GRID'5000
Plate-forme de recherche expérimentale en informatique

Groupe de réflexion

Juillet 2003

Synthèse des recommandations

Le groupe de réflexion a émis plusieurs recommandations, rassemblées ici pour la commodité du lecteur :

Recommandation 1 **: Objectif stratégique.** Nous recommandons la création dès 2003 d'un grand instrument de recherche constitué de la plate-forme expérimentale GRID'5000 Cette plate-forme est indispensable à la communauté informatique nationale pour atteindre les objectifs stratégiques de la recherche en STIC dans le domaine du calcul distribué à grande échelle, et des application dimensionnantes.

Recommandation 2 **: Création d'un GIS.** La structure proposée pour administrer la plate-forme GRID'5000 est un GIS (Groupement d'Intérêt Scientifique) regroupant le CNRS, l'INRIA, les universités associées, et peut-être le CEA. Des établissements partenaires seraient invités à se joindre au GIS dès sa création : collectivités territoriales et industriels entre autres.

Recommandation 3 **: Pérennité de l'engagement financier.** Au delà de l'achat du matériel, la plate-forme ne se constituera pas par la simple interconnexion des grappes locales. Un plan financier pluri-annuel (sur trois ans) doit être établi pour dégager les ressources nécessaires au financement des personnels chargés de gérer et maintenir la plate-forme.

Recommandation 4 **: Synergie des efforts.** Au plan national, il faut mobiliser les équipes de recherches identifiées au sein de l'ACI GRID, des nouvelles ACI Masses de Données et Sécurité, des réseaux nationaux de recherche RNRT et RNTL, et ouvrir largement à la plate-forme à toutes la communautés des chercheurs. Il faut aussi associer largement les partenaires industriels, de l'offre (constructeurs informatiques et éditeurs de logiciels) et de la demande (applications). Enfin, dans le contexte européen, il faut inscrire les développements logiciels et applicatifs au coeur des projets du prochain PCRD de la CEE.

Table des matières

1 Composition du groupe de réflexion

Le groupe de réflexion a reçu une lettre de mission de la Direction de la Recherche au Ministère de la Recherche et des Nouvelles Technologies. Voici la composition du groupe :

Michel Cosnard (Michel.Cosnard@inria.fr)
Serge Fdida (Serge.Fdida@lip6.fr)
Olivier Poch (poch@igbmc.u-strasbg.fr)
Yves Robert (coordinateur) (Yves.Robert@ens-lyon.fr)
Pierre Valiron (Pierre.Valiron@obs.ujf-grenoble.fr)

Le correspondant du groupe à la Direction de la Recherche était Antoine Petit, conseiller pour les STIC (Antoine.Petit@recherche.gouv.fr). Nous le remercions tout particulièrement pour son aide.

2 Mission du groupe de réflexion

La Direction de la Recherche a souhaité étudier la création d'un programme national visant à promouvoir la création d'une plate-forme expérimentale de recherche en informatique, constituée d'une grille de calcul de grande ampleur. Elle a confié au groupe de réflexion la réalisation d'une étude précise d'intérêt et de faisabilité. La lettre de mission du groupe de réflexion propose en particulier quatre pistes d'étude :

Intérêt scientifique Il s'agit de préciser le projet, ses objectifs scientifiques, les verrous scientifiques ou technologiques à résoudre, et les retombées attendues

Contexte international Il s'agit de comparer le projet avec des opérations analogues sur les scènes internationale et européenne

Synergie au plan national Il s'agit de positionner le projet dans le paysage français, notamment par rapport aux infrastructures existantes

Cadre opérationnel Il s'agit de décrire le fonctionnement et l'organisation générale envisagés pour ce nouveau "grand instrument"

3 Les grilles informatiques

Les grilles informatiques sont des plates-formes de calcul à grande échelle, hétérogènes et distribuées. Nous en décrivons brièvement les principes essentiels, et nous donnons quelques exemples de grilles existantes. Nous expliquons ensuite la distinction à faire entre grilles de production et grilles de recherche. Enfin, nous rappelons l'importance fondamentale des plates-formes de recherche expérimentale en informatique. Nous concluons avec une petite synthèse de cette partie.

3.1 Pourquoi des grilles informatiques ?

La puissance de calcul, les données informatiques et les capacités de stockage seront-elles, un jour, accessibles sur le même mode décentralisé que l'est l'électricité domestique dans les pays développés ? Le concept de grille informatique (le nom est emprunté au *power grid*, le réseau électrique qu'on vient d'évoquer) correspond à la réalisation de vastes réseaux mettant en commun des ressources informatiques géographiquement distantes. Les grilles de calcul ou de données permettront d'effectuer des calculs et des traitements de données à une échelle sans précédent.

L'idée de connecter et partager des ressources informatiques dispersées était déjà présente dans les années 1960. Récemment, les avancées technologiques ont donné à cette idée des perspectives relativement concrètes. Cependant, la réalisation de grilles informatiques se heurte encore à de nombreuses difficultés. Des difficultés d'ordre technique d'abord, car il s'agit de faire communiquer et coopérer des matériels distants et hétérogènes par leurs modes de fonctionnement comme par leurs performances, de créer les logiciels permettant de gérer et distribuer efficacement les ressources cumulées du réseau, de mettre au point des outils de programmation adaptés au caractère diffus et parallèle de l'exécution des tâches confiées à la grille, etc. Mais on est aussi face à des difficultés de nature sociologique, voire économique ou politique, dans la mesure où la constitution d'une grille suppose de convaincre des entités individuelles - organismes publics, entreprises privées ou individus - de mettre leurs propres ressources à disposition d'une entité collective. Cela rejaillit sur les aspects techniques. Par exemple, chacun des noeuds d'une grille peut posséder des données ou des logiciels qu'il juge confidentiels et ne souhaite pas communiquer à autrui, d'où la nécessité de garantir, par des techniques appropriées, la sécurité des échanges au sein d'une grille.

En résumé, on peut dire que **les grilles informatiques existent à l'état embryonnaire, mais que de nombreuses recherches en informatique sont encore à mener pour passer du stade des expériences pionnières à celui de l'exploitation en vraie grandeur.**

3.2 Des exemples de grilles informatiques

Un premier exemple est la grille Globus qui a relié les supercalculateurs d'une dizaine de centres de calcul du continent nord-américain. L'objectif est d'obtenir un hypercalculateur parallèle et virtuel, qui doit offrir la possibilité à chaque centre de soumettre des calculs en utilisant la puissance de tous les autres. Des infrastructures de grille sont en train d'être créées par diverses communautés de scientifiques. C'est notamment le cas pour la physique des particules en Europe (dans le cadre du projet European Data Grid), qui doit se préparer à stocker et analyser les nombreux péta-octets (1 péta-octet = 10^{15} octets, l'équivalent d'environ 100 milliards de pages de livres) de données que produira annuellement le collisionneur LHC du Cern, dont le démarrage est prévu pour 2007.

