Message Passing Interface (MPI) Programming

Aiichiro Nakano

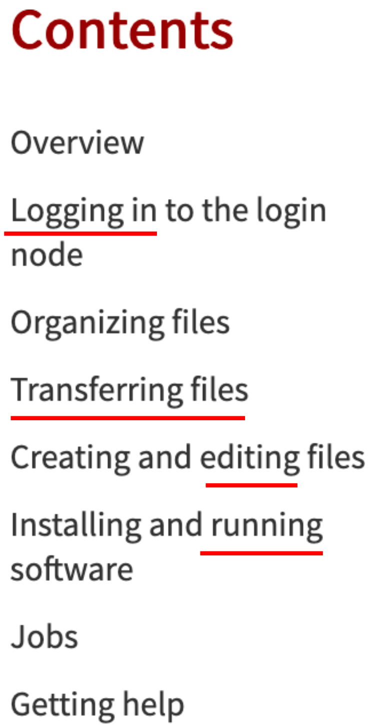
Collaboratory for Advanced Computing & Simulations
Department of Computer Science
Department of Physics & Astronomy
Department of Quantitative & Computational Biology
University of Southern California

Email: anakano@usc.edu

Preparation

Minimal knowledge required for the hands-on projects in this course:

- **Able to log in & use the Discovery computing cluster at USC Center for Advanced Research Computing (CARC) at the level of its "getting started" tutorial:**
  https://carc.usc.edu/user-information/user-guides/hpc-basics/getting-started-discovery
- **Use shell commands to interact with the operating system at the level of "Chapter 1—Introduction to the Command Line" of *Effective Computation in Physics* by Scopatz and Huff; USC students have free access to the book through Safari Online:** https://libraries.usc.edu/databases/safari-books



Contents

Overview

Logging in to the login node

Organizing files

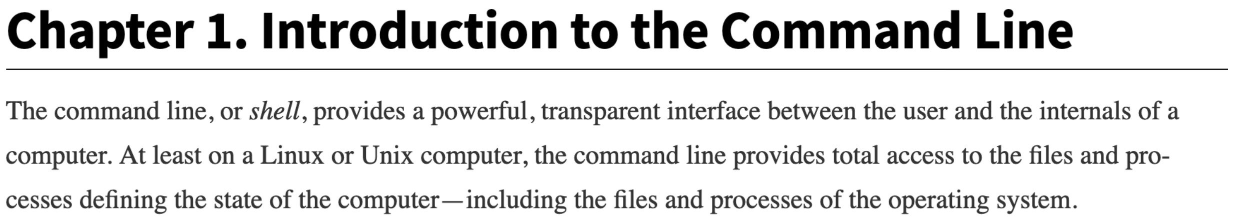
Transferring files

Creating and editing files

Installing and running software

Jobs

Getting help



Chapter 1. Introduction to the Command Line

The command line, or *shell*, provides a powerful, transparent interface between the user and the internals of a computer. At least on a Linux or Unix computer, the command line provides total access to the files and processes defining the state of the computer—including the files and processes of the operating system.

- **To be able to use CARC resources, you need to enroll in Duo two-factor authentication (2FA):** https://itservices.usc.edu/duo/enroll/

How to Use USC CARC Cluster

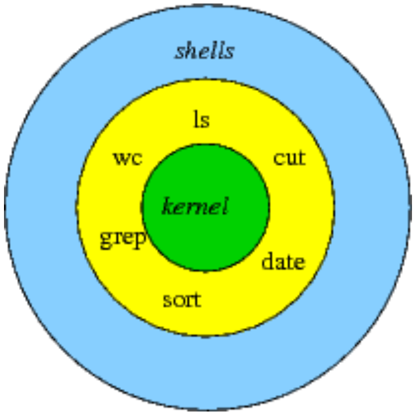
System: Intel/AMD-based computing cluster

https://carc.usc.edu

Log in

```
> ssh anakano@discovery.usc.edu
```

Alternatively, you can use `discovery2.usc.edu`



Shell is a language you speak with the operating system

To use MPI library:

Use text editor like vim, nano, emacs

If using Bash shell, add these in .bashrc

```
module purge
module load usc
```

To set up standard software environment

Type echo $0 to find which shell you are using

Compile an MPI program

```
> mpicc -o mpi_simple mpi_simple.c
```

Execute an MPI program

```
> mpirun -n 2 mpi_simple
```

To find absolute path to mpicc command

```
[anakano@discovery ~]$ which mpicc
/spack/apps/linux-centos7-x86_64/gcc-8.3.0/openmpi-4.0.2-ipm3dnvlbtxawpi4ifz7jma6jgr7mexq/bin/mpicc
[anakano@discovery ~]$ more /proc/cpuinfo
```

To find processor information

Email carc-support@usc.edu for assistance

VPN Issue

- **It is now required to use VPN (virtual private network) to access Discovery from off-campus:**
  https://itservices.usc.edu/vpn

- **Cisco AnyConnect software for VPN on Mac has a DNS (domain name system) problem, which could be bypassed using IP addresses instead of login server names (note discovery.usc.edu is a generic name for the two login servers, discovery1 and discovery2)**

  discovery1.usc.edu: 10.72.0.13

  discovery2.usc.edu: 10.72.0.14

Submit a Slurm Batch Job

Prepare a script file, mpi_simple.sl

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=2
#SBATCH --time=00:00:10
#SBATCH --output=mpi_simple.out
#SBATCH -A anakano_429
mpirun -n $SLURM_NTASKS ./mpi_simple
```

Slurm (Simple Linux Utility for Resource Management): Open-source job scheduler that allocates compute resources on clusters for queued jobs

Class project account; type `myaccount` to check all accounts

Total number of processes = ntasks-per-node × nodes

`srun --mpi=pmi2 instead of mpirun recommended by CARC`

Submit a Slurm job

```
discovery: sbatch mpi_simple.sl
Submitted batch job 63695
```

Check the status of a Slurm job

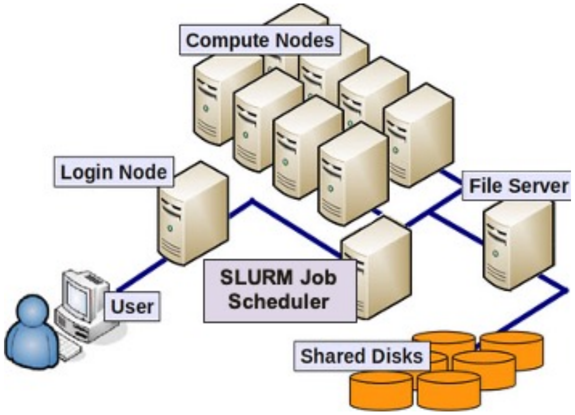
```
discovery: squeue -u anakano
JOBID PARTITION       NAME     USER ST        TIME  NODES  NODELIST(REASON)
63695      main  mpi_simple  anakano PD        0:00      1  (Resources)
```

Cancel a Slurm job

```
discovery: scancel 63695
```

Check the output

```
discovery: more mpi_simple.out
n = 777
```



For detailed explanation, see the lecture note

https://aiichironakano.github.io/cs653/02MPI.pdf

Interactive Job at CARC

When debugging your MPI program, you may want to access computing nodes interactively, so that you can edit, compile & run MPI program in real time unlike the batch job

Reserve 2 processors for 20 minutes

```
[anakano@discovery cs653]$ salloc -n 2 -t 20
salloc: Granted job allocation 63754
salloc: Waiting for resource configuration
salloc: Nodes d05-05 are ready for job
[anakano@d05-05 cs653]$ mpirun -n 2 ./mpi_simple
n = 777
[anakano@d05-05 cs653]$ exit
exit
salloc: Relinquishing job allocation 63754
[anakano@discovery cs653]$
```

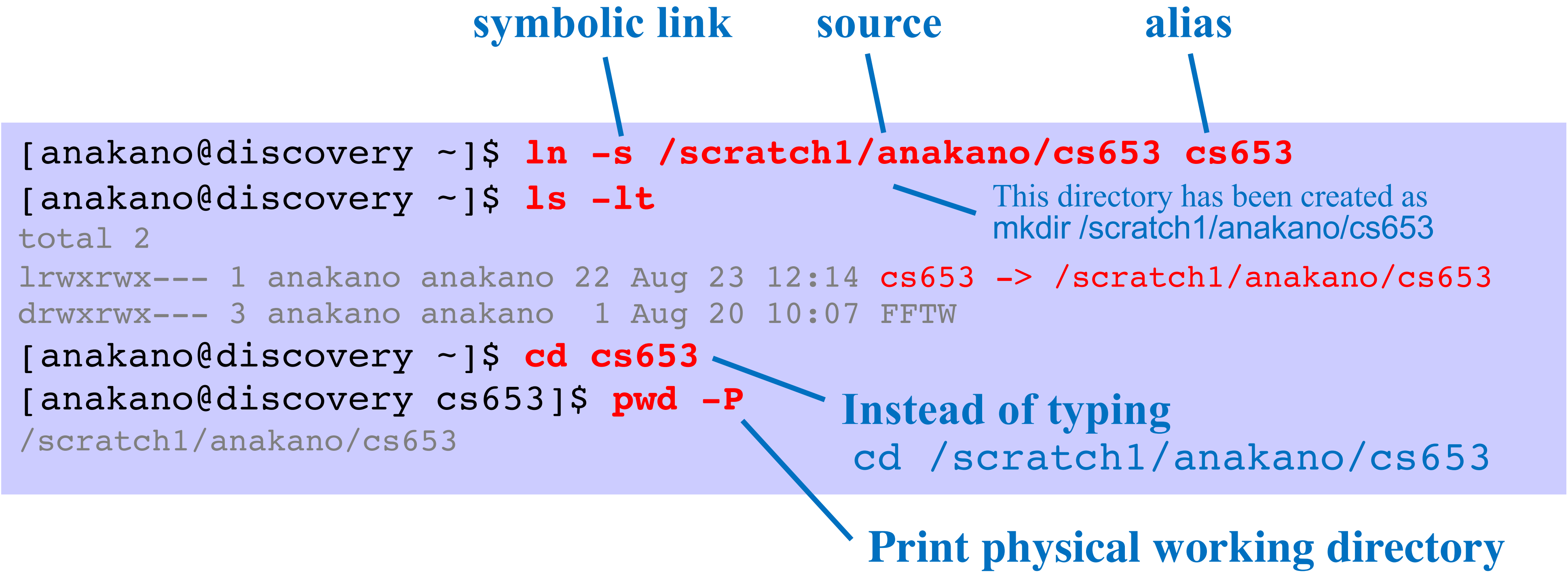
Note you are now using a computing node named d05-05

Back to the login node

Type `less /proc/cpuinfo` **to find what kind of node you got**

Symbolic Link to Work Directory

- Your home directory has very small quota (type `myquota` to confirm), so please use the scratch file system (`/scratch1/anakano` for user `anakano`) instead
- It is convenient to make a symbolic link to a directory you use often, rather than typing its long absolute path every time



