

# Quelques projets à l'interface Grilles de Production / Grilles de Recherche sur le site Toulousain

L. Broto<sup>1</sup>, M. Daydé<sup>1</sup>, D. Hagimont<sup>1</sup>, Th. Monteil<sup>2</sup>, P. Stolf<sup>1</sup>,  
R. Sharrock<sup>2</sup>, I. Touche<sup>3</sup>

13 Octobre 2009

<sup>1</sup>Institut de Recherche en Informatique de Toulouse

<sup>2</sup>LAAS-CNRS

<sup>3</sup>Laboratoire de Génie Chimique

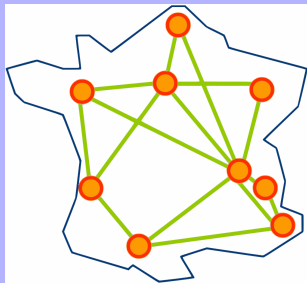
October 12, 2009

## Plan de l'exposé

- ▶ **Site Toulouse - Midi-Pyrénées de Grid'5000 : Grid'Mip**
- ▶ **Le portail d'applications scientifiques de l'INPT**
- ▶ **TLSE : site d'expertise en algèbre linéaire creuse**
- ▶ **Actions envers les industriels dans le cadre du PRAI**
- ▶ **TUNe: Administration Autonome**

## Grid'Mip : Site Toulouse Midi-Pyrénées de Grid'5000

► 9 sites



Site	Noeuds	Coeurs
Bordeaux	202	650
Grenoble	66	240
Lille	99	250
Lyon	135	270
Nancy	167	574
Orsay	342	684
Rennes	260	714
Sophia	178	568
Toulouse	137	434

Un total de 1586 noeuds, soit **4384 coeurs**.

## Configuration déployée

- ▶ Depuis 2005 Cluster Violette : 57 noeuds - bi-pro AMD Opteron 2.2Ghz, 2G de mémoire par noeud, 430G disque dur (ACI GRID)
- ▶ Fin 2007 achat de 140 nouveaux noeuds de calculs (financés par Programme PRAI Région / FEDER)
  - ▶ Cluster Pastel : 80 noeuds - bi-pro bi-coeurs AMD Opteron 2.6Ghz, 8G de mémoire par noeud, 480G disque dur
  - ▶ Cluster Capitole : 60 noeuds réservés hors charte et industriels

- ▶ Configuration totale actuelle disponible au CICT :
  - ▶ 676 coeurs, volume mémoire total de 1,2 Toctets et environ 40 Toctets de disque.
  - ▶ Puissance théorique  $\approx$  3,4 TFlops.



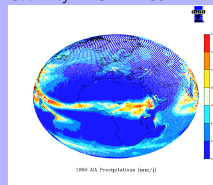
## Accès à Grid'Mip

- ▶ Infrastructure à double vocation
  - ▶ Contribuer au projet national Grid'5000
  - ▶ Favoriser une utilisation régionale dans le cadre du PRAI (Programme Régional d'Actions Innovatrices, co-financements Région et FEDER): laboratoires et partenaires industriels (avec un accent sur les PME-PMI)

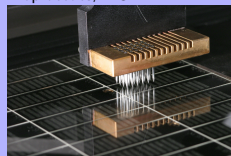
## Projets du site Toulouse/Midi-Pyrénées

- ▶ Applications de grande taille (CFD, astrophysique, génie chimique, matériaux, électromagnétisme, ...)
- ▶ Applications multi-paramétriques (expertise en matrices creuses, étude du climat et 'Global Change', Génomique fonctionnelle)
- ▶ Gestion des grilles et activités autour du middleware
  - ▶ AROMA : logiciel de gestion de ressources sur une grille de clusters
  - ▶ Outils d'administration d'autonome TUNE
  - ▶ Agents mobiles pour la gestion de grille
  - ▶ Gestion de grille et de services (sécurité, QoS, monitoring et contrôle, configuration dynamique, ...)
  - ▶ Optimisation de requêtes pour des BD distribuées de grande taille
  - ▶ Virtualisation du stockage de données sur une grille
  - ▶ Visualisation de grands volumes de données
  - ▶ Simulation de réseaux

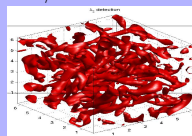
Courtesy of CERFACS



Courtesy of Lab. Biotechnologie et Bioprocédés, INSA



Courtesy of E. Climent and M. Maxey  
LGC / Brown U.



