



Quand SimGrid rencontre gLite

Action Interfaces Recherche en Grilles – Grilles de production

Frédéric Suter

13 octobre 2009

Contexte et motivations

Grilles de Production

- ▶ Infrastructures opérationnelles
- ▶ Quelques limitations connues
 - ▶ Utilisation sous-optimale des ressources
 - ▶ Taux de pannes élevé

Problématiques de recherche associées

- ▶ Optimisation des performances
- ▶ Dimensionnement des services
- ▶ Design applicatif
 - ▶ Réduction du cycle de développement
 - ▶ Prédiction de performances
 - ▶ Debugging

Freins

- ▶ Non dédiées à l'expérimentation
- ▶ Mesures non reproductibles

Contexte et motivations

Grilles de Production

- ▶ Infrastructures opérationnelles
- ▶ Quelques limitations connues
 - ▶ Utilisation sous-optimale des ressources
 - ▶ Taux de pannes élevé

Problématiques de recherche associées

- ▶ Optimisation des performances
- ▶ Dimensionnement des services
- ▶ Design applicatif
 - ▶ Réduction du cycle de développement
 - ▶ Prédiction de performances
 - ▶ Debugging

Freins

- ▶ Non dédiées à l'expérimentation → Aladdin/Grid'5000
- ▶ Mesures non reproductibles → Simulation

Objectifs

Scientifique

Définir un environnement **contrôlé** reproduisant **fidèlement** le comportement d'une grille de production pour **analyser** le fonctionnement du système **sous charge** et **évaluer** différentes stratégies d'**optimisation**

Organisationnels

- ▶ Établissement d'un **consortium** solide
 - ▶ Applications / Recherche / Production
- ▶ Définition d'une **méthodologie** d'étude innovante
- ▶ Dépôt de projet **ANR** courant 2010

Tentatives précédentes

Optorsim

- ▶ William Bell, David Cameron, Paul Millar, Luigi Capozza, Kurt Stockinger and Floriano Zini. **Optorsim : A Grid Simulator for Studying Dynamic Data Replication Strategies**. International Journal of High Performance Computing Applications, Vol. 17, No. 4, 403-416 (2003)

EDGSim

- ▶ <http://www.hep.ucl.ac.uk/~pac/EDGSim/>

Avec SimGrid

- ▶ Thomas Ferrandiz and Vania Marangozova. **Managing Scheduling and Replication in the LHC Grid**. In Proceedings of CoreGRID Workshop on Grid Middleware, July 2007.

Notre méthode

- ▶ Gestion de l'intergiciel de production **gLite**
 - ▶ **Émulation** des composants → comportement **réaliste**
 - ▶ **Déploiement** sur **Grid'5000**
- ▶ Gestion de l'infrastructure matérielle
 - ▶ **Modélisation** et **simulation** des ressources de **stockage**, **calcul** et de **communication**
 - ▶ Utilisation de **SimGrid**
- ▶ Mise en charge de l'environnement
 - ▶ Injection de **charges applicatives** et de **fautes**
- ▶ Applications cibles
 - ▶ **Validation** de l'outil proposé

Partenaires et compétences

Centre de Calcul de l'IN2P3

- ▶ **Contact** : Frédéric Suter (coordinateur)
- ▶ SimGrid et administration gLite

Image et modèles – CREATIS

- ▶ **Contact** : Tristan Glatard
- ▶ Applications et utilisateurs gLite

AlGorille – LORIA

- ▶ **Contact** : Lucas Nussbaum
- ▶ SimGrid et expertise Grid'5000 (et outils associés)

Modalis – I3S

- ▶ **Contact** : Johan Montagnat
- ▶ Observatoire de la grille et utilisateurs gLite

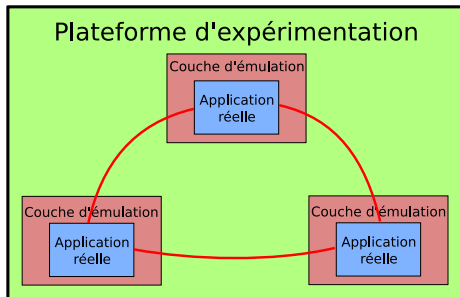
Axes d'études

- SimGrid
- gLite
- Aladdin/Grid'5000
- Traces
- Applications

SimGrid – Interposition

Idée

- ▶ Émuler en intercalant un simulateur entre l'application et la plateforme
 - ▶ Simplicité d'utilisation des simulateurs
 - ▶ Champ d'application de l'émulation



