

```

program action_mg
one-way multigrid for action minimization
Written by In-Ho Lee, September 12 (2001)
USE action, ONLY : natom_ac, npeg_ac, rk_ac, etarget, tttarget, lkntar, xnu, xnu, tau_ac, xmass, itmax_relax, &
lperpath, aperc, pamp_ac, itang, fdeltak, lclimb, exp_ac, iprint_ac, lneb, lcho, lpar, lstr, &
the_xmass_xmass, onwa, object, detar, xgamma

implicit none
include "mpif.h"
integer nvar
integer, allocatable :: npeg(:), rkag(:)
real8G, allocatable :: tauag(:)
real8G, allocatable :: xnu(:)
integer :: nlevel
integer :: nlevel1
character*100 :: character100
character*400 :: character400
character*4 :: lnt
integer, allocatable :: llag(:)
integer itemp, itemp, inate
integer myid, nproc, ierr, kount, inroot

call MPI_INIT( ierr )
call MPI_COMM_RANK( MPI_COMM_WORLD, myid, ierr )
call MPI_COMM_SIZE( MPI_COMM_WORLD, nproc, ierr )
if(myid == 0 .and. nproc > 1) print *, nproc, " processes"
if(myid == 0 .and. nproc == 1) print *, nproc, " process is also"
print *, "Process ", myid, " of ", nproc, " is alive"
if(myid == 0)then ! -----[ process id = 0
call date_and_time(date=fmnd, time=fmnt)
write(6, '(a10,2x,a8,2x,a10)') 'date,time ', fmnd, ' ',
& fmnt
endif ! -----] process id = 0

ipowell=1
ipowell=0 ! CG minimization is default for usual potentials
istart=1
nlvl=3
nlvl=4
nlvl=5
nlvl=6
nlvl=7
allocate(npeg(nlvl),rkag(nlvl)) ; allocate(llag(nlvl))
allocate(tauag(nlvl))
nvar=1

if(myid == 0)then ! -----[ process id = 0
call system_clock(itemp, inate)
open(5, file='input_ac.i', form='formatted')
read(5,*) xnu, tauag(nlvl), xmass, npeg(nlvl), rkag(nlvl), natom_ac, itmax_relax, fdeltak, xnu
if(fdeltak .lt. 0.0d0) fdeltak=0.0d0
xmass=xmass*1836.152701d0
read(5,*) iseed1, iseed2, lkntar, lclimb, itang, exp_ac, etarget, tttarget
read(5,*) lperpath, aperc, pamp_ac
read(5,*) llag(nlvl)
close(5)

---[
open(5, file='aded.i', form='formatted')
read(5,*) natom_ac, npeg(nlvl), rkag(nlvl)
read(5,*) etarget, tttarget, lkntar
read(5,*) tauag(nlvl), xnu, xnu, xgamma
read(5,*) itmax_relax, ftoi, deltat, xmass
read(5,*) lperpath, aperc, pamp_ac
read(5,*) itang, fdeltak, lclimb
read(5,*) iseed1, iseed2, exp_ac
read(5,*) llag(nlvl)
close(5)
if(xgamma < 0.0d0)then
xgamma=-1.0d0
else
xgamma=1.0d0
endif
the_xmass_xmass=.false.
if(xmass < 0.0d0)then
the_xmass=.true.
xmass=-the_xmass

```

Theory and Practice of MPI Programming

In-Ho Lee

(이인호)

Korea Research Institute of Standards and Science, Daejeon 305-340

(한국표준과학연구원)

ihlee@kriss.re.kr

<http://incredible.egloos.com>

KIAS, July 7-10, 2015

contents

Brief introduction to MPI

MPI basics

MPI practice

The 6th KIAS CAC Summer School on Parallel and Scientific Computing

7/7(Tue)

08:30 ~ 09:20	Registration		
09:20 ~ 09:30	Welcome address	CAC Director	
09:30 ~ 10:05	Research on Mechanical Response of Nanoscale Materials in UNIST Supercomputing Center	Kim, Sung Youb(UNIST)	 
10:05 ~ 10:15	Introduction to school and team project	Joo, Keehyoung(KIAS)	
10:15 ~ 10:30	Coffee break		
10:30 ~ 12:00	Theory and Practice of MPI Programming –Part 1	Lee, In-Ho(KRISS)	 
12:00 ~ 12:30	Photo		
12:30 ~ 14:00	Lunch		
14:00 ~ 15:30	Theory and Practice of MPI Programming –Part 2	Lee, In-Ho(KRISS)	 
15:30 ~ 15:45	Coffee break		
15:45 ~ 16:15	Team building		
16:15 ~ 17:45	Parallel & Scientific Computing with Python –Part 1	Joo, Keehyoung(KIAS)	 
17:45 ~ 19:00	Banquet(Pizza)		
19:00 ~	Team projects and study		

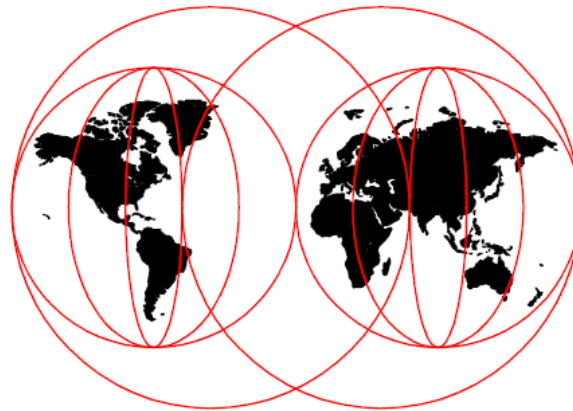
Reference



RS/6000 SP: Practical MPI Programming

*Yukiya Aoyama
Jun Nakano*

가장 중요한 참고서적
MPI를 이용한 다양한 병렬화 케이스들 소개



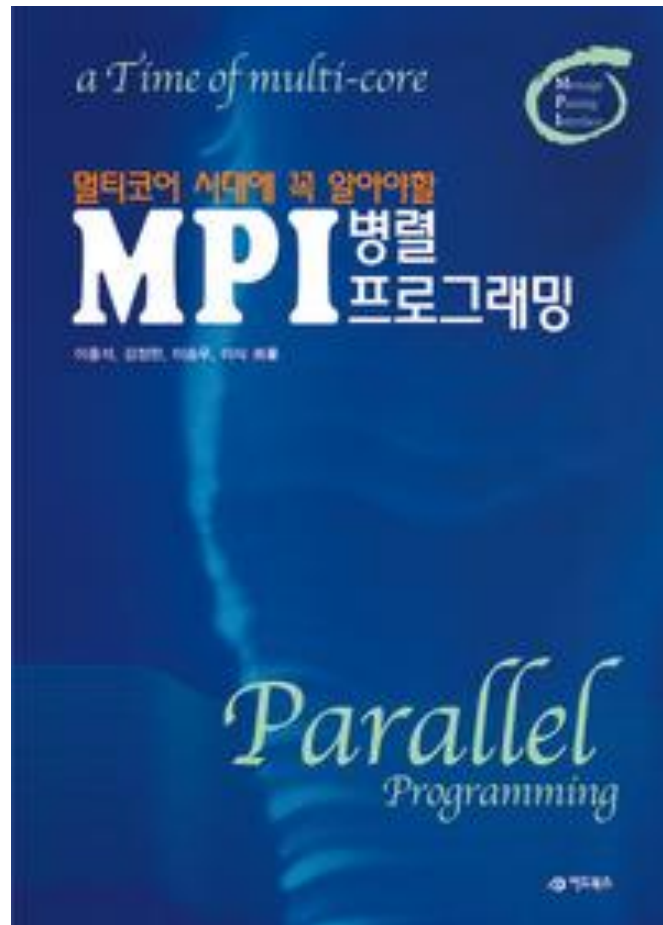
International Technical Support Organization

www.redbooks.ibm.com

SG24-5380-00

Y. Aoyama and J. Nakano, 1999

Reference



이홍석, 김정한, 이승우, 이석

<http://www.mpitutorial.com/>

Reference

Message Passing Interface (MPI)

Blaise Barney, Lawrence Livermore National Laboratory

<https://computing.llnl.gov/tutorials/mpi/>

C Language - Environment Management Routines Example

```
#include "mpi.h"
#include <stdio.h>

int main(argc,argv)
int argc;
char *argv[]; {
int numtasks, rank, rc;

rc = MPI_Init(&argc,&argv);
if (rc != MPI_SUCCESS) {
printf ("Error starting MPI program. Terminating.\n");
MPI_Abort(MPI_COMM_WORLD, rc);
}

MPI_Comm_size(MPI_COMM_WORLD,&numtasks);
MPI_Comm_rank(MPI_COMM_WORLD,&rank);
printf ("Number of tasks= %d My rank= %d\n", numtasks,rank);

/***** do some work *****/

MPI_Finalize();
}
```

Fortran - Environment Management Routines Example

```
program simple
include 'mpif.h'

integer numtasks, rank, ierr, rc

call MPI_INIT(ierr)
if (ierr .ne. MPI_SUCCESS) then
print *, 'Error starting MPI program. Terminating.'
call MPI_ABORT(MPI_COMM_WORLD, rc, ierr)
end if

call MPI_COMM_RANK(MPI_COMM_WORLD, rank, ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, numtasks, ierr)
print *, 'Number of tasks=',numtasks,' My rank=',rank

C **** do some work ****

call MPI_FINALIZE(ierr)

end
```

<http://hpc-wiki.uni-graz.at/Wiki-Seiten/MPI%20Examples.aspx>

:set paste

Reference

Google search :

Introduction to MPI – created by the PACS Training Group

Message Passing Fundamentals

Introduction to MPI – created by the PACS Training Group

All rights reserved. Do not copy or redistribute in any form.

NCSAAccess © 2001 Board of Trustees of the University of Illinois.

FORTRAN 90

Fortran 90

***A Conversion Course for Fortran 77
Programmers (The University of Manchester)***

***The University of Liverpool
Fortran 90 course notes***

두 개의 포트란 90 레퍼런스, 인터넷 자료들



fortran 90 tutorial site:incredible.egloos.com



검색

검색결과 약 443개 (0.14초)

Google.com in English 고급 검색

<http://incredible.egloos.com>

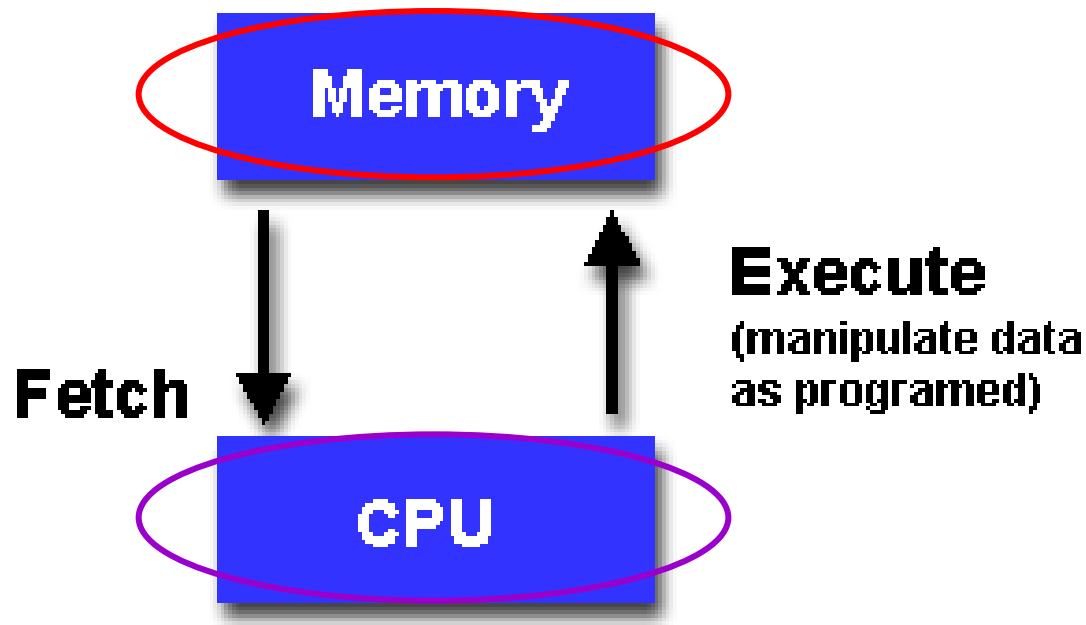
http://en.wikipedia.org/wiki/Parallel_computing

Supercomputer?

Why Parallel Computing?

- Large-scale calculations
- Fast calculations (optimization?)
- www.top500.org
- Save time - *wall clock time* (1 day vs 1 week)
- speed of light (30 cm/nanosecond)
- Performance/price

von Neumann computer



CPU gets instructions and/or data from memory, decodes the instructions and then *sequentially* performs them.

Serial $\leftrightarrow$ Parallel

associative property of addition

associative property of multiplication

commutative distributive

$$(1+2)+3 = 1+(2+3)$$

Truncation Error

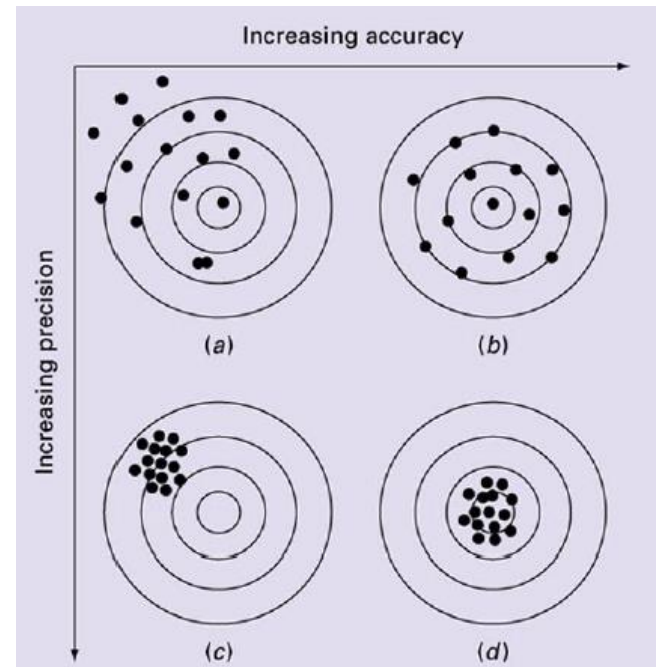
Roundoff Error

SP : 10^{-38} 10^{39}

7 decimal precision

DP : 10^{-308} 10^{308}

15 decimal precision

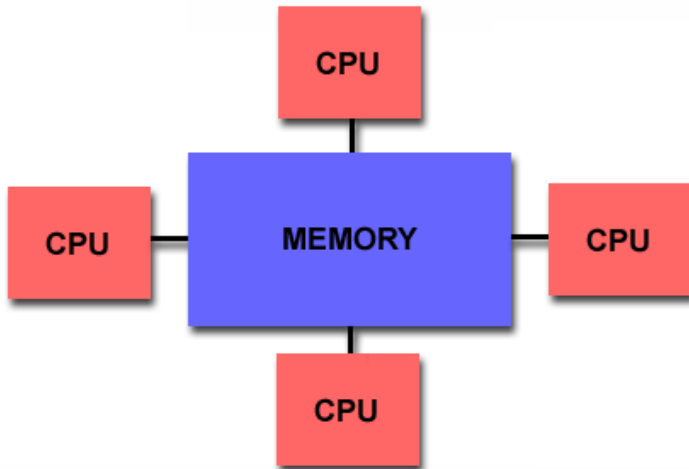


계산하는 순서가 다를 수 있다.
심지어 계산하는 CPU가 서로 다르다.

- Accuracy: how closely a computed or measured value agrees with the true values.
- Precision: how closely individual computed or measured values agree with each other.

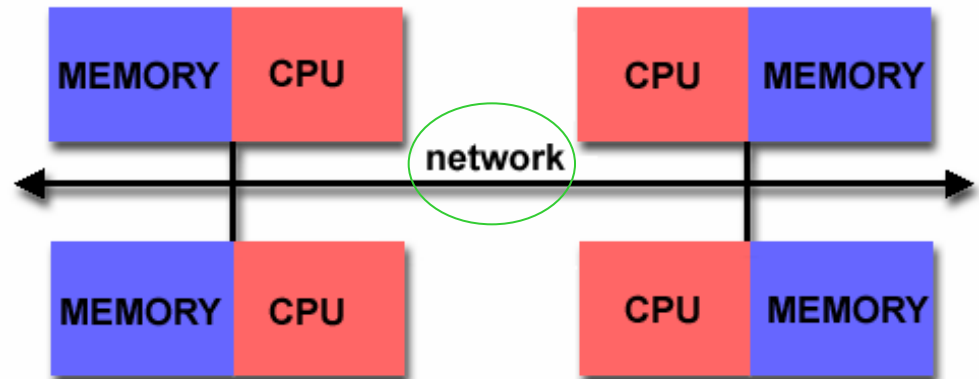
CPU & Memory

Shared Memory



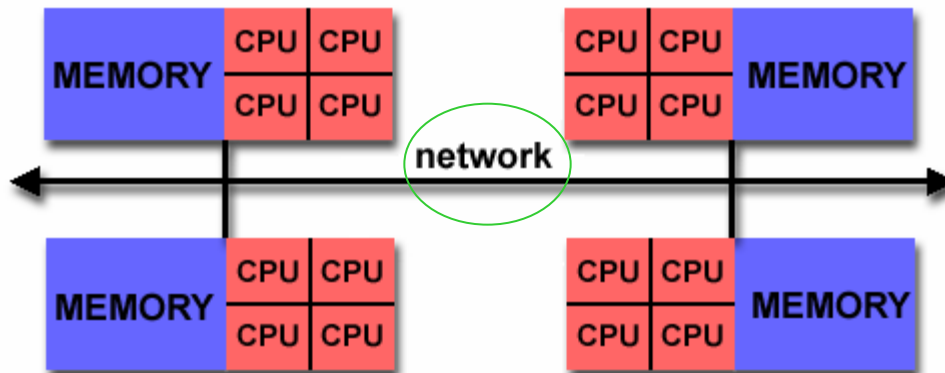
IBM SP Power3 Nodes, Regatta (Power4) servers,
SGI Power Challenge Series, Sun Enterprise Series

Distributed Memory



IBM SP (pre-Power3 nodes,
Clusters of uniprocessor nodes

Hybrid Distributed-Shared Memory

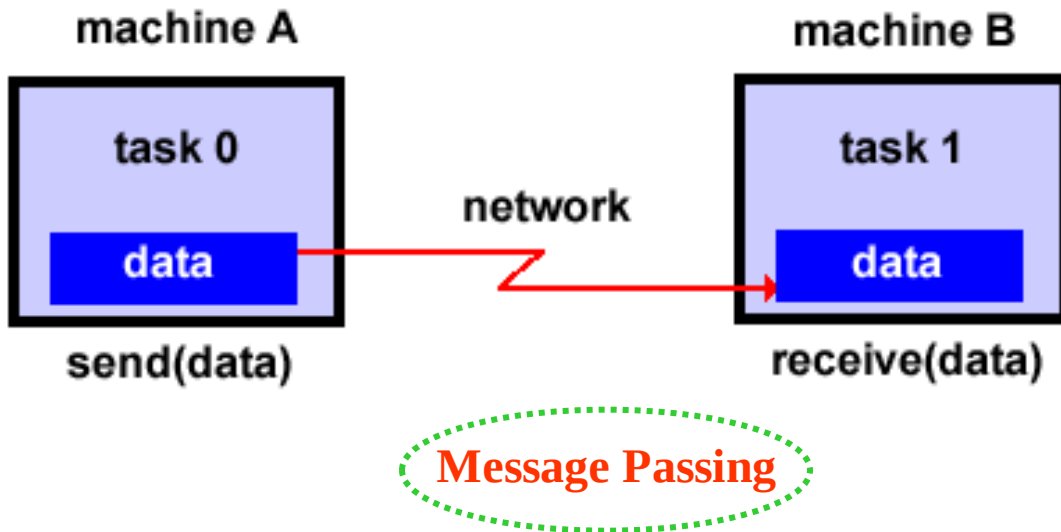


Clusters of SMPs,
IBM SP Series,
SGI Origin Series

Beowulf

- 1994 NASA, CESDIS에서 제작 **16노드**(인텔DX4), *10Mbps 이더넷*
- 지놈프로젝트, 세레라 지노믹스 사 (1200 대)

Home made?



고등과학원 cluster

Levels of parallelism

Data parallel (*fine-grained* parallel)

OpenMP and HPF Directive-based data-parallel

Parallel execution of DO loops in FORTRAN

Task parallel (*coarse-grained* parallel)

MPI, PVM (message passing)

Inter-process communication

Very general model

Hardware platforms

Great control over data location and flow in a program

Higher performance level (scalability)

Programmer has to work hard to implement!

Some General Parallel Terminology

- **Observed Speedup** (WC-serial/WC-parallel)
Scalability
- **Granularity** (coarse, fine)
- **Computation / Communication Ratio**
- **Synchronization**
- **Communications**

The purpose of parallelization is to reduce the time spent for computation. Ideally, the parallel program is p times faster than the sequential program, where p is the number of processes involved in the parallel execution, but this is not always achievable.

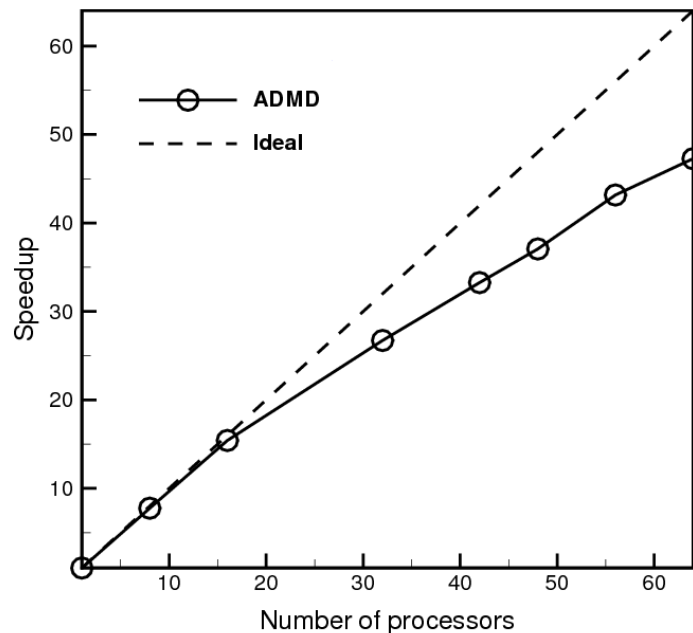
speedup

p is the number of processors.

