A. Domaine d'étude en informatique

1. Un observatoire de grille.

Une grille de calcul et de stockage combine un ensemble de composants, réseaux, processeurs et accès aux données, qui sont des systèmes complexes au sens technique du terme. La très grande taille d'un tel système distribué s'oppose à une modélisation purement analytique ; sa topologie et son état à un instant donné ne peuvent être qu'approximés ; son utilisation en production conduit à l'apparition de conditions de charge non contrôlées et parfois imprévues. La grille constitue alors un objet d'étude en tant que phénomène, certes artificiel, mais dont les lois de comportement ne sont pas connues, avec deux objectifs, respectivement à court et moyen terme.

- En gestion de production, en particulier du point de vue de la détection/récupération des erreurs, et plus généralement l'étude de la dynamique de l'usage des ressources.
- En conception, en vue d'une modélisation qui fournisse des hypothèses réalistes pour l'étude des systèmes distribués à très grande échelle, la recherche de politiques optimales et l'algorithmique applicative sur ces systèmes.

Cette activité a une complémentarité évidente avec la collaboration GRID'5000, et vise à interagir fortement avec elle.

1.1. Monitoring et recueil de traces

Fonder ces recherche renvoie à la nécessité d'études expérimentales qui, pour être fructueuses, doivent se situer à la même échelle que l'objet étudié. La mesure sur les systèmes complexes constitue une problématique scientifique en elle-même (métrologie). D'autre part, la complexité de la conception et du déploiement effectif d'un système de recueil de traces sur une grille implique de disposer, non seulement d'un accès à une grille, mais aussi d'une collaboration forte avec ses institutions gestionnaires qui permette de bâtir sur l'existant, qui est remarquablement riche.

L'objectif de cette activité est le déploiement d'une *base de traces* de fonctionnement de la grille basée sur l'enregistrement des évènements d'EGEE, accessible à la communauté (par exemple à travers un portail) et interopérable avec d'une part les environnements d'émulation/simulation classiques (par exemple les outils apparentés à Ganglia, comme gangSim), d'autre part les outils nouveaux développés dans Grid'5000. Cette base visera à la fois à fournir des données brutes correspondant aux problématiques classiques, en particulier celles liées à la localisation de données (placement, réplication, localité, transferts), d'autre part des interfaces vers des outils d'analyse statistique et de fouille de données.

Il faut ici décrire très brièvement l'existant, afin de préciser clairement l'objectif de l'activité « monitoring et recueil de traces ». On trouvera des informations plus détaillées dans [Coo05, Byr04,Coo04]. Les projets DataGrid, puis EGEE, ont été *structurellement* confrontés à une question extrêmement difficile : réaliser un système d'information distribué à très grande échelle. Le point crucial est que cette question conditionne la faisabilité même de l'intergiciel de base de la grille : du placement/ordonnancement des tâches au suivi d'exécution, les décisions ne peuvent être prises que sur la base d'un système d'information qui, à l'échelle de la grille, ne peut être que décentralisé, doit être extensible, et doit supporter l'asynchronisme. Ce système d'information doit en outre couvrir une échelle temporelle d'information qui va de données complètement statiques (capacité des ressources, par exemple en mémoire ou stockage), à des données très dynamiques (état d'avancement d'exécution d'un job), en passant par l'état des files d'attentes, la localisation des réplicas, etc..

Datagrid/EGEE ont développé R-GMA, une *architecture intégrée d'information et de monitoring* de la grille. Cet environnement implémente la proposition de standard GMA (Grid Monitoring Architecture) du GGF, qui modélise l'infrastructure d'information de la grille comme un système de médiation. Les choix d'EGEE portent sur un modèle de données relationnel, un système de médiation de type local-as-view, la prise en compte prioritaire de flots (streams), enfin une implémentation robuste de la détection des défaillances. L'extensibilité du système repose sur le déploiement d'agents autonomes d'intégration, qui fusionnent continûment des données locales, sans interdire l'accès à ces données si c'est souhaitable.