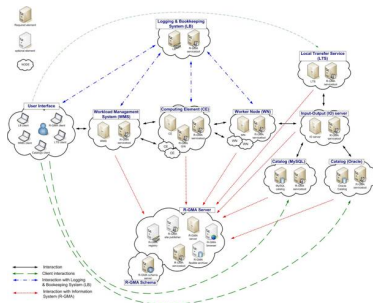
- ▶ Collaboration nécessaire avec spécialistes gLite
 - ▶ Quelles APIs pour quels modules

Description d'EGEE

- ▶ Multi-cluster large échelle
 - ▶ Comment gérer plusieurs milliers de nœuds de calcul ?
 - ▶ Gestion du routage complexe
- ▶ Comment obtenir les informations nécessaires à SimGrid ?
 - ▶ Latence et bande passante des liens de communication
 - ▶ Puissance de calcul des nœuds
 - ▶ Besoin de collaboration avec les administrateurs gLite

Organisation complexe

- ▶ Beaucoup de modules
- ▶ Problèmes de gestion de la sécurité (VOMS)
- ▶ Environnement en évolution



- ▶ Besoin de déterminer ce qui est émuleable
 - ▶ Risque majeur de ce projet

Aladdin/Grid'5000

Scientific Linux

- ▶ OS nécessaire au déploiement de gLite
- ▶ OS par défaut sur Grid'5000 : Debian
- ▶ Experiences récentes ont montré que déployer Scientific Linux n'est pas immédiat
 - ▶ Rien de bloquant toutefois

Adaptation des outils

- ▶ KaDeploy par exemple
 - ▶ Automatisation de la configuration, post-déploiement (certificats, ...)
- ▶ Interactions nécessaires avec les équipes de développement
 - ▶ A expliciter selon la criticité

Traces

Paramétrisation

- ▶ Quelles informations utiliser pour simuler la mise en charge de l'environnement ?
 - ▶ Profils des tâches
 - ▶ Types de charge
 - ▶ Profils de pannes

Injection

- ▶ Comment transformer les traces en entrées du simulateur ?
- ▶ Transformation des traces ou modification du simulateur ?

