HPC Programming

Message Passing Interface (MPI), Part III

Peter-Bernd Otte, 4.12.2018

Introduction MPI

1. Overview / Getting Started
2. Messages & Point-to-point Communication
3. Nonblocking Communication
4. Error Handling
5. Groups & Communicators
6. Collective Communication
7. Dealing with I/O
8. MPI Derived Datatypes
9. Common pitfalls and good practice ("need for speed")

MPI: different communications modes

Recap

	Blocking	Non-Blocking	note
standard send	MPI_Send	MPI_ISend	synchronous or asynchronous send (depending on message size and implementation) uses internal buffer.
synchronous send	MPI_SSend	MPI_ISSend	Only completes when the receive has started
asynchronous (buffered) send	MPI_BSend	MPI_IBSend	Completes after buffer copy (always).
ready send	MPI_RSend	MPI_IRSend	problematic: mandatory to have matching receive already listening. Not discussed in this lecture. Might be fastest solution.

„i" stands for immediate return

	Blocking	Non-Blocking	note
standard receive	MPI_Recv	MPI_IRecv	works for all sending routines.

MPI: P2P communications, Pros and Cons

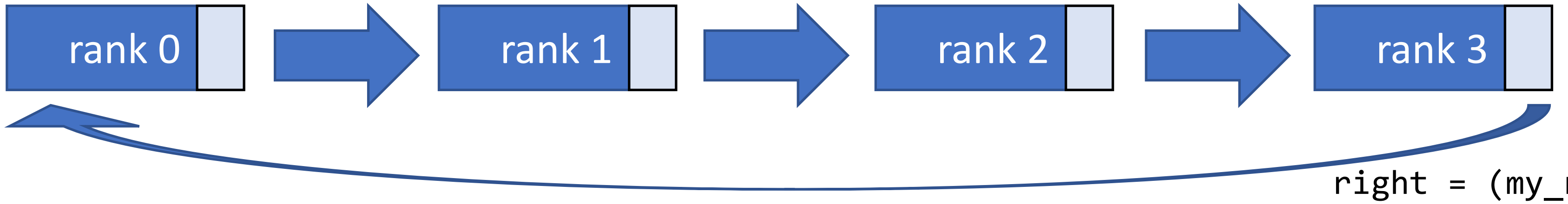
Recap

- synchronous send
  - risk of serialisation, waiting and/or deadlock
  - high latency but best bandwidth
- asynchronous send
  - no risks (except: take care of your buffers)
  - low latency but bad bandwidth
- standard send
  - risk of implementation and message dependence behaviour
  - plus risks of synchronous send

MPI: Non-Blocking Send & Receive

Recap

- to a 1D ring with 1 piece of data passing in one direction




```
right = (my_rank+1) % size;
left = (my_rank-1+size) % size;
```

- cyclic: MPI_Send(…to right…)
  MPI_Recv(…from left…)

deadlock!
All are waiting for a receiver

- non-cyclic: for rank<size-2: MPI_Send(…to right…)
  for rank>0: MPI_Recv(…from left…)

serialisation!
highest rank starts, rank 0 last

(hint: all this only true if MPI calls are synchronous sends)

MPI: Non-Blocking communication

Recap

This can be accomplished by:

- non-blocking send
    1. MPI_Isend();
    2. Different_Work();
    3. MPI_Wait(); //Waits until MPI_Isend completed / send buffer is read out

- non-blocking receive
    1. MPI_Irecv();
    2. Different_Work();
    3. MPI_Wait(); //Waits until MPI_Irecv completed / receive buffer is filled

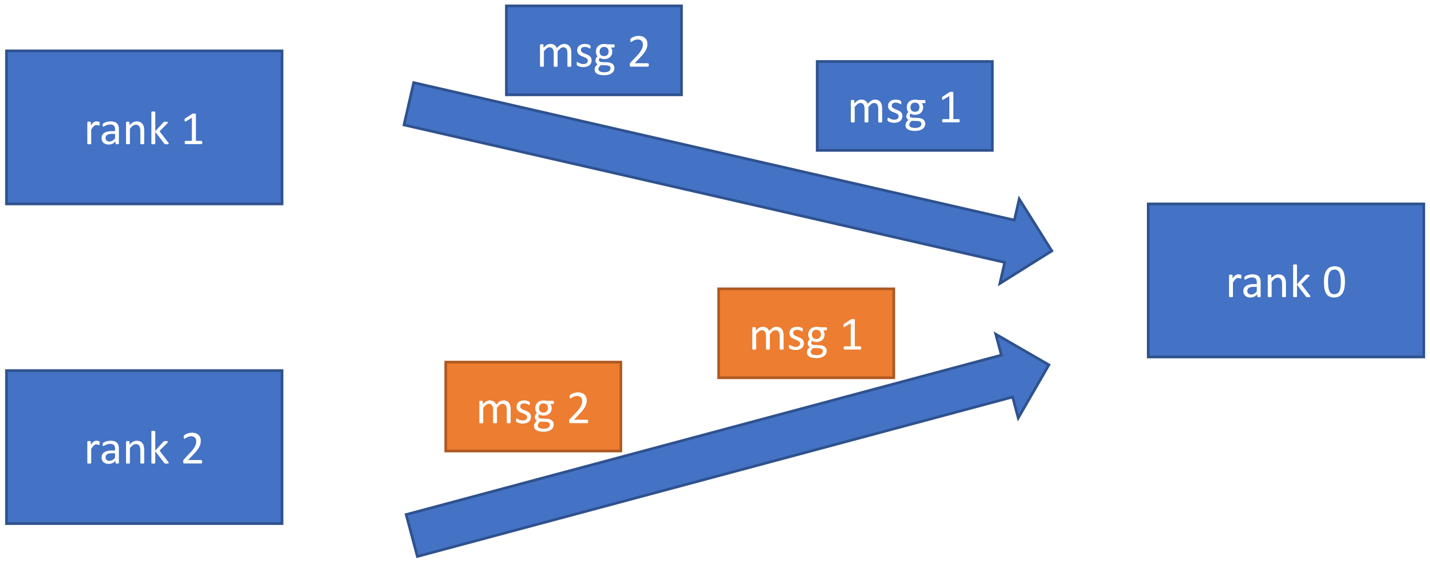
Golden MPI rule:
always <=3 lines of
MPI_* calls per task

otherwise:
check MPI reference or
wrong coding

MPI: Message Order Preservation

Recap

- Messages do not overtake, if same:
  - communicator (eg MPI_COMM_WORLD),
  - source rank and
  - destination rank
- true for: synchronous and asynchronous communications
- messages from different senders can overtake



Introduction MPI

1. Overview / Getting Started
2. Messages & Point-to-point Communication
3. Nonblocking Communication
4. Error Handling
5. Groups & Communicators
6. Collective Communication
7. Dealing with I/O
8. MPI Derived Datatypes
9. Common pitfalls and good practice (“need for speed”)

MPI: Error handling

- in short, standard behaviour:
  - MPI: abort on error
  - MPI-IO: continue and just report
  - only if error is detected by MPI, otherwise unpredictable behaviour

- in detail:
  - most important foundation: hardware error free
  - CPU, RAM & network
    - have different techniques to detect hardware errors (eg ECC-RAM, checksums in network packages)
    - you (or your system admin) are informed if hardware problem occurs
  - Change standard behaviour:
    int MPI_Comm_set_errhandler(MPI_Comm comm, MPI_Errhandler errhandler)
    int MPI_File_set_errhandler(MPI_File file, MPI_Errhandler errhandler)