```
[anakano@discovery ~]$ ln -s /scratch1/anakano/cs653 cs653
[anakano@discovery ~]$ ls -lt
total 2
lrwxrwx--- 1 anakano anakano 22 Aug 23 12:14 cs653 -> /scratch1/anakano/cs653
drwxrwx--- 3 anakano anakano  1 Aug 20 10:07 FFTW
[anakano@discovery ~]$ cd cs653
[anakano@discovery cs653]$ pwd -P
/scratch1/anakano/cs653
```

This directory has been created as
mkdir /scratch1/anakano/cs653

Instead of typing
`cd /scratch1/anakano/cs653`

Print physical working directory

File Transfer

- Use secure file transfer protocol to transfer files between your laptop and Discovery

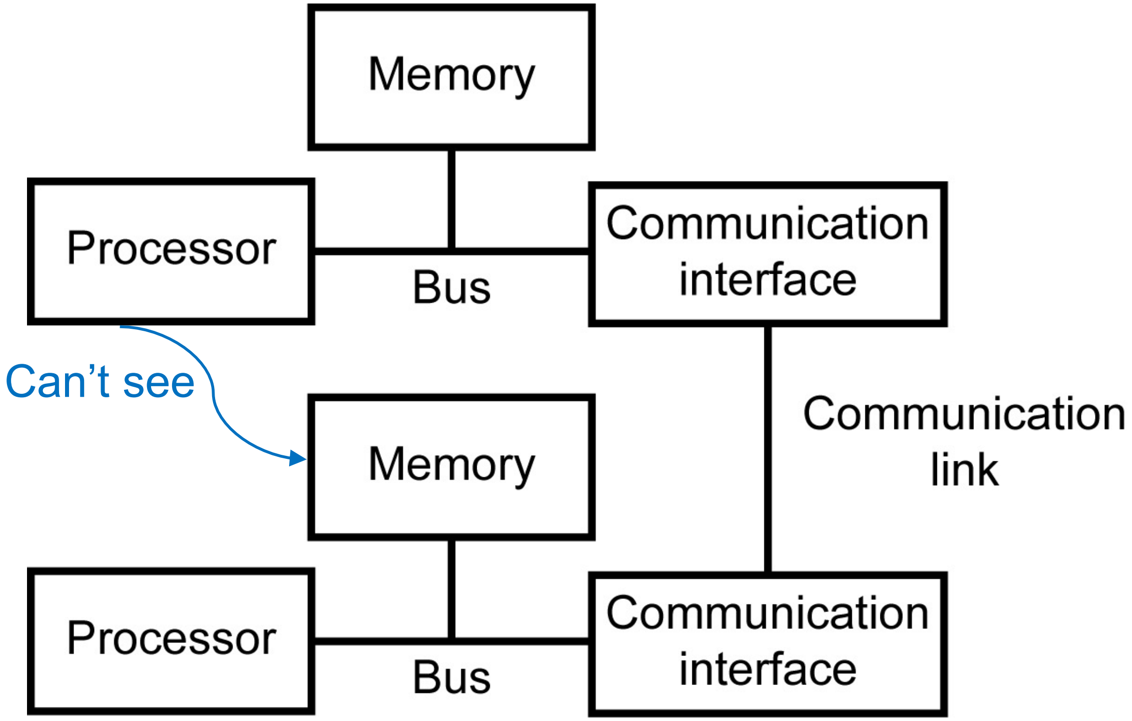
```
macbook-pro $ sftp anakano@discovery.usc.edu
Connected to discovery.usc.edu.
sftp> cd cs653
sftp> put md.*
sftp> ls
md.c    md.h    md.in
sftp> exit
macbook-pro $
```

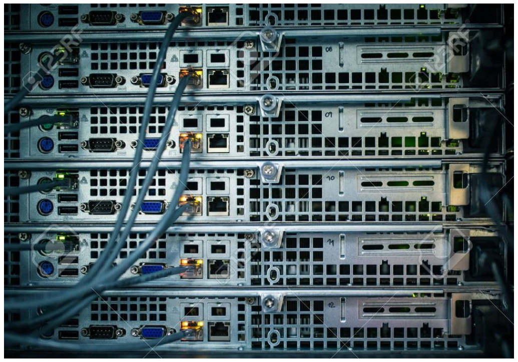
Transfer files from local computer (your laptop) to remote computer (Discovery)

Check whether the files have been transferred

- To transfer files from remote computer to local computer, use `get` instead

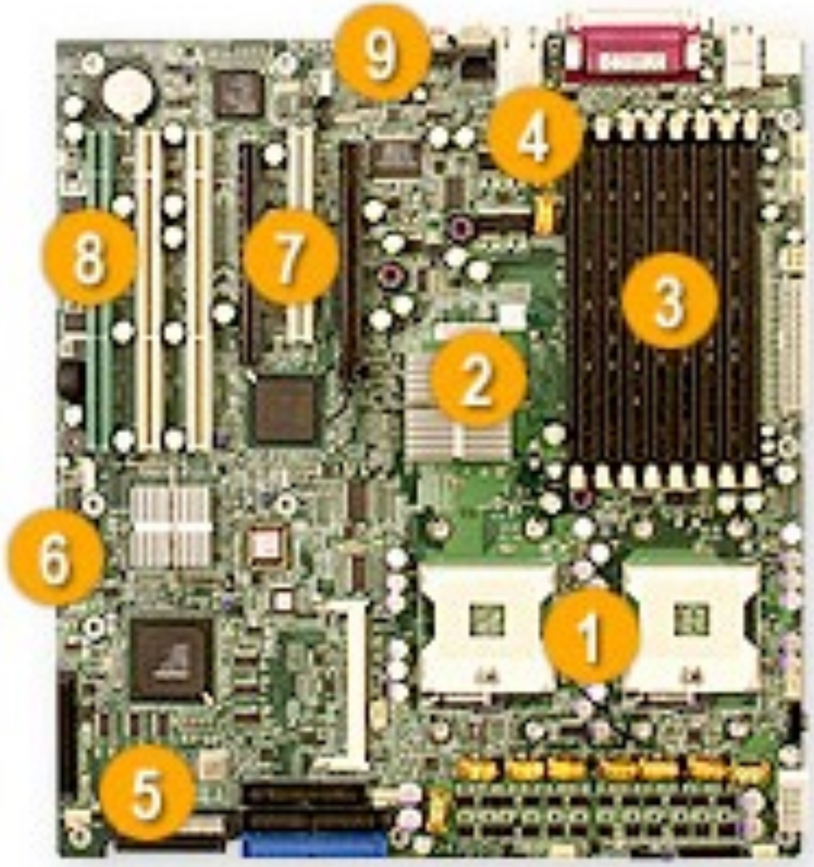
Parallel Computing Hardware

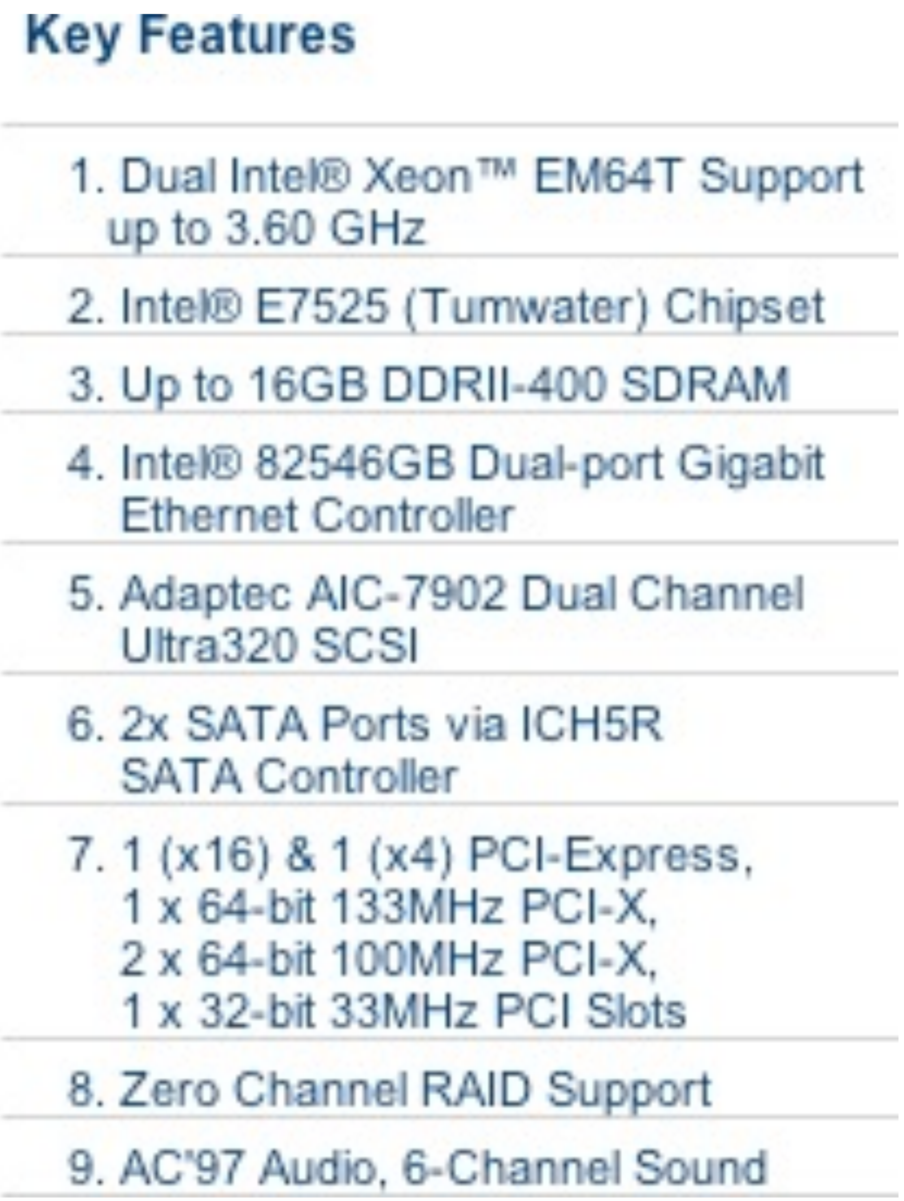




- **Processor**: Executes arithmetic & logic operations.
- **Memory**: Stores program & data.
- **Communication interface**: Performs signal conversion & synchronization between communication link and a computer.
- **Communication link**: A wire capable of carrying a sequence of bits as electrical (or optical) signals.

Motherboard





Key Features

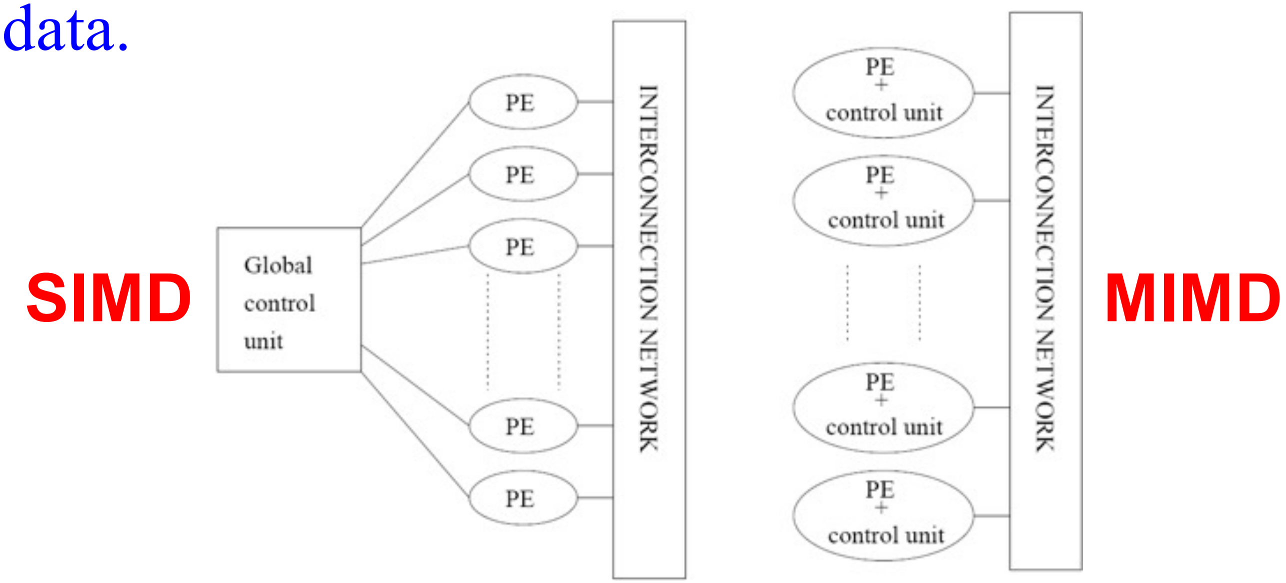
1. Dual Intel® Xeon™ EM64T Support up to 3.60 GHz
2. Intel® E7525 (Tumwater) Chipset
3. Up to 16GB DDRII-400 SDRAM
4. Intel® 82546GB Dual-port Gigabit Ethernet Controller
5. Adaptec AIC-7902 Dual Channel Ultra320 SCSI
6. 2x SATA Ports via ICH5R SATA Controller
7. 1 (x16) & 1 (x4) PCI-Express,
   1 x 64-bit 133MHz PCI-X,
   2 x 64-bit 100MHz PCI-X,
   1 x 32-bit 33MHz PCI Slots
8. Zero Channel RAID Support
9. AC'97 Audio, 6-Channel Sound

Supermicro X6DA8-G2

Parallel Computing Platforms (1)

Control structures

- **Single-instruction multiple-data (SIMD)**: A single control unit dispatches instruction to each processing element (PE).
- **Multiple-instruction multiple-data (MIMD)**: Different processing elements can execute different instructions on different data.
- **Single-program multiple-data (SPMD)**: A simple variant of MIMD; multiple instances of the same program execute on different data.

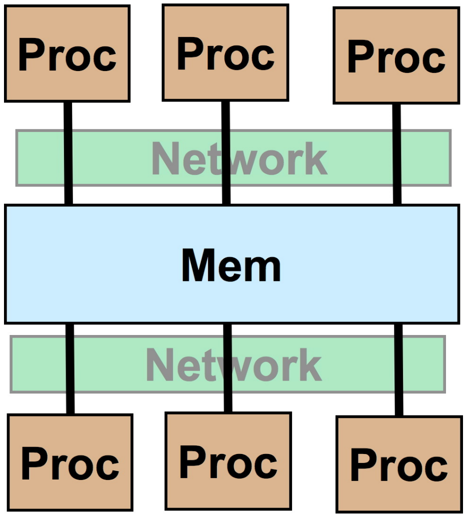


Grama’03, Chap. 2

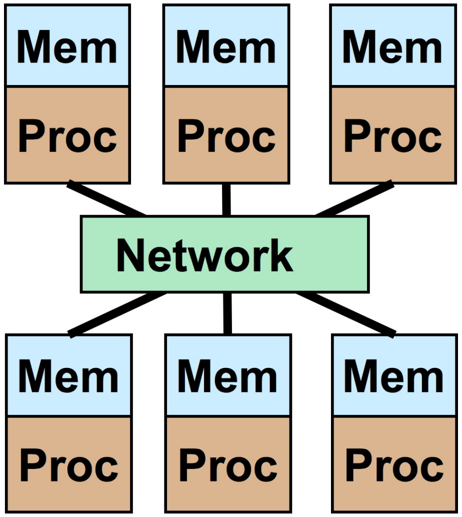
Parallel Computing Platforms (2)

Communication model

- **Shared-address-space platform (multiprocessor)**: Supports a common data space that is accessible to all processors.
  – **Uniform memory access (UMA)**: Time taken by a processor to access any memory word is identical
  – **Nonuniform memory access (NUMA)**: Time taken to access certain memory words is longer than others.
- **Message-passing platform (multicomputer)**: Consists of multiple processing nodes each with its own address space.



Multiprocessor

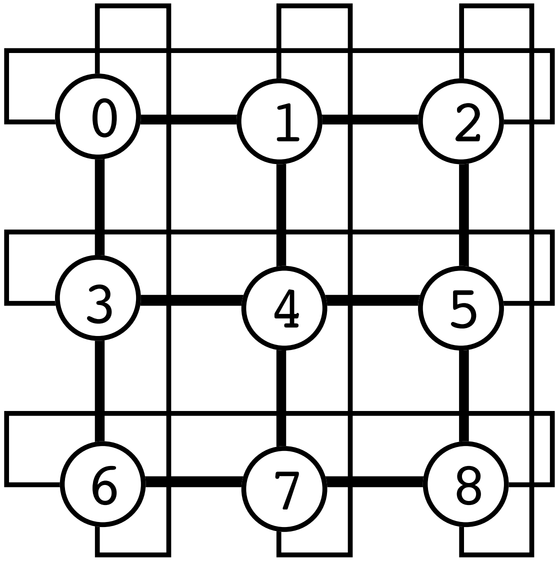


Multicomputer

Grama'03, Chap. 2

Communication Network

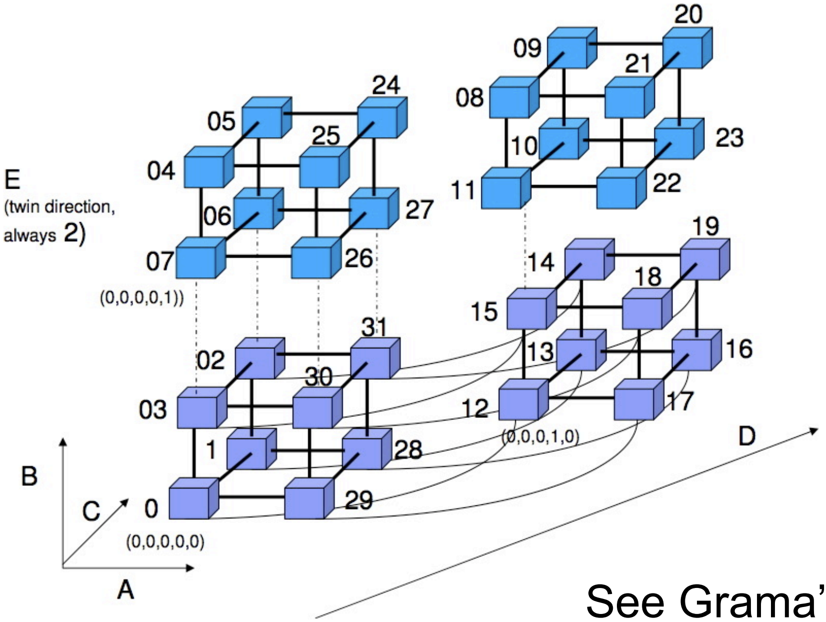
Mesh (torus)



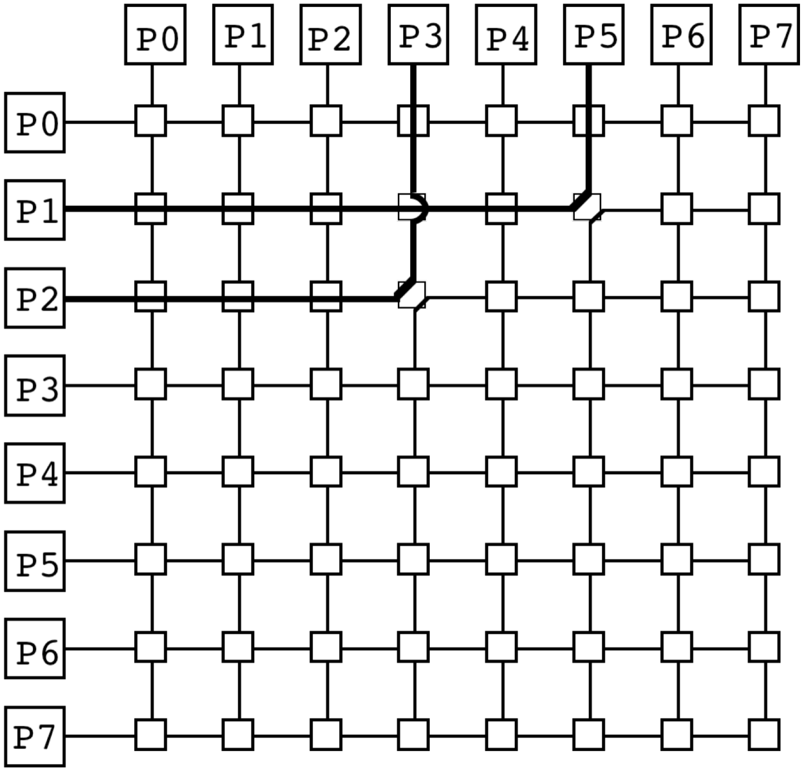
Wiring of interconnection network cables
Earth Simulator Research and Development Center

NEC Earth Simulator (640x640 crossbar)

IBM Blue Gene/Q (5D torus)



Crossbar switch



See Grama'03, Chap. 2

Parallel Programming

MPI: Message Passing Interface

- Standard programming language for multicomputers based on message passing
- Review the rest of the slides & detailed notes

  https://aiichironakano.github.io/cs653/02MPI.pdf

