Utilisation des couches de communication de LANDA pour la librairie MPI

David Gauchard

LAAS - CNRS

Toulouse - France

Grappes 2001

Plan

Le projet LANDA

Contexte

Modèle de communication: principe et implémentation

Communication locale

Communication distante et exemples d'implémentation

Implémentation sur MPICH-ADI2

Conclusions - Prospectives

Contexte

L'utilisation de machine parallèles ou de stations de travail interconnectées nécessite une gestion complexe de l'environnement système, qui alourdit la tâche du concepteur pour la réalisation de ses applications parallèles.

Les environnement parallèles ont été conçus pour aider ces utilisateurs à tirer profit des ressources qui leur sont offertes, et cela sans pour autant avoir besoin d'une grande connaissance des mécanismes sous-jacents.

Environnements parallèles

- **Cache l'architecture distribuée**
- **Aide les utilisateurs à développer des applications parallèles**
  interface graphique, analyse de trace , debugger, ...
- **Offre des bibliothèques normalisées**
  PVM, MPI, VIA

Le projet LANDA-HSN

LANDA - High Speed Network

Le projet a débuté en 1989 et a bénéficié de supports industriels (SUN, IBM, HP et Delta Partners), ainsi que de supports publics et militaires (CNRS, région Midi-Pyrénées et DRET).

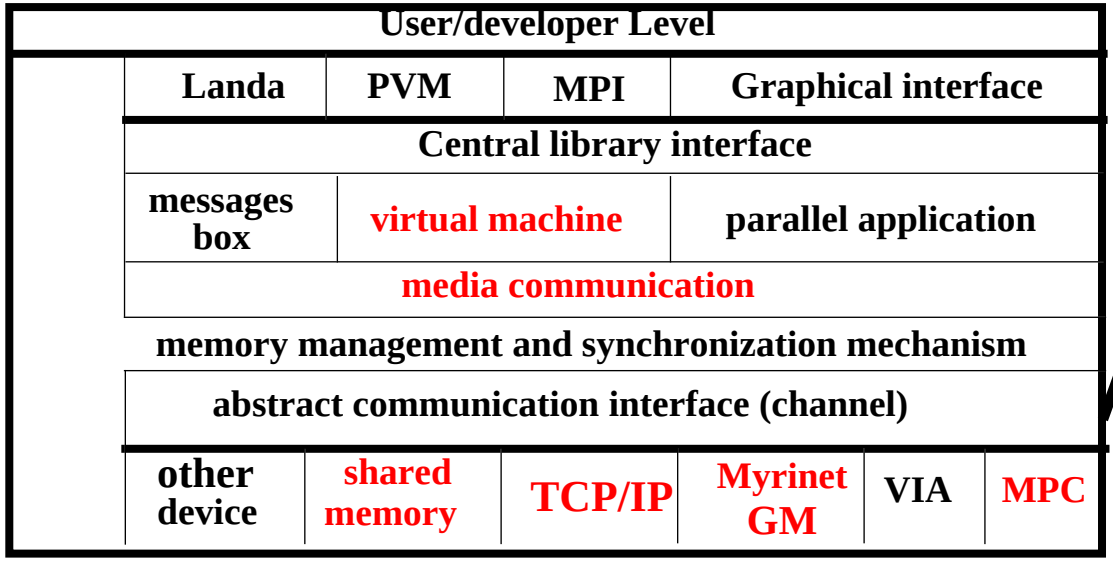
Spécificités

- **Environnement multi-sites (milieu hétérogène et multi-réseaux)**
- **Environnement complet pour le calcul haute performance**
- **Gestion intelligente des ressources:**
  - **Mémoire des différents ordinateurs**
  - **Processeurs**
  - **Réseaux (haut-débit)**
- **Création d'un outil spécifique: le Network-Analyser**
  - **Allocateur de tâches avec prediction dynamique de charge cpu et réseau**
  - **Recherche en cours sur la gestion du réseau et de la mémoire et sur de nouveaux algorithmes de placement**
- **Différentes politiques de stockages de messages**

Librairie de communication

But:

- **Une interface point-à-point abstraite simple permettant d'implémenter rapidement de nouvelles interfaces de communication (abstract communication interface)**
- **Librairie basée sur la mémoire partagée pour tirer profit des noeuds SMP**
- **Communication locale de type 0-copie (quand la sémantique le permet)**
- **Abstraction de l'hétérogénéité et mise à disposition de librairies standards (MPI)**



Interface point-à-point

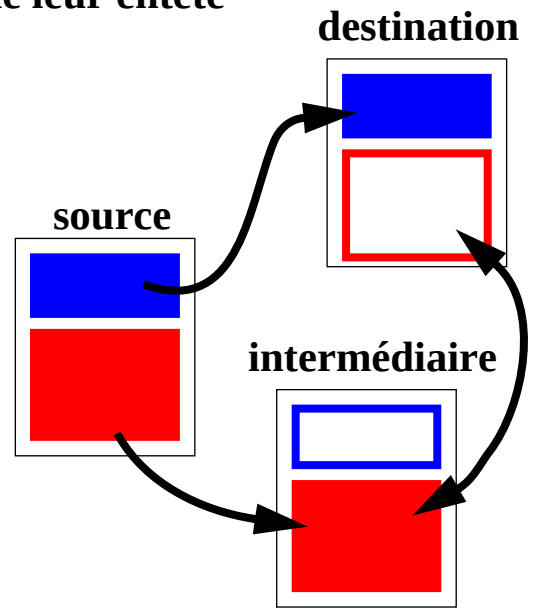
- **open/close**
- **envoi/reception bloquant ou non-bloquant**
- **test de complétion**

Principes du modèle de communication

- **Paradigme de passage de message**
  - **messages: entêtes et données**
  - **les données peuvent être localisées à un endroit différent de leur entête**
  - **l'entête est toujours reçu à destination**
  - **l'entête peut contenir les petites données**

- **Politiques de routage**

  les données sont:
  - **envoyées sur le noeud de réception**
  - **restent sur le noeud émetteur**
  - **sont stockées sur un troisième noeud**



- **Avantages**
  - **flexibilité du modèle de communication**
  - **routage dynamique**
  - **tient compte des noeuds les moins chargés**
  - **évite la duplication des données**
  - **permet l'utilisation efficace de la mémoire distribuée**

- **Inconvénients**
  - **Complexité**
  - **Overhead**

Communication locale

Allocateur distribué de mémoire partagé (adapté du malloc de Doug Lea (gnu-Libc))

Entêtes et données sont regroupées dans un buffer qui est l'unité d'allocation.

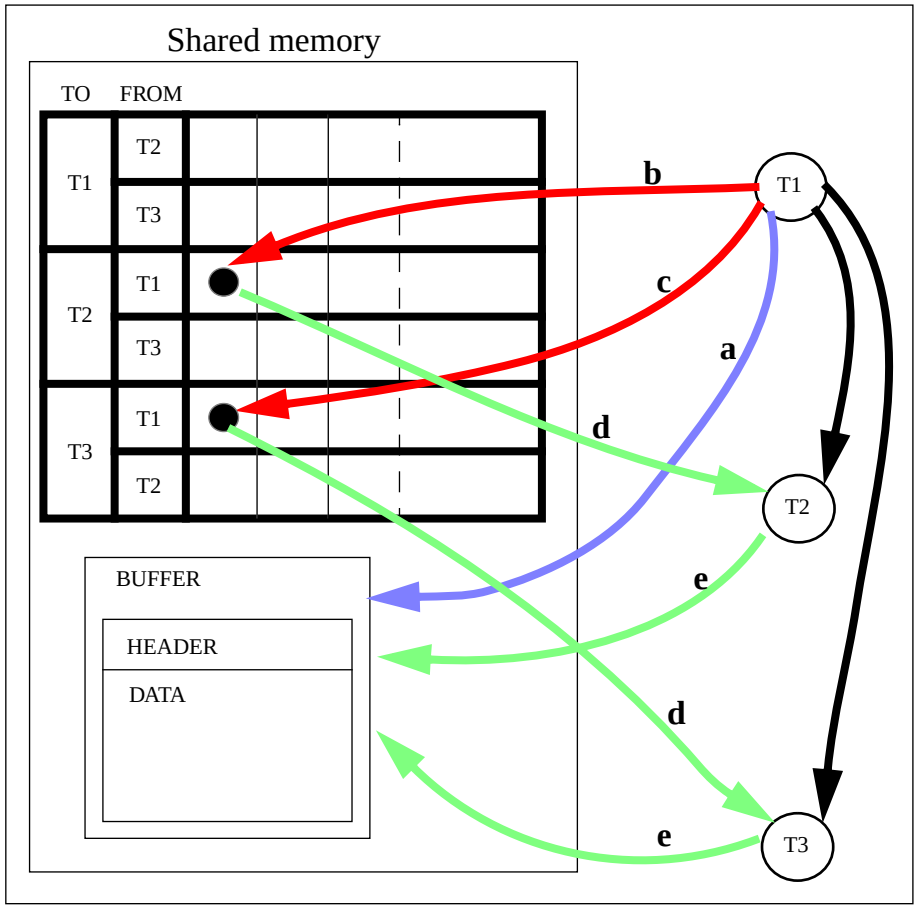
Synchronisation par spin-lock (pas d'appel système, latence basse)

