Diffusion avec MPI

Résumé: Dans ce TD, nous allons nous familiariser avec les fonctions de communications de MPI et essayer d'écrire nous-même la fonction de diffusion à partir d'envois et de receptions simples. Nous allons également nous rendre compte de la difficulté de l'écriture d'une opération de diffusion non bloquante avec MPI et ainsi souligner bon nombre de limitations de MPI .

1 Un mot d'histoire

Tout d'abord, il convient de rappeler quelques généralités sur ce qu'est MPI et sur ce que ce n'est pas. MPI n'est pas une bibliothèque de communication développée par une université ou une entreprise : c'est un standard qui est né au début des années 90 pour répondre à un besoin de clarification. En effet, à cette époque, la seule façon de faire du calcul parallèle efficace consistait à acheter une grosse machine parallèle propriétaire. Ces dernières étaient généralement livrées avec leur propre bibliothèque de communication qui n'était que rarement compatible avec celle de la machine précédente et jamais avec celle des concurrents. Il était donc très difficile de maintenir un programme à jour et un gros travail était nécessaire à chaque fois que l'on souhaitait changer de machine. Le standard MPI est donc né de la collaboration entre des universitaires et des industriels de tous domaines scientifiques. Cependant, si ce standard a permit de résoudre la majorité des problèmes que l'on pouvait avoir au moment de sa création, il souffre désormais d'un bon nombre de limitations et d'une inadéquation aux plateformes de calcul actuelles.

En effet, la mode actuelle en matière de calcul distribué et parallèle est à l'interconnection, via des liens à très haut débit, de grands centres de calculs disséminés à l'échelle d'un pays, d'un continent voire du monde. C'est ce que l'on appelle le *meta-computing*. Personne ne sait si on pourra un jour exploiter correctement de telle plateformes mais, dixit Al. Geist (responsable du projet PVM) il y a quelques jours[1] : "Il y a de toutes façons tellement d'argent investi dans ce domaine que cela aboutira forcément à quelque chose".

Mais revenons à MPI . Un des gros problème auquel on se heurte dès que l'on essaie d'utiliser un programme MPI sur une plateforme de *meta-computing* est que les bibliothèques MPI d'un centre de calcul à l'autre ne sont pas forcément identiques et rien dans le standard n'oblige une bibliothèque à perdre en performance pour être compatible avec une bibliothèque concurrente. À cela s'ajoutent la difficulté de déploiement d'un tel programme, la nécessité de prendre en compte (autant au niveau algorithmique qu'au niveau système) l'hétérogénéité des processeurs et des réseaux, le coût énorme des synchronisations et donc des opérations bloquantes, la non-tolérance aux pannes,...Même s'il serait plaisant de pouvoir programmer et utiliser ces plateformes comme une simple station de travail, nous en somme très loin actuellement et leur complexité semble rendre cette entreprise un peu utopique.

Ce portrait de MPI et du calcul parallèle actuel peut sembler un peu pessimiste mais tous ces projets ont aussi donné naissance à de très bonnes choses. Même si MPI n'est plus adapté aux plateformes de calcul telle qu'elles sont envisagées actuellement, l'aventure MPI est quand même un succès étant donné qu'elle a permis à bon nombre de personnes non informaticiennes de développer des programmes efficaces pour les plateformes de calcul parallèle classiques et de collaborer grâce à des codes portables. L'émulation créée par ces projets ambitieux de *global-computing* permet à bon nombre de scientifiques non informaticiens de résoudre des problèmes qu'ils n'auraient jamais pu espérer résoudre il y a de cela quelques années. Enfin, même si tous ces projets semblent un peu fous et extrêmement ambitieux, il est indéniable que la communauté scientifique a un besoin toujours croissant de puissance de calcul et de communication et que nous ne sommes actuellement pas en mesure de répondre à leurs besoins.

Un nouveau standard MPI -2 a été mis en place il y a quelques années et résout certains des problèmes précédemment évoqués. Il existe cependant encore très peu de bibliothèques mettant en œuvre l'intégralité du standard MPI -2.

[1] Et oui, on voit du beau monde dans les confs...

2 Introduction à l'utilisation de MPI

2.1 Initialisation et terminaison du programme

Un programme MPI commence en général par un appel de la fonction `MPI_Init` dont le prototypage est le suivant :

```
int MPI_Init(int *argc, char ***argv)
```

Cette fonction initialise les connections MPI en fonction des arguments passés à votre programme. C'est pourquoi avant de lire les arguments de votre programme, il convient de faire un appel à cette fonction.

Un programme MPI se termine généralement par l'appel de la fonction `MPI_Finalize`. Tous les processus doivent appeler ce programme avant leur terminaison. Cette opération est bloquante ne termine que lorsque toutes les opérations de communications en cours ou en attente sont terminées.

