# GRID'5000 une plate-forme d'expérimentation pour les systèmes distribués à large échelle

Philippe d'Anfray[*]
CEA, Délégation ANR-CI
Centre de Saclay
91191 Gif-sur-Yvette Cedex
Philippe.d-Anfray@cea.fr

Emmanuel Jeannot
LORIA
INRIA-Nancy Grand-Est
615, rue du Jardin Botanique
54600 Villers les Nancy
Emmanuel.Jeannot@Loria.fr

## Résumé

*Aujourd'hui, grâce à Internet, il est possible d'interconnecter des machines du monde entier pour traiter et stocker des masses de données. Cette collection hétérogène et distribuée de ressources de stockage et de calcul a donné naissance à un nouveau concept : les grilles informatiques. L'idée de mutualiser les ressources informatiques vient de plusieurs facteurs, évolution de la recherche en parallélisme qui, après avoir étudié les machines homogènes, s'est attaquée aux environnements hétérogènes puis distribués ; besoins croissants des applications qui nécessitent l'utilisation toujours plus importante de moyens informatiques forcément répartis. La notion de grille peut avoir plusieurs sens suivant le contexte : grappes de grappes, environnements de type GridRPC (appel de procédure à distance sur une grille)., réseaux pair-à-pair, systèmes de calcul sur Internet, etc... Il s'agit d'une manière générale de systèmes dynamiques, hétérogènes et distribués à large échelle. Un grand nombre de problématiques de recherche sont soulevées par les grilles informatiques. Elles touchent plusieurs domaines de l'informatique : algorithmique, programmation, intergiciels, applications, réseaux.*

*L'objectif de GRID'5000 est de construire un instrument pour réaliser des expériences en informatique dans le domaine des systèmes distribués à grande échelle (GRID). Cette plate-forme, ouverte depuis 2006 aux chercheurs de la communauté grille, regroupe un certain nombre de sites répartis sur le territoire national. Chaque site héberge une ou plusieurs grappes de processeurs. Ces grappes sont alors interconnectées via une infrastructure réseau dédiée à 10 Gb/s fournie par RENATER. À ce jour, GRID'5000 est composé de 9 sites: Lille, Rennes, Orsay, Nancy, Bordeaux, Lyon, Grenoble, Toulouse et Nice. Début 2007, GRID'5000 regroupait plus de 2500 processeurs et près de 3500 cœurs.*

## Mots clefs

grille de calcul et de données, systèmes pair à pair, réseaux à haut débit, intergiciel, applications distribuées.

[*]GIP RENATER jusqu'en Octobre 2007

## 1 Avertissement

Cette présentation de Grid'5000 comprend sous une forme résumée, certaines contributions à la partie commune du rapport scientifique (mi-2007) du projet. Ce rapport est un ouvrage collectif des membres du *steering committee* et des responsables des sites Grid'5000 : **Franck Cappello[1], Michel Daydé, Philippe d'Anfray, Frédéric Desprez, Emmanuel Jeannot, Yvon Jégou, Stéphane Lantéri, Raymond Namyst, Nouredine Melab, Pascale Primet et Olivier Richard**.

## 2 Motivations et objectifs

Les grilles et les systèmes pair-à-pair sont de plus en plus utilisés comme environnements de production dans des domaines très variés (EGEE, TeraGrid, SETI@home, Edonkey, Spyke) et constituent une source d'inspiration pour de nombreux programmes de recherche.