Exemple de communication (reception en lecture seule) entre trois tâches.

T1 crée le message,

T1 positionne une référence dans les liste de reception de T2 et T3,

T2 et T3 peuvent alors lire leur référence puis leur données.



Performances des communications locales

Test sur bi-Pentium III @ 450Mhz

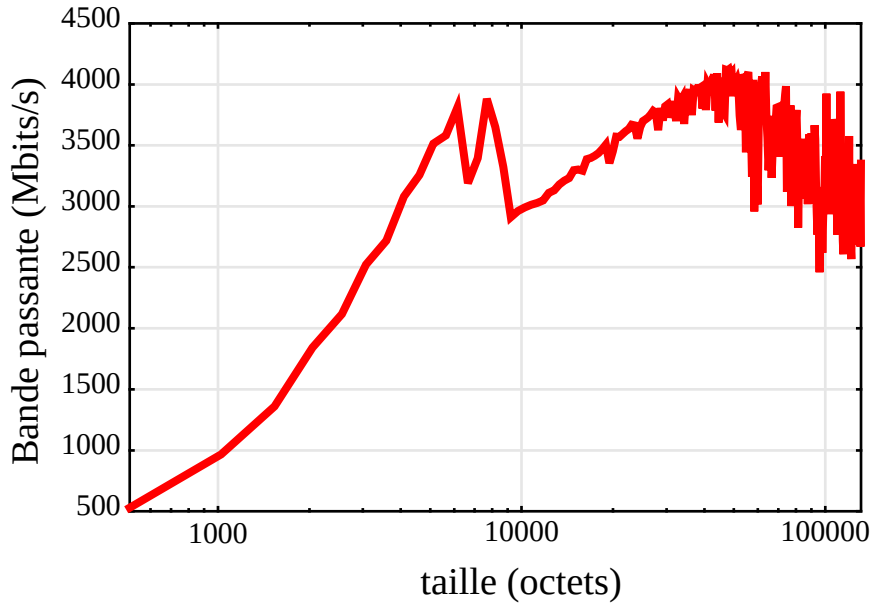
Les mesures incluent allocation et désallocation des buffers en SHM.

BP de 3Gbits/s avec paquets de 4Ko
BP maximum de 4Gbits/s

Demi-BP avec paquets de 2Ko

Latence inférieure à 4µs

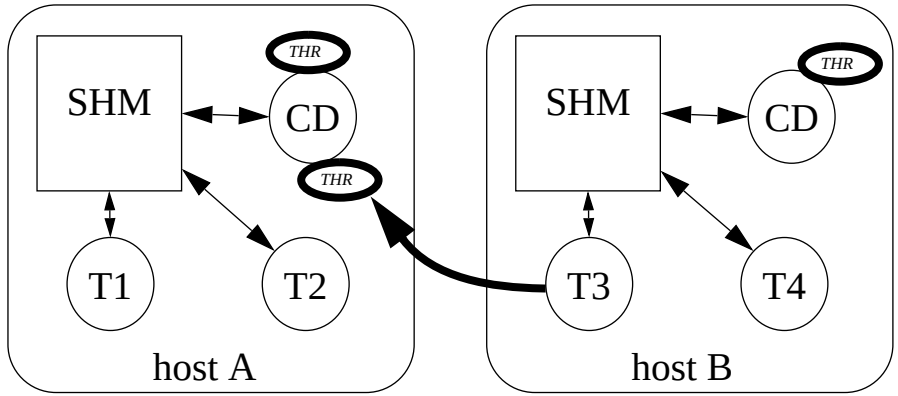
Bande passante d'un Ping-pong 1-copy (SHM + spinlock)