T_p is the execution time of the algorithm with p processors.

$$S_p = \frac{T_1}{T_p}$$

Not CPU time, but wall-clock time reference



<http://en.wikipedia.org/wiki/Speedup>

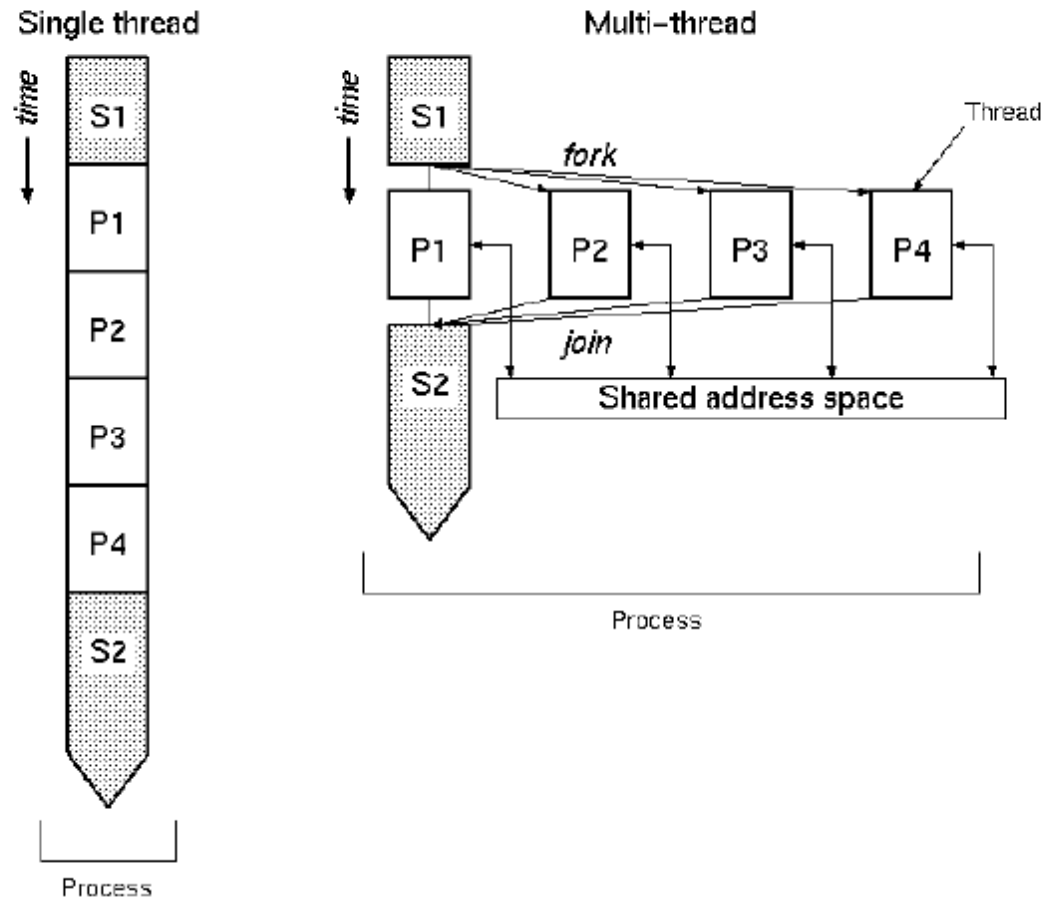
병렬화 실제 과정

- 순차 프로그램 이해
- 병렬화 가능 부분 (효율성 가정, 견적)
- Latency, bandwidth 확인 (병렬화 방식 조사)
- 프로그램 복사
- 모든 CPU에서 변수, 배열 접근 여부 확인
- 데이터 통신, 저장, 점진적으로 병렬화 시도
- 병렬화 전후 차이점 파악 (계산 결과, 실행 시간)

Embarrassingly parallel

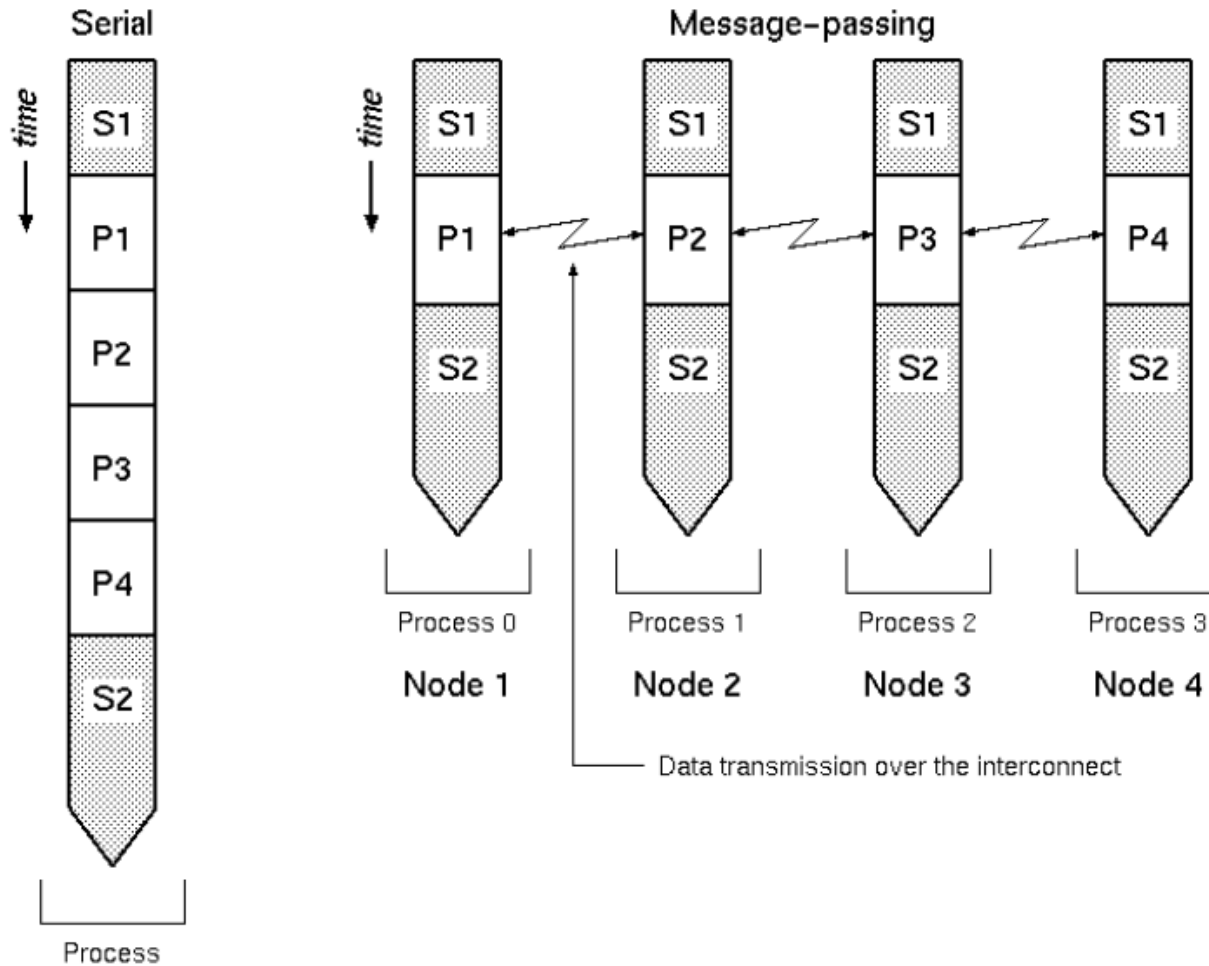
- little or no communication of results between tasks

SMP (Symmetric Multiprocessor)

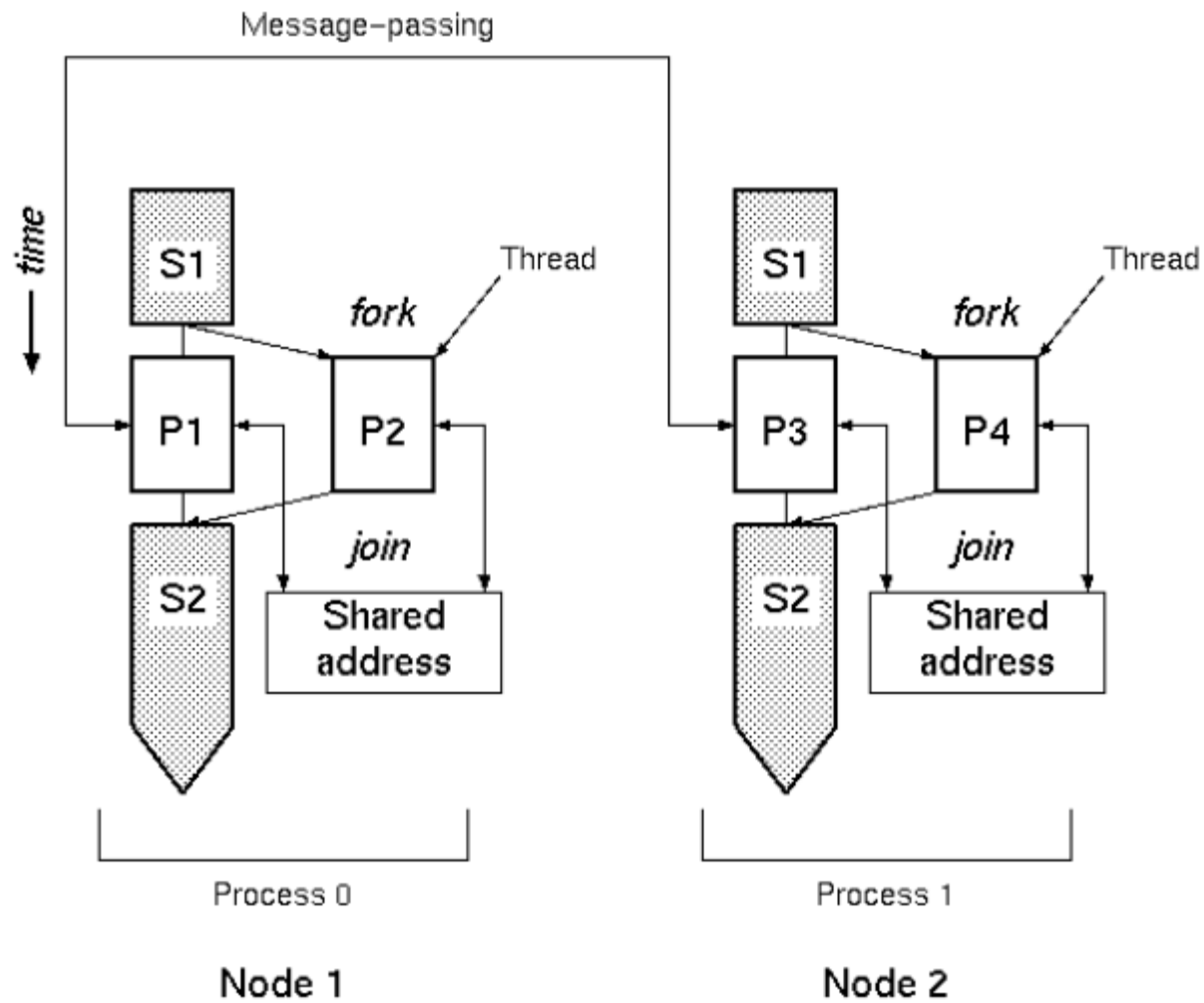


Y. Aoyama and J. Nakano

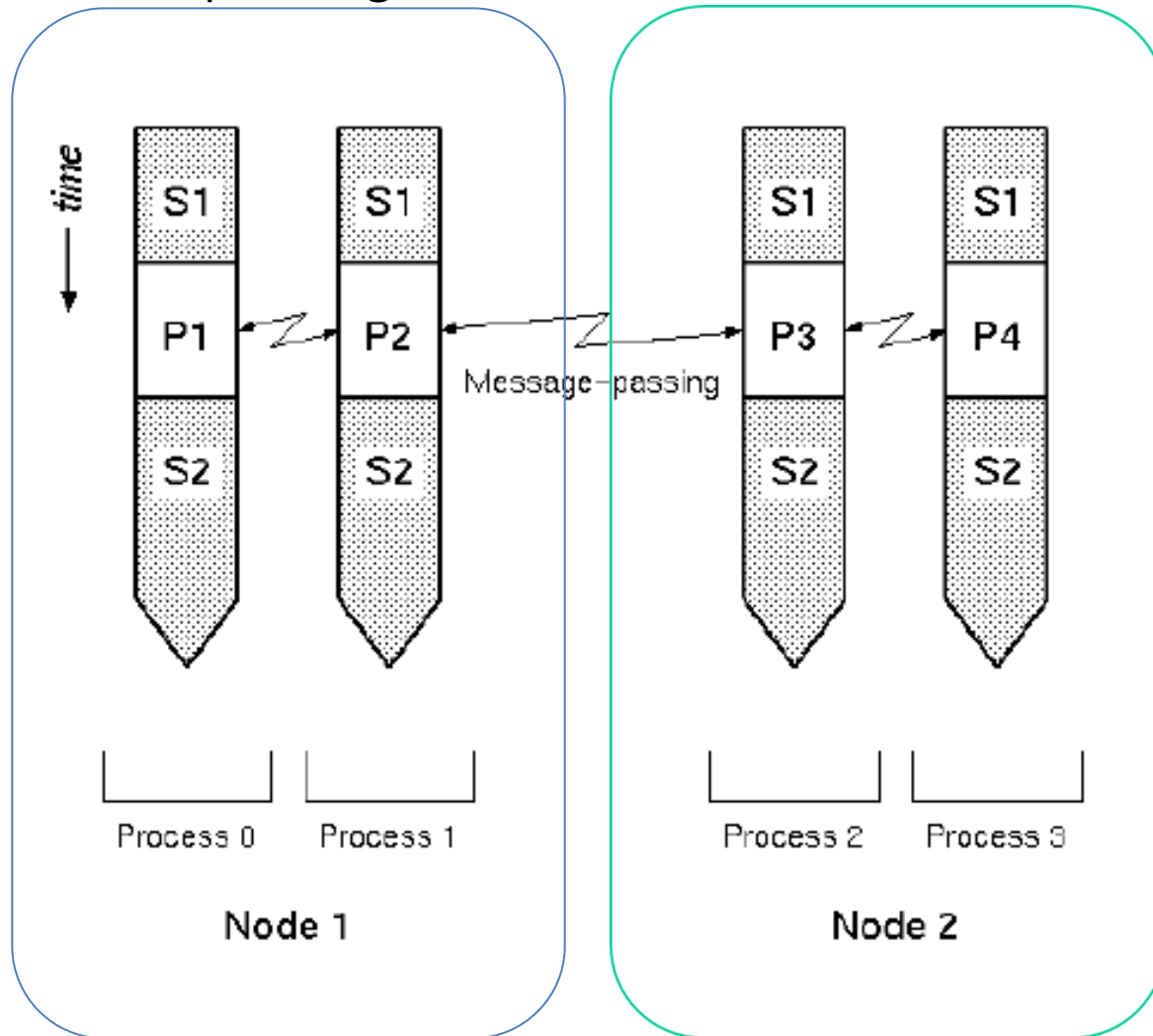
Message-Passing



Y. Aoyama and J. Nakano

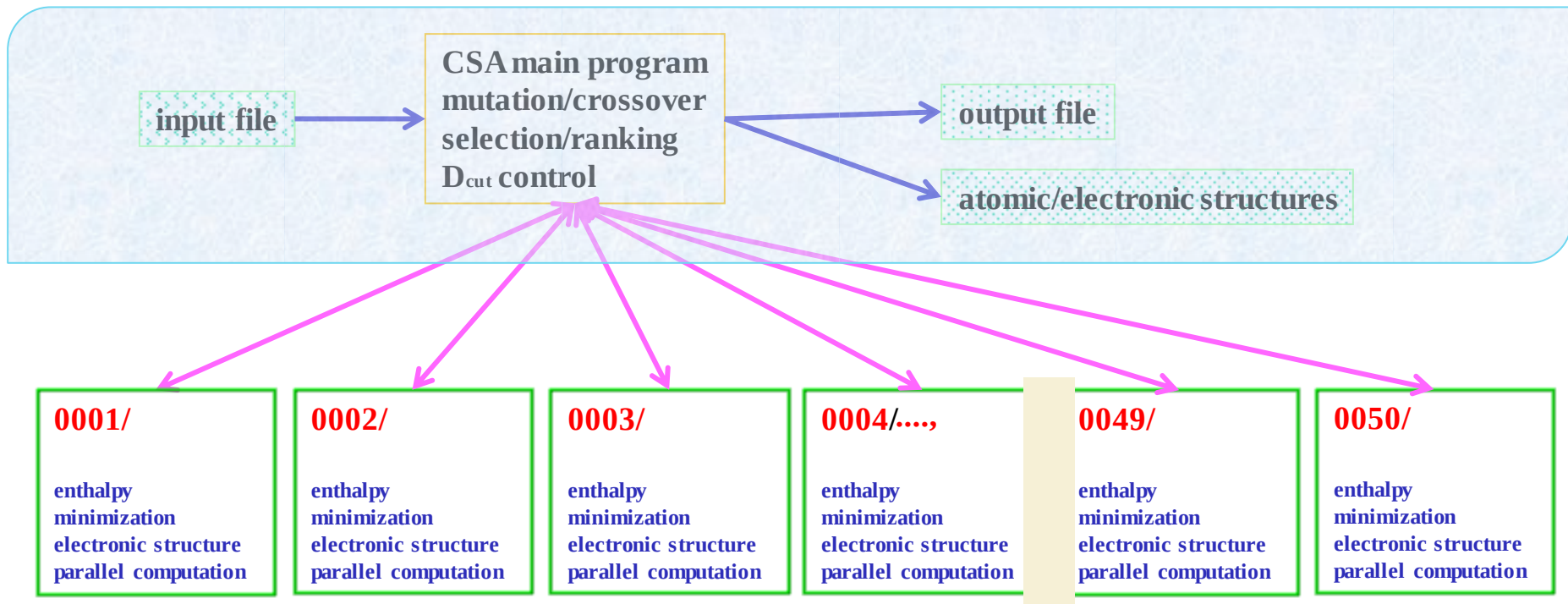


Multiple Single-Thread Processes Per Node



Y. Aoyama and J. Nakano

Parallel in parallel



기존의 병렬 프로그램을 동시에 사용하여 병렬 계산 속에 병렬 계산이 들어 갈 수 있다.
Serial main 프로그램이 여러 개의 디렉토리를 이용하여 다수의 병렬 계산들을 동시에 수행하는 방식이다.

Non-MPI, batch (1/1)

- One-serial but multiple directories
- Usual PBS-like job submission
- Data parsing over the directories

개발 완료된 프로그램을 동시에 여러 개 실행할 때 스크립트를 만들어서 실행함으로써 많은 계산들을 동시에 수행한다.

<http://incredible.egloos.com/5008405>

<http://incredible.egloos.com/4842683>

Non-MPI, master-slave, and data parsing (1/6)

- Master-slave mode
- More flexible (PBS script)
- One-serial but multiple directories
- Usual PBS-like job submission
- Data parsing over the directories

제3자의 컴퓨터 프로그램이 있고 그 내부를 잘 알지 못하는 경우,
제3자 프로그램을 새로 만들 필요가 없고 이용하기만 하면 될 경우,
또는 잘 병렬화 되어 있는 경우, 계산량이 충분히 큰 경우,
프로그램이 충분히 검정이 된 경우이고 반복적으로 사용할 수 있는 경우,
컴퓨터 자원을 독점적으로 사용하지 못할 경우,
여러 종류의 출력들을 동시에 활용하여 새로운 입력을 만들 경우,

하나의 순차 프로그램으로 사실상 병렬 프로그램이 가능하다.
유전 알고리즘과 같은 경우가 이 경우에 해당한다.
(목적 함수 계산이 매우 복잡하고 시간적으로 오래 걸리는 경우)

<http://incredible.egloos.com/4842832>

<http://incredible.egloos.com/4836430>

Non-MPI, master-slave, and data parsing (2/6)

SOLDIER.pbs (job submission)

input_file (static input)

input_file (dynamic input)

deposit/

0001/

0002/

0003/

....

0050/

Directory monitoring

Data parsing

Input generation

Job submission

call system()

inquire()

하나의 시리얼 메인 프로그램이 여러 개의 디렉토리들을 만들고 각 디렉토리에서 상당히 시간이 걸리는 제3의 프로그램을 이용해서 계산들을 진행한다. PBS를 이용해서 job을 제출한다. 제출한 job들은 계산 중으로 판단한다. 계산이 끝나면, PBS 스크립트 마지막 부분에 있는 파일생성 신호를 이용하여 계산이 끝났음을 파일형식으로 표시해 둔다. 메인 시리얼 프로그램은 각 디렉토리에서 계산 종료를 파일형식으로 확인할 수 있다. 각 디렉토리에서 계산 종료를 확인한 경우, 그 결과를 읽어 들이고, 새로운 인풋 파일을 만들어서, 그 디렉토리에서 새로운 계산을 PBS로 제출한다.

Non-MPI, master-slave, and data parsing (3/6)

```
#!/bin/sh
#PBS-l nodes=2:quad:ppn=8
#PBS-N csa_soldier
```

```
NPROCS=`wc -l < $PBS_NODEFILE`
```

```
hostname
date
```

```
cd $PBS_O_WORKDIR
```

```
# do not change filenames, e.g., stdout.log, STOP
```

```
# normal
cp INCAR_rlx INCAR
sleep 0.5
mpirun -genv I_MPI_DEBUG 5 -np $NPROCS/opt/VASP/bin/vasp.5.2.12_GRAPE_GRAPHENE_NORMAL.mpi.x > stdout.log
# accurate 550 eV
cp INCAR_rlxall INCAR
cp CONTCAR POSCAR
sleep 0.5
mpirun -genv I_MPI_DEBUG 5 -np $NPROCS/opt/VASP/bin/vasp.5.2.12_GRAPE_GRAPHENE_NORMAL.mpi.x > stdout.log
cp OUTCAR out.out1
cp CONTCAR POSCAR
# accurate 550 eV
#cp INCAR_bs INCAR
#cp CONTCAR POSCAR
#sleep 0.5
#mpirun -genv I_MPI_DEBUG 5 -np $NPROCS/opt/VASP/bin/vasp.5.2.12_GRAPE_GRAPHENE_NORMAL.mpi.x > stdout.log
```

```
STAMP=$(date +%Y%m%d_%H%M%S)_$RANDOM
echo $STAMP
cp CONTCAR /home/ihlee/csa_vasp/B28_in/deposit/CONTCAR_$STAMP
cp OUTCAR /home/ihlee/csa_vasp/B28_in/deposit/OUTCAR_$STAMP
cp EIGENVAL /home/ihlee/csa_vasp/B28_in/deposit/EIGENVAL_$STAMP
cp DOSCAR /home/ihlee/csa_vasp/B28_in/deposit/DOSCAR_$STAMP
```

```
sleep 0.5
```

```
touch STOP
```

```
echo "DONE" >> STATUS
```

한 디렉토리에서 계산이 끝날 때 그 디렉토리에 계산 종료와 관련된 정보를 적어준다.
메인 프로그램은 수시로 계산 종료를 체크한다. 파일을 읽어서 판단한다.

동일한 형식의 파일들은 시간-랜덤 넘버를 사용한 태그를 붙여서 한 곳에 보관한다.
물론, 입력, 출력 파일에 입력과 출력을 확인할 수 있는 태그가 있어야 편리하다.

! Written by In-Ho Lee, KRISS, September 11, 2013.

```
subroutine send_exe(mm,ndir,loccupied,kcmd,nwork,iseq)
  USE csa_application, OMP, ftnm, fpc
  implicit none
  integer mm,ndir,kcmd,nwork,iseq(nwork)
  logical loccupied(ndir)
  integer jd,i,ish
  real*8 r6(6),cmatrix(3,3),a1(3),a2(3),a3(3),ddg
  character*80 file_names(20),tmpname
  character*280 cmd
  real ranmar
```

```
do jd=1,ndir
  if(.not. loccupied(jd))then
    call iofilearray(jd,file_names)
    call csa_rnd_lattice_basis(mm,nwork,iseq)
    ish=ndeg-6
    if(lpbc)then
      do i=1,6
        r6(i)=qosi0(ish+i)
      enddo
      call lat2matrix(r6,cmatrix,1)
      a1(:)=cmatrix(1,:); a2(:)=cmatrix(2,:); a3(:)=cmatrix(3,:)
    endif
    if(lpbc)then
      call direct_pbc(qosi0)
    else
      call centering(qosi0)
    endif
    call write_poscar(mm,file_names(3))
    ddg=0.12d0 ; tmpname=trim(file_names(5))//'_012'
    call write_kpoints(ddg,a1,a2,a3,tmpname)
    ddg=0.06d0 ; tmpname=trim(file_names(5))//'_006'
    call write_kpoints(ddg,a1,a2,a3,tmpname)
    ddg=0.03d0 ; tmpname=trim(file_names(5))//'_003'
    call write_kpoints(ddg,a1,a2,a3,tmpname)
    ddg=0.02d0 ; tmpname=trim(file_names(5))//'_002'
    call write_kpoints(ddg,a1,a2,a3,tmpname)
    ddg=0.00d0 ; tmpname=trim(file_names(5))//'_000'
    call write_kpoints(ddg,a1,a2,a3,tmpname)
!    call sleep(1)
    call system('sleep 0.1')
    cmd='cp '//trim(file_names(3))//'_ '//trim(file_names(3))//'_ '
    cmd=trim(cmd) ; call system(cmd)
!
    cmd='cd. '//trim(file_names(1))//'_ '//qsub ./CSA_SOLDIER.pbs'
    cmd=trim(cmd) ; call system(cmd)
    loccupied(jd)=.true.
    call system('sleep 0.1')
    exit
  endif
enddo
end subroutine send_exe
```

Non-MPI, master-slave, and data parsing (4/6)

만들어진 디렉토리들에 새로운 인풋 파일 만들어주기

(기본적인 파일들은 미리 복사 해 둬)

동적으로 인풋 파일들을 만들어줌

물론, 인풋 파일에 태그를 붙여준다 출력에도 동일한 태그가 있으면 좋다

특정 디렉토리가 계산 중인지 그렇지 않은지 표시하여 둬(계산 가능, 불가능)

PBS job 제출함

제출한 job는 무조건 계산 중으로 취급한다

계산이 종료 되면 특정한 정보를 해당 디렉토리에 표시한다

메인 루틴은 이 정보를 수시로 확인하고 계산 종료를 확인한다

계산이 종료된 디렉토리는 즉각 다른 계산을 할 수 있도록 지정된다

Non-MPI, master-slave, and data parsing (5/6)

! Written by In-Ho Lee, KRISS, September 11, 2013.

```
subroutine gen_directories(ndir)
implicit none
integer ndir
character*80 string
character*280 cmd
integer isize,i

isize=4
if(ndir > 0)then
do i=1,ndir
call xnumeral(i,string,isize) ; string=trim(string)
cmd='mkdir '//trim(string) ; cmd=trim(cmd) ; call system(cmd)
cmd='cp ./CSA_SOLDIER.pbs '//trim(string)//'/' ; cmd=trim(cmd) ; call system(cmd)
cmd='cp INCAR_rlx '//trim(string)//'/' ; cmd=trim(cmd) ; call system(cmd)
cmd='cp INCAR_rlxall '//trim(string)//'/' ; cmd=trim(cmd) ; call system(cmd)
cmd='cp INCAR_bs '//trim(string)//'/' ; cmd=trim(cmd) ; call system(cmd)
cmd='cp POTCAR '//trim(string)//'/' ; cmd=trim(cmd) ; call system(cmd)
enddo
endif
if(ndir < 0)then
ndir=iabs(ndir)
do i=1,ndir
call xnumeral(i,string,isize) ; string=trim(string)
cmd='rm -rf '//trim(string)//'/' ; cmd=trim(cmd)
call system(cmd)
call sleep(1)
enddo
endif
call sleep(1)
return
end
```

!