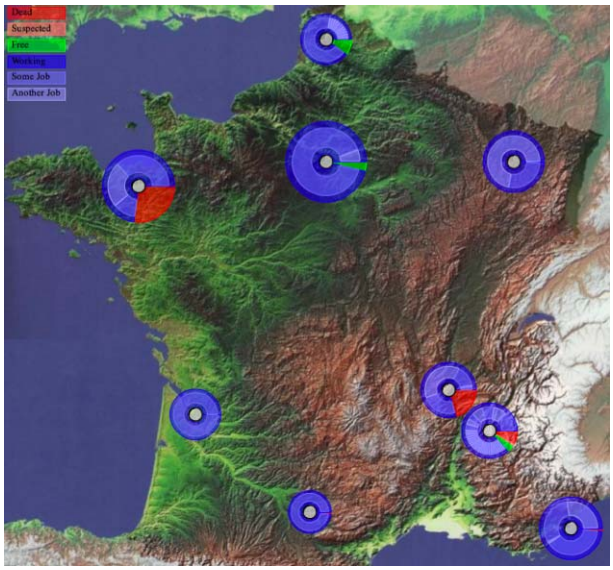



Figure 1 - *Activité de la plate-forme Grid'5000.*

[1]*head of the steering committee*

Ces systèmes distribués restent très difficiles à concevoir, à déployer, à exploiter et à optimiser ; les difficultés viennent de la complexité des logiciels mis en jeu, du grand nombre (passage à l'échelle) de ressources hétérogènes et volatiles utilisées.

La recherche dans ce domaine, s'intéresse à toute la « pile logicielle » de l'utilisateur jusqu'au matériel (hardware) : applications, environnements de programmation, environnements d'exécution, intergiciels (middleware), systèmes d'exploitation, jusqu'à la couche réseau. Les études couvrent les problématiques de performance, de sécurité, d'utilisabilité, de fiabilité et plus globalement de qualité de service.

Lors des études préparatoires de Grid'5000 [1] la plupart des recherches menées dans le domaine des grilles et des systèmes pair-à-pair étaient conduites sur des simulateurs [2][3][4], des émulateurs [5] ou encore des plates-formes de production. Néanmoins tous ces environnements présentent des limitations pour l'étude de nouveaux algorithmes ou des optimisations. Les simulateurs adressent une problématique spécifique (mécanisme, comportement) en faisant abstraction du reste du système. Il est très difficile en pratique de valider les résultats du simulateur, ce qui constitue la principale limitation. Les émulateurs autorisent une meilleure approche des facteurs qui régissent le comportement des systèmes distribués en permettant d'exécuter le code du système distribué sur une plate-forme contrôlée. Mais ces derniers ne permettent pas encore d'appréhender la variété et la complexité des environnements réels. Enfin si les plates-formes de production permettent *a priori* une meilleure approche du réel, elles ne sont pas ouvertes à des expérimentations « sur les couches basses » (système d'exploitation, protocoles réseau,...). En outre ces plates-formes, souvent très chargées, n'offrent pas un environnement permettant la reproductibilité des expériences.

La recherche dans le domaine des grilles et des systèmes pair-à-pair nécessite ainsi l'accès à une véritable plate-forme expérimentale sur laquelle les chercheurs en informatique pourront mener des expériences, observer les systèmes distribués à grande échelle, étudier les limites du système sous toutes formes de conditions expérimentales (injecter des traitements, du trafic, des fautes,...), et effectuer des mesures précises. Une telle plate-forme peut certes être constituée ponctuellement en interconnectant des grappes (*clusters*) d'universités et centres de recherche dédiées à d'autres applications, mais tout cela nécessite un long travail de préparation, l'intervention des administrateurs, etc. Comme pour les plates-formes de production, tout n'est pas possible et la reproductibilité n'est pas assurée non plus.

Suite à cette analyse, Grid'5000 a été conçu comme un grand instrument scientifique à l'instar des accélérateurs de particules ou des télescopes utilisés par les physiciens ou les astrophysiciens. Les chercheurs pourront partager un grand nombre de ressources expérimentales géographiquement distribuées. Ils pourront réserver ces ressources, les configurer, y lancer leurs expériences et effectuer des mesures précises. Enfin ces expériences pourront être reproduites, plus tard avec les mêmes conditions expérimentales. Le projet RAMP[2] aux Etats-Unis a été conduit avec des préoccupations similaires. PlanetLab [6] n'est pas vraiment une plate-forme d'expérimentation au sens de Grid'5000 car les ressources ne sont pas dédiées et les conditions réseau n'y sont pas reproductibles.