Communication distante

Un démon centralise les communications distantes: le CD (Communication Daemon)

Le schéma de communication distante est le suivant:



- **1 CD par noeud et par utilisateur**
- **Le CD reçoit toutes les données arrivant par le réseau et les passe aux tâches locales en utilisant les mécanismes locaux (SHM, mutex)**
- **Chaque tâche envoie ses messages directement au CD distant**
- **Les messages locaux ne passent pas par le CD (SHM, mutex)**

Communication Daemon

Pourquoi le CD ?

- **Les tâches utilisateur doivent être exempt de tout thread, mais les threads sont nécessaires pour faire de la communication asynchrone sur réseaux hétérogènes**
- **La latence introduite par les mécanisme locaux (4μs) est faible, la copie n'est pas nécessaire lorsque un message est transféré par le CD à une tâche.**
- **Réduction du nombre de liens d'interconnexion entre les noeuds**
- **Les tâches utilisateur n'ont qu'un seul médium de réception: la SHM**
- **Un démon comme le CD joue aussi d'autres rôles essentiels:**
  gestion de la mémoire, ressources et des tâches pour l'environnement d'exécution.

Inconvénient:

- **Le CD est un processus lourd, il ne doit en aucun cas faire de l'attente active.**

Deux implémentations

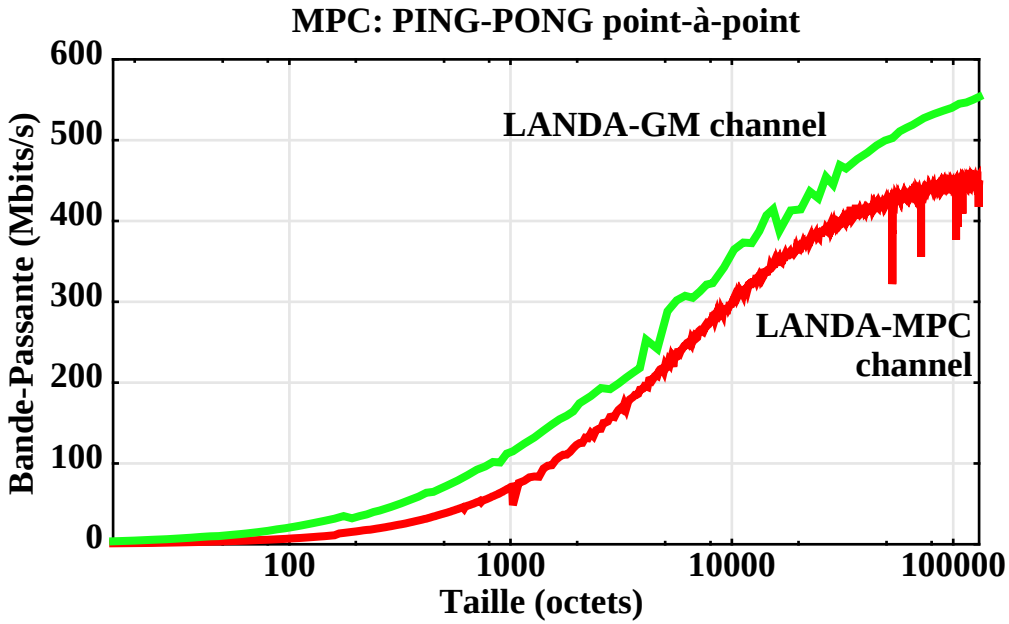
Myrinet/GM (Myricom): **Cartes Lanai4.3 PCI 32bits/33Mhz, ping-pong avec GM: 610Mbits/s - 22µs**

MPC (LIP6-ASIM): **Carte PCI 32bits/33Mhz, 495Mbits/s - 5µs (données driver PUT)**

Les tests ont été effectués sur un cluster Linux bi-PIII@450Mhz.