La flexibilité du modèle relationnel, la généricité de l'architecture, et l'extensibilité de l'implémentation, ont permis de déployer des systèmes d'information correspondant à des objectifs très variés : l'infrastructure d'information pour les fonctionnalités de base du middleware (par exemple pour le scheduling), aussi bien que des environnements de monitoring spécifiques à des applications ou de

recherche, en passant par des vues spécialisées ou synthétiques pour la gestion opérationnelle (par exemple GridIce)

Les points les plus difficiles dans le contexte de ce projet sont :

- La réalisation de vues associées aux comportements correspondant à des objectifs nouveaux. Il s'agit par exemple de la représentation du couplage entre localisation des données et du calcul, et de l'exécution des workflows complexes, de la détection d'intrusions.
- La prise de mesure elle-même dans le domaine des composantes réseau, et son association aux entités considérés juste avant. L'implantation sur un grand nombre de sites d'une infrastructure contrôlée dans le cadre d'un projet global permet de recueillir des traces d'utilisation dans un environnement réel et particulièrement demandeur pour le réseau.

Par rapport à ces questions, il est souhaitable d'explorer de façon approfondie dans quelle mesure les systèmes d'information déjà existants peuvent ou non répondre aux demandes nouvelles. Ceci pourrait contribuer à identifier d'éventuels points bloquants en particulier au niveau des requêtes à grain fin. Ces demandes en provenance de la communauté informatique constitueraient de nouveaux « use cases » pour le projet EGEE, et pourraient ainsi influer sur les développements futurs. Mais il faut aussi souligner que l'architecture R-GMA est organisée pour permettre le déploiement de systèmes d'information nouveaux.

Cette activité recherchera une collaboration forte avec le NoE CoreGrid, en particulier son *Institute on Grid Information and Monitoring Services.*

1.2. Fouille des données de la grille

Objectif: Modéliser le comportement de la grille par l'analyse des données rassemblées par l'observatoire.

La complexité des composants individuels de la grille, de leurs interactions, et la structure fondamentalement décentralisée de l'utilisation de la grille, rendent réalistes l'application d'un modèle de modélisation/optimisation de fonctionnement par découverte de connaissances et expérimentation. Les modèles et/ou indicateurs de performance ainsi définis contribueront à fournir les entrées nécessaires pour l'évaluation du passage à l'échelle dans un ensemble de domaines: placement-ordonnancement, mécanismes et environnements de distribution/réplication des données, gestion de travaux hétérogènes, algorithmique applicative,…

Le développement de ces méthodes nouvelles d'analyse des données demande un dialogue pluridisciplinaire approfondi, qui intégrera :

- du côté de la compétence STIC/fouilles de données, un noyau théorique et algorithmique commun avec des applications présentées ailleurs, au confluent de l'apprentissage statistique, de la fouille de données et de l'optimisation non-linéaire et/ou stochastique. Cette approche est décrite plus en détail dans la section 3.
- du côté de la compétence STIC/systèmes parallèles et distribués, l'expertise dans la modélisation et l'évaluation des systèmes distribués, et l'identification des paramètres spécifiques du passage à l'échelle.
- du côté de l'IN2P3, d'une part l'expertise en acquisition et gestion de grandes masses de données, qui est un acquis collectif; d'autre part, la compréhension de la complexité des questions opérationnelles, qui sont essentielles pour une réalisation efficace.

2. Accès et traitement des données

Au dessus des fonctions de base de l'intergiciel de grille, il faut construire le niveau d'intergiciel correspondant à des fonctionnalités plus avancées. Il est important de noter que, si l'intergiciel de base d'EGEE sera probablement, à l'échéance du projet, relativement stable, à l'inverse dans ces domaines, des impacts importants peuvent être attendus de l'activité du projet, au niveau d'EGEE directement, et au delà (standardisation).

Dans le domaine de l'accès et du traitement des données, on considérera les aspects suivants.

2.1. Workflows

La transition depuis le modèle d'exploitation orienté multi-séquentiel simple et purement par lots (*batch*), vers un modèle généraliste, incluant la prise en compte de *workflows* complexes et dépendants des données, émerge comme demande d'applications variées.

Cette question est particulièrement brulante actuellement, dans trois types de communautés. Il est d'ailleurs remarquable de constater qu'elles désignent ce qui est fondamentalement une problématique commune par des termes différents : *workflows* dans le monde industriel et dans la communauté systèmes parallèles et répartis, *dataflows* dans celle du web sémantique et *embarassingly parallel problems* dans celle des utilisateurs de grille. Quelques chiffres montrent la nécessité d'une grille de données en production pour la pertinence des évaluations en passage à l'échelle : un gros test de charge (expérience ATLAS) correspond à 95000 tâches et produit près de 22 To de données ; en recherche d'imagerie médicale, une image peut représenter environ une centaine de Mo, et les études statistiques ou épidémiologiques peuvent consommer un nombre virtuellement illimité de telles images.