Le projet DAS3[3] aux Pays-Bas est probablement la réalisation la plus proche de Grid'5000 au point de vue conceptuel. La plate-forme actuelle DAS3 lancée fin 2006, a été conçue pour permettre une reconfiguration dynamique du réseau d'interconnexion optique.

La figure 2 ci-dessous présente les différentes approches pour la recherche sur les systèmes distribués -plus spécialement les grilles et les systèmes pair à pair- et les environnements correspondants à la disposition des scientifiques.

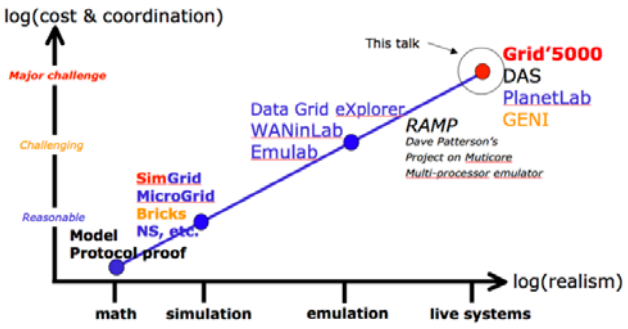
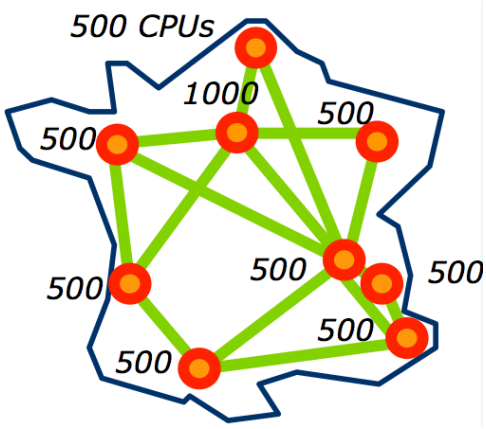



Figure 2 - *Méthodologies pour l'étude des systèmes distribués.*

L'axe vertical (coût et coordination) prend en compte les difficultés techniques ainsi que les coûts d'exploitation et de personnel pour la gestion et la coordination de la plate-forme.

# 3 Architecture de la plate-forme




[2]http://ramp.eecs.berkeley.edu

[3]http://www.starplane.org/das3/

Figure 3 - *La plate-forme Grid'5000 (projet)*

Les choix de conception de Grid'5000 prennent en compte les « retours d'expérience » des plates-formes précédentes (e-Toile, DAS1, DAS2, ...) et les besoins des chercheurs en informatique à travers la description d'un premier ensemble d'expériences. Tout cela a conduit à la réalisation d'une vaste plate-forme expérimentale - plusieurs milliers de CPUs distribués sur une dizaine de sites - offrant de nombreuses possibilités de reconfiguration et comportant une infrastructure de contrôle et de pilotage.

Les utilisateurs de la plate-forme peuvent installer, charger et exécuter leurs propres images qui peuvent inclure toute la pile logicielle depuis le système d'exploitation jusqu'au applications. Pour cette raison le système est isolé du reste de l'internet mais aucune limitation de trafic entre sites n'est introduite. Une autre contrainte est de maintenir une certaine homogénéité dans la plate-forme pour ne pas introduire une complexité trop grande et aussi pour être à même, par exemple, d'évaluer correctement les accélérations obtenues en augmentant le nombre de noeuds pour une application. Maintenir homogène environ 2/3 de la plate-forme semble un bon compromis. Enfin la reproductibilité des expériences impose d'avoir des liens réseau dédiés à la plate-forme.

