

Parallel Programming for Science and Engineering

Using MPI, OpenMP, and the PETSc library

Victor Eijkhout

1st edition 2017

Public draft - open for comments

This book will be open source under CC-BY license.

Two of the most common software systems for parallel programming in scientific computing are MPI and OpenMP. They target different types of parallelism, and use very different constructs. Thus, by covering both of them in one book we can offer a treatment of parallelism that spans a large range of possible applications.

Contents

I MPI	11
1	Getting started with MPI 12
1.1	<i>Distributed memory and message passing</i> 12
1.2	<i>History</i> 12
1.3	<i>Basic model</i> 13
1.4	<i>Making and running an MPI program</i> 14
1.5	<i>Language bindings</i> 15
1.6	<i>Review</i> 18
2	MPI topic: Functional parallelism 19
2.1	<i>The SPMD model</i> 19
2.2	<i>Starting and running MPI processes</i> 21
2.3	<i>Processor identification</i> 24
2.4	<i>Functional parallelism</i> 28
3	MPI topic: Collectives 29
3.1	<i>Working with global information</i> 29
3.2	<i>Reduction</i> 32
3.3	<i>Rooted collectives: broadcast, reduce</i> 34
3.4	<i>Rooted collectives: gather and scatter</i> 39
3.5	<i>All-to-all</i> 44
3.6	<i>Reduce-scatter</i> 46
3.7	<i>Barrier</i> 49
3.8	<i>Variable-size-input collectives</i> 50
3.9	<i>Scan operations</i> 52
3.10	<i>MPI Operators</i> 54
3.11	<i>Non-blocking collectives</i> 56
3.12	<i>Performance of collectives</i> 57
3.13	<i>Collectives and synchronization</i> 58
3.14	<i>Implementation and performance of collectives</i> 61
3.15	<i>Sources used in this chapter</i> 63
4	MPI topic: Point-to-point 72
4.1	<i>Distributed computing and distributed data</i> 72
4.2	<i>Blocking point-to-point operations</i> 73
4.3	<i>Non-blocking point-to-point operations</i> 88

4.4	<i>More about point-to-point communication</i>	98
4.5	<i>Sources used in this chapter</i>	104
5	MPI topic: Data types	116
5.1	<i>Elementary data types</i>	116
5.2	<i>Derived datatypes</i>	121
5.3	<i>Type size</i>	136
5.4	<i>More about data</i>	140
5.5	<i>Sources used in this chapter</i>	142
6	MPI topic: Communicators	145
6.1	<i>Communicator basics</i>	145
6.2	<i>Subcommunications</i>	146
6.3	<i>Duplicating communicators</i>	147
6.4	<i>Splitting a communicator</i>	150
6.5	<i>Communicators and groups</i>	153
6.6	<i>Inter-communicators</i>	153
6.7	<i>Sources used in this chapter</i>	161
7	MPI topic: Process management	162
7.1	<i>Process spawning</i>	162
7.2	<i>Socket-style communications</i>	165
7.3	<i>Sources used in this chapter</i>	167
8	MPI topic: One-sided communication	169
8.1	<i>Windows</i>	170
8.2	<i>Active target synchronization: epochs</i>	174
8.3	<i>Put, get, accumulate</i>	174
8.4	<i>Passive target synchronization</i>	182
8.5	<i>Details</i>	184
8.6	<i>Implementation</i>	186
8.7	<i>Sources used in this chapter</i>	186
9	MPI topic: File I/O	198
9.1	<i>File handling</i>	198
9.2	<i>File reading and writing</i>	200
9.3	<i>Consistency</i>	205
9.4	<i>Constants</i>	205
10	MPI topic: Topologies	207
10.1	<i>Cartesian grid topology</i>	207
10.2	<i>Distributed graph topology</i>	209
10.3	<i>Sources used in this chapter</i>	213
11	MPI topic: Shared memory	215
11.1	<i>Recognizing shared memory</i>	215
11.2	<i>Shared memory for windows</i>	215
11.3	<i>Sources used in this chapter</i>	217
12	MPI leftover topics	221
12.1	<i>Info objects</i>	221

12.2	<i>Error handling</i>	224
12.3	<i>Fortran issues</i>	226
12.4	<i>Fault tolerance</i>	226
12.5	<i>Context information</i>	226
12.6	<i>Performance</i>	227
12.7	<i>Determinism</i>	232
12.8	<i>Subtleties with processor synchronization</i>	232
12.9	<i>Multi-threading</i>	233
12.10	<i>Shell interaction</i>	235
12.11	<i>The origin of one-sided communication in ShMem</i>	236
12.12	<i>Leftover topics</i>	236
12.13	<i>Literature</i>	239
12.14	<i>Sources used in this chapter</i>	239
13	MPI Reference	241
13.1	<i>Leftover topics</i>	241
14	MPI Review	243
14.1	<i>Conceptual</i>	243
14.2	<i>Communicators</i>	243
14.3	<i>Point-to-point</i>	243
14.4	<i>Collectives</i>	246
14.5	<i>Datatypes</i>	246
14.6	<i>Theory</i>	247
II	OpenMP	249
15	Getting started with OpenMP	250
15.1	<i>The OpenMP model</i>	250
15.2	<i>Compiling and running an OpenMP program</i>	253
15.3	<i>Your first OpenMP program</i>	254
16	OpenMP topic: Parallel regions	257
16.1	<i>Nested parallelism</i>	258
16.2	<i>Cancel parallel construct</i>	260
16.3	<i>Sources used in this chapter</i>	260
17	OpenMP topic: Loop parallelism	261
17.1	<i>Loop parallelism</i>	261
17.2	<i>Loop schedules</i>	263
17.3	<i>Reductions</i>	268
17.4	<i>Collapsing nested loops</i>	268
17.5	<i>Ordered iterations</i>	269
17.6	<i>nowait</i>	269
17.7	<i>While loops</i>	270
18	OpenMP topic: Work sharing	271
18.1	<i>Sections</i>	271

18.2	<i>Single/master</i>	272
18.3	<i>Fortran array syntax parallelization</i>	273
19	OpenMP topic: Controlling thread data	274
19.1	<i>Shared data</i>	274
19.2	<i>Private data</i>	274
19.3	<i>Data in dynamic scope</i>	275
19.4	<i>Temporary variables in a loop</i>	276
19.5	<i>Default</i>	276
19.6	<i>Array data</i>	277
19.7	<i>First and last private</i>	277
19.8	<i>Persistent data through <code>threadprivate</code></i>	278
19.9	<i>Sources used in this chapter</i>	280
20	OpenMP topic: Reductions	281
20.1	<i>Built-in reduction operators</i>	283
20.2	<i>Initial value for reductions</i>	283
20.3	<i>User-defined reductions</i>	284
20.4	<i>Reductions and floating-point math</i>	285
20.5	<i>Sources used in this chapter</i>	285
21	OpenMP topic: Synchronization	286
21.1	<i>Barrier</i>	286
21.2	<i>Mutual exclusion</i>	287
21.3	<i>Locks</i>	288
21.4	<i>Example: Fibonacci computation</i>	290
22	OpenMP topic: Tasks	293
22.1	<i>Task data</i>	294
22.2	<i>Task synchronization</i>	295
22.3	<i>Task dependencies</i>	297
22.4	<i>More</i>	298
22.5	<i>Examples</i>	299
22.6	<i>Sources used in this chapter</i>	300
23	OpenMP topic: Affinity	301
23.1	<i>OpenMP thread affinity control</i>	301
23.2	<i>First-touch</i>	305
23.3	<i>Affinity control outside OpenMP</i>	306
23.4	<i>Sources used in this chapter</i>	306
24	OpenMP topic: Memory model	307
24.1	<i>Thread synchronization</i>	307
24.2	<i>Data races</i>	308
24.3	<i>Relaxed memory model</i>	309
25	OpenMP topic: SIMD processing	310
25.1	<i>Sources used in this chapter</i>	313
26	OpenMP remaining topics	314
26.1	<i>Runtime functions and internal control variables</i>	314

26.2	<i>Timing</i>	316
26.3	<i>Thread safety</i>	316
26.4	<i>Performance and tuning</i>	317
26.5	<i>Accelerators</i>	318
27	OpenMP Review	319
27.1	<i>Concepts review</i>	319
27.2	<i>Review questions</i>	320

III PETSc 329

28	PETSc basics	330
28.1	<i>What is PETSc and why?</i>	330
28.2	<i>Basics of running a PETSc program</i>	332
28.3	<i>PETSc installation</i>	334
29	PETSc objects	336
29.1	<i>Distributed objects</i>	336
29.2	<i>Scalars</i>	337
29.3	<i>Vectors</i>	338
29.4	<i>Matrices</i>	341
29.5	<i>Matrix operations</i>	344
29.6	<i>Submatrices</i>	346
29.7	<i>Shell matrices</i>	346
29.8	<i>DMDA: distributed arrays</i>	346
29.9	<i>Index sets and Vector Scatters</i>	347
29.10	<i>Options and profiling</i>	347
30	PETSc solvers	349
30.1	<i>KSP: linear system solvers</i>	349
30.2	<i>Direct solvers</i>	351
30.3	<i>Control through command line options</i>	351
31	PETSc solvers	353
31.1	<i>Error checking</i>	353
31.2	<i>Printing</i>	353
31.3	<i>Commandline options</i>	354
31.4	<i>Memory management</i>	354
32	PETSc topics	356
32.1	<i>Communicators</i>	356

IV The Rest 359

33	Exploring computer architecture	360
33.1	<i>Tools for discovery</i>	360
34	Process and thread affinity	361
34.1	<i>What does the hardware look like?</i>	362
34.2	<i>Affinity control</i>	364

35	Hybrid computing	365
35.1	<i>Discussion</i>	366
35.2	<i>Hybrid MPI-plus-threads execution</i>	367
35.3	<i>Sources used in this chapter</i>	368
36	Random number generation	370
37	Parallel I/O	371
37.1	<i>Sources used in this chapter</i>	371
38	Support libraries	372
38.1	<i>SimGrid</i>	372
38.2	<i>Other</i>	372
V	Tutorials	373
38.3	<i>Debugging</i>	375
38.4	<i>Tracing</i>	383
38.5	<i>SimGrid</i>	385
VI	Projects, index	387
39	Class projects	388
39.1	<i>A Style Guide to Project Submissions</i>	388
39.2	<i>Warmup Exercises</i>	391
39.3	<i>Mandelbrot set</i>	395
39.4	<i>Data parallel grids</i>	402
39.5	<i>N-body problems</i>	405
40	Bibliography, index, and list of acronyms	406
40.1	<i>Bibliography</i>	406
40.2	<i>List of acronyms</i>	408
40.3	<i>Index</i>	409

PART I

MPI

Chapter 1

Getting started with MPI

In this chapter you will learn the use of the main tool for distributed memory programming: the Message Passing Interface (MPI) library. The MPI library has about 250 routines, many of which you may never need. Since this is a textbook, not a reference manual, we will focus on the important concepts and give the important routines for each concept. What you learn here should be enough for most common purposes. You are advised to keep a reference document handy, in case there is a specialized routine, or to look up subtleties about the routines you use.

1.1 Distributed memory and message passing

In its simplest form, a distributed memory machine is a collection of single computers hooked up with network cables. In fact, this has a name: a *Beowulf cluster*. As you recognize from that setup, each processor can run an independent program, and has its own memory without direct access to other processors' memory. MPI is the magic that makes multiple instantiations of the same executable run so that they know about each other and can exchange data through the network.

One of the reasons that MPI is so successful as a tool for high performance on clusters is that it is very explicit: the programmer controls many details of the data motion between the processors. Consequently, a capable programmer can write very efficient code with MPI. Unfortunately, that programmer will have to spell things out in considerable detail. For this reason, people sometimes call MPI ‘the assembly language of parallel programming’. If that sounds scary, be assured that things are not that bad. You can get started fairly quickly with MPI, using just the basics, and coming to the more sophisticated tools only when necessary.

Another reason that MPI was a big hit with programmers is that it does not ask you to learn a new language: it is a library that can be interface to C/C++ or Fortran; there are even bindings to Python. A related point is that it is easy to install: there are free implementations that you can download and install on any computer that has a Unix-like operating system, even if that is not a parallel machine.

1.2 History

Before the MPI standard was developed in 1993-4, there were many libraries for distributed memory computing, often proprietary to a vendor platform. MPI standardized the inter-process communication mecha-

nisms. Other features, such as process management in PVM, or parallel I/O were omitted. Later versions of the standard have included many of these features.

Since MPI was designed by a large number of academic and commercial participants, it quickly became a standard. A few packages from the pre-MPI era, such as *Charmpp* [11], are still in use since they support mechanisms that do not exist in MPI.

1.3 Basic model

Here we sketch the two most common scenarios for using MPI. In the first, the user is working on an interactive machine, which has network access to a number of hosts, typically a network of workstations; see figure 1.1. The user types the command `mpiexec`¹ and supplies

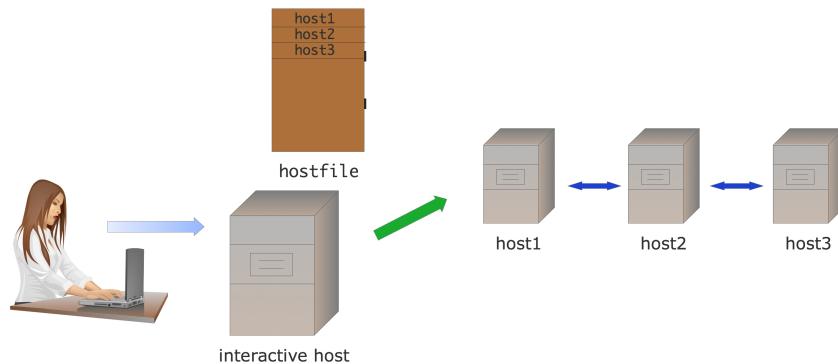


Figure 1.1: Interactive MPI setup

- The number of hosts involved,
- their names, possibly in a hostfile,
- and other parameters, such as whether to include the interactive host; followed by
- the name of the program and its parameters.

The `mpirun` program then makes an `ssh` connection to each of the hosts, giving them sufficient information that they can find each other. All the output of the processors is piped through the `mpirun` program, and appears on the interactive console.

In the second scenario (figure 1.2) the user prepares a *batch job* script with commands, and these will be run when the *batch scheduler* gives a number of hosts to the job. Now the batch script contains the `mpirun` command, and the hostfile is dynamically generated when the job starts. Since the job now runs at a time when the user may not be logged in, any screen output goes into an output file.

You see that in both scenarios the parallel program is started by the `mpirun` command using an Single Program Multiple Data (SPMD) mode of execution: all hosts execute the same program. It is possible for different hosts to execute different programs, but we will not consider that in this book.

1. A command variant is `mpirun`; your local cluster may have a different mechanism.



Figure 1.2: Batch MPI setup

There can be options and environment variables that are specific to some MPI installations, or to the network.

- *mpich* and its derivatives such as *Intel MPI* or *Cray MPI* have *mpiexec* options: <https://www.mpich.org/static/docs/v3.1/www1/mpiexec.html>

1.4 Making and running an MPI program

MPI is a library, called from programs in ordinary programming languages such as C/C++ or Fortran. To compile such a program you use your regular compiler:

```
gcc -c my_mpi_prog.c -I/path/to/mpi.h
gcc -o my_mpi_prog my_mpi_prog.o -L/path/to/mpi -lmpich
```

However, MPI libraries may have different names between different architectures, making it hard to have a portable makefile. Therefore, MPI typically has shell scripts around your compiler call:

```
mpicc -c my_mpi_prog.c
mpicc -o my_mpi_prog my_mpi_prog.o
```

MPI programs can be run on many different architectures. Obviously it is your ambition (or at least your dream) to run your code on a cluster with a hundred thousand processors and a fast network. But maybe you only have a small cluster with plain *ethernet*. Or maybe you're sitting in a plane, with just your laptop. An MPI program can be run in all these circumstances – within the limits of your available memory of course.

The way this works is that you do not start your executable directly, but you use a program, typically called *mpirun* or something similar, which makes a connection to all available processors and starts a run of your executable there. So if you have a thousand nodes in your cluster, *mpirun* can start your program once on each, and if you only have your laptop it can start a few instances there. In the latter case you will of course not get great performance, but at least you can test your code for correctness.

Python note. Load the TACC-provided python:

```
module load python
```

and run it as:

```
iexec python-mpi yourprogram.py
```

1.5 Language bindings

1.5.1 C/C++

The MPI library is written in C. Thus, its bindings are the most natural for that language.

C++ bindings existed at one point, but they were declared deprecated, and have been officially removed in the *MPI 3*. The *boost* library has its own version of MPI, but it seems not to be under further development. A recent effort at idiomatic C++ support is *MPL* <http://numbercrunch.de/blog/2015/08/mpl-a-message-passing-library/>.

1.5.2 Fortran

Fortran note. Fortran-specific notes will be indicated with a note like this.

Traditionally, *Fortran bindings* for MPI look very much like the C ones, except that each routine has a final *error return* parameter. You will find that a lot of MPI code in Fortran conforms to this.

However, in the *MPI 3* standard it is recommended that an MPI implementation providing a Fortran interface provide a module named *mpi_f08* that can be used in a Fortran program. This defines MPI functions that return an integer result, rather than having a final parameter. It also defines proper interfaces, making type checking possible: there are separate routines for each datatype, and an *Interface* block in the MPI module. If you manage to request a version that does not exist, the compiler will display a message like

```
There is no matching specific
subroutine for this generic subroutine call [MPI_Send]
```

For details see <http://mpi-forum.org/docs/mpi-3.1/mpi31-report/node409.htm>.

There are some visible implications of using the *mpi_f08* module, mostly related to the fact that some of the ‘MPI datatypes’ such as *MPI_Comm*, which were declared as *Integer* previously, are now a Fortran Type. See the following sections for details: Communicator 6.1, Datatype 5.2.1.1, Info 12.1, Op 3.10.2, Request 4.3.1, Status 4.4.2, Window 8.1.

The *mpi_f08* module solves a problem with previous *Fortran90 bindings*: Fortran90 is a strongly typed language, so it is not possible to pass argument by reference to their address, as C/C++ do with the *void** type for send and receive buffers. This was solved by having separate routines for each datatype, and providing an *Interface* block in the MPI module. If you manage to request a version that does not exist, the compiler will display a message like

```
There is no matching specific
subroutine for this generic subroutine call [MPI_Send]
```

1.5.3 Python

The `mpi4py` package [3] of *python bindings* is not defined by the MPI standards committee. Instead, it is the work of an individual, *Lisandro Dalcin*.

Notable about the Python bindings is that many communication routines exist in two variants:

- a version that can send native Python objects. These routines have lowercase names such as `bcast`; and
- a version that sends `numpy` objects; these routines have names such as `Bcast`. Their syntax can be slightly different.

The first version looks more ‘pythonic’, is easier to write, and can do things like sending python objects, but it is also decidedly less efficient since data is packed and unpacked with `pickle`. As a common sense guideline, use the `numpy` interface in the performance-critical parts of your code, and the native interface only for complicated actions in a setup phase.

Codes with `mpi4py` can be interfaced to other languages through Swig or conversion routines.

Data in `numpy` can be specified as a simple object, or `[data, (count,displ), datatype]`.

1.5.4 How to read routine prototypes

Throughout the MPI part of this book we will give the reference syntax of the routines. This typically comprises:

- The semantics: routine name and list of parameters and what they mean.
- C syntax: the routine definition as it appears in the `mpi.h` file.
- Fortran syntax: routine definition with parameters, giving in/out specification.
- Python syntax: routine name, indicating to what class it applies, and parameter, indicating which ones are optional.

These ‘routine prototypes’ look like code but they are not! Here is how you translate them.

1.5.4.1 C

The typically C routine specification in MPI looks like:

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

This means that

- The routine returns an `int` parameter. Strictly speaking you should test against `MPI_SUCCESS` (for all error codes, see section 12.2.1):

```
|| MPI_Comm comm = MPI_COMM_WORLD;
   int nprocs;
   int errorcode;
   errorcode = MPI_Comm_size( MPI_COMM_WORLD, &nprocs );
   if (errorcode!=MPI_SUCCESS) {
      printf("Routine MPI_Comm_size failed! code=%d\n",
             errorcode);
      return 1;
   }
```

However, the error codes are hardly ever useful, and there is not much your program can do to recover from an error. Most people call the routine as

```
|| MPI_Comm_world( /* parameter ... */ );
```

For more on error handling, see section [12.2](#).

- The first argument is of type `MPI_Comm`. This is not a C built-in datatype, but it behaves like one. There are many of these `MPI_something` datatypes in MPI. So you can write:

```
|| MPI_Comm my_comm =  
||     MPI_COMM_WORLD; // using a predefined value  
|| MPI_Comm_size( comm, /* remaining parameters */ );
```

- Finally, there is a ‘star’ parameter. This means that the routine wants an address, rather than a value. You would typically write:

```
|| MPI_Comm my_comm = MPI_COMM_WORLD; // using a predefined value  
|| int nprocs;  
|| MPI_Comm_size( comm, &nprocs );
```

Seeing a ‘star’ parameter usually means either: the routine has an array argument, or: the routine internally sets the value of a variable. The latter is the case here.

1.5.4.2 Fortran

The Fortran specification looks like:

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

The syntax of using this routine is close to this specification: you write

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

- Most Fortran routines have the same parameters as the corresponding C routine, except that they all have the error code as final parameter, instead of as a function result. As with C, you can ignore the value of that parameter. Just don’t forget it.
- The types of the parameters are given in the specification.
- Where C routines have `MPI_Comm` and `MPI_Request` and such parameters, Fortran has `INTEGER` parameters, or sometimes arrays of integers.

1.5.4.3 Python

The Python interface to MPI uses classes and objects. Thus, a specification like:

```
|| MPI.Comm.Send(self, buf, int dest, int tag=0)
```

should be parsed as follows.

- First of all, you need the MPI class:

```
|| from mpi4py import MPI
```

- Next, you need a Comm object. Often you will use the predefined communicator

```
|| comm = MPI.COMM_WORLD
```

- The keyword self indicates that the actual routine Send is a method of the Comm object, so you call:

```
|| comm.Send( .... )
```

- Parameters that are listed by themselves, such as buf, as positional. Parameters that are listed with a type, such as int dest are keyword parameters. Keyword parameters that have a value specified, such as int tag=0 are optional, with the default value indicated. Thus, the typical call for this routine is:

```
|| comm.Send(sendbuf, dest=other)
```

specifying the send buffer as positional parameter, the destination as keyword parameter, and using the default value for the optional tag.

Some python routines are ‘class methods’, and their specification lacks the self keyword. For instance:

```
|| MPI.Request.Waitall(type cls, requests, statuses=None)
```

would be used as

```
|| MPI.Request.Waitall(requests)
```

1.6 Review

Review 1.1. What determines the parallelism of an MPI job?

- The size of the cluster you run on.
- The number of cores per cluster node.
- The parameters of the MPI starter (mpiexec, ibrunch,...)

Review 1.2. Which languages have an object-oriented interface to MPI?

- C
- C++
- Fortran2008
- Python

Chapter 2

MPI topic: Functional parallelism

2.1 The SPMD model

MPI programs conform (mostly) to the Single Program Multiple Data (SPMD) model, where each processor runs the same executable. This running executable we call a *process*.

When MPI was first written, 20 years ago, it was clear what a processor was: it was what was in a computer on someone's desk, or in a rack. If this computer was part of a networked cluster, you called it a *node*. So if you ran an MPI program, each node would have one MPI process; figure 2.1. You could of course run

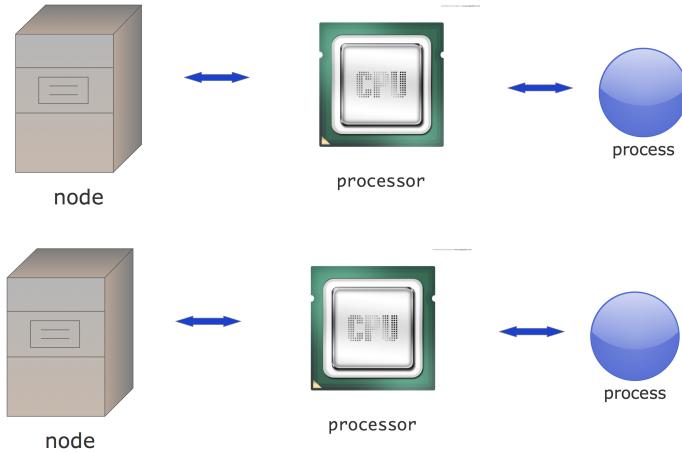


Figure 2.1: Cluster structure as of the mid 1990s

more than one process, using the *time slicing* of the Operating System (OS), but that would give you no extra performance.

These days the situation is more complicated. You can still talk about a node in a cluster, but now a node can contain more than one processor chip (sometimes called a *socket*), and each processor chip probably has multiple *cores*. Figure 2.2 shows how you could explore this using a mix of MPI between the nodes, and a shared memory programming system on the nodes.

However, since each core can act like an independent processor, you can also have multiple MPI processes per node. To MPI, the cores look like the old completely separate processors. This is the 'pure MPI' model

2. MPI topic: Functional parallelism

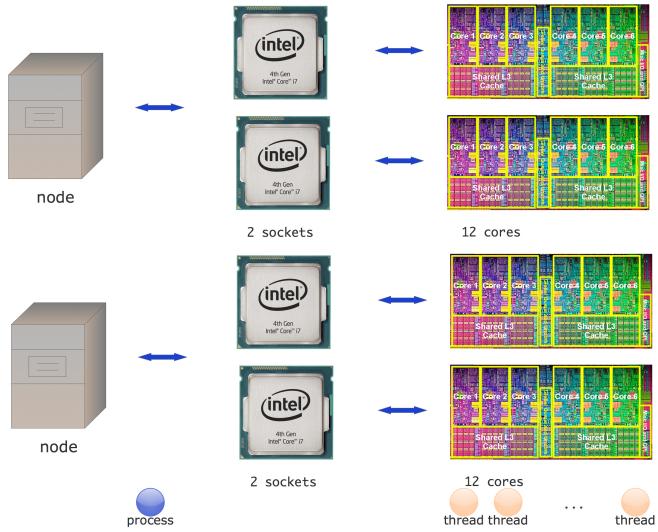


Figure 2.2: Hybrid cluster structure

of figure 2.3, which we will use in most of this part of the book. (Hybrid computing will be discussed in chapter 35.)

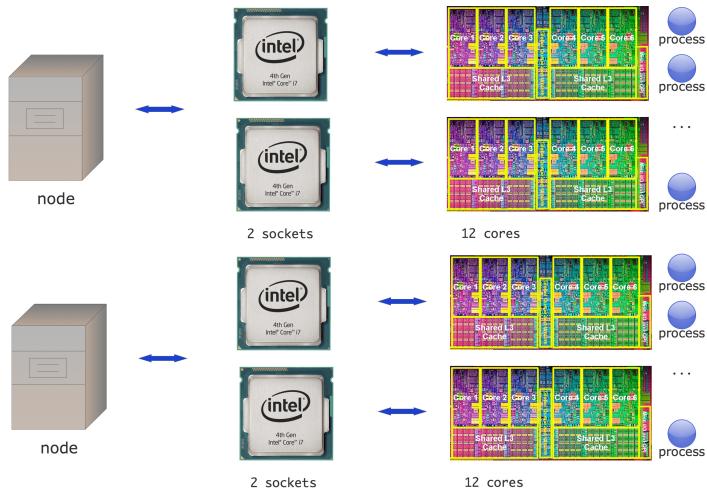


Figure 2.3: MPI-only cluster structure

This is somewhat confusing: the old processors needed MPI programming, because they were physically separated. The cores on a modern processor, on the other hand, share the same memory, and even some caches. In its basic mode MPI seems to ignore all of this: each core receives an MPI process and the programmer writes the same send/receive call no matter where the other process is located. In fact, you can't immediately see whether two cores are on the same node or different nodes. Of course, on the implementation level MPI uses a different communication mechanism depending on whether cores are on the same

socket or on different nodes, so you don't have to worry about lack of efficiency.

Remark 1 In some rare cases you may want to run in an Multiple Program Multiple Data (MPMD) mode, rather than SPMD. This can be achieved either on the OS level, using options of the `mpiexec` mechanism, or you can use MPI's built-in process management; chapter 7. Like I said, this concerns only rare cases.

2.2 Starting and running MPI processes

The SPMD model may be initially confusing. Even though there is only a single source, compiled into a single executable, the parallel run comprises a number of independently started MPI processes (see section 1.3 for the mechanism).

The following exercises are designed to give you an intuition for this one-source-many-processes setup. In the first exercise you will see that the mechanism for starting MPI programs starts up independent copies. There is nothing in the source that says 'and now you become parallel'.

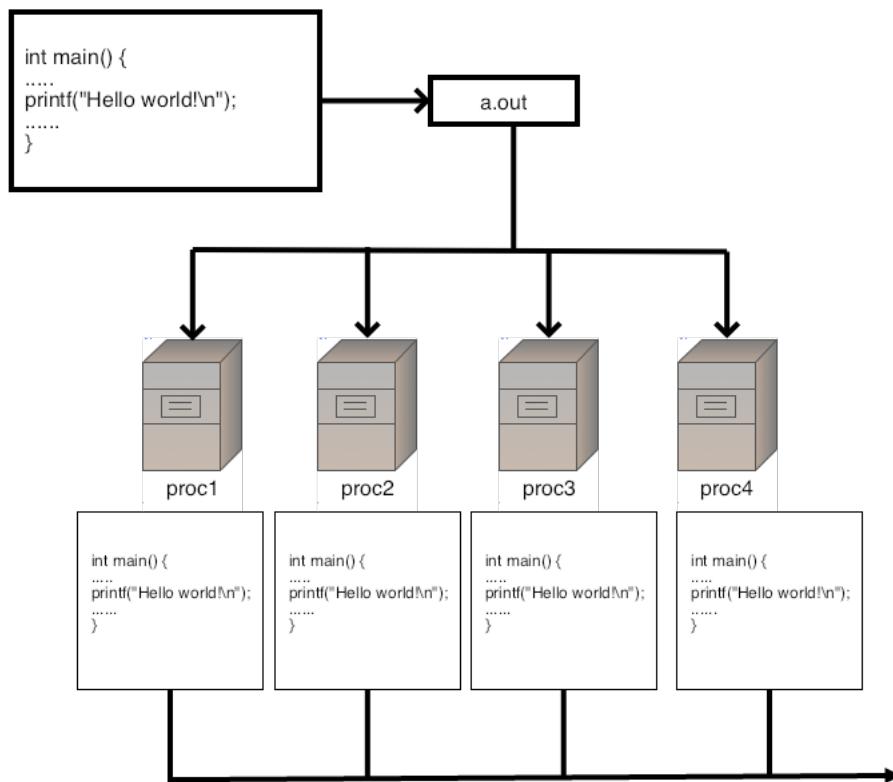


Figure 2.4: Running a hello world program in parallel

The following exercise demonstrates this point.

Exercise 2.1. Write a 'hello world' program, without any MPI in it, and run it in parallel with `mpiexec` or your local equivalent. Explain the output.

This exercise is illustrated in figure 2.4.

2.2.1 Headers

If you use MPI commands in a program file, be sure to include the proper header file, *mpi.h* or *mpif.h*.

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

For *Fortran90*, many MPI installations also have an MPI module, so you can write

```
use mpi      ! pre 3.0
use mpi_f08 ! 3.0 standard
```

The internals of these files can be different between MPI installations, so you can not compile one file against one *mpi.h* file and another file, even with the same compiler on the same machine, against a different MPI.

2.2.2 Initialization / finalization

To get a useful MPI program you need at least the calls ***MPI_Init*** and ***MPI_Finalize*** surrounding your code.

Python note. There are no initialize and finalize calls: the `import` statement performs the initialization.

Every MPI program has to start with *MPI initialization* through ***MPI_Init*** (figure 1), passing `argc` and `argv`, the arguments of a C language main program:

```
|| int main(int argc, char **argv) {
||     ....
||     return 0;
|| }
```

(It is allowed to pass `NULL` for these arguments.)

Note that the ***MPI_Init*** call is one of the few that differs between C and Fortran: the C routine takes the commandline arguments, which Fortran lacks.

This may look a bit like declaring ‘this is the parallel part of a program’, but that’s not true: again, the whole code is executed multiple times in parallel.

The regular way to conclude an MPI program is through ***MPI_Finalize*** (figure 2).

Exercise 2.2. 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.

2.2.2.1 Aborting an MPI run

Apart from ***MPI_Finalize***, which signals a successful conclusion of the MPI run, an abnormal end to a run can be forced by ***MPI_Abort*** (figure 3). This aborts execution on all processes associated with the communicator, but many implementations simply abort all processes. The `value` parameter is returned to the environment.

MPI_Init

C:
int MPI_Init(int *argc, char ***argv)

Fortran:
MPI_Init(ierror)
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

How to read routine prototypes: [1.5.4](#).

manpage 1: Routine prototype for MPI_Init

MPI_Finalize

C:
int MPI_Finalize(void)

Fortran:
MPI_Finalize(ierror)
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

How to read routine prototypes: [1.5.4](#).

manpage 2: Routine prototype for MPI_Finalize

MPI_Abort

Synopsis:
int MPI_Abort(MPI_Comm comm, int errorcode)

Input Parameters
comm : communicator of tasks to abort
errorcode : error code to return to invoking environment

Python:
MPI.Comm.Abort(self, int errorcode=0)

How to read routine prototypes: [1.5.4](#).

manpage 3: Routine prototype for MPI_Abort

2.2.2.2 Testing the initialized/finalized status

The commandline arguments `argc` and `argv` are only guaranteed to be passed to process zero, so the best way to pass commandline information is by a broadcast (section 3.3.3).

There are a few commands, such as `MPI_Get_processor_name`, that are allowed before `MPI_Init`.

If MPI is used in a library, MPI can have already been initialized in a main program. For this reason, one can test where `MPI_Init` has been called with `MPI_Initialized` (figure 4).

You can test whether `MPI_Finalize` has been called with `MPI_Finalized` (figure 5).

2.2.2.3 Information about the run

Once MPI has been initialized, the `MPI_INFO_ENV` object contains a number of key/value pairs describing run-specific information; see section 12.1.1.

2.2.2.4 Commandline arguments

The `MPI_Init` routines takes a reference to `argc` and `argv` for the following reason: the `MPI_Init` calls filters out the arguments to `mpirun` or `mpiexec`, thereby lowering the value of `argc` and eliminating some of the `argv` arguments.

On the other hand, the commandline arguments that are meant for `mpiexec` wind up in the `MPI_INFO_ENV` object as a set of key/value pairs; see section 12.1.

2.3 Processor identification

Since all processes in an MPI job are instantiations of the same executable, you'd think that they all execute the exact same instructions, which would not be terribly useful. You will now learn how to distinguish processes from each other, so that together they can start doing useful work.

2.3.1 Processor name

In the following exercise you will print out the hostname of each MPI process with `MPI_Get_processor_name` (figure 132) as a first way of distinguishing between processes.

Exercise 2.3. 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`.

The character storage is provided by the user: the character array must be at least `MPI_MAX_PROCESSOR_NAME` characters long. The actual length of the name is returned in the `resultlen` parameter.

MPI_Initialized

```
C:  
int MPI_Initialized(int *flag)  
  
Fortran:  
MPI_Initialized(flag, ierror)  
LOGICAL, INTENT(OUT) :: flag  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 4: Routine prototype for MPI_Initialized

MPI_Finalized

```
C:  
int MPI_Finalized( int *flag )  
  
Fortran:  
MPI_Finalized(flag, ierror)  
LOGICAL, INTENT(OUT) :: flag  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 5: Routine prototype for MPI_Finalized

MPI_Get_processor_name

```
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()
```

How to read routine prototypes: 1.5.4.

manpage 6: Routine prototype for MPI_Get_processor_name

2.3.2 Process and communicator properties: rank and size

First we need to introduce the concept of *communicator*, which is an abstract description of a group of processes. For now you only need to know about the existence of the `MPI_Comm` data type, and that there is a pre-defined communicator `MPI_COMM_WORLD` which describes all the processes involved in your parallel run. You will learn much more about communicators in chapter 6.

To distinguish between processes in a communicator, MPI provides two calls

1. `MPI_Comm_size` (figure 7) reports how many processes there are in all; and
2. `MPI_Comm_rank` (figure 8) states what the number of the process is that calls this routine.

If every process executes the `MPI_Comm_size` call, they all get the same result, namely the total number of processes in your run. On the other hand, if every process executes `MPI_Comm_rank`, they all get a different result, namely their own unique number, an integer in the range from zero to the the number of processes minus 1. See figure 2.5. In other words, each process can find out ‘I am process 5 out of a total of 20’.

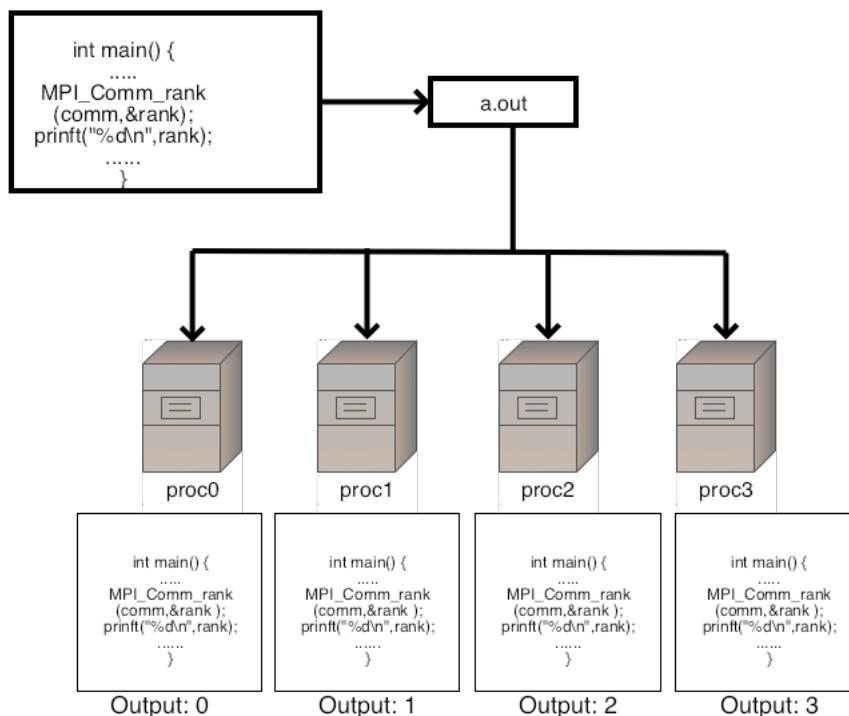


Figure 2.5: Parallel program that prints process rank

Exercise 2.4. Write a program where each process prints out a message reporting its number, and how many processes there are:

Hello from process 2 out of 5!

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.*

MPI_Comm_size

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)
```

How to read routine prototypes: 1.5.4.

manpage 7: Routine prototype for MPI_Comm_size

MPI_Comm_rank

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)
```

How to read routine prototypes: 1.5.4.

manpage 8: Routine prototype for MPI_Comm_rank

Exercise 2.5. Write a program where only the process with number zero reports on how many processes there are in total.

2.4 Functional parallelism

Being able to tell processes apart is already enough for some applications. Based on its rank, a processor can find its section of a search space. For instance, in *Monte Carlo* codes a large number of random samples is generated and some computation performed on each. (This particular example requires each MPI process to run an independent random number generator, which is not entirely trivial.)

Exercise 2.6. 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.)

As another example, in *Boolean satisfiability* problems a number of points in a search space needs to be evaluated. Knowing a process's rank is enough to let it generate its own portion of the search space. The computation of the *Mandelbrot set* can also be considered as a case of functional parallelism. However, the image that is constructed is data that needs to be kept on one processor, which breaks the symmetry of the parallel run.

Of course, at the end of a functionally parallel run you need to summarize the results, for instance printing out some total. The mechanisms for that you will learn next.

Chapter 3

MPI topic: Collectives

3.1 Working with global information

If all processes have individual data, for instance the result of a local computation, you may want to bring that information together, for instance to find the maximal computed value or the sum of all values. Conversely, sometimes one processor has information that needs to be shared with all. For this sort of operation, MPI has *collectives*.

There are various cases, illustrated in figure 3.1, which you can (sort of) motivated by considering some

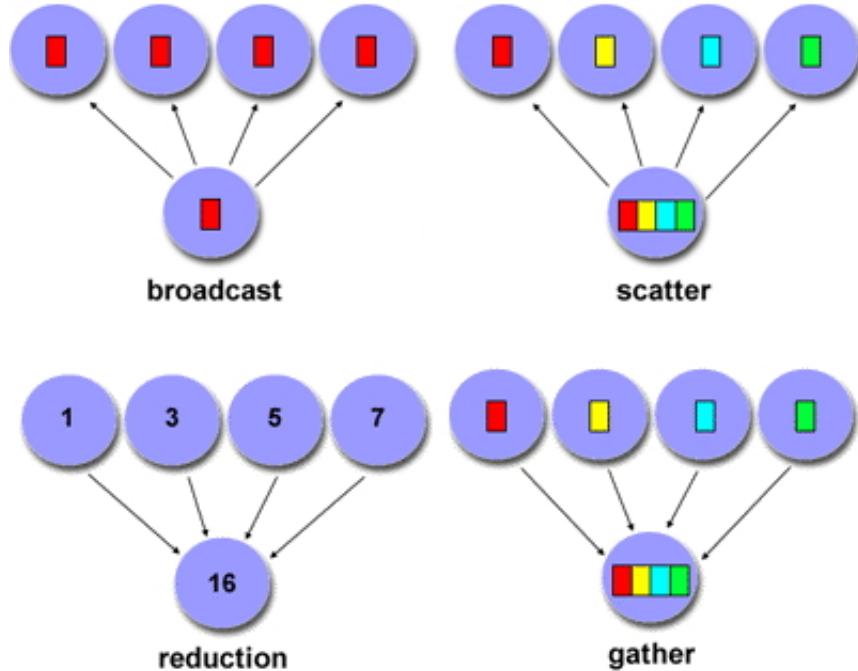


Figure 3.1: The four most common collectives

classroom activities:

3. MPI topic: Collectives

- The teacher tells the class when the exam will be. This is a *broadcast*: the same item of information goes to everyone.
- After the exam, the teacher performs a *gather* operation to collect the individual exams.
- On the other hand, when the teacher computes the average grade, each student has an individual number, but these are now combined to compute a single number. This is a *reduction*.
- Now the teacher has a list of grades and gives each student their grade. This is a *scatter* operation, where one process has multiple data items, and gives a different one to all the other processes.

This story is a little different from what happens with MPI processes, because these are more symmetric; the process doing the reducing and broadcasting is no different from the others. Any process can function as the *root process* in such a collective.

Exercise 3.1. 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.

3.1.1 Practical use of collectives

Collectives are quite common in scientific applications. For instance, if one process reads data from disc or the commandline, it can use a broadcast or a gather to get the information to other processes. Likewise, at the end of a program run, a gather or reduction can be used to collect summary information about the program run.

However, a more common scenario is that the result of a collective is needed on all processes.

Consider the computation of the *standard deviation*:

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

and assume that every process stores just one x_i value.

1. The calculation of the average μ is a reduction, since all the distributed values need to be added.
2. Now every process needs to compute $x_i - \mu$ for its value x_i , so μ is needed everywhere. You can compute this by doing a reduction followed by a broadcast, but it is better to use a so-called *allreduce* operation, which does the reduction and leaves the result on all processors.
3. The calculation of $\sum_i (x_i - \mu)$ is another sum of distributed data, so we need another reduction operation. Depending on whether each process needs to know σ , we can again use an allreduce.

For instance, if x, y are distributed vector objects, and you want to compute

$$y - (x^t y)x$$

which is part of the Gramm-Schmidt algorithm; see HPSC-?. Now you need to use an Allreduce to reduce the inner product value on all processors.

```
// compute local value
localvalue = innerproduct( x[ localpart ], y[ localpart ] );
// compute inner product on the every process
AllReduce( localvalue, reducedvalue );
// send root value to all other from the root
Broadcast( reducedvalue, root );
```

3.1.2 Synchronization

Collectives are operations that involve all processes in a communicator. A collective is a single call, and it blocks on all processors. That does not mean that all processors exit the call at the same time: because of implementational details and network latency they need not be synchronized in their execution. However, semantically we can say that a process can not finish a collective until every other process has at least started the collective.

In addition to these collective operations, there are operations that are said to be ‘collective on their communicator’, but which do not involve data movement. Collective then means that all processors must call this routine; not to do so is an error that will manifest itself in ‘hanging’ code. One such example is [MPI_Win_fence](#).

3.1.3 Collectives in MPI

We will now explain the MPI collectives in the following order.

Allreduce We use the allreduce as an introduction to the concepts behind collectives; section 3.2.1. As explained above, this routines servers many practical scenarios.

Broadcast and reduce We then introduce the concept of a root in the reduce (section 3.3.1) and broadcast (section 3.3.3) collectives.

Gather and scatter The gather/scatter collectives deal with more than a single data item.

There are more collectives or variants on the above.

- If you want to gather or scatter information, but the contribution of each processor is of a different size, there are ‘variable’ collectives; they have a *v* in the name (section 3.8).
- Sometimes you want a reduction with partial results, where each processor computes the sum (or other operation) on the values of lower-numbered processors. For this, you use a *scan* collective (section 3.9).
- If every processor needs to broadcast to every other, you use an *all-to-all* operation (section 3.5).
- A barrier is an operation that makes all processes wait until every process has reached the barrier (section 3.7).

Finally, there are some advanced topics in collectives.

- Non-blocking collectives; section 3.11.
- User-defined reduction operators; section 3.10.2.

3.2 Reduction

3.2.1 Reduce to all

Above we saw a couple of scenarios where a quantity is reduced, with all processes getting the result. The MPI call for this is `MPI_Allreduce` (figure 9) .

Example: we give each process a random number, and sum these numbers together. The result should be approximately 1/2 times the number of processes.

```
// allreduce.c
float myrandom, sumrandom;
myrandom = (float) rand() / (float) RAND_MAX;
// add the random variables together
MPI_Allreduce(&myrandom, &sumrandom,
              1, MPI_FLOAT, MPI_SUM, comm);
// the result should be approx nprocs/2:
if (procno==nprocs-1)
    printf("Result %6.9f compared to .5\n", sumrandom/nprocs);
```

For Python we illustrate both the native and the numpy variant. In the numpy variant we create an array for the receive buffer, even though only one element is used.

```
## allreduce.py
random_number = random.randint(1, nprocs*nprocs)
print("[%d] random=%d" % (procid, random_number))

max_random = comm.allreduce(random_number, op=MPI.MAX)
if procid==0:
    print("Python native:\n  max=%d" % max_random)

myrandom = np.empty(1, dtype=np.int)
myrandom[0] = random_number
allrandom = np.empty(nprocs, dtype=np.int)

comm.Allreduce(myrandom, allrandom[:1], op=MPI.MAX)
```

Exercise 3.2. 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.

3.2.2 Reduction of distributed data

One of the more common applications of the reduction operation is the *inner product* computation. Typically, you have two vectors x, y that have the same distribution, that is, where all processes store equal parts of x and y . The computation is then

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

MPI_Allreduce

```
C:  
int MPI_Allreduce(const void* sendbuf,  
                  void* recvbuf, int count, MPI_Datatype datatype,  
                  MPI_Op op, MPI_Comm comm)  
  
Semantics:  
IN sendbuf: starting address of send buffer (choice)  
OUT recvbuf: starting address of receive buffer (choice)  
IN count: number of elements in send buffer (non-negative integer)  
IN datatype: data type of elements of send buffer (handle)  
IN op: operation (handle)  
IN comm: communicator (handle)  
  
Fortran:  
MPI_Allreduce(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 native:  
recvobj = MPI.Comm.allreduce(self, sendobj, op=SUM)  
Python numpy:  
MPI.Comm.Allreduce(self, sendbuf, recvbuf, op op=SUM)
```

How to read routine prototypes: [1.5.4](#).

manpage 9: Routine prototype for MPI_Allreduce

3.2.3 Reduce in place

By default MPI will not overwrite the original data with the reduction result, but you can tell it to do so using the `MPI_IN_PLACE` specifier:

```
// allreduceinplace.c
int nrandoms = 500000;
float *myrandoms;
myrandoms = (float*) malloc(nrandoms*sizeof(float));
for (int irand=0; irand<nrandoms; irand++)
    myrandoms[irand] = (float) rand() / (float) RAND_MAX;
// add all the random variables together
MPI_Allreduce(MPI_IN_PLACE, myrandoms,
              nrandoms, MPI_FLOAT, MPI_SUM, comm);
// the result should be approx nprocs/2:
if (procno==nprocs-1) {
    float sum=0.;
    for (int i=0; i<nrandoms; i++) sum += myrandoms[i];
    sum /= nrandoms*nprocs;
    printf("Result %6.9f compared to .5\n", sum);
}
```

This has the advantage of saving half the memory.

3.2.4 Reduction operations

Several `MPI_Op` values are pre-defined. For the list, see section 3.10.1.

For use in reductions and scans it is possible to define your own operator.

```
|| MPI_Op_create( MPI_User_function *func, int commute, MPI_Op *op);
```

For more details, see section 3.10.2.

3.3 Rooted collectives: broadcast, reduce

In some scenarios there is a certain process that has a privileged status.

- One process can generate or read in the initial data for a program run. This then needs to be communicated to all other processes.
- At the end of a program run, often one process needs to output some summary information.

This process is called the *root* process, and we will now consider routines that have a root.

3.3.1 Reduce to a root

In the broadcast operation a single data item was communicated to all processes. A reduction operation with `MPI_Reduce` (figure 10) goes the other way: each process has a data item, and these are all brought together into a single item.

Here are the essential elements of a reduction operation:

```
// MPI_Reduce( senddata, recvdata..., operator,
    root, comm );
```

- There is the original data, and the data resulting from the reduction. It is a design decision of MPI that it will not by default overwrite the original data. The send data and receive data are of the same size and type: if every processor has one real number, the reduced result is again one real number.
- There is a reduction operator. Popular choices are `MPI_SUM`, `MPI_PROD` and `MPI_MAX`, but complicated operators such as finding the location of the maximum value exist. You can also define your own operators; section 3.10.2.
- There is a root process that receives the result of the reduction. Since the non-root processes do not receive the reduced data, they can actually leave the receive buffer undefined.

```
// reduce.c
float myrandom = (float) rand() / (float) RAND_MAX,
      result;
int target_proc = nprocs-1;
// add all the random variables together
MPI_Reduce(&myrandom, &result, 1, MPI_FLOAT, MPI_SUM,
            target_proc, comm);
// the result should be approx nprocs/2:
if (procno==target_proc)
    printf("Result %6.3f compared to nprocs/2=%5.2f\n",
           result, nprocs/2.);
```

Exercise 3.3. Write a program where each process computes a random number, and process 0 finds and prints the maximum generated value. Let each process print its value, just to check the correctness of your program.

Collective operations can also take an array argument, instead of just a scalar. In that case, the operation is applied pointwise to each location in the array.

Exercise 3.4. 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?

3.3.2 Reduce in place

Instead of using a send and a receive buffer in the reduction, it is possible to avoid the send buffer by putting the send data in the receive buffer. We see this mechanism in section 3.2.3 for the allreduce operation.

For the rooted call `MPI_Reduce`, it is similarly possible to use the value in the receive buffer on the root. However, on all other processes, data is placed in the send buffer and the receive buffer is null or ignored as before.

This example sets the buffer values through some pointer cleverness in order to have the same reduce call on all processes.

```
// reduceinplace.c
float mynumber, result, *sendbuf, *recvbuf;
mynumber = (float) procno;
```

3. MPI topic: Collectives

```
int target_proc = nprocs-1;
// add all the random variables together
if (procno==target_proc) {
    sendbuf = (float*)MPI_IN_PLACE; recvbuf = &result;
    result = mynumber;
} else {
    sendbuf = &mynumber;     recvbuf = NULL;
}
MPI_Reduce(sendbuf,recvbuf,1,MPI_FLOAT,MPI_SUM,
           target_proc,comm);
// the result should be nprocs*(nprocs-1)/2:
if (procno==target_proc)
    printf("Result %6.3f compared to n(n-1)/2=%5.2f\n",
           result,nprocs*(nprocs-1)/2.);
```

In Fortran the code is less elegant because you can not do these address calculations:

```
// reduceinplace.F90
call random_number(mynumber)
target_proc = ntids-1;
! add all the random variables together
if (mytid.eq.target_proc) then
    result = mytid
    call MPI_Reduce(MPI_IN_PLACE,result,1,MPI_REAL,MPI_SUM,&
                    target_proc,comm,err)
else
    mynumber = mytid
    call MPI_Reduce(mynumber,result,1,MPI_REAL,MPI_SUM,&
                    target_proc,comm,err)
end if
! the result should be ntids*(ntids-1)/2:
if (mytid.eq.target_proc) then
    write(*,'("Result ",f5.2," compared to n(n-1)/2=",f5.2)') &
        result,ntids*(ntids-1)/2.
end if
```

3.3.3 Broadcast

The broadcast call has the following structure:

```
|| MPI_Bcast( data..., root , comm);
```

The root is the process that is sending its data. Typically, it will be the root of a broadcast tree. The `comm` argument is a communicator: for now you can use `MPI_COMM_WORLD`. Unlike with send/receive there is no message tag, because collectives are blocking, so you can have only one collective active at a time.

The data in a broadcast with `MPI_Bcast` (figure 11) (or any other MPI operation for that matter) is specified as

- A buffer. In C this is the address in memory of the data. This means that you broadcast a single scalar as `MPI_Bcast(&value, ...)`, but an array as `MPI_Bcast(array, ...)`.

MPI_Reduce

```
C:  
int MPI_Reduce(  
    const void* sendbuf, void* recvbuf, int count, MPI_Datatype datatype,  
    MPI_Op op, int root, MPI_Comm comm)  
  
Fortran:  
MPI_Reduce(sendbuf, recvbuf, count, datatype, op, root, comm, ierror)  
TYPE(*), DIMENSION(..), INTENT(IN) :: sendbuf  
TYPE(*), DIMENSION(..) :: recvbuf  
INTEGER, INTENT(IN) :: count, root  
TYPE(MPI_Datatype), INTENT(IN) :: datatype  
TYPE(MPI_Op), INTENT(IN) :: op  
TYPE(MPI_Comm), INTENT(IN) :: comm  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python:  
native:  
comm.reduce(self, sendobj=None, recvobj=None, op=SUM, int root=0)  
numpy:  
comm.Reduce(self, sendbuf, recvbuf, Op op=SUM, int root=0)
```

How to read routine prototypes: 1.5.4.

manpage 10: Routine prototype for MPI_Reduce

MPI_Bcast

```
C:  
int MPI_Bcast(  
    void* buffer, int count, MPI_Datatype datatype,  
    int root, MPI_Comm comm)  
  
Fortran:  
MPI_Bcast(buffer, count, datatype, root, comm, ierror)  
TYPE(*), DIMENSION(..) :: buffer  
INTEGER, INTENT(IN) :: count, root  
TYPE(MPI_Datatype), INTENT(IN) :: datatype  
TYPE(MPI_Comm), INTENT(IN) :: comm  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python native:  
rbuf = MPI.Comm.bcast(self, obj=None, int root=0)  
Python numpy:  
MPI.Comm.Bcast(self, buf, int root=0)
```

How to read routine prototypes: 1.5.4.

manpage 11: Routine prototype for MPI_Bcast

3. MPI topic: Collectives

- The number of items and their datatype. The allowable datatypes are such things as `MPI_INT` and `MPI_FLOAT` for C, and `MPI_INTEGER` and `MPI_REAL` for Fortran, or more complicated types. See section 5 for details.

Python note. In python it is both possible to send objects, and to send more C-like buffers. The two possibilities correspond (see section 1.5.3) to different routine names; the buffers have to be created as numpy objects.

Example: in general we can not assume that all processes get the commandline arguments, so we broadcast them from process 0.

```
// init.c
if (procno==0) {
    if ( argc==1 || // the program is called without parameter
        ( argc>1 && !strcmp(argv[1], "-h") ) // user asked for help
    ) {
        printf("\nUsage: init [0-9]+\n");
        MPI_Abort(comm,1);
    }
    input_argument = atoi(argv[1]);
}
MPI_Bcast(&input_argument,1,MPI_INT,0,comm);
```

Exercise 3.5. If you give a commandline argument to a program, that argument is available as a character string as part of the `argv`, `argc` pair that you typically use as the arguments to your main program. You can use the function `atoi` to convert such a string to integer.

Write a program where process 0 looks for an integer on the commandline, and broadcasts it to the other processes. Initialize the buffer on all processes, and let all processes print out the broadcast number, just to check that you solved the problem correctly.

In python we illustrate the native and numpy variants. In the native variant the result is given as a function return; in the numpy variant the send buffer is reused.

```
## bcast.py
# first native
if procid==root:
    buffer = [ 5.0 ] * dsize
else:
    buffer = [ 0.0 ] * dsize
buffer = comm.bcast(obj=buffer, root=root)
if not reduce( lambda x,y:x and y,
               [ buffer[i]==5.0 for i in range(len(buffer)) ] ):
    print( "Something wrong on proc %d: native buffer <<%s>>" \
          % (procid,str(buffer)) )

# then with NumPy
buffer = np.arange(dsize, dtype=np.float64)
if procid==root:
    for i in range(dsize):
        buffer[i] = 5.0
comm.Bcast( buffer, root=root )
```

```

    if not all( buffer==5.0 ):
        print( "Something wrong on proc %d: numpy buffer <<%s>>" \
            % (procid,str(buffer)) )
    else:
        if procid==root:
            print("Success.")

```

For the following exercise, study figure 3.2.

Exercise 3.6. 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 3.7. Add partial pivoting to your implementation of Gauss-Jordan elimination.

Change your implementation to let each processor store multiple columns, but still do one broadcast per column. Is there a way to have only one broadcast per processor?

3.4 Rooted collectives: gather and scatter

In the **MPI_Scatter** operation, the root spreads information to all other processes. The difference with a broadcast is that it involves individual information from/to every process. Thus, the gather operation typically has an array of items, one coming from each sending process, and scatter has an array, with an individual item for each receiving process; see figure 3.4.

These gather and scatter collectives have a different parameter list from the broadcast/reduce. The broadcast/reduce involves the same amount of data on each process, so it was enough to have a buffer, datatype, and size. In the gather/scatter calls you have

- a large buffer on the root, with a datatype and size specification, and
- a smaller buffer on each process, with its own type and size specification.

Of course, since we're in SPMD mode, even non-root processes have the argument for the send buffer, but they ignore it. For instance:

```

int MPI_Scatter
    (void* sendbuf, int sendcount, MPI_Datatype sendtype,
     void* recvbuf, int recvcount, MPI_Datatype recvtype,
     int root, MPI_Comm comm)

```

3. MPI topic: Collectives

Initial:

matrix	sol	rhs	action
2 2 13	1	17	
4 5 32	1	41	
-2 -3 -16	1	-21	

Step 1:

matrix	sol	rhs	action
2 2 13	1	17	take this row
4 5 32	1	41	
-2 -3 -16	1	-21	

Step 2:

matrix	sol	rhs	action
2 2 13	1	17	take this row
↓ ↓			
4 5 32	1	41	minus × 2
-2 -3 -16	1	-21	

Step 3:

matrix	sol	rhs	action
2 2 13	1	17	take this row
0 1 6	1	7	
-2 -3 -16	1	-21	

Step 4:

matrix	sol	rhs	action
2 2 13	1	17	take this row
↓ ↓ ↓			
0 1 6	1	7	
-2 -3 -16	1	-21	plus × 1

Step 5:

matrix	sol	rhs	action
2 2 13	1	17	take this row
0 1 6	1	7	
0 -1 -3	1	-4	

Step 6:

matrix	sol	rhs	action
2 2 13	1	17	first column done
0 1 6	1	7	
0 -1 -3	1	-4	

Step 7:

matrix	sol	rhs	action
2 2 13	1	17	
0 1 6	1	7	take this row
0 -1 -3	1	-4	

Step 14:

matrix	sol	rhs	action
2 0 1	1	3	minus × 1/3
0 1 6	1	7	
↑↑↑ 0 0 3	1	3	take this row

Step 8:

matrix	sol	rhs	action
2 2 13	1	17	minus × 2
↑↑↑ 0 1 6	1	7	take this row
0 -1 -3	1	-4	

Step 15:

matrix	sol	rhs	action
2 0 0	1	2	
0 1 6	1	7	
0 0 3	1	3	take this row

Step 9:

matrix	sol	rhs	action
2 0 1	1	3	
0 1 6	1	7	take this row
0 -1 -3	1	-4	

Step 16:

matrix	sol	rhs	action
2 0 0	1	2	
0 1 6	1	7	minus × 2
↑↑↑ 0 0 3	1	3	take this row

Step 10:

matrix	sol	rhs	action
2 0 1	1	3	
0 1 6	1	7	take this row
0 -1 -3	1	-4	plus × 1

Step 17:

matrix	sol	rhs	action
2 0 0	1	2	
0 1 0	1	1	
0 0 3	1	3	take this row

Step 11:

matrix	sol	rhs	action
2 0 1	1	3	
0 1 6	1	7	
0 0 3	1	3	

Step 18:

matrix	sol	rhs	action
2 0 0	1	2	
0 1 0	1	1	
0 0 3	1	3	third column done

Step 12:

matrix	sol	rhs	action
2 0 1	1	3	
0 1 6	1	7	second column done
0 0 3	1	3	

Finished:

matrix	sol	rhs	action
2 0 0	1	2	
0 1 0	1	1	
0 0 3	1	3	

Figure 3.2: Gauss-Jordan elimination example

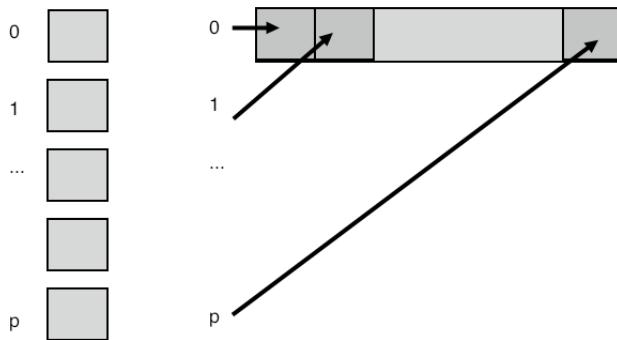


Figure 3.3: Gather collects all data onto a root

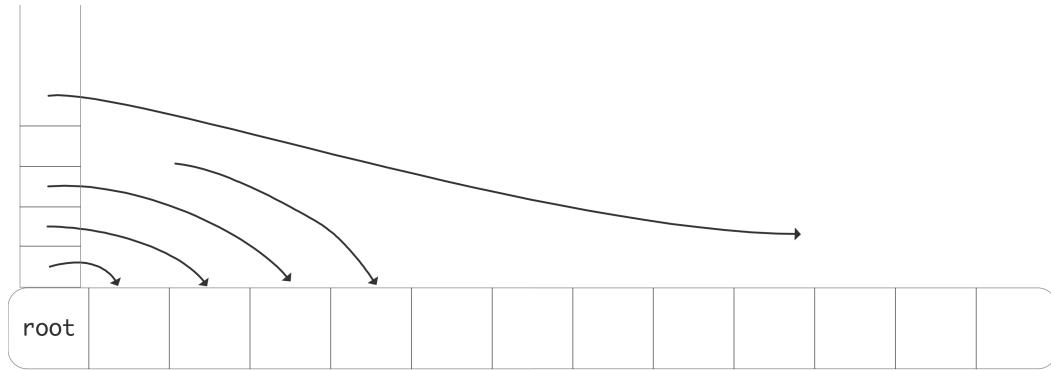


Figure 3.4: A scatter operation

The `sendcount` is not, as you might expect, the total length of the sendbuffer; instead, it is the amount of data sent to each process.

Exercise 3.8. 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.

In the gather and scatter calls, each processor has n elements of individual data. There is also a root processor that has an array of length np , where p is the number of processors. The gather call collects all this data from the processors to the root; the scatter call assumes that the information is initially on the root and it is spread to the individual processors.

The prototype for `MPI_Gather` (figure 12) has two ‘count’ parameters, one for the length of the individual send buffers, and one for the receive buffer. However, confusingly, the second parameter (which is only relevant on the root) does not indicate the total amount of information coming in, but rather the size of *each* contribution. Thus, the two count parameters will usually be the same (at least on the root); they can differ if you use different `MPI_Datatype` values for the sending and receiving processors.

Here is a small example:

```
// gather.c
// we assume that each process has a value "localsize"
```

MPI_Gather

```
C:  
int MPI_Gather(  
    const void* sendbuf, int sendcount, MPI_Datatype sendtype,  
    void* recvbuf, int recvcount, 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 recvcount: number of elements for any single receive (non-negative integer, significant only at root)  
IN recvtype: data type of recv buffer elements (significant only at root) (handle)  
IN root: rank of receiving process (integer)  
IN comm: communicator (handle)  
  
Fortran:  
MPI_Gather  
    (sendbuf, sendcount, sendtype, recvbuf, recvcount, recvtype,  
     root, comm, ierror)  
TYPE(*), DIMENSION(..), INTENT(IN) :: sendbuf  
TYPE(*), DIMENSION(..) :: recvbuf  
INTEGER, INTENT(IN) :: sendcount, recvcount, root  
TYPE(MPI_Datatype), INTENT(IN) :: sendtype, recvtype  
TYPE(MPI_Comm), INTENT(IN) :: comm  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python:  
MPI.Comm.Gather  
    (self, sendbuf, recvbuf, int root=0)
```

How to read routine prototypes: [1.5.4](#).

manpage 12: Routine prototype for MPI_Gather

```
// the root process collects these values

if (procno==root)
    localsizes = (int*) malloc( (nprocs+1)*sizeof(int) );

// everyone contributes their info
MPI_Gather(&localsize, 1, MPI_INT,
              localsizes, 1, MPI_INT, root, comm);
```

This will also be the basis of a more elaborate example in section 3.8.

The **MPI_IN_PLACE** option can be used for the send buffer on the root; the data for the root is then assumed to be already in the correct location in the receive buffer.

The **MPI_Scatter** operation is in some sense the inverse of the gather: the root process has an array of length np where p is the number of processors and n the number of elements each processor will receive.

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

3.4.1 Example

In some applications, each process computes a row or column of a matrix, but for some calculation (such as the determinant) it is more efficient to have the whole matrix on one process. You should of course only do this if this matrix is essentially smaller than the full problem, such as an interface system or the last coarsening level in multigrid.

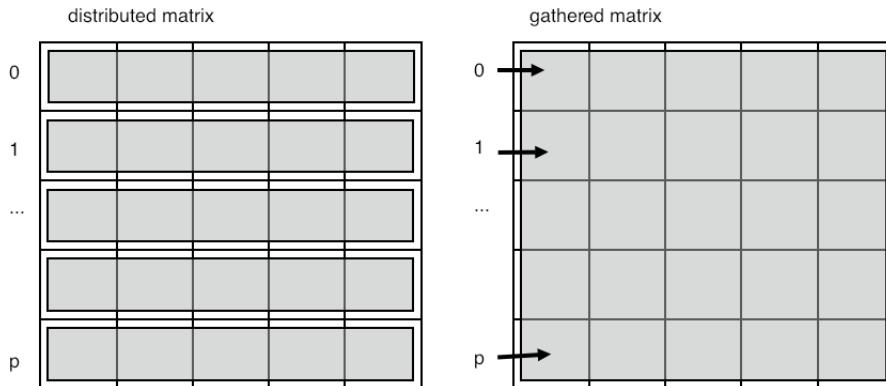


Figure 3.5: Gather a distributed matrix onto one process

Figure 3.5 pictures this. Note that conceptually we are gathering a two-dimensional object, but the buffer is of course one-dimensional. You will later see how this can be done more elegantly with the ‘subarray’ datatype; section 5.2.4.

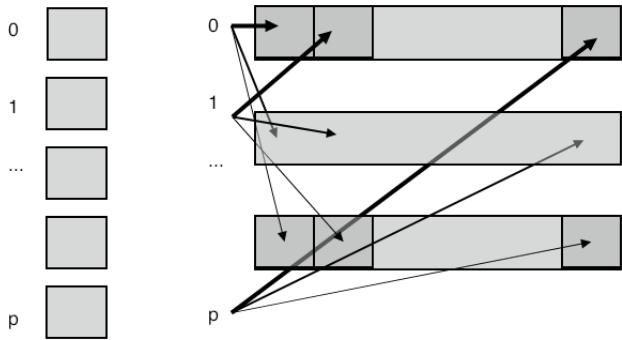


Figure 3.6: All gather collects all data onto every process

3.4.2 Allgather

The [MPI_Allgather](#) (figure 13) routine does the same gather onto every process: each process winds up with the totality of all data; figure 3.6.

This routine can be used in the simplest implementation of the *dense matrix-vector product* to give each processor the full input; see HPSC-??.

Some cases look like an all-gather but can be implemented more efficiently. Suppose you have two distributed vectors, and you want to create a new vector that contains those elements of the one that do not appear in the other. You could implement this by gathering the second vector on each processor, but this may be prohibitive in memory usage.

Exercise 3.9. Can you think of another algorithm for taking the set difference of two distributed vectors. Hint: look up ‘bucket-brigade algorithm’ in [5]. What is the time and space complexity of this algorithm? Can you think of other advantages beside a reduction in workspace?

3.5 All-to-all

The all-to-all operation [MPI_Alltoall](#) (figure 14) can be seen as a collection of simultaneous broadcasts or simultaneous gathers. The parameter specification is much like an allgather, with a separate send and receive buffer, and no root specified. As with the gather call, the receive count corresponds to an individual receive, not the total amount.

Unlike the gather call, the send buffer now obeys the same principle: with a send count of 1, the buffer has a length of the number of processes.

3.5.1 All-to-all as data transpose

The all-to-all operation can be considered as a data transpose. For instance, assume that each process knows how much data to send to every other process. If you draw a connectivity matrix of size $P \times P$, denoting

MPI_Allgather

C:

```
int MPI_Allgather(const void *sendbuf, int sendcount,
                  MPI_Datatype sendtype, void *recvbuf, int recvcount,
                  MPI_Datatype recvtype, MPI_Comm comm)
int MPI_Iallgather(const void *sendbuf, int sendcount,
                   MPI_Datatype sendtype, void *recvbuf, int recvcount,
                   MPI_Datatype recvtype, MPI_Comm comm, MPI_Request *request)
```

Fortran:

```
MPI_ALLGATHER(SENDBUF, SENDCOUNT, SENDTYPE, RECVBUF, REVCOUNT,
               RECVTYPE, COMM, IERROR)
<type>    SENDBUF (*), RECVBUF (*)
INTEGER     SENDCOUNT, SENDTYPE, REVCOUNT, RECVTYPE, COMM,
INTEGER     IERROR
MPI_IALLGATHER(SENDBUF, SENDCOUNT, SENDTYPE, RECVBUF, REVCOUNT,
                RECVTYPE, COMM, REQUEST, IERROR)
<type>    SENDBUF (*), RECVBUF (*)
INTEGER     SENDCOUNT, SENDTYPE, REVCOUNT, RECVTYPE, COMM
INTEGER     REQUEST, IERROR
```

C++ Syntax

Parameters:

sendbuf : Starting address of send buffer (choice).
sendcount: Number of elements in send buffer (integer).
sendtype: Datatype of send buffer elements (handle).
recvbuf: Starting address of recv buffer (choice).
recvcount: Number of elements received from any process (integer).
recvtype: Datatype of receive buffer elements (handle).
comm; Communicator (handle).

recvbuf: Address of receive buffer (choice).

request: Request (handle, non-blocking only).

How to read routine prototypes: [1.5.4](#).

manpage 13: Routine prototype for MPI>Allgather

MPI_Alltoall

```
int MPI_Alltoallv
      (void *sendbuf, int sendcnt, MPI_Datatype sendtype,
       void *recvbuf, int recvcnt, MPI_Datatype recvtype,
       MPI_Comm comm)
```

How to read routine prototypes: [1.5.4](#).

manpage 14: Routine prototype for MPI>Alltoall

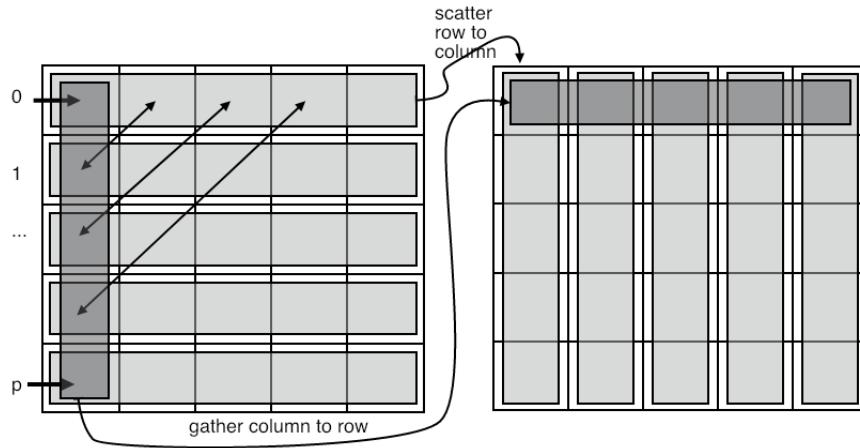


Figure 3.7: All-to-all transposes data

who-sends-to-who, then the send information can be put in rows:

$$\forall_i: C[i, j] > 0 \quad \text{if process } i \text{ sends to process } j.$$

Conversely, the columns then denote the receive information:

$$\forall_j: C[i, j] > 0 \quad \text{if process } j \text{ receives from process } i.$$

Exercise 3.10. In the initial stage of *radix sorting*, each process considers how many elements to send to every other process. Use **`MPI_Alltoall`** to derive from this how many elements they will receive from every other process.

On a larger scale, the typical application for the all-to-all operation is in the Fast Fourier Transform (FFT) algorithm, where it can take tens of percents of the running time.

3.5.2 All-to-all-v

Exercise 3.11. The actual data shuffle of a *radix sort* can be done with **`MPI_Alltoallv`**.

Finish the code of exercise 3.10.

3.6 Reduce-scatter

There are several MPI collectives that are functionally equivalent to a combination of others. You have already seen **`MPI_Allreduce`** which is equivalent to a reduction followed by a broadcast. Often such combinations can be more efficient than using the individual calls; see HPSC-??.

Here is another example: **`MPI_Reduce_scatter`** is equivalent to a reduction on an array of data (meaning a pointwise reduction on each array location) followed by a scatter of this array to the individual processes.

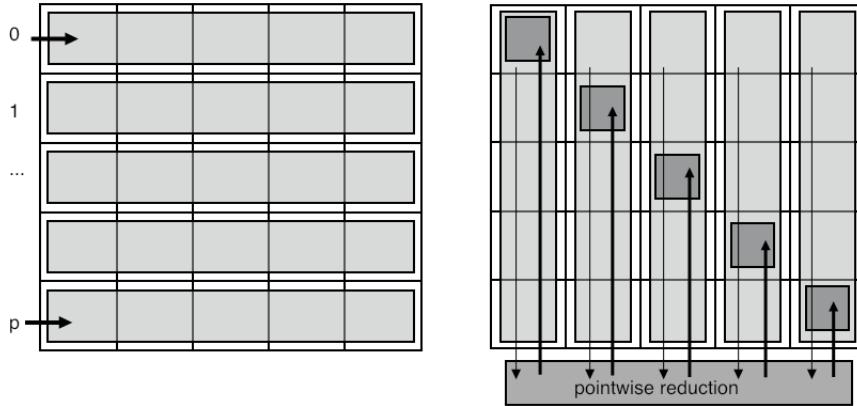


Figure 3.8: Reduce scatter

We can look at reduce-scatter as a limited form of the all-to-all data transposition discussed above (section 3.5.1). Suppose that the matrix C contains only 0/1, indicating whether or not a message is send, rather than the actual amounts. If a receiving process only needs to know how many messages to receive, rather than where they come from, it is enough to know the column sum, rather than the full column (see figure 3.8).

One important example of this command is the *sparse matrix-vector product*; see HPSC-?? for background information. Each process contains one or more matrix rows, so by looking at indices the process can decide what other processes it needs data from. The problem is for a process to find out what other processes it needs to send data to.

Using `MPI_Reduce_scatter` (figure 15) the process goes as follows:

- Each process creates an array of ones and zeros, describing who it needs data from.
- The reduce part of the reduce-scatter yields an array of requester counts; after the scatter each process knows how many processes request data from it.
- Next, the sender processes need to find out what elements are requested from it. For this, each process sends out arrays of indices.
- The big trick is that each process now knows how many of these requests will be coming in, so it can post precisely that many `MPI_Irecv` calls, with a source of `MPI_ANY_SOURCE`.

The `MPI_Reduce_scatter` command is equivalent to a reduction on an array of data, followed by a scatter of that data to the individual processes.

To be precise, there is an array `recvcounts` where `recvcounts[i]` gives the number of elements that ultimately wind up on process i . The result is equivalent to doing a reduction with a length equal to the sum of the `recvcounts[i]` values, followed by a scatter where process i receives `recvcounts[i]` values. (Since the amount of data to be scattered depends on the process, this is in fact equivalent to `MPI_Scatterv` rather than a regular scatter.)

For instance, if all `recvcounts[i]` values are 1, the sendbuffer has one element for each process, and the receive buffer has length 1.

MPI_Reduce_scatter

Semantics:

`MPI_REDUCE_SCATTER`
(`sendbuf, recvbuf, recvcounts, datatype, op, comm`)
`MPI_Reduce_scatter_block`
(`sendbuf, recvbuf, recvcount, datatype, op, comm`)

Input parameters:

`sendbuf`: starting address of send buffer (choice)
`recvcount`: element count per block (non-negative integer)
`recvcounts`: non-negative integer array (of length group size)
specifying the number of elements of the result distributed to each process.
`datatype`: data type of elements of send and receive buffers (handle)
`op`: operation (handle)
`comm`: communicator (handle)

Output parameters:

`recvbuf`: starting address of receive buffer (choice)

C:

```
int MPI_Reduce_scatter
  (const void* sendbuf, void* recvbuf, const int recvcounts[],
   MPI_Datatype datatype, MPI_Op op, MPI_Comm comm)
```

F:

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

Py:

```
comm.Reduce_scatter(sendbuf, recvbuf, recvcounts=None, Op op=SUM)
```

How to read routine prototypes: [1.5.4](#).

manpage 15: Routine prototype for MPI_Reduce_scatter

3.6.1 Examples

An important application of this is establishing an irregular communication pattern. Assume that each process knows which other processes it wants to communicate with; the problem is to let the other processes know about this. The solution is to use `MPI_Reduce_scatter` to find out how many processes want to communicate with you, and then wait for precisely that many messages with a source value of `MPI_ANY_SOURCE`

```
// reducescatter.c
// record what processes you will communicate with
int *recv_requests;
// find how many procs want to communicate with you
MPI_Reduce_scatter
    (recv_requests,&nsend_requests,counts,MPI_INT,
     MPI_SUM,comm);
// send a msg to the selected processes
for (int i=0; i<nprocs; i++)
    if (recv_requests[i]>0)
        MPI_Isend(&msg,1,MPI_INT, /*to:*/ i,0,comm,
                   mpi_requests+irequest++);
// do as many receives as you know are coming in
for (int i=0; i<nsend_requests; i++)
    MPI_Irecv(&msg,1,MPI_INT,MPI_ANY_SOURCE,MPI_ANY_TAG,comm,
               mpi_requests+irequest++);
MPI_Waitall(irequest,mpi_requests,MPI_STATUSES_IGNORE);
```

Use of `MPI_Reduce_scatter` to implement the two-dimensional matrix-vector product. Set up separate row and column communicators with `MPI_Comm_split`, use `MPI_Reduce_scatter` to combine local products.

```
MPI_Allgather(&my_x,1,MPI_DOUBLE,
              local_x,1,MPI_DOUBLE,environ.col_comm);
// bli_dgemv( BLIS_NO_TRANSPOSE,
//             BLIS_NO_CONJUGATE,
//             size_y, size_x,
//             &one,
//             local_matrix, 1, size_y,
//             local_x, 1,
//             &zero,
//             local_y, 1 );
// blas_dgemv(CblasColMajor,CblasNoTrans,
//             size_y,size_x,1.e0,
//             local_matrix,size_y,
//             local_x,1,0.e0,local_y,1);
MPI_Reduce_scatter(local_y,&my_y,&ione,MPI_DOUBLE,
                   MPI_SUM,environ.row_comm);
```

3.7 Barrier

A barrier call, `MPI_Barrier` (figure 16) is a routine that blocks all processes until they have all reached the barrier call. Thus it achieves time synchronization of the processes.

This call's simplicity is contrasted with its usefulness, which is very limited. It is almost never necessary to synchronize processes through a barrier: for most purposes it does not matter if processors are out of sync. Conversely, collectives (except the new non-blocking ones; section 3.11) introduce a barrier of sorts themselves.

3.8 Variable-size-input collectives

In the gather and scatter call above each processor received or sent an identical number of items. In many cases this is appropriate, but sometimes each processor wants or contributes an individual number of items.

Let's take the gather calls as an example. Assume that each processor does a local computation that produces a number of data elements, and this number is different for each processor (or at least not the same for all). In the regular **`MPI_Gather`** call the root processor had a buffer of size nP , where n is the number of elements produced on each processor, and P the number of processors. The contribution from processor p would go into locations $pn, \dots, (p + 1)n - 1$.

For the variable case, we first need to compute the total required buffer size. This can be done through a simple **`MPI_Reduce`** with **`MPI_SUM`** as reduction operator: the buffer size is $\sum_p n_p$ where n_p is the number of elements on processor p . But you can also postpone this calculation for a minute.

The next question is where the contributions of the processor will go into this buffer. For the contribution from processor p that is $\sum_{q < p} n_p, \dots, \sum_{q \leq p} n_p - 1$. To compute this, the root processor needs to have all the n_p numbers, and it can collect them with an **`MPI_Gather`** call.

We now have all the ingredients. All the processors specify a send buffer just as with **`MPI_Gather`**. However, the receive buffer specification on the root is more complicated. It now consists of:

```
outbuffer, array-of-outcounts, array-of-displacements, outtype
```

and you have just seen how to construct that information.

For example, in an **`MPI_Gatherv`** (figure 17) call each process has an individual number of items to contribute. To gather this, the root process needs to find these individual amounts with an **`MPI_Gather`** call, and locally construct the offsets array. Note how the offsets array has size `ntids+1`: the final offset value is automatically the total size of all incoming data. See the example below.

There are various calls where processors can have buffers of differing sizes.