Introduction MPI

1. Overview / Getting Started
2. Messages & Point-to-point Communication
3. Nonblocking Communication
4. Error Handling
5. Groups & Communicators
6. Collective Communication
7. Dealing with I/O
8. MPI Derived Datatypes
9. Common pitfalls and good practice (“need for speed”)

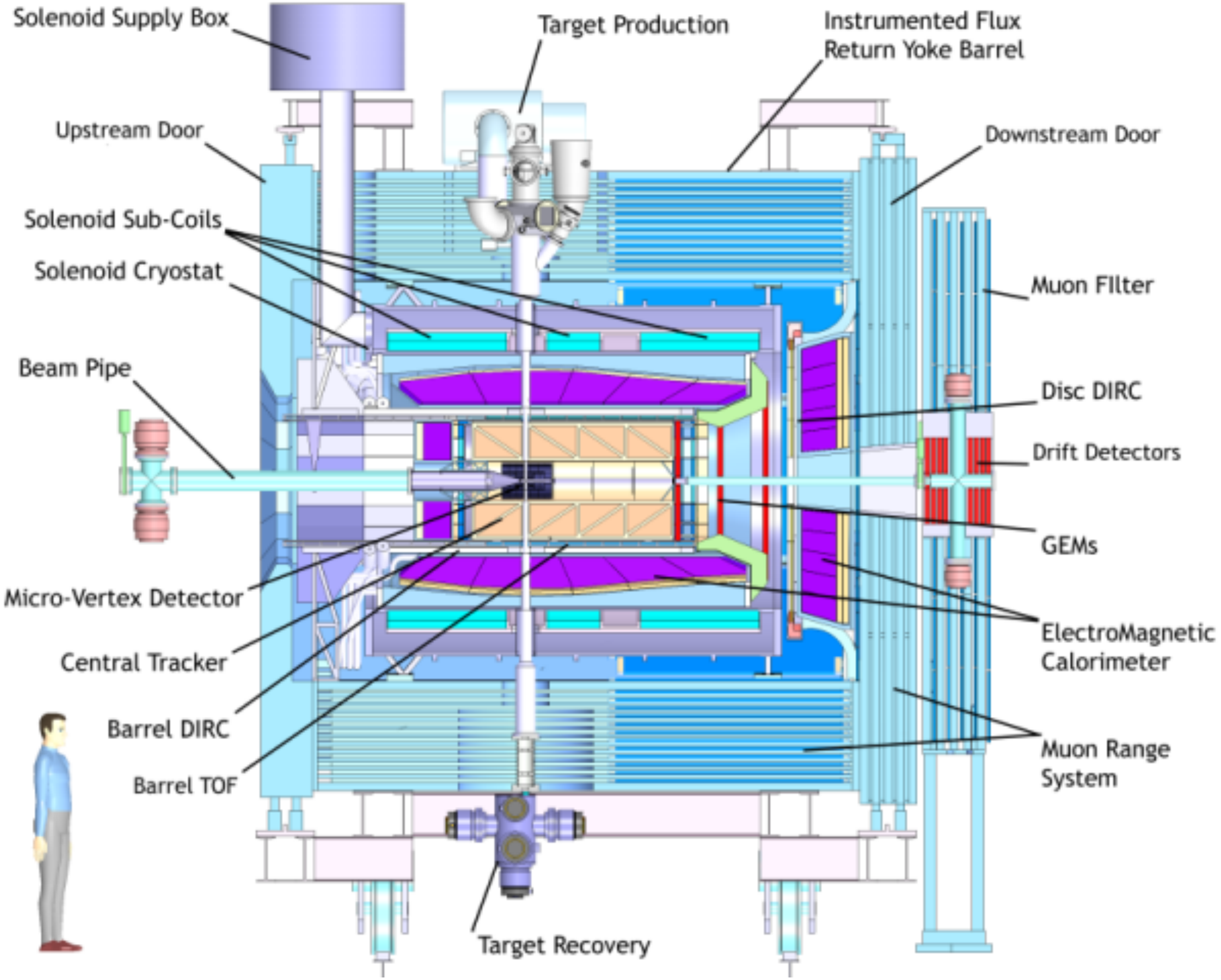
Motivation: Sub-Communicators

- Particle reconstruction
- Multiple layers in detector

- Multiple ranks working in several groups
  - code readability
  - collective communication within group

- OR: a library should NEVER use MPI_COMM_WORLD to not mix up with the main program.