# 4 Réseau RENATER

## 4.1 L'infrastructure réseau Grid'50001

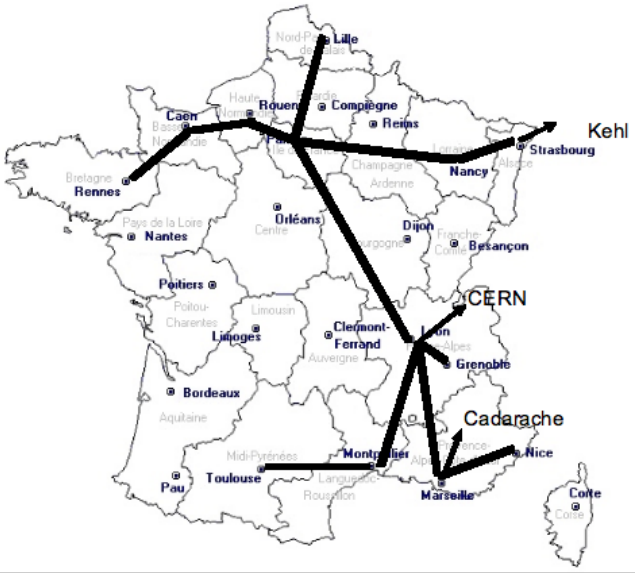



Figure 4 - *L'infrastructure projet dans RENATER 4*

L'infrastructure initiale délivrée par RENATER était basée sur un maillage complet de VPN MPLS dédiés offrant une connectivité à 1Gb/s entre les sites de Grid'5000. Cette interconnexion utilisait l'infrastructure de production de RENATER-4 (lignes louées à 2.5Gb/s). Dans un second temps, les liaisons ont été migrées vers l'infrastructure projet en fibre noire où des *lambdas* - liens à 10gb/s de niveau 1 - peuvent être configurés entre les sites Grid'5000 reliés à la fibre noire [7].

Avec cette nouvelle approche de « VPN optique » tous les sites GRID'5000 peuvent se voir, au niveau 2, à l'intérieur d'un même VLAN. La figure 4 montre l'infrastructure projet en fibre noire déployée en plus de l'infrastructure de production dans le cadre de RENATER-4.

La figure 5 ci-dessous montre les *lambdas* à 10Gb/s activés pour l'interconnexion des sites Grid'5000.

L'utilisation de la fibre noire s'est fait peu à peu en démontant les anciens VPN. Tous les sites Grid'5000 sont actuellement connectés à 10Gb/s sauf Bordeaux.

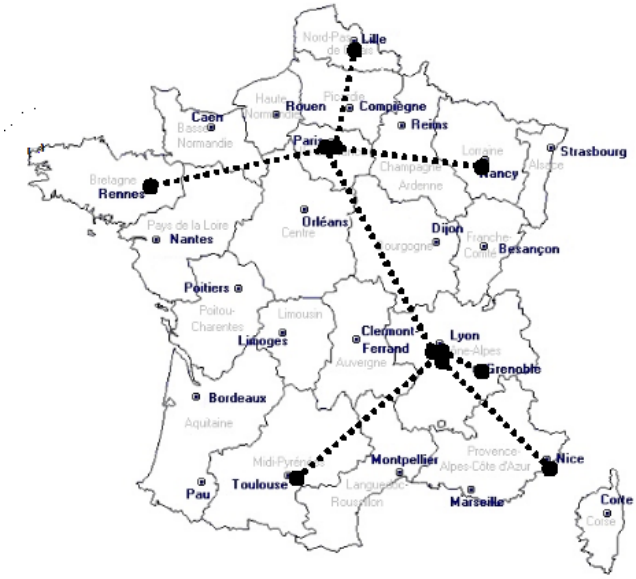



Figure 5 - *Les lambdas à 10Gb/s activés pour Grid'5000*

## 4.2 Métrologie réseau dans Grid'5000

RENATER offre un certain support aux utilisateurs de la plate-forme Grid'5000. A cet effet un URL[4] spécifique a été mis en place pour permettre de récupérer des données de métrologie sur le réseau. Il s'agit des informations relatives à l'utilisation des lambdas qui ont été configurés ainsi que les ports alloués pour l'interconnexion des sites. De même une carte « *weather map* »[5] spécifique a aussi été créée pour donner des informations en temps réel sur la charge et la disponibilité du réseau.