디렉토리들 만들기, 예를 들어 아래와 같이,

0001/

0002/

....

그리고 그곳에 필요한 인풋 및 준비 파일들
복사해 두기

Non-MPI, master-slave, and data parsing (6/6)

! Written by In-Ho Lee, KRISS, September 11, 2013.

```
subroutine jobstatus(fname,i)
implicit none
character*80 fname
integer i
logical lexist
character*20 ch

i=0
inquire(file=trim(fname),exist=lexist)
if(lexist)then
open(77,file=trim(fname),form='formatted')
do
read(77,*,end=999) ch
enddo
999 continue
close(77)
if(trim(ch) == 'DONE' .or. trim(ch) == 'done') i=1
if(trim(ch) == 'Done' .or. trim(ch) == 'DOne') i=1
if(trim(ch) == 'DOnE' ) i=1
endif
return
end
```

! Written by In-Ho Lee, KRISS, September 11, 2013.

```
subroutine jobstatus0(fname)
implicit none
character*80 fname
character*200 cmd

open(77,file=trim(fname),form='formatted')
write(77,*) 'ING'
close(77)
cmd='echo "ING" >> '//trim(fname) ; cmd=trim(cmd)
call system(cmd)
call system('sleep 0.1')
return
end
```

메인 시리얼 프로그램은 앞서 만들어진 디렉토리들 중에서
계산이 종료된 디렉토리를 수시로 확인한다.
계산 종료는 파일 형식으로 확인한다. 특정 파일의 내용을 확인하여
계산 중 또는 계산 종료를 확인한다.

메인 시리얼 프로그램은 인풋 파일 준비에 따른 실제 계산과
각 디렉토리에서 얻어진 계산 결과를 결과를 정리한다.
각 디렉토리에서는 독립적으로
제3의 프로그램을 활용한 계산이 진행될 수 있다.

MPI (message passing interface)

- The Message Passing Interface Standard (MPI) is a message passing library standard based on the consensus of the MPI Forum.
- MPI is not an IEEE or ISO standard, but has in fact, become the "industry standard" for writing message passing programs on HPC platforms.
- Various architectures
- Free implementations: MPICH, LAM, CHIMP, openmpi

Intel MPI

Infiniband, Myrinet, Quadrics

MPI (message passing interface)

- MPI 자체는 병렬 라이브러리들에 대한 표준규약이다. (125개의 서브 프로그램들로 구성되어 있다.)
- MPI는 약 40개 기관이 참여하는 MPI forum에서 관리되고 있으며, 1992년 MPI 1.0 을 시작으로 현재 MPI 2.0까지 버전업된 상태이다.
- Various architectures
 - Free implementations: MPICH, LAM, CHIMP, openmpi

MPI (message passing interface)

The MPI-1 standard was defined in Spring of 1994.

- 1 This standard specifies the names, calling sequences, and results of subroutines and functions to be called from Fortran 77 and C, respectively. All implementations of MPI must conform to these rules, thus ensuring portability. MPI programs should compile and run on any platform that supports the MPI standard.
- 2 The detailed implementation of the library is left to individual vendors, who are thus free to produce optimized versions for their machines.
- 3 Implementations of the MPI-1 standard are available for a wide variety of platforms.

An MPI-2 standard has also been defined. It provides for additional features not present in MPI-1, including tools for parallel I/O, C++ and Fortran 90 bindings, and dynamic process management. At present, some MPI implementations include portions of the MPI-2 standard but the full MPI-2 is not yet available.

Top 10 Reasons to Prefer MPI Over PVM

1. MPI has more than one freely available, quality implementation.
2. MPI defines a 3rd party profiling mechanism.
3. MPI has full asynchronous communication.
4. MPI groups are solid and efficient.
5. MPI efficiently manages message buffers.
6. MPI synchronization protects 3rd party software.
7. MPI can efficiently program MPP and clusters.
8. MPI is totally portable.
9. MPI is formally specified.
10. MPI is a de facto **standard**.

Advantages of Message Passing

- **Universality** : MP model fits well on separate processors connected by fast/slow network. Matches the hardware of most of today's parallel supercomputers as well as network of workstations (NOW)
- **Expressivity** : MP has been found to be a useful and complete model in which to express parallel algorithms. It provides the control missing from data parallel or compiler based models
- **Ease of debugging** : Debugging of parallel programs remain a challenging research area. Debugging is easier for MPI paradigm than shared memory paradigm (even if it is hard to believe)

Implementations of MPI

There are a number of freely available implementations of MPI that run on a variety of platforms:

The **MPICH** implementation runs on a wide range of platforms and operating systems including Unix/Linux, Windows NT, and Windows 2000/XP Professional.

The **LAM** implementation runs on networks of Unix/Posix workstations.

MP-MPICH runs on Unix systems, Windows NT and Windows 2000/XP Professional.

The **WMPI** implementation runs on Windows platforms running Windows 95/98, ME, NT and 2000.

MacMPI is a partial implementation of MPI for Macintosh computers.

openmpi

Intel MPI

Most parallel machine vendors have optimized versions.

MPICH_GM

MPICH_G2

<http://www.mpi.nd.edu/MPI/Mpich>

<http://www-unix.mcs.anl.gov/mpi/mpich/indexold.html>

GLOBUS:

<http://www.globus.org/mpi/>

<http://exodus.physics.ucla.edu/appleseed/>

Most current MPI implementations are supported :

MPICH/p4 : over ethernet devices

MPICH/gm : Myrinet

MVAPICH : InfiniBand

MPICH/shmem : a single-node multi-processor machine

MPICH2/PMI :

Intel MPI :

EMP :

OpenMPI :

LAM :

<http://www.mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf>

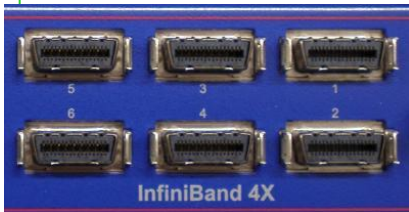
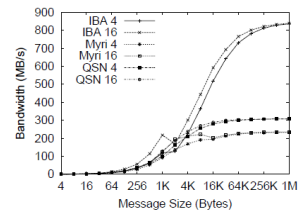
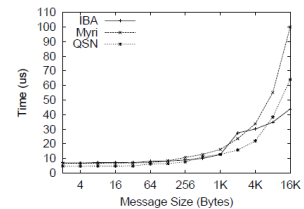
Version MPI-3.0 (September 21, 2012) is an extension of MPI-2.2.

MPICH is a freely available, portable implementation of MPI, a standard for message-passing for distributed-memory applications used in parallel computing. MPICH is Free Software and is available for most flavors of Unix-like OS (including Linux and Mac OS X).

Open MPI is a Message Passing Interface (MPI) library project combining technologies and resources from several other projects (FT-MPI, LA-MPI, LAM/MPI, and PACX-MPI).

Latency and Bandwidth

- **latency** is the time it takes to send a minimal (0 byte) message from point A to point B.
~120, ~120, ~7, ~0.5 microsecond
- **bandwidth** is the amount of data that can be communicated per unit time. ~100 Mbps, ~1 Gbps, ~1.98 Gbps, ~1 Gb/sec



높은 스�루풋, 낮은 레이턴시

Cf. CPU time Fast ethernet, Gigbit ethernet, Myrinet, Quadrics 2,
Infiniband

Computation vs communication

연결을 위해서 기본적으로 걸리는 시간 (잠복, 잠재)

Communication time ← latency, throughput

연결이 된 후 할당된 데이터 전송에 걸리는 시간: 단위시간당 처리량

Latency를 줄이기 위해서는 한 번의 연결에 집중적으로 통신하는 것이 좋다.

데이터 통신에 걸리는 시간 vs CPU를 이용한 계산시간

	Fast Ethernet	Gigabit Ethernet	<u>Myrinet</u>	Quadrics 2
latency	120 microsecond	120 microsecond	7 microsecond	0.5 microsecond
bandwidth	100 Mbps	1 Gbps	1.98 Gbps	1 Gbyte/second

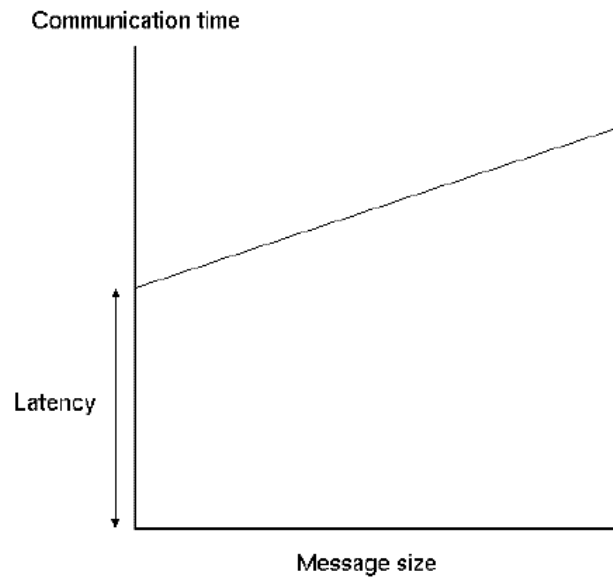


Figure 32. The Communication Time

SPMD

Start parallel job on N processors

Message passing between processors

Each processor does some calculations (*processor dependent codes are run*)

Message passing between processors

End parallel job

/usr/local/mpich/

/usr/local/mpich/bin/mpif90 a.f90

mpirun -np 8 a.out

bin/
doc/

etc/
examples/

include/
lib/

man/
sbin/

share/
www/

The 6 basic MPI routines

- MPI is used to create parallel programs based on message passing
- Usually the same program is run on multiple processors
- The 6 basic calls in MPI are:
 - MPI_INIT(ierr)
 - MPI_COMM_RANK(MPI_COMM_WORLD, **myid**, ierr)
 - MPI_COMM_SIZE(MPI_COMM_WORLD, **numprocs**, ierr)
 - MPI_Send(**buffer**, **count**, MPI_INTEGER, *destination*, tag, MPI_COMM_WORLD, ierr)
 - MPI_Recv(**buffer**, **count**, MPI_INTEGER, *source*, tag, MPI_COMM_WORLD, status, ierr)
 - call MPI_FINALIZE(ierr)

C Binding	
Format:	<code>rc = MPI_Xxxxx(parameter, ...)</code>
Example:	<code>rc = MPI_Bsend(&buf, count, type, dest, tag, comm)</code>
Error code:	Returned as "rc". MPI_SUCCESS if successful
Fortran Binding	
Format:	<code>CALL MPI_XXXXX(parameter, ..., ierr)</code> <code>call mpi_XXXXX(parameter, ..., ierr)</code>
Example:	<code>CALL MPI_BSEND(buf, count, type, dest, tag, comm, ierr)</code>
Error code:	Returned as "ierr" parameter. MPI_SUCCESS if successful

C와 FORTRAN: 일반적으로 인수의 개수가 다르다.
형식도 상이하기 때문에 반드시 포트란 언어용, C 언어용을 구분해서 사용해야 한다.

communicators

Communicator is a group of processors that can communicate with each other.

There can be many communicators.

MPI_COMM_WORLD is a **pre-defined communicator** encompassing all of the processes.

MPI_INIT: MPI 환경 초기화하기 : *유저 수준에서 바꿀 것이 사실상 없음.*

MPI_COMM_SIZE: 사용 중인 processor 숫자 반환 : *유저 수준에서 바꿀 것이 사실상 없음.*

MPI_COMM_RANK: 현 CPU의 번호 (**rank**라고 함. processor 갯수가 $nproc$ 일 때, 가능한 rank 값은 $0, 1, 2, 3, \dots, nproc-1$ 이다.) : *유저 수준에서 바꿀 것이 사실상 없음.*

두 개의 processor 간 통신: rank값들을 사용하여서 현재 processor 번호를 확인하고 준비된 데이터를 원하는 processor로 전송한다. 마찬가지로 현재의 processor번호를 확인하고 전송되어 올 데이터를 받는다. 물론, 우리는 어떤 processor로 부터 데이터가 오는지 그리고 어떤 processor가 이 데이터를 받아야 하는지 다 알고 있다.

MPI_SEND: 원하는 processor에게 데이터 전송시 사용 : 유저의 구체적인 목적이 적용됨
(원하는 데이터 형, 사이즈,...)

MPI_RECV: 원하는 processor로부터 데이터 전송받을 때 사용 : 유저의 구체적인 목적이 적용됨
(원하는 데이터 형, 사이즈,...)

MPI_FINALIZE: MPI 환경 종료하기 : *유저 수준에서 바꿀 것이 사실상 없음.*

Communicators

- Communicators
 - A parameter for most MPI calls
 - A collection of processors working on some part of a parallel job
 - `MPI_COMM_WORLD` is defined in the MPI include file as all of the processors in your job
 - Can create subsets of `MPI_COMM_WORLD`
 - Processors within a communicator are assigned numbers `0` to `n-1`

Include files

- The MPI include file
 - C: `mpi.h`
 - Fortran: `mpif.h` (a f90 module is a good place for this)
- Defines many constants used within MPI programs
- In C, defines the interfaces for the functions
- Compilers know where to find the include files

```
#include "mpi.h"  
  
#include <stdio.h>  
  
#include <math.h>
```

```
program main  
implicit none  
include 'mpif.h'
```

/usr/local/mpich/include

basex11.h

f90base/

f90choice/

mpe.h

mpe_graphics.h

mpe_log.h

mpeexten.h

mpef.h

mpetools.h

mpi.h

mpi2c++/

mpi_errno.h

mpi_fortdefs.h

mpidefs.h

mpif.h

mpio.h

mpiof.h

protofix.h

include "mpif.h" ! Constants used by MPI, e.g. constant integer value MPI_COMM_WORLD
integer istatus(MPI_STATUS_SIZE) ! MPI_STATUS_SIZE는 위에서 선언한 include문으로
불러들인 내용에서 이미 정의된 것들이다.

integer nproc, me, ierr, idestination, isource, n_array_length, itag

call MPI_Init(ierr)
call MPI_Comm_size(MPI_COMM_WORLD, nproc, ierr)
call MPI_Comm_rank(MPI_COMM_WORLD, me, ierr)

itag=19
idestination=1
call MPI_Send(real_array_user, n_array_length, MPI_REAL8, idestination, itag, MPI_COMM_WORLD, ierr)
real*8 형태의 데이터가 가야할 곳 지정해주어야 한다. 물론, 이 데이터의 크기도 보내는 곳에서 지정해줘야 한다. 특정 노드에서 정보를 보내기 때문에 위 함수는 특정 노드에서 불러져야 한다.

itag=19
isource=0
call MPI_Recv(real_array_user, n_array_length, MPI_REAL8, isource, itag, MPI_COMM_WORLD, istatus, ierr)
데이터를 받는 쪽에서는 그 형태와 크기를 알고 있어야 하며, 어디에서부터 출발했는지를 알아야 한다.

user-defined tags provided in MPI for user convenience in organizing application

Standard, Blocking Send

- MPI_Send: Sends data to another processor
- Use MPI_Receive to "get" the data
- C
 - MPI_Send(&buffer, count, datatype, *destination*, tag, communicator);
- Fortran
 - Call MPI_Send(buffer, count, datatype, *destination*, tag, communicator, ierr)
- Call blocks until message on the way

Basic C Datatypes in MPI

MPI Datatype	C datatype
MPI_CHAR	signed char
MPI_SHORT	signed short int
MPI_INT	signed int
MPI_LONG	signed long int
MPI_UNSIGNED_CHAR	unsigned char
MPI_UNSIGNED_SHORT	unsigned short int
MPI_UNSIGNED_INT	unsigned int
MPI_UNSIGNED_LONG	unsigned long int
MPI_FLOAT	float
MPI_DOUBLE	double
MPI_LONG_DOUBLE	long double
MPI_BYTE	
MPI_PACKED	

Language	Script Name	Underlying Compiler
C	mpicc	gcc
	mpigcc	gcc
	mpiicc	icc
	mpipgcc	pgcc
C++	mpiCC	g++
	mpig++	g++
	mpiicpc	icpc
	mpipgCC	pgCC
Fortran	mpif77	g77
	mpigfortran	gfortran
	mpiifort	ifort
	mpipgf77	pgf77
	mpipgf90	pgf90

-fast

-C

ifort

-CB

-check all

순차, 병렬 프로그램 개발 단계에서 활용한다.
기계적으로 에러를 찾아낼 수 있다.
프로그램 실행할 때.....

-CB -check all -warn interface -assume realloc_lhs

Compiler	Option	Example
Intel	-v	ifort -v
PGI	-v	pgf90 -v
GNU	-v --version	g++ --version

C/MPI version of “Hello, World”

```
#include <stdio.h>
#include <mpi.h>
int main(argc, argv)
int argc;
char *argv[ ];
{
    int myid, numprocs;

    MPI_Init (&argc, &argv) ;
    MPI_Comm_size(MPI_COMM_WORLD, &numprocs) ;
    MPI_Comm_rank(MPI_COMM_WORLD, &myid) ;
    printf("Hello from %d\n", myid) ;
    printf("Numprocs is %d\n", numprocs) ;
    MPI_finalize();
}
```

Fortran/MPI version of “Hello, World”

```
program hello
include 'mpif.h'
integer myid, ierr, numprocs

call MPI_INIT( ierr)
call MPI_COMM_RANK( MPI_COMM_WORLD, myid, ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, numprocs, ierr)

write (*,*) "Hello from ", myid
write (*,*) "Numprocs is", numprocs
call MPI_FINALIZE(ierr)
stop
end
```

Minimal MPI program

- Every MPI program needs these...

- C version

```
#include <mpi.h> /* the mpi include file */
/* Initialize MPI */
ierr=MPI_Init(&argc, &argv);
/* How many total PEs are there */
ierr=MPI_Finalize();
```

- Commonly Used...

- C version

```
ierr=MPI_Comm_size(MPI_COMM_WORLD, &nPEs);
/* What node am I (what is my rank?) */
ierr=MPI_Comm_rank(MPI_COMM_WORLD, &iam);
...
```

In C MPI routines are functions and return an error value

Minimal MPI program

- Every MPI program needs these...
 - Fortran version

```
include 'mpif.h'  ! MPI include file
c    Initialize MPI
      call MPI_Init(ierr)
c    Find total number of PEs
      call MPI_Comm_size(MPI_COMM_WORLD, nPES, ierr)
c    Find the rank of this node
      call MPI_Comm_rank(MPI_COMM_WORLD, iam, ierr)
      ...
      call MPI_Finalize(ierr)
```

In Fortran, MPI routines are subroutines, and last parameter is an error-handling value.

MPI “Hello, World”

- A parallel hello world program
 - Initialize MPI
 - Have each node print out its node number
 - Quit MPI

myid=0 : special

1 node calculation	: 0
2 node calculation	: 0, 1
3 node calculation,.....	: 0, 1, 2

At least one node is involved.

<https://computing.llnl.gov/tutorials/mpi/exercise.html>

Advantages of Message Passing

- Performance:
 - This is the most compelling reason why MP will remain a permanent part of parallel computing environment
 - As modern CPUs become faster, management of their caches and the memory hierarchy is the key to getting most out of them
 - MP allows a way for the programmer to explicitly associate specific data with processes and allows the compiler and cache management hardware to function fully
 - Memory bound applications can exhibit superlinear speedup when run on multiple PEs compare to single PE of MP machines

Background on MPI

- MPI - Message Passing Interface
 - Library standard defined by committee of vendors, implementers, and parallel programmer
 - Used to create parallel SPMD programs based on message passing
- Available on almost all parallel machines in C and Fortran
- About 125 routines including advanced routines
- 6 basic routines

Key Concepts of MPI

- Used to create parallel **SPMD** programs based on message passing
- Normally the same program is running on several different nodes
- Nodes communicate using message passing

Data types

- Data types
 - When sending a message, it is given a data type
 - Predefined types correspond to "normal" types
 - `MPI_REAL` , `MPI_FLOAT` -Fortran and C real
 - `MPI_DOUBLE_PRECISION`, (`mpi_real8`), `MPI_DOUBLE` – Fortran and C double
 - `MPI_INTEGER` and `MPI_INT` - Fortran and C integer
 - Can create user-defined types

Basic Communications in MPI

- Data values are transferred from one processor to another
 - One process sends the data
 - Another receives the data
- Standard, Blocking
 - Call does not return until the message buffer is free to be reused
- Standard, Nonblocking
 - Call indicates a start of send or received, and another call is made to determine if finished

Basic MPI Send and Receive

- A parallel program to send & receive data
 - Initialize MPI
 - Have processor 0 send an integer to processor 1
 - Have processor 1 receive an integer from processor 0
 - Both processors print the data
 - Quit MPI

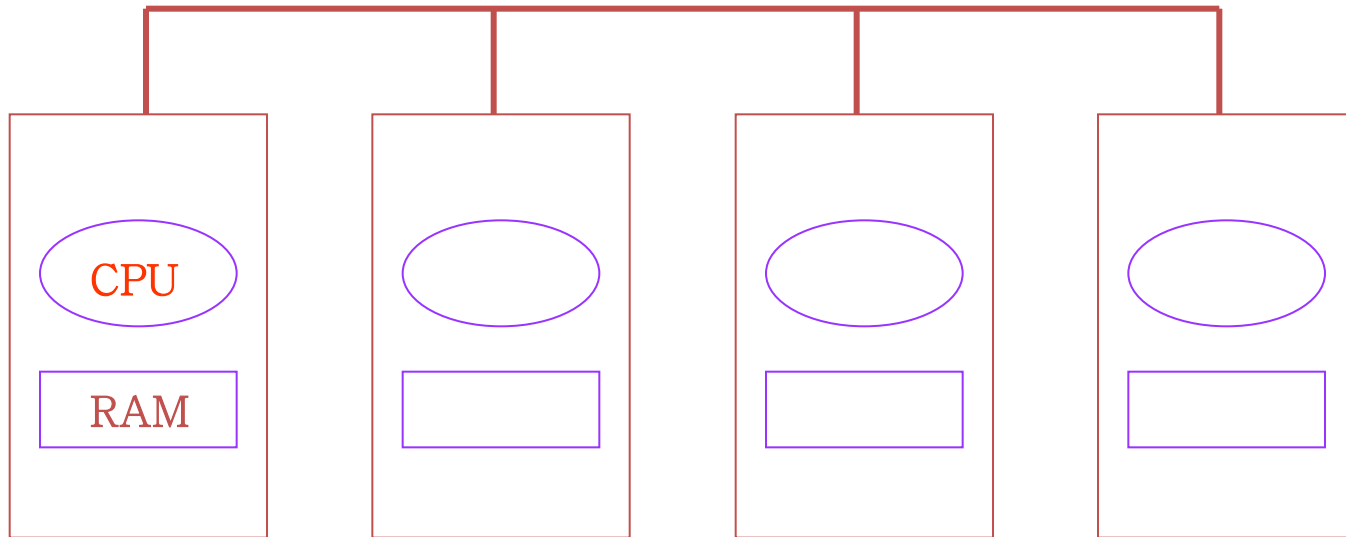
Simple Send & Receive Program

```
program send_recv
include "mpif.h"
! This is MPI send - recv program
integer myid, ierr,numprocs
integer itag, isource, idestination, kount
integer buffer
integer istatus(MPI_STATUS_SIZE)

call MPI_INIT(ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, numprocs, ierr)
itag=1234
isource=0
idestination=1
kount=1
if(myid .eq. isource)then
    buffer=5678
    Call MPI_Send(buffer, kount, MPI_INTEGER, idestination, itag, MPI_COMM_WORLD, ierr)
    write(*,*)"processor", myid, " sent ", buffer
endif
if(myid .eq. idestination)then
    Call MPI_Recv(buffer, kount, MPI_INTEGER, isource, itag, MPI_COMM_WORLD, istatus, ierr)
    write(*,*)"processor", myid, " got ", buffer
endif

call MPI_FINALIZE(ierr)
stop
end
```

A set of computing nodes connected via networks



MPI_Send

Call MPI_Send(buffer, count, datatype, *destination*, tag, communicator, ierr)

- buffer: The data
- count : Length of source array (in elements, 1 for scalars)
- datatype : Type of data, for example :
MPI_DOUBLE_PRECISION, MPI_INT, etc
- *destination* : Processor number of destination processor in communicator
- tag : Message type (arbitrary integer)
- communicator : Your set of processors
- ierr : Error return (Fortran only)

Standard, Blocking Receive

- Call blocks until message is in buffer
- C
 - `MPI_Recv(&buffer, count, datatype, source, tag, communicator, &status);`
- Fortran
 - Call `MPI_RECV(buffer, count, datatype, source, tag, communicator, istatus, ierr)`
- Status - contains information about incoming message
 - C
 - `MPI_Status status;`
 - Fortran
 - Integer `istatus(MPI_STATUS_SIZE)`

MPI_Recv

Call MPI_Recv(buffer, icount, datatype, *isource*, itag, communicator, istatus, ierr)