Myrinet (sur GM):
570Mbits/s
285Mbits avec des paquets de 5Ko
latence: 25µs (50µs paquets > 16Ko)

MPC (sur SLR/V):
450Mbits/s
225Mbits/s avec des paquets de 5Ko
latence: 65µs

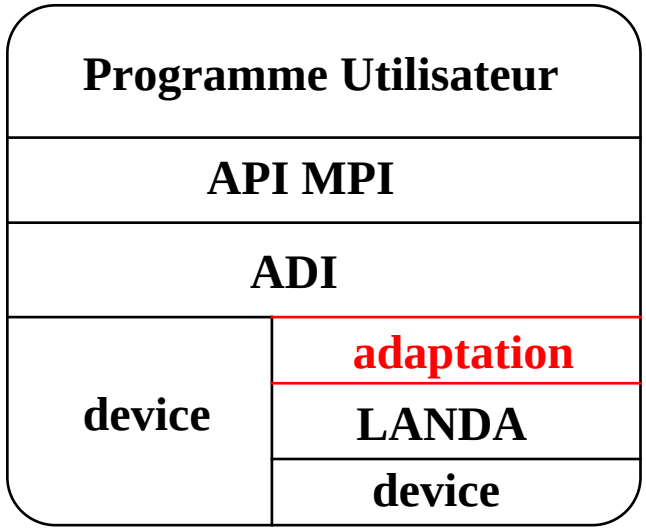


Adaptation de LANDA sous MPICH/ADI (ch2)

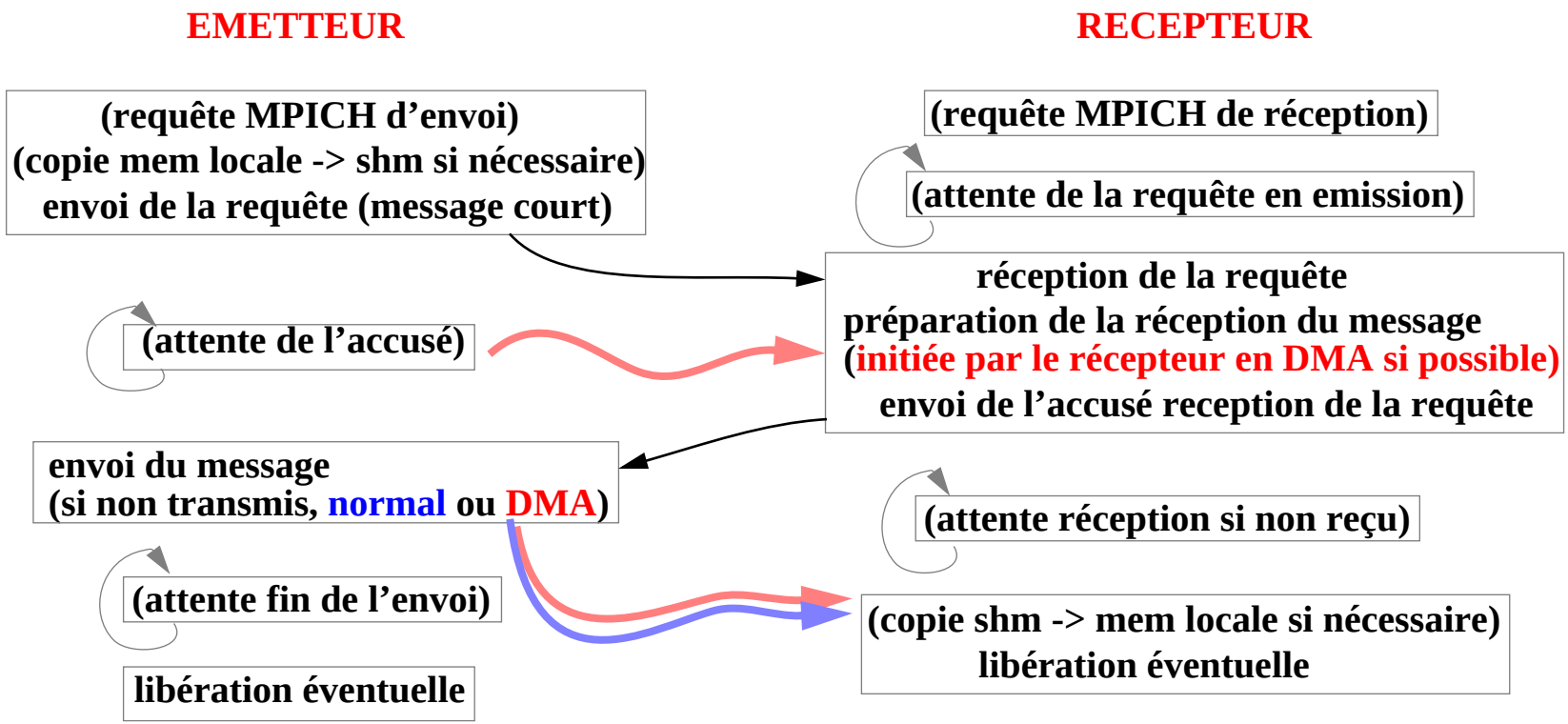
L'API LANDA est connectée à la couche ADI (Abstract Device Interface) fournie par MPICH.

