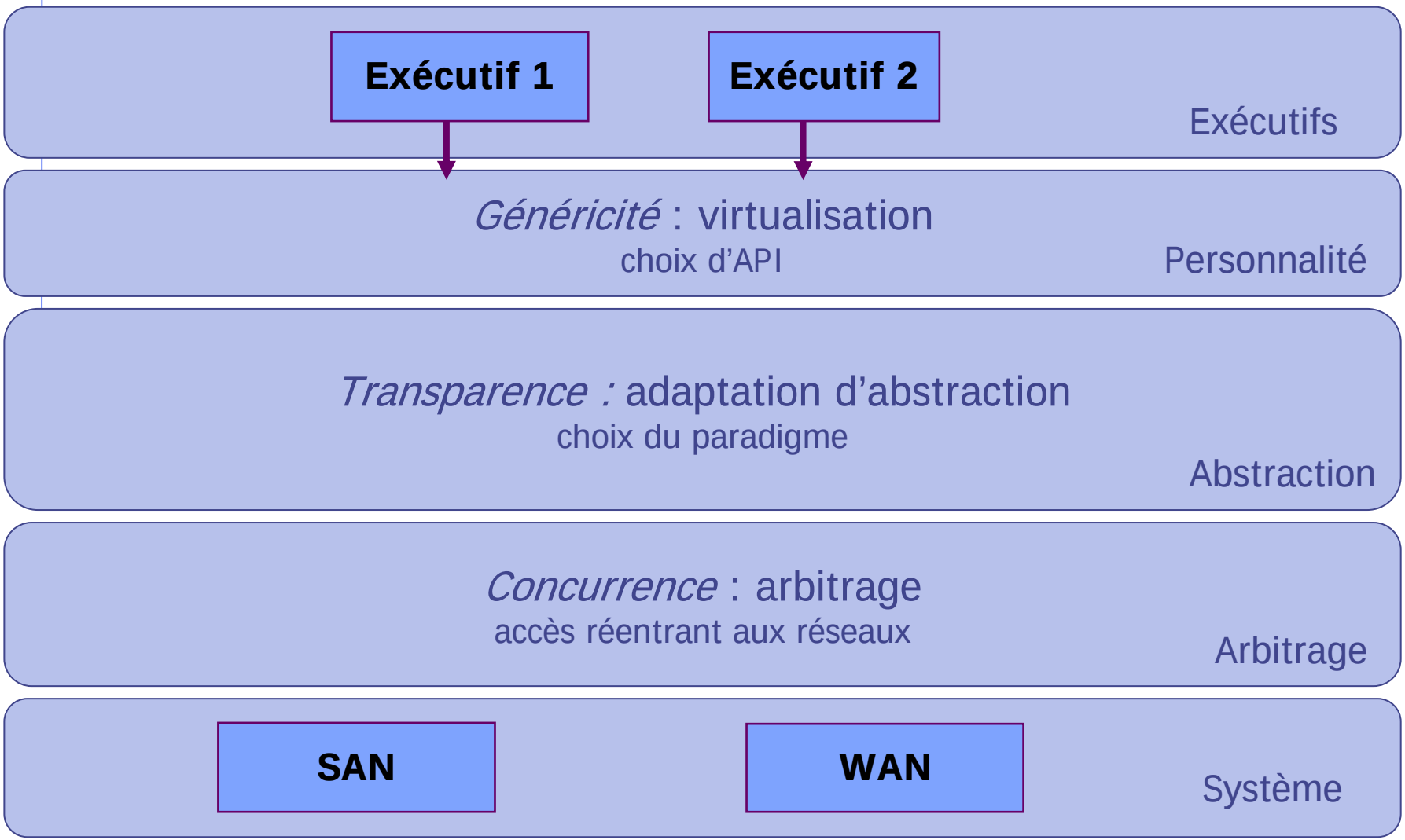


Le problème « N sur M »

- Comment gérer les communications ?
  - Large éventail d'exécutifs – répartis *et* parallèles
  - Large éventail de réseaux
  - ➢ Factoriser le code commun