## 4.3 Interaction avec Grid'5000

Le réseau d'interconnexion spécifique de la plate-forme Grid'5000 a été intégré dans la définition de l'infrastructure RENATER-4.

Les fibres noires et les équipements optiques DWDM (Dense Wavelength Division Multiplexing : multiplexage en longueur d'onde) ont été déployés pour permettre l'interconnexion des sites à 10Gb/s. Les retours des expériences réseau menées sur Grid'5000 sont particulièrement intéressants pour la conception des futures générations d'infrastructure réseau (RENATER 5, ...).

[4] http://pasillo.renater.fr/metrologie/GRID5000

[5] http://www.renater.fr/Metrologie/map-FON-GRID-5000

Le GIP RENATER soutient aussi les collaborations internationales. La connectivité de RENATER avec le réseau européen GÉANT2[6] inclut des interfaces à 10Gb/s qui peuvent être utilisées dans le cadre de projets spécifiques. Ainsi l'interconnexion avec la plate-forme DAS3 aux Pays Bas mais d'autres extensions de Grid'5000 sont planifiées vers les Etats-Unis ou le Japon.

# 5 Aspects scientifiques

## 5.1 Expériences

L'objectif principal de Grid'5000 est de permettre le déploiement, l'exécution et la collecte des résultats d'expérience de grille à grande échelle. A l'heure actuelle, environ 340 expériences ont été proposées -déjà réalisées ou planifiées-, ces expériences adressent tous les niveaux de la pile logicielle, des ressources de grille à l'utilisateur final.

Environ 50 expériences travaillent au niveau réseau. Cela inclut les recherches sur les protocoles à haut débit, la métrologie, les transferts de données (communication collectives, grands volumes de données), la simulation de l'Internet, ...
Plus de 100 expériences se focalisent sur le niveau intergiciel : test de « *middleware* », systèmes pair-à-pair, découverte de ressources, tolérance aux pannes (*fault tolerant MPI*, ...), déploiement d'applications, ...
Quelques 40 expériences s'intéressent au niveau tests et évaluation d'environnement de programmation pour la grille : *grid MPI*, *grid RPC*, par exemple l'environnement DIET[7], mais aussi objets et composants, ...

Le niveau applicatif suscite lui un très fort intérêt avec plus de 80 expériences. La principale préoccupation est l'étude des performances des applications portées sur la grille dans de nombreux domaines de la physique mais aussi la biologie et la finance, ...

Enfin environ 20 applications n'entrent dans aucune de ces catégories et concernent les systèmes d'exploitations répartis (XtremOS), les techniques de virtualisation et les outils plus spécifiquement destinés à la plate-forme : contrôle, pilotage, métrologie, injecteur de trafic, outils de déploiement et de reconfiguration...

## 5.2 Logiciels d'aide à l'expérience

Dans le cadre de Grid'5000, deux outils principaux ont été développés pour aider les utilisateurs à mener leur expériences. Le premier outil (OAR), permet de réserver la grille, le second (kadeploy) permet de configurer l'environnement. Ces deux outils ont été développés conjointement et sont très intégrés (on peut par exemple réserver et déployer un noeud en même temps).

### 5.2.1 OAR

OAR[8] est un système de batch (réservation par lot) open source, développé spécialement pour Grid'5000. Il permet aux utilisateurs de faire des réservations hiérarchiques allant de la grappe au coeur d'un processeur. Il permet de réserver à l'avance des noeuds et de suspendre l'exécution d'une application.

### 5.2.2 Kadeploy

Kadeploy[9] permet aux utilisateurs de configurer leur environnement d'expérience. Concrètement, ils peuvent créer et déployer une image d'un système (Linux, BSD, ...) qu'ils peuvent administrer et configurer et dans lequel ils peuvent installer les logiciels spécifiques à leur expérience. À tout moment, l'utilisateur peut sauvegarder son environnement et le reprendre ultérieurement.

