Introduction

- What is message passing?
  - Sending and receiving messages between *tasks* or *processes*
  - Includes performing operations on data in transit and synchronizing tasks
- Why send messages?
  - Clusters have distributed memory, i.e. each process has its own address space and no way to get at another's
- How do you send messages?
  - Programmer makes use of an Application Programming *Interface* (API)
    - In this case, MPI.
  - MPI specifies the functionality of high-level communication routines
  - MPI's functions give access to a low-level *implementation* that takes care of sockets, buffering, data copying, message routing, etc.

API for Distributed Memory Parallelism

- Assumption: processes do not see each other’s memory
- Communication speed is determined by some kind of network
  - Typical network = switch + cables + adapters + software stack…
- Key: the *implementation* of MPI (or any message passing API) can be optimized for any given network
  - Expert-level performance
  - No code changes required
  - Works in shared memory, too

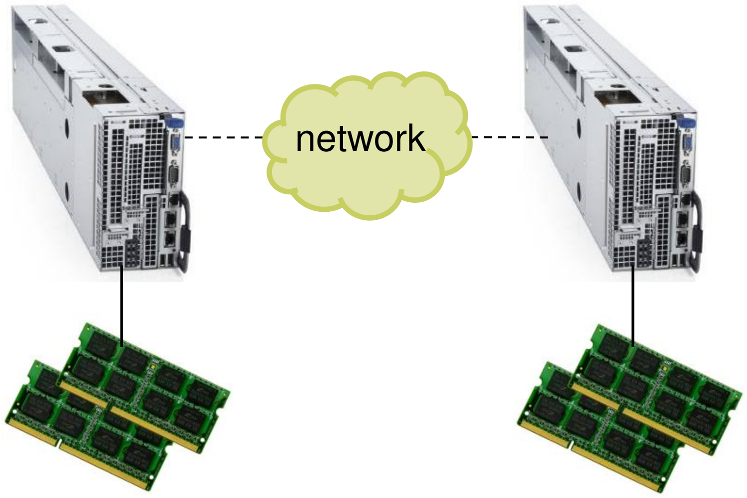


Image of Dell PowerEdge C8220X: http://www.theregister.co.uk/2012/09/19/dell_zeus_c8000_hyperscale_server/

MPI_COMM MPI Communicators

- Communicators
  - Collections of processes that can communicate with each other
  - Most MPI routines require a communicator as an argument
  - Predefined communicator MPI_COMM_WORLD encompasses all tasks
  - New communicators can be defined; any number can co-exist
- Each communicator must be able to answer two questions
  - *How many processes exist in this communicator?*
  - MPI_Comm_size returns the answer, say, N_p
  - *Of these processes, which process (numerical rank) am I?*
  - MPI_Comm_rank returns the rank of the current process within the communicator, an integer between 0 and N_p-1 inclusive
  - Typically these functions are called just after MPI_Init

MPI_COMM C Example: param.c

```
#include <mpi.h>
main(int argc, char **argv){
   int np, mype, ierr;

   ierr = MPI_Init(&argc, &argv);
   ierr = MPI_Comm_size(MPI_COMM_WORLD, &np);
   ierr = MPI_Comm_rank(MPI_COMM_WORLD, &mype);
                :
   MPI_Finalize();
}
```

MPI_COMM C++ Example: param.cc

```
#include "mpif.h"
[other includes]
int main(int argc, char *argv[]){
   int np,  mype,  ierr;
 [other declarations]
            :
          MPI::Init(argc, argv);
   np    = MPI::COMM_WORLD.Get_size();
   mype = MPI::COMM_WORLD.Get_rank();
            :
        [actual work goes here]
            :
          MPI::Finalize();
}
```

Simple MPI

Here is the basic outline of a simple MPI program :

- Include the implementation-specific header file --

  #include <mpi.h> inserts basic definitions and types
- Initialize communications –

  MPI_Init initializes the MPI environment
  MPI_Comm_size returns the number of processes
  MPI_Comm_rank returns this process's number (rank)
- Communicate to share data between processes –

  MPI_Send sends a message
  MPI_Recv receives a message
- Exit from the message-passing system --

