INTRODUCTION

MPI Users' Guide in FORTRAN

INTRODUCTION TO MESSAGE PASSING PROGRAMMING

MPI User Guide in FORTRAN

Dr Peter S. Pacheco
Department of Mathematics
University of San Francisco
San Francisco, CA 94117
March 26, 1995

Woo Chat Ming
Computer Centre
University of Hong Kong
Hong Kong
March 17, 1997

1INTRODUCTION....2

2GREETINGS !....3

2.1General MPI Programs....4
2.2Finding out About the Rest of the World....5
2.3Message : Data + Envelope....6
2.4MPI_Send and MPI_Recv....7

3AN APPLICATION....10

3.1Serial program....10
3.2Parallelizing the Trapezoid Rule....11
3.3I/O on Parallel Processors....15

4COLLECTIVE COMMUNICATION....19

4.1Tree-Structured Communication....19
4.2Broadcast....21
4.3Reduce....22
4.4Other Collective Communication Functions....24

5GROUPING DATA FOR COMMUNICATION....26

5.1The Count Parameter....26
5.2Derived Types and MPI_Type_struct....27
5.3Other Derived Datatype Constructors....31
5.4Pack/Unpack....32
5.5Deciding Which Method to Use....34

6COMMUNICATORS AND TOPOLOGIES....37

6.1Fox's Algorithm....37
6.2Communicators....38
6.3Working with Groups, Contexts, and Communicators....40
6.4MPI_Comm_split....43
6.5Topologies....44
6.6MPI_Cart_sub....47
6.7Implementation of Fox's Algorithm....48

7WHERE TO GO FROM HERE....52

7.1What We Haven't Discussed....52
7.2Implementations of MPI....53
7.3More Information on MPI....53
7.4The Future of MPI....54

8COMPILING AND RUNNING MPI PROGRAMS....55

9REFERENCE....56

1 Introduction

The Message-Passing Interface or MPI is a library of functions and macros that can be used in C, FORTRAN, and C++ programs, As its name implies, MPI is intended for use in programs that exploit the existence of multiple processors by message-passing.

MPI was developed in 1993-1994 by a group of researchers from industry, government, and academia. As such, it is one of the first standards for programming parallel processors, and it is the first that is based on message-passing.

In 1995, *A User's Guide to MPI* has been written by Dr Peter S. Pacheco. This is a brief tutorial introduction to some of the more important feature of the MPI for C programmers. It is a nicely written documentation and users in our university find it very concise and easy to read.

However, many users of parallel computer are in the scientific and engineers community and most of them use FORTRAN as their primary computer language. Most of them don't use C language proficiently. This situation occurs very frequently in Hong Kong. A a result, the "*A User's Guide to MPI" i*s translated to this guide in Fortran to address for the need of scientific programmers.

Acknowledgments. I gratefully acknowledge Dr Peter S. Pacheco for the use of C version of the user guide on which this guide is based. I would also gratefully thanks to the Computer Centre of the University of Hong Kong for their human resource support of this work. And I also thanks to all the research institution which supported the original work by Dr Pacheco.

2Greetings !

The first program that most of us saw was the "Hello, world!" program in most of introductory programming books. It simply prints the message "Hello, world!". A variant that makes some use of multiple processes is to have each process send a greeting to another process.

In MPI, the process involved in the execution of a parallel program are identified by a sequence of non-negative integers. If there are p processes executing a program, they will have ranks 0, 1,..., p-1. The following program has each process other than 0 send a message to process 0, and process 0 prints out the messages it received.

```
program greetings
include 'mpif.h'

integer my_rank
integer p
integer source
integer dest
integer tag
character*100 message
character*10 digit_string
integer size
integer status(MPI_STATUS_SIZE)
integer ierr
call MPI_Init(ierr)
call MPI_Comm_rank(MPI_COMM_WORLD, my_rank, ierr)
call MPI_Comm_size(MPI_COMM_WORLD, p, ierr)
if (my_rank  .NE.  0) then
  [1]write(digit_string,FMT="(I3)") my_rank
  message = 'Greetings from process '
+    //[2] trim(digit_string) // ' !'
  dest = 0
  tag = 0
  call MPI_Send(message, len_trim(message)[3],
+    MPI_CHARACTER, dest, tag, MPI_COMM_WORLD, ierr)
else
  do source = 1, p-1
```

[1] This line changes the binary format (integer) `my_rank` to string format `digit_string`.

[2] // is concatenation operator which combines two strings to form a third, composite string. E.g. `'This ' // 'is ' // 'a ' // 'dog.'` is equal to the string `'This is a dog.'`

[3] LEN_TRIM(STRING) returns the length of the character argument without counting trailing blank.

```
      tag = 0
      call MPI_Recv(message, 100, MPI_CHARACTER,
  +         source, tag, MPI_COMM_WORLD, status, ierr)
      write(6,FMT="(A)") message
    enddo
   endif

  call MPI_Finalize(ierr)
  end program greetings
```

The details of compiling and executing this program is in chapter 8.

When the program is compiled and run with two processes, the output should be

```
Greetings from process 1!
```

If it's run with four processes, the output should be

```
Greetings from process 1!
Greetings from process 2!
Greetings from process 3!
```

Although the details of what happens when the program is executed vary from machine to machine, the essentials are the same on all machines. Provided we run one process on each processor.

1. The user issues a directive to the operating system which has the effect of placing a copy of the executable program on each processor.

2. Each processor begins execution of its copy of the executable.

3. Different processes can execute different statements by branching within the program. Typically the branching will be based on process ranks.

So the Greetings program uses the *Single Program Multiple Data* or *SPMD* paradigm. That is, we obtain the *effect* of different programs running on different processors by taking branches within a single program on the basis of process rank : the statements executed by process 0 are different from those executed by the other processes, even though all processes are running the same program. This is the most commonly used method for writing MIMD programs, and we'll use it exclusively in this Guide.

2.1 General MPI Programs

Every MPI program must contain the preprecessor directive

```
include 'mpif.h'
```

This file, mpif.h, contains the definitions, macros and function prototypes necessary for compiling an MPI program.

Before any other MPI functions can be called, the function MPI_Init must be called, and it should only be called once. Fortran MPI routines have an IERROR argument - this contains the error code. After a program has finished using MPI library, it must call MPI_Finialize. This cleans up any "unfinished business" left by MPI - e.g. pending receives that were never completed. So a typical MPI program has the following layout.

```
  .
  .
  .
include 'mpif.h'
  .
  .
  .
call MPI_Init(ierr)
  .
  .
  .
call MPI_Finialize(ierr)
  .
  .
  .
end program
```

2.2Finding out About the Rest of the World

MPI provides the function MPI_Comm_rank, which returns the rank of a process in its second in its second argument, Its syntax is

```
CALL MPI_COMM_RANK(COMM, RANK, IERROR)
INTEGER COMM, RANK, IERROR
```

The first argument is a *communicator*. Essentially a communicator is a collection of processes that can send message to each other. For basic programs, the only communicator needed is MPI_COMM_WORLD. It is predefined in MPI and consists of all the processes running when program execution begins.

Many of the constructs in our programs also depend on the number of processes executing the program. So MPI provides the functions MPI_Comm_size for

determining this. Its first argument is a communicator. It returns the number of processes in a communicator in its second argument. Its syntax is

```
CALL MPI_COMM_SIZE(COMM, P, IERROR)
INTEGER COMM, P, IERROR
```

2.3Message : Data + Envelope

The actual message-passing in our program is carried out by the MPI functions MPI_Send and MPI_Recv. The first command sends a message to a designated process. The second receives a message from a process. These are the most basic message-passing commands in MPI. In order for the message to be successfully communicated the system must append some information to the data that the application program wishes to transmit. This additional information forms the *envelope* of the message. In MPI it contains the following information.

1. The rank of the receiver.
2. The rank of the sender.
3. A tag.
4. A communicator.

These items can be used by the receiver to distinguish among incoming messages. The *source* argument can be used to distinguish messages received from different processes. The *tag* is a user-specified integer that can be used to distinguish messages received form a single process. For example, suppose process *A* is sending two messages to process *B*; both messages contains a single real number. One of the real number is to be used in a calculation, while the other is to be printed. In order to determine which is which, *A* can use different tags for the two messages. If *B* uses the same two tags in the corresponding receives, when it receives the messages, it will "know" what to do with them. MPI guarantees that the integers 0-32767 can be used as tags. Most implementations allow much larger values.

As we noted above, a communicator is basically a collection of processes that can send messages to each other. When two processes are communicating using MPI_Send and MPI_Recv, its importance arises when separate modules of a program have been written independently of each other. For example, suppose we wish to solve a system of differential equations, and, in the course of solving the system, we need to solve a system of linear equation. Rather than writing the linear system solver from scratch, we might want to use a *library* of functions for solving linear systems that was written by someone else and that has been highly optimized for the system we're using. How do we avoid confusing the messages *we* send from process *A* to process *B* with those sent by the library functions ? Before the advent of communicators, we

would probably have to partition the set of valid tags, setting aside some of them for exclusive use by the library functions. This is tedious and it will cause problems if we try to run our program on another system : the other system's linear solver may not (probably won't) require the same set of tags. With the advent of communicators, we simply create a communicator that can be used exclusively by the linear solver, and pass it as an argument in calls to the solver. We'll discuss the details of this later. For now, we can get away with using the predefined communicator MPI_COMM_WORLD. It consists of all the processes running the program when execution begins.

2.4MPI_Send and MPI_Recv

To summarize, let's detail the syntax of MPI_Send and MPI_Recv.

```
  MPI_SEND( MESSAGE, COUNT, DATATYPE, DEST, TAG, COMM,
IERROR)
    <TYPE> MESSAGE(*)
    INTEGER COUNT, DATATYPE, DEST, TAG, COMM, IERROR

  MPI_RECV( MESSAGE, COUNT, DATATYPE, SOURCE, TAG,
    COMM, STATUS, IERROR)
    <TYPE> MESSAGE(*)
    INTEGER COUNT, DATATYPE, DEST, TAG, COMM
    INTEGER STATUS(MPI_STATUS_SIZE),IERROR
```

Most MPI functions stores an integer error code in the argument ierror. However, we will ignore these return values in most cases.

The contents of the message are stored in a block of memory referenced by the argument message. The next two arguments, count and datatype, allow the system to identify the end of the message : it contains a sequence of count values, each having *MPI* type datatype. This type is not a Fortran type, although most of the predefined types correspond Fortran types. The predefined MPI types and the corresponding FORTRAN types (if they exist) are listed in the following table.

MPI datatype	FORTRAN datatype
MPI_INTEGER	INTEGER
MPI_REAL	REAL
MPI_DOUBLE_PRECISION	DOUBLE PRECISION
MPI_COMPLEX	COMPLEX
MPI_LOGICAL	LOGICAL

MPI_CHARACTER MPI_BYTE MPI_PACKED	CHARACTER(1)

The last two types, MPI_BYTE and MPI_PACKED, don't correspond to standard Fortran types. The MPI_BYTE type can be used if you wish to force the system to perform no conversion between different data representations (e.g. on a heterogeneous network of workstations using different representations of data). We'll discuss the type MPI_PACKED later.

