Les architectures parallèles
~~~~~
Les machines parallèles de l'observatoire
~~~
MPI

Aurélia Marchand
Aurelia.Marchand@obspm.fr

Plan

-1- Introduction

Le parallélisme est la conséquence :

- besoin des applications
  - calcul scientifique
  - traitement d'images
  - bases de données

qui demandent des ressources en CPU et en temps de calcul de plus en plus importantes

- limites des architectures séquentielles
  - performance
  - capacité d'accès à la mémoire
  - tolérance aux pannes

Définition :

Les ordinateurs parallèles sont des machines qui comportent une architecture parallèle, constituée de plusieurs processeurs identiques, ou non, qui concourent au traitement d'une application.

La performance d'une architecture parallèle est la combinaison des performances de ses ressources et de leur agencement. (latence, débit)

Architectures parallèles :

=> pas de limite de mémoire

=> pas de limite de processeurs

=> accélération des calculs complexes ou coûteux en temps d'occupation CPU

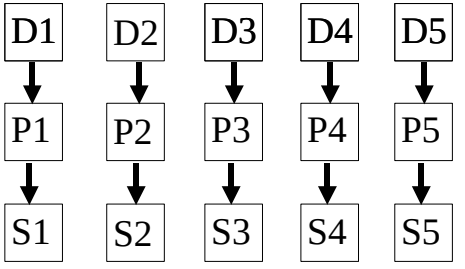
(calcul matriciel, simulation numérique, transformée de fourier...)

=> calcul répétitif sur un large ensemble de données structuré

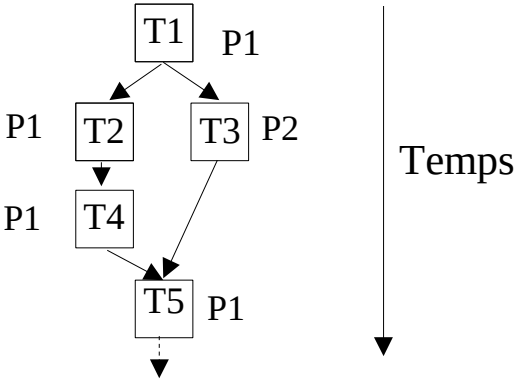
=> traitement indépendant

On peut avoir différentes sources de parallélisme :

- le parallélisme de données : la même opération est effectuée par chaque processeur sur des données différentes.



- Parallélisme de contrôle : des opérations sont réalisées simultanément sur plusieurs processeurs. Le programme présente des séquences d'opérations indépendantes qui peuvent être exécutées en parallèle.



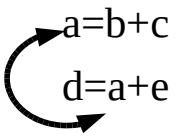
- Parallélisme de flux : Les opérations sur un même flux de données peuvent être enchaînées (pipeline)

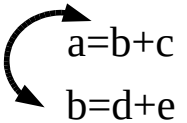
temps	Tâche 1	Tâche 2	Tâche 3	Tâche 4	Tâche 5
t1	Donnée 1				
t2	D2	D1			
t3	D3	D2	D1		
t4	D4	D3	D2	D1	
t5	D5	D4	D3	D2	D1
t6	D6	D5	D4	D3	D2

limites du parallélisme :

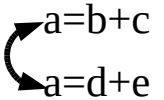
- les dépendances de données :

Dépendance de flot





Dépendance de sortie



=> les instructions doivent être indépendantes :

- exécuter dans un ordre quelconque
- simultanément

- les dépendances de contrôle :

I1	a=b+c;	I1 et I4 sont à priori indépendantes
I2	if (a<0)	mais selon la valeur de la variable 'a'
I3	{d=e+f;}	I3 peut être exécuté ou non
I4	g=d+h;	entraînant une dépendance entre I3 et I4

- les dépendances de ressources :

  nombre insuffisant de processeurs pour effectuer les instructions en parallèle alors qu'elles sont indépendantes les unes des autres.

- rapport temps de communication / temps de calcul :

Il n'est pas toujours avantageux de paralléliser une application. Les communications peuvent dans certain cas augmenter le temps d'exécution.

Mesure de performance :

débit de traitement : P= nb_op/tps

accélération : $a=t_{séquentiel}/t_{parallèle}$

efficacité : e=a/nb_proc

-2- Les architectures parallèles

-2.1- Les types d'architectures parallèles

Classification des architectures parallèles :

	1 flux d'instruction	> 1 flux d'instruction
1 flux de données	SISD	MISD (pipeline)
> 1 flux de données	SIMD	MIMD

S : Single, M : Multiple

I : Instruction, D : Data

SISD : une instruction - une donnée

machine séquentielle (modèle de Von Neumann)

SIMD : plusieurs données traitées en même temps par une seule instruction.

Utilisé dans les gros ordinateurs vectoriels.

Première machine parallèle : l'ILLIAC IV (1966)

MISD : Une donnée unique traitée par plusieurs instructions.

Architecture pipeline.

MIMD : exécution d'une instruction différente sur chaque processeurs pour des données différentes.

Pour simplifier la programmation, on exécute la même application sur tous les processeurs. Ce mode d'exécution est appelé : SPMD (Single Program Multiple Data)

Il existe deux types de machine parallèle :

- les machines parallèles à mémoire partagée
- les machines parallèles à mémoire distribuée

Les machines vectorielles

-2.2- Les machines vectorielles

Ce sont des machines multiprocesseurs, et donc parallèles.

La vectorisation consiste à effectuer une instruction sur un groupe de données plutôt que d'effectuer successivement cette opération sur chacune de ces données.

Comment fonctionne une machine vectorielle ?

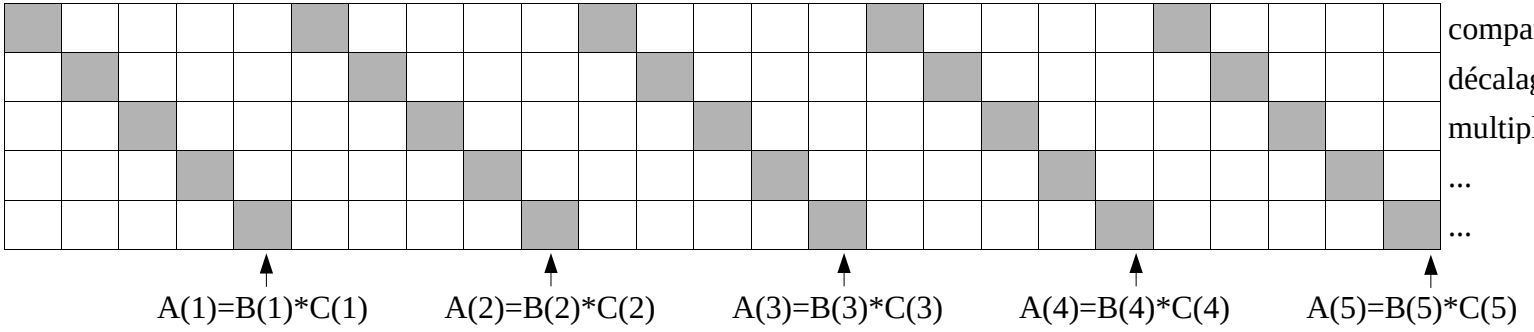
Elle effectue en parallèle, les sous-opérations d'une opération. Pour cela on décompose une instruction en plusieurs étapes de traitement élémentaire et on utilise le principe du pipeline.

Exemple :

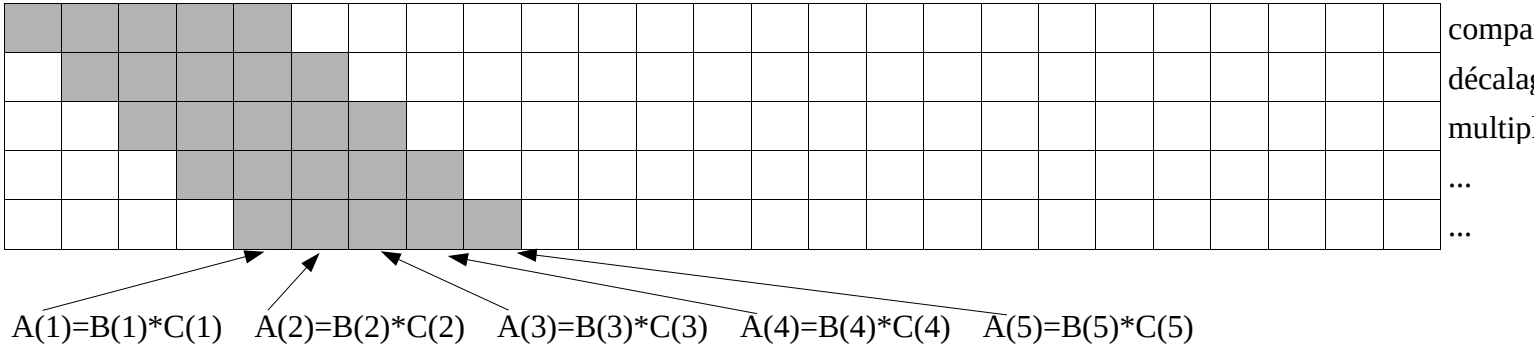
```
do i=1,n
        A(i)=B(i)*C(i)
enddo
```

L'instruction A(1)=B(1)*C(1) nécessite plusieurs sous-opérations (comparaison exposants, décalage des mantisses, multiplication des mantisses, ...)