  MPI_Finalize

Basics

Minimal Code Example: hello_mpi.c

```
#include <stdio.h>
#include "mpi.h"
main(int argc, char **argv)
{
  char message[20];
  int i, rank, size, tag = 99;
  MPI_Status status;
  MPI_Init(&argc, &argv);
  MPI_Comm_size(MPI_COMM_WORLD, &size);
  MPI_Comm_rank(MPI_COMM_WORLD, &rank);
  if (rank == 0) {
    strcpy(message, "Hello, world!");
    for (i = 1; i < size; i++)
      MPI_Send(message, 13, MPI_CHAR, i, tag, MPI_COMM_WORLD);
  } else {
    MPI_Recv(message, 20, MPI_CHAR, 0, tag, MPI_COMM_WORLD, &status);
  }
  printf("Message from process %d : %.13s\n", rank, message);
  MPI_Finalize();
}
```

Basics

Initialize and Close Environment

```
#include <stdio.h>
#include "mpi.h"
main(int argc, char **argv)
{
   char message[20];
   int i, rank, size, tag = 99;
   MPI_Status status;
   MPI_Init(&argc, &argv);
   MPI_Comm_size(MPI_COMM_WORLD, &size);
   MPI_Comm_rank(MPI_COMM_WORLD, &rank);
   if (rank == 0) {
      strcpy(message, "Hello, world!");
      for (i = 1; i < size; i++)
         MPI_Send(message, 13, MPI_CHAR, i, tag, MPI_COMM_WORLD);
   } else
      MPI_Recv(message, 20, MPI_CHAR, 0, tag, MPI_COMM_WORLD, &status);
   printf("Message from process %d : %.13s\n", rank, message);
   MPI_Finalize();
}
```

Initialize MPI environment

An implementation may also use this call as a mechanism for making the usual argc and argv command-line arguments from “main” available to all tasks (C language only).

Close MPI environment

Basics

Query Environment

```
#include <stdio.h>
#include "mpi.h"
main(int argc, char **argv)
{
  char message[20];
  int i, rank, size, tag = 99;
  MPI_Status status;
  MPI_Init(&argc, &argv);
  MPI_Comm_size(MPI_COMM_WORLD, &size);
  MPI_Comm_rank(MPI_COMM_WORLD, &rank);
  if (rank == 0) {
    strcpy(message, "Hello, world!");
    for (i = 1; i < size; i++)
      MPI_Send(message, 13, MPI_CHAR, i, tag, MPI_COMM_WORLD);
  } else
    MPI_Recv(message, 20, MPI_CHAR, 0, tag, MPI_COMM_WORLD, &status);
  printf("Message from process %d : %.13s\n", rank, message);
  MPI_Finalize();
}
```

Returns number of processes
This, like nearly all other MPI functions, must be called after MPI_Init and before MPI_Finalize. Input is the name of a communicator (MPI_COMM_WORLD is the global communicator) and output is the size of that communicator.

Returns this process' number, or rank
Input is again the name of a communicator and the output is the rank of this process in that communicator.

Basics | Pass Messages

```
#include <stdio.h>
#include "mpi.h"
main(int argc, char **argv)
{
   char message[20];
   int i, rank, size, tag = 99;
   MPI_Status status;
   MPI_Init(&argc, &argv);
   MPI_Comm_size(MPI_COMM_WORLD, &size);
   MPI_Comm_rank(MPI_COMM_WORLD, &rank);
   if (rank == 0) {
      strcpy(message, "Hello, world!");
      for (i = 1; i < size; i++)
         MPI_Send(message, 13, MPI_CHAR, i, tag, MPI_COMM_WORLD);
   } else
      MPI_Recv(message, 20, MPI_CHAR, 0, tag, MPI_COMM_WORLD, &status);
   printf("Message from process %d : %.13s\n", rank, message);
   MPI_Finalize();
}
```

Send a message
Blocking send of data in the buffer.

Receive a message
Blocking receive of data into the buffer.

