D R A F T

Document for a Standard Message-Passing Interface

Message Passing Interface Forum

May 10, 1993

This work was supported by ARPA and NSF under contract number ###, by the National Science Foundation Science and Technology Center Cooperative Agreement No. CCR-8809615, and by the Commission of the European Community through Esprit project P6643.

Contents

Abstract

The Message Passing Interface Forum (MPIF), with participation from over 40 organizations, has been meeting since January 1993 to discuss and define a set of library interface standards for message passing. MPIF is not sanctioned or supported by any official standards organization.

This is a draft of what will become the Final Report, Version 1.0, of the Message Passing Interface Forum. This document contains all the technical features proposed for the interface. This copy of the draft was processed by LaTeX on May 10, 1993.

MPIF invites comments on the technical content of MPI, as well as on the editorial presentation in the document. Comments received before July 1, 1993 will be considered in producing the final draft of Version 1.0 of the Message Passing Interface Specification.

The goal of the Message Passing Interface, simply stated, is to develop a widely used standard for writing message-passing programs. As such the interface should establishing a practical, portable, efficient, and flexible standard for message passing.

Acknowledgments

The technical development was carried out by subgroups, whose work was reviewed by the full committee. During the period of development of the Message Passing Interface (MPI), many people served in positions of responsibility:

- Jack Dongarra, David Walker, Conveners and Meeting Chairs
- Ewing Lusk, Bob Knighten, Minutes
- Marc Snir, William Gropp, Ewing Lusk, Point to Point Communications
- Al Geist, Marc Snir, Steve Otto, Collective Communications
- Rolf Hempel, Process Topologies
- Scott Berryman, Language Binding
- Steve Zenith, Formal Semantics
- William Gropp, Environmental Inquiry
- James Cownie, Profiling
- Tony Skjellum, Lyndon Clarke, Marc Snir, Richard Littlefield, Marc Sears, Communication Contexts
- Steve Otto, Editor.

Do we want an organization/company list here?

The following list includes some of the active participants in the MPI process not mentioned above:

Ed Anderson	Anne Elster	Jim Feeney
Jon Flower	Daniel Frye	Ian Glendinning
Adam Greenberg	Robert Harrison	Leslie Hart
Tom Haupt	Don Heller	Tom Henderson
C. T. Howard Ho	Steven Huss-Lederman	John Kapenga
Bob Leary	Barney Maccabe	Peter Madams
Oliver McBryan	Phil McKinley	Chuck Mosher
Dan Nessett	Paul Pierce	Sanjay Ranka
Peter Rigsbee	Ambuj Singh	Robert G. Voigt
Dennis Weeks	Stephen Wheat	

abmacca@cs.sandia.gov
amm@think.com
als@cs.umd.edu
barszcz@orville.nas.nasa.gov
bjornson@cs.yale.edu
bob@lanl.gov
ccm@arco.com
danielf@kgnvmt.vnet.ibm.com
eca@cray.com
fernando@cs.berkeley.edu
fineberg@nas.nasa.gov
hart@fsl.noaa.gov
haupt@npac.syr.edu
hep@ncube.com
jbaron@vnet.ibm.com
john@cs.wmich.edu
leary@sdsc.edu
mckinley@cps.msu.edu
peter@sun.math.usfca.edu
robison@shell.com

The following organizations made the draft available by anonymous FTP access and/or mail servers: Oak Ridge National Laboratory, These outlets were instrumental in distributing the document.

Talk about financial support here.

Chapter 1

Introduction to MPI

1.1 Overview and Goals

Message passing is a paradigm used widely on certain classes of parallel machines; especially those with distributed memory. Although there are many variations, the basic concept of processes communicating through messages is well understood. Over the last ten years, substantial progress has been made in casting significant applications in this paradigm. Each vendor has implemented their own variant. More recently, several systems have demonstrated that a message passing system can be efficiently and be portably implemented. It is thus an appropriate time to try to define both the syntax and semantics of a core of library routines that will be useful to a wide range of users and efficiently implementable on a wide range of computers.

In designing MPI we have sought to make use of the most attractive features of a number of existing message passing systems, rather than selecting one of them and adopting it as the standard. Thus, MPI has been strongly influenced by work at the IBM T. J. Watson Research Center [1, 2], Intel's NX/2 [16], Express [15], nCUBE's Vertex [13], P4 [3], and PARMACS [11, 12]. Other important contributions have come from Zipcode [17, 18], Chimp [7, 8], PVM [9, 19], and PICL [10].

One of the objectives of this paper is to promote a discussion within the concurrent computing research community of the issues that must be addressed in establishing a practical, portable, and flexible standard for message passing. This cooperative process began with a workshop on standards for message passing held in April 1992 [20].

The main advantages of establishing a message passing standard are portability and ease-of-use. In a distributed memory communication environment in which the higher level routines and/or abstractions are build upon lower level message passing routines the benefits of standardization are particularly apparent. Furthermore, the definition of a message passing standard, such as that proposed here, provides vendors with a clearly defined base set of routines that they can implement efficiently, or in some cases provide hardware support for, thereby enhancing scalability.

The goal of the Message Passing Interface simply stated is to develop a widely used standard for writing message-passing programs. As such the interface should establishing a practical, portable, efficient, and flexible standard for message passing.

A complete list of goals follow.

- Design an application programming interface (not necessarily for compilers or a system implementation library).

- Allow efficient communication: Avoid memory to memory copying and allow overlap of computation and communication and offload to communication coprocessor, where available.
- Allow (but no mandate) extensions for use in heterogeneous environment.
- Allow convenient C, Fortran 77, Fortran 90, and C++ bindings for interface.
- Assume a reliable communication interface: User need not cope with communication failures. Such failures are dealt by the underlying communication subsystem.
- Focus on a proposal that can be agreed upon in 6 months.
- Define an interface that is not too different from current practice, such as PVM, Express, P4, etc.
- Define an interface that can be quickly implemented on many vendor's platforms, with no significant changes in the underlying communication and system software.
- The interface should not contain more functions than are really necessary.
- Semantics of the interface should be language independent.

1.2 Who Should Use This Standard?

This standard is intended for use by all those who want to write portable message-passing programs in Fortran 77, C, Fortran 90, or C++. This includes individual application programmers, developers of software designed to run on parallel machines, and creators of environments, and tools. In order to be attractive to this wide audience, the standard must provide a simple, easy-to-use interface for the basic user while not semantically precluding the high-performance message-passing operations available on advanced machines.

1.3 What Platforms Are Targets For Implementation?

The attractiveness of the message-passing paradigm at least partially stems from its wide portability. Programs expressed this way may run on distributed-memory multiprocessors, networks of workstations, and combinations of all of these. In addition, shared-memory implementations are possible. The paradigm will not be made obsolete by architectures combining the shared- and distributed-memory views, or by increases in network speeds. It thus should be both possible and useful to implement this standard on a great variety of machines, including those "machines" consisting of collections of other machines, parallel or not, connected by a communication network.

The interface is suitable for use by fully general MIMD programs, as well as those written in the more restricted style of SPMD. Although no explicit support for threads is provided, the interface has been designed so as not to prejudice their use.

1.4 What Is Included In The Standard?

The standard includes (this is temporarily as inclusive as possible):

- Point-to-point communication
- Collective operations
- Process groups
- Communication contexts
- A simple way to create processes for the SPMD model
- Bindings for Fortran 77, Fortran 90, C and C++
- A model implementation
- A formal specification.
- Process topology
- A validation suite

1.5 What Is Not Included In The Standard?

The standard does not specify:

- Explicit shared-memory operations
- Operations that require more operating system support than is currently standard; for example, interrupt-driven receives, remote execution, or active messages
- Program construction tools
- Debugging facilities
- Auxiliary functions such as timers
- Explicit support for threads

There are many features that have been considered and not included in this standard. This happened for a number of reasons, one of which is the time constraint that was self imposed in finishing the standard. Features that are not included can always be offered as extensions by specific implementations.

Chapter 2

Point to Point Communication

Marc Snir
William Gropp and Ewing Lusk

2.1 Introduction

This section is a draft of the current proposal for point-to-point communication. It does not yet include a description of the Fortran 77 and C bindings.

I have tried to indicate, wherever appropriate, gaps and unresolved issues, using small type.

Discussion: The following subsections of the introduction contain general information on the design of MPI procedures. The material should be moved to a general introduction for the entire document.

2.2 Data Types

2.2.1 Handle

MPI procedures use at various places *handles*. Handles are used to access opaque objects. Such object can be created, updated and destroyed only by by calling suitable MPI procedures, and providing the handle as parameter. Opaque objects hide from the user the internal representation used for various MPI objects, thus allowing to have similar calls in C and Fortran, allowing to overcome problems with the typing rules in these languages, and allowing for future extension of their functionality. Handles are of type `void *` in C and of type `integer` in Fortran.

An opaque object can be `persistent` or `ephemeral`. A persistent object persists until destroyed by an explicit operation. An ephemeral object is good for a single use; thus an ephemeral object associated with a communication operation disappears once this operation is completed (or once this object is not needed anymore for the completion of the operation).

An opaque object is created by a call to `MPI_CREATE`, and destroyed by a call to `MPI_FREE`. Additional MPI functions are available to access and update specific opaque objects.

MPI_CREATE(handle, type, persistence)

OUT handle handle to object

IN type state value that identifies the type of object to be created (e.g., `MPI_COMMUNICATION`, `MPI_BUFFER`, `MPI_CONTEXT`, etc.).

IN persistence state value; either `MPI_PERSISTENT` or `MPI_EPHEMERAL`.

MPI_FREE(handle)

IN handle handle to object

An object can be destroyed only if there is no pending operation that is using this object; after successful return of the routine, the handle is undefined.

MPI_ASSOCIATED(handle, type)

IN handle handle to object

OUT type state

Returns the type of the object the handle is currently associated with, if such exists. Returns the special type `MPI_NULL` if the handle is not currently associated with any object.

