Communications
dans les réseaux de processeurs

Opération RUMEUR du PRC C^3

22 août 1994

Ouvrage rédigé avec le concours du Ministère de l'Enseignement Supérieur et de la Recherche (Direction de l'Information Scientifique et Technique et des Bibliothèques).

Préface

En moins de dix ans, le calcul parallèle est passé des laboratoires de recherche aux centres de calculs. Il s'agit pour l'informatique d'une véritable révolution, puisqu'elle touche profondément l'activité centrale de cette discipline : la programmation.

Ceux qui ont déjà eu l'occasion de programmer une machine parallèle savent à quel point les habitudes de la programmation séquentielle sont largement remises en cause. Pour obtenir les résultats espérés — un programme performant, ou tout simplement correct —, il faut changer de mode de pensée. Cette impression est largement confirmée par le développement de la théorie du calcul parallèle : les modèles changent, et leur étude conduit à des problèmes nouveaux.

Parmi ceux-ci, celui des communications est central. Si les programmes parallèles étaient faits de calculs sans communication, il n'y aurait rien de plus à dire que ce qui concerne les programmes séquentiels, ce qui, en soi, n'est déjà pas mal. Mais les communications sont l'essence même du calcul parallèle : la résolution d'un problème unique par plusieurs processeurs passe forcément par la distribution des données, l'échange et la comparaison de résultats intermédiaires, ou encore la diffusion des solutions de sous-problèmes ; toutes opérations qui nécessitent des communications.

L'étude des communications dans le calcul parallèle n'est pas récente, mais ce domaine a connu un essor important à l'arrivée des machines dites *massivement parallèles*. La raison en est simple. Il est physiquement possible d'interconnecter complètement un tout petit nombre de processeurs, mais dès que ce nombre croît, cela devient irréaliste. Il faut donc trouver des réseaux de communications efficaces, ainsi qu'une façon de les utiliser qui ne réduisent pas à néant les effets de l'augmentation du nombre des processeurs. Or, les machines massivement parallèles ont l'ambition de faire coopérer des centaines, voire des milliers de processeurs. On conçoit donc que ce sujet ait de l'importance.

En outre, les progrès de la technologie des processeurs n'améliorent en rien la situation. Même si les réseaux d'interconnexion, les algorithmes de routage et les dispositifs matériels de communication ont fait des progrès énormes en quelques années, le rapport entre vitesse de calcul et vitesse de communication des processeurs les plus récents est aujourd'hui plus que jamais défavorable aux communications. Il s'agit donc d'un problème fondamental, qui ne se résoudra pas de lui-même avec l'évolution de la technologie des circuits intégrés.

En France, le Programme de Recherches Coordonnées C^3 (Communication, Coopération, Concurrence) a reconnu très tôt la nécessité d'étudier ce sujet, puisqu'il l'a inscrit dans son titre lors de sa création en 1981, avant même que les premiers ordinateurs massivement parallèles ne soient conçus. Mais c'est la pratique de la programmation sur des machines parallèles qui a conduit la recherche sur ce sujet à se structurer. Dès 1986, C^3 a soutenu l'acquisition de machines parallèles à mémoire distribuée : l'Intel iPSC/1 de Rennes et le FPS T40 de Grenoble. Ces deux machines, prototypes de laboratoire plus que véritables calculateurs, ont permis à une génération de chercheurs de pratiquer le calcul parallèle et de déterminer ainsi, par l'expérience, les sujets de recherche critiques.

Les communications ont très naturellement été l'un de ces sujets. La coordination des équipes de recherche intéressées par ce problème a donné naissance en 1991 à une opération de C^3, très joliment baptisée RUMEUR par ses animateurs. RUMEUR regroupe une vingtaine de chercheurs provenant de laboratoires répartis sur toute la France — principalement Bordeaux, Grenoble, Lyon, Orsay et Sophia-Antipolis ; mais aussi Lille, Rennes et Toulouse.