Messages — Three Parameters Describe the Data

```
MPI_Send( message, 13, MPI_CHAR, i,  tag, MPI_COMM_WORLD  );

MPI_Recv( message, 20, MPI_CHAR, 0, tag, MPI_COMM_WORLD, &status);
```

Type of data, should be same for send and receive
MPI_Datatype type

Number of elements (items, not bytes)
Recv number should be greater than or equal to amount sent
int count

Address where the data start
void data*

Three Parameters Specify Routing

```
MPI_Send(  message, 13, MPI_CHAR,  i,  tag, MPI_COMM_WORLD  );

MPI_Recv(  message, 20, MPI_CHAR,  0, tag, MPI_COMM_WORLD, &status);
```

Identify process you’re communicating with by rank number
int dest/src

Arbitrary tag number, must match up (receiver can specify MPI_ANY_TAG to indicate that any tag is acceptable)
int tag

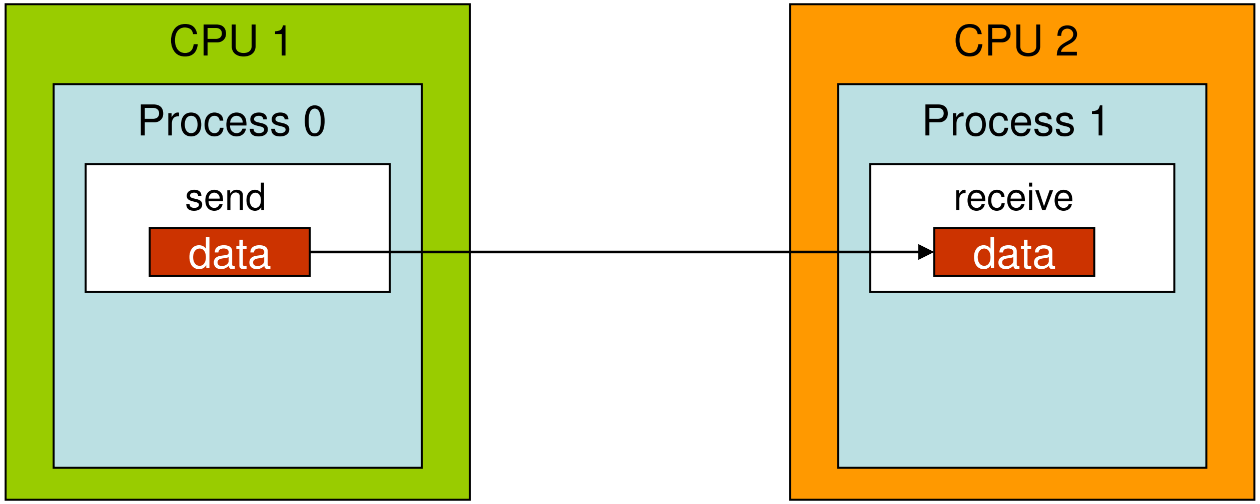
Communicator specified for send and receive must match, no wildcards
MPI_Comm comm

Returns information on received message
MPI_Status status*

Point to Point Topics

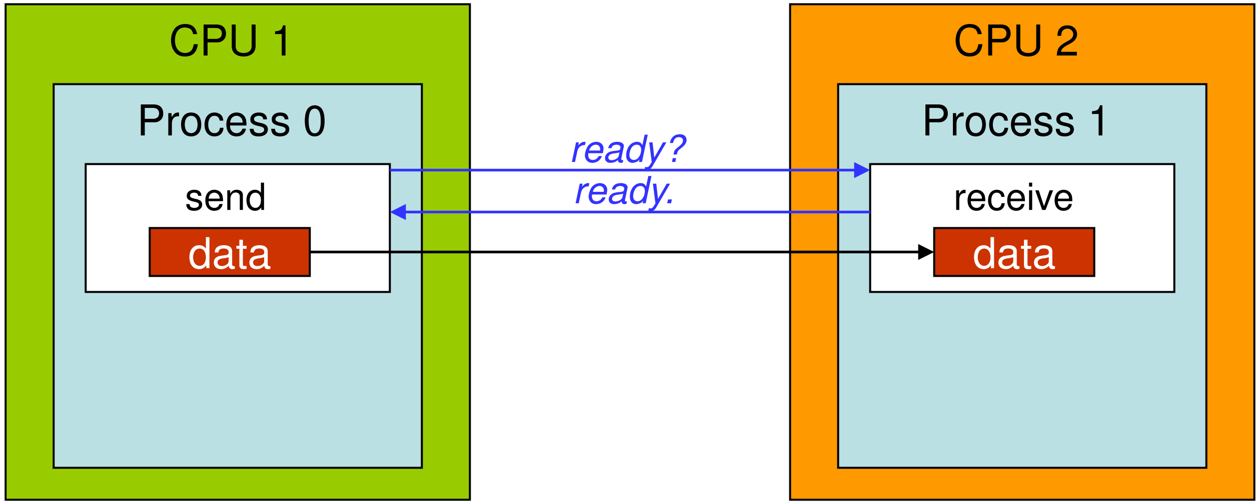
- MPI_Send and MPI_Recv: how simple are they really?
- Synchronous vs. buffered (asynchronous) communication
- Reducing overhead: ready mode, standard mode
- Combined send/receive
- Blocking vs. non-blocking send and receive
- Deadlock, and how to avoid it

Point to Point Send and Recv: Simple?