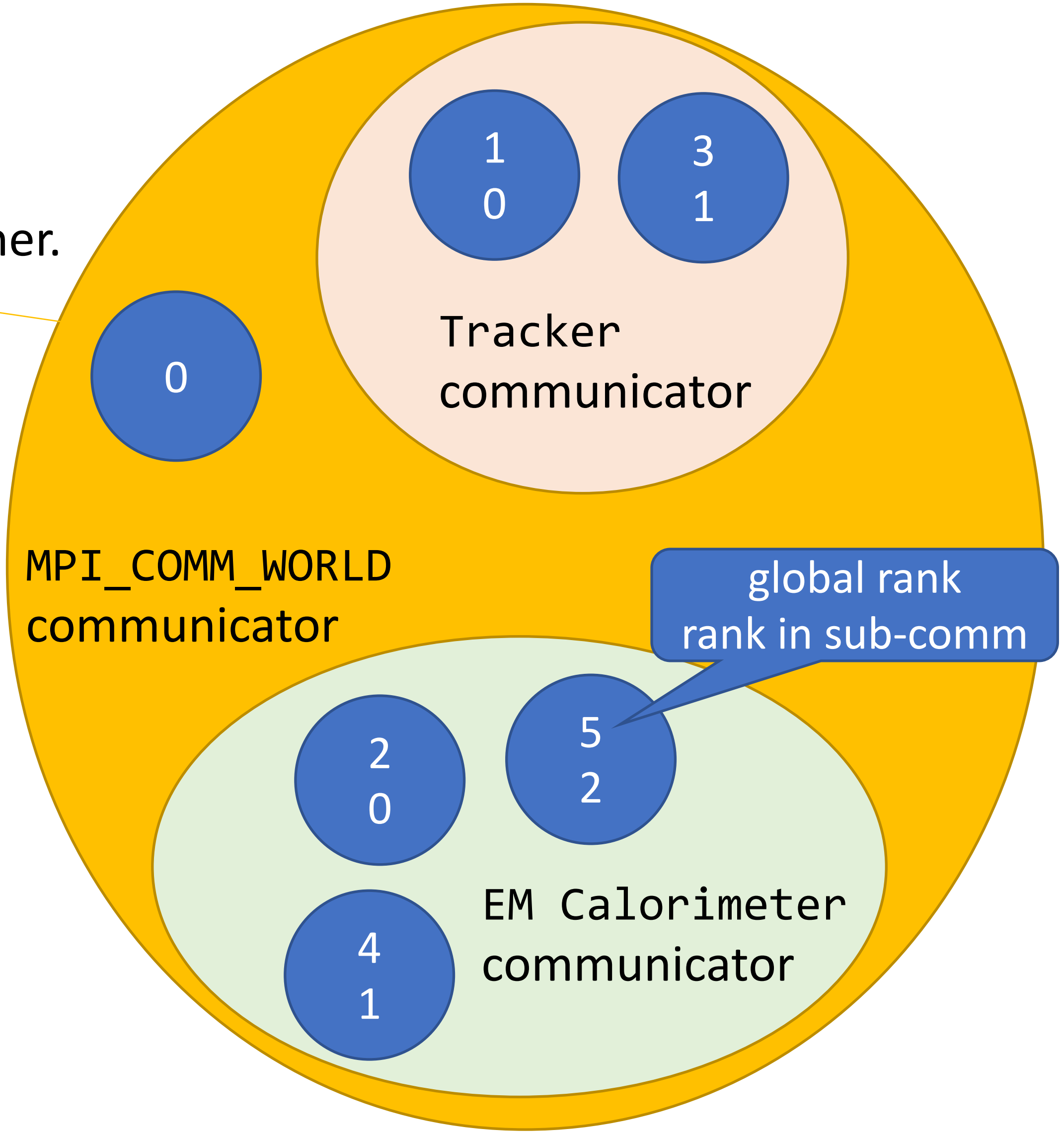
- See Exercise 5



MPI: Sub-Communicators

- MPI Communicator
  = group of processes that can send messages to each other.
- All processes are in `MPI_COMM_WORLD` communicator
- Defining sub groups (eg readability, library):
  1. MPI_Comm_split
  2. MPI_Comm_group + MPI_Comm_create

- Number of members and size in communicator:
  `MPI_Comm_size, MPI_Comm_rank`

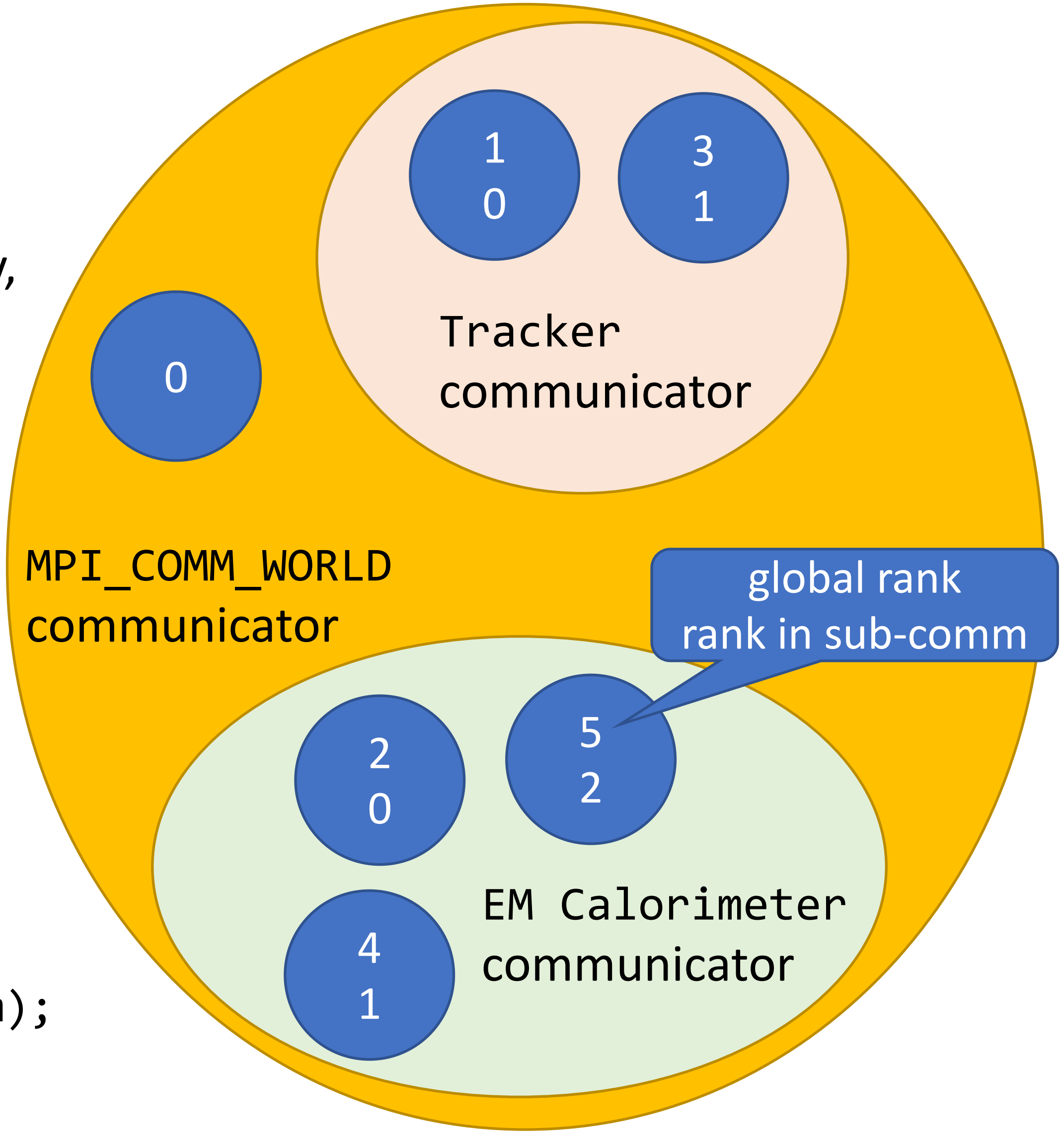


MPI: MPI_Comm_split

- Creates new communicators based on colors
- int MPI_Comm_split(MPI_Comm comm, int color, int key, MPI_Comm *newcomm)
  - ordering in new group:
    - key == 0 → as sorted in old
    - key != 0 → according to key values
  - one member group: color = MPI_UNDEFINED