Note that the amount of space allocated for the receiving buffer does not have to match the exact amount of space in the message being received. For example, when our program is run, the size of the message that process 1 sends, len_trim(message), is 28 characters, but process 0 receives the message in a buffer that has storage for 100 characters. Thsi makes sense. In general, the receiving process may not know the exact size of the message being sent. So MPI allows a message to be received as long as there is sufficient storage allocated. If there isn't sufficient storage, an overflow error occurs [4].

The arguments dest and source are, respectively, the ranks of the receiving and the sending processes. MPI allows source to be "wildcard". There is a predefined constant MPI_ANY_SOURCE that can be used if a process is ready to receive a message from *any* sending process rather than a particular sending process. There is *not* a wildcard for dest.

As we noted erlier, MPI has two mechanisms specifically designed for "partitioning the message space" : tags and communicators. The arguments tag and comm are, respectively, the tag and communicator. The tag is an integer, and for now, our only communicator is MPI_COMM_WORLD, which, as we noted earlier is predefined on all MPI systems and consists of all the processes running when execution of the program begins. There is a wildcard, MPI_ANY_TAG, that MPI_Recv can use for the tag. There is *no* wildcard for the communicator. In other words, in order for process *A* to send a message to process *B*, the argument comm that A uses in MPI_Send must be identical to the argument that *B* uses in MPI_Recv.

The last argument of MPI_Recv, status, returns information on the data that was actually received. It references a array with two elements - one for the source and one for the tags. So if, for example, the source of the receive was

MPI_ANY_SOURCE, then status will contain the rank of the process that sent the message.

3An Application

Now that we know how to send message with MPI, let's write a program that uses message-passing to calculate a definite integral with the trapezoid rule.

3.1Serial program

Recall that the trapezoid rule estimates $\int_a^b f(x)dx$ by dividing the interval [*a,b*] into n segments of equal and calculating the following sum.

$$h[f(x_0)/2 + f(x_n)/2 + \sum_{i=1}^{n-1} f(x_i)].$$

Here, $h = (b - a)/n$, and $x_i = a + ih$, $i = 0,...,n$.

By putting $f(x)$ into a subprogram, we can write a *serial* program for calculating an integral using the trapezoid rule.

```
C serial.f -- calculate definite integral using trapezoidal
C             rule.
C
C The function f(x) is hardwired.
C Input: a, b, n.
C Output: estimate of integral from a to b of f(x)
C    using n trapezoids.

   PROGRAM serial
   IMPLICIT NONE[4]
   real  integral
   real  a
   real  b
```

[4] In Fortran, if you omit to declare a variable it will not normally lead to an error when it is first used; instead it will be implicitly declared to be an integer if the first letter of its name lies in the range I-N, and will be implicitly declared to be a real variable otherwise. *This is extremely dangerous, and must be avoided at all cost.*

Fortunately, Fortran 90 provides the means to avoid this problem by instructing the compiler that all variables *must* be declared before use, and that implicit declaration is not to be allowed. This is achieved by including the statement

```
IMPLICIT NONE
```

as the first statement after the initial PROGRAM, SUBROUTINE or FUNCTION statement.

It is extremely important that this statement appears at the beginning of every program in order that implicit declarations of variables are forbidden. There are a great many stories, some apocryphal and some true, about major catastrophes in Fortran programs that would never have happened had implicit declaration not masked a programming error.

```
  integer  n
  real  h
  real  x
  integer i

  real f
  external f

  print *, 'Enter a, b, and n'
  read *,  a,  b,  n

  h = (b-a)/n
  integral = (f(a) + f(b))/2.0
  x = a
  do i = 1 , n-1
    x = x + h
    integral = integral + f(x)
  enddo
  integral =  integral*h

  print *,'With n =', n,' trapezoids, our estimate'
  print *,'of the integral from ', a, ' to ',b, ' = ' ,
 +integral
  end

C*****************************************************
  real function f(x)
  IMPLICIT NONE
  real x
C  Calculate f(x).

  f = x*x
  return
  end
C*****************************************************
```

3.2Parallelizing the Trapezoid Rule

One approach to parallelizing this program is to simply split the interval $[a,b]$ up among the processes, and each process can estimate the integral of $f(x)$ over its subinterval. In order to estimate the total integral, the processes' local calculations are added.

Suppose there are p processes and n trapezoids, and, in order to simplify the discussion, also suppose that n is evenly divisible by p. Then it is natural for the first

process to calculate the area of the first n/p trapezoids, the second process to calculate the area of the next n/p, etc. So process q will estimate the integral over the interval

$$[a + q\frac{nh}{p}, a + (q+1)\frac{nh}{p}]$$

Thus each process needs the following information.

- The number of processes, p.
- Its rank.
- The entire interval of integration, $[a,b]$.
- The number of subintervals, n.

Recall that the first two items can be found by calling the MPI functions MPI_Comm_size and MPI_Comm_Rank. The latter two items should probably be input by the user. But this can raise some difficult problems. So for our first attempt at calculating the integral, let's "hardwire" these values by simply setting their values with assignment statements.

A straightforward approach to summing the processes' individual calculations is to have each process send its *local* calculation to process 0 and have process 0 do the final addition.

With these assumptions we can write a parallel trapezoid rule program.

```
c  trap.f -- Parallel Trapezoidal Rule, first version
c
c  Input: None.
c  Output:  Estimate of the integral from a to b of f(x)
c     using the trapezoidal rule and n trapezoids.
c
c  Algorithm:
c     1.  Each process calculates "its" interval of
c         integration.
c     2.  Each process estimates the integral of f(x)
c         over its interval using the trapezoidal rule.
c     3a. Each process != 0 sends its integral to 0.
c     3b. Process 0 sums the calculations received from
c         the individual processes and prints the result.
c
c  Note:  f(x), a, b, and n are all hardwired.
c
```

```
  program trapezoidal
c
  IMPLICIT NONE
  include 'mpif.h'
c
  integer   my_rank  ! My process rank.
  integer   p        ! The number of processes.
  real      a        ! Left endpoint.
  real      b        ! Right endpoint.
  integer   n        ! Number of trapezoids.
  real      h        ! Trapezoid base length.
  real      local_a  ! Left endpoint for my process.
  real      local_b  ! Right endpoint my process.
  integer   local_n  ! Number of trapezoids for my
         ! calculation.
  real      integral ! Integral over my interval.
  real      total    ! Total integral.
  integer   source   ! Process sending integal.
  integer   dest     ! All messages go to 0.
  integer   tag
  integer   status(MPI_STATUS_SIZE)
  integer   ierr

  real      Trap

  data a, b, n, dest, tag /0.0, 1.0, 1024, 0, 50/

C  Let the system do what it needs to start up MPI.
  call MPI_INIT(ierr)

C  Get my process rank.
  call MPI_COMM_RANK(MPI_COMM_WORLD, my_rank, ierr)

C  Find out how many processes are being used.
  call MPI_COMM_SIZE(MPI_COMM_WORLD, p, ierr)


  h = (b-a)/n        ! h is the same for all processes.
  local_n = n/p      ! So is the number of trapezoids.

C  Length of each process' interval of integration =
local_n*h.
C  So my interval starts at :
  local_a = a + my_rank*local_n*h
  local_b = local_a + local_n*h
  integral = Trap(local_a, local_b, local_n, h)

C  Add up the integals calculated by each process.
```

```
  if (my_rank .EQ. 0) then
    total = integral
    do source = 1, p-1
      call MPI_RECV(integral, 1, MPI_REAL, source, tag,
  +              MPI_COMM_WORLD, status, ierr)
      total = total + integral
    enddo
  else
    call MPI_SEND(integral, 1, MPI_REAL, dest,
  +          tag, MPI_COMM_WORLD, ierr)
  endif

C  Print the result.
  if (my_rank .EQ. 0) then
    write(6,200) n
 200        format(' ','With n = ',I4,' trapezoids, our
estimate')
    write(6,300) a, b, total
 300        format(' ','of the integral from ',f6.2,' to
',f6.2,
  +             ' = ',f11.5)
  endif

C       Shut down MPI.
  call MPI_FINALIZE(ierr)
  end program trapezoidal


  real function Trap(local_a, local_b, local_n, h)
  IMPLICIT NONE
  real     local_a
  real     local_b
  integer  local_n
  real     h
  real     integral    ! Store result in integal.
  real     x
  real     i

  real     f

  integral = (f(local_a) + f(local_b))/2.0
  x = local_a
  do i = 1, local_n - 1
    x = x + h
          integral = integral + f(x)
  enddo
  integal = integral*h
  Trap = integral
```

```
return
end


real function f(x)
IMPLICIT NONE
real x

f = x*x
end
```

Observe that this program also uses the SPMD paradigm. Even though process 0 executes an essentially different set of commands from the remaining processes, it still runs the same program. The different commands are executed by branching based on the process rank.

3.3I/O on Parallel Processors

One obvious problem with our program is its lack of generality : the data, *a*, *b*, and *n*, are hardwired. The user should be able to enter these values during execution. Let's look more carefully at the problem of I/O on parallel machines.

In our greetings and trapezoid programs we assumed that process 0 could write to standard output (the terminal screen). Most parallel processors provide this much I/O. In fact, most parallel processors allow all processors to both read from standard input and write to standard output. However difficult arise when several processes are simultaneously trying to execute I/O functions. In order to understand this, let's look at an example.

Suppose we modify the trapezoid program so that each process attempts to read the values *a*, *b*, and *n* by adding the statement

```
read *, a , b, n
```

Suppose also that we run the program with two processes and the user types in

```
0 1 1024
```

What happen ? Do both processes get the data ? Does only one ? Or, even worse, does (say) process 0 get the 0 and 1, while process 1 gets the 1024 ? If all the processes get the data, what happens when we write a program, where we want process 0 gets the data, what happens to the others ? Is it even reasonable to have multiple processes reading data from a single terminal ?

On the other hand, what happens if several processes attempt to simultaneously write data to the terminal screen. Does the data from process 0 get printed first, then the data form process 1, etc ? Or does the data appear in some random order ? Or, even worse, does the data from the different processes get all mixed up - say, half a line from 0, two characters from 1, 3 characters from 0, two lines from 2, etc ?

The standard I/O commands available in Fortran (and most other languages) don't provide simple solutions to these problems, and I/O continues to be the subject of considerable research in the parallel processing community. So let's look at some not so simple solutions to these problems.

Thus far, we have assumed that process 0 can at least write to standard output. We will also assume that it can read from standard input. In most cases, we will only assume that process 0 can do I/O. It should be noted that this is a very weak assumption, since, as we noted most parallel machines allow multiple processes to carry out I/O.[5] You might want to ask your local expert whether there are any restrictions on which processes can do I/O.[6]

If only process 0 can do I/O, then we need for process 0 to send the user input to the other processes. This is readily accomplished with a short I/O function that uses MPI_Send and MPI_Recv.

```
C
*********************************************************
******
C  Function Get_data
C    Reads in the user input a, b, and n.
C    Input arguments:
C        1.  integer my_rank:  rank of current process.
C        2.  integer p:  number of processes.
C    Output parameters:
C        1.  real a:  left endpoint a.
C        2.  real b:  right endpoint b.
C        3.  integer n:  number of trapezoids.
C    Algorithm:
C        1.  Process 0 prompts user for input and
C            reads in the values.
C        2.  Process 0 sends input values to other
C            processes.
C
   subroutine Get_data(a, b, n, my_rank, p)
```

[5] The MPI function `MPI_Attr_get` can determine the rank of a process that can carry out the usual I/O functions. See [4]. But it is not important in our SP2.