```
MPI_Send(), MPI_Recv()
```

OpenMP: Open specifications for Multi Processing

- Portable application program interface (API) for shared-memory parallel programming on multiprocessors based on multithreading by compiler directives
- Review the slides

  https://aiichironakano.github.io/cs653/02-02OpenMP-slide.pdf

```
#pragma omp parallel
```

Message Passing Interface

MPI (Message Passing Interface)

A standard message passing system that enables us to write & run applications on parallel computers

Download for Unix & Windows:

```
http://www.mcs.anl.gov/mpi/mpich
```

Compile

```
> mpicc -o mpi_simple mpi_simple.c
```

Run (srun is Slurm dialect)

```
> mpirun -np 2 mpi_simple
```

MPI Programming

mpi_simple.c: Point-to-point message send & receive

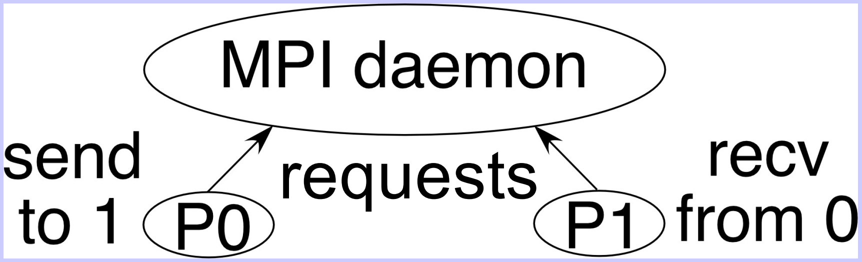
```c
#include "mpi.h"
#include <stdio.h>
int main(int argc, char *argv[]) {
  MPI_Status status;
  int myid;
  int n;
  MPI_Init(&argc, &argv);
  MPI_Comm_rank(MPI_COMM_WORLD, &myid);
  if (myid == 0) {
    n = 777;
    MPI_Send(&n, 1, MPI_INT, 1, 10, MPI_COMM_WORLD);
  }
  else {
    MPI_Recv(&n, 1, MPI_INT, 0, 10, MPI_COMM_WORLD, &status);
    printf("n = %d\n", n);
  }
  MPI_Finalize();
  return 0;
}
```

MPI rank

Matching message labels

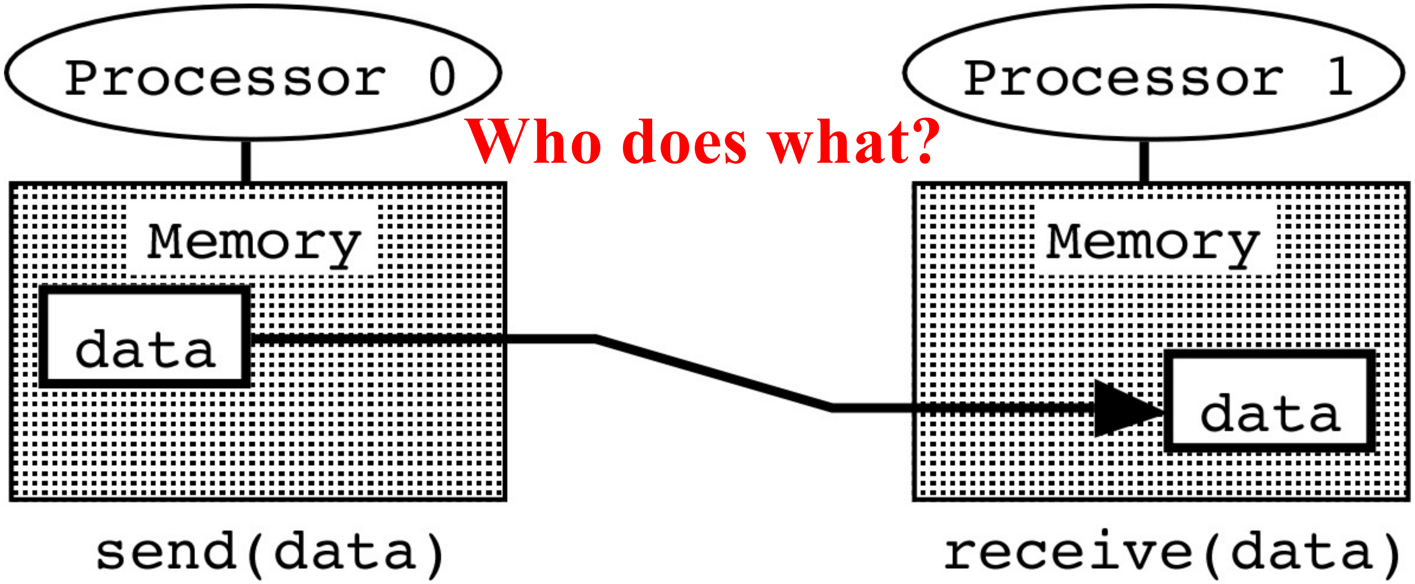
Data triplet

To/from whom



cf. p. 3 of https://aiichironakano.github.io/cs653/02MPI.pdf

Single Program Multiple Data (SPMD)



Who does what?

Process 0

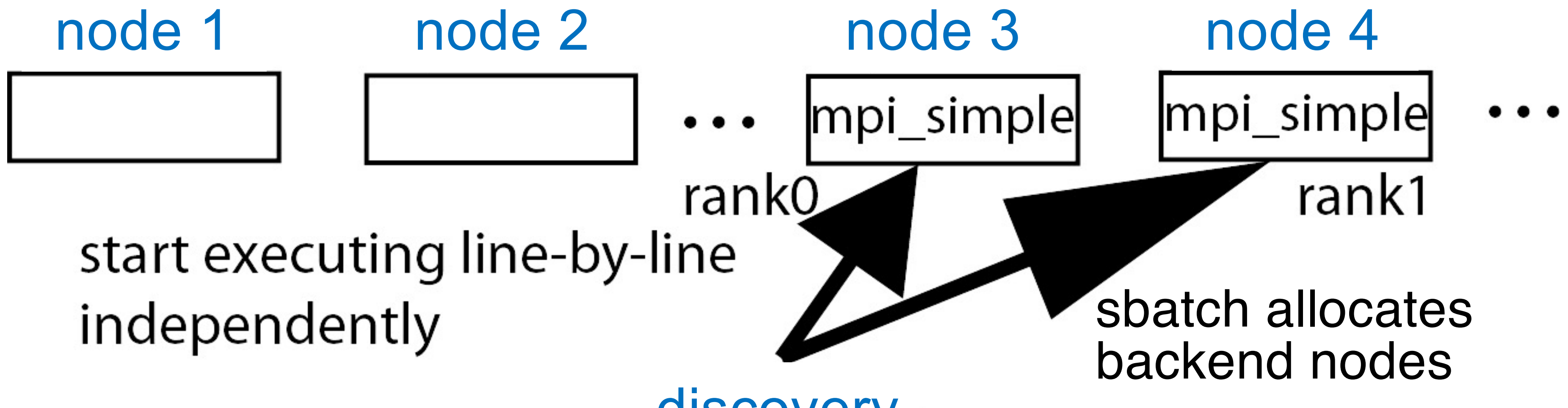
```
if (myid == 0) {
  n = 777;
  MPI_Send(&n,...);
}
else {
  MPI_Recv(&n,...);
  printf(...);
}
```

Process 1