- buffer: The data
- icount : Max. number of elements that can be received
- datatype : Type of data, for example :
MPI_DOUBLE_PRECISION, MPI_INT, etc
- *isource*: Processor number of source processor in communicator
- itag : Message type (arbitrary integer)
- communicator: Your set of processors
- istatus: Information about message
- ierr : Error return (Fortran only)

```

program hello
  implicit none
  include "mpif.h"
  character(LEN=12) :: inmsg, message
  integer i, ierr, me, nproc, itag, kount
  integer istatus(MPI_STATUS_SIZE)

  call MPI_Init(ierr)
  call MPI_Comm_size(MPI_COMM_WORLD, nproc, ierr)
  call MPI_Comm_rank(MPI_COMM_WORLD, me, ierr)
  itag = 100 ; kount=12

  if (me == 0) then
    message = "Hello, world"
    do i = 1, nproc-1
      call MPI_Send(message, kount, MPI_CHARACTER, i, itag, MPI_COMM_WORLD, ierr)
    end do
    write(*,*) "process", me, ":", message
  else
    call MPI_Recv(inmsg, kount, MPI_CHARACTER, 0, itag, MPI_COMM_WORLD, istatus, ierr)
    write(*,*) "process", me, ":", inmsg
  end if

  call MPI_Finalize(ierr)
end program hello

```

Simple Send & Receive Program

```
#include <stdio.h>
#include "mpi.h"
int main(argc,argv)
int argc;
char *argv[];
{
    int myid;
    int tag,source,destination,count;
    int buffer;
    MPI_Status status;
    MPI_Init(&argc,&argv);
    MPI_Comm_rank(MPI_COMM_WORLD,&myid);
    tag=1234;
    source=0;
    destination=1;
    count=1;
    if(myid == source){
        buffer=5678;
        MPI_Send(&buffer,count,MPI_INT,destination,tag,MPI_COMM_WORLD);
        printf("processor %d sent %d\n",myid,buffer);
    }
    if(myid == destination){
        MPI_Recv(&buffer,count,MPI_INT,source,tag,MPI_COMM_WORLD,&status);
        printf("processor %d got %d\n",myid,buffer);
    }
    MPI_Finalize();
}
```

```
subroutine Get_data(a, b, n, myid, nproc)
```

```
IMPLICIT NONE
```

```
real*8 a, b
```

```
integer n, myid, nproc
```

C

```
INCLUDE 'mpif.h'
```

```
integer source, dest, tag, istatus(MPI_STATUS_SIZE), ierr
```

```
data source /0/
```

C

```
if (myid == 0) then
```

```
print *, 'Enter a, b and n'
```

```
read *, a, b, n
```

C

```
do dest = 1, nproc-1
```

```
tag = 0
```

```
call MPI_SEND(a, 1, MPI_REAL8, dest, tag, MPI_COMM_WORLD, ierr)
```

```
tag = 1
```

```
call MPI_SEND(b, 1, MPI_REAL8, dest, tag, MPI_COMM_WORLD, ierr)
```

```
tag = 2
```

```
call MPI_SEND(n, 1, MPI_INTEGER, dest, tag, MPI_COMM_WORLD, ierr)
```

```
end do
```

```
else
```

```
tag = 0
```

```
call MPI_RECV(a, 1, MPI_REAL8, source, tag, MPI_COMM_WORLD, istatus, ierr)
```

```
tag = 1
```

```
call MPI_RECV(b, 1, MPI_REAL8, source, tag, MPI_COMM_WORLD, istatus, ierr)
```

```
tag = 2
```

```
call MPI_RECV(n, 1, MPI_INTEGER, source, tag, MPI_COMM_WORLD, istatus, ierr)
```

```
end if
```

```
return
```

```
end
```

```

PROGRAM simple_send_and_receive
INCLUDE 'mpif.h'
INTEGER myid, ierr, istatus(MPI_STATUS_SIZE)
REAL a(100)
C Initialize MPI:
  call MPI_INIT(ierr)
C Get my rank:
  call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ierr)
C Process 0 sends, process 1 receives:
  if( myid .eq. 0 )then
    call MPI_SEND( a, 100, MPI_REAL, 1, 17, MPI_COMM_WORLD, ierr)
  else if ( myid .eq. 1 )then
    call MPI_RECV( a, 100, MPI_REAL, 0, 17, MPI_COMM_WORLD, istatus, ierr )
  endif
C Terminate MPI:
  call MPI_FINALIZE(ierr)
END

```

Process 0 sends a message to process 1, and process 1 receives it.

```

include 'mpif.h'
integer myid, numprocs, ierr, ntemp
integer istatus (MPI_STATUS_SIZE)
real*8 sum
call MPI_INIT(ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, numprocs, ierr)
sum=myid
if(myid == 0)then
do i=1,numprocs-1
call MPI_RECV(ntemp,1,MPI_INTEGER, i,i, MPI_COMM+WORLD, istatus, ierr)
sum=sum+ntemp
enddo
average=sum/numprocs ; write(6,*) 'the average is ', average
else
call MPI_SEND(myid,1,MPI_INTEGER,0,myid,MPI_COMM_WORLD, ierr)
endif
call MPI_FINALIZE(ierr)
end

```

```

subroutine get_input(aa,bb,nn)
include "mpif.h"
real*8 aa,bb
integer nn,myid
integer ierr

call MPI_COMM_RANK(MPI_COMM_WORLD,myid,ierr)

if(myid == 0)then
print *, 'Enter aa, bb, nn'
read *, aa,bb,nn
endif

call MPI_BCAST(aa,1, MPI_REAL8, 0, MPI_COMM_WORLD, ierr)
call MPI_BCAST(bb,1, MPI_REAL8, 0, MPI_COMM_WORLD, ierr)
call MPI_BCAST(nn,1, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr)

end subroutine get_input

```

```

implicit none
integer n, np, i, j, ierr, master, num
real*8 h, result, a, b, integral, pi, my_a, my_range, MPI_WTime, start_time, end_time, my_result
include "mpif.h"
integer Iam, source, dest, tag, istatus(MPI_STATUS_SIZE)
data master/0/
call MPI_Init(ierr) ; call MPI_Comm_rank(MPI_COMM_WORLD, Iam, ierr)
call MPI_Comm_size(MPI_COMM_WORLD, np, ierr)
pi = acos(-1.0d0) ; a = 0.0d0 ; b = pi*1.d0/2.d0 ; dest = 0 ; tag = 123
if(Iam == master) then
print *, 'The requested number of processors =', p ; print *, 'Enter total number of increments across all processors'
read(*,*) n
start_time = MPI_Wtime()
endif
call MPI_Bcast(n, 1, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr)
h = (b-a)/n ; num = n/np ; my_range = (b-a)/np ; my_a = a + Iam*my_range
my_result = integral(my_a, num, h)
write(*, "('Process ', i2, ' has the partial result of', f10.6)") Iam, my_result
call MPI_Reduce(my_result, result, 1, MPI_REAL8, MPI_SUM, dest, MPI_COMM_WORLD, ierr)
if(Iam == master) then
print *, 'The result =', result
end_time = MPI_Wtime() ; print *, 'elapsed time is ', end_time-start_time, ' seconds'
endif
call MPI_Finalize(ierr) ; stop ; end

```

Status

- In C
 - status is a structure of type MPI_Status which contains three fields MPI_SOURCE, MPI_TAG, and MPI_ERROR
 - status.MPI_SOURCE, status.MPI_TAG, and status.MPI_ERROR contain the source, tag, and error code respectively of the received message
- In Fortran
 - status is an array of INTEGERS of length MPI_STATUS_SIZE, and the 3 constants MPI_SOURCE, MPI_TAG, MPI_ERROR are the indices of the entries that store the source, tag, & error
 - status(MPI_SOURCE), status(MPI_TAG), status(MPI_ERROR) contain respectively the source, the tag, and the error code of the received message.

```

! Program hello.ex1.f
! Parallel version using MPI calls
! Modified from basic version so that workers send back
! a message to the master, who prints out a message for each worker
program hello
implicit none
integer, parameter:: DOUBLE=kind(1.0d0), SINGLE=kind(1.0)
include "mpif.h"
character(LEN=12) :: inmsg,message
integer :: i,ierr,me,nproc,itag,iwrank
integer, dimension(MPI_STATUS_SIZE) :: istatus
!
call MPI_Init(ierr)
call MPI_Comm_size(MPI_COMM_WORLD,nproc,ierr)
call MPI_Comm_rank(MPI_COMM_WORLD,me,ierr)
itag = 100
!
if (me == 0) then
message = "Hello, world"
do i = 1,nproc-1
call MPI_Send(message,12,MPI_CHARACTER,i,itag,MPI_COMM_WORLD,ierr)
end do
write(*,*) "process", me, ":", message
do i = 1,nproc-1
call MPI_Recv(iwrank,1,MPI_INTEGER,MPI_ANY_SOURCE,itag,MPI_COMM_WORLD,istatus,ierr)
write(*,*) "process", iwrank, ":Hello, back"
end do
else
call MPI_Recv(inmsg,12,MPI_CHARACTER,0,itag,MPI_COMM_WORLD,istatus,ierr)
call MPI_Send(me,1,MPI_INTEGER,0,itag,MPI_COMM_WORLD,ierr)
end if
call MPI_Finalize(ierr)
end program hello

```

Program Example1

```
implicit none
integer n, p, i, j,num
real h, result, a, b, integral, pi
real my_a,my_range
pi = acos(-1.0) !! = 3.14159...
a = 0.0 !! lower limit of integration
b = pi*1./2. !! upper limit of integration
p = 4 !! number of processes (partitions)
n = 100000 !! total number of increments
h = (b-a)/n !! length of increment
num= n/p !! number of calculations done by each process
result = 0.0 !! stores answer to the integral
do i=0,p-1 !! sum of integrals over all processes
  my_range = (b-a)/p
  my_a = a + i*my_range
  result = result + integral(my_a,num,h)
enddo
print *,'The result =',result
stop
end

real function integral(a,n,h)
implicit none
integer n, i, j
real h, h2, aij, a
real fct, x
fct(x) = cos(x) !! kernel of the integral
integral = 0.0 !! initialize integral
h2 = h/2.
do j=0,n-1 !! sum over all "j" integrals
  aij = a+j*h !! lower limit of "j" integral
  integral = integral + fct(aij+h2)*h
enddo
return
end
```

Buffer and Deadlock

safe

```
if (rank == 0 )then
call mpi_send(sbuff,kount,MPI_REAL8, 1, itag, MPI_COMM_WORLD, ierr)
call mpi_recv(rbuff,kount,MPI_REAL8, 1, itag, MPI_COMM_WORLD, istatus,ierr)
else if (rank ==1 )then
call mpi_recv(rbuff,kount,MPI_REAL8, 0, itag, MPI_COMM_WORLD, istatus,ierr)
call mpi_send(sbuff, kount, MPI_REAL8, 0, itag, MPI_COMM_WORLD, ierr)
endif
```

Deadlock; 수렁, 교착

```
if (rank == 0 )then
call mpi_recv(rbuff,kount,MPI_REAL8, 1, itag, MPI_COMM_WORLD, istatus,ierr)
call mpi_send(sbuff,kount,MPI_REAL8, 1, itag, MPI_COMM_WORLD, ierr)
else if (rank ==1 )then
call mpi_recv(rbuff,kount,MPI_REAL8, 0, itag, MPI_COMM_WORLD, istatus, ierr)
call mpi_send(sbuff, kount, MPI_REAL8, 0, itag, MPI_COMM_WORLD, ierr)
endif
```

Buffering dependent

```
if (rank == 0 )then
call mpi_send(sbuff,kount,MPI_REAL8, 1, itag, MPI_COMM_WORLD, ierr)
call mpi_recv(rbuff,kount,MPI_REAL8, 1, itag, MPI_COMM_WORLD, istatus,ierr)
else if (rank ==1 )then
call mpi_send(sbuff, kount, MPI_REAL8, 0, itag, MPI_COMM_WORLD, ierr)
call mpi_recv(rbuff,kount,MPI_REAL8, 0, itag, MPI_COMM_WORLD, istatus, ierr)
endif
```

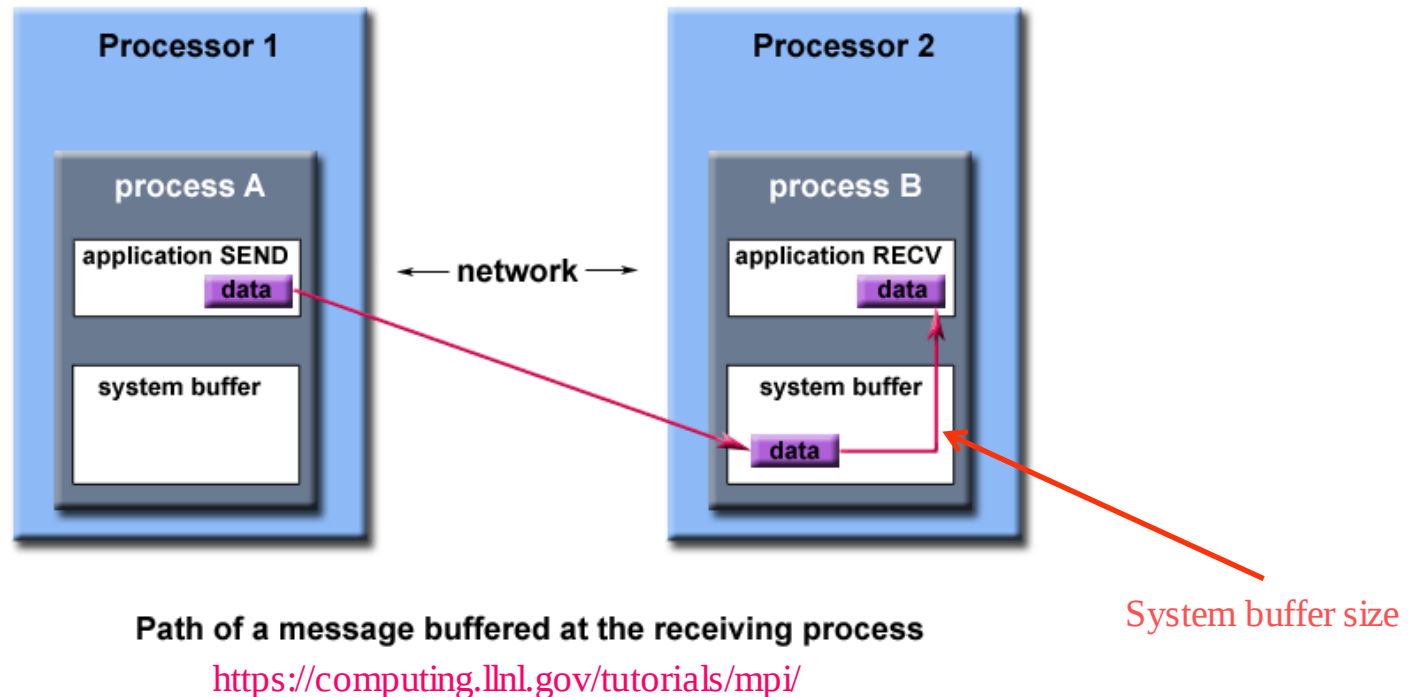
kount is large → deadlock occurs

Non-blocking communications

The programmer should be able to explain why the program does not (or does) deadlock.

Note that increasing array dimensions and message sizes have no effect on the safety of the protocol.

NCSA Access © 2001 Board of Trustees of the University of Illinois.



```

PROGRAM safe_exchange
INCLUDE 'mpif.h'
INTEGER myrank, ierr, istatus(MPI_STATUS_SIZE)
REAL a(100), b(100)
C Initialize MPI:
    call MPI_INIT(ierr)
C Get my rank:
    call MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
C Process 0 receives and sends; process 1 sends and receives
    if( myrank .eq. 0 )then
        call MPI_RECV( b, 100, MPI_REAL, 1, 19, MPI_COMM_WORLD, istatus, ierr )
        call MPI_SEND( a, 100, MPI_REAL, 1, 17, MPI_COMM_WORLD, ierr)
    else if ( myrank .eq. 1 )then
        call MPI_SEND( a, 100, MPI_REAL, 0, 19, MPI_COMM_WORLD, ierr )
        call MPI_RECV( b, 100, MPI_REAL, 0, 17, MPI_COMM_WORLD, istatus, ierr)
    endif
C Terminate MPI:
    call MPI_FINALIZE(ierr)
END

```

```
PROGRAM depends_on_buffering
INCLUDE 'mpif.h'
INTEGER myrank, ierr, istatus(MPI_STATUS_SIZE)
REAL a(100), b(100)
```

C Initialize MPI:

```
call MPI_INIT(ierr)
```

C Get my rank:

```
call MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
```

C Process 0 sends and receives; same for process 1

```
if( myrank .eq. 0 )then
call MPI_SEND( a, 100, MPI_REAL, 1, 17, MPI_COMM_WORLD, ierr)
call MPI_RECV( b, 100, MPI_REAL, 1, 19, MPI_COMM_WORLD, istatus, ierr )
else if ( myrank .eq. 1 )then
call MPI_SEND( a, 100, MPI_REAL, 0, 19, MPI_COMM_WORLD, ierr )
call MPI_RECV( b, 100, MPI_REAL, 0, 17, MPI_COMM_WORLD, istatus, ierr)
endif
```

C Terminate MPI:

```
call MPI_FINALIZE(ierr)
END
```

Again, process 0 attempts to exchange messages with process 1. This time, both processes send first, then receive. Success depends on the availability of buffering in MPI.

If the message sizes are increased, sooner or later the program will deadlock.

```

PROGRAMprobable_deadlock
INCLUDE 'mpif.h'
INTEGER myrank, ierr, istatus(MPI_STATUS_SIZE)
INTEGER n
PARAMETER (n=100000000)
REAL a(n), b(n)
C Initialize MPI:
    call MPI_INIT(ierr)
C Get my rank:
    call MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
C Process 0 sends, then receives; same for process 1
    if( myrank.eq.0 )then
        call MPI_SEND( a, n, MPI_REAL, 1, 17, MPI_COMM_WORLD, ierr)
        call MPI_RECV( b, n, MPI_REAL, 1, 19, MPI_COMM_WORLD, istatus, ierr )
    else if ( myrank.eq.1 )then
        call MPI_SEND( a, n, MPI_REAL, 0, 19, MPI_COMM_WORLD, ierr )
        call MPI_RECV( b, n, MPI_REAL, 0, 17, MPI_COMM_WORLD, istatus, ierr)
    endif
C Terminate MPI:
    call MPI_FINALIZE(ierr)
END

```

This program will deadlock under the default configuration of nearly all available MPI implementations.

```
PROGRAM simple_deadlock_avoided
INCLUDE 'mpif.h'
INTEGER myrank, ierr, status(MPI_STATUS_SIZE)
INTEGER request
REAL a(100), b(100)
```

C Initialize MPI:

```
call MPI_INIT(ierr)
```

C Get my rank:

```
call MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
```

C Process 0 posts a receive, then sends; same for process 1

```
if( myrank.eq.0 )then
```

```
call MPI_Irecv( b, 100, MPI_REAL, 1, 19, MPI_COMM_WORLD, request, ierr )
```

```
call MPI_SEND( a, 100, MPI_REAL, 1, 17, MPI_COMM_WORLD, ierr)
```

```
call MPI_WAIT( request, status, ierr )
```

```
else if ( myrank.eq.1 )then
```

```
call MPI_Irecv( b, 100, MPI_REAL, 0, 17, MPI_COMM_WORLD, request, ierr )
```

```
call MPI_SEND( a, 100, MPI_REAL, 0, 19, MPI_COMM_WORLD, ierr)
```

```
call MPI_WAIT( request, status, ierr )
```

```
endif
```

C Terminate MPI:

```
call MPI_FINALIZE(ierr)
```

```
END
```

It is always safe to order the calls of MPI_(I)SEND and MPI_(I)RECV so that a send subroutine call at one process and a corresponding receive subroutine call at the other process appear in matching order.

```
IF (myrank==0) THEN
  CALL MPI_SEND(sendbuf, ...)
  CALL MPI_RECV(recvbuf, ...)
ELSEIF (myrank==1) THEN
  CALL MPI_RECV(recvbuf, ...)
  CALL MPI_SEND(sendbuf, ...)
ENDIF
```

In this case, you can use either blocking or non-blocking subroutines.

Considering the previous options, performance, and the avoidance of deadlocks, it is recommended to use the following code.

```
IF (myrank==0) THEN
  CALL MPI_ISEND(sendbuf, ..., ireq1, ...)
  CALL MPI_Irecv(recvbuf, ..., ireq2, ...)
ELSEIF (myrank==1) THEN
  CALL MPI_ISEND(sendbuf, ..., ireq1, ...)
  CALL MPI_Irecv(recvbuf, ..., ireq2, ...)
ENDIF
CALL MPI_WAIT(ireq1, ...)
CALL MPI_WAIT(ireq2, ...)
```

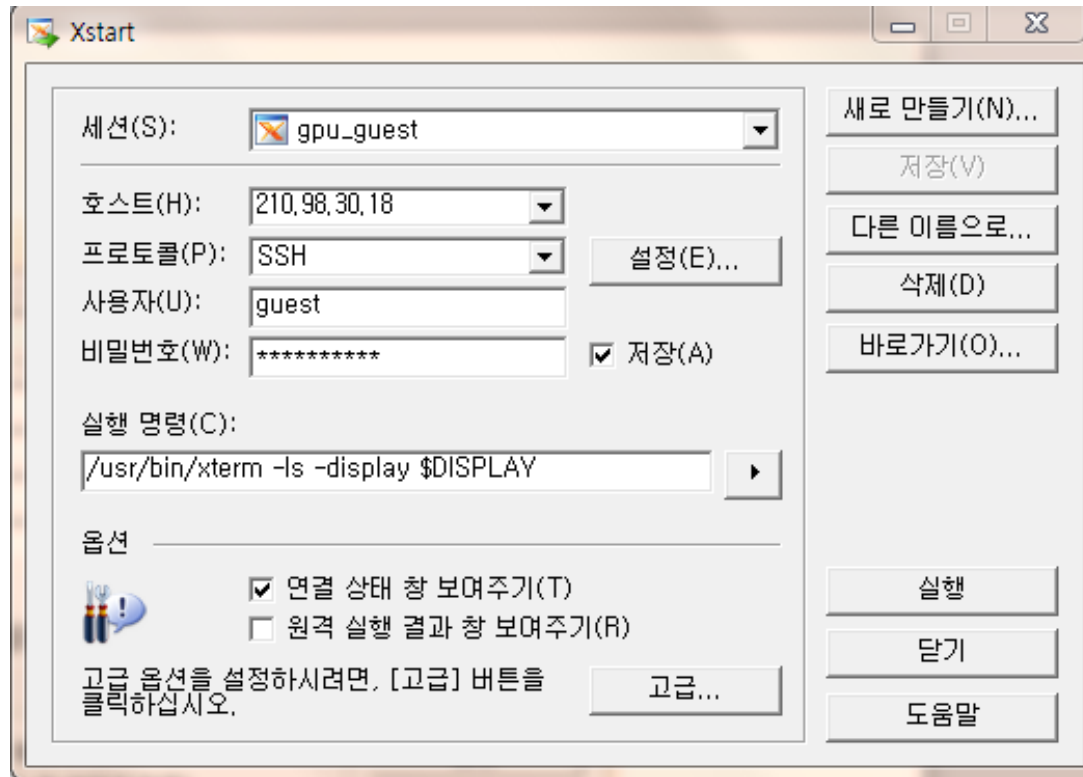
Blocking send/receive

A blocking send or receive does not return from the subroutine call until the operation has actually completed. Thus it insures that the relevant completion criteria have been satisfied before the calling process is allowed to proceed.

You are sure that the data has actually arrived and is ready for use.