Exécution sur une machine scalaire :



Exécution sur une machine vectorielle :



Nombre d'opérations :

- scalaire : k*n
- vectorielle : n+k-1

On a intérêt à vectoriser des tableaux de taille suffisamment grande pour que la vectorisation soit efficace.

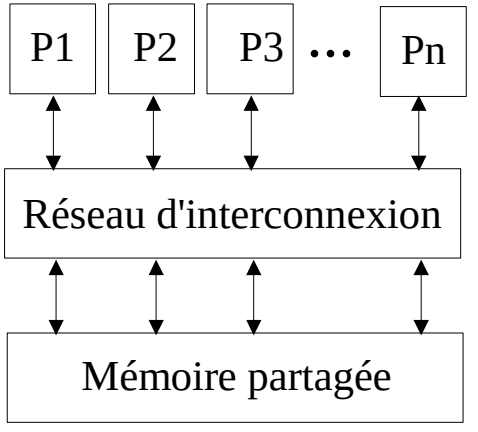
Accélération : $a = \frac{k * n}{n + k - 1}$

La vectorisation se fait sur les boucles.

Inhibition de la vectorisation :

- appel à une fonction
- tout ce qui casse la structure de la boucle (goto, if...)
- instructions d'entrées/sorties

-2.3- Les machines à mémoire partagée



Caractéristiques :

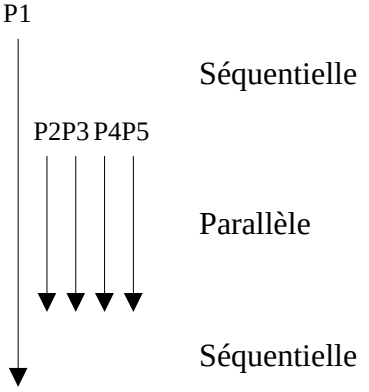
- plusieurs processeurs avec des horloges indépendantes
- une seule mémoire commune à tous les processeurs
- programmable par le standard portable OpenMP

Les machines à mémoire partagée permettent de réaliser le parallélisme de données et de contrôle.

Le programmeur n'a pas besoin de spécifier la distribution des données sur chaque processeur. Il définit seulement la partie du programme qui doit être parallélisée (directives) et doit gérer les synchronisations.

Exemple de parallélisation avec OpenMP:

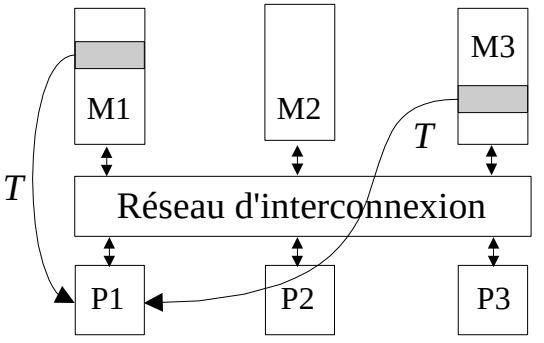
```
w=2
!$OMP PARALLEL DO PRIVATE(x), SHARED(w)
!$OMP REDUCTION(+:sum)
do i=1,100
        x=w*i
        sum=sum+x
enddo
print *,'Somme=',sum
```



Différentes organisations possibles des machines à mémoire partagée :

- UMA :
  Uniform Memory Access

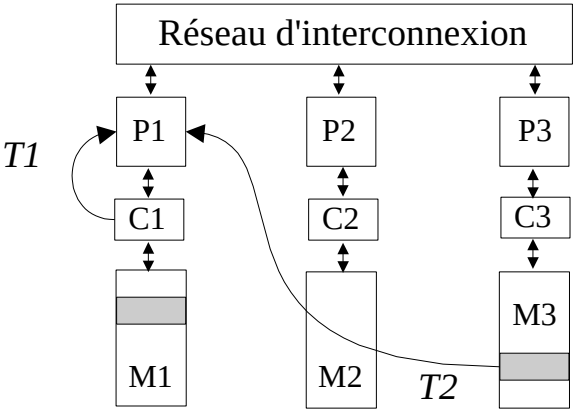
  => architecture vectorielle



T : temps d'accès à la mémoire

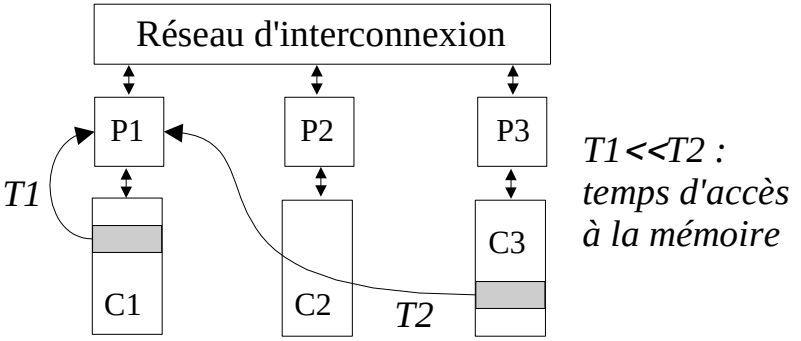
- NUMA :
  Non Uniform Memory Access
  - CRAY T3E
  - SGI Origin
  - PC multi-processeurs

- CC-NUMA :
  Cache Coherent - NUMA



T1<<T2 : temps d'accès à la mémoire

- COMA :
  Cache Only Memory Architecture



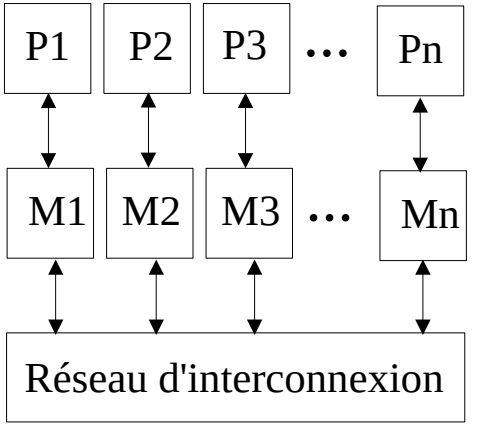
Avantages :

- simplicité d'utilisation
- portabilité au SMP (architecture homogène à mémoire partagée)
- scalabilité
- parallélisation de haut niveau

Inconvénients :

- coût
- limitation du nombre de processeurs (conflit d'accès au niveau matériel)
- la bande passante du réseau est le facteur limitant de ces architectures

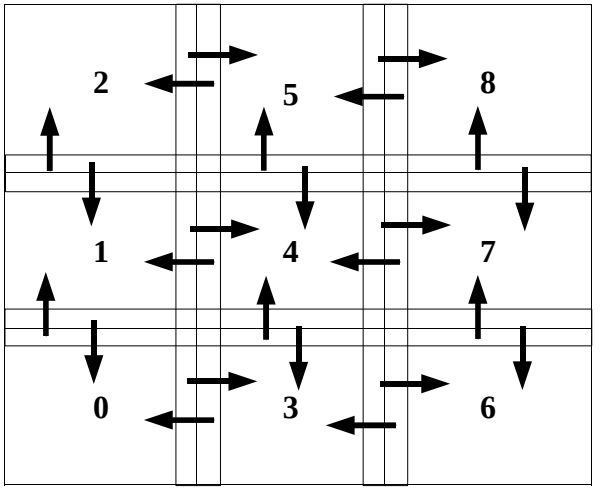
-2.4- Les machines à mémoire distribuée



Caractéristiques :

- interconnexion (réseau local rapide) de systèmes indépendants (noeuds)
- mémoire propre locale à chaque processeur
- chaque processeur exécute des instructions identiques ou non,
  sur des données identiques ou non
- Parallélisme par échange de messages

Les communications entre processeurs sont réalisés par l'appel à des fonctions de bibliothèque standard : PVM (Parallel Virtual Machine) ou MPI (Message Passing Interface).
Il est préférable de développer avec MPI qui est plus stable et qui est portable sur les machines parallèles actuelles.
MPI => MPP (massively Parallel Processing)
La décomposition de domaine permet d'accélérer le calcul :



Avantages :

- non limité théoriquement en nombre de processeurs
  => grande puissance de calcul
- réalisation aisée par la conception de cluster
  mise en réseau d'un certain nombre de machines
  système d'exploitation Linux
- partage transparent des ressources
- portabilité des applications MPI

Inconvénients :

- plus difficile à programmer que les machines à mémoire partagée
  - MPI (Message Passing Interface)
  - PVM (Parallel Virtual Machine)
- efficacité limitée
- administration du cluster

-2.5- Exemple d'architectures parallèles

- Marvel : alpha EV7, 2 à 64 processeurs, 16 Go à 512 Go de mémoire. Tore 2D, 8 Go/s par processeur. Puissance : 2.4 Gflops/processeur.

- NEC SX-6 : supercalculateur vectoriel, 8 teraflops, 64 Go de mémoire, 32 Gb/s

- Cray T3E : DEC alpha EV5, 600 Mflops, 16 à 2048 processeurs, 64 Mo à 2Go de mémoire, réseau d'interconnexion en tore 3D (bande passante de 2 à 128 Go/s).