La mise en oeuvre de worfklows sur la grille soulève la double problématique de (1) l'exécution optimale de tels workflows et (2) des techniques d'interactions des composants logiciels les composant.

1. La communauté des systèmes parallèles et distribués a réalisé de nombreux travaux portant sur l'ordonnancement de workflow qu'il convient d'intégrer et d'adapter en fonction des hypothèses propres à une infrastructure de grille de calcul/données. Schématiquement, il s'agit de passer d'un modèle de programmation, où la granularité est celle des données/instructions/tâches, à un modèle de production où la granularité se situe au niveau fichiers/services/jobs. Dans cette configuration, le gestionnaire de workflow doit être capable de réaliser un ordonnancement en prenant en compte les données à traiter, celles-ci représentant souvent un potentiel de parallélisme bien supérieur à celui du workflow lui même. Le workflow ne peut donc pas être mis en correspondance avec les ressources en fonction de sa structure seule mais en fonction des données à traitées également. En outre, le coût de transfert des données doit être finement évalué en regard du coût de calcul, la collection de traces d'exécution pouvant permettre l'estimation de ce coût sans intervention de l'utilisateur. Enfin, l'allocation des ressources doit être réalisée de manière optimale sous la contrainte que l'état de la grille à un instant donné, et en particulier le nombre de ressources disponibles, ne peut pas être connu de manière déterministe.
2. Le modèle des services web émergeant actuellement comme standard pour l'accès aux services de grille est à la fois trop simpliste et incapable de manipuler de grands volumes de données. Les tâches individuelles d'un workflow sont mal représentées dans ce modèle qui devra évoluer pour permettre l'enchaînement efficace des tâches de calcul (en limitant au maximum les échanges de données) et l'intégration d'appel à des tâches complexes pouvant nécessité un couplage fort et des performances d'exécution élevées en raison de leur granularité.

2.2. Modèles et mécanismes d'accès aux données

Nous avons pour objectif le stockage et la manipulation de gros volumes de données produits par les applications scientifiques. Outre les problèmes techniques liés à l'accès à des serveurs de données très différents suivant le domaine applicatif et rarement adaptés à une distribution à grande échelle, le volume des données à gérer conduit à des problèmes d'indexation et d'accessibilité pour des utilisateurs finaux qui, n'ayant pas nécessairement connaissance de la totalité des richesses des ensembles de données mis à leur disposition par une communauté virtuelle de grande taille, ont besoin d'outils efficace de parcourt et de recherche de données.

En outre, les données manipulées dans de nombreux domaines scientifiques ont une structuration complexe et souvent des représentation hétérogènes en fonction des sites et des conditions d'acquisition. L'utilisation de méta-données pour l'interprétation et la formalisation du contenu des données devient centrale. Si en pratique les noms de fichier ont souvent servi de méta-données pour les données contenues, cette approche, déjà peu satisfaisante, devient tout simplement irréaliste sur une grille de données en raison de la nécessité de traiter avec des formats hétérogènes, du besoin d'utilisation de ces méta-données pour l'indexation et la recherche des données dans de vastes ensembles distribués, et des très grandes communautés virtuelles d'utilisateurs qui accèdent aux données de la grille.

Seules des bases de données peuvent répondre aux besoins évoqués ci-dessus. La répartition des données sur une grille conduit immédiatement à s'interroger sur la structuration des méta-données associées et aux moyens de gérer à la fois la cohérence et la répartition des méta-données. Les thématiques principales que nous souhaiterions aborder sont