210.98.30.18

210.98.30.28



xmanager 연결할 때, 아래와 같이 하면 큰 폰트로 글씨를 표시할 수 있다.

```
/usr/bin/xterm -ls -display $DISPLAY -fn 12x24
```

```
/usr/bin/xterm -ls -display $DISPLAY -fn 12x24 -fg gray -bg black
```

guest

kias!guest

:set background=light

:set background=dark

:set paste

:syntax off

:syntax on

ls

ls -ltra

예를 들어 프로그램 실행 전후,
파일들의 생성 순서가 중요할 경우

Editors

Vi

Emacs

Pico

mkdir abc

cd abc

ls

mv aa aa1

cp aa bb

set showmatch

set paste

텍스트 복사할 때 제대로 문자들 배열이 되도록 해주는 옵션

SGE: Sun Grid Engine

pestat

```
#!/bin/bash
#$ -V
#$ -cwd
#$ -S /bin/bash
#$ -N sge_run_test

### Total used process count numbers on cpu_openmpi
#$ -pe cpu_openmpi 4

### Total used process count numbers on gpu_openmpi
### $ -pe gpu_openmpi 4

### Queue Name
### To use cpu => cpu
### To use gpu => gpu
#$ -q cpu

#$ -R yes

# This job's working directory
echo Working directory is $SGE_CWD_PATH
cd $SGE_CWD_PATH

echo Running on host `hostname`
echo Time is `date`
echo Directory is `pwd`
echo This jobs runs on the following processors:
echo `cat $PE_HOSTFILE`
# Define number of processors
NPROCS=`wc -l < $PE_HOSTFILE`
echo "nprocs=$NPROCS"
NPROCS=$NSLOTS

# your job
mpirun -np $NPROCS ./a.out > output1
```

Portable Batch System (PBS)

- Process and Resource Management
- Torque, openpbs
 - <http://www.clusterresources.com/products/torque-resource-manager.php>
- Commands (실습)
 - % qstat (Status)
 - % man qstat
 - % qstat -a
 - % qstat -u newton
 - % qstat -an
 - % qstat -q
 - % qstat -s
 - % qstat -f
 - % pestat (Node Status)
 - % qsub
 - % qdel
- PBS script (실습과 함께)

PBS Scripts

```
#!/bin/sh
#### Job name
#PBS -N JOB_NAME
#### Declare job non-rerunnable
#PBS -r n
#### Output files
#PBS -j oe
#### Mail to user
#PBS -m ae
#### Queue name (n2, n4, n16, n32, n64)
#PBS -q n32
#### Walltime limit (hh:mm:ss)
#PBS -l walltime=168:00:00

# This job's working directory
echo Working directory is $PBS_O_WORKDIR
cd $PBS_O_WORKDIR

echo Running on host `hostname`
echo Time is `date`
echo Directory is `pwd`
echo This job runs on the following processors:
echo `cat $PBS_NODEFILE`
# Define number of processors
NPROCS=`wc -l < $PBS_NODEFILE`
```

```
## your parallel job

mpiexec -np $NPROCS a.out
```

```
* PBS interactive mode
% qsub -I -q n8 -N job_name
% qsub -I -q n16 -N job_name

% echo $PBS_O_WORKDIR
% echo $PBS_NODEFILE
% cat $PBS_NODEFILE
% export | grep PBS

* PBS commands
% cat start.mat
% qsub start.mat
% qdel 57480
```

From Prof. K. Joo (KIAS), The 2nd CAC Summer School on Parallel Computing

Point to Point Communication (Fortran)

Call MPI_SEND(buf, count, datatype, dest, tag, comm, ierr);

- buf : Send buffer
- count : Number of sent elements
- datatype : MPI datatype
- dest : Destination Process ID
- tag : Message tag
- comm : Communicator

Call MPI_RECV(buf, count, datatype, source, tag, comm, status, ierr);

- buf : Receive Buffer
- source : Source Process ID
- status : Receive status

Point to Point Communication (C)

MPI_Send(buf, count, datatype, dest, tag, comm);

- buf : Send buffer
- count : Number of sent elements
- datatype : MPI datatype
- dest : Destination Process ID
- tag : Message tag
- comm : Communicator

MPI_Recv(buf, count, datatype, source, tag, comm, &status);

- buf : Receive Buffer
- source : Source Process ID
- status : Receive status

From Prof. K. Joo (KIAS), The 2nd CAC Summer School on Parallel Computing

```
mpif90 hello.f90  
mpicc hello.c
```

```
mpiexec -np 2 ./a.out
```

```
mpirun -np 3 ./a.out
```

```
qsub start.mat  
pestat  
qstat -a
```

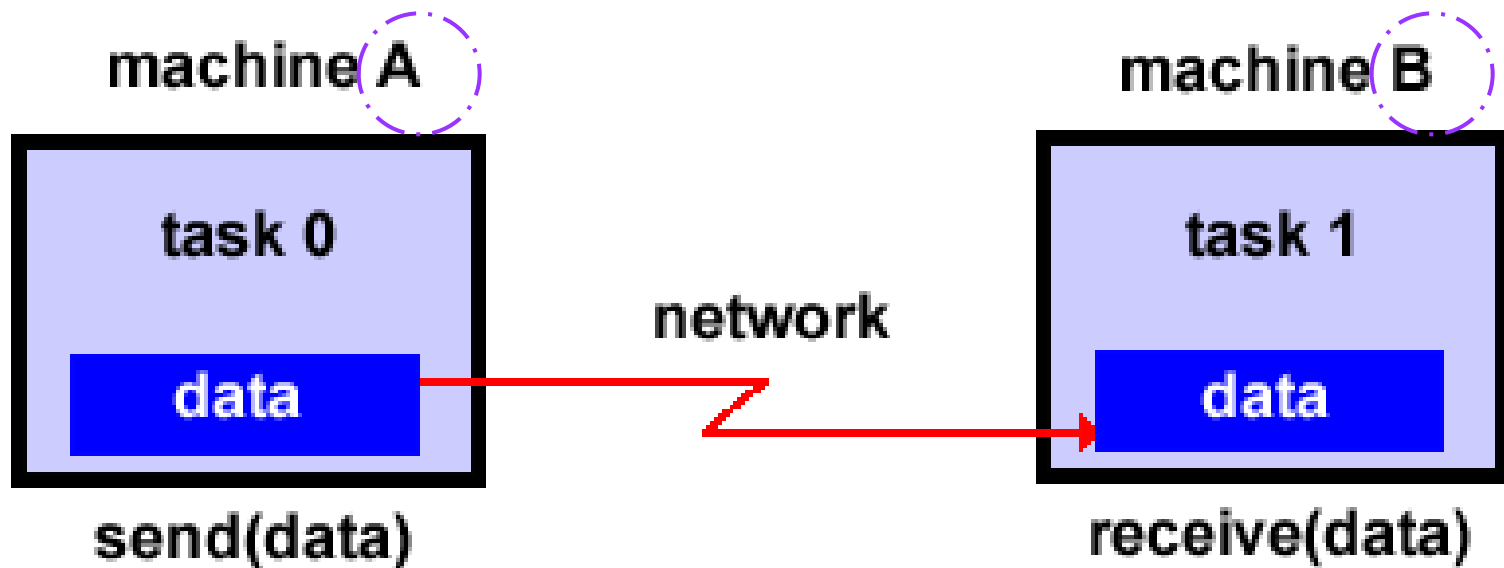
```
DOUBLE PRECISION :: start, end  
start = MPI_Wtime()
```

```
! code to be timed
```

```
end = MPI_Wtime()  
print*, 'That took ', end-start, ' seconds'
```

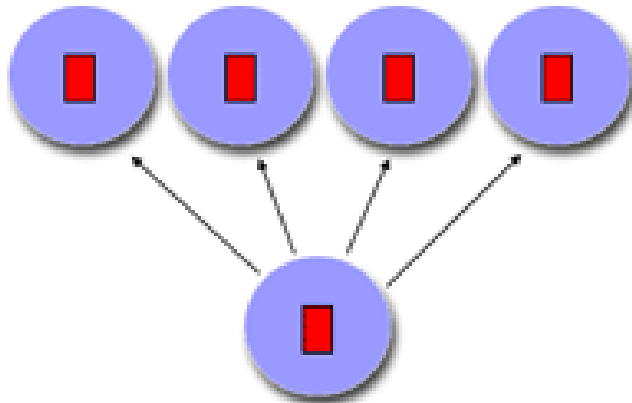
```
foreach num (`seq -s" " 41 42`)
ssh c${num} ps -ef | grep program_name | awk '{print $2}' | xargs kill -9
end
```

Point-to-point communications

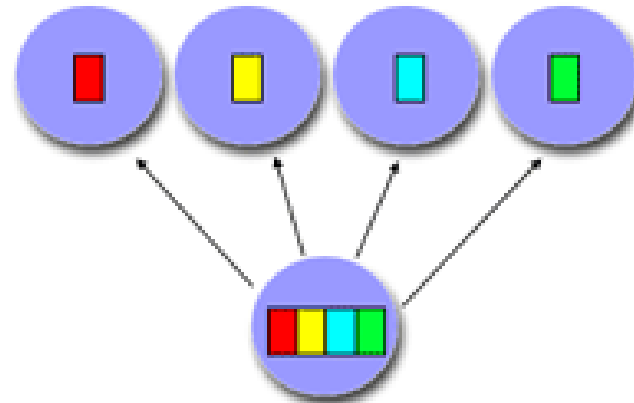


Collective communications

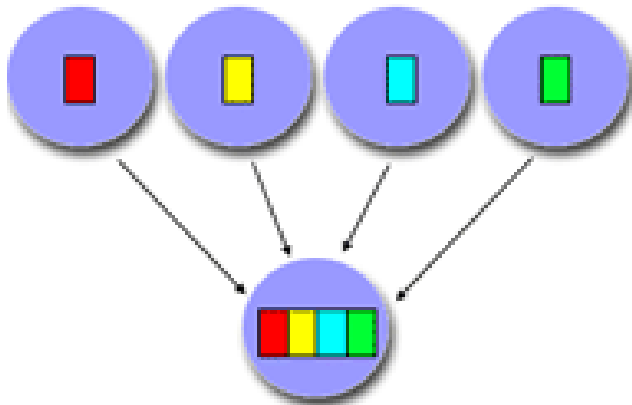
No tags



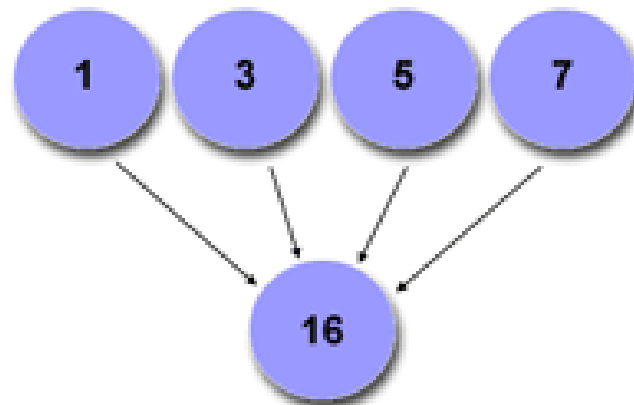
broadcast



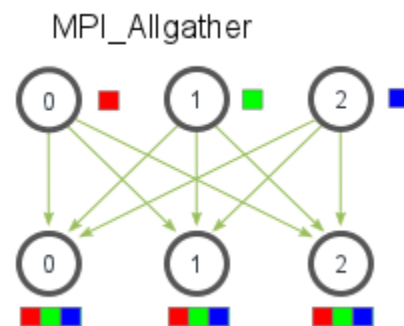
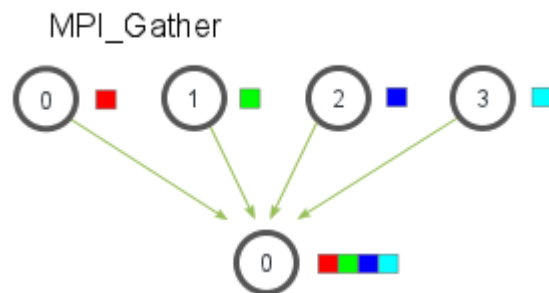
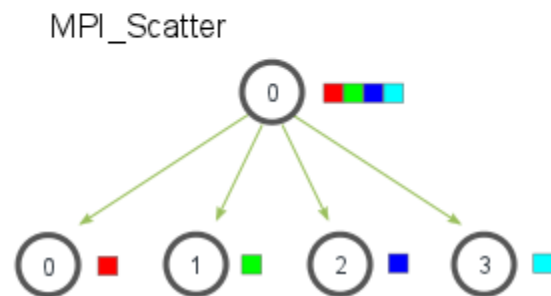
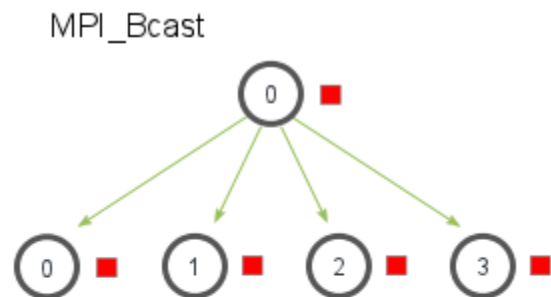
scatter



gather

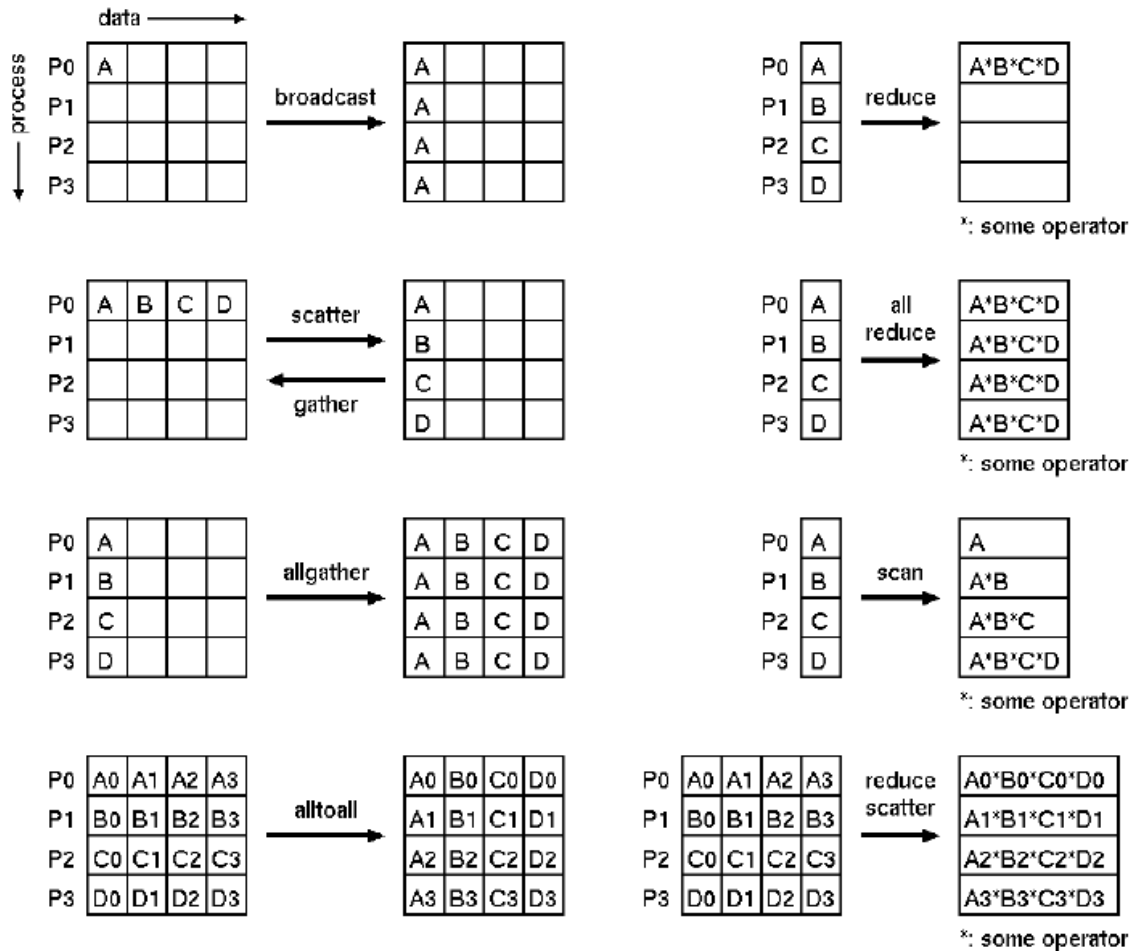


reduction



Collective communications

scatter
gather

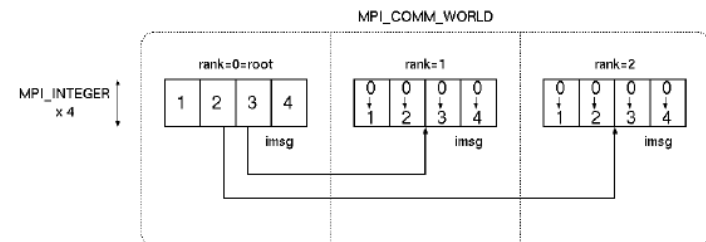


Y. Aoyama and J. Nakano

Collective communications

bcast.f

```
1      PROGRAM bcast
2      INCLUDE 'mpif.h'
3      INTEGER imsg(4)
4      CALL MPI_INIT(ierr)
5      CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
6      CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
7      IF (myrank==0) THEN
8          DO i=1,4
9              imsg(i) = i
10         ENDDO
11     ELSE
12         DO i=1,4
13             imsg(i) = 0
14         ENDDO
15     ENDIF
16     PRINT *, 'Before:', imsg
17     CALL MP_FLUSH(1)
18     CALL MPI_BCAST(imsg, 4, MPI_INTEGER,
19     &                0, MPI_COMM_WORLD, ierr)
20     PRINT *, 'After :', imsg
21     CALL MPI_FINALIZE(ierr)
22     END
```

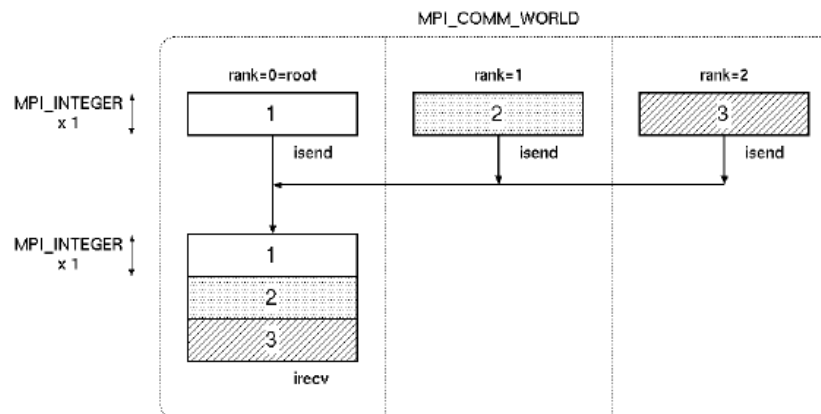


No tags

Collective communications

gather.f

```
1      PROGRAM gather
2      INCLUDE 'mpif.h'
3      INTEGER irecv(3)
4      CALL MPI_INIT(ierr)
5      CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
6      CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
7      isend = myrank + 1
8      CALL MPI_GATHER(isend, 1, MPI_INTEGER,
9      &                irecv, 1, MPI_INTEGER,
10     &                0, MPI_COMM_WORLD, ierr)
11     IF (myrank==0) THEN
12         PRINT *, 'irecv =', irecv
13     ENDIF
14     CALL MPI_FINALIZE(ierr)
15     END
```

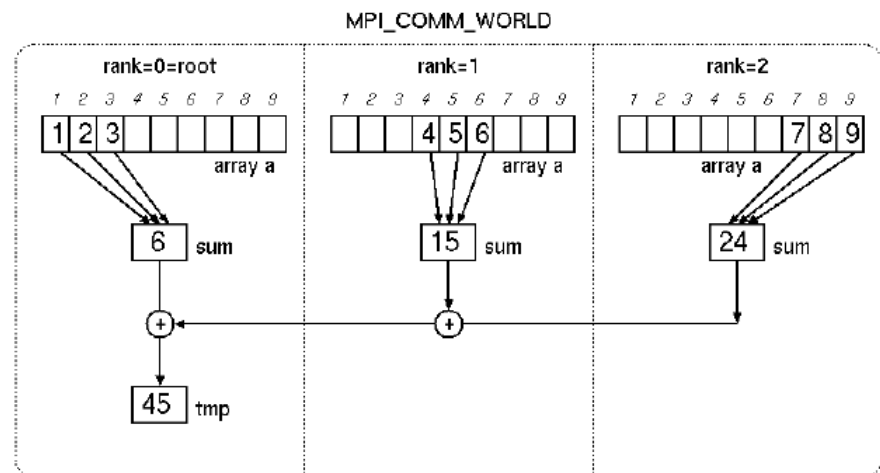


Y. Aoyama and J. Nakano

Collective communications

```
1  PROGRAM reduce
2  INCLUDE 'mpif.h'
3  REAL a(9)
4  CALL MPI_INIT(ierr)
5  CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
6  CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
7  ista = myrank * 3 + 1
8  iend = ista + 2
9  DO i=ista,iend
10     a(i) = i
11 ENDDO
12 sum = 0.0
13 DO i=ista,iend
14     sum = sum + a(i)
15 ENDDO
16 CALL MPI_REDUCE(sum, tmp, 1, MPI_REAL, MPI_SUM, 0,
17 &               MPI_COMM_WORLD, ierr)
18 sum = tmp
19 IF (myrank==0) THEN
20     PRINT *, 'sum =', sum
21 ENDIF
22 CALL MPI_FINALIZE(ierr)
23 END
```

No tags

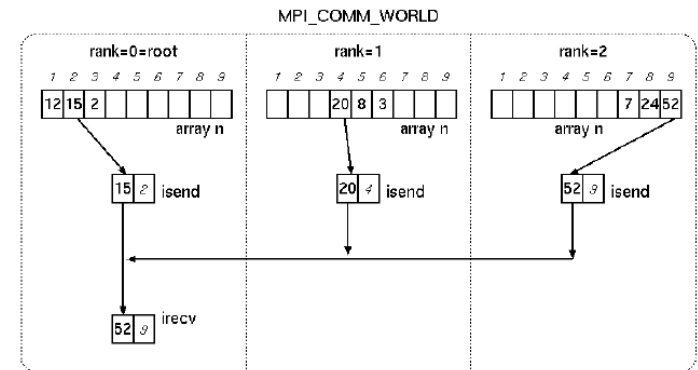


Y. Aoyama and J. Nakano

Collective communications

```
PROGRAM maxloc_p
INCLUDE 'mpif.h'
INTEGER n(9)
INTEGER isend(2), irecv(2)
DATA n /12, 15, 2, 20, 8, 3, 7, 24, 52/
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
ista = myrank * 3 + 1
iend = ista + 2
imax = -999
DO i = ista, iend
  IF (n(i) > imax) THEN
    imax = n(i)
    iloc = i
  ENDIF
ENDDO
isend(1) = imax
isend(2) = iloc

CALL MPI_REDUCE(isend, irecv, 1, MPI_2INTEGER,
&                MPI_MAXLOC, 0, MPI_COMM_WORLD, ierr)
IF (myrank == 0) THEN
  PRINT *, 'Max =', irecv(1), 'Location =', irecv(2)
ENDIF
CALL MPI_FINALIZE(ierr)
END
```



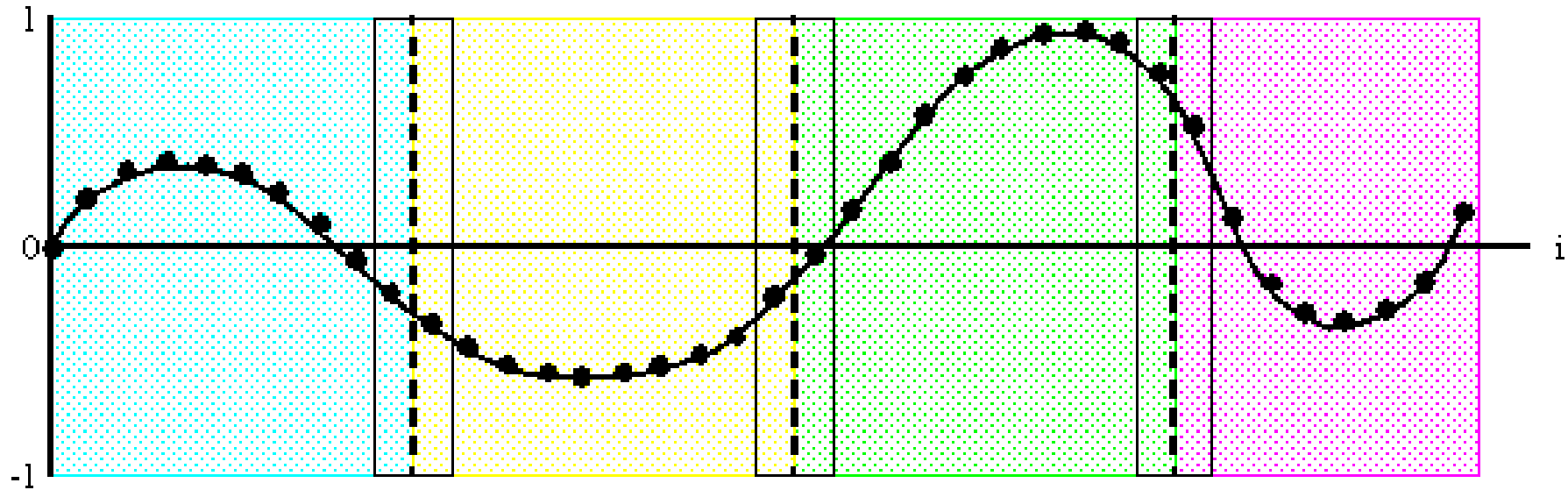
No tags

Y. Aoyama and J. Nakano

Reduce operators

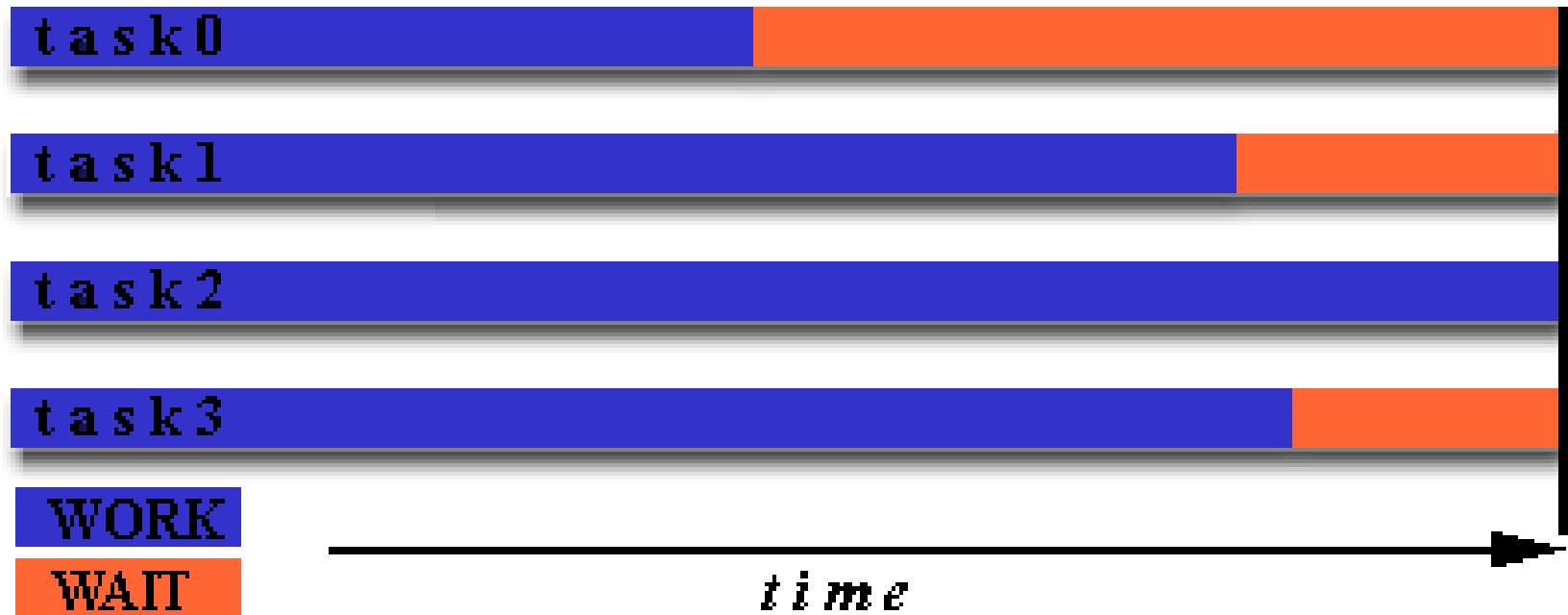
MPI Reduction Operation		C Data Types	Fortran Data Type
MPI_MAX	maximum	integer, float	integer, real, complex
MPI_MIN	minimum	integer, float	integer, real, complex
MPI_SUM	sum	integer, float	integer, real, complex
MPI_PROD	product	integer, float	integer, real, complex
MPI_LAND	logical AND	integer	logical
MPI_BAND	bit-wise AND	integer, MPI_BYTE	integer, MPI_BYTE
MPI_LOR	logical OR	integer	logical
MPI_BOR	bit-wise OR	integer, MPI_BYTE	integer, MPI_BYTE
MPI_LXOR	logical XOR	integer	logical
MPI_BXOR	bit-wise XOR	integer, MPI_BYTE	integer, MPI_BYTE
MPI_MAXLOC	max value and location	float, double and long double	real, complex, double precision
MPI_MINLOC	min value and location	float, double and long double	real, complex, double precision

Communications?