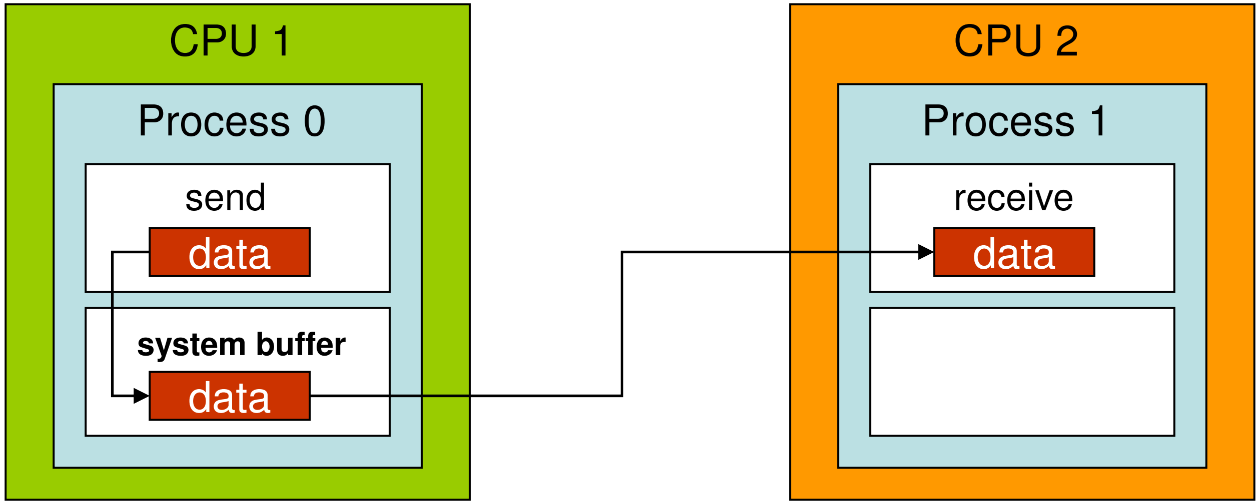
- Sending data ***from*** one point (process/task)
  to another point (process/task)
- One task sends while another receives
- But what if process 1 isn't **ready** for the message from process 0?…
- MPI provides different communication modes in order to help

Point to Point | Synchronous Send, MPI_Ssend



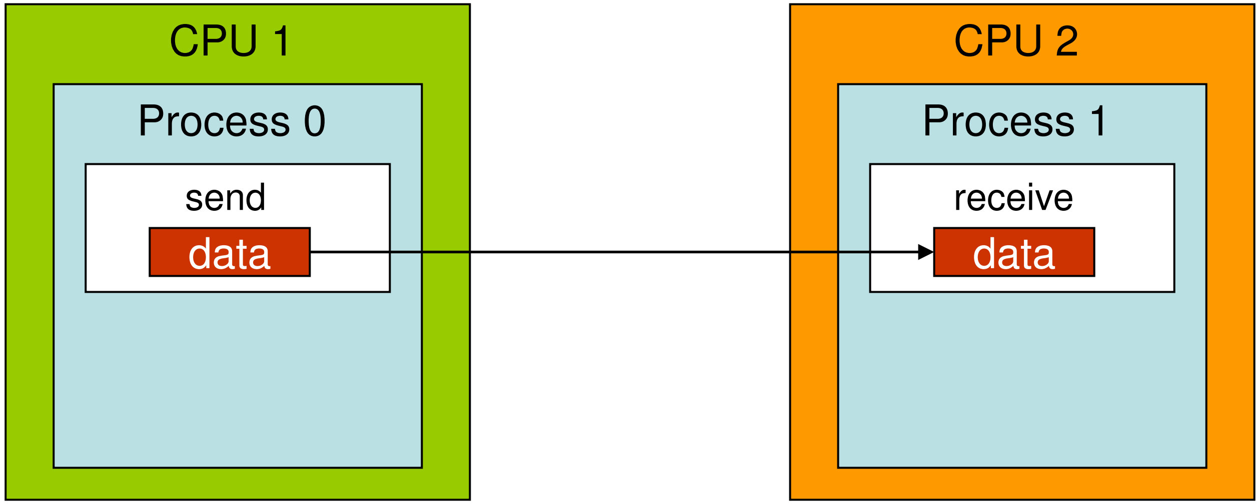
- Handshake procedure ensures both processes are ready
- It’s likely that one of the processes will end up waiting
  - If the *send* call occurs first: sender waits
  - If the *receive* call occurs first: receiver waits
- Waiting and an extra handshake? – this could be slow

