Projet de réalisation d’une grappe de 200 PC pour le calcul scientifique et les serveurs de données

Responsable scientifique : Denis Trystram.
Responsable technique : Philippe Augerat.

1 - Objectif et statut

Le projet grappe 200PC vise à :

- Promouvoir la technologie grappe au sein des utilisateurs de calcul comme solution intermédiaire entre le poste de travail et l’utilisation de grands équipements de calcul ;
- Concevoir et expérimenter une grappe de nouvelle génération intégrant dans le système communication rapide, gestion de mémoire distribuée et support pour les environnements de calcul parallèle ;
- Evaluer une grappe réaliste (200 processeurs) avec des applications réelles dans le domaine du calcul scientifique d'une part et dans le domaine des serveurs d'information d'autre part.

En dépit d’un retard pour l’acquisition de la grappe de 200 PC (appelée grappe CIMENT dans la suite) dû à un premier appel d’offre infructueux en 2001, le projet a beaucoup avancé sur l’expérimentation des grappes de PC et le développement d’outils pour l’exploitation (utilisation, administration, programmation) de telles architectures. En plus du développement de nombreuses applications (principalement en cartographie, génomique et imagerie) et d’environnements pour l’utilisation de plates-formes d’exécution réparties de grande taille (grilles de grappes et calcul pair-à-pair) des outils pour le déploiement (de systèmes, d’applications) ou pour le passage à l’échelle de mouvements de données sont maintenant opérationnels.

C'est une grappe de 225 PCs (appelée grappe icluster), don de la compagnie Hewlett Packard, qui a permis de démarrer certains développements et évaluations prévus sur la grappe CIMENT. L'usage intense du icluster par une soixantaine de projets en dehors du laboratoire ID a montré le besoin d'infrastructures de grande taille pour la recherche en informatique comme pour la production scientifique des utilisateurs de calcul. Il a prouvé, dans le même temps, que les grappes de PC sont une bonne réponse à ce besoin.

Le projet CIMENT, avec une grappe plus performante, doit permettre de répondre à des applications encore plus gourmandes en mémoire et performances réseaux.
La grappe dont l'installation sera effective à l'automne sera composée d’une centaine de bi-processeurs Pentium4 reliés par un réseau rapide Myrinet.

L'environnement logiciel choisi pour l'exploitation de la grappe est issu d'un projet RNTL baptisé CLIC auquel participe le laboratoire ID en collaboration avec les sociétés Mandrakesoft et Bull.
Cet environnement commun aux grappes de CIMENT permettra des échanges de compétences entre ingénieurs et la mise à disposition aisée des outils standards pour le parallélisme ainsi que des apports des laboratoires de recherche (logiciels Athapascan, Ka, Pajé, etc.). La première version de cet environnement logiciel sort en juillet 2002, sous forme d'une distribution Linux pour grappes. Au dessus de l'environnement CLIC, le projet de recherche CI-GRID qui démarre à l'automne doit permettre de développer une partie du middleware nécessaire à l'interconnexion des ressources de CIMENT et à l'exécution sur la grille ainsi formée, d'applications développées dans CIMENT.

2 – Partenariats

Les équipes participant à la mise en place du projet sont :

- **projet APACHE** (laboratoire I.D. - responsable : Brigitte Plateau)
- **projet SARDES** (laboratoire SARDES - responsable : Jean-Bernard Stephani)
- **projet ReMaP** (laboratoire LIP, UMR CNRS-ENS Lyon-INRIA 5568 - responsable : Frédéric Deprez)

Les équipes partenaires dans le montage de la grappe sont :

- PARIS (INRIA Rennes) : Th. Priol
- EDP IDOPT (IMAG LMC) E. Blayo et I. Charpentier
- EDP SINUS (INRIA Sophia) S. Lanteri et T. Nguyen
- Chimie quantique & astrophysique (LAG) P. Valiron
- Dynamique moléculaire (CEA/DSV – Grenoble) S. Crouzy
- Dynamique moléculaire (Scallapix, Bordeaux) E. Coulaud
- Base et traitement d’images (GRAVIR - MOVI) R. Horaud
- Arénaire (arithmétique des ordinateurs) du laboratoire LIP, ENS Lyon
- Réseaux haut-débit et applications coopératives, Laboratoire LIGIM, UCB Lyon
- Optimisation combinatoire (PRISM) : Van Dat Cung
- Bases de données réparties (LSR) : Ch. Collet
- Simulation numérique et réalité virtuelle (Gravir) : F. Faure
- Laboratoire de biométrie et Biologie évolutive (université Claude Bernard, Lyon) : L. Duret

3 - Thèmes scientifiques

Les thèmes scientifiques associés au projet couvrent un spectre assez large. On peut citer notamment:

- des recherches en matière de **support pour l'exécution de programmes parallèles** (notamment : noyaux exécutifs implantant une machine virtuelle parallèle ; gestion mémoire distribuée ; protocoles réseaux haut-débits);
- des recherches en matière d'**algorithmique parallèle et répartie** (notamment : parallélisation, gestion et allocation de ressources: partage, ordonnancement de tâches, multiplexage, réplication, optimisation, etc.; observation et supervision; coordination et synchronisation ; sûreté de fonctionnement et qualité de service);
- des recherches en matière de **modèles de programmation parallèles et répartis** (y compris aspects langages, sémantiques, génération, adaptation et optimisation de code).
- des recherches en matière d'**administration de systèmes** (configuration et reconfiguration, sécurité, authentification, comptabilisation des ressources, observation des ressources).

Dans le cadre du projet CIMENT, un des problèmes fondamentaux que nous voulons traiter est le passage à l'échelle (*scalability*). Le passage à l'échelle caractérise la capacité d'une application ou d'un système à maintenir des caractéristiques acceptables de performances, de disponibilité et de qualité de service lors d'un accroissement de la taille du système (nombre de composants, nombre d'utilisateurs, étendue géographique, etc.).