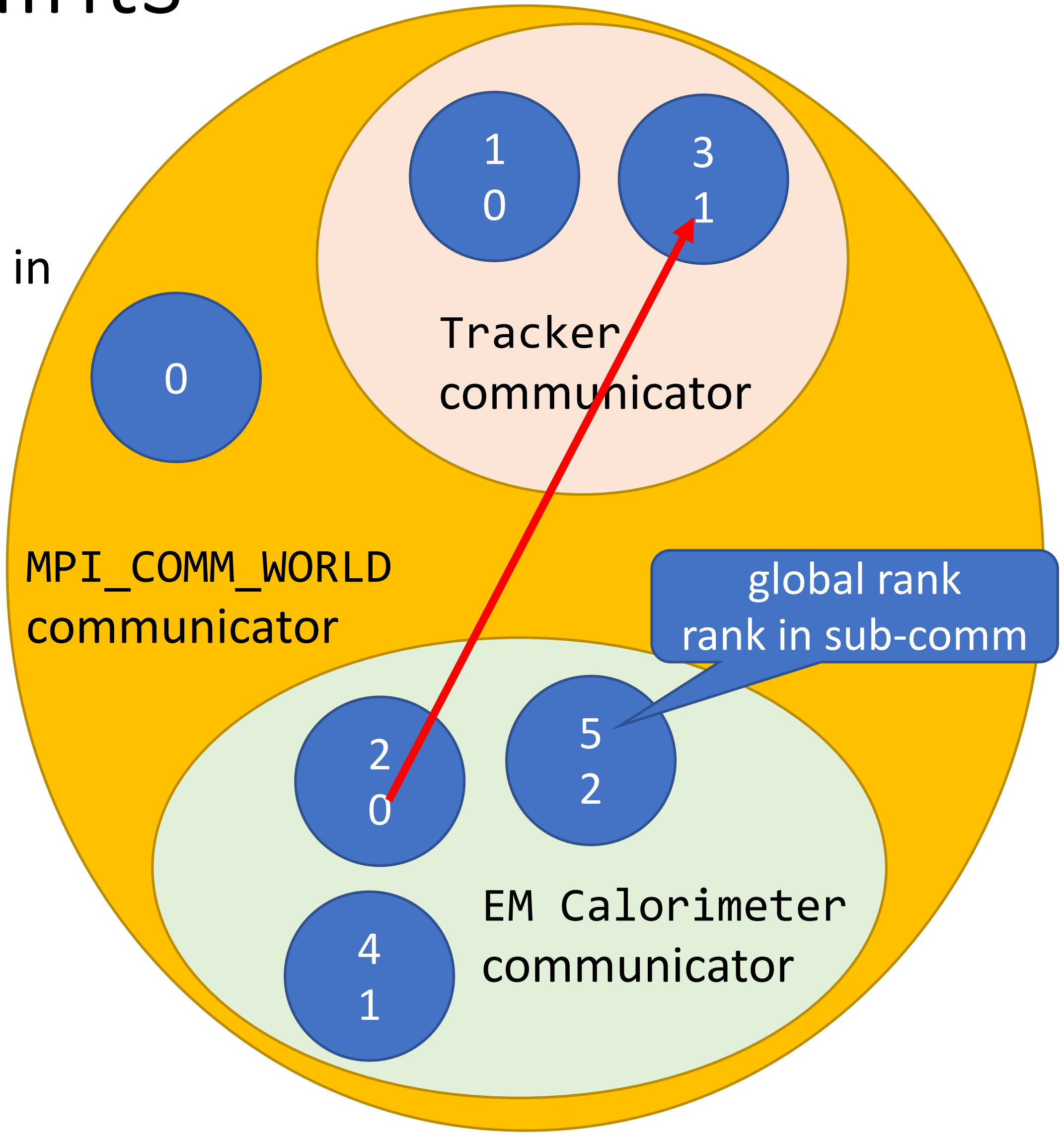
- Example:

```
MPI_Comm newcomm;
MPI_Comm_rank(MPI_COMM_WORLD, &my_rank);
mycolor = my_rank/3;
MPI_Comm_split(MPI_COMM_World, mycolor, 0, &newcomm);
MPI_Comm_rank(newcomm, &my_new_rank);
```



MPI: Sub-Communicators hints

- no difference in speed: same hardware
- Use intra-communicators to communicate between rank in different “worlds” (without MPI_COMM_WORLD ranks)
  - eg MPI_Intercomm_create()



Introduction MPI

1. Overview / Getting Started
2. Messages & Point-to-point Communication
3. Nonblocking Communication
4. Error Handling
5. Groups & Communicators
6. Collective Communication
7. Dealing with I/O
8. MPI Derived Datatypes
9. Common pitfalls and good practice (“need for speed”)

Motivation: Collective Communication

- eg matrix multiplication, helpful:
    - reading and spreading of data,
    - gather final results

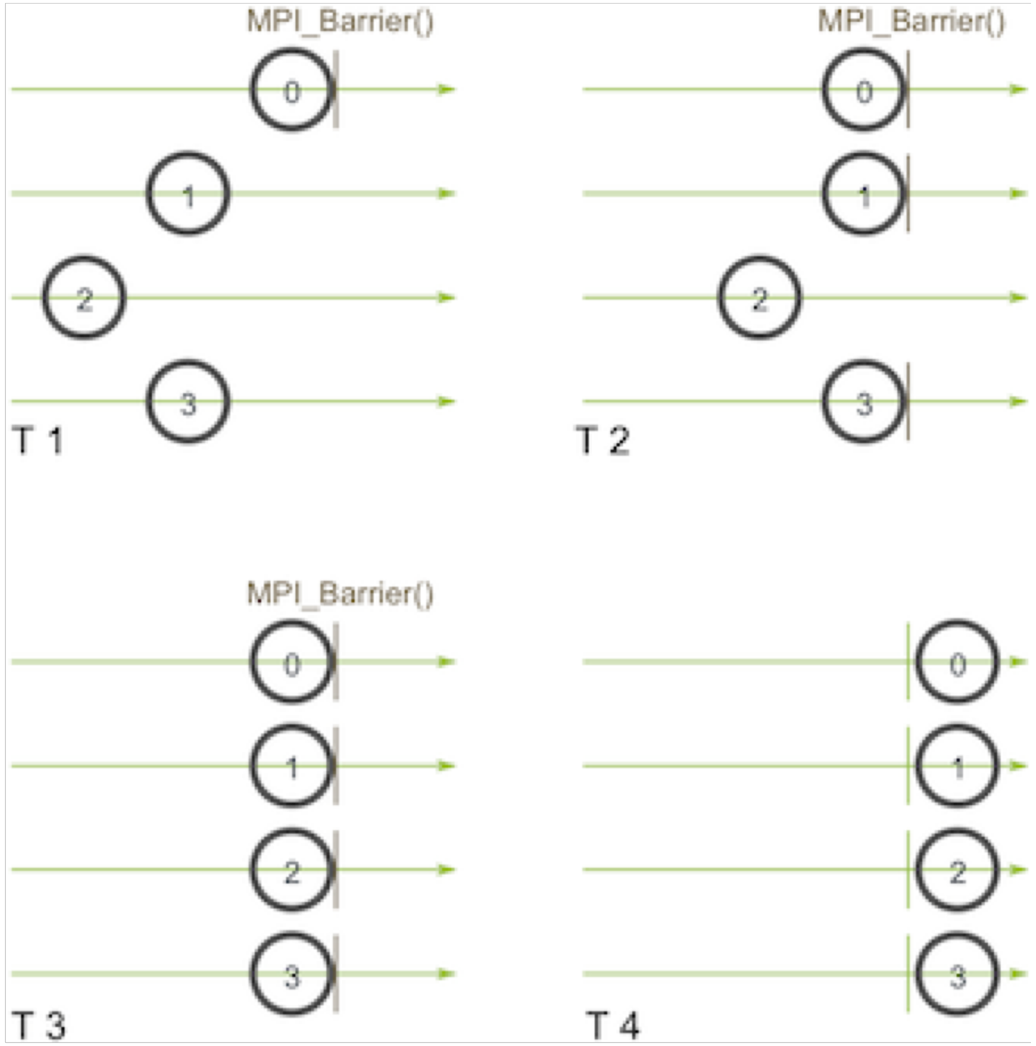
$$
\begin{bmatrix}
a_{00} & a_{01} & \cdots & a_{0,n-1} \\
a_{10} & a_{11} & \cdots & a_{1,n-1} \\
\vdots & \vdots & & \vdots \\
a_{i0} & a_{i1} & \cdots & a_{i,n-1} \\
\vdots & \vdots & & \vdots \\
a_{m-1,0} & a_{m-1,1} & \cdots & a_{m-1,n-1}
\end{bmatrix}
\begin{bmatrix}
x_0 \\ x_1 \\ \vdots \\ x_{n-1}
\end{bmatrix}
=
\begin{bmatrix}
y_0 \\ y_1 \\ \vdots \\ y_i = a_{i0}x_0 + a_{i1}x_1 + \cdots a_{i,n-1}x_{n-1} \\ \vdots \\ y_{m-1}
\end{bmatrix}
$$