[6] In our SP2, this is controlled by the environment variable `MP_STDINMODE`. By default, all processes receive the same input data from the keyboard (or standard input). See [8] for detail.

```
      IMPLICIT NONE
      real  a
      real  b
      integer    n
      integer    my_rank
      integer    p
      INCLUDE 'mpif.h'
C
      integer source
      integer dest
      integer tag
      integer   status(MPI_STATUS_SIZE)
      integer   ierr
      data source /0/
C
      if  (my_rank ==[7] 0)  then
       print *, 'Enter a, b and n'
       read *, a, b, n
C
C
        do dest = 1 ,  p-1
                tag = 0
          call MPI_SEND(a, 1, MPI_REAL , dest, tag,
     +                 MPI_COMM_WORLD, ierr )
          tag = 1
          call MPI_SEND(b, 1, MPI_REAL , dest, tag,
     +                 MPI_COMM_WORLD, ierr )
                tag = 2
          call MPI_SEND(n, 1, MPI_INTEGER, dest,
     +                 tag, MPI_COMM_WORLD, ierr )
        enddo
      else
        tag = 0
        call MPI_RECV(a, 1, MPI_REAL , source, tag,
     +          MPI_COMM_WORLD,  status, ierr )
        tag = 1
        call MPI_RECV(b, 1, MPI_REAL , source, tag,
     +          MPI_COMM_WORLD,  status, ierr )
        tag = 2
        call MPI_RECV(n, 1, MPI_INTEGER, source, tag,
     +              MPI_COMM_WORLD,  status, ierr )
      endif
      return
      end
C
C
```

[7] == means exactly the same as `.EQ.`

```
C
*************************************************************
*******
```

4Collective Communication

There are probably a few things in the trapezoid rule program that we can improve on. For example, there is the I/O issue. There are also a couple of problems we haven't discussed yet. Let's look at what happens when the program is run with eight processes.

All the processes begin executing the program (more or less) simulataneously. However, after carrying out the basic set-up tasks (calls to MPI_Init, MPI_Comm_size, and MPI_Comm_rank), processes 1-7 are idle while process 0 collects the input data. We don't want to have idle processes, but in view of our restrictions on which processes can read input data, the higher rank processes must continue to wait while 0 sends the nput data to the lower rank processes. This isn't just an I/O issue. Notice that there is a similar inefficiency at the end of the program, when process 0 does all the work of collecting andd adding the local integrals.

Of course, this is highly undesirable : the main point of parallel processing is to get multiple processes to collaborate on solving a problem. If one of the processes is doing most of the work, we might as well use a conventional, single-processor machine.

4.1Tree-Structured Communication

Let's try to improve our code. We'll begin by focusing on the distribution of the input data. How can we divide the work more evenly among the processes ? A natural solution is to imagine that we have a tree of processes, with 0 at the root.

During the first stage of data distribution, 0 sends the data to (say) 4. During the next stage, 0 sends the data to 2, while 4 sends it to 6. During the last stage, 0 sends to 1, while 2 sends to 3, 4 sends to 5, and 6 sends to 7. (see figure 4.1) So we have reduced our input distribution loop from 7 stages to 3 stages. More generally, if we have p processes, this procedure allows us to distribute the input data in $\lceil \log_2(p) \rceil$ [8]stages, rather than p-1 stages, which, if p is large, is a huge savings.

[8] The notation $\lceil x \rceil$ denotes the smallest whole number grater than or equal to x.

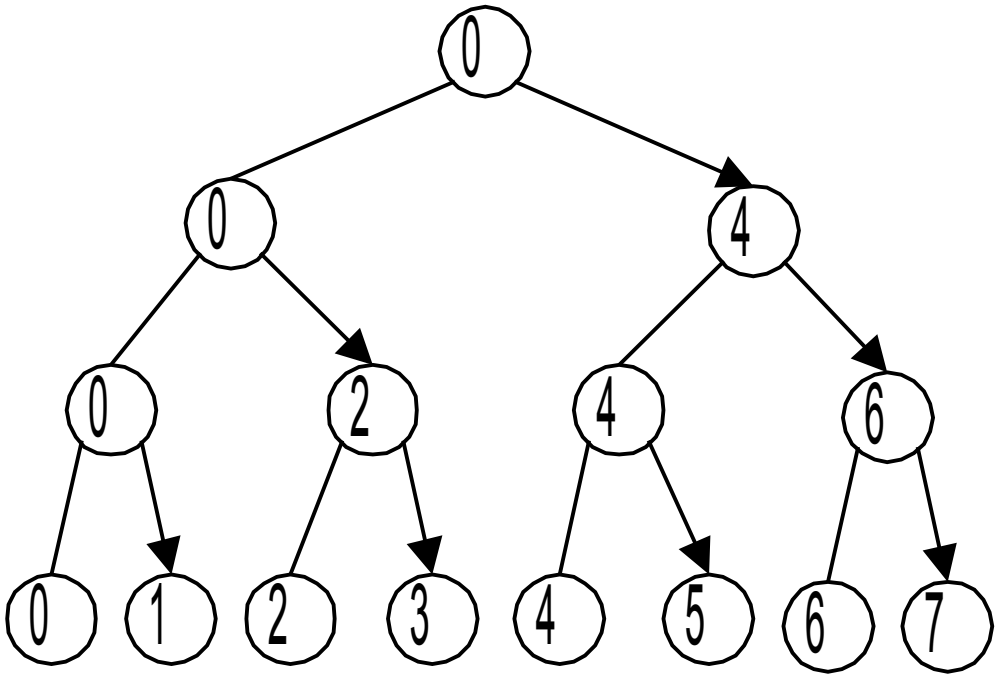


Figure 1 Processors configured as a tree

In order to modify the Get_data function to use a tree-structured distribution scheme, we need to introduce a loop with $\lceil \log_2(p) \rceil$ stages. In order to implement the loop, each process needs to calculate at each stage.

- whether it receives, and, if so, the source ; and
- whether it sends, and, if so, the destination.

As you can probably guess, these calculations can be a bit complicated, especially since there is no canonical choice of ordering. In our example, we chose :

0sends to 4.

0 sends to 2, 4 sends to 6.

0 sends to 1, 2 sends to 3, 4 sends to 5, 6 sends to 7.

We might also have chosen (for example) :

0 sends to 1.

0 sends to 2, 1 sends to 3.

0 sends to 4, 1 sends to 5, 2 sends to 6, 3 sends to 7.

Indeed, unless we know something about the underlying topology of our machine, we can't really decide which scheme is better.

So ideally we would prefer to use a function that has been specifically tailored to the machine we're using so that we won't have to worry about all these tedious details, and we won't have to modify our code every time we change machines. As you may have guess, MPI provides such a function.

4.2Broadcast

A communication pattern that involves all the processes in a communicator is a *collective communication*. As a consequence, a collective communication usually involves more than two processes. A *broadcast* is a collective communication in which a single process sends the same data top every process. In MPI the function for broadcasting data is MPI_Bcast :

```
MPI_BCAST( BUFFER, COUNT, DATATYPE, ROOT, COMM, IERROR )
<type> BUFFER(*)
INTEGER COUNT,  DATA, ROOT, COMM, IERROR
```

It simply sends a copy of the data in BUFFER on process ROOT to each process in the communicator COMM. IT should be called by all the processes in the communicator with the same arguments for ROOT and COMM. Hence a broadcast message cannot be received with MPI_Recv. The parameters COUNT and DATATYPE have the same function that they have in MPI_Send and MP_Recv : they specify the extent of the message. However, unlike the point-to-point functions, MPI insists that in collective communication COUNT and DATATYPE be the same on all the processes in the communicator [4]. The reason for this is that in some collective operations (see below), a single process will receive data from many other processes, and in order for a program to determine how much data has been received, it would need an entire *array* of return statuses.

We can rewrite the Get_data function using MPI_Bcast as follows.

```
subroutine Get_data2(a, b, n, my_rank)
real  a
real  b
integer   n
integer   my_rank
integer ierr
include 'mpif.h'
```

```
C
C
   if (my_rank .EQ. 0) then
    print *, 'Enter a, b, and n'
       read *, a, b, n
   endif
C
  call MPI_BCAST(a, 1, MPI_REAL , 0, MPI_COMM_WORLD, ierr )
  call MPI_BCAST(b, 1, MPI_REAL , 0, MPI_COMM_WORLD, ierr )
  call MPI_BCAST(n, 1, MPI_INTEGER, 0, MPI_COMM_WORLD, ierr
)
  end subroutine Get_data2
C
C
C
*********************************************************
*******
```

Certainly this version of Get_data is much more compact and readily comprehensible than the original, and if MPI_Bcast has been optimized for your system, it will also be a good deal faster.

4.3Reduce

In the trapezoid rule program after the input phase, every processor executes essentially the same commands until the final summation phase. So unless our function f(x) is fairly complicated (i.e., it requires considerably more work to evaluate over certain parts of [a,b]), this part of the program distributes the work equally among the processors. As we have already noted, this is not the case with the final summation phase, when, once again, process 0 gets a disproportionate amount of the work. However, you have probably already noticed that by reversing the arrows in figure 4.1, we can use the same idea we used in section 4.1. That is, we can distribute the work of calculating the sum among the processors as follows.

1. (a) 1 sends to 0, 3 sends to 2, 5 sends to 4, 7 sends to 6.

   (b) 0 adds its integral to that of 1, 2 adds its integral to that of 3, etc.

2. (a) 2 sends to 0, 6 sends to 4.

   (b) 0 adds, 4 adds.

3. (a) 4 sends to 0.

   (b) 0 adds.

Of course, we run into the same question that occurred when we were writing our own broadcast : is this tree structure making optimal use of the topology of our machine ? Once again, we have to answer that this depends on the machine. So, as before, we should let MPI do the work, by using an optimized function.

The "global sum" that we wish to calculate is an example of a general class of collective communication operations called *reduction operations*. In a global reduction operation, all the processes (in a communicator) contribute data which is combined using a binary operation. Typical binary operations are addition, max, min, logical and, etc. The MPI function for performing a reduction operation is

```
MPI_Reduce(OPERAND, RESULT, COUNT, DATATYEP, OP, ROOT,
COMM, IERROR)
<type> OPERAND(*), RESULT(*)
INTEGER COUNT, DATATYPE, OP, ROOT, COMM, IERROR
```

MPI_Reduce combines the operands stored in OPERAND using operation OP and stores the result in RESULT on process ROOT. Both OPERAND and RESULT refer to COUNT memory locations with type DATATYPE. MPI_Reduce must be called by all processes in the communicator COMM, and COUNT, DATATYPE, and OP must be the same on each process.