La taille de la grappe a été fixée à un ordre de grandeur de 200 PC. Atteindre cette taille nous paraît nécessaire pour mettre en évidence les problèmes de passage à l'échelle liés à une architecture de plusieurs milliers de processeurs. Ces problèmes concernent essentiellement les aspects suivants :

- l'architecture du réseau d'interconnexion c'est à dire la hiérarchie de commutation la plus appropriée pour un type de grappe devant atteindre plusieurs milliers de processeurs.
- il en est de même pour les politiques d'ordonnancement et d'équilibrage de charge d'un grand nombre de processeurs,
- pouvoir évaluer les performances de la grappe sous une charge réelle et comparer à des machines classiques nécessite de pouvoir exécuter une application de calcul scientifique réaliste. Ainsi, l'application de dynamique moléculaire développée par le projet APACHE utilise les 256 processeurs d'un Cray T3E et 25 Go de mémoire pour simuler 400 000 atomes.
- La taille de la grappe de 200 PC permettra un « passage à l'échelle » pour certains développements logiciels réalisés

par le projet ReMaP dans ce cadre, notamment en matière de systèmes de fichiers parallèles et de protocoles de communication pour les réseaux haut-débit. Il en est de même pour les outils de déploiement (fichiers, système d'exploitation, commandes parallèles) développés au sein du projet APACHE

4 - Applications visées

Il n'existe pas aujourd'hui de tests synthétiques permettant d'évaluer l'efficacité d'une grappe pour le calcul haute performance ou l'accès intensif à de grandes quantités d'information. C'est pourquoi cette évaluation n'est possible qu'à travers des tests réels issus des différents domaines d'application que sont le calcul scientifique et les serveurs de données. Le projet APACHE a particulièrement investi dans le domaine scientifique depuis plusieurs années et dispose d'une large palette d'applications opérationnelles ou qui le seront au cours du déploiement de la grappe. L'investissement du projet SARDES dans le domaine est plus récent et concerne les serveurs de données. Quelques applications issues de ces deux domaines applicatifs sont présentées plus en détail dans la suite.

4.1. Calcul Scientifique

La validation des activités de recherche dans le domaine des environnements de programmation parallèle et de l'algorithmique parallèle procède par réalisation d'applications parallèles pilotes dans différents domaines du calcul intensif. Ces réalisations se font par coopération étroite avec les spécialistes des domaines concernés. Cette coopération prend souvent la forme d'un travail de thèse co-encadré.

Équations aux Dérivées Partielles et raffinement de maillage :

La parallélisation de problèmes d'EDP se fait classiquement par des méthodes de décomposition de domaines. Dans ce cadre, une modélisation fine de certains phénomènes physiques impose de raffiner le maillage en certaines zones du domaine. Ce raffinement peut par ailleurs s'accompagner de l'utilisation d'un modèle différent (couplage de code). Ce raffinement peut être statique (e.g. dépendant d'une géométrie fixe du problème) ou bien dynamique (e.g. déplacement d'une turbulence). Ceci pose des problèmes spécifiques de répartition dynamique de la charge. Deux domaines d'application sont actuellement en cours d'étude : un problème d'évolution en océanographie qui fait intervenir des maillages structurés et des méthodes multigrilles, et un problème de turbulence en aérodynamique, qui fait intervenir des maillages non structurés. Ces travaux sont réalisés en collaboration avec le projet IMAG/INRIA IDOPT (E. Blayo et I. Charpentier) et avec l'INRIA Sophia (S. Lanteri).

Chimie et biologie : La simulation des particules, atomes et molécules, suivant les lois de la mécanique quantique ou newtonienne sont des algorithmes coûteux (de complexité égale au carré, au cube ou à la puissance 4 du nombre de particules mises en jeu). Ces simulations trouvent des applications dans de nombreux domaines : pharmacologie, structure des matériaux, astrophysique, etc. Le parallélisme est une voie incontournable pour traiter plus vite des phénomènes plus complexes afin de rendre l'approche de modélisation et de calcul cohérente avec l'approche expérimentale.

- *Dynamique moléculaire* - La dynamique est une simulation du mouvement des atomes et des molécules par calcul de leurs déplacements. Cette technique est largement utilisée pour simuler les propriétés des solides, des liquides et des gaz. Elle est également employée pour étudier les conformations des macromolécules, et pour la compréhension des mécanismes réactionnels des protéines dans les structures biologiques. Le développement des médicaments de demain sera lié à la compréhension de ces mécanismes. Nous avons développé un code, basé sur une approche de décomposition de domaine et utilisant une approximation par rayon de coupure. Les techniques et programmes développés permettent de calculer des dynamiques avec des systèmes de plus de 450 000 atomes sur des périodes de plus de 100 pico-secondes. Ce programme met en évidence l'intérêt de l'approche de programmation proposée par le projet. Ces travaux, menés conjointement avec le Laboratoire de Biologie Moléculaire et Structurale du CEA-Grenoble (S. Crouzy), ont fait l'objet de la thèse de P.E. Bernard. Ils se sont poursuivis au sein de l'action incitative INRIA SYMBIO (O. Coulaud). Les perspectives aujourd'hui sont d'étendre les fonctionnalités du code au plan de la

modélisation biologique et d'envisager des variations. Un couplage avec des codes quantiques est aussi à l'étude.

- *Chimie théorique et astrophysique* : Les problèmes de simulation numérique en chimie constituent un corpus d'applications intéressantes pour le parallélisme. Une première étude a été entreprise sur la parallélisation à grosse granularité d'un code industriel de chimie quantique ab initio faisant référence (Gaussian-94), afin de préparer la parallélisation massive du traitement complet de la corrélation électronique, qui n'a pas encore été abordée dans les codes de production existants. Ces codes de production sont utilisés à la fois pour la recherche fondamentale (en astrophysique pour l'identification de molécules dans l'espace et la modélisation de la réactivité chimique à très basse température) et dans le monde industriel pour la modélisation de la catalyse et de la synthèse de nombreux composés. Cependant la généralisation de l'emploi de la modélisation en chimie théorique est limitée par la complexité des calculs et la place mémoire occupée qui croissent comme une puissance élevée du nombre d'atomes mis en jeu, d'où le recours au parallélisme sur grappe. Ces travaux se mènent en collaboration avec le laboratoire d'astrophysique de Grenoble (P. Valiron) et font l'objet de la thèse de N. Maillard.