- L'intégration des fonctionnalités des bases de données (structuration de l'information, expressivité, typage, logiques sous-jacentes pour l'interrogation, optimisation logique) avec les contraintes et les traditions des applications scientifiques effectivement productrices de grandes masses de données, qui sont en général fondées sur des systèmes de fichiers (distribués dans le cas des grilles) couplées avec des bases de données relativement simples et centralisées. Cette thématique correspond largement à celle définie dans l'ACI Masse de Données *Gedeon (voir ci-dessous, section organisation des données)*.
- La conception d'interfaces d'interrogation intelligentes. Sans atteindre le volume d'information accessible sur le Web, les ordres de grandeur des données accessibles sur une grille peuvent demander d'aller au-delà des fonctionnalités d'interrogation classiques, en incluant une aide à l'utilisateur par une définition implicite ou explicite de ses contextes d'intérêt. Cette thématique correspond à celle de la personnalisation de l'information, (RTP personnalisation de l'information, http://www.prism.uvsq.fr/recherche/themes/sial/cnrs/ et ACI Masses de données APMD http://apmd.prism.uvsq.fr/)
- L'intégration des modes d'accès de type web sémantique sur les infrastructures de grille, permettant la prise en compte de la connaissance que l'utilisateur a du domaine applicatif pour améliorer sa recherche de données.

1.2.1. Organisation des données

Le volume de données augmentant plus vite que les capacités matérielles de stockage, les infrastructures de gestion de données sont de plus en plus distribuées à grande échelle. Ces « grilles de données » sont majoritairement utilisées dans les communautés effectuant de gros calculs qui manipulent des volumes de données très importants.

Pour gérer efficacement ces infrastructures, les systèmes de gestion de fichiers (SGF) traditionnels ont été revisités et des réponses pour la gestion de la répartition ont été apportées. Cependant l'utilisation de ces systèmes est contraignante et ils restent inadaptés à la gestion de grandes masses de données, par exemple les fonctions d'interrogation restent limitées, la duplication des données et le maintien de leur cohérence sont également souvent rudimentaires. Les systèmes de gestion de bases de données (SGDB) offrent eux divers niveaux d'abstraction, des langages de définition et de manipulation de données permettant des interrogations fines des données via les méta-données qui leurs sont associés. Cependant la structuration forte des données qu'ils imposent généralement et leur architecture souvent monolithique et peu répartis restreignent leur utilisation pour des applications de type calcul scientifique.

Nous voulons fusionner des fonctionnalités des SGFs et des SGBDs dans le but d'obtenir un système hybride gérant des grandes masses de données répartis sur une grille et ayant des fonctions élaboré d'interrogation. Nous voulons plus particulièrement nous focaliser sur la gestion de la cohérence des données et des méta-données afin d'isoler le degré nécessaire aux applications visées dans le but de proposer un système de gestion avancée de données dupliquées haute performance. Cette recherche est menée essentiellement dans le cadre de l'ACI Masses de Données Gédéon.

Le projet actuel permettra de fournir à Gédéon des informations réelles sur les utilisations de données dans des environnements de production. Cette caractérisation est importante car elle permettra d'orienter des choix architecturaux faits dans Gédéon.

3. Fouille de données scientifiques

3.1. Introduction

Il importe de noter que le système complexe constitué par la grille, d'une part, et d'autre part les systèmes complexes étudiés en physique (e.g. l'expérience Auger), en biologie (e.g. étude du comportement dynamique des macro-molécules biologiques) ou en médecine (e.g. analyse d'images médicales pour le dépistage ou le suivi d'évolution de tumeurs), présentent des caractéristiques similaires et des structures comparables.

En particulier, ces problématiques ne disposent pas d'une *vérité terrain*, interprétation exacte unique qui puisse servir de référence pour calibrer les algorithmes d'analyse, de fouille et d'interprétation des données. Le problème devient en conséquence de déterminer l'espace de recherche (classes d'interprétations - problème de sélection de modèles), les critères d'intérêt (qu'est-ce qu'une

interprétation intéressante), et un mode d'interaction pertinent (trouver un compromis entre le temps nécessaire pour proposer une solution, et la qualité de cette solution, qui permette à l'expert d'entretenir un dialogue satisfaisant avec le système - problématique des algorithmes any-time).

Un aspect fortement innovant du projet {\em GAIA} concerne ainsi la réalisation de fonctionnalités de modélisation, également pertinentes pour les utilisateurs {\em et pour les administrateurs} de la grille.

Nous situerons brièvement les dimensions stratégiques de la recherche proposée (applicative, fondamentale et prospective), avant de décrire de manière détaillée les fonctionnalités quiseront étudiées et réalisées dans le projet.

Dimension applicative.

Au delà des algorithmes répondant aux problématiques déjà identifiées par les partenaires, qui seront détaillées ci-dessous, le projet {\em GAIA} cherchera à mettre à la disposition de la communauté des utilisateurs un *Centre d'Expertise*, en s'inspirant d'initiatives comparables au niveau européen.