FFT
FD

Load Balancing



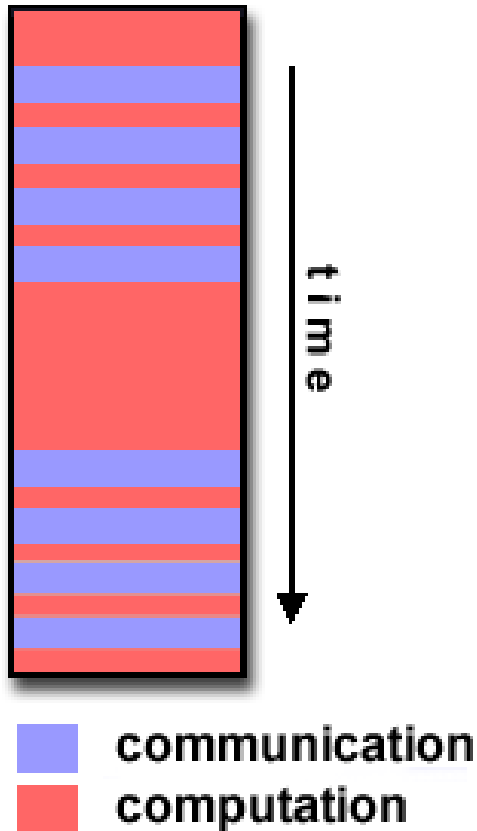
top, 각 노드에서 직접 확인, “평균하는 시간간격”을 조절함 (top→d).

Equally partition the work

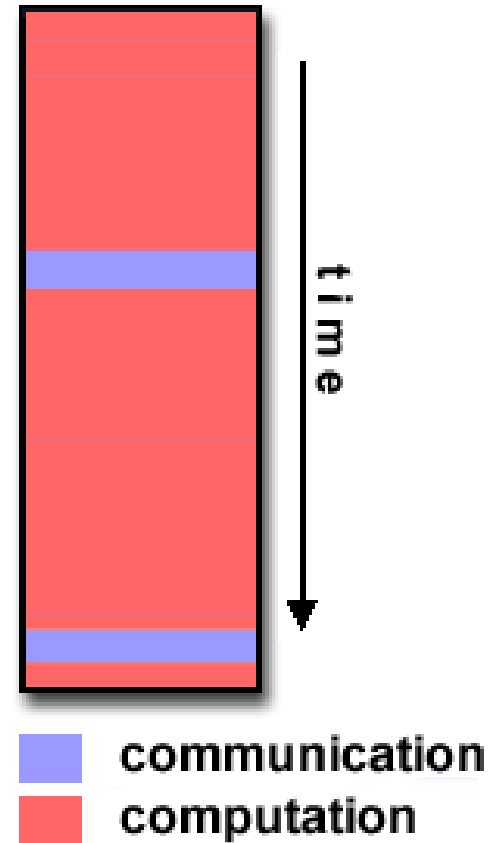
Use dynamic work assignment

Granularity

Fine-grain Parallelism



Coarse-grain Parallelism



새로운 알고리즘?

Relatively **large amounts of computational work** are done
between *communication/synchronization events*

Problem decomposition

- Domain decomposition
- Functional decomposition

Parallel programming issues

load balancing

minimizing communication

overlapping communication and computation

It is very difficult in practice to interleave communication with computation.

```
PROGRAM main
PARAMETER (n = 1000)
DIMENSION a(n)
DO i = 1, n
    a(i) = i
ENDDO
sum = 0.0
DO i = 1, n
    sum = sum + a(i)
ENDDO
PRINT *, 'sum =', sum
END
```

-

```
SUBROUTINE para_range(n1, n2, nprocs, irank, ista, iend)
  iwork1 = (n2 - n1 + 1) / nprocs
  iwork2 = MOD(n2 - n1 + 1, nprocs)
  ista = irank * iwork1 + n1 + MIN(irank, iwork2)
  iend = ista + iwork1 - 1
  IF (iwork2 > irank) iend = iend + 1
END
```

Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rank	0	0	0	0	1	1	1	1	2	2	2	3	3	3



```
SUBROUTINE para_range(n1, n2, nprocs, irank, ista, iend)
  iwork = (n2 - n1) / nprocs + 1
  ista = MIN(irank * iwork + n1, n2 + 1)
  iend = MIN(ista + iwork - 1, n2)
END
```

Iteration	1	2	3	4	5	6	7	8	8	10	11	12	13	14
Rank	0	0	0	0	1	1	1	1	2	2	2	2	3	3

Y. Aoyama and J. Nakano

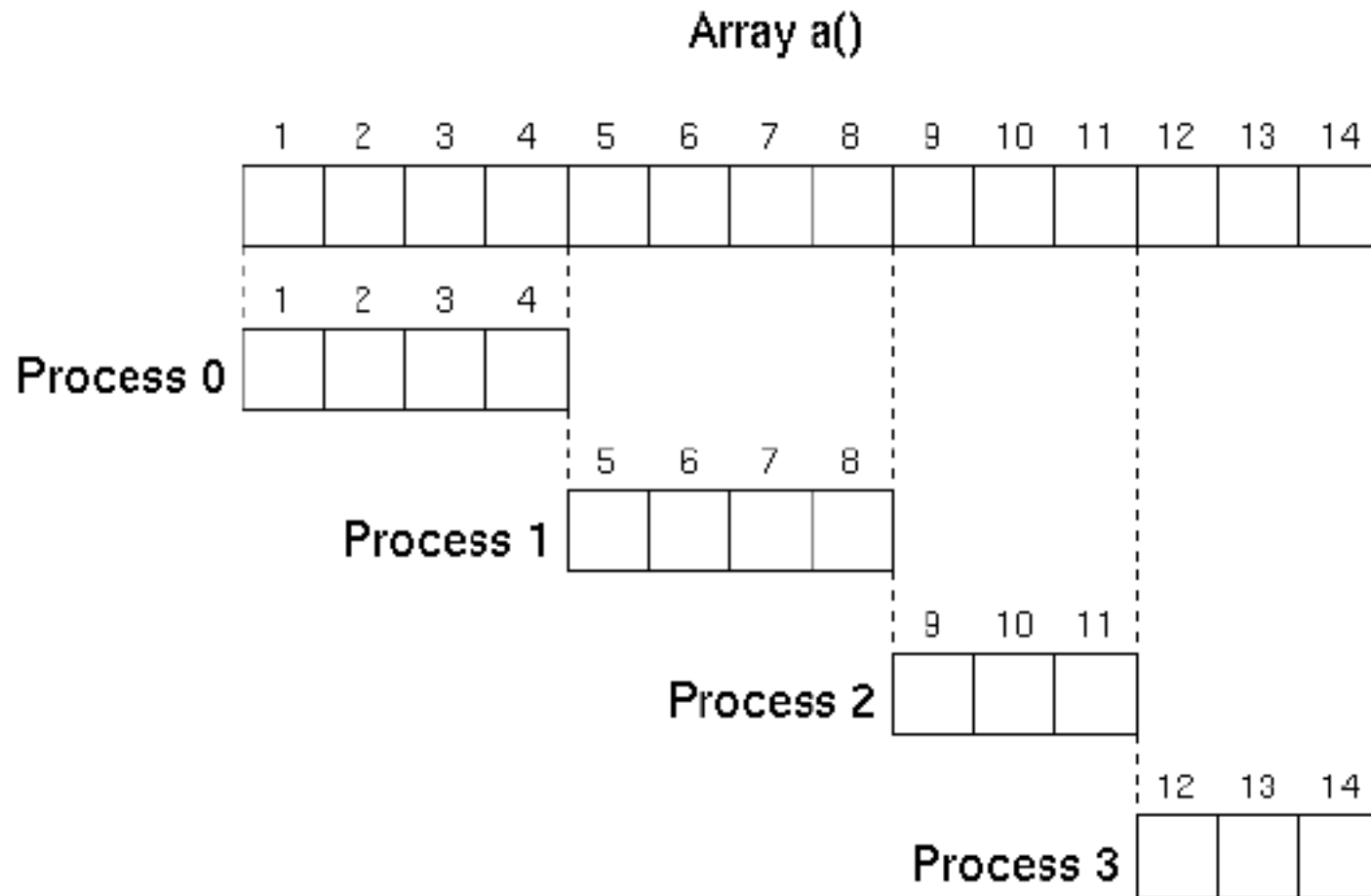
```
subroutine equal_load(n1,n2,nproc,myid,istart,ifinish)
implicit none
integer nproc,myid,istart,ifinish,n1,n2
integer iw1,iw2
iw1=(n2-n1+1)/nproc ; iw2=mod(n2-n1+1,nproc)
istart=myid*iw1+n1+min(myid,iw2)
ifinish=istart+iw1-1 ; if(iw2 > myid) ifinish=ifinish+1
! print*, n1,n2,myid,nproc,istart,ifinish
if(n2 < istart) ifinish=istart-1
return
end
```

```

PROGRAM main
INCLUDE 'mpif.h'
PARAMETER (n = 1000)
DIMENSION a(n)
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
CALL para_range(1, n, nprocs, myrank, ista, iend)
DO i = ista, iend
    a(i) = i
ENDDO
sum = 0.0
DO i = ista, iend
    sum = sum + a(i)
ENDDO
CALL MPI_REDUCE(sum, ssum, 1, MPI_REAL,
&               MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum = ssum
IF (myrank == 0) PRINT *, 'sum =', sum
CALL MPI_FINALIZE(ierr)
END

```

Y. Aoyama and J. Nakano



Y. Aoyama and J. Nakano

가장 많이 사용되는 do loop 병렬화의 방식을 아래에 표시했다.
do loop 병렬화 (3가지 방식)

1/3. block distribution

```
do i=n1,n2
```

```
.....
```

```
end do
```

-----> 아래와 같이 serial에서 parallel로 바꿉니다.

```
do i=ista,iend
```

```
.....
```

```
end do
```

여기에서 ista, iend는 노드별(irank)로 다른값이 할당된다.

```
subroutine para_range(n1,n2,nprocs,irank,ista,iend)
```

```
implicit none
```

```
integer n1,n2,nprocs,irank,ista,iend
```

```
integer iwork1,iwork2
```

```
iwork1=(n2-n1+1)/nprocs
```

```
iwork2=mod(n2-n1+1,nprocs)
```

```
ista=irank*iwork1+n1+min(irank,iwork2)
```

```
iend=ista+iwork1-1
```

```
if(iwork2> irank) iend=iend+1
```

```
end
```

Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rank	0	0	0	0	1	1	1	1	2	2	2	3	3	3

Figure 49. Block Distribution

2/3. cyclic distribution

round-robin fashion

```
do i=n1,n2
```

```
.....
```

```
end do
```

-----> 아래와 같이 serial에서 parallel로 바꿉니다.

```
do i=n1+irank, n2, nprocs
```

```
.....
```

```
end do
```

Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Rank	0	1	2	3	0	1	2	3	0	1	2	3	0	1

Figure 51. Cyclic Distribution

3/3. block-cyclic distribution

```
do i=n1,n2
```

```
.....
```

```
end do
```

-----> 아래와 같이 serial에서 parallel로 바꿉니다.

```
do ii=n1+irank*iblock, n2, nprocs*iblock
```

```
do i=ii,min(ii+iblock-1,n2)
```

```
.....
```

```
end do
```

```
end do
```

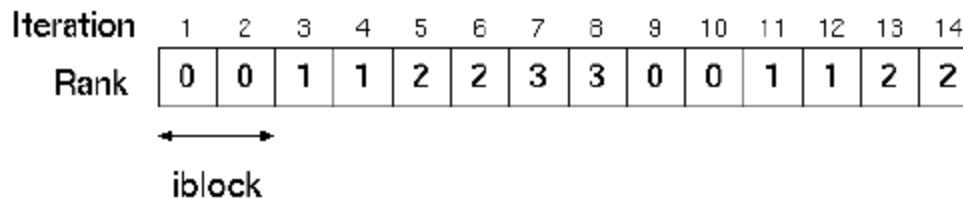


Figure 52. Block-Cyclic Distribution

```

PROGRAM main
implicit none
real sum1,ssum
integer i,ista,iend
integer ierr,n1,n2,nprocs,myrank
INCLUDE 'mpif.h'
PARAMETER (n1 = 1, n2 = 1000)
REAL, ALLOCATABLE :: a(:)

CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
CALL para_range(n1, n2, nprocs, myrank, ista, iend)
ALLOCATE (a(ista:iend))
DO i = ista, iend
a(i) = i
ENDDO
!  sum1 = 0.0
!  DO i = ista, iend
!  sum1 = sum1 + a(i)
!  ENDDO
sum1=sum(a)
DEALLOCATE (a)
CALL MPI_REDUCE(sum1, ssum, 1, MPI_REAL,MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum1 = ssum
PRINT *, 'sum1 =', sum1, myrank
CALL MPI_FINALIZE(ierr)
END

```

```

PROGRAM main
INCLUDE 'mpif.h'
PARAMETER (n1 = 1, n2 = 1000) real*8 sum, ssum
REAL, ALLOCATABLE :: a(:) real*8, allocatable ::
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
CALL para_range(n1, n2, nprocs, myrank, ista, iend)
ALLOCATE (a(ista:iend))
DO i = ista, iend
    a(i) = i
ENDDO
sum = 0.0
DO i = ista, iend
    sum = sum + a(i)

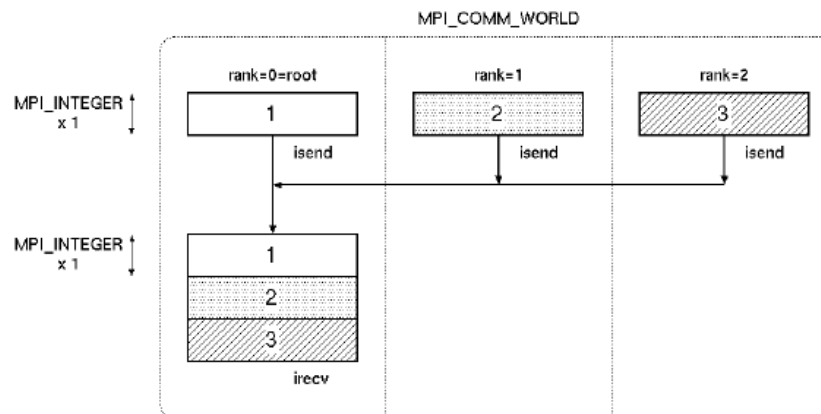
ENDDO
DEALLOCATE (a) MPI_REAL8
CALL MPI_REDUCE(sum, ssum, 1, MPI_REAL,
& MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum = ssum
PRINT *, 'sum =', sum
CALL MPI_FINALIZE(ierr)
END

```

Collective communications

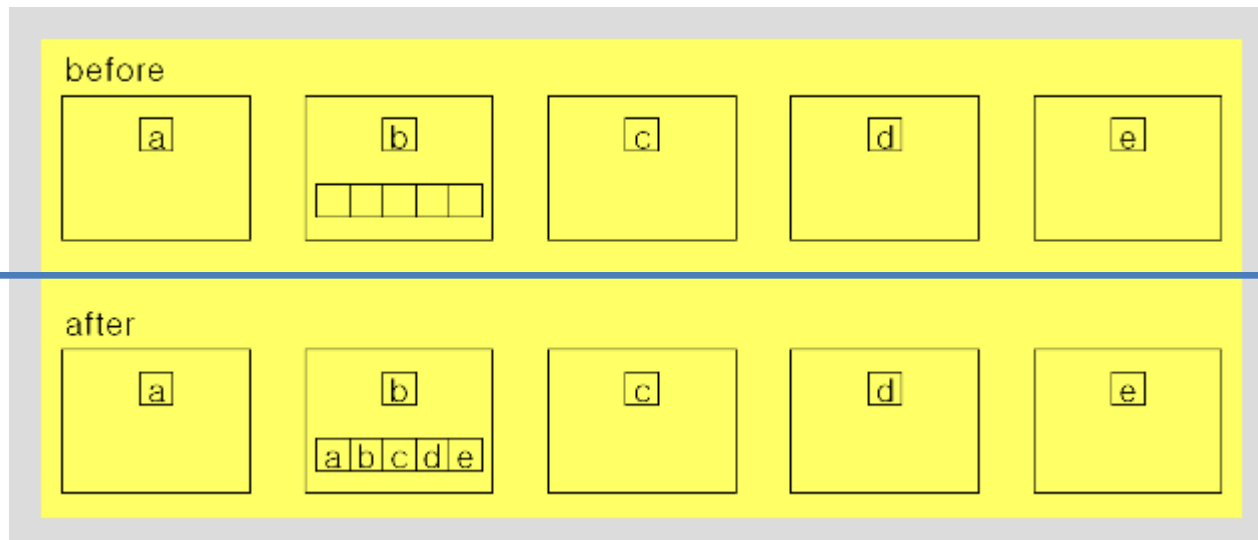
gather.f

```
1      PROGRAM gather
2      INCLUDE 'mpif.h'
3      INTEGER irecv(3)
4      CALL MPI_INIT(ierr)
5      CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
6      CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
7      isend = myrank + 1
8      CALL MPI_GATHER(isend, 1, MPI_INTEGER,
9      &                irecv, 1, MPI_INTEGER,
10     &                0, MPI_COMM_WORLD, ierr)
11     IF (myrank==0) THEN
12         PRINT *, 'irecv =', irecv
13     ENDIF
14     CALL MPI_FINALIZE(ierr)
15     END
```



Y. Aoyama and J. Nakano

gatherv



각 노드에 할당된 특정 크기(M)의 자료들을 특정 노드(0 노드)에 모두 모으고자 할 때가 있을 수 있다. 이 때, 특정 노드(0 노드)에서는 노드수(nproc)에 비례하는 배열 할당($M * nproc$)이 필요하게 된다.

<http://www.cac.cornell.edu/Ranger/MPIcc/scatterv.aspx>

gaps allowed between messages in source data

irregular message sizes allowed

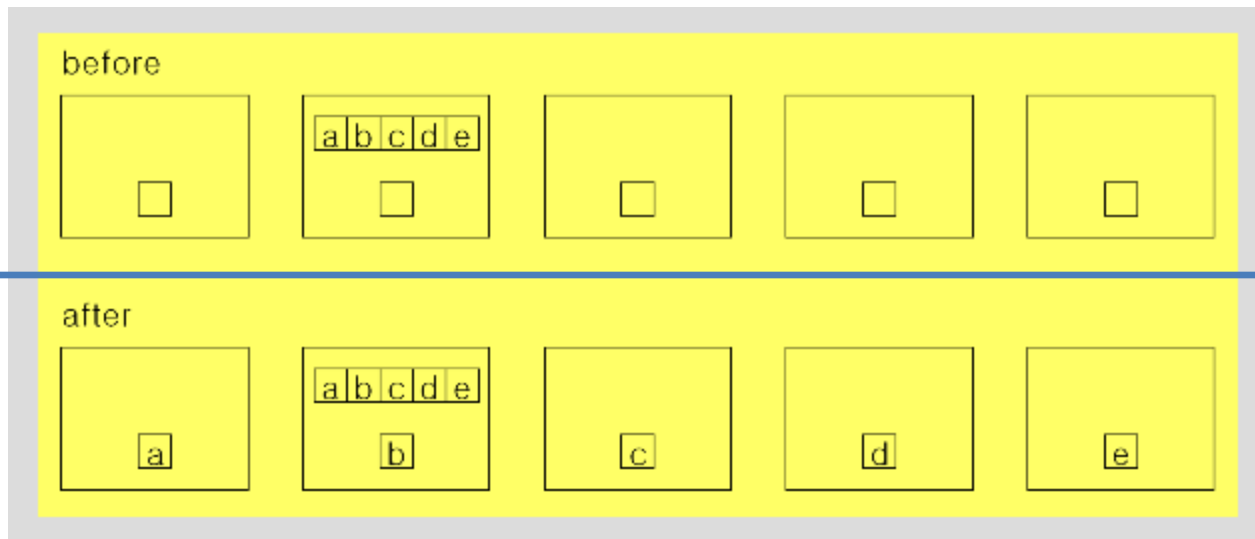
data can be distributed to processes in any order

```

!234567890
implicit none
include 'mpif.h'
integer, allocatable :: isend(:), irecv(:)
integer, allocatable :: ircnt(:), idisp(:)
integer ierr,nproc,myid
integer i,iscnt, ndsize
call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD,nproc,ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD,myid,ierr)
ndsize=10
allocate(isend(ndsize),irecv(nproc*ndsize))
allocate(ircnt(0:nproc-1),idisp(0:nproc-1))
ircnt=ndsize
idisp(0)=0
do i=1,nproc-1
  idisp(i)=idisp(i-1)+ircnt(i)
enddo
do i=1,ndsize ! node specific data with a data-size ndsize
  isend(i)=myid+1
enddo
iscnt=ndsize
call MPI_GATHERV(isend, iscnt, MPI_INTEGER, irecv, ircnt, idisp, MPI_INTEGER,0, MPI_COMM_WORLD, ierr)
if(myid == 0)then
  print*, 'irecv= ',irecv
endif
deallocate(ircnt,idisp) ; deallocate(isend,irecv)
call MPI_FINALIZE(ierr) ; stop ; end