MPI may provide predefined opaque objects and predefined, static handles to these objects. Such objects may not be destroyed.

List of handles An MPI call may need a parameter that is a *list of handles*. In C, such list will be a record with one component being the length of the list, the other components being an array of pointers. In Fortran, the list will be an array of integers, the first one of which is the length of the list.

Discussion: The mechanism for opaque objects used here follows the POSIX Fortran binding standard. An alternative choice is to have different type declarations for each type of opaque object. Then, opaque objects are created/destroyed like regular variables, rather than by MPI calls; they are still accessed and updated only via MPI functions.

2.2.2 State

MPI procedures use at various places arguments with *state* types. The values of such data type are all identified by names, and no operation is defined on them. For example, the `MPI_CREATE` routine has a state type parameter with values `MPI_PERSISTENT` and `MPI_EPHEMERAL`.

An `enumeration` declared in an included MPI.h file will be used in C for state datatypes. The Fortran 77 mechanism needs to be decided.

Discussion: Named integer constants can be used in Fortran 90, using the `PARAMETER` mechanism. The constant declarations can be made available via an `INCLUDE` file. Fortran 77 does not seem to offer any convenient mechanism. One possibility is to specify explicit integer values, and allow the use of named constants with those Fortran 77 compilers that support them conveniently. Another possibility is to use character strings, rather than integers.

2.2.3 Named constants

MPI procedures sometimes assign a special meaning to a special value of a basic type parameter; e.g. `tag` is an integer valued parameter of point-to-point communication operations, with a special `DONTCARE` value. Such parameters will have a range of regular values, which is a proper subrange of the range of values of the corresponding basic type; special values (such as DONTCARE) will be outside the regular range. The range of regular values can be queried, and sometimes set, using environment inquiry or environment setting functions (Section 7). The special values are provided by named constant, that are made available via an MPI.h include file in a C binding.

Discussion:
Need to agree on a Fortran mechanism for named constants (see the discussion above).
Implementers should try to detect illegal uses of "special values". Thus, the use of the `DONTCARE` value to tag a message sent should be flagged as an error.

2.2.4 Choice

MPI functions sometimes use parameters with a *choice* (or union) data type. I.e., distinct calls to the same routine may pass by reference actual parameters of different types. The mechanism for providing such parameters will differ from language to language. In C, a formal parameter of type `void *` will be used, with an actual pointer parameter. in Fortran, we shall cheat.

Discussion:
The Fortran 77 standard specifies that the type of actual arguments need to agree with the type of dummy arguments; no construct equivalent to C pointers is available. Thus, it would seem that there is no standard conforming mechanism to support choice parameters. However, most Fortran compiler either don't check type consistency of calls to external routines, or support a special mechanism to link foreign (e.g., C) routines. I suggest that we accept this nonconformity with Fortran 77 standard. I.e., we accept that the same routine may be passed an actual parameter of a different type at distinct calls.
Generic routines can be used in Fortran 90 to provide a standard conforming solution. This solution will be consistent with our nonstandard conforming Fortran 77 solution.

2.3 Processes

An MPI program is executed by several autonomous processes that execute each their own code, in a MIMD style. The codes executed by each process need not be identical. The processes communicate via calls to MPI communication primitives. Typically, each processor executes in its own address space, although shared-memory implementations of MPI are possible. This document specifies the behavior of a parallel program assuming that only MPI calls are used for communication. The interaction of an MPI program with other possible means of communication (e.g., shared memory) is not specified. In particular, it is assumed that message buffers at distinct processors are disjoint.

MPI does not specify the execution model for each process. A process can be sequential, or can be multithreaded, with threads possibly executing concurrently. Care has been taken to make MPI "thread-safe", by avoiding the use of implicit global states.

The initial allocation of processes to an MPI computation and their binding to physical processors is not specified by the program itself. It is expected that vendors will provide mechanisms to do so either at load time or at run time. Such mechanisms will allow to specify the initial number of required processes, the code to be executed by each initial process, and the allocation of processes to processors. Also, the current proposal does not provide for dynamic creation or deletion of processes during program execution, although it is intended to be consistent with such extension. Finally, the current proposal does not specify a naming scheme for processes. We propose to always identify processes according to their relative rank in a context (group), so that, effectively, processes are identified by consecutive integers. Absolute, system-wide unique process id's are (will be) needed only if dynamic process creation is to be supported (in such eventuality we propose to use handles to opaque *process structures* for that purpose).

2.4 Contexts

Discussion: This section contains a proposal for use of contexts that will subsume groups. It borrows heavily on the current group proposal. This proposal has not yet been discussed in MPI meetings.

A **context** consists of:

- A set of processes that currently belong to the context (possibly all processes, or a proper subset).
- A **ranking** of the processes within that context, i.e., a numbering of the processes in that context from 0 to $n - 1$, where n is the number of processes in that context.

A process may belong to several contexts at the same time.

Any interprocess communication occurs within a context, and messages sent within one context can be received only within the same context. A context is specified using a *context handle* (i.e., a handle to an opaque object that identifies a context). Context handles cannot be transferred for one process to another; they can be used only on the process where they where created.

Follows examples of possible uses for contexts.

Loosely synchronous library call interface Consider the case where a parallel application executes a "parallel call" to a library routine, i.e., where all processes transfer control to the library routine. If the library was developed separately, then one should beware of the possibility that the library code may receive by mistake messages send by the caller code, and vice-versa. To prevent such occurrence one might use a barrier synchronization before and after the parallel library call. Instead, one can allocate a different context to the library, thus preventing unwanted interference. Now, the transfer of control to the library need not be synchronized.

Functional decomposition and modular code development Often, a parallel application is developed by integrating several distinct functional modules, that is each developed separately. Each module is a parallel program that runs on a dedicated set of processes, and the computation consists of phases where modules compute separately, intermixed with global phases where all processes communicate. It is convenient to allow each module to use its own private process numbering scheme, for the intramodule computation. This is achieved by using a private module context for intramodule computation, and a global context for intermodule communication.

Collective communication MPI supports collective communication within dynamically created groups of processes. Each such group can be represented by a distinct context. This provides a simple mechanism to ensure that communication that pertains to collective communication within one group is not confused with collective communication within another group.

Lightweight gang scheduling Consider an environment where processes are multithreaded. Contexts can be used to provide a mechanism whereby all processes are time-shared between several parallel executions, and can context switch from one parallel execution to another, in a loosely synchronous manner. A thread is allocated on each process to each parallel execution, and a different context is used to identify each parallel execution. Thus, traffic from one execution cannot be confused with traffic from another execution. The blocking and unblocking of threads due to communication events provide a "lazy" context switching mechanism. This can be extended to the case where the parallel executions are spanning distinct process subsets. (MPI does not require multithreaded processes.)

Discussion: A context handle might be implemented as a pointer to a structure that consists of context label (that is carried by messages sent within this context) and a context member table, that translates process ranks within a context to absolute addresses or to routing information. Of course, other implementations are possible, including implementations that do not require each context member to store a full list of the context members.

Contexts can be used only on the process where they were created. Since the context carries information on the group of processes that belong to this context, a process can send a message within a context only to other processes that belong to that context. Thus, each process needs to keep track only of the contexts that where created at that process; the total number of contexts per process is likely to be small.

The only difference I see between this current definition of context, which subsumes the group concept, and a pared down definition, if that I assume here that process numbering is relative to the context, rather then being global, thus requiring a context member table. I argue that this is not much added overhead, and gives much additional needed functionality.

- If a new context is created by copying a previous context, then one does not need a new member table; rather, one needs just a new context label and a new pointer to the same old context member table. This holds true, in particular, for contexts that include all processes.
- A context member table makes sure that a message is sent only to a process that can execute in the context of the message. The alternative mechanism, which is checking at reception, is

less efficient, and requires that each context label be system-wide unique. This requires that, to the least, all processes in a context execute a collective agreement algorithm at the creation of this context.

- The use of relative addressing within each context is needed to support true modular development of subcomputations that execute on a subset of the processes. There is also a big advantage in using the same context construct for collective communications as well.

2.4.1 Context Operations

A global context **ALL** is predefined. All processes belong to this context when computation starts. MPI does not specify how processes are initially ranked within the context ALL. It is expected that the start-up procedure used to initiate an MPI program (at load-time or run-time) will provide information or control on this initial ranking (e.g., by specifying that processes are ranked according to their pid's, or according to the physical addresses of the executing processors, or according to a numbering scheme specified at load time).

Discussion: If we think of adding new processes at run-time, then `ALL` conveys the wrong impression, since it is just the initial set of processes.

The following operations are available for creating new contexts.

MPI_COPY_CONTEXT(newcontext, context)

Create a new context that includes all processes in the old context. The rank of the processes in the previous context is preserved. The call must be executed by all processes in the old context. It is a blocking call: No call returns until all processes have called the function. The parameters are

OUT newcontext handle to newly created context. The handle should not be associated with an object before the call.

IN context handle to old context

Discussion: I considered adding a string parameter, to provide a unique identifier to the next context. But, in an environment where processes are single threaded, this is not much help: Either all processes agree on the order they create new contexts, or the application deadlocks. A key may help in an environment where processes are multithreaded, to distinguish call from distinct threads of the same process; but it might be simpler to use a mutex algorithm at each process.

Implementation note: No communication is needed to create a new context, beyond a barrier synchronization; all processes can agree to use the same naming scheme for successive copies of the same context. Also, no new rank table is needed, just a new context label and a new pointer to the same old table.

MPI_NEW_CONTEXT(newcontext, context, key, index)

OUT newcontext handle to newly created context at calling process. This handle should not be associated with an object before the call.

IN context handle to old context

IN key integer

IN index integer

A new context is created for each distinct value of `key`; this context is shared by all processes that made the call with this key value. Within each new context the processes are ranked according to the order of the `index` values they provided; in case of ties, processes are ranked according to their rank in the old context.