Il peut être utile de savoir combien de processus participent au calcul et quel numéro on a. Les fonctions `MPI_Comm_size` et `MPI_Comm_rank` dont le prototype est le suivant. Un `MPI_Comm` est un groupe de processus MPI . Pour l'instant, nous n'utiliserons que le groupe prédéfini `MPI_COMM_WORLD` qui regroupe l'intégralité des processus MPI participant au calcul.

```
int MPI_Comm_size ( MPI_Comm comm, int *size )
int MPI_Comm_rank ( MPI_Comm comm, int *rank )
```

2.2 Communications bloquantes

Les envois et les réceptions bloquantes se font grâce aux fonctions `MPI_Send` et `MPI_Recv` dont le prototype est le suivant :

```
int MPI_Send( void *buf, int count, MPI_Datatype datatype, int dest,
              int tag, MPI_Comm comm )
int MPI_Recv( void *buf, int count, MPI_Datatype datatype, int source,
              int tag, MPI_Comm comm, MPI_tatus *status )
```

Quelques explications :

- `buf` représente l'adresse du buffer d'émission (pour `MPI_Send`) ou du buffer de réception (pour `MPI_Recv`).
- `count` est le nombre d'éléments à envoyer ou à recevoir.
- `datatype` est le type des éléments que l'on va envoyer (ou recevoir). On doit préciser le type des éléments car MPI peut effectuer des conversions de type si les architectures cibles ne représentent pas les objets de la même manière (*big indian/little indian...*). Il est possible de créer des types MPI de façon à faciliter la programmation mais nous n'utiliserons pour l'instant que le type `MPI_INT`.
- `dest` et `source` sont respectivement les indices des processeurs destinataires et sources mais vous l'aviez probablement deviné tout seul.
- `tag` est un nombre qui sert généralement à identifier un type de message. Choisissez-en un et essayez de vous y tenir...
- `comm` est le groupe de processus impliqués dans la communication. Comme cela a été signalé précédemment, nous nous contenterons de `MPI_COMM_WORLD`.
- `status` est un objet permettant de récupérer des informations sur la communication et ne nous servira pas pour l'instant.

Enfin une autre fonction qui s'avère souvent utile : la barrière de synchronisation.

```
int MPI_Barrier ( MPI_Comm comm )
```

Nous allons essayer de simuler la fonction `MPI_Bcast` dont le prototype est le suivant.

```
int MPI_Bcast ( void *buffer, int count, MPI_Datatype datatype, int root,
                MPI_Comm comm )
```

Cette fonction est bloquante. Lorsqu'un programme alterne des phases de calcul et des phases de communications, il y a donc intérêt à ce que l'équilibrage de charge soit parfait...

2.3 Communications non bloquantes

Les fonctions d'envoi et de réception non bloquantes `MPI_Isend` et `MPI_Irecv` ont le prototype suivant :

```
int MPI_Isend( void *buf, int count, MPI_Datatype datatype, int dest, int tag,
               MPI_Comm comm, MPI_Request *request )
int MPI_Irecv( void *buf, int count, MPI_Datatype datatype, int source,
               int tag, MPI_Comm comm, MPI_Request *request )
```

`request` est un *"numéro"* de communication. Cela permet de savoir si une communication est terminée ou pas, notamment grâce à la fonction `MPI_Test` ou à la fonction `MPI_Wait`.

```
int MPI_Test ( MPI_Request *request, int *flag, MPI_Status *status)
int MPI_Wait ( MPI_Request *request, MPI_Status *status)
```

`tag` reçoit la valeur vrai si la communication est effectivement terminée.

2.4 Lancement du programme

Un programme MPI minimal (`simple.c`) ainsi qu'un makefile permettant de compiler (en tapant `make...`) et de lancer des programmes MPI (en tapant `make run`) sont disponibles dans `/home/alegrand/MPI-MIM2/`. Pour pouvoir compiler tranquillement vos programmes, il vous faudra aussi *"sourcer"* les fichiers `csh/base` puis `csh/mpich` disponibles dans le même répertoire. Cela vous permettra d'avoir accès aux bibliothèques et aux headers de MPI . La bibliothèque MPI que nous allons utiliser s'appelle MPICH et est disponible sur une grande variété de plateformes.

3 Diffusion en MPI

▷ **Question 1.** Écrivez un programme où le processeur 0 diffuse aux autres processeurs un tableau d'entiers à l'aide de la fonction `MPI_Bcast`. Rajoutez un appel à la fonction `long_computation` juste après la diffusion et mesurez le temps nécessaire à ces opérations grâce à la fonction `ms_time`[2].

▷ **Question 2.** Recommencez en écrivant vous même la diffusion à l'aide d'opérations bloquantes. Que pensez-vous des performances, notamment quand vous faites varier le nombre de processus ?

▷ **Question 3.** Remplacez maintenant vos émissions bloquantes par des émissions non bloquantes. Que pensez-vous des performances ?

▷ **Question 4.** Rajoutez une barrière de synchronisation juste avant l'appel à `long_computation`. Que pensez-vous des performances ? Avez-vous une explication à proposer ?

▷ **Question 5.** Essayer de remédier au problème en utilisant autre chose qu'une barrière de synchronisation.

L'objectif de ce TD était de mettre en valeur un certain nombre de limitations de MPI . Il ne faut quand même pas oublier que ces bibliothèques sont quand même très efficaces sur des plateformes homogènes, qu'elles offrent une quantité d'opérations globales impressionnantes ainsi qu'un haut niveau d'abstraction, ce qui permet de développer très rapidement un programme parallèle.

[2] ces deux fonctions sont disponibles dans le programme MPI minimal évoqué précédemment