4.2. Serveurs de données

Les applications du coté des serveurs de données sont plus récentes. Le développement d'applications de type "mémoire d'entreprise" nécessite l'utilisation d'entrepôts de données (*Data Warehouse*) dont la taille croit régulièrement. L'exploitation de ces données en vue d'extraire des informations significatives pour la gestion de l'entreprise (*Data Mining*) est réalisée à l'aide de requêtes de complexité significative. L'exécution de ces requêtes en temps réel pour les applications d'aide à la décision requiert de paralléliser le traitement correspondant. Un autre exemple d'utilisation des architectures parallèles pour la mise en œuvre de serveurs de données est lié à l'extension du World-Wide-Web (cache, serveurs multimédia).

Caches Web :

Dans le cadre du LHPC, ReMaP développe des logiciels permettant de transformer des grappes d'ordinateurs standard en serveurs de cache Internet à hautes performances pour les têtes de réseau des boucles haut débit. Ce système, réalisé conjointement par MS&I et le LIP, utilise des technologies issues du parallélisme pour garantir son extensibilité. Le système de cache intègre des fonctionnalités d'indexation en ligne et de filtrage. Un autre aspect novateur du système réside dans le support des flux audio et vidéo par des mécanismes de cache ou de miroir pour tenir compte de la part grandissante des documents multimédia dans le

trafic Internet. Le projet inclut des phases d'expérimentation sur le réseau haut débit « Autoroutes Rhodaniennes de l'Information » de Rhône Vision Câble. Cette plate-forme permettra de vérifier l'adéquation des solutions proposées en termes de fonctionnalités et de performances et de mettre en avant les bénéfices des infrastructures à haut débit.

Cartes géographiques à la demande :

Ce travail s'effectue en collaboration avec l'UMR 8504 Géographie-Cités, elle s'est contractualisé avec les partenaires suivants : Ministère des transports sur la thématique de l'interaction spatiale, un éditeur de CD-ROM pour la production de cartes, le réseau de recherche Hypercarte pour le lissage interactif et la production de cartes animées.

La représentation de données géostatistiques, données sociales comme les indices démographiques ou économiques, nécessite des calculs importants. En effet, une carte exprime la synthèse de données statistiques. Si, par exemple, on souhaite présenter la densité de population sur le globe à partir de la base de donnée référencée (degré par degré UNED Grid ou 5' par 5' de lattitude et longitude), la technique utilisée consiste à choisir un « rayon de lissage » R et à calculer en tout point du globe la population située dans un rayon de R km autour de ce point. Les géographes souhaitent agir dynamiquement sur le paramètre pour observer l'effet du rayon de lissage sur la représentation et en déduire des propriétés sur la population étudiée. Produire une carte, avec une précision acceptable pour les géographes peut nécessiter une petite heure sur un PC standard. Cela rend impossible toute interactivité. En parallélisant l'algorithme en MPI puis en Athapascan-1 et en optimisant le code par le choix de bibliothèques adaptées, ce temps a été réduit à quelques secondes sur une architecture parallèle (cluster de quelques PC). Il faut cependant construire une interface avec le réseau internet permettant l'accès distant à la base de données géostatistiques et aux cartes construites en ligne. Une première maquette a été réalisée en 1999. Le problème sous-jacent à cette architecture d'application est de tirer parti des calculs effectués pour la génération des cartes précédentes (flux de demandes de cartes). Cela nécessite une découpe de l'application telle qu'une partie des calculs soit mise dans un cache de calcul. La gestion de ce cache devient alors le principal outil d'accélération de l'application.

Tombé de vêtement :

Ce thème de recherche se positionne dans le cadre de l'animation d'objets tridimentionnels en synthèse d'image, en particulier pour la simulation de textiles pour représenter des personnages habillés.

Afin d'obtenir un résultat réaliste, les lois fondamentales de la physique, faisant intervenir des paramètres comme la vitesse,

les forces (gravitation, ...) ou les frottements sont employées pour modéliser le mouvement de plusieurs objets interagissants. Ces modèles sont numériquement très complexes : actuellement le calcul de l'image d'une personne varie de la seconde à plusieurs minutes suivant la complexité du modèle. Un de nos objectifs est de diminuer ce temps par la parallélisation des algorithmes. Cette diminution du temps permettrait d'obtenir des animations dynamiques "temps réel" (24 images par seconde). L'objectif final est la mise en oeuvre d'animation de textiles, c'est-à-dire la visualisation de mannequins portant des habits d'une façon réaliste avec l'intégration de tous les mouvements des tissus si la personne effectue un mouvement.

A l'heure actuelle il existe différents algorithmes permettant d'obtenir des images souhaitées. Les algorithmes permettant d'accélérer le temps d'exécution se décomposent selon la structure suivante : recherche des collisions possibles entre objets, calcul d'une matrice représentant l'environnement global (positions des objets, forces , ...), résolution de systèmes linéaires pour calculer la position des points, leur vitesse et leur accélération. Les opérations coûteuses en calcul étant les résolutions des systèmes linéaires par la méthode du Gradient Conjugué, et la recherche des collisions. Des techniques de parallélisation ont déja été étudiées pour les méthodes de résolution de grands systèmes linéaires. Des approches itératives avec contrôle d'erreur ont déjà été mises en oeuvre pour la synthèse d'animation.

Génomique :

L'objectif de ce travail de recherche est de proposer une approche générique pour la reconstruction d'arbres phylogénétiques. Ce problème correspond à la filiation entre séquences biologiques (gènes, protéines, espèces vivantes, etc.). Du point de vue informatique, il s'agit de déterminer un résultat global à partir de caractéristiques locales.
La plupart des approches existantes envisagent de construire ces arbres à partir d'alignements multiples de séquences. L'idée que nous poursuivons dans ce projet consiste à effectuer ces deux phases simultanément (alignements multiples et construction de l'arbre). Ceci permet de ne pas manipuler la totalité de la structure de données dont la taille est exponentielle en le nombre de séquences.

Les problèmes biologiques ciblés concernent typiquement des milliers de séquences et donc, conduisent à de très fortes combinatoires. Le parallélisme apparaît ici de façon incontournable pour pouvoir envisager une résolution. Ces modèles sont également plus précis et ont a priori une meilleure pertinence biologique. Une première maquette est en cours de

validation sur la grappe HP-icluster disponible au laboratoire ID.