## Le portail d'applications scientifiques de l'INPT

Les applications les plus adaptées à l'architecture grille :

- ▶ Applications **multi-paramétriques** :
  - ▶ Un même calcul exécuté plusieurs fois avec des paramètres d'entrée différents
  - ▶ Diminution du temps de calcul par augmentation du nombre de cas traités simultanément car disponibilité d'un grand nombre de processeurs
- ▶ Applications **parallèles** :
  - ▶ Un seul calcul exécuté sur plusieurs noeuds
  - ▶ Diminution du temps de calcul par augmentation du nombre de noeuds et intérêt du temps de soumission réduit
- ▶ Dans tous les cas :
  - ▶ Nombre de processeurs disponibles bien plus élevé
  - ▶ Files d'attente des gestionnaires de batchs plus courtes

## Exemple d'un calcul mono-site

Sur un unique site :

- ▶ Transfert des données
  - ▶ `scp machineLocale:données siteToulouse:données`
- ▶ Connexion à la grille
  - ▶ `ssh siteToulouse`
- ▶ Réservation des ressources et lancement du calcul
  - ▶ `oarsub -l core=16,walltime=8 script`
- ▶ Récupération des résultats
  - ▶ `scp siteToulouse:résultats machineLocale:résultats`

## Exemple d'un calcul multi-site


Calcul parallèle sur deux sites :

- ▶ `scp machineLocale:données siteToulouse:données`
- ▶ `ssh siteToulouse`
- ▶ `oargridsub -d repertoire -p script -w 10  
siteToulouse:rdef="/core=8",siteSophia:rdef="/core=16"`
  - ▶ Sophia : `sleep 10*3600`
  - ▶ Toulouse : `scp siteToulouse:données siteSophia:données`
  - ▶ Toulouse : `oargridstat -l ID > machine`
  - ▶ Toulouse : `mpirun -machinefile machine -np 24 calcul`
  - ▶ Toulouse : `ssh siteSophia; oardel job; scp siteSophia:résultats  
siteToulouse:résultats`
- ▶ `scp siteToulouse:résultats machineLocale:résultats`

## Interface de soumission

- ▶ Créer une couche supplémentaire entre le chercheur et la grille pour lui permettre d'utiliser les ressources disponibles tout en s'affranchissant des contraintes techniques.
  
- ▶ Consultation de la disponibilité des ressources
- ▶ Soumission de job
  - ▶ Copie des données
  - ▶ Réservation des ressources
  - ▶ Lancement du calcul
  - ▶ Rapatriement de toutes les données sur le serveur web
  
- ▶ Consultation de l'état du job
- ▶ Récupération des données
- ▶ Relance du job si besoin pour certains codes
- ▶ Effacement des données inutiles

## Interface de soumission



Interface de soumission  
pour applications gridifiées

Centre de Compétences & Ressources Informatiques (CCR) de l'INP

Login | Recherche | Appli | Mfix | Documentation

- [Charte d'utilisation de la grille](#)
- [Statistiques d'utilisation de la grille](#)
- [Réalparh](#)
- [Vasep](#)
- [Algorithme Genetique](#)
- [Gibbs](#)
- [Mfix](#)
- [InAspid](#)
- [Iodim](#)

Pour connaître l'état de la grille :

Entrez le nombre de ressources voulu:

---

Pour soumettre un job - Etape 1

On va stocker l'archive d'un répertoire contenant les fichiers mfix.dat, fortran et si besoin .RES. Les archives supportées sont de type : .tar, .tgz, .zip.

Sélectionner l'archive à télécharger :

---

Pour soumettre un job - Etape 2

On va récupérer le nombre d'heures maximum d'une exécution, et le nombre de ressources à utiliser.  
Entrez un nombre correspondant au nombre d'heures souhaitées: (ce temps comprend aussi le transfert des données, veuillez en cas de grosse production de données à tenir compte de ce temps)

Entrez un nombre correspondant au nombre de ressources à utiliser: (Attention, vous devez avoir rajouté les paramètres NODESI, NODESJ, NODESK dans le fichier mfix.dat, avec NODESI\*NODESJ\*NODESK égal au nombre de ressources réservés)

Vous pouvez utiliser ce champ pour donner un nom à votre expérience:

---


Pour connaître l'état actuel de votre job et agir dessus :

Entrez l'identifiant du job :

---

Pour obtenir la liste des identifiants :