Ce Centre d'Expertise aura pour fonction de diriger efficacement les utilisateurs vers les experts et les approches les plus appropriées du domaine de l'apprentissage statistique et de la fouille de données.

Il favorisera également la veille et la réactivité scientifique, en facilitant la détection des grandes tendances et les évolutions des applications considérées.

Dimension fondamentale

Un axe de recherche fondamentale, en coopération avec le réseau d'excellence PASCAL (http://www.pascal-network.org), s'intéresse aux impacts du modèle de calcul et de stockage proposé par les grilles sur l'apprentissage statistique et la fouille de données, du point de vue d'une part du passage à l'échelle des algorithmes, d'autre part des opportunités nouvelles apportées par une puissance de calcul virtuellement illimitée.

- Estimation : sélection de modèles, caractérisation des distributions des données, étude de convergence;
- algorithmes any-time : intervalles de confiance, convergence en horizon fini
- apprentissage incrémental (on-line) : rafraîchissement des résultats en fonction de l'évolution des données
- détection d'anomalies/ruptures.

Les avancées fondamentales à rechercher concernent:

- La confrontation des bornes statistiques existantes au niveau non-asymptotique (bornes généralement estimées trop conservatives) et asymptotiques (estimées trop optimistes), avec la convergence empirique observée sur des données de très grande taille.\\
- La qualification des algorithmes existants en fonction des paramètres d'ordre des problèmes (volume, distribution des données, critères de spécificité ou de sensibilité cherchés), en vue de déterminer les algorithmes les plus performants dans une région des paramètres d'ordre donnée.\\
- L'étude théorique de l'apport des portages de méthodes sur des architectures très distribuées, prenant éventuellement en compte le type de distribution des données pour proposer des bornes affinées.

Dimension prospective

Les données générées et traitées dans le cadre des sciences physiques se fondent majoritairement sur des structures spatio-temporelles. Cependant, l'exemple de la bio-informatique et de la médecine montent la nécessité de faire face à des données structurées complexes. Il nous paraît absolument nécessaire d'anticiper la montée, en complexité et non pas seulement en volume, des données pertinentes dans la masse d'expériences que produira ou traitera une grille. Les annotations des données (calibration, modes de production, environnement logiciel?) sont effectuées le plus naturellement sous forme textuelle, quel que soit leur mode de conservation ultérieur. L'utilisation de telles annotations contrôle largement la possibilité de partager les données pour certaines applications (imagerie médicale), et leur réutilisabilité à long terme dans tous les cas. Un axe de recherche prospective concernera donc la jonction avec les standards de fait des données structurée (XML).

3.2. Recherche de motifs spatio-temporels

Les données spatio-temporelles (résultats de simulation, archives météorologiques, données de trafic, films médicaux, neuro-imagerie cérébrale) sont fréquemment volumineuses, décrites selon une dimension temporelle et une, deux ou trois dimensions spatiales, avec souvent une granularité élevée (par exemple, pas de temps d'une milliseconde en neuroimagerie cérébrale électro-magnétique).

Le but est de passer d'une représentation brute des données à des représentations concises et interprétables : identification de motifs pertinents, par exemple stables selon une dimension temporelle et une dimension spatiale (objectifs antagonistes) ; identification de scénarios et de succession de motifs ; catégorisation et affichage des scénarios typiques.

Le dépouillement de telles données est en général manuel et donc extrêmement fastidieux - l'expert disposant d'une caractérisation souvent implicite de ce qu'il cherche et des résultats qu'on peut considérer comme intéressants.

Un premier objectif fondamental concerne la mise au point d'approches et d'algorithmes flexibles, adaptés aux types de critères explicites disponibles (critères monotones, volume de solutions attendues) et leur localité (zones actives, zones de rupture de la corrélation).

Un aspect essentiel est celui du compromis entre la qualité et la complétude des solutions, et les ressources en temps de calcul (algorithmes any-time). Un second point concerne la prise en compte de l'objectif généralement multi-critères de l'utilisateur : les critères intéressants (e.g. généralité vs précision) sont antagonistes, et le compromis souhaité entre ces objectifs évolue au cours de la session.