L'ambition de RUMEUR est de formaliser les modèles, de conceptualiser les idées et enfin, de développer des outils. Ces derniers devraient servir aux nombreux chercheurs qui font face à des problèmes de communications. Ces outils peuvent être théoriques, en faisant appel à des techniques empruntées aux mathématiques discrètes et à la théorie des graphes, ou plus pratiques, comme des bibliothèques de communications développées pour différentes machines parallèles.

A l'initiative de RUMEUR, une école d'été a été organisée en août 1992 à Cargèse (Corse) pour des étudiants de troisième cycle, répondant ainsi à une forte demande de formation dans le domaine des communications. Le support de cours de cette école, dûment corrigé et complété, grâce aux critiques des participants, est devenu cet ouvrage.

Grâce à la compétence de ses auteurs et à leurs échanges scientifiques soutenus — entre eux, mais aussi avec la communauté scientifique internationale dans laquelle RUMEUR a pignon sur rue —, ce livre présente et synthétise les connaissances les plus récentes sur les communications dans les machines parallèles. De plus, en très peu de temps, ses auteurs ont réussi le tour de force de faire un ouvrage très pédagogique, auquel ne manquent même pas des exercices corrigés.

Unique en son genre, il intéressera donc tous ceux qui, étudiants, informaticiens chevronnés ou utilisateurs avertis de calculateurs parallèles, sont intéressés par les aspects théoriques et pratiques des communications.

Patrice Quinton

Directeur de Recherche CNRS
Directeur du Programme de Recherches Coordonnées C^3

Table des matières

Annexes

Liste des figures

Auteurs

Jean-Claude BERMOND - Directeur de Recherches CNRS
I3S Sophia-Antipolis

Pierre FRAIGNIAUD - Chargé de Recherches CNRS
LIP-IMAG Lyon

Anne GERMA - Professeur
ENST Paris

Marie-Claude HEYDEMANN - Professeur
LRI Orsay

Emmanuel LAZARD[1] - Allocataire de Recherches
LRI Orsay

Philippe MICHALLON - Allocataire de Recherches
LMC-IMAG Grenoble

André RASPAUD[2] - Maître de Conférences
LaBRI Bordeaux

Dominique SOTTEAU - Directeur de Recherches CNRS
LRI Orsay

Michel SYSKA[3] - Allocataire de Recherches
I3S Sophia-Antipolis

Denis TRYSTRAM - Professeur
LMC-IMAG Grenoble

avec la collaboration de

Eric DARROT - Allocataire de Recherches, *I3S Sophia-Antipolis*

Daniel LAFAYE DE MICHEAUX - Maître de Conférences, *I3S Sophia-Antipolis*

[1] E. Lazard est maintenant Maître de Conférences au LAMSADE (Paris-Dauphine).
[2] A. Raspaud est aujourd'hui Professeur.
[3] M. Syska a obtenu un poste de Maître de Conférences.

Avant-propos

Calcul parallèle et communications

L'importance des machines massivement parallèles — systèmes pouvant comporter plusieurs milliers de processeurs — n'est plus à démontrer aujourd'hui ; ces machines répondent à un besoin sans cesse croissant de puissance de calcul.

De ce fait, de nombreux domaines de recherche nouveaux ont émergé ; par exemple, les langages parallèles, la parallélisation automatique, les architectures distribuées, les environnements de programmation, l'analyse de complexité parallèle et l'algorithmique parallèle.

La plupart des machines existantes reposent sur des architectures à mémoire distribuée. Une des différences essentielles entre les machines traditionnelles, dites de *Von Neumann*, et les réseaux de processeurs que sont les machines à mémoire distribuée, est la gestion des communications. Par exemple, la distribution des données, le placement des processus sur les différents processeurs, et la gestion des mouvements de données requis par les calculs, sont autant de problèmes qui font intervenir des communications. Les communications sont d'autant plus importantes que la granularité du parallélisme est fine. En effet, si l'on parallélise un programme grossièrement, c'est-à-dire en identifiant des procédures indépendantes, peu de communications sont nécessaires, et le traitement reste en majorité séquentiel. Au contraire, en distribuant le plus possible les traitements élémentaires, on obtient le parallélisme potentiellement le plus efficace, mais les communications induites peuvent alors être pénalisantes. Optimiser les communications est donc la clef de l'efficacité des algorithmes parallèles.