The argument OP can take on one of the following predefined values.

Operation Name	**Meaning**
MPI_MAX	Maximum
MPI_MIN	Minimum
MPI_SUM	Sum
MPI_PROD	Product
MPI_LAND	Logical And
MPI_BAND	Bitwise And
MPI_LOR	Logical Or
MPI_BOR	Bitwist Or
MPI_LXOR	Logical Exclusive Or
MPI_BXOR	Bitwise Exclusive Or
MPI_MAXLOC	Maximum and Location of Maximum

MPI_MINLOC	Minimum and Location of Minimum

It is also possible to define additional operations. For details see [4].

As an example, let's rewrite the last few lines of the trapezoid rule program.

```
C Add up the integrals calculated by each process.
  MPI_Reduce( INTEGRAL, TOTAL, 1, MPI_REAL, MPI_SUM, 0,
  +            MPI_COMM_WORLD, ierr)
C Print the result.
```

Note that each processor calls MPI_Reduce with the same arguments. In particular, even though total only has significance on process 0, each process must supply an argument.

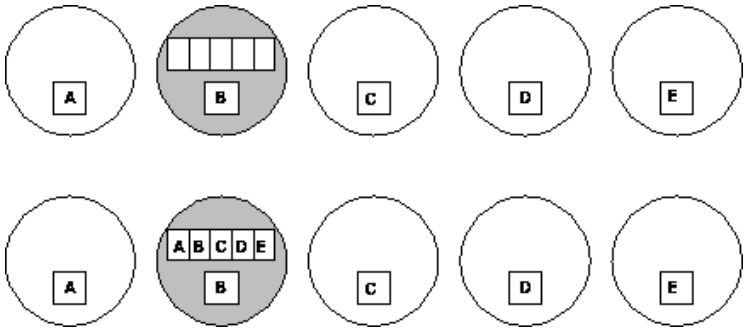
4.4Other Collective Communication Functions

MPI supplies many other collective communication functions. We briefly enumerate some of these here. For full details, see [4].

- `MPI_Barrier( COMM, IERROR)`
  `INTEGER COMM, IERROR`

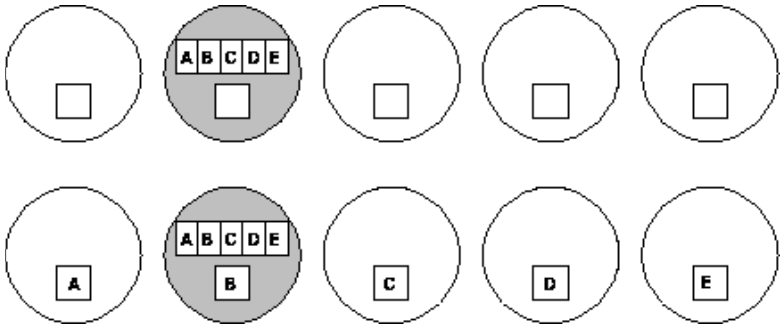
MPI_Barrier provides a mechanism for synchronizing all the processes in the communicator comm. Each process blocks (i.e., pauses) until every process in comm has called MPI_Barrier.

```
MPI_Gather(SEND_BUF, SEND_COUNT, SEND_TYPE, RECV_BUF,
RECV_COUNT, RECV_TYPE, ROOT, COMM, IERROR )
<type>SEND_BUF(*), RECV_BUF(*)
INTEGER SEND_COUNT,SEND_TYPE,RECV_COUNT,
RECV_TYPE,ROOT,COMM,IERROR
```



Each process in comm sends the contents of send_buf to process with rank root. The rpocess root concatenates the received data in process rank order in recv_buf. That is, the data from process 0 is followed by the data from process 1, which is followed by the data from process 2, etc. The recv arguments are significant only on the process with rank root. The argument recv_count indicates the number of items received from each process - not the total number received.

```
MPI_Scatter(SEND_BUF, SEND_COUNT, SEND_TYPE, RECV_BUF,
RECV_COUNT, RECV_TYPE, ROOT, COMM, IERROR )
<type>SEND_BUF(*), RECV_BUF(*)
INTEGER SEND_COUNT,SEND_TYPE,RECV_COUNT,
RECV_TYPE,ROOT,COMM,IERROR
```



The process with rank root distributes the contents of send_buf among the processes. The contents of send_buf are split into p segments each consisting of send_count items. The first segment goes to process 0, the second to process 1, etc. The send arguments are significant only on process root.

```
MPI_Allgather(SEND_BUF, SEND_COUNT, SEND_TYPE, RECV_BUF,
RECV_COUNT, RECV_TYPE, ROOT, COMM, IERROR )
<type>SEND_BUF(*), RECV_BUF(*)
INTEGER SEND_COUNT,SEND_TYPE,RECV_COUNT,
RECV_TYPE,ROOT,COMM,IERROR
```

MPI_Allgather gathers the contents of each send_buf on each process. Its effect is the same as if there were a sequence of p calls to MPI_Gather, each of which has a different process acting as root.

```
MPI_AllReduce(OPERAND, RESULT, COUNT, DATATYPE, OP, COMM,
IERROR )
<type>OPERAND(*), RESULT(*)
INTEGER COUNT, DATATYPE, OP ,COMM,IERROR
```

MPI_Allreduce stores the result of the reduce operation OP in each process' result buffer.

5Grouping Data for Communication

With current generation machines sending a message is an expensive operation. So as a rule of thumb, the fewer messages sent, the better the overall performance of the program. However, in each of our trapezoid rule programs, when we distributed the input data, we sent *a*, *b* and *n* in separate messages - whether we used MPI_Send and MPI_Recv or MPI_Bcast. So we should be able to improve the performance of the program by sending the three input values in a single message. MPI provides three mechanisms for grouping individual data items into a single message : the count parameter to the various communication routines, derived datatypes, and MPI_Pack / MPI_Unpack. We examine each of these options in turn.

5.1The Count Parameter

Recall that MPI_Send, MPI_Recv, MPI_Bcast, and MPI_Reduce all have a count and a datatype argument. These two parameters allow the user to group data items having the same basic type into a single message. In order to use this, the grouped data items must be stored in *contiguous* memory locations. Since Fortran guarantees that array elements are stored in contiguous memory locations, if we wish to send the elements of an array, or a subset of an array, we can do so in a single message. In fact, we've already done this in section 2, when we sent an array of character.

As another example, suppose we wish to send the second half of a vector containing 100 real numbers from process 0 to process 1.

```
   real vector(100)
   integer status(MPI_STATUS_SIZE)
   integer p, my_rank, ierr
   integer i
    .
    .
    .
   if (my_rank == 0) then
C         Initialize vector and send.
     tag = 47
     count = 50
     dest = 1
     call MPI_SEND(vector(51), count, MPI_REAL, dest, tag,
  +              MPI_COMM_WORLD, ierr)
   else
     tag = 47
```

```
    count = 50
    source = 0

    call MPI_RECV(vector(51), count, MPI_REAL, source, tag,
  +                  MPI_COMM_WORLD, status, ierr)

  endif
```

Unfortunately, this doesn't help us with the trapezoid rule program. The data we wish to distribute to the other processes, a, b, and n, are stored in an array. So even if we declared them one after the other in our program,

```
real a
real b
integer n
```

Fortran does *not* guarantee that they are stored in contiguous memory locations. One might be tempted to store n as a float and put the three values in an array, but this would be poor programming style and it wouldn't address the fundamental issue. In order to solve the problem we need to use one of MPI's other facilities for grouping data.

5.2Derived Types and MPI_Type_struct

It might seem that another option would be to store *a*, *b*, and *n* in a derived type[9] with three members - two reals and an integer - and try to use the datatype argument to MPI_Bcast. The difficulty here is that the type of datatype is MPI_Datatype, which is an actual type itself - not the same thing as a user-defined type in Fortran 90. For example, suppose we included the type definition

```
  type INDATA_TYPE
    real a
    real b
    integer n
  end type
```

and the variable definition

```
  type (INDATA_TYPE) indata
```

Now if we call MPI_Bcast

[9] Type definition is available in FORTRAN 90. We may access individual variables in the derived type using the operator %. E.g., indata%a = 1.0

```
call MPI_Bcast(indata, 1, INDATA_TYPE, 0, MPI_COMM_WORLD,
+               ierror)
```

the program will fail. The detail depend on the implementation of MPI that you're using. The problem here is that MPI is a *pre-existing* library of functions. That is, the MPI functions were written without knowledge of the datatypes that you define in your program. In particular, none of the MPI functions "knows" about INDATA_TYPE.

MPI provides a partial solution to this problem, by allowing the user to build MPI datatypes at execution time. In order to build an MPI datatype, one essentially specifies the layout of the data in the type - the member types and their relative locations in memory. Such a type is called a *MPI derived data type*. In order to see how this works, let's write a function that will build a MPI derived type.

```
MODULE GLOBAL[10]
  type INDATA_TYPE
    real a
    real b
    integer n
  end type INDATA_TYPE
END MODULE GLOBAL


subroutine Build_derived_type(indata, mesg_mpi_t)
use GLOBAL[11]
INCLUDE       'mpif.h'
IMPLICIT NONE

type(INDATA_TYPE)   indata
integer      mesg_mpi_t

integer      ierr

integer      block_lengths(3)
integer      displacements(3)
integer      address(4)
integer      typelist(3)

C  Build a derived datatype consisting of two real and an
integer.
```

[10] MODULE is a new kind of Fortran 90 program unit. Any program unit may use the variables and type definition in a module by the "USE" statement. This feature is intended to replace "common block" in Fortran 77 which is very inconvenient.

[11] Supra.

```
C  First specify the types.
   typelist(1) = MPI_REAL
   typelist(2) = MPI_REAL
   typelist(3) = MPI_INTEGER

C  Specify the number of elements of each type.
   block_lengths(1) = 1
   block_lengths(2) = 1
   block_lengths(3) = 1

C  Calculate the displacements of the members relative to
indata.
   call MPI_Address(indata,   address(1), ierr)
   call MPI_Address(indata%a, address(2), ierr)
   call MPI_Address(indata%b, address(3), ierr)
   call MPI_Address(indata%n, address(4), ierr)
   displacements(1) = address(2) - address(1)
   displacements(2) = address(3) - address(1)
   displacements(3) = address(4) - address(1)

C  Build the derived datatype
   call MPI_TYPE_STRUCT(3, block_lengths, displacements,
  +     typelist, mesg_mpi_t, ierr)

C  Commit it -- tell system we'll be using it for
communication.
   call MPI_TYPE_COMMIT(mesg_mpi_t, ierr)
   return
   end
```

The first three statements specify the types of the members of the MPI derived type, and the next three specifies the number of elements of each type. The next four calculate the addresses of the three members of indata. The next three statements use the calculated addresses to determine the *displacements* of the three members relative to the address of the first - which is given displacement 0. With this information, we know the types, sizes and relative locations of the members of a variable having Fortran 90 type INDATA_TYPE, and hence we can define a derived data type that corresponds to the Fortran type. This is done by calling the functions MPI_Type_struct and MPI_Type_commit.