Des grilles plus décentralisées existent aussi. Dans le contexte de l'initiative SETI@home, consacrée à la recherche d'éventuels indices de civilisations extra-terrestres parmi les signaux captés par le radio-télescope d'Arecibo, on distribue le travail d'analyse des signaux à plus d'un demi-million de volontaires, chacun acceptant de faire travailler pour SETI@home son ordinateur personnel aux moments où celui-ci est inoccupé. D'autres exemples analogues, où l'on répartit des calculs assez simples mais très volumineux sur un grand nombre d'ordinateurs personnels volontaires, sont le projet Great Internet Mersenne Prime Search, à la recherche de très grands nombres premiers, ou, en France, le Décrypthon qui a mobilisé pour quelques mois, jusqu'en mai 2002, environ 75 000 internautes volontaires afin de comparer les séquences des quelque 500 000 protéines connues du monde vivant. Bien que ces réseaux de calcul distribué ne constituent pas des grilles à proprement parler - le calcul est réparti par une autorité centrale, les ordinateurs participants sont de simples exécutants et n'ont pas la possibilité d'utiliser le réseau pour leurs propres besoins - ils illustrent la grande puissance informatique et les bénéfices que l'on peut attendre des grilles. Des puissances de plusieurs dizaines de téraflops (1 téraflops = 10^{12} flops ou 10^{12} opérations en virgule flottante par seconde) ont ainsi été obtenues dans le cadre de ces initiatives, chose inimaginable il y a seulement une dizaine d'années. Plus

proche du modèle de grille, et totalement décentralisé, le système pair-à-pair Kazaa permet l'échange de fichiers résidant sur plusieurs millions de PC domestiques.

Un dernier exemple, plus proche du grand instrument qui va nous intéresser dans ce rapport, est l'interconnexion par un réseau vraiment très rapide (VTHD) de grappes de dizaines de PC (et même centaines de PC pour le centre grenoblois), localisées dans les laboratoires de recherche de l'INRIA et du STIC-CNRS.

3.3 Grilles de production - grilles de recherche

Le concept de grille peut englober des architectures matérielles et logicielles très différentes, en fonction des objectifs recherchés. Nous identifions deux classes de plates-formes différentes, qu'il convient de séparer afin d'éviter de compromettre les performances de chacune d'elle.

3.3.1 Grilles de production

Les grilles de production sont des plates-formes applicatives, qui doivent fournir les mêmes services (heures CPU, temps de réponse) de manière constante, ininterrompue et fiable. **C'est typiquement le fonctionnement des centres de calcul qui sert de modèle pour les grilles de production.** L'utilisateur soumet son travail en mode non interactif (*batch*) et attend le résultat ; il veut que la machine qu'il utilise ait une architecture stable avec des services de qualité constante. L'administration est centralisée, effectuée par une équipe qui interagit avec les utilisateurs.

3.3.2 Grilles expérimentales de recherche

Par définition, les grilles expérimentales de recherche sont motivées par des travaux de recherche expérimentaux. Il s'agit d'une part de recherches en informatique (systèmes distribués, systèmes répartis, calcul parallèle, réseaux, protocoles, bases de données, fouille de données). Il s'agit d'autre part de recherches dans d'autres domaines (physique, biologie, calcul à haute performance, ...). Dans les deux cas, les travaux sont encore au stade fondamental, et nécessitent une première phase de validation expérimentale. Ils préludent à la réalisation d'un prototype logiciel, informatique ou applicatif, qui pourra

être déployé sur les grilles de production dans une deuxième phase (disons trois ans plus tard).

Les grilles expérimentales de recherche n'offrent pas de garantie de service, dans la mesure où des modifications peuvent être apportées sur l'infrastructure ou sur le système d'exploitation, afin d'étudier de nouveaux concepts ou algorithmes, de tester un nouvel intergiciel (*middleware*), ou encore de mettre en oeuvre un nouveau protocole réseau.

L'administration des grilles de recherche est décentralisée. Les travaux sont soumis en mode interactif : la plupart d'entre eux ne dure que quelques minutes, et la réponse doit être instantanée pour permettre au chercheur d'avancer dans sa démarche.

En résumé, **les caractéristiques des grilles de recherche sont la souplesse, le temps de réponse, et l'ouverture. Indispensable aux développeurs, ces caractéristiques peuvent aussi séduire les utilisateurs "éclairés"** (voir le Paragraphe 6.3).

3.3.3 Synthèse

Deux classes de plates-formes co-existent, avec des objectifs et des besoins différents. La reconnaissance de cette distinction est fondamentale pour le soutien d'initiatives dans le domaine.

3.4 De l'intérêt des plates-formes expérimentales

Les plates-formes informatiques expérimentales, sous toutes leurs formes, ont largement contribué à l'avancement de la recherche.

Par exemple, on ne dira jamais assez combien les arrivées des premiers réseaux de Transputers à Grenoble, et des deux premiers hypercubes à Grenoble et à Rennes, ont permis l'éclosion de bonne science en France. Malgré le retard de deux ans pris sur les américains, les équipes qui s'intéressaient à l'informatique du parallélisme ont pu réfléchir, expérimenter, innover. Non seulement la mise en oeuvre concrète des algorithmes et des environnements de programmation a t-elle permis de valider (ou d'invalider) les travaux théoriques : en retour, selon un cycle vertueux classique dans d'autres disciplines, l'expérimentation a donné naissance à de nouvelles idées, de nouveaux modèles, voire de nouvelles stratégies pour aborder les problèmes.

Au delà du monde des STIC, plusieurs applications ont été portées sur les nouvelles machines ; l'exemple le plus marquant est certainement celui

des bases de données distribuées qui sont passées dans le monde industriel, avec le succès qu'on connaît pour ORACLE, compagnie initialement liée à la machine N-CUBE.

Plus récemment, l'i-Cluster de Grenoble a permis à une large communauté de chercheurs d'expérimenter avec une grappe de plusieurs centaines de processeurs. Cette machine, que sa taille et sa souplesse de gestion (par le laboratoire ID) rendaient unique en France, a donné à plus d'une soixantaine d'équipes, pour deux-tiers issues de l'INRIA et du STIC-CNRS, et pour un tiers de disciplines applicatives, la possibilité de changer d'échelle. Les statistiques du laboratoire ID rapportent un taux d'occupation moyen entre 75% et 100% : ce taux remarquable démontre à lui seul l'intérêt de cette plate-forme pour la communauté.

Aujourd'hui, **il faut reproduire ce scénario gagnant. Il faut doter la communauté informatique d'un grand instrument, à la mesure de ses ambitions scientifiques.**

4 GRID'5000

4.1 Description matérielle

GRID'5000 constituerait la version pérenne, à grande échelle, de la plate-forme VTHD évoquée plus haut : il s'agit d'une plate-forme matérielle et logicielle, interconnectant à très haut débit une dizaine de grappes de PC de grande taille. Pour fixer un ordre de grandeur, chaque grappe pourrait comprendre 500 unités de calcul, d'où le total de 5000 qui donne le nom de code du projet GRID'5000 initialement proposé par Franck Cappello et Michel Cosnard.

Les centres qui pourraient accueillir les grappes de PC seraient typiquement les laboratoires de recherche en informatique des Réseaux Thématiques Pluridisciplinaires RTP *Calcul à hautes performances et calcul réparti* et RTP *Réseaux de communication* du STIC-CNRS, liste qui englobe des équipes issues de la plupart des unités de recherche de l'INRIA. La liste des sites pourrait comprendre Besançon, Bordeaux, Grenoble, Lille, Lyon, Nancy, Orsay, Rennes, Rocquencourt, Nice et Toulouse, soit onze villes couvrant le territoire national de façon complète.

Plusieurs de ces sites disposent déjà de grappes de PC, plus ou moins récentes. Pour ceux-ci, il s'agit de bâtir sur l'existant, et d'apporter des

compléments matériels. En particulier, le site de Grenoble dispose déjà de plusieurs centaines de processeurs, qui ont vocation à constituer une partie de la grappe du site. Le site d'Orsay va être doté d'une grosse grappe (entre 500 et 1000 unités de calcul) dans le cadre de l'ACI Masses de Données (projet Grid Explorer).