MPI: MPI_Barrier

- collective communication: always include a *synchronization point* among processes.
  - all processes must reach a point in their code before they can all begin executing again.

syntax:

MPI_Barrier(MPI_Comm comm);

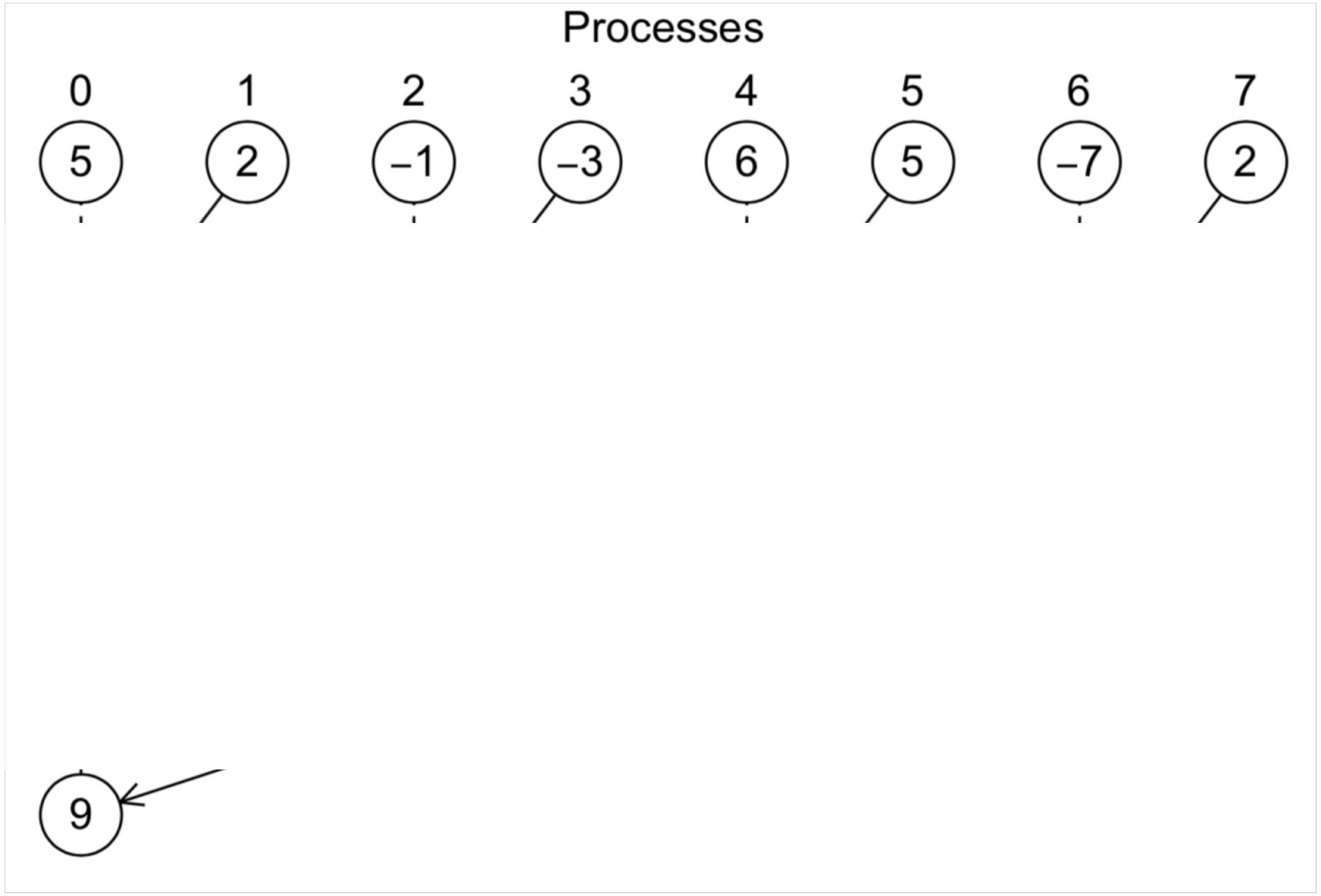


MPI: MPI_Reduction

- Reduces values on all processes to a single value
  (eg global sum)

```
int MPI_Reduce(
void *sendbuf /*in*/,
void *recvbuf /*out*/,
int count /*in*/,
MPI_Datatype datatype /*in*/,
MPI_Op operator /*in*/,
int dest_process /*in*/,
MPI_Comm comm /*in*/)
```

- hints:
  - with count>1, MPI can operate on arrays
  - sendbuf and recvbuf need to different
    (no aliasing!)