The newly created MPI datatype can be used in any of the MPI communication functions. In order to use it, we simply use the starting address of a variable of type INDATA_TYPE as the first argument, and the derived type in the datatype argument. For example, we could rewrite the Get_data function as follows.

```
subroutine Get_data3(indata, my_rank)
use global
type(INDATA_TYPE) indata
integer     my_rank
integer     mesg_mpi_t
integer     ierr
include     'mpif.h'

if (my_rank == 0) then
  print *, 'Enter a, b, and n'
  read *, indata%a, indata%b, indata%n
endif
call Build_derived_type(indata,  mesg_mpi_t)
call MPI_BCAST(indata, 1, mesg_mpi_t, 0, MPI_COMM_WORLD,
+ierr )
return
end
```

To summarize, then, we can build general MPI derived datatypes by calling MPI_Type_struct. The syntax is

```
call MPI_TYPE_STRUCT(count, array_of_block_lengths,
+array_of_displacements, array_of_types, newtype, ierror)
integer count, array_of_block_lengths(*),
integer array_of_displacements(*) , array_of_types(*)
integer array_of_types(*), newtype, ierror
```

The argument count is the number of elements in the derived type. It is also the size of the three arrays, array_of_block_lengths, array_of_displacements, and array_of_types. The array array_of_block_lengths contains the number of entries in each element of the type. So if an element of the type is an array of *m* values, then the corresponding entry in array_of_block_lengths is *m*. The array array_of_displacements contains the displacement of each element from the beginning of the message, and the array array_of_types contains the MPI datatype of each entry. The argument newtype returns a pointer to the MPI datatype created by the call to MPI_Type_struct.

Note also that newtype and the entries in array_of_types all have type MPI_Datatype. So MPI_Type_struct can be called recursively to build more complex derived datatypes.

5.3Other Derived Datatype Constructors

MPI_Type_struct is the most general datatype constructor in MPI, and as a consequence, the user must provide a *complete* description of each element of the type. If the data to be transmitted consists of a subset of the entries in an array, we shouldn't need to provide such detailed information, since all the elements have the same basic type. MPI provides three derived datatype constructors for dealing with this situation: MPI_Type_Contiguous, MPI_Type_vector and MPI_Type_indexed. The first constructor builds a derived type whose elements are contiguous entries in an array. The second builds a type whose elements are equally spaced entries of an array, and the third builds a type whose elements are arbitrary entries of an array. Note that before any derived type can be used in communication it must be *committed* with a call to MPI_Type_commit.

Details of the syntax of the additional type constructors follow.

```
MPI_Type_contiguous(count, oldtype, newtype, ierror)
integer count, oldtype, newtype, ierror
```

MPI_Type_contiguous creates a derived datatype consisting of count elements of type oldtype. The elements belong to contiguous memory locations.

```
MPI_Type_vector(count, block_length, stride,
+element_type, newtype, ierror)
integer count, blocklength, stride oldtype, newtype,
ierror
```

MPI_Type_vector creates a derived type consisting of count elements. Each element contains block_length entries of type element_type. Stride is the number of elements of type element_type between successive elements of new_type.

```
MPI_Type_indexed(count, array_of_block_lengths,
+array_of_displacements, element_type, newtype, ierror
integer count, array_of_block_lengths(*),
+array_of_displacements, element_type, newtype, ierror
```

MPI_Type_indexed creates a derived type consisting of count elements. The *i*th element (i = 1, ..., count), consists of array_of_block_lengths[i] entries of type element_type, and it is displaced array_of_displacements[i] units of type element_type from the beginning of newtype.

5.4Pack/Unpack

An alternative approach to grouping data is provided by the MPI functions MPI_Pack and MPI_Unpack. MPI_Pack allows one to explicitly store noncontiguous data in contiguous memory locations, and MPI_Unpack can be used to copy data from a contiguous buffer into noncontiguous memory locations. In order to see how they are used, let's rewrite Get_data one last time.

```
   subroutine Get_data4(a, b, n, my_rank)
   real  a
   real  b
   integer    n
   integer    my_rank
C
   INCLUDE   'mpif.h'
   integer   ierr
   character buffer(100)
   integer   position
C      in the buffer
C
   if (my_rank .EQ. 0) then
    print *,'Enter a, b, and n'
    read *, a, b, n
C
C  Now pack the data into buffer.  Position = 0
C  says start at beginning of buffer.
    position = 0
C
C  Position is in/out
    call MPI_PACK(a, 1, MPI_REAL , buffer, 100,
  +         position, MPI_COMM_WORLD, ierr )
C  Position has been incremented: it now refer-
C  ences the first free location in buffer.
C
    call MPI_PACK(b, 1, MPI_REAL , buffer, 100,
  +          position, MPI_COMM_WORLD, ierr )
C  Position has been incremented again.
C
    call MPI_PACK(n, 1, MPI_INTEGER, buffer, 100,
  +          position, MPI_COMM_WORLD, ierr )
C  Position has been incremented again.
C
C  Now broadcast contents of buffer
    call MPI_BCAST(buffer, 100, MPI_PACKED, 0,
  +         MPI_COMM_WORLD, ierr )
   else
    call MPI_BCAST(buffer, 100, MPI_PACKED, 0,
```

```
  +        MPI_COMM_WORLD, ierr )
C
C  Now unpack the contents of buffer
    position = 0
    call MPI_UNPACK(buffer, 100,  position, a, 1,
  +        MPI_REAL , MPI_COMM_WORLD, ierr )
C  Once again position has been incremented:
C  it now references the beginning of b.
C
    call MPI_UNPACK(buffer, 100,  position, b, 1,
  +        MPI_REAL , MPI_COMM_WORLD, ierr )
    call MPI_UNPACK(buffer, 100,  position, n, 1,
  +        MPI_INTEGER, MPI_COMM_WORLD, ierr )
   endif
   return
   end
C
```

In this version of Get_data process 0 uses MPI_Pack to copy a to buffer and then append b and n. After the broadcast of buffer, the remaining processes use MPI_Unpack to successively extract a, b, and n from buffer. Note that the datatype for the calls to MPI_Bcast is MPI_PACKED.

The syntax of MPI_Pack is

```
  MPI_Pack( pack_data, in_count, datatype, buffer, size,
  +position_ptr, comm, ierror)
  <type> pack_data(*), buffer(*)
  integer in_count, datatype, size, position_ptr, comm,
ierror
```

The parameter pack_data references the data to be buffered. It should consist of in_count elements, each having type datatype. The parameter position_ptr is an *in/out* parameter. On input, the data referenced by pack_data is copied into memory starting at address buffer + position_ptr. On return, position_ptr references the first location in buffer *after* the data that was copied. The parameter size contains the size *in bytes* of the memory referenced by buffer, and comm is the communicator that will be using buffer.

The syntax of MPI_Unpack is

```
  MPI_Unpack(buffer, size, position_ptr, unpack_data,
  +count, datatype, comm, ierror)
  <type> inbuf(*), outbuf(*)
  integer insize, position, outcount, datatype, comm,
ierror
```

The parameter buffer references the data to be unpacked. It consists of size bytes. The parameter position_ptr is once again an *in/out* parameter. When MPI_Unpack is called, the data starting at address buffer + position_ptr is copied into the memory referenced by unpack_data. On return, position_ptr references the first location in buffer after the data that was just copied. MPI_Unpack will copy count elements having type datatype into unpack_data. The communicator associated with buffer is comm.

5.5Deciding Which Method to Use

If the data to be sent is stored in consecutive entries of an array, then one should simply use the count and datatype arguments to the communication function(s). This approach involves no additional overhead in the form of calls to derived datatype creation functions or calls to MPI_Pack/MPI_Unpack.

If there are a large number of elements that are not in contiguous memory locations, then building a derived type will probably involve less overhead than a large number of calls to MPI_Pack/MPI_Unpack.

If the data all have the same type and are stored at regular intervals in memory (e.g., a column of a matrix), then it will almost certainly be much easier and faster to use a derived datatype than it will be to use MPI_Pack/MPI_Unpack. Furthermore, if the data all have the same type, but are stored in irregularly spaced locations in memory, it will still probably be easier and more efficient to create a derived type using MPI_Type_indexed. Finally, if the data are heterogeneous and one is repeatedly sending the same collection of data (e.g., row number, column number, matrix entry), then it will be better to use a derived type, since the overhead of creating the derived type is incurred only once, while the overhead of calling MPI_Pack/MPI_Unpack must be incurred every time the data is communicated.

This leaves the case where one is sending heterogeneous data only once, or very few times. In this case, it may be a good idea to collect some information on the cost of derived type creation and packing/unpacking the data. For example, on an nCUBE 2 running the MPICH implementation of MPI, it takes about 12 milliseconds to create the derived type used in Get_data3, while it only takes about 2 milliseconds to pack or unpack the data in Get_data4. Of course, the saving isn't as great as it seems because of the asymmetry in the pack/unpack procedure. That is, while process 0 packs the data, the other processes are idle, and the entire function won't complete until both the pack and unpack are executed. So the cost ratio is probably more like 3:1 than 6:1.

There are also a couple of situations in which the use of MPI_Pack and MPI_Unpack is preferred. Note first that it may be possible to avoid the use of

system buffering with pack, since the data is explicitly stored in a user-defined buffer. The system can exploit this by noting that the message datatype is MPI_PACKED. Also note that the user can send "variable-length"" messages by packing the number of elements at the beginning of the buffer. For example, suppose we want to send rows of a sparse matrix. If we have stored a row as a pair of arrays --- one containing the column subscripts, and one containing the corresponding matrix entries --- we could send a row from process 0 to process 1 as follows.

```
   PROGRAM SpaRow
   INCLUDE 'mpif.h'
   integer  HUGE
   parameter (HUGE = 100)
   integer   p
   integer   my_rank
   real      entries(10)
   integer   column_subscripts(10)
   integer   nonzeroes
   integer   position
   integer   row_number
   character buffer *100
   integer   status(MPI_STATUS_SIZE)
   integer   ierr
   integer   i
   data nonzeroes /10/
C
   call MPI_INIT( ierr)
   call MPI_COMM_SIZE(MPI_COMM_WORLD,  p, ierr )
   call MPI_COMM_RANK(MPI_COMM_WORLD,  my_rank, ierr )
C
   if (my_rank .EQ. 0)  then
C  Get the number of nonzeros in the row.
C  Initialize entries and column_subscripts
    do  i = 1, nonzeroes
      entries(i)           = 2*i
      column_subscripts(i) = 3*i
    enddo
C
C  Now pack the data and send
    position = 1
    call MPI_PACK( nonzeroes, 1, MPI_INTEGER, buffer, HUGE,
  +         position, MPI_COMM_WORLD, ierr )
    call MPI_PACK( row_number, 1, MPI_INTEGER, buffer,
HUGE,
  +          position, MPI_COMM_WORLD, ierr )
    call MPI_PACK(entries, nonzeroes, MPI_REAL , buffer,
  +         HUGE,  position, MPI_COMM_WORLD, ierr )
    call MPI_PACK(column_subscripts,nonzeroes,MPI_INTEGER,
  +         buffer, HUGE,  position, MPI_COMM_WORLD, ierr )
```

```
      call MPI_SEND(buffer, position, MPI_PACKED, 1, 0,
     +         MPI_COMM_WORLD, ierr )
       else
      call MPI_RECV(buffer, HUGE, MPI_PACKED, 0, 0,
     +         MPI_COMM_WORLD,  status, ierr )
      position = 1
      call MPI_UNPACK(buffer, HUGE,  position,  nonzeroes,
     +         1, MPI_INTEGER, MPI_COMM_WORLD, ierr )
      call MPI_UNPACK(buffer, HUGE,  position,  row_number,
     +         1, MPI_INTEGER, MPI_COMM_WORLD, ierr )
      call MPI_UNPACK(buffer,HUGE,  position, entries,
     +         nonzeroes, MPI_REAL , MPI_COMM_WORLD, ierr )
      call MPI_UNPACK(buffer, HUGE,  position,
     +         column_subscripts,
     +         nonzeroes, MPI_INTEGER, MPI_COMM_WORLD, ierr )
      do  i = 1, nonzeroes
        print *, entries(i), column_subscripts(i)
      enddo
     endif
C
     call MPI_FINALIZE(ierr)
     end
```

6 Communicators and Topologies

The use of communicators and topologies makes MPI different from most other message-passing systems. Recollect that, loosely speaking, a communicator is a collection of processes that can send messages to each other. A topology is a structure imposed on the processes in a communicator that allows the processes to be addressed in different ways. In order to illustrate these ideas, we will develop code to implement Fox's algorithm for multiplying two square matrices.