Pour transformer cette collection de grappes en grille expérimentale, il faut souligner l'absolue nécessité d'un réseau très rapide interconnectant les grappes. Renater (avec un réseau privé virtuel) pourrait fournir le haut débit nécessaire.

4.2 Résumé du programme scientifique

L'objectif premier du projet GRID'5000 est de **donner aux informaticiens les moyens expérimentaux pour mener à bien des recherches dans le domaine des grilles.**. Si la recherche française veut avoir un impact sur les standards logiciels, elle ne peut pas ne présenter que des résultats de simulation ! Il s'agit de lever des verrous aux plans algorithmique, simulation, intergiciels, réseaux et calcul pair à pair. Entre outils, méthodes, algorithmes et logiciels, les défis scientifiques sont nombreux.

L'objectif second est d'**associer la communauté applicative aux développements logiciels spécifiques nécessaires à la mise en oeuvre efficace d'applications dimensionnantes sur la grille**. Ici encore, des applications multi-paramétriques classiques au déploiement d'applications innovantes délocalisées sur la grande grille, les défis scientifiques sont nombreux.

Pour plus de clarté, nous ne détaillons pas davantage les recherches à mener : nous renvoyons le lecteur aux longs (et techniques) développements du Paragraphe 6.

4.3 Cadre opérationnel

Les points suivants sont discutés dans ce paragraphe :

- Structure d'administration et de supervision : organisation de la grille, supervision de l'exploitation et de la sécurité, fédération des équipes locales d'administration
- Comité scientifique : fixer les règles d'exploitation, gérer l'allocation des moyens
- Coût
- Cadre institutionnel

4.3.1 Structure d'administration et de supervision

Un premier danger à éviter est l'écueil auquel se heurte la majorité des grilles actuellement : la lourdeur administrative. Certains projets américains, financés pour trois ans pour monter une grille nationale entre universités, ont pris beaucoup de retard pour des questions de gestion.

Il est néanmoins évident qu'une structure d'administration est indispensable. Nous l'imaginons hiérarchisée, à l'image de l'infrastructure :

Niveau local Chaque grappe est administrée localement, au niveau du site qui l'héberge physiquement. Cette souplesse a un prix : elle entérine l'hétérogénéité des systèmes et des logiciels de base, qui différeront d'un site à l'autre. Ce mode de fonctionnement peut également induire une multiplication des efforts. A l'opposé, l'autonomie des grappes locales semble indispensable : celles-ci constituent en effet l'outil de travail quotidien des équipes de recherche. Et l'hétérogénéité fait partie du cahier des charges scientifique : les grilles sont, par nature, hétérogènes à tous les niveaux.

La mise en place de chaque grappe dans les sites participants dépendra de leurs ressources propres. On peut imaginer tous les scénarios, de l'installation dans la salle machine dont disposerait un grand laboratoire de recherche, à l'hébergement dans un centre de calcul voisin. Espace au sol, alimentation électrique, climatisation, accès au réseau rapide, et personnel de maintenance sont les différents paramètres qui dicteront les choix.

Niveau national Si les grappes sont gérées localement, il est indispensable de prévoir une structure administrative nationale qui pilote le projet et gère les moyens dédiés au fonctionnement national (connexion réseau, configuration en mode grille, gestion des comptes nationaux). De nombreuses questions organisationnelles se posent, et une action spécifique du STIC-CNRS, conduite par Franck Cappello, a été lancée pour proposer des solutions cohérentes. Un fonctionnement idéalisé proposerait les possibilités suivantes :

- Disponibilité de la plate-forme avec une qualité de service suffisante pour mener des expérimentations à grande échelle : outils globaux visant à faciliter le déploiement, le lancement et le contrôle d'expériences
- Reconfiguration des machines pour une expérience donnée : changement de système d'exploitation, de noyau, de pile de protocole de communication (IPV6, IPV4), des stratégies de sécurité

- Capacité de reproduction des conditions expérimentales et des résultats (en particulier une même expérience lancée depuis n'importe quel site devrait produire des résultats similaires) ; mécanismes de mesure précis et consistants, stockage, archivage, recherche, visualisation des résultats
- Système de réservation (semi)-automatisée des noeuds pour la réalisation d'expériences, système de soumission unique permettant de déployer des processus d'une manière homogène, simple et sécurisée

L'équipe d'administration nationale aura plusieurs missions :

- Définir une politique d'exploitation des grappes connectées à la grille de sorte que l'expérimentation soit possible et les résultats d'expériences pertinents
- Définir une politique de sécurité et sa mise oeuvre (firewall, accounting, certificats, etc.), qui permettent la réalisation d'expérience au niveau national tout en restant le plus possible compatible avec les politiques locales. A cet effet elle devra donner recommandations de technologies matérielles/logicielles
- Mettre en place une communication (page web) qui permette à tous les utilisateurs potentiels de savoir comment acquérir un compte, se connecter et utiliser la grappe

Interface Selon une politique à définir, l'accès à la grille nationale sera possible à partir des grappes locales. C'est le comité de pilotage scientifique, en appui sur l'équipe de gestion nationale, qui définira la politique d'attribution.

L'expérience montre que les coûts (humains et d'outils logiciels) d'administration et d'exploitation sont souvent plus importants que ceux des équipements investis. Les ressources humaines nécessaires requièrent souvent un haut niveau de qualification et sont donc plus difficiles à trouver. Les choix des outils s'avèrent essentiels car ils permettent de simplifier l'accès à la plate-forme (configuration, générateur de trafic, de pannes, etc.). Une mauvaise appréhension de ce facteur représente un risque d'échec majeur. Le temps est un facteur essentiel dans le déploiement d'une plate-forme, sa capacité d'impact et de succès.

4.3.2 Comité scientifique

Faire des expériences consomme beaucoup de temps en préparation, en réalisation et en analyse. Il y a une risque réel que la plate-forme GRID'5000

soit sous-utilisée à cause de ces deux problèmes (manque de compétence, manque de temps de la part des chercheurs). Le fonctionnement proposé ici vise à apporter une solution à ces problèmes : on peut s'inspirer du fonctionnement en cours dans d'autres disciplines (astronomie pour les télescopes, physique des particules pour les accélérateurs, etc.), et prévoir un fonctionnement au niveau national par "appel d'offre". Tous les x mois (x=6 par exemple), GRID'5000 lance un appel d'offre auprès de la communauté pour la réalisation d'expériences sur l'instrument. Les propositions sont retenues en fonctions de différents critères (pertinence scientifique, faisabilité, durée). Chaque proposition retenue se voit allouer un slot de plusieurs jours (éventuellement fragmentés) pendant lequel la plate-forme lui est attribuée. Les chercheurs préparent l'expérience en faisant des micro-tests au niveau des plates-formes locales. Pour la réalisation de l'expérience, les ingénieurs et les techniciens de GRID'5000, en concertation/collaboration avec les chercheurs, réalisent l'expérience en mobilisant toutes les ressources nécessaires.

Une telle démarche permet de rendre la plate-forme GRID'5000 très attractive, et permet aux chercheurs de se concentrer sur l'expérience en tant que telle (ceci n'empêche pas une expérience d'être interactive et de modifier les conditions en fonction des premiers résultats).