3.3. Modélisation de grands systèmes

La modélisation de grands systèmes et la recherche de politiques optimales donnent lieu à de nombreuses questions théoriques et appliquées :

- échantillonnage adéquat des aléas,
- modélisation du manque de connaissance sur le futur,
- prise en compte de l'hétérogénéité du système,
- passage à l'échelle.

Cette problématique générale sera étudiée sur une application particulière (voir Observatoire de la Grille), la modélisation du système EGEE. La compréhension affinée de ce système débouchera en effet sur l'identification des points bloquants pour le portage d'algorithmes existants sur grilles de calcul et in fine sur la conception d'algorithmes intégrant l'estimation des ressources qui leur sont nécessaires.

La modélisation théorique d'un grand système s'inspirera des approches de complexité stochastique connues sous le nom de transition de phase.

- Identification des paramètres d'ordre (taille du problème, type et structure du graphe d'interactions des composants).
- Modélisation de sous-systèmes partiels et/ou correspondant à une région particulière des paramètres d'ordre ; évaluation de l'hétérogénéité des sous-systèmes et de la confiance des modèles ; recherche de phénomènes de transition de phase, localisant les limites des modes de fonctionnement du système.

La modélisation d'un grand système précède souvent la recherche de politiques déterminant des régimes de fonctionnement optimaux du système compte-tenu des inputs et d'une fonction de coût endogène. Cependant la recherche de paramètres optimaux de processus coïncide avec l'optimisation d'une fonction en général très coûteuse, chaque itération donnant lieu à des calculs ou expérimentations intensifs.

Une alternative (surrogate optimization) consiste à s'appuyer sur le modèle disponible pour calibrer la recherche de politiques optimales. Les difficultés à résoudre mettent en jeu la qualité de l'approximation par rapport à la fonction de coût. De nouveaux critères d'erreur (compte tenu du caractère généralement non convexe de la fonction de coût) doivent en conséquence être proposés et leur convergence doit être étudiée.

3.4. Modélisation d'aléas et détection d'anomalies

La modélisation d'aléas complexes est un problème récurrent, qui est impliqué dans la modélisation de grands systèmes, et qui constitue le cœur de la recherche d'évènements rares au sein d'une masse de données.

Dans le premier cas, il est nécessaire de disposer d'une représentation fortement structurée des aléas, qu'il n'est pas facile de paramétrer. Par ailleurs, l'utilisation de critères d'identification intuitifs (recherche de l'aléa le plus proche de l'aléa réel au sens d'une distance naïve) conduit à des résultats loin de l'optimum, pour des raisons difficilement perceptibles parfois, tant la validation est alors un problème délicat : perte de variance liée à des discrétisations ; perte de corrélation temporelle par des modèles trop simplistes, conduisant à une sous-estimation de l'effet dit *effet Joseph* , et donc à des risques sous-estimés ; optimalité en un sens déconnecté de l'utilisation future de la modélisation de l'aléa ; utilisation d'une représentation facilitant les effets de surspécialisation et donc augmentant les risques, comme les représentations purement par scénarios.

La modélisation d'aléas par des structures discrètes, en particulier en vue de l'optimisation qui s'ensuit, est donc une bonne source de questions de recherche :

- élaboration de distances entre lois de probabilités conduisant à des résultats robustes et stables en recherche de politiques optimales ;
- estimation de risque de surapprentissage et recherche de bruitage pour robustification ;
- recherche de modélisations informatiquement gérables tout en préservant les corrélations temporelles et/ou spatiales et les variances ;
- génération quasi-aléatoire stratifiée (*importance sampling*), par souci de robustification face au risque.

Cette thématique concerne également le problème de la certification des résultats obtenus sur une architecture de calcul distribuée (*result checking*). La position classique du problème suppose qu'il existe une spécification explicite ou une propriété caractéristique de la correction du résultat de l'exécution d'un programme. Les phénomènes physiques ou biologiques étudiés par simulations numériques de type Monte-Carlo, ne fournissent pas nécessairement de tels critères avec une capacité de discrimination suffisante : l'objet de la simulation est précisément de définir la loi du phénomène. Il s'agit en outre d'un problème non paramétrique, la loi statistique étant souvent un mélange de lois ; les risques pris par les techniques classiques de détection d'outliers ne sont donc pas quantifiables. L'étude portera sur deux aspects complémentaires :

- élaboration d'une méthodologie générale de certification pour ce type d'application, avec pour objectif de quantifier le compromis entre les risques statistiques et la complexité de la certification ; les méthodes de test séquentiel seront particulièrement considérées ;
- définition de critères de discrimination spécifiques aux applications.

