



TEXAS ADVANCED COMPUTING CENTER

WWW.TACC.UTEXAS.EDU



TEXAS

The University of Texas at Austin

# Tutorial on MPI programming, Full Course

Victor Eijkhout [eijkhout@tacc.utexas.edu](mailto:eijkhout@tacc.utexas.edu)  
TACC training, 2018

# Justification

The MPI library is the main tool for parallel programming on a large scale. This course introduces the main concepts through lecturing and exercises.

# The SPMD model

# Overview

In this section you will learn how to think about parallelism in MPI.

Commands learned:

- MPI\_Init, MPI\_Finalize,
- MPI\_Get\_processor\_name, MPI\_Comm\_size, MPI\_Comm\_rank

# Table of Contents

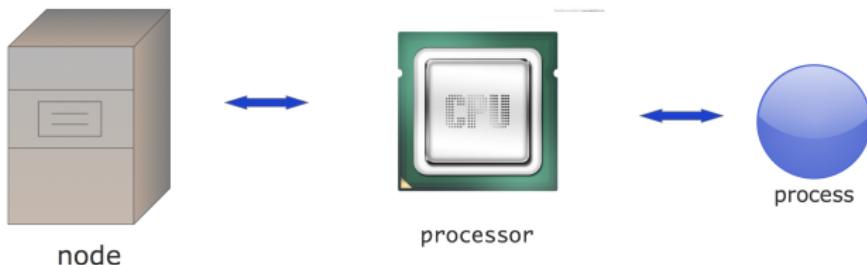
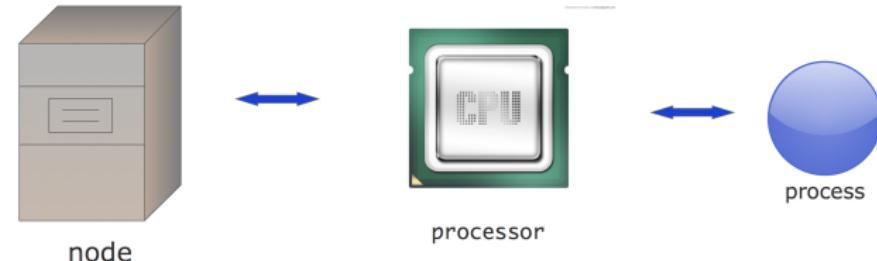
1 The MPI worldview: SPMD

2 Practicalities

3 We start learning MPI!

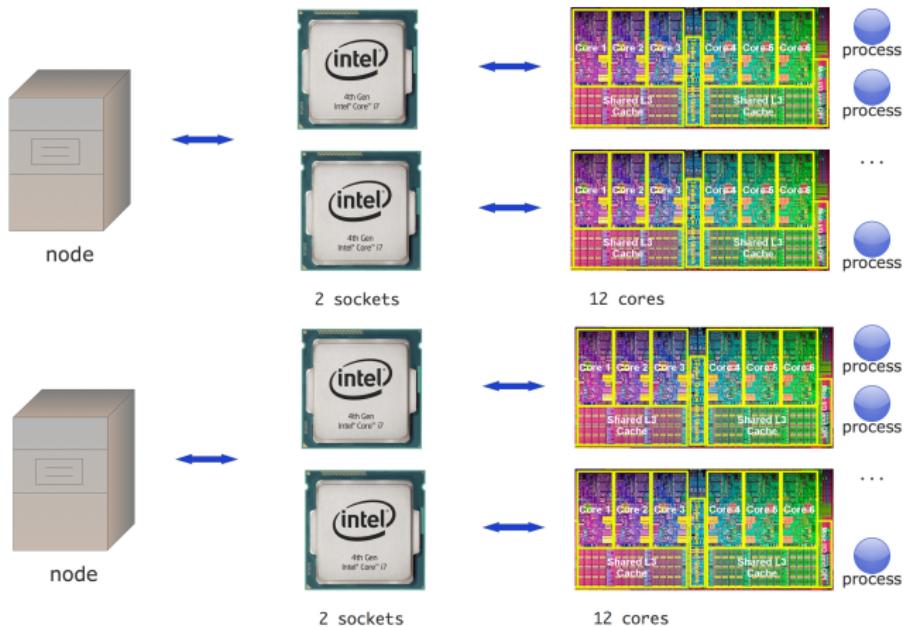
4 A practical example

# Computers when MPI was designed



One processor and one process per node;  
all communication goes through the network.

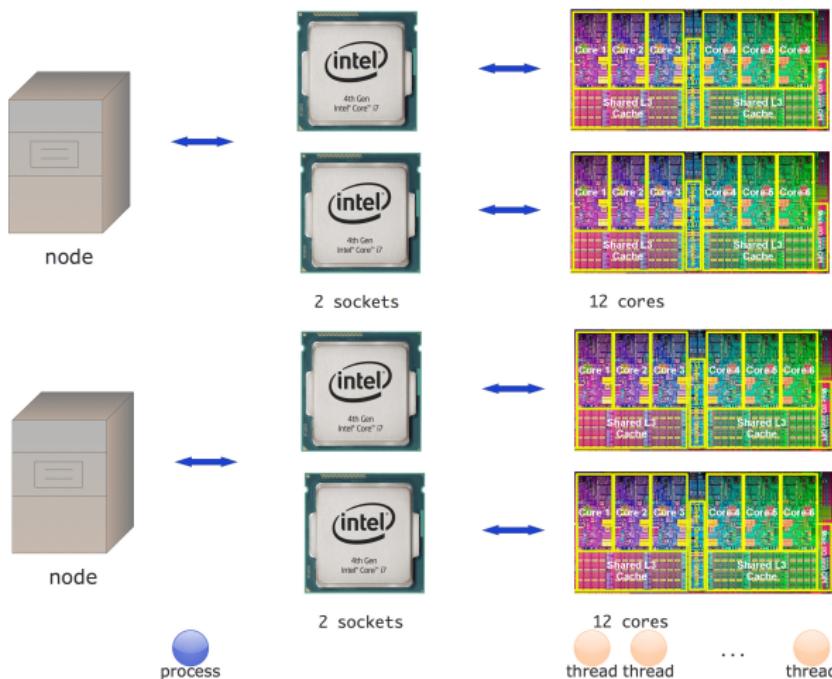
# Pure MPI



A node has multiple sockets, each with multiple cores.

Pure MPI puts a process on each core: pretend shared memory doesn't exist.

# Hybrid programming



Hybrid programming puts a process per node or per socket;  
further parallelism comes from threading.  
Not in this course...

# Terminology

'Processor' is ambiguous: is that a chip or one independent instruction processing unit?

- Socket: the processor chip
- Processor: we don't use that word
- Core: one instruction-stream processing unit
- Process: preferred terminology in talking about MPI.

# SPMD

The basic model of MPI is  
'Single Program Multiple Data':  
each process is an instance of the same program.

Symmetry: There is no 'master process', all processes are equal, start and end at the same time.

Communication calls do not see the cluster structure:  
data sending/receiving is the same for all neighbours.

# Table of Contents

1 The MPI worldview: SPMD

2 Practicalities

3 We start learning MPI!

4 A practical example

# Compiling and running

MPI compilers are usually called `mpicc`, `mpif90`, `mpicxx`.

These are not separate compilers, but scripts around the regular C/Fortran compiler. You can use all the usual flags.

Run your program with something like

```
mpiexec -n 4 hostfile ... yourprogram arguments
```

```
mpirun -np 4 hostfile ... yourprogram arguments
```

At TACC:

```
ibrun yourprog
```

the number of processes is determined by SLURM.

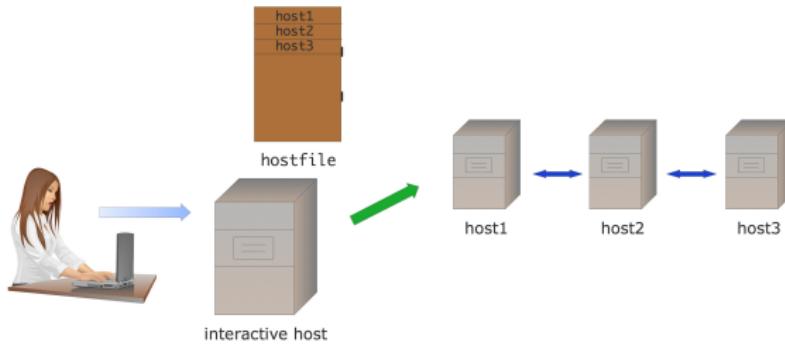
# Do I need a supercomputer?

- With `mpiexec` and such, you start a bunch of processes that execute your MPI program.
- Does that mean that you need a cluster or a big multicore?
- No! You can start a large number of MPI processes, even on your laptop. The OS will use ‘time slicing’.
- Of course it will not be very efficient...

# Cluster setup

Typical cluster:

- Login nodes, where you ssh into; usually shared with 100 (or so) other people.  
You don't run your parallel program there!
- Compute nodes: where your job is run. They are often exclusive to you: no other users getting in the way of your program.



Hostfile: the description of where your job runs. Usually generated by a job scheduler.

# How to make exercises

- Directory: exercises-mpi-c or cxx or f or f08 or p
- If a slide has a (exercisename) over it, there will be a template program exercisename.c (or F90 or py).
- Type make exercisename to compile it
- Python: setup once per session

```
module load python
```

No compilation needed. Run:

```
ibrun python yourprogram
```

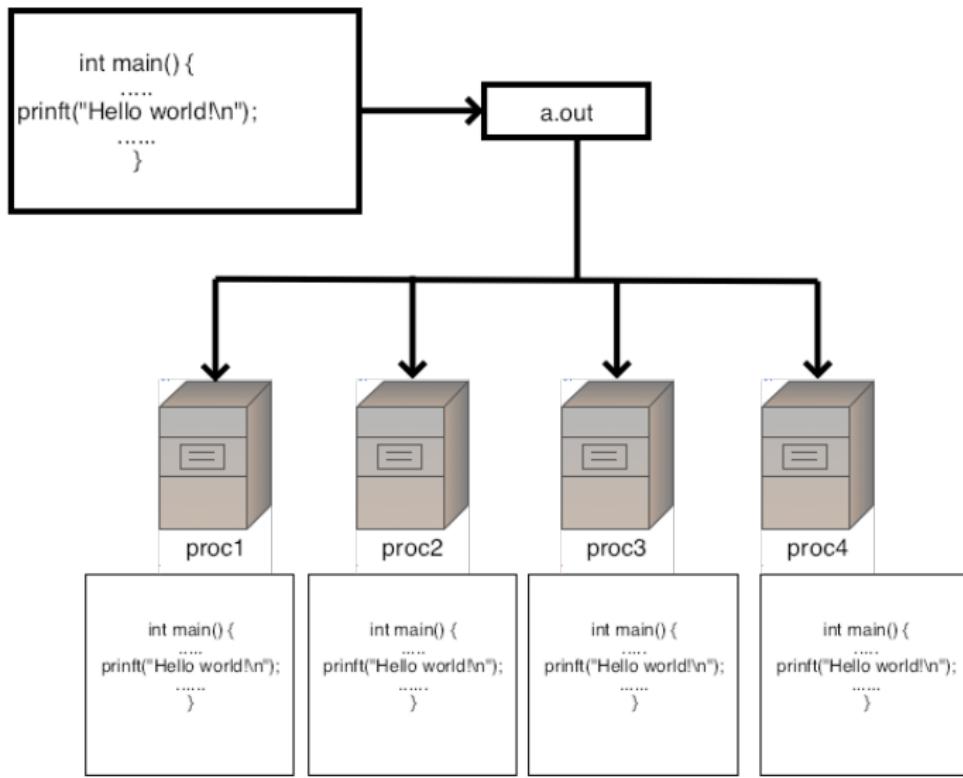
- Add an exercise of your own to the makefile: add the name to the EXERCISES

## Exercise 1 (hello)

Write a ‘hello world’ program, without any MPI in it, and run it in parallel with `mpieexec` or your local equivalent. Explain the output.

(On TACC machines such as stampede, use `ibrun`, no processor count.)

# In a picture



# Table of Contents

1 The MPI worldview: SPMD

2 Practicalities

3 We start learning MPI!

4 A practical example

# MPI definitions

You need an include file:

```
#include "mpi.h" // for C  
#include "mpif.h" ! for Fortran
```

- There are no real C++ bindings.
- True Fortran bindings as of the 2008 standard. Provided in Intel compiler:

```
module load intel/18.0.2
```

or newer. Not in gcc7.

# MPI Init / Finalize

Then put these calls around your code:

```
ierr = MPI_Init(&argc,&argv); // zeros allowed  
// your code  
ierr = MPI_Finalize();
```

and for Fortran:

```
call MPI_Init(ierr)  
! your code  
call MPI_Finalize(ierr)
```

# About error codes

MPI routines return an integer error code

- In C: function result. Can be ignored.
- In Fortran: as parameter.
- In Python: throwing exception.

There's actually not a lot you can do with an error code:  
very hard to recover from errors in parallel.  
Just ignore them ...

# Python bindings

```
module load python  
  
from mpi4py import MPI
```

Run:

```
ibrun python yourprogram.py
```

No initialization needed.

## Exercise 2

Add the commands `MPI_Init` and `MPI_Finalize` to your code. Put three different print statements in your code: one before the init, one between init and finalize, and one after the finalize. Again explain the output.

# Process identification

Every process has a number (with respect to a communicator)

```
int MPI_Comm_rank( MPI_Comm comm, int *procno )
int MPI_Comm_size( MPI_Comm comm, int *nprocs )
```

For now, the communicator will be MPI\_COMM\_WORLD.

Note: mapping of ranks to actual processes and cores is not predictable!