## 5.3 Un instrument scientifique

Les retombées de Grid'5000 en terme de quantité et de qualité de résultats scientifiques, et d'impact sur la communauté scientifique, permettent de classer la plate-forme parmi les grands instruments scientifiques.

Parmi les indicateurs, citons le nombre d'utilisateurs actifs : actuellement environ 280 répartis dans environ 60 laboratoires de part le monde ! Grid'5000 a été utilisé dans le cadre de 5 Habilitations à Diriger les Recherches, 14 thèses, des dizaines de mastères et plus de 300 publications.

Chaque site de Grid'5000 inclut dans ses programmes de recherche des expériences spécifiques. Les résultats sont publics et détaillés sur le site du projet[10].

L'activité de la plate-forme reste entre 50% et 70% de sa capacité et fonctionne de façon satisfaisante. Des pointes à 90% ont été observées, mais, à ce niveau de saturation, l'instrument n'est plus en mesure de répondre aux requêtes des utilisateurs.

Enfin, Grid'5000 est aussi utilisée dans des activités d'enseignement, et la plate-forme peut être impliquée dans des événements « *grand challenge* » comme par exemple le *Grid Plugtest* qui consiste à résoudre la plus grande instance possible du problème des N-reines sur une grille.

[6] http://www.geant2.net/

[7] http://graal.ens-lyon.fr/~diet/

[8] http://oar.imag.fr

[9] http://kadeploy.imag.fr

[10] http://www.grid5000.org

*Figure 6 - Charge de la plate-forme Grid'5000*

La figure 6 donne une idée de la charge de la plate-forme Grid'5000. Il s'agit des réservation des noeuds (2 CPU) sur une semaine. Sur la figure, une ligne par noeud, en abscisse les jours. Les barres verticales représentent les réservations des CPUs.

# 6 Conclusions et perspectives

Grid'5000 appartient à une nouvelle catégorie d'instruments de recherche pour les grilles et les systèmes pair-à-pair. Il s'agit de plates-formes à grande échelle qui peuvent être facilement reconfigurées, instrumentées et contrôlées. La plus grande différence entre Grid'5000 et les outils qui l'ont précédé est son très haut degré de reconfiguration, qui permet aux utilisateurs de redéfinir toute la pile logicielle à chaque expérimentation.

La construction de la plate-forme elle-même est un défi scientifique. Ici l'expérience d'autres communautés impliquées depuis longtemps dans la construction et l'utilisation de grands instruments scientifique (physiciens, astrophysiciens) s'avère précieuse. Typiquement, de nombreuses questions ont été soulevées concernant la qualité des expériences et des mesures effectuées sur Grid'500. Les métriques et les outils de mesure associés doivent être bien définis et compris ; les conditions expérimentales imposées avec une certaine rigueur. Il convient aussi de valider les outils d'injection de trafic ou de panne.

Grid'5000 appartient aussi à une nouvelle génération de ressources pour la recherche en informatique : des plates-formes de ressources, ouvertes et partagées par une large communauté d'utilisateurs (typiquement plusieurs centaines). Ces plates-formes sont des outils sophistiqués, maintenus par des ingénieurs experts, et offrant des logiciels et des matériels spécifiques pour la recherche.

## 6.1 Collaborations internationales

Comme nous l'avons déjà écrit, Grid'5000 et DAS3 ont beaucoup en commun au niveau de la conception et de la finalité.

Ce sont néanmoins des plates-formes complémentaires. Si dans Grid'5000, les possibilités de reconfiguration concernent la pile logicielle, le projet DAS3 s'est focalisé sur le contrôle -c'est à dire la reconfiguration- du réseau d'interconnexion.


*Figure 7: Interconnexion Das3 Grid'5000*

A partir de ce noyau, des plates-formes de dimension européenne pourraient être proposées dans le cadre du 7ème programme cadre de la commission européenne (FP7).