- In **`MPI_Scatterv`** the root process has a different amount of data for each recipient.
- In **`MPI_Gatherv`**, conversely, each process contributes a different sized send buffer to the received result; **`MPI_Allgatherv`** does the same, but leaves its result on all processes; **`MPI_Alltoallv`** does a different variable-sized gather on each process.

```
|| int MPI_Scatterv
  ||| (void* sendbuf, int *sendcounts, int *displs, MPI_Datatype sendtype,
  |||   void* recvbuf, int recvcount, MPI_Datatype recvtype,
  |||   int root, MPI_Comm comm)
```

```
// int MPI_Allgatherv
//   (void *sendbuf, int sendcount, MPI_Datatype sendtype,
//    void *recvbuf, int *recvcounts, int *displs,
//    MPI_Datatype recvtype, MPI_Comm comm)
```

3.8.1 Example of Gatherv

We use **MPI_Gatherv** to do an irregular gather onto a root. We first need an **MPI_Gather** to determine offsets.

```
// gatherv.c
// we assume that each process has an array "localdata"
// of size "localsize"

// the root process decides how much data will be coming:
// allocate arrays to contain size and offset information
if (procno==root) {
    localsizes = (int*) malloc( (nprocs+1)*sizeof(int) );
    offsets = (int*) malloc( nprocs*sizeof(int) );
}
// everyone contributes their info
MPI_Gather(&localsize, 1, MPI_INT,
              localsizes, 1, MPI_INT, root, comm);
// the root constructs the offsets array
if (procno==root) {
    offsets[0] = 0;
    for (int i=0; i<nprocs; i++)
        offsets[i+1] = offsets[i]+localsizes[i];
    alldata = (int*) malloc( offsets[nprocs]*sizeof(int) );
}
// everyone contributes their data
MPI_Gatherv(localdata, localsize, MPI_INT,
              alldata, localsizes, offsets, MPI_INT, root, comm);

## gatherv.py
# implicitly using root=0
globalsize = comm.reduce(localsize)
if procid==0:
    print("Global size=%d" % globalsize)
    collecteddata = np.empty(globalsize, dtype=np.int)
    counts = comm.gather(localsize)
    comm.Gatherv(localdata, [collecteddata, counts])
```

3.8.2 Example of Allgatherv

Prior to the actual gatherv call, we need to construct the count and displacement arrays. The easiest way is to use a reduction.

```
// allgatherv.c
MPI_Allgather
  ( &my_count, 1, MPI_INT,
```

3. MPI topic: Collectives

```
    recv_counts,1,MPI_INT, comm );
int accumulate = 0;
for (int i=0; i<nprocs; i++) {
    recv_displs[i] = accumulate; accumulate += recv_counts[i]; }
int *global_array = (int*) malloc(accumulate*sizeof(int));
MPI_Allgatherv
( my_array,procno+1,MPI_INT,
  global_array,recv_counts,recv_displs,MPI_INT, comm );
```

In python the receive buffer has to contain the counts and displacements arrays.

```
## allgatherv.py
my_count = np.empty(1,dtype=np.int)
my_count[0] = mycount
comm.Allgather( my_count,recv_counts )

accumulate = 0
for p in range(nprocs):
    recv_displs[p] = accumulate; accumulate += recv_counts[p]
global_array = np.empty(accumulate,dtype=np.float64)
comm.Allgatherv( my_array, [global_array,recv_counts,recv_displs,MPI.DOUBLE]
    )
```

3.8.3 Variable all-to-all

MPI_Alltoallv (figure 18)

3.9 Scan operations

The **MPI_Scan** operation also performs a reduction, but it keeps the partial results. That is, if processor i contains a number x_i , and \oplus is an operator, then the scan operation leaves $x_0 \oplus \dots \oplus x_i$ on processor i . This type of operation is often called a *prefix operation*; see HPSC-??.

The **MPI_Scan** (figure 19) routine is an *inclusive scan* operation. The **MPI_Op** operations do not return an error code.

In python native mode the result is a function return value.

```
## scan.py
mycontrib = 10+random.randint(1,nprocs)
myfirst = 0
mypartial = comm.scan(mycontrib)
sbuf = np.empty(1,dtype=np.int)
rbuf = np.empty(1,dtype=np.int)
sbuf[0] = mycontrib
comm.Scan(sbuf,rbuf)
```

3.9.1 Exclusive scan

Often, the more useful variant is the *exclusive scan* `MPI_Exscan` (figure 20) with the same prototype.

The result of the exclusive scan is undefined on processor 0 (None in python), and on processor 1 it is a copy of the send value of processor 1. In particular, the `MPI_Op` need not be called on these two processors.

Exercise 3.12. The exclusive definition, which computes $x_0 \oplus x_{i-1}$ on processor i , can easily be derived from the inclusive operation for operations such as `MPI_SUM` or `MPI_MULT`. Are there operators where that is not the case?

3.9.2 Use of scan operations

The `MPI_Scan` operation is often useful with indexing data. Suppose that every processor p has a local vector where the number of elements n_p is dynamically determined. In order to translate the local numbering $0 \dots n_p - 1$ to a global numbering one does a scan with the number of local elements as input. The output is then the global number of the first local variable.

Exercise 3.13. Do you use `MPI_Scan` or `MPI_Exscan` for this operation? How would you describe the result of the other scan operation, given the same input?

Exclusive scan examples:

```
// exscan.c
int my_first=0, localsize;
// localsize = ..... result of local computation .....
// find myfirst location based on the local sizes
err = MPI_Exscan(&localsize,&my_first,
                 1,MPI_INT,MPI_SUM,comm); CHK(err);

## exscan.py
localsize = 10+random.randint(1,nprocs)
myfirst = 0
mypartial = comm.exscan(localsize,0)
```

It is possible to do a *segmented scan*. Let x_i be a series of numbers that we want to sum to X_i as follows. Let y_i be a series of booleans such that

$$\begin{cases} X_i = x_i & \text{if } y_i = 0 \\ X_i = X_{i-1} + x_i & \text{if } y_i = 1 \end{cases}$$

(This is the basis for the implementation of the *sparse matrix vector product* as prefix operation; see HPSC-??.) This means that X_i sums the segments between locations where $y_i = 0$ and the first subsequent place where $y_i = 1$. To implement this, you need a user-defined operator

$$\begin{pmatrix} X \\ x \\ y \end{pmatrix} = \begin{pmatrix} X_1 \\ x_1 \\ y_1 \end{pmatrix} \bigoplus \begin{pmatrix} X_2 \\ x_2 \\ y_2 \end{pmatrix} : \begin{cases} X = x_1 + x_2 & \text{if } y_2 == 1 \\ X = x_2 & \text{if } y_2 == 0 \end{cases}$$

This operator is not commutative, and it needs to be declared as such with `MPI_Op_create`; see section 3.10.2

3.10 MPI Operators

MPI operators are used in reduction operators. Most common operators, such as sum or maximum, have been built into the MPI library, but it is possible to define new operators.

3.10.1 Pre-defined operators

The following is the list of *pre-defined operators* `MPI_OP` values.

MPI type	meaning	applies to
<code>MPI_MAX</code>	maximum	integer, floating point
<code>MPI_MIN</code>	minimum	
<code>MPI_SUM</code>	sum	integer, floating point, complex, multilanguage types
<code>MPI_PROD</code>	product	
<code>MPI_LAND</code>	logical and	C integer, logical
<code>MPI_LOR</code>	logical or	
<code>MPI_LXOR</code>	logical xor	
<code>MPI_BAND</code>	bitwise and	integer, byte, multilanguage types
<code>MPI_BOR</code>	bitwise or	
<code>MPI_BXOR</code>	bitwise xor	
<code>MPI_MAXLOC</code>	max value and location	<code>MPI_DOUBLE_INT</code> and such
<code>MPI_MINLOC</code>	min value and location	

The `MPI_MAXLOC` operation yields both the maximum and the rank on which it occurs. However, to use it the input should be an array of `real/int` structs, where the `int` is the rank of the number.

3.10.2 User-defined operators

In addition to predefined operators, MPI has the possibility of *user-defined operators* to use in a reduction or scan operation.

The routine for this is `MPI_Op_create` (figure 21), where the user function needs to have the following prototype:

```
typedef void MPI_User_function
    ( void *invec, void *inoutvec, int *len,
      MPI_Datatype *datatype);

FUNCTION USER_FUNCTION( INVEC(*), INOUTVEC(*), LEN, TYPE)
<type> INVEC(LEN), INOUTVEC(LEN)
INTEGER LEN, TYPE
```

The python equivalent of such a function receives bare buffers as arguments. Therefore, it is best to turn them first into NumPY array using `np.frombuffer`:

```
def one_norm(in_buf, inout_buf, datatype):
    typecode = MPI._typecode(datatype)
```

```

assert typecode is not None # check MPI datatype is built-in
dtype = np.dtype(typecode)

in_array = np.frombuffer(in_buf, dtype)
inout_array = np.frombuffer(inout_buf, dtype)

```

The `assert` statement accounts for the fact that this mapping of MPI datatype to NumPy dtype only works for built-in MPI datatypes.

The function has an array length argument `len`, to allow for pointwise reduction on a whole array at once. The `inoutvec` array contains partially reduced results, and is typically overwritten by the function.

There are some restrictions on the user function:

- It may not call MPI functions, except for `MPI_Abort`.
- It must be associative; it can be optionally commutative, which fact is passed to the `MPI_Op_create` call.

For example, here is an operator for finding the smallest non-zero number in an array of nonnegative integers:

```

// reductpositive.c
void reduce_without_zero(void *in,void *inout,int *len,MPI_Datatype *type) {
    // r is the already reduced value, n is the new value
    int n = *(int*)in, r = *(int*)inout;
    int m;
    if (n==0) { // new value is zero: keep r
        m = r;
    } else if (r==0) {
        m = n;
    } else if (n<r) { // new value is less but not zero: use n
        m = n;
    } else { // new value is more: use r
        m = r;
    };
    *(int*)inout = m;
}

```

Exercise 3.14. Write the reduction function to implement the *one-norm* of a vector:

$$\|x\|_1 \equiv \sum_i |x_i|.$$

You can query the commutativity of an operator with `MPI_Op_commutative` (figure 22).

A created `MPI_Op` can be freed again:

```
// int MPI_Op_free(MPI_Op *op)
```

This sets the operator to `MPI_OP_NULL`.

3.10.3 Local reduction

The application of an `MPI_Op` can be performed with the routine `MPI_Reduce_local` (figure 23) . Using this routine and some send/receive scheme you can build your own global reductions. Note that this routine does not take a communicator because it is purely local.

3.11 Non-blocking collectives

Above you have seen how the ‘Isend’ and ‘Irecv’ routines can overlap communication with computation. This is not possible with the collectives you have seen so far: they act like blocking sends or receives. However, there are also *non-blocking collectives*. These have roughly the same calling sequence as their blocking counterparts, except that they output an `MPI_Request`. You can then use an `MPI_Wait` call to make sure the collective has completed.

Such operations can be used to increase efficiency. For instance, computing

$$y \leftarrow Ax + (x^t x)y$$

involves a matrix-vector product, which is dominated by computation in the *sparse matrix* case, and an inner product which is typically dominated by the communication cost. You would code this as

```
||| MPI_Iallreduce( .... x ...., &request);
||| // compute the matrix vector product
||| MPI_Wait(request);
||| // do the addition
```

This can also be used for 3D FFT operations [8]. Occasionally, a non-blocking collective can be used for non-obvious purposes, such as the `MPI_Ibarrier` in [9].

The same calling sequence as the blocking counterpart, except for the addition of an `MPI_Request` parameter. For instance `MPI_Ibcast` (figure 24) .

Non-blocking collectives offer a number of performance advantages:

- Do two reductions (on the same communicator) with different operators simultaneously;
- do collectives on overlapping communicators simultaneously;
- overlap a non-blocking collective with a blocking one. (However, note that blocking and non-blocking don’t match: either all process call the non-blocking or all call the blocking one.)

Exercise 3.15. Revisit exercise 6.1. Let only the first row and first column have certain data, which they broadcast through columns and rows respectively. Each process is now involved in two simultaneous collectives. Implement this with non-blocking broadcasts, and time the difference between a blocking and a non-blocking solution.

`MPI_Iallreduce` (figure 25)

`MPI_Iallgather` (figure 26)

3.11.1 Non-blocking barrier

Probably the most surprising non-blocking collective is the *non-blocking barrier* `MPI_Ibarrier` (figure 27). The way to understand this is to think of a barrier not in terms of temporal synchronization, but state agreement: reaching a barrier is a sign that a process has attained a certain state, and leaving a barrier means that all processes are in the same state. The ordinary barrier is then a blocking wait for agreement, while with a non-blocking barrier:

- Posting the barrier means that a process has reached a certain state; and
- the request being fulfilled means that all processes have reached the barrier.

We can use a non-blocking barrier to good effect, utilizing the idle time that would result from a blocking barrier. In the following code fragment processes test for completion of the barrier, and failing to detect such completion, perform some local work.

```
// ibarriertest.c
for ( ; ; step++) {
    int barrier_done_flag=0;
    MPI_Test(&barrier_request,&barrier_done_flag,MPI_STATUS_IGNORE);
    if (barrier_done_flag) {
        break;
    } else {
        int flag; MPI_Status status;
        MPI_Iprobe(
            MPI_ANY_SOURCE,MPI_ANY_TAG,
            comm, &flag, &status );
    }
}
```

3.12 Performance of collectives

It is easy to visualize a broadcast as in figure 3.9: see figure 3.9. the root sends all of its data directly to

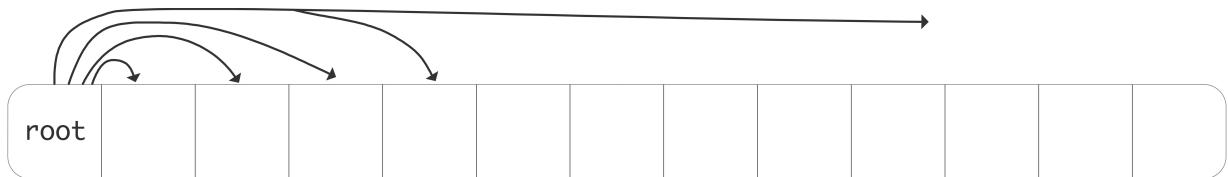


Figure 3.9: A simple broadcast

every other process. While this describes the semantics of the operation, in practice the implementation works quite differently.

The time that a message takes can simply be modeled as

$$\alpha + \beta n,$$

where α is the *latency*, a one time delay from establishing the communication between two processes, and β is the time-per-byte, or the inverse of the *bandwidth*, and n the number of bytes sent.

3. MPI topic: Collectives

Under the assumption that a processor can only send one message at a time, the broadcast in figure 3.9 would take a time proportional to the number of processors.

Exercise 3.16. What is the total time required for a broadcast involving p processes? Give α and β terms separately.

One way to ameliorate that is to structure the broadcast in a tree-like fashion. This is depicted in figure 3.10.

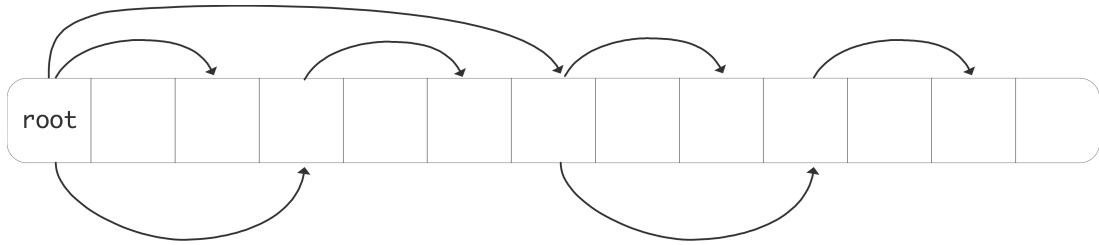


Figure 3.10: A tree-based broadcast

Exercise 3.17. How does the communication time now depend on the number of processors, again α and β terms separately.

What would be a lower bound on the α, β terms?

The theory of the complexity of collectives is described in more detail in HPSC-??; see also [1].

3.13 Collectives and synchronization

Collectives, other than a barrier, have a synchronizing effect between processors. For instance, in

```
|| MPI_Bcast( ....data... root);  
|| MPI_Send(....);
```

the send operations on all processors will occur after the root executes the broadcast. Conversely, in a reduce operation the root may have to wait for other processors. This is illustrated in figure 3.11, which gives a TAU trace of a reduction operation on two nodes, with two six-core sockets (processors) each. We see that¹:

- In each socket, the reduction is a linear accumulation;
- on each node, cores zero and six then combine their result;
- after which the final accumulation is done through the network.

We also see that the two nodes are not perfectly in sync, which is normal for MPI applications. As a result, core 0 on the first node will sit idle until it receives the partial result from core 12, which is on the second node.

While collectives synchronize in a loose sense, it is not possible to make any statements about events before and after the collectives between processors:

1. This uses mvapich version 1.6; in version 1.9 the implementation of an on-node reduction has changed to simulate shared memory.

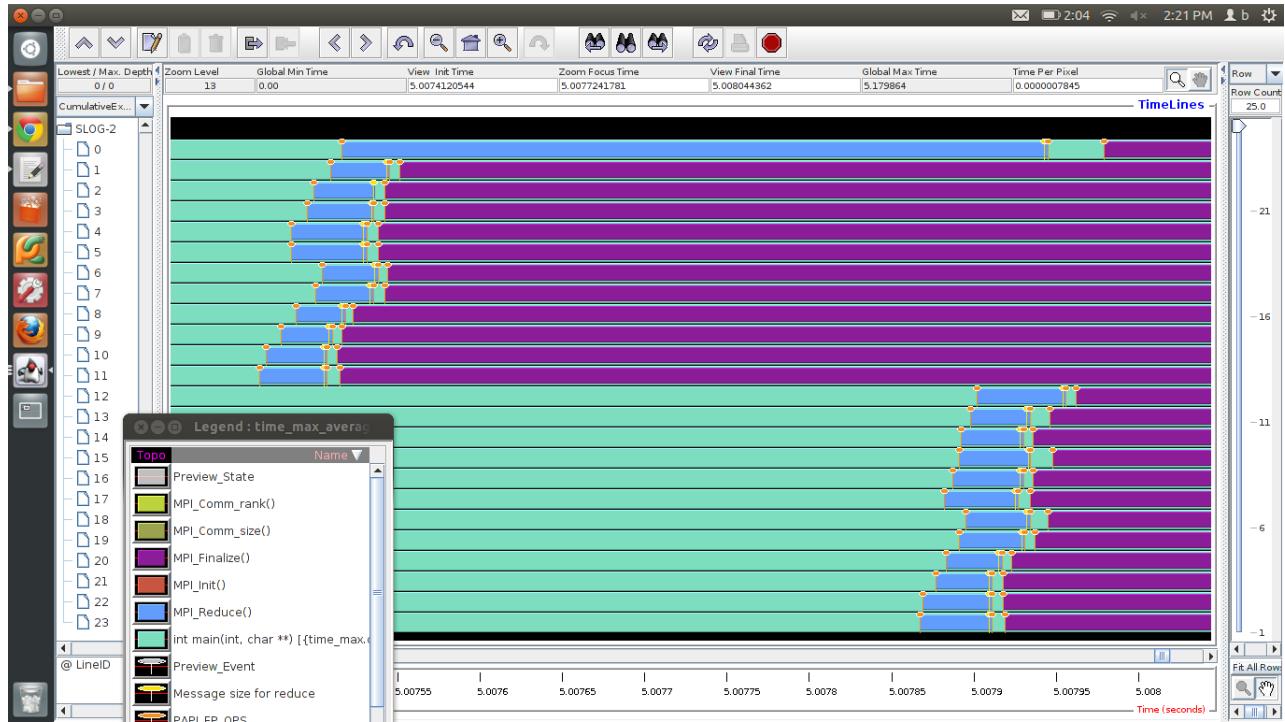


Figure 3.11: Trace of a reduction operation between two dual-socket 12-core nodes

```

|| ...event 1...
|| MPI_Bcast(...);
|| ...event 2...

```

Consider a specific scenario:

```

switch(rank) {
    case 0:
        MPI_Bcast(buf1, count, type, 0, comm);
        MPI_Send(buf2, count, type, 1, tag, comm);
        break;
    case 1:
        MPI_Recv(buf2, count, type, MPI_ANY_SOURCE, tag, comm, status);
        MPI_Bcast(buf1, count, type, 0, comm);
        MPI_Recv(buf2, count, type, MPI_ANY_SOURCE, tag, comm, status);
        break;
    case 2:
        MPI_Send(buf2, count, type, 1, tag, comm);
        MPI_Bcast(buf1, count, type, 0, comm);
        break;
}

```

Note the `MPI_ANY_SOURCE` parameter in the receive calls on processor 1. One obvious execution of this would be:

3. MPI topic: Collectives

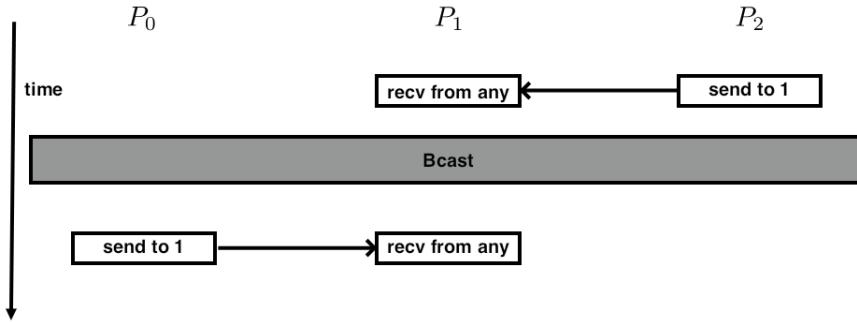
Code:

```

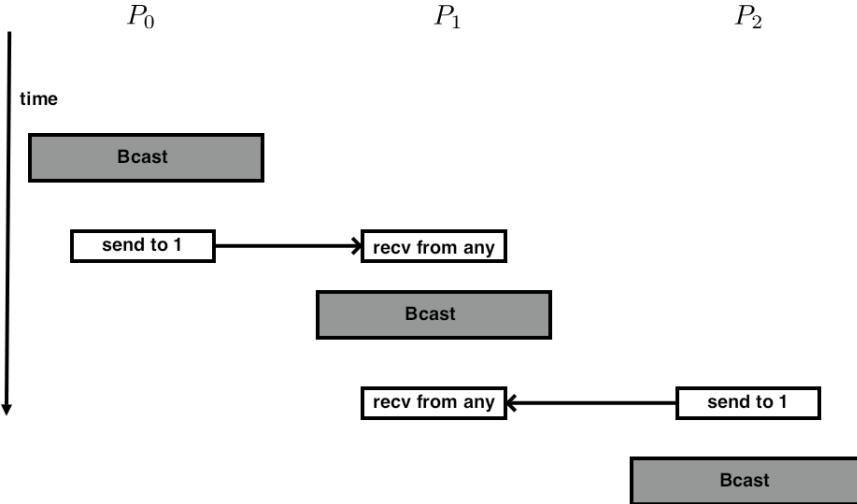
switch(rank) {
    case 0:
        MPI_Bcast(buf1, count, type, 0, comm);
        MPI_Send(buf2, count, type, 1, tag, comm);
        break;
    case 1:
        MPI_Recv(buf2, count, type, MPI_ANY_SOURCE, tag, comm, status);
        MPI_Bcast(buf1, count, type, 0, comm);
        MPI_Recv(buf2, count, type, MPI_ANY_SOURCE, tag, comm, status);
        break;
    case 2:
        MPI_Send(buf2, count, type, 1, tag, comm);
        MPI_Bcast(buf1, count, type, 0, comm);
        break;
}

```

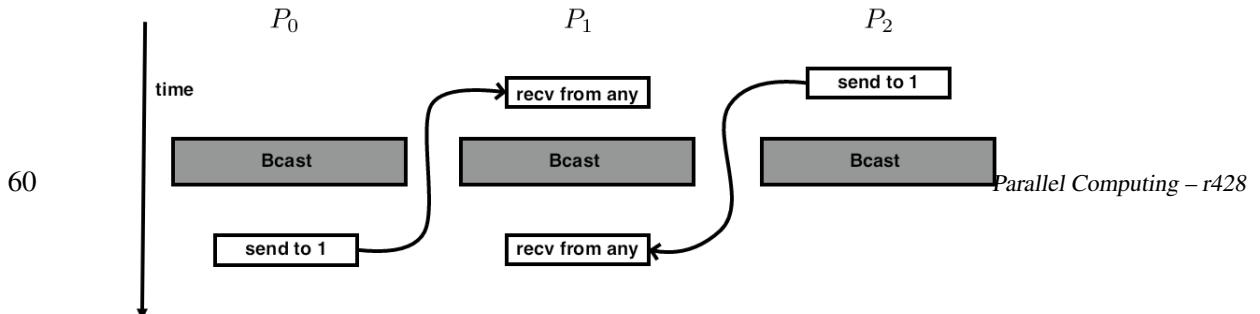
The most logical execution is:



However, this ordering is allowed too:



Which looks from a distance like:



In other words, one of the messages seems to go 'back in time'.

1. The send from 2 is caught by processor 1;
2. Everyone executes the broadcast;
3. The send from 0 is caught by processor 1.

However, it is equally possible to have this execution:

1. Processor 0 starts its broadcast, then executes the send;
2. Processor 1's receive catches the data from 0, then it executes its part of the broadcast;
3. Processor 1 catches the data sent by 2, and finally processor 2 does its part of the broadcast.

This is illustrated in figure 3.12.

3.14 Implementation and performance of collectives

In this section we will consider how collectives can be implemented in multiple ways, and the performance implications of such decisions. You can test the algorithms described here using *SimGrid* (section 38.5).

3.14.1 Broadcast

Naive broadcast Write a broadcast operation where the root does an `MPI_Send` to each other process.

What is the expected performance of this in terms of α, β ?

Run some tests and confirm.

Simple ring Let the root only send to the next process, and that one send to its neighbour. This scheme is known as a *bucket brigade*; see also section 4.2.3.

What is the expected performance of this in terms of α, β ?

Run some tests and confirm.

Pipelined ring In a ring broadcast, each process needs to receive the whole message before it can pass it on. We can increase the efficiency by breaking up the message and sending it in multiple parts. (See figure 3.13.) This will be advantageous for messages that are long enough that the bandwidth cost dominates the latency.

Assume a send buffer of length more than 1. Divide the send buffer into a number of chunks. The root sends the chunks successively to the next process, and each process sends on whatever chunks it receives.

What is the expected performance of this in terms of α, β ? Why is this better than the simple ring?

Run some tests and confirm.

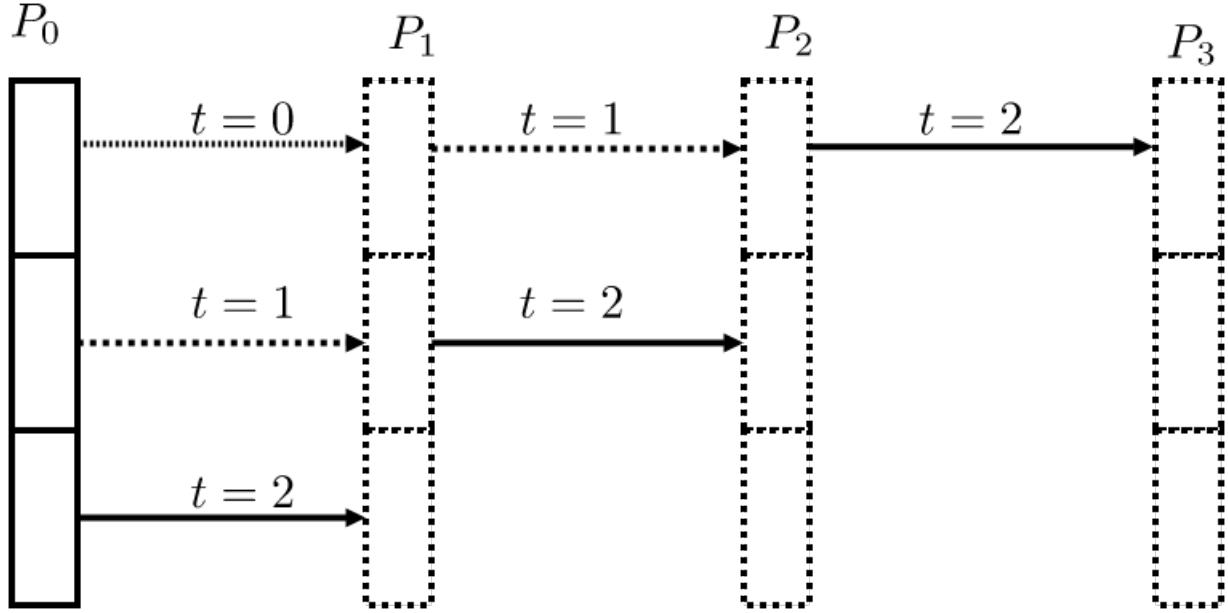


Figure 3.13: A pipelined bucket brigade

Recursive doubling Collectives such as broadcast can be *implemented* through *recursive doubling*, where the root sends to another process, then the root and the other process send to two more, those four send to four more, et cetera. However, in an actual physical architecture this scheme can be realized in multiple ways that have drastically different performance.

First consider the implementation where process 0 is the root, and it starts by sending to process 1; then they send to 2 and 3; these four send to 4–7, et cetera. If the architecture is a linear array of processors, this will lead to *contention*: multiple messages wanting to go through the same wire. (This is also related to the concept of *bisection bandwidth*.)

In the following analyses we will assume *wormhole routing*: a message sets up a path through the network, reserving the necessary wires, and performing a send in time independent of the distance through the network. That is, the send time for any message can be modeled as

$$T(n) = \alpha + \beta n$$

regardless source and destination, as long as the necessary connections are available.

Exercise 3.18. Analyze the running time of a recursive doubling broadcast as just described, with wormhole routing.

Implement this broadcast in terms of blocking MPI send and receive calls. If you have SimGrid available, run tests with a number of parameters.

The alternative, that avoids contention, is to let each doubling stage divide the network into separate halves. That is, process 0 sends to $P/2$, after which these two repeat the algorithm in the two halves of the network, sending to $P/4$ and $3P/4$ respectively.

Exercise 3.19. Analyze this variant of recursive doubling. Code it and measure runtimes on SimGrid.

Exercise 3.20. Revisit exercise 3.18 and replace the blocking calls by non-blocking `MPI_Isend` / `MPI_Irecv` calls.

Make sure to test that the data is correctly propagated.

3.15 Sources used in this chapter

Listing of code examples/mpi/c/allreducec.c:

Listing of code examples/mpi/c/allreducecp.c:

Listing of code examples/mpi/c/allreduceinplace.c:

Listing of code examples/mpi/c/reduce.c:

Listing of code examples/mpi/c/reduceinplace.c:

Listing of code XX:

Listing of code examples/mpi/c/usage.c:

Listing of code examples/mpi/c/gather.c:

Listing of code examples/mpi/c/reducescatter.c:

Listing of code examples/mpi/c/mvp2d.c:

Listing of code XX:

MPI_BARRIER

C:

```
int MPI_BARRIER( MPI_Comm comm )
```

Fortran2008:

```
MPI_BARRIER(COMM, IERROR)
Type(MPI_Comm), intent(int) :: COMM
INTEGER,intent(out) :: IERROR
```

Fortran 95:

```
MPI_BARRIER(COMM, IERROR)
INTEGER :: COMM, IERROR
```

Input parameter:

comm : Communicator (handle)

Output parameter:

Ierror : Error status (integer), Fortran only

How to read routine prototypes: [1.5.4.](#)

manpage 16: Routine prototype for MPI_BARRIER

MPI_Gatherv

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 the number of elements to receive from each process
IN displs: integer array (of length group size). Entry i specifies the displacement relative to the start of the receive buffer for process i
IN recvtype: data type of recv buffer elements (significant only at root) (handle)
IN root: rank of receiving process (integer)
IN comm: communicator (handle)

Fortran:

```
MPI_Gatherv(sendbuf, sendcount, sendtype, recvbuf, recvcounts, displs, recvtype, root, comm,
TYPE(*), DIMENSION(..), INTENT(IN) :: sendbuf
TYPE(*), DIMENSION(..) :: recvbuf
INTEGER, INTENT(IN) :: sendcount, recvcounts(*), displs(*), root
TYPE(MPI_Datatype), INTENT(IN) :: sendtype, recvtype
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
Gatherv(self, sendbuf, [recvbuf, counts], int root=0)
```

How to read routine prototypes: [1.5.4](#).

manpage 17: Routine prototype for MPI_Gatherv

MPI_Alltoallv

```
int MPI_Alltoallv
(void *sendbuf, int *sendcnts, int *sdispls, MPI_Datatype sendtype,
 void *recvbuf, int *recvcnts, int *rdispls, MPI_Datatype recvtype,
 MPI_Comm comm)
```

How to read routine prototypes: [1.5.4](#).

manpage 18: Routine prototype for MPI_Alltoallv

MPI_Scan

```
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, op=MPI.SUM)
```

How to read routine prototypes: [1.5.4](#).

manpage 19: Routine prototype for MPI_Scan

MPI_Exscan

C:

```
int MPI_Exscan(const void *sendbuf, void *recvbuf, int count,
    MPI_Datatype datatype, MPI_Op op, MPI_Comm comm)
int MPI_Iexscan(const void *sendbuf, void *recvbuf, int count,
    MPI_Datatype datatype, MPI_Op op, MPI_Comm comm,
    MPI_Request *request)
```

Fortran:

```
MPI_EXSCAN(SENDBUF, RECVBUF, COUNT, DATATYPE, OP, COMM, IERROR)
<type>    SENDBUF(*), RECVBUF(*)
INTEGER    COUNT, DATATYPE, OP, COMM, IERROR
MPI_IEXSCAN(SENDBUF, RECVBUF, COUNT, DATATYPE, OP, COMM, REQUEST, IERROR)
<type>    SENDBUF(*), RECVBUF(*)
INTEGER    COUNT, DATATYPE, OP, COMM, REQUEST, IERROR
```

Input Parameters

sendbuf: Send buffer (choice).
count: Number of elements in input buffer (integer).
datatype: Data type of elements of input buffer (handle).
op: Operation (handle).
comm: Communicator (handle).

Output Parameters

recvbuf: Receive buffer (choice).
request: Request (handle, non-blocking only).

How to read routine prototypes: [1.5.4](#).

manpage 20: Routine prototype for MPI_Exscan

MPI_Op_create

Semantics:

```
MPI_OP_CREATE( function, commute, op)
[ IN function] user defined function (function)
[ IN commute] true if commutative; false otherwise.
[ OUT op] operation (handle)
```

C:

```
int MPI_Op_create(MPI_User_function *function, int commute,
                  MPI_Op *op)
```

Fortran:

```
MPI_OP_CREATE( FUNCTION, COMMUTE, OP, IERROR)
EXTERNAL FUNCTION
LOGICAL COMMUTE
INTEGER OP, IERROR
```

Python:

```
MPI.Op.create(cls,function,bool commute=False)
```

How to read routine prototypes: [1.5.4](#).

manpage 21: Routine prototype for MPI_Op_create

MPI_Op_commutative

Semantics:

```
MPI_Op_commutative(op, commute)
IN op : handle
OUT commute : true/false
```

C:

```
int MPI_Op_commutative(MPI_Op op, int *commute)
```

Fortran:

```
MPI_OP_COMMUTATIVE( op, commute)
TYPE(MPI_Op), INTENT(IN) :: op
LOGICAL, INTENT(OUT) :: commute
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 22: Routine prototype for MPI_Op_commutative

MPI_Reduce_local

Semantics:

```
MPI_REDUCE_LOCAL( inbuf, inoutbuf, count, datatype, op)
```

Input parameters:

```
inbuf: input buffer (choice)
count: number of elements in inbuf and inoutbuf buffers
       (non-negative integer)
datatype: data type of elements of inbuf and inoutbuf buffers
       (handle)
op: operation (handle)
```

Input/output parameters:

```
inoutbuf: combined input and output buffer (choice)
```

C:

```
int MPI_Reduce_local
    (void* inbuf, void* inoutbuf, int count,
     MPI_Datatype datatype, MPI_Op op)
```

Fortran:

```
MPI_REDUCE_LOCAL(INBUF, INOUBUF, COUNT, DATATYPE, OP, IERROR)
<type> INBUF(*), INOUTBUF(*)
INTEGER :: COUNT, DATATYPE, OP, IERROR
```

How to read routine prototypes: 1.5.4.

manpage 23: Routine prototype for MPI_Reduce_local

MPI_Ibcast

C:

```
int MPI_Ibcast(
    void* buffer, int count, MPI_Datatype datatype,
    int root, MPI_Comm comm,
    MPI_Request *request
)
```

Fortran:

```
MPI_Ibcast(buffer, count, datatype, root, comm, ierror)
TYPE(*), DIMENSION(..) :: buffer
INTEGER, INTENT(IN) :: count, root
TYPE(MPI_Datatype), INTENT(IN) :: datatype
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
TYPE(MPI_Request), intent(out) :: request
```

How to read routine prototypes: 1.5.4.

manpage 24: Routine prototype for MPI_Ibcast

MPI_Iallreduce

Semantics

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

Input Parameters

```
sendbuf : starting address of send buffer (choice)
count : number of elements in send buffer (integer)
datatype : data type of elements of send buffer (handle)
op : operation (handle)
comm : communicator (handle)
```

Output Parameters

```
recvbuf : starting address of receive buffer (choice)
request : communication request (handle)
```

How to read routine prototypes: 1.5.4.

manpage 25: Routine prototype for MPI_Iallreduce

MPI_Iallgather

Semantics

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

Input Parameters

```
sendbuf : starting address of send buffer (choice)
sendcount : number of elements in send buffer (integer)
sendtype : data type of send buffer elements (handle)
recvcount : number of elements received from any process (integer)
recvtype : data type of receive buffer elements (handle)
comm : communicator (handle)
```

Output Parameters

```
recvbuf : address of receive buffer (choice)
request : communication request (handle)
```

How to read routine prototypes: 1.5.4.

manpage 26: Routine prototype for MPI_Iallgather

MPI_Ibarrier

```
C:  
int MPI_Ibarrier(MPI_Comm comm, MPI_Request *request)  
  
Input Parameters  
comm : communicator (handle)  
  
Output Parameters  
request : communication request (handle)  
  
Fortran2008:  
MPI_Ibarrier(comm, request, ierror)  
Type(MPI_Comm), intent(int) :: comm  
TYPE(MPI_Request), intent(out) :: request  
INTEGER, intent(out) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 27: Routine prototype for MPI_Ibarrier

Chapter 4

MPI topic: Point-to-point

4.1 Distributed computing and distributed data

One reason for using MPI is that sometimes you need to work on more data than can fit in the memory of a single processor. With distributed memory, each processor then gets a part of the whole data structure and only works on that.

So let's say we have a large array, and we want to distribute the data over the processors. That means that, with p processes and n elements per processor, we have a total of $n \cdot p$ elements.

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

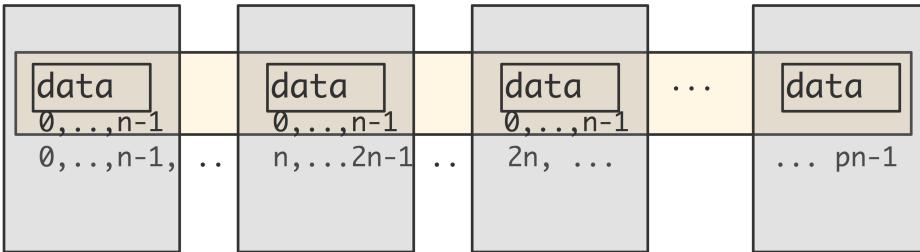


Figure 4.1: Local parts of a distributed array

We sometimes say that `data` is the local part of a *distributed array* with a total size of $n \cdot p$ elements. However, this array only exists conceptually: each processor has an array with lowest index zero, and you have to translate that yourself to an index in the global array. In other words, you have to write your code in such a way that it acts like you're working with a large array that is distributed over the processors, while actually manipulating only the local arrays on the processors.

Your typical code then looks like

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

Exercise 4.1. 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

Exercise 4.2. In the previous exercise you worked with a distributed array, computing a local quantity and combining that into a global quantity. Why is it not a good idea to gather the whole distributed array on a single processor, and do all the computation locally?

If the array size is not perfectly divisible by the number of processors, we have to come up with a division that is uneven, but not too much. You could for instance, write

```

|| int Nglobal, // is something large
|| Nlocal = Nglobal/ntids,
|| excess = Nglobal%ntids;
|| if (mytid==ntids-1)
||     Nlocal += excess;

```

Exercise 4.3. Argue that this strategy is not optimal. Can you come up with a better distribution? Load balancing is further discussed in HPSC-??.

Exercise 4.4. 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.

4.2 Blocking point-to-point operations

Suppose you have an array of numbers $x_i : i = 0, \dots, N$ and you want to compute $y_i = (x_{i-1} + x_i + x_{i+1})/3 : i = 1, \dots, N - 1$. As before (see figure 4.1), we give each processor a subset of the x_i s and y_i s. Let's define i_p as the first index of y that is computed by processor p . (What is the last index computed by processor p ? How many indices are computed on that processor?)

We often talk about the *owner computes* model of parallel computing: each processor ‘owns’ certain data items, and it computes their value.

4. MPI topic: Point-to-point

Now let's investigate how processor p goes about computing y_i for the i -values it owns. Let's assume that processor p also stores the values x_i for these same indices. Now, for many values it can compute

$$y_i = (x_{i-1} + x_i + x_{i+1})/3$$

(figure 4.3). However, there is a problem with computing the first index i_p :

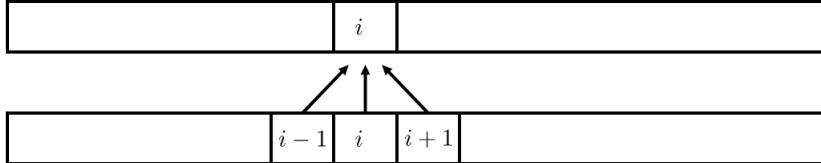


Figure 4.2: Three point averaging

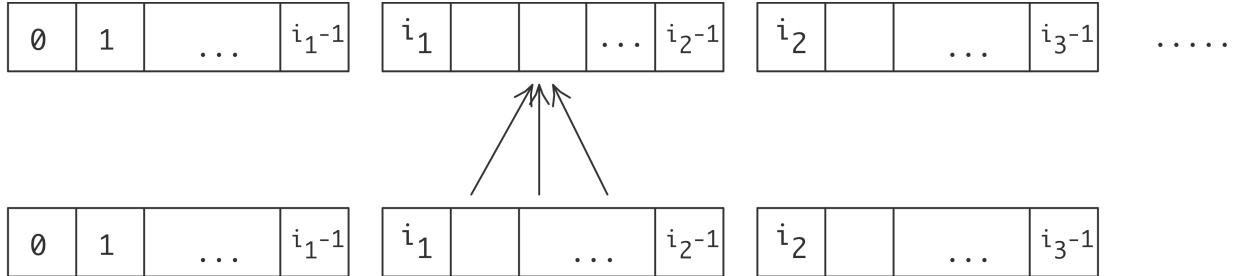


Figure 4.3: Three point averaging in parallel

$$y_{i_p} = (x_{i_p-1} + x_{i_p} + x_{i_p+1})/3$$

since x_{i_p} is not stored on processor p : it is stored on $p - 1$ (figure 4.4). There is a similar story with the last

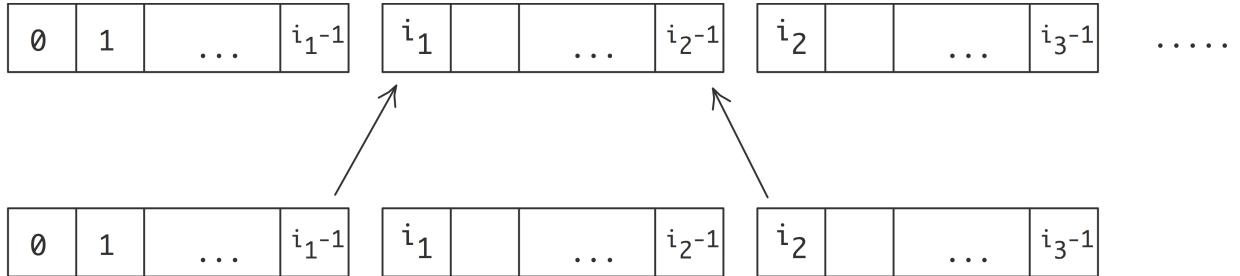


Figure 4.4: Three point averaging in parallel, case of edge points

index that p tries to compute: that involves a value that is only present on $p + 1$.

You see that there is a need for processor-to-processor, or technically *point-to-point*, information exchange. MPI realizes this through matched send and receive calls:

- One process does a send to a specific other process;
- the other process does a specific receive from that source.

4.2.1 Send example: ping-pong

A simple scenario for information exchange between just two processes is the *ping-pong*: process A sends data to process B, which sends data back to A. This means that process A executes the code

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

while process B executes

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

Since we are programming in SPMD mode, this means our program looks like:

```
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 .... );
}
```

4.2.1.1 Send call

The blocking send command is **MPI_Send** (figure 28). The features of this call are:

4.2.1.1.1 Data description The data to be sent is described by a trio of buffer/count/datatype. This is common to just about any MPI routine that involves data transfer.

In C the buffer is a memory address, so it is treated slightly differently between variables and arrays:

```
int counter;
MPI_Send( &counter, 1, MPI_INT, /* ... */ );
float point[2];
MPI_Send( point, 2, MPI_FLOAT, /* ... */ );
```

Since Fortran uses a parameter *passing by reference* mechanism, variables and arrays are treated the same:

```
integer :: counter
real(4) :: point[2]
call MPI_Send( counter, 1, MPI_INTEGER, ... )
call MPI_Send( point, 2, MPI_REAL4, ... )
```

Python sends objects, which document their own size and type, so the send call has only a single parameter describing the data:

```
comm.send( python_object, ... )
comm.Send( numpy_object, ... )
```

MPI_Send

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:
MPI.Comm.Send(self, buf, int dest, int tag=0)
```

How to read routine prototypes: [1.5.4](#).

manpage 28: Routine prototype for MPI_Send

4.2.1.1.2 *Target* Send calls have a *messsage target*: they require an explicit process rank to send to. This rank is a number from zero up to the result of `MPI_Comm_size`.

This aspect of the send call shows the symmetric nature of MPI: every target process is reached with the same send call, no matter whether it's running on the same multicore chip as the sender, or on a computational node halfway across the machine room. Of course, any self-respecting MPI implementation optimizes for the case where sender and receiver have access to the same shared memory. However, even then, there will be a copy operation from the sender buffer to the receiver buffer, so there is no actual memory sharing going on.

4.2.1.1.3 *Tag* Many applications have each sender send only one message to a given receiver. For the case where there are multiple messages between the same sender / receiver pair, the *message tag* can be used to disambiguate between the messages.

Unless otherwise needed, a tag value of zero is safe to use. If you do use tag values, you can use the key `MPI_TAG_UB` to query what the maximum value is that can be used; see section 12.5.3.

4.2.1.2 Receive call

The basic blocking receive command is `MPI_Recv` (figure 29)

4.2.1.2.1 *Data description* The receive call has the same buffer/count/data parameters as the send call. However, the `count` argument here indicates the maximum length of a message; the actual length of the received message can be determined from the status object, which is described below.

4.2.1.2.2 *Source* Mirroring the target argument of the `MPI_Send` call, `MPI_Recv` has a *message source* argument. This can be either a specific process rank, or it can be the `MPI_ANY_SOURCE` wildcard. In the latter case, the actual source can be determined after the message has been received; see below.

4.2.1.2.3 *Tag* Similar to the messsage source, the message tag of a receive call can be a specific value or a wildcard, in this case `MPI_ANY_TAG`. Again, see below.

4.2.1.2.4 *Status* In the syntax of the `MPI_Recv` command you saw one parameter that the send call lacks: the `MPI_Status` object. This serves the following purpose: the receive call can have a ‘wildcard’ behaviour, for instance specifying that the message can come from any source rather than a specific one. Similarly, the tag value can be wildcared. The status object then allows you to find out the actual message source and tag, and the message size.

See section 4.4.2 for more about the status object.

Exercise 4.5. Implement the ping-pong program. Add a timer using `MPI_Wtime`. For the `status` argument of the receive call, use `MPI_STATUS_IGNORE`.

MPI_Recv

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:

```
recvbuf = Comm.recv(self, buf=None, int source=ANY_SOURCE, int tag=ANY_TAG,
                     Status status=None)
Python numpy:
Comm.Recv(self, buf, int source=ANY_SOURCE, int tag=ANY_TAG,
          Status status=None)
```

How to read routine prototypes: [1.5.4](#).

manpage 29: Routine prototype for MPI_Recv

- 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 α, β ?

Exercise 4.6. Take your pingpong program and modify it to let half the processors be source and the other half the targets. Does the pingpong time increase?

4.2.2 Problems with blocking communication

The use of `MPI_Send` and `MPI_Recv` is known as *blocking communication*: when your code reaches a send or receive call, it blocks until the call is successfully completed. For a receive call it is clear that the receiving code will wait until the data has actually come in, but for a send call this is more subtle.

You may be tempted to think that the send call puts the data somewhere in the network, and the sending code can progress, as in figure 4.5, left. But this ideal scenario is not realistic: it assumes that somewhere

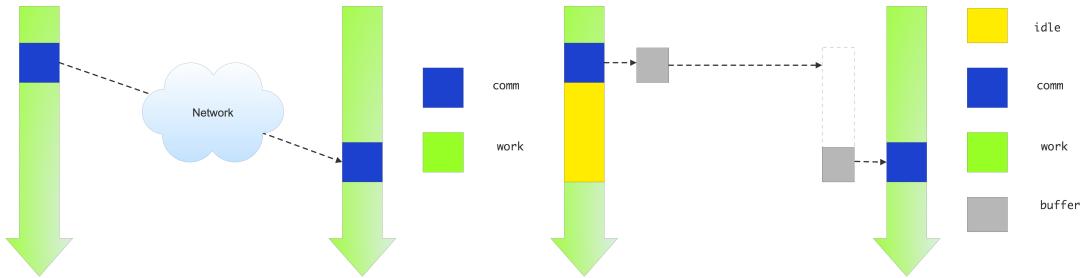


Figure 4.5: Illustration of an ideal (left) and actual (right) send-receive interaction

in the network there is buffer capacity for all messages that are in transit. This is not the case: data resides on the sender, and the sending call blocks, until the receiver has received all of it. (There is an exception for small messages, as explained in the next section.)

4.2.2.1 Deadlock

Suppose two processes need to exchange data, and consider the following pseudo-code, which purports to exchange data between processes 0 and 1:

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

Imagine that the two processes execute this code. They both issue the send call... and then can't go on, because they are both waiting for the other to issue a receive call. This is known as *deadlock*.

4.2.2.2 Eager limit

If you reverse the send and receive call, you should get deadlock, but in practice that code will often work. The reason is that MPI implementations sometimes send small messages regardless of whether the receive has been posted. This relies on the availability of some amount of available buffer space. The size under which this behaviour is used is sometimes referred to as the *eager limit*.)

The following code is guaranteed to block, since a **MPI_Recv** always blocks:

```
// recvblock.c
other = 1-procno;
MPI_Recv(&recvbuf,1,MPI_INT,other,0,comm,&status);
MPI_Send(&sendbuf,1,MPI_INT,other,0,comm);
printf("This statement will not be reached on %d\n",procno);
```

On the other hand, if we put the send call before the receive, code may not block for small messages that fall under the eager limit.

To illustrate eager and blocking behavior in **MPI_Send**, consider an example where we send gradually larger messages. From the screen output you can see what the largest message was that fell under the eager limit; after that the code hangs because of a deadlock.

```
// sendblock.c
other = 1-procno;
/* loop over increasingly large messages */
for (int size=1; size<2000000000; size*=10) {
    sendbuf = (int*) malloc(size*sizeof(int));
    recvbuf = (int*) malloc(size*sizeof(int));
    if (!sendbuf || !recvbuf) {
        printf("Out of memory\n"); MPI_Abort(comm,1);
    }
    MPI_Send(sendbuf,size,MPI_INT,other,0,comm);
    MPI_Recv(recvbuf,size,MPI_INT,other,0,comm,&status);
    /* If control reaches this point, the send call
       did not block. If the send call blocks,
       we do not reach this point, and the program will hang.
   */
    if (procno==0)
        printf("Send did not block for size %d\n",size);
    free(sendbuf); free(recvbuf);
}

// sendblock.F90
other = 1-mytid
size = 1
do
    allocate(sendbuf(size)); allocate(recvbuf(size))
    print *,size
    call MPI_Send(sendbuf,size,MPI_INTEGER,other,0,comm,err)
    call MPI_Recv(recvbuf,size,MPI_INTEGER,other,0,comm,status,err)
    if (mytid==0) then
        print *, "MPI_Send did not block for size",size
    end if
```

```

    deallocate(sendbuf); deallocate(recvbuf)
    size = size*10
    if (size>2000000000) goto 20
end do
20  continue

## sendblock.py
size = 1
while size<2000000000:
    sendbuf = np.empty(size, dtype=np.int)
    recvbuf = np.empty(size, dtype=np.int)
    comm.Send(sendbuf, dest=other)
    comm.Recv(recvbuf, source=other)
    if procid<other:
        print("Send did not block for", size)
    size *= 10

```

If you want a code to exhibit the same blocking behavior for all message sizes, you force the send call to be blocking by using **MPI_Ssend**, which has the same calling sequence as **MPI_Send**.

```

// ssendblock.c
other = 1-procno;
sendbuf = (int*) malloc(sizeof(int));
recvbuf = (int*) malloc(sizeof(int));
size = 1;
MPI_Ssend(sendbuf, size, MPI_INT, other, 0, comm);
MPI_Recv(recvbuf, size, MPI_INT, other, 0, comm, &status);
printf("This statement is not reached\n");

```

Formally you can describe deadlock as follows. Draw up a graph where every process is a node, and draw a directed arc from process A to B if A is waiting for B. There is deadlock if this directed graph has a loop.

The solution to the deadlock in the above example is to first do the send from 0 to 1, and then from 1 to 0 (or the other way around). So the code would look like:

```

if ( /* I am processor 0 */ ) {
    send(target=other);
    receive(source=other);
} else {
    receive(source=other);
    send(target=other);
}

```

The eager limit is implementation-specific. For instance, for *Intel mpi* there is a variable **I_MPI_EAGER_THRESHOLD**, for *mavapich2* it is **MV2_IBA_EAGER_THRESHOLD**, and for *OpenMPI* the **--mca** options **btl_openib_eager_limit** and **btl_openib_rndv_eager_limit**.

4.2.2.3 Serialization

There is a second, even more subtle problem with blocking communication. Consider the scenario where every processor needs to pass data to its successor, that is, the processor with the next higher rank. The

4. MPI topic: Point-to-point

basic idea would be to first send to your successor, then receive from your predecessor. Since the last processor does not have a successor it skips the send, and likewise the first processor skips the receive. The pseudo-code looks like:

```

successor = mytid+1; predecessor = mytid-1;
if ( /* I am not the last processor */
    send(target=successor);
if ( /* I am not the first processor */
    receive(source=predecessor)

```

This code does not deadlock. All processors but the last one block on the send call, but the last processor executes the receive call. Thus, the processor before the last one can do its send, and subsequently continue to its receive, which enables another send, et cetera.

In one way this code does what you intended to do: it will terminate (instead of hanging forever on a deadlock) and exchange data the right way. However, the execution now suffers from unexpected *serialization*: only one processor is active at any time, so what should have been a parallel operation becomes a sequential

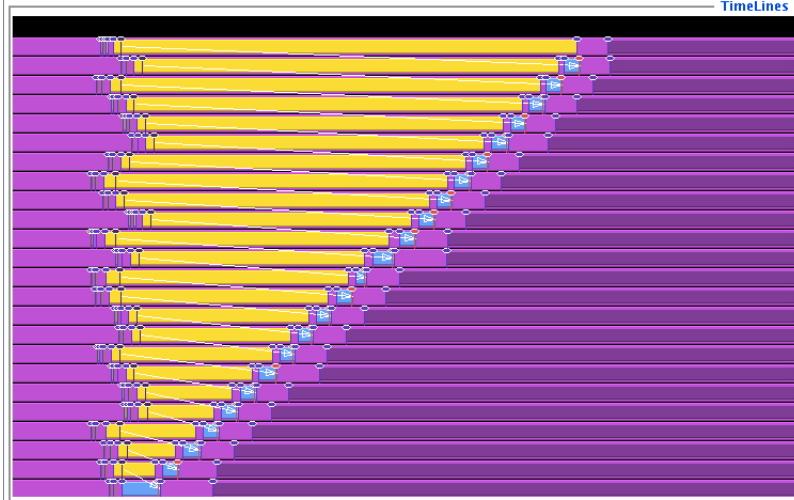


Figure 4.6: Trace of a simple send-recv code

one. This is illustrated in figure 4.6.

Exercise 4.7. (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.

Exercise 4.8. Implement the above algorithm using `MPI_Send` and `MPI_Recv` calls. Run the code, and use TAU to reproduce the trace output of figure 4.6. If you don't have TAU, can you show this serialization behaviour using timings?

It is possible to orchestrate your processes to get an efficient and deadlock-free execution, but doing so is a bit cumbersome.

Exercise 4.9. The above solution treated every processor equally. Can you come up with a solution that uses blocking sends and receives, but does not suffer from the serialization behaviour?

There are better solutions which we will explore in the next section.

4.2.3 Bucket brigade

The problem with the previous exercise was that an operation that was conceptually parallel, became serial in execution. On the other hand, sometimes the operation is actually serial in nature. One example is the *bucket brigade* operation, where a piece of data is successively passed down a sequence of processors.

Exercise 4.10. Take the code of exercise 4.8 and modify it so that the data from process zero gets propagated to every process. Specifically: compute

$$\begin{cases} x_0 = 1 & \text{on process zero} \\ x_p = x_{p-1} + (p+1)^2 & \text{on process } p \end{cases}$$

Use `MPI_Send` and `MPI_Recv`; make sure to get the order right.

4.2.4 Pairwise exchange

Above you saw that with blocking sends the precise ordering of the send and receive calls is crucial. Use the wrong ordering and you get either deadlock, or something that is not efficient at all in parallel. MPI has a way out of this problem that is sufficient for many purposes: the combined send/recv call `MPI_Sendrecv` (figure 30) .

The sendrecv call works great if every process is paired up. You would then write

```
|| sendrecv( ....from... ...to... );
```

with the right choice of source and destination. For instance, to send data to your right neighbour:

```
|| MPI_Comm_rank(comm, &procno);
|| MPI_Sendrecv( ....
||   /* from: */ procno-1
||   ....
||   /* to: */   procno+1
||   ... );
```

This scheme is correct for all processes but the first and last. MPI allows for the following solution which makes the code slightly more homogeneous:

MPI_Sendrecv

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)
```

C:

```
int MPI_Sendrecv(
    const void *sendbuf, int sendcount, MPI_Datatype sendtype,
    int dest, int sendtag,
    void *recvbuf, int recvcount, MPI_Datatype recvtype,
    int source, int recvtag,
    MPI_Comm comm, MPI_Status *status)
```

Fortran:

```
MPI_Sendrecv(sendbuf, sendcount, sendtype, dest, sendtag, recvbuf,
            recvcount, recvtype, source, recvtag, comm, status, ierror)
TYPE(*), DIMENSION(..), INTENT(IN) :: sendbuf
TYPE(*), DIMENSION(..) :: recvbuf
INTEGER, INTENT(IN) :: sendcount, dest, sendtag, recvcount, source,
recvtag
TYPE(MPI_Datatype), INTENT(IN) :: sendtype, recvtype
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Status) :: status
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
Sendrecv(self, sendbuf, int dest, int sendtag=0,
         recvbuf=None, int source=ANY_SOURCE, int recvtag=ANY_TAG,
         Status status=None)
```

How to read routine prototypes: [1.5.4](#).

manpage 30: Routine prototype for MPI`Sendrecv

```

MPI_Comm_rank( .... &mytid );
if ( /* I am not the first processor */ )
    predecessor = mytid-1;
else
    predecessor = MPI_PROC_NULL;
if ( /* I am not the last processor */ )
    successor = mytid+1;
else
    successor = MPI_PROC_NULL;
sendrecv(from=predecessor,to=successor);

```

where the sendrecv call is executed by all processors.

All processors but the last one send to their neighbour; the target value of `MPI_PROC_NULL` (figure 31) for the last processor means a ‘send to the null processor’: no actual send is done. The null processor value is also of use with the `MPI_Sendrecv` call; section 4.2.4

Likewise, receive from `MPI_PROC_NULL` succeeds without altering the receive buffer. The corresponding `MPI_Status` object has source `MPI_PROC_NULL`, tag `MPI_ANY_TAG`, and count zero.

Exercise 4.11. Revisit exercise 4.7 and solve it using `MPI_Sendrecv`.

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.

This call makes it easy to exchange data between two processors: both specify the other as both target and source. However, there need not be any such relation between target and source: it is possible to receive from a predecessor in some ordering, and send to a successor in that ordering; see figure 4.7.

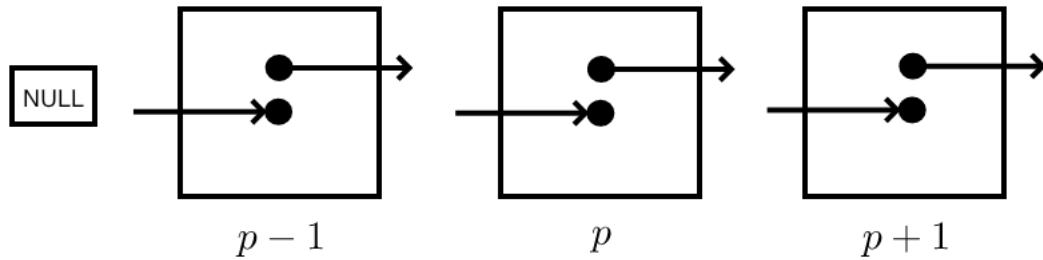


Figure 4.7: An MPI Sendrecv call

For the above three-point combination scheme you need to move data both left right, so you need two `MPI_Sendrecv` calls; see figure 4.8.

Exercise 4.12. Implement the above three-point combination scheme using `MPI_Sendrecv`; every processor only has a single number to send to its neighbour.

- Each process does one send and one receive; if a process needs to skip one or the other, you can specify `MPI_PROC_NULL` as the other process in the send or receive specification. In that case the corresponding action is not taken.
- As with the simple send/recv calls, processes have to match up: if process p specifies p' as the destination of the send part of the call, p' needs to specify p as the source of the recv part.

4. MPI topic: Point-to-point

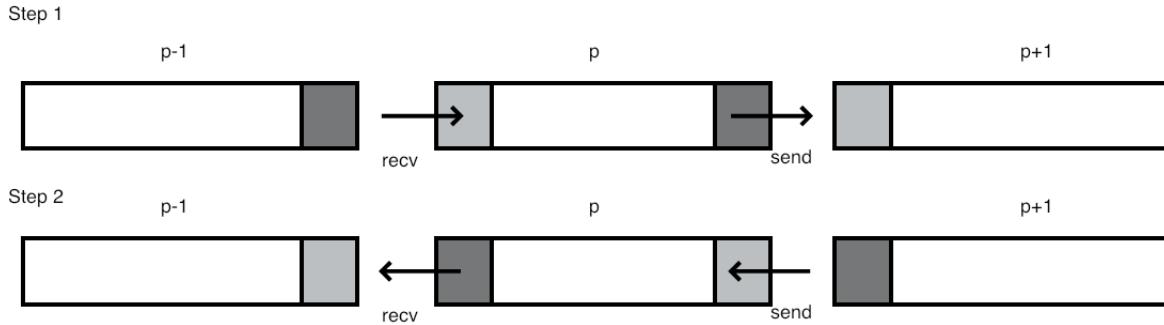


Figure 4.8: Two steps of send/recv to do a three-point combination

If the send and receive buffer have the same size, the routine `MPI_Sendrecv_replace` (figure 32) will do an in-place replacement.

The following exercise lets you implement a sorting algorithm with the send-receive call¹.

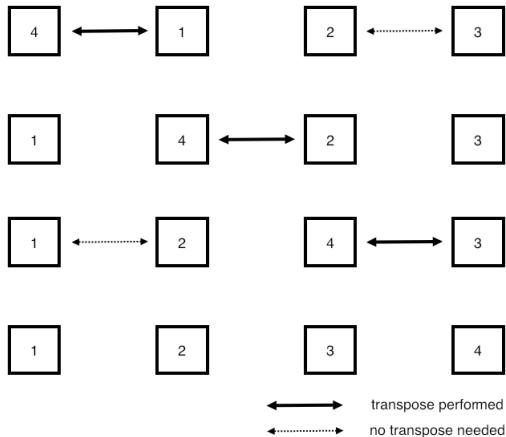


Figure 4.9: Odd-even transposition sort on 4 elements.

Exercise 4.13. A very simple sorting algorithm is *swap sort* or *odd-even transposition 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 4.9.

Implement this algorithm using `MPI_Sendrecv`. (Use `MPI_PROC_NULL` for the edge

1. There is an `MPI_Compare_and_swap` call. Do not use that.

MPI_PROC_NULL

C:
#include "mpi.h"
MPI_PROC_NULL

Fortran:
#include "mpif.h"
MPI_PROC_NULL

Python:
MPI.PROC_NULL = -1

How to read routine prototypes: 1.5.4.

manpage 31: Routine prototype for MPI·PROC·NULL

MPI_Sendrecv_replace

C:
int MPI_Sendrecv_replace(
 void *buf, int count, MPI_Datatype datatype,
 int dest, int sendtag, int source, int recvtag,
 MPI_Comm comm, MPI_Status *status)

Fortran:
MPI_SENDREREV_REPLACE (
 BUF, COUNT, DATATYPE,
 DEST, SENDTAG, SOURCE, RECVTAG,
 COMM, STATUS, IERROR)
<type> BUF(*)
INTEGER :: COUNT, DATATYPE, DEST, SENDTAG
INTEGER :: SOURCE, RECVTAG, COMM
INTEGER STATUS(MPI_STATUS_SIZE), IERROR

Input/output parameter:

buf : Initial address of send and receive buffer (choice).

Input parameters:

count : Number of elements in send and receive buffer (integer).

datatype : Type of elements to send and receive (handle).

dest : Rank of destination (integer).

sendtag : Send message tag (integer).

source : Rank of source (integer).

recvtag : Receive message tag (integer).

comm : Communicator (handle).

Output parameters:

status : Status object (status).

IERROR : Fortran only: Error status (integer).

How to read routine prototypes: 1.5.4.

manpage 32: Routine prototype for MPI·Sendrecv·replace

4. MPI topic: Point-to-point

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.

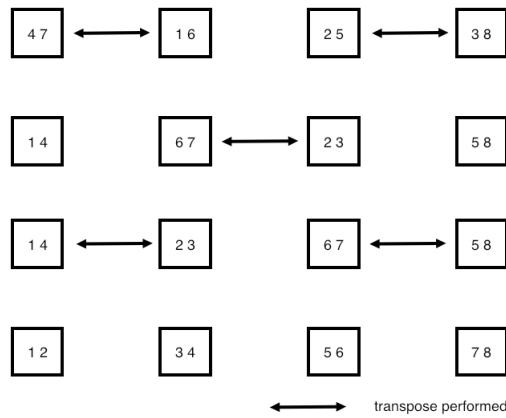


Figure 4.10: Odd-even transposition sort on 4 processes, holding 2 elements each.

Exercise 4.14. Extend this exercise to the case where each process hold an equal number of elements, more than 1. Consider figure 4.10 for inspiration. Is it coincidence that the algorithm takes the same number of steps as in the single scalar case?

4.2.5 Message status

In section 4.2.1 you saw that `MPI_Recv` has a ‘status’ argument of type `MPI_Status` that `MPI_Send` lacks. (The various `MPI_Wait`... routines also have a status argument; see section 4.3.1.) Often you specify `MPI_STATUS_IGNORE` for this argument: commonly you know what data is coming in and where it is coming from.

However, in some circumstances the recipient may not know all details of a message when you make the receive call, so MPI has a way of querying the status of the message:

- If you are expecting multiple incoming messages, it may be most efficient to deal with them in the order in which they arrive. So, instead of waiting for specific message, you would specify `MPI_ANY_SOURCE` or `MPI_ANY_TAG` in the description of the receive message. Now you have to be able to ask ‘who did this message come from, and what is in it’.
 - Maybe you know the sender of a message, but the amount of data is unknown. In that case you can overallocate your receive buffer, and after the message is received ask how big it was, or you can ‘probe’ an incoming message and allocate enough data when you find out how much data is being sent.

4.3 Non-blocking point-to-point operations

The structure of communication is often a reflection of the structure of the operation. With some regular applications we also get a regular communication pattern. Consider again the above operation:

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

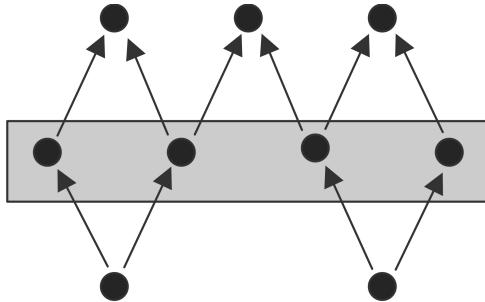


Figure 4.11: Processors with unbalanced send/receive patterns

Doing this in parallel induces communication, as pictured in figure 4.2.

We note:

- The data is one-dimensional, and we have a linear ordering of the processors.
- The operation involves neighbouring data points, and we communicate with neighbouring processors.

Above you saw how you can use information exchange between pairs of processors

- using `MPI_Send` and `MPI_Recv`, if you are careful; or
- using `MPI_Sendrecv`, as long as there is indeed some sort of pairing of processors.

However, there are circumstances where it is not possible, not efficient, or simply not convenient, to have such a deterministic setup of the send and receive calls. Figure 4.11 illustrates such a case, where processors are organized in a general graph pattern. Here, the numbers of sends and receive of a processor do not need to match.

In such cases, one wants a possibility to state ‘these are the expected incoming messages’, without having to wait for them in sequence. Likewise, one wants to declare the outgoing messages without having to do them in any particular sequence. Imposing any sequence on the sends and receives is likely to run into the serialization behaviour observed above, or at least be inefficient since processors will be waiting for messages.

4.3.1 Non-blocking send and receive calls

In the previous section you saw that blocking communication makes programming tricky if you want to avoid *deadlock* and performance problems. The main advantage of these routines is that you have full control about where the data is: if the send call returns the data has been successfully received, and the send buffer can be used for other purposes or de-allocated.

By contrast, the non-blocking calls `MPI_Isend` (figure 33) and `MPI_Irecv` (figure 34) do not wait for their counterpart: in effect they tell the runtime system ‘here is some data and please send it as follows’ or ‘here is some buffer space, and expect such-and-such data to come’. This is illustrated in figure 4.12. Issuing the `Isend`/`Irecv` call is sometimes referred to as *posting* a send/receive.

4. MPI topic: Point-to-point

MPI_Isend

C:
int MPI_Isend(void *buf,
 int count, MPI_Datatype datatype, int dest, int tag,
 MPI_Comm comm, MPI_Request *request)

Fortran:
the request parameter is an integer

Python:
request = MPI.Comm.Isend(self, buf, int dest, int tag=0)

How to read routine prototypes: [1.5.4](#).

manpage 33: Routine prototype for MPI_Isend

MPI_Irecv

C:
int MPI_Irecv(
 void* buf, int count, MPI_Datatype datatype,
 int source, int tag, MPI_Comm comm, MPI_Request *request)

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 request: request object (Request)

Fortran:
MPI_Irecv(buf, count, datatype, source, tag, comm, request, ierror)
TYPE(*), DIMENSION(..) :: buf
INTEGER, INTENT(IN) :: count, source, tag
TYPE(MPI_Datatype), INTENT(IN) :: datatype
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Request), INTENT(out) :: request
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Python native:
recvbuf = Comm.irecv(self, buf=None, int source=ANY_SOURCE, int tag=ANY_TAG,
 Request request=None)
Python numpy:
Comm.Irecv(self, buf, int source=ANY_SOURCE, int tag=ANY_TAG,
 Request status=None)

How to read routine prototypes: [1.5.4](#).

manpage 34: Routine prototype for MPI_Irecv

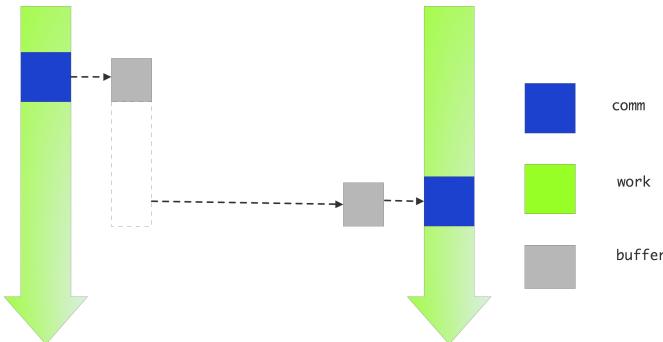


Figure 4.12: Non-blocking send

4.3.2 Request completion: wait calls

From the definition of `MPI_Isend` / `MPI_Irecv`, you seen that non-blocking routinemew yields an `MPI_Request` (figure 35) object. This request can then be used to query whether the operation has concluded. You may also notice that the `MPI_Irecv` routine does not yield an `MPI_Status` object. This makes sense: the status object describes the actually received data, and at the completion of the `MPI_Irecv` call there is no received data yet.

Waiting for the request is done with a number of routines. We first consider `MPI_Wait` (figure 36) . It takes the request as input, and gives an `MPI_Status` as output. If you don't need the status object, you can pass `MPI_STATUS_IGNORE`.

Note that the request is passed by reference, so that the wait routine can free it.

4.3.2.1 More wait calls

MPI has two types of routines for handling requests; we will start with the `MPI_Wait...` routines. These calls are blocking: when you issue such a call, your execution will wait until the specified requests have been completed. Typically you use `MPI_Waitall` to wait for all requests:

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

If you don't need the status objects, you can pass `MPI_STATUSES_IGNORE`.

Exercise 4.15. Revisit exercise 4.10 and consider replacing the blocking calls by non-blocking ones. How far apart can you put the `MPI_Isend` / `MPI_Irecv` calls and the corresponding `MPI_Waits`?

4.3.2.2 Receive status of the wait calls

The `MPI_Wait...` routines have the `MPI_Status` objects as output. If you are not interested in the status information, you can use the values `MPI_STATUS_IGNORE` for `MPI_Wait` and `MPI_Waitany`, or `MPI_STATUSES_IGNORE` for `MPI_Waitall` and `MPI_Waitsome`.

MPI_Request

```
C:  
MPI_Request request ;  
  
Fortran2008:  
Type(MPI_Request) :: request
```

How to read routine prototypes: [1.5.4](#).

[manpage 35: Routine prototype for MPI_Request](#)

MPI_Wait

```
Semantics:  
MPI_Wait( request, status)  
  
INOUT request: request object (handle)  
OUT status: status objects (handle)  
  
C:  
int MPI_Wait(  
    MPI_Request *requests,  
    MPI_Status *statuses)  
  
Fortran:  
MPI_Wait( request, status, ierror)  
TYPE(MPI_Request), INTENT(INOUT) :: requests  
TYPE(MPI_Status), INTENT(OUT) :: statuses  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python:  
MPI.Request.Wait(type cls, request, status=None)
```

Use `MPI_STATUS_IGNORE` to ignore

How to read routine prototypes: [1.5.4](#).

[manpage 36: Routine prototype for MPI_Wait](#)

Exercise 4.16. 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.)

There is a second motivation for the `Irecv` calls: if your hardware supports it, the communication can progress while your program can continue to do useful work:

```
// start non-blocking communication
MPI_Isend( ... ); MPI_Irecv( ... );
// do work that does not depend on incoming data
.....
// wait for the Isend/Irecv calls to finish
MPI_Wait( ... );
// now do the work that absolutely needs the incoming data
....
```

This is known as *overlapping computation and communication*, or *latency hiding*.

Unfortunately, a lot of this communication involves activity in user space, so the solution would have been to let it be handled by a separate thread. Until recently, processors were not efficient at doing such multi-threading, so true overlap stayed a promise for the future. Some network cards have support for this overlap, but it requires a non-trivial combination of hardware, firmware, and MPI implementation.

Exercise 4.17. Take your code of exercise 4.16 and modify it to use latency hiding.

Operations that can be performed without needing data from neighbours should be performed in between the `MPI_Isend` / `MPI_Irecv` calls and the corresponding `MPI_Wait` calls.

Remark 2 There is nothing special about a non-blocking or synchronous message. The `MPI_Recv` call can match any of the send routines you have seen so far (but not `MPI_Sendrecv`), and conversely a message sent with `MPI_Send` can be received by `MPI_Irecv`.

4.3.2.3 Buffer issues in non-blocking communication

While the use of non-blocking routines prevents deadlock, it introduces two new problems:

- When the send call returns, the actual send may not have been executed, so the send buffer may not be safe to overwrite. When the recv call returns, you do not know for sure that the expected data is in it. Thus, you need a mechanism to make sure that data was actually sent or received.
- With a blocking send call, you could repeatedly fill the send buffer and send it off.

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

To send multiple messages with non-blocking calls you have to allocate multiple buffers.

```

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

// irecvloop.c
MPI_Request requests =
(MPI_Request*) malloc( 2*nprocs*sizeof(MPI_Request) );
recv_buffers = (int*) malloc( nprocs*sizeof(int) );
send_buffers = (int*) malloc( nprocs*sizeof(int) );
for (int p=0; p<nprocs; p++) {
    int left_p = (p-1) % nprocs,
        right_p = (p+1) % nprocs;
    send_buffer[p] = nprocs-p;
    MPI_Isend(sendbuffer+p, 1, MPI_INT, right_p, 0, requests+2*p);
    MPI_Irecv(recvbuffer+p, 1, MPI_INT, left_p, 0, requests+2*p+1);
}
/* your useful code here */
MPI_Waitall(2*nprocs, requests, MPI_STATUSES_IGNORE);

```

4.3.3 More about non-blocking communication

4.3.3.1 Asynchronous progress

Above we saw that `I send`/`I recv` calls can overlap communication and computation. However, for this to happen we need for the MPI implementation to make *asynchronous progress*: the message needs to make its way through the network while the application is busy computing. However, communication of this sort can typically not be off-loaded to the network card, so different mechanisms are needed.

This can happen in a number of ways:

- Compute nodes may have a dedicated communications processor. The *Intel Paragon* was of this design; modern multicore processors are a more efficient realization of this idea.
- The MPI library may reserve a core or thread for communications processing. This is implementation dependent; for instance, *Intel MPI* has a number of `I_MPI_ASYNC_PROGRESS_...` variables.
- Absent such dedicated resources, the application can force MPI to make progress by occasional calls to a *polling* routine such as `MPI_Iprobe`.

A similar problem arises with passive target synchronization: it is possible that the origin process may hang until the target process makes an MPI call.

4.3.3.2 Wait and test calls

There are several wait calls: you can wait for a single request, all outstanding requests, or any of the outstanding requests.

`MPI_Wait` waits for a single request. If you are indeed waiting for a single nonblocking communication to complete, this is the right routine. If you are waiting for multiple requests you could call this routine in a loop.

```

|| for (p=0; p<nrequests ; p++) // Not efficient!
||   MPI_Wait(request[p],&(status[p]));

```

However, this would be inefficient if the first request is fulfilled much later than the others: your waiting process would have lots of idle time. In that case, use one of the following routines.

MPI_Waitall (figure 37) allows you to wait for a number of requests, and it does not matter in what sequence they are satisfied. Using this routine is easier to code than the loop above, and it could be more efficient.

The ‘waitall’ routine is good if you need all nonblocking communications to be finished before you can proceed with the rest of the program. However, sometimes it is possible to take action as each request is satisfied. In that case you could use **MPI_Waitany** (figure 38) and write:

```

|| for (p=0; p<nrequests; p++) {
||   MPI_Waitany(nrequests,request_array,&index,&status);
||   // operate on buffer[index]
|| }

```

Note that this routine takes a single status argument, passed by reference, and not an array of statuses!

Finally, **MPI_Waitsome** is very much like **MPI_Waitany**, except that it returns multiple numbers, if multiple requests are satisfied. Now the status argument is an array of **MPI_Status** objects.

Figure 4.13 shows the trace of a non-blocking execution using **MPI_Waitall**.

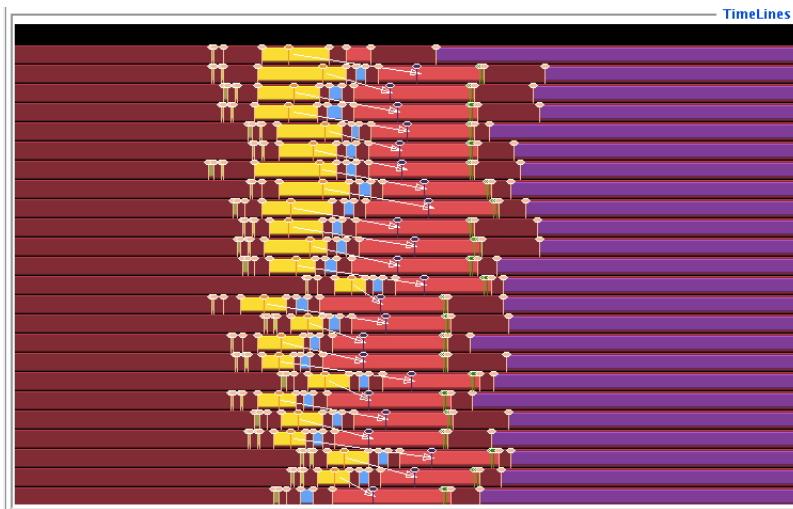


Figure 4.13: A trace of a nonblocking send between neighbouring processors

4.3.3.3 Implementing polling with Waitany

The **MPI_Waitany** routine can be used to implement *polling*: occasionally check for incoming messages while other work is going on.

4. MPI topic: Point-to-point

```

// irecv_source.c
if (procno==nprocs-1) {
    int *recv_buffer;
    MPI_Request *request; MPI_Status status;
    recv_buffer = (int*) malloc((nprocs-1)*sizeof(int));
    request = (MPI_Request*) malloc((nprocs-1)*sizeof(MPI_Request));

    for (int p=0; p<nprocs-1; p++) {
        ierr = MPI_Irecv(recv_buffer+p,1,MPI_INT, p,0,comm,
                           request+p); CHK(ierr);
    }
    for (int p=0; p<nprocs-1; p++) {
        int index, sender;
        MPI_Waitany(nprocs-1,request,&index,&status); //MPI_STATUS_IGNORE;
        if (index!=status.MPI_SOURCE)
            printf("Mismatch index %d vs source %d\n",index,status.MPI_SOURCE);
        printf("Message from %d: %d\n",index,recv_buffer[index]);
    }
} else {
    ierr = MPI_Send(&procno,1,MPI_INT, nprocs-1,0,comm); CHK(ierr);
}

## irecv_source.py
if procid==nprocs-1:
    receive_buffer = np.empty(nprocs-1,dtype=np.int)
    requests = [ None ] * (nprocs-1)
    for sender in range(nprocs-1):
        requests[sender] = comm.Irecv(receive_buffer[sender:sender+1],source=
sender)
    # alternatively: requests = [ comm.Irecv(s) for s in .... ]
    status = MPI.Status()
    for sender in range(nprocs-1):
        ind = MPI.Request.Waitany(requests,status=status)
        if ind!=status.Get_source():
            print("sender mismatch: %d vs %d" % (ind,status.Get_source()))
            print("received from",ind)
    else:
        mywait = random.randint(1,2*nprocs)
        print("[%d] wait for %d seconds" % (procid,mywait))
        time.sleep(mywait)
        mydata = np.empty(1,dtype=np.int)
        mydata[0] = procid
        comm.Send([mydata,MPI.INT],dest=nprocs-1)

```

Each process except for the root does a blocking send; the root posts **MPI_Irecv** from all other processors, then loops with **MPI_Waitany** until all requests have come in. Use **MPI_SOURCE** to test the index parameter of the wait call.

Note the **MPI_STATUS_IGNORE** parameter: we know everything about the incoming message, so we do not need to query a status object. Contrast this with the example in section 39.3.

Python note. In python creating the array for the returned requests is somewhat tricky.

```

## irecvloop.py
requests = [ None ] * (2*nprocs)
sendbuffer = np.empty( nprocs, dtype=np.int )
recvbuffer = np.empty( nprocs, dtype=np.int )

for p in range(nprocs):
    left_p = (p-1) % nprocs
    right_p = (p+1) % nprocs
    requests[2*p] = comm.Isend( sendbuffer[p:p+1], dest=
        left_p )
    requests[2*p+1] = comm.Irecv( recvbuffer[p:p+1], source=
        right_p )
MPI.Request.Waitall(requests)

```

Fortran note. The `index` parameter is the index in the array of requests, so it uses *1-based indexing*.

```

// irecv_source.F90
if (mytid==ntids-1) then
  do p=1,ntids-1
    print *, "post"
    call MPI_Irecv(recv_buffer(p),1,MPI_INTEGER,p-1,0,comm
      ,&
      requests(p),err)
  end do
  do p=1,ntids-1
    call MPI_Waitany(ntids-1,requests,index,
      MPI_STATUS_IGNORE,err)
    write(*,'("Message from",i3,":",i5)' ) index,
    recv_buffer(index)
  end do

```

4.3.3.4 Test: non-blocking request wait

The `MPI_Wait...` routines are blocking. Thus, they are a good solution if the receiving process can not do anything until the data (or at least some data) is actually received. The `MPI_Test...` calls are themselves non-blocking: they test for whether one or more requests have been fulfilled, but otherwise immediately return. This can be used in the *manager-worker model*: the manager process creates tasks, and sends them to whichever worker process has finished its work. (This uses a receive from `MPI_ANY_SOURCE`, and a subsequent test on the `MPI_SOURCE` field of the receive status.) While waiting for the workers, the manager can do useful work too, which requires a periodic check on incoming message.