# About routine prototypes: C/C++

Prototype:

```
int MPI_Comm_size(MPI_Comm comm, int *nprocs)
```

Use:

```
MPI_Comm comm = MPI_COMM_WORLD;  
int nprocs;  
int errorcode;  
errorcode = MPI_Comm_size( comm, &nprocs );
```

(but forget about that error code most of the time)

# About routine prototypes: Fortran

## Prototype

```
MPI_Comm_size(comm, size, ierror)
INTEGER, INTENT(IN) :: comm
INTEGER, INTENT(OUT) :: size
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

## Use:

```
integer :: comm = MPI_COMM_WORLD
integer :: size
CALL MPI_Comm_size( comm, size, ierr )
```

- Final parameter always error parameter. Do not forget!
- Most MPI\_... types are INTEGER.

# About routine prototypes: Python

Prototype:

```
# object method  
MPI.Comm.Send(self, buf, int dest, int tag=0)  
# class method  
MPI.Request.Waitall(type cls, requests, statuses=None)
```

Use:

```
from mpi4py import MPI  
comm = MPI.COMM_WORLD  
comm.Send(sendbuf, dest=other)  
MPI.Request.Waitall(requests)
```

Semantics:

MPI\_COMM\_SIZE(comm, size)

IN comm: communicator (handle)

OUT size: number of processes in the group of comm (integer)

C:

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

Fortran:

MPI\_Comm\_size(comm, size, ierror)

TYPE(MPI\_Comm), INTENT(IN) :: comm

INTEGER, INTENT(OUT) :: size

INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Python:

```
MPI.Comm.Get_size(self)
```

Semantics:

MPI\_COMM\_RANK(comm, rank)

IN comm: communicator (handle)

OUT rank: rank of the calling process in group of comm (integer)

C:

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

Fortran:

```
MPI_Comm_rank(comm, rank, ierror)
```

TYPE(MPI\_Comm), INTENT(IN) :: comm

INTEGER, INTENT(OUT) :: rank

INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Python:

```
MPI.Comm.Get_rank(self)
```

## Exercise 3 (commrank)

Write a program where each process prints out message reporting its number, and how many processes there are.

Write a second version of this program, where each process opens a unique file and writes to it. *On some clusters this may not be advisable if you have large numbers of processors, since it can overload the file system.*

## Exercise 4 (commrank)

Write a program where only the process with number zero reports on how many processes there are in total.

## Exercise (optional) 5

Now use the command `MPI_Get_processor_name` in between the init and finalize statement, and print out on what processor your process runs. Confirm that you are able to run a program that uses two different nodes.

(The character buffer needs to be allocated by you, it is not created by MPI, with size at least `MPI_MAX_PROCESSOR_NAME`.) TACC nodes have a hostname `cRRR-CNN`, where RRR is the rack number, C is the chassis number in the rack, and NN is the node number within the chassis. Communication is faster inside a rack than between racks!

C:

```
int MPI_Get_processor_name(char *name, int *resultlen)
name : buffer char[MPI_MAX_PROCESSOR_NAME]
```

Fortran:

```
MPI_Get_processor_name(name, resultlen, ierror)
CHARACTER(LEN=MPI_MAX_PROCESSOR_NAME), INTENT(OUT) :: name
INTEGER, INTENT(OUT) :: resultlen
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
MPI.Get_processor_name()
```

# In a picture

Four processes on two nodes (idev -N 2 -n 4)

```
Program:  
number <- MPI_Comm_rank  
  
name <- MPI_Get_processor_name
```

```
Program:  
number <- MPI_Comm_rank  
0  
name <- MPI_Get_processor_name  
c111.tacc.utexas.edu
```

```
Program:  
number <- MPI_Comm_rank  
1  
name <- MPI_Get_processor_name  
c111.tacc.utexas.edu
```

```
Program:  
number <- MPI_Comm_rank  
2  
name <- MPI_Get_processor_name  
c222.tacc.utexas.edu
```

```
Program:  
number <- MPI_Comm_rank  
3  
name <- MPI_Get_processor_name  
c222.tacc.utexas.edu
```

c111.tacc.utexas.edu

c222.tacc.utexas.edu

# Table of Contents

1 The MPI worldview: SPMD

2 Practicalities

3 We start learning MPI!

4 A practical example

# Functional Parallelism

Parallelism by letting each process do a different thing.

Example: divide up a search space.

Each process knows its rank, so it can find its part of the search space.

## Exercise 6 (prime)

Is the number  $N = 2,000,000,111$  prime? Let each process test a range of integers, and print out any factor they find. You don't have to test all integers  $< N$ : any factor is at most  $\sqrt{N} \approx 45,200$ .

(Hint: `i%0` probably gives a runtime error.)

# Collectives

# Overview

In this section you will learn ‘collective’ operations, that combine information from all processes.

Commands learned:

- MPI\_Bcast, MPI\_Reduce, MPI\_Gather, MPI\_Scatter
- MPI\_All... variants, MPI\_....v variants
- MPI\_Barrier, MPI\_Alltoall, MPI\_Scan

# Table of Contents

5 Concepts

6 Basic collectives

7 Advanced collectives

# Collectives

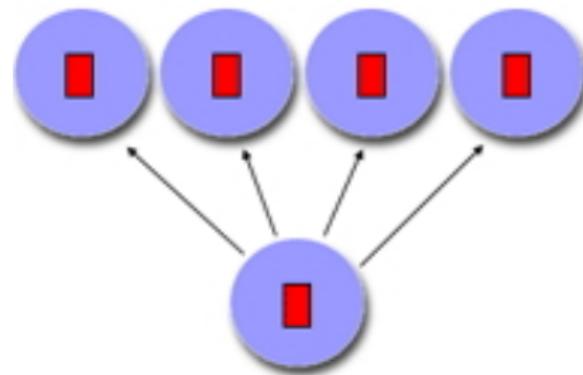
Gathering and spreading information:

- Every process has data, you want to bring it together;
- One process has data, you want to spread it around.

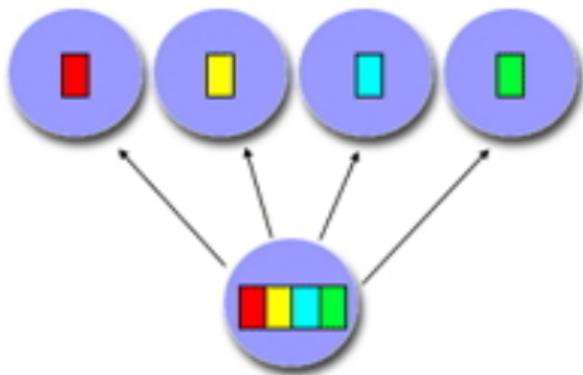
Root process: the one doing the collecting or disseminating.

Basic cases:

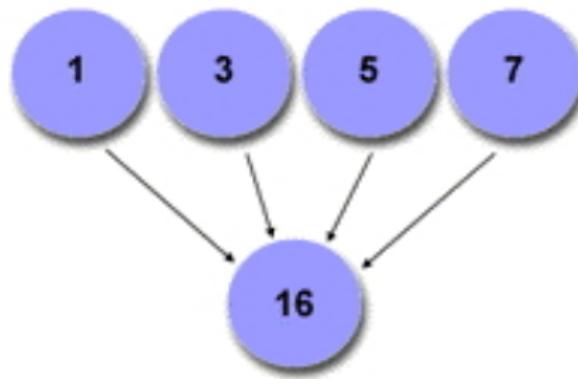
- Collect data: gather.
- Collect data and compute some overall value (sum, max): reduction.
- Send the same data to everyone: broadcast.
- Send individual data to each process: scatter.



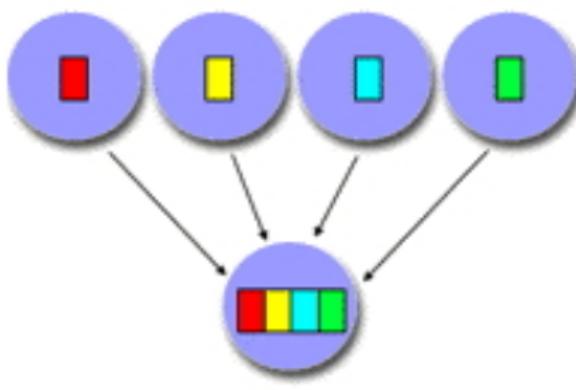
broadcast



scatter



reduction



gather

## Exercise 7

How would you realize the following scenarios with MPI collectives?

- Let each process compute a random number. You want to print the maximum of these numbers to your screen.
- Each process computes a random number again. Now you want to scale these numbers by their maximum.
- Let each process compute a random number. You want to print on what processor the maximum value is computed.

## More collectives

- Instead of a root, collect to all: MPI\_All...  
• Scatter individual data, but also individual size: MPI\_Scatterv  
• Everyone broadcasts: all-to-all  
• Scan: like a reduction, but with partial results

...and more

# Table of Contents

5 Concepts

6 Basic collectives

7 Advanced collectives

# Motivation for allreduce

Standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_i^N (x_i - \mu)^2} \quad \text{where} \quad \mu = \frac{\sum_i^N x_i}{N}$$

and assume that every processor stores just one  $x_i$  value.

How do we compute this?

# Motivation for allreduce

Standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_i^N (x_i - \mu)^2} \quad \text{where} \quad \mu = \frac{\sum_i^N x_i}{N}$$

and assume that every processor stores just one  $x_i$  value.

How do we compute this?

- ➊ The calculation of the average  $\mu$  is a reduction.

# Motivation for allreduce

Standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_i^N (x_i - \mu)^2} \quad \text{where} \quad \mu = \frac{\sum_i^N x_i}{N}$$

and assume that every processor stores just one  $x_i$  value.

How do we compute this?

- ① The calculation of the average  $\mu$  is a reduction.
- ② Every process needs to compute  $x_i - \mu$  for its value  $x_i$ , so use allreduce operation, which does the reduction and leaves the result on all processors.

# Motivation for allreduce

Standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_i^N (x_i - \mu)^2} \quad \text{where} \quad \mu = \frac{\sum_i^N x_i}{N}$$

and assume that every processor stores just one  $x_i$  value.

How do we compute this?

- ① The calculation of the average  $\mu$  is a reduction.
- ② Every process needs to compute  $x_i - \mu$  for its value  $x_i$ , so use allreduce operation, which does the reduction and leaves the result on all processors.
- ③  $\sum_i (x_i - \mu)$  is another sum of distributed data, so we need another reduction operation. Might as well use alreduce.

## Another Allreduce

Example: normalizing a vector

$$y \leftarrow x / \|x\|$$

- Vectors  $x, y$  are distributed: every process has certain elements
- The norm calculation is an all-reduce: every process gets same value
- Every process scales its part of the vector.

## Reduction to single process

Regular reduce: great for printing out summary information at the end of your job.

# Allreduce syntax

```
int MPI_Allreduce(  
    const void* sendbuf,  
    void* recvbuf, int count, MPI_Datatype datatype,  
    MPI_Op op, MPI_Comm comm)
```

- All processes have send and recv buffer
- (No root argument)
- count is number of items in the buffer: 1 for scalar.
- MPI\_Datatype is MPI\_INT, MPI\_REAL8 et cetera.
- MPI\_Op is MPI\_SUM, MPI\_MAX et cetera.

# Elementary datatypes

C	Fortran	meaning
MPI_CHAR	MPI_CHARACTER	only for text
MPI_SHORT	MPI_BYTE	8 bits
MPI_INT	MPI_INTEGER	like the C/F types
MPI_FLOAT	MPI_REAL	
MPI_DOUBLE	MPI_DOUBLE_PRECISION	
	MPI_COMPLEX	
	MPI_LOGICAL	
		internal use
		MPI_Aint
		MPI_Offset

A bunch more.

# MPI operators

MPI operator	description
MPI_MAX	maximum
MPI_MIN	minimum
MPI_SUM	sum
MPI_PROD	product
MPI_LAND	logical and
MPI_BAND	bitwise and
MPI_LOR	logical or
MPI_BOR	bitwise or
MPI_LXOR	logical xor
MPI_BXOR	bitwise xor

A couple more.

# Buffers in C

General principle: buffer argument is address in memory of the data.

- Buffer is void pointer:
- write `&x` or `(void*)&x` for scalar
- write `x` or `(void*)x` for array

# Buffers in Fortran

General principle: buffer argument is address in memory of the data.

- Fortran always passes by reference:
- write  $x$  for scalar
- write  $x$  for array

# Buffers in Python

For many routines there are two variants:

- lowercase: can send Python objects;  
**output is return result**

```
result = comm.recv(...)
```

this uses pickle: slow.

- uppercase: communicates numpy objects;  
input and output are function argument.

```
result = np.empty(.....)
comm.Recv(result, ...)
```

basically wrapper around C code: fast

## Exercise 8 (randommax)

Let each process compute a random number, and compute the sum of these numbers using the MPI\_Allreduce routine.

(The operator is MPI\_SUM for C/Fortran, or MPI.SUM for Python.)

Each process then scales its value by this sum. Compute the sum of the scaled numbers and check that it is 1.

# Random numbers in C

```
// Initialize the random number generator  
srand(procno*(double)RAND_MAX/nprocs);  
// compute a random number  
randomfraction = (rand() / (double)RAND_MAX);
```

# Random numbers in Fortran

```
integer :: randsize
integer,allocatable,dimension(:) :: randseed
real :: random_value

call random_seed(size=randsize)
allocate(randseed(randsize))
do i=1,randsize
    randseed(i) = 1023*procno
end do
call random_seed(put=randseed)
```

## Exercise (optional) 9

Create on each process an array of length 2 integers, and put the values 1,2 in it on each process. Do a sum reduction on that array. Can you predict what the result should be? Code it. Was your prediction right?

# Reduction to root

```
int MPI_Reduce  
  (void *sendbuf, void *recvbuf,  
   int count, MPI_Datatype datatype,  
   MPI_Op op, int root, MPI_Comm comm)
```

- **Buffers:** sendbuf, recvbuf are ordinary variables/arrays.
- Every process has data in its sendbuf,  
Root combines it in recvbuf (ignored on non-root processes).
- count is number of items in the buffer: 1 for scalar.
- MPI\_Op is MPI\_SUM, MPI\_MAX et cetera.

# Broadcast

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

- All processes call with the same argument list
- root is the rank of the process doing the broadcast
- Each process allocates buffer space;  
root explicitly fills in values,  
all others receive values through broadcast call.
- Datatype is MPI\_FLOAT, MPI\_INT et cetera, different between C/Fortran.
- comm is usually MPI\_COMM\_WORLD

# Allreduce vs Reduce/Bcast

Allreduce is equivalent to reduce and broadcast.

- One line less code.
- Gives the implementation more possibilities for optimization.
- Is actually twice as fast: allreduce same time as reduce.

# Gauss-Jordan elimination

<https://youtu.be/aQYuwatlWME>

## Exercise 10 (jordan)

The Gauss-Jordan algorithm for solving a linear system with a matrix  $A$  (or computing its inverse) runs as follows:

for pivot  $k = 1, \dots, n$

    let the vector of scalings  $\ell_i^{(k)} = A_{ik}/A_{kk}$

    for row  $r \neq k$

        for column  $c = 1, \dots, n$

$$A_{rc} \leftarrow A_{rc} - \ell_r^{(k)} A_{rc}$$

where we ignore the update of the righthand side, or the formation of the inverse.

Let a matrix be distributed with each process storing one column. Implement the Gauss-Jordan algorithm as a series of broadcasts: in iteration  $k$  process  $k$  computes and broadcasts the scaling vector  $\{\ell_i^{(k)}\}_i$ . Replicate the right-hand side on all processors.

## Exercise (optional) 11

Bonus exercise: can you extend your program to have multiple columns per processor?

# Gather/Scatter