4.3.3 Coût

Deux ingénieurs par grappe locale, et trois ingénieurs au niveau national sont indispensables à la bonne marche du projet. Un total de 23 ingénieurs sur trois ans est un chiffre ambitieux mais réaliste. Tous les organismes nationaux (ministère, EPST, programmes ACI, RNTL, RNRT, etc), , régionaux (régions, conseils généraux), locaux (unités de recherche, laboratoires), ainsi que des partenariats industriels, devront être sollicités.

Le coût humain peut s'évaluer ainsi : 23 ingénieurs x 3 ans x 40 Kilo-euros en coût chargé = 2760 Kilo-euros. Le coût matériel de la plate-forme s'estime à 2 Keuros par CPU x 5000, soit 10 Mega-euros. Cette estimation inclut le réseau d'interconnexion interne à chaque grappe, mais pas l'interconnexion des grappes par le réseau rapide national. Un total de 15 Mega-euros pour l'opération complète semble constituer une estimation raisonnable. Toutefois, ce chiffre correspond à la totalité des coûts, alors qu'une partie plus ou moins importante de l'opération peut s'appuyer sur l'existant :

– sur le plan matériel, plusieurs sites disposent déjà de dizaines ou centaine de CPU. L'ACI Masses de données vient d'allouer un budget de

plus d'un million d'euros pour le projet GridExplorer, dont une large part sera consacrée à l'acquisition de la grappe sur le site d'Orsay.
- sur le plan humain, tout dépend de la politique scientifique des institutions concernées ; le CNRS, l'INRIA et les universités partenaires peuvent mettre à disposition des ingénieurs de recherche titulaires sur les postes d'administration au niveau tant national que local.

Il est indispensable d'établir un plan de financement complet, inventoriant les moyens existants, les moyens mis à disposition du projet par les institutions, et les moyens à apporter sur budget propre. La pérennité de l'investissement, et ainsi même que la viabilité de l'opération sur plusieurs années, sont en jeu.

4.3.4 Cadre institutionnel

S'agissant d'une action à priorité de recherche scientifique et dans laquelle seront essentiellement impliqués des organismes publics, nous proposons la création d'un GIS (Groupement d'Intérêt Scientifique) qui aurait pour mission :

- la coordination des principaux acteurs et la définition des objectifs de GRID'5000
- le suivi des activités scientifique et leur évaluation (nomination et suivi du comité scientifique)
- le suivi budgétaire et administratif

Ce GIS pourrait être créé par le CNRS, l'INRIA et les Universités associées. Le CEA pourrait être invité à le rejoindre. Un club des partenaires serait créé composé de 2 collèges, le collège des collectivités locales et régionales, et le collège des industriels.

Le GIS serait piloté par un comité de pilotage composé des directeurs des organismes fondateurs et assisté du conseil scientifique, composé pour moitié au moins de membres extérieurs aux organismes. Les clubs des partenaires pourraient être représentés au conseil scientifique. Le comité de pilotage nommerait le directeur de GRID'5000. Celui-ci serait assisté d'un conseil de groupement, en charge du suivi interne, et d'un comité technique, en charge de la coordination technique entre les sites.

En ce qui concerne la propriété des équipement locaux (grappes de calcul), elle serait commune aux établissements financeurs. Ces équipements seraient mis à disposition gratuite du GIS. Le réseau d'interconnexion serait financé en dehors du GIS et mis à disposition gratuite du GIS. Le GIS n'aurait donc

aucune charge en terme d'acquisition de la plate-forme, mais devrait veiller à son bon fonctionnement et à la facilité d'utilisation.

5 Contexte international et synergie nationale

5.1 Contexte international

Nous décrivons successivement le projet américain Teragrid, le projet néerlandais DAS-2 et le projet de grille hongroise, en synthétisant au passage quelques éléments comparatifs.

5.1.1 Teragrid

TeraGrid est un projet visant à bâtir la plus puissante infrastructure de calcul du monde, et délivrera 20 Téraflops de puissance de calcul (en fait, Teragrid devra se contenter de la deuxième position, après l'Earth Simulator japonais). L'architecture est répartie sur cinq sites, interconnectés par un réseau à 40 Gigabits, et peut traiter un volume de données de l'ordre d'un Petabyte. Des environnements de visualisation, et des intergiciels pour le calcul sur grille, seront développés pour faciliter l'accès des utilisateurs aux ressources.

Les sites sont les grands centres du super-calcul : National Center for Supercomputing Applications (NCSA-UIUC), San Diego Supercomputer Center (SDSC), Argonne National Laboratory, Center for Advanced Computing Research (CACR-CalTech), et Pittsburgh Supercomputer Center (PSC).

Commentaires Le projet Teragrid est un projet d'envergure, qui s'inscrit plus dans la problématique des grilles de production, et du super-calcul que dans la recherche expérimentale en informatique.

5.1.2 DAS-2 (Pays-Bas)

DAS-2 (Distributed ASCI Supercomputer) comprend 200 noeuds Dual Pentium-III nodes. Cette machine est constituée de cinq grappes, interconnectées par le réseau SurfNet, le *backbone* universitaire hollandais, alors que les PC des cinq grappes sont eux reliés localement par un réseau Myrinet.

Les cind grappes sont localisées au sein de cinq universités : Delft, Leiden, Utrec et deux à Amsterdam.

Commentaires Le projet DAS-2 est par certains cotés une version miniature du projet GRID'5000. Des grappes de PC sont reliées par un réseau rapide, et plusieurs équipes joignent leurs efforts de recherche, soit en s'appuyant sur les standards Globus et MPI2, soit en développant de nouveaux intergiciels (voir par exemple lesprojets Albatross et Ibis de H. Bal).

5.1.3 Grille hongroise

La grille hongroise (Hungarian Supercomputing Grid) interconnecte plusieurs super-calculateurs situés dans différentes villes. Ceux-ci sont reliés par le réseau universitaire hongrois. L'originalité du projet est due au choix du logiciel *Condor* pour déployer les programmes sur la grille. Condor est couplé à P-GRADE, un environnement de développement hongrois. Les services de base sont confiés à Globus.

D'autre part, mentionnons le projet *Hungarian cluster grid initiative*, qui vise à interconnecter les 99 grappes universitaires de Hongrie en une grille spécialisée. Chaque grappe contient 20 PC et un serveur. De jour, les grappes sont utilisées indépendamment pour des travaux pratiques d'étudiants ; la nuit elles sont reliées à travers le réseau universitaire hongrois (2.5 Gbit/sec) et se transformen en grille expérimentale de recherche.

Commentaires Les deux projets sont intéressants. L'architecture de la grille hongroise est classique (une collection de ressources hétérogènes, supercalculateurs et grappes, reliées par un réseau rapide) et s'inspire des projets de la NPACI américaine. Le réseau éducatif est original et pourrait donner des idées pour la gestion de GRID'5000.

5.2 Synergie au plan national

Nous décrivons successivement le projet DEISA, le projet E-TOILE et le projet COREGRID, en synthétisant au passage quelques éléments comparatifs.

5.2.1 Le projet intégré européen DEISA

Piloté par l'IDRIS, un consortium de sept centres de calcul nationaux s'est constitué, dans le but de déployer et opérer un super-calculateur réparti européen, qui disposera d'une capacité de calcul globale de quelques dizaines de Téraflops.

L'idée "est de donner une image unique à un ensemble de supercalculateurs. Cette image unique inclura, non seulement la capacité d'accéder de manière transparente à la totalité des ressources de calcul, mais aussi une gestion globale des données, par la mise en place des système de fichiers répartis à l'échelle continentale. Ainsi, une application s'exécutant sur un site pourra - si tout se passe comme prévu - utiliser de manière transparente et efficace des données qui se trouvent n'importe où ailleurs."