Pseudo-code:

```

while ( not done ) {
  // create new inputs for a while
  ....
  // see if anyone has finished
  MPI_Test( .... &index, &flag );
  if ( flag ) {
    // receive processed data and send new
  }
}

```

`MPI_Test` (figure 39) `MPI_Testany` (figure 40) `MPI_Testall` (figure 41)

Exercise 4.18. Read section HPSC-?? and give pseudo-code for the distributed sparse matrix-vector product using the above idiom for using `MPI_Test...` calls. Discuss the advantages and disadvantages of this approach. The answer is not going to be black and white: discuss when you expect which approach to be preferable.

4.3.3.5 More about requests

Every non-blocking call allocates an `MPI_Request` object. Unlike `MPI_Status`, an `MPI_Request` variable is not actually an object, but instead it is an (opaque) pointer. This means that when you call, for instance, `MPI_Irecv`, MPI will allocate an actual request object, and return its address in the `MPI_Request` variable.

Correspondingly, calls to `MPI_Wait...` or `MPI_Test` free this object, setting the handle to `MPI_REQUEST_NULL`. Thus, it is wise to issue wait calls even if you know that the operation has succeeded. For instance, if all receive calls are concluded, you know that the corresponding send calls are finished and there is no strict need to wait for their requests. However, omitting the wait calls would lead to a memory leak.

Another way around this is to call `MPI_Request_free`,

```
// int MPI_Request_free(MPI_Request *request)
```

which sets the request variable to `MPI_REQUEST_NULL`, and marks the object for deallocation after completion of the operation.

You can inspect the status of a request without freeing the request object with `MPI_Request_get_status`:

```
int MPI_Request_get_status(
    MPI_Request request,
    int *flag,
    MPI_Status *status
);
```

4.4 More about point-to-point communication

4.4.1 Message probing

MPI receive calls specify a receive buffer, and its size has to be enough for any data sent. In case you really have no idea how much data is being sent, and you don't want to overallocate the receive buffer, you can use a 'probe' call.

The routine `MPI_Probe` (figure 42) (and `MPI_Iprobe`, for which see section 4.3.3.1), accepts a message, but does not copy the data. Instead, when probing tells you that there is a message, you can use `MPI_Get_count` to determine its size, allocate a large enough receive buffer, and do a regular receive to have the data copied.

```
// probe.c
if (procno==receiver) {
    MPI_Status status;
    MPI_Probe(sender,0,comm,&status);
```

```

    int count;
    MPI_Get_count (&status, MPI_FLOAT, &count);
    float recv_buffer[count];
    MPI_Recv(recv_buffer, count, MPI_FLOAT, sender, 0, comm, MPI_STATUS_IGNORE);
} else if (procno==sender) {
    float buffer[buffer_size];
    ierr = MPI_Send(buffer, buffer_size, MPI_FLOAT, receiver, 0, comm); CHK(ierr);
}

```

There is a problem with the `MPI_Probe` call in a multithreaded environment: the following scenario can happen.

1. A thread determines by probing that a certain message has come in.
2. It issues a blocking receive call for that message...
3. But in between the probe and the receive call another thread has already received the message.
4. ... Leaving the first thread in a blocked state with no message to receive.

This is solved by `MPI_Mprobe` (figure 43), which after a successful probe removes the message from the *matching queue*: the list of messages that can be matched by a receive call. The thread that matched the probe now issues an `MPI_Mrecv` (figure 44) call on that message through an object of type `MPI_Message`.

4.4.2 The Status object and wildcards

With some receive calls you know everything about the message in advance: its source, tag, and size. In other cases you want to leave some options open, and inspect the message for them after it was received. To do this, the receive call has a `MPI_Status` (figure 45) parameter.

This status is a property of the actually received message, so `MPI_Irecv` does not have a status parameter, but `MPI_Wait` does.

The `MPI_Status` object is a structure with the following freely accessible members:

4.4.2.1 Source

In some applications it makes sense that a message can come from one of a number of processes. In this case, it is possible to specify `MPI_ANY_SOURCE` as the source. To find out the source where the message actually came from, you would use the `MPI_SOURCE` field of the status object that is delivered by `MPI_Recv` or the `MPI_Wait...` call after an `MPI_Irecv`.

```

MPI_Recv(recv_buffer+p, 1, MPI_INT, MPI_ANY_SOURCE, 0, comm,
         &status);
sender = status.MPI_SOURCE;

```

The source of a message can be obtained as the `MPI_SOURCE` (figure 46) member of the status structure.

There are various scenarios where receiving from ‘any source’ makes sense. One is that of the *master-worker model*. The master task would first send data to the worker tasks, then issues a blocking wait for the data of whichever process finishes first.

4. MPI topic: Point-to-point

4.4.2.2 Tag

If a processor is expecting more than one message from a single other processor, message tags are used to distinguish between them. In that case, a value of `MPI_ANY_TAG` can be used, and the actual tag of a message can be retrieved as the `MPI_TAG` (figure 47) member in the status structure.

4.4.2.3 Error

Any *errors* during the receive operation can be found as the `MPI_ERROR` (figure 48) member of the status structure.

4.4.2.4 Count

If the amount of data received is not known a priori, the *count* of elements received can be found by `MPI_Get_count` (figure 49) :

```
// MPI_Get_count(&recv_status, MPI_INT, &recv_count);
```

This may be necessary since the *count* argument to `MPI_Recv` is the buffer size, not an indication of the actually expected number of data items.

Note that unlike the above this is not directly a member of the status structure.

4.4.2.5 Example: receiving from any source

Using the `MPI_ANY_SOURCE` specifier. We retrieve the actual source from the `MPI_Status` object through the `MPI_SOURCE` field.

```
// anysource.c
if (procno==nprocs-1) {
    int *recv_buffer;
    MPI_Status status;

    recv_buffer = (int*) malloc((nprocs-1)*sizeof(int));

    for (int p=0; p<nprocs-1; p++) {
        err = MPI_Recv(recv_buffer+p, 1, MPI_INT, MPI_ANY_SOURCE, 0, comm,
                       &status); CHK(err);
        int sender = status.MPI_SOURCE;
        printf("Message from sender=%d: %d\n",
               sender, recv_buffer[p]);
    }
} else {
    float randomfraction = (rand() / (double)RAND_MAX);
    int randomwait = (int) ( nprocs * randomfraction );
    printf("process %d waits for %e/%d=%d\n",
           procno, randomfraction, nprocs, randomwait);
    sleep(randomwait);
    err = MPI_Send(&randomwait, 1, MPI_INT, nprocs-1, 0, comm); CHK(err);
}
```

```
## anysource.py
rstatus = MPI.Status()
comm.Recv(rbuf, source=MPI.ANY_SOURCE, status=rstatus)
print("Message came from %d" % rstatus.Get_source())
```

In sections 39.3 and 4.3.3.4 we explained the manager-worker model, and how it offers an opportunity for inspecting the `MPI_SOURCE` field of the `MPI_Status` object describing the data that was received.

4.4.3 Synchronous and asynchronous communication

It is easiest to think of blocking as a form of synchronization with the other process, but that is not quite true. Synchronization is a concept in itself, and we talk about *synchronous* communication if there is actual coordination going on with the other process, and *asynchronous* communication if there is not. Blocking then only refers to the program waiting until the user data is safe to reuse; in the synchronous case a blocking call means that the data is indeed transferred, in the asynchronous case it only means that the data has been transferred to some system buffer. The four possible cases are illustrated in figure 4.14.

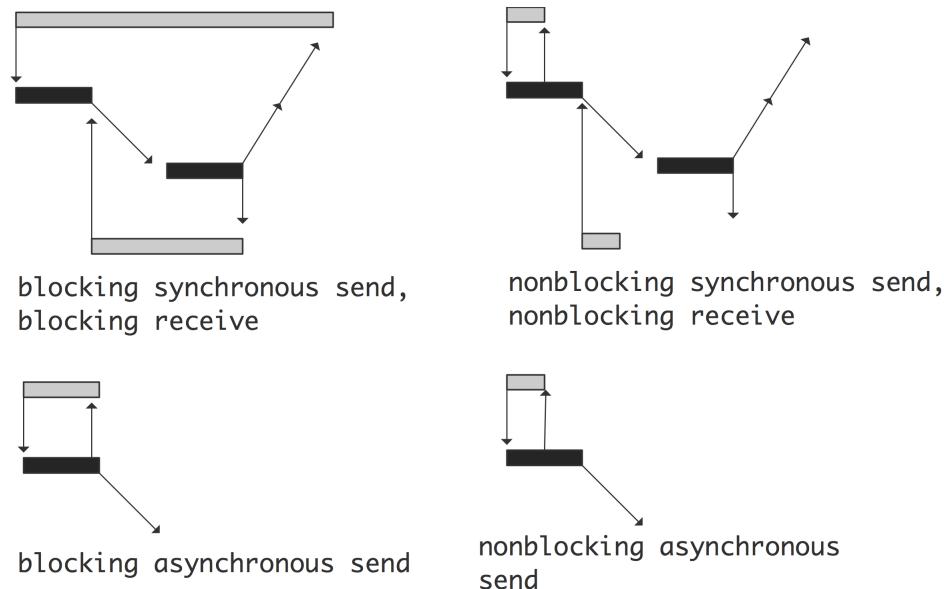


Figure 4.14: Blocking and synchronicity

MPI has a number of routines for synchronous communication, such as `MPI_Ssend`.

```
// ssendblock.c
other = 1-procno;
sendbuf = (int*) malloc(sizeof(int));
recvbuf = (int*) malloc(sizeof(int));
size = 1;
MPI_Ssend(sendbuf, size, MPI_INT, other, 0, comm);
MPI_Recv(recvbuf, size, MPI_INT, other, 0, comm, &status);
printf("This statement is not reached\n");
```

4.4.4 Persistent communication

An `MPI_Isend` or `MPI_Irecv` call has an `MPI_Request` parameter. This is an object that gets created in the send/recv call, and deleted in the wait call. You can imagine that this carries some overhead, and if the same communication is repeated many times you may want to avoid this overhead by reusing the request object.

To do this, MPI has *persistent communication*:

- You describe the communication with `MPI_Send_init`, which has the same calling sequence as `MPI_Isend`, or `MPI_Recv_init`, which has the same calling sequence as `MPI_Irecv`.
- The actual communication is performed by calling `MPI_Start`, for a single request, or `MPI_Startall` for an array or requests.
- Completion of the communication is confirmed with `MPI_Wait` or similar routines as you have seen in the explanation of non-blocking communication.
- The wait call does not release the request object: that is done with `MPI_Request_free`.

The calls `MPI_Send_init` (figure 50) and `MPI_Recv_init` (figure 51) for creating a persistent communication have the same syntax as those for non-blocking sends and receives. The difference is that they do not start an actual communication, they only create the request object.

Given these request object, a communication (both send and receive) is then started with `MPI_Start` (figure 52) for a single request or `MPI_Startall` (figure 53) for multiple requests, given in an array.

These are equivalent to starting an `MPI_Isend` or `MPI_Irecv`; correspondingly, it is necessary to issue an `MPI_Wait...` call (section 4.3.1) to determine their completion.

After a request object has been used, possibly multiple times, it can be freed; see 4.3.3.5.

In the following example a ping-pong is implemented with persistent communication.

```
// persist.c
if (procno==src) {
    MPI_Send_init(send,s,MPI_DOUBLE,tgt,0,comm,requests+0);
    MPI_Recv_init(recv,s,MPI_DOUBLE,tgt,0,comm,requests+1);
    printf("Size %d\n",s);
    t[cnt] = MPI_Wtime();
    for (int n=0; n<NEXPERIMENTS; n++) {
        MPI_Startall(2,requests);
        MPI_Waitall(2,requests,MPI_STATUSES_IGNORE);
    }
    t[cnt] = MPI_Wtime()-t[cnt];
    MPI_Request_free(requests+0); MPI_Request_free(requests+1);
} else if (procno==tgt) {
    for (int n=0; n<NEXPERIMENTS; n++) {
        MPI_Recv(recv,s,MPI_DOUBLE,src,0,comm,MPI_STATUS_IGNORE);
        MPI_Send(recv,s,MPI_DOUBLE,src,0,comm);
    }
}

## persist.py
sendbuf = np.ones(size,dtype=np.int)
recvbuf = np.ones(size,dtype=np.int)
if procid==src:
```

```

print("Size:",size)
times[isize] = MPI.Wtime()
for n in range(nexperiments):
    requests[0] = comm.Isend(sendbuf[0:size], dest=tgt)
    requests[1] = comm.Irecv(recvbuf[0:size], source=tgt)
    MPI.Request.Waitall(requests)
    sendbuf[0] = sendbuf[0]+1
    times[isize] = MPI.Wtime()-times[isize]
elif procid==tgt:
    for n in range(nexperiments):
        comm.Recv(recvbuf[0:size], source=src)
        comm.Send(recvbuf[0:size], dest=src)

```

As with ordinary send commands, there are the variants `MPI_Bsend_init`, `MPI_Ssend_init`, `MPI_Rsend_init`

4.4.5 Buffered communication

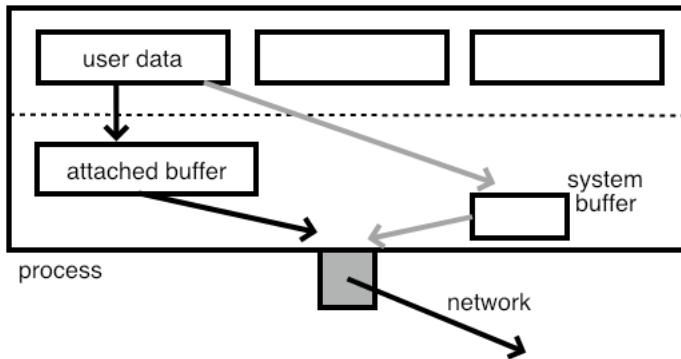


Figure 4.15: User communication routed through an attached buffer

By now you have probably got the notion that managing buffer space in MPI is important: data has to be somewhere, either in user-allocated arrays or in system buffers. Using *buffered communication* is yet another way of managing buffer space.

1. You allocate your own buffer space, and you attach it to your process. This buffer is not a send buffer: it is a replacement for buffer space used inside the MPI library or on the network card; figure 4.15. If high-bandwidth memory is available, you could create your buffer there.
2. You use the `MPI_Bsend` call for sending, using otherwise normal send and receive buffers;
3. You detach the buffer when you're done with the buffered sends.

4.4.5.1 Bufferend send calls

Section 4.3 discusses non-blocking communication, where multiple sends will be under way. Since these sends are under way with possibly no receive having been posted, the send buffers can not be reused. It would be possible to reuse the send buffers if MPI had enough internal buffer space. For this, there is the

buffered send mode, where you first give MPI internal buffer space; subsequently only a single send buffer is needed. The `MPI_Bsend` (figure 54) call is non-blocking.

There can be only one buffer per process, attached with `MPI_Buffer_attach` (figure 55). Its size should be enough for all outstanding `MPI_Bsend` calls that are simultaneously outstanding, plus `MPI_BSEND_OVERHEAD`. You can compute the needed size of the buffer with `MPI_Pack_size`; see section 5.4.3.

The possible error codes are

- `MPI_SUCCESS` the routine completed successfully.
- `MPI_ERR_BUFFER` The buffer pointer is invalid; this typically means that you have supplied a null pointer.
- `MPI_ERR_INTERN` An internal error in MPI has been detected.

The buffer is detached with `MPI_Buffer_detach`:

```
|| int MPI_Buffer_detach(  
||   void *buffer, int *size);
```

This returns the address and size of the buffer; the call blocks until all buffered messages have been delivered.

You can force delivery by

```
|| MPI_Buffer_detach( &b, &n );  
|| MPI_Buffer_attach( b, n );
```

The asynchronous version is `MPI_Ibsend`, the persistent (see section 4.4.4) call is `MPI_Bsend_init`.

4.4.5.2 Persistent buffered communication

There is a persistent variant `MPI_Bsend_init` (figure 56) of buffered sends, as with regular sends (section 4.4.4).

4.5 Sources used in this chapter

Listing of code examples/mpi/c/recvblock.c:

Listing of code examples/mpi/c/sendblock.c:

Listing of code XX:

Listing of code examples/mpi/c/ssendblock.c:

Listing of code examples/mpi/c/irecvloop.c:

Listing of code examples/mpi/c/waitforany.c:

Listing of code XX:

Listing of code examples/mpi/c/probe.c:

Listing of code examples/mpi/c/anysource.c:

Listing of code examples/mpi/c/ssendblock.c:

Listing of code examples/mpi/c/persist.c:

MPI_Waitall

Semantics:

```
MPI_WAITALL( count, array_of_requests, array_of_statuses)
IN count: lists length (non-negative integer)
INOUT array_of_requests: array of requests (array of handles)
OUT array_of_statuses: array of status objects (array of Status)
```

C:

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

Fortran:

```
MPI_Waitall(count, array_of_requests, array_of_statuses, ierror)
INTEGER, INTENT(IN) :: count
TYPE(MPI_Request), INTENT(INOUT) :: array_of_requests(count)
TYPE(MPI_Status) :: array_of_statuses(*)
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
MPI.Request.Waitall(type cls, requests, statuses=None)
```

Use MPI_STATUSES_IGNORE to ignore

How to read routine prototypes: [1.5.4.](#)

manpage 37: Routine prototype for MPI_Waitall

MPI_Waitany

Semantics:

```
int MPI_Waitany(
    int count, MPI_Request array_of_requests[], int *index,
    MPI_Status *status)
```

IN count: list length (non-negative integer)
INOUT array_of_requests: array of requests (array of handles)
OUT index: index of handle for operation that completed (integer)
OUT status: status object (Status)

C:
`MPI_Waitany(count, array_of_requests, index, status, ierror)`

Fortran:
`INTEGER, INTENT(IN) :: count
TYPE(MPI_Request), INTENT(INOUT) :: array_of_requests(count)
INTEGER, INTENT(OUT) :: index
TYPE(MPI_Status) :: status
INTEGER, OPTIONAL, INTENT(OUT) :: ierror`

Python:
`MPI.Request.Waitany(requests,status=None)
class method, returns index`

How to read routine prototypes: 1.5.4.

manpage 38: Routine prototype for MPI_Waitany

MPI_Test

C:
`int MPI_Test(MPI_Request *request, int *flag, MPI_Status *status)`

Input Parameters
`request : MPI request (handle)`

Output Parameters
`flag : true if operation completed (logical)
status : status object (Status). May be MPI_STATUS_IGNORE.`

Python:
`request.Test()`

How to read routine prototypes: 1.5.4.

manpage 39: Routine prototype for MPI_Test

MPI_Testany

C:
int MPI_Testany(
 int count, MPI_Request array_of_requests[],
 int *index, int *flag, MPI_Status *status)

Fortran:

```
MPI_Testany(count, array_of_requests, index, flag, status, ierror)
INTEGER, INTENT(IN) :: count
TYPE(MPI_Request), INTENT(INOUT) :: array_of_requests(count)
INTEGER, INTENT(OUT) :: index
LOGICAL, INTENT(OUT) :: flag
TYPE(MPI_Status) :: status
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 40: Routine prototype for MPI_Testany

MPI_Testall

Semantics:
MPI_TESTALL(count, array_of_requests, flag, array_of_statuses)
IN countlists length (non-negative integer)
INOUT array_of_requestsarray of requests (array of handles)
OUT flag(logical)
OUT array_of_statusesarray of status objects (array of Status)

C:

```
int MPI_Testall(
    int count, MPI_Request array_of_requests[],
    int *flag, MPI_Status array_of_statuses[])
```

Fortran:

```
MPI_Testall(count, array_of_requests, flag, array_of_statuses, ierror)
INTEGER, INTENT(IN) :: count
TYPE(MPI_Request), INTENT(INOUT) :: array_of_requests(count)
LOGICAL, INTENT(OUT) :: flag
TYPE(MPI_Status) :: array_of_statuses(*)
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 41: Routine prototype for MPI_Testall

MPI_Probe

```
int MPI_Probe
  ( int source, int tag, MPI_Comm comm,
    MPI_Status *status )
int MPI_Iprobe
  (int source, int tag, MPI_Comm comm, int *flag,
   MPI_Status *status)

Input parameters:
source : source rank, or MPI_ANY_SOURCE (integer)
tag    : tag value or MPI_ANY_TAG (integer)
comm   : communicator (handle)

Output parameter:
flag   : True if a message with the specified
         source, tag, and communicator is available
status : message status
```

How to read routine prototypes: 1.5.4.

manpage 42: Routine prototype for MPI_Probe

MPI_Mprobe

```
int MPI_Mprobe(int source, int tag, MPI_Comm comm,
               MPI_Message *message, MPI_Status *status)

Input Parameters:
source - rank of source or MPI_ANY_SOURCE (integer)
tag    - message tag or MPI_ANY_TAG (integer)
comm   - communicator (handle)

Output Parameters:
message - returned message (handle)
status  - status object (status)
```

How to read routine prototypes: 1.5.4.

manpage 43: Routine prototype for MPI_Mprobe

MPI_Mrecv

```
int MPI_Mrecv(void *buf, int count, MPI_Datatype type,
              MPI_Message *message, MPI_Status *status)

Input Parameters:
count      - Number of elements to receive (nonnegative integer).
datatype  - Datatype of each send buffer element (handle).
message   - Message (handle).

Output Parameters:
buf       - Initial address of receive buffer (choice).
status    - Status object (status).
IERROR    - Fortran only: Error status (integer).

MPI_MRECV(BUF, COUNT, DATATYPE, MESSAGE, STATUS, IERROR)
           <type>   BUF(*)
INTEGER     COUNT, DATATYPE, MESSAGE
INTEGER     STATUS(MPI_STATUS_SIZE), IERROR
```

How to read routine prototypes: [1.5.4](#).

manpage 44: Routine prototype for MPI_Mrecv

MPI_Status

```
C:
MPI_Status status;

Fortran:
integer :: status(MPI_STATUS_SIZE)

Fortran2008:
type(MPI_Status) :: recv_status

Python:
MPI.Status() # returns object
```

How to read routine prototypes: [1.5.4](#).

manpage 45: Routine prototype for MPI_Status

MPI_SOURCE

Semantics:
MPI_SOURCE is the name of an integer field
in an MPI_STATUS structure

C:
status.MPI_SOURCE // is int

F:
status_object%MPI_SOURCE ! is integer

Python:
status.Get_source() # returns int

How to read routine prototypes: 1.5.4.

manpage 46: Routine prototype for MPI_SOURCE

MPI_TAG

C:
int status.MPI_TAG;

F:
integer :: MPI_TAG

Python:
status.Get_tag() # returns int

How to read routine prototypes: 1.5.4.

manpage 47: Routine prototype for MPI_TAG

MPI_ERROR

C:
int status.MPI_ERROR;

F:
integer :: MPI_ERROR

Python:
status.Get_error() # returns int

How to read routine prototypes: 1.5.4.

manpage 48: Routine prototype for MPI_ERROR

MPI_Get_count

```
// C:  
int MPI_Get_count(MPI_Status *status, MPI_Datatype datatype,  
                  int *count)  
  
! Fortran:  
MPI_Get_count( INTEGER status(MPI_STATUS_SIZE), INTEGER datatype,  
                INTEGER count, INTEGER ierror)  
  
Python:  
status.Get_count( Datatype datatype=BYTE )
```

How to read routine prototypes: 1.5.4.

manpage 49: Routine prototype for MPI·Get·count

MPI_Send_init

```
C:  
int MPI_Send_init(  
    const void* buf, int count, MPI_Datatype datatype,  
    int dest, int tag, MPI_Comm comm, MPI_Request *request)  
  
Fortran:  
MPI_Send_init(buf, count, datatype, dest, tag, comm, request, ierror)  
TYPE(*), DIMENSION(..), INTENT(IN), ASYNCHRONOUS :: buf  
INTEGER, INTENT(IN) :: count, dest, tag  
TYPE(MPI_Datatype), INTENT(IN) :: datatype  
TYPE(MPI_Comm), INTENT(IN) :: comm  
TYPE(MPI_Request), INTENT(OUT) :: request  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python:  
MPI.Comm.Send_init(self, buf, int dest, int tag=0)
```

Semantics:

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

How to read routine prototypes: 1.5.4.

manpage 50: Routine prototype for MPI·Send·init

MPI_Recv_init

```
C:
int MPI_Recv_init(
    void* buf, int count, MPI_Datatype datatype,
    int source, int tag, MPI_Comm comm, MPI_Request *request)

Fortran:
MPI_Recv_init(buf, count, datatype, source, tag, comm, request,
ierror)
TYPE(*), DIMENSION(..), ASYNCHRONOUS :: buf
INTEGER, INTENT(IN) :: count, source, tag
TYPE(MPI_Datatype), INTENT(IN) :: datatype
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Request), INTENT(OUT) :: request
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Python:
MPI.Comm.Recv_init(
    self, buf, int source=ANY_SOURCE, int tag=ANY_TAG)

Semantics:
OUT buf: initial address of receive buffer (choice)
IN count: number of elements received (non-negative integer)
IN datatype: type of each element (handle)
IN source: rank of source or MPI_ANY_SOURCE (integer)
IN tag: message tag or MPI_ANY_TAG (integer)
IN com: mcommunicator (handle)
OUT request: communication request (handle)
```

How to read routine prototypes: 1.5.4.

manpage 51: Routine prototype for MPI_Recv_init

MPI_Start

```
C:
int MPI_Start(MPI_Request request)

Fortran:
MPI_Start(request, ierror)
TYPE(MPI_Request), INTENT(INOUT) :: request
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

MPI_START(REQUEST, IERROR)
INTEGER REQUESTS, IERROR
```

```
Python:
MPI.Prequest.Start(type cls, request)
```

```
Semantics:
INOUT request : request (handle)
```

How to read routine prototypes: 1.5.4.

manpage 52: Routine prototype for MPI_Start

MPI_Startall

C:
int MPI_Startall(int count, MPI_Request array_of_requests[])

Fortran:
MPI_Startall(count, array_of_requests, ierror)
INTEGER, INTENT(IN) :: count
TYPE(MPI_Request), INTENT(INOUT) :: array_of_requests(count)
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
MPI_STARTALL(COUNT, ARRAY_OF_REQUESTS, IERROR)
INTEGER COUNT, ARRAY_OF_REQUESTS(*), IERROR

Python:
MPI.Prequest.Startall(type cls, requests)

Semantics:
IN countlist length (non-negative integer)
INOUT array_of_requestsarray of requests (array of handle)

How to read routine prototypes: 1.5.4.

manpage 53: Routine prototype for MPI_Startall

MPI_Bsend

C:
int MPI_Bsend
(const void *buf, int count, MPI_Datatype datatype,
 int dest, int tag,MPI_Comm comm)

Input Parameters
buf : initial address of send buffer (choice)
count : number of elements in send buffer (nonnegative integer)
datatype : datatype of each send buffer element (handle)
dest : rank of destination (integer)
tag : message tag (integer)
comm : communicator (handle)

How to read routine prototypes: 1.5.4.

manpage 54: Routine prototype for MPI_Bsend

MPI_Buffer_attach

int MPI_Buffer_attach(void *buffer, int size);

Input arguments:
buffer : initial buffer address (choice)
size : buffer size, in bytes (integer)

How to read routine prototypes: 1.5.4.

manpage 55: Routine prototype for MPI_Buffer_attach

MPI_Bsend_init

Synopsis

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

Input Parameters

buf : initial address of send buffer (choice)
count : number of elements sent (integer)
datatype : type of each element (handle)
dest : rank of destination (integer)
tag : message tag (integer)
comm : communicator (handle)

Output Parameters

request : communication request (handle)

How to read routine prototypes: [1.5.4.](#)

manpage 56: Routine prototype for MPI_Bsend_init

Chapter 5

MPI topic: Data types

In the examples you have seen so far, every time data was sent, it was as a contiguous buffer with elements of a single type. In practice you may want to send heterogeneous data, or non-contiguous data.

- Communicating the real parts of an array of complex numbers means specifying every other number.
- Communicating a C structure of Fortran type with more than one type of element is not equivalent to sending an array of elements of a single type.

The datatypes you have dealt with so far are known as *elementary datatypes*; irregular objects are known as *derived datatypes*.

5.1 Elementary data types

MPI has a number of elementary data types, corresponding to the simple data types of programming languages. The names are made to resemble the types of C and Fortran, for instance `MPI_FLOAT` and `MPI_DOUBLE` versus `MPI_REAL` and `MPI_DOUBLE_PRECISION`.

MPI calls accept arrays of elements:

```
|| double x[20];
|| MPI_Send( x, 20, MPI_DOUBLE, . . . . . )
```

so for a single element you need to take its address:

```
|| double x;
|| MPI_Send( &x, 1, MPI_DOUBLE, . . . . . )
```

5.1.1 C/C++

MPI_CHAR	only for text data, do not use for small integers
MPI_UNSIGNED_CHAR	
MPI_SIGNED_CHAR	
MPI_SHORT	
MPI_UNSIGNED_SHORT	
MPI_INT	
MPI_UNSIGNED	
MPI_LONG	
MPI_UNSIGNED_LONG	
MPI_LONG_LONG_INT	
MPI_FLOAT	
MPI_DOUBLE	
MPI_LONG_DOUBLE	
MPI_BYTE	
MPI_PACKED	

There is some, but not complete, support for *C99* types.

5.1.2 Fortran

MPI_CHARACTER	Character(Len=1)
MPI_INTEGER	
MPI_INTEGER1	
MPI_INTEGER2	
MPI_INTEGER4	
MPI_INTEGER8	
MPI_INTEGER16	
MPI_REAL	
MPI_DOUBLE_PRECISION	
MPI_REAL2	
MPI_REAL4	
MPI_REAL8	
MPI_COMPLEX	
MPI_DOUBLE_COMPLEX	Complex(Kind=Kind(0.d0))
MPI_LOGICAL	
MPI_PACKED	

Not all these types need be supported, for instance MPI_INTEGER16 may not exist, in which case it will be equivalent to MPI_DATATYPE_NULL.

The default integer type MPI_INTEGER is equivalent to INTEGER(KIND=MPI_INTEGER_KIND).

Addresses have type MPI_Aint or INTEGER(KIND=MPI_ADDRESS_KIND) in Fortran. The start of the address range is given in MPI_BOTTOM.

There is also MPI_COUNT_KIND, MPI_OFFSET_KIND.

5. MPI topic: Data types

5.1.2.1 Fortran90 kind-defined types

If your Fortran code uses `KIND` to define scalar types with specified precision, these do not in general correspond to any predefined MPI datatypes. Hence the following routines exist to make *MPI equivalences of Fortran scalar types*: `MPI_Type_create_f90_integer` (figure 57) `MPI_Type_create_f90_real` (figure 58) `MPI_Type_create_f90_complex` (figure 59).

Examples:

```
|| INTEGER ( KIND = SELECTED_INTEGER_KIND(15) ) , &
||   DIMENSION(100) :: array INTEGER :: root , integertype , error

|| CALL MPI_Type_create_f90_integer( 15 , integertype , error )
|| CALL MPI_Bcast ( array , 100 ,
||   & integertype , root ,
||   & MPI_COMM_WORLD , error )

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

|| COMPLEX ( KIND = SELECTED_REAL_KIND(15 ,300) ) , &
||   DIMENSION(100) :: array
|| CALL MPI_Type_create_f90_complex( 15 , 300 , complextyp , error )
```

5.1.3 Python

mpi4py type	NumPy type
<code>MPI.INT</code>	<code>np.intc</code>
<code>MPI.LONG</code>	<code>np.int</code>
<code>MPI.FLOAT</code>	<code>np.float32</code>
<code>MPI.DOUBLE</code>	<code>np.float64</code>

5.1.4 Byte addressing type

So far we have mostly been taking about datatypes in the context of sending them. The `MPI_Aint` type is not so much for sending, as it is for describing the size of objects, such as the size of an `MPI_Win` object; section 8.1.

See also the `MPI_Sizeof` and `MPI_Get_address` routines.

5.1.4.1 Fortran

The equivalent of `MPI_Aint` in Fortran is an integer of kind `MPI_ADDRESS_KIND`:

```
|| integer(kind=MPI_ADDRESS_KIND) :: winsize
```

Fortran lacks a `sizeof` operator to query the sizes of datatypes. Since sometimes exact byte counts are necessary, for instance in one-sided communication, Fortran can use the `MPI_Sizeof` (figure 60) routine.

Example usage in `MPI_Win_create`:

MPI_Type_create_f90_integer

C:

```
int MPI_Type_create_f90_integer(int r, MPI_Datatype *newtype);
```

Fortran:

```
MPI_TYPE_CREATE_F90_INTEGER( INTEGER R, INTEGER NEWTYPE, INTEGER IERROR)
```

Input Parameter

r : Precision, in decimal digits (integer).

Output Parameters

newtype : New data type (handle).

IERROR : Fortran only: Error status (integer).

How to read routine prototypes: 1.5.4.

manpage 57: Routine prototype for MPI_Type_create_f90_integer

MPI_Type_create_f90_real

C:

```
int MPI_Type_create_f90_real(int p, int r, MPI_Datatype *newtype)
```

Fortran:

```
MPI_TYPE_CREATE_F90_REAL (P, R, NEWTYPE, IERROR)
```

Input Parameters

p : Precision, in decimal digits (integer).

r : Decimal exponent range (integer).

Output Parameters

newtype : New data type (handle).

IERROR : Fortran only: Error status (integer).

Either p or r, but not both, may be omitted from calls to SELECTED_REAL_KIND. Similarly, either argument to MPI_Type_create_f90_real may be set to MPI_UNDEFINED.

How to read routine prototypes: 1.5.4.

manpage 58: Routine prototype for MPI_Type_create_f90_real

MPI_Type_create_f90_complex

C:
int MPI_Type_create_f90_real(int p, int r, MPI_Datatype *newtype)

Fortran:
MPI_TYPE_CREATE_F90_REAL (P, R, NEWTYPE, IERROR)

Input Parameters
p : Precision, in decimal digits (integer).
r : Decimal exponent range (integer).

Output Parameters
newtype : New data type (handle).
IERROR : Fortran only: Error status (integer).

Either p or r, but not both, may be omitted from calls to
SELECTED_REAL_KIND. Similarly, either argument to
MPI_Type_create_f90_complex may be set to MPI_UNDEFINED.

How to read routine prototypes: [1.5.4](#).

manpage 59: Routine prototype for MPI_Type_create_f90_complex

MPI_Sizeof

Synopsis:

MPI_Sizeof(v,size) – Returns the size, in bytes, of the given type

Fortran:
MPI_SIZEOF(V, SIZE, IERROR)
<type> V
INTEGER SIZE, IERROR

Input parameter:
V : A Fortran variable of numeric intrinsic type (choice).

Output parameter:
size : Size of machine representation of that type (integer).
ierror (Fortran) : Error status (integer); NOTE: NOT OPTIONAL!

How to read routine prototypes: [1.5.4](#).

manpage 60: Routine prototype for MPI_Sizeof

```

|| call MPI_Sizeof(windowdata,window_element_size,ierr)
|| window_size = window_element_size*500
|| call MPI_Win_create( windowdata,window_size,window_element_size,... );

```

This routine is deprecated in *MPI-4*: use of `storage_size` and/or `c_sizeof` is recommended.

5.1.4.2 Python

The `MPI_Win_create` routine needs a displacement in bytes. Here is a good way for finding the size of `numpy` datatypes:

```

|| numpy.dtype('i').itemsize

```

5.2 Derived datatypes

MPI allows you to create your own data types, somewhat (but not completely...) analogous to defining structures in a programming language. MPI data types are mostly of use if you want to send multiple items in one message.

There are two problems with using only elementary datatypes as you have seen so far.

- MPI communication routines can only send multiples of a single data type: it is not possible to send items of different types, even if they are contiguous in memory. It would be possible to use the `MPI_BYTE` data type, but this is not advisable.
- It is also ordinarily not possible to send items of one type if they are not contiguous in memory. You could of course send a contiguous memory area that contains the items you want to send, but that is wasteful of bandwidth.

With MPI data types you can solve these problems in several ways.

- You can create a new *contiguous data type* consisting of an array of elements of another data type. There is no essential difference between sending one element of such a type and multiple elements of the component type.
- You can create a *vector data type* consisting of regularly spaced blocks of elements of a component type. This is a first solution to the problem of sending non-contiguous data.
- For not regularly spaced data, there is the *indexed data type*, where you specify an array of index locations for blocks of elements of a component type. The blocks can each be of a different size.
- The *struct data type* can accomodate multiple data types.

And you can combine these mechanisms to get irregularly spaced heterogeneous data, et cetera.

5.2.1 Basic calls

The typical sequence of calls for creating a new datatype is as follows:

```

|| MPI_Datatype newtype;
|| MPI_Type_<sometype>(< oldtype specifications >, &newtype );
|| MPI_Type_commit( &newtype );
|| /* code that uses your new type */
|| MPI_Type_free( &newtype );

```

5.2.1.1 Datatype objects

MPI derived data types are stored in variables of type `MPI_Datatype` (figure 61).

5.2.1.2 Create calls

The `MPI_Datatype` variable gets its value by a call to one of the following routines:

- `MPI_Type_contiguous` for contiguous blocks of data; section 5.2.2;
- `MPI_Type_vector` for regularly strided data; section 5.2.3;
- `MPI_Type_create_subarray` for subsets out higher dimensional block; section 5.2.4;
- `MPI_Type_create_struct` for heterogeneous irregular data; section 5.2.6;
- `MPI_Type_indexed` and `MPI_Type_hindexed` for irregularly strided data; section 5.2.5.

These calls take an existing type, whether elementary or also derived, and produce a new type.

5.2.1.3 Commit and free

It is necessary to call `MPI_Type_commit` (figure 62) on a new data type, which makes MPI do the indexing calculations for the data type.

When you no longer need the data type, you call `MPI_Type_free` (figure 63).

- The definition of the datatype identifier will be changed to `MPI_DATATYPE_NULL`.
- Any communication using this data type, that was already started, will be completed successfully.
- Datatypes that are defined in terms of this data type will still be usable.

5.2.2 Contiguous type

The simplest derived type is the ‘contiguous’ type, constructed with `MPI_Type_contiguous` (figure 64).

A contiguous type describes an array of items of an elementary or earlier defined type. There is no difference between sending one item of a contiguous type and multiple items of the constituent type. This is illustrated

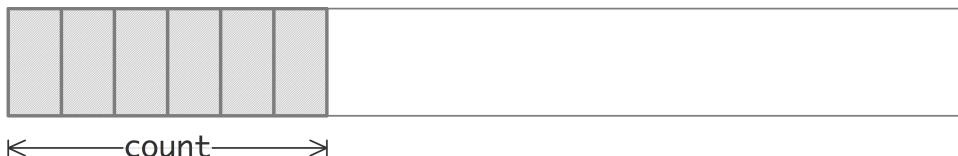


Figure 5.1: A contiguous datatype is built up out of elements of a constituent type

in figure 5.1.

```
// contiguous.c
MPI_Datatype newvectortype;
if (procno==sender) {
    MPI_Type_contiguous(count,MPI_DOUBLE,&newvectortype);
    MPI_Type_commit(&newvectortype);
    MPI_Send(source,1,newvectortype,receiver,0,comm);
```

MPI_Datatype

```
C:  
MPI_Datatype datatype ;  
  
Fortran:  
Type(MPI_Datatype) datatype
```

How to read routine prototypes: [1.5.4](#).

manpage 61: Routine prototype for MPI_Datatype

MPI_Type_commit

```
C:  
int MPI_Type_commit(MPI_Datatype *datatype)  
  
Fortran:  
MPI_Type_commit(datatype, ierror)  
TYPE(MPI_Datatype), INTENT(INOUT) :: datatype  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 62: Routine prototype for MPI_Type_commit

MPI_Type_free

```
int MPI_Type_free (MPI_Datatype *datatype)
```

How to read routine prototypes: [1.5.4](#).

manpage 63: Routine prototype for MPI_Type_free

```

    MPI_Type_free(&newvectortype);
} else if (procno==receiver) {
    MPI_Status recv_status;
    int recv_count;
    MPI_Recv(target,count,MPI_DOUBLE, sender, 0, comm,
              &recv_status);
    MPI_Get_count(&recv_status,MPI_DOUBLE,&recv_count);
    ASSERT(count==recv_count);
}

// contiguous.F90
integer :: newvectortype
if (mytid==sender) then
    call MPI_Type_contiguous(count,MPI_DOUBLE_PRECISION,newvectortype,err)
    call MPI_Type_commit(newvectortype,err)
    call MPI_Send(source,1,newvectortype, receiver, 0, comm,err)
    call MPI_Type_free(newvectortype,err)
else if (mytid==receiver) then
    call MPI_Recv(target,count,MPI_DOUBLE_PRECISION, sender, 0, comm,&
                  recv_status,err)
    call MPI_Get_count(recv_status,MPI_DOUBLE_PRECISION,recv_count,err)
    !ASSERT(count==recv_count);
end if

```

5.2.3 Vector type

The simplest non-contiguous datatype is the ‘vector’ type, constructed with `MPI_Type_vector` (figure 65)

A vector type describes a series of blocks, all of equal size, spaced with a constant stride. This is illustrated

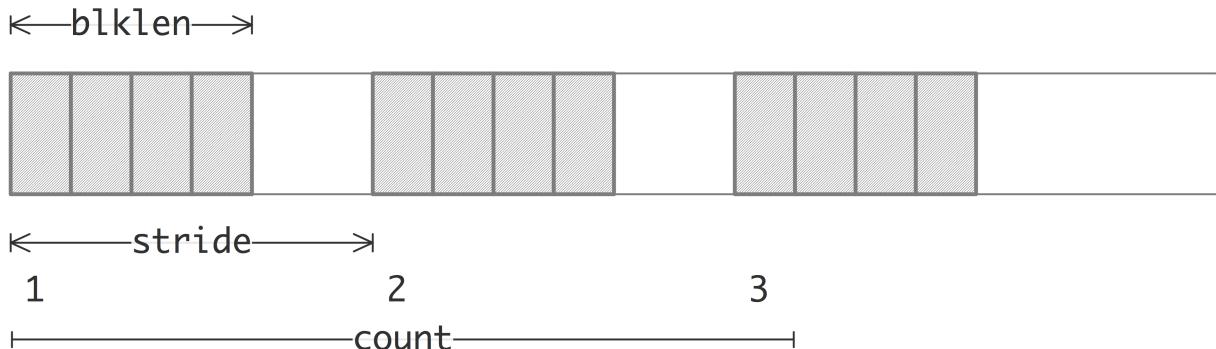


Figure 5.2: A vector datatype is built up out of strided blocks of elements of a constituent type

in figure 5.2.

The vector datatype gives the first non-trivial illustration that datatypes can be *different on the sender and receiver*. If the sender sends b blocks of length l each, the receiver can receive them as b_1 contiguous

MPI_Type_contiguous

Semantics:

```
MPI_TYPE_CONTIGUOUS
    (count, oldtype, newtype)
IN count: replication count (non-negative integer)
IN oldtype: old datatype (handle)
OUT newtype: new datatype (handle)
```

C:

```
int MPI_Type_contiguous
    (int count, MPI_Datatype oldtype, MPI_Datatype *newtype)
```

Fortran:

```
MPI_Type_contiguous
    (count, oldtype, newtype, ierror)
INTEGER, INTENT(IN) :: count
TYPE(MPI_Datatype), INTENT(IN) :: oldtype
TYPE(MPI_Datatype), INTENT(OUT) :: newtype
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
Create_contiguous(self, int count)
```

How to read routine prototypes: [1.5.4](#).

manpage 64: Routine prototype for MPI_Type_contiguous

MPI_Type_vector

Semantics:

```
MPI_TYPE_VECTOR(count, blocklength, stride, oldtype, newtype)
IN count: number of blocks (non-negative integer)
IN blocklength: number of elements in each block (non-negative integer)
IN stride: number of elements between start of each block (integer)
IN oldtype: old datatype (handle)
OUT newtype: new datatype (handle)
```

C:

```
int MPI_Type_vector
    (int count, int blocklength, int stride,
     MPI_Datatype oldtype, MPI_Datatype *newtype)
```

Fortran:

```
MPI_Type_vector(count, blocklength, stride, oldtype, newtype, ierror)
INTEGER, INTENT(IN) :: count, blocklength, stride
TYPE(MPI_Datatype), INTENT(IN) :: oldtype
TYPE(MPI_Datatype), INTENT(OUT) :: newtype
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
MPI.Datatype.Create_vector(self, int count, int blocklength, int stride)
```

How to read routine prototypes: [1.5.4](#).

manpage 65: Routine prototype for MPI_Type_vector

5. MPI topic: Data types

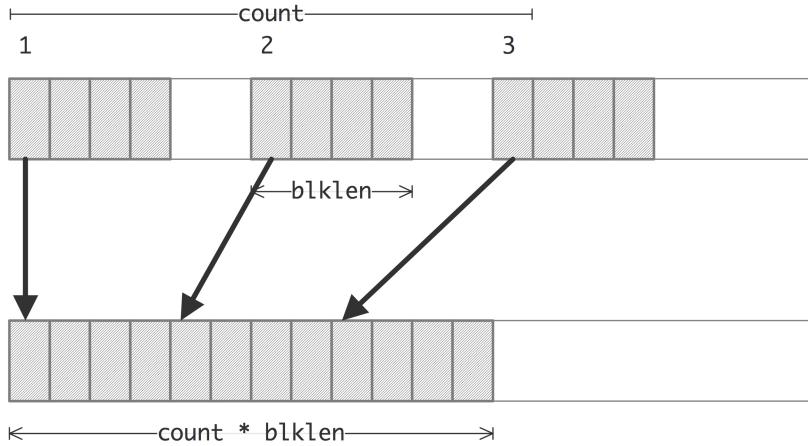


Figure 5.3: Sending a vector datatype and receiving it as elementary or contiguous

elements, either as a contiguous datatype, or as a contiguous buffer of an elementary type; see figure 5.3. In this case, the receiver has no knowledge of the stride of the datatype on the sender.

In this example a vector type is created only on the sender, in order to send a strided subset of an array; the receiver receives the data as a contiguous block.

```
// 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);
}

// vector.F90
integer :: newvectortype
ALLOCATE(source(stride*count))
ALLOCATE(target(stride*count))
if (mytid==sender) then
    call MPI_Type_vector(count,1,stride,MPI_DOUBLE_PRECISION,&
                        newvectortype,err)
    call MPI_Type_commit(newvectortype,err)
    call MPI_Send(source,1,newvectortype,receiver,0,comm,err)
    call MPI_Type_free(newvectortype,err)
```

```

|| else if (mytid==receiver) then
  call MPI_Recv(target,count,MPI_DOUBLE_PRECISION, sender, 0, comm, &
    recv_status,err)
  call MPI_Get_count(recv_status,MPI_DOUBLE_PRECISION,recv_count,err)
end if

## vector.py
source = np.empty(stride*count,dtype=np.float64)
target = np.empty(count,dtype=np.float64)
if procid==sender:
  newvectortype = MPI.DOUBLE.Create_vector(count,1,stride)
  newvectortype.Commit()
  comm.Send([source,1,newvectortype],dest=the_other)
  newvectortype.Free()
elif procid==receiver:
  comm.Recv([target,count,MPI.DOUBLE],source=the_other)

```

Figure 5.4 indicates one source of irregular data: with a matrix on *column-major storage*, a column is stored

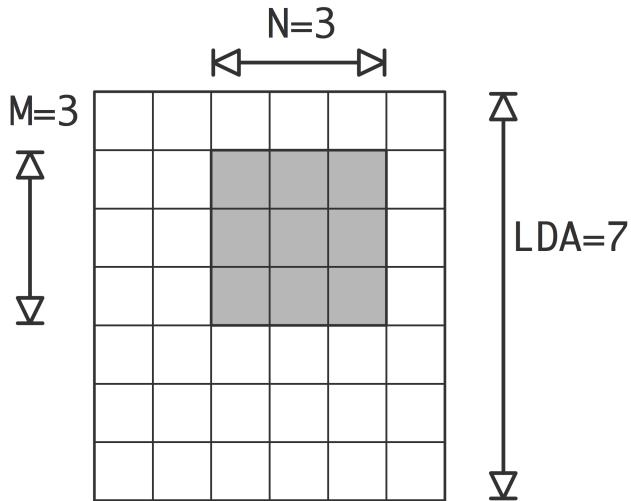


Figure 5.4: Memory layout of a row and column of a matrix in column-major storage

in contiguous memory. However, a row of such a matrix is not contiguous; its elements being separated by a *stride* equal to the column length.

Exercise 5.1. How would you describe the memory layout of a submatrix, if the whole matrix has size $M \times N$ and the submatrix $m \times n$?

As an example of this datatype, consider the example of transposing a matrix, for instance to convert between C and Fortran arrays (see section HPSC-??). Suppose that a processor has a matrix stored in C, row-major, layout, and it needs to send a column to another processor. If the matrix is declared as

```
|| int M, N; double mat[M][N]
```

then a column has M blocks of one element, spaced N locations apart. In other words:

5. MPI topic: Data types

```

|| MPI_Datatype MPI_column;
|| MPI_Type_vector(
    /* count= */ M, /* blocklength= */ 1, /* stride= */ N,
    MPI_DOUBLE, &MPI_column );

```

Sending the first column is easy:

```

|| MPI_Send( mat, 1, MPI_column, ... );

```

The second column is just a little trickier: you now need to pick out elements with the same stride, but starting at $A[0][1]$.

```

|| MPI_Send( &(mat[0][1]), 1, MPI_column, ... );

```

You can make this marginally more efficient (and harder to read) by replacing the index expression by $\text{mat}+1$.

Exercise 5.2. Suppose you have a matrix of size $4N \times 4N$, and you want to send the elements $A[4*i][4*j]$ with $i, j = 0, \dots, N - 1$. How would you send these elements with a single transfer?

Exercise 5.3. Allocate a matrix on processor zero, using Fortran column-major storage.
Using P sendrecv calls, distribute the rows of this matrix among the processors.

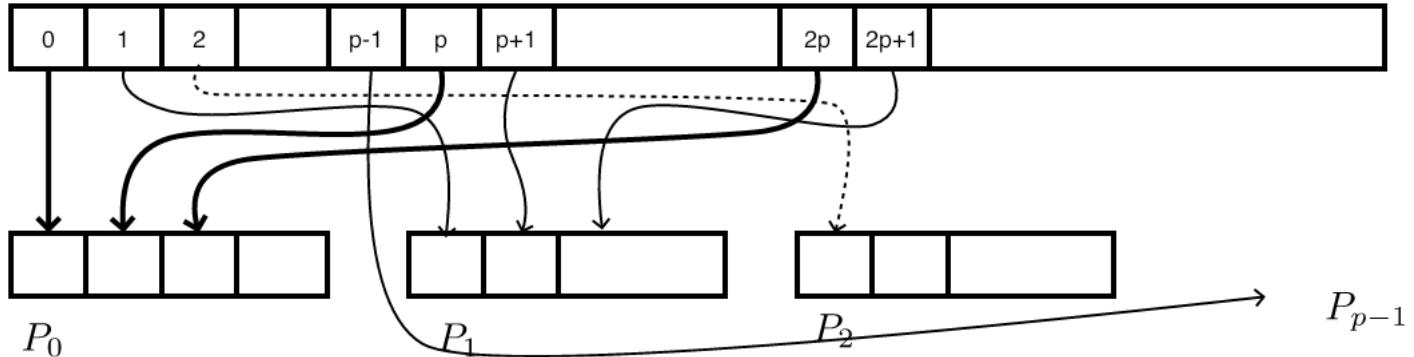


Figure 5.5: Send strided data from process zero to all others

Exercise 5.4. 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 5.5. Write code to compare the time it takes to send a strided subset from an array:
copy the elements by hand to a smaller buffer, or use a vector data type. What do you find? You may need to test on fairly large arrays.

5.2.4 Subarray type

The vector datatype can be used for blocks in an array of dimension more than 2 by using it recursively. However, this gets tedious. Instead, there is an explicit subarray type `MPI_Type_create_subarray` (figure 66) This describes the dimensionality and extent of the array, and the starting point (the ‘upper left corner’) and extent of the subarray. The possibilities for the `order` parameter are `MPI_ORDER_C` and `MPI_ORDER_FORTRAN`.

Exercise 5.6.

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.

Fortran note. Subarrays are naturally supported in Fortran through array sections:

```
// section.F90
integer,parameter :: siz=20
real,dimension(siz,siz) :: matrix = [ ((j+(i-1)*siz,i=1,siz)
, j=1,siz) ]
real,dimension(2,2) :: submatrix
if (procno==0) then
  call MPI_Send(matrix(1:2,1:2),4,MPI_REAL,1,0,comm)
else if (procno==1) then
  call MPI_Recv(submatrix,4,MPI_REAL,0,0,comm,
  MPI_STATUS_IGNORE)
  if (submatrix(2,2)==22) then
    print *, "Yay"
  else
    print *, "nay...."
  end if
end if
```

5.2.5 Indexed type

The indexed datatype, constructed with `MPI_Type_indexed` (figure 67) can send arbitrarily located elements from an array of a single datatype. You need to supply an array of index locations, plus an array of blocklengths with a separate blocklength for each index. The total number of elements sent is the sum of the blocklengths.

The following example picks items that are on prime number-indexed locations.

```
// indexed.c
displacements = (int*) malloc(count*sizeof(int));
blocklengths = (int*) malloc(count*sizeof(int));
source = (int*) malloc(totalcount*sizeof(int));
target = (int*) malloc(targetbuffersize*sizeof(int));
MPI_Datatype newvectortype;
if (procno==sender) {
  MPI_Type_indexed(count,blocklengths,displacements,MPI_INT,&newvectortype);
  MPI_Type_commit(&newvectortype);
  MPI_Send(source,1,newvectortype,the_other,0,comm);
  MPI_Type_free(&newvectortype);
```

MPI_Type_create_subarray

```

Semantics:
MPI_TYPE_CREATE_SUBARRAY(
    ndims, array_of_sizes, array_of_subsizes,
    array_of_starts, order, oldtype, newtype)
IN ndims: number of array dimensions (positive integer)
IN array_of_sizes: number of elements of type oldtype in each dimension
    of the full array (array of positive integers)
IN array_of_subsizes: number of elements of type oldtype in each
    dimension of the subarray (array of positive integers)
IN array_of_starts: starting coordinates of the subarray in each
    dimension (array of non-negative integers)
IN order: array storage order flag (state)
IN oldtype: array element datatype (handle)
OUT newtype: new datatype (handle)

C:
int MPI_Type_create_subarray(
    int ndims, const int array_of_sizes[],
    const int array_of_subsizes[], const int array_of_starts[],
    int order, MPI_Datatype oldtype, MPI_Datatype *newtype)

Fortran:
MPI_Type_create_subarray(ndims, array_of_sizes, array_of_subsizes,
    array_of_starts, order, oldtype, newtype, ierror)
INTEGER, INTENT(IN) :: ndims, array_of_sizes(ndims),
    array_of_subsizes(ndims), array_of_starts(ndims), order
TYPE(MPI_Datatype), INTENT(IN) :: oldtype
TYPE(MPI_Datatype), INTENT(OUT) :: newtype
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Python:
MPI.Datatype.Create_subarray
    (self, sizes, subsizes, starts, int order=ORDER_C)

```

How to read routine prototypes: [1.5.4](#).

manpage 66: Routine prototype for MPI_Type_create_subarray

MPI_Type_indexed

Semantics:

```
count [in] number of blocks --
      also number of entries in indices and blocklens
blocklens [in] number of elements in each block
      (array of nonnegative integers)
indices [in] displacement of each block in multiples of old_type
      (array of integers)
old_type [in] old datatype (handle)
newtype [out] new datatype (handle)
```

C:

```
int MPI_Type_indexed(int count,
                      const int array_of_blocklengths[],
                      const int array_of_displacements[],
                      MPI_Datatype oldtype, MPI_Datatype
                      *newtype)
```

Fortran:

```
MPI_Type_indexed(count, array_of_blocklengths, array_of_displacements,
                  oldtype, newtype, ierror)
INTEGER, INTENT(IN) :: count, array_of_blocklengths(count),
array_of_displacements(count)
TYPE(MPI_Datatype), INTENT(IN) :: oldtype
TYPE(MPI_Datatype), INTENT(OUT) :: newtype
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
MPI.Datatype.Create_vector(self, blocklengths,displacements )
```

How to read routine prototypes: 1.5.4.

manpage 67: Routine prototype for MPI_Type_indexed

5. MPI topic: Data types

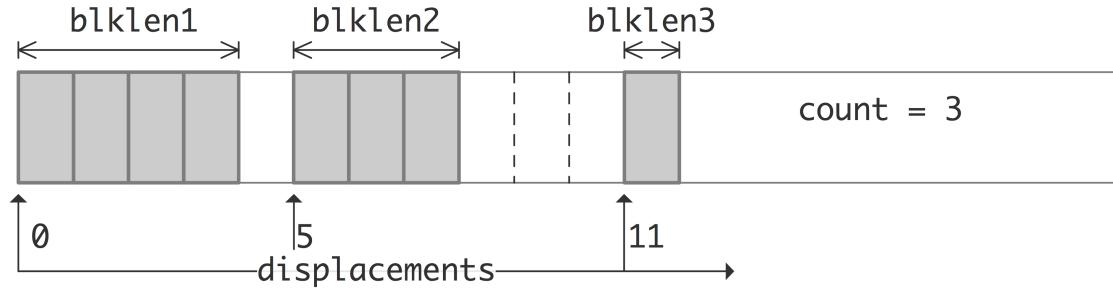


Figure 5.6: The elements of an MPI Indexed datatype

```

} else if (procno==receiver) {
    MPI_Status recv_status;
    int recv_count;
    MPI_Recv(target,targetbuffersize,MPI_INT,the_other,0,comm,
              &recv_status);
    MPI_Get_count(&recv_status,MPI_INT,&recv_count);
    ASSERT(recv_count==count);
}

// indexed.F90
integer :: newvectortype;
ALLOCATE(indices(count))
ALLOCATE(blocklengths(count))
ALLOCATE(source(totalcount))
ALLOCATE(targt(count))
if (mytid==sender) then
    call MPI_Type_indexed(count,blocklengths,indices,MPI_INT,&
                           newvectortype,err)
    call MPI_Type_commit(newvectortype,err)
    call MPI_Send(source,1,newvectortype,receiver,0,comm,err)
    call MPI_Type_free(newvectortype,err)
else if (mytid==receiver) then
    call MPI_Recv(targt,count,MPI_INT,receiver,0,comm,&
                  recv_status,err)
    call MPI_Get_count(recv_status,MPI_INT,recv_count,err)
    !   ASSERT(recv_count==count);
end if

## indexed.py
displacements = np.empty(count,dtype=np.int)
blocklengths = np.empty(count,dtype=np.int)
source = np.empty(totalcount,dtype=np.float64)
target = np.empty(count,dtype=np.float64)
if procid==sender:
    newindextype = MPI.DOUBLE.Create_indexed(blocklengths,displacements)
    newindextype.Commit()
    comm.Send([source,1,newindextype],dest=the_other)
    newindextype.Free()
elif procid==receiver:
    comm.Recv([target,count,MPI.DOUBLE],source=the_other)

```

You can also `MPI_Type_create_hindexed` which describes blocks of a single old type, but with index locations in bytes, rather than in multiples of the old type.

```
int MPI_Type_create_hindexed
  (int count, int blocklens[], MPI_Aint indices[],
   MPI_Datatype old_type, MPI_Datatype *newtype)
```

A slightly simpler version, `MPI_Type_create_hindexed_block` (figure 68) assumes constant block length.

There is an important difference between the `hindexed` and the above `MPI_Type_indexed`: that one described offsets from a base location; these routines describes absolute memory addresses. You can use this to send for instance the elements of a linked list. You would traverse the list, recording the addresses of the elements with `MPI_Get_address` (figure 69).

In C++ you can use this to send an `std::vector`, that is, a vector object from the *C++ standard library*, if the component type is a pointer.

5.2.6 Struct type

The structure type, created with `MPI_Type_create_struct` (figure 70), can contain multiple data types. (The routine `MPI_Type_struct` is deprecated with MPI 3.) The specification contains a ‘count’ parameter

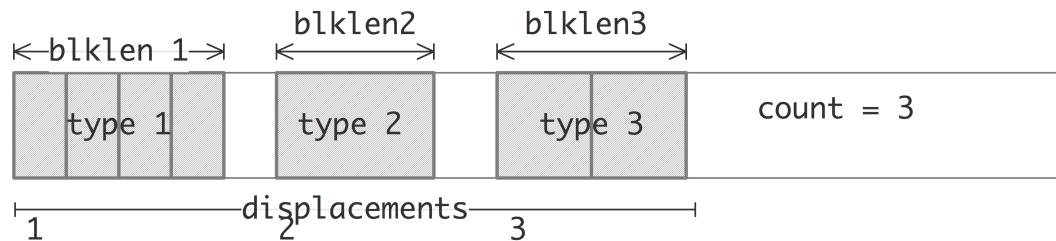


Figure 5.7: The elements of an MPI Struct datatype

that specifies how many blocks there are in a single structure. For instance,

```
struct {
  int i;
  float x,y;
} point;
```

has two blocks, one of a single integer, and one of two floats. This is illustrated in figure 5.7.

count The number of blocks in this datatype. The `blocklengths`, `displacements`, `types` arguments have to be at least of this length.

blocklengths array containing the lengths of the blocks of each datatype.

displacements array describing the relative location of the blocks of each datatype.

types array containing the datatypes; each block in the new type is of a single datatype; there can be multiple blocks consisting of the same type.

MPI_Type_create_hindexed_block

```
int MPI_Type_create_hindexed_block
    (int count, int blocklength,
     const MPI_Aint array_of_displacements[],
     MPI_Datatype oldtype, MPI_Datatype *newtype)

Input Parameters:
count : length of array of displacements (integer)
blocklength : size of block (integer)
array_of_displacements : array of displacements (array of integer)
oldtype : old datatype (handle)

Output Parameter:
newtype : new datatype (handle)
```

How to read routine prototypes: [1.5.4](#).

manpage 68: Routine prototype for MPI_Type_create_hindexed_block

MPI_Get_address

```
C:
int MPI_Get_address
    (void *location,
     MPI_Aint *address
    );

Input Parameters:
location : location in caller memory (choice)

Output parameters:
address : address of location (address)
```

How to read routine prototypes: [1.5.4](#).

manpage 69: Routine prototype for MPI_Get_address

MPI_Type_create_struct

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

How to read routine prototypes: [1.5.4](#).

manpage 70: Routine prototype for MPI_Type_create_struct

In this example, unlike the previous ones, both sender and receiver create the structure type. With structures it is no longer possible to send as a derived type and receive as a array of a simple type. (It would be possible to send as one structure type and receive as another, as long as they have the same *datatype signature*.)

```

// struct.c
struct object {
    char c;
    double x[2];
    int i;
};

MPI_Datatype newstructuretype;
int structlen = 3;
int blocklengths[structlen]; MPI_Datatype types[structlen];
MPI_Aint displacements[structlen];

/*
 * where are the components relative to the structure?
 */
MPI_Aint current_displacement=0;

// one character
blocklengths[0] = 1; types[0] = MPI_CHAR;
displacements[0] = (size_t)&(myobject.c) - (size_t)&myobject;

// two doubles
blocklengths[1] = 2; types[1] = MPI_DOUBLE;
displacements[1] = (size_t)&(myobject.x) - (size_t)&myobject;

// one int
blocklengths[2] = 1; types[2] = MPI_INT;
displacements[2] = (size_t)&(myobject.i) - (size_t)&myobject;

MPI_Type_create_struct(structlen, blocklengths, displacements, types, &
    newstructuretype);
MPI_Type_commit(&newstructuretype);
if (procno==sender) {
    MPI_Send(&myobject, 1, newstructuretype, the_other, 0, comm);
} else if (procno==receiver) {
    MPI_Recv(&myobject, 1, newstructuretype, the_other, 0, comm, MPI_STATUS_IGNORE);
}
MPI_Type_free(&newstructuretype);

// struct.F90
Type object
    character :: c
    real*8,dimension(2) :: x
    integer :: i
end type object
type(object) :: myobject
integer,parameter :: structlen = 3
type(MPI_Datatype) :: newstructuretype
integer,dimension(structlen) :: blocklengths
type(MPI_Datatype),dimension(structlen) :: types;
MPI_Aint,dimension(structlen) :: displacements

```

5. MPI topic: Data types

```
MPI_Aint :: base_displacement, next_displacement
if (procno==sender) then
    myobject%c = 'x'
    myobject%x(0) = 2.7; myobject%x(1) = 1.5
    myobject%i = 37

    !! component 1: one character
    blocklengths(1) = 1; types(1) = MPI_CHAR
    call MPI_Get_address(myobject,base_displacement)
    call MPI_Get_address(myobject%c,next_displacement)
    displacements(1) = next_displacement-base_displacement

    !! component 2: two doubles
    blocklengths(2) = 2; types(2) = MPI_DOUBLE
    call MPI_Get_address(myobject%x,next_displacement)
    displacements(2) = next_displacement-base_displacement

    !! component 3: one int
    blocklengths(3) = 1; types(3) = MPI_INT
    call MPI_Get_address(myobject%i,next_displacement)
    displacements(3) = next_displacement-base_displacement

if (procno==sender) then
    call MPI_Send(myobject,1,newstructuretype,receiver,0,comm)
else if (procno==receiver) then
    call MPI_Recv(myobject,1,newstructuretype, sender, 0, comm, MPI_STATUS_IGNORE)
end if
call MPI_Type_free(newstructuretype)
```

Note the displacement calculations in this example, which involve some not so elegant pointer arithmetic. (Alternatively, you could use `MPI_Get_address`, which is the only way address calculations can be done in Fortran.)

It would have been incorrect to write

```
|| displacement[0] = 0;
|| displacement[1] = displacement[0] + sizeof(char);
```

since you do not know the way the compiler lays out the structure in memory¹.

If you want to send more than one structure, you have to worry more about padding in the structure. You can solve this by adding an extra type `MPI_UB` for the ‘upper bound’ on the structure:

```
|| displacements[3] = sizeof(myobject); types[3] = MPI_UB;
|| MPI_Type_create_struct(struclen+1,.....);
```

5.3 Type size

The space that MPI takes for a structure type can be queried in a variety of ways. First of all `MPI_Type_size` (figure 71) counts the datatype size as the number of bytes occupied by the data in a type. That means that

1. Homework question: what does the language standard say about this?

in an *MPI* vector datatype it does not count the gaps.

```
// typesize.c
MPI_Type_vector(count,bs,stride,MPI_DOUBLE,&newtype);
MPI_Type_commit(&newtype);
MPI_Type_size(newtype,&size);
ASSERT( size==(count*bs)*sizeof(double) );
MPI_Type_free(&newtype);
```

On the other hand, the *datatype extent*, measured with `MPI_Type_get_extent` (figure 72) is strictly the distance from the first to the last data item of the type, that is, with counting the gaps in the type.

```
MPI_Type_vector(count,bs,stride,MPI_DOUBLE,&newtype);
MPI_Type_commit(&newtype);
MPI_Type_get_extent(newtype,&lb,&asize);
ASSERT( lb==0 );
ASSERT( asize==((count-1)*stride+bs)*sizeof(double) );
MPI_Type_free(&newtype);
```

(There is a deprecated function `MPI_Type_extent` with the same functionality.)

The *subarray datatype* need not start at the first element of the buffer, so the extent is an overstatement of how much data is involved. The routine `MPI_Type_get_true_extent` (figure 73) returns the lower bound, indicating where the data starts, and the extent from that point.

```
// trueextent.c
int sender = 0, receiver = 1, the_other = 1-procno,
count = 4;
int sizes[2] = {4,6}, subsizes[2] = {2,3}, starts[2] = {1,2};
MPI_Datatype subarraytype;
if (procno==sender) {
    MPI_Type_create_subarray
        (2,sizes,subsizes,starts,MPI_ORDER_C,MPI_DOUBLE,&subarraytype);
    MPI_Type_commit(&subarraytype);

    MPI_Aint true_lb,true_extent,extent;
    //    MPI_Type_get_extent(subarraytype,&extent);
    MPI_Type_get_true_extent
        (subarraytype,&true_lb,&true_extent);
    MPI_Aint
        comp_lb =
            ( starts[0]*sizes[1]+starts[1] ) *sizeof(double),
        comp_extent =
            ( (starts[0]+subsizes[0]-1)*sizes[1] + starts[1]+subsizes[1] )
            *sizeof(double) - comp_lb;
    //    ASSERT(extent==true_lb+extent);
    ASSERT(true_lb==comp_lb);
    ASSERT(true_extent==comp_extent);

    MPI_Send(source,1,subarraytype,the_other,0,comm);
    MPI_Type_free(&subarraytype);
```

The size of a datatype is not always statically known, for instance if the Fortran `KIND` keyword is used. The translation of datatypes in the source language can be translated to MPI types with `MPI_Type_match_size`

MPI_Type_size

Semantics:

```
int MPI_Type_size(
    MPI_Datatype datatype,
    int *size
);
```

datatype: [in] datatype to get information on (handle)
size: [out] datatype size in bytes

How to read routine prototypes: [1.5.4](#).

manpage 71: Routine prototype for MPI_Type_size

MPI_Type_get_extent

Semantics:

```
int MPI_Type_get_extent(
    MPI_Datatype datatype,
    MPI_Aint *lb, MPI_Aint *extent
);
```

datatype: [in] datatype to get information on (handle)
lb: [out] lower bound of datatype (integer)
extent: [out] extent of datatype (integer)

How to read routine prototypes: [1.5.4](#).

manpage 72: Routine prototype for MPI_Type_get_extent

MPI_Type_get_true_extent

Semantics:

`MPI_Type_get_true_extent (datatype,true_lb,true_extent)`

Input argument:

datatype: Data type for which information is wanted (handle).

Output arguments:

true_lb: True lower bound of data type (integer).

true_extent: True size of data type (integer).

C:

```
int MPI_Type_get_true_extent(
    MPI_Datatype datatype,
    MPI_Aint *true_lb, MPI_Aint *true_extent)
int MPI_Type_get_true_extent_x(
    MPI_Datatype datatype,
    MPI_Count *true_lb, MPI_Count *true_extent)
```

Fortran

```
MPI_TYPE_GET_TRUE_EXTENT(DATATYPE, TRUE_LB, TRUE_EXTENT, IERROR)
    INTEGER      DATATYPE, IERROR
    INTEGER(KIND=MPI_ADDRESS_KIND) TRUE_LB, TRUE_EXTENT
MPI_TYPE_GET_TRUE_EXTENT_X(DATATYPE, TRUE_LB, TRUE_EXTENT, IERROR)
    INTEGER      DATATYPE, IERROR
    INTEGER(KIND=MPI_COUNT_KIND) TRUE_LB, TRUE_EXTENT
```

How to read routine prototypes: [1.5.4](#).

manpage 73: Routine prototype for MPI_Type_get_true_extent

(figure 74) where the `typeclass` argument is one of `MPI_TYPECLASS_REAL`, `MPI_TYPECLASS_INTEGER`, `MPI_TYPECLASS_COMPLEX`.

```
// typematch.c
float x5;
double x10;
int s5,s10;
MPI_Datatype mpi_x5,mpi_x10;

MPI_Type_match_size(MPI_TYPECLASS_REAL,sizeof(x5),&mpi_x5);
MPI_Type_match_size(MPI_TYPECLASS_REAL,sizeof(x10),&mpi_x10);
MPI_Type_size(mpi_x5,&s5);
MPI_Type_size(mpi_x10,&s10);
```

In Fortran, the size of the datatype in the language can be obtained with `MPI_Szef` (note the non-optional error parameter!).

```
// matchkind.F90
call MPI_Szef(x10,s10,ierr)
call MPI_Type_match_size(MPI_TYPECLASS_REAL,s10,mpi_x10)
call MPI_Type_size(mpi_x10,s10)
print *, "10 positions supported, MPI type size is",s10
```

5.4 More about data

5.4.1 Datatype signatures

With the primitive types it pretty much went without saying that if the sender sends an array of doubles, the receiver had to declare the datatype also as doubles. With derived types that is no longer the case: the sender and receiver can declare a different datatype for the send and receive buffer, as long as these have the same *datatype signature*.

The signature of a datatype is the internal representation of that datatype. For instance, if the sender declares a datatype consisting of two doubles, and it sends four elements of that type, the receiver can receive it as two elements of a type consisting of four doubles.

You can also look at the signature as the form ‘under the hood’ in which MPI sends the data.

5.4.2 Big data types

The `size` parameter in MPI send and receive calls is of type integer, meaning that it’s maximally $2^{31} - 1$. These day computers are big enough that this is a limitation. Derived types offer some way out: to send a *big data type* of 10^{40} elements you would

- create a contiguous type with 10^{20} elements, and
- send 10^{20} elements of that type.

This often works, but it’s not perfect. For instance, the routine `returns` the total number of basic elements sent (as opposed to `MPI_Get_count` which would return the number of elements of the derived type). Since its output argument is of integer type, it can’t store the right value.

The MPI 3 standard has addressed this as follows.

- To preserve backwards compatibility, the `size` parameter keeps being of type integer.
- The trick with sending elements of a derived type still works, but
- There are new routines that can return the correct information about the total amount of data; for instance, `MPI_Get_elements_x` returns its result as a `MPI_Count`.

5.4.3 Packing

One of the reasons for derived datatypes is dealing with non-contiguous data. In older communication libraries this could only be done by *packing* data from its original containers into a buffer, and likewise unpacking it at the receiver into its destination data structures.

MPI offers this packing facility, partly for compatibility with such libraries, but also for reasons of flexibility. Unlike with derived datatypes, which transfers data atomically, packing routines add data sequentially to the buffer and unpacking takes them sequentially.

This means that one could pack an integer describing how many floating point numbers are in the rest of the packed message. Correspondingly, the unpack routine could then investigate the first integer and based on it unpack the right number of floating point numbers.

MPI offers the following:

- The `MPI_Pack` command adds data to a send buffer;
- the `MPI_Unpack` command retrieves data from a receive buffer;
- the buffer is sent with a datatype of `MPI_PACKED`.

With `MPI_Pack` data elements can be added to a buffer one at a time. The `position` parameter is updated each time by the packing routine.

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

Conversely, `MPI_Unpack` retrieves one element from the buffer at a time. You need to specify the MPI datatype.

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

A packed buffer is sent or received with a datatype of `MPI_PACKED`. The sending routine uses the `position` parameter to specify how much data is sent, but the receiving routine does not know this value a priori, so has to specify an upper bound.

```
// pack.c
if (procno==sender) {
    MPI_Pack(&nstarts, 1, MPI_INT, buffer, buflen, &position, comm);
    for (int i=0; i<nstarts; i++) {
        double value = rand() / (double) RAND_MAX;
```

```

    MPI_Pack(&value, 1, MPI_DOUBLE, buffer, buflen, &position, comm);
}
MPI_Pack(&nstarts, 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, &nstarts, 1, MPI_INT, comm);
    for (int i=0; i<nstarts; i++) {
        MPI_Unpack(buffer, buflen, &position, &xrecv_value, 1, MPI_DOUBLE, comm);
    }
    MPI_Unpack(buffer, buflen, &position, &irecv_value, 1, MPI_INT, comm);
    ASSERT(irecv_value==nstarts);
}

```

You can precompute the size of the required buffer with `MPI_Pack_size` (figure 75) Add one time `MPI_BSEND_OVERHEAD`.

Exercise 5.7. Suppose you have a ‘structure of arrays’

```

struct aos {
    int length;
    double *reals;
    double *imags;
};

```

with dynamically created arrays. Write code to send and receive this structure.

5.5 Sources used in this chapter

[Listing of code examples/mpi/c/contiguous.c:](#)

[Listing of code XX:](#)

[Listing of code examples/mpi/c/vector.c:](#)

[Listing of code XX:](#)

[Listing of code XX:](#)

[Listing of code examples/mpi/c/indexed.c:](#)

[Listing of code XX:](#)

[Listing of code examples/mpi/c/struct.c:](#)

[Listing of code examples/mpi/f/struct.F90:](#)

[Listing of code examples/mpi/c/vectortypesize.c:](#)

[Listing of code examples/mpi/c/vectortypeextent.c:](#)

[Listing of code examples/mpi/c/trueextent.c:](#)

MPI_Type_match_size

Synopsis:

```
int MPI_Type_match_size
    (int typeclass, int size, MPI_Datatype *datatype)
```

Input Parameters

typeclass : generic type specifier (integer)
size : size, in bytes, of representation (integer)

Output Parameters

datatype : datatype with correct type, size (handle)

Notes

typeclass is one of:

- MPI_TYPECLASS_REAL,
- MPI_TYPECLASS_INTEGER and
- MPI_TYPECLASS_COMPLEX.

How to read routine prototypes: 1.5.4.

manpage 74: Routine prototype for MPI_Type_match_size

MPI_Pack_size

C:

```
int MPI_Pack_size
    (int incount, MPI_Datatype datatype, MPI_Comm comm, int *size)
```

Input parameters:

incount : Count argument to packing call (integer).
datatype : Datatype argument to packing call (handle).
comm : Communicator argument to packing call (handle).

Output parameters:

size : Upper bound on size of packed message, in bytes (integer).

Fortran:

```
MPI_PACK_SIZE(INCOUNT, DATATYPE, COMM, SIZE, IERROR)
```

input parameters:

INTEGER :: INCOUNT, DATATYPE, COMM
INTEGER :: SIZE, IERROR

How to read routine prototypes: 1.5.4.

manpage 75: Routine prototype for MPI_Pack_size

Listing of code examples/mpi/c/typematch.c:

Listing of code examples/mpi/c/typematch.c:

Listing of code examples/mpi/pack/pack.c:

Chapter 6

MPI topic: Communicators

6.1 Communicator basics

A communicator is an object describing a group of processes. In many applications all processes work together closely coupled, and the only communicator you need is `MPI_COMM_WORLD`, the group describing all processes that your job starts with. This group is fixed: it can neither be extended with additional processes, nor can it contract, for instance to eliminate crashed processes. But some flexibility in process handling may still be needed. For instance, there are circumstances where you want one subset of processes to operate independently of another subset. Examples are:

- If processors are organized in a 2×2 grid, you may want to do broadcasts inside a row or column.
- For an application that includes a producer and a consumer part, it makes sense to split the processors accordingly.

In this section we will see mechanisms for defining new communicators that are subsets of `MPI_COMM_WORLD`. Chapter 7 discusses dynamic process management, which, while not extending `MPI_COMM_WORLD` does extend the set of available processes.

6.1.1 Communicator types

There are three predefined communicators:

- `MPI_COMM_WORLD` comprises all processes that were started together by *mpirun* (or some related program).
- `MPI_COMM_SELF` is the communicator that contains only the current process.
- `MPI_COMM_NULL` is the invalid communicator. Routines that construct communicators can give this as result if an error occurs.

If you don't want to write `MPI_COMM_WORLD` repeatedly, you can assign that value to a variable of type `MPI_Comm` (figure 76).

Examples:

```
// C:  
#include <mpi.h>  
MPI_Comm comm = MPI_COMM_WORLD;
```

```
!! Fortran 2008 interface
use mpi_f08
Type(MPI_Comm) :: comm = MPI_COMM_WORLD

!! Fortran legacy interface
#include <mpif.h>
Integer :: comm = MPI_COMM_WORLD
```

6.2 Subcommunications

In many scenarios you divide a large job over all the available processors. However, your job has two or more parts that can be considered as jobs by themselves. In that case it makes sense to divide your processors into subgroups accordingly.

Suppose for instance that you are running a simulation where inputs are generated, a computation is performed on them, and the results of this computation are analyzed or rendered graphically. You could then consider dividing your processors in three groups corresponding to generation, computation, rendering.

As long as you only do sends and receives, this division works fine. However, if one group of processes needs to perform a collective operation, you don't want the other groups involved in this. Thus, you really want the three groups to be really distinct from each other.

In order to make such subsets of processes, MPI has the mechanism of taking a subset of `MPI_COMM_WORLD` and turning that subset into a new communicator.

Now you understand why the MPI collective calls had an argument for the communicator: a collective involves all processes of that communicator. By making a communicator that contains a subset of all available processes, you can do a collective on that subset.

6.2.1 Scenario: distributed linear algebra

For *scalability* reasons, matrices should often be distributed in a 2D manner, that is, each process receives a subblock that is not a block of columns or rows. This means that the processors themselves are, at least logically, organized in a 2D grid. Operations then involve reductions or broadcasts inside rows or columns. For this, a row or column of processors needs to be in a subcommunicator.

6.2.2 Scenario: climate model

A climate simulation code has several components, for instance corresponding to land, air, ocean, and ice. You can imagine that each needs a different set of equations and algorithms to simulate. You can then divide your processes, where each subset simulates one component of the climate, occasionally communicating with the other components.

6.2.3 Scenario: quicksort

The popular quicksort algorithm works by splitting the data into two subsets that each can be sorted individually. If you want to sort in parallel, you could implement this by making two subcommunicators, and sorting the data on these, creating recursively more subcommunicators.

6.2.4 Shared memory

There is an important application of communicator splitting in the context of one-sided communication, grouping processes by whether they access the same shared memory area; see section 11.1.

6.3 Duplicating communicators

With `MPI_Comm_dup` you can make an exact duplicate of a communicator. This may seem pointless, but it is actually very useful for the design of software libraries. Image that you have a code

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

and suppose that the library has receive calls. Now it is possible that the receive in the library inadvertently catches the message that was sent in the outer environment.

In section 12.7 it was explained that MPI messages are non-overtaking. This may lead to confusing situations, witness the following. First of all, here is code where the library stores the communicator of the calling program:

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

This models a main program that does a simple message exchange, and it makes two calls to library routines. Unbeknown to the user, the library also issues send and receive calls, and they turn out to interfere.

Here

- The main program does a send,

6. MPI topic: Communicators

- the library call `function_start` does a send and a receive; because the receive can match either send, it is paired with the first one;
- the main program does a receive, which will be paired with the send of the library call;
- both the main program and the library do a wait call, and in both cases all requests are successfully fulfilled, just not the way you intended.

To prevent this confusion, the library should duplicate the outer communicator with `MPI_Comm_dup` (figure 77) and send all messages with respect to its duplicate. Now messages from the user code can never reach the library software, since they are on different communicators.