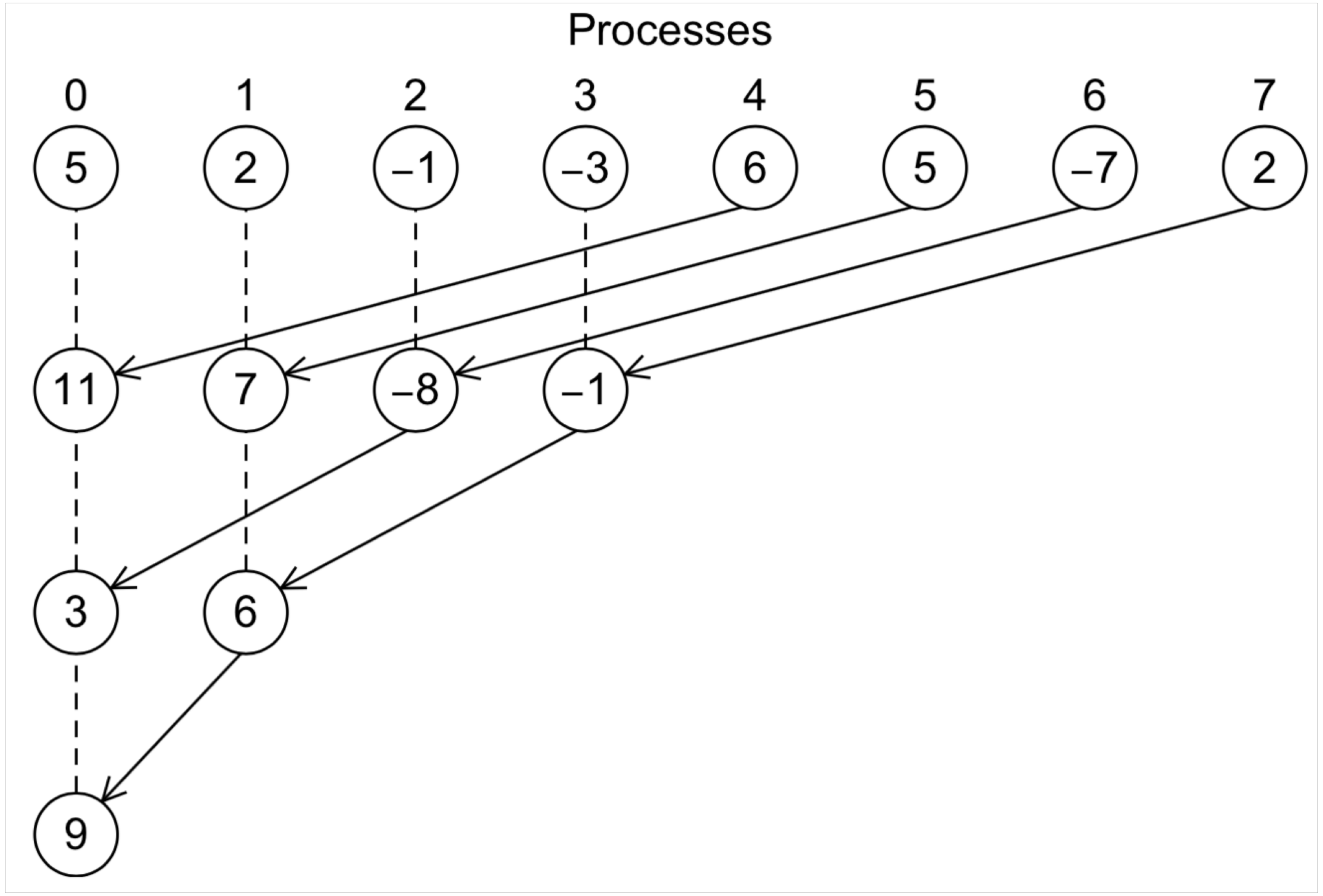


MPI: MPI_Reduction

- steps = $\lceil \log_2(N) \rceil$

```
int MPI_Reduce(
void *sendbuf /*in*/,
void *recvbuf /*out*/,
int count /*in*/,
MPI_Datatype datatype /*in*/,
MPI_Op operator /*in*/,
int dest_process /*in*/,
MPI_Comm comm /*in*/)
```

- hint:
  - with count>1, MPI can operate on arrays
  - sendbuf and recvbuf need to different (no aliasing!)



MPI: MPI_Reduce

Worked out example:

```
int local_n, n;

local_n = my_rank;

MPI_Reduce(&local_n /*send_buf*/, &n /*recv_buf*/, 1 /*count*/, MPI_INT,
   MPI_SUM, 0 /*dest_process*/, MPI_COMM_WORLD);
printf("sum of all local_n: %f", n);
```

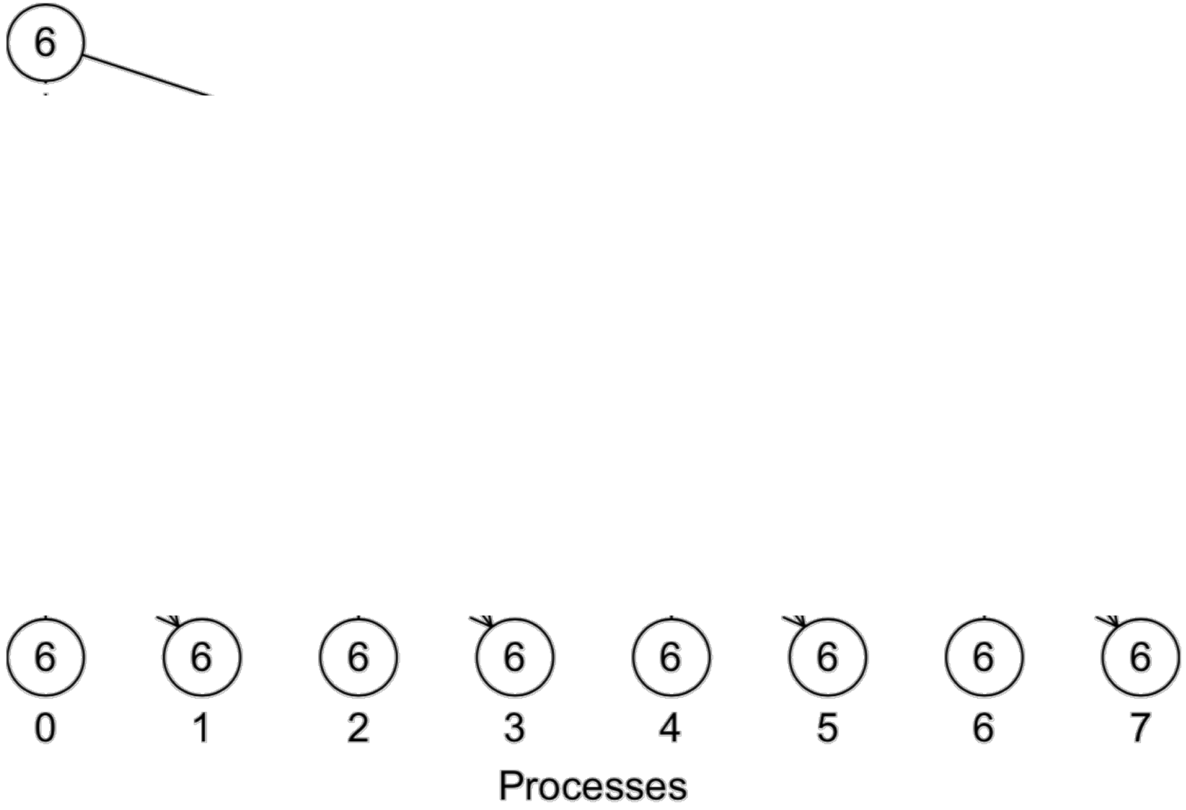
MPI: Reduction Operators

Operation	Meaning
MPI_MAX	Returns the maximum element.
MPI_MIN	Returns the minimum element.
MPI_SUM	Sums the elements.
MPI_PROD	Multiplies all elements.
MPI_LAND	Performs a logical and across the elements.
MPI_LOR	Performs a logical or across the elements.
MPI_BAND	Performs a bitwise and across the bits of the elements.
MPI_BOR	Performs a bitwise or across the bits of the elements.
MPI_MAXLOC	Returns the maximum value and the rank of the process that owns it.
MPI_MINLOC	Returns the minimum value and the rank of the process that owns it.

MPI: P2P ⇔ Collective Communication

- ALL processes in communicator must call SAME collective function at the same time.
- Arguments in all ranks must fit:
  - eg. same dest_process, datatype, operator, comm
  - depending on function
- Only rank dest_process may use recvbuf (but all ranks have to provide such argument)
- MPI_Reduce calls matched solely on:
  - the communicator and
  - the order on which they are called.
  - No helping tags or sender id available.

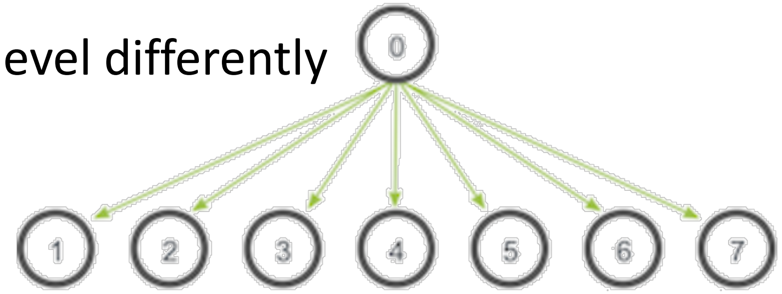
MPI: Broadcast



Broadcasts a message from the process "sending_rank" to all other processes of the communicator

- MPI_Bcast(
  void *data,
  int count,
  MPI_Datatype datatype,
  int sending_rank,
  MPI_Comm comm)

- Hint: All ranks have to call this function
- Might be implemented on hardware level differently (MPI implementation should know)