This call is blocking: No call returns until all processes in the old context executed the call.

Particular uses of this function are:

(i) Reordering processes: All processes provide the same `key` value, and provide their index in the new order.

(ii) Splitting a context into subcontexts, while preserving the old relative order among processes: All processes provide the same `index` value, and provide a key identifying their new subcontext.

MPI_RANK(rank, context)

OUT rank integer

IN context context handle

Return the rank of the calling process within the specified context.

MPI_SIZE(size, context)

OUT size integer

IN context context handle

Return the number of processes that belong to the specified context.

Usage note Use of contexts for libraries: Each library may provide an initialization routine that is to be called by all processes, and that generate a context for the use of that library.

Use of contexts for functional decomposition: A harness program, running in the context `ALL` generates a subcontext for each module and then starts the submodule within the corresponding context.

Use of contexts for collective communication: A context is created for each group of processes where collective communication is to occur.

Use of contexts for context-switching among several parallel executions: A preamble code is used to generate a different context for each execution; this preamble code needs to use a mutual exclusion protocol to make sure each thread claims the right context.

Discussion: If process handles are made explicit in MPI, then an additional function needed is **MPI_PROCESS(process, context, rank)**, which returns a handle to the process identified by the `rank` and `context` parameters.

A possible addition is a function of the form **MPI_CREATE_CONTEXT(newcontext, list_of_process_handles)** which creates a new context out of an explicit list of members (and rank them in their order of occurrence in the list). This, coupled with a mechanism for requiring the spawning of new processes to the computation, will allow to create a new all inclusive context that includes the additional processes. However, I oppose the idea of requiring dynamic process creation as part of MPI. Many implementers want to run MPI in an environment where processes are statically allocated at load-time.

2.4.2 Error Handling

It is assumed that MPI is implemented on top of an error-free communication subsystem: A message sent is always received correctly, and the user does not need to check for transmission errors, time-outs, and the likes. In other words, MPI does not provide mechanisms to deal with failures in the underlying communication subsystem – it is the responsibility of the MPI implementer to insulate the user from such errors (or to reflect them as global program failures). The same holds true for node failures.

Of course, MPI programs may still be erroneous. A **program error** can occur when an MPI call is called with an incorrect parameter (non-existing destination in a send operation, buffer too small in a receive operation, etc.) This type of error would occur in any implementation. In addition, a **resource error** may occur when a program exceeds the amount of available system resources (number of pending messages, system buffers, etc.). The occurrence of this type of error depends on the amount of available resources in the system and the resource allocation mechanism used; this may differ from system to system. The recommended implementation profile provides several mechanisms to alleviate the portability problem this represents. One can also write **safe** programs, that are not subject to resource errors.

All MPI procedure calls return an error parameter that indicates successful completion of the operation, or the error condition that occurred, otherwise.

The recommended implementation profile in a POSIX environment is for any MPI routine that encounters a recoverable error to store an error number in a global variable (*errno* in a C environment) and generate an *MPI error signal*, using a special signal value. The default handler for this signal terminates the execution of all involved processes, with a suitable error message being returned to the user. However, the user can provide his or her own signal handling routine. In particular, the user can specify a "noop" signal handler, thus relegating all error handling to the user code, using the error parameters returned by the MPI calls.

MPI calls may initiate operations that continue asynchronously after the call returned. Thus, the operation may return with a code indicating successful completion, yet later cause an error exception to be raised. If there is a subsequent call that relates to the same operation (e.g., a call that verifies that an asynchronous operation has completed) then the error parameter associated with this call will be used to indicate the nature of the error. In a few cases, the error may occur after all calls that relate to the operation have completed, so that no error parameter can be used to indicate the nature of the error (e.g., an error in a send with the ready mode). In such cases, an error will be undetected, if the user disabled the MPI error signal.

Discussion: The alternative choice is to have fatal and non-fatal signals.

2.5 Messages

A message consists of an *envelope* and *data*.

2.5.1 Data

The data part of a message consists of a sequence of values, each of a basic datatype in the host language. Thus, in Fortran, a message consists of a sequence of values that are each of type `INTEGER`, `REAL`, `DOUBLE PRECISION`, `COMPLEX`, `LOGICAL`, or (length 1) `CHARACTER`. A message may mix values of different types.

Discussion: May also need `DOUBLE COMPLEX` in Fortran.

2.5.2 Envelope

The following information is associated with each message:

source The rank the sending process

destination The rank of the receiving process

tag User defined

context handle

The range of valid values for the **source** and **destination** fields is `0 ... n-1`, where `n` is the number of processes in the current context. The ranges of valid values for `tag` is implementation dependent, and can be found by calling a suitable query function, as described in Section 7. `Context` should be a context shared by both source and destination.

The `tag` field can be arbitrarily set by the application, and can be used to distinguish different messages.

The actual mechanism used to associate an envelope with a message is implementation dependent; some of the information (e.g., **sender** or **receiver**) may be implicit, and need not be explicitly carried by a message.

2.6 Data Buffers

The basic point to point communication operations are **send** and **receive**. A **send** operation creates a message; the message data is assembled from the **send buffer**. A **receive** operation consumes a message; the message data is moved into the **receive buffer**. The specification of send or receive buffers uses the same syntax.

A buffer consists of a sequence **buffer components**. Each component consists of a sequence variables of the same basic type. There are three basic types of buffer components:

block A sequence of contiguous values of the same basic type, specified by

- **start** Initial element
- **len** Number of elements (`len` $\geq$ `0`)
- **datatype** Type of elements

vector A sequence of equally spaced and equally sized blocks of elements of the same basic type, specified by

- **start** Initial element
- **len** Number of elements (`len` $\geq$ `0`)
- **stride** Number of elements between the start of each block
- **lenblk** Number of elements in each block (`lenblk` $\leq$ `stride`)
- **datatype** Type of elements

Note that a constant stride becomes contiguous when `stride = lenblk`. A vector buffer component can be an arbitrary submatrix of a two-dimensional matrix.

indexed A sequence of elements of the same basic type, specified by

- **start** initial element
- **list_of_indices** List of displacements of the elements in the buffer components, relative to the initial element.
- **datatype** Type of elements

For example, if a Fortran array is declared as `double precision a(10)`, then the tuple <`(a(3), (0,3,6,2), DOUBLE`> specifies a buffer component with entries `a(3) a(6), a(9), a(5)`.

> **Discussion:** Do we allow entries to be repeated in an indexed buffer component?
> Do we allow different buffer components to overlap?
> Do we require in an vector buffer component that `len` be a multiple of `lenblk`?

A buffer is described by an opaque object accessed via a **buffer handle**. Such object is created and destroyed via calls to `MPI_CREATE` and `MPI_FREE`. It is associated with successive buffer components by calling in succession one of the functions `MPI_ADD_BLOCK`, `MPI_ADD_VECTOR` or `MPI_ADD_INDEX`, in order to append a component to the buffer associated with a buffer handle. A buffer object can be destroyed only if there is no pending communication operation using it. After a buffer object is destroyed the associated buffer handle is undefined.

MPI_ADD_BLOCK(buffer, start, len, datatype)

Append a block component to buffer. The parameters are:

INOUT buffer buffer handle

IN start buffer component initial element (choice type)

IN len Number of elements (integer)

IN datatype datatype identifier

MPI_ADD_VEC(buffer, len, stride, lenblk, datatype)

Append a vector buffer component to buffer. The parameters are:

INOUT buffer buffer handle

IN start buffer component initial element (choice type)

IN len Number of elements (integer)

IN stride Number of elements between the start of each block (integer)

IN lenblk Number of elements in each block (integer)

IN datatype datatype identifier

MPI_ADD_INDEX(buffer, start, list_of_indices)

Append an indexed buffer component to buffer. The parameters are:

INOUT buffer buffer handle

start initial position for indexing (choice type)

list_of_indices list of relative indices of entries (array of integers)

Consider, for example, the following fragment of Fortran code

```
DOUBLE PRECISION A(10,20)
INTEGER B, C(5,10)
INTEGER BH
...
CALL MPI_CREATE(BH, MPI_BUFFER, MPI_PERSISTENT)
CALL MPI_ADD_BLOCK (BH, B, 1, MPI_INT)
CALL MPI_ADD_VEC (BH, A(1,3), 11, 4, 2, MPI_DOUBLE)
CALL MPI_ADD_INDEX(BH, C(3,7), (4,2,1), MPI_INT)
```

Then the buffer associated with the handle `BH` consists of the sequence of variables `B, A(1,3), A(2,3), A(5,3), A(6,3), A(9,3), A(10,3), A(3,4), A(4,4), A(7,4), A(8,4), A(1,5), C(2,8), C(5,7), C(4,7)`.

A message created from this buffer will consist of a sequence of one integer, followed by eleven double precision reals, followed by three integers.

A buffer handle can be used for communication, even if it is not associated with any variables (i.e., even if it was set by any `MPI_ADD_xxx` call). Such handle is associated with an empty buffer, and a message created from it contains no data.

Discussion:

The main modifications w.r.t. the proposal of Gropp and Lusk is measuring length in elements, rather than bytes. Seem more natural, since type is known. Also, object creation uses a generic function, rather than a function specific to buffer descriptors. As result, I gave up on the size parameter. This may not be such a loss: it is not clear that specifying a maximum length at buffer object creation is useful, since indices of arbitrary size may need to be stored in the object.

2.6.1 Data Conversion

The types and the locations of the entries used to create a message is solely determined from the information in the buffer descriptor, using the storage association rules specified by the host language and its implementation; the type and the locations of these entries do not depend on the declaration for the corresponding variables in the calling program. It is not required that the data types specified in a buffer descriptor match the data types of the corresponding elements in the host program. However, in case of mismatches, the correspondence between entries in the host program and entries in a message created with the buffer descriptor may be implementation dependent. No data conversion occurs when data is moved from a sender buffer into a message.