```
// 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();
};

## commdup.py
class Library():
    def __init__(self, comm):
        # wrong: self.comm = comm
        self.comm = comm.Dup()
        self.other = self.comm.Get_size() - self.comm.Get_rank() - 1
        self.requests = [None] * 2
    def communication_start(self):
        sendbuf = np.empty(1, dtype=np.int); sendbuf[0] = 37
        recvbuf = np.empty(1, dtype=np.int)
        self.requests[0] = self.comm.Isend(sendbuf, dest=other, tag=2)
        self.requests[1] = self.comm.Irecv(recvbuf, source=other)
    def communication_end(self):
        MPI.Request.Waitall(self.requests)

mylibrary = Library(comm)
my_requests[0] = comm.Isend(sendbuffer, dest=other, tag=1)
mylibrary.communication_start()
my_requests[1] = comm.Irecv(recvbuffer, source=other)
MPI.Request.Waitall(my_requests, my_status)
mylibrary.communication_end()
```

MPI_Comm

C:
MPI_Comm commvariable; // datatype is defined in mpi.h

Fortran with 2008 support:
Type(MPI_Comm) :: commvariable

Fortran legacy interface:
Integer :: commvariable

Python:
MPI.Comm : class of communicators
MPI.Intracomm
MPI.Intercomm : subclasses of the MPI.Comm class.
MPI.Comm.Is_inter()
MPI.Comm.Is_intra() : convenience functions, not part of the MPI standard.

How to read routine prototypes: 1.5.4.

manpage 76: Routine prototype for MPI_Comm

MPI_Comm_dup

Semantics:
MPI_COMM_DUP(comm, newcomm)
IN comm: communicator (handle)
OUT newcomm: copy of comm (handle)

C:
int MPI_Comm_dup(MPI_Comm comm, MPI_Comm *newcomm)

F:
MPI_Comm_dup(comm, newcomm, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Comm), INTENT(OUT) :: newcomm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Py:
newcomm = oldcomm.Dup(Info info=None)

How to read routine prototypes: 1.5.4.

manpage 77: Routine prototype for MPI_Comm_dup

Newly created communicators should be released again with `MPI_Comm_free`. Note how the preceding example does this in a C++ destructor.

6.4 Splitting a communicator

Above we saw several scenarios where it makes sense to divide `MPI_COMM_WORLD` into disjoint subcommunicators. The command `MPI_Comm_split` (figure 78) uses a ‘colour’ to define these subcommunicators: all processes in the old communicator with the same colour wind up in a new communicator together. The old communicator still exists, so processes now have two different contexts in which to communicate.

The ranking of processes in the new communicator is determined by a ‘key’ value. Most of the time, there is no reason to use a relative ranking that is different from the global ranking, so the `MPI_Comm_rank` value of the global communicator is a good choice.

Here is one example of communicator splitting. Suppose your processors are in a two-dimensional grid:

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

You can now create a communicator per column:

```
|| MPI_Comm column_comm;
|| MPI_Comm_split( MPI_COMM_WORLD, proc_j, mytid, &column_comm );
```

and do a broadcast in that column:

```
|| MPI_Bcast( data, /* tag: */ 0, column_comm );
```

Because of the SPMD nature of the program, you are now doing in parallel a broadcast in every processor column. Such operations often appear in *dense linear algebra*.

The `MPI_Comm_split` routine has a ‘key’ parameter, which controls how the processes in the new communicator are ordered. By supplying the rank from the original communicator you let them be arranged in the same order.

One application of communicator splitting is setting up a processor grid, with the possibility of using MPI solely within one row or column; see figure 6.1.

Exercise 6.1. 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 2×3 processor grid you should find:

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

MPI_Comm_split

Semantics:

```
MPI_COMM_SPLIT(comm, color, key, newcomm)
IN comm: communicator (handle)
IN color: control of subset assignment (integer)
IN key: control of rank assignment (integer)
OUT newcomm: new communicator (handle)
```

C:

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

F:

```
MPI_Comm_split(comm, color, key, newcomm, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, INTENT(IN) :: color, key
TYPE(MPI_Comm), INTENT(OUT) :: newcomm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
MPI_COMM_SPLIT(COMM, COLOR, KEY, NEWCOMM, IERROR)
INTEGER COMM, COLOR, KEY, NEWCOMM, IERROR
```

Py:

```
newcomm = comm.Split(int color=0, int key=0)
```

How to read routine prototypes: [1.5.4.](#)

manpage 78: Routine prototype for MPI`Comm`split

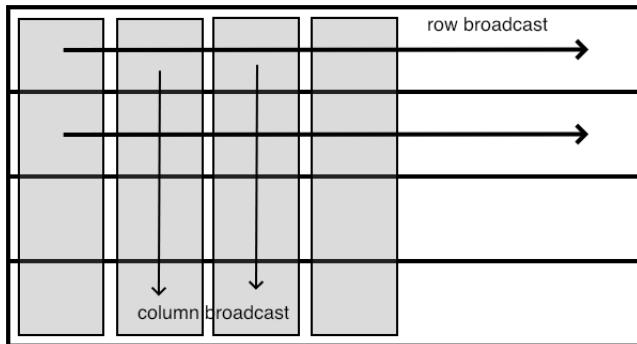


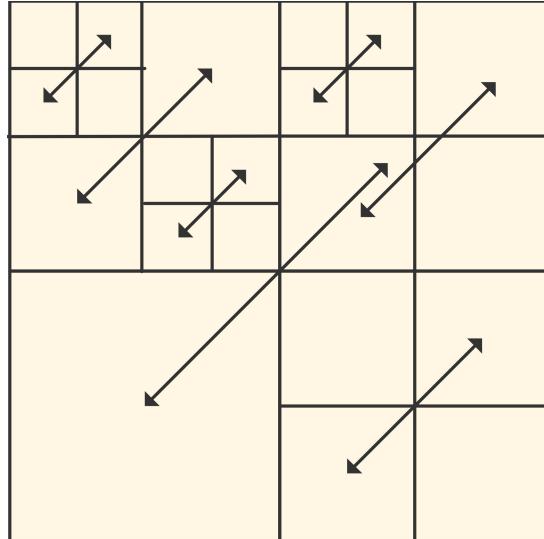
Figure 6.1: Row and column broadcasts in subcommunicators

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.

As another example of communicator splitting, consider the recursive algorithm for *matrix transposition*. Processors are organized in a square grid. The matrix is divided on 2×2 block form.

Exercise 6.2. Implement a recursive algorithm for matrix transposition:



- Swap blocks (1, 2) and (2, 1); then
- Divide the processors into four subcommunicators, and apply this algorithm recursively on each;
- If the communicator has only one process, transpose the matrix in place.

6.5 Communicators and groups

You saw in section 6.4 that it is possible derive communicators that have a subset of the processes of another communicator. There is a more general mechanism, using `MPI_Group` (figure 79) objects.

Using groups, it takes three steps to create a new communicator:

1. Access the `MPI_Group` of a communicator object using `MPI_Comm_group` (figure 80).
2. Use various routines, discussed next, to form a new group.
3. Make a new communicator object from the group with `MPI_Group`, using `MPI_Comm_create` (figure 81)

Creating a new communicator from a group is collective on the old communicator. There is also a routine `MPI_Comm_create_group` that only needs to be called on the group that constitutes the new communicator.

6.5.1 Process groups

Groups are manipulated with `MPI_Group_incl`, `MPI_Group_excl`, `MPI_Group_difference` and a few more.

You can name your communicators with `MPI_Comm_set_name`, which could improve the quality of error messages when they arise.

```
||| MPI_Comm_group (comm, group, ierr)
||| MPI_Comm_create (MPI_Comm comm, MPI_Group group, MPI_Comm newcomm, ierr)
||| MPI_Group_union (group1, group2, newgroup, ierr)
||| MPI_Group_intersection (group1, group2, newgroup, ierr)
||| MPI_Group_difference (group1, group2, newgroup, ierr)
||| MPI_Group_incl (group, n, ranks, newgroup, ierr)
||| MPI_Group_excl (group, n, ranks, newgroup, ierr)
||| MPI_Group_size (group, size, ierr)
||| MPI_Group_rank (group, rank, ierr)
```

6.6 Inter-communicators

In several scenarios it may be desirable to have a way to communicate between communicators. For instance, an application can have clearly functionally separated modules (preprocessor, simulation, postprocessor) that need to stream data pairwise. In another example, dynamically spawned processes (section 7.1) get their own value of `MPI_COMM_WORLD`, but still need to communicate with the process(es) that spawned them. In this section we will discuss the *inter-communicator* mechanism that serves such use cases.

Communicating between disjoint communicators can of course be done by having a communicator that overlaps them, but this would be complicated: since the ‘inter’ communication happens in the overlap communicator, you have to translate its ordering into those of the two worker communicators. It would be easier to express messages directly in terms of those communicators, and this is what happens in an *inter-communicator*.

A call to `MPI_Intercomm_create` (figure 82) involves the following communicators:

MPI_Group

C:
MPI_Group groupvariable; // datatype is defined in mpi.h

Fortran with 2008 support:
Type(MPI_Group) :: groupvariable

Fortran legacy interface:
Integer :: groupvariable

Python:
MPI.Group is a class

How to read routine prototypes: [1.5.4](#).

manpage 79: Routine prototype for MPI_Group

MPI_Comm_group

Synopsis
int MPI_Comm_group(MPI_Comm comm, MPI_Group *group)

Input Parameters:
comm : Communicator (handle)

Output Parameters
group : Group in communicator (handle)

How to read routine prototypes: [1.5.4](#).

manpage 80: Routine prototype for MPI_Comm_group

MPI_Comm_create

Synopsis

```
MPI_Comm_create( MPI_Comm comm, MPI_Group group, MPI_Comm *newcomm )
```

Input parameters:

comm : Communicator (handle).

group : Group, which is a subset of the group of comm (handle).

Output parameters:

newcomm : New communicator (handle).

C:

```
int MPI_Comm_create(MPI_Comm comm, MPI_Group group, MPI_Comm *newcomm)
```

Fortran90:

```
MPI_COMM_CREATE(COMM, GROUP, NEWCOMM, IERROR)
INTEGER      COMM, GROUP, NEWCOMM, IERROR
```

Fortran2008:

```
MPI_Comm_create(comm, group, newcomm, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Group), INTENT(IN) :: group
TYPE(MPI_Comm), INTENT(OUT) :: newcomm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 81: Routine prototype for MPI_Comm`create

MPI_Intercomm_create

Synopsis:

```
int MPI_Intercomm_create
    (MPI_Comm local_comm, int local_leader,
     MPI_Comm peer_comm, int remote_leader,
     int tag, MPI_Comm *newintercomm
    );
```

Input parameters:

```
local_comm : Local (intra)communicator
local_leader : Rank in local_comm of leader (often 0)
peer_comm : Communicator used to communicate between a designated process in
            the other communicator. Significant only at the process in local_comm
            with rank local_leader.
remote_leader : Rank in peer_comm of remote leader (often 0)
tag : Message tag to use in constructing intercommunicator; if multiple
      MPI_IntercommCreates are being made, they should use different tags
      (more precisely, ensure that the local and remote leaders are using
      different tags for each MPI_IntercommCreate).
```

Output Parameter:

```
comm_out : Created intercommunicator
```

How to read routine prototypes: [1.5.4](#).

manpage 82: Routine prototype for MPI_Intercomm_create

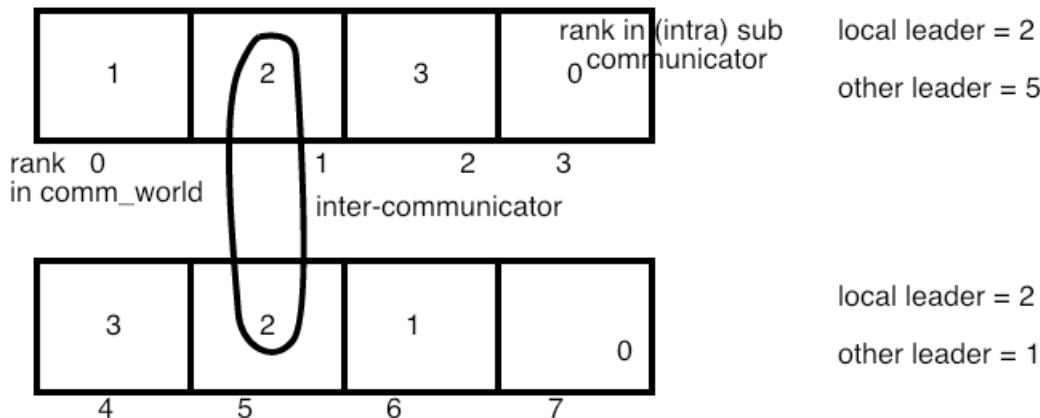


Figure 6.2: Illustration of ranks in an inter-communicator setup

- Two local communicators, which in this context are known as *intra-communicators*: one process in each will act as the local leader, connected to the remote leader;
- The *peer communicator*, often `MPI_COMM_WORLD`, that contains the local communicators;
- An *inter-communicator* that allows the leaders of the subcommunicators to communicate with the other subcommunicator.

Even though the intercommunicator connects only two processes, it is collective on the peer communicator.

6.6.1 Inter-communicator point-to-point

The local leaders can now communicate with each other.

- The sender specifies as target the local number of the other leader in the other sub-communicator;
- Likewise, the receiver specifies as source the local number of the sender in its sub-communicator.

In one way, this design makes sense: processors are referred to in their natural, local, numbering. On the other hand, it means that each group needs to know how the local ordering of the other group is arranged. Using a complicated `key` value makes this difficult.

```
// intercomm.c
if (i_am_local_leader) {
    if (color==0) {
        interdata = 1.2;
        int inter_target = local_number_of_other_leader;
        printf("[%d] sending interdata %e to %d\n",
               procno, interdata, inter_target);
        MPI_Send(&interdata, 1, MPI_DOUBLE, inter_target, 0, intercomm);
    } else {
        MPI_Status status;
        MPI_Recv(&interdata, 1, MPI_DOUBLE, MPI_ANY_SOURCE, MPI_ANY_TAG, intercomm, &
                 status);
        int inter_source = status.MPI_SOURCE;
        printf("[%d] received interdata %e from %d\n",
               procno, interdata, inter_source);
    }
}
```

```
    if (inter_source!=local_number_of_other_leader)
        fprintf(stderr,
                "Got inter communication from unexpected %d; s/b %d\n",
                inter_source, local_number_of_other_leader);
}
}
```

6.6.2 Inter-communicator collectives

The intercommunicator can be used in collectives such as a broadcast.

- In the sending group, the root process passes `MPI_ROOT` as ‘root’ value; all others use `MPI_PROC_NULL`.
 - In the receiving group, all processes use a ‘root’ value that is the rank of the root process in the root group. Note: this is not the global rank!

Gather and scatter behave similarly; the allgather is different: all send buffers of group A are concatenated in rank order, and places on all processes of group B.

Inter-communicators can be used if two groups of process work asynchronously with respect to each other; another application is fault tolerance (section 12.4).

```
if (color==0) { // sending group: the local leader sends
if (i_am_local_leader)
    root = MPI_ROOT;
else
    root = MPI_PROC_NULL;
} else { // receiving group: everyone indicates leader of other group
root = local_number_of_other_leader;
}
if (DEBUG) fprintf(stderr, "[%d] using root value %d\n", procno, root);
MPI_Bcast(&bcast_data, 1, MPI_INT, root, intercomm);
```

6.6.3 Inter-communicator querying

Some of the operations you have seen before for *intra-communicators* behave differently with inter-communicator:

- `MPI_Comm_size` returns the size of the local group, not the size of the inter-communicator.
 - `MPI_Comm_rank` returns the rank in the local group.
 - `MPI_Comm_group` returns the local group.

Test whether a communicator is intra or inter: `MPI_Comm_test_inter` (figure 83).

MPI_Comm_compare works for inter-communicators.

`MPI_Comm_remote_size` (figure 88) `MPI_Comm_remote_group` (figure 85)

Virtual topologies cannot be created with an intercommunicator. To set up virtual topologies, first transform the intercommunicator to an intracommunicator with the function `MPI_Intercomm_merge` (figure 86).

MPI_Comm_test_inter

```
MPI_COMM_TEST_INTER(comm, flag)
IN comm : communicator (handle)
OUT flag : (logical)

int MPI_Comm_test_inter(MPI_Comm comm, int *flag)

MPI_COMM_TEST_INTER(COMM, FLAG, IERROR)
INTEGER COMM, IERROR
LOGICAL FLAG
```

How to read routine prototypes: [1.5.4](#).

manpage 83: Routine prototype for MPI_Comm`test`inter

MPI_Comm_remote_size

Semantics:

```
MPI_COMM_REMOTE_SIZE(comm, size)
IN comm: inter-communicator (handle)
OUT size: number of processes in the remote group of comm (integer)

C:
int MPI_Comm_remote_size(MPI_Comm comm, int *size)
```

Fortran:

```
MPI_Comm_remote_size(comm, size, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, INTENT(OUT) :: size
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
Intercomm.Get_remote_size(self)
```

How to read routine prototypes: [1.5.4](#).

manpage 84: Routine prototype for MPI_Comm`remote`size

MPI_Comm_remote_group

Semantics:

```
MPI_COMM_REMOTE_GROUP(comm, group)
IN comm: inter-communicator (handle)
OUT group: group of processes in the remote group of comm
```

C:

```
int MPI_Comm_remote_group(MPI_Comm comm, MPI_Group *group)
```

Fortran:

```
MPI_Comm_remote_group(comm, group, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Group), INTENT(OUT) :: group
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
Intercomm.Get_remote_group(self)
```

How to read routine prototypes: [1.5.4](#).

[manpage 85: Routine prototype for MPI_Comm::remote_group](#)

MPI_Intercomm_merge

Synopsis:

```
int MPI_Intercomm_merge
    (MPI_Comm intercomm, int high,
     MPI_Comm *newintracomm)
```

Input Parameters:

intercomm : Intercommunicator (handle)
high : Used to order the groups within comm (logical) when creating the new
communicator. This is a boolean value; the group that sets high true
has its processes ordered after the group that sets this value to
false. If all processes in the intercommunicator provide the same
value, the choice of which group is ordered first is arbitrary.

Output Parameters:

```
newintracomm : Created intracommunicator (handle)
```

How to read routine prototypes: [1.5.4](#).

[manpage 86: Routine prototype for MPI_Intercomm::merge](#)

6.7 Sources used in this chapter

Listing of code examples/mpi/c/wrongcatchlib.cxx:

Listing of code examples/mpi/c/rightcatchlib.c:

Listing of code XX:

Listing of code XX:

Chapter 7

MPI topic: Process management

In this course we have up to now only considered the SPMD model of running MPI programs. In some rare cases you may want to run in an MPMD mode, rather than SPMD. This can be achieved either on the OS level, using options of the *mpiexec* mechanism, or you can use MPI's built-in process management. Read on if you're interested in the latter.

7.1 Process spawning

The first version of MPI did not contain any process management routines, even though the earlier *PVM* project did have that functionality. Process management was later added with MPI-2.

Unlike what you might think, newly added processes do not become part of `MPI_COMM_WORLD`; rather, they get their own communicator, and an *inter-communicator* (section 6.6) is established between this new group and the existing one. The first routine is `MPI_Comm_spawn` (figure 87), which tries to fire up multiple copies of a single named executable. Errors in starting up these codes are returned in an array of integers, or if you're feeling sure of yourself, specify `MPI_ERRCODES_IGNORE`.

It is not immediately clear whether there is opportunity for spawning new executables; after all, `MPI_COMM_WORLD` contains all your available processors. You can probably tell your job starter to reserve space for a few extra processes, but that is installation-dependent (see below). However, there is a standard mechanism for querying whether such space has been reserved. The attribute `MPI_UNIVERSE_SIZE`, retrieved with `MPI_Attr_get` (section 12.5.3), will tell you to the total number of hosts available.

If this option is not supported, you can determine yourself how many processes you want to spawn. If you exceed the hardware resources, your multi-tasking operating system (which is some variant of Unix for almost everyone) will use *time-slicing* to start the spawned processes, but you will not gain any performance.

Here is an example of a work manager.

```
// spawn_manager.c
MPI_Comm_size(MPI_COMM_WORLD, &world_size);
MPI_Comm_rank(MPI_COMM_WORLD, &manager_rank);

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

MPI_Comm_spawn

Semantics:

```
MPI_COMM_SPAWN(command, argv, maxprocs, info, root, comm,
                intercomm, array_of_errcodes)
```

IN command: name of program to be spawned
 (string, significant only at root)
 IN argv: arguments to command
 (array of strings, significant only at root)
 IN maxprocs: maximum number of processes to start
 (integer, significant only at root)
 IN info: a set of key-value pairs telling the runtime system where and
 how to start the processes (handle, significant only at root)
 IN root: rank of process in which previous arguments are examined
 (integer)
 IN comm: intracommunicator containing group of spawning processes
 (handle)
 OUT intercomm: intercommunicator between original group and the
 newly spawned group (handle)
 OUT array_of_errcodes: one code per process (array of integer)

C:

```
int MPI_Comm_spawn(const char *command, char *argv[], int maxprocs,
                    MPI_Info info, int root, MPI_Comm comm,
                    MPI_Comm *intercomm, int array_of_errcodes[])
```

Fortran:

```
MPI_Comm_spawn(command, argv, maxprocs, info, root, comm, intercomm,
               array_of_errcodes, ierror)
CHARACTER(LEN=*) , INTENT(IN) :: command, argv(*)
INTEGER, INTENT(IN) :: maxprocs, root
TYPE(MPI_Info), INTENT(IN) :: info
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(MPI_Comm), INTENT(OUT) :: intercomm
INTEGER :: array_of_errcodes(*)
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
MPI.Intracomm.Spawn(self,
                      command, args=None, int maxprocs=1, Info info=INFO_NULL,
                      int root=0, errcodes=None)
returns an intracommunicator
```

How to read routine prototypes: [1.5.4](#).

manpage 87: Routine prototype for MPI`Comm`spawn

7. MPI topic: Process management

```
if (!flag) {
    if (manager_rank==0) {
        printf("This MPI does not support UNIVERSE_SIZE.\nHow many processes
total?");
        scanf("%d", &universe_size);
    }
    MPI_Bcast(&universe_size, 1, MPI_INTEGER, 0, MPI_COMM_WORLD);
} else {
    universe_size = *universe_sizep;
    if (manager_rank==0)
        printf("Universe size deduced as %d\n", universe_size);
}
ASSERT(universe_size>world_size, "No room to start workers");
int nworkers = universe_size-world_size;
const char *worker_program = "spawn_worker";
int errorcodes[nworkers];
MPI_Comm everyone; /* intercommunicator */
MPI_Comm_spawn(worker_program, MPI_ARGV_NULL, nworkers,
                MPI_INFO_NULL, 0, MPI_COMM_WORLD, &everyone,
                errorcodes);

## spawn_manager.py
try :
    universe_size = comm.Get_attr(MPI.UNIVERSE_SIZE)
    if universe_size is None:
        print("Universe query returned None")
        universe_size = nprocs + 4
    else:
        print("World has {} ranks in a universe of {}".format(nprocs,
                                                               universe_size))
except :
    print("Exception querying universe size")
    universe_size = nprocs + 4
nworkers = universe_size - nprocs

itercomm = comm.Spawn("./spawn_worker.py", maxprocs=nworkers)
```

You could start up a single copy of this program with

```
mpirun -np 1 spawn_manager
```

but with a hostfile that has more than one host.

TACC note. Intel mpi requires you to pass an option `-usize` to `mpiexec` indicating the size of the comm universe. With the TACC jobs starter `ibrun` do the following:

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

The spawned program looks very much like a regular MPI program, with its own initialization and finalize calls.

```
// spawn_worker.c
```

```
|| MPI_Comm_size(MPI_COMM_WORLD, &nworkers);
|| MPI_Comm_rank(MPI_COMM_WORLD, &workerno);
|| MPI_Comm_get_parent(&parent);
|| ASSERTm(parent!=MPI_COMM_NULL, "No parent!");

|| MPI_Comm_remote_size(parent, &remotesize);
|| if (workerno==0) {
||     printf("Deducing %d workers and %d parents\n", nworkers, remotesize);
|| }
|| // ASSERTm(nworkers==size-1, "nworkers mismatch. probably misunderstanding");

|| ## spawn_worker.py
|| parentcomm = comm.Get_parent()
|| nparents = parentcomm.Get_remote_size()
```

Spawned processes wind up with a value of `MPI_COMM_WORLD` of their own, but managers and workers can find each other regardless. The spawn routine returns the intercommunicator to the parent; the children can find it through `MPI_Comm_get_parent`. The number of spawning processes can be found through `MPI_Comm_remote_size` (figure 88) on the parent communicator.

7.1.1 MPMD

Instead of spawning a single executable, you can spawn multiple with `MPI_Comm_spawn_multiple`.

7.2 Socket-style communications

It is possible to establish connections with running MPI programs that have their own world communicator.

- The *server* process establishes a port with `MPI_Open_port`, and calls `MPI_Comm_accept` to accept connections to its port.
- The *client* process specifies that port in an `MPI_Comm_connect` call. This establishes the connection.

7.2.1 Server calls

The server calls `MPI_Open_port` (figure 89), yielding a port name. Port names are generated by the system and copied into a character buffer of length at most `MPI_MAX_PORT_NAME`.

The server then needs to call `MPI_Comm_accept` (figure 90) prior to the client doing a connect call. This is collective over the calling communicator. It returns an intercommunicator that allows communication with the client.

The port can be closed with `MPI_Close_port`.

MPI_Comm_remote_size

Semantics:

`MPI_COMM_REMOTE_SIZE(comm, size)`
IN `comm`: inter-communicator (handle)
OUT `size`: number of processes in the remote group of `comm` (integer)

C:

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

Fortran:

```
MPI_Comm_remote_size(comm, size, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, INTENT(OUT) :: size
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
Intercomm.Get_remote_size(self)
```

How to read routine prototypes: [1.5.4](#).

manpage 88: Routine prototype for MPI_Comm::remote_size

MPI_Open_port

C:

```
#include <mpi.h>
int MPI_Open_port(MPI_Info info, char *port_name)
```

Input parameters:

`info` : Options on how to establish an address (handle). No options currently supported.

Output parameters:

`port_name` : Newly established port (string).

How to read routine prototypes: [1.5.4](#).

manpage 89: Routine prototype for MPI_Open_port

MPI_Comm_accept

Synopsis:

```
int MPI_Comm_accept
      (const char *port_name, MPI_Info info, int root,
       MPI_Comm comm, MPI_Comm *newcomm)
```

Input parameters:

`port_name` : Port name (string, used only on root).

`info` : Options given by root for the accept (handle, used only on root). No options currently supported.

`root` : Rank in `comm` of root node (integer).

`comm` : Intracommunicator over which call is collective (handle).

Output parameters:

`newcomm` : Intercommunicator with client as remote group (handle)

How to read routine prototypes: [1.5.4](#).

manpage 90: Routine prototype for MPI_Comm::accept

7.2.2 Client calls

After the server has generated a port name, the client needs to connect to it with `MPI_Comm_connect` (figure 91), again specifying the port through a character buffer.

If the named port does not exist (or has been closed), `MPI_Comm_connect` raises an error of class `MPI_ERR_PORT`.

The client can sever the connection with `MPI_Comm_disconnect`

The connect call is collective over its communicator.

7.2.3 Published service names

`MPI_Publish_name` (figure 92)

`MPI_Unpublish_name`

Unpublishing a non-existing or already unpublished service gives an error code of `MPI_ERR_SERVICE`.

`MPI_Comm_join`

MPI provides no guarantee of fairness in servicing connection attempts. That is, connection attempts are not necessarily satisfied in the order in which they were initiated, and competition from other connection attempts may prevent a particular connection attempt from being satisfied.

7.3 Sources used in this chapter

Listing of code examples/mpi/c:

Listing of code XX:

MPI_Comm_connect

Synopsis

```
int MPI_Comm_connect
    (const char *port_name, MPI_Info info, int root,
     MPI_Comm comm, MPI_Comm * newcomm)
```

Input Parameters

port_name : network address (string, used only on root)
info : implementation-dependent information (handle, used only on root)
root : rank in comm of root node (integer)
comm : intracommunicator over which call is collective (handle)

Output Parameters

newcomm : intercommunicator with server as remote group (handle)

How to read routine prototypes: 1.5.4.

manpage 91: Routine prototype for MPI_Comm_connect

MPI_Publish_name

Synopsis:

```
MPI_Publish_name(service_name, info, port_name)
```

Input parameters:

service_name : a service name to associate with the port (string)
info : implementation-specific information (handle)
port_name : a port name (string)

C:

```
int MPI_Publish_name
    (char *service_name, MPI_Info info, char *port_name)
```

Fortran77:

```
MPI_PUBLISH_NAME(SERVICE_NAME, INFO, PORT_NAME, IERROR)
INTEGER INFO, IERROR
CHARACTER*(*) SERVICE_NAME, PORT_NAME
```

How to read routine prototypes: 1.5.4.

manpage 92: Routine prototype for MPI_Publish_name

Chapter 8

MPI topic: One-sided communication

Above, you saw point-to-point operations of the two-sided type: they require the co-operation of a sender and receiver. This co-operation could be loose: you can post a receive with `MPI_ANY_SOURCE` as sender, but there had to be both a send and receive call. In this section, you will see one-sided communication routines where a process can do a ‘put’ or ‘get’ operation, writing data to or reading it from another processor, without that other processor’s involvement.

In one-sided MPI operations, also known as Remote Direct Memory Access (RDMA) or Remote Memory Access (RMA) operations, there are still two processes involved: the *origin*, which is the process that originates the transfer, whether this is a ‘put’ or a ‘get’, and the *target* whose memory is being accessed. Unlike with two-sided operations, the target does not perform an action that is the counterpart of the action on the origin.

That does not mean that the origin can access arbitrary data on the target at arbitrary times. First of all, one-sided communication in MPI is limited to accessing only a specifically declared memory area on the target: the target declares an area of user-space memory that is accessible to other processes. This is known as a *window*. Windows limit how origin processes can access the target’s memory: you can only ‘get’ data from a window or ‘put’ it into a window; all the other memory is not reachable from other processes.

The alternative to having windows is to use *distributed shared memory* or *virtual shared memory*: memory is distributed but acts as if it shared. The so-called Partitioned Global Address Space (PGAS) languages such as Unified Parallel C (UPC) use this model. The MPI RMA model makes it possible to lock a window which makes programming slightly more cumbersome, but the implementation more efficient.

Within one-sided communication, MPI has two modes: active RMA and passive RMA. In *active RMA*, or *active target synchronization*, the target sets boundaries on the time period (the ‘epoch’) during which its window can be accessed. The main advantage of this mode is that the origin program can perform many small transfers, which are aggregated behind the scenes. Active RMA acts much like asynchronous transfer with a concluding `MPI_Waitall`.

In *passive RMA*, or *passive target synchronization*, the target process puts no limitation on when its window can be accessed. (PGAS languages such as UPC are based on this model: data is simply read or written at will.) While intuitively it is attractive to be able to write to and read from a target at arbitrary time, there are problems. For instance, it requires a remote agent on the target, which may interfere with execution of

the main thread, or conversely it may not be activated at the optimal time. Passive RMA is also very hard to debug and can lead to strange deadlocks.

8.1 Windows

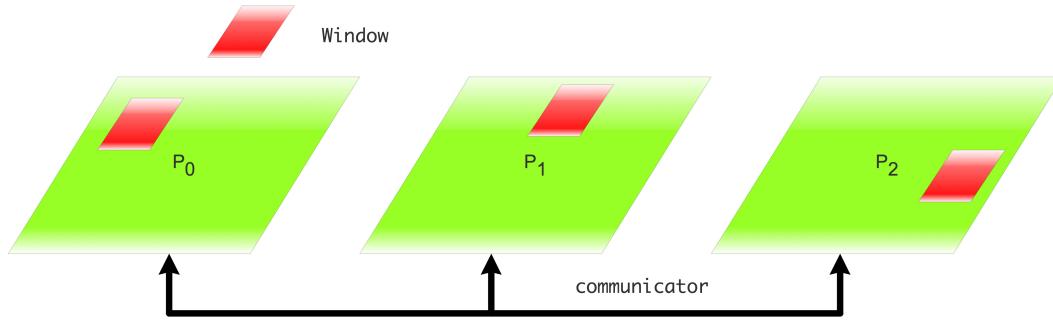


Figure 8.1: Collective definition of a window for one-sided data access

In one-sided communication, each processor can make an area of memory, called a *window*, available to one-sided transfers. This is stored in a variable of type `MPI_Win` (figure 93). A process can put an arbitrary item from its own memory to the window of another process, or get something from the other process' window in its own memory.

A window can be characterized as follows:

- The window is defined on a communicator, so the create call is collective; see figure 8.1.
- The window size can be set individually on each process. A zero size is allowed, but since window creation is collective, it is not possible to skip the create call.
- The datatype can also be set individually on each process. This makes it possible to use a derived type on one process, for instance for copying strided data into a contiguous buffer.
- The window is the target of data in a put operation, or the source of data in a get operation; see figure 8.2.
- There can be memory associated with a window, so it needs to be freed explicitly.

The typical calls involved are:

```

MPI_Info info;
MPI_Win window;
MPI_Win_allocate( /* size info */, info, comm, &memory, &window );
// do put and get calls
MPI_Win_free( &window );

```

8.1.1 Window creation and allocation

The memory for a window is at first sight ordinary data in user space. There are multiple ways you can associate data with a window:

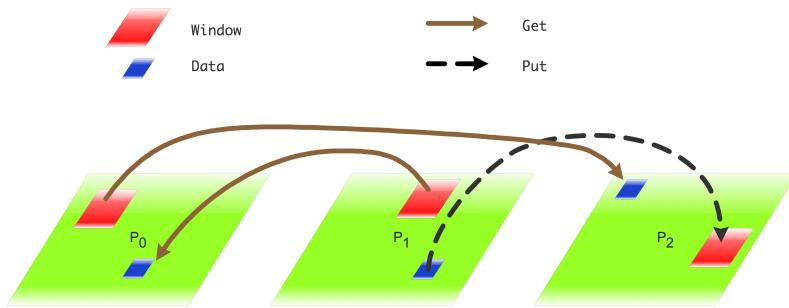


Figure 8.2: Put and get between process memory and windows

1. You can pass a user buffer to `MPI_Win_create`. This buffer can be an ordinary array, or it can be created with `MPI_Alloc_mem`.
2. You can let MPI do the allocation, so that MPI can perform various optimizations regarding placement of the memory. The user code then receives the pointer to the data from MPI. This can again be done in two ways:
 - Use `MPI_Win_allocate` to create the data and the window in one call.
 - If a communicator is on a shared memory (see section 11.1) you can create a window in that shared memory with `MPI_Win_allocate_shared`.
3. Finally, you can create a window with `MPI_Win_create_dynamic` which postpones the allocation; see section 8.5.2.

First of all, `MPI_Win_create` (figure 94) creates a window from a pointer to memory. The data array must not be `PARAMETER` or `static const`.

The size parameter is measured in bytes. In C this is easily done with the `sizeof` operator; for doing this calculation in Fortran, see section 12.3.1.

Next, one can obtain the memory from MPI by using `MPI_Win_allocate` (figure 95), which has the data pointer as output. Note the `void*` in the C prototype; it is still necessary to pass a pointer to a pointer:

```
double *window_data;
MPI_Win_allocate( ... &window_data ... );
```

The routine `MPI_Alloc_mem` (figure 96) performs only the allocation part of `MPI_Win_allocate`, after which you need to `MPI_Win_create`:

This memory is freed with

```
MPI_Free_mem()
```

These calls reduce to `malloc` and `free` if there is no special memory area; SGI is an example where such memory does exist.

There will be more discussion of window memory in section 8.5.2.

MPI_Win

```
C:  
MPI_Win win ;  
  
Fortran:  
Type(MPI_Win) win
```

How to read routine prototypes: [1.5.4](#).

manpage 93: Routine prototype for MPI_Win

MPI_Win_create

```
C:  
int MPI_Win_create  
    (void *base, MPI_Aint size, int disp_unit,  
     MPI_Info info, MPI_Comm comm, MPI_Win *win)  
  
Fortran:  
MPI_Win_create(base, size, disp_unit, info, comm, win, ierror)  
TYPE(*), DIMENSION(..), ASYNCHRONOUS :: base  
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(IN) :: size  
INTEGER, INTENT(IN) :: disp_unit  
TYPE(MPI_Info), INTENT(IN) :: info  
TYPE(MPI_Comm), INTENT(IN) :: comm  
TYPE(MPI_Win), INTENT(OUT) :: win  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python:  
MPI.Win.Create  
    (memory, int disp_unit=1,  
     Info info=INFO_NULL, Intracomm comm=COMM_SELF)
```

How to read routine prototypes: [1.5.4](#).

manpage 94: Routine prototype for MPI_Win_create

MPI_Win_allocate

Semantics:

```
MPI_WIN_ALLOCATE(size, disp_unit, info, comm, baseptr, win)
```

Input parameters:

size: size of local window in bytes (non-negative integer)
 disp_unit local unit size for displacements, in bytes (positive integer)

info: info argument (handle)

comm: intra-communicator (handle)

Output parameters:

baseptr: address of local allocated window segment (choice)

win: window object returned by the call (handle)

C:

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

Fortran:

```
MPI_Win_allocate
  (size, disp_unit, info, comm, baseptr, win, ierror)
USE, INTRINSIC :: ISO_C_BINDING, ONLY : C_PTR
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(IN) :: size
INTEGER, INTENT(IN) :: disp_unit
TYPE(MPI_Info), INTENT(IN) :: info
TYPE(MPI_Comm), INTENT(IN) :: comm
TYPE(C_PTR), INTENT(OUT) :: baseptr
TYPE(MPI_Win), INTENT(OUT) :: win
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

[manpage 95: Routine prototype for MPI_Win_allocate](#)

MPI_Alloc_mem

```
int MPI_Alloc_mem(MPI_Aint size, MPI_Info info, void *baseptr)
```

How to read routine prototypes: [1.5.4](#).

[manpage 96: Routine prototype for MPI_Alloc_mem](#)

8.2 Active target synchronization: epochs

One-sided communication has an obvious complication over two-sided: if you do a put call instead of a send, how does the recipient know that the data is there? This process of letting the target know the state of affairs is called ‘synchronization’, and there are various mechanisms for it. First of all we will consider *active target synchronization*. Here the target knows when the transfer may happen (the *communication epoch*), but does not do any data-related calls.

In this section we look at the first mechanism, which is to use a *fence* operation: `MPI_Win_fence` (figure 97). This operation is collective on the communicator of the window. It is comparable to `MPI_Wait` calls for non-blocking communication.

The use of fences is somewhat complicated. The interval between two fences is known as an *epoch*. You can give various hints to the system about this epoch versus the ones before and after through the `assert` parameter.

```
|| MPI_Win_fence( (MPI_MODE_NOPUT | MPI_MODE_NOPRECEDE), win);
|| MPI_Get( /* operands */, win);
|| MPI_Win_fence(MPI_MODE_NOSUCCEED, win);
```

In between the two fences the window is exposed, and while it is you should not access it locally. If you absolutely need to access it locally, you can use an RMA operation for that. Also, there can be only one remote process that does a put; multiple accumulate accesses are allowed.

Fences are, together with other window calls, collective operations. That means they imply some amount of synchronization between processes. Consider:

```
|| MPI_Win_fence( ... win ... ); // start an epoch
if (mytid==0) // do lots of work
else // do almost nothing
MPI_Win_fence( ... win ... ); // end the epoch
```

and assume that all processes execute the first fence more or less at the same time. The zero process does work before it can do the second fence call, but all other processes can call it immediately. However, they can not finish that second fence call until all one-sided communication is finished, which means they wait for the zero process.

As a further restriction, you can not mix `MPI_Get` with `MPI_Put` or `MPI_Accumulate` calls in a single epoch. Hence, we can characterize an epoch as an *access epoch* on the origin, and as an *exposure epoch* on the target.

Assertions are an integer parameter: you can combine assertions by adding them or using logical-or. The value zero is always correct. For further information, see section 8.3.7.

8.3 Put, get, accumulate

We will now look at the first three routines for doing one-sided operations: the Put, Get, and Accumulate call. (We will look at so-called ‘atomic’ operations in section 8.3.9.) These calls are somewhat similar to a

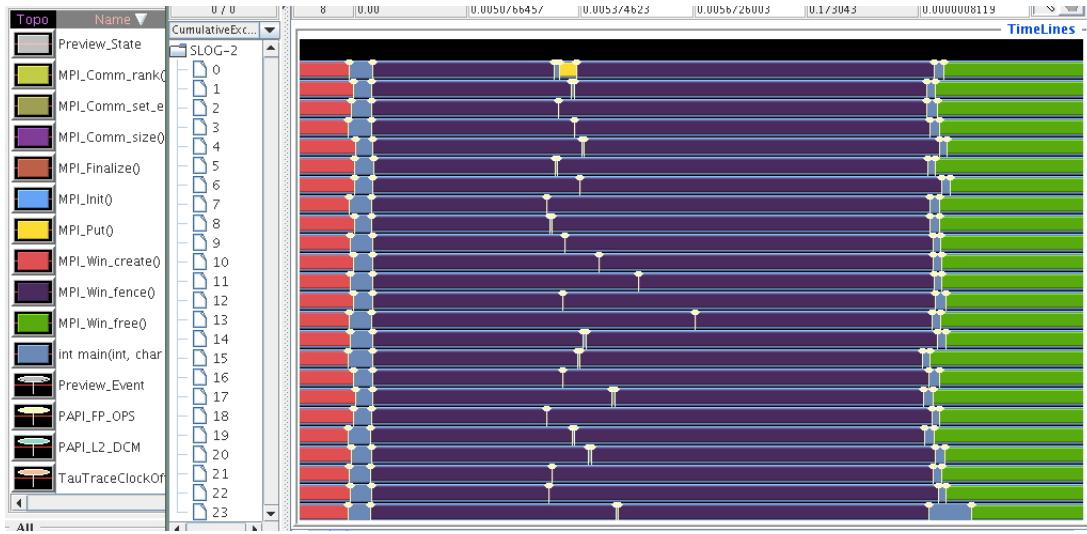


Figure 8.3: A trace of a one-sided communication epoch where process zero only originates a one-sided transfer

Send, Receive and Reduce, except that of course only one process makes a call. Since one process does all the work, its calling sequence contains both a description of the data on the origin (the calling process) and the target (the affected other process).

As in the two-sided case, `MPI_PROC_NULL` can be used as a target rank.

8.3.1 Put

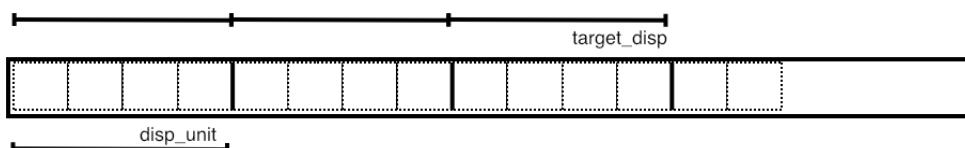
The `MPI_Put` (figure 98) call can be considered as a one-sided send. As such, it needs to specify

- the target rank
- the data to be sent from the origin, and
- the location where it is to be written on the target.

The description of the data on the origin is the usual trio of buffer/count/datatype. However, the description of the data on the target is more complicated. It has a count and a datatype, but instead of an address it has a *displacement unit* with respect to the start of the window on the target. This displacement can be given in bytes, so its type is `MPI_Aint`, but strictly speaking it is a multiple of the displacement unit that was specified in the window definition.

Specifically, data is written starting at

$$\text{window_base} + \text{target_disp} \times \text{disp_unit}.$$



Fortran note. The disp_unit variable is declared as

```
// integer(kind=MPI_ADDRESS_KIND) :: displacement
```

Specifying a literal constant, such as 0, can lead to bizarre runtime errors.

Here is a single put operation. Note that the window create and window fence calls are collective, so they have to be performed on all processors of the communicator that was used in the create call.

```
// putblock.c
MPI_Win_create(&other_number, 1, sizeof(int),
               MPI_INFO_NULL, comm, &the_window);
MPI_Win_fence(0, the_window);
if (mytid==0) {
    MPI_Put( /* data on origin: */   &my_number, 1, MPI_INT,
            /* data on target: */   1, 0,           1, MPI_INT,
            the_window);
    sleep(.5);
}
MPI_Win_fence(0, the_window);
if (mytid==1)
    printf("I got the following: %d\n", other_number);
```

Exercise 8.1. Revisit exercise 4.7 and solve it using **MPI_Put**.

Exercise 8.2. Write code where process 0 randomly writes in the window on 1 or 2.

8.3.2 Get

The **MPI_Get** (figure 99) call is very similar.

Example:

```
// getfence.c
MPI_Win_create(&other_number, 2*sizeof(int), sizeof(int),
               MPI_INFO_NULL, comm, &the_window);
MPI_Win_fence(0, the_window);
if (procno==0) {
    MPI_Get( /* data on origin: */   &my_number, 1, MPI_INT,
            /* data on target: */   other, 1,           1, MPI_INT,
            the_window);
}
MPI_Win_fence(0, the_window);
```

We make a null window on processes that do not participate.

```
## getfence.py
if procid==0 or procid==nprocs-1:
    win_mem = np.empty( 1, dtype=np.float64 )
    win = MPI.Win.Create( win_mem, comm=comm )
else:
    win = MPI.Win.Create( None, comm=comm )

# put data on another process
win.Fence()
```

```

||| if procid==0 or procid==nprocs-1:
    putdata = np.empty( 1,dtype=np.float64 )
    putdata[0] = mydata
    print("[%d] putting %e" % (procid,mydata))
    win.Put( putdata,other )
    win.Fence()

```

8.3.3 Put and get example: halo update

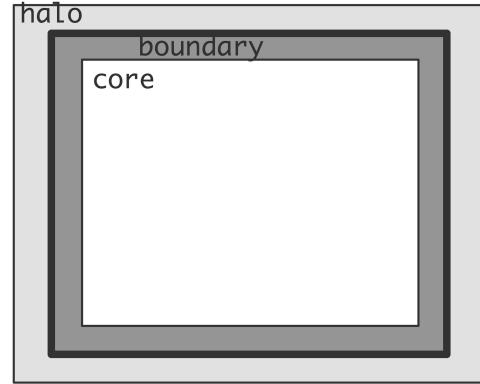
As an example, let's look at *halo update*. The array A is updated using the local values and the halo that comes from bordering processors, either through Put or Get operations.

In a first version we separate computation and communication. Each iteration has two fences. Between the two fences in the loop body we do the `MPI_Put` operation; between the second and and first one of the next iteration there is only computation, so we add the `NOPRECEDE` and `NOSUCCEED` assertions. (For much more about assertions, see section 8.3.7 below.) The `NOSTORE` assertion states that the local window was not updated: the Put operation only works on remote windows.

```

||| for ( .... ) {
    update(A);
    MPI_Win_fence(MPI_MODE_NOPRECEDE, win);
    for(i=0; i < toneighbors; i++)
        MPI_Put( ... );
    MPI_Win_fence((MPI_MODE_NOSTORE | MPI_MODE_NOSUCCEED), win);
}

```



Next, we split the update in the core part, which can be done purely from local values, and the boundary, which needs local and halo values. Update of the core can overlap the communication of the halo.

```

||| for ( .... ) {
    update_boundary(A);
    MPI_Win_fence((MPI_MODE_NOPUT | MPI_MODE_NOPRECEDE), win);
    for(i=0; i < fromneighbors; i++)
        MPI_Get( ... );
    update_core(A);
    MPI_Win_fence(MPI_MODE_NOSUCCEED, win);
}

```

The `NOPRECEDE` and `NOSUCCEED` assertions still hold, but the `Get` operation implies that instead of `NOSTORE` in the second fence, we use `NOPUT` in the first.

8.3.4 Accumulate

A third one-sided routine is `MPI_Accumulate` (figure 100) which does a reduction operation on the results that are being put.

Accumulate is a reduction with remote result. As with `MPI_Reduce`, the order in which the operands are accumulated is undefined. The same predefined operators are available, but no user-defined ones. There is one extra operator: `MPI_REPLACE`, this has the effect that only the last result to arrive is retained.

Exercise 8.3. Implement an ‘all-gather’ operation using one-sided communication: each processor stores a single number, and you want each processor to build up an array that contains the values from all processors. Note that you do not need a special case for a processor collecting its own value: doing ‘communication’ between a processor and itself is perfectly legal.

Exercise 8.4.

Implement a shared counter:

- One process maintains a counter;
- Iterate: all others at random moments update this counter.
- When the counter is no longer positive, everyone stops iterating.

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

8.3.5 Ordering and coherence of RMA operations

There are few guarantees about what happens inside one epoch.

- No ordering of Get and Put/Accumulate operations: if you do both, there is no guarantee whether the Get will find the value before or after the update.
- No ordering of multiple Puts. It is safer to do an Accumulate.

The following operations are well-defined inside one epoch:

- Instead of multiple Put operations, use Accumulate with `MPI_REPLACE`.
- `MPI_Get_accumulate` with `MPI_NO_OP` is safe.
- Multiple Accumulate operations from one origin are done in program order by default. To allow reordering, for instance to have all reads happen after all writes, use the info parameter when the window is created; section 8.5.4.

8.3.6 Request-based operations

Analogous to `MPI_Isend` there are request based one-sided operations: `MPI_Rput` (figure 101) and similarly `MPI_Rget` and `MPI_Raccumulate` and `MPI_Rget_accumulate`.

These only apply to passive target synchronization. Any `MPI_Win_flush...` call also terminates these transfers.

8.3.7 Assertions

The `MPI_Win_fence` call, as well `MPI_Win_start` and such, take an argument through which assertions can be passed about the activity before, after, and during the epoch. The value zero is always allowed, by you can make your program more efficient by specifying one or more of the following, combined by bitwise OR in C/C++ or IOR in Fortran.

- `MPI_Win_start` Supports the option:
 - `MPI_MODE_NOCHECK` the matching calls to `MPI_Win_post` have already completed on all target processes when the call to `MPI_Win_start` is made. The nocheck option can be specified in a start call if and only if it is specified in each matching post call. This is similar to the optimization of “ready-send” that may save a handshake when the handshake is implicit in the code. (However, ready-send is matched by a regular receive, whereas both start and post must specify the nocheck option.)
- `MPI_Win_post` supports the following options:
 - `MPI_MODE_NOCHECK` the matching calls to `MPI_Win_start` have not yet occurred on any origin processes when the call to `MPI_Win_post` is made. The nocheck option can be specified by a post call if and only if it is specified by each matching start call.
 - `MPI_MODE_NOSTORE` the local window was not updated by local stores (or local get or receive calls) since last synchronization. This may avoid the need for cache synchronization at the post call.
 - `MPI_MODE_NOPUT` the local window will not be updated by put or accumulate calls after the post call, until the ensuing (wait) synchronization. This may avoid the need for cache synchronization at the wait call.
- `MPI_Win_fence` supports the following options:
 - `MPI_MODE_NOSTORE` the local window was not updated by local stores (or local get or receive calls) since last synchronization.
 - `MPI_MODE_NOPUT` the local window will not be updated by put or accumulate calls after the fence call, until the ensuing (fence) synchronization.
 - `MPI_MODE_NOPRECEDE` the fence does not complete any sequence of locally issued RMA calls. If this assertion is given by any process in the window group, then it must be given by all processes in the group.
 - `MPI_MODE_NOSUCCEED` the fence does not start any sequence of locally issued RMA calls. If the assertion is given by any process in the window group, then it must be given by all processes in the group.
- `MPI_Win_lock` supports the following option:
 - `MPI_MODE_NOCHECK` no other process holds, or will attempt to acquire a conflicting lock, while the caller holds the window lock. This is useful when mutual exclusion is achieved by other means, but the coherence operations that may be attached to the lock and unlock calls are still required.

8.3.8 More active target synchronization

The ‘fence’ mechanism (section 8.2) uses a global synchronization on the communicator of the window. As such it is good for applications where the processes are largely synchronized, but it may lead to performance

8. MPI topic: One-sided communication

inefficiencies if processors are not in step with each other. There is a mechanism that is more fine-grained, by using synchronization only on a processor group. This takes four different calls, two for starting and two for ending the epoch, separately for target and origin.

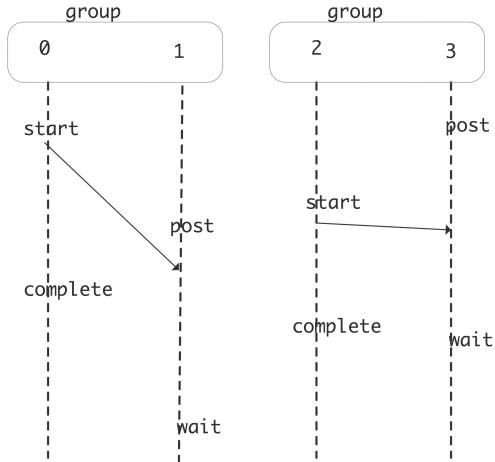


Figure 8.4: Window locking calls in fine-grained active target synchronization

You start and complete an *exposure epoch* with:

```
// int MPI_Win_post(MPI_Group group, int assert, MPI_Win win)
// int MPI_Win_wait(MPI_Win win)
```

In other words, this turns your window into the *target* for a remote access.

You start and complete an *access epoch* with:

```
// int MPI_Win_start(MPI_Group group, int assert, MPI_Win win)
// int MPI_Win_complete(MPI_Win win)
```

In other words, these calls border the access to a remote window, with the current processor being the *origin* of the remote access.

In the following snippet a single processor puts data on one other. Note that they both have their own definition of the group, and that the receiving process only does the post and wait calls.

```
// postwaitwin.c
if (procno==origin) {
    MPI_Group_incl(all_group, 1, &target, &two_group);
    // access
    MPI_Win_start(two_group, 0, the_window);
    MPI_Put( /* data on origin: */    &my_number, 1, MPI_INT,
            /* data on target: */   target, 0, 1, MPI_INT,
            the_window);
    MPI_Win_complete(the_window);
}
```

```

    if (procno==target) {
        MPI_Group_incl(all_group, 1, &origin, &two_group);
        // exposure
        MPI_Win_post(two_group, 0, the_window);
        MPI_Win_wait(the_window);
    }
}

```

Both pairs of operations declare a *group of processors*; see section 6.5.1 for how to get such a group from a communicator. On an origin processor you would specify a group that includes the targets you will interact with, on a target processor you specify a group that includes the possible origins.

8.3.9 Atomic operations

One-sided calls are said to emulate shared memory in MPI, but the put and get calls are not enough for certain scenarios with shared data. Consider the scenario where:

- 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.

The problem is that reading and updating the pointer is not an *atomic operation*, so it is possible that multiple processes get hold of the same value; conversely, multiple updates of the pointer may lead to work descriptors being skipped. These inconsistent views of the data are called a *race condition*.

In MPI 3 some atomic routines have been added. Both `MPI_Fetch_and_op` (figure 102) and `MPI_Get_accumulate` (figure 103) atomically retrieve data from the window indicated, and apply an operator, combining the data on the target with the data on the origin.

Both routines perform the same operations: return data before the operation, then atomically update data on the target, but `MPI_Get_accumulate` is more flexible in data type handling. The more simple routine, `MPI_Fetch_and_op`, which operates on only a single element, allows for faster implementations, in particular through hardware support.

Exercise 8.5. Redo exercise 8.4 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?

Example. A root process has a table of data; the other processes do atomic gets and update of that data using passive target synchronization through `MPI_Win_lock`.

```

// 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);
}

## passive.py
if procid==repository:
    # repository process creates a table of inputs
    # and associates it with the window
    win_mem = np.empty( ninputs,dtype=np.float32 )
    win = MPI.Win.Create( win_mem,comm=comm )
else:
    # everyone else has an empty window
    win = MPI.Win.Create( None,comm=comm )
if procid!=repository:
    contribution = np.empty( 1,dtype=np.float32 )
    contribution[0] = 1.*procid
    table_element = np.empty( 1,dtype=np.float32 )
    win.Lock( repository,lock_type=MPI_LOCK_EXCLUSIVE )
    win.Fetch_and_op( contribution,table_element,repository
        ,0,MPI.SUM)
    win.Unlock( repository )

```

Finally, **MPI_Compare_and_swap** (figure 104) swaps the origin and target data if the target data equals some comparison value.

8.4 Passive target synchronization

In *passive target synchronization* only the origin is actively involved: the target makes no calls whatsoever. This means that the origin process remotely locks the window on the target, performs a one-sided transfer, and releases the window by unlocking it again.

During an access epoch, also called an *passive target epoch* in this case (the concept of ‘exposure epoch’ makes no sense with passive target synchronization), a process can initiate and finish a one-sided transfer. Typically it will lock the window with **MPI_Win_lock** (figure 105) :

```

||| if (rank == 0) {
    MPI_Win_lock (MPI_LOCK_EXCLUSIVE, 1, 0, win);
    MPI_Put (outbuf, n, MPI_INT, 1, 0, n, MPI_INT, win);
    MPI_Win_unlock (1, win);
}

```

The two lock types are:

- **MPI_LOCK_SHARED** which should be used for Get calls: since multiple processors are allowed to read from a window in the same epoch, the lock can be shared.

- `MPI_LOCK_EXCLUSIVE` which should be used for Put and Accumulate calls: since only one processor is allowed to write to a window during one epoch, the lock should be exclusive.

These routines make MPI behave like a shared memory system; the instructions between locking and unlocking the window effectively become *atomic operations*.

To lock the windows of all processes in the group of the windows, use `MPI_Win_lock_all` (figure 106) .

To unlock a window, use `MPI_Win_unlock` (figure 107) and `MPI_Win_unlock_all`.

8.4.1 Completion and consistency

In one-sided transfer one should keep straight the multiple instances of the data, and the various *completions* that effect their *consistency*.

- The user data. This is the buffer that is passed to a Put or Get call. For instance, after a Put call, but still in an access epoch, the user buffer is not safe to reuse. Making sure the buffer has been transferred is called *local completion*.
- The window data. While this may be publicly accessible, it is not necessarily always consistent with internal copies.
- The remote data. Even a successful Put does not guarantee that the other process has received the data. A successful transfer is a *remote completion*.

You can force the remote completion, that is, update on the target with `MPI_Win_unlock` or some variant of it, concluding the epoch.

There is also `MPI_Win_flush` (figure 108) or `MPI_Win_flush_all`, which has to come inside the *passive target epoch*.

Local completion, again: during the epoch, is done with with `MPI_Win_flush_local` (figure 109) or `MPI_Win_flush_local_all`. After these, buffers involved in the call can be reused.

Finally, there is `MPI_Win_sync` which synchronizes private and public copies of the window.

8.4.2 Atomic shared memory operations

The above example is of limited use. Suppose processor zero has a data structure `work_table` with items that need to be processed. A counter `first_work` keeps track of the lowest numbered item that still needs processing. You can imagine the following *master-worker* scenario:

- Each process connects to the master,
- inspects the `first_work` variable,
- retrieves the corresponding work item, and
- increments the `first_work` variable.

It is important here to avoid a *race condition* (see section HPSC-??) that would result from a second process reading the `first_work` variable before the first process could have updated it. Therefore, the reading and updating needs to be an *atomic operation*.

Unfortunately, you can not have a put and get call in the same access epoch. For this reason, MPI version 3 has added certain atomic operations, such as `MPI_Fetch_and_op`.

Exercise 8.6.

- 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?)

8.5 Details

8.5.1 More about window memory

8.5.2 Memory models

You may think that the window memory is the same as the buffer you pass to `MPI_Win_create` or that you get from `MPI_Win_allocate`. This is not necessarily true, and the actual state of affairs is called the *memory model*. There are two memory models:

- Under the *unified* memory model, the buffer in process space is indeed the window memory, or at least they are kept *coherent*. This means that after *completion* of an epoch you can read the window contents from the buffer. To get this, the window needs to be created with `MPI_Win_allocate_shared`.
- Under the *separate* memory model, the buffer in process space is the *private window* and the target of put/get operations is the *public window* and the two are not the same and are not kept coherent. Under this model, you need to do an explicit get to read the window contents.

See also section 8.5.5.

8.5.3 Dynamically attached memory

In section 8.1.1 we looked at simple ways to create a window and its memory.

It is also possible to have windows where the size is dynamically set. Create a dynamic window with `MPI_Win_create_dynamic` (figure 110) and attach memory to the window with `MPI_Win_attach` (figure 111). The memory is released with `MPI_Win_detach` (figure 112).

8.5.4 Window usage hints

The following keys can be passed as info argument:

- *no_locks*: if set to true, passive target synchronization (section 8.4) will not be used on this window.
- *accumulate_ordering*: a comma-separated list of the keywords `rar`, `raw`, `war`, `waw` can be specified. This indicates that reads or writes from `MPI_Accumulate` or `MPI_Get_accumulate` can be reordered, subject to certain constraints.
- *accumulate_ops*: the value `same_op` indicates that concurrent Accumulate calls use the same operator; `same_op_no_op` indicates the same operator or `MPI_NO_OP`.

8.5.5 Window information

The `MPI_Info` parameter can be used to pass implementation-dependent information; see section 12.1.

A number of attributes are stored with a window when it is created.

Obtaining a pointer to the start of the window area:

```
|| void *base;
|| MPI_Win_get_attr(win, MPI_WIN_BASE, &base, &flag)
```

Obtaining the size and *window displacement unit*:

```
|| MPI_Aint *size;
|| MPI_Win_get_attr(win, MPI_WIN_SIZE, &size, &flag),
|| int *disp_unit;
|| MPI_Win_get_attr(win, MPI_WIN_DISP_UNIT, &disp_unit, &flag),
```

The type of create call used:

```
|| int *create_kind;
|| MPI_Win_get_attr(win, MPI_WIN_CREATE_FLAVOR, &create_kind, &flag)
```

with possible values:

- `MPI_WIN_FLAVOR_CREATE` if the window was create with `MPI_Win_create`;
- `MPI_WIN_FLAVOR_ALLOCATE` if the window was create with `MPI_Win_allocate`;
- `MPI_WIN_FLAVOR_DYNAMIC` if the window was create with `MPI_Win_create_dynamic`. In this case the base is `MPI_BOTTOM` and the size is zero;
- `MPI_WIN_FLAVOR_SHARED` if the window was create with `MPI_Win_allocate_shared`;

The window model:

```
|| int *memory_model;
|| MPI_Win_get_attr(win, MPI_WIN_MODEL, &memory_model, &flag);
```

with possible values:

- `MPI_WIN_SEPARATE`,
- `MPI_WIN_UNIFIED`,

Get the group of processes associated with a window:

```
int MPI_Win_get_group(MPI_Win win, MPI_Group *group)
MPI_Win_get_group(win, group, ierror)
TYPE(MPI_Win), INTENT(IN) :: win
TYPE(MPI_Group), INTENT(OUT) :: group
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

int MPI_Win_set_info(MPI_Win win, MPI_Info info)
MPI_Win_set_info(win, info, ierror)
TYPE(MPI_Win), INTENT(IN) :: win
TYPE(MPI_Info), INTENT(IN) :: info
```

```
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

int MPI_Win_get_info(MPI_Win win, MPI_Info *info_used)
MPI_Win_get_info(win, info_used, ierror)
TYPE(MPI_Win), INTENT(IN) :: win
TYPE(MPI_Info), INTENT(OUT) :: info_used
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

8.6 Implementation

You may wonder how one-sided communication is realized¹. Can a processor somehow get at another processor's data? Unfortunately, no.

Active target synchronization is implemented in terms of two-sided communication. Imagine that the first fence operation does nothing, unless it concludes prior one-sided operations. The Put and Get calls do nothing involving communication, except for marking with what processors they exchange data. The concluding fence is where everything happens: first a global operation determines which targets need to issue send or receive calls, then the actual sends and receive are executed.

Exercise 8.7. Assume that only Get operations are performed during an epoch. Sketch how these are translated to send/receive pairs. The problem here is how the senders find out that they need to send. Show that you can solve this with an **MPI_Reduce_scatter** call.

The previous paragraph noted that a collective operation was necessary to determine the two-sided traffic. Since collective operations induce some amount of synchronization, you may want to limit this.

Exercise 8.8. Argue that the mechanism with window post/wait/start/complete operations still needs a collective, but that this is less burdensome.

Passive target synchronization needs another mechanism entirely. Here the target process needs to have a background task (process, thread, daemon,...) running that listens for requests to lock the window. This can potentially be expensive.

8.7 Sources used in this chapter

Listing of code examples/mpi/c/putblock.c:

Listing of code examples/mpi/c/getfence.c:

Listing of code examples/mpi/c/postwaittwo.c:

Listing of code examples/mpi/c/fetchop.c:

1. For more on this subject, see [14].

MPI_Win_fence

Semantics:

MPI_WIN_FENCE(assert, win)
IN assert: program assertion (integer)
IN win: window object (handle)

C:

int MPI_Win_fence(int assert, MPI_Win win)

F:

MPI_Win_fence(assert, win, ierror)
INTEGER, INTENT(IN) :: assert
TYPE(MPI_Win), INTENT(IN) :: win
INTEGER, OPTIONAL, INTENT(OUT) :: ierror

Python:

win.Fence(self, int assertion=0)

How to read routine prototypes: [1.5.4.](#)

manpage 97: Routine prototype for MPI_Win_fence

MPI_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
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
win.Put(self, origin, int target_rank, target=None)
```

How to read routine prototypes: [1.5.4](#).

manpage 98: Routine prototype for MPI_Put

MPI_Get

```
C:  
int MPI_Get(  
    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_Get(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  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror  
  
Python:  
win.Get(self, origin, int target_rank, target=None)
```

How to read routine prototypes: [1.5.4](#).

manpage 99: Routine prototype for MPI`Get

8. MPI topic: One-sided communication

MPI_Accumulate

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 (nonnegative).
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

```
MPI_Raccumulate: RMA request
IERROR (Fortran only): Error status (integer).
```

Fortran:

```
MPI_ACCUMULATE
    (ORIGIN_ADDR, ORIGIN_COUNT, ORIGIN_DATATYPE,
     TARGET_RANK,TARGET_DISP, TARGET_COUNT, TARGET_DATATYPE,
     OP, WIN, IERROR)
<type> ORIGIN_ADDR(*)
INTEGER(KIND=MPI_ADDRESS_KIND) :: TARGET_DISP
INTEGER :: ORIGIN_COUNT, ORIGIN_DATATYPE,
            TARGET_RANK, TARGET_COUNT,TARGET_DATATYPE,
            OP, WIN, IERROR
MPI_RACCUMULATE
    (ORIGIN_ADDR, ORIGIN_COUNT, ORIGIN_DATATYPE,
     TARGET_RANK,TARGET_DISP, TARGET_COUNT, TARGET_DATATYPE,
     OP, WIN, REQUEST, IERROR)
<type> ORIGIN_ADDR(*)
INTEGER(KIND=MPI_ADDRESS_KIND) :: TARGET_DISP
INTEGER :: ORIGIN_COUNT, ORIGIN_DATATYPE, TARGET_RANK, TARGET_COUNT, TARGET_DATATYPE,
            OP, WIN, REQUEST, IERROR
```

Python:

```
MPI.Win.Accumulate(self, origin, int target_rank, target=None, Op op=SUM)
```

How to read routine prototypes: [1.5.4](#).

manpage 100: Routine prototype for MPI_Accumulate

MPI_Rput

C:

```
int MPI_Rput(
    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, MPI_Request *request)
```

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)
OUT request: RMA request (handle)

How to read routine prototypes: 1.5.4.

manpage 101: Routine prototype for MPI_Rput

MPI_Fetch_and_op

Semantics:

```
MPI_FETCH_AND_OP(origin_addr, result_addr, datatype, target_rank,
                  target_disp, op, win)
IN origin_addr: initial address of buffer (choice)
OUT result_addr: initial address of result buffer (choice)
IN datatype: datatype of the entry in origin, result, and target buffers
(handle)
IN target_rank: rank of target (non-negative integer)
IN target_disp: displacement from start of window to beginning of target
buffer (non-negative integer)
IN op: reduce operation (handle)
IN win: window object (handle)
```

C:

```
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)
```

Fortran:

```
MPI_Fetch_and_op(origin_addr, result_addr, datatype, target_rank,
                  target_disp, op, win, ierror)
TYPE(*), DIMENSION(..), INTENT(IN), ASYNCHRONOUS :: origin_addr
TYPE(*), DIMENSION(..), ASYNCHRONOUS :: result_addr
TYPE(MPI_Datatype), INTENT(IN) :: datatype
INTEGER, INTENT(IN) :: target_rank
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(IN) :: target_disp
TYPE(MPI_Op), INTENT(IN) :: op
TYPE(MPI_Win), INTENT(IN) :: win
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 102: Routine prototype for MPI_Fetch_and_op

MPI_Get_accumulate

C:

```
int MPI_Get_accumulate
  (const void *origin_addr, int origin_count, MPI_Datatype origin_datatype,
   void *result_addr, int result_count, MPI_Datatype result_datatype,
   int target_rank,
   MPI_Aint target_disp, int target_count, MPI_Datatype target_datatype,
   MPI_Op op, MPI_Win win)
```

Input Parameters

origin_addr : initial address of buffer (choice)
origin_count : number of entries in buffer (nonnegative integer)
origin_datatype : datatype of each buffer entry (handle)

result_addr : initial address of result buffer (choice)
result_count : number of entries in result buffer (non-negative integer)
result_datatype : datatype of each entry in result buffer (handle)
target_rank : rank of target (nonnegative integer)
target_disp : displacement from start of window to beginning
 of target buffer (nonnegative integer)
target_count : number of entries in target buffer (nonnegative integer)
target_datatype : datatype of each entry in target buffer (handle)
op : predefined reduce operation (handle)
win : window object (handle)

How to read routine prototypes: 1.5.4.

manpage 103: Routine prototype for MPI`Get`accumulate

MPI_Compare_and_swap

C:

```
int MPI_Compare_and_swap
  (const void *origin_addr, const void *compare_addr,
   void *result_addr, MPI_Datatype datatype,
   int target_rank, MPI_Aint target_disp,
   MPI_Win win)
```

Input Parameters

origin_addr : initial address of buffer (choice)
compare_addr : initial address of compare buffer (choice)
result_addr : initial address of result buffer (choice)
datatype : datatype of the entry in origin, result, and target buffers (handle)
target_rank : rank of target (nonnegative integer)
target_disp : displacement from start of window to beginning
 of target buffer (non-negative integer)
win : window object (handle)

How to read routine prototypes: 1.5.4.

manpage 104: Routine prototype for MPI`Compare`and`swap

MPI_Win_lock

C:

```
int MPI_Win_lock(int lock_type, int rank, int assert, MPI_Win win)
```

Input Parameters:

lock_type - Indicates whether other processes may access the target window at the same time (if MPI_LOCK_SHARED) or not (MPI_LOCK_EXCLUSIVE)

rank - rank of locked window (nonnegative integer)

assert - Used to optimize this call; zero may be used as a default. (integer)

win - window object (handle)

Python:

```
MPI.Win.Lock(self,  
             int rank, int lock_type=LOCK_EXCLUSIVE, int assertion=0)
```

How to read routine prototypes: [1.5.4](#).

manpage 105: Routine prototype for MPI_Win_lock

MPI_Win_lock_all

C:

```
int MPI_Win_lock( int assert, MPI_Win win)
```

Input Parameters:

assert - Used to optimize this call; zero may be used as a default. (integer)

win - window object (handle)

How to read routine prototypes: [1.5.4](#).

manpage 106: Routine prototype for MPI_Win_lock_all

MPI_Win_unlock

C:

Py:

```
MPI.Win.Unlock(self, int rank)  
MPI.Win.Unlock_all(self)
```

How to read routine prototypes: [1.5.4](#).

manpage 107: Routine prototype for MPI_Win_unlock

MPI_Win_flush

Synopsis

`MPI_WIN_FLUSH(rank, win)`

Input arguments:

`rank` : rank of target window (non-negative integer)

`win` : window object (handle)

C:

`int MPI_Win_flush(int rank, MPI_Win win)`

Fortran:

`MPI_Win_flush(rank, win, ierror)`

`INTEGER, INTENT(IN) :: rank`

`TYPE(MPI_Win), INTENT(IN) :: win`

`INTEGER, OPTIONAL, INTENT(OUT) :: ierror`

`MPI_WIN_FLUSH(RANK, WIN, IERROR)`

`INTEGER RANK, WIN, IERROR`

Synopsis:

`MPI_WIN_FLUSH_ALL(win)`

Input arguments:

`win` : window object (handle)

C:

`int MPI_Win_flush_all(MPI_Win win)`

Fortran:

`MPI_Win_flush_all(win, ierror)`

`TYPE(MPI_Win), INTENT(IN) :: win`

`INTEGER, OPTIONAL, INTENT(OUT) :: ierror`

`MPI_WIN_FLUSH_ALL(WIN, IERROR)`

`INTEGER WIN, IERROR`

How to read routine prototypes: [1.5.4](#).

manpage 108: Routine prototype for MPI_Win_flush

MPI_Win_flush_local

Synopsis:

`MPI_WIN_FLUSH_LOCAL(rank, win)`

Input arguments:

rank: rank of target window (non-negative integer)

win : window object (handle)

C:

```
int MPI_Win_flush_local(int rank, MPI_Win win)
```

Fortran:

```
MPI_Win_flush_local(rank, win, ierror)
INTEGER, INTENT(IN) :: rank
TYPE(MPI_Win), INTENT(IN) :: win
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
MPI_WIN_FLUSH_LOCAL(RANK, WIN, IERROR)
INTEGER RANK, WIN, IERROR
```

Synopsis:

`MPI_WIN_FLUSH_LOCAL_ALL(win)`

Input arguments:

win : window object (handle)

C:

```
int MPI_Win_flush_local_all(MPI_Win win)
```

Fortran:

```
MPI_Win_flush_local_all(win, ierror)
TYPE(MPI_Win), INTENT(IN) :: win
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
MPI_WIN_FLUSH_LOCAL_ALL(WIN, IERROR)
INTEGER WIN, IERROR
```

How to read routine prototypes: [1.5.4](#).

manpage 109: Routine prototype for MPI_Win_flush_local

MPI_Win_create_dynamic

`int MPI_Win_create_dynamic(MPI_Info info, MPI_Comm comm, MPI_Win *win)`

Input Parameters

info : info argument (handle)

comm : communicator (handle)

Output Parameters

win : window object returned by the call (handle)

How to read routine prototypes: [1.5.4](#).

manpage 110: Routine prototype for MPI_Win_create_dynamic

MPI_Win_attach

Semantics:

`MPI_Win_attach(win, base, size)`

Input Parameters:

`win : window object (handle)
base : initial address of memory to be attached
size : size of memory to be attached in bytes`

C:

`int MPI_Win_attach(MPI_Win win, void *base, MPI_Aint size)`

Fortran:

`MPI_Win_attach(win, base, size, ierror)
TYPE(MPI_Win), INTENT(IN) :: win
TYPE(*), DIMENSION(..), ASYNCHRONOUS :: base
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(IN) :: size
INTEGER, OPTIONAL, INTENT(OUT) :: ierror`*How to read routine prototypes: [1.5.4](#).*

manpage 111: Routine prototype for MPI_Win_attach

MPI_Win_detach

Semantics:

`MPI_Win_detach(win, base)`

Input parameters:

`win : window object (handle)
base : initial address of memory to be detached`

C:

`int MPI_Win_detach(MPI_Win win, const void *base)`

Fortran:

`MPI_Win_detach(win, base, ierror)
TYPE(MPI_Win), INTENT(IN) :: win
TYPE(*), DIMENSION(..), ASYNCHRONOUS :: base
INTEGER, OPTIONAL, INTENT(OUT) :: ierror`*How to read routine prototypes: [1.5.4](#).*

manpage 112: Routine prototype for MPI_Win_detach

Chapter 9

MPI topic: File I/O

File input and output in parallel is a little more complicated than sequentially.

- There is nothing against every process opening an existing file for reading, and using an individual file pointer to get its unique data.
- ... but having every process open the same file for output is probably not a good idea.
- Based on the process rank it is easy enough to have every process create a unique file, but that can put a lot of strain on the file system, and it means you may have to post-process to get all the data in one file.

Wouldn't it be nice if there was a way to open one file in parallel, and have every process read from and write to its own location? That's where *MPI/O* comes in. In fact, MPI-IO is more flexible than that, since it uses MPI *derived datatypes* for both the source data (that is, in memory) and target data (that is, on disk). Thus, in one call that is collective on a communicator each process can address data that is not contiguous in memory, and place it in locations that are not contiguous on disc.

There are dedicated libraries for file I/O, such as *hdf5*, *netcdf*, or *silo*. However, these often add header information to a file that may not be understandable to post-processing applications. With MPI I/O you are in complete control of what goes to the file. (A useful tool for viewing your file is the unix utility *od*.)

9.1 File handling

MPI has its own file handle: `MPI_File` (figure 113).

You open a file with `MPI_File_open` (figure 114). This routine is collective, even if only certain processes will access the file with a read or write call. Similarly, `MPI_File_close` is collective.

Python note. Note the slightly unusual syntax for opening a file: even though the file is opened on a communicator, it is a class method for the `MPI.File` class, rather than for the communicator object. The latter is passed in as an argument.

File access modes:

- `MPI_MODE_RDONLY`: read only,
- `MPI_MODE_RDWR`: reading and writing,
- `MPI_MODE_WRONLY`: write only,

MPI_File

C:
MPI_File file ;

Fortran:
Type(MPI_File) file

How to read routine prototypes: 1.5.4.

manpage 113: Routine prototype for MPI_File

MPI_File_open

Semantics:
MPI_FILE_OPEN(comm, filename, amode, info, fh)
IN comm: communicator (handle)
IN filename: name of file to open (string)
IN amode: file access mode (integer)
IN info: info object (handle)
OUT fh: new file handle (handle)

C:
int MPI_File_open
(MPI_Comm comm, char *filename, int amode,
 MPI_Info info, MPI_File *fh)

Fortran:
MPI_FILE_OPEN(COMM, FILENAME, AMODE, INFO, FH, IERROR)
CHARACTER(*) FILENAME
INTEGER COMM, AMODE, INFO, FH, IERROR

Python:
Open(type cls, Intracomm comm, filename,
 int amode=MODE_RDONLY, Info info=INFO_NULL)

How to read routine prototypes: 1.5.4.

manpage 114: Routine prototype for MPI_File_open

- `MPI_MODE_CREATE`: create the file if it does not exist,
- `MPI_MODE_EXCL`: error if creating file that already exists,
- `MPI_MODE_DELETE_ON_CLOSE`: delete file on close,
- `MPI_MODE_UNIQUE_OPEN`: file will not be concurrently opened elsewhere,
- `MPI_MODE_SEQUENTIAL`: file will only be accessed sequentially,
- `MPI_MODE_APPEND`: set initial position of all file pointers to end of file.

These modes can be added or bitwise-or'ed.

You can delete a file with `MPI_File_delete`.

Buffers can be flushed with `MPI_File_sync`.

9.2 File reading and writing

The basic file operations, in between the open and close calls, are the POSIX-like calls

- `MPI_File_seek` (figure 115) . The `whence` parameter can be:
 - `MPI_SEEK_SET` The pointer is set to offset.
 - `MPI_SEEK_CUR` The pointer is set to the current pointer position plus offset.
 - `MPI_SEEK_END` The pointer is set to the end of the file plus offset.
- `MPI_File_read`
- `MPI_File_write`

For thread safety it is good to combine seek and read/write operations:

- `MPI_File_read_at`: combine read and seek.
- `MPI_File_write_at`: combine write and seek

Collective calls, performed by all processes in the communicator:

- `MPI_File_read_all`
- `MPI_File_write_all`
- `MPI_File_read_at_all`
- `MPI_File_write_at_all`

Using a shared file pointer the operations are:

- `MPI_File_read_shared`
- `MPI_File_write_shared`

Writing to and reading from a parallel file is rather similar to sending a receiving:

- The process uses an elementary data type or a derived datatype to describe what elements in an array go to file, or are read from file.
- In the simplest case, your read or write that data to the file using an offset, or first having done a seek operation.
- But you can also set a ‘file view’ to describe explicitly what elements in the file will be involved.

File accesses:

- `MPI_File_read_ordered`
- `MPI_File_write_ordered`

MPI_File_seek

MPI_File_seek - Updates individual file pointers (noncollective)

C:

```
#include <mpi.h>
int MPI_File_seek(MPI_File fh, MPI_Offset offset,int whence)
```

Fortran 2008:

```
USE mpi_f08
MPI_File_seek(fh, offset, whence, ierror)
  TYPE(MPI_File), INTENT(IN) :: fh
  INTEGER(KIND=MPI_OFFSET_KIND), INTENT(IN) :: offset
  INTEGER, INTENT(IN) :: whence
  INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Fortran 90:

```
USE MPI
! or the older form: INCLUDE 'mpif.h'
MPI_FILE_SEEK(FH, OFFSET, WHENCE, IERROR)
  INTEGER    FH, WHENCE, IERROR
  INTEGER(KIND=MPI_OFFSET_KIND)      OFFSET
```

Input parameters:

fh : File handle (handle).
offset : File offset (integer).
whence : Update mode (integer).

Output parameters:

IERROR : Fortran only: Error status (integer)

How to read routine prototypes: [1.5.4](#).

manpage 115: Routine prototype for MPI`File`seek

9.2.1 Individual file pointers, contiguous writes

After the collective open call, each rank holds an *individual file pointer* each rank can individually position the pointer somewhere in the shared file. Let's explore this modality.

The simplest way of writing a data to file is much like a send call: a buffer is specified with the usual count/datatype specification, and a target location in the file is given. The routine `MPI_File_write_at` (figure 116) gives this location in absolute terms with a parameter of type `MPI_Offset`, which counts bytes.

Figure 9.1: Writing at an offset



Exercise 9.1. Create a buffer of length `nwords=3` on each process, and write these buffers as a sequence to one file with `MPI_File_write_at`.

Instead of giving the position in the file explicitly, you can also use a `MPI_File_seek` call to position the file pointer, and write with `MPI_File_write` at the pointer location. The write call itself also *advances the file pointer* so separate calls for writing contiguous elements need no seek calls with `MPI_SEEK_CUR`.

Exercise 9.2. Rewrite the code of exercise 9.1 to use a loop where each iteration writes only one item to file. Note that no explicit advance of the file pointer is needed.

Exercise 9.3. Construct a file with the consecutive integers $0, \dots, WP$ where W some integer, and P the number of processes. Each process p writes the numbers $p, p + W, p + 2W, \dots$. Use a loop where each iteration

1. writes a single number with `MPI_File_write`, and
2. advanced the file pointer with `MPI_File_seek` with a *whence* parameter of `MPI_SEEK_CUR`.

9.2.2 File views

The previous mode of writing is enough for writing simple contiguous blocks in the file. However, you can also access non-contiguous areas in the file. For this you use `MPI_File_set_view` (figure 117). This call is collective, even if not all processes access the file.

- The `etyp`e describes the data type of the file, it needs to be the same on all processes.
- The `filetyp`e describes how this process sees the file, so it can differ between processes.
- The `disp` displacement parameters is measured in bytes. It can differ between processes. On sequential files such as tapes or network streams it does not make sense to set a displacement; for those the `MPI_DISPLACEMENT_CURRENT` value can be used.

MPI_File_write_at

```
MPI_File_write_at(fh,offset,buf,count,datatype)

Semantics:
Input Parameters
fh : File handle (handle).
offset : File offset (integer).
buf : Initial address of buffer (choice).
count : Number of elements in buffer (integer).
datatype : Data type of each buffer element (handle).

Output Parameters:
status : Status object (status).

C:
int MPI_File_write_at
    (MPI_File fh, MPI_Offset offset, const void *buf,
     int count, MPI_Datatype datatype, MPI_Status *status)

Fortran:
MPI_FILE_WRITE_AT
    (FH,  OFFSET,  BUF,  COUNT,  DATATYPE,  STATUS,  IERROR)
<type>    BUF(*)
INTEGER :: FH, COUNT, DATATYPE, STATUS(MPI_STATUS_SIZE), IERROR
INTEGER(KIND=MPI_OFFSET_KIND) :: OFFSET

Python:
MPI.File.Write_at(self, Offset offset, buf, Status status=None)
```

How to read routine prototypes: [1.5.4](#).

manpage 116: Routine prototype for MPI_File::write::at

MPI_File_set_view

Semantics:

```
MPI_FILE_SET_VIEW(fh, disp, etype, filetype, datarep, info)
INOUT fh: file handle (handle)
IN disp: displacement (integer)
IN etype: elementary datatype (handle)
IN filetype: filetype (handle)
IN datarep: data representation (string)
IN info: info object (handle)
```

C:

```
int MPI_File_set_view
    (MPI_File fh,
     MPI_Offset disp, MPI_Datatype etype, MPI_Datatype filetype,
     char *datarep, MPI_Info info)
```

Fortran:

```
MPI_FILE_SET_VIEW(FH, DISP, ETYPE, FILETYPE, DATAREP, INFO, IERROR)
INTEGER FH, ETYPE, FILETYPE, INFO, IERROR
CHARACTER(*) DATAREP
INTEGER(KIND=MPI_OFFSET_KIND) DISP
```

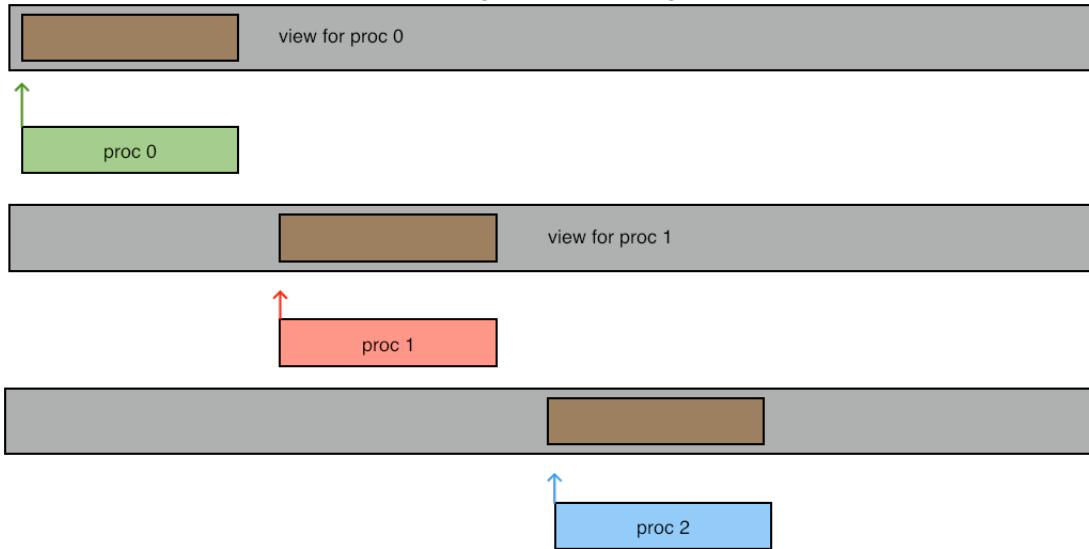
Python:

```
mpifile = MPI.File.Open( .... )
mpifile.Set_view
    (self,
     Offset disp=0, Datatype etype=None, Datatype filetype=None,
     datarep=None, Info info=INFO_NULL)
```

How to read routine prototypes: 1.5.4.

manpage 117: Routine prototype for MPI_File_set_view

Figure 9.2: Writing at a view



Exercise 9.4. Write a file in the same way as in exercise 9.1, but now use `MPI_File_write` and use `MPI_File_set_view` to set a view that determines where the data is written.