Sur le plan applicatif, DEISA vise "des projets ambitieux dans le domaine de la physique ou la chimie computationnelles".

Commentaires Clairement, le projet DEISA s'inscrit dans la problématique des grilles de production. Au plan national, il a davantage vocation à interagir avec le CINES et les méso-centres qu'avec GRID'5000.

5.2.2 Le projet RNTL e-Toile

Piloté par l'UREC du CNRS, le projet RNTL e-Toile réunit des laboratoires (projets Reso, Remap, Apache, Paris communs au CNRS et à l'INRIA, équipes du PRISM et de l'IBCP au CNRS), et des industriels (SUNlabs France, EDF, CEA, et France-Telecom R&D via VTHD).

Les objectifs scientifiques du projet sont :

- Agréger des ressources haute performance sur un réseau haut débit
- Développer de nouveaux composants logiciels, et les expérimenter
- Promouvoir le concept de grille active
- Tester et optimiser des applications a'forts besoins de calcul

Pour l'infrastructure matérielle, le projet n'utilise pas de ressources nouvelles mais s'appuie sur des ressources existantes, mises à disposition par les partenaires du projet. L'infrastructure réseau est celle de VTHD, qui a été étendu pour le projet.

Les intergiciels développés sont :

- Mémoire distribuée virtuelle (projet PARIS)
- Interface de communication Madeleine (LIP)

– Accès et transfert de fichier à haute performance (ID)
– Optimisation de l'Allocation de Ressources (PRISM)
– Tester et optimiser des applications a'forts besoins de calcul
– Monitoring et sécurité (CS-SI)

Les applications vont de la simulation numérique (EDF) à la physique à haute énergie (CEA) et à la bio-informatique (IBCP), en passant par l'optimisation combinatoire (PRISM).

Commentaires Sur le plan matériel, le projet E-TOILE utilise des ressources existantes non dédiées et connaît les difficultés inhérentes à cette approche (faible disponibilité, voire disparition des ressources). Sur le plan logiciel, le projet E-TOILE peut être vu comme un embryon du projet GRID'5000, en ce sens qu'il réunit plusieurs équipes pour développer des intergiciels pour la grille. Un consortium à plus grand échelle, et s'appuyant davantage sur les équipes du RTP Grilles, semble indispensable pour réussir dans cette voie ; il est clair que le projet GRID'5000 devra s'appuyer sur l'expérience du projet E-TOILE.

5.2.3 L'infrastructure de recherche EGEE

Le projet EGEE, *Enabling Grids for E-science and industry in Europe*, va être déposé comme infrastructure de recherche devant la communauté européenne. Les trois objectifs principaux sont les suivants :

– intégrer les développements technologiques européens liés à la grille
– établir une infrastructure de grille à l'échelle européenne pour la science et l'industrie, avec pour priorité la prise en compte de l'hétérogénéité et de l'inter-opérabilité
– permettre le déploiement d'applications scientifiques et industrielles sur la grille

Commentaires Le projet EGEE vise à la réalisation d'une infrastructure européenne, et GRID'5000 pourrait naturellement faire partie de cette infrastructure. Les objectifs applicatifs du projet sont très (trop ?) ambitieux : faire co-exister applications scientifiques novatrices et déploiements industriels semble prématuré.

5.2.4 Le réseau d'excellence européen CoreGrid

Le réseau d'excellence européen CoreGrid, piloté par l'INRIA, a pour intitulé complet : *CoreGRID : Foundations, Software Infrastructures and Applications for large scale distributed, Grid and Peer-to-Peer Technologies.*

Le but est de rassembler les meilleures équipes européennes actives en calcul distribué à grande échelle (grille et pair-à-pair) afin de conduire un effort de recherche coordonné dans ce domaine. Sur le plan français, les principales équipes apportent leur contribution. Les activités du consortium doivent inclure notamment :

- Architectures logicielles pour les plates-formes distribuées à grande échelle
- Infrastructures pair-à-pair (sécurité, accounting, confidentialité)
- Intergiciels adaptatifs pour les systèmes de grille et P2P
- Systèmes d'exploitation "orientés grille"
- Algorithmes distribués (systèmes de nommage, groupes)
- Modèles de programmation
- Mécanismes d'auto-configuration et d'auto-management
- Fouille de données à grande échelle

Commentaires Le projet CoreGrid inscrit dans une perspective européenne la problématique de recherche liée à GRID'5000. La disponibilité de la plate-forme est essentielle pour permettre à la communauté française de jouer pleinement son rôle.

6 Programme scientifique détaillé

Nous présentons ici en détail toutes les recherches que la plate-forme GRID'500 permettrait de mener aux différentes communautés scientifiques impliquées.

6.1 Programme scientifique pour la communauté *Grille*

Voici la liste des laboratoires (et des responsables de projets ou d'équipes) du RTP Grilles :

- OASIS/I3S Nice (I. Attali, D. Caromel)
- LIFL Lille (J-M. Geib, E.G. Talbi)

– LaBRI Bordeaux (R. Namyst, J. Roman)
– LIB Besançon (H. Guyennet, L. Philippe))
– LRI Orsay (F. Cappello)
– LORIA Nancy (E. Jeannot)
– IRISA Rennes (L. Bougé, T. Priol)
– ID Grenoble (J-F. Méhaut, B. Plateau)
– LIP Lyon (F. Desprez, P. Primet)
– PRISM Versailles (V-D. Cung)
– IRIT/CERFACS Toulouse (M. Dayde)

Toutes ces équipes ont exprimé leur volonté affirmée de participer au projet GRID'5000. Pour chacune d'entre elle, l'accès souple et interactif à une plate-forme expérimentale est vitale, pour valider leurs développements logiciels. **Si la recherche française veut avoir un impact sur les standards logiciels, elle ne peut pas ne présenter que des résultats de simulation !**

Les problématiques scientifiques relevant du calcul sur grilles informatiques sont bien connues. Il s'agit essentiellement de faire communiquer et coopérer des matériels distants et hétérogènes (par leurs modes de fonctionnement comme par leurs performances), de créer les logiciels permettant de gérer et distribuer efficacement les ressources disponibles, et de concevoir des outils de programmation adaptés au caractère distribué et parallèle de l'exécution.

Les travaux expérimentaux qui accompagnent ces problématiques peuvent se diviser en trois domaines :
– Expérimentations algorithmiques en vraie grandeur
– Déploiement d'intergiciels et de composants
– Emulation de systèmes pair-à-pair à très grande échelle

6.1.1 Algorithmes

Les principaux verrous scientifiques et technologiques sont les suivants :
– Adaptativité des logiciels à la dynamicité des plates-formes (même si on ne va pas jusqu'à l'exploitation de jachères de calcul)
– Monitoring des processeurs et des liens de communication en temps réel (la plate-forme ne sera jamais dédiée à une application à un instant donné)
– Gestion de l'hétérogénéité, à tous les niveaux : tout est hétérogène, des machines elles-mêmes aux réseaux, en passant par les systèmes

d'exploitation
– Modélisation de la plate-forme pour le passage à l'échelle : les algorithmes et heuristiques doivent pouvoir fonctionner pour un très grand nombre de machines avec des réseaux très hiérarchiques

Pour donner une illustration concrète, prenons l'exemple d'un simple produit de deux matrices, composant de base des nombreuses applications numériques. Sur une machine parallèle homogène classique, l'équi-répartition des données sur les processeurs donnera une solution optimale, en allouant des sous-carrés de données à chaque ressource de calcul. Sur une grappe hétérogène, décider quel sous-rectangle de données sera alloué à chaque processeur, avec la double contrainte d'un équilibrage des charges et d'une minimisation du volume de communications, est un problème combinatoire difficile (NP-complet). Sur une architecture hiérarchique (grappe de grappes), aucune solution n'est connue. Bien sûr on déborde ici du cadre strict d'un produit de matrices sur la grille : pour tout application parallèle, l'équilibrage de charge est bien plus compliqué qu'avant. Se pose aussi à l'algorithmicien le problème de la modélisation fine des communications avec partage des liens, et de réservation de bande passante.