Consider the following fragment of Fortran code

```
REAL A(100)
INTEGER BH
...
CALL MPI_CREATE(BH, MPI_BUFFER, MPI_PERSISTENT)
CALL MPI_ADD_BLOCK (BH, A(1), 1, MPI_REAL)
CALL MPI_ADD_BLOCK (BH, A(2), 1, MPI_INT)
CALL MPI_ADD_BLOCK (BH, A(3), 1, MPI_LOGICAL)
CALL MPI_ADD_BLOCK (BH, A(4), 1, MPI_COMPLEX)
CALL MPI_ADD_BLOCK (BH, A(6), 1, MPI_DOUBLE)
CALL MPI_ADD_BLOCK(BH, A(8), 4, MPI_CHAR)
```

A message created from this buffer will consist of a sequence containing one real, one integer, one logical, one complex, one double, and four characters. No data conversion occurs when values are copied from the sender buffer to the message. Thus, the first entry in the message is a real number equal to `A(1)`; the second entry is an integer number that happens to have the same binary representation as `A(2)`; the third entry is a logical value that happens to have the same binary representation as `A(3)`; the fourth entry in a complex number with a binary representation identical to `A(4), A(5)`; the fifth entry is a double precision value with a binary representation identical to `A(6), A(7)`; and if words have four bytes then the last four entries are bytes that make the binary representation of `A(8)`, in the byte order of the executing machine.

The correspondence between the first seven entries of the array `A` and the first five entries of the message created from this buffer is determined by the rules of Fortran 77 on storage association: Each variable of type `INTEGER`, `REAL`, or `LOGICAL` occupy one *numeric storage unit*; a variable of type `DOUBLE` or `COMPLEX` occupy two numeric storage units. Thus, the same correspondence will hold for any implementation. However, different implementations may have different binary encodings of integer, real and logical values, so that the actual values transferred by the message may differ.

The correspondence between the entry `A(8)` of the array, and the last four character entries in the message is implementation dependent, since the Fortran language does not specify a correspondence between character storage units and numeric storage units (an array of characters cannot be equivalenced with an array of integers). Different results may occur in big-endians or small-endians machines, or in 32 bit or 64 bit machines.

The same holds, symmetrically, when a message is received. Entries are moved from the message into the receiver memory according to the information provided by the buffer descriptor, with no regard to the way the corresponding variables are declared in the receiving program.

When data is moved in a homogeneous environment between nodes having the same architecture, then no data conversion occur at any point during data transfer. Assume, in the previous example, that an identically declared buffer descriptor is used to receive data into an identically declared array at the receiving process. Then, when these two nodes communicate, the values in the first eight entries of array `A` of the sender will be copied into the first eight entries of array `A` at the receiver (assuming that reals occupy the same storage as four bytes). In particular, in a homogeneous environment, it is possible to communicate using only character typed messages.

When data is moved into a heterogeneous environment between nodes having distinct architectures, data conversion may occur during the transfer: Each entry is converted from the data representation used on the sending node to the data representation used in the receiving node.

Consider the following fragment of Fortran code.

```
REAL X, Y
CHARACTER (4) Z
INTEGER BH
...
CALL MPI_CREATE(BH, MPI_BUFFER, MPI_PERSISTENT)
CALL MPI_ADD_BLOCK (BH, X, 1, MPI_REAL)
CALL MPI_ADD_BLOCK (BH, Y, 4, MPI_CHAR)
CALL MPI_ADD_BLOCK (BH, Z, 4, MPI_CHAR)
```

Assume that the same arrays `X, Y, Z` and buffer handle `BH` are created at two processes A and B; this handle is used by A to create and send a message for B, by B to receive this message. Further assume that both A and B run on distinct nodes, with possibly different 32 bit architectures. Then `X` at process B is assigned the value of `X` at process A (up to rounding errors that may occur during conversion); `Z` on process B is assigned the character string value of `Z` on process A; if both nodes use ASCII encoding, then no conversion is required for the characters. On the other hand, variable `Y` on process B may be allocated a value that differs from the value of variable `Y` on process A. This may occur if the two nodes use a different byte sequence (little-endian vs big-endian), or use a different binary representation for reals (IEEE vs HEX). Thus, in order to ensure correct execution in a heterogeneous environment, it is important that the types of values in a message match the types of the corresponding values in the sending and in the receiving program.

2.7 Receive Criteria

The selection of a message by a receive operation is done uniquely according to the value of the message envelope. The receive operation specifies an **envelope pattern**; a message can be received by that receive operation only if its envelope matches that pattern. A pattern specifies values for the `source`, `tag` and `context` fields of the message envelope. In addition, the value for the `dest` field is set, implicitly, to be equal to the receiving process id. The receiver may specify a **DONTCARE** value for `source`, or `tag`, indicating that any source and/or tag are acceptable. It cannot specify a DONTCARE value for `context` or `dest`. Thus, a message can be received by a receive operation only if it is addressed to the receiving task, has a matching context, has matching source unless source=DONTCARE in the pattern, and has a matching tag unless tag=DONTCARE in the pattern.

The length of the received message must be less or equal the length of the receive buffer. I.e., all incoming data must fit, without truncation, into the receive buffer. It is erroneous to receive a message which length exceed the receive buffer, and the outcome of program where this occurs is undetermined.

2.8 Communication Mode

A sending operation can occur in one of two modes:

REGULAR The send may start whether or not a matching receive has been posted.

READY The send may start only if a matching receive has been posted.

A **ready send** can start only if a matching receive is already posted; otherwise the operation is erroneous and its outcome is undefined. In some systems, this will allow to optimize communication and avoid a hand-shaking operation that is otherwise required.

Discussion: I deleted the symmetric ready receive. Will revive it if there is a requirement for it.

2.9 Communication Handles

An opaque communication object identifies various properties of a communication operation, such as the buffer descriptor that is associated with it, its context, the tag and destination parameters to be used for a send, or the tag and source parameters to be used for a receive. In addition, this object stores information about the status of the last communication operation that was performed with this object. This object is accessed using a communication handle.

One can consider communication operations to consist of the following suboperations:

INIT(operation, params, handle) Process provides all relevant parameters for its participation in the communication operation (type of operation, data buffer, tag, participants, etc.). A handle is created that identifies the operation.

START(handle) The communication operation is started

COMPLETE(handle) The communication operation is completed.

FREE(handle) The handle, and associated resources are freed.

Correct invocation of these suboperations is a sequence of the form

INIT (START COMPLETE)* FREE.

I.e., a handle needs be created before communication occurs; it can be reused only after the previous use has completed; and it needs to be freed eventually (of course, one can assume that all handles are freed at program termination, by default).

The above scenario pertains to *persistent* handles. One can also create *ephemeral* handles. Such handle persists only until the communication operation is completed, at which point it is destroyed. Thus, correct invocation of suboperations with an ephemeral handle is **INIT START COMPLETE.**

A user may directly invokes these suboperations. This would allow to amortize the overhead of setting up a communication over many successive uses of the same handle, and allows to overlap communication and computation. Simpler communication operations combine several of these suboperations into one operation, thus simplifying the use of communication primitives. Thus, one only needs to specify precisely the semantics of these suboperations in order to specify the semantics of MPI point to point communication operations.

We say that a communication operation (send or receive) is **posted** once a **start** suboperation was invoked; the operation is **completed** once the **complete** suboperation completes. A send and a receive operation **match** if the receive pattern specified by the receive matches the message envelope created by the send.

2.9.1 Communication Object Creation

An object for a send operation is created by a call to **MPI_INIT_SEND**. A call to **MPI_INIT_RECV** is similarly used for creating an object for a receive operation. The creation of a communication object is a local operation that need not involve communication with a remote process.

MPI_INIT_SEND (handle, buffer_handle, dest, tag, context, mode, persistence)

Creates a send communication object. Parameters are

OUT handle message handle. The handle should not be associated with any object before the call.

IN buffer_handle handle to send buffer descriptor

IN dest rank in context of destination (integer)

IN tag user tag for messages sent with this handle (integer)

IN context handle to context of messages sent with this handle

IN mode send mode (state type, with two values: `MPI_REGULAR` and `MPI_READY`)

IN persistence handle persistence (state type, with two values: `MPI_PERSISTENT` and `MPI_EPHEMERAL`)

MPI_INIT_RECV (handle, buffer_handle, source, tag, context, persistence)

Create a receive handle. Parameters are

OUT handle message handle. The handle should not be associated with any object before the call.

IN buffer_handle handle to receive buffer descriptor.

IN source rank in context of source, or DONTCARE (integer).

IN tag user tag for messages received with this handle, or DONTCARE (integer).

IN context handle to context of messages received with this handle.

IN persistence handle persistence (state type, with two values: `MPI_PERSISTENT` and `MPI_EPHEMERAL`)

See Section 2.5.2 for a discussion of source, tag and context.

Discussion: I have not included proposals for partially specified message handles, that some peoples seem to desire.

I have merged all handle setup into one call.

2.9.2 Communication Start

MPI_START(handle)

IN handle communication handle

The `MPI_START` function starts the execution of a communication operation (send or receive). A sender should not update the send buffer after a send operation has started and until it is completed. A receiver should not access the receive buffer after a receive operation was started and until it is completed. A program that does not satisfy this condition is erroneous and its outcome is undetermined.

2.9.3 Communication Completion

MPI_WAIT (handle, return_status_handle)

IN handle communication handle

OUT return_handle handle that is associated with return status object.

A call to **MPI_WAIT** returns when the send operation identified by `handle` is complete. The completion of a send operation indicates that the sender is now free to update the locations in the send buffer, or any other location that can be referenced by the send operation. However, it does not indicate that the message has been received; rather it may have been buffered by the communication subsystem.

The completion of a receive operation indicates that the receiver is now free to access the locations in the receive buffer, which contain the received message, or any other location that can be referenced by the receive operation. It does not indicate that the matching send operation has completed.