- SGI origin 3000 : numa, 2 à 1024 processeurs, 512 Mo à 1 To de mémoire, 0.8 à 1 Gflops

- HP superdome : 2.2 à 3 Gflops, 2 à 64 processeurs, 4 Go à 256 Go de mémoire, crossbar, 12 à 51 Go/s

- IBM SP2 : Scalable POWERparallel, RS/6000 à 480 Mflops, 256 Mo de mémoire, ou 540 Mflops, 1 à 2 Go de mémoire vive. Communication par passage de messages à travers un réseau cross-bar à deux niveaux; bande passante de 150 Mo/s.

- SP 4 : 4 à 5.2 Gflops, 1 à 32 processeurs par noeuds, 8, 16 ou 32 noeuds

- compaq ES45 : 1 à 2 Gflops, 4 processeurs EV68, 32 Go de mémoire, configuration trucluster

-2.6- Les grilles de calcul

Agréger une importante quantité de ressources et fournir une grande puissance de calcul sur des sites différents.

Pour avoir accès à une grille il faut faire partie d'une VO (Virtual Organisation)

Groupement d'utilisateurs selon leur besoins et leur intérêt, par exemple astro.vo.eu-egee.org

Avantages :

- disponibilité des ressources
- taux d'occupation de ces ressources (grille de station de travail, seti@home)
- partage de ressources spécifiques et de logiciels
- dispersion géographique
- fiabilité, continuité

-3- Les machines parallèles de l'observatoire

L'Observatoire possède deux machines parallèles :

- Une machine parallèle à mémoire partagée : MPOPM
  16 processeurs, 64 Go de mémoire partagée

- Un cluster de PCs : siolino
  100 coeurs, 256 Go de mémoire distribuée

+ un cluster privé de 88 coeurs, 176 Go de mémoire distribuée

-3.1- MPOPM : Machine Parallèle de l'Observatoire de Paris-Meudon

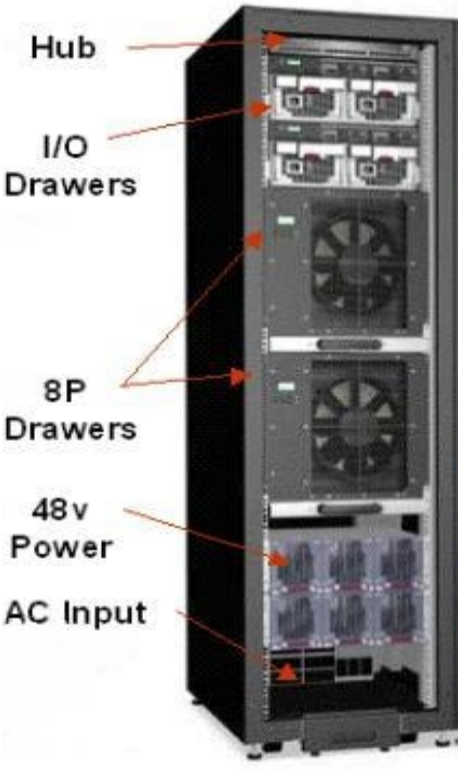
L'Observatoire possède depuis Septembre 2003 une machine parallèle à mémoire partagée.

Cette machine est dédiée à la simulation numérique :

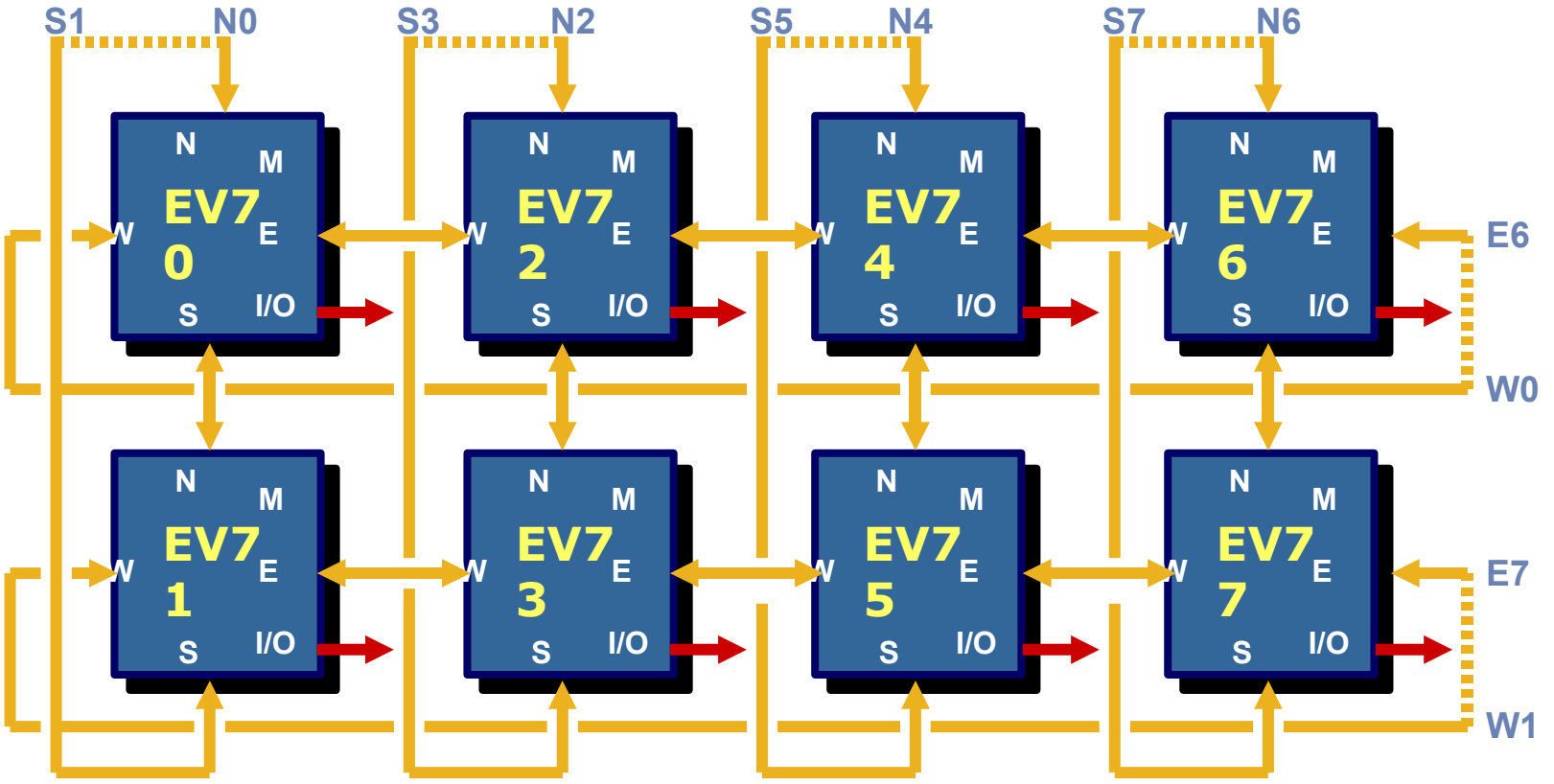
- Formation des structures cosmiques ;
- Chimie des systèmes complexes ;
- Hydrodynamique relative multidimensionnelle.

MPOPM est une machine HP de type Marvel pourvu de 16 processeurs Alpha EV7 à 1.15 GHz et 64 Go de mémoire. Elle possède un RAID de 2 To.

Liste de diffusion MPOPM : mpopm@sympa.obspm.fr



L'interconnexion des processeurs repose sur une architecture torique 2D.



-3.2- Cluster de l'Observatoire : SIOLINO

-3.2.1- Configuration

L'observatoire possède depuis Juin 2004 un cluster de 100 processeurs : siolino.

C'est un cluster de type beowulf qui est constitué de :

- 1 machine frontale (siolino) : bi-opteron, 1.6 Ghz, un disque /home de 1.2 To, un disque /poubelle de 1.6 To (ces 2 disques sont montés NFS sur les serveurs)
- 15 serveurs de calcul :
  - quadri[2..6] : quadri-opteron cadencé à 2.2 Ghz, avec 16 Go de mémoire partagée par quadri opteron et 32 Go pour quadri5
  - quadri[7..9] : quadri-opteron bi-coeurs cadencé à 2.2 Ghz, avec 16 Go de mémoire partagée par quadri opteron
  - quadri[10..16] : bi-opteron quadri-coeurs cadencé à 2.33 Ghz, avec 16 Go de mémoire partagée par bi opteron

=> 256 Go de mémoire distribuée pour le cluster.

Chaque serveur communique avec les autres par une connexion ethernet gigabit.

siolino est la machine interface avec le réseau de l'observatoire

Le protocole d'échange de message est MPICH.

-3.2.2- Utilisation du cluster

Avoir un compte utilisateur sur chacune des machines du cluster (siolino, quadri[1..16]).

L'édition, la mise au point, la compilation, le lancement de programmes(batch) se fait sur la frontale.

Serveurs : exécution de calculs parallèles lancés depuis la frontale et lecture/écriture de fichier

Logiciels d'édition :

vi, emacs, nedit

-3.2.3- Langages et compilateurs

langage	***compilateur***
Fortran 77	mpif77
Fortran 90	mpif90 prog.f90 -o prog
Langage C	mpicc
Langage C++	mpiCC

-3.2.4- Démarrer la machine virtuelle MPI(MPICH)

Ajouter dans votre fichier~/.bash_profile

```
PATH=/usr/local/absoft/bin:${PATH}
PATH=/usr/local/mpich-2/bin:${PATH}
export PATH
```

Créer un fichier ~/.rhosts qui contient l'ensemble des machines du cluster :

```
siolino nom_utilisateur
quadri1 nom_utilisateur
quadri2 nom_utilisateur
quadri3 nom_utilisateur
quadri4 nom_utilisateur
...
```

Créer un fichier ~/.mpd.conf :

```
password=USERpasswd
```

Créer un fichier ~/.mpdpasswd vide.