```
int MPI_Gather(  
    void *sendbuf, int sendcnt, MPI_Datatype sendtype,  
    void *recvbuf, int recvcnt, MPI_Datatype recvtype,  
    int root, MPI_Comm comm  
);  
int MPI_Scatter  
(void* sendbuf, int sendcount, MPI_Datatype sendtype,  
 void* recvbuf, int recvcount, MPI_Datatype recvtype,  
 int root, MPI_Comm comm)
```

- Compare buffers to ▶ reduce
- Scatter: the sendcount / Gather: the recvcount:  
this is not, as you might expect, the total length of the buffer; instead, it is  
the amount of data to/from each process.

Also: MPI\_Allgather

## Exercise 12

Let each process compute a random number. You want to print the maximum value and on what processor it is computed. What collective(s) do you use?  
Write a short program.

# Table of Contents

5 Concepts

6 Basic collectives

7 Advanced collectives

# Scan

Scan or ‘parallel prefix’: reduction with partial results

- Useful for indexing operations:
- Each process has an array of  $n_p$  elements;
- My first element has global number  $\sum_{q < p} n_q$ .

C:

```
int MPI_Scan(const void* sendbuf, void* recvbuf,
             int count, MPI_Datatype datatype, MPI_Op op, MPI_Comm comm)
IN sendbuf: starting address of send buffer (choice)
OUT recvbuf: starting address of receive buffer (choice)
IN count: number of elements in input buffer (non-negative integer)
IN datatype: data type of elements of input buffer (handle)
IN op: operation (handle)
IN comm: communicator (handle)
```

Fortran:

```
MPI_Scan(sendbuf, recvbuf, count, datatype, op, comm, ierror)
TYPE(*), DIMENSION(..), INTENT(IN) :: sendbuf
TYPE(*), DIMENSION(..) :: recvbuf
INTEGER, INTENT(IN) :: count
TYPE(MPI_Datatype), INTENT(IN) :: datatype
TYPE(MPI_Op), INTENT(IN) :: op
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
res = Intracomm.scan( sendobj=None, recvobj=None, op=MPI_SUM)
res = Intracomm.exscan( sendobj=None, recvobj=None,
```

## V-type collectives

- Gather/scatter but with individual sizes
- Requires displacement in the gather/scatter buffer

C:

```
int MPI_Gatherv(
    const void* sendbuf, int sendcount, MPI_Datatype sendtype,
    void* recvbuf, const int recvcounts[], const int displs[],
    MPI_Datatype recvtype, int root, MPI_Comm comm)
```

Semantics:

IN sendbuf: starting address of send buffer (choice)

IN sendcount: number of elements in send buffer (non-negative integer)

IN sendtype: data type of send buffer elements (handle)

OUT recvbuf: address of receive buffer (choice, significant only at root)

IN recvcounts: non-negative integer array (of length group size) containing t

IN displs: integer array (of length group size). Entry i specifies the displa

IN recvtype: data type of recv buffer elements (significant only at root) (ha

IN root: rank of receiving process (integer)

IN comm: communicator (handle)

Fortran:

```
MPI_Gatherv(sendbuf, sendcount, sendtype, recvbuf, recvcounts, displs, recvty
```

TYPE(\*), DIMENSION(..), INTENT(IN) :: sendbuf

TYPE(\*), DIMENSION(..) :: recvbuf

INTEGER, INTENT(IN) :: sendcount, recvcounts(\*), displs(\*), root

TYPE(MPI\_Datatype), INTENT(IN) :: sendtype, recv

TYPE(MPI\_Comm), INTENT(IN) :: comm

## All-to-all

- Every process does a scatter;
- (equivalently: every process gather)
- each individual data, but amounts are identical
- Example: data transposition in FFT

## All-to-allv

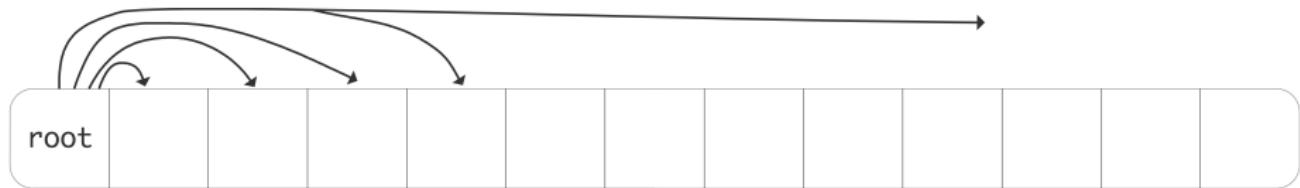
- Every process does a scatter or gather;
- each individual data and individual amounts.
- Example: radix sort by least-significant digit.

# Barrier

- Synchronize processes:
- each process waits at the barrier until all processes have reached the barrier
- **This routine is almost never needed**
- One conceivable use: timing

# Naive realization of collectives

Broadcast:

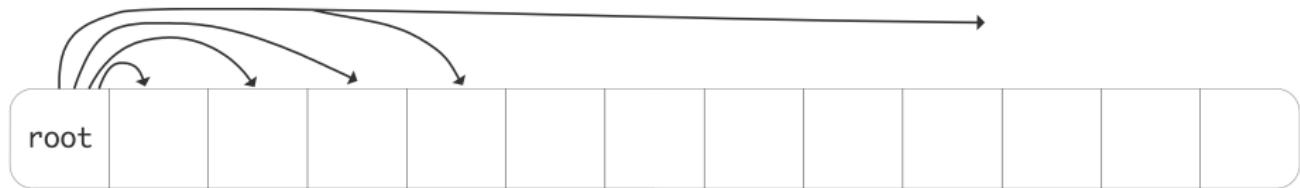


Single message:

$$\alpha = \text{message startup} \approx 10^{-6} \text{ s}, \quad \beta = \text{time per word} \approx 10^{-9} \text{ s}$$

# Naive realization of collectives

Broadcast:



Single message:

$$\alpha = \text{message startup} \approx 10^{-6} \text{ s}, \quad \beta = \text{time per word} \approx 10^{-9} \text{ s}$$

- Time for message of  $n$  words:

$$\alpha + \beta n$$

# Naive realization of collectives

Broadcast:



Single message:

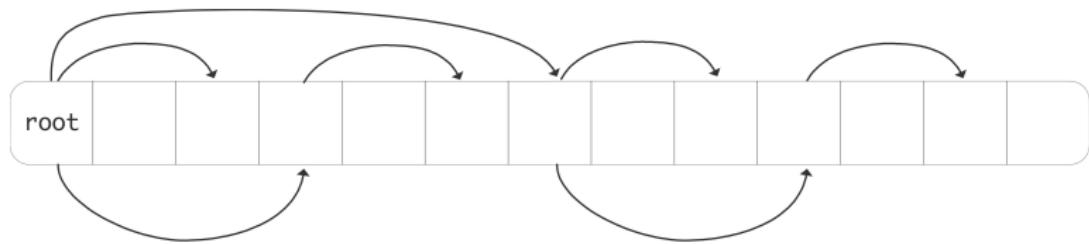
$$\alpha = \text{message startup} \approx 10^{-6} \text{ s}, \quad \beta = \text{time per word} \approx 10^{-9} \text{ s}$$

- Time for message of  $n$  words:

$$\alpha + \beta n$$

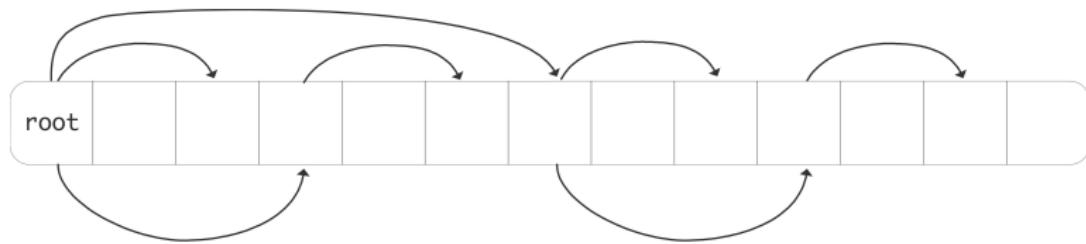
- Time for collective? Can you improve on that?

## Better implementation of collective



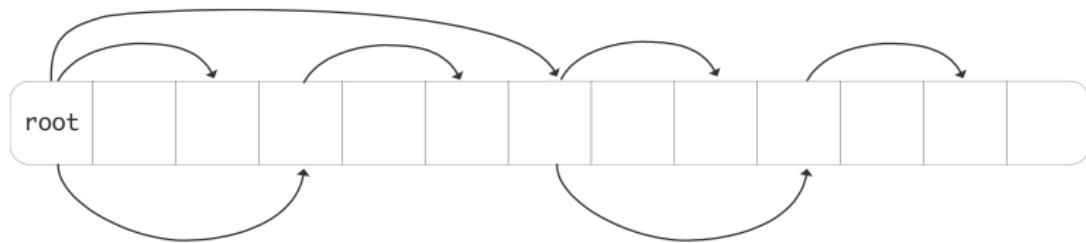
- What is the running time now?

## Better implementation of collective



- What is the running time now?
- Can you come up with lower bounds on the  $\alpha, \beta$  terms? Are these achieved here?

## Better implementation of collective



- What is the running time now?
- Can you come up with lower bounds on the  $\alpha, \beta$  terms? Are these achieved here?
- How about the case of really long buffers?

# Point-to-point communication

# Overview

This section concerns direct communication between two processes.  
Discussion of distributed work, deadlock and other parallel phenomena.

Commands learned:

- MPI\_Send, MPI\_Recv, MPI\_Sendrecv, MPI\_Isend, MPI\_Irecv
- MPI\_Wait...
- **Mention of MPI\_Test, MPI\_B/S/Rsend.**

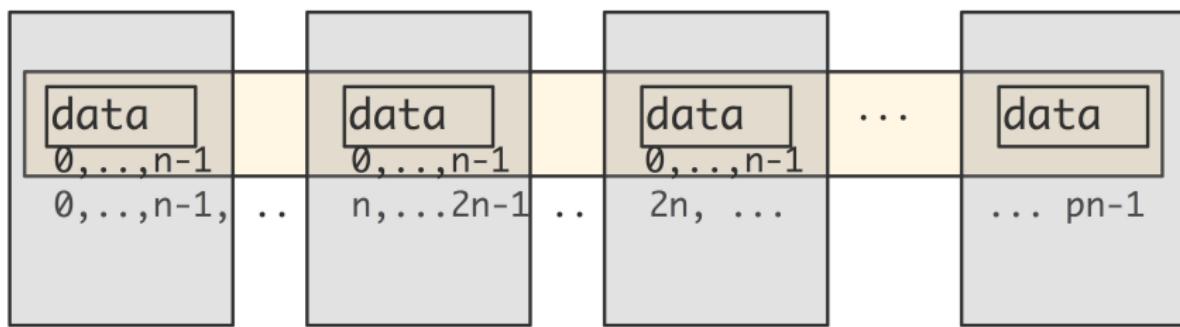
# Table of Contents

- 8 Distributed data
- 9 Local information exchange
- 10 Blocking communication
- 11 Pairwise exchange
- 12 Irregular exchanges: non-blocking communication
- 13 Communicating other than in pairs

# Distributed data

Distributed array: each process stores disjoint local part

```
int n;  
double data[n];
```



Local numbering  $0, \dots, n_{\text{local}}$ ;  
global numbering is ‘in your mind’.

# Local and global indexing

Every local array starts at 0 (Fortran: 1);  
you have to translate that yourself to global numbering:

```
int myfirst = .....;
for (int ilocal=0; ilocal<nlocal; ilocal++) {
    int iglobal = myfirst+ilocal;
    array[ilocal] = f(iglobal);
}
```

# Load balancing

If the distributed array is not perfectly divisible:

```
int Nglobal, // is something large
    Nlocal = Nglobal/nprocs,
    excess = Nglobal%nprocs;
if (procno==nprocs-1)
    Nlocal += excess;
```

This gives a load balancing problem. Better solution?

(for future reference)

Let

$$f(i) = \lfloor iN/p \rfloor$$

and give process  $i$  the points  $f(i)$  up to  $f(i+1)$ .

Result:

$$\lfloor N/p \rfloor \leq f(i+1) - f(i) \leq \lceil N/p \rceil$$

# Inner product calculation

Given vectors  $x, y$ :

$$x^t y = \sum_{i=0}^{N-1} x_i y_i$$

Start out with a distributed vector.

- Wrong way: collect the vector on one process and evaluate.
- Right way: compute local part, then collect local sums.

```
local_inprod = 0;  
for (i=0; i<localsize; i++)  
    local_inprod += x[i]*y[i];  
MPI_Allreduce( &local_inprod, &global_inprod, 1,MPI_DOUBLE ... )
```

## Exercise (optional) 13 (inprod)

Implement an inner product routine: let  $x$  be a distributed vector of size  $N$  with elements  $x[i] = i$ , and compute  $x^t x$ . As before, the right value is  $(2N^3 + 3N^2 + N)/6$ .

Use the inner product value to scale to vector so that it has norm 1. Check that your computation is correct.

## Exercise (optional) 14

Implement a (very simple-minded) Fourier transform: if  $f$  is a function on the interval  $[0, 1]$ , then the  $n$ -th Fourier coefficient is

$$f_n \hat{=} \int_0^1 f(t) e^{-t/\pi} dt$$

which we approximate by

$$f_n \hat{=} \sum_{i=0}^{N-1} f(ih) e^{-in/\pi}$$

- Make one distributed array for the  $e^{-inh}$  coefficients,
- make one distributed array for the  $f(ih)$  values
- calculate a couple of coefficients

# Table of Contents

- 8 Distributed data
- 9 Local information exchange
- 10 Blocking communication
- 11 Pairwise exchange
- 12 Irregular exchanges: non-blocking communication
- 13 Communicating other than in pairs

# Motivation

Partial differential equations:

$$-\Delta u = -u_{xx}(\bar{x}) - u_{yy}(\bar{x}) = f(\bar{x}) \text{ for } \bar{x} \in \Omega = [0, 1]^2 \text{ with } u(\bar{x}) = u_0 \text{ on } \delta\Omega.$$

Simple case:

$$-u_{xx} = f(x).$$

Finite difference approximation:

$$\frac{2u(x) - u(x+h) - u(x-h)}{h^2} = f(x, u(x), u'(x)) + O(h^2),$$

## Motivation (continued)

### Equations

$$\begin{cases} -u_{i-1} + 2u_i - u_{i+1} = h^2 f(x_i) & 1 < i < n \\ 2u_1 - u_2 = h^2 f(x_1) + u_0 \\ 2u_n - u_{n-1} = h^2 f(x_n) + u_{n+1}. \end{cases}$$

$$\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} h^2 f_1 + u_0 \\ h^2 f_2 \\ \vdots \end{pmatrix} \quad (1)$$

So we are interested in sparse/banded matrices.

# Matrix vector product

Most common operation: matrix vector product

$$y \leftarrow Ax, \quad A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \end{pmatrix}$$

- Component operation:  $y_i = 2x_i - x_{i-1} - x_{i+1}$
- Parallel execution: each process has range of  $i$ -coordinates
- So we need a point-to-point mechanism

# Operating on distributed data

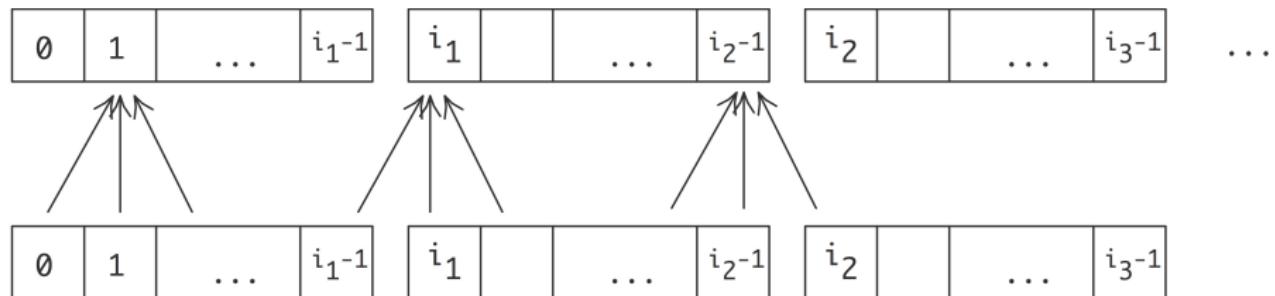
Array of numbers  $x_i: i = 0, \dots, N$

compute

$$y_i = (x_{i-1} + x_i + x_{i+1})/3: i = 1, \dots, N-1$$

'owner computes'

This leads to communication:



so we need a point-to-point mechanism.

# MPI point-to-point mechanism

- Two-sided communication
- Matched send and receive calls
- One process sends to a specific other process
- Other process does a specific receive.

# Ping-pong

A sends to B, B sends back to A

What is the code for A? For B?

# Ping-pong

A sends to B, B sends back to A

Process A executes the code