Enfin la collaboration avec l'initiative de recherche nationale Japonaise sur les grilles NAREGI[11] a conduit à réaliser - été 2007 - l'interconnexion de Grid'5000 (site de Lyon) avec des noeuds de NAREGI au Japon à travers un lien à 1Gb/s via les réseaux GÉANT2 et SINET.

Le projet d'interconnexion Grid'5000, DAS3 ouvrira de nouveaux horizons aux chercheurs des deux communautés tout en offrant une plate-forme de plus grande taille propice à de nouvelles études : interopérabilité, sécurité, ... Cette interconnexion est effective, depuis l'été 2007 à travers un lien dédié à 10Gb/s de GÉANT2 mis en oeuvre dans le cadre d'une collaboration de RENATER et du réseau hollandais SurfNet[12]

## Remerciements

Beaucoup de personnes ont contribué au succès du projet Grid'5000 initié par Michel Cosnard au niveau national. Thierry Priol et Brigitte Plateau sont respectivement directeur de l'ACI Grid et présidente du comité scientifique. Dany Vandromme, membre du *steering committee* est le directeur du GIP RENATER. Pierre Neyron est le responsable du comité technique. Nous remercions encore le ministère de la recherche, l'ACI Grid et l'ACI masses de données, l'INRIA, le CNRS, le GIP RENATER, les conseils régionaux d'Aquitaine, Bretagne, Ile de France et Provence Alpe Côte d'Azur, le conseil général des Alpes Maritimes ainsi que les universités Paris Sud à Orsay, Joseph Fourier à Grenoble, Nice-Sophia Antipolis, Rennes 1, l'Institut National Polytechnique de Toulouse / INSA / FERIA / Paul Sabatier à Toulouse, Bordeaux 1, Lille 1 / GENOPOLE, et l'Ecole Normale Supérieure de Lyon.

[11] http://www.naregi.org/index_e.html

[12] http://www.surfnet.nl/

## Bibliographie


[1] Franck Cappello, Eddy Caron, Michel Dayde, Frederic Desprez, Emmanuel Jeannot, Yvon Jegou, Stephane Lanteri, Julien Leduc, Nouredine Melab, Guillaume Mornet, Raymond Namyst, Pascale Primet, and Olivier Richard. Grid'5000: a large scale, reconfigurable, controlable and monitorable Grid platform. In Grid'2005 Workshop, Seattle, USA, November 13-14 2005. IEEE/ACM.

[2] Atsuko Takefusa, Satoshi Matsuoka, Kento Aida, Hidemoto Nakada, and Umpei Nagashima. Overview of a performance evaluation system for global computing scheduling algorithms. In HPDC '99: Proceedings of the The Eighth IEEE International Symposium on High Performance Distributed Computing, page 11, Washington, DC, USA, 1999. IEEE Computer Society.

[3] Henri Casanova, Arnaud Legrand, and Loris Marchal. Scheduling distributed applications: the simgrid simulation framework. In Proceedings of the third IEEE International Symposium on Cluster Computing and the Grid (CCGrid'03), may 2003.

[4] C. Dumitrescu and I. Foster. Gangsim: A simulator for grid scheduling studies. In Proceedings of the IEEE International Symposium on Cluster Computing and the Grid (CCGrid'05), Cardiff, UK, may 2005.

[5] Xin Liu, Huaxia Xia, and Andrew Chien. Validating and scaling the microgrid: A scientific instrument for grid dynamics. The Journal of Grid Computing, Volume 2(2):141 – 161, 2004.

[6] Brent Chun, David Culler, Timothy Roscoe, Andy Bavier, Larry Peterson, Mike Wawrzoniak, and Mic Bowman. PlanetLab: An Overlay Testbed for Broad-Coverage Services. ACM SIGCOMM Computer Communication Review, 33(3):00–00, July 2003.

[7] Philippe d'Anfray, Franck Simon. RENATER dark fiber project architecture. In HPDC 2006, IEEE International Symposium on High Performance Distributed Computing, poster session, Paris, June 19-23 2006 .