3.5. Modélisation non linéaire boîte noire ou boîte grise

Dans d'innombrables problèmes d'ingénierie, on est amené à construire des modèles pour concevoir et optimiser des systèmes, et ce sujet correspond à l'une des activités de la division Systèmes du L2S, et qui sera ici abordé en particulier sous l'angle nouveau pour nous de la modélisation d'une grille de calcul. En général, les modèles détaillés à base de connaissance ne sont pas disponibles, et quand ils le sont leur simulation s'avère trop complexe pour qu'on puisse les utiliser de façon répétitive à des fins d'optimisation. Ceci conduit au développement de modèles non linéaires simples du comportement entrée-sortie de systèmes. On parle alors de modèles boîte noire, qui peuvent devenir des modèles boîte grise si l'on est capable d'exploiter des informations a priori. Une fois construits, ces modèles peuvent être utilisés pour optimiser des comportements tout en tenant compte des exigences de robustesse des solutions.

Deux approches peuvent être envisagées pour la construction de ces modèles simples. La première consiste à faire tourner des gros codes de simulation de modèles à base de connaissance. On parle alors d'expérimentation numérique. La seconde consiste à construire des prototypes sur lesquels on expérimente les conséquences sur les performances de variations des facteurs sur lesquels on peut jouer. Dans les deux cas il s'agit d'un travail long et coûteux, et il est particulièrement important de développer des méthodes permettant de construire des modèles simples et efficaces à partir d'un nombre d'expériences qui soit le plus petit possible. L'engouement pour les réseaux de neurones dans ce contexte semble refluer au bénéfice de méthodes scientifiquement plus ambitieuses comme les *Support Vector Machines* ou plus généralement les méthodes à noyaux reproduisants. Le

krigeage fournit un cadre probabiliste très adapté pour résoudre les problèmes cruciaux pour ces méthodes à noyaux que sont le choix d'une structure de noyau, l'estimation des paramètres de cette structure et la caractérisation de l'incertitude sur les prédictions fournies par le modèle. Nos résultats méthodologiques récents sur les problèmes multivariables (plusieurs sorties dépendant de plusieurs facteurs) sont très prometteurs et doivent être exploités. Il en est de même de nos résultats sur la prise en compte d'information a priori pour arriver à des modèles de type boîte grise, dont on peut attendre de meilleures performances qu'avec des modèles purement boîte noire. Notre participation à ce projet nous donnera en particulier l'occasion d'envisager la modélisation boite noire ou grise du système complexe que constitue une grille de calcul à partir de l'observation de son comportement.

4. Bibliographie

[Coo05] Andrew W. Cooke, Alasdair J. G. Gray, Werner Nutt: Stream Integration Techniques for Grid Monitoring. *J. Data Semantics* 2: 136-175 (2005)

[Byr04] Rob Byrom, Brian A. Coghlan, Andrew W. Cooke, Roney Cordenonsi, Linda Cornwall, Ari Datta, Abdeslem Djaoui, Laurence Field, Steve Fisher, Steve Hicks, Stuart Kenny, James Magowan, Werner Nutt, David O'Callaghan, Manfred Oevers, Norbert Podhorszki, John Ryan, Manish Soni, Paul Taylor, Antony J. Wilson, Xiaomei Zhu: The CanonicalProducer: An Instrument Monitoring Component of the Relational Grid Monitoring Architecture (R-GMA). ISPDC/HeteroPar 2004: 232-237

[Coo04] Andrew W. Cooke, Alasdair J. G. Gray, Werner Nutt, James Magowan, Manfred Oevers, Paul Taylor, Roney Cordenonsi, Rob Byrom, Linda Cornwall, Abdeslem Djaoui, Laurence Field, Steve Fisher, Steve Hicks, Jason Leake, R. Middleton, Antony J. Wilson, Xiaomei Zhu, Norbert Podhorszki, Brian A. Coghlan, Stuart Kenny, David O'Callaghan, John Ryan: The Relational Grid Monitoring Architecture: Mediating Information about the Grid. *J. Grid Comput*. 2(4): 323-339 (2004)