```
MPI_Send( /* to: */ B ..... );
MPI_Recv( /* from: */ B ... );
```

Process B executes

```
MPI_Recv( /* from: */ A ... );
MPI_Send( /* to: */ A ..... );
```

# Ping-pong in MPI

Remember SPMD:

```
if ( /* I am process A */ ) {  
    MPI_Send( /* to: */ B ..... );  
    MPI_Recv( /* from: */ B ... );  
} else if ( /* I am process B */ ) {  
    MPI_Recv( /* from: */ A ... );  
    MPI_Send( /* to: */ A ..... );  
}
```

C:

```
int MPI_Send(  
    const void* buf, int count, MPI_Datatype datatype,  
    int dest, int tag, MPI_Comm comm)
```

Semantics:

IN buf: initial address of send buffer (choice)  
IN count: number of elements in send buffer (non-negative integer)  
IN datatype: datatype of each send buffer element (handle)  
IN dest: rank of destination (integer)  
IN tag: message tag (integer)  
IN comm: communicator (handle)

Fortran:

```
MPI_Send(buf, count, datatype, dest, tag, comm, ierror)  
TYPE(*), DIMENSION(..), INTENT(IN) :: buf  
INTEGER, INTENT(IN) :: count, dest, tag  
TYPE(MPI_Datatype), INTENT(IN) :: datatype  
TYPE(MPI_Comm), INTENT(IN) :: comm  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python native:

```
MPI.Comm.send(self, obj, int dest, int tag=0)
```

Python, numpy:

C:

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

Semantics:

OUT buf: initial address of receive buffer (choice)  
IN count: number of elements in receive buffer (non-negative integer)  
IN datatype: datatype of each receive buffer element (handle)  
IN source: rank of source or MPI\_ANY\_SOURCE (integer)  
IN tag: message tag or MPI\_ANY\_TAG (integer)  
IN comm: communicator (handle)  
OUT status: status object (Status)

Fortran:

```
MPI_Recv(buf, count, datatype, source, tag, comm, status, ierror)  
TYPE(*), DIMENSION(..) :: buf  
INTEGER, INTENT(IN) :: count, source, tag  
TYPE(MPI_Datatype), INTENT(IN) :: datatype  
TYPE(MPI_Comm), INTENT(IN) :: comm  
TYPE(MPI_Status) :: status  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python native:

Eijkhout: MPI intro

## Status object

- Receive call can have various wildcards:  
`MPI_ANY_SOURCE, MPI_ANY_TAG`
- Receive buffer size is actually upper bound, not exact
- Use status object to retrieve actual description of the message  
use `MPI_STATUS_IGNORE` if the above does not apply

## Exercise 15 (pingpong)

Implement the ping-pong program. Add a timer using `MPI_Wtime`. For the status argument of the receive call, use `MPI_STATUS_IGNORE`.

- Run multiple ping-pongs (say a thousand) and put the timer around the loop. The first run may take longer; try to discard it.
- Run your code with the two communicating processes first on the same node, then on different nodes. Do you see a difference?
- Then modify the program to use longer messages. How does the timing increase with message size?

For bonus points, can you do a regression to determine  $\alpha, \beta$ ?

C:

```
double MPI_Wtime(void);
```

Fortran:

```
DOUBLE PRECISION MPI_WTIME()
```

Python:

```
MPI.Wtime()
```

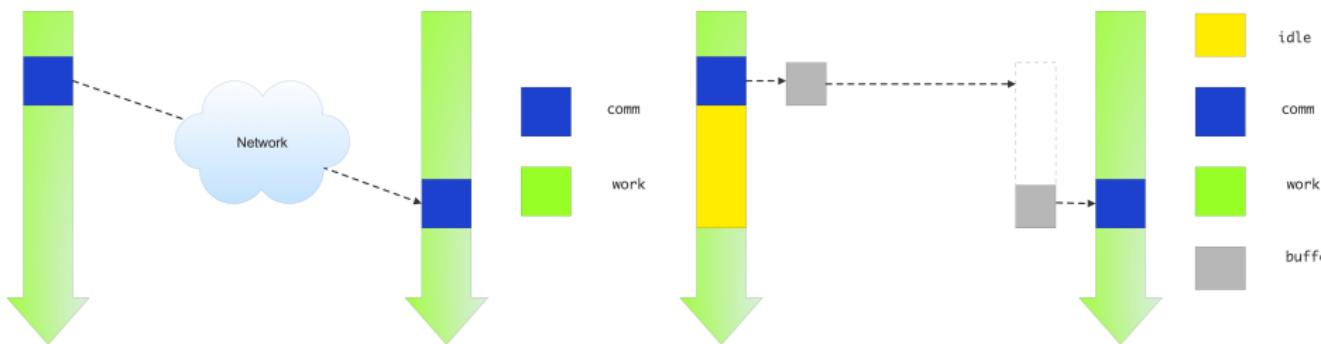
# Table of Contents

- 8 Distributed data
- 9 Local information exchange
- 10 Blocking communication
- 11 Pairwise exchange
- 12 Irregular exchanges: non-blocking communication
- 13 Communicating other than in pairs

# Blocking send/recv

`MPI_Send` and `MPI_Recv` are *blocking* operations:

- The process waits ('blocks') until the operation is concluded.
- A send can not complete until the receive executes.



Ideal vs actual send/recv behaviour.

# Deadlock

```
other = 1-procno; /* if I am 0, other is 1; and vice versa */
receive(source=other);
send(target=other);
```

A subtlety.

This code may actually work:

```
other = 1-procno; /* if I am 0, other is 1; and vice versa */
send(target=other);
receive(source=other);
```

Small messages get sent even if there is no corresponding receive.  
(Often a system parameter)

# Protocol

Communication is a ‘rendez-vous’ or ‘hand-shake’ protocol:

- Sender: ‘I have data for you’
- Receiver: ‘I have a buffer ready, send it over’
- Sender: ‘Ok, here it comes’
- Receiver: ‘Got it.’

Small messages bypass this: ‘eager’ send.

Definition of ‘small message’ controlled by environment variables.

## Exercise 16 (serialsend)

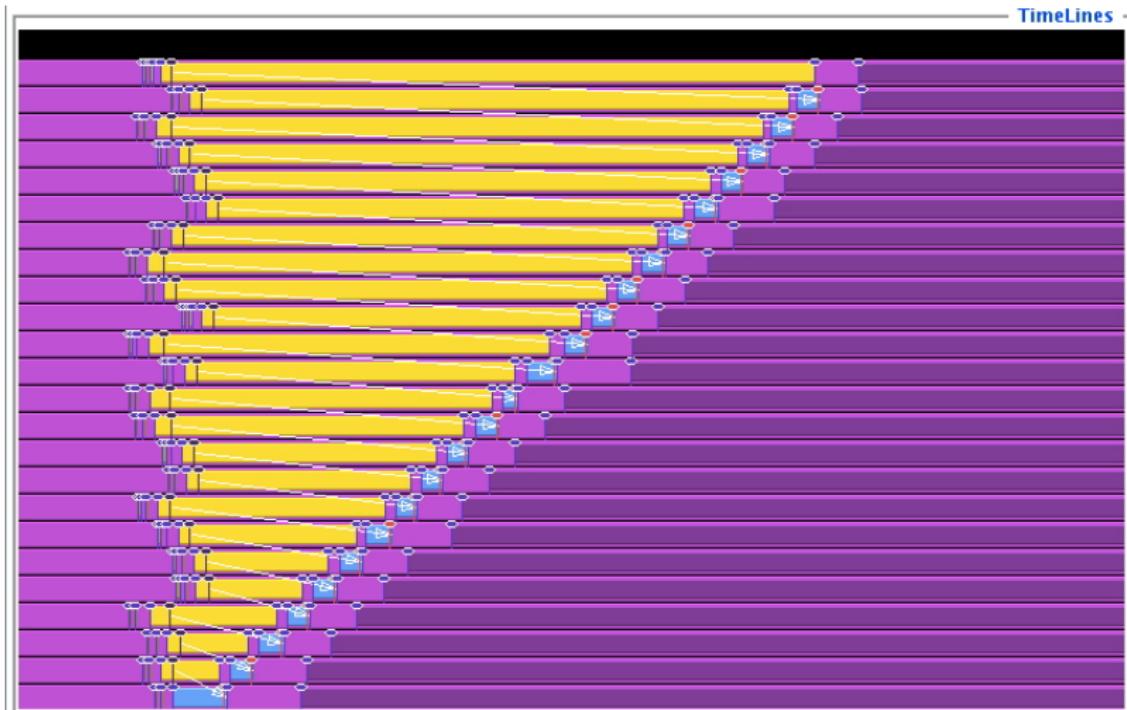
(Classroom exercise) Each student holds a piece of paper in the right hand – keep your left hand behind your back – and we want to execute:

- ① Give the paper to your right neighbour;
- ② Accept the paper from your left neighbour.

Including boundary conditions for first and last process, that becomes the following program:

- ① If you are not the rightmost student, turn to the right and give the paper to your right neighbour.
- ② If you are not the leftmost student, turn to your left and accept the paper from your left neighbour.

# TAU trace: serialization



## The problem here...

Here you have a case of a program that computes the right output, just way too slow.

Beware! Blocking sends/receives can be trouble.  
(How would you solve this particular case?)

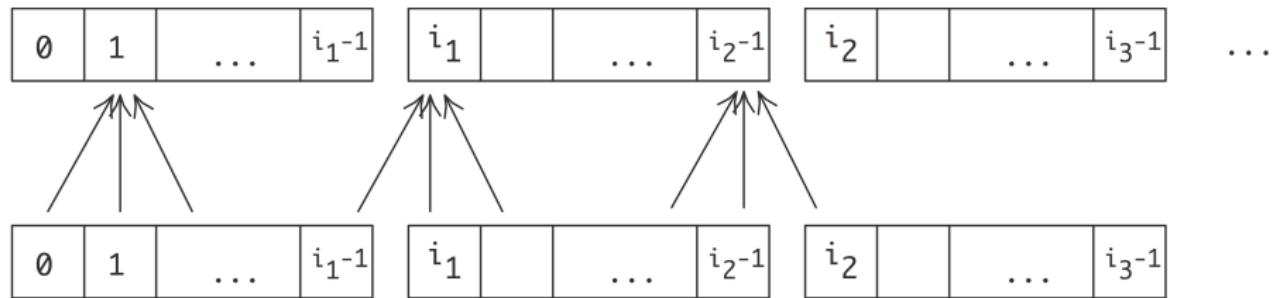
# Table of Contents

- 8 Distributed data
- 9 Local information exchange
- 10 Blocking communication
- 11 Pairwise exchange
- 12 Irregular exchanges: non-blocking communication
- 13 Communicating other than in pairs

# Operating on distributed data

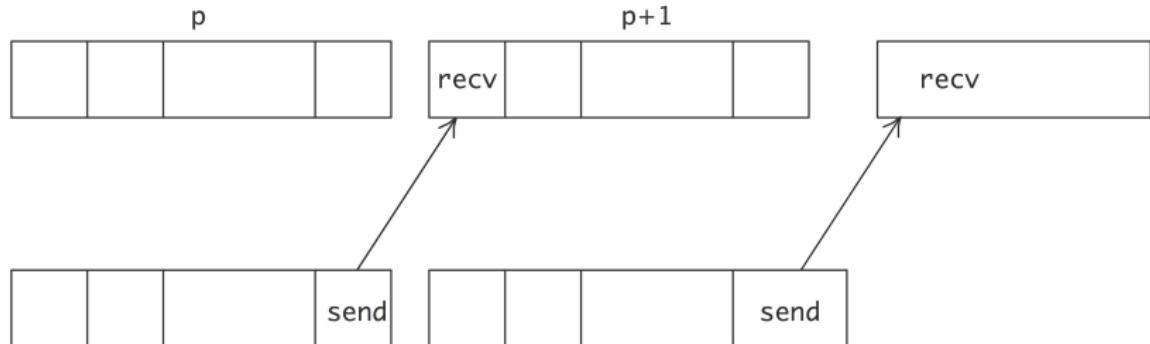
Take another look:

$$y_i = x_{i-1} + x_i + x_{i+1} : i = 1, \dots, N-1$$



- One-dimensional data and linear process numbering;
- Operation between neighbouring indices: communication between neighbouring processes.

# First step



First do all the data movement to the right.

- Each process does a send and receive
- So everyone does the send, then the receive? We just saw the problem with that.
- Better solution coming up!

# Sendrecv

Instead of separate send and receive: use

Semantics:

`MPI_SENDRECV(`

`sendbuf, sendcount, sendtype, dest, sendtag,`  
  `recvbuf, recvcount, recvtype, source, recvtag,`  
  `comm, status)`

IN `sendbuf`: initial address of send buffer (choice)

IN `sendcount`: number of elements in send buffer (non-negative integer)

IN `sendtype`: type of elements in send buffer (handle)

IN `dest`: rank of destination (integer)

IN `sendtag`: send tag (integer)

OUT `recvbuf`: initial address of receive buffer (choice)

IN `recvcount`: number of elements in receive buffer (non-negative integer)

IN `recvtype`: type of elements in receive buffer (handle)

IN `source`: rank of source or `MPI_ANY_SOURCE` (integer)

IN `recvtag`: receive tag or `MPI_ANY_TAG` (integer)

IN `comm`: communicator (handle)

OUT `status`: status object (`Status`)

## Sendrecv with incomplete pairs