The call returns a handle to an opaque object that contains information on the completed operation – the **return status** object.

MPI_STATUS (handle, flag, return_handle)

IN handle communication handle

OUT flag logical

OUT return_status_handle handle that is associated with return status object.

A call to **MPI_STATUS** returns `flag=true` if the operation identified by `handle` is complete, In such case, the return handle points to an opaque object that contains information on the completed information. It returns `flag=false`, otherwise. In such case, the value of the return handle is undefined.

Implementation notes:

A call to `MPI_WAIT` blocks only the executing thread. If the executing process is multi-threaded, then other threads within the process can be scheduled for execution.

The use of a blocking receive operation (`MPI_WAIT`) allows the operating system to deschedule the blocked thread and schedule another thread for execution, if such is available. The use of a nonblocking receive operation (`MPI_STATUS`) allows the user to schedule alternative activities within a single thread of execution.

The intended implementation of `MPI_STATUS` is for that operation to return as soon as possible. However, if repeatedly called for an operation that is enabled, it must eventually succeed.

The return status object for a send operation carries no information. The return status object for a receive operation carries information on the source, tag and length of the received message. These fields are required because the receive operation may have specified `DONTCARE` in either source or tag field, and the message may have been shorter than the receive buffer.

MPI_RETURN_STAT(handle, len, source, tag)

IN handle handle to return status object

OUT len difference between length of receive buffer and length of received message, in bytes. Thus, the value returned is zero if the received message matches the the receive buffer, positive if it is shorter (integer).

OUT source rank of message sender in message context (integer).

OUT tag tag of received message (integer).

Discussion:
I put the difference between message buffer and message length as the value returned, rather than length of received message, so that it might be easy to test for exact match.

The use of a return status object, rather than a list of parameters may simplify the use of MPI routines, if the values stored in the object are seldom checked. A predefined return status object should be provided, to ease programming.

2.9.4 Multiple Completions

It is convenient to be able to wait for the completion of any or all the operations in a set, rather than having to wait for specific message. A call to `MPI_WAITANY` or `MPI_STATUSANY` can be used to wait for the completion of one out of several operations; a call to `MPI_WAITALL` can be used to wait for all pending operations in a list.

MPI_WAITANY (list_of_handles, index, return_status_handle)

Blocks until one of the operations associated with the communication handles in the array has completed. Returns the index of that handle in the array, and returns the status of that operation in the object associated with the return_status_handle. The parameters are:

IN list_of_handles list of handles to communication objects.

OUT index index of handle for operation that completed (integer).

OUT return_status_handle handle that is associated with return status object. Set to return status of operation that completed.

The successful execution of `MPI_WAITANY(list_of_handles, index, return_status_handle)` is equivalent to the successful execution of `MPI_WAIT(handle[i], return_status_handle)`, where `i` is the value returned by `index` and `handle[i]` is the `i`-th handle in the list.

If more then one operation is enabled and can terminate, one is arbitrarily chosen (subject to the restrictions on operation termination order, see Section 2.13).

```
MPI_WAITANY ( list_of_handles, index, return_status_handle)
is

   {MPI_WAIT_RECV (handle[1], \return\_status\_handle); index = 0} || ...
||
   {MPI_WAIT_RECV (handle[n], \return\_status\_handle); index = n-1}
```

("||" indicates choice; one of the alternatives is chosen, nondeterministically.)

MPI_STATUSANY (list_of_handles, index, return_status_handle)

Causes either one or none of the operations associated with the communication handles to return. In the former case, it has the same return semantics as a call to `MPI_WAIT_ANY`. In the later case, it returns a value of -1 in `index` and `return_status_handle` is undefined. The parameters are:

IN list_of_handles list of handles to communication objects.

OUT index index of handle for operation that completed, or -1 if none completed (integer).

OUT return_status_handle handle that is associated with return status object. Set to return status of operation that completed, if any; undefined when `index = -1`.

MPI_WAITALL(list_of_handles, list_of_return_status_handles)

Blocks until all communication operations associated with handles in the list complete, and return the status of all these operations. The parameters are:

IN list_of_handles list of handles to communication objects.

OUT list_of_return_status_handles Must have the same length as the first list. Each return status object is set to the return status of the corresponding operation in the first list.

2.10 Blocking Communication

Blocking send and receive operations combine all communication suboperations into one call. The operation returns only when the communication completes and no communication object persists after the call completed. However, the buffer descriptor object needs be created ahead of the call.

We use the following naming convention for such operations:

$$\begin{bmatrix} - \\ \mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{SEND} \\ \mathbf{RECV} \end{bmatrix}$$

The first letter (void or **R**) indicates the start mode (regular or ready).

MPI_SEND (buffer_handle, dest, tag, context) is

```
MPI_INIT_SEND(handle, buffer_handle, dest, tag, context, REGULAR, EPHEMERAL)
MPI_START(handle)
MPI_WAIT(handle, null)
```

MPI_RSEND (buffer_handle, dest, tag, context) is

```
MPI_INIT_SEND(handle, buffer_handle, dest, tag, context, READY, EPHEMERAL)
MPI_START(handle)
MPI_WAIT(handle, null)
```

MPI_RECV(buffer_handle, source, tag, context, return_status_handle) is

```
MPI_INIT_RECV(handle, buffer_handle, source, tag, context, EPHEMERAL)
MPI_START(handle)
MPI_WAIT(handle,return_status_handle)
```

Implementation note:

While these functions can be implemented via calls to functions that implement suboperations, as described in this subsection, an efficient implementation may optimize away these multiple calls, provided it does not change the behavior of correct programs.

2.11 Nonblocking Communication

Nonblocking send and receive operations combine the first two suboperations (`INIT` and `START`) into one call. They use ephemeral communication objects, so that the operation is completed, and the associated resources are freed, by using one of the functions `MPI_WAIT`, `MPI_STATUS`, `MPI_WAITANY`, `MPI_STATUSANY`, or `MPI_WAITALL`. Here, too, a buffer object has to be created ahead of the communication initiation operation.

We use the same naming convention as for blocking operations: a prefix of **R** indicates the `READY` mode. In addition, a prefix of **I** is used to indicate *immediate* (i.e., nonblocking) execution.

MPI_ISEND (handle, buffer_handle, dest, tag, context) is

```
MPI_INIT_SEND(handle, buffer_handle, dest, tag, context, REGULAR, EPHEMERAL)
MPI_START(handle)
```

MPI_IRSEND (handle, buffer_handle, dest, tag, context) is

```
MPI_INIT_SEND(handle, buffer_handle, dest, tag, context, READY, EPHEMERAL)
MPI_START(handle)
```

MPI_IRECV(handle, buffer_handle, source, tag, context, return_status_handle) is

```
MPI_INIT_RECV(handle, buffer_handle, source, tag, context, EPHEMERAL)
MPI_START(handle)
```

2.12 Block Sending Operations

The most frequent type of buffer used is a contiguous buffer of numeric storage units, i.e., a contiguous buffer of words that may contain either INTEGER, REAL or LOGICAL values (in FORTRAN). In a homogeneous environment such messages can be used to send arbitrary sequences of contiguous items, where each item occupies an integer number of words.

We specialize the functions in the two previous subsections to this case, thus avoiding the need for the creation of a buffer descriptor object. We use the same naming scheme used in the previous subsections, and append a **B** in the function name, for `BLOCK`.

MPI_SENDB (start, len, dest, tag, context) is

```
MPI_CREATE(buffer_handle, MPI_BUFFER, MPI_EPHEMERAL)
MPI_ADD(buffer_handle, start, len, MPI_REAL)
MPI_SEND (buffer_handle, len, dest, tag, context)
```

MPI_RSENDB (buffer_handle, dest, tag, context) is

```
MPI_CREATE(buffer_handle, MPI_BUFFER, MPI_EPHEMERAL)
MPI_ADD(buffer_handle, start, len, MPI_REAL)
MPI_RSEND (buffer_handle, len, dest, tag, context)
```

MPI_RECVB(buffer_handle, source, tag, context, return_status_handle) is

```
MPI_CREATE(buffer_handle, MPI_BUFFER, MPI_EPHEMERAL)
MPI_ADD(buffer_handle, start, len, MPI_REAL)
MPI_RECV(buffer_handle, source, tag, context, return_status_handle)
```

MPI_ISENDB (handle, buffer_handle, dest, tag, context) is

```
MPI_CREATE(buffer_handle, MPI_BUFFER, MPI_EPHEMERAL)
MPI_ADD(buffer_handle, start, len, MPI_REAL)
MPI_ISEND(handle, buffer_handle, dest, tag, context)
```

MPI_IRSENDB (handle, buffer_handle, dest, tag, context) is

```
MPI_CREATE(buffer_handle, MPI_BUFFER, MPI_EPHEMERAL)
MPI_ADD(buffer_handle, start, len, MPI_REAL)
MPI_IRSEND(handle, buffer_handle, dest, tag, context)
```

MPI_IRECVB(handle, buffer_handle, source, tag, context, return_status_handle) is

```
MPI_CREATE(buffer_handle, MPI_BUFFER, MPI_EPHEMERAL)
MPI_ADD(buffer_handle, start, len, MPI_REAL)
MPI_IRECV(handle, buffer_handle, source, tag, context)
```

Discussion:
I use word count, rather than byte count. I believe word messages are much more prevalent than byte messages, and it's a blessing not to have to multiply by 4 for each message. Byte messages are still available the hard way. Also, word messages fit well with the Fortran numeric storage unit concept.

If we are to add more functions, my next addition would be a vector send operation.

2.13 Correctness

Discussion: The material in this section has not yet been discussed by MPIF. Some or all of it is likely to move to Section 6. It is incorporated here for completeness.