MPI: Broadcast

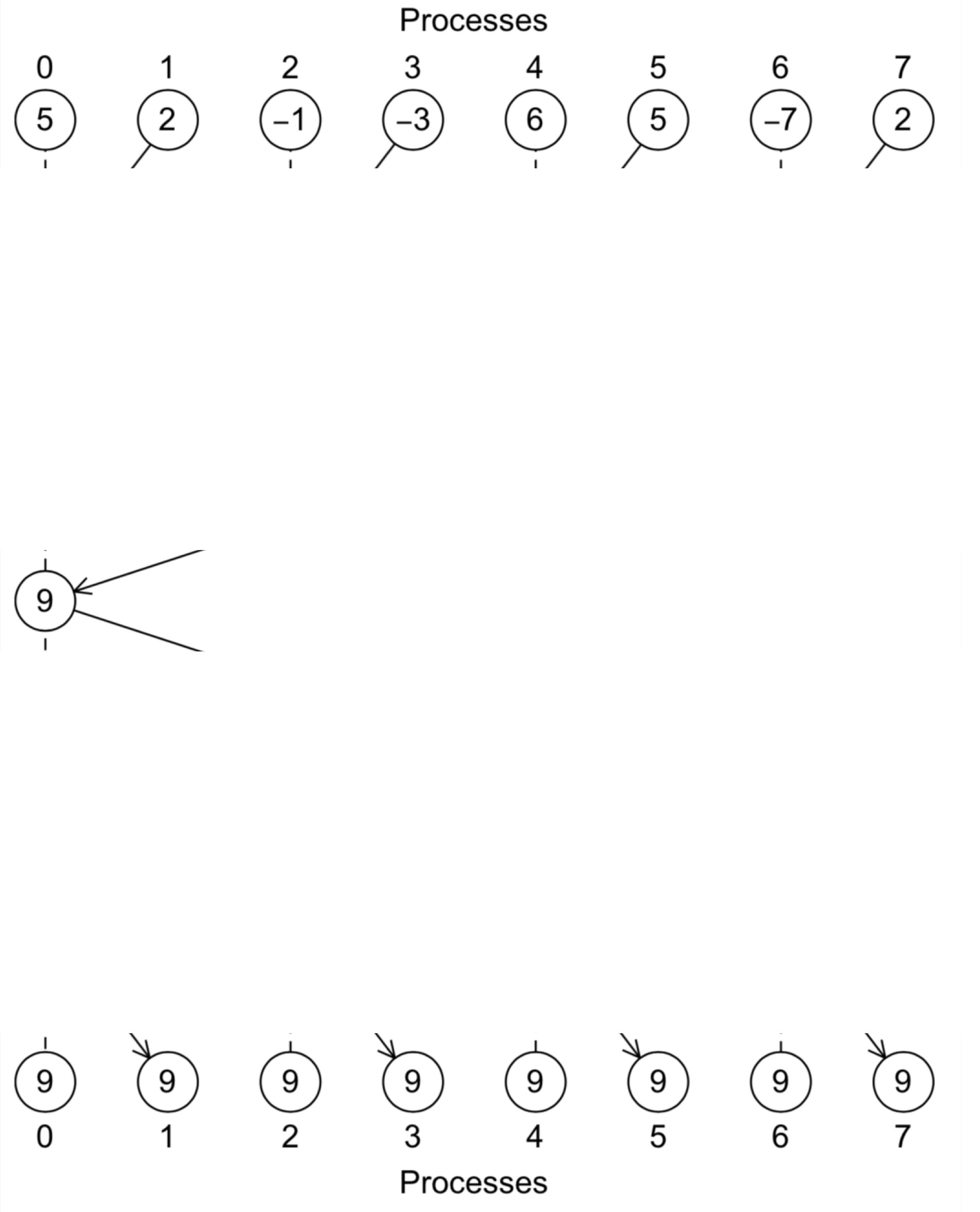
Worked out example, sending 2 variables:

```
int my_rank, my_size, *n;
double *a;

if (!my_rank) {
        printf(”Enter a and n:\n“);
        scanf(“%lf %d”, a, n);
}
MPI_Bcast(a, 1, MPI_DOUBLE, 0, MPI_COMM_WORLD);
MPI_Bcast(n, 1, MPI_DOUBLE, 0, MPI_COMM_WORLD);
```

MPI: MPI_Allreduce

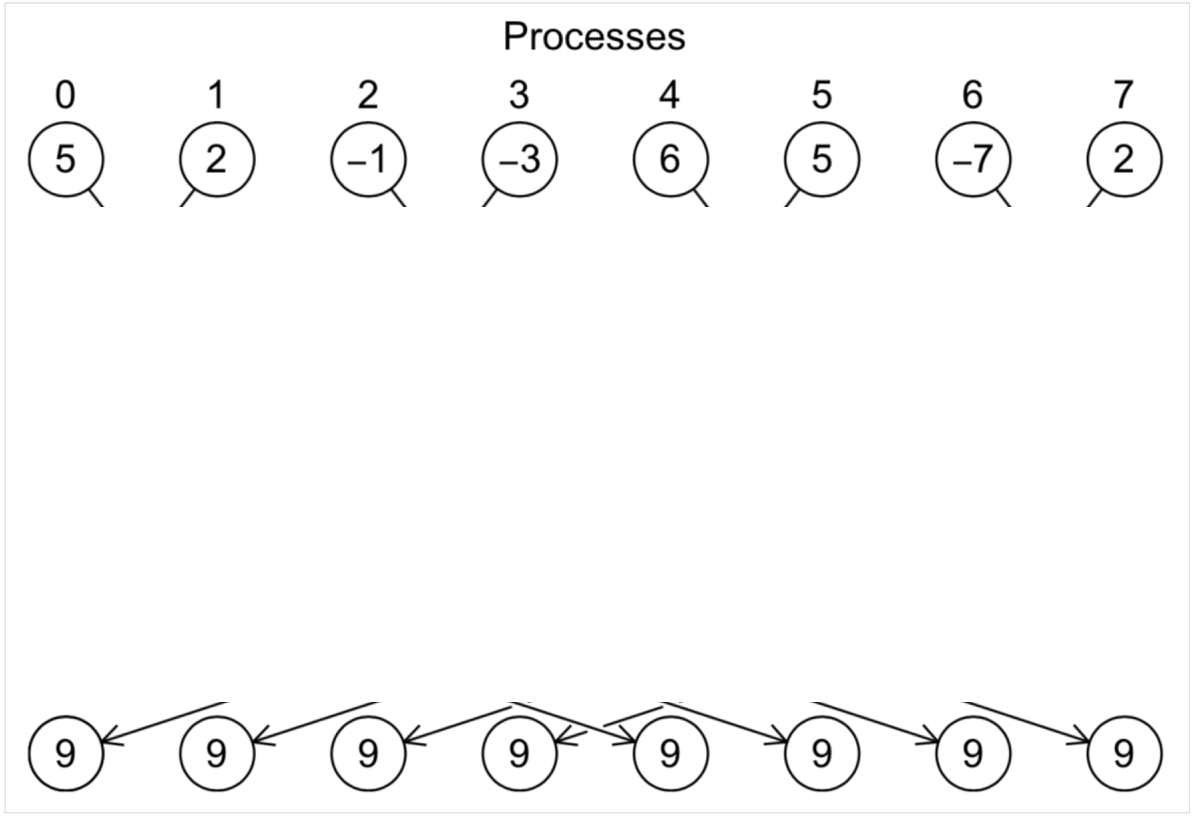
- Combines values from all processes and distributes the result back to all processes
  eg: all processes need global sum
  1. compute global sum (MPI_Reduce) + Broadcast