```
MPI_Comm_rank( .... &procno );
if ( /* I am not the first process */ )
    predecessor = procno-1;
else
    predecessor = MPI_PROC_NULL;

if ( /* I am not the last process */ )
    successor = procno+1;
else
    successor = MPI_PROC_NULL;

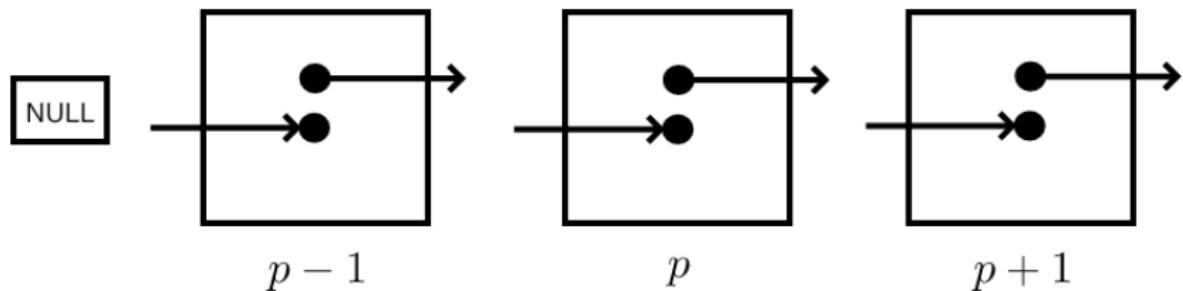
sendrecv(from=predecessor,to=successor);
```

## Exercise 17 (rightsend)

Revisit exercise 16 and solve it using MPI\_Sendrecv

# SPMD picture

What does process  $p$  do?

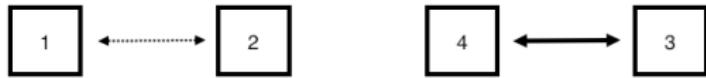
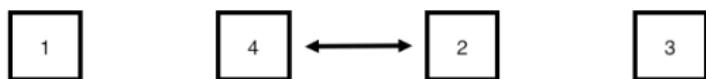


## Exercise 18 (sendrecv)

Implement the above three-point combination scheme using MPI\_Sendrecv; every processor only has a single number to send to its neighbour.

If you have TAU installed, make a trace. Does it look different from the serialized send/recv code? If you don't have TAU, run your code with different numbers of processes and show that the runtime is essentially constant.

# Odd-even transposition sort



↔ transpose performed  
↔ no transpose needed

Odd-even transposition sort on 4 elements.

## Exercise (optional) 19

A very simple sorting algorithm is *exchange sort*: pairs of processors compare data, and if necessary exchange. The elementary step is called a compare-and-swap: in a pair of processors each sends their data to the other; one keeps the minimum values, and the other the maximum. For simplicity, in this exercise we give each processor just a single number.

The exchange sort algorithm is split in even and odd stages, where in the even stage, processors  $2i$  and  $2i+1$  compare and swap data, and in the odd stage, processors  $2i+1$  and  $2i+2$  compare and swap. You need to repeat this  $P/2$  times, where  $P$  is the number of processors; see figure 115.

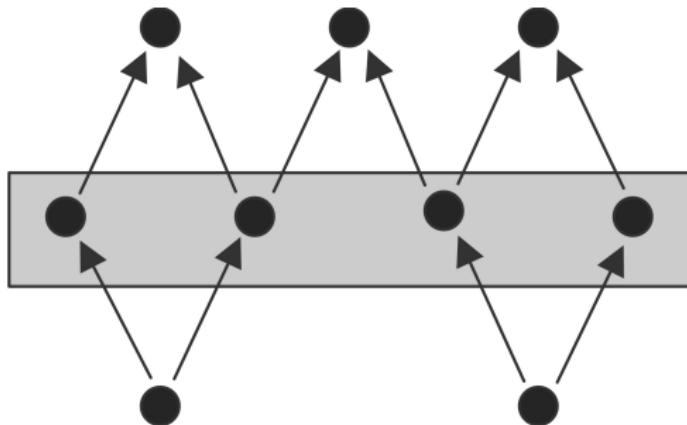
Implement this algorithm using `MPI_Sendrecv`. (Use `MPI_PROC_NULL` for the edge cases if needed.) Use a gather call to print the global state of the distributed array at the beginning and end of the sorting process.

# Table of Contents

- 8 Distributed data
- 9 Local information exchange
- 10 Blocking communication
- 11 Pairwise exchange
- 12 Irregular exchanges: non-blocking communication
- 13 Communicating other than in pairs

# Sending with irregular connections

Graph operations:

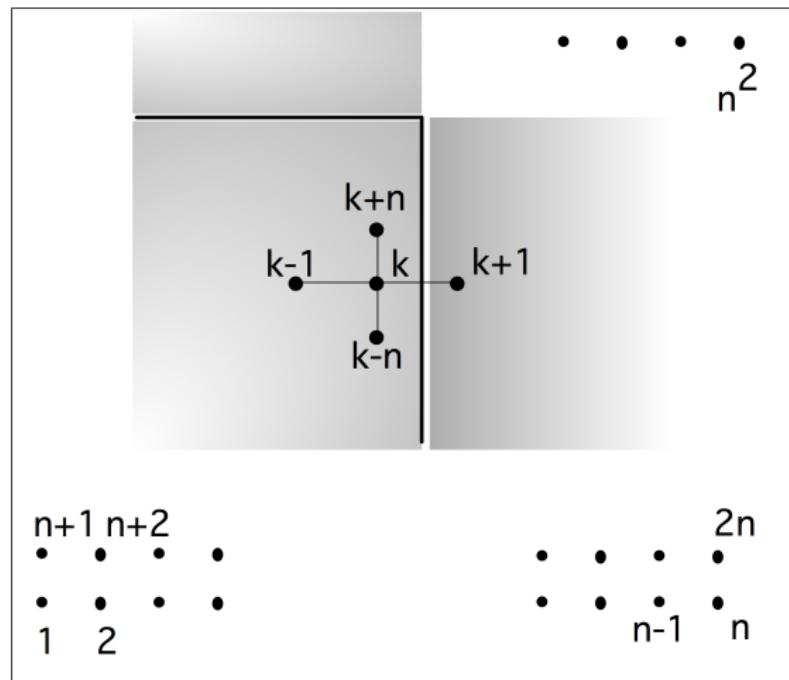


# Table of Contents

- 8 Distributed data
- 9 Local information exchange
- 10 Blocking communication
- 11 Pairwise exchange
- 12 Irregular exchanges: non-blocking communication
- 13 Communicating other than in pairs

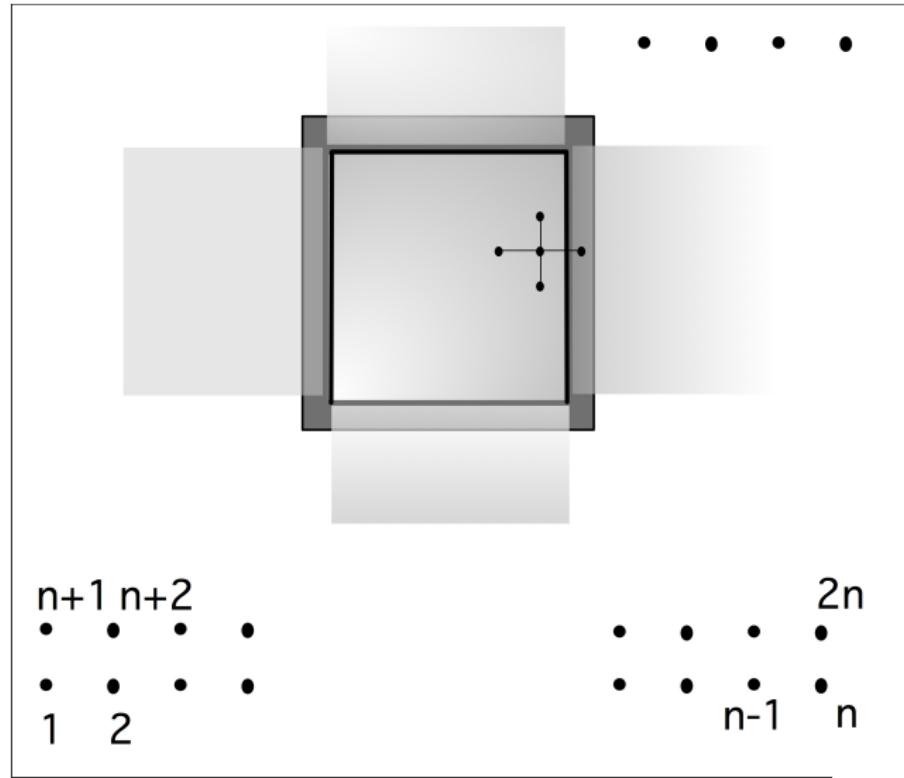
## PDE, 2D case

A difference stencil applied to a two-dimensional square domain, distributed over processors. A cross-processor connection is indicated  $\Rightarrow$  complicated to express pairwise



# Halo region

The halo region of a process, induced by a stencil



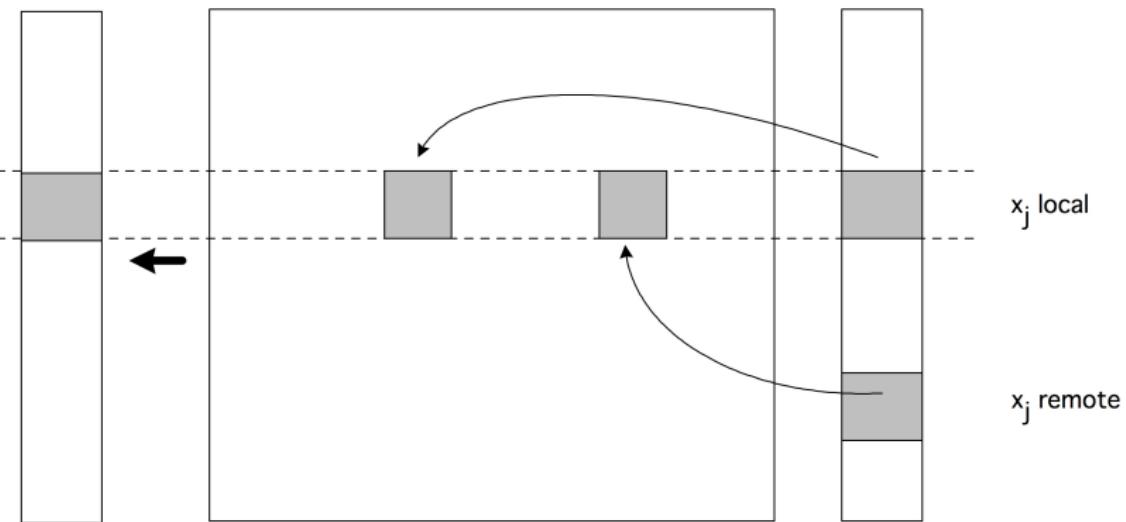
# PDE matrix

$$A = \left( \begin{array}{cccc|ccc|c} 4 & -1 & & & 0 & -1 & & & 0 \\ -1 & 4 & -1 & & & -1 & & & \\ \ddots & \ddots & \ddots & & & \ddots & & & \\ & \ddots & \ddots & -1 & & & \ddots & & \\ \hline 0 & & -1 & 4 & 0 & 0 & & -1 & \\ -1 & & 0 & & 4 & -1 & & -1 & \\ & -1 & & & -1 & 4 & -1 & & -1 \\ \hline & \overset{\uparrow}{k-n} & \ddots & & k-1 & \overset{\uparrow}{k} & \overset{\uparrow}{k+1} & & \overset{\uparrow}{k+n} \\ & & & & & & -1 & 4 & \\ \hline & & & & \ddots & & & & \ddots \end{array} \right)$$

# Matrices in parallel

$$y \leftarrow Ax$$

and  $A, x, y$  all distributed:



## How do you approach this?

- It is very hard to figure out a send/receive sequence that does not deadlock or serialize
- Even if you manage that, you may have process idle time.

Instead:

- Declare 'this data needs to be sent' or 'these messages are expected', and
- then wait for them collectively.

# Non-blocking send/recv

```
// start non-blocking communication  
MPI_Isend( ... ); MPI_Irecv( ... );  
// wait for the Isend/Irecv calls to finish in any order  
MPI_Wait( ... );
```

# Syntax

Very much like blocking `▶ send` and `▶ recv`:

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

The `MPI_Request` can be tested:

```
int MPI_Waitall(int count, MPI_Request array_of_requests[],  
    MPI_Status array_of_statuses[])
```

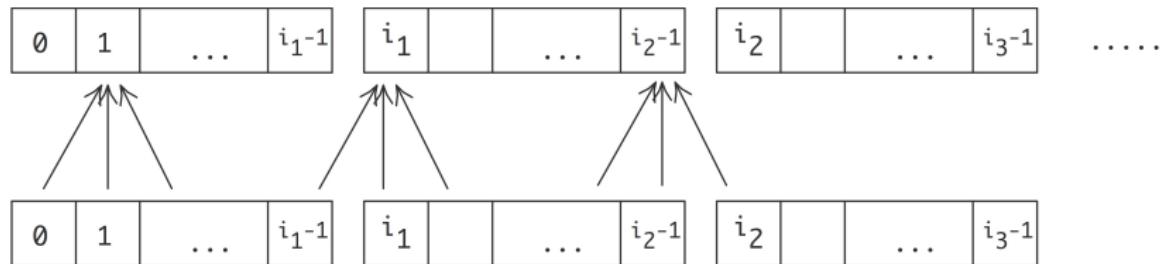
- ignore status: `MPI_STATUSES_IGNORE`
- also `MPI_Wait`, `MPI_Waitany`, `MPI_Waitsome`

## Exercise 20 (isendirecv)

Now use nonblocking send/receive routines to implement the three-point averaging operation

$$y_i = (x_{i-1} + x_i + x_{i+1})/3 : i = 1, \dots, N-1$$

on a distributed array. (Hint: use MPI\_PROC\_NULL at the ends.)



(Can you think of a different way of handling the end points?)

# Comparison

- Obvious: blocking vs non-blocking behaviour.
- Buffer reuse: when a blocking call returns, the buffer is safe for reuse;
- A buffer in a non-blocking call can only be reused after the wait call.

# Buffer use in blocking/non-blocking case

Blocking:

```
double *buffer;
for ( ... p ... ) {
    buffer = // fill in the data
    MPI_Send( buffer, ... /* to: */ p );
```

Non-blocking:

```
double **buffers;
for ( ... p ... ) {
    buffers[p] = // fill in the data
    MPI_Isend( buffers[p], ... /* to: */ p );
```

# Latency hiding

Other motivation for non-blocking calls:

overlap of computation and communication, provided hardware support.

Also known as 'latency hiding'.

Example: three-point combination operation (see above):

- ① Start communication for edge points,
- ② Do local operations while communication goes on,
- ③ Wait for edge points from neighbour processes
- ④ Incorporate incoming data.

## Exercise 21 (isendirecvarray)

Now use nonblocking send/receive routines to implement the three-point averaging operation

$$y_i = (x_{i-1} + x_i + x_{i+1}) / 3 : i = 1, \dots, N - 1$$

on a distributed array. (Hint: use `MPI_PROC_NULL` at the ends.)

Write your code so that it can achieve latency hiding.

## Test: non-blocking wait

- Post non-blocking receives
- test for incoming messages
- if nothing comes in, do local work

```
while (1) {  
    MPI_Test( /* from: */ ANY_SOURCE, &flag );  
    if (flag)  
        // do something with incoming message  
    else  
        // do local work  
}
```

## More sends and receive

- MPI\_Bsend, MPI\_Ibsend: buffered send
- MPI\_Ssend, MPI\_Issend: synchronous send
- MPI\_Rsend, MPI\_Irsend: ready send
- Persistent communication: repeated instance of same proc/data description.

too obscure to go into.

# Advanced topics

- One-sided communication: ‘just’ put/get the data somewhere
- Derived data types: send strided/irregular/inhomogeneous data
- Sub-communicators: work with subsets of `MPI_COMM_WORLD`
- I/O: efficient file operations
- Non-blocking collectives
- Graph topology and neighbourhood collectives

# One-sided communication

# Overview

This section concerns one-sided operations, which allows 'shared memory' type programming.

Commands learned:

- MPI\_Put, MPI\_Get, MPI\_Accumulate
- Active target synchronization MPI\_Win\_create, MPI\_Win\_fence
- MPI\_Post/Wait/Start/Complete
- Passive target synchronization MPI\_Win\_lock/unlock
- Atomic operations: MPI\_Fetch\_and\_op

# Table of Contents

14

Active target synchronization

15

Passive target synchronization

16

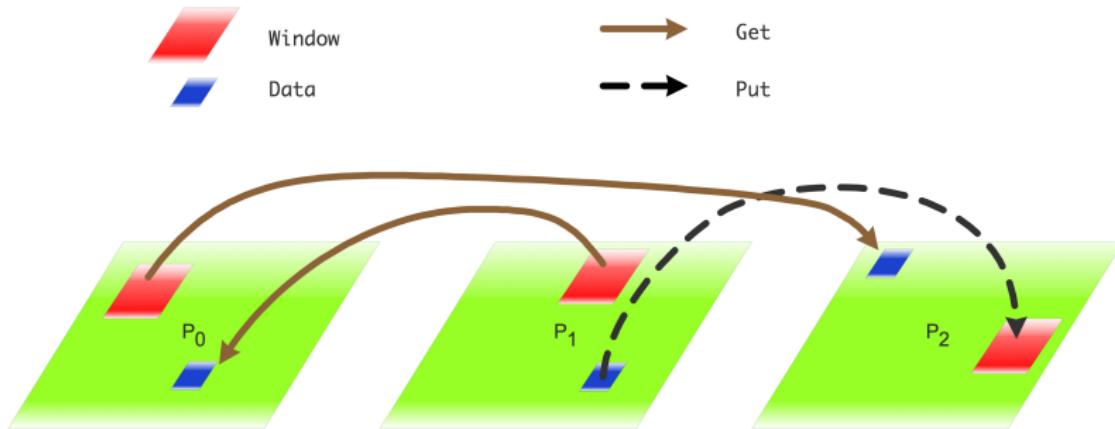
More

# Motivation

With two-sided messaging, you can not just put data on a different processor: the other has to expect it and receive it.

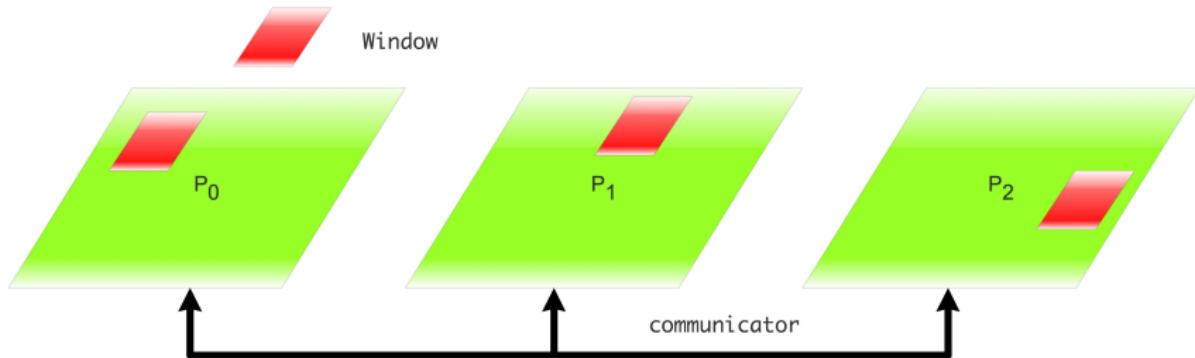
- Sparse matrix: it is easy to know what you are receiving, not what you need to send. Usually solved with complicated preprocessing step.
- Neuron simulation: spiking neuron propagates information to neighbours. Uncertain when this happens.
- Other irregular data structures: distributed hash tables.

# One-sided concepts



- A process has a window that other processes can access.
- Origin: process doing a one-sided call; target: process being accessed.
- One-sided calls: `MPI_Put`, `MPI_Get`, `MPI_Accumulate`.
- Various synchronization mechanisms.

# Window creation



```
MPI_Win_create (void *base, MPI_Aint size,  
    int disp_unit, MPI_Info info, MPI_Comm comm, MPI_Win *win)
```

- **size:** in bytes
- **disp\_unit:** sizeof(type)
- **Also:** MPI\_Win\_allocate, can use dedicated fast memory.

Also call MPI\_Win\_free when done. This is important!

# Window allocation

Instead of passing buffer, let MPI allocate:

```
int MPI_Win_allocate
    (MPI_Aint size, int disp_unit, MPI_Info info,
     MPI_Comm comm, void *baseptr, MPI_Win *win)
```

# Active target synchronization

All processes call MPI\_Win\_fence. Epoch is between fences:

```
MPI_Win_fence(MPI_MODE_NOPRECEDE, win);  
if (procno==producer)  
    MPI_Put( /* operands */, win);  
MPI_Win_fence(MPI_MODE_NOSUCCEED, win);
```

Second fence indicates that one-sided communication is concluded:  
target knows that data has been put.

C:

```
int MPI_Put(
    const void *origin_addr, int origin_count, MPI_Datatype origin_datatype,
    int target_rank, MPI_Aint target_disp, int target_count, MPI_Datatype target_datatype,
    MPI_Win win)
```

Semantics:

IN origin\_addr: initial address of origin buffer (choice)  
IN origin\_count: number of entries in origin buffer (non-negative integer)  
IN origin\_datatype: datatype of each entry in origin buffer (handle)  
IN target\_rank: rank of target (non-negative integer)  
IN target\_disp: displacement from start of window to target buffer (non-negative integer)  
IN target\_count: number of entries in target buffer (non-negative integer)  
IN target\_datatype: datatype of each entry in target buffer (handle)  
IN win: window object used for communication (handle)

Fortran:

```
MPI_Put(origin_addr, origin_count, origin_datatype,
         target_rank, target_disp, target_count, target_datatype, win, ierror)
TYPE(*), DIMENSION(..), INTENT(IN), ASYNCHRONOUS :: origin_addr
INTEGER, INTENT(IN) :: origin_count, target_rank, target_count
TYPE(MPI_Datatype), INTENT(IN) :: origin_datatype, target_datatype
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(IN) :: target_disp
TYPE(MPI_Win), INTENT(IN) :: win
```

## Exercise 22 (rightput)

Revisit exercise 16 and solve it using MPI\_Put

## Exercise 23 (randomput)

Write code where process 0 randomly writes in the window on 1 or 2.

## Exercise (optional) 24 (randomput )

Replace `MPI_Win_create` by `MPI_Win_allocate`.

C:

```
int MPI_Accumulate
  (const void *origin_addr, int origin_count,MPI_Datatype origin_datatype,
   int target_rank,MPI_Aint target_disp, int target_count,MPI_Datatype target_datatype,
   MPI_Op op, MPI_Win win)
int MPI_Raccumulate
  (const void *origin_addr, int origin_count,MPI_Datatype origin_datatype,
   int target_rank,MPI_Aint target_disp, int target_count,MPI_Datatype target_datatype,
   MPI_Op op, MPI_Win win,MPI_Request *request)
```

#### Input Parameters

origin\_addr : Initial address of buffer (choice).

origin\_count : Number of entries in buffer (nonnegative integer).

origin\_datatype : Data type of each buffer entry (handle).

target\_rank : Rank of target (nonnegative integer).

target\_disp : Displacement from start of window to beginning of target buffer.

target\_count : Number of entries in target buffer (nonnegative integer).

target\_datatype : Data type of each entry in target buffer (handle).

op : Reduce operation (handle).

win : Window object (handle).

#### Output Parameter

## Exercise (optional) 25 (countdown)

Implement a shared counter:

- One process maintains a counter;
- Iterate: all others at random moments update this counter.
- When the counter is zero, everyone stops iterating.

The problem here is data synchronization: does everyone see the counter the same way?

# Race conditions in one-sided communication

- One process stores a table of work descriptors, and a pointer to the first unprocessed descriptor;
- Each process reads the pointer, reads the corresponding descriptor, and increments the pointer; and
- A process that has read a descriptor then executes the corresponding task.

# Atomic operations

Race condition problem in read/write:  
result of MPI\_Get is only known after the fence.

```
int MPI_Fetch_and_op
  (const void *origin_addr, void *result_addr,
   MPI_Datatype datatype,
   int target_rank, MPI_Aint target_disp,
   MPI_Op op, MPI_Win win)
```

```
// passive.cxx
if (procno==repository) {
    // Repository processor creates a table of inputs
    // and associates that with the window
}
if (procno!=repository) {
    float contribution=(float)procno,table_element;
    int loc=0;
    MPI_Win_lock(MPI_LOCK_EXCLUSIVE,repository,0,the_window);
    // read the table element by getting the result from adding zero
    err = MPI_Fetch_and_op
        (&contribution,&table_element,MPI_FLOAT,
         repository,loc,MPI_SUM,the_window); CHK(err);
    MPI_Win_unlock(repository,the_window);
}
```

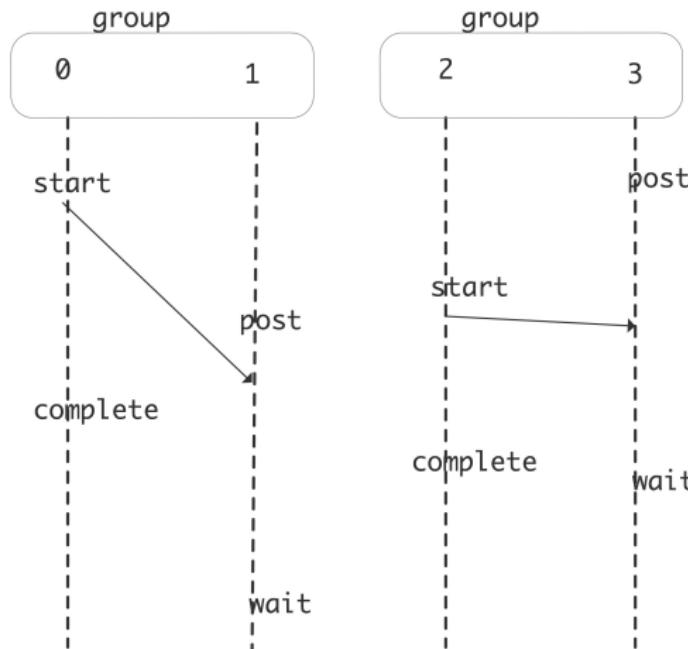
## Exercise (optional) 26

Redo exercise 25 using MPI\_Fetch\_and\_op. The problem is again to make sure all processes have the same view of the shared counter.

Does it work to make the fetch-and-op conditional? Is there a way to do it unconditionally? What should the ‘break’ test be, seeing that multiple processes can update the counter at the same time?

# A second active synchronization

Use Post, Wait, Start, Complete calls



More fine-grained than fences.

# Table of Contents

14

Active target synchronization

15

Passive target synchronization

16

More

# Passive target synchronization

Lock a window on the target:

```
MPI_Win_lock (int locktype, int rank, int assert, MPI_Win win)  
MPI_Win_unlock (int rank, MPI_Win win)
```

## Exercise 27 (onesidedbuild)

- Let each process have an empty array of sufficient length and a stack pointer that maintains the first free location.
- Now let each process randomly put data in a free location of another process' array.
- Use window locking. (Why is active target synchronization not possible?)

# Table of Contents

14 Active target synchronization

15 Passive target synchronization

16 More

# Shared memory interface

- Use `MPI_Comm_split_type` to find processes on the same shared memory
- Use `MPI_Win_allocate_shared` to create a window between processes on the same shared memory
- Use `MPI_Win_shared_query` to get pointer to another process' window data.
- You can now use `memcpy` instead of `MPI_Put`.

# Justification

MPI basic concepts suffice for many applications, especially on medium scale. Various advanced concepts let your code deal with unusual scenarios or very large scale runs.

# Derived Datatypes

# Overview

In this section you will learn about derived data types.

Commands learned:

- MPI\_Type\_contiguous/vector/indexed/struct  
MPI\_Type\_create\_subarray
- MPI\_Pack/Unpack
- F90 types

# Table of Contents

17 Discussion

18 Datatypes

19 Subarray type

20 Packed data

# Motivation: datatypes in MPI

All examples so far:

- contiguous buffer
- elements of single type

We need data structures with gaps, or heterogeneous types.

- Send real or imaginary parts out of complex array.
- Gather/scatter cyclicly.
- Send struct or Type data.

MPI allows for recursive construction of data types.

# Datatype topics

- Elementary types: built-in.
- Derived types: user-defined.
- Packed data: not really a datatype.

# Table of Contents

17 Discussion

18 Datatypes

19 Subarray type

20 Packed data

# Elementary datatypes

C/C++	Fortran
MPI_CHAR	MPI_CHARACTER
MPI_UNSIGNED_CHAR	
MPI_SIGNED_CHAR	MPI_LOGICAL
MPI_SHORT	
MPI_UNSIGNED_SHORT	
MPI_INT	MPI_INTEGER
MPI_UNSIGNED	
MPI_LONG	
MPI_UNSIGNED_LONG	
MPI_FLOAT	MPI_REAL
MPI_DOUBLE	MPI_DOUBLE_PRECISION
MPI_LONGDOUBLE	MPI_COMPLEX MPI_DOUBLE_COMPLEX

# How to use derived types

Create, commit, use, free:

```
MPI_datatype newtype;  
MPI_Type_xxx( ... oldtype ... &newtype);  
MPI_Type_commit ( &newtype );  
  
// code using the new type  
  
MPI_Type_free ( &newtype );
```

The `oldtype` can be elementary or derived.

Recursively constructed types.

# Contiguous type

```
int MPI_Type_contiguous(  
    int count, MPI_Datatype old_type, MPI_Datatype *new_type_p)
```



This one is indistinguishable from just sending `count` instances of the `old_type`.

# Example: non-contiguous data

Matrix in column storage:

- Columns are contiguous
- Rows are not contiguous

Logical:

(1,1)	(1,2)	
(2,1)		
(3,1)		

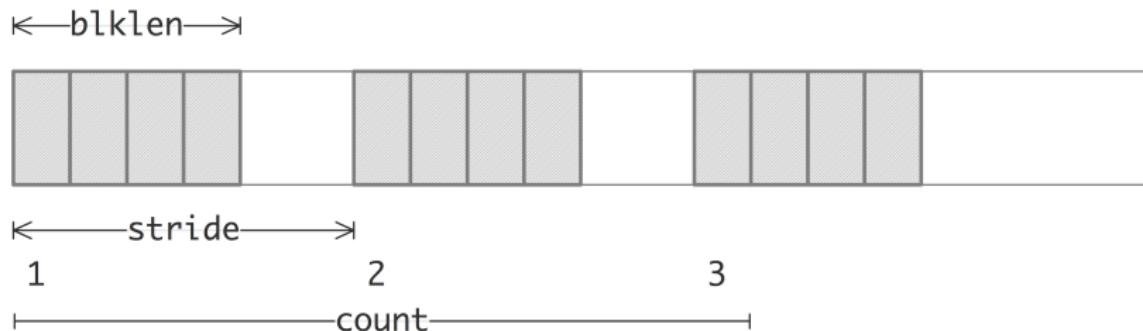


Physical:

(1,1)	(2,1)	(3,1)	...	(1,2)	...
-------	-------	-------	-----	-------	-----

# Vector type

```
int MPI_Type_vector(  
    int count, int blocklength, int stride,  
    MPI_Datatype old_type, MPI_Datatype *newtype_p  
) ;
```



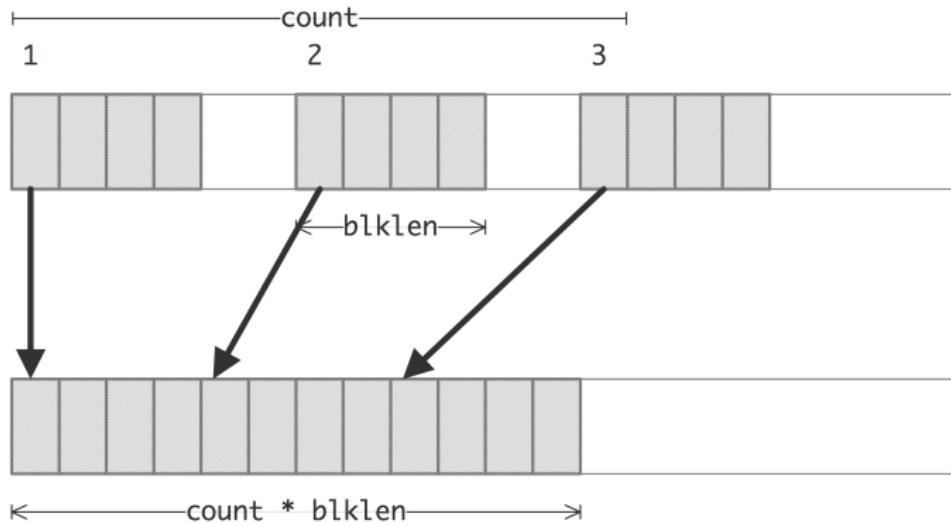
Used to pick a regular subset of elements from an array.

```
// vector.c
source = (double*) malloc(stride*count*sizeof(double));
target = (double*) malloc(count*sizeof(double));
MPI_Datatype newvectortype;
if (procno==sender) {
    MPI_Type_vector(count,1,stride,MPI_DOUBLE,&newvectortype);
    MPI_Type_commit(&newvectortype);
    MPI_Send(source,1,newvectortype,the_other,0,comm);
    MPI_Type_free(&newvectortype);
} else if (procno==receiver) {
    MPI_Status recv_status;
    int recv_count;
    MPI_Recv(target,count,MPI_DOUBLE,the_other,0,comm,
             &recv_status);
    MPI_Get_count(&recv_status,MPI_DOUBLE,&recv_count);
    ASSERT(recv_count==count);
}
```

# Different send and receive types

Sender type: vector

receiver type: contiguous or elementary



Receiver has no knowledge of the stride of the sender.

## Exercise 28 (stridesend)

Let processor 0 have an array  $x$  of length  $10P$ , where  $P$  is the number of processors. Elements  $0, P, 2P, \dots, 9P$  should go to processor zero,  $1, P+1, 2P+1, \dots$  to processor 1, et cetera. Code this as a sequence of send/recv calls, using a vector datatype for the send, and a contiguous buffer for the receive.

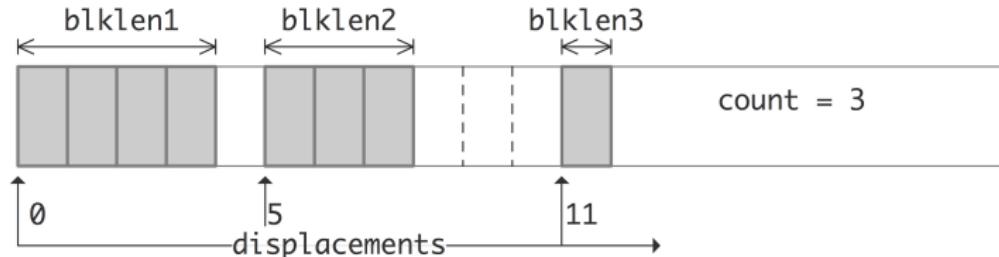
For simplicity, skip the send to/from zero. What is the most elegant solution if you want to include that case?

For testing, define the array as  $x[i] = i$ .

## Exercise 29

Allocate a matrix on processor zero, using Fortran column-major storage.  
Using  $P$  sendrecv calls, distribute the rows of this matrix among the  
processors.

# Indexed type

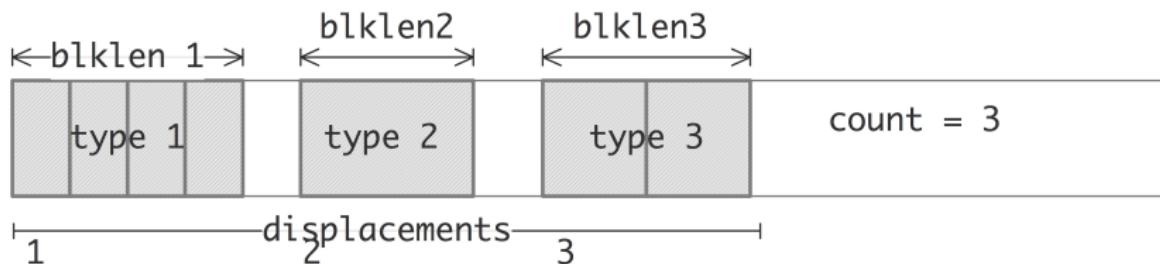


```
int MPI_Type_indexed(  
    int count, int blocklens[], int displacements[],  
    MPI_Datatype old_type, MPI_Datatype *newtype);
```

Also hindexed with byte offsets.

# Heterogeneous: Structure type

```
int MPI_Type_create_struct(  
    int count, int blocklengths[], MPI_Aint displacements[],  
    MPI_Datatype types[], MPI_Datatype *newtype);
```



This gets very tedious...

# Table of Contents

17 Discussion

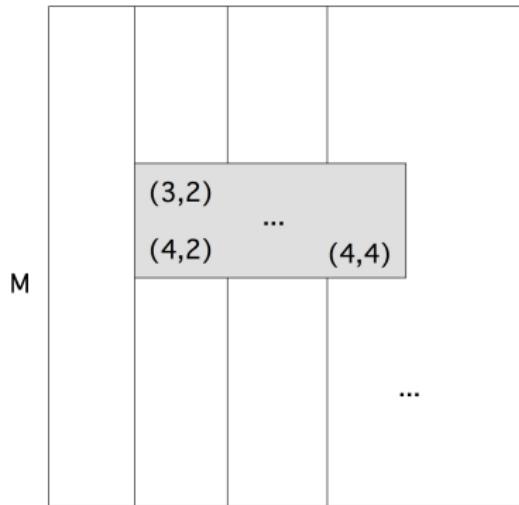
18 Datatypes

19 Subarray type

20 Packed data

# Submatrix storage

Logical:



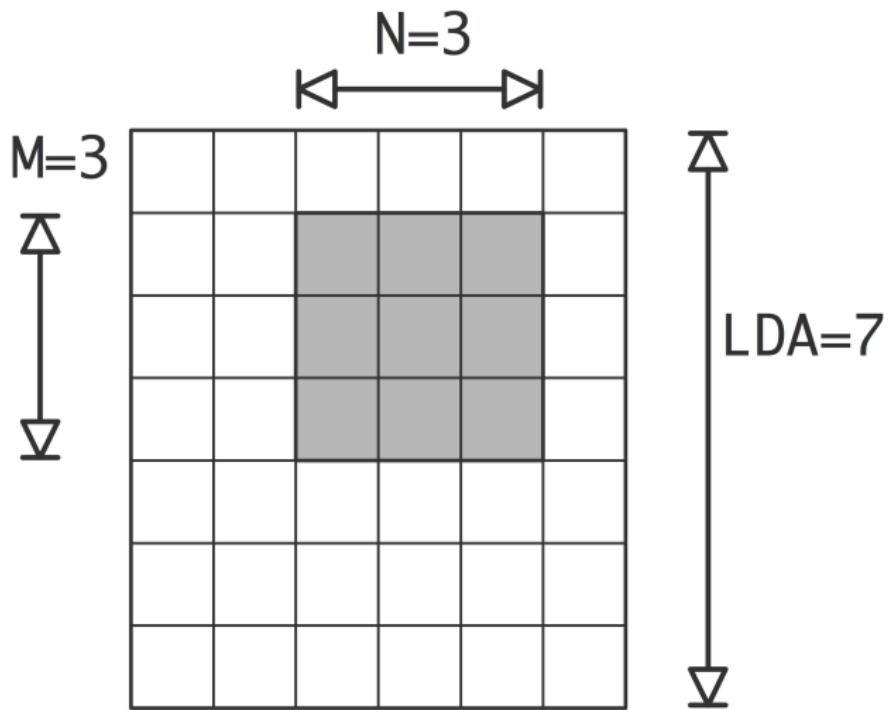
Physical:



- Location of first element
- Stride, blocksize

# BLAS/Lapack storage

Three parameter description:



## Subarray type

- Vector type is convenient for 2D subarrays,
- it gets tedious in higher dimensions.
- Better solution: MPI\_Type\_create\_subarray

```
MPI_TYPE_CREATE_SUBARRAY(  
    ndims, array_of_sizes, array_of_subsizes,  
    array_of_starts, order, oldtype, newtype)
```

Subtle: data does not start at the buffer start

## Exercise 30 (cubegather)

Assume that your number of processors is  $P = Q^3$ , and that each process has an array of identical size. Use `MPI_Type_create_subarray` to gather all data onto a root process. Use a sequence of send and receive calls; `MPI_Gather` does not work here.

This is one of the rare cases where you should use `ibrun -np 27` and such: normally you use `ibrun` without process count argument.

# Fortran 'kind' types

Check out MPI\_Type\_create\_f90\_integer MPI\_Type\_create\_f90\_real  
MPI\_Type\_create\_f90\_complex

Example:

```
REAL ( KIND = SELECTED_REAL_KIND(15 ,300) ) , &
DIMENSION(100) :: array
CALL MPI_Type_create_f90_real( 15 , 300 , realtype , error )
```

# Table of Contents

17 Discussion

18 Datatypes

19 Subarray type

20 Packed data

# Packing into buffer

```
int MPI_Pack(  
    void *inbuf, int incount, MPI_Datatype datatype,  
    void *outbuf, int outcount, int *position,  
    MPI_Comm comm);  
  
int MPI_Unpack(  
    void *inbuf, int insize, int *position,  
    void *outbuf, int outcount, MPI_Datatype datatype,  
    MPI_Comm comm);
```

## Example

```
// pack.c
if (procno==sender) {
    MPI_Pack(&nsends,1,MPI_INT,buffer,buflen,&position,comm);
    for (int i=0; i<nsends; i++) {
        double value = rand()/(double)RAND_MAX;
        MPI_Pack(&value,1,MPI_DOUBLE,buffer,buflen,&position,comm);
    }
    MPI_Pack(&nsends,1,MPI_INT,buffer,buflen,&position,comm);
    MPI_Send(buffer,position,MPI_PACKED,other,0,comm);
} else if (procno==receiver) {
    int irecv_value;
    double xrecv_value;
    MPI_Recv(buffer,buflen,MPI_PACKED,other,0,comm,MPI_STATUS_IGNORE);
    MPI_Unpack(buffer,buflen,&position,&nsends,1,MPI_INT,comm);
    for (int i=0; i<nsends; i++) {
        MPI_Unpack(buffer,buflen,&position,&xrecv_value,1,MPI_DOUBLE,comm);
    }
    MPI_Unpack(buffer,buflen,&position,&irecv
    ASSERT(irecv_value==nsends);
```

# Communicator manipulations

# Overview

In this section you will learn about various subcommunicators.

Commands learned:

- MPI\_Comm\_dup, discussion of library design
- MPI\_Comm\_split
- discussion of groups
- discussion of inter/intra communicators.

# Sub-computations

Simultaneous groups of processes, doing different tasks, but loosely interacting:

- Simulation pipeline: produce input data, run simulation, post-process.
- Climate model: separate groups for air, ocean, land, ice.
- Quicksort: split data in two, run quicksort independently on the halves.
- Process grid: do broadcast in each column.

New communicators are formed recursively from `MPI_COMM_WORLD`.

# Communicator duplication

Simplest new communicator: identical to a previous one.