Point to Point | Buffered Send, MPI_Bsend



- Message data are copied to a system-controlled block of memory
- Process 0 continues executing other tasks without waiting
- When process 1 is ready, it fetches the message from the remote system buffer and stores it in the appropriate memory location
- Must be preceded with a call to MPI_Buffer_attach

Point to Point Ready Send, MPI_Rsend

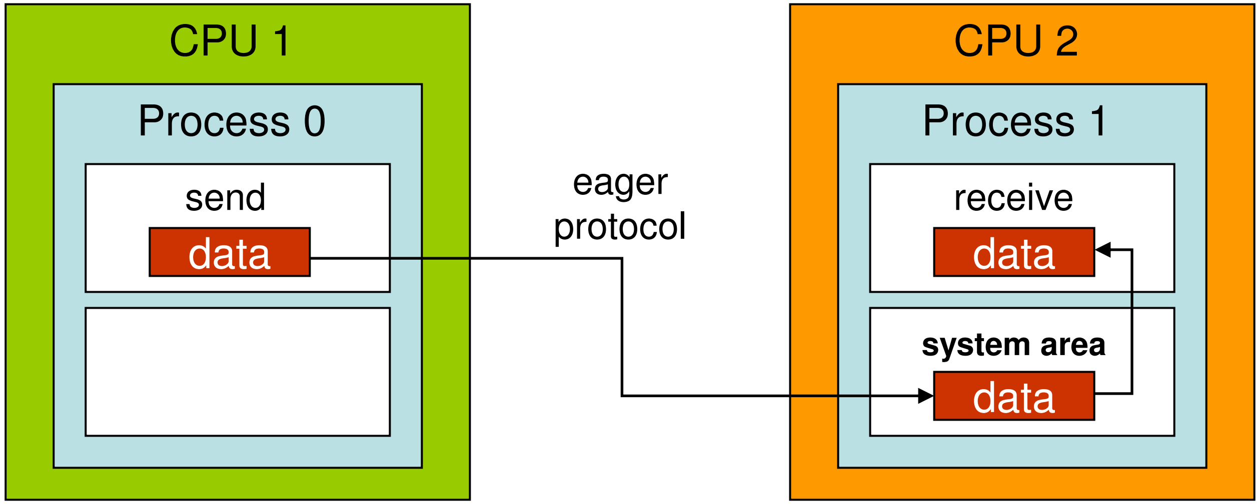


- Process 0 just assumes process 1 is ready! The message is sent!
- Truly simple communication, no extra handshake or copying
- But an error is generated if process 1 is unable to receive
- Only useful when logic dictates that the receiver *must* be ready

Point to Point Overhead

- **System overhead**
  Buffered send has more system overhead due to the extra copy operation.
- **Synchronization overhead**
  Synchronous send has no extra copying but more waiting, because a handshake must arrive before the send can occur.
- **MPI_Send**
  Standard mode tries to trade off between the types of overhead.
  - Large messages use the "rendezvous protocol" to avoid extra copying: a handshake procedure establishes direct communication.
  - Small messages use the "eager protocol" to avoid synchronization cost: the message is quickly copied to a small system buffer on the receiver.

Point to Point | Standard Send, Eager Protocol



- Message goes a system-controlled area of memory *on the receiver*
- Process 0 continues executing other tasks; when process 1 is ready to receive, the system simply copies the message from the system buffer into the appropriate memory location controlled by process
- *Does not* need to be preceded with a call to MPI_Buffer_attach

Point to Point MPI_Sendrecv

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

- Good for two-way communication between a pair of nodes, in which each one sends and receives a message
- However, destination and source need not be the same (ring, e.g.)
- Equivalent to blocking send + blocking receive
- Send and receive use the same communicator but have distinct tags

Point to Point Send and Recv: So Many Choices

The communication mode indicates how the message should be *sent*.