Les difficultés précédentes sont seulement liées à l'hétérogénéité. Les aspects dynamiques viennent encore compliquer la donne : les ressources ne sont pas dédiées, leur charge évolue. Il faut disposer d'une "photographie" de la plate-forme en termes de capacités de calcul/mémoire/réseau. Pour l'instant, il n'y a pas de "standard" pour récupérer ces informations. Le développement d'un tel logiciel fait parti des recherches en cours (voir le paragraphe suivant).

On conçoit bien la difficulté algorithmique liée au déploiement d'applications complexes sur la grille, le besoin de concevoir des heuristiques d'ordonnancement distribué à très grande échelle, et la nécessité de les valider par expérimentation. Terminons par une question méthodologique : qu'est-ce qu'un benchmark ? Qu'est ce qu'une experience ? Comment planifier et réaliser une expérience reproductible ?

6.1.2 Intergiciels et composants

Les principaux défis sont les suivants :
– Assurer une communication efficace entre composants logiciels, notamment grâce à une bonne compatibilité entre les aspects parallèle et distribué de l'exécution, et à l'adaptation des protocoles de communication aux infrastructures hétérogènes et aux très hauts débits

- Optimiser la répartition des tâches, et ceci suivant différents modes, voire de manière dynamique : les exécutions pourront être équilibrées a priori ou non (par exemple, la tendance actuelle du calcul scientifique vers le multiéchelle multimodèle auto-adaptatif va dans le sens d'un besoin de redéploiements dynamiques de l'exécution
- Développer des outils de vérification formelle et de certification qui permettront de ne pas renoncer à une exécution sure
- Développer un véritable langage pour le calcul scientifique à grande échelle
- Développer des modèles de composants logiciels distribués permettant de construire des applications réparties sur les grilles (idéalement, ceci se ferait de la même manière que des applications à forte mobilité des utilisateurs et des codes).
- Réussir à faire vraiment coopérer un grand ensemble de machines pour étudier le comportement d'intergiciels pas vraiment stables (ex : pb de choix d'OS, de gestionnaires de batch, de politiques de sécurité)

Une solution "simple" aux problèmes de déploiement serait de disposer d'un logiciel de bas niveau qui se charge de la localisation des ressources, de leur allocation, des problèmes de sécurité (authentification, autorisation). Globus (surtout dans la version OGSA) n'est pas encore assez mature, et beaucoup trop limité pour la plupart des travaux développés. Un problème également avec Globus est l'aspect "boite noire" : il sera très difficile d'avoir un retour sur son comportement et ses choix. La communauté s'inscrit résolument dans la recherche d'une alternative européenne interne (projet de réseau d'excellence CoreGrid. Il ne s'agit évidemment pas de re-développer une panoplie logicielle complète *from scratch* mais de proposer un ensemble cohérent et performant qui unifie et étend des briques existantes. Le but final est d'assurer le contrôle d'un ensemble massif, hétérogène et fluctuant de ressources pour que celles-ci coopèrent à un même calcul (la dénomination calcul étant prise au sens large "processing").

6.1.3 Calcul pair à pair

La connexion à grande échelle des ordinateurs aux réseaux de communication et les mécanismes qui permettent leur interopérabilité motivent l'étude des systèmes Pair à Pair (P2P). Par rapport aux systèmes répartis classiques, ces nouveaux systèmes possèdent deux caractéristiques fondamentales : a) l'échelle du nombre de ressources (mille fois plus) et b) la com-

plexité des ressources qui sont toutes des ordinateurs capables d'exécuter des programmes complexes. Pour étudier scientifiquement ces systèmes et comprendre/contrôler les phénomènes nouveaux qui apparaissent, il est nécessaire d'associer trois techniques : l'expérimentation in-situ (sur les grilles et systèmes déployés), la simulation à partir de modèles mathématiques et l'expérimentation par un nouveau moyen dans le domaine des grilles : l'émulation.

En plus des "testbeds" réalistes (XtremWeb), et des simulateurs, il y a un besoin d'émulateurs de systèmes P2P à grande échelle permettant de conduire des expériences dans des conditions reproductibles. Plusieurs initiatives ont été lancées sur l'analyse théorique (tolérance aux pannes dans les systèmes P2P). mais la simulation et l'émulation à grande échelle restent pour le moment inexploitées. La notion de grande échelle commence à être étudiée par exemple dans le projet "Petascale Virtual Machine" à Oak Ridge National Laboratory : ce projet vise à l'étude d'algorithmes parallèles et distribués pour l'échelle de 100000 machines. Cette étude est étroitement liée à celle d'un simulateur fonctionnant sur un cluster Linux et capable de simuler 100000 noeuds.

Il s'agit de permettre de configurer virtuellement les noeuds de la plateforme de taille significative, pour en faire des *Grilles virtuelles* et tester leurs systèmes, algorithmes, applications dans des conditions expérimentales reproductibles (performance, sécurité, pannes, dynamicité). Cela permettrait aussi de dériver des simulateurs et des modèles théoriques réalistes. Pour le calcul P2P, l'objectif serait d'émuler des systèmes à 10K voire 100K noeuds. Il serait également possible d'émuler un réseau haut débit connectant 3 ou 4 machines parallèles à 128 et 256 processeurs (chacune recevant une charge propre réaliste).

L'intégration de l'émulateur dans la plate-forme GRID'5000 permettrait de combiner l'émulation et l'expérimentation dans des conditions réelles. Des sondes placées sur la plate-forme Grid permettraient d'alimenter l'émulateur en valeurs réalistes pour les paramètres d'émulation. Inversement des mécanismes développés et testés en émulation pourraient être évalués en conditions réalistes.

Un émulateur permet d'exécuter des expériences en utilisant une application réelle, dans des conditions reproductibles et un environnement contrôler finement. Il permet aussi, par virtualisation maîtrisée, de rendre compte de phénomène à grande échelle alors que les expériences sont menées à plus petite échelle. Les systèmes qui sont émulés sont généralement trop complexes pour qu'une analyse théorique permette leur étude. Ils sont aussi trop

dépendants de paramètres temporels et structurels fins pour que la simulation puisse rendre compte systématiquement de leur comportement. Leur observation in-situ comporte d'autres limitations (charge, configuration) qui empêchent la généralisation des résultats. L'objectif est de concevoir des mécanismes système permettant de reproduire, sur n noeuds de la plateforme d'émulation, le comportement de m noeuds d'un système P2P (avec $m \gg n$) avec une précision de reproduction connue. La dilatation temporelle devra être maîtrisée et contrôlée de sorte que les événements inter noeuds interviennent à des instants compatibles avec un fonctionnement réel.