## Interface de soumission



Interface de soumission  
pour applications gridifiées



Centre de Compétences & Ressources Informatiques (CCRI) de l'INP

Login : itouche | Appli : Mfix | Documentation

- Charte d'utilisation de la grille
- Statistiques d'utilisation de la grille
- testrun
- Yasp
- Algorithme Genetique
- folche
- Mfix
- trAccel
- testim

Le job avait les caractéristiques suivantes :

Nom : test  
Id de l'interface : 1205318063162 - id de OAR : 91591  
Lancement à Toulouse par itouche  
Temps maximal : 1 - Nombre de ressources : 4  
Soumission : 2008-03-12 11:36:18 - Demarrage : 2008-03-12 11:36:19 - Fin : 2008-03-12 11:45:38  
Temps d'attente : 0:0:1 - Temps horaire de l'exécution : 0:9:19 - Temps Cpu de l'exécution : 0:37:16  
Localisation des données : /home/toulouse/itouche/interface\_web/mfix/job/1205318063162/run  
Etat actuel du job : Terminated, copie, non efface  
Relance du job : false

Vous pouvez visualiser vos fichiers d'entrée :

- [mfix.dat](#)
- [fichier](#)

Vous pouvez visualiser vos résultats :

- [La sortie standard OAR](#)
- [La sortie erreur OAR](#)

Vous pouvez télécharger l'archive du répertoire run : [run.tgz](#)

Une fois que vous avez téléchargé vos résultats, vous pouvez effacer toutes traces de votre job. Attention, cette action est irréversible.

1205318063162 - user: itouche

Vous pouvez relancer votre job. Pour cela, veuillez modifier votre fichier mfix.dat et redonner le nombre d'heure souhaité, et le nombre de processeurs

Sélectionnez le fichier mfix.dat:

Entrez un nombre correspondant au nombre d'heures souhaité:

Entrez un nombre correspondant au nombre de processeurs souhaité:

## Mise en place

- ▶ 6 mois :
  - ▶ Structure du portail
  - ▶ Intégration de 4 applications
  
- ▶ Fonctionnement des applications
  - ▶ Définir le type de l'application : séquentielle, multi-paramétrique, parallèle
  - ▶ Modifications : multi-paramétrique, point de reprise...
  - ▶ Voir les contraintes en terme de compilateur, librairies...
  - ▶ Définir l'intérêt/possibilité du multi-site
  
- ▶ Fonctionnement de la grille
  - ▶ Trouver les compilateurs/librairies installés sur les différents sites
  - ▶ Machine dédiée à la compilation ou non
  - ▶ Utilisation d'une tâche par noeud ou une tâche par processeur

## VASP

- ▶ Etude quantique de défauts dans les réseaux métalliques
- ▶ Application parallèle
- ▶ Problèmes de compilation
- ▶ Mono-site uniquement
  
- ▶ Cas test :
  - ▶ CALMIP : sur 4 processeurs : 44h 54mn.

- ▶ GridMIP :

Noeuds	Durée
4	43h 03mn
8	23h 52mn
16	14h 03mn
24	20h 05mn
32	28h 47mn

## Isoturb

- ▶ Simulation numérique des écoulements diphasiques en génie des procédés
- ▶ Application parallèle
- ▶ Compilateur et librairie libres et portables
- ▶ Possibilité d'exécution multi-site
- ▶ Cas test :
  - ▶ IDRIS : sur 4 processeurs : 3h 40mn.
  - ▶ GridMIP :

Noeuds	Durée	Speed up	Efficacité
1	20h 25mn		
4	12h 32mn	1,63	0,40
8	07h 40mn	2,66	0,33
16	02h 58mn	6,87	0,43
24	02h 23mn	8,62	0,36
32	18h 01mn	1,13	0,04

## Algorithme Génétique

- ▶ Etude des problèmes de gestion ou d'ordonnancement d'ateliers discontinus
- ▶ Application séquentielle
- ▶ Architecture du code
  - ▶ Initialisation
  - ▶ Boucle
    - ▶ Génération des paramètres
    - ▶ Calculs
  - ▶ Calcul de moyennes
- ▶ Application multi-paramétrique!
- ▶ Installation sur 7 sites
- ▶ 400h en séquentiel
- ▶ 4h sur 100 noeuds

## Gibbs

- ▶ Simulation moléculaire pour la prédiction de propriétés d'équilibre liquide-vapeur
- ▶ Application séquentielle
- ▶ Architecture du calcul
  - ▶ Phase d'équilibrage (1 à 2 millions de configurations)
  - ▶ Génération de quelques millions de configurations
  - ▶ Moyenne statistique
- ▶ Application presque multi-paramétrique!
- ▶ Cas test avec 9.5 millions de configurations
  - ▶ 21h 37mn en séquentiel
  - ▶ 7h 38mn sur 8 noeuds
    - ▶ Equilibrage - 1.5 millions : 4h 32mn
    - ▶ Production - 8 fois 1 million : 3h 06mn maximum

