THESE DE DOCTORAT

Spécialité : Informatique

Soutenue publiquement par **Olivier GLÜCK**, le 12 juillet 2002.

« Optimisations de la bibliothèque de communication MPI pour machines parallèles de type grappe de PCs sur une primitive d'écriture distante »

Cette thèse a démarré en octobre 1999. Elle a été préparée au Laboratoire d'Informatique de Paris 6 (LIP6), dans le thème Architecture des Systèmes Intégrés et Micro-électroniques (ASIM), sous la direction du Professeur Alain Greiner, au sein de l'Ecole Doctorale d'Informatique, Télécommunications et Electronique de Paris, et financée par une allocation de thèse du Ministère de la Recherche.

Le travail qui y est présenté s'inscrit dans le cadre du projet de recherche MPC (Multi-PC) démarré en 1995 à l'Université Pierre et Marie Curie. Le but de ce projet est la réalisation d'une machine parallèle à faible coût : les nœuds de calcul sont des PC standard achetés dans le commerce auxquels s'ajoutent des composantes aussi bien matérielles que logicielles.

Cette thèse présente des optimisations de la bibliothèque de communication MPI pour machines parallèles de type « grappe de PCs », disposant d'un réseau de communication qui fournit une primitive d'écriture en mémoire distante (*Remote DMA*). Ce mécanisme de communication est implanté de manière très efficace au niveau matériel. Notre objectif est de faire bénéficier les applications de la très faible latence matérielle de ce réseau spécifique, en minimisant le temps de traversée des couches logicielles qui sépare l'appel à une primitive de communication au niveau applicatif de la prise en compte du transfert par le matériel réseau. Ce manuscrit de thèse présente, dans ce cadre, une implémentation optimisée de MPI au-dessus d'une primitive d'écriture distante. La machine MPC du LIP6 constitue notre plate-forme expérimentale.

Notre premier travail a été d'étudier les caractéristiques d'une écriture distante et de définir une interface de programmation générique d'écriture en mémoire distante sur laquelle nous avons construit nos couches de communication. Cette interface permet de porter facilement notre implémentation de MPI sur n'importe quelle plate-forme matérielle disposant d'une telle primitive de communication (le réseau Myrinet par exemple). Elle fournit une abstraction suffisante des caractéristiques spécifiques du réseau cible. Nous avons ensuite réalisé l'interface entre cette couche logicielle générique et le matériel réseau de notre plate-forme expérimentale (la machine MPC), constitué d'une carte PCI équipée de deux composants VLSI conçus au laboratoire. Cette interface rend accessibles les registres internes du contrôleur réseau afin de poster des ordres d'écriture en mémoire distante.

Nous avons alors réalisé une première implémentation de MPI (MPI-MPC1) au-dessus de la primitive d'écriture distante générique que nous avions définie. Après l'étude des services que nous devons fournir au niveau applicatif, nous proposons deux modes de transfert des données de l'application. Nous décrivons comment faire remonter les informations de signalisation des événements réseaux au niveau MPI. Nous évaluons les performances en termes de débit et de latence à l'aide d'un « *ping-pong* » MPI, sur notre plate-forme expérimentale. L'analyse de ces performances montre que les limites de cette première implémentation sont liées au fait que la primitive d'écriture distante se trouve dans le système d'exploitation et que la signalisation est réalisée par interruptions matérielles provenant du contrôleur réseau.

Nous proposons une deuxième implémentation de MPI (MPI-MPC2) qui utilise une primitive d'écriture distante en mode utilisateur et une signalisation par scrutation des ressources réseaux. L'accès en mode utilisateur à l'interface réseau pose des problèmes de sécurité et de partage des ressources de la carte réseau entre les différents processus de l'application. Nous proposons des mécanismes génériques permettant ce partage des ressources et nous décrivons comment nous avons appliqué ces solutions à la primitive d'écriture distante de la machine MPC. Nous analysons les problèmes liés à la scrutation des ressources réseaux. Nous réalisons de nouvelles mesures de performance avec MPI-MPC2. En comparant celles-ci à celles obtenues avec MPI-MPC1, nous constatons que la discontinuité des tampons de l'application en mémoire physique est pénalisante pour le transfert des messages de grande taille.

En effet, un inconvénient de la primitive d'écriture en mémoire distante réside dans le fait qu'elle utilise des adresses physiques pour réaliser ses transferts : le contrôleur réseau accède directement à la mémoire physique (DMA) sur le nœud émetteur et sur le nœud récepteur pour transférer les données. Il est donc nécessaire d'effectuer des conversions d'adresses (virtuelles/physiques) lors du transfert des données.

Pour palier à ce problème de discontinuité des données, lié aux caractéristiques des réseaux fournissant une primitive d'écriture distante, nous avons imaginé une méthode originale permettant de se ramener à une situation dans laquelle les régions mémoires de chaque processus de l'application correspondent à des zones de mémoire physique contiguës. Il s'agit de contourner les principes de gestion de la mémoire virtuelle telle qu'elle est faite classiquement dans les systèmes d'exploitation. L'intérêt de la solution proposée est qu'elle n'entraîne aucune modification, non seulement du système d'exploitation et de la librairie C, mais surtout de l'application. La méthode proposée est totalement transparente pour l'utilisateur. Cette solution permet de réduire au maximum le coût de traduction des adresses virtuelles fournies par l'application en adresses physiques utilisables par le contrôleur réseau. Elle a débouchée sur une troisième implémentation de MPI : MPI-MPC3. Nous avons suivi la même procédure de mesures que pour les deux premières. Alors que la deuxième apporte un gain significatif au niveau de la latence des transferts, la troisième montre une nette amélioration du débit qui permet d'approcher très significativement le débit matériel maximum du réseau cible.

Enfin, nous étudions l'impact de divers facteurs sur les performances obtenues au niveau applicatif. Des mesures de performances sur des applications réelles, utilisant les trois implémentations décrites précédemment, ont été réalisées. Cela nous a permis de vérifier que les conclusions générales qui ont pu être tirées de la comparaison des trois implémentations de MPI dans le cas du ping-pong restent valables dans le cas d'une application réelle.

Mots-clés : **machine parallèle, grappes de PCs, bibliothèque de communication, passage de messages, MPI, écriture distante, DMA, communication en mode utilisateur, gestion mémoire, adresse virtuelle/physique, traduction d'adresse**.

Composition du jury :
M. Bernard LECUSSAN : rapporteur, Professeur SupAero
M. Loïc PRYLLI : rapporteur, chargé de recherche CNRS, ENS Lyon
M. Jean-Marie CHESNEAUX : Professeur , directeur de l'IST
M. Paul FEAUTRIER : Professeur, ENS Lyon
M. Claude GIRAULT : Professeur, thème SRC du LIP6
M. Daniel MILLOT : Maître de conférences, Institut National des Télécommunications
M. Alain GREINER : Professeur, directeur du thème ASIM du LIP6