Les systèmes à émuler seront des systèmes distribués à grande échelle comme Gnutella, Freenet, ou XtremWeb, ou des parties de ce type de systèmes, comme les mécanismes de recherche de ressource Pastry, Tapestry, CAN, Chord. Ce thème comporte deux aspects : (i) étudier les mécanismes nécessaire à la reproduction de conditions et (ii) la virtualisation.

Les conditions à reproduire concernent les caractéristiques matérielles, de volatilité et de charge des noeuds qui composent le système à émuler et le réseau. Il faut être capable de reproduire certains paramètres d'un environnement existant avec une précision connue. Ceci suppose la conception de sondes capables d'extraire ces caractéristiques et de mécanismes d'émulation capables d'appliquer ces caractéristiques sur des ressources génériques. Les sondes de volatilité et de charge existent déjà pour les noeuds du système dans XtremWeb. Les sondes concernant le réseau sont à étudier. Le système NWS pourra être pris comme référence pour cette partie. Les mécanismes d'émulation sont à étudier. Il s'agit de reproduire les caractéristiques des noeuds (taille mémoire - inférieure à celle de la machine d'émulation, vitesse CPU - inférieure à celle de la machine d'émulation) et du réseau (débit, latence, contention, etc.) de façon logicielle et d'être capable connaître la précision de reproduction.

La virtualisation maîtrisée consiste à émuler sur un nombre de ressources limité (100 noeuds), un système composé de plusieurs milliers de noeuds. Comme il s'agit d'exécuter l'application complète, la virtualisation se traduit par une dilatation temporelle : la durée d'exécution est supérieure à celle du système dans un environnement réel. La virtualisation maîtrisée consiste à contrôler de façon précise le temps des événements dans l'environnement virtuelle (temps virtuel). La recherche étudiera les mécanismes à mettre en oeuvre pour imposer la temporalité des événements dans l'environnement virtuel afin qu'ils interviennent à des instants compatibles (interdépendances) et crédibles (performances) relativement à un environnement réel.

6.2 Programme scientifique pour la communauté *Réseaux*

6.2.1 Introduction

Les plates-formes de recherche et d'expérimentation en réseaux ont joué un rôle crucial dans le développement de l'Internet. Une politique active de déploiement de telles plates-formes a été conduite aux Etats-Unis depuis l'émergence de l'Internet et se poursuit aujourd'hui (décision prise en Février 2003) avec un investissement de 10M$ de la NSF afin de prolonger cette action dans la suite d'Internet 2. Ces plates-formes ont permis très tôt aux chercheurs américains de tester, valider, expérimenter leurs idées avant de les transférer dans des contextes opérationnels, mais aussi d'identifier de nouvelles voies de recherche. Elles ont largement permis de consolider la suprématie des principaux acteurs industriels. Cette vision a intégrée la nécessité de l'expérimentation de manière identique à la démarche reconnue dans de nombreuses autres disciplines.

6.2.2 Intérêt scientifique

Le domaine des réseaux poursuit son expansion avec l'émergence de nouvelles technologies et l'expression de nouveaux besoins. De manière analogue à l'expérimentation conduite sur Internet afin d'en valider les fondements, il faudra disposer de plates-formes pour ces environnements. Cependant, les problèmes adressés demandent des plates-formes d'un type nouveau. Les principales solutions développées dans le passé (vBNS ou Internet 2 par exemple) ont ciblé essentiellement des besoins de réseaux grande distance, plus performants. Aujourd'hui, des formes nouvelles sont nécessaires pour aborder les défis futurs. Ce sont par exemple, des systèmes tels que les réseaux spontanés, les réseaux de capteurs (la NSF vient de lancer un programme de 35M$ sur ce thème), les réseaux "overlays", les réseaux ambiants qui sont autant de solutions capables de modifier profondément notre vision sur la manière de construire des réseaux. De même, ces plates-formes se diversifient dans leur forme afin d'atteindre des objectifs de facteur d'échelle par exemple ou de reconfiguration plus dynamique du système. Ce constat conduit à des solutions reposant sur des " clusters " partagés (Emulab : `http://www.emulab.net`) ou un grand émulateur réseau (PlanetLab : `http://www.planet-lab.org`). Ces deux plates-formes sont essentiellement utilisées pour l'expérimentation de protocoles réseaux. De même, le périmètre de la grille s'élargit régulièrement pour rencontrer des problématiques spécifiquement

réseaux (Qualité de service, multipoint, mobilité, reconfiguration, etc.). De fait, une mutualisation de plate-forme à usage des chercheurs des communautés *Réseaux* et *Grille* est envisageable. En particulier, les aspects émulation de réseaux, métrologie ou mobilité constituent des exigences pour l'expérimentation en réseau, mais que l'on retrouve très souvent comme besoin pour la grille.

En résumé, de nouveaux besoins émergent dans le domaine des télécommunications et suscitent une transformation des réseaux expérimentaux actuels vers des plates-formes originales : grappes, réseaux "overlay", réseaux de capteurs, réseaux ambiant ou fédération de certains de ces réseaux. De nombreuses similitudes existent avec la grille informatique.

Enfin, la plate-forme GRID'5000 pourrait être vue comme un grand émulateur de l'Internet : dans ce cas, les noeuds de calcul seront utilisés soit comme de vrais clients ou serveurs, soit comme des systèmes autonomes de l'Internet (routeurs logiciels émulant latence et pertes de paquets). Les capacités de reconfiguration dynamique d'un tel environnement seraient une véritable innovation.

6.2.3 Contexte international

Les plates-formes ont toujours été très présentes aux Etats-Unis dans le domaine des réseaux informatiques. Cela a été déterminant pour établir le rôle de leader de l'industrie américaine dans ce domaine. L'expérimentation fait partie de la culture de la recherche Internet aux Etats-Unis, depuis l'émergence des fondements de ce réseau, dont le développement expérimental a été fortement soutenu initialement par la DARPA puis la NSF. Depuis, il y a toujours eu des plates-formes en réseaux aux Etats-Unis, fournissant les infrastructures expérimentales des réseaux opérationnels à venir. Arpanet est l'ancêtre des réseaux expérimentaux, produit de la recherche sur les réseaux à commutation de paquets, qui a donné naissance à une infrastructure commerciale. Dartnet fut construit en 1991, en réponse au succès commercial d'Arpanet, qui ne permettait plus des expérimentations mettant en cause la stabilité du réseau. Des résultats remarquables furent produit directement ou indirectement : IP multicast, les protocoles ST-II et RSVP, le mécanismesd'ordonnancement CBQ, les applications de vidéoconférence vic, vat, sdr, etc. Dartnet se termina en 1996 et contribua largement à former une culture de coopération dans cette communauté scientifique. Le pro-

gramme MAGIC (Gigabit Testbed) fut soutenu par le DARPA de 1992 à 1999. Il cibla plus particulièrement les réseaux à très haut débit et les applications distribuées temps-réel et interactives. Plusieurs plates-formes ont été développées dans le cadre du programme NGI (Next generation Internet) de 1994 à 2000. Les problèmes abordés étaient concentrés sur les couches basses (transmission) de l'architecture. De nombreuses technologies ont été ainsi développées et déployées telles que IP/WDM ou les réseaux optiques.

Emulab, cité précédemment inaugure une nouvelle génération de plateforme. Il s'agit d'une grappe de 168 machines concentrée sur un site, permettant à une communauté de se partager dynamiquement ces ressources d'émulation de réseau (il est possible de modifier des mécanismes et protocoles sans les contraintes de développement réel et avec un plus grand réalisme qu'une simulation).