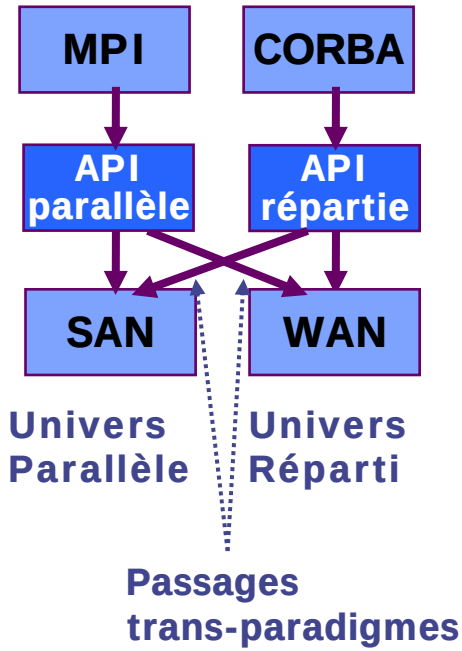


Modèle de communication

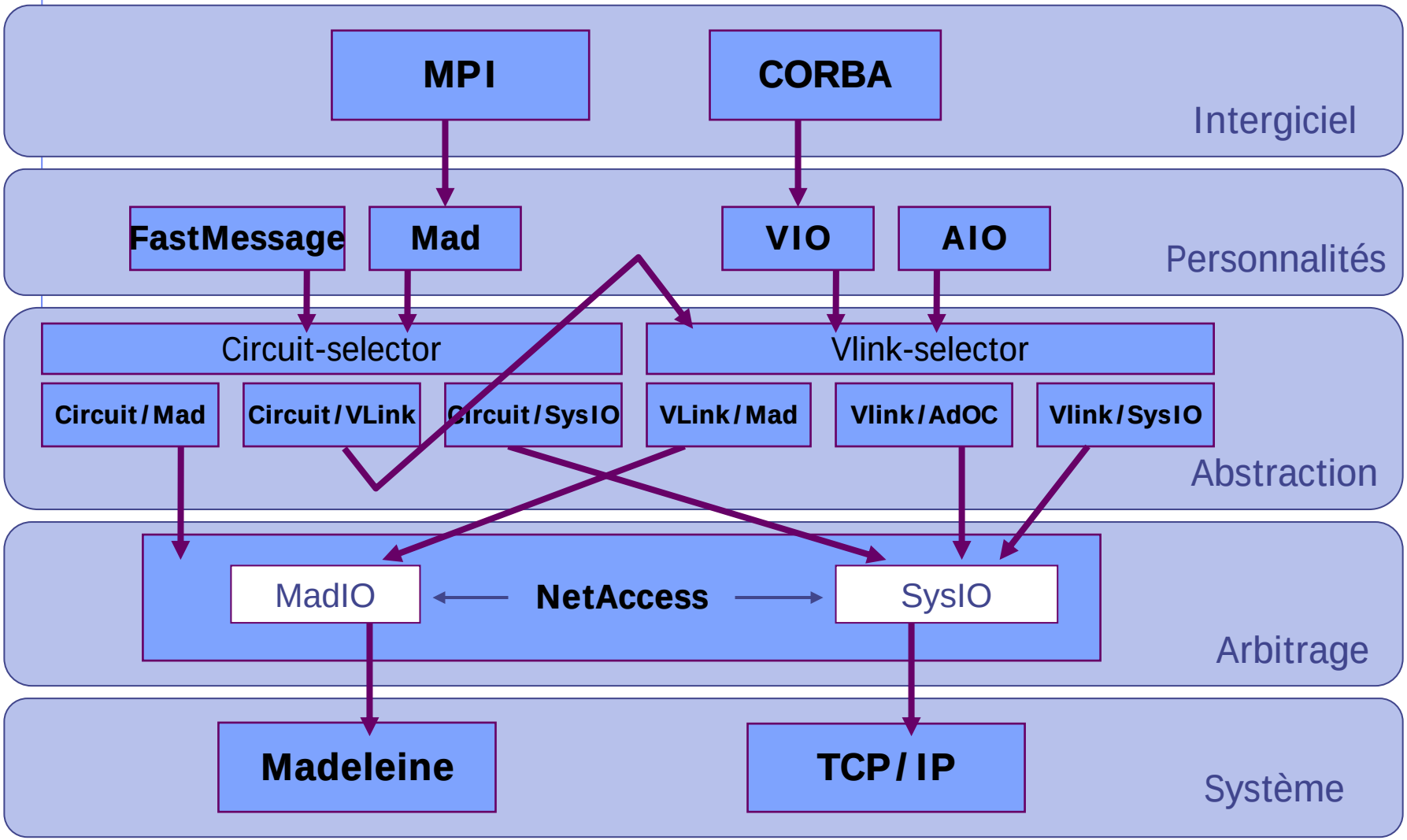


Modèle de communication multi-paradigmes

- Notre approche
  - Différencier les paradigmes
    - Paradigme « parallèle »
    - Paradigme « système reparti »
  - Interface unifiée à l'intérieur d'un paradigme
- Solution hybride
  - Deux interfaces génériques : parallèle+répartie