Modifier les droits d'accès à ces fichiers : chmod 600 ~/.mpd*

Se connecter sur un des quadri-opteron (ex : rsh quadri1)

```
mpdboot -n $nb_machine -f hostfile –rsh=rsh
```

Le fichier hostfile doit contenir les noms des machines du cluster :

```
quadri1
quadri2
quadri3
quadri4
...
```

Ce sont l'ensemble des machines sur lesquelles seront exécutés les processus MPI.

exemple :

```
mpdboot -n 4 -f hostfile –rsh=rsh
```

-3.2.5- Exécuter une application MPI

La soumission d'un job se fait sur la frontale (siolino) et le job est exécuté sur les serveurs (quadri[1..9]).

En interactif :

quadri1 > mpirun -np NB_PROC prog : exécution sur NB_PROC processeurs.

En batch :

Le gestionnaire de batch est openPBS.

Il existe trois queues :

- small : pour les jobs inférieurs à 1h. Chaque utilisateur peut exécuter 50 jobs en même temps.
- middle : pour les jobs plus longs, mais limités à 30h et 30 coeurs.
- long : pour les jobs encore plus longs, mais illimités en temps de calcul. 30 coeurs.

Chaque utilisateur peut exécuter au maximun 50 jobs à la fois sur l'ensemble des queues.

Créer un fichier mon_prog.pbs qui précise les ressources nécessaires et les attributs du job à exécuter :

```
#!/bin/sh
### Job name
#PBS -N nom_prog
### Declare job non-rerunable
#PBS -r n
### Mail to user
#PBS -m ae
### Queue name (small, log, verylong)
#PBS -q small
### Define number of processors
### Number of nodes (node property ev7 wanted)
#PBS -l nodes=1:ppn=4

NPROCS=4
```

```
# Job's working directory
echo working directory is $PBS_O_WORKDIR
nb=`echo $PBS_JOBID | cut -d"." -f1`
cd $PBS_O_WORKDIR
echo Running on host `hostname`
echo Time is `date`
echo Directory is `pwd`
echo This job run on the following processors :
echo `cat $PBS_NODEFILE`

# Run the parallel MPI executable
time mpirun -np $NPROCS nom_prog

# Copy nom_prog.err et nom_prog.log dans $PBS_O_WORKDIR
```

soumettre le job :

```
siolino > qsub nom_prog.pbs
```

Voir l'occupation des queues :

```
siolino > qstat -a
Job id          Name    User    Time Use        S       Queue
52.siolino      prog1   user1   00:19:35        R       small
53.siolino      prog2   user1   00:05:34        R       long
54.siolino      prog1   user1   00:04:15        R       small
55.siolino      prog3   user1   00:00:00        Q       small
56.siolino      job1    user2   00:02:04        R       small
```

Statut : R (en exécution), Q (en attente), S (suspendu), E (sortie d'exécution)

Supprimer un job de la file d'attente :

```
siolino > qdel Job_ID
```

-3.2.6-Arrêter la machine MPI

Arrêt du cluster : mpdallexit

```
quadri1 > mpdallexit
```

http://sio.obspm.fr/fichiersHTML/calcul.html

-3.2.7- Bibliothèques

- LAPACK (Linear Algebra PACKage) :
  librairie Fortran de calcul numérique dédiée à l'algèbre linéaire.

- SCALAPACK (SCAlable LAPACK) :
  inclut un sous-ensemble des routines LAPACK, optimisé pour les architectures à mémoire distribuée

- BLACS (Basic Linear Algebra Communication Subprogram) :
  propose une interface pour l'utilisation du transfert de messages pour les applications en algèbre linéaire.

- FFTW : transformée de Fourier

-4- MPI

-4.1- Historique

En Juin 1994, le forum MPI (Message Passing Interface) définit un ensemble de sous-programmes de la bibliothèque d'échange de messages MPI . Une quarantaine d'organisations y participent. Cette norme, MPI-1, vise à la fois la portabilité et la garantie de bonnes performances.

En Juillet 1997, la norme MPI-2 est publiée. Elle permet la gestion dynamique de processus, la copie mémoire à mémoire, la définition de types dérivés et MPI-IO

-4.2- Définition

Une application MPI est un ensemble de processus qui exécutent des instructions indépendamment les unes des autres et qui communiquent via l'appel à des procédures de la bibliothèque MPI.

On peut définir plusieurs catégories de procédure :

- environnement
- communications point à point
- communications collectives
- types de données dérivées
- topologies
- groupes et communicateurs

-4.3- Environnement

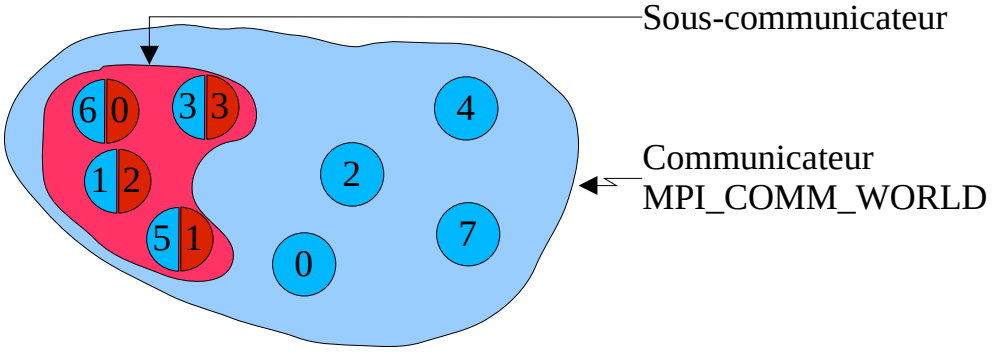
MPI_INIT() permet d'initialiser l'environnement MPI :

```
integer, intent(out) :: code
call MPI_INIT(code)
```

MPI_FINALIZE() permet de désactiver cet environnement :

```
integer, intent(out) :: code
call MPI_FINALIZE(code)
```

MPI_INIT initialise le communicateur par défaut, MPI_COMM_WORLD qui contient tous les processus actifs. Un communicateur est un ensemble de processus qui peuvent communiquer entre eux. Chaque processeur possède un rang unique dans la machine. Les N processeurs sont numérotés de 0 à N-1. Un processus peut appartenir à plusieurs communicateurs avec un rang différent.



La procédure MPI_COMM_SIZE() permet de connaître le nombre de processeurs gérés par un communicateur :

```
integer :: code, nbproc
call MPI_COMM_SIZE(MPI_COMM_WORLD,nbproc,code)
```

Le rang du processus est donné par la procédure MPI_COMM_RANK() :

```
integer :: code, rang
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
```

Exemple :

```
Program SizeRang
implicit none
include 'mpif.h'
integer :: code, rang, nbproc
call MPI_INIT(code)
call MPI_COMM_SIZE(MPI_COMM_WORLD,nbproc,code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
call MPI_FINALIZE(code)
if (rang==0) print *,'Nombre de processus : ',nbproc
print *,'Je suis le processus de rang ',rang
end program SizeRang
```

```
> mpirun -np 3 SizeRang
 Je suis le processus de rang 1
 Nombre de processus : 3
 Je suis le processus de rang 0
 Je suis le processus de rang 2
```

-4.4- Communication point à point

-4.4.1- Définition

Une communication point à point est effectuée entre deux processus, un émetteur et un récepteur, identifiés par leur rang dans le communicateur au sein duquel se fait la communication.

L'enveloppe du message est constituée :

- du rang du processus émetteur
- du rang du processus récepteur
- du nom du communicateur
- du type de données
- de la longueur du message
- de l'étiquette du message

-4.4.2- Les procédures standard de communication

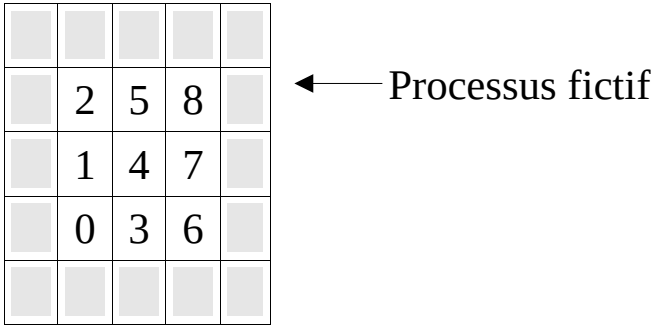
Les deux procédures standards d'envoi et de réception de message sont :

```
MPI_SEND(valeur,taille,type,destination,etiquette,comm,code)
MPI_RECV(valeur,taille,type,source,etiquette,comm,statut,code)
```

Pour la procédure MPI_RECV, le rang du processus récepteur peut être MPI_ANY_SOURCE.

La valeur de l'étiquette MPI_ANY_TAG

On peut également effectuer des communications avec un processus fictif de rang MPI_PROC_NULL.



Il existe d'autres modes de communication : synchrone, bufferisé

Les communications peuvent également être bloquantes ou non bloquantes.