MPI: MPI_Allreduce

- Combines values from all processes and distributes the result back to all processes
  eg: all processes need global sum
  1. compute global sum (MPI_Reduce) + Broadcast
  2. exchange partial sums (better!, "butterfly")
- MPI has optimal performance with

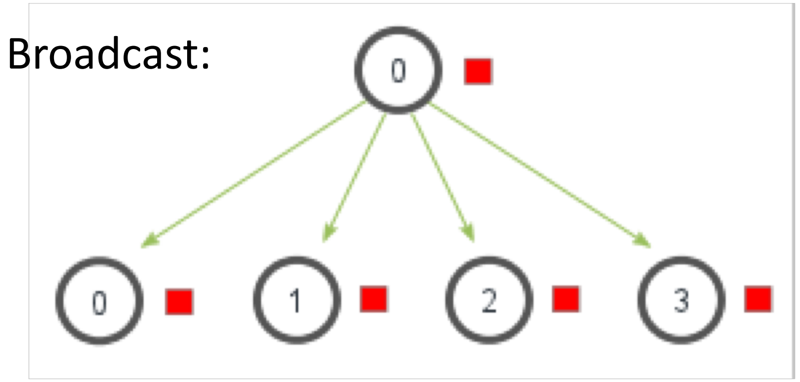
```
int MPI_Allreduce(
void *sendbuf /*in*/,
void *recvbuf /*out*/,
int count /*in*/,
MPI_Datatype datatype /*in*/,
MPI_Op operator /*in*/,
MPI_Comm comm /*in*/)
```

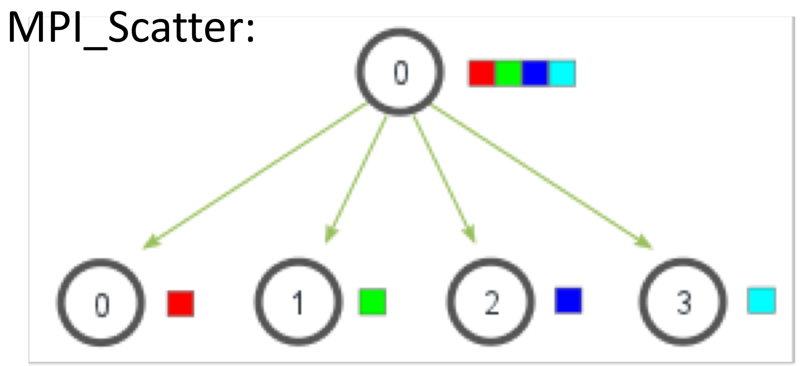


MPI: MPI_Scatter

Sends data from one process to all other processes in a communicator

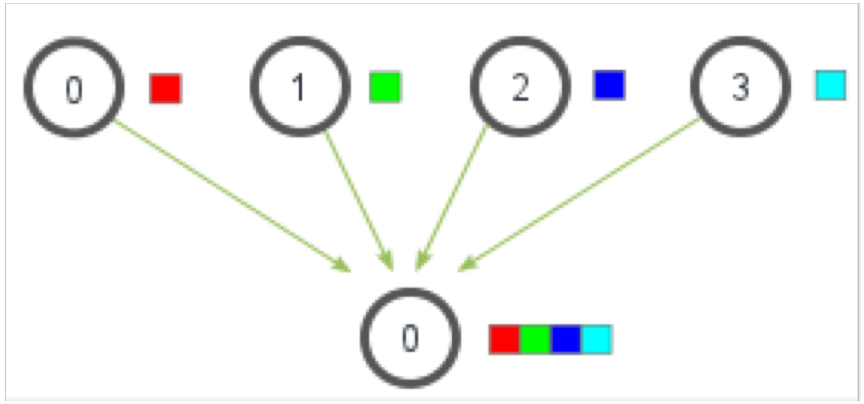
- Selective distribution of data to processes
- MPI_Scatter(
  void* send_data /*in*/,
  int send_count /*in*/,
  MPI_Datatype send_datatype /*in*/,
  void* recv_data /*out*/,
  int recv_count /*in*/,
  MPI_Datatype recv_datatype /*in*/,
  int src_proc /*in*/,
  MPI_Comm comm /*in*/)





MPI: MPI_Gather

Gathers together values from a group of processes



- MPI_Gather(
  void* send_data /*in*/,
  int send_count /*in*/,
  MPI_Datatype send_datatype /*in*/,
  void* recv_data /*out*/,
  int recv_count /*in*/,
  MPI_Datatype recv_datatype /*in*/,
  int dest_proc /*in*/,
  MPI_Comm comm /*in*/)

Set up your workbench

- Connect 2 times via SSH to Mogon2 / HIMster2
    1. Use the first SSH connection for editing (gedit, vi, vim, nano, geany) and
       module load mpi/OpenMPI/3.1.1-GCC-7.3.0
       compiling: mpicc -o ExecutableName SourceFileName.c
    2. Use the second connection for the interactive execution on the nodes (no execution on the head node!):
       salloc -p parallel -N 1 --time=01:30:00 -A m2_himkurs --reservation=himkurs -C skylake
       module load mpi/OpenMPI/3.1.1-GCC-7.3.0
       mpirun -n 2 ./ExecutableName
- Download the files via: wget https://www.hi-mainz.de/fileadmin/user_upload/IT/lectures/WiSe2018/HPC/files/MPI-03.zip && unzip MPI-03.zip

Hints:

- If the reservation with salloc –p parallel fails, try:
    - salloc -p devel -n 4 -A m2_him_exp
- The reserved resources with salloc can’t be overwritten with mpirun
    - Resources(salloc) => Resources(mpirun)
- Possible to check reservation with: squeue -u USERNAME

Exercise 5: Msg passing in two rings

Learning objectives:

- Using Sub-Communicators, similar to example 4 but with two rings)

Steps:

1. Download the skeleton from lecture webpage:
   - wget https://www.hi-mainz.de/fileadmin/user_upload/IT/lectures/WiSe2018/HPC/files/MPI-04.zip && unzip MPI-04.zip
2. Change to program from example 4 and create 2 Sub-Communicators: The first 1/3 of ranks belong to group 0, the second 2/3 belong to group 1. Within each group, establish a ring and pass around
   1. the rank within the sub-communicator
   2. the rank within MPI_COMM_WORLD

   Sum up theses values locally and print finally the results for each rank.

Reminder: ranks within a ring passes on information to their neighbour (1D-cyclic boundary condition).

- each rank sets its local send buffer to “my_rank”/ ”my_subcomm_rank”
- Each rank does “my_size”/”my_subcomm_size” times:
- receives data from previous rank and stores this in a local buffer,
- increment the local sum by the just received local buffer
- sends the new buffer to the next in the ring

- Use non-blocking communication

Bonus: Use MPI_Allreduce to get to the same results.

Exercise 6: Scatter and Gather data

Learning objectives:

- First usage of MPI_Scatter and MPI_Gather

Steps:

1. Download the skeleton from lecture webpage:
    - wget https://www.hi-mainz.de/fileadmin/user_upload/IT/lectures/WiSe2018/HPC/files/MPI-06.zip && unzip MPI-06.zip
2. Complete to program to scatter data to all processes
   and
   to gather lather the results.

Bonus: Use MPI_Allgather finally.

Solutions