5 – Environnements pour grappes et grilles de grappes

5.1 - Outils d'exploitation pour grappes de PC

Dans le cadre du projet RNTL CLIC, le laboratoire ID participe à la réalisation d'une distribution Linux pour grappe de PCs.
La généralisation de l'usage des grappes passe en effet par l'offre d'interfaces de programmation et d'outils d'administration qui permettront de voir (accéder, administrer, programmer) une grappe de PC comme s'il s'agissait d'une seule machine.
Outre leur intégration en un ensemble homogène et interopérable, la fourniture des standards de programmation parallèle (MPI, PVM), la distribution grappe doit offrir les fonctions suivantes qui font défaut ou ne sont que partiellement fournies dans les distributions actuelles :

- le support de réseaux hautes performances, des architectures IA32 et IA64 monoprocesseurs et multiprocesseurs ;
- des outils de configuration avancée (gestion des noeuds, modularité), les outils d'administration et de surveillance intégrés via une interface unique ;
- des outils d'exploitation permettant de paramétrer finement la partition des ressources de la grappe, les enchaînement d'exécution, les entrés/sorties, etc ;
- un choix d'environnements de programmation parallèle, de déboguage et d'analyse de performances.

Le laboratoire contribue au projet RNTL par le développement d'outils d'administration (déploiement, visualisation système, …), un travail de recherche sur l'amélioration des performances des environnements de programmation (MPI, OpenMP, Athapascan) et enfin la validation de la distribution ainsi construite sur des grappes en production.

5.2 - Metacomputing

L'évolution probable des architectures de type grappes consiste à faire fonctionner ces grappes entre elles et donc à construire et exploiter des grilles de grappes, potentiellement hétérogènes. C'est la problématique appelée métacomputing. Le laboratoire travaille dans différents domaines dans ce cadre.

Transfert de fichiers haut débit : dans le cadre du projet européen DataGrid, le laboratoire travaille à l'élaboration de

mécanismes permettant le transfert de fichiers sur des liens haut débit à grande distance, comme ceux offerts par le projet VTHD. Ces travaux portent à la fois sur la conception d'un protocole ainsi que sur l'implantation de ce protocole au sein d'un système d'exploitation.

Ordonnancement multi-niveaux : la nature hiérarchique des architectures de métacomputing pose de nouveaux problèmes d'ordonnancement tant théoriques que pratiques. Le laboratoire mène à la fois des recherches d'heuristiques adaptées à ces architectures et il développe également l'infrastructure logicielle qui permettra de les implanter efficacement.

Couplage de code : un autre aspect du métacomputing consiste à faire coopérer différents codes existants afin de faire de nouvelles applications scientifiques. C'est notamment le cas de l'application de dynamique moléculaire déjà présentée dans la partie applicative de ce rapport où, encore dans le cadre de VTHD, des composants s'exécutent à Nancy et coopèrent avec d'autres qui s'exécutent eux à Grenoble. Les recherches menées dans ce cadre étudient l'adéquation d'Athapascan-1 à cette problématique de couplage de code.

Enfin, En complément aux approches classiques du métacomputing comme mise en relation d'un faible nombre de ressources de calculs de forte puissance, nous nous intéressons à l'exploitation d'un très grand nombre de machines de faible capacité (typiquement les ordinateurs personnels) pendant leurs périodes d'inactivité. Cette approche est parfois appelée calcul « pair-à-pair ».

Dans ce cadre nous étudions deux facettes de cette problématique d'exploitation. La première concerne l'étude de ce type de système pour le stockage de l'information où nous tentons d'identifier les différents paramètres pertinents permettant de caractériser les performances pour la fonction à réaliser.

La deuxième concerne les capacités de calculs qui cumulées constituent un formidable potentiel. A l'heure actuelle une seule approche de type tâches indépendantes à gros grains et à faible volume de données à traiter semble réaliste. Des études sont donc à mener pour en étendre le domaine d'application. Un projet ACI-GRID vient tout juste de démarrer pour développer une plate-forme d'expérimentation dans ce cadre.

6 - Mise en œuvre et ouverture du projet

En 1999, l'unité de recherche INRIA Rhône-Alpes a fait équiper électriquement / thermiquement une salle afin d'accueillir les différentes grappes. Une première grappe de 14 PC bi-processeurs y a été installée en 1999. En 2000, une grappe de 100 PCs

« entrée de gamme » a été installée puis complétée en 2001 par 125 machines identiques. Cette grappe, nommée I-cluster, est équipée d'un réseau standard « fast ethernet ». Il s'agit d'un don de HP.

La grappe CIMENT de 100 PCs bi-processeurs haut de gamme reliés par un réseau haut débit faible latence verait être installée en automne 2002.

Une cinquantaine de personnes sont concernées, au sein du projet, par l'utilisation de la grappe pour le développement et l'exploitation d'outils et d'applications.

Les équipes concernées ont une action continue et visible autour des serveurs de calculs et de données. Par ailleurs, le projet est largement ouvert à la communauté Rhône-Alpine et nationale.
Plusieurs projets nationaux (RNTL CLIC, RNRT VTHD, ACI GRID, …) utilisent ou utiliseront l'infrastructure du projet.
L'action CIMENT a aussi contribué au démarrage de collaborations industrielles (HP, Bull, Microsoft, Mandrakesoft, Polyspace, Yxendis,…) qui fournissent en retour des investissements conséquents (ingénieurs, boursiers, matériel).

L'usage intense du icluster par une trentaine de projets en dehors du laboratoire ID a montré le besoin d'infrastructures de grande taille pour la recherche en informatique comme pour la production scientifique des utilisateurs de calcul. Il a prouvé, dans le même temps, que les grappes de PC sont une bonne réponse à ce besoin. Le projet CIMENT, avec une grappe plus performante, doit permettre de répondre à des applications encore plus gourmandes en mémoire et performances réseaux.

Un projet pour la création d'une grille de grappes de PC qui intègrerait dans une même structure administrative plusieurs grappes installées dans des laboratoires utilisateurs ainsi que le I-cluster et la grappe CIMENT vient d'être soutenu dans le cadre de l'ACI GRID. Ce projet doit permettre la mutualisation des ressources de calcul et la diffusion d'expertise en matière d'administration et de programmation des architectures de type grappe.

Références :

1 P.Augerat, S. Derr, S. Martin, and C. Robert.
Outils d'exploitation de grappe de pc.
In *Journées réseaux 2001*, December 2001.