```

scatterv



<http://www.cac.cornell.edu/Ranger/MPIcc/gathervscatterv.aspx>

Memory

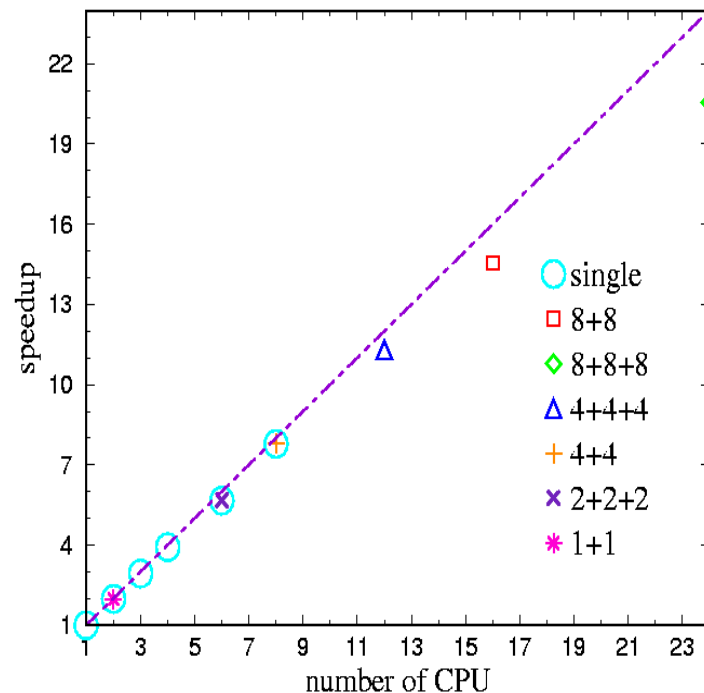
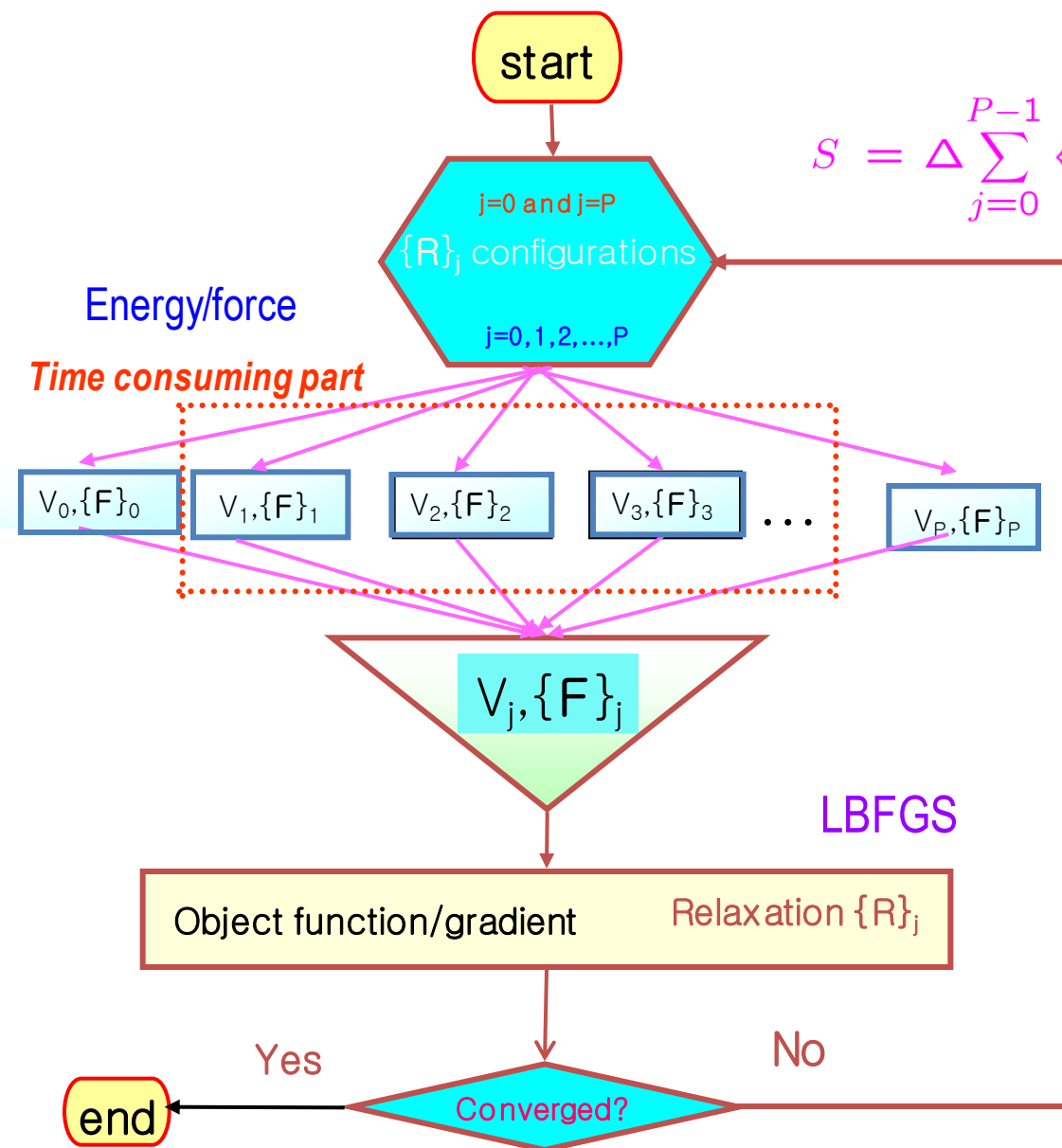
Processors

CPU

top

Parallel ADMD

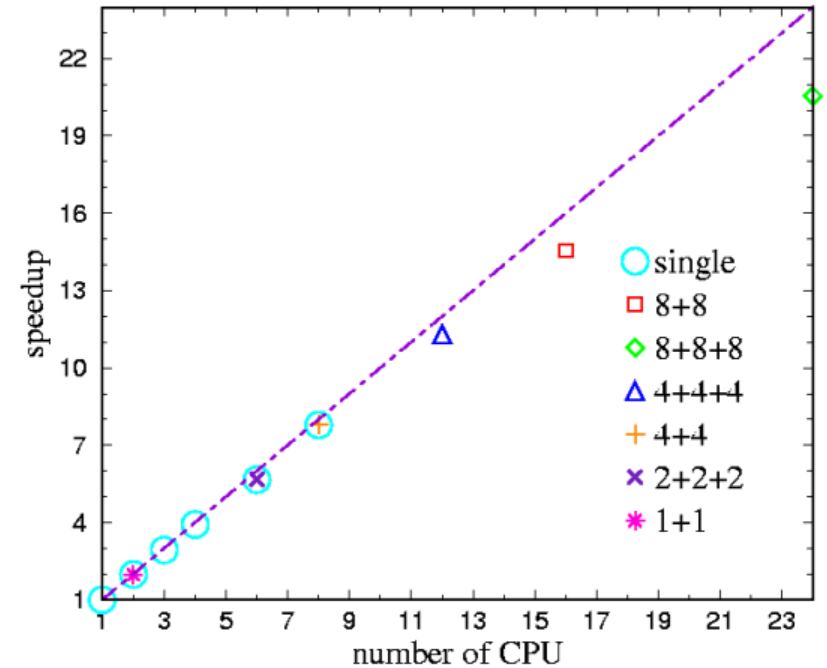
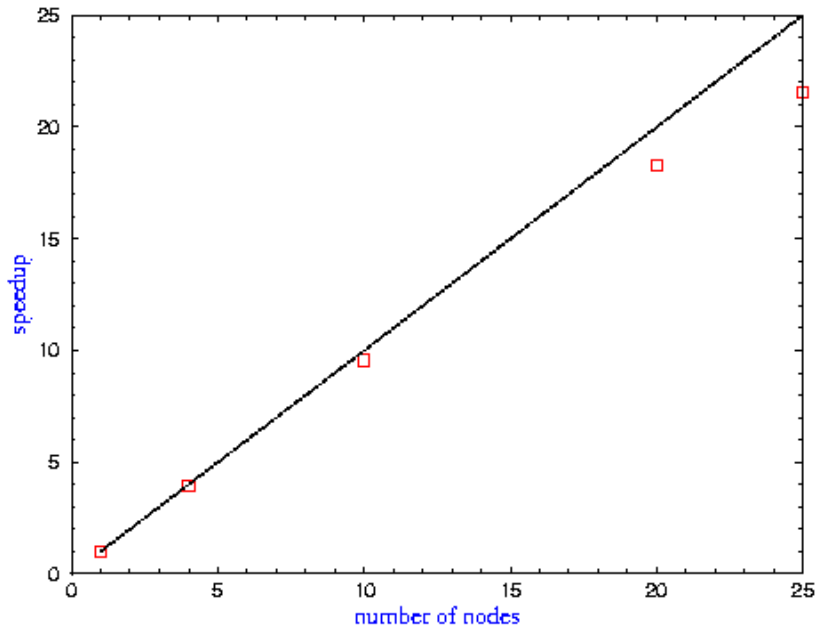
$$S = \Delta \sum_{j=0}^{P-1} \left\{ \sum_{I=1}^N \frac{M_I}{2} \left(\frac{\vec{R}_I^j - \vec{R}_I^{j+1}}{\Delta} \right)^2 - V(\{\vec{R}_I^j\}) \right\}$$



{Communication}/{CPU} ratio

Distributed computing

Observed speedup



top 각 노드에서 직접 확인

GLOBUS; MPICH_G2

```
program main
include 'mpif.h'
integer ierr
call MPI_INIT(ierr)
print*, 'hello world'
call MPI_FINALIZE(ierr)
end
```

```
mpif77 hello_world_parallel_1.f
mpiexec -n 2 ./a.out
mpiexec -np 2 ./a.out
```

ASCII vs. binary

> 10 speedup

Format	C	FORTRAN
ASCII	<code>fprintf()</code>	<code>open(6,file='test',form='formatted')</code> <code>write(6,*)</code>
Binary	<code>fwrite()</code>	<code>open(6,file='test',form='unformatted')</code> <code>write(6)</code>


```

!234567890
program d_access
implicit none
include 'mpif.h'
integer myid, nproc, ierr, iroot, kount
real*8 time_start,time_end
integer nsites
real*8, allocatable :: spin_lattice(:),tspin_lattice(:)
real*8 before,after
integer kdum,kk
call MPI_INIT( ierr )
call MPI_COMM_RANK( MPI_COMM_WORLD, myid, ierr )
call MPI_COMM_SIZE( MPI_COMM_WORLD, nproc, ierr )
if(myid == 0 .and. nproc > 1) print *, nproc," processes are alive"
if(myid == 0 .and. nproc ==1) print *, nproc," process is alive"
time_start=MPI_WTIME()

```

```

nsites=100
allocate(spin_lattice(nsites))
allocate(tspin_lattice(nsites))
spin_lattice=0.d0
tspin_lattice=0.d0
before=float(myid)
print*, before, myid,' node,in the memory'
open(97,file='fort.97',access='direct',recl=8*(nsites+1))
write(97,rec=myid+1) before,(spin_lattice(kdum),kdum=1,nsites)
close(97)

```

포트란 프로그램 실행 중(on the fly) 특정 파일 지우기는 아래와 같이 실행하면 된다.

OPEN(11,FILE='del')

CLOSE(11,STATUS='DELETE')! 파일 지우기를 실행함.

MPI 인 경우 0 번 노드에서만 지워야한다. 모든 노드가 다 지울 수 없다.

```

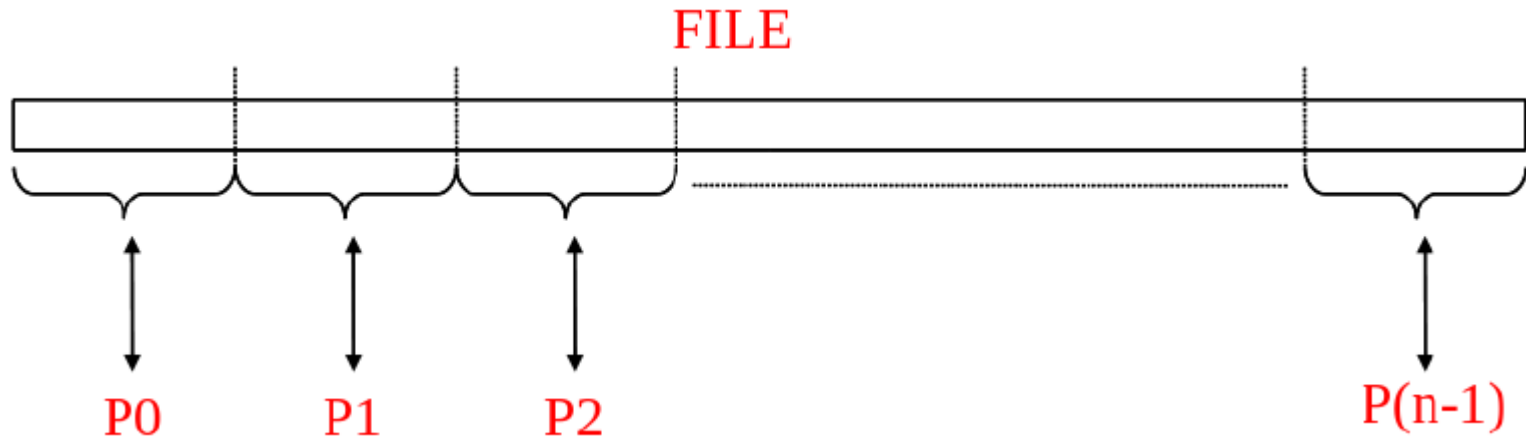
call MPI_BARRIER( MPI_COMM_WORLD, ierr )
if(myid == 0)then ! -----[ process id = 0
open(97,file='fort.97',access='direct',recl=8*(nsites+1))
before=2.d222
do kk=1,nproc
read(97,rec=kk) after,(tspin_lattice(kdum),kdum=1,nsites)
print*, after, kk-1,' node,in the file'
! if(before > after)then
! before=after
! spin_lattice=tspin_lattice
! endif
enddo
close(97)
! print*, before,' before'
endif ! -----=== } process id =0

deallocate(spin_lattice)
deallocate(tspin_lattice)
time_end=MPI_WTIME()
if(myid == 0) then ! -----=== { process id =0
write(6,'(4(f14.5,1x,a))' ) (time_end-time_start),'s', (time_end-time_start)/60.d0,'m', (time_end-time_start)/3600.d0,'h',
(time_end-time_start)/3600.d0/24.d0,'d'
endif ! -----=== } process id =0
call MPI_FINALIZE(ierr)
stop
end program d_access

```

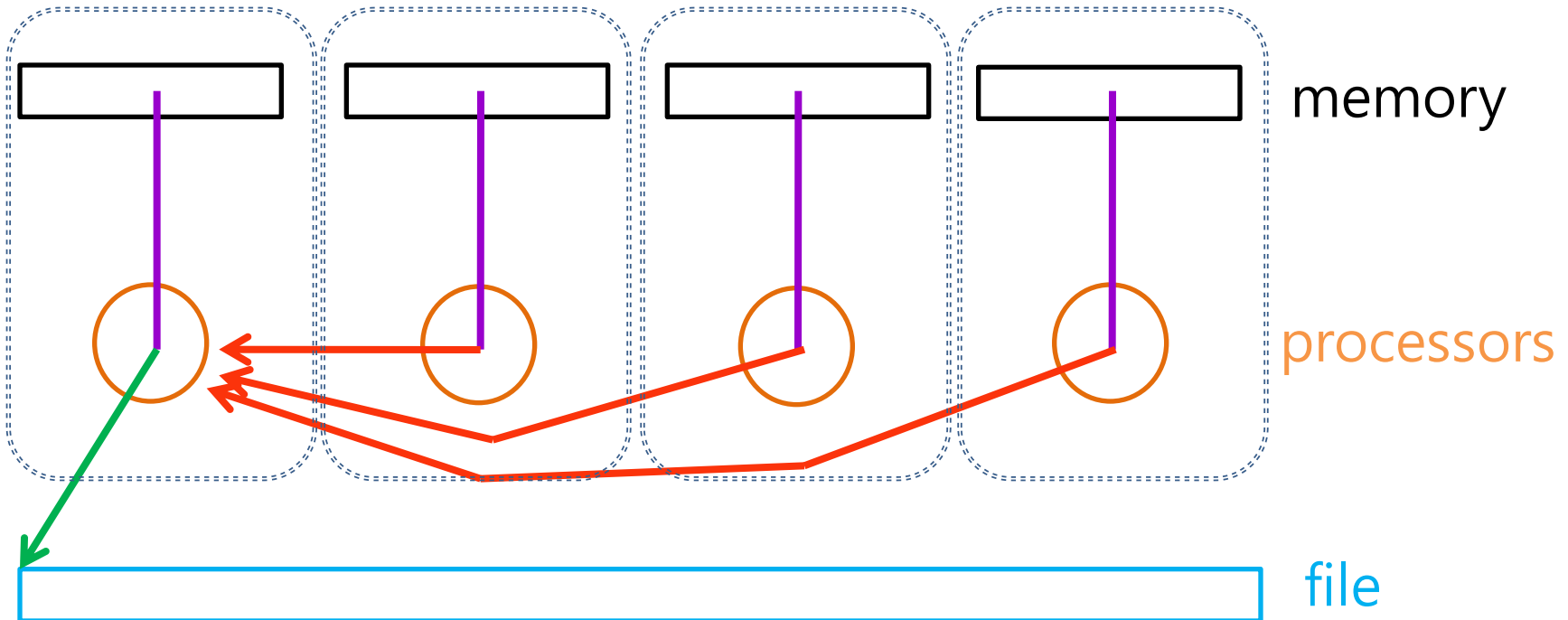
- What is Parallel I/O?

Multiple processes of a parallel program accessing data (reading or writing) from a common file.



Parallel I/O (1/3)

Sequential I/O from an parallel program

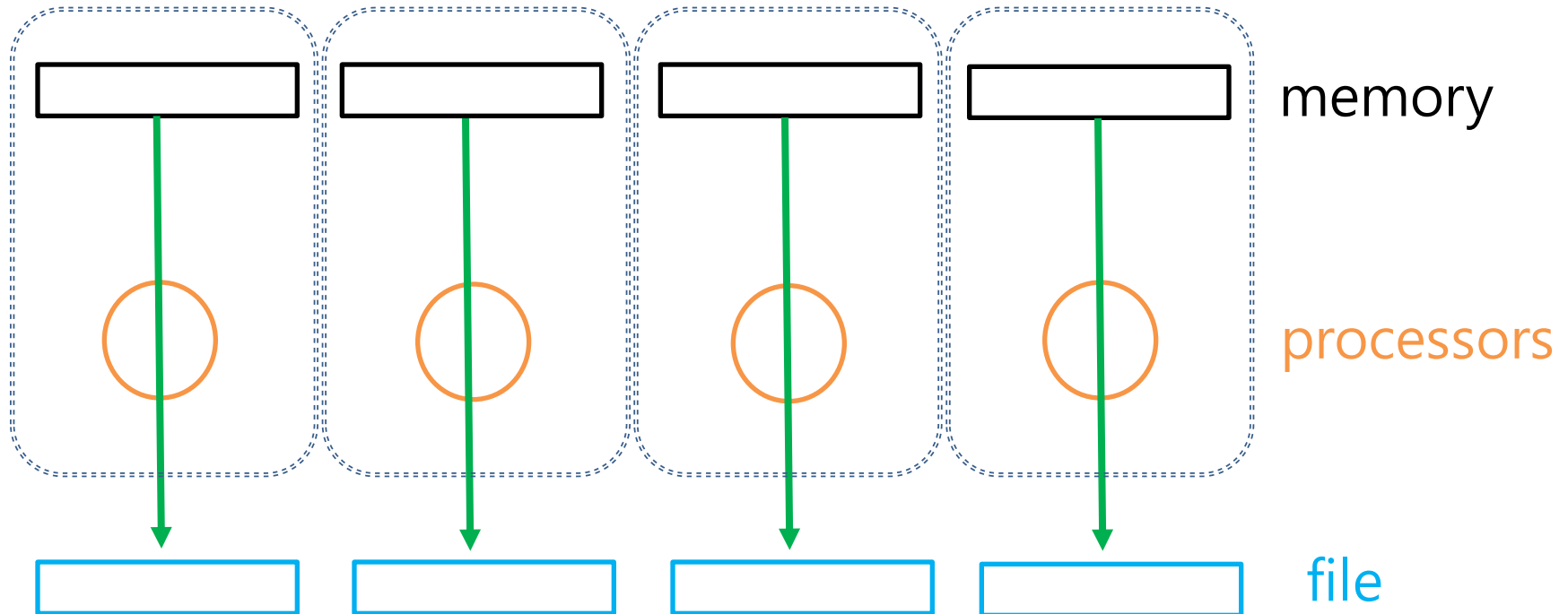


Send : write

Recv : read

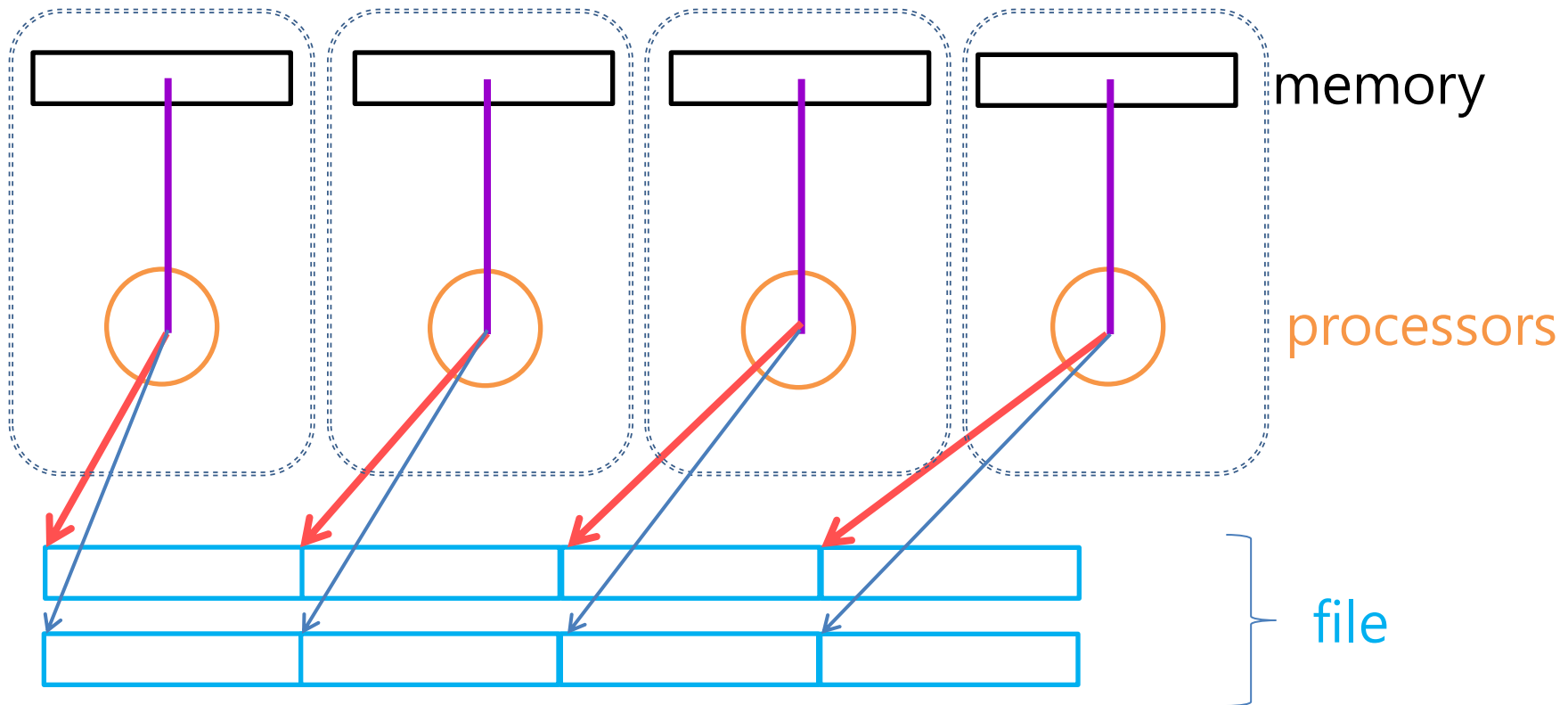
Parallel I/O (2/3)

Parallel I/O to multiple files



Parallel I/O (3/3)

Parallel I/O to a single file



```

! ! example of parallel MPI write into multiple files
PROGRAM main
  ! Fortran 90 users can (and should) use
  !   use mpi
  ! instead of include 'mpif.h' if their MPI implementation provides a
  ! mpi module.
  include 'mpif.h'

  integer ierr, i, myrank, BUFSIZE, thefile
  parameter (BUFSIZE=100)
  integer buf(BUFSIZE)
  character*12 ofname

  call MPI_INIT(ierr)
  call MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)

  do i = 1, BUFSIZE
    buf(i) = myrank * BUFSIZE + i
  enddo

  write(ofname,'(a8,i4.4)') 'testfile',myrank

  open(unit=11,file=ofname,form='unformatted')
  write(11) buf

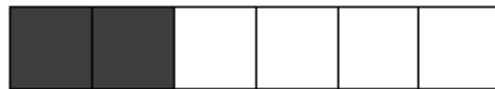
  call MPI_FINALIZE(ierr)

END PROGRAM main

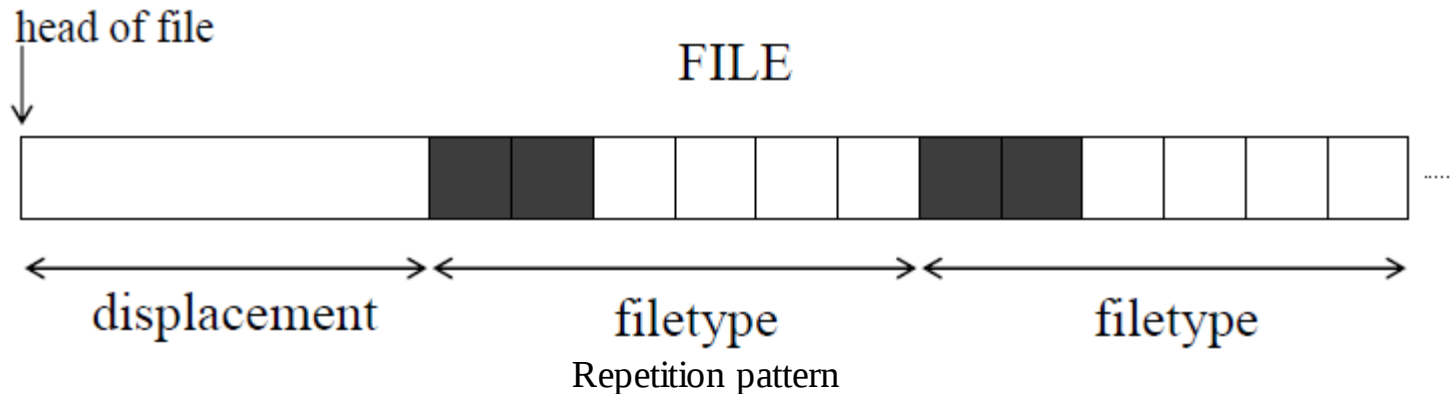
```