Rq : MPI_SEND est implémenté de façon bloquante ce qui peut entraîner des situations de verrouillage si deux processus s'envoient mutuellement un message à un instant donné. Il faut dans ce cas intervertir MPI_SEND et MPI_RECV pour l'un des processus, ou utiliser la procédure MPI_SENDRECV

-4.4.3- Types de données

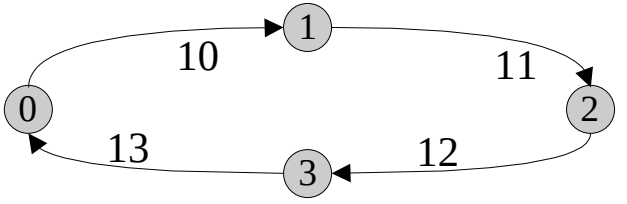
```
MPI_INTEGER
MPI_REAL
MPI_DOUBLE_PRECISION
MPI_COMPLEX
MPI_LOGICAL
MPI_CHARACTER
```

-4.4.4- Exemples

```
Program CommPaP
include 'mpif.h'
integer :: code, rang, val
integer, parameter :: tag=100
integer, dimension(MPI_STATUS_SIZE) :: statut
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
if (rang==0) then
        val=999
        call MPI_SEND(val,1,MPI_INTEGER,1,tag,MPI_COMM_WORLD,code)
elseif (rang==1) then
        print *,"La valeur de la variable val est ",val
        call MPI_RECV(val,1,MPI_INTEGER,0,tag,MPI_COMM_WORLD,statut,code)
        print *,"Le processus 1 a recu la valeur ",val," du processus 0"
endif
call MPI_FINALIZE(code)
end program CommPaP
```

```
> mpirun -np 2 CommPaP
 La valeur de la variable val est 0
 Le processus 1 a recu la valeur 999 du processus 0
```

Communication en anneaux




```
apres=mod(rang+1,nbproc)
avant=mod(nbproc+rang-1,nbproc)
if (mod(rang,2)==0) then
call MPI_SEND(rang+10,1,MPI_INTEGER,apres,tag,MPI_COMM_WORLD,code)
call MPI_RECV(val,1,MPI_INTEGER,avant,tag,MPI_COMM_WORLD,statut,code)
else
call MPI_RECV(val,1,MPI_INTEGER,avant,tag,MPI_COMM_WORLD,statut,code)
call MPI_SEND(rang+10,1,MPI_INTEGER,apres,tag,MPI_COMM_WORLD,code)
endif
print *,'Moi processus ',rang," j'ai recu ",val,' du processus',avant
```

```
> mpirun -np 4 CommAnneau
Moi processus 1 j'ai recu 10 du processus 0
Moi processus 2 j'ai recu 11 du processus 1
Moi processus 0 j'ai recu 13 du processus 3
Moi processus 3 j'ai recu 12 du processus 2
```

Communication SEND_RECV

```
apres=mod(rang+1,nbproc)
avant=mod(nbproc+rang-1,nbproc)
call MPI_SENDRECV(rang+10,1,MPI_INTEGER,apres,tag,
                  val,    1,MPI_INTEGER,avant,tag,MPI_COMM_WORLD,statut,code)
print *,'Moi processus ',rang,'j'ai recu ',val,' du processus',avant
```

-4.4.5- Mode de communications

Synchroneous : l'envoi du message se termine lorsque la réception est achevée.

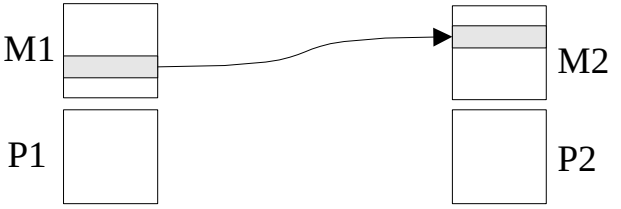
Buffered : le programmeur doit effectuer une recopie temporaire du message. L'envoi se termine lorsque la recopie est achevée. L'envoi est ainsi découplé de la réception.

Standard : MPI effectue une recopie temporaire ou non suivant la taille du message.

Si il y a une recopie, l'envoi est découplé de la réception dans le cas contraire, l'envoi se termine lorsque la réception est, elle même, terminée. Cela peut entraîner des situations de blocage. MPI_SEND est remplacé implicitement par un MPI_BSEND pour les messages de petite taille.

Ready : l'envoi ne peut être effectué que si la réception à été initiée. (utilisé pour les applications clients-serveurs)

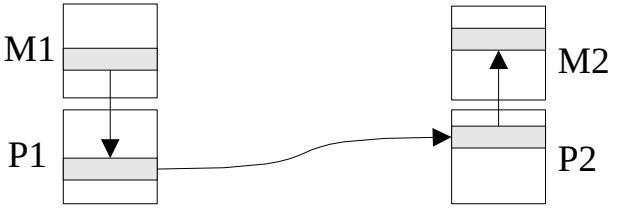
Envoi bloquant : sans recopie temporaire du message. L'envoi et la réception sont couplées.

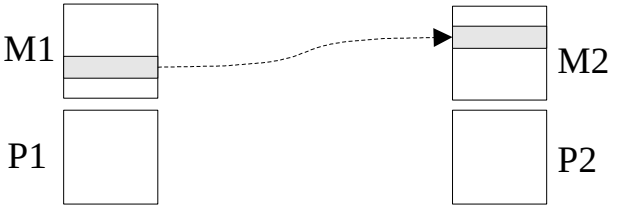


Envoi non bloquant :

- avec recopie temporaire du message. L'envoi et la réception ne sont pas couplées.

- sans recopie temporaire du message. L'envoi et la réception sont couplées. Attention à ne pas modifier la donnée avant qu'elle n'ait quittée la mémoire du processus émetteur. Ce sont des tâches de fond.





modes	bloquant	Non bloquante
Envoi standard	MPI_SEND	MPI_ISEND
Envoi synchroneous	MPI_SSEND	MPI_ISSEND
Envoi buffered	MPI_BSEND	MPI_IBSEND
Réception	MPI_RECV	MPI_IRECV

L'utilisation des procédures MPI_ISSEND et MPI_IRECV permettent de recouvrir les communications par des calculs => gains de performance.

Pour les procédures non bloquantes, il existe des procédures qui permettent de synchroniser les processus (MPI_WAIT, MPI_WAITALL), ou de tester si une requête est bien terminée (MPI_TEST, MPI_TESTALL)

-4.5- Communications collectives

Les communications collectives permettent par l'appel à une procédure de faire plusieurs communications point à point.

Ces communications sont :

- effectuées sur l'ensemble des processus appartenant à un communicateur.
- Synchronisées : l'appel à la procédure se termine lorsque la communication a été achevée pour l'ensemble des processus.

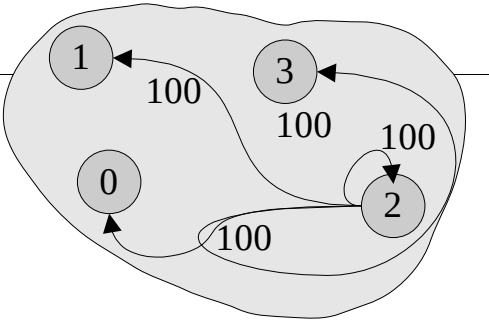
-4.5.1- Synchronisation des processus

La procédure MPI_BARRIER() réalise une synchronisation globale.

```
interger, intent(out) :: code
call MPI_BARRIER(MPI_COMM_WORLD,code)
```

-4.5.2- Diffusion générale

On envoi la même donnée à chaque processus.

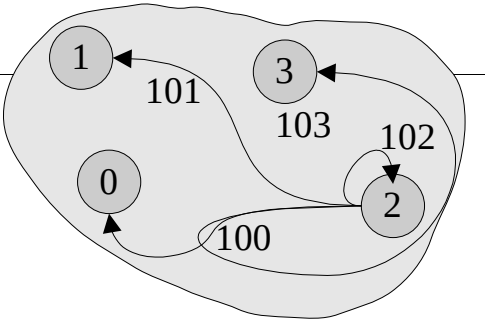



```
Program bcast
include 'mpif.h'
integer valeur, rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
if (rang==2) valeur=100
call MPI_BCAST(valeur,1,MPI_INTEGER,2,MPI_COMM_WORLD,code)
print *,'Le processus ',rang,' a recu la valeur ',valeur,' du processus 2'
call MPI_FINALIZE(code)
end Program bcast
```

```
 > mpirun -np 4 bcast
 Le processus 2 a recu la valeur 100 du processus 2
 Le processus 3 a recu la valeur 100 du processus 2
 Le processus 0 a recu la valeur 100 du processus 2
 Le processus 1 a recu la valeur 100 du processus 2
```

-4.5.3- Diffusion sélective

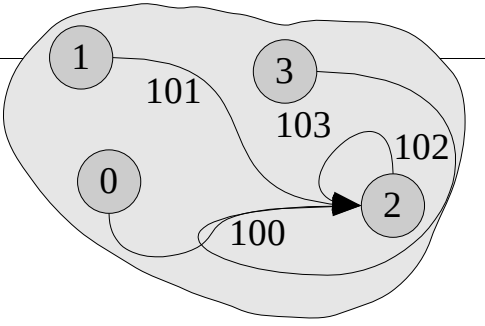
Envoi d'éléments contigus en mémoire à chaque processus.




```
Program scatter
include 'mpif.h'
integer valeurs(4), val, rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
if (rang==2) valeurs(:)=(/(100+i,i=0,3)/)
call MPI_SCATTER(valeurs,1,MPI_INTEGER,val,1,MPI_INTEGER,2,MPI_COMM_WORLD,code)
print *,'Le processus ',rang,' a recu la valeur ',val,' du processus 2'
call MPI_FINALIZE(code)
end Program scatter
```

```
 > mpirun -np 4 scatter
 Le processus 2 a recu la valeur 102 du processus 2
 Le processus 3 a recu la valeur 103 du processus 2
 Le processus 0 a recu la valeur 100 du processus 2
 Le processus 1 a recu la valeur 101 du processus 2
```