## Bilan

- ▶ 16 utilisateurs
- ▶ 7 applications
- ▶ 63 218h CPU pour 2007
- ▶ 28 425h CPU pour 2008

## TLSE : Tests for Large Systems of Equations

Site d'expertise en algèbre linéaire creuse <http://gridtlse.org>

- ▶ Aide les utilisateurs à sélectionner logiciels / paramètres sur leur problème via des procédures d'expertise
- ▶ A partir de scénarios d'expertise spécifiés par les spécialistes + logiciels, collections de matrices, ...
- ▶ Aussi plate-forme de test pour les développeurs
- ▶ Financé par l'ACI GRID au travers du **Projet GRID-TLSE** et actuellement par le Projets ANR LEGO 2006-2009, ANR SOLSTICE 2007-2009 et le Projet France-Japon **CNRS / JST REDIMPS**



**LEGO**

# The GRID-TLSE Platform

Execution of a straightforward scenario

User expert request

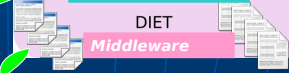
Request for expertise

Workflow (set of experiments in WEAVER format)

XML description of experiments

Solvers

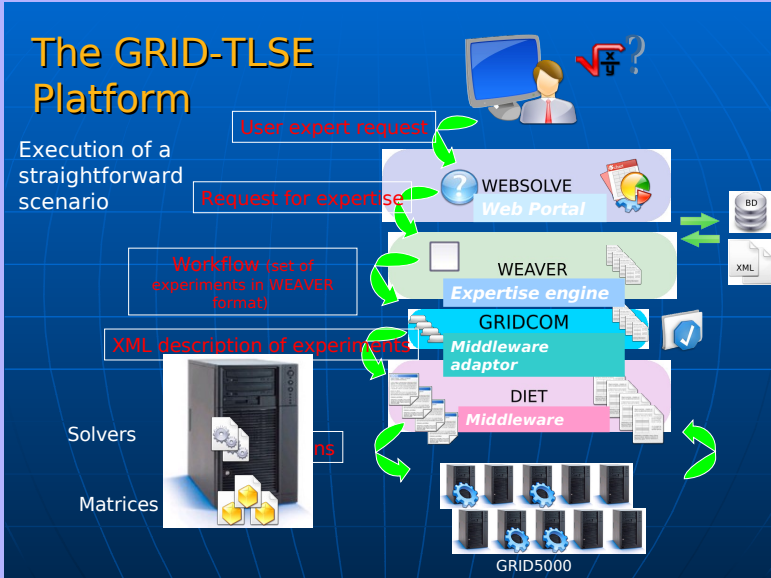
Matrices



GRID5000



ns



## Description des procédures d'expertise

- ▶ On ne va pas demander aux experts d'algèbre linéaire de déployer ou d'appeler des services sur la grille
- ▶ Nous avons introduit :
  - ▶ une interface graphique de haut niveau pour la description des scénarios appelée **GEOS**
  - ▶ basée sur une description sémantique des logiciels, des paramètres de contrôle, des résultats et des matrices à partir de méta-données définies dans **PRUNE**

## Scénarios d'expertise

- ▶ Description de type data-flow
- ▶ Hiérarchisés : un scénario peut en appeler un autre
- ▶ Analyse / exécution d'un scénario peut nécessiter plusieurs étapes
- ▶ introduction de :
  - ▶ Caractéristiques : nombre de flops, mémoire, . . .
  - ▶ Opérateurs: Transformation, Filtering, Generation;
  - ▶ ...

## Description graphique des scénarios

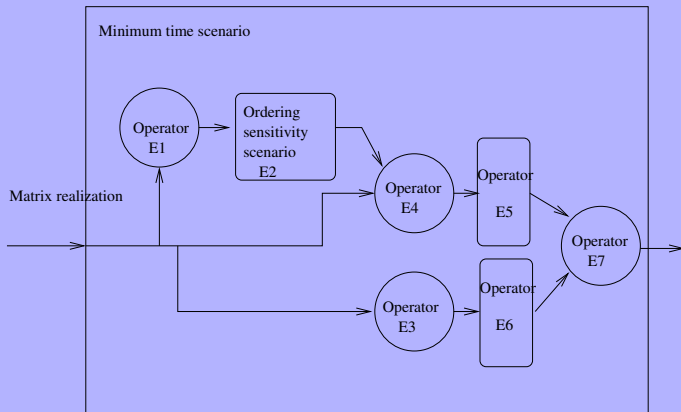


Figure: Minimum Time Scenario.