2.13.1 Order

MPI preserves the order of messages between any fixed pair of processes. In other words, if process A executes two successive send `start` suboperations, process B executes two successive receive `start` operations, and both receives match either sends, then the first receive will receive the message sent by the first send, and the second receive will receive the message sent by the second send. Thus, if a two messages from the same source can satisfy a pending receive, the first message sent is accepted; if a message can satisfy two pending receives, the first receive posted is satisfied.

The last paragraph assumes that the send `start` operations are ordered by the program order at process A, and the receive `start` operations are ordered by the program order at process B. If a process is multithreaded and the operations are executed by distinct threads, then the semantics of the threaded system may not define an order between the two operations, in which case the condition is void.

2.13.2 Progress and Fairness

We can model the execution of MPI programs as an interaction between executing processes that execute each their own program, and the **communication subsystem**. The communication subsystem may have various constraints on the amount of resources it can use. E.g.:

Bounds on the number and total sizes of active communication objects. Such bound can be global, per node, or per pair of communicating nodes.

Bounds on the number and total sizes of messages buffered in the system. Such bound can, again, be global, per node, or per pair of communicating node. In addition, a message may be buffered at the sender, at the receiver, at both, or perhaps at another place altogether.

Thus, it will be difficult to set rules on resource management of the communication subsystem. However, it is generally expected that implementers will provide information on the mechanism used for resource allocation, and that query and set functions will allow to query and possibly control the amount of available resources.

We provide in this section a set of minimal requirements on the communication subsystem. Programs that execute on any subsystem that fulfils these minimal requirements are **safe** and will port to any MPI implementation. **Unsafe** programs may execute on some MPI implementations, depending on the amount of available resources and the implementation used for the MPI communication subsystem. Finally **erroneous** programs never execute correctly. (While it is desirable to detect erroneous programs, it is not possible to do so at compile time, and often prohibitive to do so a run time. Thus, the document does not specify a behavior for erroneous programs, although the desired behavior is to return a useful error message.)

1. If a process executes an `INIT` operation, then the operation eventually succeeds, or a *resource exception* occurs. The standard does not specify when a resource exception is allowed to occur. It is expected that an operational definition will be made available, in the form of test programs that have to execute with no resource exceptions. It is highly desirable to have generous bounds on the number of concurrently active communication objects each process may have, so that, in practice, `INIT` operations will always be guaranteed to succeed.

2. Each process can initiate a communication operation for each active communication object. I.e. correct `START` operations always succeed (eventually).

3. A send operation is **enabled** if the sending process has issued a `COMPLETE` operation and the receiving process has issued a `START` operation for a matching receive. Symmetrically, a receive operation is **enabled** if the receiving process has issued a `COMPLETE` operation and the sending process has issued a `START` operation for a matching send. An enabled operation may become **disabled** either because it completes successfully or, in the case of a receive, because the matching message is successfully received by another receive operation.

   An enabled operation either completes successfully or becomes permanently disabled.

4. A `FREE` operation always succeeds (eventually).

The four conditions guarantee progress in the communication subsystem. The third condition guarantee (weak) fairness among competing communication requests.

Examples (involving two processors with ranks 0 and 1)

The following program is safe, and should always succeed.

```
CALL MPI_RANK(rank, context)
IF (rank.EQ.0) THEN
   CALL MPI_SENDB(sendbuf, len, 1, tag, context)
   CALL MPI_RECVB(recvbuf, len, 1, tag, context)
ELSE    ! rank.EQ.1
   CALL MPI_RECVB(recvbuf, len, 0, tag, context)
   CALL MPI_SENDB(sendbuf, len, 0, tag, context)
END IF
```

The following program is erroneous, and should always fail.

```
CALL MPI_RANK(rank, context)
IF (rank.EQ.0) THEN
   CALL MPI_RECVB(recvbuf, len, 1, tag, context)
   CALL MPI_SENDB(sendbuf, len, 1, tag, context)
ELSE    ! rank.EQ.1
   CALL MPI_RECVB(recvbuf, len, 0, tag, context)
   CALL MPI_SENDB(sendbuf, len, 0, tag, context)
END IF
```

The receive operation of the first process must complete before its send, and can complete only if the matching receive of the second processor is executed; the receive operation of the second process must complete before its send and can complete only if the matching receive of the first process is executed. This program will deadlock.

The following program is unsafe, and may succeed or fail, depending on implementation.

```
CALL MPI_RANK(rank, context)
```

```
IF (rank.EQ.0) THEN
   CALL MPI_SENDB(sendbuf, len, 1, tag, context)
   CALL MPI_RECVB(recvbuf, len, 1, tag, context)
ELSE    ! rank.EQ.1
   CALL MPI_SENDB(sendbuf, len, 0, tag, context)
   CALL MPI_RECVB(recvbuf, len, 0, tag, context)
END IF
```

The message sent by each process has to be copied out before the send operation returns and the receive operation starts. For the program to complete, it is necessary that at least one of the two messages sent is buffered out of either processes' address space. Thus, this program can succeed only if the communication system has sufficient buffer space to buffer `len` words of data.

If additional requirements will become part of the standard (e.g., bounds on the minimal number of concurrently active handles that need be supported, then further programs become safe.

Chapter 3

Collective Communication

Al Geist
Marc Snir

3.1 Introduction

This section is a draft of the current proposal for collective communication. Collective communication is defined to be communication that involves a group of processes. Examples are broadcast and global sum. A collective operation is executed by having all processes in the group call the communication routine, with matching parameters. Routines can (but are not required to) return as soon as their participation in the collective communication is complete. The completion of a call indicates that the caller is now free to access the locations in the communication buffer, or any other location that can be referenced by the collective operation. However, it does not indicate that other processes in the group have started the operation (unless otherwise indicated in the description of the operation). However, the successful completion of a collective communication call may depend on the execution of a matching call at all processes in the group.

The syntax and semantics of the collective operations is defined so as to be consistent with the syntax and semantics of the point to point operations.

The reader is referred to the point-to-point communication section of the current MPI draft for information concerning communication buffers and their manipulations. The context section describes the formation, manipulation, and query functions (such as group size) that are available for groups and group objects.

The collective communication routines are built above the point-to-point routines. While vendors may optimize certain collective routines for their architectures, a complete library of the collective communication routines can be written entirely using point-to-point communication functions. We are using naive implementations of the collective calls in terms of point to point operations in order to provide an operational definition of their semantics.

The following communication functions are proposed.

- Broadcast from one member to all members of a group.
- Barrier across all group members
- Gather data from all group members to one member.

- Scatter data from one member to all members of a group.
- Global operations such as sum, max, min, etc., were the result is known by all group members and a variation where the result is known by only one member. The ability to have user defined global operations.
- Simultaneous shift of data around the group, the simplest example being all members sending their data to (rank+1) with wrap around.
- Scan across all members of a group (also called parallel prefix).
- Broadcast from all members to all members of a group.
- Scatter data from all members to all members of a group (also called complete exchange or index).

To simplify the collective communication interface it is designed with two layers. The low level routines have all the generality of, and make use of, the communication buffer routines of the point-to-point section which allows arbitrarily complex messages to be constructed. The second level routines are similar to the upper level point-to-point routines in that they send only a contiguous buffer.

3.2 Group Functions

A full description of the group formation and manipulation functions can be found in the context chapter of the MPI document. Here we describe only those group functions that are used in the semantic description of the collective communication routines.

An initial group containing all processes is supplied by default in MPI. MPI provides a procedure that returns the handle to this initial group. The processes in the inital group each have a unique rank in the group represented by integers (0, 1, 2, ..., [number-of-processes - 1].

MPI_ALLGROUP(group) Return the descriptor of the inital group containing all processes.

OUT group handle to descriptor object of initial group.

MPI_RANK(group, rank) Return the rank of the calling process within the specified group.

IN group group handle

OUT rank integer

MPI_GSIZE(group, size) Return the number of processes that belong to the specified group.

IN group group handle

OUT size integer

3.3 Communication Functions

The proposed communication functions are divided into two layers. The lowest level uses the same buffer descriptor objects available in point-to-point to create noncontiguous, multiple data type messages. The second level is similar to the block send/receive point-to-point operations in that it supports only contiguous buffers of data. For each communication operation, we list these two level of calls together.

3.4 Synchronization

Barrier synchronization

MPI_BARRIER(group)

MPI_BARRIER blocks the calling process until all group members have called it; the call returns at any process only after all group members have entered the call.

IN group group handle

`MPI_BARRIER( group )` is

```
MPI_CREATE_BUFFER(buffer_handle, MPI_BUFFER, MPI_PERSISTENT);
MPI_GSIZE( group, &size );
MPI_RANK( group, &rank );
if (rank==0)
{
   for (i=1; i < size; i++)
      MPI_RECV(buffer_handle, i, tag, group, return_handle);
   for (i=1; i < size; i++)
      MPI_SEND(buffer_handle, i, tag, group);
}
else
{
   MPI_SEND(buffer_handle, 0, tag, group);
   MPI_RECV(buffer_handle, 0, tag, group, return_handle);
}
MPI_FREE(buffer_handle);
```

3.5 Data move functions

Broadcast

MPI_BCAST(buffer_handle, group, root)

`MPI_BCAST` broadcasts a message from the process with rank `root` to all other processes of the group. It is called by all members of group using the same arguments for `group, and root`. On return the contents of the buffer of the process with rank `root` is contained in the buffer of the calling process.

INOUT buffer_handle Handle for buffer where from message is sent or received.

IN group group handle

IN root rank of broadcast root (integer)

MPI_BCASTC(buf, len, type, group, root)

MPI_BCASTC behaves like broadcast, restricted to a block buffer. It is called by all processes with the same arguments for `len, group` and `root`.