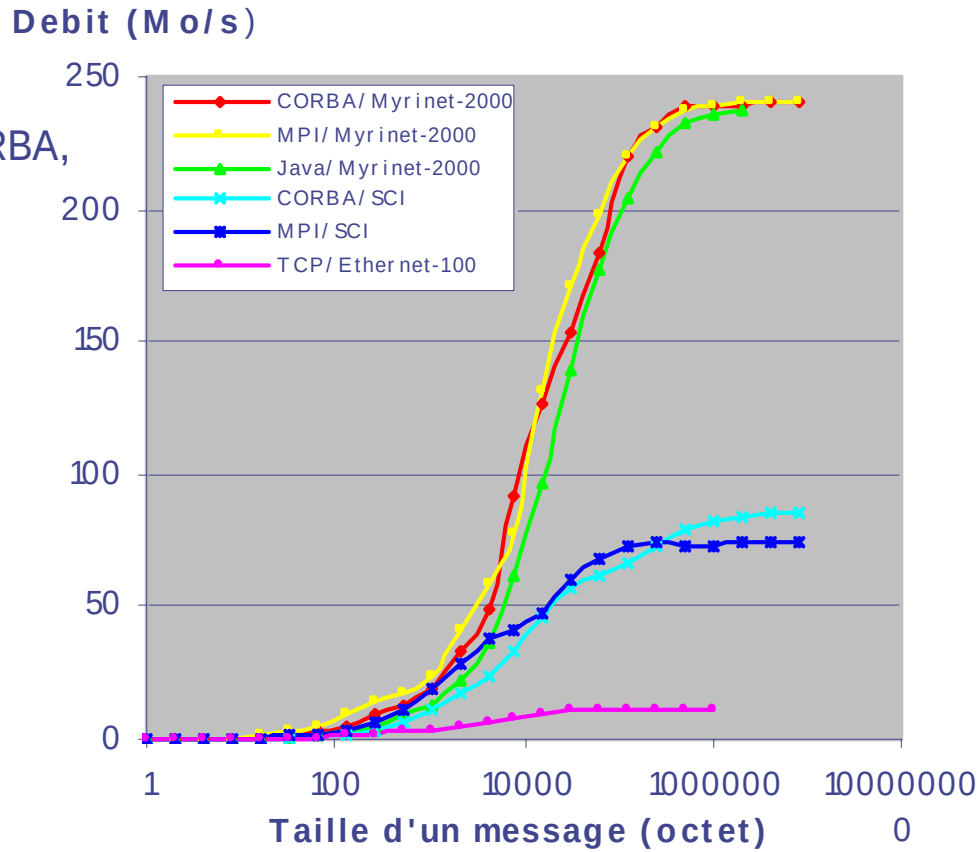


Modèle de communication de PadicoTM



Quelques résultats sur SAN

- Débits
  - Myrinet-2000 : 240MB/s
  - SCI : 86MB/s
  - 96% du débit matériel pour CORBA, MPI et Java
- Latence
  - CORBA/Myrinet-2000: 18 µs
  - MPI/Myrinet-2000: 11 µs
  - Java/Myrinet-2000: 40 µs
  - CORBA/Ethernet: 150 µs
- Réseaux récents
  - Myrinet 10G/MX
    - Latence : ~ 5 µs
    - Débit : ~ 1 000 Mo/s
  - InfiniBand/IB
    - Latence : ~ 5 µs
    - Débit : ~ 800 Mo/s