Une couche d'adaptation entre l'ADI et LANDA est nécessaire:

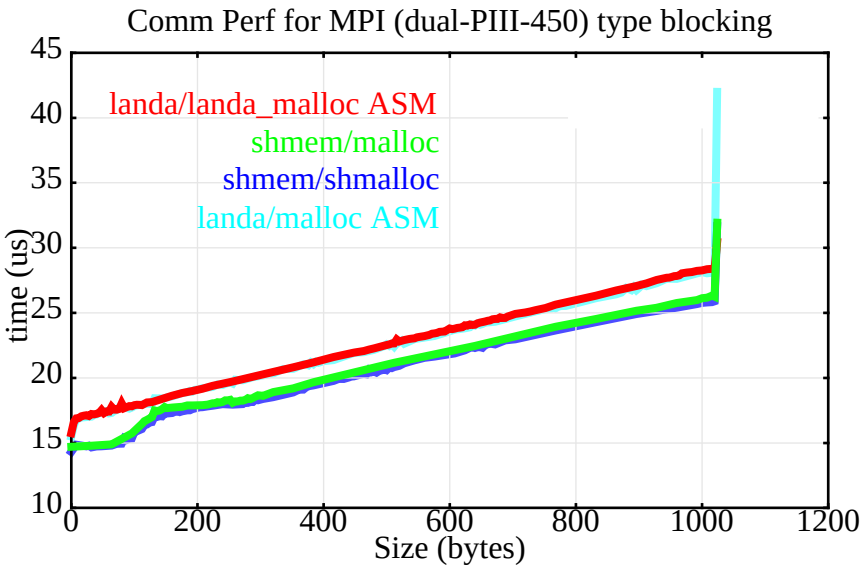
- **Une couche générique adaptable appelée ch2 est fournie dans la distribution de MPICH**
- **La méthode d'adaptation est inspirée de GM-MPICH**
- **Implémentation des "channels" courts bloquants et rendez-vous non-bloquant**
- **Généricité de l'implémentation: Les envois et/ou réceptions à distance par accès direct (SHM, DMA) doivent rester possible sans perte de performance**