2 P. Augerat, C. Martin, and B. de Oliveira Stein.
Scalable monitoring tools for grids and clusters.
In *10th Euromicro Workshop on Parallel, Distributed and Network-Based Processing*. IEEE Computer Society Press, January 2002.

3 Philippe Augerat, Camille Goudeseune, Hank Kaczmarski, Bruno Raffin, Benjamin Schaeffer, Luciano Soares, and Marcelo Knorich Zuffo.
Commodity clusters for immersive projection environments.
Siggraph 2002 Course, July 2002.

4 E. Bampis, F. Guinand, and D. Trystram.
Some models for scheduling parallel programs with communication delays.
Discrete Applied Mathematics, (72):5-24, 1997.

5 M. Béguin, J-M. Vincent, and B. Ycart.
An interacting particule model for load transferring.
In *IFIP WG 7.3*, Lausanne, October 1996.

6 A. Ben-Abdallah, A. S. Charão, I. Charpentier, and B. Plateau.
Ahpik: A Parallel Multithreaded Framework Using Adaptivity and Domain Decomposition Methods for Solving PDE Problems.
In *Proceedings of the 13th International Conference on Domain Decomposition Methods, October 2000*. CNME UPS, Barcelone, October 2001.

7 P-E. Bernard, B. Plateau, and D. Trystram.
Using threads for developing applications: Molecular dynamics as a case study.
In Trobec, editor, *Parallel Numerics 96*, pages 3-16, Gozd Martuljek, Slovenia, September 1996.

8 P.-E. Bernard and D. Trystram.
Report on a Parallel Molecular Dynamics Implementation.
In *Parallel Computing*, Bonn, Germany, May 1997.

9 Pierre-Eric Bernard, Thierry Gautier, and Denis Trystram.
Large scale simulation of parallel molecular dynamics.
In *Proceedings of Second Merged Symposium IPPS/SPDP 13th International Parallel Processing Symposium and 10th Symposium on Parallel and Distributed Processing*, San Juan, Puerto Rico, April 1999.

10 E. Blayo, L. Debreu, G. Mounié, and D. Trystram.
Dynamic load-balancing for adaptive mesh ocean circulation model.
Engineering Simulations, 22(2):8-24, 2000.

11 Eric Blayo, Laurent Debreu, Grégory Mounié, and Denis Trystram.
Topic 03 - dynamic load balancing for ocean circulation model with adaptive meshing.
In Patrick Amestoy, Philippe Berger, Michel Daydé, Iain Duff, Valérie Frayssé, Luc Giraud, and Daniel Ruiz, editors, *Euro-Par' 99 Parallel Processing - 5th International Euro-Par Conference*, number 1685 in Lecture Notes in Computer Science, pages 303-312, September 1999.

12 J. Blazewicz, M. Drozdowski, F. Guinand, and D. Trystram.
Scheduling a divisible task in a two-dimensional toroidal mesh.
DAMATH: Discrete Applied Mathematics and Combinatorial Operations Research and Computer Science, 94:35-50, 1999.

13 J. Blazewicz, K. Ecker, B. Plateau, and D. Trystram, editors.
Handbook on Parallel and Distributed Processing.
International Handbooks on Information Systems. Springer Verlag, 2000.

14 J. Blazewicz, F. Guinand, B. Penz, and D. Trystram.
Scheduling complete trees on two uniform processors with integer speed ratios and communication delays.
Parallel Processing Letters, 10(4):267-277, 2000.

15 J. Blazewicz, M. Machowiak, G. Mounié, and D. Trystram.
Suboptimal approach to scheduling malleable tasks.
Computational Methods in Science and Technology, Scientific Publishers OWN, 6:25-40, 2000.

16 J. Blazewicz, M. Machowiak, G. Mounié, and D. Trystram.
Approximation algorithms for scheduling independent malleable tasks.
In *Europar 2001*, number 2150 in LNCS, pages 191-196. Springer-Verlag, 2001.

17 B. Braschi and D. Trystram.
A new insight into the Coffman-Graham algorithm.

Siam Journal of Computing, 1994.

18 J. Briat and A. Carissimi. Intégration de *threads* et communications: une étude de cas. In *11 Rencontres francophones du parallélisme, des architectures et des systèmes*, Rennes, France, June 1999.

19 J. Briat, I. Ginzburg, and M. Pasin.
ATHAPASCAN-0B : un noyau exécutif parallèle.
Lettre du Calculateur Parallèle, 10(3):273-293, 1998.

20 J. Briat, I. Ginzburg, M. Pasin, and B. Plateau.
Athapascan runtime : Efficiency for irregular problems.
In *Proceedings of the Europar'97 Conference*, pages 590-599, Passau, Germany, August 1997. Springer Verlag.

21 F. Capello, D. Litaize, J-F. Mehaut, C. Morin, S. Petiton, and D. Trystram.
Metacomputing : vers une nouvelle dimension pour les calcul à haute performance.