Communication Mode	Blocking Routines	Non-Blocking Routines
Synchronous	MPI_Ssend	MPI_Issend
Ready	MPI_Rsend	MPI_Irsend
Buffered	MPI_Bsend	MPI_Ibsend
Standard	MPI_Send	MPI_Isend
	MPI_Recv	MPI_Irecv
	MPI_Sendrecv	
	MPI_Sendrecv_replace	

Note: the receive routine does not specify the communication mode -- it is simply blocking or non-blocking.

Point to Point Communication Modes

Mode	Pros	Cons
Synchronous – sending and receiving tasks must 'handshake'.	- Safest, therefore most portable - No need for extra buffer space - SEND/RECV order not critical	Synchronization overhead
Ready- assumes that a 'ready to receive' message has already been received.	- Lowest total overhead - No need for extra buffer space - Handshake not required	RECV *must* precede SEND
Buffered – move data to a buffer so process does not wait.	- Decouples SEND from RECV - No sync overhead on SEND - Programmer controls buffer size	Buffer copy overhead
Standard – defined by the implementer; meant to take advantage of the local system.	- Good for many cases - Small messages go right away - Large messages must sync - Compromise position	Your program may not be suitable

Point to Point Blocking vs. Non-Blocking

MPI_Send, MPI_Recv

A ***blocking*** call suspends execution of the process until the message buffer being sent/received is safe to use.

MPI_Isend, MPI_Irecv

A ***non-blocking*** call just initiates communication; the status of data transfer and the success of the communication must be verified later by the programmer (MPI_Wait or MPI_Test).

Point to Point | One-Way Blocking/Non-Blocking

- Blocking send, non-blocking recv

```
IF (rank==0) THEN
  ! Do my work, then send to rank 1
  CALL MPI_SEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,ie)
ELSEIF (rank==1) THEN
  CALL MPI_IRECV (recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,req,ie)
  ! Do stuff that doesn't yet need recvbuf from rank 0
  CALL MPI_WAIT (req,status,ie)
  ! Do stuff with recvbuf
ENDIF
```

- Non-blocking send, non-blocking recv

```
IF (rank==0) THEN
  ! Get sendbuf ready as soon as possible
  CALL MPI_ISEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,req,ie)
  ! Do other stuff that doesn’t involve sendbuf
ELSEIF (rank==1) THEN
  CALL MPI_IRECV (recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,req,ie)
ENDIF
CALL MPI_WAIT (req,status,ie)
```

Point to Point Two-Way Communication: Deadlock!

- **Deadlock 1**

```
IF (rank==0) THEN
  CALL MPI_RECV (recvbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,status,ie)
  CALL MPI_SEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,ie)
ELSEIF (rank==1) THEN
  CALL MPI_RECV (recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,status,ie)
  CALL MPI_SEND (sendbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,ie)
ENDIF
```

- **Deadlock 2**

```
IF (rank==0) THEN
  CALL MPI_SSEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,ie)
  CALL MPI_RECV (recvbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,status,ie)
ELSEIF (rank==1) THEN
  CALL MPI_SSEND (sendbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,ie)
  CALL MPI_RECV (recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,status,ie)
ENDIF
```

  - MPI_Send has same problem for count*MPI_REAL > 12K (the MVAPICH2 “eager threshold”; it’s 256K for Intel MPI)

Point to Point Deadlock Solutions

- Solution 1

```
IF (rank==0) THEN
  CALL MPI_SEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,ie)
  CALL MPI_RECV (recvbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,status,ie)
ELSEIF (rank==1) THEN
  CALL MPI_RECV (recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,status,ie)
  CALL MPI_SEND (sendbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,ie)
ENDIF
```

- Solution 2

```
IF (rank==0) THEN
  CALL MPI_SENDRECV (sendbuf,count,MPI_REAL,1,tag, &
                recvbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,status,ie)
ELSEIF (rank==1) THEN
  CALL MPI_SENDRECV (sendbuf,count,MPI_REAL,0,tag, &
                recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,status,ie)
ENDIF
```