```
int MPI_Comm_dup(MPI_Comm comm, MPI_Comm *newcomm)
```

This is useful for library writers:

```
MPI_Isend(...); MPI_Irecv(...);
// library call
MPI_Waitall(...);
```

# Use of a library

```
library my_library(comm);
MPI_Isend(&sdata,1,MPI_INT,other,1,comm,&(request[0]));
my_library.communication_start();
MPI_Irecv(&rdata,1,MPI_INT,other,MPI_ANY_TAG,
          comm,&(request[1]));
MPI_Waitall(2,request,status);
my_library.communication_end();
```

# Use of a library

```
int library::communication_start() {  
    int sdata=6,rdata;  
    MPI_Isend(&sdata,1,MPI_INT,other,2,comm,&(request[0]));  
    MPI_Irecv(&rdata,1,MPI_INT,other,MPI_ANY_TAG,  
              comm,&(request[1]));  
    return 0;  
}  
  
int library::communication_end() {  
    MPI_Status status[2];  
    MPI_Waitall(2,request,status);  
    return 0;  
}
```

# Wrong way

```
// commdup_wrong.cxx
class library {
private:
    MPI_Comm comm;
    int procno, nprocs, other;
    MPI_Request *request;
public:
    library(MPI_Comm incomm) {
        comm = incomm;
        MPI_Comm_rank(comm, &procno);
        other = 1 - procno;
        request = new MPI_Request[2];
    };
    int communication_start();
    int communication_end();
};
```

# Right way

```
// commdup_right.cxx
class library {
private:
    MPI_Comm comm;
    int procno, nprocs, other;
    MPI_Request *request;
public:
    library(MPI_Comm incomm) {
        MPI_Comm_dup(incomm, &comm);
        MPI_Comm_rank(comm, &procno);
        other = 1 - procno;
        request = new MPI_Request[2];
    };
    ~library() {
        MPI_Comm_free(&comm);
    }
    int communication_start();
    int communication_end();
};
```

# Disjoint splitting

Split a communicator in multiple disjoint others.

Give each process a ‘colour’, group processes by colour:

```
int MPI_Comm_split(MPI_Comm comm, int color, int key,  
                    MPI_Comm *newcomm)
```

(key determines ordering: use rank unless you want special effects)

## Row/column example

```
MPI_Comm_rank( MPI_COMM_WORLD, &procno );
proc_i = procno % proc_column_length;
proc_j = procno / proc_column_length;

MPI_Comm column_comm;
MPI_Comm_split( MPI_COMM_WORLD, proc_j, procno, &column_comm );

MPI_Bcast( data, ... column_comm );
```

## Exercise 31 (procgrid)

Organize your processes in a grid, and make subcommunicators for the rows and columns. For this compute the row and column number of each process.

In the row and column communicator, compute the rank. For instance, on a  $2 \times 3$  processor grid you should find:

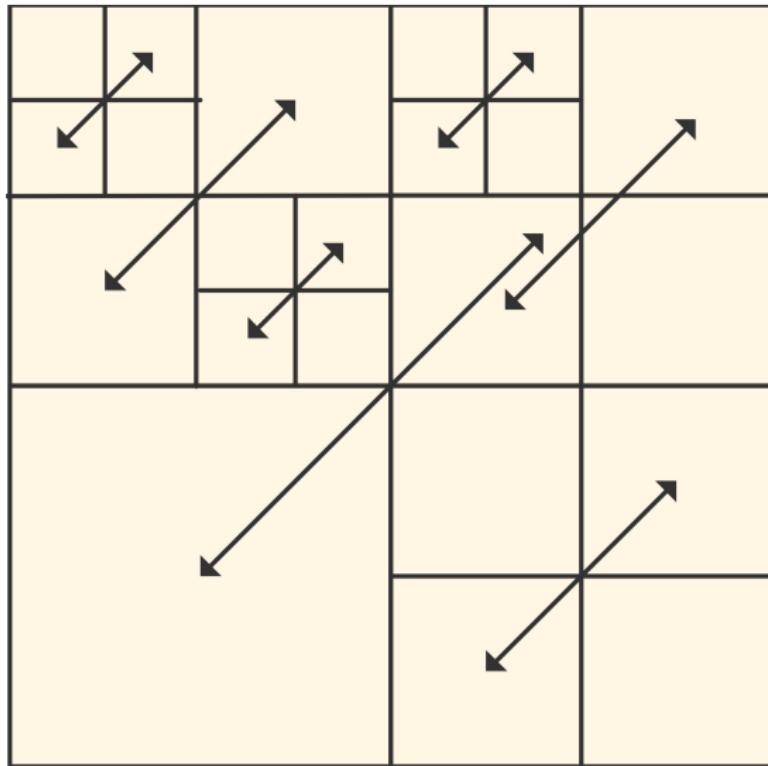
Global ranks:	Ranks in row:	Ranks in column:
0 1 2	0 1 2	0 0
3 4 5	0 1 2	1 1

Check that the rank in the row communicator is the column number, and the other way around.

Run your code on different number of processes, for instance a number of rows and columns that is a power of 2, or that is a prime number. This is one occasion where you could use `ibrun -np 9`; normally you would *never* put a processor count on `ibrun`.

## Exercise 32

Implement a recursive algorithm for matrix transposition:



## More

- Non-disjoint subcommunicators through process groups.
- Intra-communicators and inter-communicators.
- Process topologies: cartesian and graph.

# MPI File I/O

# Overview

This section discusses parallel I/O. What is the problem with regular I/O in parallel?

Commands learned:

- MPI\_File\_open/write/close
- parallel file pointer routines: MPI\_File\_set\_view/write\_at

# The trouble with parallel I/O

- Multiple process reads from one file: no problem.
- Multiple writes to one file: big problem.
- Everyone writes to separate file: stress on the file system, and requires post-processing.

# MPI I/O

- Part of MPI since MPI-2
- Joint creation of one file from bunch of processes.
- You could also use hdf5, netcdf, silo ...

# The usual bits

```
MPI_File mpifile;
MPI_File_open(comm, "blockwrite.dat",
              MPI_MODE_CREATE | MPI_MODE_WRONLY, MPI_INFO_NULL,
              &mpifile);
if (procno==0) {
    MPI_File_write
        (mpifile, output_data, nwords, MPI_INT, MPI_STATUS_IGNORE);
}
MPI_File_close(&mpifile);
```

# How do you make it unique for a process?

```
MPI_File_write_at  
  (mpifile, offset, output_data, nwords,  
   MPI_INT, MPI_STATUS_IGNORE);
```

or

```
MPI_File_set_view  
  (mpifile,  
   offset, datatype,  
   MPI_INT, "native", MPI_INFO_NULL);  
MPI_File_write // no offset, we have a view  
  (mpifile, output_data, nwords, MPI_INT, MPI_STATUS_IGNORE);
```

## Exercise 33 (blockwrite)

The given code works for one writing process. Compute a unique offset for each process (in bytes!) so that all the local arrays are placed in the output file in sequence.

## Exercise 34 (viewwrite)

Solve the previous exercise by using `MPI_File_write` (that is, without offset), but by using `MPI_File_set_view` to specify the location.

## Exercise 35 (scatterwrite)

Now write the local arrays cyclically to the file: with 5 processes and 3 elements per process the file should contain

```
1 4 7 10 13 | 2 5 8 11 14 | 3 6 9 12 15
```

Do this by defining a vector derived type and setting that as the file view.

# Process management

# Overview

This section discusses processes management; intra communicators.

Commands learned:

- MPI\_Comm\_spawn, MPI\_COMM\_UNIVERSE
- MPI\_Comm\_get\_parent, MPI\_Comm\_remote\_size

# Process management

- PVM was a precursor of MPI: could dynamically create new processes.
- It took MPI a while to catch up.
- MPI\_COMM\_UNIVERSE: space for creating more processes outside MPI\_COMM\_WORLD.
- New processes have their own MPI\_COMM\_WORLD.
- Communication between the two communicators: ‘inter communicator’ (the old time is ‘intra communicator’)

# Space for processes

Probably a machine dependent component.

Intel MPI at TACC:

```
MY_MPIRUN_OPTIONS="-usize 8" ibrun -np 4 spawn_manager
```

Discover size of the universe:

```
MPI_Attr_get(MPI_COMM_WORLD, MPI_UNIVERSE_SIZE,  
             (void*)&universe_sizep, &flag);
```

# Manager program

```
MPI_Attr_get(MPI_COMM_WORLD, MPI_UNIVERSE_SIZE,
              (void*)&universe_sizep, &flag);
MPI_Comm everyone; /* intercommunicator */
int nworkers = universe_size-world_size;
MPI_Comm_spawn(worker_program, /* executable */
               MPI_ARGV_NULL, nworkers,
               MPI_INFO_NULL, 0, MPI_COMM_WORLD, &everyone,
               errorcodes);
```

# Worker program

```
MPI_Comm_size(MPI_COMM_WORLD,&nworkers);  
MPI_Comm parent;  
MPI_Comm_get_parent(&parent);  
MPI_Comm_remote_size(parent, &remotesize);
```

# Process topologies

# Overview

This section discusses graph topologies.

Commands learned:

- MPI\_Dist\_graph\_create, MPI\_DIST\_GRAPH
- MPI\_Neighbor\_...

# Process topologies

- Processes don't communicate at random
- Example: Cartesian grid, each process 4 (or so) neighbours
- MPI can optimize for locality
- Express operations in terms of topology.

# Illustration

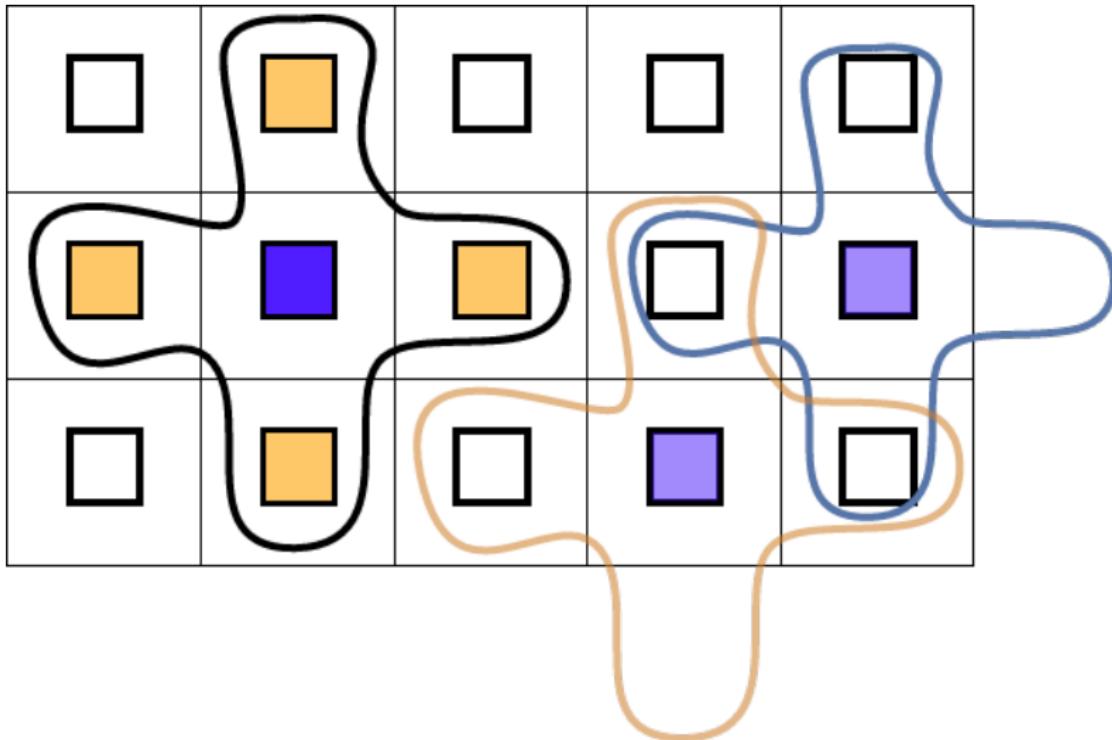


Illustration of a distributed graph topology where each node has four neighbours

# Create graph topology

```
int MPI_Dist_graph_create  
  (MPI_Comm comm_old, int n, const int sources[],  
   const int degrees[], const int destinations[], const int we-  
   MPI_Info info, int reorder,  
   MPI_Comm *comm_dist_graph)
```

- **n**sources **how many processes described?** (Usually 1)
- **sources** **the processes being described** (Usually `MPI_Comm_rank` value)
- **degrees** **how many processes to send to**
- **destinations** **their ranks**
- **weights**: just set to  $\equiv 1$
- **info**: `MPI_INFO_NULL` will do
- **reorder**: 1 if dynamically reorder processes

# Neighbourhood collectives

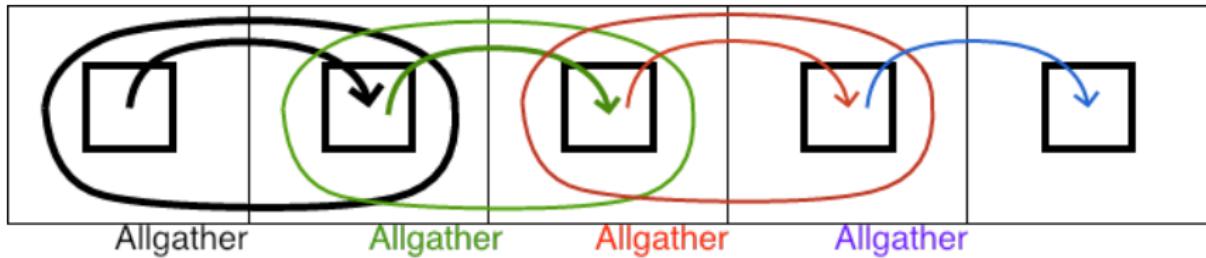
```
int MPI_Neighbor_allgather  
  (const void *sendbuf, int sendcount, MPI_Datatype sendtype,  
   void *recvbuf, int recvcount, MPI_Datatype recvtype,  
   MPI_Comm comm)
```

Like an ordinary Allgather, but  
the receive buffer has a length degree.

## Exercise 36 (rightgraph)

Revisit exercise 16 and solve it using `MPI_Dist_graph_create`. Use figure 221 for inspiration.

# Inspiring picture for the previous exercise



Solving the right-send exercise with neighbourhood collectives