Progression à base de tâches des communications asynchrones

Florian Reynier

Progression à base de tâches des communications asynchrones

Auteur : Florian Reynier[1,2]
Directeurs de thèse : Emmanuel Jeannot[2] et Alexandre Denis[2]
Encadrant CEA : Julien Jaeger[1]
[1]CEA, DAM, DIF, F-91297 Arpajon, France
[2]Inria Bordeaux – Sud-Ouest, France

DE LA RECHERCHE À L'INDUSTRIE

INVENTEURS DU MONDE NUMÉRIQUE

université de BORDEAUX

CONTEXTE

CONTRIBUTIONS

LES COMMUNICATIONS MPI

1

- **Standard très utilisé en calcul haute performance**
  - Communication par messages
  - Programmation sur mémoire distribuée
- **Communications non bloquantes**
  - L'appel rend directement la main
  - Communications en "tâches de fond"
  - Gain de temps processeur
- **Opérations collectives**
  - Opérations basique sur plus de deux noeuds
  - Broadcast, reduce, gather ...

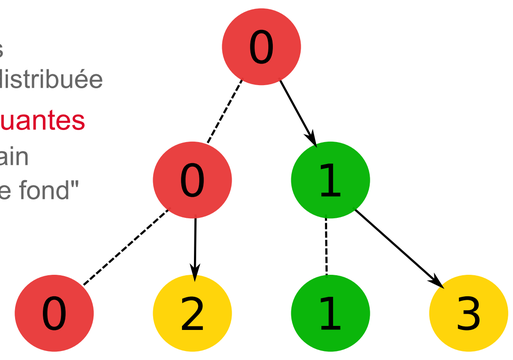


fig1: MPI_Bcast sur 4 noeuds

PROBLEMATIQUE

2

- **Recouvrir les communications par du calcul**
  - Paralléliser les communications et le calcul
  - Minimiser leur impact mutuel
- **Comment bien gérer les ressources disponibles?**
  - Répartir la communication et le calcul
  - Quelle priorité ?
  - Quelle politique de répartition ?

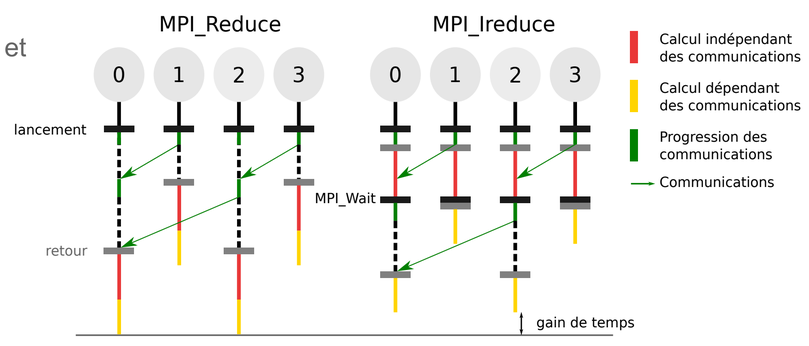
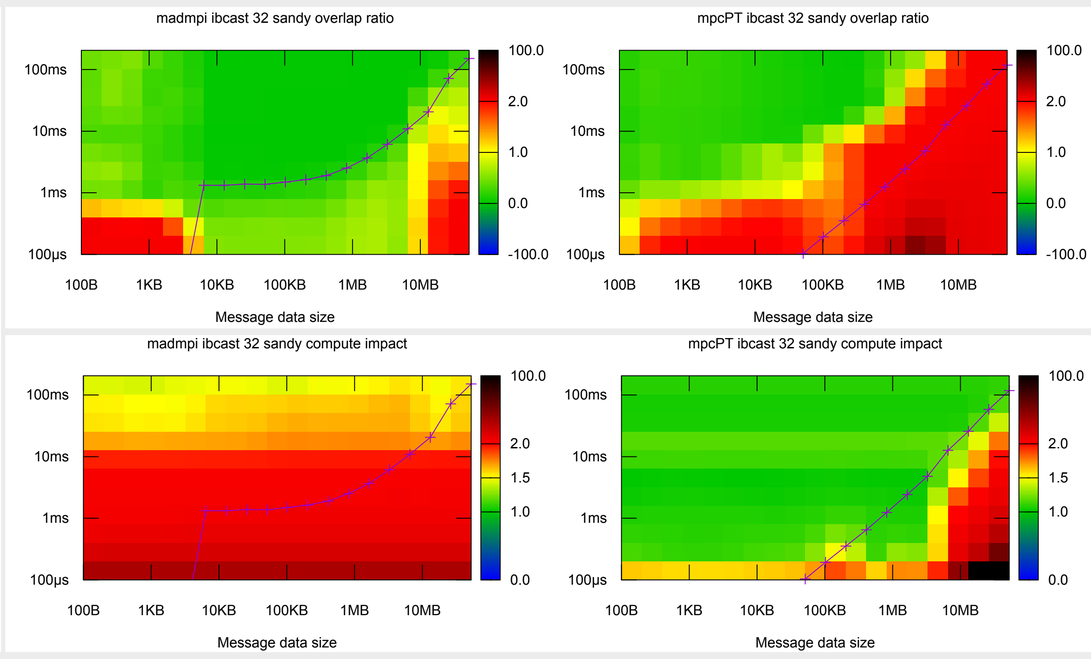


fig2: Comparaison d'un reduce bloquant / non bloquant

BENCHMARKING DES COLLECTIVES NON BLOQUANTES

3

- **Synchronisation des horloges**
  - Paralléliser les communications et le calcul
  - Minimiser leur impact mutuel
- **Mesurer une collective**
  - Paralléliser les communications et le calcul
  - Minimiser leur impact mutuel