INOUT buffer Starting address of buffer (choice type)

IN len Number of entries in buffer (integer)

IN type data type of buffer

IN group group handle

in root rank of broadcast root (integer)

`MPI_BCAST( buffer_handle, group, root )` is

```
MPI_GSIZE( group, &size );
MPI_RANK( group, &rank );
MPI_IRECV(handle, buffer_handle, root, tag, group, return_handle);
if (rank==root)
   for (i=0; i < size; i++)
      MPI_SEND(buffer_handle, i, tag, group);
MPI_WAIT(handle)
```

Circular shift

MPI_CSHIFT(inbuf, outbuf, group, shift)

Process with rank `i` sends the data in its input buffer to process with rank (`i` + `shift`) mod `group_size`, who receives the data in its output buffer. All processes make the call with the same values for `group`, and `shift`. The `shift` value can be positive, zero, or negative.

IN inbuf handle to input buffer descriptor

IN outbuf handle to output buffer descriptor

IN group handle to group

IN shift integer

MPI_CSHIFT1(buf, group, shift)

Process with rank `i` sends the data in its buffer to process with rank (`i` + `shift`) mod `group_size`, who receives the data in the same buffer. All processes make the call with the same values for `group`, and `shift`. The `shift` value can be positive, zero, or negative.

INOUT buf handle to buffer descriptor

IN group handle to group

IN shift integer

MPI_CSHIFTC(inbuf, outbuf, len, type, group, shift)

Behaves like `MPI_CSHIFT`, with buffers restricted to be blocks of numeric units. All processes make the call with the same values for `len, group`, and `shift`.

IN inbuf initial location of input buffer

OUT outbuf initial location of output buffer

IN len number of entries in input (and output) buffers

IN type data type of buffer

IN group handle to group

IN shift integer

MPI_CSHIFTC1(buf, len, type, group, shift)

Behaves like `MPI_CSHIFT1`, with buffers restricted to be blocks of numeric units. All processes make the call with the same values for `len, group`, and `shift`.

INOUT buf initial location of buffer

IN len number of entries in input (and output) buffers

IN type data type of buffer

IN group handle to group

IN shift integer

`MPI_CSHIFT( inbuf, outbuf, group, shift)` is

```
MPI_GSIZE( group, &size );
MPI_RANK( group, &rank );
MPI_ISEND( handle, inbuf, mod(rank+shift, size), tag, group);
MPI_RECV( outbuf, mod(rank-shift,size), tag, group, return_handle)
MPI_WAIT(handle);
```

End-off shift

MPI_EOSHIFT(inbuf, outbuf, group, shift)

Process with rank `i`, $\max(\texttt{0}, -\texttt{shift}) \le \texttt{i} < \min(\texttt{size}, \texttt{size} - \texttt{shift})$, sends the data in its input buffer to process with rank `i+ shift`, who receives the data in its output buffer. The output buffer of processes which do not receive data is left unchanged. All processes make the call with the same values for `group`, and `shift`.

IN inbuf handle to input buffer descriptor

IN outbuf handle to output buffer descriptor

IN group handle to group

IN shift integer

MPI_EOSHIFT1(buf, group, shift)

Process with rank `i`, $\max(0, -\mathtt{shift}) \leq \mathtt{i} < \mathtt{min}(\mathtt{size}, \mathtt{size} - \mathtt{shift})$, sends the data in its buffer to process with rank `i+ shift`, who receives the data in the same buffer. The output buffer of processes which do not receive data is left unchanged. All processes make the call with the same values for `group`, and `shift`.

INOUT buf handle to buffer descriptor

IN group handle to group

IN shift integer

MPI_EOSHIFTC(inbuf, outbuf, len, type, group, shift)

Behaves like `MPI_EOSHIFT`, with buffers restricted to be blocks of numeric units. All processes make the call with the same values for `len, group`, and `shift`.

IN inbuf initial location of input buffer

OUT outbuf initial location of output buffer

IN len number of entries in input (and output) buffers

IN type data type of buffer

IN group handle to group

IN shift integer

MPI_EOSHIFTC1(buf, len, type, group, shift)

Behaves like `MPI_EOSHIFT1`, with buffer restricted to be blocks of numeric units. All processes make the call with the same values for `len, group`, and `shift`.

INOUT buf initial location of buffer

IN len number of entries in buffer

IN type data type of buffer

IN group handle to group

IN shift integer

Gather

MPI_GATHER(inbuf, list_of_outbufs, group, root, return_status)

Each process (including the root process) sends the content of its input buffer to the root process. The root process places all the incoming messages in the location specified by the output buffer handle corresponding to the sender's rank. For example, the root places the data from process with rank 3 in the location specified by the third buffer descriptor in the list of outbufs. The list_of_outbufs argument is ignored for all non-root processes. The routine is called by all members of group using the same arguments for `group`, and `root`. The input buffer of each process may have a different length.

IN inbuf handle to input buffer descriptor

IN list_of_outbufs list of buffer descriptor handles (root)

IN group group handle

IN root rank of receiving process (integer)

OUT return_status return status handle

Discussion: Do we want the collective routines to have return status handles? And if so what information do we want the handle to be able to return?

MPI_GATHERC(inbuf, outbuf, inlen, type, group, root)

`MPI_GATHERC` behaves like `MPI_GATHER` restricted to block buffers, and with the additional restriction that all input buffers should have the same length. All processes should provided the same values for `inlen`, `group`, and `root` .

IN inbuf first variable of input buffer (choice)

OUT outbuf first variable of output buffer – significant only at root (matches type)

IN inlen Number of (word) variables in input buffer (integer)

IN type data type of buffer

IN group group handle

IN root rank of receiving process (integer)

`MPI_GATHERC( inbuf, outbuf, inlen, type, group, root)` is

```
MPI_GSIZE( &size, group);
MPI_RANK( &rank, group);
MPI_ISENDC(handle, inbuf, inlen, root, tag, group);
if (rank==root)
   for (i=0; i < size; i++)
   {
```

```
      MPI_RECVC(outbuf, inlen, i, tag, group, return_status);
      outbuf += inlen;
   }
MPI_WAIT(handle);
```

Scatter

MPI_SCATTER(list_of_inbufs, outbuf, group, root, return_status)

The root process sends the content of its `i`-th input buffer to the process with rank `i`; each process (including the root process) stores the incoming message in its output buffer. The routine is called by all members of the group using the same arguments for `group`, and `root`.

IN list_of_inbufs list of buffer descriptor handles

IN outbuf buffer descriptor handle

IN group handle

IN root rank of sending process (integer)

OUT return_status return status handle

`MPI_SCATTER( list_of_inbufs, outbuf, group, root, return_status)` is

```
MPI_GSIZE( group, &size );
MPI_RANK( group, &rank );
MPI_IRECV(handle, outbuf, root, tag, group);
if (rank=root)
   for (i=0; i < size; i++)
      MPI_SEND(inbuf[i], i, tag, group);
MPI_WAIT(handle, return_status);
```

MPI_SCATTERC(inbuf, outbuf, len, type, group, root)

`MPI_SCATTERC` behaves like `MPI_SCATTER` restricted to block buffers, and with the additional restriction that all output buffers have the same length. The input buffer block of the root process is partitioned into `n` consecutive blocks, each consisting of `len` words. The `i`-th block is sent to the `i`-th process in the group and stored in its output buffer. The routine is called by all members of the group using the same arguments for `group, len`, and `root`.

IN inbuf first entry in input buffer – significant only at root (choice).

OUT outbuf first entry in output buffer (choice).

IN len number of entries to be stored in output buffer (integer)

IN type data type of buffer

IN group handle

IN root rank of sending process (integer)

MPI_SCATTERC(inbuf, outbuf, outlen, type, group, root) is

```
MPI_GSIZE( &size, group);
MPI_RANK( &rank, group);
MPI_IRECVC( handle, outbuf, outlen, type, root, tag, group, return_handle);
if (rank=root)
   for (i=0; i < size; i++)
   {
      MPI_SENDC(inbuf, outlen, type, i, tag, group);
      inbuf += outlen;
   }
MPI_WAIT(handle);
```

All-to-all scatter

MPI_ALLSCATTER(list_of_inbufs, list_of_outbufs, group, return_status)

Each process in the group sends its `i`-th buffer in its input buffer list to the process with rank `i` (itself included); each process places the incoming messages in the location specified by output buffer handle corresponding to the rank of the sender. For example, each process places the data from process with rank 3 in the location specified by the third buffer descriptor in the list of outbufs. The routine is called by all members of the group using the same arguments for `group`.

IN list_of_inbufs list of buffer descriptor handles

IN list_of_outbufs list of buffer descriptor handles

IN group handle

OUT return_status return status handle

MPI_ALLSCATTERC(inbuf, outbuf, len, type, group)

`MPI_ALLSCATTERC` behaves like `MPI_ALLSCATTER` restricted to block buffers, and with the additional restriction that all blocks sent from one process to another have the same length. The input buffer block of each process is partitioned into `n` consecutive blocks, each consisting of `len` words. The `i`-th block is sent to the `it`-th process in the group. Each process concatenates the incoming messages, in the order of the senders' ranks, and store them in its output buffer. The routine is called by all members of the group using the same arguments for `group`, and `len`.

IN inbuf first entry in input buffer (matches type).

OUT outbuf first entry in output buffer (matches type).

IN len number of entries sent from each process to each other (integer).

IN type data type of buffer

IN group handle

`MPI_ALLSCATTERC( inbuf, outbuf, len, type, group)` is

```
MPI_GSIZE( group, &size );
MPI_RANK( group, &rank );
for (i=0; i < rank; i++)
   {
    MPI_IRECVC(recv_handles[i], outbuf, len, type, tag, group, return_handle);
    outbuf += len;
   }
for (i=0; i < size; i++)
   {
    MPI_ISENDC(send_handle[i], inbuf, len, type, i, tag, group);
    inbuf += len;
   }
MPI_WAITALL(send_handle);
MPI_WAITALL(recv_handle);
```

All-to-all broadcast

MPI_ALLCAST(inbuf, list_of_outbufs, group, return_status)

Each process in the group broadcasts its input buffer to all processes (including itself); each process places the incoming messages in the location specified by output buffer handle corresponding to the rank of the sender. For example, each process places the data from process with rank 3 in the location specified by the third buffer descriptor in the list of outbufs. The routine is called by all members of the group using the same arguments for `group`.