Specified by a triplet (*displacement*, *etype*, and *filetype*) passed to **MPI_File_set_view**

 *etype* = MPI_INT



filetype = two MPI_INTs followed by
a gap of four MPI_INTs



```
! example of parallel MPI write into a single file, in Fortran
PROGRAM main
! Fortran 90 users can (and should) use
!   use mpi
! instead of include 'mpif.h' if their MPI implementation provides a
! mpi module.
include 'mpif.h'
```

```
integer ierr, i, myrank, BUFSIZE, thefile
parameter (BUFSIZE=100)
integer buf(BUFSIZE)
integer(kind=MPI_OFFSET_KIND) disp
```

```
call MPI_INIT(ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
```

```
do i = 0, BUFSIZE
  buf(i) = myrank * BUFSIZE + i
enddo
call MPI_FILE_OPEN(MPI_COMM_WORLD, 'testfile', &
  MPI_MODE_WRONLY + MPI_MODE_CREATE, &
  MPI_INFO_NULL, thefile, ierr)
```

```
! assume 4-byte integers
disp = myrank * BUFSIZE * 4
call MPI_FILE_SET_VIEW(thefile, disp, MPI_INTEGER, &
  MPI_INTEGER, 'native', &
  MPI_INFO_NULL, ierr)
call MPI_FILE_WRITE(thefile, buf, BUFSIZE, MPI_INTEGER, &
  MPI_STATUS_IGNORE, ierr)
call MPI_FILE_CLOSE(thefile, ierr)
call MPI_FINALIZE(ierr)
```

```
END PROGRAM main
```

MPI_MODE_RDWR
MPI_MODE_RDONLY
MPI_MODE_WRONLY
MPI_MODE_CREATE

| : C
+ : Fortran

MPI_MODE_CREATE | MPI_MODE_WRONLY
MPI_MODE_WRONLY + MPI_MODE_CREATE

http://www.mcs.anl.gov/research/projects/mpi/usingmpi2/examples/starting/io3f_f90.htm

```

program write_individual_pointer
use mpi
implicit none
integer ier,nproc,myid,ifile,intsize
integer mode,info,ietype,ifiletype
integer istatus(MPI_STATUS_SIZE)
integer (kind=MPI_OFFSET_KIND) :: idisp
integer, parameter :: kount=100
integer ibuf(kount)
integer i
real*8 aa1,aa2

call MPI_INIT(ier)
call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ier)
call MPI_COMM_SIZE(MPI_COMM_WORLD, nproc, ier)

mode=ior(MPI_MODE_CREATE,MPI_MODE_WRONLY)
info=0
call MPI_TYPE_EXTENT(MPI_INTEGER,intsize,ier)

aa1=MPI_WTIME()
do i=1,kount
  ibuf(i)=myid*kount+i
enddo

CALL MPI_FILE_OPEN(MPI_COMM_WORLD, 'test',mode, info, ifile, ier)
idisp=myid* kount* intsize
ietype=MPI_INTEGER
ifiletype=MPI_INTEGER
CALL MPI_FILE_SET_VIEW(ifile,idisp,ietype,ifiletype,'native',info, ier)
CALL MPI_FILE_WRITE(ifile,myid,1,MPI_INTEGER,istatus, ier)

write(6,*) 'hello form myid',myid,'i wrote:', myid,.'

CALL MPI_FILE_CLOSE(ifile,ier)

aa2=MPI_WTIME()

call MPI_FINALIZE(ier)
end program write_individual_pointer

```

```

program write_individual_pointer
use mpi
implicit none
integer ier,nproc,myid,ifile,intsize
integer mode,info,ietype,ifiletype
integer istatus(MPI_STATUS_SIZE)
integer (kind=MPI_OFFSET_KIND) :: idisp
integer, parameter :: kount=100
integer ibuf(kount)
integer i,itest
real*8 aa1,aa2

call MPI_INIT(ier)
call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ier)
call MPI_COMM_SIZE(MPI_COMM_WORLD, nproc, ier)

mode=MPI_MODE_RDONLY
info=0

call MPI_TYPE_EXTENT(MPI_INTEGER,intsize,ier)

aa1=MPI_WTIME()
do i=1,kount
  ibuf(i)=myid*kount+i
enddo

CALL MPI_FILE_OPEN(MPI_COMM_WORLD, 'test',mode, info, ifile, ier)
idisp=myid* kount* intsize
ietype=MPI_INTEGER
ifiletype=MPI_INTEGER
CALL MPI_FILE_SET_VIEW(ifile,idisp,ietype,ifiletype,'native',info, ier)
CALL MPI_FILE_READ(ifile,itest,1,MPI_INTEGER,istatus, ier)

write(6,*) 'hello form myid',myid,'i read:', itest,','

CALL MPI_FILE_CLOSE(ifile,ier)

aa2=MPI_WTIME()

call MPI_FINALIZE(ier)
end program write_individual_pointer

```

```
integer isize  
character*9 string,fname
```

```
isize=4  
call xnumeral( myid, string, isize)  
fname='conf1'//trim(string) ! confi10000, confi10001, confi10002  
open(29,file=fname,form='formatted')  
write(29,*) nparticle,mm  
write(29,*) tau,beta,ss_lowest  
do ia=1,nparticle  
do j=0,mm  
write(29,*) (qqq(jcomp,j,ia),jcomp=1,idims)  
enddo  
enddo  
do ia=1,nparticle  
write(29,*) iiq(ia)  
enddo  
close(29)
```

```
PROGRAM env
INCLUDE 'mpif.h'
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
PRINT *, 'nprocs =', nprocs, 'myrank =', myrank
CALL MPI_FINALIZE(ierr)
END
```

```
$ mpxlf env.f
** env === End of Compilation 1 ===
1501-510 Compilation successful for file env.f.
$ export MP_STDOUTMODE=ordered
$ export MP_LABELIO=yes
$ a.out -procs 3
0: nprocs = 3 myrank = 0
1: nprocs = 3 myrank = 1
2: nprocs = 3 myrank = 2
```

From Dr. Jongsoo Kim (KASI), The 1st CAC Summer School on Parallel Computing

```

PROGRAM bcast
INCLUDE 'mpif.h'
INTEGER imsg(4)
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
IF (myrank==0) THEN
DO i=1,4
imsg(i) = i
ENDDO
ELSE
DO i=1,4
imsg(i) = 0
ENDDO
ENDIF
PRINT *, 'Before:', imsg
CALL MP_FLUSH(1)
CALL MPI_BCAST(imsg, 4, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr)
PRINT *, 'After :', imsg
CALL MPI_FINALIZE(ierr)
END

```

```

$ a.out -procs 3
0: Before: 1 2 3 4
1: Before: 0 0 0 0
2: Before: 0 0 0 0
0: After : 1 2 3 4
1: After : 1 2 3 4
2: After : 1 2 3 4

```

```
PROGRAM gather
INCLUDE 'mpif.h'
INTEGER irecv(3)
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
isend = myrank + 1
CALL MPI_GATHER(isend, 1, MPI_INTEGER, irecv, 1, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr)
IF (myrank==0) THEN
PRINT *, 'irecv =', irecv
ENDIF
CALL MPI_FINALIZE(ierr)
END
```

```
$ a.out -procs 3
0: irecv = 1 2 3
```

```
mpirun -np 3 ./a.out
```

```

/*gather*/

#include <mpi.h>

#include <stdio.h>

void main (int argc, char *argv[]){

    int i, nprocs, myrank ;

    int isend, irecv[3];

    MPI_Init(&argc, &argv);

    MPI_Comm_size(MPI_COMM_WORLD, &nprocs);

    MPI_Comm_rank(MPI_COMM_WORLD, &myrank);

    isend = myrank + 1;

    MPI_Gather(&isend,1,MPI_INT,irecv,1,MPI_INT,0, MPI_COMM_WORLD);

    if(myrank == 0) {

        printf(" irecv = ");

        for(i=0; i<3; i++)

            printf(" %d", irecv[i]); printf("\n");

    }

    MPI_Finalize();

}

```

```

PROGRAM reduce
INCLUDE 'mpif.h'
REAL a(9)
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
ista = myrank * 3 + 1
iend = ista + 2
DO i=ista,iend
  a(i) = i
ENDDO
sum = 0.0
DO i=ista,iend
  sum = sum + a(i)
ENDDO
CALL MPI_REDUCE(sum, tmp, 1, MPI_REAL, MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum = tmp
IF (myrank==0) THEN
  PRINT *, 'sum = ', sum
ENDIF
CALL MPI_FINALIZE(ierr)
END

```

\$ a.out -procs 3

0: sum = 45.00000000

```

PROGRAM maxloc_p
INCLUDE 'mpif.h'
INTEGER n(9)
INTEGER isend(2), irecv(2)
DATA n /12, 15, 2, 20, 8, 3, 7, 24, 52/
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
ista = myrank * 3 + 1
iend = ista + 2
imax = -999
DO i = ista, iend
  IF (n(i) > imax) THEN
    imax = n(i)
    iloc = i
  ENDIF
ENDDO
isend(1) = imax
isend(2) = iloc
CALL MPI_REDUCE(isend, irecv, 1, MPI_INTEGER, MPI_MAXLOC, 0, MPI_COMM_WORLD, ierr)
IF (myrank == 0) THEN
  PRINT *, 'Max =', irecv(1), 'Location =', irecv(2)
ENDIF
CALL MPI_FINALIZE(ierr)
END

```

\$ a.out -procs 3

0: Max = 52 Location = 9

```
SUBROUTINE para_range(n1,n2,nprocs,irank,ista,iend)
iwork = (n2 - n1) / nprocs + 1
ista = MIN(irank * iwork + n1, n2 + 1)
iend = MIN(ista + iwork - 1, n2)
END
```

```
PROGRAM main
PARAMETER (n = 1000)
DIMENSION a(n)
DO i = 1, n
a(i) = i
ENDDO
sum = 0.0
DO i = 1, n
sum = sum + a(i)
ENDDO
PRINT *, 'sum =', sum
END
```

```

PROGRAM main
INCLUDE 'mpif.h'
PARAMETER (n = 1000)
DIMENSION a(n)
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
CALL para_range(1, n, nprocs, myrank, ista, iend)
DO i = ista, iend
a(i) = i
ENDDO
sum = 0.0
DO i = ista, iend
sum = sum + a(i)
ENDDO
CALL MPI_REDUCE(sum, ssum, 1, MPI_REAL, MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum = ssum
IF (myrank == 0) PRINT *, 'sum =', sum
CALL MPI_FINALIZE(ierr)
END

```

```
DO i = n1, n2  
  computation  
ENDDO
```

```
DO i = n1 + myrank, n2, nprocs  
  computation  
ENDDO
```

```
PROGRAM main  
INCLUDE 'mpif.h'  
PARAMETER (n = 1000)  
DIMENSION a(n)  
CALL MPI_INIT(ierr)  
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)  
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)  
DO i = 1 + myrank, n, nprocs  
  a(i) = i  
ENDDO  
sum = 0.0  
DO i = 1 + myrank, n, nprocs  
  sum = sum + a(i)  
ENDDO  
CALL MPI_REDUCE(sum, ssum, 1, MPI_REAL, MPI_SUM, 0, MPI_COMM_WORLD, ierr)  
sum = ssum  
PRINT *, 'sum = ', sum  
CALL MPI_FINALIZE(ierr)  
END
```

```
DO i = n1, n2  
  computation  
ENDDO
```

```
DO ii = n1 + myrank * iblock, n2, nprocs * iblock  
  DO i = ii, MIN(ii + iblock - 1, n2)  
    computation  
  ENDDO  
ENDDO
```

shrink

```
PROGRAM main
INCLUDE 'mpif.h'
PARAMETER (n1 = 1, n2 = 1000)
REAL, ALLOCATABLE :: a(:)
CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
CALL para_range(n1, n2, nprocs, myrank, ista, iend)
ALLOCATE (a(ista:iend))
DO i = ista, iend
  a(i) = i
ENDDO
sum = 0.0
DO i = ista, iend
  sum = sum + a(i)
ENDDO
DEALLOCATE (a)
CALL MPI_REDUCE(sum, ssum, 1, MPI_REAL, MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum = ssum
PRINT *, 'sum = ', sum
CALL MPI_FINALIZE(ierr)
END
```

shrink

```
PROGRAM main
implicit none
real sum1,ssum
integer i,ista,iend
integer ierr,n1,n2,nprocs,myrank
INCLUDE 'mpif.h'
PARAMETER (n1 = 1, n2 = 1000)
REAL, ALLOCATABLE :: a(:)

CALL MPI_INIT(ierr)
CALL MPI_COMM_SIZE(MPI_COMM_WORLD, nprocs, ierr)
CALL MPI_COMM_RANK(MPI_COMM_WORLD, myrank, ierr)
CALL para_range(n1, n2, nprocs, myrank, ista, iend)
ALLOCATE (a(ista:iend))
DO i = ista, iend
  a(i) = i
ENDDO
! sum1 = 0.0
! DO i = ista, iend
! sum1 = sum1 + a(i)
! ENDDO
sum1=sum(a)
DEALLOCATE (a)
CALL MPI_REDUCE(sum1, ssum, 1, MPI_REAL,MPI_SUM, 0, MPI_COMM_WORLD, ierr)
sum1 = ssum
PRINT *, 'sum1 =',sum1, myrank
CALL MPI_FINALIZE(ierr)
END
```

```
subroutine equal_load(n1,n2,nproc,myid,istart,ifinish)
!   Written by In-Ho Lee, KRISS, September (2006)
implicit none
integer nproc,myid,istart,ifinish,n1,n2
integer iw1,iw2
iw1=(n2-n1+1)/nproc ; iw2=mod(n2-n1+1,nproc)
istart=myid*iw1+n1+min(myid,iw2)
ifinish=istart+iw1-1 ; if(iw2 > myid) ifinish=ifinish+1
!   print*, n1,n2,myid,nproc,istart,ifinish
if(n2 < istart) ifinish=istart-1
return
end
```

```

!234567890
implicit none
include 'mpif.h'
integer, allocatable :: isend(:), irecv(:)
integer, allocatable :: ircnt(:), idisp(:)
integer ierr,nproc,myid
integer i,iscnt
integer ndsize

call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD,nproc,ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD,myid,ierr)

ndsize=10
allocate(isend(ndsize),irecv(nproc*ndsize))
allocate(ircnt(0:nproc-1),idisp(0:nproc-1))

ircnt=ndsize
idisp(0)=0
do i=1,nproc-1
  idisp(i)=idisp(i-1)+ircnt(i)
enddo

do i=1,ndsize      ! node specific data with a data-size ndsize
  isend(i)=myid+1
enddo

iscnt=ndsize
call MPI_GATHERV(isend, iscnt, MPI_INTEGER, irecv, ircnt, idisp, MPI_INTEGER,0, MPI_COMM_WORLD, ierr)

if(myid==0)then
  print*, 'irecv= ',irecv
endif

deallocate(ircnt,idisp)
deallocate(isend,irecv)
call MPI_FINALIZE(ierr)
stop
end

irecv= 1 1 1 1 1
1 1 1 1 1 2
2 2 2 2 2 2
2 2 2 3 3 3
3 3 3 3 3 3
3 4 4 4 4 4
4 4 4 4 4

```

!234567890

```
program integration
implicit none
include 'mpif.h'
integer i,n
integer myid,nproc,ierr,kount,iroot
real*8 ff,xx,xsum,pi,hh,xpi
character*8 fnnd ; character*10 fnnt
integer itemp,itemq,irate
ff(xx)= 4.0d0/(1.0d0+xx*xx)
call MPI_init(ierr)
call MPI_comm_size(MPI_COMM_WORLD,nproc,ierr)
call MPI_comm_rank(MPI_COMM_WORLD,myid,ierr)
if(myid == 0 .and. nproc > 1) print *, nproc," processes are alive"
if(myid == 0 .and. nproc == 1) print *, nproc," process is alive"
if(myid == 0)then ! ----[ process id = 0
call date_and_time(date=fnnd,time=fnnt)
write(6,'(a10,2x,a8,2x,a10)') 'date,time ', fnnd,fnnt
endif ! ----] process id = 0
if( myid == 0) then
write(6,*) 'number of intervals?'
read(5,*) n
write(6,*) 'number of intervals:',n
end if
iroot=0
call MPI_bcast(n,1,MPI_INTEGER, iroot, MPI_COMM_WORLD, ierr)
hh=1.d0/float(n)
xsum=0.0d0
do i=myid +1, n, nproc
xx=(float(i)-0.5d0)*hh
xsum=xsum+ff(xx)
end do
xpi=hh*xsum
call MPI_reduce(xpi,pi, 1, MPI_DOUBLE_PRECISION, MPI_SUM, 0, MPI_COMM_WORLD,ierr)
if( myid == 0 )then
write(6,'(a,f18.8)') 'estimated pi value', pi
end if
call MPI_finalize(ierr)
end program integration
```

```

program main
include "mpif.h"
double precision PI25DT
parameter      (PI25DT = 3.141592653589793238462643d0)
double precision mypi, pi, h, sum, x, f, a
double precision starttime, endtime
integer n, myid, numprocs, i, ierr
c              function to integrate
f(a) = 4.d0 / (1.d0 + a*a)

call MPI_INIT(ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, numprocs, ierr)

10 if ( myid .eq. 0 ) then
    print *, 'Enter the number of intervals: (0 quits)'
    read(*,*) n
endif
c              broadcast n
starttime = MPI_WTIME()
call MPI_BCAST(n,1,MPI_INTEGER,0,MPI_COMM_WORLD,ierr)
c              check for quit signal
if ( n .le. 0 ) goto 30
c              calculate the interval size
h = 1.0d0/n
sum = 0.0d0
do 20 i = myid+1, n, numprocs
    x = h * (dble(i) - 0.5d0)
    sum = sum + f(x)
20 continue
mypi = h * sum
c              collect all the partial sums
call MPI_REDUCE(mypi,pi,1,MPI_DOUBLE_PRECISION,MPI_SUM,0,
&              MPI_COMM_WORLD,ierr)
c              node 0 prints the answer.
endtime = MPI_WTIME()
if (myid .eq. 0) then
    print *, 'pi is ', pi, ' Error is', abs(pi - PI25DT)
    print *, 'time is ', endtime-starttime, ' seconds'
endif
goto 10
30 call MPI_FINALIZE(ierr)
stop
end

```

$$\int_0^1 \frac{4}{1+x^2} dx = \pi$$

```

SUBROUTINE PAR_BLAS2(m, n, a, b, c, comm)
REAL a(m), b(m,n) ! local slice of array
REAL c(n) !result
REAL sum(n)
INTEGER n, comm, i, j, ierr

! local sum
DO j= 1, n
sum(j) = 0.0
DO i = 1, m
sum(j) = sum(j) + a(i)*b(i,j)
END DO
END DO

! global sum
CALL MPI_ALLREDUCE(sum, c, n, MPI_REAL, MPI_SUM, comm, ierr)

! return result at all nodes
RETURN
END

```

dot_product: compute a scalar product

```
subroutine dot_product(global,x,y,n)
implicit none
include "mpif.h"
integer n,i,ierr
double precision global,x(n),y(n)
double precision tmp,local
local = 0.0d0
global = 0.0d0
do i=1,n
local = local + x(i)*y(i)
enddo
call MPI_ALLREDUCE(local,tmp,1,MPI_DOUBLE_PRECISION,
> MPI_SUM, MPI_COMM_WORLD, ierr)
global = tmp
return
end
```

```

SUBROUTINE PAR_BLAS2(m, n, a, b, c, comm)
REAL a(m), b(m,n) ! local slice of array
REAL c(n) ! result
REAL sum(n)
INTEGER n, comm, i, j, ierr

! local sum
DO j= 1, n
sum(j) = 0.0
DO i = 1, m
sum(j) = sum(j) + a(i)*b(i,j)
END DO
END DO

! global sum
CALL MPI_REDUCE(sum, c, n, MPI_REAL, MPI_SUM,0, comm, ierr)

! return result at node zero (and garbage at the other nodes)
RETURN
END

```

!234567890

```
program equal_load_sum
implicit none
include 'mpif.h'
integer nn
real*8, allocatable :: aa(:)
integer nproc,myid,ierr,istart,ifinish
integer i
real*8 xsum,xxsum
```

```
nn=10000
```

```
call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, nproc, ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, myid, ierr)
```

```
call equal_load(1,nn,nproc,myid,istart,ifinish)
```

allocate(aa(istart:ifinish)) ! 단순한 인덱스의 분할 뿐만아니라 메모리의 분할이 이루어지고 있다. 노드별로

```
do i=istart,ifinish
aa(i)=float(i)
enddo
```

```
xsum=0.0d0
do i=istart,ifinish
xsum=xsum+aa(i)
enddo
```

```
call MPI_REDUCE(xsum,xxsum,1,MPI_DOUBLE_PRECISION,MPI_SUM,0,MPI_COMM_WORLD,ierr)
xsum=xxsum
```

```
if(myid == 0)then
write(6,*) xsum,' xsum'
endif
```

```
deallocate(aa)
call MPI_FINALIZE(ierr)
end program equal_load_sum
```

메모리의 낭비가 있기는 하지만 뭔가 쉽게 쉽게 가자는 분위기로 만들 수 있는 방식.

```
allocate(exqq(natom_ac,3,0:np_ac), exforce(natom_ac,3,0:np_ac), exvofqj(0:np_ac))
if(myid == 0)then  ! ----{ PROCESS ID = 0
exqq=qq
endif  ! ----} PROCESS ID = 0
iroot=0 ; kount=3*natom_ac*(np_ac+1)
call MPI_BCAST(exqq, kount,MPI_REAL8, iroot,MPI_COMM_WORLD,ierr)
n1=0 ; n2=np_ac
call equal_load(n1,n2,nproc,myid,jstart,jfinish)
!
exforce=0.0d0 ; exvofqj=0.0d0
do j=jstart,jfinish
call xtinker(exqq(1,1,j),exforce(1,1,j),exvofqj(j),natom_ac,isequence,isymbol,iattyp,ii12)
enddo
iroot=0 ; kount=3*natom_ac*(np_ac+1)
call MPI_REDUCE(exforce,force_ac,kount,MPI_DOUBLE_PRECISION,MPI_SUM,iroot,MPI_COMM_WORLD,ierr)
iroot=0 ; kount=(np_ac+1)
call MPI_REDUCE(exvofqj,vofqj,kount,MPI_DOUBLE_PRECISION,MPI_SUM,iroot,MPI_COMM_WORLD,ierr)
!
deallocate(exforce, exvofqj, exqq)

real*8 aa(n,n), bb(n,n)

do j=1,n
do i=1,n
aa(i,j)=0.d0
bb(i,j)=0.0d0
enddo
enddo

do j=1,n
do i=1+myid,n, nproc
aa(i,j)=...
enddo
enddo

call MPI_REDUCE(aa,bb,n*n,MPI_DOUBLE_PRECISION, MPI_SUM, 0, MPI_COMM_WORLD, ierr)
```

<http://incredible.egloos.com/3880690>

MPI Implementations

- Most parallel machine vendors have optimized versions
- Others:
 - <http://www.mpi.nd.edu/MPI/Mpich>
 - <http://www-unix.mcs.anl.gov/mpi/mpich/>
 - [indexold.html](#)
 - GLOBUS:
 - <http://www.globus.org/mpi/>
 - <http://exodus.physics.ucla.edu/appleseed/>

추가 참고 문헌

“parallel computing” on the search engine

<https://computing.llnl.gov/tutorials/mpi/>

<http://condor.cc.ku.edu/~grobe/docs/intro-MPI.shtml>

<http://www-unix.mcs.anl.gov/mpi/>

http://www.llnl.gov/computing/tutorials/workshops/workshop/parallel_comp/MAIN.html#Whatis

<http://www.nas.nasa.gov/Groups/SciCon/Tutorials/MPIintro/toc.html>

<http://www-unix.mcs.anl.gov/mpi/tutorial/mpibasics/>

<http://www-unix.mcs.anl.gov/mpi/usingmpi/examples/main.htm>

http://people.sc.fsu.edu/~jburkardt/f_src/f_src.html

<http://people.sc.fsu.edu/~jburkardt/index.html>

Summary

- 새로운 패러다임 (병렬계산, 성능/가격)
- 통신의 중요성 (latency + bandwidth)
- MPI 프로그램의 기본 이해 (SPMD)
- 실제 문제 재설계 필요
- *점진적 병렬화 작업* (serial → parallel)
- 서로 다른 방식으로의 병렬화 모색
- 실질적 시간 절약 효과를 검증 해야 함