Soumission de jobs sur EGEE

Personal Workstation



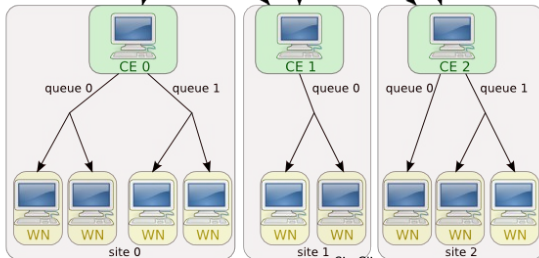
User Interface



Resource Broker



Computing Element



Working Node



Latency
R

Soumission de jobs sur EGEE

Personal Workstation

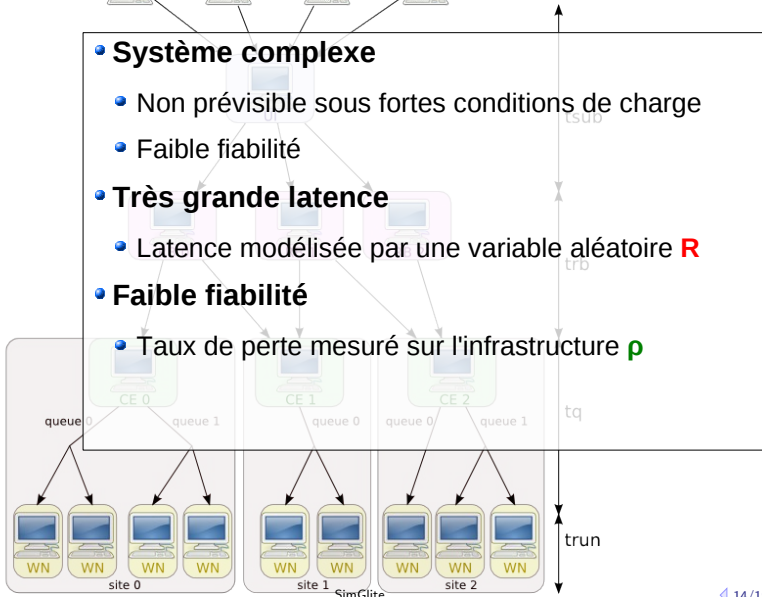


User Interface

Resource Broker

Computing Element

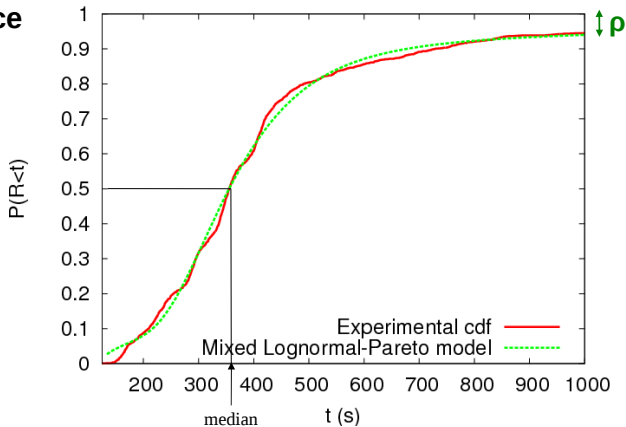
Working Node



Modéliser latence et pertes

- **Estimation de la latence**

- D'après les traces
- Loi de probabilité (heavy tailed p.d.f.)

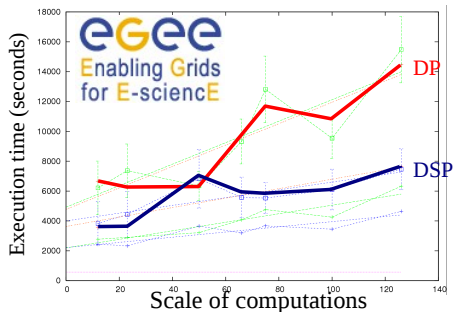
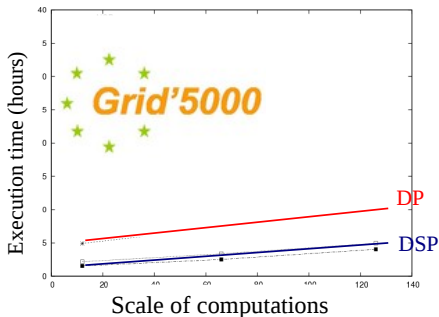


- La loi de probabilité observée correspond à une loi mixed Lognormal – Pareto

Estimation des performances

• Comparaison de performances entre plates-formes

- Même workflow exécuté sur des ressources dédiées (G5K) et en production (EGEE)
- Deux modes de parallélisation (DP et DSP)



- Le modèle, appliqué a posteriori, explique les variations de charge en production

Applications

Modèle

- ▶ Applications décrites par des chaînes de traitements (workflows)
- ▶ Description structurée facilitant la simulation
 - ▶ Description des services (paramètres d'entrée, fichiers, dépendances)
 - ▶ Description des dépendances entre services (transferts de données et structures de contrôle)

Benchmarks pour l'évaluation

- ▶ Comparaison des performances simulées VS observées en production
 - ▶ Fréquence et nature des erreurs
 - ▶ Statistique des temps d'ordonnancement, de mise en queue, de transferts de données et d'exécution
- ▶ Différents modes de soumission de tâches
 - ▶ Soumission directe gLite
 - ▶ Pré-filtrage des sites
 - ▶ Tâches pilotes
- ▶ Stratégies de réplication des données

Applications : exemples envisagés

Simulateur IRM (SiMRI)

- ▶ parallélisme de données sur les coupes et les volumes de l'image
- ▶ MPI pour la simulation d'une coupe

Simulateur images US (FIELD-II)

- ▶ parallélisme de données sur les lignes d'une coupe, les coupes et les volumes
- ▶ code Matlab

Simulateurs TEP et radiothérapie (Sorteo et GATE)

- ▶ simulations Monte-Carlo
- ▶ divisible load

Segmentation d'images cardiaques

- ▶ Balayages de paramètres sur des BDD d'images
- ▶ Parallélisme de données + MPI envisageable

Forces du projet

- ▶ Partenariat Recherche / Production
 - ▶ Recherche : Simulation et gestion de Grid'5000
 - ▶ Production : Administration Glite, traces et applications
 - ▶ Interactions nécessaires (mais à définir) pour tous les axes d'études
 - ▶ Approche innovante mais potentiellement risquée
 - ▶ Mélange émulation et simulation
 - ▶ Risques liés à la complexité de l'environnement de production
 - ▶ Outil final à fort potentiel
 - ▶ Aide au utilisateurs actuels d'EGEE (développement, calibration, ...)
 - ▶ Démonstrateur pour la communauté recherche (ordonnancement, réplication, déploiement, ...)
- ⇒ peut faciliter le transfert Recherche → Production