6.1 Fox's Algorithm

We assume that the factor matrices A = (a_{ij}) and B = (b_{ij}) have order n. We also assume that the number of processes, p, is a perfect square, whose square root evenly divides n. Say $p = q^2$, and $\bar{n} = n/q$. In Fox's algorithm the factor matrices are partitioned among the processes in a *block checkerboard* fashion. So we view our processes as a virtual two-dimensional q x q grid, and each process is assigned an $\bar{n}$ x $\bar{n}$ submatrix of each of the factor matrices. More formally, we have a mapping

$$\phi: \{0,1,\ldots,p-1\} \rightarrow \{(s,t): 0 \le s,t \le q-1\}$$

that is both one-to-one and onto. This defines our grid of processes: process i belongs to the row and column given by $\phi(i)$. Further, the process with rank $\phi^{-1}(s,t)$ is assigned the submatrices

$$A_{st} = \begin{pmatrix} a_{s*\bar{n},t*\bar{n}} & \cdots & a_{s*\bar{n},(t+1)*\bar{n}-1} \\ \cdots & & \cdots \\ a_{(s+1)*\bar{n}-1,t*\bar{n}} & \cdots & a_{(s+1)*\bar{n}-1,(t+1)*\bar{n}-1} \end{pmatrix},$$

and

$$B_{st} = \begin{pmatrix} b_{s*\bar{n},t*\bar{n}} & \cdots & b_{s*\bar{n},(t+1)*\bar{n}-1} \\ \cdots & & \cdots \\ b_{(s+1)*\bar{n}-1,t*\bar{n}} & \cdots & b_{(s+1)*\bar{n}-1,(t+1)*\bar{n}-1} \end{pmatrix}.$$

For example, if $p = 9$, $\phi(x) = (x/3, x \bmod 3)$, and $n = 6$, then A would be partitioned as follows.

Process 0	Process 1	Process 2
$A_{00} = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$	$A_{01} = \begin{pmatrix} a_{02} & a_{03} \\ a_{12} & a_{13} \end{pmatrix}$	$A_{02} = \begin{pmatrix} a_{04} & a_{05} \\ a_{14} & a_{15} \end{pmatrix}$
Process 3	Process 4	Process 5
$A_{10} = \begin{pmatrix} a_{20} & a_{21} \\ a_{30} & a_{31} \end{pmatrix}$	$A_{11} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix}$	$A_{12} = \begin{pmatrix} a_{24} & a_{25} \\ a_{34} & a_{35} \end{pmatrix}$
Process 6	Process 7	Process 8
$A_{20} = \begin{pmatrix} a_{40} & a_{41} \\ a_{50} & a_{51} \end{pmatrix}$	$A_{21} = \begin{pmatrix} a_{42} & a_{43} \\ a_{52} & a_{53} \end{pmatrix}$	$A_{22} = \begin{pmatrix} a_{44} & a_{45} \\ a_{54} & a_{55} \end{pmatrix}$

In Fox's algorithm, the block submatrices, A_{rs} and B_{st}, $s = 0, 1, \ldots, q - 1$, are multiplied and accumulated on process $\phi^{-1}(r,t)$. The basic algorithm is:

```
do step = 0, q - 1
        1. Choose a submatrix of A from each row of
              processes.
        2. In each row of processes broadcast the
              submatrix chosen in that row to the other
              processes in that row.
        3. On each process, multiply the newly received
              submatrix of A by the submatrix of B
              currently residing on the process.
        4. On each process, send the submatrix of B to
           the process directly above. (On processes in
           the first row, send the submatrix to the last
           row.)
enddo
```

The submatrix chosen in the rth row is $A_{r,u}$, , where

$$u = (r + \texttt{step}) \bmod q.$$

6.2Communicators

If we try to implement Fox's algorithm, it becomes apparent that our work will be greatly facilitated if we can treat certain subsets of processes as a communication universe --- at least on a temporary basis. For example, in the pseudo-code

```
2. In each row of processes broadcast the submatrix
   chosen in that row to the other processes in that
   row.
```

it would be useful to treat each row of processes as a communication universe, while in the statement

```
4. On each process, send the submatrix of B to the
   process directly above. (On processes in the first
   row, send the submatrix to the last row.)
```

it would be useful to treat each column of processes as a communication universe.

The mechanism that MPI provides for treating a subset of processes as a "communication"' universe is the *communicator*. Up to now, we've been loosely defining a communicator as a collection of processes that can send messages to each other. However, now that we want to construct our own communicators, we will need a more careful discussion.

In MPI, there are two types of communicators: *intra-communicators* and *inter-communicators*.

Intra-communicators are essentially a collection of processes that can send messages to each other *and* engage in collective communication operations. For example, MPI_COMM_WORLD is an intra-communicator, and we would like for each row and each column of processes in Fox's algorithm to form an intra-communicator. Inter-communicators, as the name implies, are used for sending messages between processes belonging to *disjoint* intra-communicators. For example, an inter-communicator would be useful in an environment that allowed one to dynamically create processes: a newly created set of processes that formed an intra-communicator could be linked to the original set of processes (e.g., MPI_COMM_WORLD) by an inter-communicator. We will only discuss intra-communicators. The interested reader is referred to [4] for details on the use of inter-communicators.

A minimal (intra-)communicator is composed of

- a *Group*, and
- a *Context*.

A group is an ordered collection of processes. If a group consists of p processes, each process in the group is assigned a unique *rank*, which is just a nonnegative integer in the range 0, 1, ... , p - 1. A context can be thought of as a system-defined tag that is attached to a group. So two processes that belong to the same group and that use the same context can communicate. This pairing of a group with a context is the most basic form of a communicator. Other data can be associated to a communicator. In

particular, a structure or topology can be imposed on the processes in a communicator, allowing a more natural addressing scheme. We'll discuss topologies in section 6.5.

6.3Working with Groups, Contexts, and Communicators

To illustrate the basics of working with communicators, let's create a communicator whose underlying group consists of the processes in the first row of our virtual grid. Suppose that MPI_COMM_WORLD consists of p processes, where $q^2 = p$. Let's also suppose that $\phi(x) = (x/q, x \bmod q)$. So the first row of processes consists of the processes with ranks 0, 1, ..., q - 1. (Here, the ranks are in MPI_COMM_WORLD.) In order to create the group of our new communicator, we can execute the following code.

```
      PROGRAM ComCrt
      INCLUDE 'mpif.h'
      IMPLICIT NONE
      integer, parameter[12] :: MAX_PROCS = 100
      integer        p
      real           p_real
      integer        q
      integer        my_rank
      integer        MPI_GROUP_WORLD
      integer        first_row_group
      integer        first_row_comm
      integer        process_ranks(0:MAX_PROCS-1)
      integer        proc
      integer        test
      integer        sum
      integer        my_rank_in_first_row
      integer        ierr
C
C
      test = 0
      call MPI_INIT( ierr)
      call MPI_COMM_SIZE(MPI_COMM_WORLD,  p, ierr )
      call MPI_COMM_RANK(MPI_COMM_WORLD,  my_rank, ierr )
C
      p_real = p
      q =  sqrt(p_real)
```

[12] An integer with the attribute "parameter" means it is a integer symbolic constant and cannot be changed at run time.

```
C
C  Make a list of the processes in the new communicator.
   do  proc = 0,  q-1
    process_ranks(proc) = proc
   enddo
C
C  Get the group underlying MPI_COMM_WORLD
   call MPI_COMM_GROUP(MPI_COMM_WORLD, MPI_GROUP_WORLD, ierr
)
C
C  Create the new group
   call MPI_GROUP_INCL(MPI_GROUP_WORLD, q, process_ranks,
  +                   first_row_group, ierr)
C
C  Create the new communicator
   call MPI_COMM_CREATE(MPI_COMM_WORLD, first_row_group,
  +      first_row_comm, ierr)
C
```

This code proceeds in a fairly straightforward fashion to build the new communicator. First it creates a list of the processes to be assigned to the new communicator. Then it creates a group consisting of these processes. This required two commands: first get the group associated with MPI_COMM_WORLD, since this is the group from which the processes in the new group will be taken; then create the group with MPI_Group_incl. Finally, the actual communicator is created with a call to MPI_Comm_create. The call to MPI_Comm_create implicitly associates a context with the new group. The result is the communicator first_row_comm. Now the processes in first_row_comm can perform collective communication operations. For example, process 0 (in first_row_group) can broadcast A_{00} to the other processes in first_row_group.

```
   integer my_rank_in_first_row
   real, allocatable[13] , dimension( :,: )[14] ::[15] A_00

   if (my_rank < q)  then
     call MPI_COMM_RANK(first_row_comm,
  +                     my_rank_in_first_row, ierr)
     !  Allocate space for A_00, order n_bar.
```

[13] Allocatable is an attribute of a variable. It means that the array can be dynamically allocated at run time.

[14] "dimension(:, :)" means the array is two dimensional and the size of the two dimension is unknown at compile time. The array have to be allocated at run time.

[15] :: is a separator between variable attributes and variable names.

```
  allocate (A_00(n_bar,n_bar))[16]
  if (my_rank_in_first_row == 0) then
    ! initialize A_00
  endif
  call MPI_BCAST( A_00, n_bar*n_bar, MPI_INTEGER, 0,
+                    first_row_comm, ierr)
endif
```

Groups and communicators are *opaque objects*. From a practical standpoint, this means that the details of their internal representation depend on the particular implementation of MPI, and, as a consequence, they cannot be directly accessed by the user. Rather the user accesses a *handle* that references the opaque object, and the opaque objects are manipulated by special MPI functions, for example, MPI_Comm_create, MPI_Group_incl, and MPI_Comm_group.