Technique et Science Informatique, 19(6):877-902, 2000.
22 A. Carissimi and M. Pasin.
Athapascan: An experience on mixing mpi communications and threads.
In Vassil Alexandreov and Jack Dongarra, editors, *Proceedings of 5th European PVM/MPI Users' Group Meeting*, LNCS 1497, pages 137-144, Liverpool, UK, September 1998. Springer Verlag.
23 Gerson G. H. Cavalheiro, François Galilée, and Jean-Louis Roch.
Athapascan-1: Parallel Programming with Asynchronous Tasks.
In *Proceedings of the Yale Multithreaded Programming Workshop*, Yale, USA, june 1998.
24 Gerson-Geraldo-Homrich Cavalheiro, Yves Denneulin, and Jean-Louis. Roch.
A general modular specification for distributed schedulers.
In *Proceedings of EuroPar'98*, Southampton, England., September 1998.
25 Gerson-Geraldo-Homrich Cavalheiro, Matthias Doreille, François Galilée, Thierry Gautier, and Jean-Louis Roch.
Scheduling parallel programs on non-uniform memory architecture s.
In *HPCA Conference - Workshop on ``Parallel Computing for Irregular Applications WPCIA1''*, Orlando, USA, January 1999.
26 A.S. Charão, I. Charpentier, and B. Plateau.
A framework for parallel multithreaded implementation of domain decomposition methods.
In *Proceedings of Parallel Computing'99*, Delft, The Netherlands, August 1999.
27 A. S. Charão, I. Charpentier, and B. Plateau.
Programmation par objet et utilisation de processus légers pour les méthodes de décomposition de domaine.
Technique et Science Informatiques, 1999.
to appear.
28 J. Chassin de Kergommeaux and B. de Oliveira Stein.
Pajé: an extensible environment for visualizing multi-threaded programs executions.
In *Proceedings of EuroPar2000, Munich, Germany*, 2000.
Accepté.
29 J. Chassin de Kergommeaux and B. de Oliveira Stein.
Pajé, an interactive visualization tool for tuning multi-threaded parallel applications.
Parallel Computing, 2000.
30 J. Chassin de Kergommeaux and A. Fagot.
Execution replay of parallel procedural programs.
Journal of Systems Architecture, 46(10):835-849, July 2000.
31 J. Chassin de Kergommeaux, É. Maillet, and J.-M. Vincent.
Parallel Program Development for Cluster Computing: Methodology, Tools and Integrated Environments, chapter 6: Monitoring Parallel Programs for Performance Tuning in Distributed Environments.
Nova Science, 2000.
32 J. Chassin de Kergommeaux, É. Maillet, and J.-M. Vincent.
Monitoring parallel programs for performance tuning in cluster environments.
In J. Cunha, P. Kacsuck, and Winter S., editors, *Parallel Program Development for Cluster Computing: Methodology, Tools and Integrated Environments*, volume 5 of *Advances in Computation: Theory and Practice*, chapter 6, pages 131-150. Nova Science, 2001.
33 J. Chassin de Kergommeaux, M. Ronsse, and K. De Bosschere.
Mpl*: efficient record/replay of nondeterministic features of message passing libraries.
In *Proc. EuroPVM/MPI'99*. Springer Verlag, September 1999.
34 O. Coulaud and T. Gautier.
Architecture distribuée pour la simulation moléculaire distribuée.
In J. Roman J.-L. Pazat, S. Rajopadhye, editor, *iHPerf'2000: Applications parallèles hautes performances : Analyse, Conception et Utilisation de Grappes Homogènes ou Hétérogènes de Calculateurs*, pages 221-226. CNRS, Aussois, December 2000.
35 Carolina Cruz-Neira, Christopher Just, Kevin Meinert, Allen Bierbaum, Patrick Hartling, and Bruno Raffin.
Open source virtual reality.
IEEE VR 2002 Tutorial, March 2002.
36 V.-D. Cung, P. Fraigniaud, T. Gautier, and D. Trystram.
De l'algorithme au support.
In D. Barth, J. Chassin de Kergommeaux, J.-L. Roch, and J. Roman, editors, *ICaRE'97: conception et mise en oeuvre d'applications parallèles irrégulières de grande taille*. CNRS, December 1997.
37 Yves Denneulin and Pierre Lombard.
Towards single image system: Preemptive migration of system threads with the sci network.
In *Proceedings of the CLUSTER'2000 conference*, Chemnitz, Germany, November 2000.
38 Yves Denneulin and Jean-François Méhaut.
Customizable thread scheduling directed by priorities.
In *Proceedings of Workshop on Multi-Threaded Execution, Architecture and Compilation (MTEAC 99), joint with HPCA'5*, Orlando, USA, January 1999.

39 E. D'Hollander, G. Joubert, F. Peters, and D. Trystram, editors.
Parallel Computing: state-of-the-art and perspectives.
Number 11 in Advances in Parallel Computing. North Holland, 1996.
40 M. Doreille, B. Dumitrescu, J.-L. Roch, and D. Trystram.
Influence of Scheduling on Actual High-performance Computing Applications.
In *Proceedings of PPAM'97 - 2nd International Conference on Parallel Processing and Applied Mathematics*, pages 41-55, Zakopane, Poland, 1997.
41 M. Doreille, B. Dumitrescu, J.-L. Roch, and D. Trystram.
Two-dimensional block partitionings for the parallel sparse Cholesky factorization.
Numerical Algorithms, 16(1):17, February 1998.
42 P.-F. Dutot and D. Trystram.
Scheduling on hierarchical clusters using malleable tasks.
In *Proceedings of the 13th annual ACM symposium on Parallel Algorithms and Architectures - SPAA 2001*, pages 199-208, Crete Island, July 2001.
43 Pierre-François Dutot.
Ordonnancement de chaînes de tâches malléables.
In *14èmes Rencontres Francophones du Parallélisme*, pages 35-41, avril 2002.
44 E. Edi, D. Trystram, and J.-M. Vincent.
Amelioration de performances de serveur de requetes dynamiques.
In *6ieme colloque africain sur la recherche en Informatique - CARI*, Yaounde, Cameroun, oct 2002.
45 D. El Baz and B. Plateau, editors.
Multithreads, volume 10 of *Calculateurs Parallèles Réseaux et Systèmes répartis*. HERMES, July 1998.
46 Luis Gustavo Fernandes, Nicolas Maillard, and Yves Denneulin.
Parallelizing a dense matching region growing algorithm for an image interpoliation application.
In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'2001)*, Las Vegas, June 2001.
47 F. Galilée, Jean-Louis Roch, Gerson-Geraldo-Homrich Cavalheiro, and Mahtias Doreille.
Athapascan-1: On-line building data flow graph in a parallel language.
In IEEE, editor, *Pact'98*, pages 88-95, Paris, France, October 1998.
48 A. Goldman, G. Mounié, and D. Trystram.
1-optimality of static bsp computations: scheduling independent chains as a case study.
Theoretical Computer Science.
to appear.
49 A. Goldman and D. Trystram.
Algorithms for the message exchange problem.
In *Proceedings of the International Conference on Parallel Computing in Electrical Engineering*, pages 153-158, Byalistok, Poland, September 1998.
50 A. Goldman and D. Trystram.
Efficient parallel algorithm for solving the knapsack problem on hypercube.
Journal of Parallel and Distributed Computing - JPDC, to appear.
51 A. Goldman, D. Trystram, and J. Peters.
Exchange of messages of different sizes.
In *Proceedings of Irregular'98*, number 1457 in Lecture Notes in Computer Science, pages 194-205, Berkeley, USA, August 1998.
52 C. Guilloud, P. Augerat, J. Chassin de Kergommeaux, and B. Stein.
Outil visuel d'administration système pour grappe de processeurs.
In *RenPar'13*, pages 163-168, April 2001.
53 F. Guinand, G. Parmentier, and D. Trystram.
Construction of phylogenetic trees on parallel clusters.
In *Fourth SIAM International Conference on Parallel Processing and Applied Mathematics - PPAM 01*, Naleczow, Poland, September 2001.
Article invité.
54 F. Guinand, C. Rapine, and D. Trystram.
Worst case analysis of lawler's algorithm for scheduling trees with communication delays.
IEEE Trans. on Parallel and Distributed Systems, 8(10), 1997.
55 F. Guinand and D. Trystram.
Scheduling uet trees with communication delays on two processors.
RAIRO, Operational Research, 34(2):131-144, 2000.
56 Christian Guinet, Emmanuel Romagnoli, and Yves Denneulin.
A scheduler of parallel and sequential jobs with dependencies for clusters and grids.
In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'2001)*, Las Vegas, June 2001.
57 Claude-Pierre Jeannerod and Nicolas Maillard.
Using Computer Algebra To Diagonalize Some Kane Matrices.
Journal of Physics A: Mathematics General, 33:2857-2870, 2000.
58 G. Joubert, D. Trystram, F. Peters, and D. Evans, editors.
Parallel Computing: Trends and perspectives.
Number 7 in Advances in Parallel Computing. North Holland, 1994.
59 P. Kacsuck, J. Chassin de Kergommeaux, É. Maillet, and J.-M. Vincent.
Parallel Program Development for Cluster Computing: Methodology, Tools and Integrated Environments, chapter 14: The Tape/PVM Monitor