```
if (myid == 0) {
  n = 777;
  MPI_Send(&n,...);
}
else {
  MPI_Recv(&n,...);
  printf(...);
}
```

Single Program Multiple Data (SPMD)

What really happens?



start executing line-by-line independently

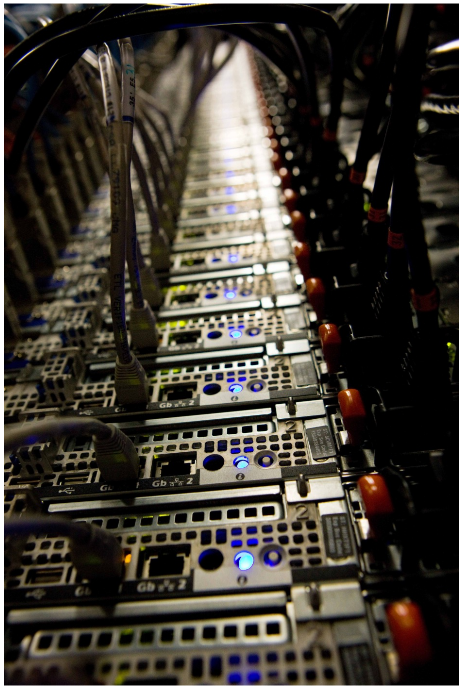
sbatch allocates backend nodes

discovery

```
%mpirun -n  2 mpi_simple
```

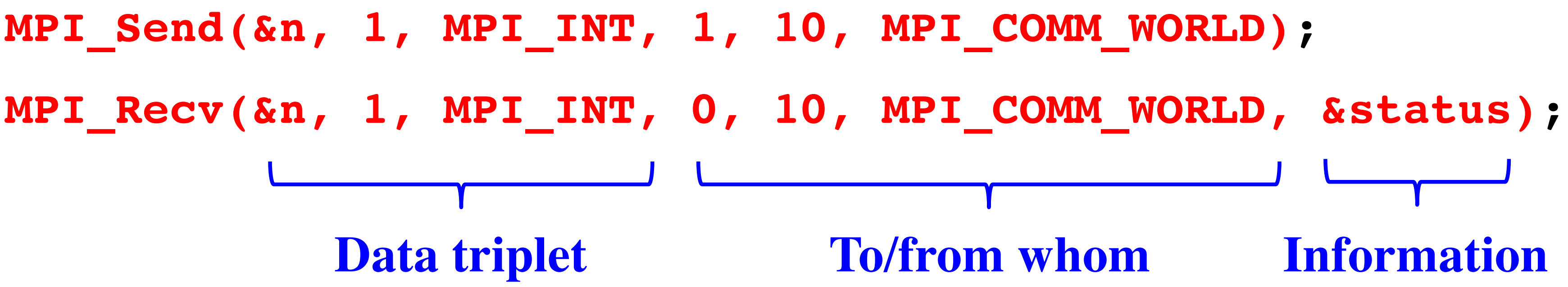
```
%ssh discovery.usc.edu
```

My laptop



MPI Minimal Essentials

We only need `MPI_Send()` & `MPI_Recv()`
within `MPI_COMM_WORLD`



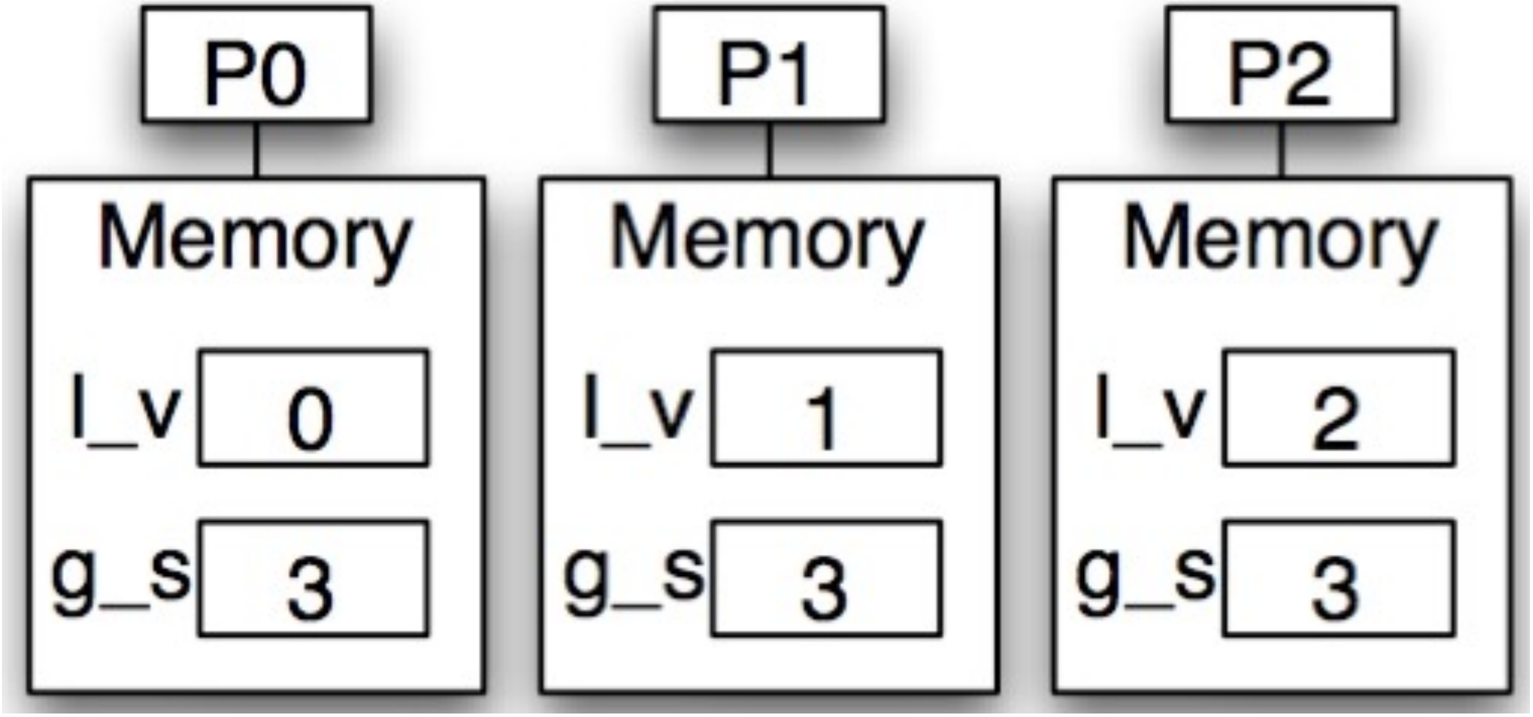
```
MPI_Send(&n, 1, MPI_INT, 1, 10, MPI_COMM_WORLD);
MPI_Recv(&n, 1, MPI_INT, 0, 10, MPI_COMM_WORLD, &status);
```

Global Operation

All-to-all reduction: Each process contributes a partial value to obtain the global summation. In the end, all the processes will receive the calculated global sum.

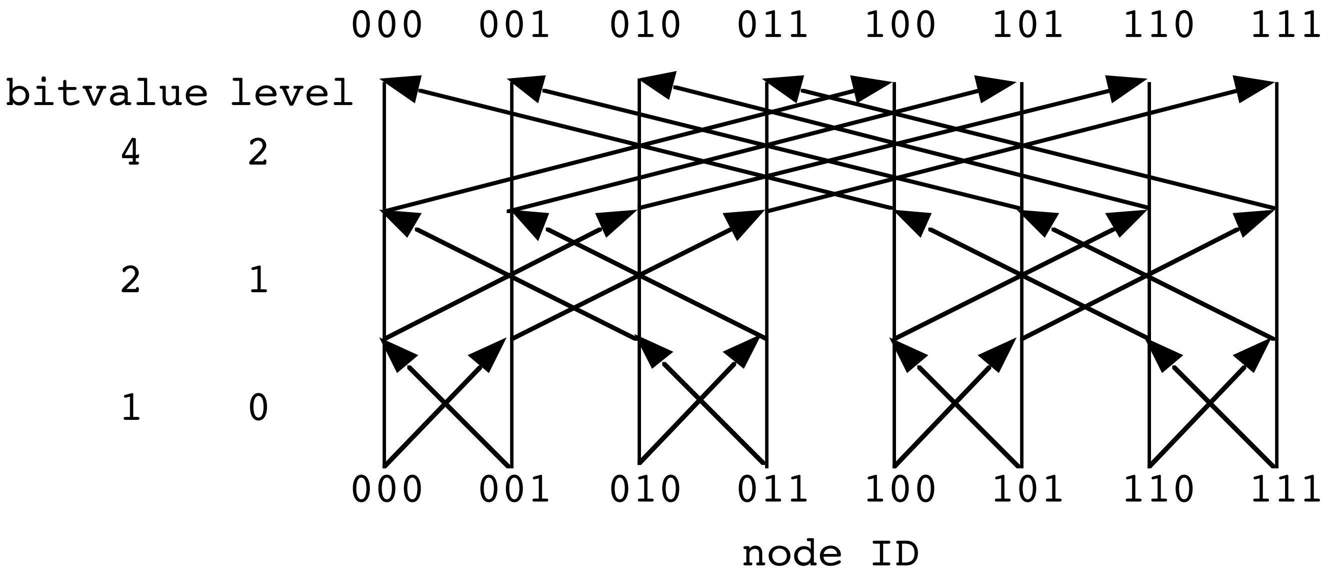
```
MPI_Allreduce(&local_value, &global_sum, 1, MPI_INT, MPI_SUM,
MPI_COMM_WORLD)
```

```
int l_v, g_s;  // local variable & global sum
l_v = myid;    // myid is my MPI rank
MPI_Allreduce(&l_v, &g_s, 1, MPI_INT, MPI_SUM, MPI_COMM_WORLD);
```



Hypercube Algorithm

Hypercube algorithm: Communication of a reduction operation is structured as a series of pairwise exchanges, one with each neighbor in a hypercube (**butterfly**) structure. Allows a computation requiring all-to-all communication among p processes to be performed in $\log_2 p$ steps.

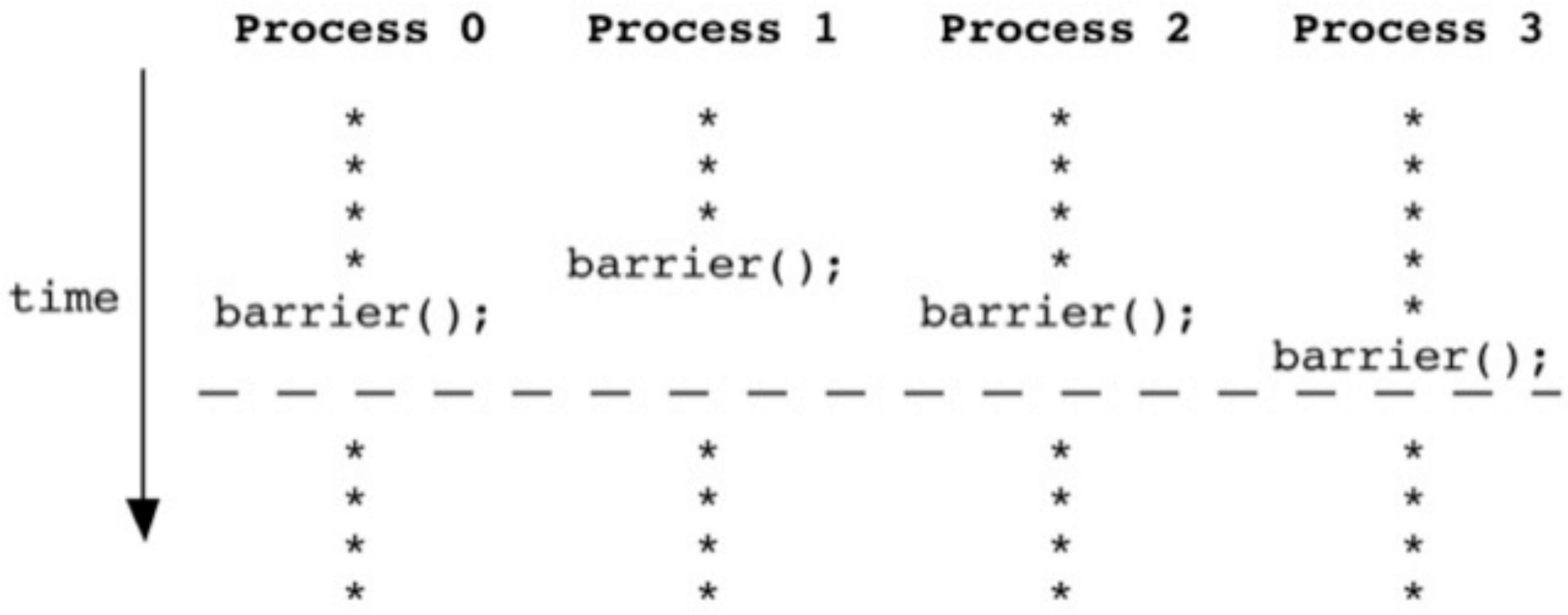


Butterfly network