Contexts are not explicitly used in any MPI functions. Rather they are implicitly associated with groups when communicators are created. The syntax of the commands we used to create first_row_comm is fairly self-explanatory. The first command

```
MPI_Comm_group(comm, group, ierror)
integer comm, group, ierror
```

simply returns the group underlying the communicator comm.

The second command

```
MPI_Group_incl(old_group, new_group_size,
+ranks_in_old_group, new_group, ierror)
integer old_group, new_group_size,
integer ranks_in_old_group(*), new_group, ierror
```

creates a new group from a list of processes in the existing group old_group. The number of processes in the new group is new_group_size, and the processes to be included are listed in ranks_in_old_group. Process 0 in new_group has rank ranks_in_old_group(0) in old_group, process 1 in new_group has rank ranks_in_old_group(1) in old_group, etc.

The final command

```
MPI_Comm_create(old_comm, new_group, new_comm, ierror)
integer old_comm, new_group, new_comm, ierror
```

[16] The function "Allocate " allocates memory space for an allocatable array.

associates a context with the group new_group and creates the communicator new_comm. All of the processes in new_group belong to the group underlying old_comm.

There is an extremely important distinction between the first two functions and the third. MPI_Comm_group and MPI_Group_incl, are both *local* operations. That is, there is *no* communication among processes involved in their execution. However, MPI_Comm_create *is* a collective operation. *All* the processes in old_comm must call MPI_Comm_create with the same arguments. The *Standard* [4] gives three reasons for this:

1. It allows the implementation to layer MPI_Comm_create on top of regular collective communications.

2. It provides additional safety.

3. It permits implementations to avoid communication related to context creation.

Note that since MPI_Comm_create is collective, it will behave, in terms of the data transmitted, as if it synchronizes. In particular, if several communicators are being created, they must be created in the same order on all the processes.

6.4MPI_Comm_split

In our matrix multiplication program we need to create multiple communicators --- one for each row of processes and one for each column. This would be an extremely tedious process if p were large and we had to create each communicator using the three functions discussed in the previous section. Fortunately, MPI provides a function, MPI_Comm_split that can create several communicators simultaneously. As an example of its use, we'll create one communicator for each row of processes.

```
   integer my_row_comm
   integer my_row

C  my_rank is rank in MPI_COMM_WORLD.
C  q*q = p
   my_row = my_rank/q
   call MPI_COMM_SPLIT(MPI_COMM_WORLD, my_row, my_rank,
  +      my_row_comm, ierr)
```

The single call to MPI_Comm_split creates q new communicators, all of them having the same name, my_row_comm. For example, if $p = 9$, the group underlying my_row_comm will consist of the processes 0, 1, and 2 on processes 0,

1, and 2. On processes 3, 4, and 5, the group underlying my_row_comm will consist of the processes 3, 4, and 5, and on processes 6, 7, and 8 it will consist of processes 6, 7, and 8.

The syntax of MPI_Comm_split is

```
MPI_COMM_SPLIT(old_comm, split_key, rank_key,
+      new_comm, ierror)
integer old_comm, split_key, rank_key, new_comm, ierror
```

It creates a new communicator for each value of split_key. Processes with the same value of split_key form a new group. The rank in the new group is determined by the value of rank_key. If process *A* and process *B* call MPI_Comm_split with the same value of split_key, and the rank_key argument passed by process *A* is less than that passed by process *B*, then the rank of *A* in the group underlying new_comm will be less than the rank of process *B*. If they call the function with the same value of rank_key, the system will arbitrarily assign one of the processes a lower rank.

MPI_Comm_split is a collective call, and it must be called by all the processes in old_comm. The function can be used even if the user doesn't wish to assign every process to a new communicator. This can be accomplished by passing the predefined constant MPI_UNDEFINED as the split_key argument. Processes doing this will have the predefined value MPI_COMM_NULL returned in new_comm.

6.5Topologies

Recollect that it is possible to associate additional information --- information beyond the group and context --- with a communicator. This additional information is said to be *cached* with the communicator, and one of the most important pieces of information that can be cached with a communicator is a topology. In MPI, a *topology* is just a mechanism for associating different addressing schemes with the processes belonging to a group. Note that MPI topologies are *virtual* topologies --- there may be no simple relation between the process structure defined by a virtual topology, and the actual underlying physical structure of the parallel machine.

There are essentially two types of virtual topologies that can be created in MPI --- a *cartesian* or *grid* topology and a *graph* topology. Conceptually, the former is subsumed by the latter. However, because of the importance of grids in applications, there is a separate collection of functions in MPI whose purpose is the manipulation of virtual grids.

In Fox's algorithm we wish to identify the processes in MPI_COMM_WORLD with the coordinates of a square grid, and each row and each column of the grid needs to form its own communicator. Let's look at one method for building this structure.

We begin by associating a square grid structure with MPI_COMM_WORLD. In order to do this we need to specify the following information.

1. The number of dimensions in the grid. We have 2.

2. The size of each dimension. In our case, this is just the number of rows and the number of columns. We have q rows and q columns.

3. The periodicity of each dimension. In our case, this information specifies whether the first entry in each row or column is "adjacent" to the last entry in that row or column, respectively. Since we want a "circular"' shift of the submatrices in each column, we want the second dimension to be periodic. It's unimportant whether the first dimension is periodic.

4. Finally, MPI gives the user the option of allowing the system to optimize the mapping of the grid of processes to the underlying physical processors by possibly reordering the processes in the group underlying the communicator. Since we don't need to preserve the ordering of the processes in MPI_COMM_WORLD, we should allow the system to reorder.

Having made all these decisions, we simply execute the following code.

```
integer       grid_comm
integer       dim_sizes(0:1)
logical       wrap_around(0:1)
logical       reorder[17] = .TRUE.

dim_sizes(0) = q
dim_sizes(1) = q
wrap_around(0) = .TRUE.
wrap_around(1) = .TRUE.
call MPI_CART_CREATE(MPI_COMM_WORLD, 2, dim_sizes,
+      wrap_around, reorder,  grid_comm, ierr)
```

After executing this code, the communicator grid_comm will contain all the processes in MPI_COMM_WORLD (possibly reordered), and it will have a two-dimensional cartesian coordinate system associated. In order for a process to determine its coordinates, it simply calls the function MPI_Cart_coords :

```
integer   coordinates(0:1)
integer   my_grid_rank

call MPI_COMM_RANK(grid_comm,  my_grid_rank, ierr)
call MPI_CART_COORDS(grid_comm, my_grid_rank, 2,
```

[17] This syntax initialize the variable.

```
+          coordinates, ierr)
```

Notice that we needed to call MPI_Comm_rank in order to get the process rank in grid_comm. This was necessary because in our call to MPI_Cart_create we set the reorder flag to .TRUE. , and hence the original process ranking in MPI_COMM_WORLD may have been changed in grid_comm.

The "inverse" to MPI_Cart_coords is MPI_Cart_rank.

```
call MPI_CART_RANK(grid_comm, coordinates, grid_rank,
+          ierr)
integer grid_comm, coordinates(*), grid_rank, ierr
```

Given the coordinates of a process, MPI_Cart_rank returns the rank of the process in its third parameter process_rank.

The syntax of MPI_Cart_create is

```
call MPI_CART_CREATE(old_comm, number_of_dims, dim_sizes,
+      periods, reorder,  cart_comm, ierror)
integer old_comm, number_of_dims, dim_sizes(*)
logical periods(*), reorder
integer cart_comm, ierror
```

MPI_Cart_create creates a new communicator, cart_comm by caching a cartesian topology with old_comm. Information on the structure of the cartesian topology is contained in the parameters number_of_dims, dim_sizes, and periods. The first of these, number_of_dims, contains the number of dimensions in the cartesian coordinate system. The next two, dim_sizes and periods, are arrays with order equal to number_of_dims. The array dim_sizes specifies the order of each dimension, and periods specifies whether each dimension is circular or linear.

The processes in cart_comm are ranked in *row-major* order. That is, the first row consists of processes 0, 1, ... , dim_sizes(0)-1, the second row consists of processes dim_sizes(0), dim_sizes(0)+1, ... , 2*dim_sizes(0)-1, etc. Thus it may be advantageous to change the relative ranking of the processes in old_comm. For example, suppose the *physical* topology is a 3 x 3 grid, and the processes (numbers) in old_comm are assigned to the processors (grid squares) as follows.

3	4	5
0	1	2
6	7	8

Clearly, the performance of Fox's algorithm would be improved if we re-numbered the processes. However, since the user doesn't know what the exact mapping of processes to processors is, we must let the system do it by setting the reorder parameter to .TRUE. .

Since MPI_Cart_create constructs a new communicator, it is a collective operation.

The syntax of the address information functions is

```
  MPI_Cart_rank(comm, coordinates, rank, ierror)
  integer comm, coordinates(*), rank, ierror

  MPI_Cart_coords(comm, rank, number_of_dims, coordinates,
 +  ierror)
  integer comm, rank, number_of_dims, coordinates(*),
ierror
```

MPI_Cart_rank returns the rank in the cartesian communicator comm of the process with cartesian coordinates coordinates. So coordinates is an array with order equal to the number of dimensions in the cartesian topology associated with comm. MPI_Cart_coords is the inverse to MPI_Cart_rank: it returns the coordinates of the process with rank rank in the cartesian communicator comm. Note that both of these functions are local.

6.6MPI_Cart_sub

We can also partition a grid into grids of lower dimension. For example, we can create a communicator for each row of the grid as follows.

```
  logical varying_coords(0:1)
  integer row_comm

  varying_coords(0) = .FALSE.
  varying_coords(1) = .TRUE.
  call MPI_CART_SUB(grid_comm, varying_coords, row_comm,
ierr)
```

The call to MPI_Cart_sub creates q new communicators. The varying_coords argument is an array of boolean. It specifies whether each dimension "belongs" to the new communicator. Since we're creating communicators for the rows of the grid, each new communicator consists of the processes obtained by fixing the row coordinate and letting the column coordinate *vary*. Hence we assigned varying_coords(0) the value .FALSE. --- the first coordinate doesn't vary --- and we assigned varying_coords(1) the value .TRUE. --- the second coordinate varies. On each

process, the new communicator is returned in row_comm. In order to create the communicators for the columns, we simply reverse the assignments to the entries in varying_coords.

```
  integer col_comm

  varying_coords(0) = .TRUE.
  varying_coords(1) = .FALSE.
  call MPI_CART_SUB(grid_comm, varying_coords, row_comm,
ierr)
```

Note the similarity of MPI_Cart_sub to MPI_Comm_split. They perform similar functions --- they both partition a communicator into a collection of new communicators. However, MPI_Cart_sub can only be used with a communicator that has an associated cartesian topology, and the new communicators can only be created by fixing (or varying) one or more dimensions of the old communicators. Also note that MPI_Cart_sub is, like MPI_Comm_split, a collective operation.

6.7Implementation of Fox's Algorithm

To complete our discussion, let's write the code to implement Fox's algorithm. First, we'll write a function that creates the various communicators and associated information. Since this requires a large number of variables, and we'll be using this information in other functions, we'll put it into a Fortran 90 derived type to facilitate passing it among the various functions.