and the PROVE Visualization Tool.
Nova Science, 2000.
60 P. Kacsuck, J. Chassin de Kergommeaux, É. Maillet, and J.-M. Vincent.
The tape/pvm monitor and the PROVE visualization tool.
In J. Cunha, P. Kacsuck, and Winter S., editors, *Parallel Program Development for Cluster Computing: Methodology, Tools and Integrated Environments*, volume 5 of *Advances in Computation: Theory and Practice*, chapter 14, pages 291-303. Nova Science, 2001.
61 T. Kalinowski, I. Kort, and D. Trystram.
List scheduling of general task graphs under LogP Model.
Parallel Computing, Special Issue on Scheduling Parallel and Distributed systems, 26:1109-1128, 2000.
62 J.-C. Konig, P.S. Rao, and D. Trystram.
Analysis of gossiping algorithms with restricted biffers.
Parallel Algorithms and Applications, 13:117-133, 1998.
63 J.-C. Konig and J.-L. Roch.
Machines virtuelles et techniques d'ordonnancement.
In D. Barth, J. Chassin de Kergommeaux, J.-L. Roch, and J. Roman, editors, *ICaRE'97: conception et mise en oeuvre d'applications parallèles irrégulières de grande taille*. CNRS, December 1997.
64 Erricos Kontoghiorghes, Amed Sameh, and Denis Trystram, editors.
Parallel Computing, volume 28 of *Special Issue on Parallel matrix algorithms and applications*.
Elsevier, 2002.
65 I. Kort and D. Trystram.
Assessing LogP Model Parameters for the IBM-SP.
In *Proceedings of EuroPar'98*, Southampton, England, September 1998.
66 I. Kort and D. Trystram.
Some Results on Scheduling Flat Trees in LogP Model.
Journal of Information Systems and Operational Research (INFOR), 37(1), 1999.
67 D. Kranzlmüller, J. Chassin de Kergommeaux, and Ch Schaubschl ger.
Correction of monitor intrusion for testing nondeterministic mpi-programs.
In *Proc. Euro-Par'99*, pages 154-158. Springer Verlag, August 1999.
68 Wieslaw Kubiak, Bernard Penz, and Denis Trystram.
Scheduling independent chains on uniform processors.
J. of Scheduling, to appear.
69 C. Labbé, S. Martin, and J.-M. Vincent.
A reconfigurable hardware tool for high speed network simulation.
In *10th International Conference for Computer Performance Evaluation, TOOLS'98*, September 1998.
70 C. Labbé, F. Reblewski, and J.-M. Vincent.
Performance evaluation of high speed network protocols by emulation on a versatile architecture.
RAIRO Recherche Operationnelle - Operational Research, 32(3):253-270, 1998.
71 B. Le Cun, S. Rajopadhye, J.-L. Roch, and C. Roucairol.
Algorithmique parallèle.
In J. Roman J.-L. Pazat, S. Rajopadhye, editor, *iHPerf'2000: Applications parallèles hautes performances : Analyse, Conception et Utilisation de Grappes Homogènes ou Hétérogènes de Calculateurs*, pages 37-73. CNRS, Aussois, France, December 2000.
72 R. Lepère, G. Mounié, C. Rapine, and D. Trystram.
A general method for designing approximation algorithms for scheduling malleable tasks.
In *ECCO, the 13th European Conference on Combinatorial Optimization*, Capri, Italy, May 2000.
73 R. Lepère, G. Mounié, and D. Trystram.
An approximation algorithm for scheduling trees of malleable tasks.
European Journal of Operational Research.
2002.
74 R. Lepère, G. Mounié, D. Trystram, and B. Robic.
Malleable tasks: an efficient model for solving actual parallel applications.
In E. D'Hollander et al., editor, *Parallel Computing - fundamentals and Applications. Proceedings of PARCO'99, Delft*, pages 598-605. Imperial College Press, 2000.
75 R. Lepère, D. Trystram, and G.J. Woeginger.
Approximation scheduling for malleable tasks under precedence constraints.
In *9th Annual European Symposium on Algorithms - ESA 2001*, number 2161 in LNCS, pages 146-157. Springer-Verlag, 2001.
76 Renaud Lepère and Grégory Mounié.
Ordonnancement de tâches malléables. une alternative efficace pour la programmation d'applications parallèles.
Réseaux et Systèmes Répartis, Calculateur Parallèle, 2001.
77 Renaud Lepère and Denis Trystram.
A new clustering algorithm for scheduling with large communication delays.
In *16th IEEE-ACM annual International Parallel and Distributed Processing Symposium, IPDPS 02*, Fort Lauderdale, apr 2002.
78 Pierre Lombard and Yves Denneulin.
Towards single image system: preemptive migration