Objectif : identifier la combinaison ordering / factorisation minimisant le temps d'exécution.

## TLSE Project : Conclusion

- ▶ Ouvert au public depuis plus d'un an
- ▶ Utilisé en tant que plate-forme d'échanges et de test dans plusieurs projets (ANR SOLSTICE, REDIMPS, ...)
- ▶ Coopération avec JAEA : solveurs et machines japonaises



The screenshot shows the GRID-TLSE website. At the top, a blue banner contains the text "TEST FOR LARGE SYSTEMS OF EQUATIONS". Below the banner is a navigation menu with links: Home | Overview | People | Events | Links | Open Positions. On the left side, there is a sidebar with a logo for "GRID-TLSE" and a user login section showing "Logged as : the" and a "Sign out" link. Below the login section is a vertical menu with expandable items: Menu, Welcome, My home, Expert, Administrate, Matrices, Expertises, Groups, Resources, Tools, and BIB-TLSE. The main content area on the right features a section titled "What is GRID-TLSE ?" with a dashed line separator. The text below explains the site's purpose: "This web site aims to provide tools and software for sparse matrices. It will allow the comparative analysis of a number of direct solvers (free or commercially distributed) on user-submitted problems, as well as on matrices from collections available on the site. The site will provide user assistance in choosing the right solver for its problems and appropriate values for the control parameters of the selected solver. The computations are carried over a computational grid. It also includes a bibliography on sparse matrices and access to collections of sparse matrices." Below this text, a section titled "You will be able to :" lists several capabilities: perform experiments on your own sparse matrix, quickly evaluate sparse direct solvers and obtain statistics on solving sparse linear systems, use the site as a platform for cooperative work on sparse matrices (features are available for creating a private work group), consult the database of bibliography references on sparse linear algebra, and consult the database of collections of sparse matrices.

## Objectifs : accès à des moyens technologiques

- ▶ Mutualisation d'outils software / hardware (pics de charge, diminution de coûts, ...)
- ▶ Logiciels de calcul (MSC NASTRAN/PATRAN, SAMCEF, LS-DYNA, ...) ou autres (visualisation, design, ...) utilisés ponctuellement
- ▶ Logiciels très onéreux, à acquérir et à maintenir → pb de l'amortissement économique des investissements
- ▶ Idem pour le hardware: besoins ponctuels de puissance de machine (pics de calcul ou de stockage)

## Emergence de communautés

- ▶ Favorise la création de tissu/réseau de moyens et de compétences
  - ▶ Hardware
  - ▶ Modélisation
  - ▶ Essais
  - ▶ Expertise sur les outils et logiciels

*A fédérer autour de plateformes communes*
- ▶ Mise en place moyens de communiquer (partager / analyser) les modèles, résultats, ...
- ▶ Développer la notion de simplification des échanges inter entreprise, labo, ... afin de développer/valider des modèles numériques, des logiciels, ...

## Mutualisation de logiciels

- ▶ Offres groupées entre CALMIP (SGI ALTIX à 256 noeuds) et Grid'Mip → disponibilité maximale
- ▶ Gibbs, FLUENT, VAPS, Isoturb pour la communauté académique
- ▶ SAMCEF, FLUENT pour les industriels (Hyperworks à venir)

## Criblage Virtuel : Projet avec les industriels de la semence

- ▶ Discussion au sein d'un consortium de partenaires :
  - ▶ Laboratoires de recherches / organismes publics : CERFACS, IRIT, THEOGONE, CICT
  - ▶ Industriels : LIMAGRAIN, RAGT, (ARVALIS ?), UPETEC, Orange Business Services
- ▶ Objectif : améliorer la mise au point des nouvelles semences **Criblage Virtuel** et déployer une infrastructure de grille opérationnelle
- ▶ Mieux comprendre / prédire lien entre génotype et phénotype
- ▶ Grands volumes de données / infrastructures distribuées à grande échelle
- ▶ Fouille de données, classification, assimilation de données, multi-agents adaptatifs, algorithmes génétiques, réseaux de neurones, ...
- ▶ Financement pour la phase exploratoire
- ▶ But : préparer un FUI ou une ANR avec plus de partenaires