```
a000 + a001 + a010 + a011 + a100 + a101 + a110 + a111
=  ((a000 + a001) + (a010 + a011))
+  ((a100 + a101) + (a110 + a111))
②                 ①         ⓪
```

Barrier

```
<A>;
barrier();
<B>;
```



MPI_Barrier(MPI_Comm communicator)

Useful for debugging (but would slow down the program)

MPI Communication

MPI communication functions:

1. Point-to-point

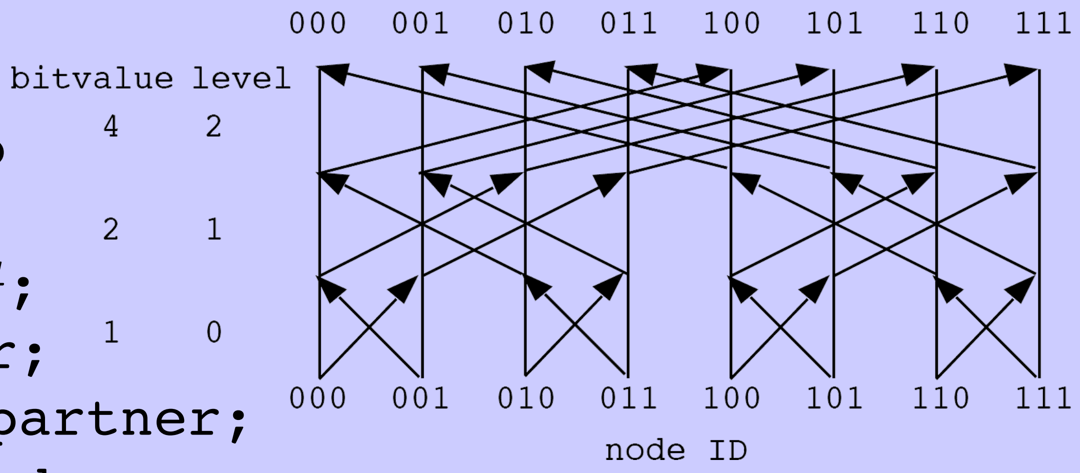
```
MPI_Send()
MPI_Recv()
```

2. Global

```
MPI_Allreduce()
MPI_Barrier()
MPI_Bcast()
```

Hypercube Template

```
procedure hypercube(myid, input, log$_2$P, output)
begin
  mydone := input;
  for $l$ := 0 to log$_2$P-1 do
  begin
    partner := myid XOR $2^l$;
    send mydone to partner;
    receive hisdone from partner;
    mydone = mydone OP hisdone
  end
  output := mydone
end
```



level l	2^l bitvalue
0	001
1	010
2	100

Associative operator
(*e.g.*, sum, max)

(*a* OP *b*) OP *c* = *a* OP (*b* OP *c*)

Exclusive OR

a	b		a XOR b
0	0	b	0
0	1		1
1	0	$\bar{b}$	1
1	1		0

abcdefg XOR 0000100 = abcd$\bar{e}$fg

In C, ^ (caret operator) is bitwise XOR applied to int

Driver for Hypercube Test

```
#include "mpi.h"
#include <stdio.h>
int nprocs;  /* Number of processes */
int myid;    /* My rank */

double global_sum(double partial) {
  /* Implement your own global summation here */
}

int main(int argc, char *argv[]) {
  double partial, sum, avg;
  MPI_Init(&argc, &argv);
  MPI_Comm_rank(MPI_COMM_WORLD, &myid);
  MPI_Comm_size(MPI_COMM_WORLD, &nprocs);
  partial = (double) myid;
  printf(”Rank %d has %le\n", myid, partial);
  sum = global_sum(partial);
  if (myid == 0) {
    avg = sum/nprocs;
    printf("Global average = %d\n", avg);
  }
  MPI_Finalize();
  return 0;
}
```

Who am I?

How big is the world? (see p. 5 in lecture note)

Sample Slurm Script

Run two MPI runs in a single Slurm job

```
#!/bin/bash
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=4
#SBATCH --time=00:00:59
#SBATCH --output=global.out
#SBATCH -A anakano_429

mpirun -n $SLURM_NTASKS ./global_avg
mpirun -n 4 ./global
```

`mpicc —o global_avg global_avg.c`

Total number of processors
= ntasks-per-node (4) × nodes (2) = 8

- **Type** `sbatch global_avg.sl` **in the directory where the executable** `global_avg` **resides, or** `cd` **(change directory) to where it is**

Output of `global.c`

- 4-processor job

```
Rank 0 has 0.000000e+00
Rank 1 has 1.000000e+00
Rank 2 has 2.000000e+00
Rank 3 has 3.000000e+00
Global average = 1.500000e+00
```

- 8-processor job

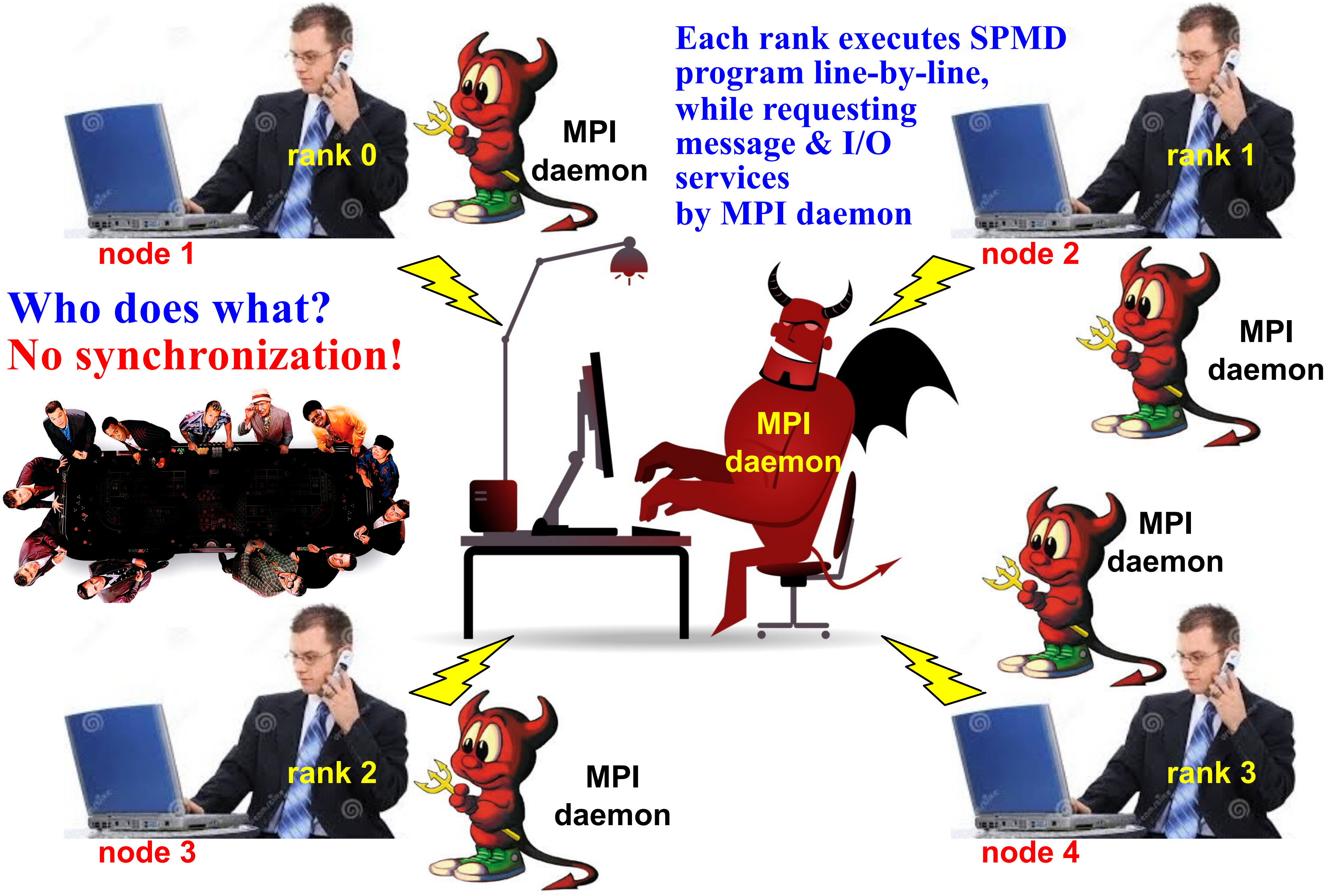
```
Rank 0 has 0.000000e+00
Rank 1 has 1.000000e+00
Rank 2 has 2.000000e+00
Rank 3 has 3.000000e+00
Rank 5 has 5.000000e+00
Rank 6 has 6.000000e+00
Rank 4 has 4.000000e+00
Rank 7 has 7.000000e+00
Global average = 3.500000e+00
```

Actual output is random order in ranks — Why?

References on Hypercube Algorithms

1. https://en.wikipedia.org/wiki/Hypercube_(communication_pattern)
2. I. Foster, *Designing and Building Parallel Programs* (Addison-Wesley, 1995) Chap. 11 — Hypercube algorithms: https://www.mcs.anl.gov/~itf/dbpp/text/node123.html

Distributed-Memory Parallel Computing



Each rank executes SPMD program line-by-line, while requesting message & I/O services by MPI daemon

Who does what?

No synchronization!

Communicator

mpi_comm.c: Communicator = process group + context

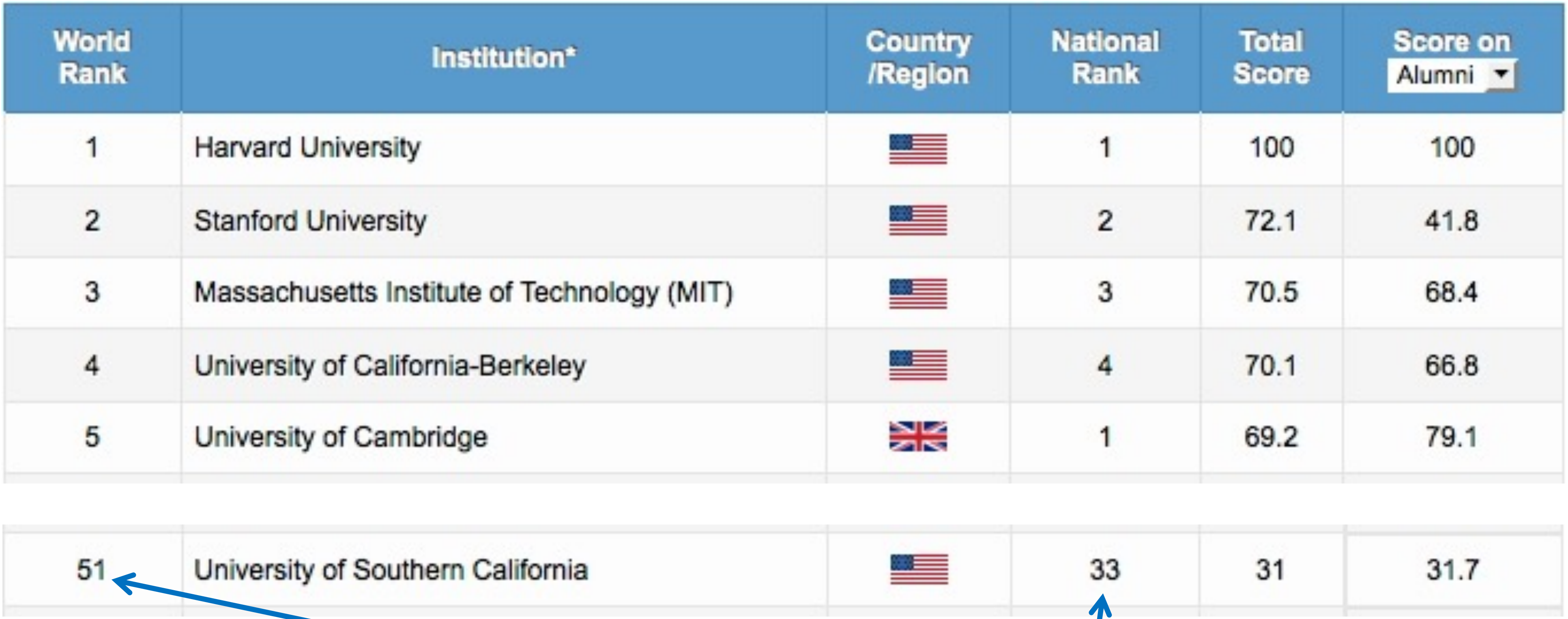
```
#include "mpi.h"
#include <stdio.h>
#define N 64
int main(int argc, char *argv[]) {
  MPI_Comm world, workers;
  MPI_Group world_group, worker_group;
  int myid, nprocs;
  int server, n = -1, ranks[1];
  MPI_Init(&argc, &argv);
  world = MPI_COMM_WORLD;
  MPI_Comm_rank(world, &myid);
  MPI_Comm_size(world, &nprocs);
  server = nprocs-1;
  MPI_Comm_group(world, &world_group);
  ranks[0] = server;
  MPI_Group_excl(world_group, 1, ranks, &worker_group);
  MPI_Comm_create(world, worker_group, &workers);
  MPI_Group_free(&worker_group);
  if (myid != server)
    MPI_Allreduce(&myid, &n, 1, MPI_INT, MPI_SUM, workers);
  printf("process %2d: n = %6d\n", myid, n);
  MPI_Comm_free(&workers);
  MPI_Finalize();
  return 0;
}
```

Usage

- **Avoid accidental match of unintended Send-Receive pairs**
- **Global operations in a subgroup of processes**

Code at https://aiichironakano.github.io/cs596/src/mpi/
For detail, see p. 4 in https://aiichironakano.github.io/cs596/02MPI.pdf

Example: Ranks in Different Groups



World Rank	Institution*	Country /Region	National Rank	Total Score	Score on Alumni ▾
1	Harvard University	🇺🇸	1	100	100
2	Stanford University	🇺🇸	2	72.1	41.8
3	Massachusetts Institute of Technology (MIT)	🇺🇸	3	70.5	68.4
4	University of California-Berkeley	🇺🇸	4	70.1	66.8
5	University of Cambridge	🇬🇧	1	69.2	79.1
51	University of Southern California	🇺🇸	33	31	31.7