You can get very creative effects by setting the view to a derived datatype.

Fortran note. In Fortran you have to assure that the displacement parameter is of ‘kind’. In particular, you can not specify a literal zero ‘0’ as the displacement; use `0_MPI_OFFSET_KIND` instead.

More: `MPI_File_set_size` `MPI_File_get_size` `MPI_File_preeallocate` `MPI_File_get_view`

9.3 Consistency

It is possible for one process to read data previously written by another process. For this it is of course necessary to impose a temporal order, for instance by using `MPI_Barrier`, or using a zero-byte send from the writing to the reading process.

However, the file also needs to be declared atomic: `MPI_File_set_atomicity`.

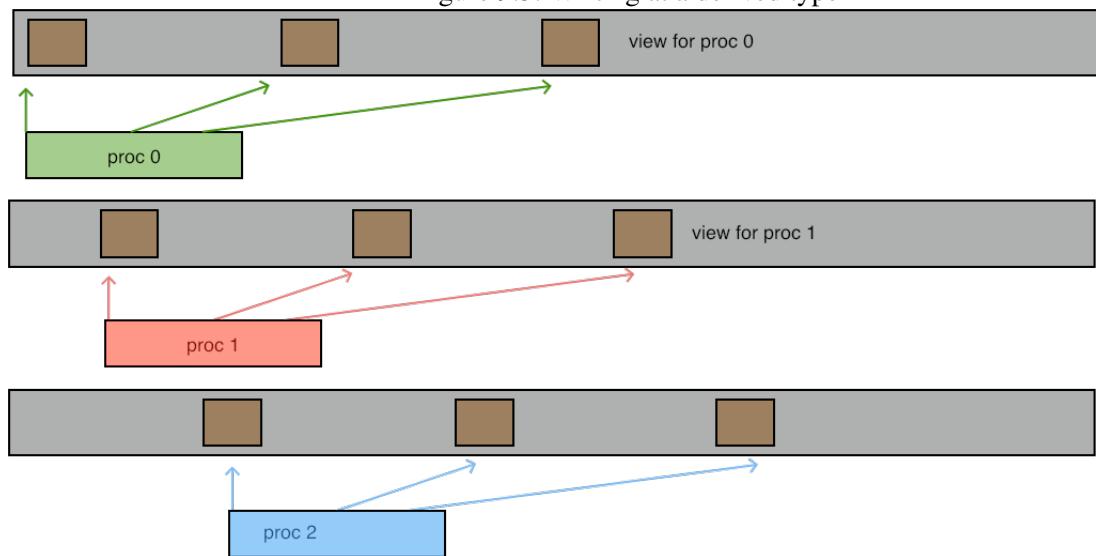
9.4 Constants

`MPI_SEEK_SET` used to be called `SEEK_SET` which gave conflicts with the C++ library. This had to be circumvented with

```
make CPPFLAGS="-DMPICH_IGNORE_CXX_SEEK -DMPICH_SKIP_MPICXX"
```

and such.

Figure 9.3: Writing at a derived type



Chapter 10

MPI topic: Topologies

A communicator describes a group of processes, but the structure of your computation may not be such that every process will communicate with every other process. For instance, in a computation that is mathematically defined on a Cartesian 2D grid, the processes themselves act as if they are two-dimensionally ordered and communicate with N/S/E/W neighbours. If MPI had this knowledge about your application, it could conceivably optimize for it, for instance by renumbering the ranks so that communicating processes are closer together physically in your cluster.

The mechanism to declare this structure of a computation to MPI is known as a *virtual topology*. The following types of topology are defined:

- `MPI_UNDEFINED`: this value holds for communicators where no topology has explicitly been specified.
- `MPI_CART`: this value holds for Cartesian topologies, where processes act as if they are ordered in a multi-dimensional ‘brick’; see section 10.1.
- `MPI_GRAPH`: this value describes the graph topology that was defined in *MPI 1*; section 10.2.4. It is unnecessarily burdensome, since each process needs to know the total graph, and should therefore be considered obsolete; the type `MPI_DIST_GRAPH` should be used instead.
- `MPI_DIST_GRAPH`: this value describes the distributed graph topology where each process only describes the edges in the process graph that touch itself; see section 10.2.

These values can be discovered with the routine `MPI_Topo_test` (figure 118) .

10.1 Cartesian grid topology

A *Cartesian grid* is a structure, typically in 2 or 3 dimensions, of points that have two neighbours in each of the dimensions. Thus, if a Cartesian grid has sizes $K \times M \times N$, its points have coordinates (k, m, n) with $0 \leq k < K$ et cetera. Most points have six neighbours $(k \pm 1, m, n), (k, m \pm 1, n), (k, m, n \pm 1)$; the exception are the edge points. A grid where edge processors are connected through *wraparound connections* is called a *periodic grid*.

The most common use of Cartesian coordinates is to find the rank of process by referring to it in grid terms. For instance, one could ask ‘what are my neighbours offset by $(1, 0, 0), (-1, 0, 0), (0, 1, 0)$ et cetera’.

While the Cartesian topology interface is fairly easy to use, as opposed to the more complicated general graph topology below, it is not actually sufficient for all Cartesian graph uses. Notably, in a so-called *star stencil*, such as the *nine-point stencil*, there are diagonal connections, which can not be described in a single step. Instead, it is necessary to take a separate step along each coordinate dimension. In higher dimensions this is of course fairly awkward.

Thus, even for Cartesian structures, it may be advisable to use the general graph topology interface.

10.1.1 Cartesian routines

The cartesian topology is specified by giving `MPI_Cart_create` the sizes of the processor grid along each axis, and whether the grid is periodic along that axis.

```
int MPI_Cart_create(
    MPI_Comm comm_old, int ndims, int *dims, int *periods,
    int reorder, MPI_Comm *comm_cart)
```

Each point in this new communicator has a coordinate and a rank. They can be queried with `MPI_Cart_coords` and `MPI_Cart_rank` respectively.

```
int MPI_Cart_coords(
    MPI_Comm comm, int rank, int maxdims,
    int *coords);
int MPI_Cart_rank(
    MPI_Comm comm, int *coords,
    int *rank);
```

Note that these routines can give the coordinates for any rank, not just for the current process.

```
// cart.c
MPI_Comm comm2d;
ndim = 2; periodic[0] = periodic[1] = 0;
dimensions[0] = idim; dimensions[1] = jdim;
MPI_Cart_create(comm, ndim, dimensions, periodic, 1, &comm2d);
MPI_Cart_coords(comm2d, procno, ndim, coord_2d);
MPI_Cart_rank(comm2d, coord_2d, &rank_2d);
printf("I am %d: (%d, %d); originally %d\n", rank_2d, coord_2d[0], coord_2d[1],
procno);
```

The `reorder` parameter to `MPI_Cart_create` indicates whether processes can have a rank in the new communicator that is different from in the old one.

Strangely enough you can only shift in one direction, you can not specify a shift vector.

```
int MPI_Cart_shift(MPI_Comm comm, int direction, int displ, int *source,
                   int *dest)
```

If you specify a processor outside the grid the result is `MPI_PROC_NULL`.

```
char mychar = 65+procno;
MPI_Cart_shift(comm2d, 0, +1, &rank_2d, &rank_right);
MPI_Cart_shift(comm2d, 0, -1, &rank_2d, &rank_left);
```

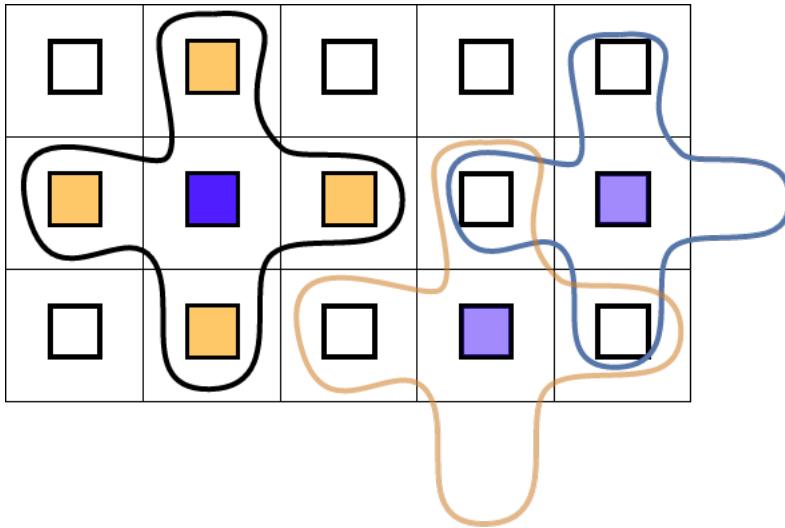


Figure 10.1: Illustration of a distributed graph topology where each node has four neighbours

```

MPI_Cart_shift(comm2d, 1, +1, &rank_2d, &rank_up);
MPI_Cart_shift(comm2d, 1, -1, &rank_2d, &rank_down);
int irequest = 0; MPI_Request *requests = malloc(8*sizeof(MPI_Request));
MPI_Isend(&mychar, 1, MPI_CHAR, rank_right, 0, comm, requests+irequest++);
MPI_Isend(&mychar, 1, MPI_CHAR, rank_left, 0, comm, requests+irequest++);
MPI_Isend(&mychar, 1, MPI_CHAR, rank_up, 0, comm, requests+irequest++);
MPI_Isend(&mychar, 1, MPI_CHAR, rank_down, 0, comm, requests+irequest++);
MPI_Irecv( indata+idata++, 1, MPI_CHAR, rank_right, 0, comm, requests+irequest
    ++);
MPI_Irecv( indata+idata++, 1, MPI_CHAR, rank_left, 0, comm, requests+irequest
    ++);
MPI_Irecv( indata+idata++, 1, MPI_CHAR, rank_up, 0, comm, requests+irequest
    ++);
MPI_Irecv( indata+idata++, 1, MPI_CHAR, rank_down, 0, comm, requests+irequest
    ++);

```

10.2 Distributed graph topology

In many calculations on a grid (using the term in its mathematical, Finite Element Method (FEM), sense), a grid point will collect information from grid points around it. Under a sensible distribution of the grid over processes, this means that each process will collect information from a number of neighbour processes. The number of neighbours is dependent on that process. For instance, in a 2D grid (and assuming a five-point stencil for the computation) most processes communicate with four neighbours; processes on the edge with three, and processes in the corners with two.

Such a topology is illustrated in figure 10.1.

MPI's notion of *graph topology*, and the *neighbourhood collectives*, offer an elegant way of expressing such communication structures. There are various reasons for using graph topologies over the older, simpler methods.

- MPI is allowed to reorder the ranks, so that network proximity in the cluster corresponds to proximity in the structure of the code.
- Ordinary collectives could not directly be used for graph problems, unless one would adopt a subcommunicator for each graph neighbourhood. However, scheduling would then lead to deadlock or serialization.
- The normal way of dealing with graph problems is through non-blocking communications. However, since the user indicates an explicit order in which they are posted, congestion at certain processes may occur.
- Collectives can pipeline data, while send/receive operations need to transfer their data in its entirety.
- Collectives can use spanning trees, while send/receive uses a direct connection.

Thus the minimal description of a process graph contains for each process:

- Degree: the number of neighbour processes; and
- the ranks of the processes to communicate with.

However, this ignores that communication is not always symmetric: maybe the processes you receive from are not the ones you send to. Worse, maybe only one side of this duality is easily described. Therefore, there are two routines:

- **`MPI_Dist_graph_create_adjacent`** assumes that a process knows both who it is sending to, and who will send to it. This is the most work for the programmer to specify, but it is ultimately the most efficient.
- **`MPI_Dist_graph_create`** specifies on each process only what it is the source for; that is, who this process will be sending to. Consequently, some amount of processing – including communication – is needed to build the converse information, the ranks that will be sending to a process.

10.2.1 Graph creation

There are two creation routines for process graphs. These routines are fairly general in that they allow any process to specify any part of the topology. In practice, of course, you will mostly let each process describe its own neighbour structure.

The routine **`MPI_Dist_graph_create_adjacent`** assumes that a process knows both who it is sending to, and who will send to it. This means that every edge in the communication graph is represented twice, so the memory footprint is double of what is strictly necessary. However, no communication is needed to build the graph.

The second creation routine, **`MPI_Dist_graph_create`** (figure 119), is probably easier to use, especially in cases where the communication structure of your program is symmetric, meaning that a process sends to the same neighbours that it receives from. Now you specify on each process only what it is the source for; that is, who this process will be sending to.¹. Consequently, some amount of processing – including communication – is needed to build the converse information, the ranks that will be sending to a process.

1. I disagree with this design decision. Specifying your sources is usually easier than specifying your destinations.

MPI_Topo_test

```
int MPI_Topo_test(MPI_Comm comm, int *status)

status:
MPI_UNDEFINED
MPI_CART
MPI_GRAPH
MPI_DIST_GRAPH
```

How to read routine prototypes: 1.5.4.

manpage 118: Routine prototype for MPI'Topo'test

MPI_Dist_graph_create

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

Input Parameters:
comm_old : input communicator (handle)
n : number of source nodes for which this process specifies edges (non-negative integer)
sources : array containing the n source nodes for which this process specifies edges (array)
degrees : array specifying the number of destinations for each source node in the source noo
destinations : destination nodes for the source nodes in the source
node array (array of
non-negative
integers)
weights : weights for source to destination edges (array of
non-negative integers or MPI_UNWEIGHTED)
info : hints on optimization and interpretation of weights (handle)
reorder : the process may be reordered (true) or not (false) (logical)

Output Parameters:
comm_dist_graph : communicator with distributed graph topology added (handle)
```

How to read routine prototypes: 1.5.4.

manpage 119: Routine prototype for MPI'Dist'graph'create

Figure 10.1 describes the common five-point stencil structure. If we let each process only describe itself, we get the following:

- nsources = 1 because the calling process describes on node in the graph: itself.
- sources is an array of length 1, containing the rank of the calling process.
- degrees is an array of length 1, containing the degree (probably: 4) of this process.
- destinations is an array of length the degree of this process, probably again 4. The elements of this array are the ranks of the neighbour nodes; strictly speaking the ones that this process will send to.
- weights is an array declaring the relative importance of the destinations. For an *unweighted graph* use `MPI_UNWEIGHTED`.
- reorder (int in C, LOGICAL in Fortran) indicates whether MPI is allowed to shuffle ranks to achieve greater locality.

The resulting communicator has all the processes of the original communicator, with the same ranks. Its main point is usage in the so-called ‘neighbour collectives’.

10.2.2 Neighbour collectives

We can now use the graph topology to perform a gather or allgather `MPI_Neighbor_allgather` (figure 120) that combines only the processes directly connected to the calling process.

The neighbour collectives have the same argument list as the regular collectives, but they apply to a graph communicator.

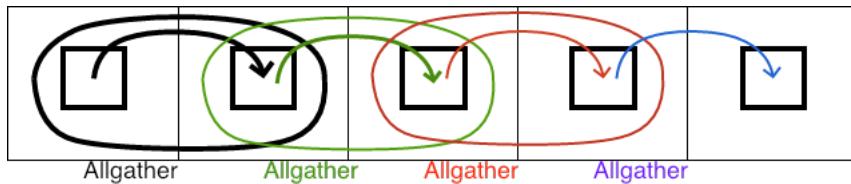


Figure 10.2: Solving the right-send exercise with neighbourhood collectives

Exercise 10.1. Revisit exercise 4.7 and solve it using `MPI_Dist_graph_create`. Use figure 10.2 for inspiration.

The other collective is `MPI_Neighbor_alltoall`.

The vector variants are `MPI_Neighbor_allgatherv` and `MPI_Neighbor_alltoallv`.

There is a heterogenous (multiple datatypes) variant: `MPI_Neighbor_alltoallw`.

For unclear reasons there is no `MPI_Neighbor_allreduce`.

10.2.3 Query

Statistics query: `MPI_Dist_graph_neighbors_count`

10.2.4 Graph topology (deprecated)

The original *MPI 1* had a graph topology interface `MPI_Graph_create` which required each process to specify the full process graph. Since this is not scalable, it should be considered deprecated. Use the distributed graph topology (section 10.2) instead.

10.3 Sources used in this chapter

Listing of code XX:

Listing of code XX:

MPI_Neighbor_allgather

Synopsis

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

Input Parameters:

sendbuf : starting address of the send buffer (choice)
sendcount : number of elements sent to each neighbor (non-negative integer)
sendtype : data type of send buffer elements (handle)
recvcount : number of elements received from each neighbor (non-negative integer)
recvtype : data type of receive buffer elements (handle)
comm : communicator (handle)

Output Parameters

recvbuf : starting address of the receive buffer (choice)

How to read routine prototypes: [1.5.4](#).

manpage 120: Routine prototype for MPI_Neighbor_allgather

Chapter 11

MPI topic: Shared memory

The one-sided MPI calls (chapter 8) can be used to emulate shared memory. In this chapter we will look at the ways MPI can interact with the presence of actual shared memory. Many MPI implementations have optimizations that detect shared memory and can exploit it, but that is not exposed to the programmer. The *MPI 3* standard added routines that do give the programmer that knowledge.

11.1 Recognizing shared memory

MPI's one-sided routines take a very symmetric view of processes: each process can access the window of every other process (within a communicator). Of course, in practice there will be a difference in performance depending on whether the origin and target are actually on the same shared memory, or whether they can only communicate through the network. For this reason MPI makes it easy to group processes by shared memory domains using `MPI_Comm_split_type` (figure 121).

Here the `split_type` parameter has to be from the following (short) list:

- `MPI_COMM_TYPE_SHARED`: split the communicator into subcommunicators of processes sharing a memory area.

In the following example, `CORES_PER_NODE` is a platform-dependent constant:

```
// commsplittype.c
MPI_Info info;
MPI_Comm_split_type(MPI_COMM_WORLD, MPI_COMM_TYPE_SHARED, procno, info, &
    sharedcomm);
MPI_Comm_size(sharedcomm, &new_nprocs);
MPI_Comm_rank(sharedcomm, &new_procno);

ASSERT(new_procno<CORES_PER_NODE);
```

11.2 Shared memory for windows

Processes that exist on the same physical shared memory should be able to move data by copying, rather than through MPI send/receive calls – which of course will do a copy operation under the hood. In order to do such user-level copying:

1. We need to create a shared memory area with `MPI_Win_allocate_shared`, and
2. We need to get pointers to where a process' area is in this shared space; this is done with `MPI_Win_shared_query`.

11.2.1 Pointers to a shared window

The first step is to create a window (in the sense of one-sided MPI; section 8.1) on the processes on one node. Using the `MPI_Win_allocate_shared` (figure 122) call presumably will put the memory close to the socket on which the process runs.

```
// sharedbulk.c
MPI_Aint window_size; double *window_data; MPI_Win node_window;
if (onnode_procid==0)
    window_size = sizeof(double);
else window_size = 0;
MPI_Win_allocate_shared
( window_size, sizeof(double), MPI_INFO_NULL,
  nodecomm,
  &window_data, &node_window );
```

The memory allocated by `MPI_Win_allocate_shared` is contiguous between the processes. This makes it possible to do address calculation. However, if a cluster node has a Non-Uniform Memory Access (NUMA) structure, for instance if two sockets have memory directly attached to each, this would increase latency for some processes. To prevent this, the key `alloc_shared_noncontig` can be set to `true` in the `MPI_Info` object.

```
// numa.c
MPI_Info window_info;
MPI_Info_create(&window_info);
MPI_Info_set(window_info, "alloc_shared_noncontig", "true");
MPI_Win_allocate_shared( window_size, sizeof(double), window_info,
                        nodecomm,
                        &window_data, &node_window );
MPI_Info_free(&window_info);
```

Let's now consider a scenario where you spawn two MPI ranks per node, and the node has 100G of memory. Using the above option to allow for non-contiguous window allocation, you hope that the windows of the two ranks are placed 50G apart. However, if you print out the addresses, you will find that they are placed considerably closer together. For a small windows that distance may be as little as 4K, the size of a *small page*.

The reason for this mismatch is that an address that you obtain with the ampersand operator in C is not a *physical address*, but a *virtual address*. The translation of where pages are placed in physical memory is determined by the *page table*.

11.2.2 Querying the shared structure

Even though the window created above is shared, that doesn't mean it's contiguous. Hence it is necessary to retrieve the pointer to the area of each process that you want to communicate with: `MPI_Win_shared_query` (figure 123).

```
||| MPI_Aint window_size0; int window_unit; double *win0_addr;
||| MPI_Win_shared_query( node_window, 0,
|||                         &window_size0, &window_unit, &win0_addr );
```

11.2.3 Heat equation example

As an example, which consider the 1D heat equation. On each process we create a local area of three point:

```
// sharedshared.c
MPI_Win_allocate_shared(3, sizeof(int), info, sharedcomm, &shared_baseptr, &
                        shared_window);
```

11.2.4 Shared bulk data

In applications such as *ray tracing*, there is a read-only large data object (the objects in the scene to be rendered) that is needed by all processes. In traditional MPI, this would need to be stored redundantly on each process, which leads to large memory demands. With MPI shared memory we can store the data object once per node. Using as above `MPI_Comm_split_type` to find a communicator per NUMA domain, we store the object on process zero of this node communicator.

Exercise 11.1. Let the ‘shared’ data originate on process zero in `MPI_COMM_WORLD`. Then:

- create a communicator per shared memory domain;
- create a communicator for all the processes with number zero on their node;
- broadcast the shared data to the processes zero on each node.

11.3 Sources used in this chapter

Listing of code XX:

Listing of code examples/mpi/c/sharedbulk.c:

Listing of code examples/mpi/c/numa.c:

Listing of code code/mpi/shared.c:

Listing of code XX:

MPI_Comm_split_type

C:

```
int MPI_Comm_split_type(
    MPI_Comm comm, int split_type, int key,
    MPI_Info info, MPI_Comm *newcomm)
```

Fortran:

```
MPI_Comm_split_type(comm, split_type, key, info, newcomm, ierror)
TYPE(MPI_Comm), INTENT(IN) :: comm
INTEGER, INTENT(IN) :: split_type, key
TYPE(MPI_Info), INTENT(IN) :: info
TYPE(MPI_Comm), INTENT(OUT) :: newcomm
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Python:

```
MPI.Comm.Split_type(
    self, int split_type, int key=0, Info info=INFO_NULL)
```

How to read routine prototypes: [1.5.4](#).

manpage 121: Routine prototype for MPI_Comm::split_type

MPI_Win_allocate_shared

Semantics:

```
MPI_WIN_ALLOCATE_SHARED(size, disp_unit, info, comm, baseptr, win)
```

Input parameters:

size: size of local window in bytes (non-negative integer)
disp_unit local unit size for displacements, in bytes (positive integer)
info: info argument (handle)
comm: intra-communicator (handle)

Output parameters:

baseptr: address of local allocated window segment (choice)
win: window object returned by the call (handle)

C:

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

Fortran:

```
MPI_Win_allocate_shared  
    (size, disp_unit, info, comm, baseptr, win, ierror)  
USE, INTRINSIC :: ISO_C_BINDING, ONLY : C_PTR  
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(IN) :: size  
INTEGER, INTENT(IN) :: disp_unit  
TYPE(MPI_Info), INTENT(IN) :: info  
TYPE(MPI_Comm), INTENT(IN) :: comm  
TYPE(C_PTR), INTENT(OUT) :: baseptr  
TYPE(MPI_Win), INTENT(OUT) :: win  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 122: Routine prototype for MPI_Win_allocate'shared

MPI_Win_shared_query

Semantics:

`MPI_WIN_SHARED_QUERY(win, rank, size, disp_unit, baseptr)`

Input arguments:

`win`: shared memory window object (handle)
`rank`: rank in the group of window `win` (non-negative integer)
or `MPI_PROC_NULL`

Output arguments:

`size`: size of the window segment (non-negative integer)
`disp_unit`: local unit size for displacements,
in bytes (positive integer)
`baseptr`: address for load/store access to window segment (choice)

C:

```
int MPI_Win_shared_query  
    (MPI_Win win, int rank, MPI_Aint *size, int *disp_unit,  
     void *baseptr)
```

Fortran:

```
MPI_Win_shared_query(win, rank, size, disp_unit, baseptr, ierror)  
USE, INTRINSIC :: ISO_C_BINDING, ONLY : C_PTR  
TYPE(MPI_Win), INTENT(IN) :: win  
INTEGER, INTENT(IN) :: rank  
INTEGER(KIND=MPI_ADDRESS_KIND), INTENT(OUT) :: size  
INTEGER, INTENT(OUT) :: disp_unit  
TYPE(C_PTR), INTENT(OUT) :: baseptr  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 123: Routine prototype for MPI_Win_shared_query

Chapter 12

MPI leftover topics

12.1 Info objects

Certain MPI routines can accept `MPI_Info` (figure 124) objects. These contain key-value pairs that can offer system or implementation dependent information.

Create an info object with `MPI_Info_create` (figure 125) and delete it with `MPI_Info_free` (figure 126)

Keys are then set with `MPI_Info_set` (figure 127), and they can be queried with `MPI_Info_get` (figure 128). Note that the output of the ‘get’ routine is not allocated: it is a buffer that is passed. The maximum length of a key is given by the parameter `MPI_MAX_INFO_KEY`. You can delete a key with `MPI_Info_delete` (figure 129).

There is a straightforward duplication of info objects: `MPI_Info_dup` (figure 130)

You can also query the number of keys in an info object with `MPI_Info_get_nkeys` (figure 131), after which the keys can be queried in succession with `MPI_Info_get_nthkey`

12.1.1 Environment information

The object `MPI_INFO_ENV` is predefined, containing:

- command Name of program executed.
- argv Space separated arguments to command.
- maxprocs Maximum number of MPI processes to start.
- soft Allowed values for number of processors.
- host Hostname.
- arch Architecture name.
- wdir Working directory of the MPI process.
- file Value is the name of a file in which additional information is specified.
- thread_level Requested level of thread support, if requested before the program started execution.

Note that these are the requested values; the running program can for instance have lower thread support.

MPI_Info

```
C  
MPI_Info info ;  
  
Fortran:  
Type(MPI_Info) info  
  
Python:
```

How to read routine prototypes: [1.5.4](#).

manpage 124: Routine prototype for MPI_Info

MPI_Info_create

```
MPI_INFO_CREATE(info)  
OUT info info object created (handle)  
  
C:  
int MPI_Info_create(MPI_Info *info)  
  
Fortran legacy:  
MPI_INFO_CREATE(INFO, IERROR)  
INTEGER INFO, IERROR
```

How to read routine prototypes: [1.5.4](#).

manpage 125: Routine prototype for MPI_Info_create

MPI_Info_free

```
MPI_INFO_FREE(info)  
INOUT info info object (handle)  
int MPI_Info_free(MPI_Info *info)  
MPI_INFO_FREE(INFO, IERROR)  
INTEGER INFO, IERROR
```

How to read routine prototypes: [1.5.4](#).

manpage 126: Routine prototype for MPI_Info_free

MPI_Info_set

```
MPI_INFO_SET(info, key, value)  
INOUT info info object (handle)  
IN key key (string)  
IN value value (string)  
int MPI_Info_set(MPI_Info info, char *key, char *value)  
MPI_INFO_SET(INFO, KEY, VALUE, IERROR)  
INTEGER INFO, IERROR  
CHARACTER(*) KEY, VALUE
```

How to read routine prototypes: [1.5.4](#).

manpage 127: Routine prototype for MPI_Info_set

MPI_Info_get

```
MPI_INFO_GET(info, key, valuelen, value, flag)
IN infoinfo object (handle)
IN keykey (string)
IN valuelenlength of value arg (integer)
OUT valuevalue (string)
OUT flagtrue if key defined, false if not (boolean)
int MPI_Info_get(MPI_Info info, char *key, int valuelen, char *value,
int *flag)
MPI_INFO_GET(INFO, KEY, VALUELEN, VALUE, FLAG, IERROR)
INTEGER INFO, VALUELEN, IERROR
CHARACTER(*) KEY, VALUE
LOGICAL FLAG
```

How to read routine prototypes: [1.5.4](#).

manpage 128: Routine prototype for MPI_Info_get

MPI_Info_delete

```
MPI_INFO_DELETE(info, key)
INOUT infoinfo object (handle)
IN keykey (string)
int MPI_Info_delete(MPI_Info info, char *key)
MPI_INFO_DELETE(INFO, KEY, IERROR)
INTEGER INFO, IERROR
CHARACTER(*) KEY
```

How to read routine prototypes: [1.5.4](#).

manpage 129: Routine prototype for MPI_Info_delete

MPI_Info_dup

```
MPI_INFO_DUP(info, newinfo)
IN infoinfo object (handle)
OUT newinfoinfo object (handle)
int MPI_Info_dup(MPI_Info info, MPI_Info *newinfo)
MPI_INFO_DUP(INFO, NEWINFO, IERROR)
INTEGER INFO, NEWINFO, IERROR
```

How to read routine prototypes: [1.5.4](#).

manpage 130: Routine prototype for MPI_Info_dup

MPI_Info_get_nkeys

```
MPI_INFO_GET_NKEYS(info, nkeys)
IN infoinfo object (handle)
OUT nkeysnumber of defined keys (integer)
int MPI_Info_get_nkeys(MPI_Info info, int *nkeys)
MPI_INFO_GET_NKEYS(INFO, NKEYS, IERROR)
INTEGER INFO, NKEYS, IERROR
```

How to read routine prototypes: [1.5.4](#).

manpage 131: Routine prototype for MPI_Info_get_nkeys

12.1.2 Communicator and window information

MPI has a built-in possibility of attaching information to *communicators* and *windows* using the calls `MPI_Comm_get_info`, `MPI_Comm_set_info`, `MPI_Win_get_info`, `MPI_Win_set_info`.

Copying a communicator with `MPI_Comm_dup` would cause the info to be copied; to attach new information to the copy there is `MPI_Comm_dup_with_info`.

12.2 Error handling

Errors in normal programs can be tricky to deal with; errors in parallel programs can be even harder. This is because in addition to everything that can go wrong with a single executable (floating point errors, memory violation) you now get errors that come from faulty interaction between multiple executables.

A few examples of what can go wrong:

- MPI errors: an MPI routine can abort for various reasons, such as receiving much more data than its buffer can accommodate. Such errors, as well as the more common type mentioned above, typically cause your whole execution to abort. That is, if one incarnation of your executable aborts, the MPI runtime will kill all others.
- Deadlocks and other hanging executions: there are various scenarios where your processes individually do not abort, but are all waiting for each other. This can happen if two processes are both waiting for a message from each other, and this can be helped by using non-blocking calls. In another scenario, through an error in program logic, one process will be waiting for more messages (including non-blocking ones) than are sent to it.

12.2.1 Error codes

- `MPI_ERR_ARG`: an argument was invalid that is not covered by another error code.
- `MPI_ERR_BUFFER`: The buffer pointer is invalid; this typically means that you have supplied a null pointer.
- `MPI_ERR_COMM`: invalid communicator. A common error is to use a null communicator in a call.
- `MPI_ERR_INTERN`: An internal error in MPI has been detected.
- `MPI_ERR_INFO`: invalid info object.
- `MPI_ERR_OTHER`: an error occurred; use `MPI_Error_string` to retrieve further information about this error.
- `MPI_ERR_PORT`: invalid port; this applies to `MPI_Comm_connect` and such.
- `MPI_ERR_SERVICE`: invalid service in `MPI_Unpublish_name`.
- `MPI_SUCCESS`: no error; MPI routine completed successfully.

12.2.2 Error handling

The MPI library has a general mechanism for dealing with errors that it detects. The default behaviour, where the full run is aborted, is equivalent to your code having the following call:

```
|| MPI_Comm_set_errhandler(MPI_COMM_WORLD, MPI_ERRORS_ARE_FATAL);
```

Remark 3 The routine `MPI_Errhandler_set` is deprecated.

Another simple possibility is to specify

```
|| MPI_Comm_set_errhandler(MPI_COMM_WORLD, MPI_ERRORS_RETURN);
```

which gives you the opportunity to write code that handles the error return value. The values `MPI_ERRORS_ARE_FATAL` and `MPI_ERRORS_RETURN` are of type `MPI_Errhandler`.

In most cases where an MPI error occurs a complete abort is the sensible thing, since there are few ways to recover. Alternatively, you could compare the return code to `MPI_SUCCESS` and print out debugging information:

```
|| int ierr;
   ierr = MPI_Something();
   if (ierr!=MPI_SUCCESS) {
       // print out information about what your programming is doing
       MPI_Abort();
   }
```

For instance,

```
Fatal error in MPI_Waitall:
See the MPI_ERROR field in MPI_Status for the error code
```

You could then retrieve the `MPI_ERROR` field of the status, and print out an error string with `MPI_Error_string` or maximal size `MPI_MAX_ERROR_STRING`:

```
|| MPI_Comm_set_errhandler(MPI_COMM_WORLD, MPI_ERRORS_RETURN);
   ierr = MPI_Waitall(2*ntrids-2, requests, status);
   if (ierr!=0) {
       char errtxt[MPI_MAX_ERROR_STRING];
       for (int i=0; i<2*ntrids-2; i++) {
           int err = status[i].MPI_ERROR;
           int len=MPI_MAX_ERROR_STRING;
           MPI_Error_string(err,errtxt,&len);
           printf("Waitall error: %d %s\n",err,errtxt);
       }
       MPI_Abort(MPI_COMM_WORLD, 0);
   }
```

One cases where errors can be handled is that of *MPI file I/O*: if an output file has the wrong permissions, code can possibly progress without writing data, or writing to a temporary file.

MPI operators (`MPI_Op`) do not return an error code. In case of an error they call `MPI_Abort`; if `MPI_ERRORS_RETURN` is the error handler, errors may be silently ignore.

You can create your own error handler with `MPI_Comm_create_errhandler`, which is then installed with `MPI_Comm_set_errhandler`. You can retrieve the error handler with `MPI_Comm_get_errhandler`.

12.3 Fortran issues

MPI is typically written in C, what if you program *Fortran*?

See section 5.1.2.1 for MPI types corresponding to *Fortran90* types.

12.3.1 Assumed-shape arrays

Use of other than contiguous data, for instance `A(1:N:2)`, was a problem in MPI calls, especially non-blocking ones. In that case it was best to copy the data to a contiguous array. This has been fixed in MPI3.

- Fortran routines have the same prototype as C routines except for the addition of an integer error parameter.
- The call for `MPI_Init` in Fortran does not have the commandline arguments; they need to be handled separately.
- The routine `MPI_Sizeof` is only available in Fortran, it provides the functionality of the C/C++ operator `sizeof`.

12.4 Fault tolerance

Processors are not completely reliable, so it may happen that one ‘breaks’: for software or hardware reasons it becomes unresponsive. For an MPI program this means that it becomes impossible to send data to it, and any collective operation involving it will hang. Can we deal with this case? Yes, but it involves some programming.

First of all, one of the possible MPI error return codes (section 12.2) is `MPI_ERR_COMM`, which can be returned if a processor in the communicator is unavailable. You may want to catch this error, and add a ‘replacement processor’ to the program. For this, the `MPI_Comm_spawn` can be used (see 7.1 for details). But this requires a change of program design: the communicator containing the new process(es) is not part of the old `MPI_COMM_WORLD`, so it is better to set up your code as a collection of inter-communicators to begin with.

12.5 Context information

See also section 12.5.3.

12.5.1 Processor name

You can query the *hostname* of a processor with `MPI_Get_processor_name` (figure 132). This name need not be unique between different processor ranks.

You have to pass in the character storage: the character array must be at least `MPI_MAX_PROCESSOR_NAME` characters long. The actual length of the name is returned in the `resultlen` parameter.

12.5.2 Version information

For runtime determination, The *MPI version* is available through two parameters `MPI_VERSION` and `MPI_SUBVERSION` or the function `MPI_Get_version` (figure 133) .

12.5.3 Attributes

Some runtime (or installation dependent) values are available as attributes through `MPI_Attr_get` (figure 134) .

Attributes are:

- `MPI_TAG_UB` Upper bound for *tag value*. Note that `MPI_TAG_UB` is the key, not the actual upper bound!
- `MPI_HOST` Host process rank, if such exists, `MPI_PROC_NULL`, otherwise.
- `MPI_IO` rank of a node that has regular I/O facilities (possibly myrank). Nodes in the same communicator may return different values for this parameter.
- `MPI_WTIME_IS_GLOBAL` Boolean variable that indicates whether clocks are synchronized.

Also:

- `MPI_UNIVERSE_SIZE`: the total number of processes that can be created. This can be more than the size of `MPI_COMM_WORLD` if the host list is larger than the number of initially started processes. See section 7.1.
Python: `mpi4py.MPI.UNIVERSE_SIZE`.
- `MPI_APPNUM`: if MPI is used in MPMD mode, or if `MPI_Comm_spawn_multiple` is used, this attribute reports the how-many program we are in.

12.6 Performance

In most of this book we talk about functionality of the MPI library. There are cases where a problem can be solved in more than one way, and then we wonder which one is the most efficient. In this section we will explicitly address performance. We start with two sections on the mere act of measuring performance.

12.6.1 Tools interface

Recent versions of MPI have a standardized way of reading out performance variables: the *tools interface*. However, since this is installation-dependent, you first need to query how much of the tools interface is provided.

```
// mpit.c
MPI_Init_thread(&argc, &argv, MPI_THREAD_SINGLE, &tlevel);
MPI_T_init_thread(MPI_THREAD_SINGLE, &tlevel);

MPI_Comm_size(MPI_COMM_WORLD, &nprocs);
MPI_Comm_rank(MPI_COMM_WORLD, &procid);

int npvar;
MPI_T_pvar_get_num(&npvar);
```

MPI_Get_processor_name

```
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()
```

How to read routine prototypes: [1.5.4.](#)

manpage 132: Routine prototype for MPI`Get`processor`name

MPI_Get_version

```
Semantics:  
MPI_GET_VERSION( version, subversion )  
  OUT version version number (integer)  
  OUT subversion subversion number (integer)  
  
C:  
int MPI_Get_version(int *version, int *subversion)  
  
Fortran:  
MPI_GET_VERSION(VERSION, SUBVERSION, IERROR)  
  INTEGER VERSION, SUBVERSION, IERROR
```

How to read routine prototypes: [1.5.4.](#)

manpage 133: Routine prototype for MPI`Get`version

MPI_Attr_get

```
int MPI_Attr_get(  
    MPI_Comm comm, int keyval, void *attribute_val, int *flag)  
  
Python:  
MPI.Comm.Get_attr(self, int keyval)
```

How to read routine prototypes: [1.5.4.](#)

manpage 134: Routine prototype for MPI`Attr`get

12.6.2 Timing

MPI has a wall clock timer: `MPI_Wtime` (figure 135) which gives the number of seconds from a certain point in the past. (Note the absence of the error parameter in the fortran call.)

```
// pingpong.c
int src = 0,tgt = nprocs/2;
double t, send=1.1,recv;
if (procno==src) {
    t = MPI_Wtime();
    for (int n=0; n<NEXPERIMENTS; n++) {
        MPI_Send(&send,1,MPI_DOUBLE,tgt,0,comm);
        MPI_Recv(&recv,1,MPI_DOUBLE,tgt,0,comm,MPI_STATUS_IGNORE);
    }
    t = MPI_Wtime()-t; t /= NEXPERIMENTS;
    printf("Time for pingpong: %e\n",t);
} else if (procno==tgt) {
    for (int n=0; n<NEXPERIMENTS; n++) {
        MPI_Recv(&recv,1,MPI_DOUBLE,src,0,comm,MPI_STATUS_IGNORE);
        MPI_Send(&recv,1,MPI_DOUBLE,src,0,comm);
    }
}
```

The timer has a resolution of `MPI_Wtick` (figure 136).

Timing in parallel is a tricky issue. For instance, most clusters do not have a central clock, so you can not relate start and stop times on one process to those on another. You can test for a global clock as follows:

```
int *v,flag;
MPI_Attr_get( comm, MPI_WTIME_IS_GLOBAL, &v, &flag );
if (mytid==0) printf("Time synchronized? %d->%d\n",flag,*v);
```

Normally you don't worry about the starting point for this timer: you call it before and after an event and subtract the values.

```
t = MPI_Wtime();
// something happens here
t = MPI_Wtime()-t;
```

If you execute this on a single processor you get fairly reliable timings, except that you would need to subtract the overhead for the timer. This is the usual way to measure timer overhead:

```
t = MPI_Wtime();
// absolutely nothing here
t = MPI_Wtime()-t;
```

12.6.2.1 Global timing

However, if you try to time a parallel application you will most likely get different times for each process, so you would have to take the average or maximum. Another solution is to synchronize the processors by using a *barrier*:

```
|| MPI_Barrier(comm)
|| t = MPI_Wtime();
|| // something happens here
|| MPI_Barrier(comm)
|| t = MPI_Wtime() - t;
```

Exercise 12.1. This scheme also has some overhead associated with it. How would you measure that?

12.6.2.2 Local timing

Now suppose you want to measure the time for a single send. It is not possible to start a clock on the sender and do the second measurement on the receiver, because the two clocks need not be synchronized. Usually a *ping-pong* is done:

```
|| if ( proc_source ) {
||   MPI_Send( /* to target */ );
||   MPI_Recv( /* from target */ );
|| else if ( proc_target ) {
||   MPI_Recv( /* from source */ );
||   MPI_Send( /* to source */ );
|| }
```

Exercise 12.2. Why is it generally not a good idea to use processes 0 and 1 for the source and target processor? Can you come up with a better guess?

No matter what sort of timing you are doing, it is good to know the accuracy of your timer. The routine `MPI_Wtick` gives the smallest possible timer increment. If you find that your timing result is too close to this ‘tick’, you need to find a better timer (for CPU measurements there are cycle-accurate timers), or you need to increase your running time, for instance by increasing the amount of data.

12.6.3 Profiling

MPI allows you to write your own profiling interface. To make this possible, every routine `MPI_Something` calls a routine `PMPPI_Something` that does the actual work. You can now write your `MPI_...` routine which calls `PMPPI_...`, and inserting your own profiling calls. As you can see in figure 12.1, normally only the PMPPI routines show up in the stack trace.

Does the standard mandate this?

12.6.4 Programming for performance

We outline some issues pertaining to performance.

Eager limit Short blocking messages are handled by a simpler mechanism than longer. The limit on what is considered ‘short’ is known as the *eager limit* (section 4.2.2.2), and you could tune your code by increasing its value. However, note that a process may likely have a buffer accommodating eager sends for every single other process. This may eat into your available memory.

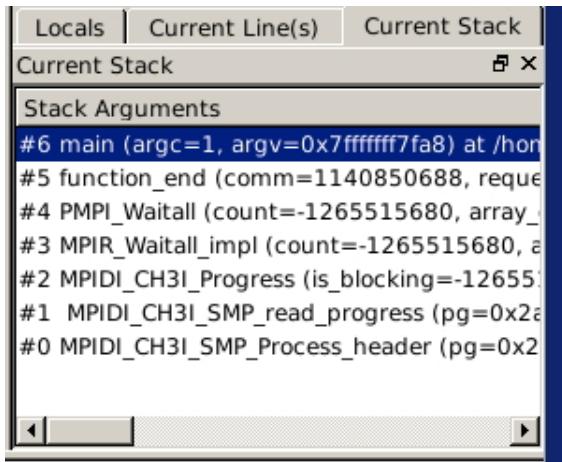


Figure 12.1: A stack trace, showing the MPI calls.

Blocking versus non-blocking The issue of *blocking* versus *non-blocking* communication is something of a red herring. While non-blocking communication allows *latency hiding*, we can not consider it an alternative to blocking sends, since replacing non-blocking by blocking calls will usually give *deadlock*.

Still, even if you use non-blocking communication for the mere avoidance of deadlock or serialization (section 4.2.2.3), bear in mind the possibility of overlap of communication and computation. This also brings us to our next point.

Looking at it the other way around, in a code with blocking sends you may get better performance from non-blocking, even if that is not structurally necessary.

Progress MPI is not magically active in the background, especially if the user code is doing scalar work that does not involve MPI. As sketched in section 4.3.3.1, there are various ways of ensuring that latency hiding actually happens.

Persistent sends If a communication between the same pair of processes, involving the same buffer, happens regularly, it is possible to set up a *persistent communication*. See section 4.4.4.

Buffering MPI uses internal buffers, and the copying from user data to these buffers may affect performance. For instance, derived types (section 5.2) can typically not be streamed straight through the network (this requires special hardware support [12]) so they are first copied. Somewhat surprisingly, we find that *buffered communication* (section 4.4.5) does not help. Perhaps MPI implementors have not optimized this mode since it is so rarely used.

This is issue is extensively investigated in [4].

Graph topology and neighborhood collectives Load balancing and communication minimization are important in irregular applications. There are dedicated programs for this (*ParMetis*, *Zoltan*), and libraries such as *PETSc* may offer convenient access to such capabilities.

In the declaration of a *graph topology* (section 10.2) MPI is allowed to reorder processes, which could be used to support such activities. It can also serve for better message sequencing when *neighbourhood collectives* are used.

Network issues In the discussion so far we have assumed that the network is a perfect conduit for data. However, there are issues of port design, in particular caused by *oversubscription* that adversely affect performance. While in an ideal world it may be possible to set up routine to avoid this, in the actual practice of a supercomputer cluster, *network contention* or *message collision* from different user jobs is hard to avoid.

Offloading and onloading There are different philosophies of *network card design*: *Mellanox*, being a network card manufacturer, believes in off-loading network activity to the Network Interface Card (NIC), while *Intel*, being a processor manufacturer, believes in ‘on-loading’ activity to the process. There are argument either way.

Either way, investigate the capabilities of your network.

12.7 Determinism

MPI processes are only synchronized to a certain extent, so you may wonder what guarantees there are that running a code twice will give the same result. You need to consider two cases: first of all, if the two runs are on different numbers of processors there are already numerical problems; see HPSC-??.

Let us then limit ourselves to two runs on the same set of processors. In that case, MPI is deterministic as long as you do not use wildcards such as `MPI_ANY_SOURCE`. Formally, MPI messages are ‘non-overtaking’: two messages between the same sender-receiver pair will arrive in sequence. Actually, they may not arrive in sequence: they are *matched* in sequence in the user program. If the second message is much smaller than the first, it may actually arrive earlier in the lower transport layer.

12.8 Subtleties with processor synchronization

Blocking communication involves a complicated dialog between the two processors involved. Processor one says ‘I have this much data to send; do you have space for that?’, to which processor two replies ‘yes, I do; go ahead and send’, upon which processor one does the actual send. This back-and-forth (technically known as a *handshake*) takes a certain amount of communication overhead. For this reason, network hardware will sometimes forgo the handshake for small messages, and just send them regardless, knowing that the other process has a small buffer for such occasions.

One strange side-effect of this strategy is that a code that should *deadlock* according to the MPI specification does not do so. In effect, you may be shielded from your own programming mistake! Of course, if you then run a larger problem, and the small message becomes larger than the threshold, the deadlock will suddenly occur. So you find yourself in the situation that a bug only manifests itself on large problems, which are usually harder to debug. In this case, replacing every `MPI_Send` with a `MPI_Ssend` will force the handshake, even for small messages.

Conversely, you may sometimes wish to avoid the handshake on large messages. MPI as a solution for this: the `MPI_Rsend` ('ready send') routine sends its data immediately, but it needs the receiver to be ready for this. How can you guarantee that the receiving process is ready? You could for instance do the following (this uses non-blocking routines, which are explained below in section 4.3.1):

```
if ( receiving ) {
    MPI_Irecv() // post non-blocking receive
    MPI_Barrier() // synchronize
} else if ( sending ) {
    MPI_Barrier() // synchronize
    MPI_Rsend() // send data fast
```

When the barrier is reached, the receive has been posted, so it is safe to do a ready send. However, global barriers are not a good idea. Instead you would just synchronize the two processes involved.

Exercise 12.3. Give pseudo-code for a scheme where you synchronize the two processes through the exchange of a blocking zero-size message.

12.9 Multi-threading

Hybrid MPI/threaded codes need to replace `MPI_Init` by `MPI_Init_thread` (figure 179). With the required parameter the user requests a certain level of support, and MPI reports the actual capabilities in the provided parameter.

The following constants are defined:

- `MPI_THREAD_SINGLE`: each MPI process can only have a single thread.
- `MPI_THREAD_FUNNELED`: an MPI process can be multithreaded, but all MPI calls need to be done from a single thread.
- `MPI_THREAD_SERIALIZED`: a process can sustain multiple threads that make MPI calls, but these threads can not be simultaneous: they need to be for instance in an OpenMP *critical section*.
- `MPI_THREAD_MULTIPLE`: processes can be fully generally multi-threaded.

These values are monotonically increasing.

After the initialization call, you can query the support level with `MPI_Query_thread` (figure 138).

In case more than one thread performs communication, `MPI_Is_thread_main` (figure 139) can determine whether a thread is the main thread:

MPI_Wtime

```
C:  
double MPI_Wtime(void);  
  
Fortran:  
DOUBLE PRECISION MPI_WTIME()  
  
Python:  
MPI.Wtime()
```

How to read routine prototypes: [1.5.4](#).

manpage 135: Routine prototype for MPI_Wtime

MPI_Wtick

```
C:  
double MPI_Wtick(void);  
  
Fortran:  
DOUBLE PRECISION MPI_WTICK()  
  
Python  
MPI.Wtick()
```

How to read routine prototypes: [1.5.4](#).

manpage 136: Routine prototype for MPI_Wtick

MPI_Init_thread

```
C:  
int MPI_Init_thread(int *argc, char ***argv, int required, int *provided)  
  
Fortran:  
MPI_Init_thread(required, provided, ierror)  
INTEGER, INTENT(IN) :: required  
INTEGER, INTENT(OUT) :: provided  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 137: Routine prototype for MPI_Init_thread

MPI_Query_thread

```
C:  
int MPI_Query_thread(int *provided)  
  
Fortran:  
MPI_Query_thread(provided, ierror)  
INTEGER, INTENT(OUT) :: provided  
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 138: Routine prototype for MPI_Query_thread

12.10 Shell interaction

MPI programs are not run directly from the shell, but are started through an *ssh tunnel*. We briefly discuss ramifications of this.

12.10.1 Standard input

Letting MPI processes interact with the environment is not entirely straightforward. For instance, *shell input redirection* as in

```
mpirun -np 2 mpiprogram < someinput
```

may not work.

Instead, use a script `programscript` that has one parameter:

```
#!/bin/bash
mpirunprogram < $1
```

and run this in parallel:

```
mpirun -np 2 programscript someinput
```

12.10.2 Standard out and error

The `stdout` and `stderr` streams of an MPI process are returned through the ssh tunnel. Thus they can be caught as the `stdout/err of mpirun`.

```
// outerr.c
fprintf(stdout,"This goes to std out\n");
fprintf(stderr,"This goes to std err\n");
```

12.10.3 Process status

The return code of `MPI_Abort` is returned as the *processes status of mpirun*. Running

```
// abort.c
if (procno==nprocs-1)
    MPI_Abort(comm, 37);
```

as

```
mpirun -np 4 ./abort ; \
echo "Return code from ${MPIRUN} is <<$$?>>"
```

gives

```
TACC: Starting up job 3760534
TACC: Starting parallel tasks...
application called MPI_Abort(MPI_COMM_WORLD, 37) - process 3
TACC: MPI job exited with code: 37
TACC: Shutdown complete. Exiting.
Return code from ibrun is <<37>>
```

12.10.4 Multiple program start

The sort of script of section 12.10.1 can also be used to implement *MPMD* runs: we let the script start one of a number of programs. For this, we use the fact that the MPI rank is known in the environment as `PMI_RANK`. Use a script `mpmdscript`:

```
#!/bin/bash
if [ ${PMI_RANK} -eq 0 ] ; then
    ./programmaster
else
    ./programworker
fi
```

which is then run in parallel:

```
mpirun -np 25 mpmdscript
```

12.11 The origin of one-sided communication in ShMem

The Cray *T3E* had a library called *shmem* which offered a type of shared memory. Rather than having a true global address space it worked by supporting variables that were guaranteed to be identical between processors, and indeed, were guaranteed to occupy the same location in memory. Variables could be declared to be shared a ‘symmetric’ pragma or directive; their values could be retrieved or set by `shmem_get` and `shmem_put` calls.

12.12 Leftover topics

12.12.1 MPI constants

MPI has a number of built-in *constants*. These do not all behave the same.

- Some are *compile-time* constants. Examples are `MPI_VERSION` and `MPI_MAX_PROCESSOR_NAME`. Thus, they can be used in array size declarations, even before `MPI_Init`.
- Some *link-time* constants get their value by MPI initialization, such as `MPI_COMM_WORLD`. Such symbols, which include all predefined handles, can be used in initialization expressions.

- Some link-time symbols can not be used in initialization expressions, such as `MPI_BOTTOM` and `MPI_STATUS_IGNORE`.

For symbols, the binary realization is not defined. For instance, `MPI_COMM_WORLD` is of type `MPI_Comm`, but the implementation of that type is not specified.

See Annex A of the 3.1 standard for full lists.

The following are the compile-time constants:

- `MPI_MAX_PROCESSOR_NAME`
- `MPI_MAX_LIBRARY_VERSION_STRING`
- `MPI_MAX_ERROR_STRING`
- `MPI_MAX_DATAREP_STRING`
- `MPI_MAX_INFO_KEY`
- `MPI_MAX_INFO_VAL`
- `MPI_MAX_OBJECT_NAME`
- `MPI_MAX_PORT_NAME`
- `MPI_VERSION`
- `MPI_SUBVERSION`
- `MPI_STATUS_SIZE` (Fortran only)
- `MPI_ADDRESS_KIND` (Fortran only)
- `MPI_COUNT_KIND` (Fortran only)
- `MPI_INTEGER_KIND` (Fortran only)
- `MPI_OFFSET_KIND` (Fortran only)
- `MPI_SUBARRAYS_SUPPORTED` (Fortran only)
- `MPI_ASYNC_PROTECTS_NONBLOCKING` (Fortran only)

The following are the link-time constants:

- `MPI_BOTTOM`
- `MPI_STATUS_IGNORE`
- `MPI_STATUSES_IGNORE`
- `MPI_ERRCODES_IGNORE`
- `MPI_IN_PLACE`
- `MPI_ARGV_NULL`
- `MPI_ARGVS_NULL`
- `MPI_UNWEIGHTED`
- `MPI_WEIGHTS_EMPTY`

Assorted constants:

- `MPI_PROC_NULL`
- `MPI_ANY_SOURCE`
- `MPI_ANY_TAG`
- `MPI_UNDEFINED`
- `MPI_BSEND_OVERHEAD`
- `MPI_KEYVAL_INVALID`
- `MPI_LOCK_EXCLUSIVE`

- `MPI_LOCK_SHARED`
- `MPI_ROOT`

(This section was inspired by <http://blogs.cisco.com/performance/mpi-outside-of-c-and-fortran>)

12.12.2 32-bit size issues

The `size` parameter in MPI routines is defined as an `int`, meaning that it is limited to 32-bit quantities. There are ways around this, such as sending a number of `MPI_Type_contiguous` blocks that add up to more than 2^{31} .

12.12.3 Python issues

12.12.3.1 Byte calculations

The `MPI_Win_create` routine needs a displacement in bytes. Here is a good way for finding the size of `numpy` datatypes:

```
// numpy.dtype('i').itemsize
```

12.12.3.2 Arrays of objects

Objects of type `MPI_Status` or `MPI_Request` often need to be created in an array, for instance when looping through a number of `Irecv` calls. In that case the following idiom may come in handy:

```
// requests = [ None ] * nprocs
for p in range(nprocs):
    requests[p] = comm.Irecv( ... )
```

12.12.4 Cancelling messages

In section 39.3 we showed a master-worker example where the master accepts in arbitrary order the messages from the workers. Here we will show a slightly more complicated example, where only the result of the first task to complete is needed. Thus, we issue an `MPI_Recv` with `MPI_ANY_SOURCE` as source. When a result comes, we broadcast its source to all processes. All the other workers then use this information to cancel their message with an `MPI_Cancel` operation.

```
// cancel.c
fprintf(stderr, "get set, go!\n");
if (procno==nprocs-1) {
    MPI_Status status;
    MPI_Recv(dummy, 0, MPI_INT, MPI_ANY_SOURCE, 0, comm,
             &status);
    first_tid = status.MPI_SOURCE;
    MPI_Bcast(&first_tid, 1, MPI_INT, nprocs-1, comm);
    fprintf(stderr, "[%d] first msg came from %d\n", procno, first_tid);
} else {
    float randomfraction = (rand() / (double) RAND_MAX);
    int randomwait = (int) (nprocs * randomfraction);
```

```

MPI_Request request;
fprintf(stderr, "[%d] waits for %e/%d=%d\n",
        procno, randomfraction, nprocs, randomwait);
sleep(randomwait);
MPI_Isend(dummy, 0, MPI_INT, nprocs-1, 0, comm,
            &request);
MPI_Bcast(&first_tid, 1, MPI_INT, nprocs-1, comm
            );
if (procno!=first_tid) {
    MPI_Cancel(&request);
    fprintf(stderr, "[%d] canceled\n", procno);
}
}

```

After the cancelling operation it is still necessary to call **MPI_Request_free**, **MPI_Wait**, or **MPI_Test** in order to free the request object.

The **MPI_Cancel** operation is local, so it can not be used for *non-blocking collectives* or one-sided transfers.

12.12.5 Constants

MPI constants such as **MPI_COMM_WORLD** or **MPI_INT** are not necessarily statitally defined, such as by a `#define` statement: the best you can say is that they have a value after **MPI_Init** or **MPI_Init_thread**. That means you can not transfer a compiled MPI file between platforms, or even between compilers on one platform. However, a working MPI source on one MPI implementation will also work on another.

12.13 Literature

Online resources:

- MPI 1 Complete reference:
<http://www.netlib.org/utk/papers/mpi-book/mpi-book.html>
- Official MPI documents:
<http://www.mpi-forum.org/docs/>
- List of all MPI routines:
<http://www.mcs.anl.gov/research/projects/mpi/www/www3/>

Tutorial books on MPI:

- Using MPI [7] by some of the original authors.

12.14 Sources used in this chapter

Listing of code XX:

Listing of code XX:

Listing of code XX:

MPI_Is_thread_main

C:

```
int MPI_Is_thread_main(int *flag)
```

Fortran:

```
MPI_Is_thread_main(flag, ierror)
LOGICAL, INTENT(OUT) :: flag
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: [1.5.4](#).

manpage 139: Routine prototype for MPI_Is_thread_main

Chapter 13

MPI Reference

This section gives reference information and illustrative examples of the use of MPI. While the code snippets given here should be enough, full programs can be found in the repository for this book <https://bitbucket.org/VictorEijkhout/parallel-computing-book>.

13.1 Leftover topics

13.1.1 MPI constants

MPI has a number of built-in *constants*. These do not all behave the same.

- Some are *compile-time* constants. Examples are `MPI_VERSION` and `MPI_MAX_PROCESSOR_NAME`. Thus, they can be used in array size declarations, even before `MPI_Init`.
- Some *link-time* constants get their value by MPI initialization, such as `MPI_COMM_WORLD`. Such symbols, which include all predefined handles, can be used in initialization expressions.
- Some link-time symbols can not be used in initialization expressions, such as `MPI_BOTTOM` and `MPI_STATUS_IGNORE`.

For symbols, the binary realization is not defined. For instance, `MPI_COMM_WORLD` is of type `MPI_Comm`, but the implementation of that type is not specified.

See Annex A of the 3.1 standard for full lists.

The following are the compile-time constants:

```
MPI_MAX_PROCESSOR_NAME
MPI_MAX_LIBRARY_VERSION_STRING
MPI_MAX_ERROR_STRING
MPI_MAX_DATAREP_STRING
MPI_MAX_INFO_KEY
MPI_MAX_INFO_VAL
MPI_MAX_OBJECT_NAME
MPI_MAX_PORT_NAME
MPI_VERSION
MPI_SUBVERSION
MPI_STATUS_SIZE (Fortran only)
MPI_ADDRESS_KIND (Fortran only)
```

```
MPI_COUNT_KIND (Fortran only)
MPI_INTEGER_KIND (Fortran only)
MPI_OFFSET_KIND (Fortran only)
MPI_SUBARRAYS_SUPPORTED (Fortran only)
MPI_ASYNC_PROTECTS_NONBLOCKING (Fortran only)
```

The following are the link-time constants:

```
MPI_BOTTOM
MPI_STATUS_IGNORE
MPI_STATUSES_IGNORE
MPI_ERRCODES_IGNORE
MPI_IN_PLACE
MPI_ARGV_NULL
MPI_ARGVS_NULL
MPI_UNWEIGHTED
MPI_WEIGHTS_EMPTY
```

Assorted constants:

```
C type: const int (or unnamed enum)
Fortran type: INTEGER

MPI_PROC_NULL
MPI_ANY_SOURCE
MPI_ANY_TAG
MPI_UNDEFINED
MPI_BSEND_OVERHEAD
MPI_KEYVAL_INVALID
MPI_LOCK_EXCLUSIVE
MPI_LOCK_SHARED
MPI_ROOT
```

(This section was inspired by <http://blogs.cisco.com/performance/mpi-outside-of-c-and-fortran>)

Chapter 14

MPI Review

For all true/false questions, if you answer that a statement is false, give a one-line explanation.

14.1 Conceptual

Exercise 14.1. True or false: `mpicc` is a compiler.

Exercise 14.2. What is the function of a hostfile?

14.2 Communicators

1. True or false: in each communicator, processes are numbered consecutively from zero.
2. If a process is in two communicators, it has the same rank in both.

14.3 Point-to-point

1. Describe a deadlock scenario involving three processors.
2. True or false: a message sent with `MPI_Isend` from one processor can be received with an `MPI_Recv` call on another processor.
3. True or false: a message sent with `MPI_Send` from one processor can be received with an `MPI_Irecv` on another processor.
4. Why does the `MPI_Irecv` call not have an `MPI_Status` argument?
5. What is the relation between the concepts of ‘origin’, ‘target’, ‘fence’, and ‘window’ in one-sided communication.
6. What are the three routines for one-sided data transfer?
7. In the following fragments assume that all buffers have been allocated with sufficient size. For each fragment note whether it deadlocks or not. Discuss performance issues.

```
// block1.c
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm);
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Recv(rbuffer,buflen,MPI_INT,p,0,comm,MPI_STATUS_IGNORE);

// block2.c
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Recv(rbuffer,buflen,MPI_INT,p,0,comm,MPI_STATUS_IGNORE);
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm);

// block3.c
int ireq = 0;
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Isend(sbuffers[p],buflen,MPI_INT,p,0,comm,&(requests[ireq++]));
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Recv(rbuffer,buflen,MPI_INT,p,0,comm,MPI_STATUS_IGNORE);
MPI_Waitall(nprocs-1,requests,MPI_STATUSES_IGNORE);

// block4.c
int ireq = 0;
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Irecv(rbuffers[p],buflen,MPI_INT,p,0,comm,&(requests[ireq++]));
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm);
MPI_Waitall(nprocs-1,requests,MPI_STATUSES_IGNORE);

// block5.c
int ireq = 0;
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Irecv(rbuffers[p],buflen,MPI_INT,p,0,comm,&(requests[ireq++]));
MPI_Waitall(nprocs-1,requests,MPI_STATUSES_IGNORE);
for (int p=0; p<nprocs; p++)
    if (p!=procid)
        MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm);
```

Fortran codes:

```

// block1.F90
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm,ierr)
  end if
end do
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Recv(rbuffer,buflen,MPI_INT,p,0,comm,MPI_STATUS_IGNORE
      ,ierr)
  end if
end do

// block2.F90
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Recv(rbuffer,buflen,MPI_INT,p,0,comm,MPI_STATUS_IGNORE
      ,ierr)
  end if
end do
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm,ierr)
  end if
end do

// block3.F90
ireq = 0
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Isend(sbuffers(1,p+1),buflen,MPI_INT,p,0,comm,&
      requests(ireq+1),ierr)
    ireq = ireq+1
  end if
end do
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Recv(rbuffer,buflen,MPI_INT,p,0,comm,MPI_STATUS_IGNORE
      ,ierr)
  end if
end do
call MPI_Waitall(nprocs-1,requests,MPI_STATUSES_IGNORE,ierr)

// block4.F90
ireq = 0
do p=0,nprocs-1
  if (p/=procid) then
    call MPI_Irecv(rbuffers(1,p+1),buflen,MPI_INT,p,0,comm,&
      requests(ireq+1),ierr)
    ireq = ireq+1
  end if
end do

```

```
|| do p=0,nprocs-1
||   if (p/=procid) then
||     call MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm,ierr)
||   end if
|| end do
|| call MPI_Waitall(nprocs-1,requests,MPI_STATUSES_IGNORE,ierr)

|| // block5.F90
|| ireq = 0
|| do p=0,nprocs-1
||   if (p/=procid) then
||     call MPI_Irecv(rbuffers(1,p+1),buflen,MPI_INT,p,0,comm,&
||                   requests(ireq+1),ierr)
||     ireq = ireq+1
||   end if
|| end do
|| call MPI_Waitall(nprocs-1,requests,MPI_STATUSES_IGNORE,ierr)
|| do p=0,nprocs-1
||   if (p/=procid) then
||     call MPI_Send(sbuffer,buflen,MPI_INT,p,0,comm,ierr)
||   end if
|| end do
```

14.4 Collectives

1. MPI collectives can be divided into (a) rooted vs rootless (b) using uniform buffer lengths vs variable length buffers (c) blocking vs non-blocking. Give examples of each type.
2. True or false: an **MPI_Scatter** call puts the same data on each process.
3. Given a distributed array, with every processor storing

```
|| double x[N]; // N can vary per processor
```

give the approximate MPI-based code that computes the maximum value in the array, and leaves the result on every processor.

4. With data as in the previous question, given the code for normalizing the array.

14.5 Datatypes

1. Give two examples of MPI derived datatypes. What parameters are used to describe them?
2. Give a practical example where the sender uses a different type to send than the receiver uses in the corresponding receive call. Name the types involved.
3. Fortran only. True or false?
 - (a) Array indices can be different between the send and receive buffer arrays.
 - (b) It is allowed to send an array section.
 - (c) You need to *Reshape* a multi-dimensional array to linear shape before you can send it.
 - (d) An allocatable array, when dimensioned and allocated, is treated by MPI as if it were a normal static array, when used as send buffer.

- (e) An allocatable array is allocated if you use it as the receive buffer: it is filled with the incoming data.
- 4. Fortran only: how do you handle the case where you want to use an allocatable array as receive buffer, but it has not been allocated yet, and you do not know the size of the incoming data?

14.6 Theory

- 1. Give a simple model for the time a send operation takes.
- 2. Give a simple model for the time a broadcast of a single scalar takes.

PART II

OPENMP

Chapter 15

Getting started with OpenMP

This chapter explains the basic concepts of OpenMP, and helps you get started on running your first OpenMP program.

15.1 The OpenMP model

We start by establishing a mental picture of the hardware and software that OpenMP targets.

15.1.1 Target hardware

Modern computers have a multi-layered design. Maybe you have access to a cluster, and maybe you have learned how to use MPI to communicate between cluster nodes. OpenMP, the topic of this chapter, is concerned with a single *cluster node* or *motherboard*, and getting the most out of the available parallelism available there.

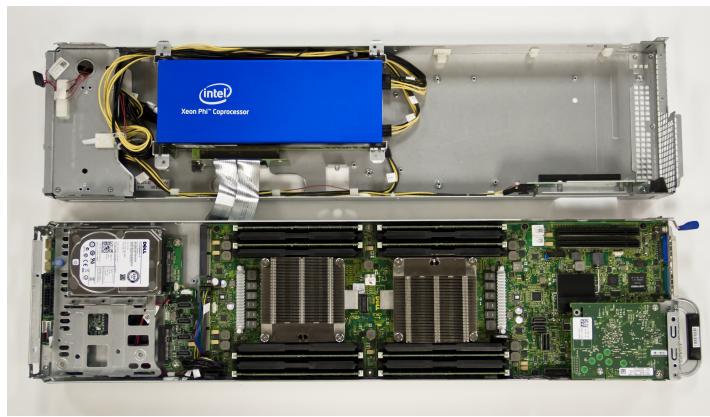


Figure 15.1: A node with two sockets and a co-processor

Figure 15.1 pictures a typical design of a node: within one enclosure you find two sockets: single processor chips. Your personal laptop of computer will probably have one socket, most supercomputers have nodes

with two or four sockets (the picture is of a *Stampede node* with two sockets)¹, although the recent *Intel Knight's Landing* is again a single-socket design.

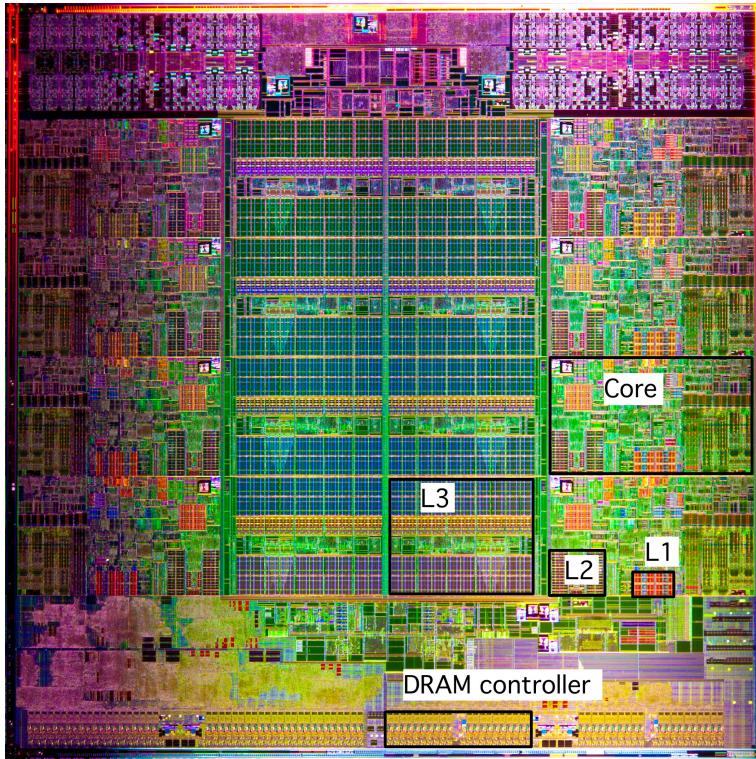


Figure 15.2: Structure of an Intel Sandybridge eight-core socket

To see where OpenMP operates we need to dig into the sockets. Figure 15.2 shows a picture of an *Intel Sandybridge* socket. You recognize a structure with eight cores: independent processing units, that all have access to the same memory. (In figure 15.1 you saw four memory banks attached to each of the two sockets; all of the sixteen cores have access to all that memory.)

To summarize the structure of the architecture that OpenMP targets:

- A node has up to four sockets;
- each socket has up to 60 cores;
- each core is an independent processing unit, with access to all the memory on the node.

15.1.2 Target software

OpenMP is based on two concepts: the use of *threads* and the *fork/join model* of parallelism. For now you can think of a thread as a sort of process: the computer executes a sequence of instructions. The fork/join model says that a thread can split itself ('fork') into a number of threads that are identical copies. At some point these copies go away and the original thread is left ('join'), but while the *team of threads*

1. In that picture you also see a co-processor: OpenMP is increasingly targeting those too.

created by the fork exists, you have parallelism available to you. The part of the execution between fork and join is known as a *parallel region*.

Figure 15.3 gives a simple picture of this: a thread forks into a team of threads, and these threads themselves can fork again.

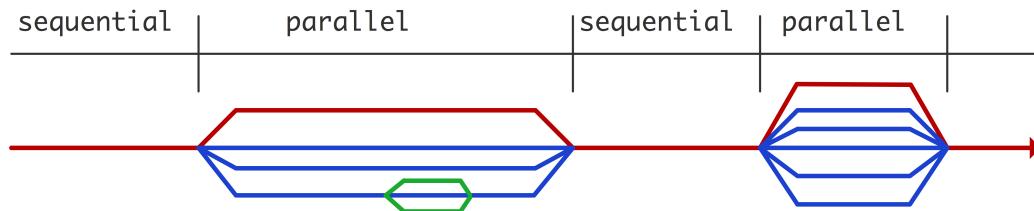


Figure 15.3: Thread creation and deletion during parallel execution

The threads that are forked are all copies of the *master thread*: they have access to all that was computed so far; this is their *shared data*. Of course, if the threads were completely identical the parallelism would be pointless, so they also have private data, and they can identify themselves: they know their thread number. This allows you to do meaningful parallel computations with threads.

This brings us to the third important concept: that of *work sharing* constructs. In a team of threads, initially there will be replicated execution; a work sharing construct divides available parallelism over the threads.

So there you have it: OpenMP uses teams of threads, and inside a parallel region the work is distributed over the threads with a work sharing construct. Threads can access shared data, and they have some private data.

An important difference between OpenMP and MPI is that parallelism in OpenMP is dynamically activated by a thread spawning a team of threads. Furthermore, the number of threads used can differ between parallel regions, and threads can create threads recursively. This is known as as *dynamic mode*. By contrast, in an MPI program the number of running processes is (mostly) constant throughout the run, and determined by factors external to the program.

15.1.3 About threads and cores

OpenMP programming is typically done to take advantage of *multicore* processors. Thus, to get a good speedup you would typically let your number of threads be equal to the number of cores. However, there is nothing to prevent you from creating more threads: the operating system will use *time slicing* to let them all be executed. You just don't get a speedup beyond the number of actually available cores.

On some modern processors there are *hardware threads*, meaning that a core can actually let more than one thread be executed, with some speedup over the single thread. To use such a processor efficiently you would let the number of OpenMP threads be $2\times$ or $4\times$ the number of cores, depending on the hardware.

15.1.4 About thread data

In most programming languages, visibility of data is governed by rules on the *scope of variables*: a variable is declared in a block, and it is then visible to any statement in that block and blocks with a *lexical scope*

contained in it, but not in surrounding blocks:

```
|| main () {
// no variable 'x' define here
{
    int x = 5;
    if (somecondition) { x = 6; }
    printf("x=%e\n", x); // prints 5 or 6
}
printf("x=%e\n", x); // syntax error: 'x' undefined
}
```

In C, you can redeclare a variable inside a nested scope:

```
|| {
int x;
if (something) {
    double x; // same name, different entity
}
x = ... // this refers to the integer again
}
```

Doing so makes the outer variable inaccessible.

Fortran has simpler rules, since it does not have blocks inside blocks.

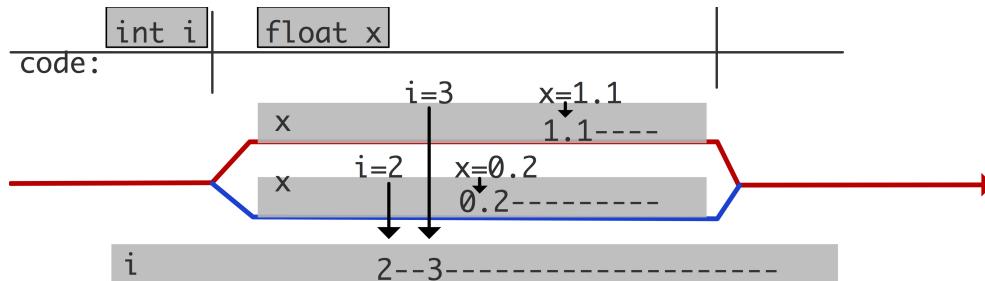


Figure 15.4: Locality of variables in threads

In OpenMP the situation is a bit more tricky because of the threads. When a team of threads is created they can all see the data of the master thread. However, they can also create data of their own. This is illustrated in figure 15.4. We will go into the details later.

15.2 Compiling and running an OpenMP program

15.2.1 Compiling

Your file or Fortran module needs to contain

```
|| #include "omp.h"
```

in C, and

```
|| use omp_lib
```

or

```
|| #include "omp_lib.h"
```

for Fortran.

OpenMP is handled by extensions to your regular compiler, typically by adding an option to your commandline:

```
# gcc
gcc -o foo foo.c -fopenmp
# Intel compiler
icc -o foo foo.c -openmp
```

If you have separate compile and link stages, you need that option in both.

When you use the openmp compiler option, a *cpp* variable `_OPENMP` will be defined. Thus, you can have conditional compilation by writing

```
|| #ifdef _OPENMP
    ...
|| #else
    ...
|| #endif
```

15.2.2 Running an OpenMP program

You run an OpenMP program by invoking it the regular way (for instance `./a.out`), but its behaviour is influenced by some *OpenMP environment variables*. The most important one is `OMP_NUM_THREADS`:

```
export OMP_NUM_THREADS=8
```

which sets the number of threads that a program will use. See section 26.1 for a list of all environment variables.

15.3 Your first OpenMP program

In this section you will see just enough of OpenMP to write a first program and to explore its behaviour. For this we need to introduce a couple of OpenMP language constructs. They will all be discussed in much greater detail in later chapters.

15.3.1 Directives

OpenMP is not magic, so you have to tell it when something can be done in parallel. This is mostly done through *directives*; additional specifications can be done through library calls.

In C/C++ the *pragma* mechanism is used: annotations for the benefit of the compiler that are otherwise not part of the language. This looks like:

```
|| #pragma omp somedirective clause(value,othervalue)
||     parallel statement;

|| #pragma omp somedirective clause(value,othervalue)
|| {
||     parallel statement 1;
||     parallel statement 2;
|| }
```

with

- the `#pragma omp sentinel` to indicate that an OpenMP directive is coming;
- a directive, such as `parallel`;
- and possibly clauses with values.
- After the directive comes either a single statement or a block in *curly braces*.

Directives in C/C++ are case-sensitive. Directives can be broken over multiple lines by escaping the line end.

The sentinel in Fortran looks like a comment:

```
|| !$omp directive clause(value)
||     statements
|| !$omp end directive
```

The difference with the C directive is that Fortran can not have a block, so there is an explicit *end-of directive* line.

If you break a directive over more than one line, all but the last line need to have a continuation character, and each line needs to have the sentinel:

```
|| !$OMP parallel do &
|| %OMP    copyin(x),copyout(y)
```

The directives are case-insensitive. In *Fortran fixed-form* source files, `C$omp` and `*$omp` are allowed too.

15.3.2 Parallel regions

The simplest way to create parallelism in OpenMP is to use the `parallel` pragma. A block preceded by the `omp parallel` pragma is called a *parallel region*; it is executed by a newly created team of threads. This is an instance of the *Single Program Multiple Data (SPMD)* model: all threads execute the same segment of code.

```

|| #pragma omp parallel
{
    // this is executed by a team of threads
}

```

We will go into much more detail in section 16.

15.3.3 An actual OpenMP program!

Exercise 15.1. Write a program that contains the following lines:

```

printf("There are %d processors\n",omp_get_num_procs());
|| #pragma omp parallel
    printf("There are %d threads\n",
           /* !!!! something missing here !!!! */ );

```

The first print statement tells you the number of available cores in the hardware. Your assignment is to supply the missing function that reports the number of threads used.

Compile and run the program. Experiment with the OMP_NUM_THREADS environment variable. What do you notice about the number of lines printed?

Exercise 15.2. Extend the program from exercise 15.1. Make a complete program based on these lines:

```

int tsum=0;
|| #pragma omp parallel
    tsum += /* the thread number */
printf("Sum is %d\n",tsum);

```

Compile and run again. (In fact, run your program a number of times.) Do you see something unexpected? Can you think of an explanation?

15.3.4 Code and execution structure

Here are a couple of important concepts:

Definition 1

structured block An OpenMP directive is followed by an structured block; in C this is a single statement, a compound statement, or a block in braces; In Fortran it is delimited by the directive and its matching ‘end’ directive.

A structured block can not be jumped into, so it can not start with a labeled statement, or contain a jump statement leaving the block.

construct An OpenMP construct is the section of code starting with a directive and spanning the following structured block, plus in Fortran the end-directive. This is a lexical concept: it contains the statements directly enclosed, and not any subroutines called from them.

region of code A region of code is defined as all statements that are dynamically encountered while executing the code of an OpenMP construct. This is a dynamic concept: unlike a ‘construct’, it does include any subroutines that are called from the code in the structured block.

Chapter 16

OpenMP topic: Parallel regions

The simplest way to create parallelism in OpenMP is to use the `parallel` pragma. A block preceded by the `omp parallel` pragma is called a *parallel region*; it is executed by a newly created team of threads. This is an instance of the *SPMD* model: all threads execute the same segment of code.

```
|| #pragma omp parallel
|| {
||     // this is executed by a team of threads
|| }
```

It would be pointless to have the block be executed identically by all threads. One way to get a meaningful parallel code is to use the function `omp_get_thread_num`, to find out which thread you are, and execute work that is individual to that thread. There is also a function `omp_get_num_threads` to find out the total number of threads. Both these functions give a number relative to the current team; recall from figure 15.3 that new teams can be created recursively.

For instance, if you program computes

```
|| result = f(x) + g(x) + h(x)
```

you could parallelize this as

```
|| double result, fresult, gresult, hresult;
|| #pragma omp parallel
|| {
||     int num = omp_get_thread_num();
||     if (num==0) fresult = f(x);
||     else if (num==1) gresult = g(x);
||     else if (num==2) hresult = h(x);
|| }
|| result = fresult + gresult + hresult;
```

The first thing we want to do is create a team of threads. This is done with a *parallel region*. Here is a very simple example:

```
|| // hello.c
|| #pragma omp parallel
|| {
||     int t = omp_get_thread_num();
||     printf("Hello world from %d!\n", t);
|| }
```

or in Fortran

```
// hellocount.F90
 !$omp parallel
 nthreads = omp_get_num_threads()
 mythread = omp_get_thread_num()
 write(*,'("Hello from",i3," out of",i3)') mythread,nthreads
 !$omp end parallel
```

This code corresponds to the model we just discussed:

- Immediately preceding the parallel block, one thread will be executing the code. In the main program this is the *initial thread*.
- At the start of the block, a new *team of threads* is created, and the thread that was active before the block becomes the *master thread* of that team.
- After the block only the master thread is active.
- Inside the block there is team of threads: each thread in the team executes the body of the block, and it will have access to all variables of the surrounding environment. How many threads there are can be determined in a number of ways; we will get to that later.

Exercise 16.1. Make a full program based on this fragment. Insert different print statements before, inside, and after the parallel region. Run this example. How many times is each print statement executed?

You see that the `parallel` directive

- Is preceded by a special marker: a `#pragma omp` for C/C++, and the `!$OMP sentinel` for Fortran;
- Is followed by a single statement or a block in C/C++, or followed by a block in Fortran which is delimited by an `!$omp end` directive.

Directives look like *cpp directives*, but they are actually handled by the compiler, not the preprocessor.

Exercise 16.2. Take the ‘hello world’ program above, and modify it so that you get multiple messages to your screen, saying

```
Hello from thread 0 out of 4!
Hello from thread 1 out of 4!
```

and so on. (The messages may very well appear out of sequence.)

What happens if you set your number of threads larger than the available cores on your computer?

Exercise 16.3. What happens if you call `omp_get_thread_num` and `omp_get_num_threads` outside a parallel region?