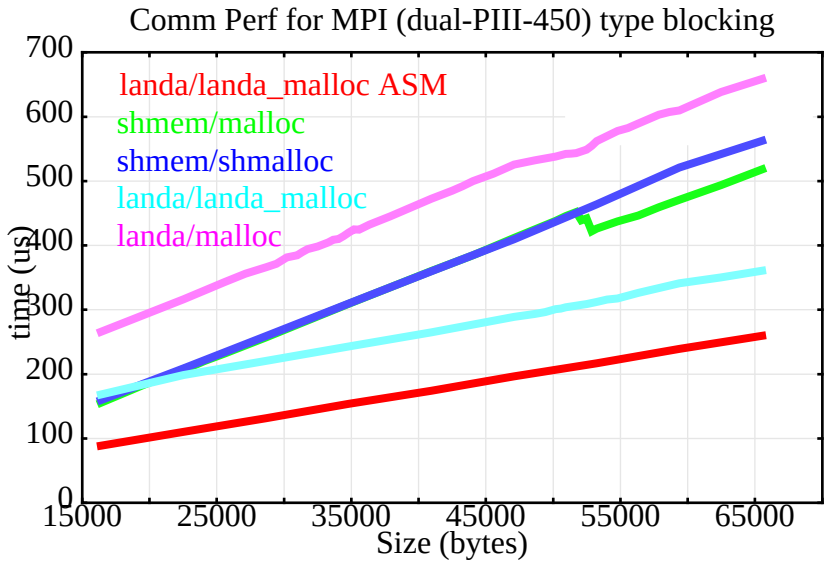


LANDA-MPICH: Mécanisme simplifié



Performances comparées





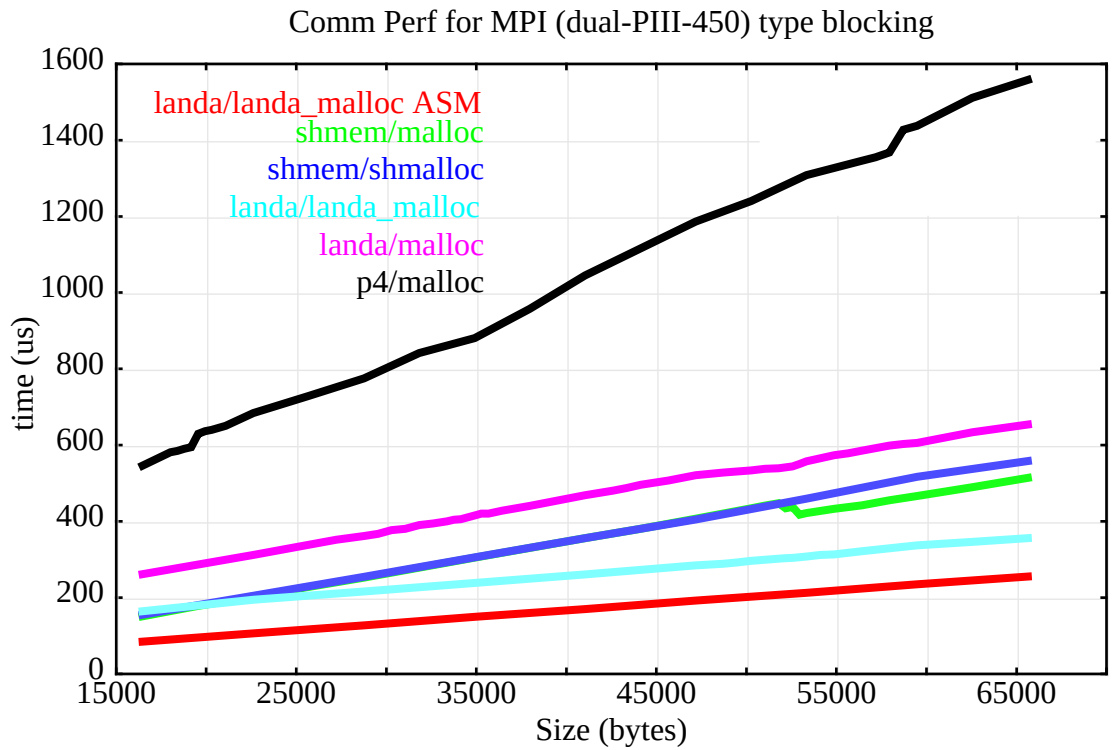
Pour les petits paquets, LANDA offre une latence supérieure de quelques microsecondes.

Le temps de recopie des petites paquets entre la mémoire locale et la SHM est négligeable.

LANDA est pénalisé par les sémaphores IPC.

Le temps de recopie des gros paquets n'est plus négligeable (différence malloc/landa_malloc)

Performances comparées



Conclusion et Prospectives

La couche de communication LANDA, avec les sémaphores actifs et la mémoire partagée semble prometteuse. L'adjonction de device type réseau avec ou sans accès DMA reste à faire.

Concernant les petits paquets, une optimisation du portage de l'ADI2 est en cours et devrait permettre de gagner le retard face à MPICH-SHM.

Une intégration de cette nouvelle version du noyau de communication de LANDA dans l'environnement gestionnaire d'applications avec interfaces graphiques, ainsi que l'adaptation de la couche Myrinet et TCP/IP sont en cours et devrait voir le jours d'ici quelques mois.