of system threads with the sci network.
In *proceedings of the IEEE international conference on cluster computing, CLUSTER'2000*, pages 250-257, Chemnitz, Germany, November 2000.
79 Pierre Lombard and Yves Denneulin.
A freeze/unfreeze mechanism for the linuxthreads library.
In *Proceedings of the 9th Euromicro workshop on parallel and distributed processing*, pages 84-88, Mantova, Italy, February 2001.
80 Pierre Lombard and Yves Denneulin.
nfsp: a distributed nfs server for clusters of workstations.
In *Proc. of the International Parallel and Distributed Processing Symposium*, Fort Lauderdale, Florida, April 2002.
81 Z. Mahjoub, W. Nasri, and D. Trystram.
Parallelization of a divide and conquer algorithm for solving triangular matrix inversion on heterogeneous platforms.
In *6ieme colloque africain sur la recherche en Informatique - CARI*, Yaounde, Cameroun, oct 2002.
82 C. Martin and O. Richard.
Parallel lancher for cluster of pc.
In London Imperial College Press, editor, *PARCO 2001, World Scientific*, 2001.
83 Grégory Mounié, Christophe Rapine, and Denis Trystram.
Efficient approximation algorithms for scheduling malleable tasks.
In *Eleventh ACM Symposium on Parallel Algorithms and Architectures (SPAA'99)*, pages 23-32, June 1999.
84 F. Ottogalli and J-M Vincent.
Mise en cohérence et analyse de traces logicielles multi-niveaux.
Calculateurs parallèles, 1999.
85 F. Ottogalli and J-M Vincent.
Mise en cohérence et analyse de traces multi-niveaux.
In *première Conférence Française sur les Systèmes d'Exploitation (CSFE'1)*, Rennes, France, June 1999.
86 F.-G. Ottogalli, C. Labbé, V. Olive, B. Oliveira Stein, J. Chassin de Kergommeaux, and J.-M. Vincent.
Visualisation of distributed applications for performance debugging.
In V. Alexandrov, J. Dongarra, B. Juliano, R. Renner, and C.J. Kenneth Tan, editors, *ICCS'01: International Conference in Computational Science*, LNCS 2074, pages 831-840, Berlin, Heidelberg, 2001. Springer.
87 B. Penz, C. Rapine, and D. Trystram.
Sensitivity analysis of scheduling algorithms.
European Journal of Operational research, 134:606-615, 2001.
88 B. Plateau and D. Trystram.
Parallel and distributed computing: State-of-the-art and emerging trends.
In J. et al. Blazewicz, editor, *International Handbooks on Information Systems*, pages 1-12. Springer Verlag, 2000.
89 C. Rapine R. Lepere.
An asymptotic approximation algorithm for the scheduling problem with duplication.
In *STACS'02*, Antibes, France, march 2002.
90 Bruno Raffin.
Des grappes de pc pour la réalité virtuelle.
In *Imagina'02*, Monaco, 2002.
Conférencier invité.
91 C. Rapine and D. Trystram.
Iterative approaches for the clustering problem.
In *EuroPar96*, Lyon, August 1996. LNCS-Springer Verlag.
92 Christophe Rapine, Isaac Scherson, and Denis Trystram.
On-line scheduling of parallelizable jobs.
In Springer verlag, editor, *Proceedings of EUROPAR'98*, number 1470 in LNCS, pages 322-327, Southampton, England, September 1998.
93 B. Richard and P. Augerat.
I-cluster: intense computing with untapped resources.
In *Proc. of the Fourth International Conference on Massively Parallel Computing Systems*, Ischia, Italy, 2002.
Invited talk.
94 B. Richard, P. Augerat, S. Derr, S. Martin, and C. Robert.
I-cluster : reaching top500 performance using mainstream hardware.
Technical Report HPL-2001-206, HP Laboratories technical report, August 2001.
95 Jean-Louis Roch.
Ordonnancement de programmes parallèles sur grappes : théorie versus pratique.
In *Actes du Congrés International Université Mohamm V*, pages 131-144, Rabat, Maroc, 28-31 Mai 2001.
96 M. Ronsse, J. Chassin de Kergommeaux, and K. De Bosschere.
Execution replay for an mpi-based multi-threaded runtime system.
In *Proc. ParCo99*, August 1999.
presented at the conference. To appear in the proceedings.
97 M. Ronsse, J. Chassin de Kergommeaux, and K. De Bosschere.
Execution replay for an MPI-based multi-threaded runtime system.
In E. H. D'Hollander, G. R. Joubert, F. J. Peters, and H. J. Sips, editors, *Parallel Computing: Fundamentals and Applications*, Proceedings of the International Conference ParCo99, pages 656-663. Imperial College Press, 2000.

98 M. Ronsse, K. De Bosschere, and J. Chassin de Kergommeaux.
Execution replay and debugging.
In M. Ducassé, editor, *Proc. of the Fourth International Workshop on Automated Debugging, AADEBUG 2000*, pages 5-18, Munich, 2000. TUM-IRISA.
Invited talk.
99 Andre Schiper and Denis Trystram, editors.
Numéro Spécial 9ièmes rencontres Francophones du Parallélisme, volume 17 of *TSI (Technique et Sciences Informatiques*.
Hermes, 1998.
100 D Trystram.
Implementation of parallel applications: an experience of 15 years at imag.
In G. Oska et al., editor, *Proceedings of the International workshop Parallel Numerics (ParNum2000)*, pages 9-20, Bratislava, September 2000.
101 D. Trystram.
Scheduling on hierarchical clusters using mt.
In *Workshop on Scheduling and Communication, IPDPS 2001*, San Francisco, April 2001.
Article invité.
102 D. Trystram and W. Zimmermann.
On multi-broadcast and scheduling receive-graphs under logp with long messages.
In S. Jaehnichen and X. Zhou, editors, *The Fourth International Workshop on Advanced Parallel Processing Technologies - APPT 01*, pages 37-48, Ilmenau, Germany, September 2001.
103 F. Zara.
Simulation physique de textiles sur grappe de processeurs.
In *Association Française d'Informatique Graphique, AFIP 2001*, Limoges, November 2001.