```
MPI_Comm_rank(world, &usc_world);
MPI_Comm_rank(us, &usc_national);
```

Rank is relative in each communicator!

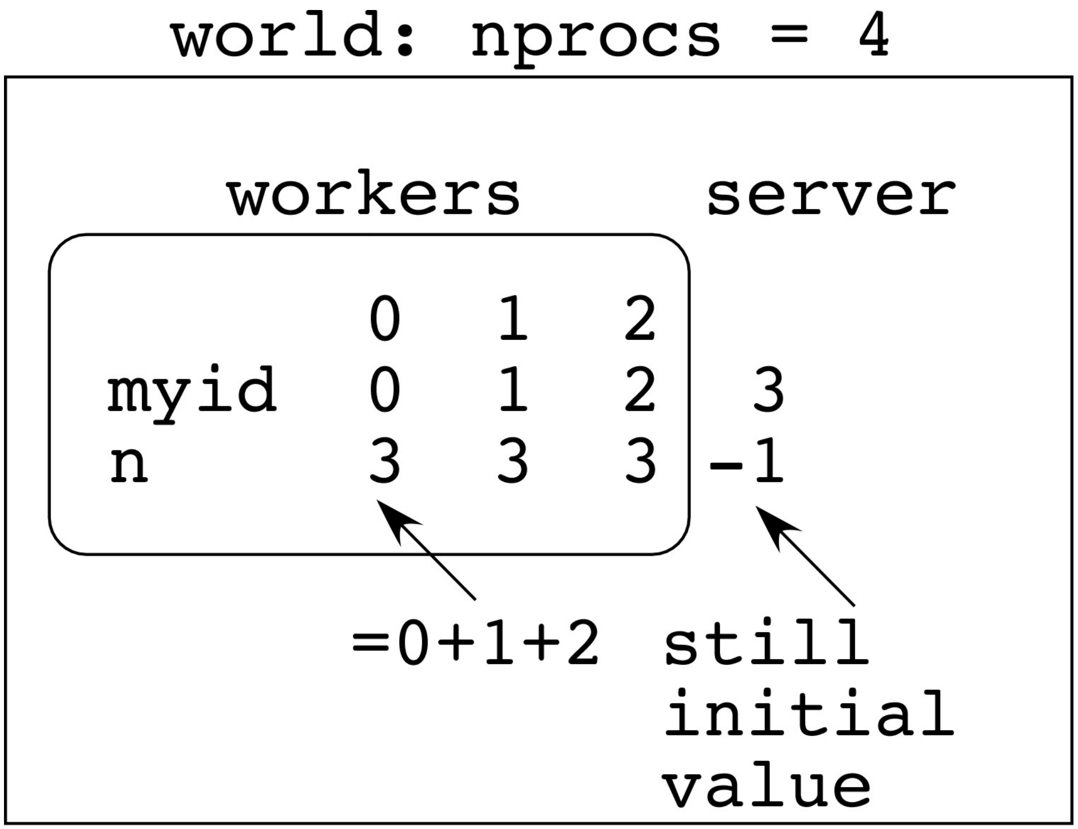
Output from mpi_comm.c

Slurm script

```
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=2
...
mpirun -n $SLURM_NTASKS./mpi_comm
```

```
process  3: n =      -1
process  0: n =       3
process  1: n =       3
process  2: n =       3
```

What Has Happened?



world: nprocs = 4

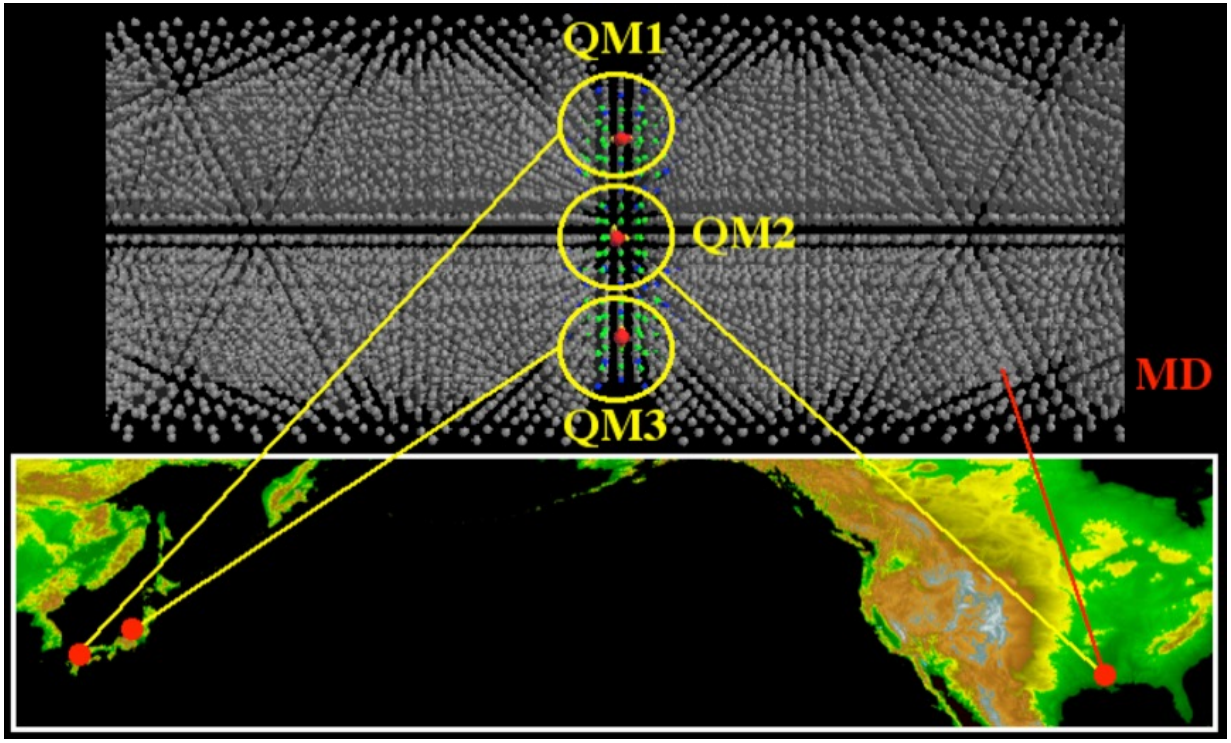
	workers			server
	0	1	2	
myid	0	1	2	3
n	3	3	3	-1

=0+1+2

still initial value

Grid Computing & Communicators

H. Kikuchi *et al*., “Collaborative simulation Grid: multiscale quantum-mechanical/classical atomistic simulations on distributed PC clusters in the US & Japan, *IEEE/ACM SC02*

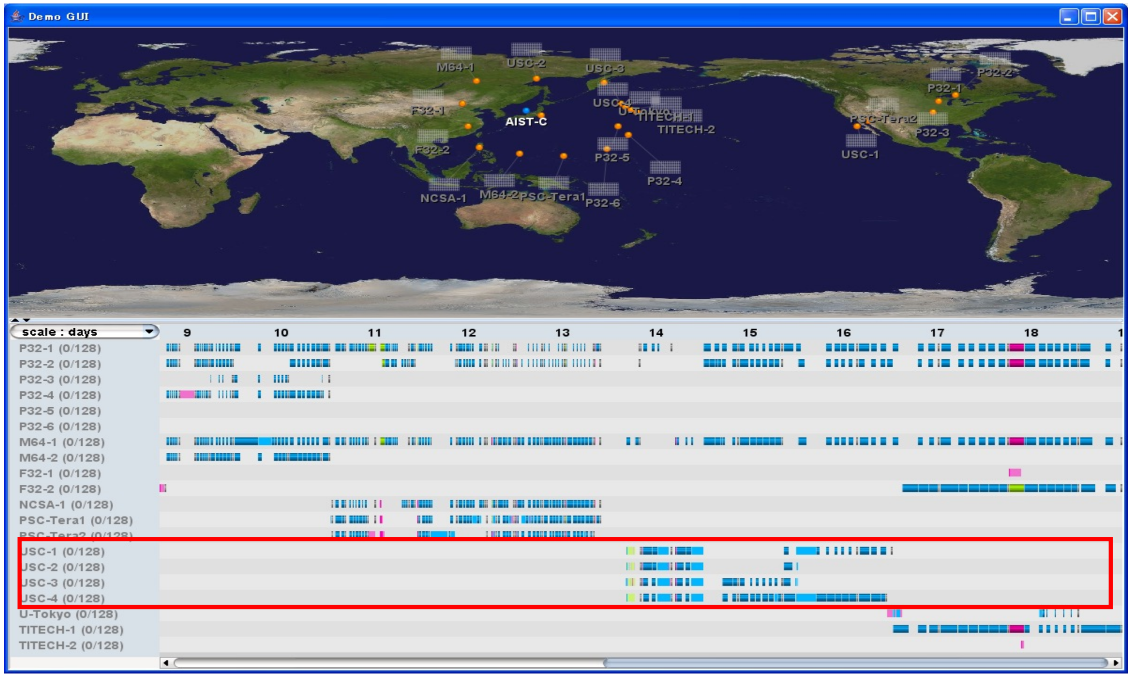


Communicator = a nice migration path to distributed computing

- **Single MPI program run with the Grid-enabled MPI implementation, MPICH-G2**
- **Processes are grouped into MD & QM groups by defining multiple MPI communicators as subsets of MPI_COMM_WORLD; a machine file assigns globally distributed processors to the MPI processes**

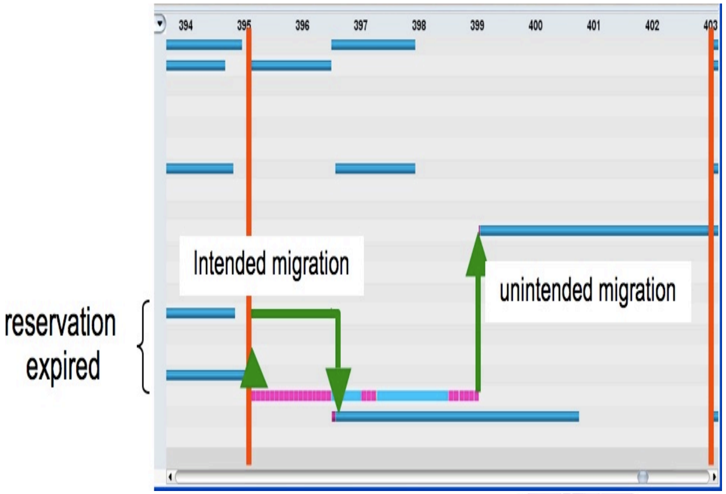
Global Grid QM/MD

- *One of the largest (153,600 cpu-hrs) sustained Grid supercomputing* **at 6 sites in the US (USC, Pittsburgh, Illinois) & Japan (AIST, U Tokyo, Tokyo IT)**



USC

Automated resource migration & fault recovery



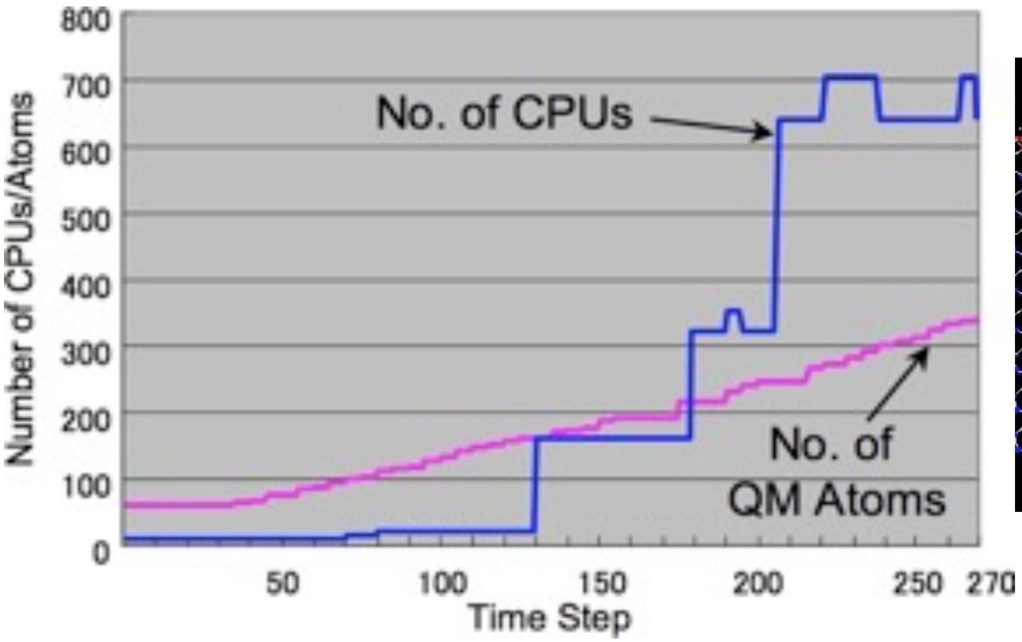
東京大学 THE UNIVERSITY OF TOKYO — AIST — SC06 POWERFUL BEYOND IMAGINATION — USC — NCSA — PITTSBURGH SUPERCOMPUTING CENTER

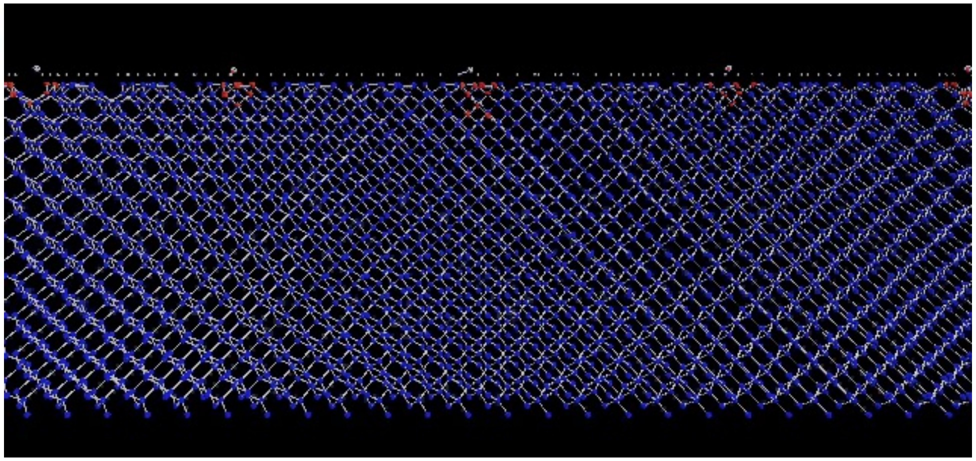
Takemiya *et al*., "Sustainable adaptive Grid supercomputing: multiscale simulation of semiconductor processing across the Pacific," *IEEE/ACM SC06*

Sustainable Grid Supercomputing

- **Sustained (> months) supercomputing (> 10^3 CPUs) on a Grid of geographically distributed supercomputers**
- **Hybrid Grid remote procedure call (GridRPC) + message passing (MPI) programming**
- **Dynamic allocation of computing resources on demand & automated migration due to reservation schedule & faults**

Ninf-G GridRPC: ninf.apgrid.org; MPICH: www.mcs.anl.gov/mpi

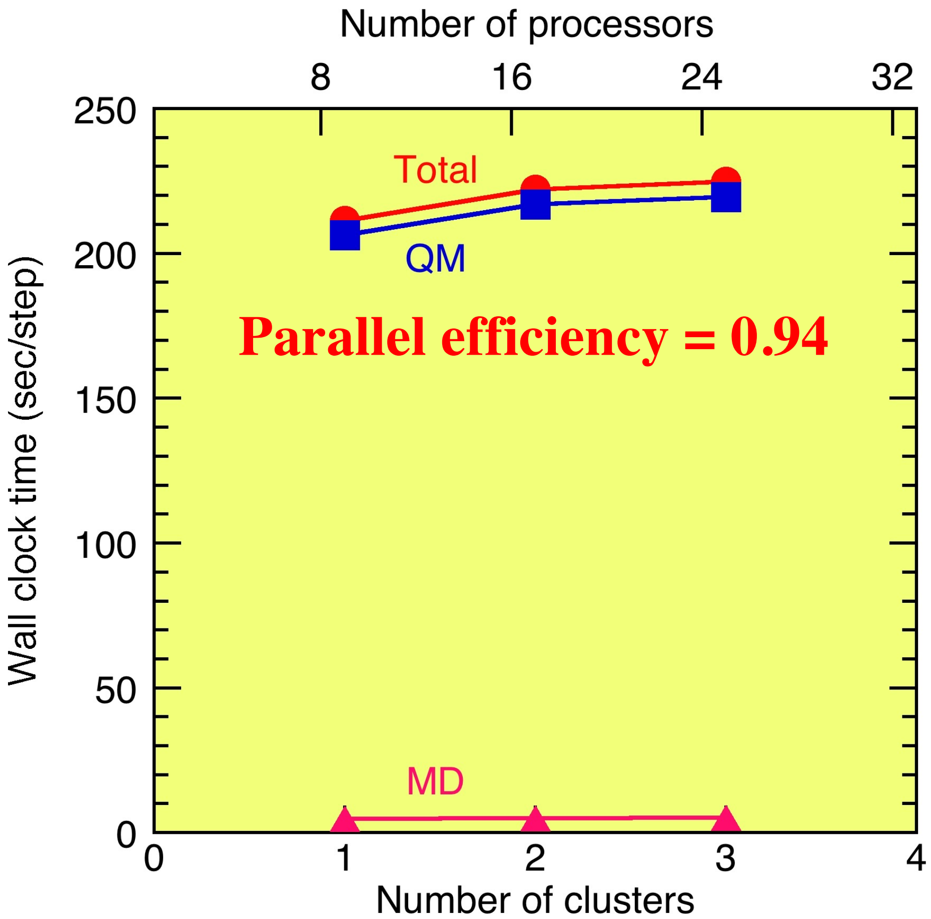




Multiscale QM/MD simulation of high-energy beam oxidation of Si

Computation-Communication Overlap

H. Kikuchi *et al*., "Collaborative simulation Grid: multiscale quantum-mechanical/classical atomistic simulations on distributed PC clusters in the US & Japan, *IEEE/ACM SC02*



$$\frac{\text{Earth's circumference}}{\text{Light speed}} = \frac{40{,}000\,[\mathrm{km}] = 4\times10^{7}[\mathrm{m}]}{3\times10^{8}[\mathrm{m}]} = 0.1\ \mathrm{s} = 100\ \mathrm{ms}$$

Try on Discovery:

```
traceroute www.u-tokyo.ac.jp
```

vs. `ping hpc-transfer.usc.edu`

- **How to overcome 200 ms latency & 1 Mbps bandwidth?**
- **Computation-communication overlap: To hide the latency, the communications between the MD & QM processors have been overlapped with the computations using asynchronous messages**

Synchronous Message Passing

`MPI_Send()`: (blocking), synchronous

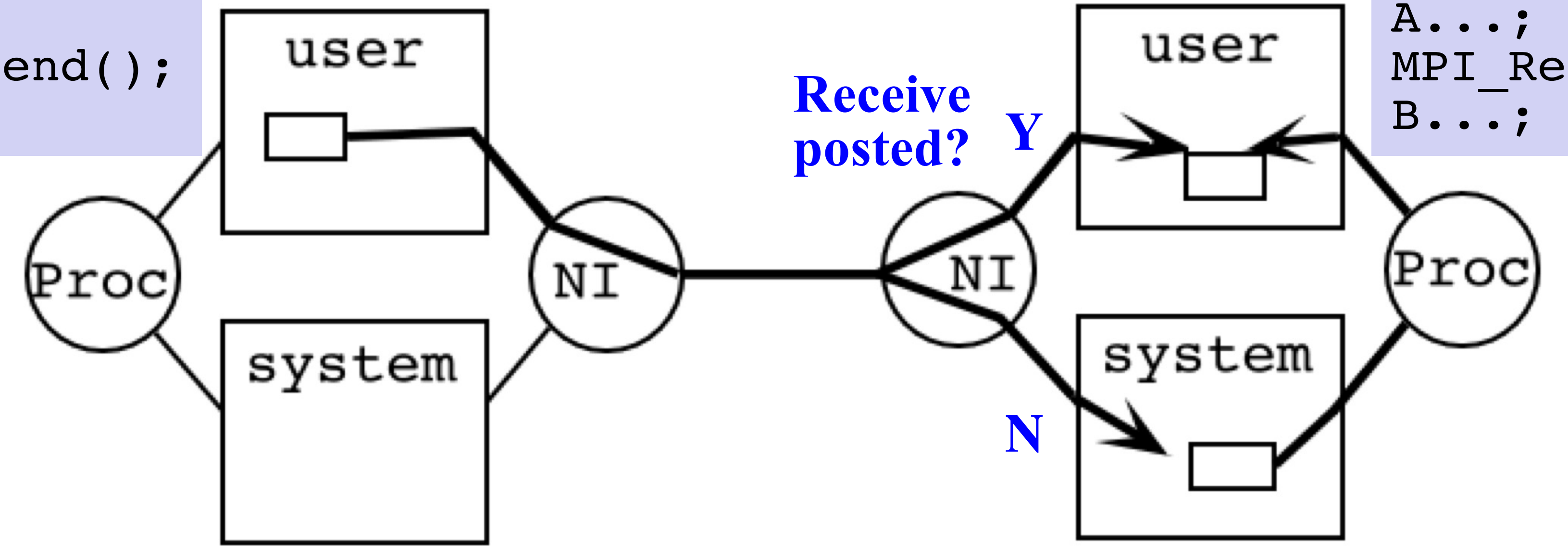
- Safe to modify original data immediately on return
- Depending on implementation, it may return whether or not a matching receive has been posted, or it may block (especially if no buffer space available)

`MPI_Recv()`: blocking, synchronous

- Blocks for message to arrive
- Safe to use data on return

Experienced a lot of blocking on iPSC/860 with 12 MB user & 4 MB system memory per node

```
A...;
MPI_Send();
B...;
```



```
A...;
MPI_Recv();
B...;
```

Asynchronous Message Passing

Allows computation-communication overlap

`MPI_Isend()`: non-blocking, asynchronous

- Returns immediately whether or not a matching receive has been posted
- Not safe to modify original data immediately (use `MPI_Wait()` system call)

`MPI_Irecv()`: non-blocking, asynchronous

- Does not block for message to arrive
- Cannot use data before checking for completion with `MPI_Wait()`

`MPI_Irecv()` is just a "request" for data delivery, when a matching message arrives

```
A...;
MPI_Isend();
B...;
MPI_Wait();
C...;  // Reuse the send buffer
```

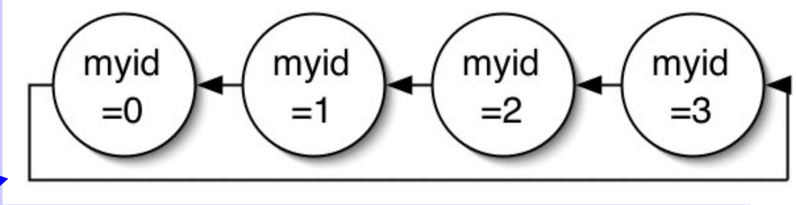
```
A...;
MPI_Irecv();
B...;  // Indep. of received message
MPI_Wait();
C...;  // Use the received message
```

See p. 6 in https://aiichironakano.github.io/cs653/02MPI.pdf

Program irecv_mpi.c