-4.5.4- Collecte de données réparties



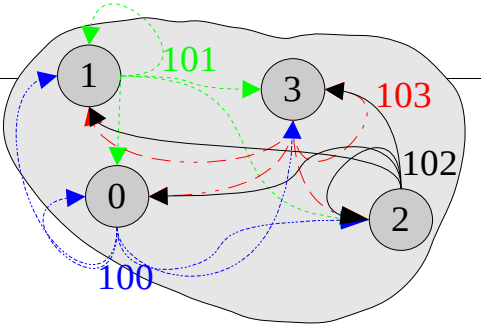

```
Program gather
include 'mpif.h'
integer valeurs(4), val, rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
val=100+rang
call MPI_GATHER(val,1,MPI_INTEGER,valeurs,1,MPI_INTEGER,2,MPI_COMM_WORLD,code)
if (rang==2 ) print *,'Le processus 2 a recu les valeurs ',valeurs
call MPI_FINALIZE(code)
end Program gather
```

```
 > mpirun -np 4 gather
 Le processus 2 a recu les valeurs 100 101 102 103
```

Les données sont rangées dans le tableau valeurs selon le rang des processus émetteurs, par ordre croissant.

-4.5.5- Collecte générale

Collecte par l'ensemble des processus de données réparties.

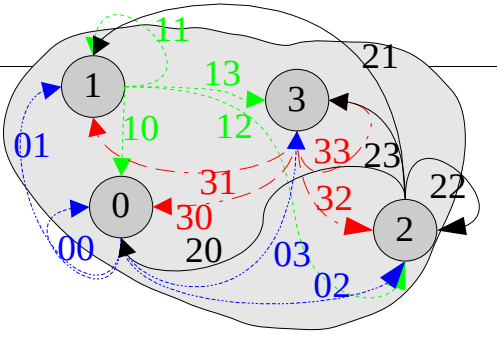



```
Program allgather
include 'mpif.h'
integer valeurs(4), val, rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
val=100+rang
call MPI_ALLGATHER(val,1,MPI_INTEGER,valeurs,1,MPI_INTEGER,MPI_COMM_WORLD,code)
print *,'Le processus ',rang,' a recu les valeurs ',valeurs
call MPI_FINALIZE(code)
end Program allgather
```

```
 > mpirun -np 4 allgather
 Le processus 2 a recu les valeurs 100 101 102 103
 Le processus 1 a recu les valeurs 100 101 102 103
 Le processus 3 a recu les valeurs 100 101 102 103
 Le processus 0 a recu les valeurs 100 101 102 103
```

-4.5.6- Alltoall

Diffusion sélective de données réparties, par tous les processus.




```
Program alltoall
include 'mpif.h'
integer vals(4), rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
valeurs(:)=(/(10*rang+i,i=0,3)/)
call MPI_ALLTOALL(vals,1,MPI_INTEGER,vals,1,MPI_INTEGER,MPI_COMM_WORLD,code)
print *,'Le processus ',rang,' a recu les valeurs ',vals
call MPI_FINALIZE(code)
end Program alltoall
```

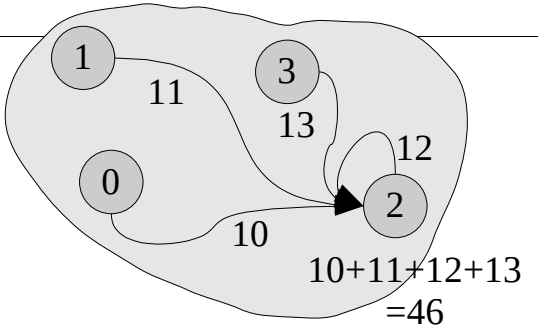
```
 > mpirun -np 4 allgather
 Le processus 2 a recu les valeurs 2 12 22 32
 Le processus 1 a recu les valeurs 1 11 21 31
 Le processus 3 a recu les valeurs 3 13 23 33
 Le processus 0 a recu les valeurs 0 10 20 30
```

-4.5.7- Les opérations de réduction

Ces procédures effectuent, en plus du transfère de données, une opération sur ces données (somme, produit, minimum, maximum...) : MPI_REDUCE , MPI_ALLREDUCE (MPI_REDUCE+MPI_BCAST: diffusion du résultat)

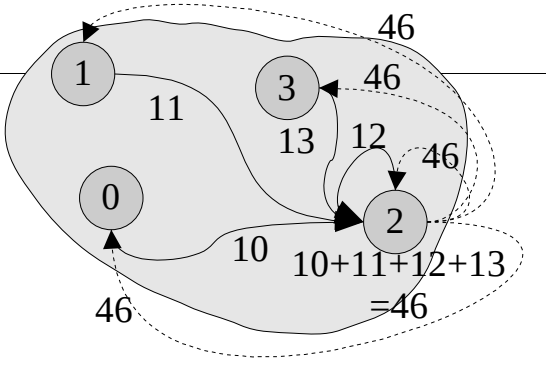
Principales opérations de réduction pré-définies :

MPI_SUM	Somme des données
MPI_PROD	Produit des données
MPI_MAX	Recherche du maximum
MPI_MIN	Recherche du minimum
MPI_MAXLOC	Recherche de l'indice du minimum
MPI_MINLOC	Recherche de l'indice du maximum
MPI_LAND	ET logique
MPI_LOR	OU logique
MPI_LXOR	OU exclusif logique




```
Program reduce
include 'mpif.h'
integer somme, val, rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
val=10+rang
call MPI_REDUCE(val,somme,1,MPI_INTEGER,MPI_SUM,2,MPI_COMM_WORLD,code)
print *,'Pour le processus ',rang,' la somme est de ',somme
call MPI_FINALIZE(code)
end Program reduce
```

```
 > mpirun -np 4 reduce
 Pour le processus 0 la somme est de 0
 Pour le processus 1 la somme est de 0
 Pour le processus 3 la somme est de 0
 Pour le processus 2 la somme est de 46
```




```
Program allreduce
include 'mpif.h'
integer somme, val, rang, code
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
val=10+rang
call MPI_ALLREDUCE(val,somme,1,MPI_INTEGER,MPI_SUM,MPI_COMM_WORLD,code)
print *,'Pour le processus ',rang,' la somme est de ',somme
call MPI_FINALIZE(code)
end Program allreduce
```

```
 > mpirun -np 4 allreduce
 Pour le processus 0 la somme est de 46
 Pour le processus 1 la somme est de 46
 Pour le processus 3 la somme est de 46
 Pour le processus 2 la somme est de 46
```

-4.6- Types de données dérivées

Il existe des types de données pré-définis par MPI : MPI_INTEGER, MPI_REAL...

On peut en définir de nouveau en utilisant les procédures MPI_TYPE_CONTIGUOUS, MPI_TYPE_VECTOR, MPI_TYPE_HVECTOR.

Ces types sont activés par la procédure MPI_TYPE_COMMIT et détruits par la procédure MPI_TYPE_FREE.

-4.6.1- Types contigus

MPI_TYPE_CONTIGUOUS crée une structure de données à partir d'un ensemble homogène de données de type pré-défini qui sont contiguës en mémoire.

```
Program colonne
include 'mpif.h'
integer :: code, type_colonne, tag, rang, i, j, tab(3,4), statut(MPI_STATUS_SIZE)
call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
do i=1,3 do j=1,4
```

```
tab(i,j)=3*(j-1)+i-1+(100*rang+1)
enddo enddo
call MPI_TYPE_CONTIGUOUS(3,MPI_INTEGER,type_colonne,code)
call MPI_TYPE_COMMIT(type_colonne,code)
if (rang==0) then
        call MPI_SEND(tab(1,2),1,type_colonne,1,tag,MPI_COMM_WORLD,code)
elseif (rang==1) then
        call MPI_RECV(tab(2,3),1,type_colonne,0,tag,MPI_COMM_WORLD,statut,code)
endif
call MPI_TYPE_FREE(type_colonne,code)
call MPI_FINALIZE(code)
end Program colonne
```

1	4	7	10
2	5	8	11
3	6	9	12

tab processus 0

101	104	107	110
102	105	108	111
103	106	109	112

tab processus 1, t1

101	104	107	6
102	105	4	111
103	106	5	112

tab processus 1, t2

-4.6.2- Types avec un pas constant

MPI_TYPE_VECTOR crée une structure de données à partir d'un ensemble homogène de données de type pré-défini qui sont distantes d'un pas constant en mémoire.

```
MPI_TYPE_VECTOR(nb_donnees,taille,pas,ancien_type,nouveau_type,code)
```

```
Program ligne
include 'mpif.h'
integer :: code, type_ligne, tag, rang, i, j, tab(3,4), statut(MPI_STATUS_SIZE)

call MPI_INIT(code)
call MPI_COMM_RANK(MPI_COMM_WORLD,rang,code)
do i=1,3
        do j=1,4
                tab(i,j)=3*(j-1)+i-1+(100*rang+1)
        enddo
enddo
call MPI_TYPE_VECTOR(4,1,3,MPI_INTEGER,type_ligne,code)
```

```
call MPI_TYPE_COMMIT(type_ligne,code)
if (rang==0) then
        call MPI_SEND(tab(1,1),1,type_ligne,1,tag,MPI_COMM_WORLD,code)
elseif (rang==1) then
        call MPI_RECV(tab(3,1),1,type_ligne,0,tag,MPI_COMM_WORLD,statut,code)
endif
call MPI_TYPE_FREE(type_ligne,code)
call MPI_FINALIZE(code)
end Program ligne
```