Les plates-formes de recherche ont joué un rôle majeur dans le processus d'innovation et de valorisation aux Etats-Unis. Une politique de soutien continue est menée sur ce thème, de la création d'Arpanet (ancêtre d'Internet) à nos jours (budget de 10M$ alloué par la NSF en Février 2003).

La situation a toujours été plus confuse en Europe dans la mesure où la prise de conscience de l'intérêt à accorder aux plates-formes est arrivée tardivement et, leur développement souvent contraint par le cadre de financement de durée réduite des programmes de recherche. Citons pour illustration les réseaux de la recherche européen TEN-155 puis GEANT qui ont fourni une base pour l'expérimentation des projets IST. Chaque pays européen a adopté des choix allant d'un soutien faible à inexistant jusqu'à une action active (par exemple SuperJanet au Royaume-Uni). Notons que de nombreux laboratoires et chercheurs européens ont développé cette culture de lien entre théorie et expérimentation. Ils sont en grande partie organisés en réseaux (ENET : réseau européen IST, ENEXT réseau d'excellence FP6) avec une forte représentation française.

En France, le soutien aux plates-formes a démarré à l'initiative de quelques chercheurs du CNRS et de l'INRIA avec la plate-forme MIRIHADE en 1996 (LIP6-LAAS-INRIA Sophia). Depuis, le relais a été pris avec des plates-formes fournies dans RENATER (@irs/@irs++ par exemple) ou par France Telecom (VTHD/VTHD++), l'ensemble étant coordonné dans le cadre du RNRT. A nouveau, le problème de la pérennité de ces solutions se pose de manière aiguë et justifie un effort spécifique dans cette direction.

Les plates-formes de recherche en réseau en France sont assez

récentes (Mirihade, 1996). Un effort significatif a été développé par le RNRT récemment (@irs, VTHD) mais la pérennité de ces plates-formes est largement compromise.

6.3 Programme scientifique pour les communautés applicatives

6.3.1 Applications multi-paramétriques

Un premier bénéfice pour la communauté applicative est la disponibilité et la souplesse de la grille expérimentale. Il est actuellement impossible de mobiliser une fraction significative de la mémoire d'un grand centre sur une seule application sans délais d'attente prohibitifs. Il est également impossible de créer un fichier temporaire de quelques Téra-octets, ou de lancer un job durant plusieurs jours. La grille expérimentale pourrait permettre le déploiement rapide d'applications utilisateurs multi-paramétriques. Le passage à l'échelle devrait se contenter de développements assez simples sur les outils grille, sans impliquer de grosses contraintes côté applications ou utilisateurs. Le "rendement" obtenu sera très significatif. Avec 5000 processeurs à 2 Ghz, la puissance théorique de la grille atteint 10 Gflops, soit 5 fois environ le total de la puissance disponible au CINES et à l'IDRIS.

Le bénéfice immédiat serait ainsi de "rentabiliser" la grille pour des applications scientifiques très variées hors STIC. Le multi-paramétrique couvre des champs immenses, partout où il faut faire tourner un modèle un peu lourd pour explorer des espaces de paramètres multi-dimensionnels. Cela peut concerner la chimie ou l'astrochimie (surfaces de potentiel, chemins réactifs, criblage de modèles de composés ou de médicaments, etc), l'astrophysique (fit de modèles sur les centaines de milliers d'objets, abaques de modèles pour la cosmologie, etc), la physique des particules, etc.

Dans un second temps, on peut imaginer des applications faiblement couplées, avec des approches du type décomposition hiérarchiques de domaines. Cela peut s'imaginer pour des applications régulières (mécanique des fluides, hydrodynamique, magnétohydrodynamique...) ou pour des modélisations particulaires (agrégation de particules, comme par exemple dans les modèles de formation des planétésimaux dans un disque protoplanétaire).

Ces expérimentations de première et deuxième génération apporteront certainement aussi une stimulation pour le développement d'outils grille d'usage général. A titre d'exemple, les expérimentations menées par les as-

trophysiciens dans le cadre de la grille grenobloise CIMENT à Grenoble ont mis en évidence deux besoins : (i) un ordonnanceur de boucle hétérogène irrégulière distribuée, et (ii) un outil pour découvrir automatiquement la connexité des espaces disques. Un protopype efficace a été mis au point sur quelques dizaines de noeuds, mais le passage à l'échelle représente un problème de recherche en soi.

Les exemples précédents illustrent également l'émergence d'un besoin d'applications *auto-adaptatives* capables de s'adapter à des conditions irrégulières et à une topologie matérielle non explicitée. De telles applications auto-adaptatives pourraient également fournir un champ d'application très riche aux développements STIC.

Pour toutes ces applications, nous anticipons la coexistence possible d'une sous-grille *interactive* et d'une ou plusieurs sous-grilles *utilisateurs batch* avec un partitionnement dynamique, en généralisant par exemple les expérimentations grenobloises.

6.3.2 Vers des applications innovantes délocalisées sur la grande grille ?

Il serait séduisant de pouvoir également déployer des applications utilisateur "natives" sur la plate-forme GRID'5000 entière, qui ne se limitent pas à un scaling d'applications existantes. Il y a trop d'inconnues aujourd'hui pour pouvoir dresser un inventaire, et le meilleure stratégie pourrait être de lancer un appel à idées national pour identifier rapidement quelques équipes porteuses d'un projet et désireuses de tenter l'aventure, puis de renouveler périodiquement (annuellement par exemple) cet appel à idées en le couplant avec l'un des workshops de retour d'expérience technique et scientifique sur la grande grille.

Les spécificités majeures de la "grande grille" sont son caractère fortement hiérarchisé, une latence point à point toujours inférieure à une fraction de seconde, une puissance de traitement considérable, et la possibilité de faire des points de reprise par entrées-sorties externes pour des traitement de très longue durée. Les caractéristiques ci-dessus semblent compatibles avec des modélisations cognitives, soit bottom-up en couplant des réseaux de neurones pour expérimenter sur des hiérarchies d'assemblées de neurones "à la Jean Pierre Changeux", soit top-down avec la modélisation de processus cognitifs d'apprentissage sur des modèles plus globaux. Et la puissance globale semble correspondre à la capacité d'échanges bruts d'informations au niveau d'un

fragment de cortex (simulation temps réel que quelques millions à dizaines de millions de neurones avec toutes leurs synapses). Une telle mise en oeuvre représenterait une véritable rupture par rapport à ce qui s'est fait jusqu'à présent.

6.3.3 Gestion des données

Il ne faut pas oublier que les applications des grilles ne sont pas uniquement tournées vers le calcul. La gestion de données distribuées est d'une première importance. On va aussi certainement devoir travailler sur des algorithmes adaptés à de nouvelles applications (comme par exemple la recherche de motifs en parallèle pour la bioinformatique).

L'espace disque qui sera disponible sur la plate-forme GRID'5000, et sa gestion, sont des facteurs très importants pour le dimensionnement de chacune des grappes. Une seule application utilisateur, en astrophysique, en chimie, en mécanique des fluides, etc, capable de dévorer 100 000 heures CPU par jour, brasse forcément beaucoup de données, en entrée, en espace temporaire, et en sortie. Les besoins se chiffreront en dizaines de Téraoctets par application au minimum. Il faudra déployer une fraction de cet espace sur la grille pour chaque run, et identifier pour chaque application un site de référence pour le stockage. **Sans stockage il n'y aura pas ou peu d'ouverture ambitieuse hors STIC.**