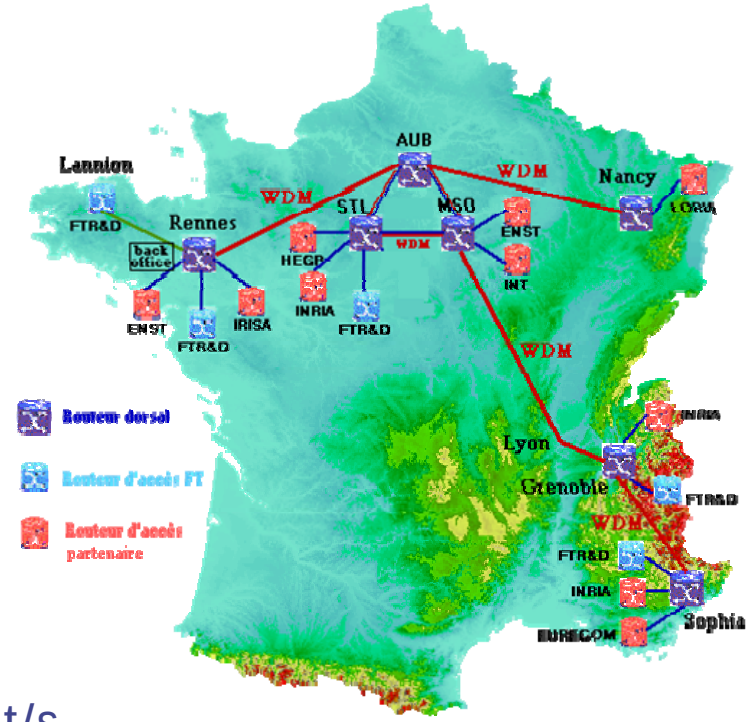


Débit d'exécutifs sur PadicoTM
sur Myrinet-2000, SCI et Fast-Ethernet

Quelques résultats sur WAN

- Flux parallèles
  - Réseaux VTHD (Rennes-Nice)
  - Sans (1 flux) : 8 Mo/s (64 Mb/s)
  - Avec (4 flux) : 12 Mo/s (96 Mb/s)
  - Pour tous les exécutifs
    - Surcoût négligeable

- Compression adaptative (AdOC)
  - Projet Algorille (Loria)
  - Lien ADSL 608 kbit/s
    - Sans compression : 580 kbit/s
    - Avec compression : 800 à 1200 kbit/s
  - Pour tous les exécutifs
    - Surcoût négligeable



Le réseau VTHD

Bilan

- Un modèle de communication multi-paradigmes
  - Modèle à trois couches
    - Arbitrage, abstraction & personnalité
  - Découplage des interfaces de programmation des interfaces réseaux
  - PadicoTM, une implémentation du modèle
    - Utilise Marcel & Madeleine, logiciels développés par le projet RUNTIME (LaBRI)
    - Validé avec un grand nombre d'exécutifs
- Bénéfice
  - Modèle de programmation multi-exécutifs comme GridCCM
  - Transparence dans l'utilisation des grilles informatiques
    - Réseau privé, pare-feu, ...

Perspectives

- Collaboration avec Alexandre Denis
  - CR INRIA dans le projet RUNTIME (LaBRI)
- Adaptabilité
  - Changement à chaud d'une décision
- Politiques à définir pour
  - Affectation des communications aux réseaux
  - Choix d'utilisation d'un réseau
    - Ex: chiffrement, multi-flux, compression, ...

Bilan & Perspectives

Bilan

- Vers un environnement de programmation des grilles
  - Objectif : définir un modèle de programmation efficace, simple et découplé des ressources
  - 3 axes de recherche
    - Modèle de programmation
    - Support exécutif
    - Modèle de déploiement
  - Réalisation de prototypes pour la validation expérimentale sur des applications

- Réalisation d'un environnement de programmation : Padico
  - PaCO++ - une implémentation portable des objets parallèles CORBA
    - ~ 28000 lignes de Java/Python et C++
  - GridCCM - une implémentation portable des composants parallèles CORBA
    - En cours de développement
    - Basé sur PaCO++
  - PadicoTM - une plate-forme d'intégration d'intergiciels communicants
    - ~ 31 000 lignes de C/C++, 2300 lignes de Java
  - ADAGE - un outil de déploiement automatique d'applications sur grilles de calcul
    - ~ 35000 lignes de C/C++.

Perspectives

- Continuer à découpler les modèles de programmation des ressources
  - Modélisation
    - Coordination entre composants
    - Algorithmique à base de composants logiciels
  - Conception
    - De nouveaux types de ports
      - Protocoles
  - Déploiement
    - Stratégie de placement des éléments
      - Qualité de service
    - Support des aspects dynamique de l'application & des ressources
      - Modèle de composant
      - Interconnexion avec les modèles d'adaptation

Remerciements

- Merci à tous !
- Et plus particulièrement à
  - Thierry Priol, Alexandre Denis, André Ribes, Sébastien Lacour, Hinde Bouziane, Boris Daix, Zsolt Németh, Gabriel Lopez, Landry Breuil
  - Gabriel Antoniu & Luc Bougé
  - Aux membres du projet PARIS
  - Aux personnes des différents services de l'IRISA