1	4	7	10
2	5	8	11
3	6	9	12

tab processus 0

101	104	107	110
102	105	108	111
103	106	109	112

tab processus 1, t1

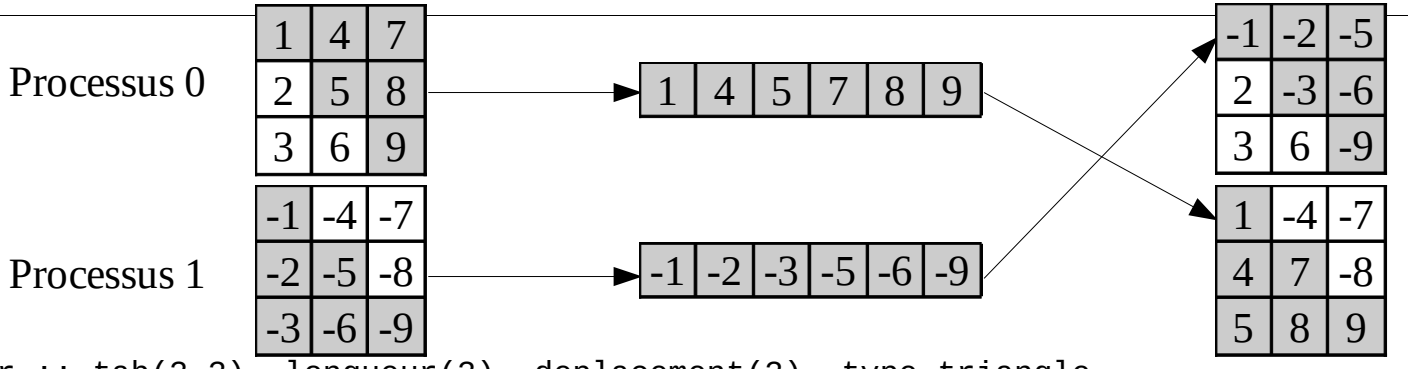
101	104	107	110
102	105	108	111
1	4	7	10

tab processus 1, t2

Le pas passé en paramètre de la procédure MPI_TYPE_HVECTOR est donné en octet.

-4.6.3- Type avec un pas variable

MPI_TYPE_INDEXED permet de définir un type qui comporte des blocs de données de type pré-défini, de taille variable, repérés par un déplacement par rapport au début de la structure. Ce déplacement est donné en nombre d'occurrence du type pré-définit. MPI_TYPE_HINDEXED : déplacement en octet.




```
integer :: tab(3,3), longueur(3), deplacement(3), type_triangle
if (rang==0) then longueur(:)=(/1,2,3/); deplacement=(/0,3,6/); endif
if (rang==1) then longueur(:)=(/3,2,1/); deplacement=(/0,4,8/); endif
call MPI_TYPE_INDEXED(3,longueur,deplacement,MPI_INTEGER,type_triangle,code)
call MPI_TYPE_COMMIT(type_triangle,code)
call MPI_SENDRECV_REPLACE(tab,1,type_triangle,mod(rang+1,2),tag,
                          mod(rang+1,2),tag,MPI_COMM_WORLD,statut,code)
```

-4.6.4- Type de données hétérogènes

MPI_TYPE_STRUCT est le constructeur de type le plus général. Il permet de faire la même chose que MPI_TYPE_INDEXED mais sur des données hétérogènes : le champ type de données est un tableau de type de données. Les déplacements sont donnés en octet, ils peuvent être calculés grâce à la procédure MPI_ADDRESS.

```
Integer, dimension(4) :: types, longueur, adresse, deplacement
integer :: type_personne

Type personne
        character(len=20)         :: nom
        character(len=30)         :: prenom
        integer                   :: age
        real                              :: taille
end type personne
type(personne) :: p
types = (/MPI_CHARACTER,MPI_CHARACTER,MPI_INTEGER,MPI_REAL/)
longueur = (/20,30,1,1/)
```

```
call MPI_ADDRESS(p%nom,addresse(1),code)
call MPI_ADDRESS(p%prenom,addresse(2),code)
call MPI_ADDRESS(p%age,addresse(3),code)
call MPI_ADDRESS(p%taille,addresse(4),code)
do i=1,4
        deplacement(i)=adresse(i)-adresse(1)
enddo
call MPI_TYPE_STRUCT(4,longueur,deplacement,types,type_personne,code)
call MPI_TYPE_COMMIT(type_personne,code)
if (rang==0) then
        p%nom='Dupond'; p%prenom='Pierre'; p%age=35; p%taille=1.85
        print *,"Le processus 0 envoi : ",p
        call MPI_SEND(p%type,1,type_personne,1,tag,MPI_COMM_WORLD,code)
else
        call MPI_RECV(p%type,1,type_personne,0,tag,MPI_COMM_WORLD,statut,code)
        print *,"Le processus 1 a recu : ",p
endif
call MPI_TYPE_FREE(type_personne,code)
```

-4.6.5- Quelques procédures utiles

Taille d'un type de données :

1	4	7
2	5	8
3	6	9

```
integer, intent(in)  :: type
integer, intent(out) :: taille,code
MPI_TYPE_SIZE(type,taille,code)
```

```
Taille type = 20 (5*4)
```

```
MPI_TYPE_EXTEND : taille d'un type aligné en mémoire
```

Adresse de début et de fin du type de données :

```
MPI_TYPE_LB(type,adresse,code)
MPI_TYPE_UB(type,adresse,code)
```

```
Borne inférieure : 0
Borne supérieure : 32 (8*4)
```

-4.7- Topologies

Pour les applications qui nécessite une décomposition de domaine il est intéressant de pouvoir définir une topologie qui organise les processus de façon régulière.

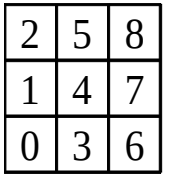
Il existe deux types de topologie :

- cartésienne

  définition d'une grille

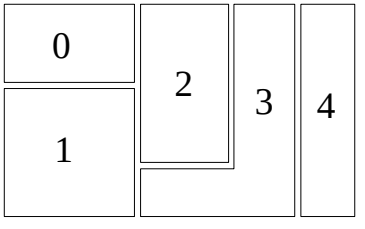
  périodique ou non

  identification des processus par leurs coordonnées



- graphe

  généralisation à des topologies plus complexes.



-4.7.1- Topologies cartésiennes

Une topologie cartésienne est définit par l'appel de la procédure MPI_CART_CREATE. Elle crée un nouveau communicateur.

```
Integer, parameter                :: nb_dim=2
integer, dimension(nb_dim)        :: dims
logical, dimension(nb_dim)        :: period
integer                           :: nom_comm, code
logical                           :: reorganise
call MPI_CART_CREATE(MPI_COMM_WORLD,nb_dim,dims,period,reorganise,nom_comm,code)
```

reorganise : false => le processus conserve le rang qu'il avait dans l'ancien communicateur.

Le choix du nombre de processus pour chaque dimension peut être laissé au soin du processus en utilisant la procédure MPI_DIMS_CREATE :

```
integer, intent(in)                         :: nb_proc, nb_dim
integer, dimension(nb_dim), intent(out)     :: dims
integer, intent(out)                        :: code
call MPI_DIMS_CREATE(nb_proc,nb_dim,dims,code)
```

La procédure MPI_CART_COORDS fournit les coordonnées d'un processus de rang donné dans la grille

```
integer, intent(in)                          :: nom_comm, rang, nb_dim
integer, dimension(nb_dim), intent(out)      :: coords
integer, intent(out)                         :: code
call MPI_CART_COORDS(nom_comm,rang,nb_dim,coords,code)
```

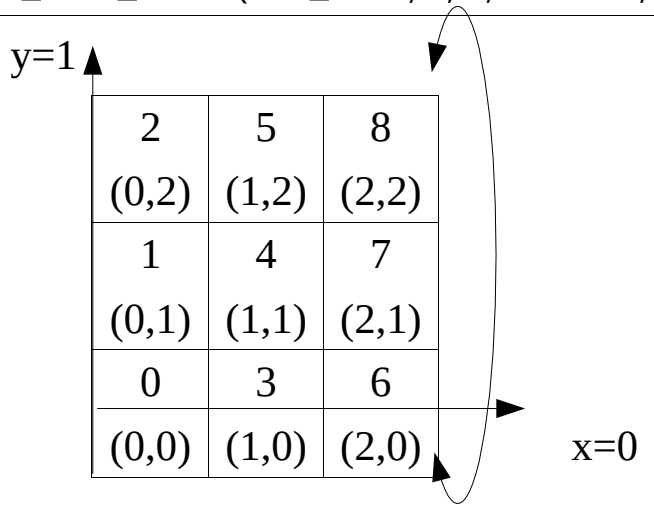
La procédure MPI_CART_RANK permet de connaître le rang du processus associé aux coordonnées dans la grille.

```
integer, intent(in)                          :: nom_comm
integer, dimension(nb_dim), intent(out)      :: coords
integer, intent(out)                         :: rang, code
call MPI_CART_RANK(nom_comm,coords,rang,code)
```

La procédure MPI_CART_SHIFT permet de connaître le rang des voisins d'un processus dans une direction donnée.