- **Ratio d'overhead**
  - Mesure la paralellisation du calcul et des communication
  - $overhead = \frac{t_{measured} - max(t_{comm}, t_{coll})}{min(t_{comm}, t_{coll})}$
  - **ratio = 0:** recouvrement parfait
    ratio = 1: comm et calcul serialisé
    ratio > 1: perte de temps
- **Ratio d'impact sur le calcul**
  - Mesure l'impact de la parallelisation sur le temps de calcul
  - $impact = \frac{t_{calc}}{t_{calc\text{-}pure}}$
  - **ratio = 1:** aucun impact sur le calcul
    ratio > 1: le calcul est impacté

NEW MADELEINE[2]

4

- **Une bibliothèque de communication**
  - Conçue sur le principe des communications asynchrones
  - Optimisation reposant sur la gestion dynamique de paquets
  - Modulable: permet d'experimenter diverses stratégies d'ordonnancement

Une bibliothèque de communication

  - Implante le modèle à base de tâches
  - Les tâches intermédiaires sont créées en fonction des évènements réseaux

- **MadMPI: une implémentation MPI basée sur New Madeleine**
  - Permet le fonctionnement d'applcation MPI avec NewMadeleine
  - Basé sur quatres opérations basiques du standard, commune avec NewMadeleine:
    - MPI_Isend, MPI_Irecv
    - MPI_Wait, MPI_Test
  - Implémentation de collectives non bloquantes: MPI_Ibcast, MPI_Ireduce

MPC[1]

5

- **Un environnement d'éxécution unifié**
  - Implémentation OpenMP, MPI, Pthread
  - Implémentation MPI basée sur des threads
  - Threads utilisateurs: gérés par un ordananceur propre à MPC
  - Optimisé pour le fonctionnement simultanés des modèles de programmation

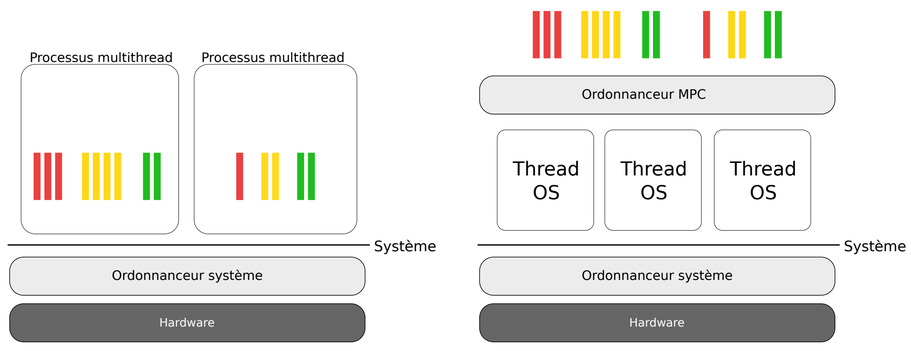


fig3: fonctionnement multimodèle classique / MPC

- **Ajout d'un moteur de tâches au sein de MPC**
  - Basé sur des tâches légères
  - Gestion de dépendances pour les collectives
  - Moteur basé sur les threads utilisateurs MPC

BENCHMARKING DES COLLECTIVES NON BLOQUANTES

6

- **Définition d'une tâche**
  - Une tâche peut être définie comme une portion de code à éxécuter
  - Contrairement au thread, elle n'est pas liée à un contexte d'exécution
  - Dans notre cas: des tâches légères
- **Un modèle léger et souple**
  - Les tâches ne sont liées à aucun coeur, elles peuvent être plus finement réparties dynamiquement à l'exécution.
  - Utile dans le cas où le code n'est pas très lourd pour le processeur, ce qui est le cas pour des communications

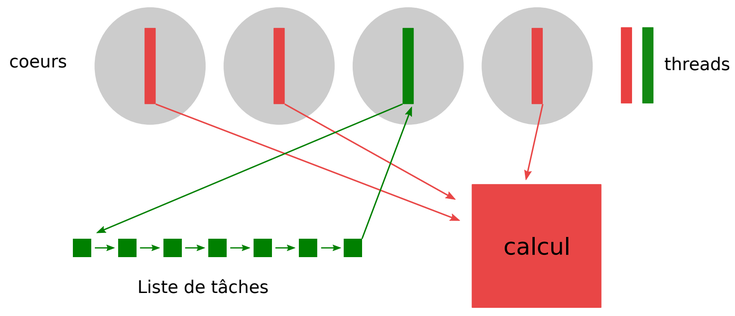


fig4: Modèle à base au sein d'un noeud

[1] Pérache (M.), Jourdren (H.) et Namyst (R.). – MPC : A Unified Parallel Runtime for Clusters of NUMA Machines. – In Springer (édité par), the 14th International Euro-Par Conference, LNCS, volume 5168, pp. 78–88, Las Palmas de Gran Canaria, Spain, août 2008
[2] Aumage (O.), Brunet (E.), Furmento (N.), Namyst (R.) - NewMadeleine: a Fast Communication Scheduling Engine for High Performance Networks - Workshop on Communication Architecture for Clusters (CAC 2007), workshop held in conjunction with IPDPS