THESE DE DOCTORAT

Spécialité : Informatique

Présentée par **Olivier GLÜCK**

Pour obtenir le titre de docteur de l'Université Pierre et Marie Curie (Paris VI)

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

Soutenue publiquement le 12 juillet 2002 devant le jury composé de :

M. Bernard LECUSSAN	Rapporteur
M. Loïc PRYLLI	Rapporteur
M. Jean-Marie CHESNEAUX	Examinateur
M. Paul FEAUTRIER	Examinateur
M. Claude GIRAULT	Examinateur
M. Daniel MILLOT	Examinateur
M. Alain GREINER	Directeur de thèse

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, financée par une allocation de thèse du Ministère de la Recherche.

Résumé

Le travail présenté dans cette thèse 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 réalisées dans le cadre du projet.

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 Laboratoire d'Informatique de Paris VI constitue notre plate-forme expérimentale mais nous avons construit nos couches de communication au-dessus d'une API (*Applications Programming Interface*) générique d'écriture en mémoire distante permettant de porter facilement notre implémentation de MPI sur n'importe quelle plate-forme matérielle disposant d'une primitive d'écriture en mémoire distante.

Nous étudions l'impact de divers facteurs sur les performances obtenues au niveau applicatif et nous proposons des solutions optimisées pour implanter l'environnement de programmation parallèle MPI sur la primitive d'écriture distante. Plus particulièrement, nous décrivons des mécanismes pour réaliser les communications en mode utilisateur en éliminant les appels système et la signalisation par interruption matérielle. Un inconvénient de la primitive d'écriture en mémoire distante est 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. Les principales difficultés concernent le partage des ressources réseau entre plusieurs processus utilisateur et le problème des conversions d'adresses. Nous proposons une solution pour réduire au maximum le coût de traduction des adresses virtuelles fournies par l'application en adresses physiques utilisables par le contrôleur réseau.

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**.

« Optimizations of the Message Passing Interface communication library for PCs clusters using a remote write communication primitive »

Abstract

This Ph.D Thesis is a part of the MPC (Multi-PC) research project started in 1995 at Pierre et Marie Curie University, Paris. The goal was to design a low cost and high performance parallel computer. The MPC parallel computer consists of several processing nodes interconnected by a gigabit High Speed Link network.

This work presents how the Message Passing Interface (MPI) communication library can be optimized for a parallel computer made of clusters of workstations, providing a remote-write communication primitive. From the hardware point of view, this communication mechanism is very efficient. Our goal is to minimize the overhead of communication software layers used by applications for accessing the high speed network. This thesis focuses on an efficient and optimized implementation of MPI built on a simple Remote Direct Memory Access hardware primitive. For experimental purposes, the MPC parallel computer of LIP6 laboratory was used. However, our communication software layers were built over a generic remote-write API in order to port easily our MPI implementation on every hardware platform using a remote write primitive.

We study the impact of several factors on application performances and we propose efficient mechanisms to implement the Message Passing Interface on a remote DMA communication primitive. Precisely, we describe solutions to eliminate system calls and interrupts during communications. A drawback of the remote-write primitive is that it uses physical memory addresses for sending data: the network controller accesses directly the host memory on the sender node and the receiver node. The major difficulty of this work deals with the user-level accesses to the network interface by several processes and the address translations. We propose a mechanism to significantly reduce the overhead due to the translations of virtual addresses supplied by the applications in physical addresses used by the network controller.

Keywords : **parallel computer, PCs clusters, communication library, message passing, MPI, remote write, Direct Memory Access, user-level communication, memory management, virtual/physical address, address translation.**