```
||  omp_get_thread_limit
```

`OMP_WAIT_POLICY` values: ACTIVE, PASSIVE

16.1 Nested parallelism

What happens if you call a function from inside a parallel region, and that function itself contains a parallel region?

```
|| int main() {
|| ...
|| #pragma omp parallel
|| {
|| ...
|| func(...)
|| ...
|| }
|| } // end of main
|| void func(...) {
|| #pragma omp parallel
|| {
|| ...
|| }
|| }
```

By default, the nested parallel region will have only one thread. To allow nested thread creation, set

```
OMP_NESTED=true
or
omp_set_nested(1)
```

Exercise 16.4. Test nested parallelism by writing an OpenMP program as follows:

1. Write a subprogram that contains a parallel region.
2. Write a main program with a parallel region; call the subprogram both inside and outside the parallel region.
3. Insert print statements
 - (a) in the main program outside the parallel region,
 - (b) in the parallel region in the main program,
 - (c) in the subprogram outside the parallel region,
 - (d) in the parallel region inside the subprogram.

Run your program and count how many print statements of each type you get.

Writing subprograms that are called in a parallel region illustrates the following point: directives are evaluated with respect to the *dynamic scope* of the parallel region, not just the lexical scope. In the following example:

```
|| #pragma omp parallel
|| {
||   f();
|| }
|| void f() {
|| #pragma omp for
||   for ( .... ) {
||     ...
||   }
|| }
```

the body of the function `f` falls in the dynamic scope of the parallel region, so the for loop will be parallelized.

If the function may be called both from inside and outside parallel regions, you can test which is the case with `omp_in_parallel`.

The amount of nested parallelism can be set:

```
OMP_NUM_THREADS=4,2
```

means that initially a parallel region will have four threads, and each thread can create two more threads.

```
OMP_MAX_ACTIVE_LEVELS=123
```

```
omp_set_max_active_levels( n )
n = omp_get_max_active_levels()
```

```
OMP_THREAD_LIMIT=123
```

```
n = omp_get_thread_limit()
```

```
omp_set_max_active_levels
omp_get_max_active_levels
omp_get_level
omp_get_active_level
omp_get_ancestor_thread_num
```

```
omp_get_team_size(level)
```

16.2 Cancel parallel construct

```
|| !$omp cancel construct [if (expr)]
```

where construct is `parallel`, `sections`, `do` or `taskgroup`

16.3 Sources used in this chapter

Listing of code XX:

Listing of code XX:

Chapter 17

OpenMP topic: Loop parallelism

17.1 Loop parallelism

Loop parallelism is a very common type of parallelism in scientific codes, so OpenMP has an easy mechanism for it. OpenMP parallel loops are a first example of OpenMP ‘worksharing’ constructs (see section 18 for the full list): constructs that take an amount of work and distribute it over the available threads in a parallel region.

The parallel execution of a loop can be handled a number of different ways. For instance, you can create a parallel region around the loop, and adjust the loop bounds:

```
#pragma omp parallel
{
    int threadnum = omp_get_thread_num(),
        numthreads = omp_get_num_threads();
    int low = N*threadnum/numthreads,
        high = N*(threadnum+1)/numthreads;
    for (i=low; i<high; i++)
        // do something with i
}
```

A more natural option is to use the `parallel for` pragma:

```
#pragma omp parallel
#pragma omp for
for (i=0; i<N; i++) {
    // do something with i
}
```

This has several advantages. For one, you don’t have to calculate the loop bounds for the threads yourself, but you can also tell OpenMP to assign the loop iterations according to different schedules (section 17.2).

Figure 17.1 shows the execution on four threads of

```
#pragma omp parallel
{
    code1();
#pragma omp for
    for (i=1; i<=4*N; i++) {
```

```

    code2();
}
code3();
}
}

```

The code before and after the loop is executed identically in each thread; the loop iterations are spread over the four threads.

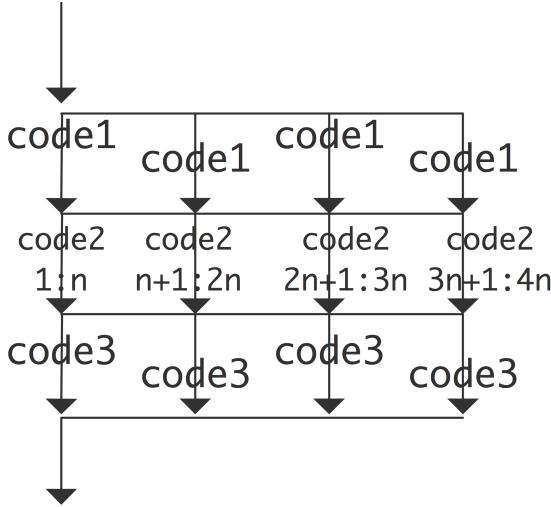


Figure 17.1: Execution of parallel code inside and outside a loop

Note that the `parallel do` and `parallel for` pragmas do not create a team of threads: they take the team of threads that is active, and divide the loop iterations over them.

This means that the `omp for` or `omp do` directive needs to be inside a parallel region. It is also possible to have a combined `omp parallel for` or `omp parallel do` directive.

If your parallel region only contains a loop, you can combine the pragmas for the parallel region and distribution of the loop iterations:

```

#pragma omp parallel for
for (i=0; ....

```

Exercise 17.1. Compute π by *numerical integration*. We use the fact that π is the area of the unit circle, and we approximate this by computing the area of a quarter circle using *Riemann sums*.

- Let $f(x) = \sqrt{1 - x^2}$ be the function that describes the quarter circle for $x = 0 \dots 1$;
- Then we compute

$$\pi/4 \approx \sum_{i=0}^{N-1} \Delta x f(x_i) \quad \text{where } x_i = i\Delta x \text{ and } \Delta x = 1/N$$

Write a program for this, and parallelize it using OpenMP parallel for directives.

1. Put a `parallel` directive around your loop. Does it still compute the right result? Does the time go down with the number of threads? (The answers should be no and no.)
2. Change the `parallel` to `parallel for` (or `parallel do`). Now is the result correct? Does execution speed up? (The answers should now be no and yes.)
3. Put a `critical` directive in front of the update. (Yes and very much no.)
4. Remove the `critical` and add a clause `reduction(+:quarterpi)` to the `for` directive. Now it should be correct and efficient.

Use different numbers of cores and compute the speedup you attain over the sequential computation. Is there a performance difference between the OpenMP code with 1 thread and the sequential code?

Remark 4 *In this exercise you may have seen the runtime go up a couple of times where you weren't expecting it. The issue here is false sharing; see HPSC-?? for more explanation.*

There are some restrictions on the loop: basically, OpenMP needs to be able to determine in advance how many iterations there will be.

- The loop can not contain `break`, `return`, `exit` statements, or `goto` to a label outside the loop.
- The `continue` (C) or `cycle` (F) statement is allowed.
- The index update has to be an increment (or decrement) by a fixed amount.
- The loop index variable is automatically private, and not changes to it inside the loop are allowed.

17.2 Loop schedules

Usually you will have many more iterations in a loop than there are threads. Thus, there are several ways you can assign your loop iterations to the threads. OpenMP lets you specify this with the `schedule` clause.

```
|| #pragma omp for schedule(....)
```

The first distinction we now have to make is between static and dynamic schedules. With static schedules, the iterations are assigned purely based on the number of iterations and the number of threads (and the `chunk` parameter; see later). In dynamic schedules, on the other hand, iterations are assigned to threads that are unoccupied. Dynamic schedules are a good idea if iterations take an unpredictable amount of time, so that *load balancing* is needed.

Figure 17.2 illustrates this: assume that each core gets assigned two (blocks of) iterations and these blocks take gradually less and less time. You see from the left picture that thread 1 gets two fairly long blocks, whereas thread 4 gets two short blocks, thus finishing much earlier. (This phenomenon of threads having unequal amounts of work is known as *load imbalance*.) On the other hand, in the right figure thread 4 gets block 5, since it finishes the first set of blocks early. The effect is a perfect load balancing.

17. OpenMP topic: Loop parallelism

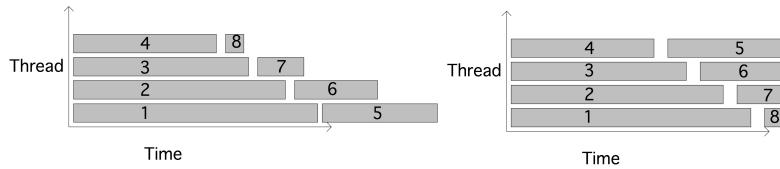


Figure 17.2: Illustration static round-robin scheduling versus dynamic

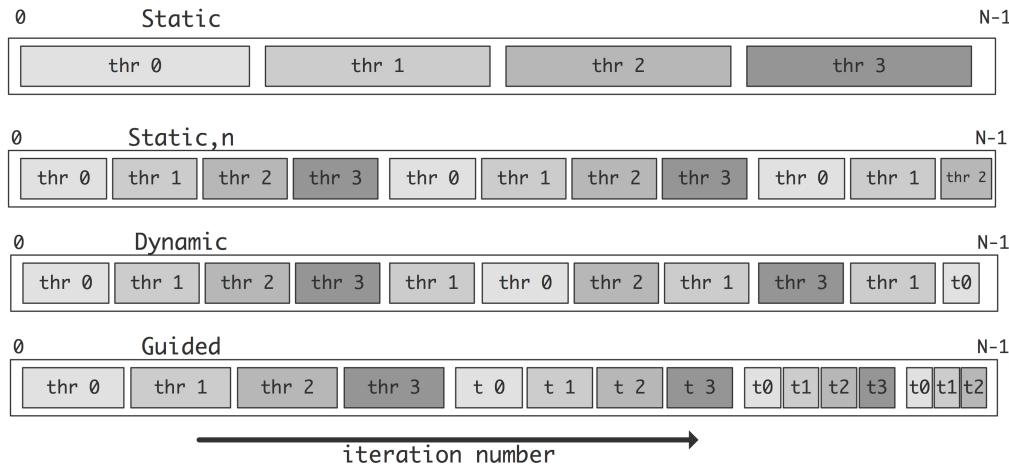


Figure 17.3: Illustration of the scheduling strategies of loop iterations

The default static schedule is to assign one consecutive block of iterations to each thread. If you want different sized blocks you can define a chunk size:

```
|| #pragma omp for schedule(static[,chunk])
```

(where the square brackets indicate an optional argument). With static scheduling, the compiler will split up the loop iterations at compile time, so, provided the iterations take roughly the same amount of time, this is the most efficient at runtime.

The choice of a chunk size is often a balance between the low overhead of having only a few chunks, versus the load balancing effect of having smaller chunks.

Exercise 17.2. Why is a chunk size of 1 typically a bad idea? (Hint: think about cache lines, and read HPSC-??.)

In dynamic scheduling OpenMP will put blocks of iterations (the default chunk size is 1) in a task queue, and the threads take one of these tasks whenever they are finished with the previous.

```
|| #pragma omp for schedule(static[,chunk])
```

While this schedule may give good load balancing if the iterations take very differing amounts of time to execute, it does carry runtime overhead for managing the queue of iteration tasks.

Finally, there is the guided schedule, which gradually decreases the chunk size. The thinking here is

that large chunks carry the least overhead, but smaller chunks are better for load balancing. The various schedules are illustrated in figure 17.3.

If you don't want to decide on a schedule in your code, you can specify the `runtime` schedule. The actual schedule will then at runtime be read from the `OMP_SCHEDULE` environment variable. You can even just leave it to the runtime library by specifying `auto`.

Exercise 17.3. We continue with exercise 17.1. We add ‘adaptive integration’: where needed, the program refines the step size¹. This means that the iterations no longer take a predictable amount of time.

1. It doesn't actually do this in a mathematically sophisticated way, so this code is more for the sake of the example.

```
|| for (i=0; i<
      nsteps;
      i++) {
double
  x = i*h,
  x2 = (i
         +1)*h,
  y = sqrt
    (1-x*x),
  y2 =
    sqrt(1-
      x2*x2),
  slope = (
    y-y2)/h;
if (slope
    >15)
  slope =
    15;
int
  samples =
    1+(int)
  slope,
  is;
for (is=0;
      is<
      samples;
      is++) {
double
  hs = h/
  samples,
  xs = x+
  is*hs,
  ys =
  sqrt(1-
  xs*xs);
  quarterpi
    += hs*
  ys;
  nsamples
   ++;
}
}
pi = 4*
  quarterpi
;
```

1. Use the `omp parallel for` construct to parallelize the loop. As in the previous lab, you may at first see an incorrect result. Use the `reduction` clause to fix this.
2. Your code should now see a decent speedup, using up to 8 cores. However, it is possible to get completely linear speedup. For this you need to adjust the schedule.

Start by using `schedule(static, n)`. Experiment with values for n .

When can you get a better speedup? Explain this.

3. Since this code is somewhat dynamic, try `schedule(dynamic)`. This will actually give a fairly bad result. Why? Use `schedule(dynamic, n)` instead, and experiment with values for n .
4. Finally, use `schedule(guided)`, where OpenMP uses a heuristic. What results does that give?

Exercise 17.4. Program the *LU factorization* algorithm without pivoting.

```

|| for k=1, n:
||   A[k, k] = 1./A[k, k]
||   for i=k+1, n:
||     A[i, k] = A[i, k]/A[k, k]
||     for j=k+1, n:
||       A[i, j] = A[i, j] - A[i, k]*A[k, j]

```

1. Argue that it is not possible to parallelize the outer loop.
2. Argue that it is possible to parallelize both the i and j loops.
3. Parallelize the algorithm by focusing on the i loop. Why is the algorithm as given here best for a matrix on row-storage? What would you do if the matrix was on column storage?
4. Argue that with the default schedule, if a row is updated by one thread in one iteration, it may very well be updated by another thread in another. Can you find a way to schedule loop iterations so that this does not happen? What practical reason is there for doing so?

The schedule can be declared explicitly, set at runtime through the `OMP_SCHEDULE` environment variable, or left up to the runtime system by specifying `auto`. Especially in the last two cases you may want to enquire what schedule is currently being used with `omp_get_schedule`.

```
// int omp_get_schedule(omp_sched_t * kind, int * modifier);
```

Its mirror call is `omp_set_schedule`, which sets the value that is used when schedule value `runtime` is used. It is in effect equivalent to setting the environment variable `OMP_SCHEDULE`.

```
// void omp_set_schedule (omp_sched_t kind, int modifier);
```

Type	environment variable OMP_SCHEDULE=	clause <code>schedule(...)</code>	modifier default
static	static[,n]	static[,n]	$N/n\text{threads}$
dynamic	dynamic[,n]	dynamic[,n]	1
guided	guided[,n]	guided[,n]	

Here are the various schedules you can set with the `schedule` clause:

affinity Set by using value `omp_sched_affinity`

auto The schedule is left up to the implementation. Set by using value `omp_sched_auto`

dynamic value: 2. The modifier parameter is the *chunk size*; default 1. Set by using value `omp_sched_dynamic`

guided Value: 3. The modifier parameter is the *chunk* size. Set by using value `omp_sched_guided`
runtime Use the value of the `OMP_SCHEDULE` environment variable. Set by using value `omp_sched_runtime`
static value: 1. The modifier parameter is the *chunk* size. Set by using value `omp_sched_static`

17.3 Reductions

So far we have focused on loops with independent iterations. Reductions are a common type of loop with dependencies. There is an extended discussion of reductions in section 20.

17.4 Collapsing nested loops

In general, the more work there is to divide over a number of threads, the more efficient the parallelization will be. In the context of parallel loops, it is possible to increase the amount of work by parallelizing all levels of loops instead of just the outer one.

Example: in

```
|| for ( i=0; i<N; i++ )  
||   for ( j=0; j<N; j++ )  
||     A[i][j] = B[i][j] + C[i][j]
```

all N^2 iterations are independent, but a regular `omp_for` directive will only parallelize one level. The `collapse` clause will parallelize more than one level:

```
#pragma omp for collapse(2)  
for ( i=0; i<N; i++ )  
  for ( j=0; j<N; j++ )  
    A[i][j] = B[i][j] + C[i][j]
```

It is only possible to collapse perfectly nested loops, that is, the loop body of the outer loop can consist only of the inner loop; there can be no statements before or after the inner loop in the loop body of the outer loop. That is, the two loops in

```
|| for ( i=0; i<N; i++ ) {  
||   y[i] = 0.;  
||   for ( j=0; j<N; j++ )  
||     y[i] += A[i][j] * x[j]  
|| }
```

can not be collapsed.

Exercise 17.5. Can you rewrite the preceding code example so that it can be collapsed? Do timing tests to see if you can notice the improvement from collapsing.

17.5 Ordered iterations

Iterations in a parallel loop that are execution in parallel do not execute in lockstep. That means that in

```
#pragma omp parallel for
for ( ... i ... ) {
    ... f(i) ...
    printf("something with %d\n", i);
}
```

it is not true that all function evaluations happen more or less at the same time, followed by all print statements. The print statements can really happen in any order. The `ordered` clause coupled with the `ordered` directive can force execution in the right order:

```
#pragma omp parallel for ordered
for ( ... i ... ) {
    ... f(i) ...
    #pragma omp ordered
    printf("something with %d\n", i);
}
```

Example code structure:

```
#pragma omp parallel for shared(y) ordered
for ( ... i ... ) {
    int x = f(i)
    #pragma omp ordered
    y[i] += f(x)
    z[i] = g(y[i])
}
```

There is a limitation: each iteration can encounter only one `ordered` directive.

17.6 nowait

The implicit barrier at the end of a work sharing construct can be cancelled with a `nowait` clause. This has the effect that threads that are finished can continue with the next code in the parallel region:

```
#pragma omp parallel
{
    #pragma omp for nowait
    for (i=0; i<N; i++) { ... }
    // more parallel code
}
```

In the following example, threads that are finished with the first loop can start on the second. Note that this requires both loops to have the same schedule. We specify the static schedule here to have an identical scheduling of iterations over threads:

```
#pragma omp parallel
{
    x = local_computation()
#pragma omp for schedule(static) nowait
    for (i=0; i<N; i++) {
        x[i] = ...
    }
#pragma omp for schedule(static)
    for (i=0; i<N; i++) {
        y[i] = ... x[i] ...
    }
}
```

17.7 While loops

OpenMP can only handle ‘for’ loops: *while loops* can not be parallelized. So you have to find a way around that. While loops are for instance used to search through data:

```
while ( a[i] !=0 && i<imax ) {
    i++;
    // now i is the first index for which \n{a[i]} is zero.
```

We replace the while loop by a for loop that examines all locations:

```
result = -1;
#pragma omp parallel for
for (i=0; i<imax; i++) {
    if (a[i] !=0 && result<0) result = i;
}
```

Exercise 17.6. Show that this code has a race condition.

You can fix the race condition by making the condition into a critical section; section 21.2.1. In this particular example, with a very small amount of work per iteration, that is likely to be inefficient in this case (why?). A more efficient solution uses the `lastprivate` pragma:

```
result = -1;
#pragma omp parallel for lastprivate(result)
for (i=0; i<imax; i++) {
    if (a[i] !=0) result = i;
}
```

You have now solved a slightly different problem: the `result` variable contains the *last* location where `a[i]` is zero.

Chapter 18

OpenMP topic: Work sharing

The declaration of a *parallel region* establishes a team of threads. This offers the possibility of parallelism, but to actually get meaningful parallel activity you need something more. OpenMP uses the concept of a *work sharing construct*: a way of dividing parallelizable work over a team of threads. The work sharing constructs are:

- `for` (for C) or `do` (for Fortran). The threads divide up the loop iterations among themselves; see [17.1](#).
- `sections` The threads divide a fixed number of sections between themselves; see section [18.1](#).
- `single` The section is executed by a single thread; section [18.2](#).
- `task` See section [22](#).
- `workshare` Can parallelize Fortran array syntax; section [18.3](#).

18.1 Sections

A parallel loop is an example of independent work units that are numbered. If you have a pre-determined number of independent work units, the `sections` is more appropriate. In a `sections` construct can be any number of `section` constructs. These need to be independent, and they can be execute by any available thread in the current team, including having multiple sections done by the same thread.

```
#pragma omp sections
{
    #pragma omp section
        // one calculation
    #pragma omp section
        // another calculation
}
```

This construct can be used to divide large blocks of independent work. Suppose that in the following line, both `f(x)` and `g(x)` are big calculations:

```
|| y = f(x) + g(x)
```

You could then write

```

double y1,y2;
#pragma omp sections
{
#pragma omp section
    y1 = f(x)
#pragma omp section
    y2 = g(x)
}
y = y1+y2;

```

Instead of using two temporaries, you could also use a critical section; see section 21.2.1. However, the best solution is have a reduction clause on the `sections` directive:

```

|| y = f(x) + g(x)

```

You could then write

```

y = 0;
#pragma omp sections reduction(+:y)
{
#pragma omp section
    y += f(x)
#pragma omp section
    y += g(x)
}

```

18.2 Single/master

The `single` and `master` pragma limit the execution of a block to a single thread. This can for instance be used to print tracing information or doing I/O operations.

```

#pragma omp parallel
{
#pragma omp single
    printf("We are starting this section!\n");
    // parallel stuff
}

```

Another use of `single` is to perform initializations in a parallel region:

```

int a;
#pragma omp parallel
{
#pragma omp single
    a = f(); // some computation
#pragma omp sections
    // various different computations using a
}

```

The point of the `single` directive in this last example is that the computation needs to be done only once, because of the shared memory. Since it's a work sharing construct there is an *implicit barrier* after it, which guarantees that all threads have the correct value in their local memory (see section 24.3).

Exercise 18.1. What is the difference between this approach and how the same computation would be parallelized in MPI?

The `master` directive, also enforces execution on a single thread, specifically the master thread of the team, but it does not have the synchronization through the implicit barrier.

Exercise 18.2. Modify the above code to read:

```
||| int a;
||| #pragma omp parallel
||| {
|||     #pragma omp master
|||     a = f(); // some computation
|||     #pragma omp sections
|||         // various different computations using a
||| }
```

This code is no longer correct. Explain.

Above we motivated the `single` directive as a way of initializing shared variables. It is also possible to use `single` to initialize private variables. In that case you add the `copyprivate` clause. This is a good solution if setting the variable takes I/O.

Exercise 18.3. Give two other ways to initialize a private variable, with all threads receiving the same value. Can you give scenarios where each of the three strategies would be preferable?

18.3 Fortran array syntax parallelization

The `parallel do` directive is used to parallelize loops, and this applies to both C and Fortran. However, Fortran also has implied loops in its *array syntax*. To parallelize array syntax you can use the `workshare` directive.

The `workshare` directive exists only in Fortran. It can be used to parallelize the implied loops in *array syntax*, as well as *forall* loops.

Chapter 19

OpenMP topic: Controlling thread data

In a parallel region there are two types of data: private and shared. In this sections we will see the various way you can control what category your data falls under; for private data items we also discuss how their values relate to shared data.

19.1 Shared data

In a parallel region, any data declared outside it will be shared: any thread using a variable `x` will access the same memory location associated with that variable.

Example:

```
|| int x = 5;
|| #pragma omp parallel
|| {
||     x = x+1;
||     printf("shared: x is %d\n", x);
|| }
```

All threads increment the same variable, so after the loop it will have a value of five plus the number of threads; or maybe less because of the data races involved. See HPSC-?? for an explanation of the issues involved; see 21.2.1 for a solution in OpenMP.

Sometimes this global update is what you want; in other cases the variable is intended only for intermediate results in a computation. In that case there are various ways of creating data that is local to a thread, and therefore invisible to other threads.

19.2 Private data

In the C/C++ language it is possible to declare variables inside a *lexical scope*; roughly: inside curly braces. This concept extends to OpenMP parallel regions and directives: any variable declared in a block following an OpenMP directive will be local to the executing thread.

Example:

```
|| int x = 5;
|| #pragma omp parallel
|| {
||     int x; x = 3;
||     printf("local: x is %d\n", x);
|| }
```

After the parallel region the outer variable `x` will still have the value 5: there is no *storage association* between the private variable and global one.

The Fortran language does not have this concept of scope, so you have to use a `private` clause:

```
|| !$OMP parallel private(x)
```

The `private` directive declares data to have a separate copy in the memory of each thread. Such private variables are initialized as they would be in a main program. Any computed value goes away at the end of the parallel region. (However, see below.) Thus, you should not rely on any initial value, or on the value of the outer variable after the region.

```
|| int x = 5;
|| #pragma omp parallel private(x)
|| {
||     x = x+1; // dangerous
||     printf("private: x is %d\n", x);
|| }
|| printf("after: x is %d\n", x); // also dangerous
```

Data that is declared private with the `private` directive is put on a separate *stack per thread*. The OpenMP standard does not dictate the size of these stacks, but beware of *stack overflow*. A typical default is a few megabyte; you can control it with the environment variable `OMP_STACKSIZE`. Its values can be literal or with suffixes:

```
123 456k 567K 678m 789M 246g 357G
```

A normal *Unix process* also has a stack, but this is independent of the OpenMP stacks for private data. You can query or set the Unix stack with `ulimit`:

```
[] ulimit -s
64000
[] ulimit -s 8192
[] ulimit -s
8192
```

The Unix stack can grow dynamically as space is needed. This does not hold for the OpenMP stacks: they are immediately allocated at their requested size. Thus it is important not too make them too large.

19.3 Data in dynamic scope

Functions that are called from a parallel region fall in the *dynamic scope* of that parallel region. The rules for variables in that function are as follows:

- Any variables locally defined to the function are private.
- static variables in C and save variables in Fortran are shared.
- The function arguments inherit their status from the calling environment.

19.4 Temporary variables in a loop

It is common to have a variable that is set and used in each loop iteration:

```
|| #pragma omp parallel for
|| for ( ... i ... ) {
||   x = i*h;
||   s = sin(x); c = cos(x);
||   a[i] = s+c;
||   b[i] = s-c;
|| }
```

By the above rules, the variables `x`, `s`, `c` are all shared variables. However, the values they receive in one iteration are not used in a next iteration, so they behave in fact like private variables to each iteration.

- In both C and Fortran you can declare these variables private in the parallel for directive.
- In C, you can also redefine the variables inside the loop.

Sometimes, even if you forget to declare these temporaries as private, the code may still give the correct output. That is because the compiler can sometimes eliminate them from the loop body, since it detects that their values are not otherwise used.

19.5 Default

- Loop variables in an `omp for` are private;
- Local variables in the parallel region are private.

You can alter this default behaviour with the `default` clause:

```
|| #pragma omp parallel default(shared) private(x)
|| { ... }
|| #pragma omp parallel default(private) shared(matrix)
|| { ... }
```

and if you want to play it safe:

```
|| #pragma omp parallel default(none) private(x) shared(matrix)
|| { ... }
```

- The `shared` clause means that all variables from the outer scope are shared in the parallel region; any private variables need to be declared explicitly. This is the default behaviour.
- The `private` clause means that all outer variables become private in the parallel region. They are not initialized; see the next option. Any shared variables in the parallel region need to be declared explicitly. This value is not available in C.

- The `firstprivate` clause means all outer variables are private in the parallel region, and initialized with their outer value. Any shared variables need to be declared explicitly. This value is not available in C.
- The `none` option is good for debugging, because it forces you to specify for each variable in the parallel region whether it's private or shared. Also, if your code behaves differently in parallel from sequential there is probably a data race. Specifying the status of every variable is a good way to debug this.

19.6 Array data

The rules for arrays are slightly different from those for scalar data:

1. Statically allocated data, that is with a syntax like

```
// alloc3.c
int array[100];
||| integer,dimension(:) :: array(100)
```

can be shared or private, depending on the clause you use.

2. Dynamically allocated data, that is, created with `malloc` or `allocate`, can only be shared.

Example of the first type: in

```
// alloc3.c
int array[nthreads];
{
    int t = 2;
    array += t;
    array[0] = t;
}
```

each thread gets a private copy of the array, properly initialized.

On the other hand, in

```
// alloc1.c
int *array = (int*) malloc(nthreads*sizeof(int));
#pragma omp parallel firstprivate(array)
{
    int t = omp_get_thread_num();
    array += t;
    array[0] = t;
}
```

each thread gets a private pointer, but all pointers point to the same object.

19.7 First and last private

Above, you saw that private variables are completely separate from any variables by the same name in the surrounding scope. However, there are two cases where you may want some *storage association* between a private variable and a global counterpart.

First of all, private variables are created with an undefined value. You can force their initialization with `firstprivate`.

```

|| int t=2;
|| #pragma omp parallel firstprivate(t)
|| {
||     t += f( omp_get_thread_num() );
||     g(t);
|| }
```

The variable `t` behaves like a private variable, except that it is initialized to the outside value.

Secondly, you may want a private value to be preserved to the environment outside the parallel region. This really only makes sense in one case, where you preserve a private variable from the last iteration of a parallel loop, or the last section in an `sections` construct. This is done with `lastprivate`:

```

|| #pragma omp parallel for \
||         lastprivate(tmp)
|| for (i=0; i<N; i++) {
||     tmp = .....
||     x[i] = .... tmp ....
|| }
|| ..... tmp ....
```

19.8 Persistent data through `threadprivate`

Most data in OpenMP parallel regions is either inherited from the master thread and therefore shared, or temporary within the scope of the region and fully private. There is also a mechanism for *thread-private data*, which is not limited in lifetime to one parallel region. The `threadprivate` pragma is used to declare that each thread is to have a private copy of a variable:

```
|| #pragma omp threadprivate(var)
```

The variable needs be:

- a file or static variable in C,
- a static class member in C++, or
- a program variable or common block in Fortran.

19.8.1 Thread private initialization

If each thread needs a different value in its `threadprivate` variable, the initialization needs to happen in a parallel region.

In the following example a team of 7 threads is created, all of which set their thread-private variable. Later, this variable is read by a larger team: the variables that have not been set are undefined, though often simply zero:

```
// threadprivate.c
#include <stdlib.h>
#include <stdio.h>
#include <omp.h>

static int tp;

int main(int argc, char **argv) {

#pragma omp threadprivate(tp)

#pragma omp parallel num_threads(7)
tp = omp_get_thread_num();

#pragma omp parallel num_threads(9)
printf("Thread %d has %d\n", omp_get_thread_num(), tp);

return 0;
}
```

On the other hand, if the thread private data starts out identical in all threads, the `copyin` clause can be used:

```
#pragma omp threadprivate(private_var)

private_var = 1;
#pragma omp parallel copyin(private_var)
private_var += omp_get_thread_num()
```

If one thread needs to set all thread private data to its value, the `copyprivate` clause can be used:

```
#pragma omp parallel
{
...
#pragma omp single copyprivate(private_var)
private_var = read_data();
...
}
```

19.8.2 Thread private example

The typical application for thread-private variables is in *random number generation*. A random number generator needs saved state, since it computes each next value from the current one. To have a parallel generator, each thread will create and initialize a private ‘current value’ variable. This will persist even when the execution is not in a parallel region; it gets updated only in a parallel region.

Exercise 19.1. Calculate the area of the *Mandelbrot set* by random sampling. Initialize the random number generator separately for each thread; then use a parallel loop to evaluate the points. Explore performance implications of the different loop scheduling strategies.

Fortran note. Named common blocks can be made thread-private with the syntax

```
|| $!OMP threadprivate( /blockname/ )
```

Threadprivate variables require `OMP_DYNAMIC` to be switched off.

19.9 Sources used in this chapter

Listing of code XX:

Listing of code XX:

Listing of code XX:

Chapter 20

OpenMP topic: Reductions

Parallel tasks often produce some quantity that needs to be summed or otherwise combined. In section 16 you saw an example, and it was stated that the solution given there was not very good.

The problem in that example was the *race condition* involving the `result` variable. The simplest solution is to eliminate the race condition by declaring a *critical section*:

```
double result = 0;
#pragma omp parallel
{
    double local_result;
    int num = omp_get_thread_num();
    if (num==0)    local_result = f(x);
    else if (num==1) local_result = g(x);
    else if (num==2) local_result = h(x);
    #pragma omp critical
        result += local_result;
}
```

This is a good solution if the amount of serialization in the critical section is small compared to computing the functions f, g, h . On the other hand, you may not want to do that in a loop:

```
double result = 0;
#pragma omp parallel
{
    double local_result;
    #pragma omp for
    for (i=0; i<N; i++) {
        local_result = f(x, i);
    #pragma omp critical
        result += local_result;
    } // end of for loop
}
```

Exercise 20.1. Can you think of a small modification of this code, that still uses a critical section, that is more efficient? Time both codes.

The easiest way to effect a reduction is of course to use the `reduction` clause. Adding this to an `omp for` or an `omp sections` construct has the following effect:

- OpenMP will make a copy of the reduction variable per thread, initialized to the identity of the reduction operator, for instance 1 for multiplication.
- Each thread will then reduce into its local variable;
- At the end of the loop, the local results are combined, again using the reduction operator, into the global variable.

This is one of those cases where the parallel execution can have a slightly different value from the one that is computed sequentially, because floating point operations are not associative. See HPSC-?? for more explanation.

If your code can not be easily structure as a reduction, you can realize the above scheme by hand by ‘duplicating’ the global variable and gather the contributions later. This example presumes three threads, and gives each a location of their own to store the result computed on that thread:

```
double result, local_results[3];
#pragma omp parallel
{
    int num = omp_get_thread_num();
    if (num==0)      local_results[num] = f(x)
    else if (num==1) local_results[num] = g(x)
    else if (num==2) local_results[num] = h(x)
}
result = local_results[0]+local_results[1]+local_results[2]
```

While this code is correct, it may be inefficient because of a phenomemon called *false sharing*. Even though the threads write to separate variables, those variables are likely to be on the same *cacheline* (see HPSC-?? for an explanation). This means that the cores will be wasting a lot of time and bandwidth updating each other’s copy of this cacheline.

False sharing can be prevent by giving each thread its own cacheline:

```
double result, local_results[3][8];
#pragma omp parallel
{
    int num = omp_get_thread_num();
    if (num==0)      local_results[num][1] = f(x)
    // et cetera
}
```

A more elegant solution gives each thread a true local variable, and uses a critical section to sum these, at the very end:

```
double result = 0;
#pragma omp parallel
{
    double local_result;
    local_result = .....
#pragma omp critical
    result += local_result;
}
```

20.1 Built-in reduction operators

Arithmetic reductions: $+$, $*$, $-$, \max , \min

Logical operator reductions in C: $\&$ $\&\&$ $|$ $||$ \wedge

Logical operator reductions in Fortran: $.and.$ $.or.$ $.eqv.$ $.neqv.$ $.iand.$ $.ior.$ $.ieor.$

Exercise 20.2. The maximum and minimum reductions were not added to OpenMP until version 3.1. Write a parallel loop that computes the maximum and minimum values in an array. Discuss the various options. Do timings to evaluate the speedup that is attained and to find the best option.

20.2 Initial value for reductions

The treatment of initial values in reductions is slightly involved.

```
x = init_x
#pragma omp parallel for reduction(min:x)
for (int i=0; i<N; i++)
    x = min(x, data[i]);
```

Each thread does a partial reduction, but its initial value is not the user-supplied `init_x` value, but a value dependent on the operator. In the end, the partial results will then be combined with the user initial value. The initialization values are mostly self-evident, such as zero for addition and one for multiplication. For `min` and `max` they are respectively the maximal and minimal representable value of the result type.

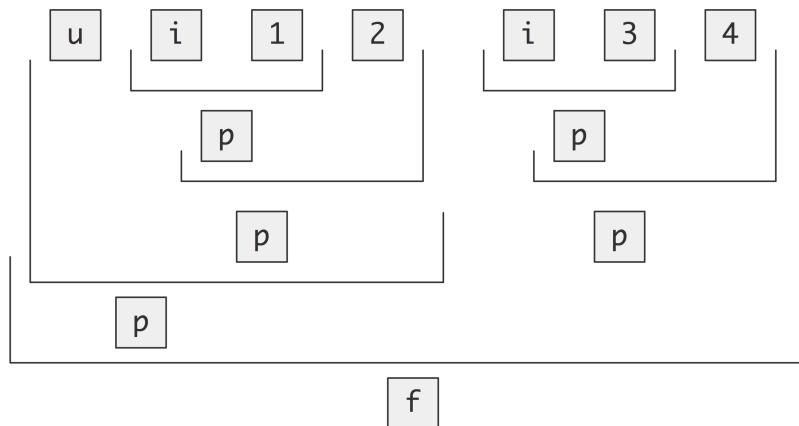


Figure 20.1: Reduction of four items on two threads, taking into account initial values.

Figure 20.1 illustrates this, where `1, 2, 3, 4` are four data items, `i` is the OpenMP initialization, and `u` is the user initialization; each `p` stands for a partial reduction value. The figure is based on execution using two threads.

Exercise 20.3. Write a program to test the fact that the partial results are initialized to the unit of the reduction operator.

20.3 User-defined reductions

With *user-defined reductions*, the programmer specifies the function that does the elementwise comparison. This takes two steps.

1. You need a function of two arguments that returns the result of the comparison. You can do this yourself, but, especially with the C++ standard library, you can use functions such as `std::vector::insert`.
2. Specifying how this function operates on two variables `omp_out` and `omp_in`, corresponding to the partially reduced result and the new operand respectively. The new partial result should be left in `omp_out`.
3. Optionally, you can specify the value to which the reduction should be initialized.

This is the syntax of the definition of the reduction, which can then be used in multiple `reduction` clauses.

```
|| #pragma omp declare reduction
  ( identifier : typelist : combiner )
  [initializer(initializer-expression)]
```

where:

identifier is a name; this can be overloaded for different types, and redefined in inner scopes.

typelist is a list of types.

combiner is an expression that updates the internal variable `omp_out` as function of itself and `omp_in`.

initializer sets `omp_priv` to the identity of the reduction; this can be an expression or a brace initializer.

For instance, recreating the maximum reduction would look like this:

```
// ireduct.c
int mymax(int r,int n) {
    // r is the already reduced value
    // n is the new value
    int m;
    if (n>r) {
        m = n;
    } else {
        m = r;
    }
    return m;
}
#pragma omp declare reduction \
(rwz:int:omp_out=mymax(omp_out,omp_in)) \
initializer(omp_priv=INT_MIN)
m = INT_MIN;
#pragma omp parallel for reduction(rwz:m)
for (int idata=0; idata<ndata; idata++)
    m = mymax(m,data[idata]);
```

Exercise 20.4. Write a reduction routine that operates on an array of non-negative integers, finding the smallest nonzero one. If the array has size zero, or entirely consists of zeros, return -1.

Support for C++ iterators

```
|| #pragma omp declare reduction (merge : std::vector<int>
||   : omp_out.insert(omp_out.end(), omp_in.begin(), omp_in.end()))
```

20.4 Reductions and floating-point math

The mechanisms that OpenMP uses to make a reduction parallel go against the strict rules for floating point expression evaluation in C; see HPSC-?. OpenMP ignores this issue: it is the programmer's job to ensure proper rounding behaviour.

20.5 Sources used in this chapter

Listing of code XX:

Chapter 21

OpenMP topic: Synchronization

In the constructs for declaring parallel regions above, you had little control over in what order threads executed the work they were assigned. This section will discuss *synchronization* constructs: ways of telling threads to bring a certain order to the sequence in which they do things.

- `critical`: a section of code can only be executed by one thread at a time; see [21.2.1](#).
- `atomic` Update of a single memory location. Only certain specified syntax patterns are supported. This was added in order to be able to use hardware support for atomic updates.
- `barrier`: section [21.1](#).
- `ordered`: section [17.5](#).
- `locks`: section [21.3](#).
- `flush`: section [24.3](#).
- `nowait`: section [17.6](#).

21.1 Barrier

A barrier defines a point in the code where all active threads will stop until all threads have arrived at that point. With this, you can guarantee that certain calculations are finished. For instance, in this code snippet, computation of `y` can not proceed until another thread has computed its value of `x`.

```
#pragma omp parallel
{
    int mytid = omp_get_thread_num();
    x[mytid] = some_calculation();
    y[mytid] = x[mytid]+x[mytid+1];
}
```

This can be guaranteed with a `barrier` pragma:

```
#pragma omp parallel
{
    int mytid = omp_get_thread_num();
    x[mytid] = some_calculation();
    #pragma omp barrier
    y[mytid] = x[mytid]+x[mytid+1];
}
```

Apart from the barrier directive, which inserts an explicit barrier, OpenMP has *implicit barriers* after a load sharing construct. Thus the following code is well defined:

```
#pragma omp parallel
{
    #pragma omp for
    for (int mytid=0; mytid<number_of_threads; mytid++)
        x[mytid] = some_calculation();
    #pragma omp for
    for (int mytid=0; mytid<number_of_threads-1; mytid++)
        y[mytid] = x[mytid]+x[mytid+1];
}
```

You can also put each parallel loop in a parallel region of its own, but there is some overhead associated with creating and deleting the team of threads in between the regions.

21.1.1 Implicit barriers

At the end of a parallel region the team of threads is dissolved and only the master thread continues. Therefore, there is an *implicit barrier at the end of a parallel region*.

There is some *barrier behaviour* associated with `omp for` loops and other *worksharing constructs* (see section 18.3). For instance, there is an *implicit barrier* at the end of the loop. This barrier behaviour can be cancelled with the `nowait` clause.

You will often see the idiom

```
#pragma omp parallel
{
    #pragma omp for nowait
    for (i=0; i<N; i++)
        a[i] = // some expression
    #pragma omp for
    for (i=0; i<N; i++)
        b[i] = ..... a[i] .....
```

Here the `nowait` clause implies that threads can start on the second loop while other threads are still working on the first. Since the two loops use the same schedule here, an iteration that uses `a[i]` can indeed rely on it that that value has been computed.

21.2 Mutual exclusion

Sometimes it is necessary to let only one thread execute a piece of code. Such a piece of code is called a *critical section*, and OpenMP has several mechanisms for realizing this.

The most common use of critical sections is to update a variable. Since updating involves reading the old value, and writing back the new, this has the possibility for a *race condition*: another thread reads the current value before the first can update it; the second thread updates to the wrong value.

Critical sections are an easy way to turn an existing code into a correct parallel code. However, there are disadvantages to this, and sometimes a more drastic rewrite is called for.

21.2.1 critical and atomic

There are two pragmas for critical sections: `critical` and `atomic`. The second one is more limited but has performance advantages.

The typical application of a critical section is to update a variable:

```
|| #pragma omp parallel
|| {
||   int mytid = omp_get_thread_num();
||   double tmp = some_function(mytid);
||   #pragma omp critical
||     sum += tmp;
|| }
```

Exercise 21.1. Consider a loop where each iteration updates a variable.

```
|| #pragma omp parallel for shared(result)
||   for ( i ) {
||     result += some_function_of(i);
|| }
```

Discuss qualitatively the difference between:

- turning the update statement into a critical section, versus
- letting the threads accumulate into a private variable `tmp` as above, and summing these after the loop.

Do an Ahmdal-style quantitative analysis of the first case, assuming that you do n iterations on p threads, and each iteration has a critical section that takes a fraction f .

Assume the number of iterations n is a multiple of the number of threads p . Also assume the default static distribution of loop iterations over the threads.

A `critical` section works by acquiring a lock, which carries a substantial overhead. Furthermore, if your code has multiple critical sections, they are all mutually exclusive: if a thread is in one critical section, the other ones are all blocked.

On the other hand, the syntax for `atomic` sections is limited to the update of a single memory location, but such sections are not exclusive and they can be more efficient, since they assume that there is a hardware mechanism for making them critical.

The problem with `critical` sections being mutually exclusive can be mitigated by naming them:

```
|| #pragma omp critical (optional_name_in_parens)
```

21.3 Locks

OpenMP also has the traditional mechanism of a *lock*. A lock is somewhat similar to a critical section: it guarantees that some instructions can only be performed by one process at a time. However, a critical section is indeed about code; a lock is about data. With a lock you make sure that some data elements can only be touched by one process at a time.

One simple example of the use of locks is generation of a *histogram*. A histogram consists of a number of bins, that get updated depending on some data. Here is the basic structure of such a code:

```

|| int count[100];
|| float x = some_function();
|| int ix = (int)x;
|| if (ix>=100)
||     error();
|| else
||     count[ix]++;

```

It would be possible to guard the last line:

```

|| #pragma omp critical
||     count[ix]++;

```

but that is unnecessarily restrictive. If there are enough bins in the histogram, and if the `some_function` takes enough time, there are unlikely to be conflicting writes. The solution then is to create an array of locks, with one lock for each `count` location.

Create/destroy:

```

|| void omp_init_lock(omp_lock_t *lock);
|| void omp_destroy_lock(omp_lock_t *lock);

```

Set and release:

```

|| void omp_set_lock(omp_lock_t *lock);
|| void omp_unset_lock(omp_lock_t *lock);

```

Since the set call is blocking, there is also

```

|| omp_test_lock();

```

Unsetting a lock needs to be done by the thread that set it.

Lock operations implicitly have a flush.

Exercise 21.2. In the following code, one process sets array A and then uses it to update B; the other process sets array B and then uses it to update A. Argue that this code can deadlock. How could you fix this?

```

|| #pragma omp parallel shared(a, b, nthreads, locka, lockb)
|| #pragma omp sections nowait
|| {
||     #pragma omp section
||     {
||         omp_set_lock(&locka);
||         for (i=0; i<N; i++)
||             a[i] = ...
|
||         omp_set_lock(&lockb);
||         for (i=0; i<N; i++)
||             b[i] = ... a[i] ...
||         omp_unset_lock(&lockb);
||         omp_unset_lock(&locka);
||     }

```

```
#pragma omp section
{
    omp_set_lock(&lockb);
    for (i=0; i<N; i++)
        b[i] = ...

    omp_set_lock(&locka);
    for (i=0; i<N; i++)
        a[i] = .. b[i] ..
    omp_unset_lock(&locka);
    omp_unset_lock(&lockb);
}
/* end of sections */
} /* end of parallel region */
```

21.3.1 Nested locks

A lock as explained above can not be locked if it is already locked. A *nested lock* can be locked multiple times by the same thread before being unlocked.

- `omp_init_nest_lock`
- `omp_destroy_nest_lock`
- `omp_set_nest_lock`
- `omp_unset_nest_lock`
- `omp_test_nest_lock`

lock—)

21.4 Example: Fibonacci computation

The *Fibonacci sequence* is recursively defined as

$$F(0) = 1, \quad F(1) = 1, \quad F(n) = F(n - 1) + F(n - 2) \text{ for } n \geq 2.$$

We start by sketching the basic single-threaded solution. The naive code looks like:

```
int main() {
    value = new int[nmax+1];
    value[0] = 1;
    value[1] = 1;
    fib(10);
}

int fib(int n) {
    int i, j, result;
    if (n>=2) {
        i=fib(n-1); j=fib(n-2);
        value[n] = i+j;
    }
    return value[n];
}
```

However, this is inefficient, since most intermediate values will be computed more than once. We solve this by keeping track of which results are known:

```
...
done = new int[nmax+1];
for (i=0; i<=nmax; i++)
    done[i] = 0;
done[0] = 1;
done[1] = 1;
...
int fib(int n) {
    int i, j;
    if (!done[n]) {
        i = fib(n-1); j = fib(n-2);
        value[n] = i+j; done[n] = 1;
    }
    return value[n];
}
```

The OpenMP parallel solution calls for two different ideas. First of all, we parallelize the recursion by using tasks (section 22):

```
int fib(int n) {
    int i, j;
    if (n>=2) {
#pragma omp task shared(i) firstprivate(n)
        i=fib(n-1);
#pragma omp task shared(j) firstprivate(n)
        j=fib(n-2);
#pragma omp taskwait
        value[n] = i+j;
    }
    return value[n];
}
```

This computes the right solution, but, as in the naive single-threaded solution, it recomputes many of the intermediate values.

A naive addition of the done array leads to data races, and probably an incorrect solution:

```
int fib(int n) {
    int i, j, result;
    if (!done[n]) {
#pragma omp task shared(i) firstprivate(n)
        i=fib(n-1);
#pragma omp task shared(i) firstprivate(n)
        j=fib(n-2);
#pragma omp taskwait
        value[n] = i+j;
        done[n] = 1;
    }
    return value[n];
}
```

For instance, there is no guarantee that the `done` array is updated later than the `value` array, so a thread can think that `done[n-1]` is true, but `value[n-1]` does not have the right value yet.

One solution to this problem is to use a lock, and make sure that, for a given index `n`, the values `done[n]` and `value[n]` are never touched by more than one thread at a time:

```
int fib(int n)
{
    int i, j;
    omp_set_lock( &(dolock[n]) );
    if (!done[n]) {
        #pragma omp task shared(i) firstprivate(n)
        i = fib(n-1);
        #pragma omp task shared(j) firstprivate(n)
        j = fib(n-2);
        #pragma omp taskwait
        value[n] = i+j;
        done[n] = 1;
    }
    omp_unset_lock( &(dolock[n]) );
    return value[n];
}
```

This solution is correct, optimally efficient in the sense that it does not recompute anything, and it uses tasks to obtain a parallel execution.

However, the efficiency of this solution is only up to a constant. A lock is still being set, even if a value is already computed and therefore will only be read. This can be solved with a complicated use of critical sections, but we will forego this.

Chapter 22

OpenMP topic: Tasks

Tasks are a mechanism that OpenMP uses under the cover: if you specify something as being parallel, OpenMP will create a ‘block of work’: a section of code plus the data environment in which it occurred. This block is set aside for execution at some later point.

Let’s look at a simple example using the `task` directive.

Code	Execution
<code>x = f();</code>	the variable <code>x</code> gets a value
<code>#pragma omp task</code> <code>{ y = g(x); }</code>	a task is created with the current value of <code>x</code>
<code>z = h();</code>	the variable <code>z</code> gets a value

The thread that executes this code segment creates a task, which will later be executed, probably by a different thread. The exact timing of the execution of the task is up to a *task scheduler*, which operates invisible to the user.

The task mechanism allows you to do things that are hard or impossible with the loop and section constructs. For instance, a *while loop* traversing a *linked list* can be implemented with tasks:

Code	Execution
<code>p = head_of_list();</code>	one thread traverses the list
<code>while (!end_of_list(p)) {</code>	
<code>#pragma omp task</code>	a task is created,
<code>process(p);</code>	one for each element
<code>p = next_element(p);</code>	the generating thread goes on without waiting
<code>}</code>	the tasks are executed while more are being generated.

The way tasks and threads interact is different from the worksharing constructs you’ve seen so far. Typically, one thread will generate the tasks, adding them to a queue, from which all threads can take and execute them. This leads to the following idiom:

```
|| #pragma omp parallel
|| #pragma omp single
|| {
||   ...
|| #pragma omp task
||   { ... }
```

```
|| } ...
```

1. A parallel region creates a team of threads;
2. a single thread then creates the tasks, adding them to a queue that belongs to the team,
3. and all the threads in that team (possibly including the one that generated the tasks)

With tasks it becomes possible to parallelize processes that did not fit the earlier OpenMP constructs. For instance, if a certain operation needs to be applied to all elements of a linked list, you can have one thread go down the list, generating a task for each element of the list.

Another concept that was hard to parallelize earlier is the ‘while loop’. This does not fit the requirement for OpenMP parallel loops that the loop bound needs to be known before the loop executes.

Exercise 22.1. Use tasks to find the smallest factor of a large number (using $2999 \cdot 3001$ as test case): generate a task for each trial factor. Start with this code:

```
int factor=0;
#pragma omp parallel
#pragma omp single
for (int f=2; f<4000; f++) {
    // see if 'f' is a factor
    if (N%f==0) { // found factor!
        factor = f;
    }
}
if (factor>0)
    break;
}
if (factor>0)
    printf("Found a factor: %d\n", factor);
```

- Turn the factor finding block into a task.
- Run your program a number of times:

```
for i in `seq 1 1000` ; do ./taskfactor ; done | grep -v 2999
```

Does it find the wrong factor? Why? Try to fix this.

- Once a factor has been found, you should stop generating tasks. Let tasks that should not have been generated, meaning that they test a candidate larger than the factor found, print out a message.

22.1 Task data

Treatment of data in a task is somewhat subtle. The basic problem is that a task gets created at one time, and executed at another. Thus, if shared data is accessed, does the task see the value at creation time or at execution time? In fact, both possibilities make sense depending on the application, so we need to discuss the rules when which possibility applies.

The first rule is that shared data is shared in the task, but private data becomes `firstprivate`. To see the distinction, consider two code fragments. In the first example:

```
|| int count = 100;
|| #pragma omp parallel
|| #pragma omp single
{
    while (count>0) {
# pragma omp task
    {
        int countcopy = count;
        if (count==50) {
            sleep(1);
            printf("%d,%d\n", count, countcopy);
        } // end if
    } // end task
    count--;
} // end while
} // end single
```

the variable `count` is declared outside the parallel region and is therefore shared. When the print statement is executed, all tasks will have been generated, and so `count` will be zero. Thus, the output will likely be `0, 50`.

In the second example:

```
|| #pragma omp parallel
|| #pragma omp single
{
    int count = 100;
    while (count>0) {
# pragma omp task
    {
        int countcopy = count;
        if (count==50) {
            sleep(1);
            printf("%d,%d\n", count, countcopy);
        } // end if
    } // end task
    count--;
} // end while
} // end single
```

the `count` variable is private to the thread creating the tasks, and so it will be `firstprivate` in the task, preserving the value that was current when the task was created.

22.2 Task synchronization

Even though the above segment looks like a linear set of statements, it is impossible to say when the code after the `task` directive will be executed. This means that the following code is incorrect:

```
|| x = f();
|| #pragma omp task
|| { y = g(x); }
|| z = h(y);
```

Explanation: when the statement computing z is executed, the task computing y has only been scheduled; it has not necessarily been executed yet.

In order to have a guarantee that a task is finished, you need the `taskwait` directive. The following creates two tasks, which can be executed in parallel, and then waits for the results:

Code	Execution
<code>x = f();</code>	the variable x gets a value
<code>#pragma omp task { y1 = g1(x); }</code>	two tasks are created with the current value of x
<code>#pragma omp task { y2 = g2(x); }</code>	
<code>#pragma omp taskwait</code>	the thread waits until the tasks are finished
<code>z = h(y1)+h(y2);</code>	the variable z is computed using the task results

The `task` pragma is followed by a structured block. Each time the structured block is encountered, a new task is generated. On the other hand `taskwait` is a standalone directive; the code that follows is just code, it is not a structured block belonging to the directive.

Another aspect of the distinction between generating tasks and executing them: usually the tasks are generated by one thread, but executed by many threads. Thus, the typical idiom is:

```

|| #pragma omp parallel
|| #pragma omp single
{
|| // code that generates tasks
}

```

This makes it possible to execute loops in parallel that do not have the right kind of iteration structure for a `omp parallel for`. As an example, you could traverse and process a linked list:

```

|| #pragma omp parallel
|| #pragma omp single
{
|| while (!tail(p)) {
||     p = p->next();
||     #pragma omp task
||         process(p)
||     }
||     #pragma omp taskwait
}

```

One task traverses the linked list creating an independent task for each element in the list. These tasks are then executed in parallel; their assignment to threads is done by the task scheduler.

You can indicate task dependencies in several ways:

1. Using the ‘task wait’ directive you can explicitly indicate the *join* of the *forked* tasks. The instruction after the wait directive will therefore be dependent on the spawned tasks.
2. The `taskgroup` directive, followed by a structured block, ensures completion of all tasks created in the block, even if recursively created.

3. Each OpenMP task can have a depend clause, indicating what *data dependency* of the task. By indicating what data is produced or absorbed by the tasks, the scheduler can construct the dependency graph for you.

Another mechanism for dealing with tasks is the `taskgroup`: a task group is a code block that can contain task directives; all these tasks need to be finished before any statement after the block is executed.

A task group is somewhat similar to having a `taskwait` directive after the block. The big difference is that that `taskwait` directive does not wait for tasks that are recursively generated, while a `taskgroup` does.

22.3 Task dependencies

It is possible to put a partial ordering on tasks through use of the `depend` clause. For example, in

```
#pragma omp task
x = f()
#pragma omp task
y = g(x)
```

it is conceivable that the second task is executed before the first, possibly leading to an incorrect result. This is remedied by specifying:

```
#pragma omp task depend(out:x)
x = f()
#pragma omp task depend(in:x)
y = g(x)
```

Exercise 22.2. Consider the following code:

```
for i in [1:N]:
    x[0,i] = some_function_of(i)
    x[i,0] = some_function_of(i)

for i in [1:N]:
    for j in [1:N]:
        x[i,j] = x[i-1,j]+x[i,j-1]
```

- Observe that the second loop nest is not amenable to OpenMP loop parallelism.
- Can you think of a way to realize the computation with OpenMP loop parallelism? Hint: you need to rewrite the code so that the same operations are done in a different order.
- Use tasks with dependencies to make this code parallel without any rewriting: the only change is to add OpenMP directives.

Tasks dependencies are used to indicated how two uses of one data item relate to each other. Since either use can be a read or a write, there are four types of dependencies.

RaW (Read after Write) The second task reads an item that the first task writes. The second task has to be executed after the first:

```
|| ... omp task depend(OUT:x)
||   foo(x)
|| ... omp task depend( IN:x)
||   foo(x)
```

WaR (Write after Read) The first task reads and item, and the second task overwrites it. The second task has to be executed second to prevent overwriting the initial value:

```
|| ... omp task depend( IN:x)
||   foo(x)
|| ... omp task depend(OUT:x)
||   foo(x)
```

WaW (Write after Write) Both tasks set the same variable. Since the variable can be used by an intermediate task, the two writes have to be executed in this order:

```
|| ... omp task depend(OUT:x)
||   foo(x)
|| ... omp task depend(OUT:x)
||   foo(x)
```

RaR (Read after Read) Both tasks read a variable. Since neither tasks has an ‘out’ declaration, they can run in either order.

```
|| ... omp task depend( IN:x)
||   foo(x)
|| ... omp task depend( IN:x)
||   foo(x)
```

22.4 More

22.4.1 Scheduling points

Normally, a task stays tied to the thread that first executes it. However, at a *task scheduling point* the thread may switch to the execution of another task created by the same team.

- There is a scheduling point after explicit task creation. This means that, in the above examples, the thread creating the tasks can also participate in executing them.
- There is a scheduling point at `taskwait` and `taskyield`.

On the other hand a task created with them `untied` clause on the task pragma is never tied to one thread. This means that after suspension at a scheduling point any thread can resume execution of the task. If you do this, beware that the value of a thread-id does not stay fixed. Also locks become a problem.

Example: if a thread is waiting for a lock, with a scheduling point it can suspend the task and work on another task.

```
|| while (!omp_test_lock(lock))
|| #pragma omp taskyield
|| ;
```

22.4.2 Task cancelling

It is possible (in *OpenMP version 4*) to cancel tasks. This is useful when tasks are used to perform a search: the task that finds the result first can cancel any outstanding search tasks.

The directive `cancel` takes an argument of the surrounding construct (`parallel`, `for`, `sections`, `taskgroup`) in which the tasks are cancelled.

Exercise 22.3. Modify the prime finding example.

22.5 Examples

22.5.1 Fibonacci

As an example of the use of tasks, consider computing an array of Fibonacci values:

```
// taskgroup0.c
for (int i=2; i<N; i++)
{
    fibo_values[i] = fibo_values[i-1]+fibo_values[i-2];
```

If you simply turn each calculation into a task, results will be unpredictable (confirm this!) since tasks can be executed in any sequence. To solve this, we put dependencies on the tasks:

```
// taskgroup2.c
for (int i=2; i<N; i++)
#pragma omp task \
depend(out:fibo_values[i]) \
depend(in:fibo_values[i-1],fibo_values[i-2])
{
    fibo_values[i] = fibo_values[i-1]+fibo_values[i-2];
```

22.5.2 Binomial coefficients

Exercise 22.4. An array of binomial coefficients can be computed as follows:

```
// binomial1.c
for (int row=1; row<=n; row++)
    for (int col=1; col<=row; col++)
        if (row==1 || col==1 || col==row)
            array[row][col] = 1;
        else
            array[row][col] = array[row-1][col-1] + array[row-1][col];
```

Putting a single task group around the double loop, and use `depend` clauses to make the execution satisfy the proper dependencies.

22.5.3 Tree traversal

OpenMP tasks are a great way of handling trees.

22.5.3.1 Post-order traversal

In *post-order tree traversal* you visit the subtrees before visiting the root. This is the traversal that you use to find summary information about a tree, for instance the sum of all nodes, and the sums of nodes of all subtrees:

```
for all children c do
    compute the sum  $s_c$ 
```

$$s \leftarrow \sum_c s_c$$

Another example is matrix factorization:

$$S = A_{33} - A_{31}A_{11}^{-1}A_{13} - A_{32}A_{22}^{-1}A_{23}$$

where the two inverses A_{11}^{-1}, A_{22}^{-1} can be computed independently and recursively.

22.5.3.2 Pre-order traversal

If a property needs to propagate from the root to all subtrees and nodes, you can use *pre-order tree traversal*:

```
Update node value  $s$ 
for all children c do
    update c with the new value  $s$ 
```

22.6 Sources used in this chapter

Listing of code XX:

Listing of code XX:

Listing of code XX:

Chapter 23

OpenMP topic: Affinity

23.1 OpenMP thread affinity control

The matter of thread affinity becomes important on *multi-socket nodes*; see the example in section 23.2.

Thread placement can be controlled with two environment variables:

- the environment variable `OMP_PROC_BIND` describes how threads are bound to *OpenMP places*; while
- the variable `OMP_PLACES` describes these places in terms of the available hardware.
- When you're experimenting with these variables it is a good idea to set `OMP_DISPLAY_ENV` to true, so that OpenMP will print out at runtime how it has interpreted your specification. The examples in the following sections will display this output.

23.1.1 Thread binding

The variable `OMP_PLACES` defines a series of places to which the threads are assigned.

Example: if you have two sockets and you define

```
OMP_PLACES=sockets
```

then

- thread 0 goes to socket 0,
- thread 1 goes to socket 1,
- thread 2 goes to socket 0 again,
- and so on.

On the other hand, if the two sockets have a total of sixteen cores and you define

```
OMP_PLACES=cores  
OMP_PROC_BIND=close
```

then

- thread 0 goes to core 0, which is on socket 0,
- thread 1 goes to core 1, which is on socket 0,

- thread 2 goes to core 2, which is on socket 0,
- and so on, until thread 7 goes to core 7 on socket 0, and
- thread 8 goes to core 8, which is on socket 1,
- et cetera.

The value `OMP_PROC_BIND=close` means that the assignment goes successively through the available places. The variable `OMP_PROC_BIND` can also be set to `spread`, which spreads the threads over the places. With

```
OMP_PLACES=cores
OMP_PROC_BIND=spread
```

you find that

- thread 0 goes to core 0, which is on socket 0,
- thread 1 goes to core 8, which is on socket 1,
- thread 2 goes to core 1, which is on socket 0,
- thread 3 goes to core 9, which is on socket 1,
- and so on, until thread 14 goes to core 7 on socket 0, and
- thread 15 goes to core 15, which is on socket 1.

So you see that `OMP_PLACES=cores` and `OMP_PROC_BIND=spread` very similar to `OMP_PLACES=sockets`. The difference is that the latter choice does not bind a thread to a specific core, so the operating system can move threads about, and it can put more than one thread on the same core, even if there is another core still unused.

The value `OMP_PROC_BIND=master` puts the threads in the same place as the master of the team. This is convenient if you create teams recursively. In that case you would use the `proc_bind` clause rather than the environment variable, set to `spread` for the initial team, and to `master` for the recursively created team.

23.1.2 Effects of thread binding

Let's consider two example program. First we consider the program for computing π , which is purely compute-bound.

#threads	close/cores	spread/sockets	spread/cores
1	0.359	0.354	0.353
2	0.177	0.177	0.177
4	0.088	0.088	0.088
6	0.059	0.059	0.059
8	0.044	0.044	0.044
12	0.029	0.045	0.029
16	0.022	0.050	0.022

We see pretty much perfect speedup for the `OMP_PLACES=cores` strategy; with `OMP_PLACES=sockets` we probably get occasional collisions where two threads wind up on the same core.

Next we take a program for computing the time evolution of the *heat equation*:

$$t = 0, 1, 2, \dots : \forall_i : x_i^{(t+1)} = 2x_i^{(t)} - x_{i-1}^{(t)} - x_{i+1}^{(t)}$$

This is a bandwidth-bound operation because the amount of computation per data item is low.

#threads	close/cores	spread/sockets	spread/cores
1	2.88	2.89	2.88
2	1.71	1.41	1.42
4	1.11	0.74	0.74
6	1.09	0.57	0.57
8	1.12	0.57	0.53
12	0.72	0.53	0.52
16	0.52	0.61	0.53

Again we see that `OMP_PLACES=sockets` gives worse performance for high core counts, probably because of threads winding up on the same core. The thing to observe in this example is that with 6 or 8 cores the `OMP_PROC_BIND=spread` strategy gives twice the performance of `OMP_PROC_BIND=close`.

The reason for this is that a single socket does not have enough bandwidth for all eight cores on the socket. Therefore, dividing the eight threads over two sockets gives each thread a higher available bandwidth than putting all threads on one socket.

23.1.3 Place definition

There are three predefined values for the `OMP_PLACES` variable: `sockets`, `cores`, `threads`. You have already seen the first two; the `threads` value becomes relevant on processors that have hardware threads. In that case, `OMP_PLACES=cores` does not tie a thread to a specific hardware thread, leading again to possible collisions as in the above example. Setting `OMP_PLACES=threads` ties each OpenMP thread to a specific hardware thread.

There is also a very general syntax for defining places that uses a

`location:number:stride`

syntax. Examples:

- `OMP_PLACES="0:8:1, 8:8:1"`

is equivalent to `sockets` on a two-socket design with eight cores per socket: it defines two places, each having eight consecutive cores. The threads are then placed alternating between the two places, but not further specified inside the place.

- The setting `cores` is equivalent to

`OMP_PLACES="0, 1, 2, ..., 15"`

- On a four-socket design, the specification

`OMP_PLACES="0:4:8:4:1"`

states that the place $0, 8, 16, 24$ needs to be repeated four times, with a stride of one. In other words, thread 0 winds up on core 0 of some socket, the thread 1 winds up on core 1 of some socket, et cetera.

23.1.4 Binding possibilities

Values for `OMP_PROC_BIND` are: `false`, `true`, `master`, `close`, `spread`.

- `false`: set no binding
- `true`: lock threads to a core
- `master`: collocate threads with the master thread
- `close`: place threads close to the master in the places list
- `spread`: spread out threads as much as possible

This effect can be made local by giving the `proc_bind` clause in the `parallel` directive.

A safe default setting is

```
export OMP_PROC_BIND=true
```

which prevents the operating system from *migrating a thread*. This prevents many scaling problems.

Good examples of *thread placement* on the *Intel Knight's Landing*: <https://software.intel.com/en-us/articles/process-and-thread-affinity-for-intel-xeon-phi-processors-x200>

As an example, consider a code where two threads write to a shared location.

```
// sharing.c
#pragma omp parallel
{ // not a parallel for: just a bunch of reps
    for (int j = 0; j < reps; j++) {
#pragma omp for schedule(static,1)
        for (int i = 0; i < N; i++) {
#pragma omp atomic
            a++;
        }
    }
}
```

There is now a big difference in runtime depending on how close the threads are. We test this on a processor with both cores and hyperthreads. First we bind the OpenMP threads to the cores:

```
OMP_NUM_THREADS=2 OMP_PLACES=cores OMP_PROC_BIND=close ./sharing
run time = 4752.231836usec
sum = 80000000.0
```

Next we force the OpenMP threads to bind to hyperthreads inside one core:

```
OMP_PLACES=threads OMP_PROC_BIND=close ./sharing
run time = 941.970110usec
```

```
sum = 80000000.0
```

Of course in this example the inner loop is pretty much meaningless and parallelism does not speed up anything:

```
OMP_NUM_THREADS=1 OMP_PLACES=cores OMP_PROC_BIND=close ./sharing
run time = 806.669950usec
sum = 80000000.0
```

However, we see that the two-thread result is almost as fast, meaning that there is very little parallelization overhead.

23.2 First-touch

The affinity issue shows up in the *first-touch* phenomenon. Memory allocated with `malloc` and like routines is not actually allocated; that only happens when data is written to it. In light of this, consider the following OpenMP code:

```
double *x = (double*) malloc(N*sizeof(double));

for (i=0; i<N; i++)
    x[i] = 0;

#pragma omp parallel for
for (i=0; i<N; i++)
    .... something with x[i] ...
```

Since the initialization loop is not parallel it is executed by the master thread, making all the memory associated with the socket of that thread. Subsequent access by the other socket will then access data from memory not attached to that socket.

Exercise 23.1. Finish the following fragment and run it with first all the cores of one socket, then all cores of both sockets. (If you know how to do explicit placement, you can also try fewer cores.)

```
for (int i=0; i<nlocal+2; i++)
    in[i] = 1.;

for (int i=0; i<nlocal; i++)
    out[i] = 0.;

for (int step=0; step<nsteps; step++) {
#pragma omp parallel for schedule(static)
    for (int i=0; i<nlocal; i++) {
        out[i] = (in[i]+in[i+1]+in[i+2])/3.;
    }
#pragma omp parallel for schedule(static)
    for (int i=0; i<nlocal; i++)
        in[i+1] = out[i];
    in[0] = 0; in[nlocal+1] = 1;
}
```

Exercise 23.2. How do the OpenMP dynamic schedules relate to this?

C++ valarray does initialization, so it will allocate memory on thread 0.

You could move pages with `move_pages`.

By regarding affinity, in effect you are adopting an SPMD style of programming. You could make this explicit by having each thread allocate its part of the arrays separately, and storing a private pointer as `threadprivate` [13]. However, this makes it impossible for threads to access each other's parts of the distributed array, so this is only suitable for total *data parallel* or *embarrassingly parallel* applications.

23.3 Affinity control outside OpenMP

There are various utilities to control process and thread placement.

Process placement can be controlled on the Operating system level by `numactl` (the TACC utility `tacc_affinity` is a wrapper around this) on Linux (also `taskset`); Windows `start/affinity`.

Corresponding system calls: `pbing` on Solaris, `sched_setaffinity` on Linux, `SetThreadAffinityMask` on Windows.

Corresponding environment variables: `SUNW_MP_PROCBIND` on Solaris, `KMP_AFFINITY` on Intel.

The *Intel compiler* has an environment variable for affinity control:

```
export KMP_AFFINITY=verbose,scatter
```

values: `none`, `scatter`, `compact`

For *gcc*:

```
export GOMP_CPU_AFFINITY=0,8,1,9
```

For the *Sun compiler*:

```
SUNW_MP_PROCBIND
```

23.4 Sources used in this chapter

Listing of code XX:

Chapter 24

OpenMP topic: Memory model

24.1 Thread synchronization

Let's do a *producer-consumer* model¹. This can be implemented with sections, where one section, the producer, sets a flag when data is available, and the other, the consumer, waits until the flag is set.

```
#pragma omp parallel sections
{
    // the producer
    #pragma omp section
    {
        ... do some producing work ...
        flag = 1;
    }
    // the consumer
    #pragma omp section
    {
        while (flag==0) { }
        ... do some consuming work ...
    }
}
```

One reason this doesn't work, is that the compiler will see that the flag is never used in the producing section, and that is never changed in the consuming section, so it may optimize these statements, to the point of optimizing them away.

The producer then needs to do:

```
... do some producing work ...
#pragma omp flush
#pragma atomic write
flag = 1;
#pragma omp flush(flag)
```

and the consumer does:

1. This example is from Intel's excellent OMP course by Tim Mattson

```
|| #pragma omp flush(flag)
|| while (flag==0) {
||     #pragma omp flush(flag)
|| }
|| #pragma omp flush
```

This code strictly speaking has a *race condition* on the `flag` variable. It is better to use an `atomic` pragma here: the producer has

```
|| #pragma atomic write
||     flag = 1;
```

and the consumer:

```
|| while (1) {
||     #pragma omp flush(flag)
||     #pragma omp atomic read
||     flag_read = flag
||     if (flag_read==1) break;
|| }
```

24.2 Data races

OpenMP, being based on shared memory, has a potential for *race conditions*. These happen when two threads access the same data item. The problem with race conditions is that programmer convenience runs counter to efficient execution. For this reason, OpenMP simply does not allow some things that would be desirable.

For a simple example:

Code:

```
// race.c
#pragma omp parallel for shared(counter)
for (int i=0; i<count; i++)
    counter++;
printf("Counter should be %d, is %d\n",
       count, counter);
```

Output:

The basic rule about multiple-thread access of a single data item is:

Any memory location that is *written* by one thread, can not be *read* by another thread in the same parallel region, if no synchronization is done.

To start with that last clause: any workshare construct ends with an *implicit barrier*, so data written before that barrier can safely be read after it.

As an illustration of a possible problem:

```
c = d = 0;
#pragma omp sections
{
```

```
|| #pragma omp section
|| { a = 1; c = b; }
|| #pragma omp section
|| { b = 1; d = a; }
|| }
```

Under any reasonable interpretation of parallel execution, the possible values for `c`, `d` are 1, 1 0, 1 or 1, 0. This is known as *sequential consistency*: the parallel outcome is consistent with a sequential execution that interleaves the parallel computations, respecting their local statement orderings. (See also HPSC-??.)

However, without synchronization, threads are allowed to maintain a value for a variable locally that is not the same as the stored value. In this example, that means that the thread executing the first section need not write its value of `a` to memory, and likewise `b` in the second thread, so 0, 0 is in fact a possible outcome.

In order to resolve multiple accesses:

1. Thread one reads the variable.
2. Thread one flushes the variable.
3. Thread two flushes the variable.
4. Thread two reads the variable.

24.3 Relaxed memory model

`flush`

- There is an implicit flush of all variables at the start and end of a *parallel region*.
- There is a flush at each barrier, whether explicit or implicit, such as at the end of a *work sharing*.
- At entry and exit of a *critical section*
- When a *lock* is set or unset.

Chapter 25

OpenMP topic: SIMD processing

You can declare a loop to be executable with *vector instructions* with `simd`

The `simd` pragma has the following clauses:

- `safelen(n)`: limits the number of iterations in a SIMD chunk. Presumably useful if you combine `parallel` for `simd`.
- `linear`: lists variables that have a linear relation to the iteration parameter.
- `aligned`: specifies alignment of variables.

If your SIMD loop includes a function call, you can declare that the function can be turned into vector instructions with `declare simd`

If a loop is both multi-threadable and vectorizable, you can combine directives as `pragma omp parallel for simd`.

Compilers can be made to report whether a loop was vectorized:

```
LOOP BEGIN at simdf.c(61,15)
    remark #15301: OpenMP SIMD LOOP WAS VECTORIZED
LOOP END
```

with such options as `-Qvec-report=3` for the Intel compiler.

Performance improvements of these directives need not be immediately obvious. In cases where the operation is bandwidth-limited, using `simd` parallelism may give the same or worse performance as thread parallelism.

The following function can be vectorized:

```
// tools.c
#pragma omp declare simd
double cs(double x1,double x2,double y1,double y2) {
    double
        inprod = x1*x2+y1*y2,
        xnorm = sqrt(x1*x1 + x2*x2),
        ynorm = sqrt(y1*y1 + y2*y2);
    return inprod / (xnorm*ynorm);
}
```

```

||#pragma omp declare simd uniform(x1,x2,y1,y2) linear(i)
||double csa(double *x1,double *x2,double *y1,double *y2, int i) {
||    double
||        inprod = x1[i]*x2[i]+y1[i]*y2[i],
||        xnorm = sqrt(x1[i]*x1[i] + x2[i]*x2[i]),
||        ynorm = sqrt(y1[i]*y1[i] + y2[i]*y2[i]);
||    return inprod / (xnorm*ynorm);
||}

```

Compiling this the regular way

```

# parameter 1(x1): %xmm0
# parameter 2(x2): %xmm1
# parameter 3(y1): %xmm2
# parameter 4(y2): %xmm3

movaps    %xmm0, %xmm5      5 <- x1
movaps    %xmm2, %xmm4      4 <- y1
mulsd    %xmm1, %xmm5      5 <- 5 * x2 = x1 * x2
mulsd    %xmm3, %xmm4      4 <- 4 * y2 = y1 * y2
mulsd    %xmm0, %xmm0      0 <- 0 * 0 = x1 * x1
mulsd    %xmm1, %xmm1      1 <- 1 * 1 = x2 * x2
addsd    %xmm4, %xmm5      5 <- 5 + 4 = x1*x2 + y1*y2
mulsd    %xmm2, %xmm2      2 <- 2 * 2 = y1 * y1
mulsd    %xmm3, %xmm3      3 <- 3 * 3 = y2 * y2
addsd    %xmm1, %xmm0      0 <- 0 + 1 = x1*x1 + x2*x2
addsd    %xmm3, %xmm2      2 <- 2 + 3 = y1*y1 + y2*y2
sqrtsd   %xmm0, %xmm0      0 <- sqrt(0) = sqrt( x1*x1 + x2*x2 )
sqrtsd   %xmm2, %xmm2      2 <- sqrt(2) = sqrt( y1*y1 + y2*y2 )

```

which uses the scalar instruction `mulsd`: multiply scalar double precision.

With a `declare simd` directive:

```

movaps    %xmm0, %xmm7
movaps    %xmm2, %xmm4
mulpd    %xmm1, %xmm7
mulpd    %xmm3, %xmm4

```

which uses the vector instruction `mulpd`: multiply packed double precision, operating on 128-bit SSE2 registers.

Compiling for the *Intel Knight's Landing* gives more complicated code:

```

# parameter 1(x1): %xmm0
# parameter 2(x2): %xmm1
# parameter 3(y1): %xmm2
# parameter 4(y2): %xmm3

```

```

vmulpd    %xmm3, %xmm2, %xmm4          4 <- y1*y2
vmulpd    %xmm1, %xmm1, %xmm5          5 <- x1*x2
vbroadcastsd .L_2i10floatpacket.0(%rip), %zmm21
movl      $3, %eax                     set accumulator EAX
vbroadcastsd .L_2i10floatpacket.5(%rip), %zmm24
kmovw     %eax, %k3                   set mask k3
vmulpd    %xmm3, %xmm3, %xmm6          6 <-y1*y1 (stall)
vfmadd231pd %xmm0, %xmm1, %xmm4          4 <- 4 + x1*x2 (no
vfmadd213pd %xmm5, %xmm0, %xmm0          0 <- 0 + 0*5 = x1 +
vmovaps   %zmm21, %zmm18             #25.26 c7
vmovapd   %zmm0, %zmm3{ %k3}{z}        #25.26 c11
vfmadd213pd %xmm6, %xmm2, %xmm2          #24.29 c13
vpcmpgtq  %zmm0, %zmm21, %k1{ %k3}      #25.26 c13
vscalefpd .L_2i10floatpacket.1(%rip){1to8}, %zmm0, %zmm3{ %k1} #25.26 c15
vmovaps   %zmm4, %zmm26             #25.26 c15
vmovapd   %zmm2, %zmm7{ %k3}{z}        #25.26 c17
vpcmpgtq  %zmm2, %zmm21, %k2{ %k3}      #25.26 c17
vscalefpd .L_2i10floatpacket.1(%rip){1to8}, %zmm2, %zmm7{ %k2} #25.26 c19
vrsqrt28pd %zmm3, %zmm16{ %k3}{z}      #25.26 c19
vpxorq    %zmm4, %zmm4, %zmm26{ %k3}    #25.26 c19
vrsqrt28pd %zmm7, %zmm20{ %k3}{z}      #25.26 c21
vmulpd    {rn-sae}, %zmm3, %zmm16, %zmm19{ %k3}{z} #25.26 c27 stall 2
vscalefpd .L_2i10floatpacket.2(%rip){1to8}, %zmm16, %zmm17{ %k3}{z} #25.26 c
vmulpd    {rn-sae}, %zmm7, %zmm20, %zmm23{ %k3}{z} #25.26 c29
vscalefpd .L_2i10floatpacket.2(%rip){1to8}, %zmm20, %zmm22{ %k3}{z} #25.26 c
vfnmadd231pd {rn-sae}, %zmm17, %zmm19, %zmm18{ %k3} #25.26 c33 stall 1
vfnmadd231pd {rn-sae}, %zmm22, %zmm23, %zmm21{ %k3} #25.26 c35
vfmadd231pd {rn-sae}, %zmm19, %zmm18, %zmm19{ %k3} #25.26 c39 stall 1
vfmadd231pd {rn-sae}, %zmm23, %zmm21, %zmm23{ %k3} #25.26 c41
vfmadd213pd {rn-sae}, %zmm17, %zmm17, %zmm18{ %k3} #25.26 c45 stall 1
vfnmadd231pd {rn-sae}, %zmm19, %zmm19, %zmm3{ %k3} #25.26 c47
vfmadd213pd {rn-sae}, %zmm22, %zmm22, %zmm21{ %k3} #25.26 c51 stall 1
vfnmadd231pd {rn-sae}, %zmm23, %zmm23, %zmm7{ %k3} #25.26 c53
vfmadd213pd %zmm19, %zmm18, %zmm3{ %k3} #25.26 c57 stall 1
vfmadd213pd %zmm23, %zmm21, %zmm7{ %k3} #25.26 c59
vscalefpd .L_2i10floatpacket.3(%rip){1to8}, %zmm3, %zmm3{ %k1} #25.26 c63 st
vscalefpd .L_2i10floatpacket.3(%rip){1to8}, %zmm7, %zmm7{ %k2} #25.26 c65
vfixupimmpd $112, .L_2i10floatpacket.4(%rip){1to8}, %zmm0, %zmm3{ %k3} #25.2
vfixupimmpd $112, .L_2i10floatpacket.4(%rip){1to8}, %zmm2, %zmm7{ %k3} #25.2
vmulpd    %xmm7, %xmm3, %xmm0          #25.26 c71
vmovaps   %zmm0, %zmm27             #25.26 c79
vmovaps   %zmm0, %zmm25             #25.26 c79

```

```
vrcp28pd  {sae}, %zmm0, %zmm27{%k3}          #25.26 c81
vfnmadd213pd {rn-sae}, %zmm24, %zmm27, %zmm25{%k3} #25.26 c89 stall 3
vfmadd213pd {rn-sae}, %zmm27, %zmm25, %zmm27{%k3} #25.26 c95 stall 2
vcmpdd    $8, %zmm26, %zmm27, %k1{%k3}          #25.26 c101 stall 2
vmulpd    %zmm27, %zmm4, %zmm1{%k3}{z}          #25.26 c101
kortestw   %k1, %k1                                #25.26 c103
je         ..B1.3        # Prob 25%                #25.26 c105
vdivpd    %zmm0, %zmm4, %zmm1{%k1}                #25.26 c3 stall 1
vmovaps   %xmm1, %xmm0                            #25.26 c77
ret

|| #pragma omp declare simd uniform(op1) linear(k) notinbranch
|| double SqrtMul(double *op1, double op2, int k) {
||     return (sqrt(op1[k]) * sqrt(op2));
|| }
```

25.1 Sources used in this chapter

Listing of code XX:

Chapter 26

OpenMP remaining topics

26.1 Runtime functions and internal control variables

OpenMP has a number of settings that can be set through *environment variables*, and both queried and set through *library routines*. These settings are called *Internal Control Variables (ICVs)*: an OpenMP implementation behaves as if there is an internal variable storing this setting.

The runtime functions are:

- `omp_set_num_threads`
- `omp_get_num_threads`
- `omp_get_max_threads`
- `omp_get_thread_num`
- `omp_get_num_procs`
- `omp_in_parallel`
- `omp_set_dynamic`
- `omp_get_dynamic`
- `omp_set_nested`
- `omp_get_nested`
- `omp_get_wtime`
- `omp_get_wtick`
- `omp_set_schedule`
- `omp_get_schedule`
- `omp_set_max_active_levels`
- `omp_get_max_active_levels`
- `omp_get_thread_limit`
- `omp_get_level`
- `omp_get_active_level`
- `omp_get_ancestor_thread_num`
- `omp_get_team_size`

Here are the OpenMP *environment variables*:

- `OMP_CANCELLATION` Set whether cancellation is activated
- `OMP_DISPLAY_ENV` Show OpenMP version and environment variables

- `OMP_DEFAULT_DEVICE` Set the device used in target regions
- `OMP_DYNAMIC` Dynamic adjustment of threads
- `OMP_MAX_ACTIVE_LEVELS` Set the maximum number of nested parallel regions
- `OMP_MAX_TASK_PRIORITY` Set the maximum task priority value
- `OMP_NESTED` Nested parallel regions
- `OMP_NUM_THREADS` Specifies the number of threads to use
- `OMP_PROC_BIND` Whether threads may be moved between CPUs
- `OMP_PLACES` Specifies on which CPUs the threads should be placed
- `OMP_STACKSIZE` Set default thread stack size
- `OMP_SCHEDULE` How threads are scheduled
- `OMP_THREAD_LIMIT` Set the maximum number of threads
- `OMP_WAIT_POLICY` How waiting threads are handled; ICV *wait-policy-var*. Values: ACTIVE for keeping threads spinning, PASSIVE for possibly yielding the processor when threads are waiting.

There are 4 ICVs that behave as if each thread has its own copy of them. The default is implementation-defined unless otherwise noted.

- It may be possible to adjust dynamically the number of threads for a parallel region. Variable: `OMP_DYNAMIC`; routines: `omp_set_dynamic`, `omp_get_dynamic`.
- If a code contains *nested parallel regions*, the inner regions may create new teams, or they may be executed by the single thread that encounters them. Variable: `OMP_NESTED`; routines `omp_set_nested`, `omp_get_nested`. Allowed values are TRUE and FALSE; the default is false.
- The number of threads used for an encountered parallel region can be controlled. Variable: `OMP_NUM_THREADS`; routines `omp_set_num_threads`, `omp_get_max_threads`.
- The schedule for a parallel loop can be set. Variable: `OMP_SCHEDULE`; routines `omp_set_schedule`, `omp_get_schedule`.

Non-obvious syntax:

```
export OMP_SCHEDULE="static,100"
```

Other settings:

- `omp_get_num_threads`: query the number of threads active at the current place in the code; this can be lower than what was set with `omp_set_num_threads`. For a meaningful answer, this should be done in a parallel region.
- `omp_get_thread_num`
- `omp_in_parallel`: test if you are in a parallel region (see for instance section 16).
- `omp_get_num_procs`: query the physical number of cores available.

Other environment variables:

- `OMP_STACKSIZE` controls the amount of space that is allocated as per-thread stack; the space for private variables.
- `OMP_WAIT_POLICY` determines the behaviour of threads that wait, for instance for *critical section*:
 - ACTIVE puts the thread in a *spin-lock*, where it actively checks whether it can continue;
 - PASSIVE puts the thread to sleep until the Operating System (OS) wakes it up.

The ‘active’ strategy uses CPU while the thread is waiting; on the other hand, activating it after the wait is instantaneous. With the ‘passive’ strategy, the thread does not use any CPU while waiting, but activating it again is expensive. Thus, the passive strategy only makes sense if threads will be waiting for a (relatively) long time.

- `OMP_PROC_BIND` with values `TRUE` and `FALSE` can bind threads to a processor. On the one hand, doing so can minimize data movement; on the other hand, it may increase load imbalance.

26.2 Timing

OpenMP has a wall clock timer routine `omp_get_wtime`

```
|| double omp_get_wtime(void);
```

The starting point is arbitrary and is different for each program run; however, in one run it is identical for all threads. This timer has a resolution given by `omp_get_wtick`.

Exercise 26.1. Use the timing routines to demonstrate speedup from using multiple threads.

- Write a code segment that takes a measurable amount of time, that is, it should take a multiple of the tick time.
- Write a parallel loop and measure the speedup. You can for instance do this

```
|| for (int use_threads=1; use_threads<=nthreads;
       use_threads++) {
    #pragma omp parallel for num_threads(use_threads)
    for (int i=0; i<nthreads; i++) {
        ....
    }
    if (use_threads==1)
        time1 = tend-tstart;
    else // compute speedup
```

- In order to prevent the compiler from optimizing your loop away, let the body compute a result and use a reduction to preserve these results.

26.3 Thread safety

With OpenMP it is relatively easy to take existing code and make it parallel by introducing parallel sections. If you’re careful to declare the appropriate variables shared and private, this may work fine. However, your code may include calls to library routines that include a *race condition*; such code is said not to be *thread-safe*.

For example a routine

```
|| static int isave;
int next_one() {
    int i = isave;
    isave += 1;
    return i;
```

```
    }
    ...
    for ( .... ) {
        int ivalue = next_one();
    }
```

has a clear race condition, as the iterations of the loop may get different `next_one` values, as they are supposed to, or not. This can be solved by using an `critical` pragma for the `next_one` call; another solution is to use an `threadprivate` declaration for `isave`. This is for instance the right solution if the `next_one` routine implements a *random number generator*.

26.4 Performance and tuning

The performance of an OpenMP code can be influenced by the following.

Amdahl effects Your code needs to have enough parts that are parallel (see HPSC-??). Sequential parts may be sped up by having them executed redundantly on each thread, since that keeps data locally.

Dynamism Creating a thread team takes time. In practice, a team is not created and deleted for each parallel region, but creating teams of different sizes, or resize thread creation, may introduce overhead.

Load imbalance Even if your program is parallel, you need to worry about load balance. In the case of a parallel loop you can set the `schedule` clause to `dynamic`, which evens out the work, but may cause increased communication.

Communication Cache coherence causes communication. Threads should, as much as possible, refer to their own data.

- Threads are likely to read from each other's data. That is largely unavoidable.
- Threads writing to each other's data should be avoided: it may require synchronization, and it causes coherence traffic.
- If threads can migrate, data that was local at one time is no longer local after migration.
- Reading data from one socket that was allocated on another socket is inefficient; see section 23.2.

Affinity Both data and execution threads can be bound to a specific locale to some extent. Using local data is more efficient than remote data, so you want to use local data, and minimize the extent to which data or execution can move.

- See the above points about phenomena that cause communication.
- Section 23.1.1 describes how you can specify the binding of threads to places. There can, but does not need, to be an effect on affinity. For instance, if an OpenMP thread can migrate between hardware threads, cached data will stay local. Leaving an OpenMP thread completely free to migrate can be advantageous for load balancing, but you should only do that if data affinity is of lesser importance.
- Static loop schedules have a higher chance of using data that has affinity with the place of execution, but they are worse for load balancing. On the other hand, the `nowait` clause can alleviate some of the problems with static loop schedules.

Binding You can choose to put OpenMP threads close together or to spread them apart. Having them close together makes sense if they use lots of shared data. Spreading them apart may increase bandwidth. (See the examples in section [23.1.2](#).)

Synchronization Barriers are a form of synchronization. They are expensive by themselves, and they expose load imbalance. Implicit barriers happen at the end of worksharing constructs; they can be removed with `nowait`.

Critical sections imply a loss of parallelism, but they are also slow as they are realized through *operating system* functions. These are often quite costly, taking many thousands of cycles. Critical sections should be used only if the parallel work far outweighs it.

26.5 Accelerators

In OpenMP 4.0 there is support for offloading work to an *accelerator* or *co-processor*:

```
|| #pragma omp target [clauses]
```

with clauses such as

- `data`: place data
- `update`: make data consistent between host and device

Chapter 27

OpenMP Review

27.1 Concepts review

27.1.1 Basic concepts

- process / thread / thread team
- threads / cores / tasks
- directives / library functions / environment variables

27.1.2 Parallel regions

execution by a team

27.1.3 Work sharing

- loop / sections / single / workshare
- implied barrier
- loop scheduling, reduction
- sections
- single vs master
- (F) workshare

27.1.4 Data scope

- shared vs private, C vs F
- loop variables and reduction variables
- default declaration
- firstprivate, lastprivate

27.1.5 Synchronization

- barriers, implied and explicit
- nowait
- critical sections
- locks, difference with critical

27.1.6 Tasks

- generation vs execution
- dependencies

27.2 Review questions

27.2.1 Directives

What do the following program output?