Point to Point More Deadlock Solutions

- Solution 3

```
IF (rank==0)   THEN
  CALL MPI_IRECV (recvbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,req,ie)
  CALL MPI_SEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,ie)
ELSEIF (rank==1) THEN
  CALL MPI_IRECV (recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,req,ie)
  CALL MPI_SEND (sendbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,ie)
ENDIF
CALL MPI_WAIT (req,status)
```

- Solution 4

```
IF (rank==0) THEN
  CALL MPI_BSEND (sendbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,ie)
  CALL MPI_RECV (recvbuf,count,MPI_REAL,1,tag,MPI_COMM_WORLD,status,ie)
ELSEIF (rank==1) THEN
  CALL MPI_BSEND (sendbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,ie)
  CALL MPI_RECV(recvbuf,count,MPI_REAL,0,tag,MPI_COMM_WORLD,status,ie)
ENDIF
```

Point to Point | Two-way Communications: Summary

	Task 0	Task 1
Deadlock 1	Recv/Send	Recv/Send
Deadlock 2	Send/Recv	Send/Recv
Solution 1	Send/Recv	Recv/Send
Solution 2	Sendrecv	Sendrecv
Solution 3	Irecv/Send, Wait	(I)recv/Send, (Wait)
Solution 4	Bsend/Recv	(B)send/Recv

Collective Motivation

- What if one task wants to send to *everyone*?

```
if (mytid == 0) {
  for (tid=1; tid<ntids; tid++) {
    MPI_Send( (void*)a, /* target= */ tid, … );
  }
} else {
  MPI_Recv( (void*)a, 0, … );
}
```

- Implements a very naive, serial broadcast
- Too primitive
  - Leaves no room for the OS / switch to optimize
  - Leaves no room for more efficient algorithms
- Too slow

Collective Topics

- Overview
- Barrier and Broadcast
- Data Movement Operations
- Reduction Operations

Collective Overview

- Collective calls involve ALL processes within a communicator
- There are 3 basic types of collective communications:
  - Synchronization (MPI_Barrier)
  - Data movement (MPI_Bcast/Scatter/Gather/Allgather/Alltoall)
  - Collective computation (MPI_Reduce/Allreduce/Scan)
- Programming considerations & restrictions
  - Blocking operation (also non-blocking in MPI-3)
  - No use of message tag argument
  - Collective operations within subsets of processes require separate grouping and new communicator

Collective | Barrier Synchronization, Broadcast

- *Barrier* blocks until all processes in comm have called it
  - Useful when measuring communication/computation time

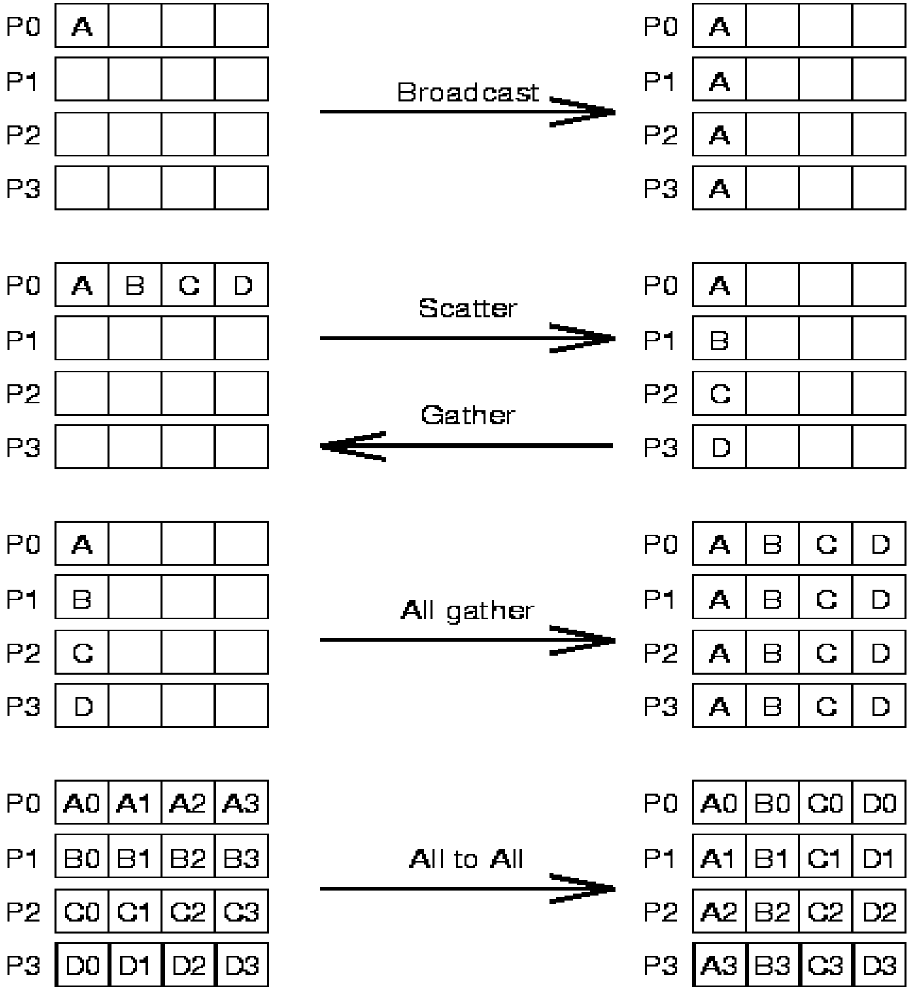
```
mpi_barrier(comm, ierr)
MPI_Barrier(comm)
```