Notice that since each of our communicators has an associated topology, we constructed them using the topology construction functions --- MPI_Cart_create and MPI_Cart_sub --- rather than the more general communicator construction functions MPI_Comm_create and MPI_Comm_split.

```
  program myfox
  include 'mpif.h'
  IMPLICIT NONE
  type GRID_INFO_TYPE
    integer p        ! Total number of processes.
    integer comm     ! Communicator for the entire grid.
    integer row_comm ! Communicator for my row.
    integer col_comm ! Communicator for my col.
    integer q        ! Order of grid.
    integer my_row   ! My row number.
    integer my_col   ! My column number.
    integer my_rank  ! My rank in the grid communicator.
  end type GRID_INFO_TYPE
```

```
TYPE (GRID_INFO_TYPE) :: grid_info
integer my_rank, ierr
real, allocatable, dimension(:,:) :: A,B,C
integer n, n_bar

call MPI_INIT(ierr)
call Setup_grid(grid_info)

call MPI_Comm_rank(MPI_COMM_WORLD, my_rank, ierr)
if (my_rank == 0)  then
 print *, 'What is the order of the matrices?'
 read *, n
endif

call MPI_BCAST(n,1,MPI_INTEGER, 0, MPI_COMM_WORLD,ierr)
n_bar = n/(grid_info%q)
!  Allocate local storage for local matrix.
allocate( A(n_bar,n_bar) )
allocate( B(n_bar,n_bar) )
allocate( C(n_bar,n_bar) )

A = 1.0  [18]
B = 2.0
call Fox(n,grid_info,A,B,C,n_bar)
print *, C

contains[19]
subroutine Setup_grid(grid)
 TYPE (GRID_INFO_TYPE), intent(inout)[20] :: grid

 integer old_rank
 integer dimensions(0:1)
 logical periods(0:1)
 integer coordinates(0:1)
 logical varying_coords(0:1)
 integer ierr

 !  Set up Global Grid Information.
 call MPI_Comm_size(MPI_COMM_WORLD, grid%p,   ierr)
 call MPI_Comm_rank(MPI_COMM_WORLD, old_rank, ierr )
```

[18] This is Fortran 90 array syntax. It assigns every element of array A to be 1.0 .

[19] In Fortran 90, it is permissible to include a procedure as an integral part of a program unit. The program unit can invoke the internal procedure. This inclusion is done by the statement "contains".

[20] intent(inout) is an attribute of dummy argument. It informs the compiler that this dummy argument may be used for read and write inside the subroutine.

```
  grid%q = int(sqrt(dble(grid%p)))
  dimensions(0) = grid%q
  dimensions(1) = grid%q
  periods(0) = .TRUE.
  periods(1) = .TRUE.
  call MPI_Cart_create(MPI_COMM_WORLD, 2,
+         dimensions, periods, .TRUE. , grid%comm, ierr)
  call MPI_Comm_rank  (grid%comm, grid%my_rank, ierr )
  call MPI_Cart_coords(grid%comm, grid%my_rank, 2,
+         coordinates, ierr )
  grid%my_row = coordinates(0)
  grid%my_col = coordinates(1)

  !  Set up row and column communicators.
  varying_coords(0) = .FALSE.
  varying_coords(1) = .TRUE.
  call MPI_Cart_sub(grid%comm,varying_coords,
+         grid%row_comm,ierr)
  varying_coords(0) = .TRUE.
  varying_coords(1) = .FALSE.
  call MPI_Cart_sub(grid%comm,varying_coords,
+         grid%col_comm,ierr)
 end subroutine Setup_grid

 subroutine Fox(n,grid,local_A,local_B,loca l_C,n_bar)
  integer, intent(in)               :: n, n_bar
  TYPE(GRID_INFO_TYPE), intent(in)  :: grid
  real, intent(in) , dimension(:,:) :: local_A, local_B
  real, intent(out), dimension (:,:) :: local_C

  real temp_A( SIZE(A,DIM=1),SIZE(A,DIM=2) )
  integer step, source, dest, request
  integer status(MPI_STATUS_SIZE), bcast_root

  local_C = 0.0
  source = mod( (grid%my_row + 1), grid%q )
  dest   = mod( (grid%my_row + grid%q -1), (grid%q) )
  temp_A = 0.0

  do step = 0, grid%q -1
    bcast_root = mod( (grid%my_row + step), (grid%q) )
    if (bcast_root == grid%my_col)  then
      call MPI_BCAST(local_A,n_bar*n_bar,MPI_REAL,
+                 bcast_root, grid%row_comm, ierr)
      call sgemm('N','N',n_bar,n_bar,n_bar,1.0,
+ local_A,n_bar,local_B,n_bar,1.0,local_C,n_bar)
    else
      call MPI_BCAST(temp_A,n_bar*n_bar,MPI_REAL,
```

```
  +                 bcast_root, grid%row_comm, ierr)
       call sgemm('N','N',n_bar,n_bar,n_bar,1.0,
  +
temp_A,n_bar,local_B,n_bar,1.0,local_C,n_bar)
      endif
    call MPI_Send(local_B,n_bar*n_bar,MPI_REAL,dest,  0,
  +        grid%col_comm, ierr)
    call MPI_Recv(local_B,n_bar*n_bar,MPI_REAL,source,0,
  +        grid%col_comm, status, ierr )

    enddo
  end subroutine Fox

  end program myfox
```

7Where To Go From Here

7.1What We Haven't Discussed

MPI is a large library. The Standard [4] is over 200 pages long and it defines more than 125 functions. As a consequence, this *Guide* has covered only a small fraction of MPI, and many readers will fail to find a discussion of functions that they would find very useful in their applications. So we briefly list some of the more important ideas in MPI that we have not discussed here.

1. **Communication Modes.** We have used only the *standard* communication mode for send. This means that it is up to the system to decide whether the message is buffered. MPI provides three other communication modes: *buffered, synchronous*, and *ready*. In buffered mode, the user explicitly controls the buffering of outgoing messages. In synchronous mode, a send will not complete until a matching receive is posted. In ready mode, a send may be started only if a matching receive has already been posted. MPI provides three additional send functions for these modes.

2. **Nonblocking Communication.** We have used only *blocking* sends and receives (MPI_Send and MPI_Recv.) For the send, this means that the call won't return until the message data and envelope have been buffered or sent --- i.e., until the memory referenced in the call to MPI_Send is available for re-use. For the receive, this means that the call won't return until the data has been received into the memory referenced in the call to MPI_Recv. Many applications can improve their performance by using *nonblocking* communication. This means that the calls to send/receive may return before the operation completes. For example, if the machine has a separate communication processor, a non-blocking send could simply notify the communication processor that it should begin composing and sending the message. MPI provides nonblocking sends in each of the four modes and a nonblocking receive. It also provides various utility functions for determining the completion status of a non-blocking operation.

3. **Inter-communicators.** Recollect that MPI provides two types of communicators: intra-communicators and inter-communicators. Inter-communicators can be used for point-to-point communications between processes belonging to distinct intra-communicators.

There are many other functions available to users of MPI. If we haven't discussed a facility you need, please consult the *Standard* [4] to determine whether it is part of MPI.

7.2Implementations of MPI

If you don't have an implementation of MPI, there are three versions that are freely available by anonymous ftp from the following sites.

- Argonne National Lab/Mississippi State University. The address is info.mcs.anl.gov, and the directory is pub/mpi.
- Edinburgh University. The address is ftp.epcc.ed.ac.uk, and the directory is pub/chimp/release.
- Ohio Supercomputer Center. The address is tbag.osc.edu, and the directory is pub/lam.

All of these run on networks of UNIX workstations. The Argonne/Mississippi State and Edinburgh versions also run on various parallel processors. Check the "README" files to see if your machine(s) are supported.

7.3More Information on MPI

There is an MPI FAQ available by anonymous ftp at

- Mississippi State University. The address is ftp.erc.msstate.edu, and the file is pub/mpi/faq.

There are also numerous web pages devoted to MPI. A few of these are

- http://www.epm.ornl.gov/~walker/mpi. The Oak Ridge National Lab MPI web page.
- http://www.erc.msstate.edu/mpi. The Mississippi State MPI web page.
- http://www.mcs.anl.gov/mpi. The Argonne MPI web page.

Each of these sites contains a wealth of information about MPI. Of particular note, the Mississippi State page contains a bibliography of papers on MPI, and the Argonne page contains a collection of test MPI programs.

The *MPI Standard* [4] is currently available from each of the sites above. This is, of course, the definitive statement of what MPI is. So if you're not clear on something, this is the final arbiter. It also contains a large number of nice examples of uses of the various MPI functions. So it is considerably more than just a reference. Currently, several members of the MPI Forum are working on an annotated version of the MPI standard [5].

The book [2] is a tutorial introduction to MPI. It provides numerous complete examples of MPI programs.

The book [6] contains a tutorial introduction to MPI (on which this guide is based). It also contains a more general introduction to parallel processing and the programming of message-passing machines.

The Usenet newsgroup, comp.parallel.mpi, provides information on updates to all of these documents and software.

7.4The Future of MPI

As it is currently defined, MPI fails to specify two critical concepts: I/O and the creation/destruction of processes. Work has already been started on the development of both I/O facilities and dynamic process creation. Information on the former can be obtained from http://lovelace.nas.nasa.gov/MPI-IO/mpi-io.html, and information on the latter can be found on the Argonne MPI web page. Significant developments are invariably posted to comp.parallel.mpi.

8 Compiling and Running MPI Programs

This section is intended to give the outline of how to compile and run a program in the IBM SP2.

MPI program written in Fortran or Fortran 90 can be compiled using the following command :

```
mpif77 program.f
```

By default, the program will be running on 4 processors of the SP2. The program can be invoked by the name of executable

```
a.out
```

The number of processes is controled by the environment variable MP_PROCS. The web page http://www.hku.hk/cc/sp2/technical/setenv.html has manuals for setting the environment variable.

9 Reference

[1] Geoffrey Fox, et al., *Solving Problems on Concurrent Processors*, Englewood Cliffs, NJ, Prentice--Hall, 1988.

[2] William Gropp, Ewing Lusk, and Anthony Skjellum, *Using MPI: Portable Parallel Programming with the Message-Passing Interface*, Cambridge, MA, MIT Press, 1994.

[3] Brian W. Kernighan and Dennis M. Ritchie, *The C Programming Language*, 2nd ed., Englewood Cliffs, NJ, Prentice--Hall, 1988.

[4] Message Passing Interface Forum, *MPI: A Message-Passing Interface Standard*, International Journal of Supercomputer Applications, vol 8, nos 3/4, 1994. Also available as Technical Report CS-94-230, Computer Science Dept., University of Tennessee, Knoxville, TN, 1994.

[5] Steve Otto, et al., *MPI Annotated Reference Manual*, Cambridge, MA, MIT Press, to appear.

[6] Peter S. Pacheco, *Programming Parallel with MPI*, San Francisco, CA, Morgan Kaufmann. 1997.

[7] Peter S. Pacheco, *A User's Guide to MPI*, 1995.