```
int main() {
    printf("procs %d\n",
        omp_get_num_procs());
    printf("threads %d\n",
        omp_get_num_threads());
    printf("num %d\n",
        omp_get_thread_num());
    return 0;
}
```

```
int main() {
#pragma omp parallel
{
    printf("procs %d\n",
        omp_get_num_procs());
    printf("threads %d\n",
        omp_get_num_threads());
    printf("num %d\n",
        omp_get_thread_num());
}
return 0;
}
```

```
Program main
use omp_lib
print *, "Procs:", &
omp_get_num_procs()
print *, "Threads:", &
omp_get_num_threads()
print *, "Num:", &
omp_get_thread_num()
End Program
```

```
Program main
use omp_lib
!$OMP parallel
print *, "Procs:", &
omp_get_num_procs()
print *, "Threads:", &
omp_get_num_threads()
print *, "Num:", &
omp_get_thread_num()
!$OMP end parallel
End Program
```

27.2.2 Parallelism

Can the following loops be parallelized? If so, how? (Assume that all arrays are already filled in, and that there are no out-of-bounds errors.)

```
// variant #1
for (i=0; i<N; i++) {
    x[i] = a[i]+b[i+1];
    a[i] = 2*x[i] + c[i+1];
}
```

```
// variant #3
for (i=1; i<N; i++) {
    x[i] = a[i]+b[i+1];
    a[i] = 2*x[i-1] + c[i+1];
}
```

```
// variant #2
for (i=0; i<N; i++) {
    x[i] = a[i]+b[i+1];
    a[i] = 2*x[i+1] + c[i+1];
}
```

```
// variant #4
for (i=1; i<N; i++) {
    x[i] = a[i]+b[i+1];
    a[i+1] = 2*x[i-1] + c[i+1];
}
```

```
! variant #1
do i=1,N
    x(i) = a(i)+b(i+1)
    a(i) = 2*x(i) + c(i+1)
end do
```

```
! variant #3
do i=2,N
    x(i) = a(i)+b(i+1)
    a(i) = 2*x(i-1) + c(i+1)
end do
```

```
! variant #2
do i=1,N
    x(i) = a(i)+b(i+1)
    a(i) = 2*x(i+1) + c(i+1)
end do
```

```
! variant #3
do i=2,N
    x(i) = a(i)+b(i+1)
    a(i+1) = 2*x(i-1) + c(i+1)
end do
```

27.2.3 Data and synchronization

27.2.3.1

What is the output of the following fragments? Assume that there are four threads.

```
// variant #1
int nt;
#pragma omp parallel
{
    nt = omp_get_thread_num();
    printf("thread number: %d\n", nt
        );
}
```

```
// variant #2
int nt;
#pragma omp parallel private(nt)
{
    nt = omp_get_thread_num();
    printf("thread number: %d\n", nt
        );
}
```

```
// variant #3
int nt;
#pragma omp parallel
{
    #pragma omp single
    {
        nt = omp_get_thread_num();
        printf("thread number: %d\n",
            nt);
    }
}
```

```
// variant #4
int nt;
#pragma omp parallel
{
    #pragma omp master
    {
        nt = omp_get_thread_num();
        printf("thread number: %d\n",
            nt);
    }
}
```

```
// variant #5
int nt;
#pragma omp parallel
{
    #pragma omp critical
    {
        nt = omp_get_thread_num();
        printf("thread number: %d\n",
            nt);
    }
}
```

```
! variant #1
integer nt
!$OMP parallel
    nt = omp_get_thread_num()
    print *, "thread number:", nt
!$OMP end parallel
```

```
! variant #2
integer nt
!$OMP parallel private(nt)
    nt = omp_get_thread_num()
    print *, "thread number:", nt
!$OMP end parallel
```

```
! variant #3
integer nt
!$OMP parallel
!$OMP single
    nt = omp_get_thread_num()
    print *, "thread number:", nt
!$OMP end single
!$OMP end parallel
```

```

! variant #4
integer nt
!$OMP parallel
!$OMP master
    nt = omp_get_thread_num()
    print *, "thread number:", nt
!$OMP end master
!$OMP end parallel

```

```

! variant #5
integer nt
!$OMP parallel
!$OMP critical
    nt = omp_get_thread_num()
    print *, "thread number:", nt
!$OMP end critical
!$OMP end parallel

```

27.2.3.2

The following is an attempt to parallelize a serial code. Assume that all variables and arrays are defined. What errors and potential problems do you see in this code? How would you fix them?

```

#pragma omp parallel
{
    x = f();
    #pragma omp for
    for (i=0; i<N; i++)
        y[i] = g(x, i);
    z = h(y);
}

```

```

!$OMP parallel
x = f()
!$OMP do
do i=1,N
    y(i) = g(x, i)
end do
!$OMP end do
z = h(y)
!$OMP end parallel

```

27.2.3.3

Assume two threads. What does the following program output?

```
int a;
#pragma omp parallel private(a) {
    ...
    a = 0;
    #pragma omp for
    for (int i = 0; i < 10; i++)
    {
        #pragma omp atomic
        a++; }
    #pragma omp single
    printf("a=%e\n", a);
}
```

27.2.4 Reductions

27.2.4.1

Is the following code correct? Is it efficient? If not, can you improve it?

```
#pragma omp parallel shared(r)
{
    int x;
    x = f(omp_get_thread_num());
    #pragma omp critical
    r += f(x);
}
```

27.2.4.2

Compare two fragments:

```
// variant 1
#pragma omp parallel reduction(+:s)
#pragma omp for
for (i=0; i<N; i++)
    s += f(i);
```

```
// variant 2
#pragma omp parallel
#pragma omp for reduction(+:s)
for (i=0; i<N; i++)
    s += f(i);
```

```
! variant 1
!$OMP parallel reduction(+:s)
!$OMP do
do i=1,N
    s += f(i);
```

```
end do
!$OMP end do
!$OMP end parallel
```

```
! variant 2
!$OMP parallel
!$OMP do reduction(+:s)
  do i=1,N
```

```
s += f(i);
end do
!$OMP end do
!$OMP end parallel
```

Do they compute the same thing?

27.2.5 Barriers

Are the following two code fragments well defined?

```
#pragma omp parallel
{
#pragma omp for
for (mytid=0; mytid<nthreads;
     mytid++)
    x[mytid] = some_calculation();
#pragma omp for
for (mytid=0; mytid<nthreads-1;
     mytid++)
    y[mytid] = x[mytid]+x[mytid+1];
}
```

```
#pragma omp parallel
{
#pragma omp for
for (mytid=0; mytid<nthreads;
     mytid++)
    x[mytid] = some_calculation();
#pragma omp for nowait
for (mytid=0; mytid<nthreads-1;
     mytid++)
    y[mytid] = x[mytid]+x[mytid+1];
}
```

27.2.6 Data scope

The following program is supposed to initialize as many rows of the array as there are threads.

```
int main() {
    int i, icount, iarray[100][100];
    icount = -1;
#pragma omp parallel private(i)
    {
#pragma omp critical
        { icount++; }
        for (i=0; i<100; i++)
            iarray[icount][i] = 1;
    }
    return 0;
}
```

```
Program main
integer :: i, icount, iarray
(100,100)
icount = 0
!$OMP parallel private(i)
!$OMP critical
icount = icount + 1
!$OMP end critical
do i=1,100
    iarray(icount,i) = 1
end do
!$OMP end parallel
End program
```

Describe the behaviour of the program, with argumentation,

- as given;
- if you add a clause `private(icount)` to the `parallel` directive;
- if you add a clause `firstprivate(icount)`.

What do you think of this solution:

```
#pragma omp parallel private(i)
    shared(icount)
    {
#pragma omp critical
        { icount++;
```

```
        for (i=0; i<100; i++)
            iarray[icount][i] = 1;
    }
    return 0;
```

{}

```

!$OMP parallel private(i) shared(
    icount)
!$OMP critical
    icount = icount+1
    do i=1,100
        iarray(icount,i) = 1
    end do
!$OMP critical
!$OMP end parallel

```

27.2.7 Tasks

Fix two things in the following example:

```

#pragma omp parallel
#pragma omp single
{
    int x,y,z;
#pragma omp task
    x = f();
#pragma omp task
    y = g();
#pragma omp task
    z = h();
    printf("sum=%d\n",x+y+z);
}

```

```

integer :: x,y,z
 !$OMP parallel
 !$OMP single

 !$OMP task
    x = f()
 !$OMP end task

 !$OMP task
    y = g()
 !$OMP end task

 !$OMP task
    z = h()
 !$OMP end task

 print *, "sum=",x+y+z
 !$OMP end single
 !$OMP end parallel

```

27.2.8 Scheduling

Compare these two fragments. Do they compute the same result? What can you say about their efficiency?

```

#pragma omp parallel
#pragma omp single
{
    for (i=0; i<N; i++) {
        #pragma omp task
        x[i] = f(i)
    }
    #pragma omp taskwait
}

```

```

#pragma omp parallel
#pragma omp for schedule(dynamic)
{
    for (i=0; i<N; i++) {
        x[i] = f(i)
    }
}

```

How would you make the second loop more efficient? Can you do something similar for the first loop?

PART III

PETSC

Chapter 28

PETSc basics

28.1 What is PETSc and why?

PETSc is a library with a great many uses, but for now let's say that it's primarily a library for dealing with the sort of linear algebra that comes from discretized Partial Differential Equations (PDEs). On a single processor, the basics of such computations can be coded out by a grad student during a semester course in numerical analysis, but on large scale issues get much more complicated and a library becomes indispensable.

PETSc's prime justification is then that it helps you realize scientific computations at large scales, meaning large problem sizes on large numbers of processors.

There are two points to emphasize here:

- Linear algebra with dense matrices is relatively simple to formulate. For sparse matrices the amount of logistics in dealing with nonzero patterns increases greatly. PETSc does most of that for you.
- Linear algebra on a single processor, even a multicore one, is manageable; distributed memory parallelism is much harder, and distributed memory sparse linear algebra operations are doubly so. Using PETSc will save you many, many, Many! hours of coding over developing everything yourself from scratch.

28.1.1 What is in PETSc?

The routines in PETSc (of which there are hundreds) can roughly be divided in these classes:

- Basic linear algebra tools: dense and sparse matrices, both sequential and parallel, their construction and simple operations.
- Solvers for linear systems, and to a lesser extent nonlinear systems; also time-stepping methods.
- Profiling and tracing: after a successful run, timing for various routines can be given. In case of failure, there are traceback and memory tracing facilities.

28.1.2 Programming model

PETSc, being based on MPI, uses the SPMD programming model (section 2.1), where all processes execute the same executable. Even more than in regular MPI codes, this makes sense here, since most PETSc objects are collectively created on some communicator, often `MPI_COMM_WORLD`. With the object-oriented design (section 28.1.3) this means that a PETSc program almost looks like a sequential program.

```
|| MatMult (A, x, y);      // y <- Ax
|| VecCopy (y, res);       // r <- y
|| VecAXPY (res, -1., b); // r <- r - b
```

This is sometimes called *sequential semantics*.

28.1.3 Design philosophy

PETSc has an object-oriented design, even though it is written in C. There are classes of objects, such `Mat` for matrices and `Vec` for Vectors, but there is also the `KSP` (for "Krylov SSpace solver") class of linear system solvers, and `PetscViewer` for outputting matrices and vectors to screen or file.

Part of the object-oriented design is the polymorphism of objects: after you have created a `Mat` matrix as sparse or dense, all methods such as `MatMult` (for the matrix-vector product) take the same arguments: the matrix, and an input and output vector.

This design where the programmer manipulates a ‘handle’ also means that the internal of the object, the actual storage of the elements, is hidden from the programmer. This hiding goes so far that even filling in elements is not done directly but through function calls:

```
|| VecSetValue (i, j, v, mode)
|| MatSetValue (i, j, v, mode)
|| MatSetValues (ni, is, nj, js, v, mode)
```

28.1.4 Language support

28.1.4.1 C/C++

PETSc is implemented in C, so there is a natural interface to C. There is no separate C++ interface.

28.1.4.2 Fortran

A `Fortran90` interface exists. The `Fortran77` interface is only of interest for historical reasons.

To use Fortran, include both a module and a cpp header file:

```
#include "petsc/finclude/petscXXX.h"
use petscXXX
```

(here XXX stands for one of the PETSc types, but including `petsc.h` and using `use petsc` gives inclusion of the whole library.)

Variables can be declared with their type (**Vec**, **Mat**, **KSP** et cetera), but internally they are Fortran *Type* objects so they can be declared as such.

Example:

```
#include "petsc/finclude/petscvec.h"
use petscvec
Vec b
type(tVec) x
```

28.1.4.3 Python

A *python* interface was written by Lisandro Dalcin, and requires separate installation, based on already defined PETSC_DIR and PETSC_ARCH variables. This can be downloaded at <https://bitbucket.org/petsc/petsc4py/src/master/>, with documentation at <https://www.mcs.anl.gov/petsc/petsc4py-current/docs/>.

28.1.5 Documentation

PETSc comes with a manual in pdf form and web pages with the documentation for every routine. The starting point is the web page <https://www.mcs.anl.gov/petsc/documentation/index.html>.

There is also a mailing list with excellent support for questions and bug reports.

TACC note. For questions specific to using PETSc on TACC resources, submit tickets to the *TACC* or *XSEDE portal*.

28.2 Basics of running a PETSc program

28.2.1 Compilation

A PETSc compilation needs a number of include and library paths, probably too many to specify interactively. The easiest solution is to create a makefile:

```
include ${PETSC_DIR}/lib/petsc/conf/variables
include ${PETSC_DIR}/lib/petsc/conf/rules
program : program.o
${CLINKER} -o $@ $^ ${PETSC_LIB}
```

The two include lines provide the compilation rule and the library variable. If you want to write your own compiler rule, use

```
include ${PETSC_DIR}/lib/petsc/conf/variables
%.o : %.c
${CC} -c $^ ${PETSC_CC_INCLUDES}
program : program.o
${CLINKER} -o $@ $^ ${PETSC_LIB}
```

(The `PETSC_CC_INCLUDES` variable contains all paths for compilation of C programs; correspondingly there is `PETSC_FC_INCLUDES` for Fortran source.)

The build process assumes that variables `PETSC_DIR` and `PETSC_ARCH` have been set. These depend on your local installation. Usually there will be one installation with debug settings and one with production settings. Develop your code with the former: it will do memory and bound checking. Then recompile and run your code with the optimized production installation.

TACC note. On TACC clusters, a petsc installation is loaded by commands such as

```
module load petsc/3.11
```

Use `module avail petsc` to see what configurations exist. The basic versions are

```
# development
module load petsc/3.11-debug
# production
module load petsc/3.11
```

Other installations are real versus complex, or 64bit integers instead of the default 32.

The command

```
module spider petsc
```

tells you all the available petsc versions. The listed modules have a naming convention such as `petsc/3.11-i64debug` where the 3.11 is the PETSc release (minor patches are not included in this version; TACC aims to install only the latest patch, but generally several versions are available), and `i64debug` describes the debug version of the installation with 64bit integers.

28.2.2 Running

PETSc programs use MPI for parallelism, so they are started like any other MPI program:

```
mpirun -np 5 -machinefile mf \
        your_petsc_program option1 option2 option3
```

TACC note. On TACC clusters, use `ibrun`.

28.2.3 Startup

PETSc has an call that initializes both PETSc and MPI, so normally you would replace `MPI_Init` by `PetscInitialize` (figure 140). Unlike with MPI, you do not want to use a NULL value for the `argc`, `argv` arguments, since PETSc makes extensive use of commandline options; see section 29.10.

Python note. The following works if you don't need commandline options.

```
from petsc4py import PETSc
```

To pass commandline arguments to PETSc, do:

```
import sys
from petsc4py import init
init(sys.argv)
from petsc4py import PETSc
```

After initialization, you can use `MPI_COMM_WORLD` or `PETSC_COMM_WORLD`:

```
||| MPI_Comm comm = PETSC_COMM_WORLD;
||| MPI_Comm_rank(comm, &mytid);
||| MPI_Comm_size(comm, &ntrids);
```

Python note.

```
comm = PETSC.COMM_WORLD
nprocs = comm.getSize(self)
procno = comm.getRank(self)
```

28.3 PETSc installation

PETSc has a large number of installation options. These can roughly be divided into:

1. Options to describe the environment in which PETSc is being installed, such as the names of the compilers or the location of the MPI library;
2. Options to specify the type of PETSc installation: real versus complex, 32 versus 64-bit integers, et cetera;
3. Options to specify additional packages to download.

For an existing installation, you can find the options used in the file

```
$PETSC_DIR/$PETSC_ARCH/lib/petsc/conf/configure.log
```

28.3.1 Environment options

Compilers, compiler options, MPI.

While it is possible to specify `-download_mpich`, this should only be done on machines that you are certain do not already have an MPI library, such as your personal laptop. Supercomputer clusters are likely to have an optimized MPI library, and letting PETSc download its own will lead to degraded performance.

28.3.2 Variants

- Scalars: the option `-with-scalar-type` has values `real`, `complex`; `-with-precision` has values `single`, `double`, `__float128`, `__fp16`.

PetscInitialize

How to read routine prototypes: 1.5.4.

manpage 140: Routine prototype for PetscInitialize

Chapter 29

PETSc objects

29.1 Distributed objects

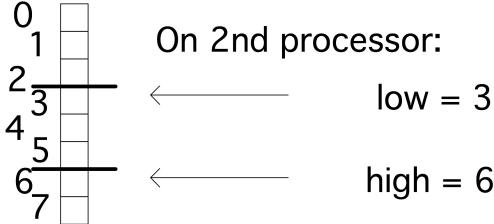
PETSc is based on the SPMD model, and all its objects act like they exist in parallel, spread out over all the processes. Therefore, prior to discussing specific objects in detail, we briefly discuss how PETSc treats distributed objects.

For a matrix or vector you need to specify the size. This can be done two ways:

- you specify the global size and PETSc distributes the object over the processes, or
- you specify on each process the local size

If you specify both the global size and the local sizes, PETSc will check for consistency.

For example, if you have a vector of N components, or a matrix of N rows, and you have P processes, each process will receive N/P components or rows if P divides evenly in N . If P does not divide evenly, the excess is spread over the processes.



The way the distribution is done is by contiguous blocks: with 10 processes and 1000 components in a vector, process 0 gets the range $0 \dots 99$, process 1 gets $1 \dots 199$, et cetera. This simple scheme suffices for many cases, but PETSc has facilities for more sophisticated load balancing.

29.1.1 Support for distributions

Once an object has been created and distributed, you do not need to remember the size or the distribution yourself: you can query these with calls such as `VecGetSize`, `VecGetLocalSize`.

The corresponding matrix routines `MatGetSize`, `MatGetLocalSize` give both information for the distributions in i and j direction, which can be independent. Since a matrix is distributed by rows, `MatGetOwnershipRange` only gives a row range.

```
// split.c
N = 100; n = PETSC_DECIDE;
PetscSplitOwnership(comm, &n, &N);
PetscPrintf(comm, "Global %d, local %d\n", N, n);

N = PETSC_DECIDE; n = 10;
PetscSplitOwnership(comm, &n, &N);
PetscPrintf(comm, "Global %d, local %d\n", N, n);
```

While PETSc objects are implemented using local memory on each process, conceptually they act like global objects, with a global indexing scheme. Thus, each process can query which elements out of the global object are stored locally. For vectors, the relevant routine is `VecGetOwnershipRange`, which returns two parameters, `low` and `high`, respectively the first element index stored, and one-more-than-the-last index stored.

This gives the idiom:

```
|| VecGetOwnershipRange(myvector, &low, &high);
|| for (int myidx=low; myidx<high; myidx++)
||   // do something at index myidx
```

These conversions between local and global size can also be done explicitly, using the `PetscSplitOwnership` (figure 141) routine. This routine takes two parameter, for the local and global size, and whichever one is initialized to `PETSC_DECIDE` gets computed from the other.

29.2 Scalars

Unlike programming languages that explicitly distinguish between single and double precision numbers, PETSc has only a single scalar type: `PetscScalar`. The precision of this is determined at installation time. In fact, a `PetscScalar` can even be a complex number if the installation specified that the scalar type is complex.

Even in applications that use complex numbers there can be quantities that are real: for instance, the norm of a complex vector is a real number. For that reason, PETSc also has the type `PetscReal`. There is also an explicit `PetscComplex`.

Integers in PETSc are likewise of a size determined at installation time: `PetscInt` can be 32 or 64 bits. Furthermore, there is a `PetscErrorCode` type for catching the return code of PETSc routines.

29.2.1 Complex

Numbers of type `PetscComplex` have a precision matching `PetscReal`.

Form a complex number as

```
|| PetscComplex x = 1.0 + 2.0 * PETSC_I;
```

29.2.2 MPI Scalars

For MPI calls, `MPIU_SCALAR` is the MPI type corresponding to the current `PetscScalar`.

For MPI calls, `MPIU_COMPLEX` is the MPI type corresponding to the current `PetscComplex`.

29.3 Vectors

Vectors are objects with a linear index. The elements of a vector are floating point numbers or complex numbers (see section 29.2), but not integers: for that see section 29.9.

29.3.1 Vector construction

Constructing a vector takes a number of steps. First of all, the vector object needs to be created on a communicator with `VecCreate` (figure 142)

Python note. In python, `PETSc.Vec()` creates an object with null handle, so a subsequent `create()` call is needed. In C and Fortran, the vector type is a keyword; in Python it is a member of `PETSc.Vec.Type`.

The corresponding routine `VecDestroy` (figure 143) deallocates data and zeros the pointer.

The vector type needs to be set with `VecSetType` (figure 144).

The most common vector types are:

- `VECSEQ` for sequential vectors, that is, living on a single process;
- `VECMPI` for a vector distributed over the communicator.

Once you have created one vector, you can make more like it by `VecDuplicate`.

29.3.2 Vector layout

Next in the creation process the vector size is set with `VecSetSizes` (figure 145). Since a vector is typically distributed, this involves the global size and the sizes on the processors. Setting both is redundant, so it is possible to specify one and let the other be computed by the library. This is indicated by setting it to `PETSC_DECIDE`.

The size is queried with `VecGetSize` (figure 146).

Each processor gets a contiguous part of the vector. Use `VecGetOwnershipRange` to query the first index on this process, and the first one of the next process.

In general it is best to let PETSc take care of memory management of matrix and vector objects, including allocating and freeing the memory. However, in cases where PETSc interfaces to other applications it maybe desirable to create a `Vec` object from an already allocated array: `VecCreateSeqWithArray` and `VecCreateMPIWithArray`.

29.3.3 Vector operations

There are many routines operating on vectors that you need to write scientific applications. Examples are: norms, vector addition (including Basic Linear Algebra Subprograms (BLAS)-type ‘AXPY’ routines), pointwise scaling, inner products. A large number of such operations are available in PETSc through single function calls to `VecXYZ` routines.

For debugging purposes, the `VecView` (figure 147) routine can be used to display vectors on screen as ascii output, but the routine call also use more general `PetscViewer` objects, for instance to dump a vector to file.

Exercise 29.1. Use the routines `VecDot` (figure 148), `VecScale` (figure 149) and `VecNorm` (figure 150) to compute the inner product of vectors x, y , scale the vector x , and check its norm:

$$\begin{aligned} p &\leftarrow x^t y \\ x &\leftarrow x/p \\ n &\leftarrow \|x\|_2 \end{aligned}$$

29.3.4 Vector elements

Setting elements of a traditional array is simple. Setting elements of a distributed array is harder. First of all, `VecSet` sets the vector to a constant value.

In the general case, setting elements in a PETSc vector is done through a function `VecSetValue` (figure 151) for setting elements that uses global numbering; any process can set any elements in the vector.

There is also a routine `VecSetValues` (figure 152) for setting multiple elements. This is mostly useful for setting dense subblocks of a block matrix.

```

|| i = 1; v = 3.14;
|| VecSetValue(x,i,v,INSERT_VALUES);
|| ii[0] = 1; ii[1] = 2; vv[0] = 2.7; vv[1] = 3.1;
|| VecSetValues(x,2,ii,vv,INSERT_VALUES);

|| call VecSetValue(x,i,v,INSERT_VALUES,ierr,e)
|| ii(1) = 1; ii(2) = 2; vv(1) = 2.7; vv(2) = 3.1
|| call VecSetValues(x,2,ii,vv,INSERT_VALUES,ierr,e)

```

Using `VecSetValue` for specifying a local vector element corresponds to simple insertion in the local array. However, an element that belongs to another process needs to be transferred. This is done in two calls: `VecAssemblyBegin` (figure 153) and `VecAssemblyEnd`.

(If you know the MPI library, you’ll recognize that the first call corresponds to posting non-blocking send and receive calls; the second then contains the wait calls. Thus, the existence of these separate calls make *latency hiding* possible.)

```

|| VecAssemblyBegin(myvec);
|| // do work that does not need the vector myvec
|| VecAssemblyEnd(myvec);

```

Elements can either be inserted (`INSERT_VALUES`) or added (`ADD_VALUES`). You can not immediately mix these modes; to do so you need to call `VecAssemblyBegin / VecAssemblyEnd`.

PetscSplitOwnership *How to read routine prototypes: 1.5.4.*

manpage 141: Routine prototype for PetscSplitOwnership

VecCreate *How to read routine prototypes: 1.5.4.*

manpage 142: Routine prototype for VecCreate

VecDestroy *How to read routine prototypes: 1.5.4.*

manpage 143: Routine prototype for VecDestroy

VecSetType *How to read routine prototypes: 1.5.4.*

manpage 144: Routine prototype for VecSetType

VecSetSizes *How to read routine prototypes: 1.5.4.*

manpage 145: Routine prototype for VecSetSizes

VecGetSize *How to read routine prototypes: 1.5.4.*

manpage 146: Routine prototype for VecGetSize

VecView *How to read routine prototypes: 1.5.4.*

manpage 147: Routine prototype for VecView

VecDot *How to read routine prototypes: 1.5.4.*

manpage 148: Routine prototype for VecDot

VecScale *How to read routine prototypes: 1.5.4.*

manpage 149: Routine prototype for VecScale

VecNorm *How to read routine prototypes: 1.5.4.*

manpage 150: Routine prototype for VecNorm

VecSetValue *How to read routine prototypes: 1.5.4.*

manpage 151: Routine prototype for VecSetValue

VecSetValues *How to read routine prototypes: 1.5.4.*

manpage 152: Routine prototype for VecSetValues

VecAssemblyBegin *How to read routine prototypes: 1.5.4.*

manpage 153: Routine prototype for VecAssemblyBegin

29.3.4.1 Explicit element access

Since the vector routines cover a large repertoire of operations, you hardly ever need to access the actual elements. Should you still need those elements, you can use `VecGetArray` (figure 154) for general access or `VecGetArrayRead` (figure 154) for read-only.

PETSc insists that you properly release this pointer again with `VecRestoreArray` (figure 155) or `VecRestoreArrayRead` (figure 155).

Note that in a distributed running context you can only get the array of local elements. Accessing the elements from another process requires explicit communication; see section 29.9.

```

||| PetscScalar *in_array,*out_array;
||| VecGetArrayRead(in,&in_array);
||| VecGetArray(out,&out_array);
||| VecGetLocalSize(in,&localsize);
||| for (int i=0; i<localsize; i++)
|||   out_array[i] = 2*in_array[i];
||| VecRestoreArrayRead(in,&in_array);
||| VecRestoreArray(out,&out_array);

```

29.4 Matrices

PETSc matrices come in a number of types, sparse and dense being the most important ones. Another possibility is to have the matrix in operation form, where only the action $y \leftarrow Ax$ is defined.

29.4.1 Matrix creation

Creating a matrix also starts by specifying a communicator on which the matrix lives collectively: `MatCreate` (figure 156)

Set the matrix type with `MatSetType` (figure 157). The main choices are between sequential versus distributed and dense versus sparse, giving types: `MATMPIIDENSE`, `MATMPIAIJ`, `MATSEQDENSE`, `MATSEQAIJ`.

Distributed matrices are partitioned by block rows: each process stores a *block row*, that is, a contiguous set of matrix rows. It stores all elements in that block row. In order for a matrix-vector product to be executable, both the input and output vector need to be partitioned conforming to the matrix.

While for dense matrices the block row scheme is not scalable, for matrices from PDEs it makes sense. There, a subdivision by matrix blocks would lead to many empty blocks.

Just as with vectors, there is a local and global size; except that that now applies to rows and columns. Set sizes with `MatSetSizes` (figure 158) and subsequently query them with `MatSizes` (figure 159). The concept of local column size is tricky: since a process stores a full block row you may expect the local column size to be the full matrix size, but that is not true. The exact definition will be discussed later, but for square matrices it is a safe strategy to let the local row and column size to be equal.

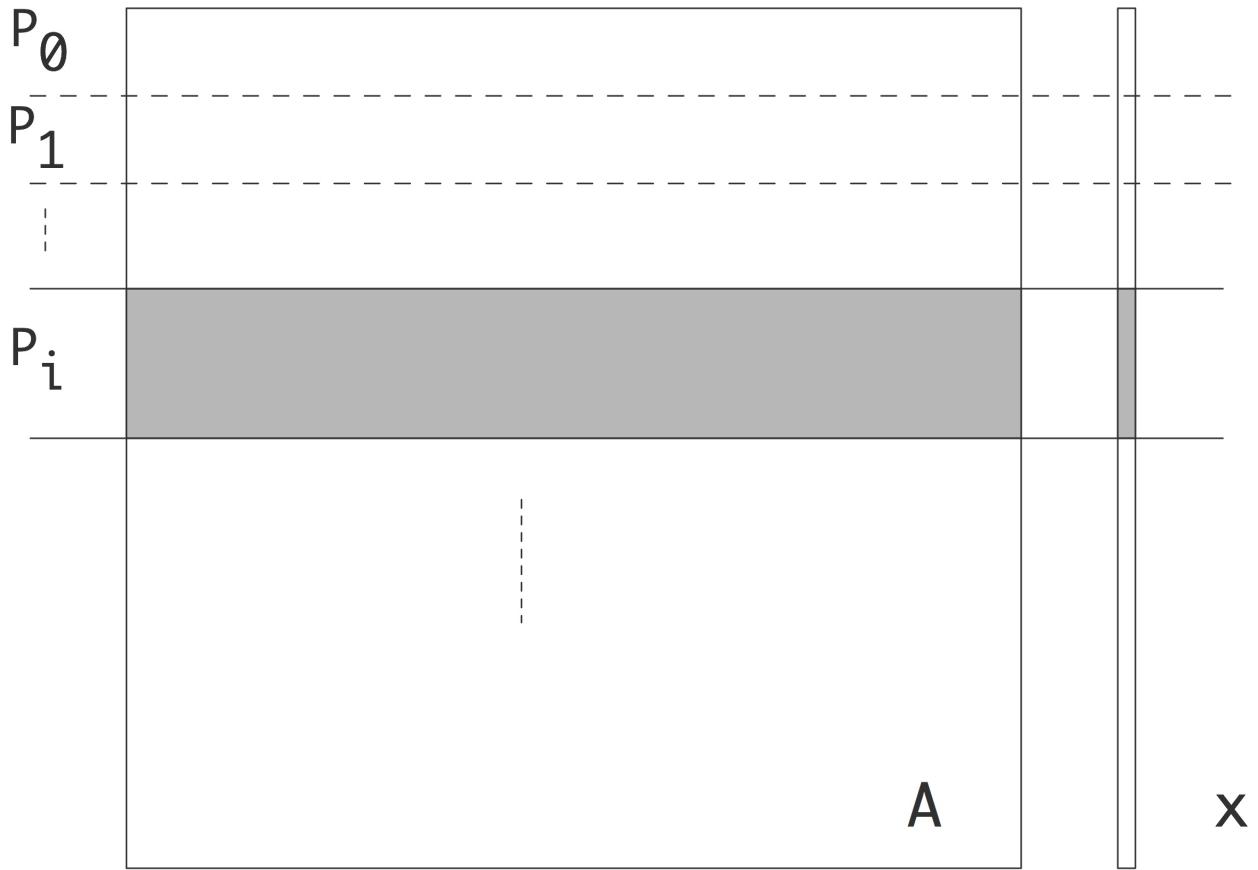


Figure 29.1: Matrix partitioning by block rows

29.4.2 Nonzero structure

In case of a dense matrix, once you have specified the size and the number of MPI ranks, it is simple to determine how much space PETSc needs to allocate for the matrix. For a sparse matrix this is more complicated, since the matrix can be anywhere between completely empty and completely filled in. It would be possible to have a dynamic approach where, as elements are specified, the space grows; however, repeated allocations and re-allocations are inefficient. For this reason PETSc puts a small burden on the programmer: you need to specify a bound on how many elements the matrix will contain.

We explain this by looking at some cases. First we consider a matrix that only lives on a single process. You would then use `MatSeqAIJSetPreallocation` (figure 160) . In the case of a tridiagonal matrix you would specify that each row has three elements:

```
// MatSeqAIJSetPreallocation(A, 3, NULL);
```

If the matrix is less regular you can use the third argument to give an array of explicit row lengths:

```
// int *rowlengths;  
// allocate, and then:
```

```

|| for (int row=0; row<nrows; row++)
    rowlengths[row] = // calculation of row length
|| MatSeqAIJSetPreallocation(A, NULL, rowlengths);

```

In case of a distributed matrix you need to specify this bound with respect to the block structure of the matrix. As illustrated in figure 29.2, a matrix has a diagonal part and an off-diagonal part. The diagonal part

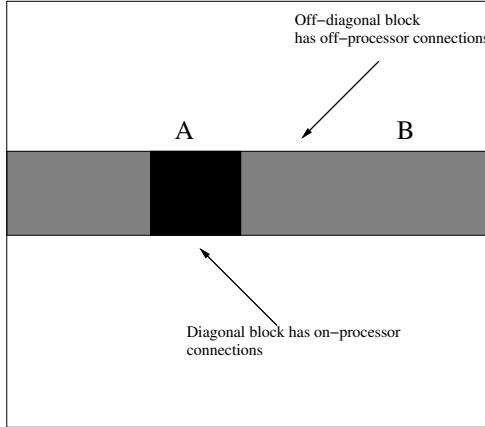


Figure 29.2: The diagonal and off-diagonal parts of a matrix

describes the matrix elements that couple elements of the input and output vector that live on this process. The off-diagonal part contains the matrix elements that are multiplied with elements not on this process, in order to compute elements that do live on this process.

The preallocation specification now has separate parameters for these diagonal and off-diagonal parts: with **MatMPIAIJSetPreallocation** (figure 160) you specify for both either a global upper bound on the number of nonzeros, or a detailed listing of row lengths. For the matrix of the *Laplace equation*, this specification would seem to be:

```

|| MatMPIAIJSetPreallocation(A, 3, NULL, 2, NULL);

```

However, this is only correct if the block structure from the parallel division equals that from the lines in the domain. In general it may be necessary to use values that are an overestimate. It is then possible to contract the storage by copying the matrix.

Specifying bounds on the number of nonzeros is often enough, and not too wasteful. However, if many rows have fewer nonzeros than these bounds, a lot of space is wasted. In that case you can replace the NULL arguments by an array that lists for each row the number of nonzeros in that row.

29.4.3 Matrix elements

You can set a single matrix element with **MatSetValue** (figure 161) or a block of them, where you supply a set of i and j indices, using **MatSetValues**.

After setting matrix elements, the matrix needs to be assembled. This is where PETSc moves matrix elements to the right processor, if they were specified elsewhere. As with vectors this takes two calls:

MatAssemblyBegin (figure 162) and **MatAssemblyEnd** (figure 162) which can be used to achieve *latency hiding*.

Elements can either be inserted (`INSERT_VALUES`) or added (`ADD_VALUES`). You can not immediately mix these modes; to do so you need to call **MatAssemblyBegin / MatAssemblyEnd** with a value of `MAT_FLUSH_ASSEMBLY`.

PETSc sparse matrices are very flexible: you can create them empty and then start adding elements. However, this is very inefficient in execution since the OS needs to reallocate the matrix every time it grows a little. Therefore, PETSc has calls for the user to indicate how many elements the matrix will ultimately contain.

```
|| MatSetOption(A, MAT_NEW_NONZERO_ALLOCATION_ERR, PETSC_FALSE)
```

29.4.3.1 Element access

If you absolutely need access to the matrix elements, there are routines such as **MatGetRow** (figure 163). With this, any rank can request, using global row numbering, the contents of a row that it owns. (Requesting elements that are not local requires the different mechanism of taking submatrices; section 29.6.)

Since PETSc is geared towards *sparse matrices*, this returns not only the element values, but also the column numbers, as well as the mere number of stored columns. If any of these three return values are not needed, they can be unrequested by setting the parameter passed to `NULL`.

PETSc insists that you properly release the row again with **MatRestoreRow** (figure 163).

It is also possible to retrieve the full **CRS!** (**CRS!**) contents of the local matrix with

```
|| MatGetArray  
|| MatRestoreArray
```

29.5 Matrix operations

29.5.1 Matrix-vector operations

In the typical application of PETSc, solving large sparse linear systems of equations with iterative methods, matrix-vector operations are most important. Foremost there is the matrix-vector product **MatMult** (figure 164) and the transpose product **MatMultTranspose** (figure 164). (In the complex case, the transpose product is not the Hermitian matrix product; for that use **MatMultHermitianTranspose**.)

For the BLAS gemv semantics $y \leftarrow \alpha Ax + \beta ey$, **MatMultAdd** computes $y \leftarrow z \leftarrow Ax + y$.

29.5.2 Matrix-matrix operations

There is a number of matrix-matrix routines such as **MatMatMult**.

VecGetArray	<i>How to read routine prototypes: 1.5.4.</i> manpage 154: Routine prototype for VecGetArray
VecRestoreArray	<i>How to read routine prototypes: 1.5.4.</i> manpage 155: Routine prototype for VecRestoreArray
MatCreate	<i>How to read routine prototypes: 1.5.4.</i> manpage 156: Routine prototype for MatCreate
MatSetType	<i>How to read routine prototypes: 1.5.4.</i> manpage 157: Routine prototype for MatSetType
MatSetSizes	<i>How to read routine prototypes: 1.5.4.</i> manpage 158: Routine prototype for MatSetSizes
MatSizes	<i>How to read routine prototypes: 1.5.4.</i> manpage 159: Routine prototype for MatSizes
MatSeqAIJSetPreallocation	<i>How to read routine prototypes: 1.5.4.</i> manpage 160: Routine prototype for MatSeqAIJSetPreallocation
MatSetValue	<i>How to read routine prototypes: 1.5.4.</i> manpage 161: Routine prototype for MatSetValue
MatAssemblyBegin	<i>How to read routine prototypes: 1.5.4.</i> manpage 162: Routine prototype for MatAssemblyBegin
MatGetRow	<i>How to read routine prototypes: 1.5.4.</i> manpage 163: Routine prototype for MatGetRow
MatMult	<i>How to read routine prototypes: 1.5.4.</i> manpage 164: Routine prototype for MatMult

29.6 Submatrices

Given a parallel matrix, there are two routines for extracting submatrices:

- `MatCreateSubMatrix` creates a single parallel submatrix.
- `MatCreateSubMatrices` creates a sequential submatrix on each rank.

29.7 Shell matrices

In many scientific applications, a matrix stands for some operator, and we are not intrinsically interested in the matrix elements, but only in the action of the matrix on a vector. In fact, under certain circumstances it is more convenient to implement a routine that computes the matrix action than to construct the matrix explicitly.

Maybe surprisingly, solving a linear system of equations can be handled this way. The reason is that PETSc's iterative solvers (section 30.1) only need the matrix-times-vector (and perhaps the matrix-transpose-times-vector) product.

PETSc supports this mode of working. The routine `MatCreateShell` (figure 165) declares the argument to be a matrix given in operator form. The next step is then to add the custom multiplication routine, which will be invoked by `MatMult: MatShellSetOperation` (figure 166)

The routine that implements the actual product should have the same prototype as `MatMult`, accepting a matrix and two vectors. The key to realizing your own product routine lies in the 'context' argument to the create routine. With `MatShellSetContext` (figure 167) you pass a pointer to some structure that contains all contextual information you need. In your multiplication routine you then retrieve this with `MatShellGetContext` (figure 168) .

Setting the context means passing a pointer (really: an address) to some allocated structure

```
|| struct matrix_data mystruct;
|| MatShellSetContext( A, &mystruct );
```

The routine prototype has this argument as a `void*` but it's not necessary to cast it to that. Getting the context means that a pointer to your structure needs to be set

```
|| struct matrix_data *mystruct;
|| MatShellGetContext( A, &mystruct );
```

Somewhat confusingly, the Get routine also has a `void*` argument, even though it's really a pointer variable.

29.8 DMDA: distributed arrays

`DMDACreate2d` (figure 169)

29.9 Index sets and Vector Scatters**29.10 Options and profiling**

Python note. In Python, do not specify the initial hyphen of an option name.

```
hasn = PETSc.Options().hasName("n")
```

MatCreateShell *How to read routine prototypes: [1.5.4](#).*

manpage 165: Routine prototype for MatCreateShell

MatShellSetOperation *How to read routine prototypes: [1.5.4](#).*

manpage 166: Routine prototype for MatShellSetOperation

MatShellSetContext *How to read routine prototypes: [1.5.4](#).*

manpage 167: Routine prototype for MatShellSetContext

MatShellGetContext *How to read routine prototypes: [1.5.4](#).*

manpage 168: Routine prototype for MatShellGetContext

DMDACreate2d *How to read routine prototypes: [1.5.4](#).*

manpage 169: Routine prototype for DMDACreate2d

Chapter 30

PETSc solvers

Probably the most important activity in PETSc is solving a linear system. This is done through a solver object: an object of the class `KSP`. (This stands for Krylov SPace solver.) The solution routine `KSPSolve` takes a matrix and a right-hand-side and gives a solution; however, before you can call this some amount of setup is needed.

There two very different ways of solving a linear system: through a direct method, essentially a variant of Gaussian elimination; or through an iterative method that makes successive approximations to the solution. In PETSc there are only iterative methods. We will show how to achieve direct methods later. The default linear system solver in PETSc is fully parallel, and will work on many linear systems, but there are many settings and customizations to tailor the solver to your specific problem.

30.1 KSP: linear system solvers

30.1.1 Math background

The main solution method for linear systems

$$?_x : Ax = b$$

in PETSc is through so-called iterative solution methods. Instead of directly computing the solution to the system, they compute a sequence of approximations that, with luck, converges to the true solution. Ideally, the sequence would stop when the distance to the true solution is small enough, but computing this is clearly not feasible. Instead, in each step the *residual*

$$r_i = Ax_i - b$$

is computed, and the iteration is stopped if this is small enough.

30.1.2 Solver objects

Create a KSP object:

After this, the basic scenario is:

```
|| KSP solver;
|| KSPSetOperators(solver,A,A);
|| KSPSolve(solver,rhs,sol);
```

Since neither solution nor solution speed is guaranteed, an iterative solver is subject to some tolerances:

- a relative tolerance for when the residual has been reduced enough;
- an absolute tolerance for when the residual is objectively small;
- a divergence tolerance that stops the iteration if the residual grows by too much; and
- a bound on the number of iterations, regardless any progress the process may still be making.

30.1.3 Why did my solver stop? Did it work?

On return of the `KSPSolve` routine there is no guarantee that the system was successfully solved. Therefore, you need to invoke `KSPGetConvergedReason` (figure 172) to get a `KSPConvergedReason` parameter that indicates what state the solver stopped in:

- The iteration can have successfully converged; this corresponds to `reason > 0`;
- the iteration can have diverged, or otherwise failed: `reason < 0`;
- or the iteration may have stopped at the maximum number of iterations while still making progress; `reason = 0`.

For more detail, `KSPReasonView` can print out the reason in readable form; for instance

```
|| KSPReasonView(solver,PETSC_VIEWER_STDOUT_WORLD);
```

(This can also be activated with the `-ksp_converged_reason` commandline option.)

In case of successful convergence, you can use `KSPGetIterationNumber` to report how many iterations were taken.

30.1.4 Choice of iterator

There are many iterative methods, and it takes a few function calls to fully specify them:

Here are some values:

- `KSPCG`: only for symmetric positive definite systems.
- `KSPGMRES`: a minimization method that works fairly generally; has high memory demands.
- `KSPBCGS`: a quasi-minimization method; uses less memory than GMRES.

30.1.5 Preconditioners

Another part of an iterative solver is the *preconditioner*; think of it as an approximation to the inverse.

```
|| PC prec;
|| KSPGetPC(solver,&prec);
|| PCSetType(prec,PCILU);
```

Some popular types:

- `PCILU`: an approximate factorization
- `PCSPAI`: an approximate inverse
- `PCASM`: additive Schwarz method

30.2 Direct solvers

PETSc has some support for direct solvers, that is, variants of LU decomposition. In a sequential context, the `PCLU` preconditioner can be used for this: a direct solver is equivalent to an iterative method that stops after one preconditioner application. This can be forced by specifying a KSP type of `KSPREONLY`.

Distributed direct solvers are more complicated. PETSc does not have this implemented in its basic code, but it becomes available by configuring PETSc with the `scalapack` library.

You need to specify which package provides the LU factorization:

```
|| PCFactorSetMatSolverType(pc, <solvertype> )
```

where *solvertype* can be mumps, superlu, umfpack, or a number of others. Note that availability of these packages depends on how PETSc was installed on your system.

30.3 Control through command line options

From the above you may get the impression that there are lots of calls to be made to set up a PETSc linear system and solver. And what if you want to experiment with different solvers, does that mean that you have to edit a whole bunch of code? Fortunately, there is an easier way to do things. If you call the routine `KSPSetFromOptions` (figure 174) with the *solver* as argument, PETSc will look at your command line options and take those into account in defining the solver. Thus, you can either omit setting options in your source code, or use this as a way of quickly experimenting with different possibilities. Example:

```
myprogram -ksp_type gmres -ksp_type_gmres_restart 20 -ksp_max_it 200 \
           -pc_type ilu -pc_type_ilu_levels 3
```

KSPCreate *How to read routine prototypes: 1.5.4.*

manpage 170: Routine prototype for KSPCreate

KSPSetTolerances *How to read routine prototypes: 1.5.4.*

manpage 171: Routine prototype for KSPSetTolerances

KSPGetConvergedReason *How to read routine prototypes: 1.5.4.*

manpage 172: Routine prototype for KSPGetConvergedReason

KSPSetType *How to read routine prototypes: 1.5.4.*

manpage 173: Routine prototype for KSPSetType

KSPSetFromOptions *How to read routine prototypes: 1.5.4.*

manpage 174: Routine prototype for KSPSetFromOptions

Chapter 31

PETSc solvers

31.1 Error checking

PETSc performs a good amount of runtime error checking. Some of this is for internal consistency, but it can also detect certain mathematical errors. To facilitate error reporting, the following scheme is used.

1. Every PETSc routine is a function returning a parameter of type `PetscErrorCode`.
2. Calling the macro `CHKERRQ` on the error code will cause an error to be printed and the current routine to be terminated. Recursively this gives a traceback of where the error occurred.
3. You can effect your own error return by using `SETERRQ`.

```
// PetscErrorCode ierr;  
// ierr = AnyPetscRoutine( arguments ); CHKERRQ(ierr);
```

Fortran note. In the main program, use `CHKERRA` and `SETERRA`. Also beware that these error ‘commands’ are macros, and after expansion may interfere with *Fortran line length*.

31.2 Printing

Printing screen output in parallel is tricky. If two processes execute a print statement at more or less the same time there is no guarantee as to in what order they may appear on screen. (Even attempts to have them print one after the other may not result in the right ordering.) Furthermore, lines from multi-line print actions on two processes may wind up on the screen interleaved.

PETSc has two routines that fix this problem. First of all, often the information printed is the same on all processes, so it is enough if only one process, for instance process 0, prints it. This is done with `PetscPrintf` (figure 175).

If all processes need to print, you can use `PetscSynchronizedPrintf` (figure 176) that forces the output to appear in process order.

To make sure that output is properly flushed from all system buffers use `PetscSynchronizedFlush` (figure 177) where for ordinary screen output you would use `stdout` for the file.

Python note. Since the print routines use the python `print` call, they automatically include the trailing newline. You don’t have to specify it as in the C calls.

31.3 Commandline options

PETSc has as large number of commandline options, most of which we will discuss later. For now we only mention `-log_summary` which will print out profile of the time taken in various routines. For these options to be parsed, it is necessary to pass `argc`, `argv` to the `PetscInitialize` call.

31.4 Memory management

`PetscNew` to allocate, and `PetscFree` to free. Allocated memory is aligned to `PETSC_MEMALIGN`. The state of memory allocation can be written to file or standard out with `PetscMallocDump`. The commandline option `-malloc_dump` outputs all not-freed memory during `PetscFinalize`.

`PetscMalloc` is deprecated.

PetscPrintf

How to read routine prototypes: 1.5.4.

manpage 175: Routine prototype for PetscPrintf

PetscSynchronizedPrintf

How to read routine prototypes: 1.5.4.

manpage 176: Routine prototype for PetscSynchronizedPrintf

PetscSynchronizedFlush

How to read routine prototypes: 1.5.4.

manpage 177: Routine prototype for PetscSynchronizedFlush

Chapter 32

PETSc topics

32.1 Communicators

PETSc has a ‘world’ communicator, which by default equals `MPI_COMM_WORLD`. If you want to run PETSc on a subset of processes, you can assign a subcommunicator to the variable `PETSC_COMM_WORLD` in between the calls to `MPI_Init` and `PetscInitialize`.

PetscComm

How to read routine prototypes: [1.5.4](#).

manpage 178: Routine prototype for PetscComm

PART IV

THE REST

Chapter 33

Exploring computer architecture

There is much that can be said about computer architecture. However, in the context of parallel programming we are mostly concerned with the following:

- How many networked nodes are there, and does the network have a structure that we need to pay attention to?
- On a compute node, how many sockets (or other Non-Uniform Memory Access (NUMA) domains) are there?
- For each socket, how many cores and hyperthreads are there? Are caches shared?

33.1 Tools for discovery

An easy way for discovering the structure of your parallel machine is to use tools that are written especially for this purpose.

33.1.1 Intel cpuinfo

The *Intel compiler suite* comes with a tool *cpuinfo* that reports on the structure of the node you are running on. It reports on the number of *packages*, that is: sockets, cores, and threads.

33.1.2 hwloc

The open source package *hwloc* does similar reporting to *cpuinfo*, but it has been ported to many platforms. Additionally, it can generate ascii and pdf graphic renderings of the architecture.

Chapter 34

Process and thread affinity

In the preceding chapters we mostly considered all MPI nodes or OpenMP threads as being in one flat pool. However, for high performance you need to worry about *affinity*: the question of which process or thread is placed where, and how efficiently they can interact.

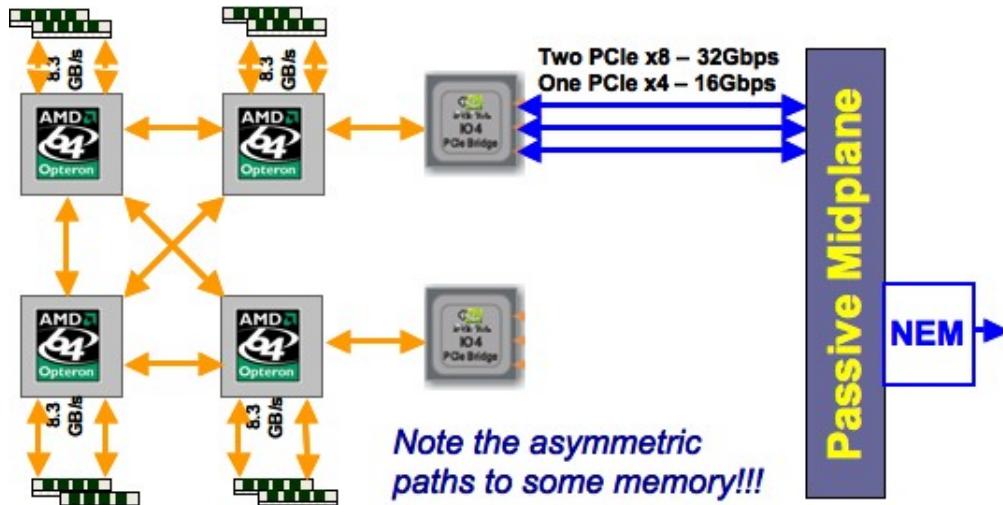


Figure 34.1: The NUMA structure of a Ranger node

Here are some situations where affinity becomes a concern.

- In pure MPI mode processes that are on the same node can typically communicate faster than processes on different nodes. Since processes are typically placed sequentially, this means that a scheme where process p interacts mostly with $p + 1$ will be efficient, while communication with large jumps will be less so.
- If the cluster network has a structure (*processor grid* as opposed to *fat-tree*), placement of processes has an effect on program efficiency. MPI tries to address this with *graph topology*; section 10.2.
- Even on a single node there can be asymmetries. Figure 34.1 illustrates the structure of the four sockets of the *Ranger* supercomputer (no longer in production). Two cores have no direct connection.

This asymmetry affects both MPI processes and threads on that node.

- Another problem with multi-socket designs is that each socket has memory attached to it. While every socket can address all the memory on the node, its local memory is faster to access. This asymmetry becomes quite visible in the *first-touch* phenomenon; section 23.2.
- If a node has fewer MPI processes than there are cores, you want to be in control of their placement. Also, the operating system can migrate processes, which is detrimental to performance since it negates data locality. For this reason, utilities such as `numactl` (and at TACC `tacc_affinity`) can be used to *pin a thread* or process to a specific core.
- Processors with *hyperthreading* or *hardware threads* introduce another level of worry about where threads go.

34.1 What does the hardware look like?

If you want to optimize affinity, you should first know what the hardware looks like. The `hwloc` utility is valuable here [6] (<https://www.open-mpi.org/projects/hwloc/>).

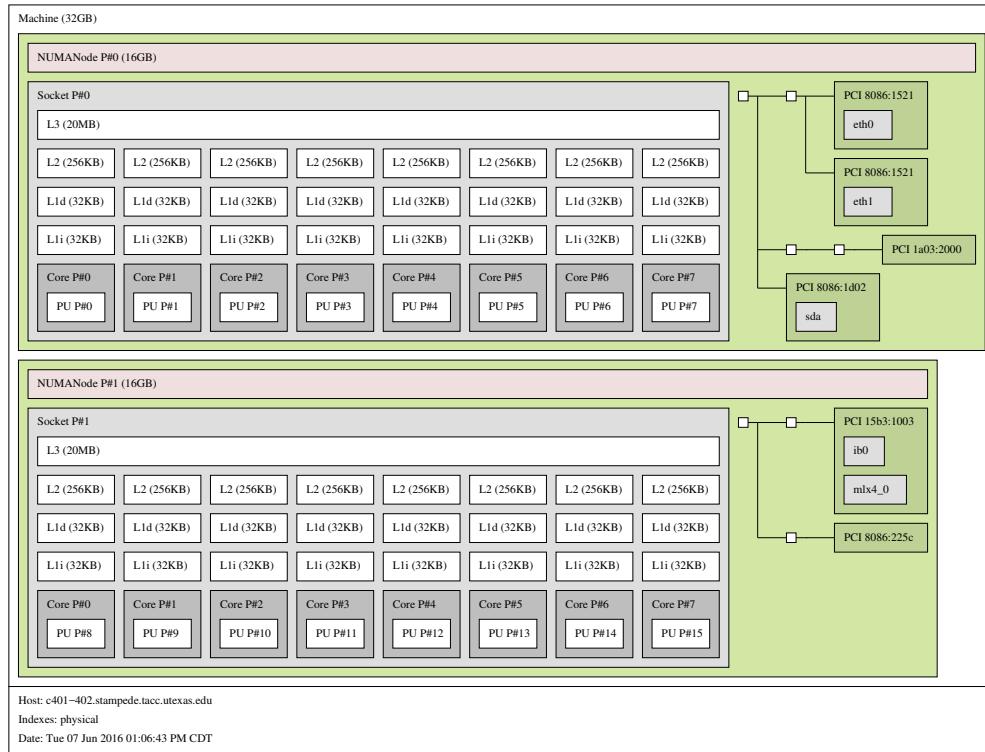


Figure 34.2: Structure of a Stampede compute node

Figure 34.2 depicts a *Stampede compute node*, which is a two-socket *Intel SandyBridge* design; figure 34.3 shows a *Stampede largemem node*, which is a four-socket design. Finally, figure 34.4 shows a *Lonestar5* compute node, a two-socket design with 12-core *Intel Haswell* processors with two hardware threads each.

34.1. What does the hardware look like?

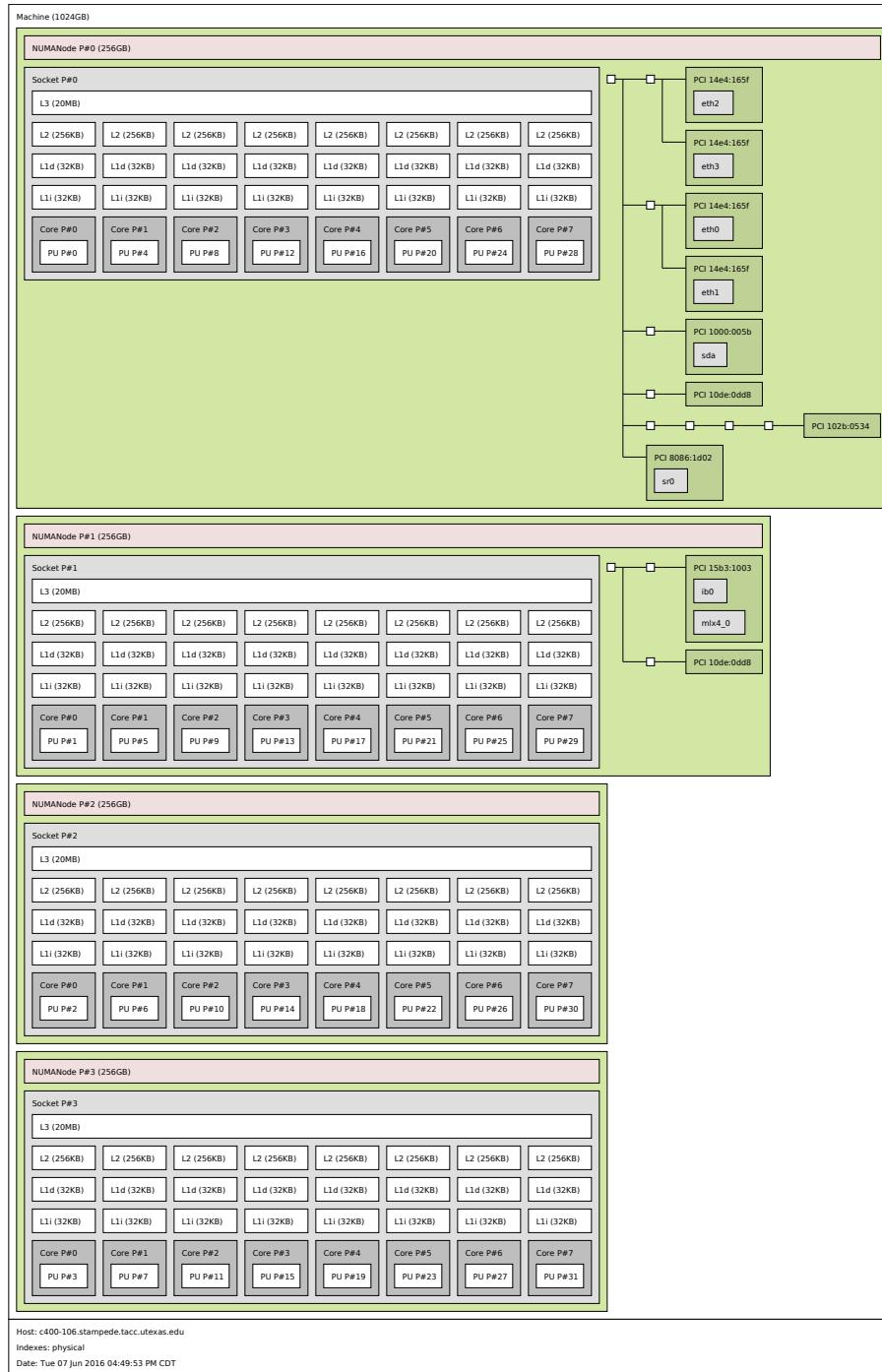


Figure 34.3: Structure of a Stampede largemem four-socket compute node

34. Process and thread affinity

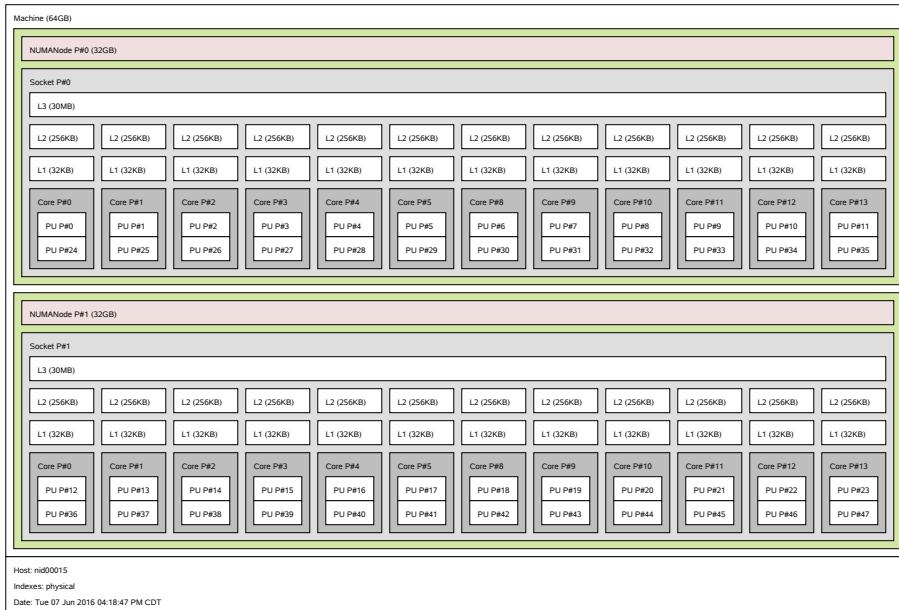


Figure 34.4: Structure of a Lonestar5 compute node

34.2 Affinity control

See chapter 23 for OpenMP affinity control.

Chapter 35

Hybrid computing

So far, you have learned to use MPI for distributed memory and OpenMP for shared memory parallel programming. However, distributed memory architectures actually have a shared memory component, since each cluster node is typically of a multicore design. Accordingly, you could program your cluster using MPI for inter-node and OpenMP for intra-node parallelism.

Say you use 100 cluster nodes, each with 16 cores. You could then start 1600 MPI processes, one for each core, but you could also start 100 processes, and give each access to 16 OpenMP threads.

In your slurm scripts, the first scenario would be specified `-N 100 -n 1600`, and the second as

```
#$SBATCH -N 100  
#$SBATCH -n 100  
  
export OMP_NUM_THREADS=16
```

There is a third choice, in between these extremes, that makes sense. A cluster node often has more than one socket, so you could put one MPI process on each socket, and use a number of threads equal to the number of cores per socket.

The script for this would be:

```
#$SBATCH -N 100  
#$SBATCH -n 200  
  
export OMP_NUM_THREADS=8  
ibrun tacc_affinity yourprogram
```

The `tacc_affinity` script unsets the following variables:

```
export MV2_USE_AFFINITY=0  
export MV2_ENABLE_AFFINITY=0  
export VIADEV_USE_AFFINITY=0  
export VIADEV_ENABLE_AFFINITY=0
```

35. Hybrid computing

If you don't use `tacc_affinity` you may want to do this by hand, otherwise `mvapich2` will use its own affinity rules.

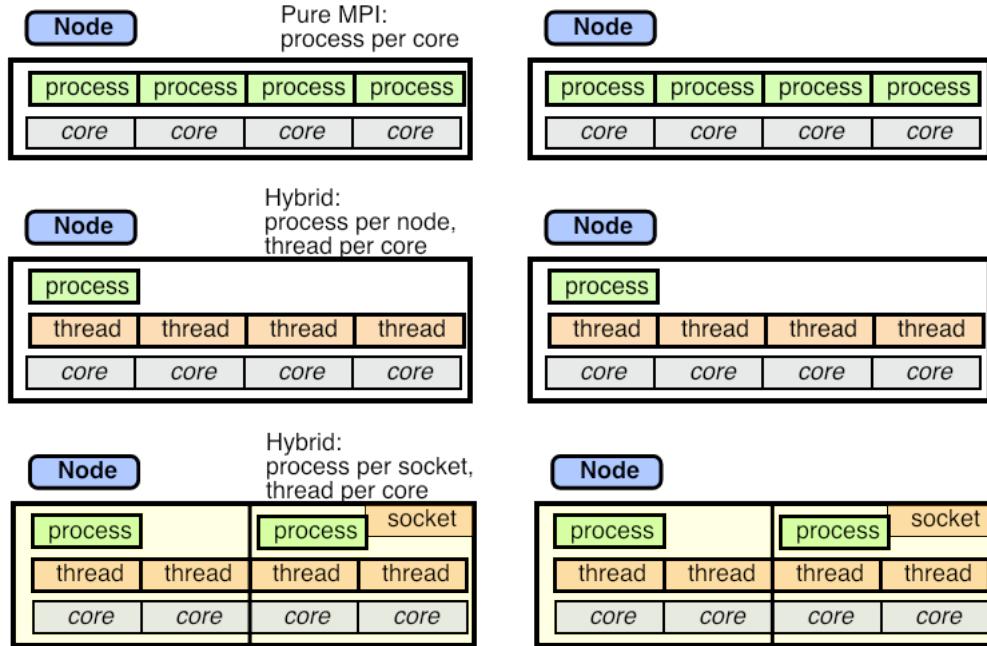


Figure 35.1: Three modes of MPI/OpenMP usage on a multi-core cluster

Figure 35.1 illustrates these three modes: pure MPI with no threads used; one MPI process per node and full multi-threading; two MPI processes per node, one per socket, and multiple threads on each socket.

35.1 Discussion

The performance implications of the pure MPI strategy versus hybrid are subtle.

- First of all, we note that there is no obvious speedup: in a well balanced MPI application all cores are busy all the time, so using threading can give no immediate improvement.
- Both MPI and OpenMP are subject to Amdahl's law that quantifies the influence of sequential code; in hybrid computing there is a new version of this law regarding the amount of code that is MPI-parallel, but not OpenMP-parallel.
- MPI processes run unsynchronized, so small variations in load or in processor behaviour can be tolerated. The frequent barriers in OpenMP constructs make a hybrid code more tightly synchronized, so load balancing becomes more critical.
- On the other hand, in OpenMP codes it is easier to divide the work into more tasks than there are threads, so statistically a certain amount of load balancing happens automatically.
- Each MPI process has its own buffers, so hybrid takes less buffer overhead.

Exercise 35.1. Review the scalability argument for 1D versus 2D matrix decomposition in HPSC-?. Would you get scalable performance from doing a 1D decomposition (for instance, of the rows) over MPI processes, and decomposing the other directions (the columns) over OpenMP threads?

Another performance argument we need to consider concerns message traffic. If let all threads make MPI calls (see section 35.2) there is going to be little difference. However, in one popular hybrid computing strategy we would keep MPI calls out of the OpenMP regions and have them in effect done by the master thread. In that case there are only MPI messages between nodes, instead of between cores. This leads to a decrease in message traffic, though this is hard to quantify. The number of messages goes down approximately by the number of cores per node, so this is an advantage if the average message size is small. On the other hand, the amount of data sent is only reduced if there is overlap in content between the messages.

Limiting MPI traffic to the master thread also means that no buffer space is needed for the on-node communication.

35.2 Hybrid MPI-plus-threads execution

In hybrid execution, the main question is whether all threads are allowed to make MPI calls. To determine this, replace the `MPI_Init` call by `MPI_Init_thread` (figure 179). Here the required and provided parameters can take the following values:

`MPI_THREAD_SINGLE` Only a single thread will execute.

`MPI_THREAD_FUNNELLED` The program may use multiple threads, but only the main thread will make MPI calls.

The main thread is usually the one selected by the `master` directive, but technically it is the only that executes `MPI_Init_thread`. If you call this routine in a parallel region, the main thread may be different from the master.

`MPI_THREAD_SERIAL` The program may use multiple threads, all of which may make MPI calls, but there will never be simultaneous MPI calls in more than one thread.

`MPI_THREAD_MULTIPLE` Multiple threads may issue MPI calls, without restrictions.

The *mvapich* implementation of MPI does have the required threading support, but you need to set this environment variable:

```
export MV2_ENABLE_AFFINITY=0
```

Another solution is to run your code like this:

```
ibrun tacc_affinity <my_multithreaded_mpi_executable>
```

The *mpirun* program usually propagates *environment variables*, so the value of `OMP_NUM_THREADS` when you call *mpirun* will be seen by each MPI process.

- It is possible to use blocking sends in threads, and let the threads block. This does away with the need for polling.
- You can not send to a thread number: use the MPI message tag to send to a specific thread.

Exercise 35.2. Consider the 2D heat equation and explore the mix of MPI/OpenMP parallelism:

- Give each node one MPI process that is fully multi-threaded.
- Give each core an MPI process and don't use multi-threading.

Discuss theoretically why the former can give higher performance. Implement both schemes as special cases of the general hybrid case, and run tests to find the optimal mix.

```
// thread.c
MPI_Init_thread(&argc,&argv,MPI_THREAD_MULTIPLE,&threading);
comm = MPI_COMM_WORLD;
MPI_Comm_rank(comm,&procno);
MPI_Comm_size(comm,&nprocs);

if (procno==0) {
    switch (threading) {
        case MPI_THREAD_MULTIPLE : printf("Glorious multithreaded MPI\n"); break;
        case MPI_THREAD_SERIALIZED : printf("No simultaneous MPI from threads\n");
            break;
        case MPI_THREAD_FUNNELED : printf("MPI from main thread\n"); break;
        case MPI_THREAD_SINGLE : printf("no threading supported\n"); break;
    }
}
MPI_Finalize();
```

35.3 Sources used in this chapter

Listing of code XX:

MPI_Init_thread

C:

```
int MPI_Init_thread(int *argc, char ***argv, int required, int *provided)
```

Fortran:

```
MPI_Init_thread(required, provided, ierror)
INTEGER, INTENT(IN) :: required
INTEGER, INTENT(OUT) :: provided
INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

How to read routine prototypes: 1.5.4.

manpage 179: Routine prototype for MPI_Init_thread

Chapter 36

Random number generation

Here is how you initialize the random number generator uniquely on each process:

C:

```
// Initialize the random number generator
srand((int)(mytid*(double)RAND_MAX/ntids));
// compute a random number
randomfraction = (rand() / (double)RAND_MAX);
```

Fortran:

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

call random_seed(size=randsize)
allocate(randseed(randsize))
randseed(:) = 1023*mytid
call random_seed(put=randseed)
```

Chapter 37

Parallel I/O

Parallel I/O is a tricky subject. You can try to let all processors jointly write one file, or to write a file per process and combine them later. With the standard mechanisms of your programming language there are the following considerations:

- On clusters where the processes have individual file systems, the only way to write a single file is to let it be generated by a single processor.
- Writing one file per process is easy to do, but
 - You need a post-processing script;
 - if the files are not on a shared file system (such as Lustre), it takes additional effort to bring them together;
 - if the files are on a shared file system, writing many files may be a burden on the metadata server.
- On a shared file system it is possible for all files to open the same file and set the file pointer individually. This can be difficult if the amount of data per process is not uniform.

Illustrating the last point:

```
// pseek.c
FILE *pfile;
pfile = fopen("pseek.dat", "w");
fseek(pfile, procid*sizeof(int), SEEK_CUR);
fseek(pfile, procid*sizeof(char), SEEK_CUR);
fprintf(pfile, "%d\n", procid);
fclose(pfile);
```

MPI also has its own portable I/O: *MPI I/O*, for which see chapter 9.

Alternatively, one could use a library such as *hdf5*.

37.1 Sources used in this chapter

Listing of code XX:

Chapter 38

Support libraries

There are many libraries related to parallel programming to make life easier, or at least more interesting, for you.

38.1 SimGrid

SimGrid [10] is a simulator for distributed systems. It can for instance be used to explore the effects of architectural parameters. It has been used to simulate large scale operations such as **HPL!** (**HPL!**) [2].

38.2 Other

ParaMesh

Global Arrays

Hdf5 and Silo

PART V

TUTORIALS

here are some tutorials

38.3 Debugging

When a program misbehaves, *debugging* is the process of finding out *why*. There are various strategies of finding errors in a program. The crudest one is debugging by print statements. If you have a notion of where in your code the error arises, you can edit your code to insert print statements, recompile, rerun, and see if the output gives you any suggestions. There are several problems with this:

- The edit/compile/run cycle is time consuming, especially since
- often the error will be caused by an earlier section of code, requiring you to edit, compile, and rerun repeatedly. Furthermore,
- the amount of data produced by your program can be too large to display and inspect effectively, and
- if your program is parallel, you probably need to print out data from all processors, making the inspection process very tedious.

For these reasons, the best way to debug is by the use of an interactive *debugger*, a program that allows you to monitor and control the behaviour of a running program. In this section you will familiarize yourself with *gdb*, which is the open source debugger of the *GNU* project. Other debuggers are proprietary, and typically come with a compiler suite. Another distinction is that *gdb* is a commandline debugger; there are graphical debuggers such as *ddd* (a frontend to *gdb*) or *DDT* and *TotalView* (debuggers for parallel codes). We limit ourselves to *gdb*, since it incorporates the basic concepts common to all debuggers.

In this tutorial you will debug a number of simple programs with *gdb* and *valgrind*. The files can be found in the repository in the directory `tutorials/debug_tutorial_files`.

38.3.1 Step 0: compiling for debug

You often need to recompile your code before you can debug it. A first reason for this is that the binary code typically knows nothing about what variable names corresponded to what memory locations, or what lines in the source to what instructions. In order to make the binary executable know this, you have to include the *symbol table* in it, which is done by adding the `-g` option to the compiler line.

Usually, you also need to lower the *compiler optimization level*: a production code will often be compiled with flags such as `-O2` or `-Xhost` that try to make the code as fast as possible, but for debugging you need to replace this by `-O0` ('oh-zero'). The reason is that higher levels will reorganize your code, making it hard to relate the execution to the source¹.

38.3.2 Invoking *gdb*

There are three ways of using *gdb*: using it to start a program, attaching it to an already running program, or using it to inspect a *core dump*. We will only consider the first possibility.

Here is an example of how to start *gdb* with a program that has no arguments (Fortran users, use `hello.F`):

```
tutorials/gdb/c/hello.c
```

1. Typically, actual code motion is done by `-O3`, but at level `-O2` the compiler will inline functions and make other simplifications.

```

%% cc -g -o hello hello.c
# regular invocation:
%% ./hello
hello world
# invocation from gdb:
%% gdb hello
GNU gdb 6.3.50-20050815 # ..... version info
Copyright 2004 Free Software Foundation, Inc. .... copyright info ....
(gdb) run
Starting program: /home/eijkhout/tutorials/gdb/hello
Reading symbols for shared libraries +. done
hello world

Program exited normally.
(gdb) quit
%%

```

Important note: the program was compiled with the *debug flag* `-g`. This causes the *symbol table* (that is, the translation from machine address to program variables) and other debug information to be included in the binary. This will make your binary larger than strictly necessary, but it will also make it slower, for instance because the compiler will not perform certain optimizations².

To illustrate the presence of the symbol table do

```

%% cc -g -o hello hello.c
%% gdb hello
GNU gdb 6.3.50-20050815 # ..... version info
(gdb) list

```

and compare it with leaving out the `-g` flag:

```

%% cc -o hello hello.c
%% gdb hello
GNU gdb 6.3.50-20050815 # ..... version info
(gdb) list

```

For a program with commandline input we give the arguments to the `run` command (Fortran users use `say.F`):

`tutorials/gdb/c/say.c`

```

%% cc -o say -g say.c
%% ./say 2

```

2. Compiler optimizations are not supposed to change the semantics of a program, but sometimes do. This can lead to the nightmare scenario where a program crashes or gives incorrect results, but magically works correctly with compiled with debug and run in a debugger.