- *Broadcast* sends data from root to all processes in comm
  - Again, blocks until all tasks have called it

```
mpi_bcast(data, count, type, root, comm, ierr)
MPI_Bcast(data, count, type, root, comm)
```

Collective Data Movement

- Broadcast
- Scatter/Gather
- Allgather
- Alltoall



Broadcast

P0	A				P0	A			
P1					P1	A			
P2					P2	A			
P3					P3	A			

Scatter
Gather

P0	A	B	C	D	P0	A			
P1					P1	B			
P2					P2	C			
P3					P3	D			

All gather

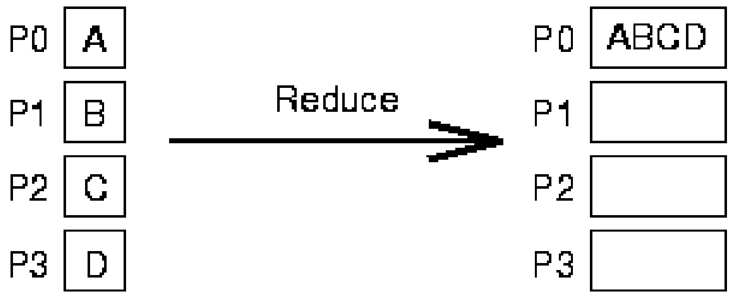
P0	A				P0	A	B	C	D
P1	B				P1	A	B	C	D
P2	C				P2	A	B	C	D
P3	D				P3	A	B	C	D

All to All

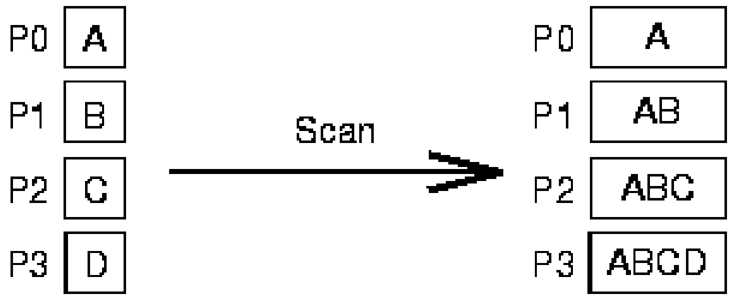
P0	A0	A1	A2	A3	P0	A0	B0	C0	D0
P1	B0	B1	B2	B3	P1	A1	B1	C1	D1
P2	C0	C1	C2	C3	P2	A2	B2	C2	D2
P3	D0	D1	D2	D3	P3	A3	B3	C3	D3

Collective Reduction Operations

- Reduce



- Scan (Prefix)



Collective Reduction Operations

Name	Meaning
MPI_MAX	Maximum
MPI_MIN	Minimum
MPI_SUM	Sum
MPI_PROD	Product
MPI_LAND	Logical and
MPI_BAND	Bit-wise and
MPI_LOR	Logical or
MPI_BOR	Bit-wise or
MPI_LXOR	Logical xor
MPI_BXOR	Logical xor
MPI_MAXLOC	Max value and location
MPI_MINLOC	Min value and location

Collective

The Collective Collection!