```
#include "mpi.h"
#include <stdio.h>
#define N 1000
int main(int argc, char *argv[]) {
  MPI_Status status;
  MPI_Request request;
  int send_buf[N], recv_buf[N];
  int send_sum = 0, recv_sum = 0;
  long myid, left, Nnode, msg_id, i;
  MPI_Init(&argc, &argv);
  MPI_Comm_rank(MPI_COMM_WORLD, &myid);
  MPI_Comm_size(MPI_COMM_WORLD, &Nnode);
  left  = (myid + Nnode - 1) % Nnode;
  for (i=0; i<N; i++) send_buf[i] = myid*N + i;
  MPI_Irecv(recv_buf, N, MPI_INT, MPI_ANY_SOURCE, 777, MPI_COMM_WORLD,
            &request); /* Post a receive */
  /* Perform tasks that don't use recv_buf */
  MPI_Send(send_buf, N, MPI_INT, left, 777, MPI_COMM_WORLD);
  for (i=0; i<N; i++) send_sum += send_buf[i];
  MPI_Wait(&request, &status); /* Complete the receive */
  /* Now it's safe to use recv_buf */
  for (i=0; i<N; i++) recv_sum += recv_buf[i];
  printf("Node %d: Send %d Recv %d\n", myid, send_sum, recv_sum);
  MPI_Finalize();
  return 0;
}
```

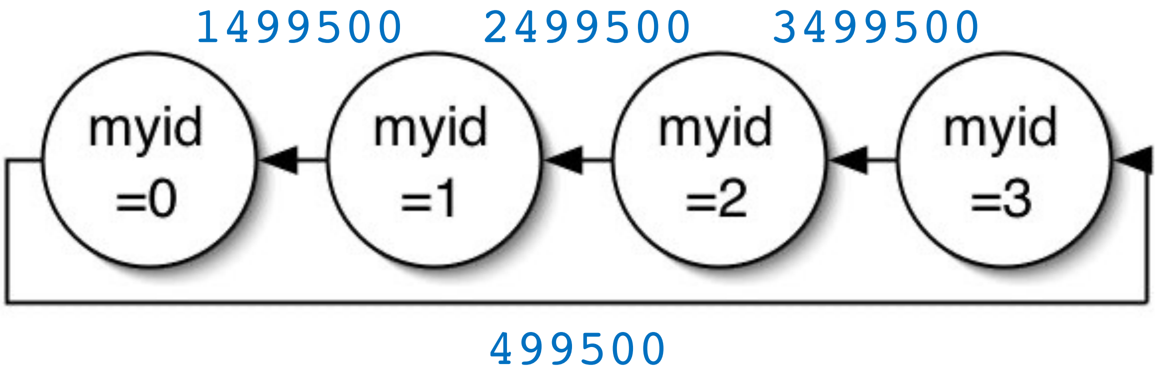
Wrap-around/torus
***via* modulo (%) operator**
(*cf.* periodic boundary condition)



Code at https://aiichironakano.github.io/cs596/src/mpi/

Output from `irecv_mpi.c`

```
Node 1: Send 1499500 Recv 2499500
Node 3: Send 3499500 Recv 499500
Node 0: Send 499500 Recv 1499500
Node 2: Send 2499500 Recv 3499500
```



Multiple Asynchronous Messages

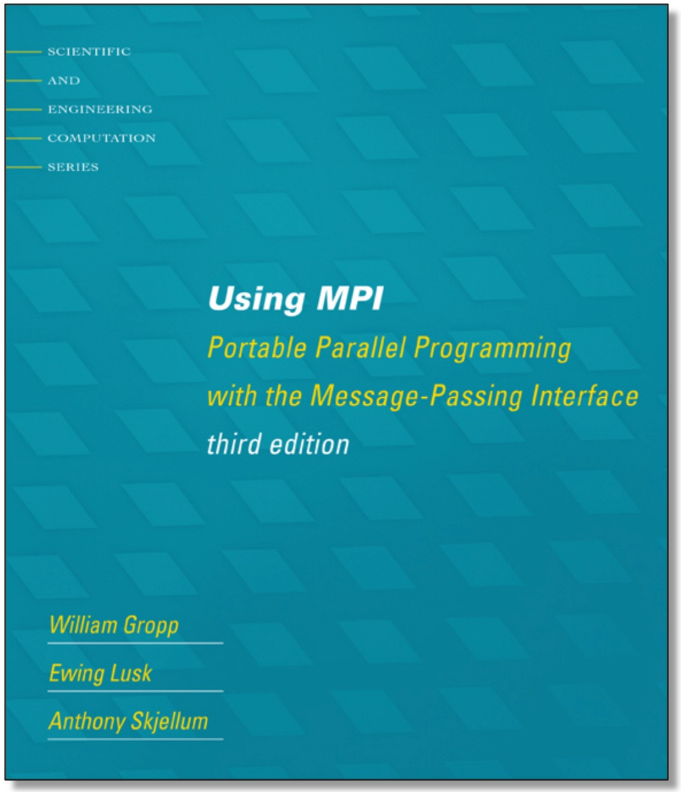
```
MPI_Request requests[N_message];
MPI_Status statuses[N_message];
MPI_Status status;
int index;

/* Wait for all messages to complete */
MPI_Waitall(N_message, requests, statuses);

/* Wait for any specified messages to complete */
MPI_Waitany(N_message, requests, &index, &status);
```

returns the index (∈ [0,N_message-1]) of the message that completed

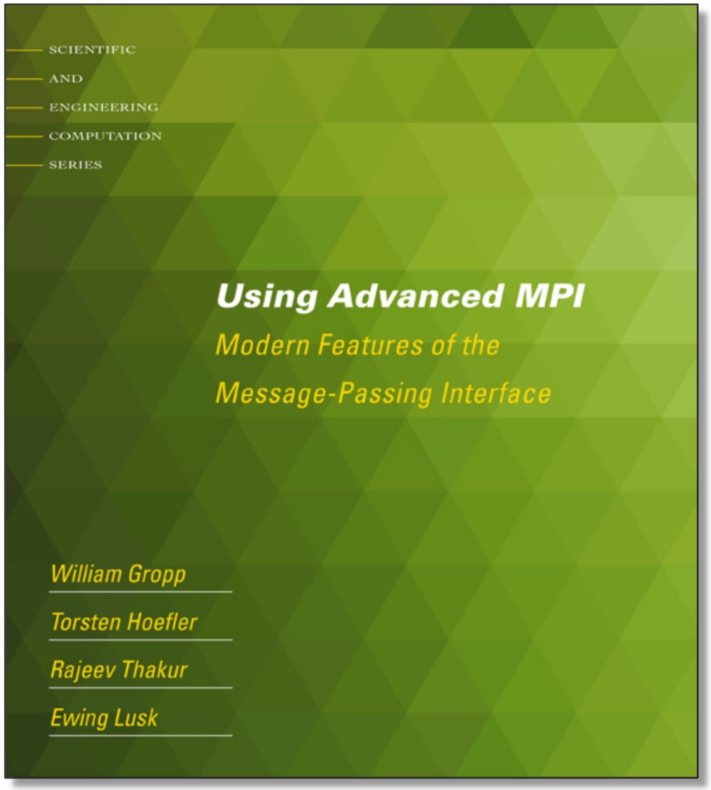
Polling MPI_Irecv

```
int flag;

/* Post an asynchronous receive */
MPI_Irecv(recv_buf, N, MPI_INT, MPI_ANY_SOURCE, 777,
          MPI_COMM_WORLD, &request);

/* Perform tasks that don't use recv_buf */
...

/* Polling */
MPI_Test(&request, &flag, &status); /* Check completion */
if (flag) { /* True if message received */
  /* Now it's safe to use recv_buf */
  ...
}
```

Where to Go from Here



Basic MPI



Advanced MPI, including MPI-3

- **Complete MPI reference at** http://www.netlib.org/utk/papers/mpi-book/mpi-book.html
- **MPI is evolving (MPI-2 to MPI-3) to include advanced features like remote memory access (`MPI_Put()` & `MPI_Get()`; *cf.* `sftp`), parallel I/O and dynamic process management**
- **Various versions of MPI standard are specified at** https://www.mpi-forum.org/docs/

See ATPESC 2002 lecture on scalable MPI

MPI Basics: Recap

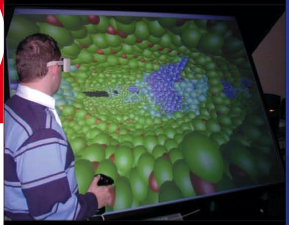
- **Parallel computing = Who does what**
- **Single program multiple data (SPMD) programming: Do it with MPI rank (who am I) & selection constructs (`if`, *etc*.)**
- **Only need `MPI_Send() & MPI_Recv()` within communicators to implement any distributed-memory parallel computing**
- **Asynchronous message passing (`MPI_Isend() & MPI_Irecv()`) to overlap computation & communication**
- **You can survive professionally only with a few global communication functions, *e.g*., `MPI_Allreduce()`, `MPI_Barrier() & MPI_Bcast()`**

Start using MPI for your research & projects!

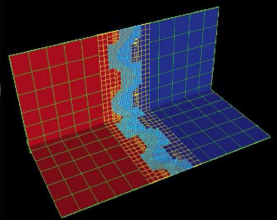
20 Years-Unleashing the Power of HPC

SC2001

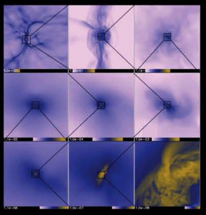
2001 Chair
Charles Slocomb
Denver, CO



A discrete particle simulation of 1.5 billion atoms



Adaptive mesh simulation of advecting sinusoidal density contours



Adaptive mesh simulation of star formation

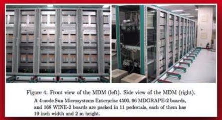
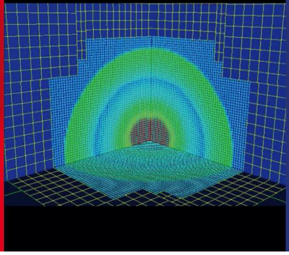


Figure 4: Front view of the MDM (left). Side view of the MDM (right). A 4-node Sun Microsystems Enterprise 4500, 96 MDGRAPE-2 boards, and 168 WINE-2 boards are packed in 11 pedestals, each of them has 19 inch width and 2 m height.

The MDM system



Adaptive mesh simulation of a spherical shock

2001

Notable Systems first mentioned this year in the proceedings:

- SGI Origin 3000
- Sun Fire 6000
- ASCI White
- Blue Horizon
- ASCI Blue Mountain

A WINE-2 system board

Notable Processors:

- MIPS R 12000
- Intel Pentium 4
- Intel Itanium

Noteworthy Architecture Topics:

- Cache coherence through snooping
- Application speedups through custom on-the-fly FPGA function units
- Interactive program steering
- Grid-enabled parallel computing

Notable Programming Languages:

- HDL
- PThreads

Research Machines:

- CPlant

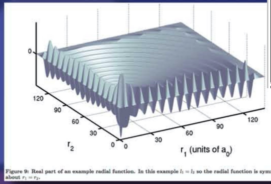
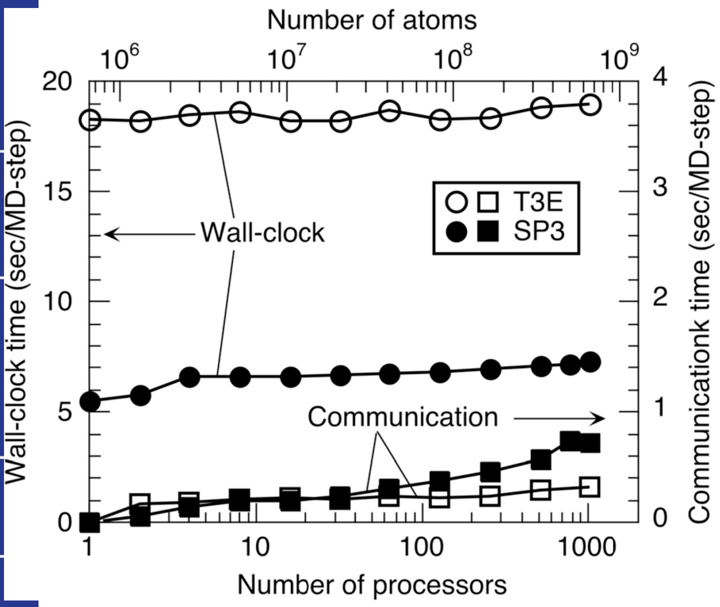
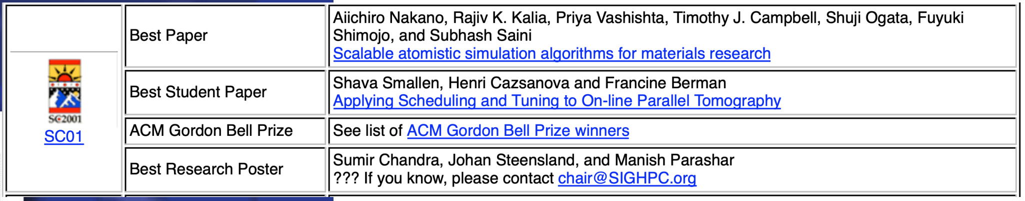


Figure 9: Real part of an example radial function. In this example $l_1 = l_2$ so the radial function is symmetric about $r_1 = r_2$.

Solution of a three body quantum mechanics problem





SC01	Best Paper	Aiichiro Nakano, Rajiv K. Kalia, Priya Vashishta, Timothy J. Campbell, Shuji Ogata, Fuyuki Shimojo, and Subhash Saini Scalable atomistic simulation algorithms for materials research
	Best Student Paper	Shava Smallen, Henri Cazsanova and Francine Berman Applying Scheduling and Tuning to On-line Parallel Tomography
	ACM Gordon Bell Prize	See list of ACM Gordon Bell Prize winners
	Best Research Poster	Sumir Chandra, Johan Steensland, and Manish Parashar ??? If you know, please contact chair@SIGHPC.org