```

hello world
hello world
%% gdb say
.... the usual messages ...
(gdb) run 2
Starting program: /home/eijkhout/tutorials/gdb/c/say 2
Reading symbols for shared libraries +. done
hello world
hello world

Program exited normally.

```

38.3.3 Finding errors

Let us now consider some programs with errors.

38.3.3.1 C programs

tutorials/gdb/c/square.c

```

%% cc -g -o square square.c
%% ./square
5000
Segmentation fault

```

The *segmentation fault* (other messages are possible too) indicates that we are accessing memory that we are not allowed to, making the program abort. A debugger will quickly tell us where this happens:

```

%% gdb square
(gdb) run
50000

Program received signal EXC_BAD_ACCESS, Could not access memory.
Reason: KERN_INVALID_ADDRESS at address: 0x0000000000eb4a
0x00007fff824295ca in __svfscanf_l ()

```

Apparently the error occurred in a function `__svfscanf_l`, which is not one of ours, but a system function. Using the `backtrace` (or `bt`, also `where` or `w`) command we quickly find out how this came to be called:

```

(gdb) backtrace
#0 0x00007fff824295ca in __svfscanf_l ()
#1 0x00007fff8244011b in fscanf ()
#2 0x0000000100000e89 in main (argc=1, argv=0x7fff5fbfc7c0) at square.c:7

```

We take a close look at line 7, and see that we need to change nmax to &nmax.

There is still an error in our program:

```
(gdb) run  
50000  
  
Program received signal EXC_BAD_ACCESS, Could not access memory.  
Reason: KERN_PROTECTION_FAILURE at address: 0x000000010000f000  
0x0000000100000ebe in main (argc=2, argv=0x7fff5fbfc7a8) at square1.c:9  
9           squares[i] = 1. / (i * i); sum += squares[i];
```

We investigate further:

```
(gdb) print i  
$1 = 11237  
(gdb) print squares[i]  
Cannot access memory at address 0x10000f000
```

and we quickly see that we forgot to allocate squares.

By the way, we were lucky here: this sort of memory errors is not always detected. Starting our programm with a smaller input does not lead to an error:

```
(gdb) run  
50  
Sum: 1.625133e+00  
  
Program exited normally.
```

38.3.3.2 Fortran programs

Compile and run the following program:

tutorials/gdb/f/square.F

It should abort with a message such as ‘Illegal instruction’. Running the program in gdb quickly tells you where the problem lies:

```
(gdb) run  
Starting program: tutorials/gdb//fsquare  
Reading symbols for shared libraries +++. done  
  
Program received signal EXC_BAD_INSTRUCTION, Illegal instruction/operand.  
0x0000000100000da3 in square () at square.F:7  
7           sum = sum + squares(i)
```

We take a close look at the code and see that we did not allocate squares properly.

38.3.4 Memory debugging with Valgrind

Insert the following allocation of `squares` in your program:

```
squares = (float *) malloc( nmax*sizeof(float) );
```

Compile and run your program. The output will likely be correct, although the program is not. Can you see the problem?

To find such subtle memory errors you need a different tool: a memory debugging tool. A popular (because open source) one is *valgrind*; a common commercial tool is *purify*.

tutorials/gdb/c/square1.c Compile this program with `cc -o square1 square1.c` and run it with `valgrind square1` (you need to type the input value). You will lots of output, starting with:

```
%% valgrind square1
==53695== Memcheck, a memory error detector
==53695== Copyright (C) 2002-2010, and GNU GPL'd, by Julian Seward et al.
==53695== Using Valgrind-3.6.1 and LibVEX; rerun with -h for copyright info
==53695== Command: a.out
==53695==
10
==53695== Invalid write of size 4
==53695==   at 0x100000EB0: main (square1.c:10)
==53695==   Address 0x10027e148 is 0 bytes after a block of size 40 alloc'd
==53695==     at 0x1000101EF: malloc (vg_replace_malloc.c:236)
==53695==   by 0x100000E77: main (square1.c:8)
==53695==
==53695== Invalid read of size 4
==53695==   at 0x100000EC1: main (square1.c:11)
==53695==   Address 0x10027e148 is 0 bytes after a block of size 40 alloc'd
==53695==     at 0x1000101EF: malloc (vg_replace_malloc.c:236)
==53695==   by 0x100000E77: main (square1.c:8)
```

Valgrind is informative but cryptic, since it works on the bare memory, not on variables. Thus, these error messages take some exegesis. They state that a line 10 writes a 4-byte object immediately after a block of 40 bytes that was allocated. In other words: the code is writing outside the bounds of an allocated array. Do you see what the problem in the code is?

Note that valgrind also reports at the end of the program run how much memory is still in use, meaning not properly freed.

If you fix the array bounds and recompile and rerun the program, valgrind still complains:

```
==53785== Conditional jump or move depends on uninitialized value(s)
==53785==   at 0x10006FC68: __ dtoa (in /usr/lib/libSystem.B.dylib)
==53785==   by 0x10003199F: __ vfprintf (in /usr/lib/libSystem.B.dylib)
==53785==   by 0x1000738AA: vfprintf_l (in /usr/lib/libSystem.B.dylib)
==53785==   by 0x1000A1006: printf (in /usr/lib/libSystem.B.dylib)
==53785==   by 0x1000000EF3: main (in ./square2)
```

Although no line number is given, the mention of `printf` gives an indication where the problem lies. The reference to an ‘uninitialized value’ is again cryptic: the only value being output is `sum`, and that is not uninitialized: it has been added to several times. Do you see why valgrind calls `is uninitialized` all the same?

38.3.5 Stepping through a program

Often the error in a program is sufficiently obscure that you need to investigate the program run in detail. Compile the following program

tutorials/gdb/c/roots.c and run it:

```
%% ./roots
sum: nan
```

Start it in `gdb` as follows:

```
%% gdb roots
GNU gdb 6.3.50-20050815 (Apple version gdb-1469) (Wed May 5 04:36:56 UTC 2005)
Copyright 2004 Free Software Foundation, Inc.

...
(gdb) break main
Breakpoint 1 at 0x100000ea6: file root.c, line 14.
(gdb) run
Starting program: tutorials/gdb/c/roots
Reading symbols for shared libraries +. done

Breakpoint 1, main () at roots.c:14
14          float x=0;
```

Here you have done the following:

- Before calling `run` you set a *breakpoint* at the main program, meaning that the execution will stop when it reaches the main program.
- You then call `run` and the program execution starts;
- The execution stops at the first instruction in `main`.

If execution is stopped at a breakpoint, you can do various things, such as issuing the `step` command:

```
Breakpoint 1, main () at roots.c:14
14          float x=0;
(gdb) step
15          for (i=100; i>-100; i--)
(gdb)
16          x += root(i);
(gdb)
```

(if you just hit return, the previously issued command is repeated). Do a number of steps in a row by hitting return. What do you notice about the function and the loop?

Switch from doing step to doing next. Now what do you notice about the loop and the function?

Set another breakpoint: break 17 and do cont. What happens?

Rerun the program after you set a breakpoint on the line with the sqrt call. When the execution stops there do where and list.

- If you set many breakpoints, you can find out what they are with info breakpoints.
- You can remove breakpoints with delete n where n is the number of the breakpoint.
- If you restart your program with run without leaving gdb, the breakpoints stay in effect.
- If you leave gdb, the breakpoints are cleared but you can save them: save breakpoints <file>. Use source <file> to read them in on the next gdb run.

38.3.6 Inspecting values

Run the previous program again in gdb: set a breakpoint at the line that does the sqrt call before you actually call run. When the program gets to line 8 you can do print n. Do cont. Where does the program stop?

If you want to repair a variable, you can do set var=value. Change the variable n and confirm that the square root of the new value is computed. Which commands do you do?

If a problem occurs in a loop, it can be tedious keep typing cont and inspecting the variable with print. Instead you can add a condition to an existing breakpoint: the following:

```
condition 1 if (n<0)
```

or set the condition when you define the breakpoint:

```
break 8 if (n<0)
```

Another possibility is to use ignore 1 50, which will not stop at breakpoint 1 the next 50 times.

Remove the existing breakpoint, redefine it with the condition n<0 and rerun your program. When the program breaks, find for what value of the loop variable it happened. What is the sequence of commands you use?

38.3.7 Parallel debugging

Debugging parallel programs is harder than sequential programs, because every sequential bug may show up, plus a number of new types, caused by the interaction of the various processes.

Here are a few possible parallel bugs:

- Processes can *deadlock* because they are waiting for a message that never comes. This typically happens with blocking send/receive calls due to an error in program logic.
- If an incoming message is unexpectedly larger than anticipated, a memory error can occur.

-
- A collective call will hang if somehow one of the processes does not call the routine.

There are few low-budget solutions to parallel debugging. The main one is to create an xterm for each process. We will describe this next. There are also commercial packages such as *DDT* and *TotalView*, that offer a GUI. They are very convenient but also expensive. The *Eclipse* project has a parallel package, *Eclipse PTP*, that includes a graphic debugger.

38.3.7.1 MPI debugging with gdb

You can not run parallel programs in gdb, but you can start multiple gdb processes that behave just like MPI processes! The command

```
mpirun -np <NP> xterm -e gdb ./program
```

create a number of xterm windows, each of which execute the commandline `gdb ./program`. And because these xterms have been started with `mpirun`, they actually form a communicator.

38.3.8 Further reading

A good tutorial: <http://www.dirac.org/linux/gdb/>.

Reference manual: http://www.ofb.net-gnu/gdb/gdb_toc.html.

38.4 Tracing

38.4.1 TAU profiling and tracing

TAU <http://www.cs.uoregon.edu/Research/tau/home.php> is a utility for profiling and tracing your parallel programs. Profiling is the gathering and displaying of bulk statistics, for instance showing you which routines take the most time, or whether communication takes a large portion of your runtime. When you get concerned about performance, a good profiling tool is indispensable.

Tracing is the construction and displaying of time-dependent information on your program run, for instance showing you if one process lags behind others. For understanding a program's behaviour, and the reasons behind profiling statistics, a tracing tool can be very insightful.

TAU works by adding *instrumentation* to your code: in effect it is a source-to-source translator that takes your code and turns it into one that generates run-time statistics. Doing this instrumentation is fortunately simple: start by having this code fragment in your makefile:

```
ifdef TACC_TAU_DIR  
    CC = tau_cc.sh  
else  
    CC = mpicc  
endif  
  
% : %.c  
<TAB>${CC} -o $@ $^
```

It is a good move to create a directory for your tracing and profiling information. You can use two separate directories, but there is no harm in using the same for both:

```
mkdir tau_trace  
mkdir tau_profile  
export PROFILEDIR=tau_profile  
export TRACEDIR=tau_trace  
  
export TAU_PROFILE=1  
export TAU_TRACE=1  
  
mpirun myprogram
```

After this, you view profiling information with `paraprof`

```
paraprof tau_profile
```

Viewing the traces takes a few steps:

```
cd tau_trace  
rm -f tau.trc tau.edf  
tau_treemerge.pl
```

```
tau2slog2 tau.trc tau.edf -o yourprogram.slog2
```

The slog2 file can then be viewed with jumpshot:

```
jumpshot yourprogram.slog2
```

38.5 SimGrid

Many readers of this book will have access to some sort of parallel machine so that they can run simulations, maybe even some realistic scaling studies. However, not many people will have access to more than one cluster type so that they can evaluate the influence of the *interconnect*. Even then, for didactic purposes one would often wish for interconnect types (fully connected, linear processor array) that are unlikely to be available.

In order to explore architectural issues pertaining to the network, we then resort to a simulation tool, *SimGrid*.

Installation

Compilation You write plain MPI files, but compile them with the *SimGrid compiler* `smpicc`.

Running SimGrid has its own version of `mpirun`: `smpirun`. You need to supply this with options:

- `-np 123456` for the number of (virtual) processors;
- `-hostfile simgridhostfile` which lists the names of these processors. You can basically make these up, but are defined in:
- `-platform arch.xml` which defines the connectivity between the processors.

For instance, with a hostfile of 8 hosts, a linearly connected network would be defined as:

```
<?xml version='1.0'?>
<!DOCTYPE platform SYSTEM "http://simgrid.gforge.inria.fr/simgrid/simgrid.d
.

<platform version="4">

<zone id="first zone" routing="Floyd">
  <!-- the resources -->
  <host id="host1" speed="1Mf"/>
  <host id="host2" speed="1Mf"/>
  <host id="host3" speed="1Mf"/>
  <host id="host4" speed="1Mf"/>
  <host id="host5" speed="1Mf"/>
  <host id="host6" speed="1Mf"/>
  <host id="host7" speed="1Mf"/>
  <host id="host8" speed="1Mf"/>
  <link id="link1" bandwidth="125MBps" latency="100us"/>
  <!-- the routing: specify how the hosts are interconnected -->
  <route src="host1" dst="host2"><link_ctn id="link1"/></route>
  <route src="host2" dst="host3"><link_ctn id="link1"/></route>
  <route src="host3" dst="host4"><link_ctn id="link1"/></route>
  <route src="host4" dst="host5"><link_ctn id="link1"/></route>
```

```
<route src="host5" dst="host6"><link_ctn id="link1"/></route>
<route src="host6" dst="host7"><link_ctn id="link1"/></route>
<route src="host7" dst="host8"><link_ctn id="link1"/></route>
</zone>

</platform>
```

(such files are easily generated with a shell script).

The Floyd designation of the routing means that any route using the transitive closure of the paths given can be used. It is also possible to use `routing="Full"` which requires full specification of all pairs that can communicate.

PART VI

PROJECTS, INDEX

Chapter 39

Class projects

39.1 A Style Guide to Project Submissions

Here are some guidelines for how to submit assignments and projects. As a general rule, consider programming as an experimental science, and your writeup as a report on some tests you have done: explain the problem you're addressing, your strategy, your results.

Structure of your writeup Most of the projects in this book use a scientific question to allow you to prove your coding skills. That does not mean that turning in the code is sufficient, nor code plus sample output. Turn in a writeup in pdf form that was generated from a text processing program such (preferably) \LaTeX (for a tutorial, see HPSC-??).

Your writeup should have:

- Foremost, a short description of the purpose of your project and your results;
- An explanation of your algorithms or solution strategy;
- Relevant fragments of your code;
- A scientific discussion of what you observed,
- Any code-related observations.
- If applicable: graphs, both of application quantities and performance issues. (For parallel runs possibly TAU plots; see 38.4).

Observe, measure, hypothesize, deduce Your project may be a scientific investigation of some phenomenon. Formulate hypotheses as to what you expect to observe, report on your observations, and draw conclusions.

Quite often your program will display unexpected behaviour. It is important to observe this, and hypothesize what the reason might be for your observed behaviour.

In most applications of computing machinery we care about the efficiency with which we find the solution. Thus, make sure that you do measurements. In general, make observations that allow you to judge whether your program behaves the way you would expect it to.

Including code If you include code samples in your writeup, make sure they look good. For starters, use a mono-spaced font. In L^AT_EX, you can use the `verbatim` environment or the `verbatiminput` command. In that section option the source is included automatically, rather than cut and pasted. This is to be preferred, since your writeup will stay current after you edit the source file.

Including whole source files makes for a long and boring writeup. The code samples in this book were generated as follows. In the source files, the relevant snippet was marked as

```
... boring stuff
//snippet samplex
    .. interesting! ..
//snippet end
... more boring stuff
```

The files were then processed with the following command line (actually, included in a makefile, which requires doubling the dollar signs):

```
for f in *.{c,cxx,h} ; do
    cat $x | awk 'BEGIN {f=0}
                    /snippet end/ {f=0}
                    f==1 {print $0 > file}
                    /snippet/ && !/end/ {f=1; file=$2}
    '
done
```

which gives (in this example) a file `samplex`. Other solutions are of course possible.

Code formatting Included code snippets should be readable. At a minimum you could indent the code correctly in an editor before you include it in a `verbatim` environment. (Screenshots of your terminal window are a decidedly suboptimal solution.) But it's better to use the `listing` package which formats your code, include syntax coloring. For instance,

```
\lstset{language=C++} % or Fortran or so
\begin{lstlisting}
for (int i=0; i<N; i++)
    s += 1;
\end{lstlisting}

|| for (int i=0; i<N; i++)
    s += 1;
```

Running your code A single run doesn't prove anything. For a good report, you need to run your code for more than one input dataset (if available) and in more than one processor configuration. When you choose problem sizes, be aware that an average processor can do a billion operations per second: you need

to make your problem large enough for the timings to rise above the level of random variations and startup phenomena.

When you run a code in parallel, beware that on clusters the behaviour of a parallel code will always be different between one node and multiple nodes. On a single node the MPI implementation is likely optimized to use the shared memory. This means that results obtained from a single node run will be unrepresentative. In fact, in timing and scaling tests you will often see a drop in (relative) performance going from one node to two. Therefore you need to run your code in a variety of scenarios, using more than one node.

Reporting scaling If you do a scaling analysis, a graph reporting runtimes should not have a linear time axis: a logarithmic graph is much easier to read. A speedup graph can also be informative.

Some algorithms are mathematically equivalent in their sequential and parallel versions. Others, such as iterative processes, can take more operations in parallel than sequentially, for instance because the number of iterations goes up. In this case, report both the speedup of a single iteration, and the total improvement of running the full algorithm in parallel.

Repository organization If you submit your work through a repository, make sure you organize your submissions in subdirectories, and that you give a clear name to all files. Object files and binaries should not be in a repository since they are dependent on hardware and things like compilers.

39.2 Warmup Exercises

We start with some simple exercises.

39.2.1 Hello world

For background, see section 2.3.

First of all we need to make sure that you have a working setup for parallel jobs. The example program `helloworld.c` does the following:

```
// helloworld.c
MPI_Init(&argc,&argv);
MPI_Comm_size(MPI_COMM_WORLD,&ntids);
MPI_Comm_rank(MPI_COMM_WORLD,&mytid);
printf("Hello, this is processor %d out of %d\n",mytid,ntids);
MPI_Finalize();
```

Compile this program and run it in parallel. Make sure that the processors do *not* all say that they are processor 0 out of 1!

39.2.2 Collectives

It is a good idea to be able to collect statistics, so before we do anything interesting, we will look at MPI collectives; section 3.1.

Take a look at `time_max.cxx`. This program sleeps for a random number of seconds:

```
// time_max.cxx
wait = (int) ( 6.*rand() / (double) RAND_MAX );
tstart = MPI_Wtime();
sleep(wait);
tstop = MPI_Wtime();
jitter = tstop-tstart-wait;
```

and measures how long the sleep actually was:

```
if (mytid==0)
sendbuf = MPI_IN_PLACE;
else sendbuf = (void*)&jitter;
MPI_Reduce(sendbuf, (void*)&jitter, 1, MPI_DOUBLE, MPI_MAX, 0, comm);
```

In the code, this quantity is called ‘jitter’, which is a term for random deviations in a system.

Exercise 39.1. Change this program to compute the average jitter by changing the reduction operator.

Exercise 39.2. Now compute the standard deviation

$$\sigma = \sqrt{\frac{\sum_i (x_i - m)^2}{n}}$$

where m is the average value you computed in the previous exercise.

- Solve this exercise twice: once by following the reduce by a broadcast operation and once by using an Allreduce.
- Run your code both on a single cluster node and on multiple nodes, and inspect the TAU trace. Some MPI implementations are optimized for shared memory, so the trace on a single node may not look as expected.
- Can you see from the trace how the allreduce is implemented?

Exercise 39.3. Finally, use a gather call to collect all the values on processor zero, and print them out. Is there any process that behaves very differently from the others?

39.2.3 Linear arrays of processors

In this section you are going to write a number of variations on a very simple operation: all processors pass a data item to the processor with the next higher number.

- In the file `linear-serial.c` you will find an implementation using blocking send and receive calls.
- You will change this code to use non-blocking sends and receives; they require an `MPI_Wait` call to finalize them.
- Next, you will use `MPI_Sendrecv` to arrive at a synchronous, but deadlock-free implementation.
- Finally, you will use two different one-sided scenarios.

In the reference code `linear-serial.c`, each process defines two buffers:

```
// linear-serial.c
int my_number = mytid, other_number=-1.;
```

where `other_number` is the location where the data from the left neighbour is going to be stored.

To check the correctness of the program, there is a gather operation on processor zero:

```
int *gather_buffer=NULL;
if (mytid==0) {
    gather_buffer = (int*) malloc(ntids*sizeof(int));
    if (!gather_buffer) MPI_Abort(comm,1);
}
MPI_Gather(&other_number,1,MPI_INT,
           gather_buffer,1,MPI_INT, 0,comm);
if (mytid==0) {
    int i,error=0;
    for (i=0; i<ntids; i++)
```

```
    if (gather_buffer[i]!=i-1) {
        printf("Processor %d was incorrect: %d should be %d\n",
               i,gather_buffer[i],i-1);
        error =1;
    }
    if (!error) printf("Success!\n");
    free(gather_buffer);
}
```

39.2.3.1 Coding with blocking calls

Passing data to a neighbouring processor should be a very parallel operation. However, if we code this naively, with MPI_Send and MPI_Recv, we get an unexpected serial behaviour, as was explained in section 4.2.2.

```
if (mytid<ntids-1)
MPI_Ssend( /* data: */ &my_number,1,MPI_INT,
           /* to: */ mytid+1, /* tag: */ 0, comm);
if (mytid>0)
MPI_Recv( /* data: */ &other_number,1,MPI_INT,
           /* from: */ mytid-1, 0, comm, &status);
```

(Note that this uses an Ssend; see section 12.8 for the explanation why.)

Exercise 39.4. Compile and run this code, and generate a TAU trace file. Confirm that the execution is serial. Does replacing the Ssend by Send change this?

Let's clean up the code a little.

Exercise 39.5. First write this code more elegantly by using MPI_PROC_NULL.

39.2.3.2 A better blocking solution

The easiest way to prevent the serialization problem of the previous exercises is to use the MPI_Sendrecv call. This routine acknowledges that often a processor will have a receive call whenever there is a send. For border cases where a send or receive is unmatched you can use MPI_PROC_NULL.

Exercise 39.6. Rewrite the code using MPI_Sendrecv. Confirm with a TAU trace that execution is no longer serial.

Note that the Sendrecv call itself is still blocking, but at least the ordering of its constituent send and recv are no longer ordered in time.

39.2.3.3 Non-blocking calls

The other way around the blocking behaviour is to use Isend and Irecv calls, which do not block. Of course, now you need a guarantee that these send and receive actions are concluded; in this case, use MPI_Waitall.

Exercise 39.7. Implement a fully parallel version by using `MPI_Isend` and `MPI_Irecv`.

39.2.3.4 One-sided communication

Another way to have non-blocking behaviour is to use one-sided communication. During a Put or Get operation, execution will only block while the data is being transferred out of or into the origin process, but it is not blocked by the target. Again, you need a guarantee that the transfer is concluded; here use `MPI_Win_fence`.

Exercise 39.8. Write two versions of the code: one using `MPI_Put` and one with `MPI_Get`.

Make TAU traces.

Investigate blocking behaviour through TAU visualizations.

Exercise 39.9. If you transfer a large amount of data, and the target processor is occupied, can you see any effect on the origin? Are the fences synchronized?

39.3 Mandelbrot set

If you've never heard the name *Mandelbrot set*, you probably recognize the picture; figure 39.1 Its formal

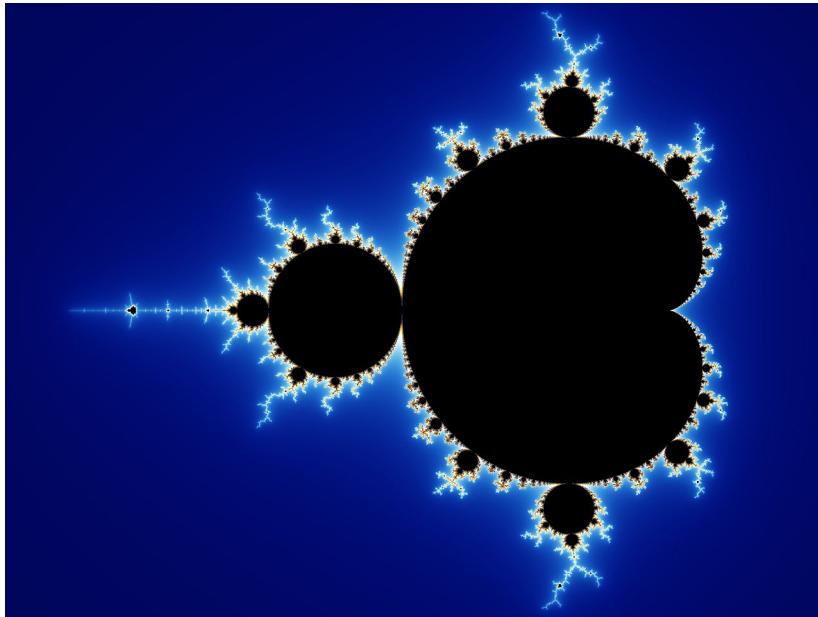


Figure 39.1: The Mandelbrot set

definition is as follows:

A point c in the complex plane is part of the Mandelbrot set if the series x_n defined by

$$\begin{cases} x_0 = 0 \\ x_{n+1} = x_n^2 + c \end{cases}$$

satisfies

$$\forall n : |x_n| \leq 2.$$

It is easy to see that only points c in the bounding circle $|c| < 2$ qualify, but apart from that it's hard to say much without a lot more thinking. Or computing; and that's what we're going to do.

In this set of exercises you are going to take an example program `mandel_main.cxx` and extend it to use a variety of MPI programming constructs. This program has been set up as a *manager-worker* model: there is one manager processor (for a change this is the last processor, rather than zero) which gives out work to, and accepts results from, the worker processors. It then takes the results and constructs an image file from them.

39.3.1 Invocation

The `mandel_main` program is called as

```
mpirun -np 123 mandel_main steps 456 iters 789
```

where the `steps` parameter indicates how many steps in x, y direction there are in the image, and `iters` gives the maximum number of iterations in the `belong` test.

If you forget the parameter, you can call the program with

```
mandel_serial -h
```

and it will print out the usage information.

39.3.2 Tools

The driver part of the Mandelbrot program is simple. There is a `circle` object that can generate coordinates

```
// mandel.h
class circle {
public :
    circle(int pxls,int bound,int bs);
    void next_coordinate(struct coordinate& xy);
    int is_valid_coordinate(struct coordinate xy);
    void invalid_coordinate(struct coordinate& xy);
```

and a global routine that tests whether a coordinate is in the set, at least up to an iteration bound. It returns zero if the series from the given starting point has not diverged, or the iteration number in which it diverged if it did so.

```
int belongs(struct coordinate xy,int itbound) {
    double x=xy.x, y=xy.y; int it;
    for (it=0; it<itbound; it++) {
        double xx,yy;
        xx = x*x - y*y + xy.x;
        yy = 2*x*y + xy.y;
        x = xx; y = yy;
        if (x*x+y*y>4.) {
            return it;
        }
    }
    return 0;
}
```

In the former case, the point could be in the Mandelbrot set, and we colour it black, in the latter case we give it a colour depending on the iteration number.

```
if (iteration==0)
    memset(colour,0,3*sizeof(float));
```

```

else {
    float rfloat = ((float) iteration) / workcircle->infty;
    colour[0] = rfloat;
    colour[1] = MAX((float)0, (float)(1-2*rfloat));
    colour[2] = MAX((float)0, (float)(2*(rfloat-.5)));
}

```

We use a fairly simple code for the worker processes: they execute a loop in which they wait for input, process it, return the result.

```

void queue::wait_for_work(MPI_Comm comm, circle *workcircle) {
    MPI_Status status; int ntids;
    MPI_Comm_size(comm, &ntids);
    int stop = 0;

    while (!stop) {
        struct coordinate xy;
        int res;

        MPI_Recv(&xy, 1, coordinate_type, ntids-1, 0, comm, &status);
        stop = !workcircle->is_valid_coordinate(xy);
        if (stop) break; //res = 0;
        else {
            res = belongs(xy, workcircle->infty);
        }
        MPI_Send(&res, 1, MPI_INT, ntids-1, 0, comm);
    }
    return;
}

```

A very simple solution using blocking sends on the manager is given:

```

// mandel_serial.cxx
class serialqueue : public queue {
private :
    int free_processor;
public :
    serialqueue(MPI_Comm queue_comm, circle *workcircle)
        : queue(queue_comm, workcircle) {
        free_processor=0;
    };
    /**
     * The 'addtask' routine adds a task to the queue. In this
     * simple case it immediately sends the task to a worker
     * and waits for the result, which is added to the image.
     */
}

```

```
This routine is only called with valid coordinates;
the calling environment will stop the process once
an invalid coordinate is encountered.

*/
int addtask(struct coordinate xy) {
    MPI_Status status; int contribution, err;

    err = MPI_Send(&xy, 1, coordinate_type,
        free_processor, 0, comm); CHK(err);
    err = MPI_Recv(&contribution, 1, MPI_INT,
        free_processor, 0, comm, &status); CHK(err);

    coordinate_to_image(xy, contribution);
    total_tasks++;
    free_processor = (free_processor+1)% (ntids-1);

    return 0;
};
```

Exercise 39.10. Explain why this solution is very inefficient. Make a trace of its execution that bears this out.

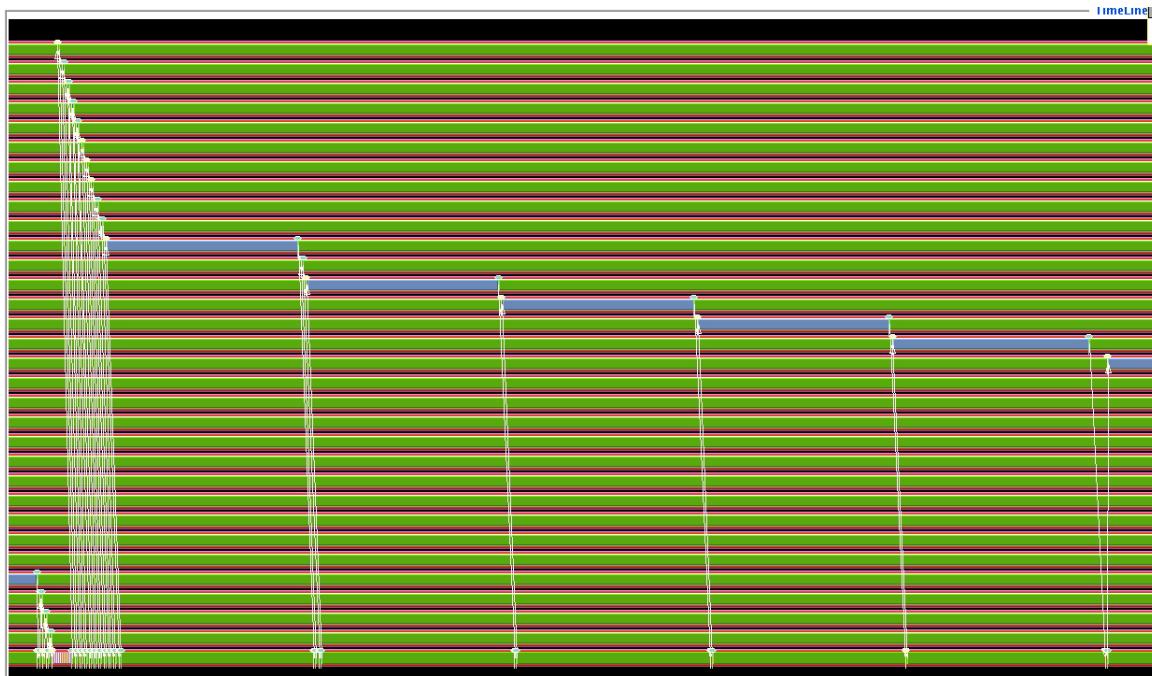


Figure 39.2: Trace of a serial Mandelbrot calculation

39.3.3 Bulk task scheduling

The previous section showed a very inefficient solution, but that was mostly intended to set up the code base. If all tasks take about the same amount of time, you can give each process a task, and then wait on them all to finish. A first way to do this is with non-blocking sends.

Exercise 39.11. Code a solution where you give a task to all worker processes using non-blocking sends and receives, and then wait for these tasks with `MPI_Waitall` to finish before you give a new round of data to all workers. Make a trace of the execution of this and report on the total time.

You can do this by writing a new class that inherits from `queue`, and that provides its own `addtask` method:

```
// mandel_bulk.cxx
class bulkqueue : public queue {
public :
    bulkqueue(MPI_Comm queue_comm, circle *workcircle)
        : queue(queue_comm, workcircle) {
```

You will also have to override the `complete` method: when the `circle` object indicates that all coordinates have been generated, not all workers will be busy, so you need to supply the proper `MPI_Waitall` call.

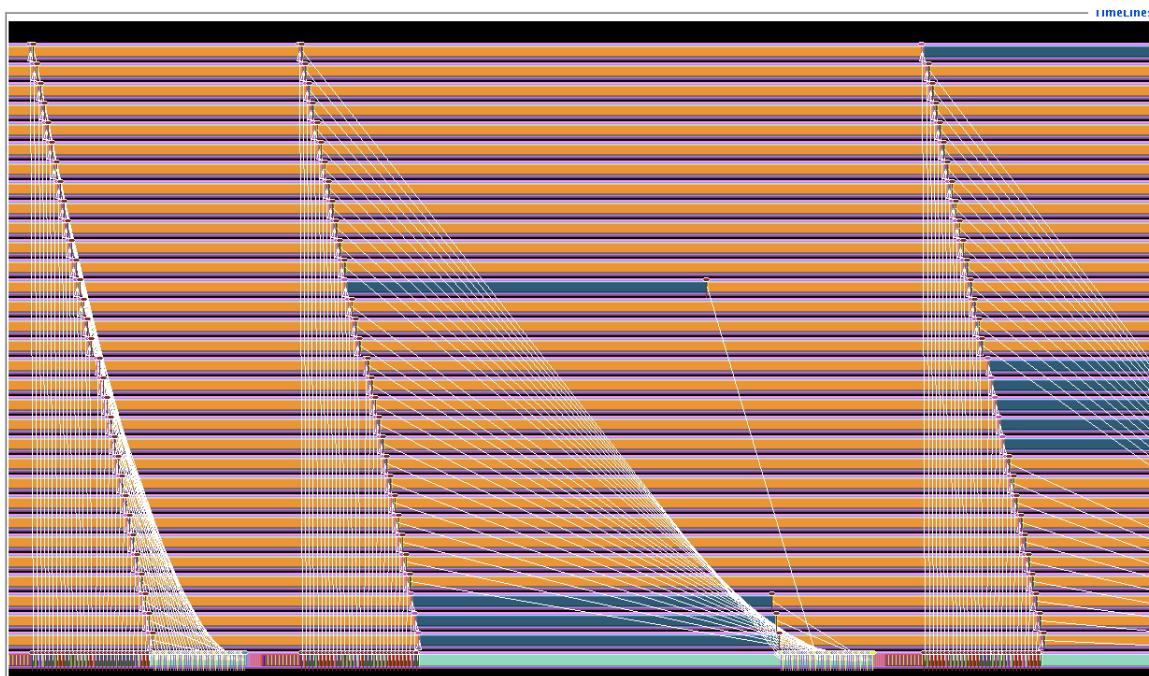


Figure 39.3: Trace of a bulk Mandelbrot calculation

39.3.4 Collective task scheduling

Another implementation of the bulk scheduling of the previous section would be through using collectives.

Exercise 39.12. Code a solution which uses scatter to distribute data to the worker tasks, and gather to collect the results. Is this solution more or less efficient than the previous?

39.3.5 Asynchronous task scheduling

At the start of section 39.3.3 we said that bulk scheduling mostly makes sense if all tasks take similar time to complete. In the Mandelbrot case this is clearly not the case.

Exercise 39.13. Code a fully dynamic solution that uses MPI_Probe or MPI_Waitany.

Make an execution trace and report on the total running time.

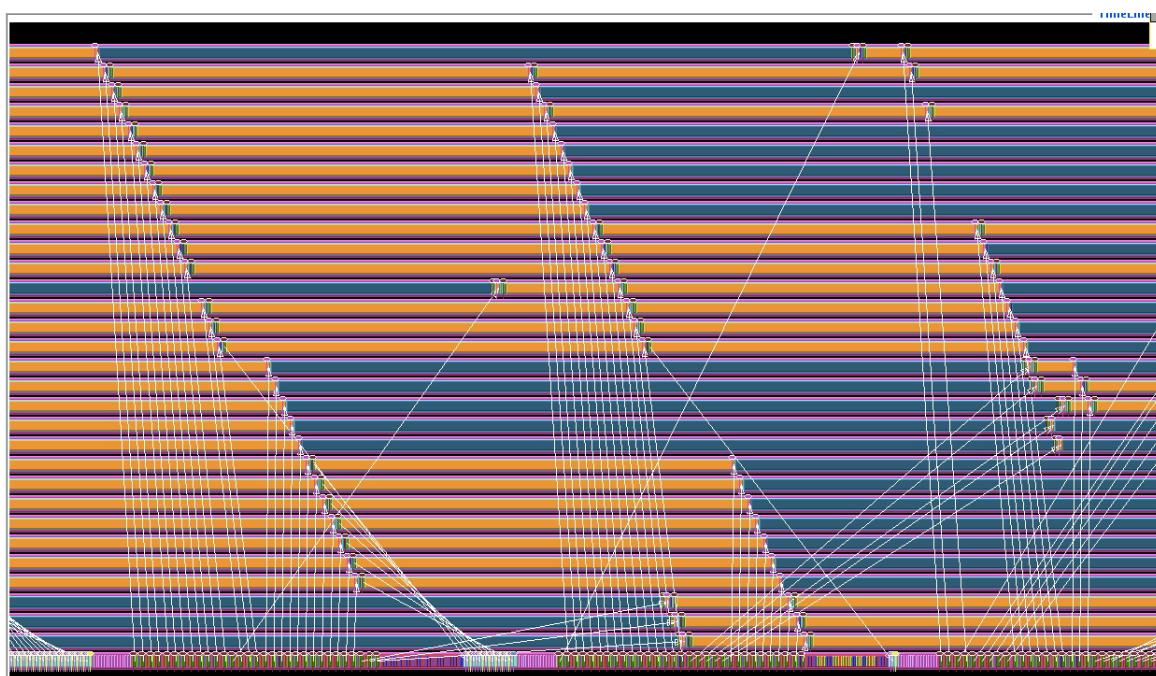


Figure 39.4: Trace of an asynchronous Mandelbrot calculation

39.3.6 One-sided solution

Let us reason about whether it is possible (or advisable) to code a one-sided solution to computing the Mandelbrot set. With active target synchronization you could have an exposure window on the host to which the worker tasks would write. To prevent conflicts you would allocate an array and have each worker write to a separate location in it. The problem here is that the workers may not be sufficiently synchronized because of the differing time for computation.

Consider then passive target synchronization. Now the worker tasks could write to the window on the manager whenever they have something to report; by locking the window they prevent other tasks from

interfering. After a worker writes a result, it can get new data from an array of all coordinates on the manager.

It is hard to get results into the image as they become available. For this, the manager would continuously have to scan the results array. Therefore, constructing the image is easiest done when all tasks are concluded.

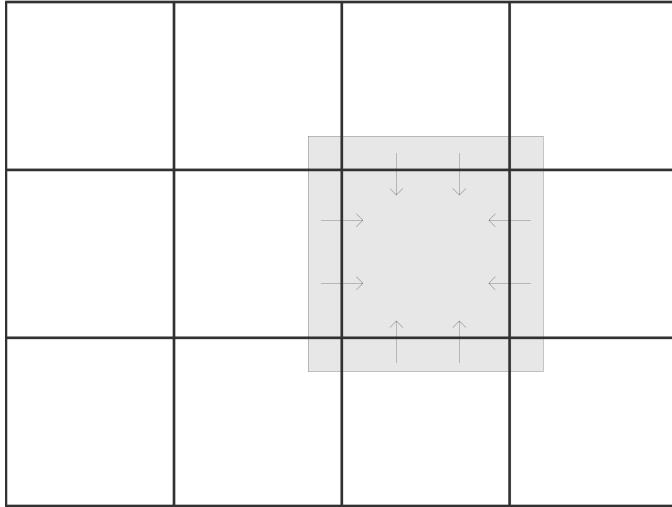


Figure 39.5: A grid divided over processors, with the ‘ghost’ region indicated

39.4 Data parallel grids

In this section we will gradually build a semi-realistic example program. To get you started some pieces have already been written: as a starting point look at `code/mpi/c/grid.cxx`.

39.4.1 Description of the problem

With this example you will investigate several strategies for implementing a simple iterative method. Let’s say you have a two-dimensional grid of datapoints $G = \{g_{ij} : 0 \leq i < n_i, 0 \leq j < n_j\}$ and you want to compute G' where

$$g'_{ij} = 1/4 \cdot (g_{i+1,j} + g_{i-1,j} + g_{i,j+1} + g_{i,j-1}). \quad (39.1)$$

This is easy enough to implement sequentially, but in parallel this requires some care.

Let’s divide the grid G and divide it over a two-dimension grid of $p_i \times p_j$ processors. (Other strategies exist, but this one scales best; see section HPSC-??.) Formally, we define two sequences of points

$$0 = i_0 < \dots < i_{p_i} < i_{p_i+1} = n_i, \quad 0 < j_0 < \dots < j_{p_j} < j_{p_j+1} = n_j$$

and we say that processor (p, q) computes g_{ij} for

$$i_p \leq i < i_{p+1}, \quad j_q \leq j < j_{q+1}.$$

From formula (39.1) you see that the processor then needs one row of points on each side surrounding its part of the grid. A picture makes this clear; see figure 39.5. These elements surrounding the processor’s own part are called the *halo* or *ghost region* of that processor.

The problem is now that the elements in the halo are stored on a different processor, so communication is needed to gather them. In the upcoming exercises you will have to use different strategies for doing so.

39.4.2 Code basics

The program needs to read the values of the grid size and the processor grid size from the commandline, as well as the number of iterations. This routine does some error checking: if the number of processors does not add up to the size of MPI_COMM_WORLD, a nonzero error code is returned.

```
ierr = parameters_from_commandline
      (argc, argv, comm, &ni, &nj, &pi, &pj, &nit);
if (ierr) return MPI_Abort(comm,1);
```

From the processor parameters we make a processor grid object:

```
processor_grid *pgrid = new processor_grid(comm,pi,pj);
```

and from the numerical parameters we make a number grid:

```
number_grid *grid = new number_grid(pgrid,ni,nj);
```

Number grids have a number of methods defined. To set the value of all the elements belonging to a processor to that processor's number:

```
grid->set_test_values();
```

To set random values:

```
grid->set_random_values();
```

If you want to visualize the whole grid, the following call gathers all values on processor zero and prints them:

```
grid->gather_and_print();
```

Next we need to look at some data structure details.

The definition of the `number_grid` object starts as follows:

```
class number_grid {
public:
    processor_grid *pgrid;
    double *values,*shadow;
```

where `values` contains the elements owned by the processor, and `shadow` is intended to contain the values plus the ghost region. So how does `shadow` receive those values? Well, the call looks like

```
grid->build_shadow();
```

and you will need to supply the implementation of that. Once you've done so, there is a routine that prints out the shadow array of each processor

```
grid->print_shadow();
```

This routine does the sequenced printing that you implemented in exercise ??.

In the file `code/mpi/c/grid_impl.cxx` you can see several uses of the macro `INDEX`. This translates from a two-dimensional coordinate system to one-dimensional. Its main use is letting you use (i, j) coordinates for indexing the processor grid and the number grid: for processors you need the translation to the linear rank, and for the grid you need the translation to the linear array that holds the values.

A good example of the use of `INDEX` is in the `number_grid::relax` routine: this takes points from the shadow array and averages them into a point of the values array. (To understand the reason for this particular averaging, see HPSC-?? and HPSC-??.) Note how the `INDEX` macro is used to index in a `ilength × jlength` target array `values`, while reading from a $(\text{ilength} + 2) \times (\text{jlength} + 2)$ source array `shadow`.

```
for (i=0; i<ilength; i++) {
    for (j=0; j<jlength; j++) {
        int c=0;
        double new_value=0.;
        for (c=0; c<5; c++) {
            int ioff=i+1+ioffsets[c], joff=j+1+joffsets[c];
            new_value += coefficients[c] *
                shadow[ INDEX(ioff, joff, ilength+2, jlength+2) ];
        }
        values[ INDEX(i, j, ilength, jlength) ] = new_value/8.;
    }
}
```

39.5 N-body problems

N-body problems describe the motion of particles under the influence of forces such as gravity. There are many approaches to this problem, some exact, some approximate. Here we will explore a number of them.

For background reading see HPSC-??.

39.5.1 Solution methods

It is not in the scope of this course to give a systematic treatment of all methods for solving the N-body problem, whether exactly or approximately, so we will just consider a representative selection.

1. Full N^2 methods. These compute all interactions, which is the most accurate strategy, but also the most computationally demanding.
2. Cutoff-based methods. These use the basic idea of the N^2 interactions, but reduce the complexity by imposing a cutoff on the interaction distance.
3. Tree-based methods. These apply a coarsening scheme to distant interactions to lower the computational complexity.

39.5.2 Shared memory approaches

39.5.3 Distributed memory approaches

Chapter 40

Bibliography, index, and list of acronyms

40.1 Bibliography

- [1] Ernie Chan, Marcel Heimlich, Avi Purkayastha, and Robert van de Geijn. Collective communication: theory, practice, and experience. *Concurrency and Computation: Practice and Experience*, 19:1749–1783, 2007.
- [2] Tom Cornebize, Franz C Heinrich, Arnaud Legrand, and Jérôme Vienne. Emulating High Performance Linpack on a Commodity Server at the Scale of a Supercomputer. working paper or preprint, December 2017.
- [3] Lisandro Dalcin. MPI for Python, homepage. <https://mpi4py.bitbucket.io/>.
- [4] Victor Eijkhout. Performance of MPI sends of non-contiguous data. *arXiv e-prints*, page arXiv:1809.10778, Sep 2018.
- [5] Eijkhout, Victor with Robert van de Geijn and Edmond Chow. *Introduction to High Performance Scientific Computing*. lulu.com, 2011. <http://www.tacc.utexas.edu/~eijkhout/istc/istc.html>.
- [6] Brice Goglin. Managing the Topology of Heterogeneous Cluster Nodes with Hardware Locality (hwloc). In *International Conference on High Performance Computing & Simulation (HPCS 2014)*, Bologna, Italy, July 2014. IEEE.
- [7] W. Gropp, E. Lusk, and A. Skjellum. *Using MPI*. The MIT Press, 1994.
- [8] Torsten Hoefer, Prabhanjan Kambadur, Richard L. Graham, Galen Shipman, and Andrew Lumsdaine. A case for standard non-blocking collective operations. In *Proceedings, Euro PVM/MPI*, Paris, France, October 2007.
- [9] Torsten Hoefer, Christian Siebert, and Andrew Lumsdaine. Scalable communication protocols for dynamic sparse data exchange. *SIGPLAN Not.*, 45(5):159–168, January 2010.
- [10] INRIA. SimGrid homepage. <http://simgrid.gforge.inria.fr/>.
- [11] L. V. Kale and S. Krishnan. Charm++: Parallel programming with message-driven objects. In *Parallel Programming using C++*, G. V. Wilson and P. Lu, editors, pages 175–213. MIT Press, 1996.
- [12] M. Li, H. Subramoni, K. Hamidouche, X. Lu, and D. K. Panda. High performance mpi datatype support with user-mode memory registration: Challenges, designs, and benefits. In *2015 IEEE International Conference on Cluster Computing*, pages 226–235, Sept 2015.
- [13] Zhenying Liu, Barbara Chapman, Tien-Hsiung Weng, and Oscar Hernandez. Improving the performance of openmp by array privatization. In *Proceedings of the OpenMP Applications and Tools 2003*

-
- International Conference on OpenMP Shared Memory Parallel Programming*, WOMPAT'03, pages 244–259, Berlin, Heidelberg, 2003. Springer-Verlag.
- [14] R. Thakur, W. Gropp, and B. Toonen. Optimizing the synchronization operations in MPI one-sided communication. *Int'l Journal of High Performance Computing Applications*, 19:119–128, 2005.

40.2 List of acronyms

AVX	Advanced Vector Extensions	OS	Operating System
BLAS	Basic Linear Algebra Subprograms	PGAS	Partitioned Global Address Space
BSP	Bulk Synchronous Parallel	PDE	Partial Differential Equation
CAF	Co-array Fortran	PRAM	Parallel Random Access Machine
CUDA	Compute-Unified Device Architecture	RDMA	Remote Direct Memory Access
DAG	Directed Acyclic Graph	RMA	Remote Memory Access
DSP	Digital Signal Processing	SAN	Storage Area Network
FEM	Finite Element Method	SaaS	Software as-a Service
FPU	Floating Point Unit	SFC	Space-Filling Curve
FFT	Fast Fourier Transform	SIMD	Single Instruction Multiple Data
FSA	Finite State Automaton	SIMT	Single Instruction Multiple Thread
GPU	Graphics Processing Unit	SM	Streaming Multiprocessor
HPC	High-Performance Computing	SMP	Symmetric Multi Processing
HPF	High Performance Fortran	SOR	Successive Over-Relaxation
ICV	Internal Control Variable	SP	Streaming Processor
MIC	Many Integrated Cores	SPMD	Single Program Multiple Data
MPMD	Multiple Program Multiple Data	SPD	symmetric positive definite
MIMD	Multiple Instruction Multiple Data	SSE	SIMD Streaming Extensions
MPI	Message Passing Interface	TLB	Translation Look-aside Buffer
MTA	Multi-Threaded Architecture	UMA	Uniform Memory Access
NIC	Network Interface Card	UPC	Unified Parallel C
NUMA	Non-Uniform Memory Access	WAN	Wide Area Network

40.3 Index

Bold reference: defining passage; italic reference: illustration.

40.3.1 General Index

- active target synchronization, 169, 174
- address
 - physical, 216
 - virtual, 216
- affinity, 361
 - process and thread, 361–364
 - thread
 - on multi-socket nodes, 301
- all-to-all, 31
- allocate
 - and private/shared data, 277
- allreduce, 30
- argc, 22, 24
- argv, 22, 24
- atomic operation, 181, 183
- atomic operations, 183
- bandwidth, 57
 - bisection, 62
- barrier, 229
 - implicit, 308
 - non-blocking, 57
- batch
 - job, 13
 - scheduler, 13
- Beowulf cluster, 12
- block row, 341
- Boolean satisfiability, 28
- boost, 15
- breakpoint, 380
- broadcast, 30
- btl_openib_eager_limit, 81
- btl_openib_rndv_eager_limit, 81
- bucket brigade, 61, 83
- C
 - MPI bindigs, see MPI, C/C++ bindings
- C++
 - MPI bindigs, see MPI, C/C++ bindings
 - standard library, 133
 - vector, 133
 - C++ iterators
 - in OMP reduction, 284
 - C99, 117
 - c(sizeof, 121
 - cacheline, 282
 - Charmpp, 13
 - chunk, 267, 268
 - chunk, 268
 - client, 165
 - collectives, 29
 - neighbourhood, 210, 232
 - non-blocking, 56
 - cancelling, 239
 - column-major storage, 127
 - communication
 - asynchronous, 101
 - blocking, 79–83
 - vs non-blocking, 231
 - buffered, 103, 231
 - non-blocking, 89–98
 - one-sided, 169–186
 - one-sided, implementation of, 186
 - persistent, 102–103, 231
 - synchronous, 101
 - two-sided, 104
 - communicator, 26, 146–158
 - info object, 224
 - inter, 153, 157, 162
 - intra, 157, 158
 - peer, 157
 - compare-and-swap, 86
 - compiler, 136
 - optimization level, 375
 - completion, 183

- local, 183
- remote, 183
- construct, 256
- contention, 62
- contiguous
 - data type, 121
- core, 19, 251
- core dump, 375
- cpp, 254
- cpuinfo, 360
- Cray
 - MPI, 14
 - T3E, 236
- critical section
 - flush at, 309
- critical section, 233, 281, 287, 315
- curly braces, 255
- Dalcin
 - Lisandro, 16
- data dependency, 297
- data race, see race condition
- datatype, 116–142
 - big, 140–141
 - derived, 121–198
 - different on sender and receiver, 124
 - elementary, 116–121
 - in C, 117
 - in Fortran, 117
 - in Python, 118
 - signature, 135, 140
- datatypes
 - derived, 116
 - elementary, 116
- ddd, 375
- DDT, 375, 382
- deadlock, 79, 89, 231, 233, 381
- debug flag, 376
- debugger, 375
- debugging, 375–382
 - parallel, 381
- dense linear algebra, 150
- directive
 - end-of, 255
- directives, 255, 255
- cpp, 258
- distributed array, 72
- distributed shared memory, 169
- doubling
 - recursive, see recursive doubling
- dynamic mode, 252
- eager limit, 80–81, 230
- Eclipse, 382
 - PTP, 382
- environment variables, 367
- epoch, 174
 - access, 174, 180
 - communication, 174
 - completion, 184
 - exposure, 174, 180
 - passive target, 182, 183
- error return, 15
- ethernet, 14
- false sharing, 263, 282
- fat-tree, 361
- fence, 174
- Fibonacci sequence, 290–292
- file
 - pointer
 - advance by write, 202
 - individual, 202
- first-touch, 305, 362
- fork/join model, 251, 296
- Fortran
 - 1-based indexing, 97
 - 90
 - bindings, 15
 - array syntax, 273
 - assumed-shape arrays in MPI, 226
 - fixed-form source, 255
 - forall loops, 273
 - Fortran90, 22
 - line length, 353
 - MPI bindings, see MPI, Fortran bindings
 - MPI equivalences of scalar types, 118
 - MPI issues, 226

Fortran2008
 MPI bindings, see MPI, Fortran2008 bindings

Fortran77
 PETSc interface, 331

Fortran90
 PETSc interface, 331

Fortran90 types
 in MPI, 226

gather, 30

Gauss-Jordan algorithm, 39

gcc
 thread affinity, 306

gdb, 375–382

gemv, 344

ghost region, 402

GNU, 375
 gdb, see gdb

graph
 topology, 210, 232
 unweighted, 212

graph topology, 361

grid
 Cartesian, 207
 periodic, 207
 processor, 361

group, 180

group of
 processors, 181

halo, 402
 update, 177

handshake, 232

hdf5, 198, 371

heat equation, 303

histogram, 288

hostname, 226

hwloc, 360

hyperthreading, 362

I/O
 in OpenMP, 272

I_MPI_ASYNC_PROGRESS_..., 94

I_MPI_EAGER_THRESHOLD, 81

ibrun, 164, 333

implicit barrier, 287
 after single directive, 272

indexed
 data type, 121

inner product, 32

input redirection
 shell, 235

instrumentation, 383

Intel, 232
 compiler
 thread affinity, 306
 compiler suite, 360
 Haswell, 362
 Knight's Landing, 251, 311
 thread placement, 304
 MPI, 14, 94
 mpi, 81, 164
 Paragon, 94
 SandyBridge, 362
 Sandybridge, 251

interconnect, 385

Internal Control Variable (ICV), 314–316

jumpshot, 384

KIND, 137

KMP_AFFINITY, 306

Laplace equation, 343

latency, 57
 hiding, 93, 231, 339
 latency hiding, 344

lexical scope, 274

linked list, 293

listing, 389

load balancing, 263

load imbalance, 263

lock, 288, 288–290
 flush at, 309
 nested, 290

Lonestar5, 362

LU factorization, 267

Lustre, 371

malloc
 and private/shared data, 277

malloc, 305

manager-worker, 395

manager-worker model, 97

Mandelbrot set, 28, 279, 395

master-worker, 183

master-worker model, 99

matching, 232

matching queue, 99

matrix
 sparse, 56, 344
 transposition, 152

matrix-vector product
 dense, 44
 sparse, 47

Mellanox, 232

memory
 coherent, 184

message
 collision, 232
 count, 100
 source, 77
 status, 88, 99
 error, 100
 source, 99
 tag, 100
 tag, 77, 367

messsage
 target, 77

Monte Carlo codes, 28

motherboard, 250

move_pages, 306

MPI
 1, 207, 213
 3, 133, 141, 181, 215
 C++ bindings removed, 15
 Fortran2008 interface, 15

C/C++ bindings, 15

constants, 236–238, 241–242
 compile-time, 236, 241
 link-time, 236, 241

datatype

extent, 137

size, 136

subarray, 137

vector, 137

Fortran bindings, 15

Fortran issues, see Fortran, MPI issues

Fortran2008 bindings, 15

I/O, 225, 371

initialization, 22

Python bindings, 16

Python issues, 238

semantics, 232

tools interface, 227

version, 227

MPI-4, 121

mpi.h, 22

mpi.h, 16

MPI/O, 198–205

mpi4py, 16

mpi_f08, 15

mpich, 14

mpieexec, 13, 21, 24, 162
 options, 14

mpieexec, 21

mpif.h, 22

mpirun, 13, 14, 24, 145
 and environment variables, 367

MPL, 15

mulpd, 311

mulsd, 311

multicore, 252

Multiple Program Multiple Data (MPMD), 236

MV2_IBA_EAGER_THRESHOLD, 81

mvapich, 367

mvapich2, 81, 366

nested parallelism, 258–260

netcdf, 198

network
 card, 232
 contention, 232
 port
 oversubscription, 232

node, 19

cluster, 250
norm
 one, 55
np.frombuffer, 54
numactl, 362
numerical integration, 262
numpy, 16, 121, 238

od, 198
offloading
 vs onloading, 232
omp
 barrier
 implicit, 287
 for
 barrier behaviour, 287
 reduction, 281–285
 user-defined, 284–285
OMP_NUM_THREADS, 367
onloading, see offloading, vs onloading
OpenMP
 accelerator support in, 318
 co-processor support in, 318
 compiling, 253–254
 environment variables, 254, 314–316
 library routines, 314–316
 places, 301
 running, 254
 tasks, 293–300
 data, 294–295
 dependencies, 297–298
 synchronization, 295–297
 version 4, 299
OpenMPI, 81
operating system, 318
operator, 54–56
 predefined, 54
 user-defined, 54
origin, 169, 180
owner computes, 73

package, 360
packing, 141
page
 small, 216
 table, 216
parallel
 data, 306
 embarrassingly, 306
parallel region, 252, 257–260, 271
 barrier at the end of, 287
 dynamic scope, 259, 275
 flush at, 309
parallel regions
 nested, 315
paraprof, 383
ParMetis, 232
pass-by-reference
 in Fortran, 75
passive target synchronization, 169, 181, 182
pbng, 306
persistent communication, see communication, persistent
PETSc, 232
PETSC_ARCH, 332
PETSC_DIR, 332
pin a thread, 362
ping-pong, 75, 230
PMI_RANK, 236
point-to-point, 74
polling, 94, 95
posting, 89
pragma, see for list see under ‘omp’, 255
preconditioner, 350
prefix operation, 52
process, 19
processes status of, 235
producer-consumer, 307
progress
 asynchronous, 94
purify, 379
PVM, 13, 162
Python
 MPI bindigs, see MPI, Python bindings
 PETSc interface, 332

race condition, 181, 183, 281, 287, 308, 316
 in OpenMP, 308–309

random number generation, 279
random number generator, 317
Ranger, 361
rar, 184
raw, 184
ray tracing, 217
recursive doubling, 62
redirection, see shell, input redirection
reduction, 30
region of code, 256
register
 SSE2, 311
residual, 349
Riemann sums, 262
RMA
 active, 169
 passive, 169
root, 34
root process, 30

scalapack, 351
scan, 31
 exclusive, 53
 inclusive, 52
scatter, 30
sched_setaffinity, 306
schedule
 clause, 267
scope
 lexical, 252
 of variables, 252
SEEK_SET, 205
segmentation fault, 377
segmented scan, 53
send
 buffered, 104
sentinel, 255, 258
sequential
 semantics, 331
sequential consistency, 309
serialization, 82
server, 165
SetThreadAffinityMask, 306
shared data, 252

shmem, 236
silo, 198
SimGrid, 61, 385–386
 compiler, 385
Single Program Multiple Data (SPMD), 255, 257
sizeof, 171, 226
smpicc, 385
smpirun, 385
socket, 19, 216, 250, 365
sort
 odd-even transposition, 86
 radix, 46
 swap, 86
sorting
 radix, 46
sparse matrix vector product, 53
spin-lock, 315
ssh, 13
stack, 315
 overflow, 275
 per thread, 275
Stampede
 compute node, 362
 largemem node, 362
 node, 251
standard deviation, 30
start/affinity, 306
status
 of received message, 88
stderr, 235
stdout, 235
stdout/err of, 235
stencil
 nine-point, 208
 star, 208
storage association, 275, 277
storage_size, 121
stride, 127
struct
 data type, 121
structured block, 256
Sun
 compiler, 306

SUNW_MP_PROCBIND, 306
symbol table, 375, 376
synchronization
 in OpenMP, 286–292

TACC
 portal, 332

tacc_affinity, 362, 365

tag
 bound on value, 227

target, 169, 180
 active synchronization, see active target synchronization
 passive synchronization, see passive target synchronization

task
 scheduler, 293
 scheduling point, 298

taskset, 306

TAU, 383–384

thread
 affinity, 301–305
 migrating a, 304
 private data, 278

thread-safe, 316–317

threads, 251
 hardware, 252, 362
 initial, 258
 master, 252, 258
 team of, 251, 258

time slicing, 19, 252

time-slicing, 162

timing
 MPI, 229–230

topology
 virtual, 207

TotalView, 375, 382

tree
 traversal
 post-order, 300
 pre-order, 300

tunnel
 ssh, 235

ulimit, 275

Unix
 process, 275

valarray, 306

valgrind, 379–380

vector
 data type, 121
 instructions, 310

verbatim, 389

virtual shared memory, 169

wall clock, 229

war, 184

waw, 184

while loop, 293

while loops, 270

window, 169–174
 consistency, 183
 displacement unit, 175, 185
 info object, 224
 memory, see also memory model
 model, 184
 separate, 184
 unified, 184
 memory allocation, 170–171
 private, 184
 public, 184

work sharing, 252

work sharing construct, 271

workshare
 flush after, 309

worksharing constructs
 implied barriers at, 287

wormhole routing, 62

wraparound connections, 207

XSEDE
 portal, 332

Zoltan, 232

40.3.2 Index of MPI commands

0_MPI_OFFSET_KIND, 205
accumulate_ops, 184
accumulate_ordering, 184
alloc_shared_noncontig, 216

hwloc, 362

KSPSolve, 350

MPI.SUM, 32
MPI_Abort, 22, 55, 225, 235
MPI_Accumulate, 174, 178, 184
MPI_ADDRESS_KIND, 118, 237
MPI_Aint, 117, 118, 175
 in Fortran, 118
MPI_Allgather, 44
MPI_Allgatherv, 50
MPI_Alloc_mem, 171, 171
MPI_Allreduce, 32, 32, 46
MPI_Alltoall, 44, 46
MPI_Alltoallv, 46, 50, 52
MPI_ANY_SOURCE, 47, 49, 59, 77, 88, 97, 99, 100,
 169, 232, 237, 238
MPI_ANY_TAG, 77, 85, 88, 100, 237
MPI_APPNUM, 227
MPI_ARGV_NULL, 237
MPI_ARGVS_NULL, 237
MPI_ASYNC_PROTECTS_NONBLOCKING, 237
MPI_Attr_get, 162, 227
MPI_BAND, 54
MPI_Barrier, 49, 205, 229
MPI_Bcast, 36, 36
MPI_BOR, 54
MPI_BOTTOM, 117, 185, 237, 241
MPI_Bsend, 103, 104, 104
MPI_Bsend_init, 103, 104, 104
MPI_BSEND_OVERHEAD, 104, 142, 237
MPI_Buffer_attach, 104
MPI_Buffer_detach, 104
MPI_BXOR, 54
MPI_BYTE, 117, 121
MPI_Cancel, 238, 239

MPI_CART, 207
MPI_Cart_coords, 208
MPI_Cart_create, 208
MPI_Cart_rank, 208
MPI_CHAR, 117
MPI_CHARACTER, 117
MPI_Close_port, 165
MPI_Comm, 15, 17, 26, 145, 237, 241
MPI_Comm_accept, 165, 165
MPI_Comm_compare, 158
MPI_Comm_connect, 165, 167, 167, 224
MPI_Comm_create, 153
MPI_Comm_create_errhandler, 225
MPI_Comm_create_group, 153
MPI_Comm_disconnect, 167
MPI_Comm_dup, 147, 148, 224
MPI_Comm_dup_with_info, 224
MPI_Comm_free, 150
MPI_Comm_get_errhandler, 225
MPI_Comm_get_info, 224
MPI_Comm_get_parent, 165
MPI_Comm_group, 153, 158
MPI_Comm_join, 167
MPI_COMM_NULL, 145
MPI_Comm_rank, 26, 26, 150, 158
MPI_Comm_remote_group, 158
MPI_Comm_remote_size, 158, 165
MPI_COMM_SELF, 145
MPI_Comm_set_errhandler, 224, 225
MPI_Comm_set_info, 224
MPI_Comm_set_name, 153
MPI_Comm_size, 26, 26, 77, 158
MPI_Comm_spawn, 162, 226
MPI_Comm_spawn_multiple, 165, 227
MPI_Comm_split, 49, 150, 150
MPI_Comm_split_type, 215, 217
MPI_Comm_test_inter, 158
MPI_COMM_TYPE_SHARED, 215
MPI_COMM_WORLD, 26, 36, 145, 146, 153, 157,
 162, 165, 217, 226, 227, 236, 237, 239,
 241, 331, 334

MPI_Compare_and_swap, 182
MPI_COMPLEX, 117
MPI_Count, 141
MPI_COUNT_KIND, 117, 237
MPI_Datatype, 41, 122, 122
MPI_DATATYPE_NULL, 117, 122
MPI_DISPLACEMENT_CURRENT, 202
MPI_DIST_GRAPH, 207
MPI_Dist_graph_create, 210, 210, 212
MPI_Dist_graph_create_adjacent, 210
MPI_Dist_graph_neighbors_count, 212
MPI_DOUBLE, 116, 117
MPI_DOUBLE_COMPLEX, 117
MPI_DOUBLE_INT, 54
MPI_DOUBLE_PRECISION, 116, 117
MPI_ERR_ARG, 224
MPI_ERR_BUFFER, 104, 224
MPI_ERR_COMM, 224, 226
MPI_ERR_INFO, 224
MPI_ERR_INTERN, 104, 224
MPI_ERR_OTHER, 224
MPI_ERR_PORT, 167, 224
MPI_ERR_SERVICE, 167, 224
MPI_ERRCODES_IGNORE, 162, 237
MPI_Errhandler, 225
MPI_ERROR, 100, 225
MPI_Error_string, 224, 225
MPI_ERRORS_ARE_FATAL, 225
MPI_ERRORS_RETURN, 225, 225
MPI_Exscan, 53, 53
mpi_f08, 15
MPI_Fetch_and_op, 181, 181, 183
MPI_File, 198
MPI_File_close, 198
MPI_File_delete, 200
MPI_File_get_size, 205
MPI_File_get_view, 205
MPI_File_open, 198
MPI_File_preallocate, 205
MPI_File_read, 200
MPI_File_read_all, 200
MPI_File_read_at, 200
MPI_File_read_at_all, 200
MPI_File_read_ordered, 200
MPI_File_read_shared, 200
MPI_File_seek, 200, 202
MPI_File_set_atomicity, 205
MPI_File_set_size, 205
MPI_File_set_view, 202, 205
MPI_File_sync, 200
MPI_File_write, 200, 202, 205
MPI_File_write_all, 200
MPI_File_write_at, 200, 202, 202
MPI_File_write_at_all, 200
MPI_File_write_ordered, 200
MPI_File_write_shared, 200
MPI_Finalize, 22, 22, 24
MPI_Finalized, 24
MPI_FLOAT, 38, 116, 117
MPI_Gather, 41, 50, 51, 129
MPI_Gatherv, 50, 50, 51
MPI_Get, 174, 176
MPI_Get_accumulate, 178, 181, 181, 184
MPI_Get_address, 118, 133, 136
MPI_Get_count, 98, 100, 140
MPI_Get_elements, 140
MPI_Get_elements_x, 141
MPI_Get_processor_name, 24, 24, 226
MPI_Get_version, 227
MPI_GRAPH, 207
MPI_Graph_create, 213
MPI_Group, 153, 153
MPI_Group_difference, 153
MPI_Group_excl, 153
MPI_Group_incl, 153
MPI_HOST, 227
MPI_Iallgather, 56
MPI_Iallreduce, 56
MPI_Ibarrier, 56, 57
MPI_Ibcast, 56
MPI_Ibsend, 104
MPI_IN_PLACE, 34, 43, 237
MPI_Info, 185, 216, 221
MPI_Info_create, 221
MPI_Info_delete, 221
MPI_Info_dup, 221

MPI_INFO_ENV, 24, 221
MPI_Info_free, 221
MPI_Info_get, 221
MPI_Info_get_nkeys, 221
MPI_Info_get_nthkey, 221
MPI_Info_set, 221
MPI_Init, 22, 22, 24, 233, 236, 239, 241, 333, 356
 in Fortran, 226
MPI_Init_thread, 233, 239, 367, 367
MPI_Initialized, 24
MPI_INT, 38, 117, 239
MPI_INTEGER, 38, 117
MPI_INTEGER1, 117
MPI_INTEGER16, 117
MPI_INTEGER2, 117
MPI_INTEGER4, 117
MPI_INTEGER8, 117
MPI_INTEGER_KIND, 237
MPI_Intercomm_create, 153
MPI_Intercomm_merge, 158
MPI_IO, 227
MPI_Iprobe, 94, 98
MPI_Irecv, 47, 63, 89, 91, 93, 96, 98, 99, 102
MPI_Is_thread_main, 233
MPI_Isend, 63, 89, 91, 102, 178
MPI_KEYVAL_INVALID, 237
MPI_LAND, 54
MPI_LOCK_EXCLUSIVE, 183, 237
MPI_LOCK_SHARED, 182, 238
MPI_LOGICAL, 117
MPI_LONG, 117
MPI_LONG_DOUBLE, 117
MPI_LONG_LONG_INT, 117
MPI_LOR, 54
MPI_LXOR, 54
MPI_MAX, 35, 54
MPI_MAX_DATAREP_STRING, 237
MPI_MAX_ERROR_STRING, 225, 237
MPI_MAX_INFO_KEY, 221, 237
MPI_MAX_INFO_VAL, 237
MPI_MAX_LIBRARY_VERSION_STRING, 237
MPI_MAX_OBJECT_NAME, 237
MPI_MAX_PORT_NAME, 165, 237
MPI_MAX_PROCESSOR_NAME, 24, 24, 226, 236, 237, 241
MPI_MAXLOC, 54, 54
MPI_Message, 99
MPI_MIN, 54
MPI_MINLOC, 54
MPI_MODE_APPEND, 200
MPI_MODE_CREATE, 200
MPI_MODE_DELETE_ON_CLOSE, 200
MPI_MODE_EXCL, 200
MPI_MODE_NOCHECK, 179
MPI_MODE_NOPRECEDE, 179
MPI_MODE_NOPUT, 179
MPI_MODE_NOSTORE, 179
MPI_MODE_NOSUCCEED, 179
MPI_MODE_RDONLY, 198
MPI_MODE_RDWR, 198
MPI_MODE_SEQUENTIAL, 200
MPI_MODE_UNIQUE_OPEN, 200
MPI_MODE_WRONLY, 198
MPI_Mprobe, 99
MPI_Mrecv, 99
MPI_MULT, 53
MPI_Neighbor_allgather, 212
MPI_Neighbor_allgatherv, 212
MPI_Neighbor_allreduce, 212
MPI_Neighbor_alltoall, 212
MPI_Neighbor_alltoallv, 212
MPI_Neighbor_alltoallw, 212
MPI_NO_OP, 178, 184
MPI_Offset, 202
MPI_OFFSET_KIND, 117, 205, 237
MPI_OP, 54
MPI_Op, 34, 52, 53, 55, 56, 225
MPI_Op_commutative, 55
MPI_Op_create, 53, 54, 55
MPI_OP_NULL, 55
MPI_Open_port, 165, 165
MPI_ORDER_C, 129
MPI_ORDER_FORTRAN, 129
MPI_Pack, 141
MPI_Pack_size, 104, 142
MPI_PACKED, 117, 141

MPI_Probe, 98, 99
MPI_PROC_NULL, 85, 85, 86, 93, 158, 175, 208, 227, 237, 393
MPI_PROD, 35, 54
MPI_Publish_name, 167
MPI_Put, 174, 175, 176, 177
MPI_Query_thread, 233
MPI_Raccumulate, 178
MPI_REAL, 38, 116, 117
MPI_REAL2, 117
MPI_REAL4, 117
MPI_REAL8, 117
MPI_Recv, 77, 77, 79, 80, 83, 88, 89, 93, 99, 100, 238
MPI_Recv_init, 102, 102
MPI_Reduce, 34, 35, 50, 178
MPI_Reduce_local, 56
MPI_Reduce_scatter, 46, 47, 47, 49, 186
MPI_REPLACE, 178, 178
MPI_Request, 17, 56, 91, 98, 102
MPI_Request_free, 98, 102, 239
MPI_Request_get_status, 98
MPI_REQUEST_NULL, 98
MPI_Rget, 178
MPI_Rget_accumulate, 178
MPI_ROOT, 158, 238
MPI_Rput, 178
MPI_Rsend, 233
MPI_Rsend_init, 103
MPI_Scan, 52, 52, 53
MPI_Scatter, 39, 43
MPI_Scatterv, 47, 50
MPI_SEEK_CUR, 200, 202
MPI_SEEK_END, 200
MPI_SEEK_SET, 200, 205
MPI_Send, 61, 75, 77, 79–81, 83, 88, 89, 93, 233
MPI_Send_init, 102, 102
MPI_Sendrecv, 83, 85, 86, 89, 93, 393
MPI_Sendrecv_replace, 86
MPI_SHORT, 117
MPI_SIGNED_CHAR, 117
MPI_Sizeof, 118, 118, 140, 226
MPI_SOURCE, 96, 97, 99, 99–101
MPI_Ssend, 81, 101, 233
MPI_Ssend_init, 103
MPI_Start, 102, 102
MPI_Startall, 102, 102
MPI_Status, 77, 85, 88, 91, 95, 98, 99, 99–101
MPI_STATUS_IGNORE, 77, 88, 91, 96, 237, 241
MPI_STATUS_SIZE, 237
MPI_STATUSES_IGNORE, 91, 237
MPI_SUBARRAYS_SUPPORTED, 237
MPI_SUBVERSION, 227, 237
MPI_SUCCESS, 16, 104, 224, 225
MPI_SUM, 32, 35, 50, 53, 54
MPI_TAG, 100
MPI_TAG_UB, 77, 227, 227
MPI_Test, 98, 98, 239
MPI_Testall, 98
MPI_Testany, 98
MPI_THREAD_FUNNELED, 233
MPI_THREAD_FUNNELLED, 367
MPI_THREAD_MULTIPLE, 233, 367
MPI_THREAD_SERIAL, 367
MPI_THREAD_SERIALIZED, 233
MPI_THREAD_SINGLE, 233, 367
MPI_Topo_test, 207
MPI_Type_commit, 122
MPI_Type_contiguous, 122, 122, 238
MPI_Type_create_f90_complex, 118
MPI_Type_create_f90_integer, 118
MPI_Type_create_f90_real, 118
MPI_Type_create_hindexed, 133
MPI_Type_create_hindexed_block, 133
MPI_Type_create_struct, 122, 133
MPI_Type_create_subarray, 122, 129, 129
MPI_Type_extent, 137
MPI_Type_free, 122
MPI_Type_get_extent, 137
MPI_Type_get_true_extent, 137
MPI_Type_hindexed, 122
MPI_Type_indexed, 122, 129, 133
MPI_Type_match_size, 140
MPI_Type_size, 136
MPI_Type_struct, 133
MPI_Type_vector, 122, 124

MPI_TYPECLASS_COMPLEX, 140
MPI_TYPECLASS_INTEGER, 140
MPI_TYPECLASS_REAL, 140
MPI_UB, 136
MPI_UNDEFINED, 207, 237
MPI_UNIVERSE_SIZE, 162, 227
MPI_Unpack, 141
MPI_Unpublish_name, 167, 224
MPI_UNSIGNED, 117
MPI_UNSIGNED_CHAR, 117
MPI_UNSIGNED_LONG, 117
MPI_UNSIGNED_SHORT, 117
MPI_UNWEIGHTED, 212, 237
MPI_VERSION, 227, 236, 237, 241
MPI_Wait, 56, 91, 91, 94, 99, 102, 174, 239
MPI_Wait..., 88, 91, 99, 102
MPI_Waitall, 91, 95, 95, 169
MPI_Waitany, 91, 95, 95, 96
MPI_Waitsome, 91, 95
MPI_WEIGHTS_EMPTY, 237
MPI_Win, 118, 170
MPI_Win_allocate, 171, 171, 184, 185
MPI_Win_allocate_shared, 171, 184, 185,
 216, 216
MPI_Win_attach, 184
MPI_Win_complete, 180
MPI_Win_create, 118, 121, 171, 171, 184, 185,
 238
MPI_Win_create_dynamic, 171, 184, 185
MPI_Win_detach, 184
MPI_Win_fence, 31, 174, 179, 394
MPI_WIN_FLAVOR_ALLOCATE, 185
MPI_WIN_FLAVOR_CREATE, 185
MPI_WIN_FLAVOR_DYNAMIC, 185
MPI_WIN_FLAVOR_SHARED, 185
MPI_Win_flush, 183
MPI_Win_flush..., 178
MPI_Win_flush_all, 183
MPI_Win_flush_local, 183
MPI_Win_flush_local_all, 183
MPI_Win_get_info, 224
MPI_Win_lock, 179, 181, 182
MPI_Win_lock_all, 183
MPI_Win_post, 179, 180
MPI_WIN_SEPARATE, 185
MPI_Win_set_info, 224
MPI_Win_shared_query, 216, 216
MPI_Win_start, 179, 180
MPI_Win_sync, 183
MPI_WIN_UNIFIED, 185
MPI_Win_unlock, 183, 183
MPI_Win_unlock_all, 183
MPI_Win_wait, 180
MPI_Wtick, 229, 230
MPI_Wtime, 77, 229
MPI_WTIME_IS_GLOBAL, 227, 229
no_locks, 184
numactl, 306
omp_get_wtick, 316
omp_get_wtime, 316
PetscInitialize, 356
PMPI_..., 230
same_op, 184
same_op_no_op, 184
tacc_affinity, 306

40.3.3 Index of OpenMP commands

_OPENMP, 254

omp

- atomic, 288, 308
- barrier, 286
- cancel, 299
- critical, 288, 317
- declare simd, 310
- flush, 289, 309
- lastprivate, 270
- master, 272, 273, 367
- ordered, 269
- parallel, 255, 257, 304
- parallel for, 261
- private, 275
- section, 271
- sections, 271, 278
- simd, 310, 310
- single, 272
- task, 293, 296, 297
- taskgroup, 296, 297
- taskwait, 296--298
- taskyield, 298
- threadprivate, 278, 306, 317
- workshare, 273

omp clause

- aligned, 310
- collapse, 268
- copyin, 279
- copyprivate, 273, 279
- default, 276
 - firstprivate, 277
 - none, 277
 - private, 276
 - shared, 276
- depend, 297
- firstprivate, 278, 294
- lastprivate, 278
- linear, 310
- nowait, 269, 287, 317
- ordered, 269
- private, 275

proc_bind, 302, 304

reduction, 281, 284

safelen(n), 310

schedule, 317

- auto, 265
- chunk, 264
- guided, 264
- runtime, 265
- untied, 298

OMP_CANCELLATION, 314

OMP_DEFAULT_DEVICE, 315

omp_destroy_nest_lock, 290

OMP_DISPLAY_ENV, 301, 314

OMP_DYNAMIC, 280, 315, 315

omp_get_active_level, 314

omp_get_ancestor_thread_num, 314

omp_get_dynamic, 314, 315

omp_get_level, 314

omp_get_max_active_levels, 314

omp_get_max_threads, 314, 315

omp_get_nested, 314, 315

omp_get_num_procs, 314, 315

omp_get_num_threads, 257, 314, 315

omp_get_schedule, 267, 314, 315

omp_get_team_size, 314

omp_get_thread_limit, 314

omp_get_thread_num, 257, 314, 315

omp_get_wtick, 314

omp_get_wtime, 314

omp_in, 284

omp_in_parallel, 260, 314, 315

omp_init_nest_lock, 290

OMP_MAX_ACTIVE_LEVELS, 315

OMP_MAX_TASK_PRIORITY, 315

OMP_NESTED, 315, 315

OMP_NUM_THREADS, 254, 315, 315

omp_out, 284

OMP_PLACES, 301, 303, 315

omp_priv, 284

OMP_PROC_BIND, 301, 304, 315, 316

omp_sched_affinity, 267

INDEX

omp_sched_auto, 267
omp_sched_dynamic, 267
omp_sched_guided, 268
omp_sched_runtime, 268
omp_sched_static, 268
OMP_SCHEDULE, 265, 267, 268, 315, 315
omp_set_dynamic, 314, 315
omp_set_max_active_levels, 314
omp_set_nest_lock, 290
omp_set_nested, 314, 315
omp_set_num_threads, 314, 315
omp_set_schedule, 267, 314, 315
OMP_STACKSIZE, 275, 315, 315
omp_test_nest_lock, 290
OMP_THREAD_LIMIT, 315
omp_unset_nest_lock, 290
OMP_WAIT_POLICY, 258, 315, 315

schedule, 263

wait-policy-var, 315

40.3.4 Index of PETSc commands

-download_mpich, 334
-with-precision, 334
-with-scalar-type, 334

ADD_VALUES, 339, 344

CHKERRA, 353
CHKERRQ, 353

DMCreate2d, 346

INSERT_VALUES, 339, 344

KSP, 349
KSPConvergedReason, 350
KSPGetConvergedReason, 350
KSPGetIterationNumber, 350
KSPReasonView, 350
KSPSetFromOptions, 351

MAT_FLUSH_ASSEMBLY, 344
MatAssemblyBegin, 344, 344
MatAssemblyEnd, 344, 344
MatCreate, 341
MatCreateShell, 346
MatCreateSubMatrices, 346
MatCreateSubMatrix, 346
MatGetRow, 344
MatMatMult, 344
MATMPIAIJ, 341
MatMPIAIJSetPreallocation, 343
MATMPIDENSE, 341
MatMult, 344, 346
MatMultAdd, 344
MatMultHermitianTranspose, 344
MatMultTranspose, 344
MatRestoreRow, 344
MATSEQAIJ, 341
MatSeqAIJSetPreallocation, 342
MATSEQDENSE, 341
MatSetSizes, 341
MatSetType, 341
MatSetValue, 343

MatSetValues, 343
MatShellGetContext, 346
MatShellSetContext, 346
MatShellSetOperation, 346
MatSizes, 341
MPIU_COMPLEX, 338
MPIU_SCALAR, 338

PETSC_ARCH, 333
PETSC_CC_INCLUDES, 333
PETSC_COMM_WORLD, 334
PETSC_DECIDE, 337, 338
PETSC_DIR, 333
PETSC_FC_INCLUDES, 333
PETSC_MEMALIGN, 354
PetscComplex, 337, 337, 338
PetscErrorCode, 337, 353
PetscFinalize, 354
PetscFree, 354
PetscInitialize, 333, 354
PetscInt, 337
PetscMalloc, 354
PetscMallocDump, 354
PetscNew, 354
PetscPrintf, 353
PetscReal, 337, 337
PetscScalar, 337, 338
PetscSplitOwnership, 337
PetscSynchronizedFlush, 353
PetscSynchronizedPrintf, 353

SETERRA, 353
SETERRQ, 353

VecAssemblyBegin, 339, 339
VecAssemblyEnd, 339
VecCreate, 338
VecCreateMPIWithArray, 338
VecCreateSeqWithArray, 338
VecDestroy, 338
VecDot, 339
VecDuplicate, 338

INDEX

VecGetArray, [341](#)
VecGetArrayRead, [341](#)
VecGetOwnershipRange, [338](#)
VecGetSize, [338](#)
VECMPI, [338](#)
VecNorm, [339](#)
VecRestoreArray, [341](#)
VecRestoreArrayRead, [341](#)
VecScale, [339](#)
VECSEQ, [338](#)
VecSet, [339](#)
VecSetSizes, [338](#)
VecsetType, [338](#)
VecSetValue, [339](#)
VecSetValues, [339](#)
VecView, [339](#)

