Distributed-Memory Parallel Computers

- Scalable interconnection network.
- Some memory physically local, some remote
- Major issues: network latency and bandwidth, message-passing overhead, data distribution.
Logical model
Physical model
Interconnection Network Topology

• Very interesting theoretical topic.
• For typical users of modern parallel machines, the details of interconnection topology are not a first-order concern*.

* Your mileage may vary.
Message passing programming model

- N nodes connected by an *interconnection network*
- Nodes are processor(s)/memory/network
- K processes are distributed to a subset of N nodes
- Processes have separate address spaces, code, data, stack, …
- Processes communicate by explicitly sending/receiving messages.
- A message is an array of, possibly typed, bytes
- Message passing is *two-sided*.
- Flexible, very scalable model
- Incremental parallelization and debugging can be difficult.
What is MPI?

- De facto standard API for explicit message-passing MIMD/SPMD programming.
- Many implementations over many networks
- Developed in mid 90’s by consortium, reflecting lessons learned from machine-specific libraries and PVM.
- Focused on: homogeneous MPPs, high performance, library writing, portability.
Six-Function MPI

- MPI_Init
- MPI_Finalize
- MPI_Comm_rank
- MPI_Comm_size
- MPI_Send
- MPI_Recv

Initialize MPI
Close it down
Get my process #
How many total?
Send message
Receive message
MPI “Hello World” in Fortran77

```fortran
implicit none
include 'mpif.h'
integer myid, numprocs, ierr

call mpi_init( ierr )
call mpi_comm_rank( mpi_comm_world, myid, ierr )
call mpi_comm_size( mpi_comm_world, numprocs, ierr )

print *, "hello from ", myid, " of ", numprocs

call mpi_finalize( ierr )
stop
end
```
MPI “Hello World” in C

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

int main(int argc, char *argv[
{
    int myid, numprocs, namelen;
    char processor_name[MPI_MAX_PROCESSOR_NAME];

    MPI_Init(&argc,&argv);
    MPI_Comm_size(MPI_COMM_WORLD,&numprocs);
    MPI_Comm_rank(MPI_COMM_WORLD,&myid);
    MPI_Get_processor_name(processor_name, &namelen);
    printf("hello from %s: process %d of %d\n",
        processor_name, myid, numprocs);
    MPI_Finalize();
}
Function MPI_Send

MPI_SEND(buf, count, datatype, dest, tag, comm)

IN   buf     initial address of send buffer (choice)
IN   count   number of entries to send (integer)
IN   datatype datatype of each entry (handle)
IN   dest    rank of destination (integer)
IN   tag     message tag (integer)
IN   comm    communicator (handle)

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

MPI_SEND(BUF, COUNT, DATATYPE, DEST, TAG, COMM, IERR)
<type> BUF(*)
    INTEGER COUNT, DATATYPE, DEST, TAG, COMM, IERR
## Function MPI_Recv

MPI_RECV(buf, count, datatype, source, tag, comm, status)

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OUT</td>
<td>buf</td>
</tr>
<tr>
<td></td>
<td>initial address of receive buffer (choice)</td>
</tr>
<tr>
<td>IN</td>
<td>count</td>
</tr>
<tr>
<td></td>
<td>max number of entries to receive (integer)</td>
</tr>
<tr>
<td>IN</td>
<td>datatype</td>
</tr>
<tr>
<td></td>
<td>datatype of each entry (handle)</td>
</tr>
<tr>
<td>IN</td>
<td>dest</td>
</tr>
<tr>
<td></td>
<td>rank of source (integer)</td>
</tr>
<tr>
<td>IN</td>
<td>tag</td>
</tr>
<tr>
<td></td>
<td>message tag (integer)</td>
</tr>
<tr>
<td>IN</td>
<td>comm</td>
</tr>
<tr>
<td></td>
<td>communicator (handle)</td>
</tr>
<tr>
<td>OUT</td>
<td>status</td>
</tr>
<tr>
<td></td>
<td>return status (Status)</td>
</tr>
</tbody>
</table>

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

```c
MPI_RECV(BUF, COUNT, DATATYPE, SOURCE, TAG, COMM, STATUS, IERR)
```

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;type&gt;</td>
<td>BUF(*)</td>
</tr>
<tr>
<td></td>
<td>INTEGER COUNT, DATATYPE, SOURCE, TAG, COMM</td>
</tr>
<tr>
<td></td>
<td>INTEGER STATUS (MPI_STATUS_SIZE), IERR</td>
</tr>
</tbody>
</table>