Les chercheurs en parallélisme étudient donc les problèmes liés aux communications afin d'accélérer les traitements, tant du point de vue du logiciel que de l'architecture du matériel. Il est indispensable d'aborder cette étude de façon globale. En effet, une connaissance approfondie des modèles et des mécanismes élémentaires de communication des machines est requise pour proposer des solutions efficaces, applicables aux systèmes réels.

Mais l'étude des communications fait appel à des connaissances variées : en architecture (circuits intégrés, cartes et liens de communication), en algorithmique (communications globales, routage, interblocage, mise en œuvre d'applications, etc.) et en théorie des graphes (coloration, plongement, cheminement, etc.). Cette grande diversité montre combien il est difficile d'avoir une vision synthétique du domaine. La synthèse des résultats actuels exposée dans ce livre, tend à offrir au lecteur un cadre général pour l'étude de divers aspects des systèmes parallèles : architecture, environnements, langages, algorithmique et programmation.

Jusqu'à présent, les communications ont été étudiées de manière dispersée par des chercheurs ne s'intéressant qu'à des problèmes très spécifiques (développement, dans un modèle donné, d'un schéma de communication donné, sur une architecture donnée). L'objet de ce livre est au contraire de présenter de manière unifiée les recherches sur les communications, sans pour autant en gommer les particularités.

Description de l'ouvrage

Ce livre est organisé en six chapitres, complétés par deux annexes.

Dans le chapitre 1 sont définis les termes de théorie des graphes utilisés dans cet ouvrage. On y trouve les définitions de familles de graphes classiques utilisés dans les architectures (anneau, graphe complet, hypercube, grille, etc.), ainsi que leurs propriétés élémentaires. Ce chapitre contient aussi une étude théorique du problème du nombre maximum de sommets d'un graphe de degré maximum et de diamètre fixés.

Le chapitre 2 présente d'autres graphes, bien adaptés pour servir de support aux réseaux (graphes de de Bruijn et de Kautz, graphe *shuffle-exchange*, etc.). Ces graphes ont un petit degré et un diamètre logarithmique en le nombre de sommets. Le chapitre 2 introduit aussi les réseaux multiétages qui, contrairement aux précédents, ont la particularité de ne pas être des réseaux statiques mais de pouvoir réaliser, grâce aux commutateurs qui en sont la base, un grand nombre de permutations des entrées sur les sorties.

Les modèles de communication et les différents types de routage sont abordés au chapitre 3. Une classification des machines parallèles et les principes architecturaux des machines à mémoire distribuée y sont donnés. De même, les principaux modes de commutation sont présentés (*circuit-switched*, *store-and-forward*, *wormhole*, *virtual cut-through*, etc.). Ceci conduit à une modélisation des réseaux et des temps de communication (temps constant et linéaire). Les modes de communications principaux sont comparés. Enfin, le problème de l'interblocage lors du routage est posé et quelques solutions y sont apportées.

Le chapitre 4 décrit les principaux types de communications globales (diffusion, échange total, distribution et rassemblement). Ce chapitre est découpé en trois parties, suivant les principales hypothèses décrites au chapitre précédent. Les deux premières parties traitent du mode *store-and-forward*, sous l'hypothèse temps constant, puis sous l'hypothèse temps linéaire. La troisième partie traite du mode de routage *wormhole*. Dans chacune des parties sont présentés des bornes inférieures et des algorithmes déduits de méthodologies d'étude des différents protocoles. Ces résultats sont systématiquement illustrés : sur l'hypercube en mode *store-and-forward* et sur la grille en mode *wormhole*.