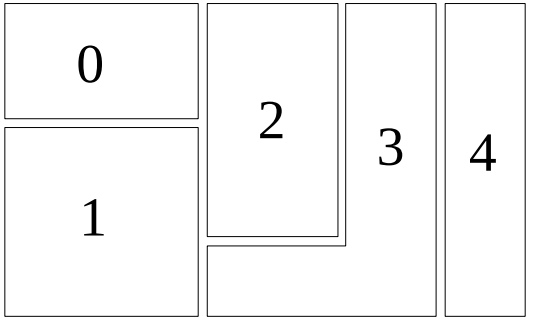
```
integer, intent(in)       :: nom_comm, direction, pas
integer, intent(out)      :: avant, apres, code
call MPI_CART_SHIFT(nom_comm,direction,pas,avant,apres,code)

call MPI_CART_SHIFT(nom_comm,0,1,gauche,droite,code)
call MPI_CART_SHIFT(nom_comm,1,1,dessous,dessus,code)
```



-4.7.2- Topologies graphes

Décomposition de domaine sur une grille non régulière : un processus peut avoir un nombre quelconque de voisins.



No proc	Liste des voisins	index
0	1, 2	2
1	0, 2, 3	5
2	0, 1, 3	8
3	1, 2, 4	11
4	3	12

La liste des voisins et le tableau d'index permet de définir les voisins d'un processus.

```
integer :: nom_comm, nb_proc, reorg, code
integer, dimension(nb_proc) :: index
integer, dimension(nb_max_voisin) :: voisins
index =   (/  2,   5,     8,    11,  12/)
voisins = (/1,2, 0,2,3, 0,1,3, 1,2,4, 3/)
call MPI_GRAPH_CREATE(MPI_COMM_WORLD,nb_proc,index,voisins,reorg,nom_comm,code)
```

La procédure MPI_GRAPH_NEIGHBORS_COUNT permet de connaître le nombre de voisins d'un processus donné.

```
integer, intent(in)   :: nom_comm, rang
integer, intent(out)  :: nb_voisins, code

MPI_GRAPH_NEIGHBORS_COUNT(nom_comm,rang,nb_voisins,code)
```

La procédure MPI_GRAPH_NEIGHBORS donne la liste des voisins d'un processus.

```
integer, intent(in)   :: nom_comm, rang, nb_voisins
integer, intent(out)  :: code
integer, dimension(nb_max_voisin), intent(out) :: voisins

MPI_GRAPH_NEIGHBORS(nom_comm,rang,nb_voisins,voisins,code)
```

-4.8- Gestion des groupes de processus

Pour construire l'espace de communication, on utilise un communicateur qui est constitué :

- d'un groupe de processus,
- d'un contexte de communication, i.e une propriété des communicateurs qui permet de partager l'espace de communication, gérer les communications point-à-point et collectives de telle sorte qu'elles n'interfèrent pas.

Il y a deux manières de construire un communicateur :

- à partir d'un autre communicateur,
- par l'intermédiaire d'un groupe de processus.

Le communicateur par défaut est MPI_COMM_WORLD. Il contient tous les processus actifs.

Il est créé avec MPI_INIT et détruit avec MPI_FINALIZE.

-4.8.1-Communicateur issu d'un communicateur

On peut créer des sous-ensembles de communication, on partage le communicateur initial en plusieurs sous-communicateurs distincts mais de même nom.

Pour cela on définit :

- une couleur associant à chaque processus le numéro du communicateur auquel il appartiendra
- une clé permettant de numéroter les processus dans chaque communicateur.

```
Integer, intent(in) :: nom_comm ! Communicateur courant à partager
integer, intent(in) :: couleur, cle
integer, intent(out) :: new_comm, code
MPI_COMM_SPLIT (nom_comm,couleur,cle,new_comm,code)
```

L'intérêt de ce découpage de communicateur est de pouvoir effectuer une communication collective que sur une partie des processus de l'application.

Exemple : définition de deux communicateurs content les processus pairs et impairs de MPI_COMM_WORLD

Rang dans MPI_COMM_WORLD	0	1	2	3	4	5	6	7
couleur	0	1	0	1	0	1	0	1
clé	5	2	3	3	8	0	3	7
Rang dans le nouveau communicateur	2	1	0	2	3	0	1	3

La fonction de destruction d'un communicateur est MPI_COMM_FREE :

```
integer, intent(inout) :: nom_comm
integer, intent(out)   :: code
MPI_COMM_FREE(nom_comm,code)
```

-4.8.2-Communicateur issu d'un groupe

Un groupe est un ensemble ordonné de processus identifiés par leurs rangs.

Initialement tous les processus appartiennent au groupe associé au communicateur MPI_COMM_WORLD.

Tout communicateur est associé à un groupe. La procédure MPI_COMM_GROUP permet de connaître ce groupe.

```
Integer, intent(in)  :: nom_comm
integer, intent(out) :: groupe, code
MPI_COMM_GROUP(nom_comm,groupe,code)
```

A partir de ce groupe on peut construire un sous-groupe. Pour cela on doit au préalable définir un vecteur contenant les rangs des processus qui formeront ce sous-groupe. Ensuite on utilise MPI_GROUP_INCL :

```
integer, intent(in)                          :: groupe,nb_proc_grp
integer, dimension(nb_proc_grp), intent(in) :: rangs
integer, intent(out)                         :: newgroupe,code
MPI_GROUP_INCL (groupe,nb_proc_grp,rangs,newgroupe,code)
```

On peut également utiliser la procédure MPI_GROUP_EXCL, qui elle exclue du sous-groupe les processus qui ne sont pas dans le tableau rangs :

```
integer, intent(in)                          :: groupe, nb_proc_grp
integer, dimension(nb_proc_grp), intent(in) :: rangs
integer, intent(out)                         :: newgroupe, code
MPI_GROUP_EXCL (groupe,nb_proc_grp,rangs,newgroupe,code)
```

Une fois le sous-groupe crée, on construit le communicateur associé à l'aide de MPI_COMM_CREATE :

```
integer, intent(in)  :: nom_comm, newgroupe
integer, intent(out) :: new_comm, code
MPI_COMM_CREATE (nom_comm,newgroup,new_comm,code)
```

Il reste à libérer le sous-groupe crée avec MPI_GROUP_FREE :

```
integer, intent(inout) :: groupe
integer, intent(out)   :: code
MPI_GROUP_FREE (groupe, code)
```

Comme pour le communicateur on dispose de procédures pour déterminer le nombre d'éléments du groupe :

```
integer, intent(in)  :: groupe
integer, intent(out) :: taille, code
MPI_GROUP_SIZE(groupe,taille,code)
```

et le rang de ses processus :

```
integer, intent(in)  :: groupe
integer, intent(out) :: rang, code
MPI_GROUP_RANK(groupe,rang,code)
```

Opérations sur les groupes :

```
integer, intent(in)  :: groupe1, groupe2
integer, intent(out) :: newgroup, resultat, code
MPI_GROUP_COMPARE(groupe1,groupe2,resultat,code)
MPI_GROUP_UNION(groupe1,groupe2,newgroup,code)
MPI_GROUP_INTERSECTION(groupe1,groupe2,newgroup,code)
MPI_GROUP_DIFFERENCE(groupe1,groupe2,newgroup,code)
```

-4.9- Quelques procédures

-4.9.1-Temps de communication

```
MPI_WTIME() ou MPI_WTICK()
```

temps (sec) = calcul+préparation (latence+surcoût)+transfert

latence : temps d'initialisation des paramètres réseaux

surcoût : temps de préparation du message lié à l'implémentation MPI et au mode de transfert.

-4.9.2- Nom d'un processeur

```
Integer, intent(out) :: longueur, code
character(MPI_MAX_PROCESSOR_NAME), intent(out) :: nom
if (rang==0) then
        MPI_GET_PROCESSOR_NAME(nom,longueur,code)
        print *,'le nom du processeur 0 est : ',nom(1:longueur)
endif
```

```
le nom du processeur 0 est : quadri1.obspm.fr
```

-4.10- MPI-2

Une nouvelle norme est disponible depuis Juillet 1997 : MPI-2.

Elle permet :

- la gestion dynamique des processus,
- la communication de mémoire à mémoire,
- entrées/sorties parallèles.

-4.11- Bibliographie

- Les spécifications de la norme MPI :
  a message passing interface standard, Mars 1994.
  ftp://ftp.irisa.fr/pub/netlib/mpi/drafts/draft-final.ps

- Marc Snir & al. MPI :
  The Complete Reference. Second Edition. MIT Press, 1998.
  Volume 1, The MPI core. Volume 2, MPI-2

- MPI subroutine reference :
  - http://www.ccr.jussieu.fr/ccr/Documentation/Calcul/ppe.html/d3d80mst.html
  - http://www.lam-mpi.org/tutorials/bindings/

Index des procédures MPI