Le problème du plongement de graphes dans le contexte des communications est motivé dans la première partie du chapitre 5. La deuxième partie, la plus importante du chapitre, traite de la notion classique de plongement dans l'aspect statique. Les principaux résultats connus sont donnés (complexité, paramètres des plongements de structures simples dans l'hypercube et les graphes de de Bruijn) et résumés dans le tableau final. La dernière partie aborde les aspects dynamiques de la notion de plongement.

Au chapitre 6 sont présentées deux applications typiques du calcul numérique utilisant des communications : le produit matriciel et le calcul de la transformée de Fourier discrète. Ces deux exemples illustrent la mise en œuvre de schémas de communication présentés dans les chapitres précédents.

Dans l'annexe A, une série de fiches synthétiques sur les réseaux usuels récapitule pour chacun ses propriétés principales (ordre, degré, nombre d'arêtes ou d'arcs, etc.).

L'annexe B commence par un rapide rappel sur les architectures des processeurs et une définition du vocabulaire nécessaire à la compréhension de tout document traitant des performances d'un processeur ou d'une machine parallèle. Puis, les processeurs les plus utilisés ainsi que les machines parallèles les plus récentes sont présentés. La description de ces machines est principalement axée sur leur réseau d'interconnexion et leurs propriétés liées aux communications.

Chaque chapitre contient une bibliographie et est illustré par des exercices dont énoncés et solutions sont regroupés à la fin du chapitre.

Quelques conventions et notations

Toutes les définitions, théorèmes, propositions, corollaires, etc. sont référencés indifféremment, dans leur ordre d'apparition au fil du texte. Un théorème est un résultat de portée générale (non nécessairement difficile à démontrer). Une proposition, au contraire, concerne des cas particuliers. Nous avons remplacé la plupart du temps les termes anglo-saxons par leur traduction française, cependant, les termes non traduits sont écrits en italiques. Nous rappelons ci-dessous des notations habituelles :

- $|X|$ désigne le cardinal de l'ensemble X.

- $\ln n$ désigne le logarithme népérien de n, et $\log_b n$ désigne le logarithme en base b de n.

- C_n^p et $\binom{n}{p}$ représentent le nombre de combinaisons de p éléments, pris parmi n éléments.

- Soient f et g deux fonctions réelles positives de variable entière :

  - La notation $f = O(g)$ signifie qu'il existe un réel positif c et un entier M tel que, pour tout x, $x \geq M$, on a $f(x) \leq cg(x)$.
  - La notation $f = \Omega(g)$ signifie qu'il existe un réel positif c et un entier M tel que, pour tout x, $x \geq M$, on a $f(x) \geq cg(x)$.
  - La notation $f = \Theta(g)$ est équivalente à $f = O(g)$ et $g = O(f)$.

Jean de Rumeur

Cet ouvrage est écrit par une équipe de chercheurs spécialistes de domaines différents mais complémentaires (théorie des graphes, placement et ordonnancement, algorithmique numérique, architecture, etc.). Ces auteurs ont uni leurs compétences, en tirant profit de plusieurs années de collaboration au sein de l'opération RUMEUR du PRC C^3. Ce livre fait suite à un support de cours rédigé pour une école d'été, organisée en août 1992 à l'Institut d'Etudes Scientifiques de Cargèse, pour des étudiants de troisième cycle. Il a été modifié et complété grâce aux nombreuses remarques et critiques faites au cours de cette école, ainsi que dans les cours de DEA dispensés par les auteurs.

Remerciements

Les auteurs tiennent à remercier particulièrement :

- leurs laboratoires de recherche, qui ont mis à leur disposition les moyens matériels nécessaires à la réalisation de cet ouvrage ;

- le Programme de Recherches Coordonnées C^3, qui a financé les réunions de RUMEUR qui ont permis aux auteurs de collaborer régulièrement ;

- le CIMPA, qui a aidé à organiser l'école d'été de Cargèse en 1992 ;

- les collègues et les étudiants qui les ont aidés dans la tâche ingrate de relecture, en particulier : Dominique Barth, Christophe Calvin, Charles Delorme, Odile Favaron, Rabah Harbane, Jean-Claude König, Stéphane Pérennes et Jean-François Saclé.