IN inbuf buffer descriptor handle for input buffer

IN list_of_outbufs list of buffer descriptor handles

IN group handle

OUT return_status return status handle

MPI_ALLCASTC(inbuf, outbuf, len, type, group)

`MPI_ALLCASTC` behaves like `MPI_ALLCAST` restricted to block buffers, and with the additional restriction that all blocks sent from one process to another have the same length. Each process concatenates the incoming messages, in the order of the senders' ranks, and store them in its output buffer. The routine is called by all members of the group using the same arguments for `group`, and `len`.

IN inbuf first entry in input buffer (choice). root (integer)

OUT outbuf first entry in output buffer (choice).

IN len number of entries sent from each process to each other (including itself).

IN type data type of buffer

IN group group handle

MPI_ALLCASTC(inbuf, outbuf, len, type, group) is

```
MPI_GSIZE( group, &size );
MPI_RANK( group, &rank );
for (i=0; i < rank; i++)
   {
    MPI_IRECVC(recv_handles[i], outbuf, len, type, tag, group, return_handle);
    outbuf += len;
   }
for (i=0; i < size; i++)
   {
    MPI_ISENDC(send_handle[i], inbuf, len, type, i, tag, group);
   }
MPI_WAITALL(send_handle);
MPI_WAITALL(recv_handle);
```

3.6 Global Compute Operations

Reduce

MPI_REDUCE(inbuf, outbuf, group, root, op)

Combines the values provided in the input buffer of each process in the group, using the operation op, and returns the combined value in the output buffer of the process with rank root. Each process can provide one value, or a sequence of values, in which case the combine operation is executed pointwise on each entry of the sequence. For example, if the operation is max and input buffers contains two floating point numbers, then outbuf(1) = global max(inbuf(1)) and outbuf(2) = global max(inbuf(2)). All input buffers should define sequences of equal length of entries of types that match the type of the operands of op. The output buffer should define a sequence of the same length of entries of types that match the type of the result of op. (Note that, here as for all other communication operations, the type of entries inserted in a message depend on the information provided by the input buffer descriptor, and not on the declarations of these variables in the calling program. The types of the variables in the calling program need not match the types defined by the buffer descriptor, but in such case the outcome of a reduce operation may be implementation dependent.)

The operation defined by op is associative and commutative, and the implementation can take advantage of associativity and commutativity in order to change order of evaluation. The routine is called by all group members using the same arguments for group, root and op.

IN inbuf handle to input buffer

IN outbuf handle to output buffer – significant only at root

IN group handle to group

IN root rank of root process (integer)

IN op operation

The buffer descriptor contains data type information so that the correct form of the operation can be performed. We list below the operations which are supported.

MPI_MAX maximum

MPI_MIN minimum

MPI_MIN minimum

MPI_SUM sum

MPI_PROD product

MPI_AND and (logical or bit-wise integer)

MPI_OR or (logical or bit-wise integer)

MPI_XOR xor (logical or bit-wise integer)

MPI_MAXLOC rank of process with maximum value

MPI_MINLOC rank of process with minimum value

MPI_REDUCEC(inbuf, outbuf, len, type, group, root, op)

Is same as `MPI_REDUCE`, restricted to a block buffer.

IN inbuf first location in input buffer

OUT outbuf first location in output buffer – significant only at root

IN len number of entries in input and output buffer (integer)

IN type data type of buffer

IN group handle to group

IN root rank of root process (integer)

IN op operation

MPI_USER_REDUCE(inbuf, outbuf, group, root, function)

Same as the reduce operation function above except that a user supplied function is used. `function` is an associative and commutative function with two arguments. The types of the two arguments and of the returned value of the function, and the types of all entries in the input and output buffers all agree. The output buffer has the same length as the input buffer.

IN inbuf handle to input buffer

IN outbuf handle to output buffer – significant only at root

IN group handle to group

IN root rank of root process (integer)

MPI_USER_REDUCEC(inbuf, outbuf, len, type, group, root, function)

Is same as MPI__USER_REDUCE, restricted to a block buffer.

IN inbuf first location in input buffer

OUT outbuf first location in output buffer – significant only at root

IN len number of entries in input and output buffer (integer)

IN type data type of buffer

IN group handle to group

IN root rank of root process (integer)

IN function user provided function

Discussion:
Do we also want a version of reduce that broadcasts the result to all processes in the group? (This can be achieved by a reduce followed by a broadcast, but a combined function may be somewhat more efficient.) These would be respectively:

MPI_GOP(inbuf, outbuf, group, op)

MPI_GOPC(inbuf, outbuf, len, type, group, op)

MPI_USER_GOP(inbuf, outbuf, group, function)

MPI_USER_GOPC(inbuf, outbuf, len, type, group, function)

Do we want a user provided *function* (two IN parameters, one OUT value), or a user provided procedure that overwrites the second input (ie. one IN param, one INOUT param, the equivalent of C `a op= b` type assignment)? The second choice may allow a more efficient implementation, without changing the semantics of the MPI call.

Scan

MPI_SCAN(inbuf, outbuf, group, op)

MPI_SCAN is used to perform a parallel prefix with respect to an associative reduction operation on data distributed across the group. The operation returns in the output buffer of the process with rank `i` the reduction of the values in the input buffers of processes with ranks `0,...,i`. The type of operations supported and their semantic, and the constraints on input and output buffers are as for MPI_REDUCE.

IN inbuf handle to input buffer

IN outbuf handle to output buffer

IN group handle to group

IN op operation

MPI_SCANC(inbuf, outbuf, len, type, group, op) Same as `MPI_SCAN`, restricted to block buffers.

IN inbuf first input buffer element (choice)

OUT outbuf first output buffer element (choice)

IN len number of entries in input and output buffer (integer)

IN type data type of buffer

IN group handle to group

IN op operation

MPI_USER_SCAN(inbuf, outbuf, group, function)

Same as the scan operation function above except that a user supplied function is used. `function` is an associative and commutative function with two arguments. The types of the two arguments and of the returned values all agree.

IN inbuf handle to input buffer

IN outbuf handle to output buffer

IN group handle to group

IN function user provided function

MPI_USER_SCANC(inbuf, outbuf, len, type, group, function)

Is same as `MPI_USER_SCAN`, restricted to a block buffer.

IN inbuf first location in input buffer

OUT outbuf first location in output buffer

IN len number of entries in input and output buffer (integer)

IN type data type of buffer

IN group handle to group

IN function user provided function

Discussion:

Do we want scan operations executed by segments? (The HPF definition of prefix and suffix operation might be handy – in addition to the scanned vector of values there is a mask that tells where segments start and end.)

3.7 Correctness

Discussion: This is still very preliminary

The semantics of the collective communication operations can be derived from their operational definition in terms of point-to-point communication. It is assumed that messages pertaining to one operation cannot be confused with messages pertaining to another operation. Also messages pertaining to two distinct occurrences of the same operation cannot be confused, if the two occurrences have distinct parameters. The relevant parameters for this purpose are `group`, `root` and `op`. messages pertaining to another occurrence of the same operation, with different parameters. The implementer can, of course, use another, more efficient implementation, as long as it has the same effect.

Discussion:

This statement does not yet apply to the current, incomplete and somewhat careless definitions I provided in this draft.

The definition above means that messages pertaining to a collective communication carry information identifying the operation itself, and the values of the `group` and, where relevant, `root` or `op` parameters. Is this acceptable?

A few examples:

```
MPI_BCASTC(buf, len, type, group, 0);
MPI_BCASTC(buf, len, type, group, 1);
```

Two consecutive broadcasts, in the same group, with the same tag, but different roots. Since the operations are distinguishable, messages from one broadcast cannot be confused with messages from the other broadcast; the program is safe and will execute as expected.

```
MPI_BCASTC(buf, len, type, group, 0);
MPI_BCASTC(buf, len, type, group, 0);
```

Two consecutive broadcasts, in the same group, with the same tag and root. Since point-to-point communication preserves the order of messages here, too, messages from one broadcast will not be confused with messages from the other broadcast; the program is safe and will execute as intended.

```
MPI_RANK(&rank, group)
if (rank==0)
  {
```

```
   MPI_BCASTC(buf, len, type, group, 0);
   MPI_SENDC(buf, len, type, 2, tag, group);
  }
elseif (rank==1)
  {
   MPI_RECVC(buf, len, type, MPI_DONTCARE, tag, group);
   MPI_BCASTC(buf, len, type, group, 0);
   MPI_RECVC(buf, len, type, MPI_DONTCARE, tag, group);
  }
else
  {
   MPI_SENDC(buf, len, type, 2, tag, group);
   MPI_BCASTC(buf, len, type, group, 0);
  }
```

Process zero executes a broadcast followed by a send to process one; process two executes a send to process one, followed by a broadcast; and process one executes a receive, a broadcast and a receive. A possible outcome is for the operations to be matched as illustrated by the diagram below.

```
    0                       1                      2

                / - >  receive            / - send
              /                         /
broadcast   /         broadcast       /   broadcast
           /                         /
  send   -             receive  < -
```

The reason is that broadcast is not a synchronous operation; the call at a process may return before the other processes have entered the broadcast. Thus, the message sent by process zero can arrive to process one before the message sent by process two, and before the call to broadcast on process one.

Chapter 4

Process Topologies

Rolf Hempel

4.1 Introduction

A process group in MPI is a collection of n processes. Each process in the group is assigned a rank between 0 and n-1. In the majority of parallel applications this linear ordering does not reflect the communication structure of the processes which is usually determined by the underlying problem geometry and the numerical algorithm used. Often the processes are arranged in topological patterns as simple as a two– or three–dimensional grid.

For the user who programs an application on such a process structure the most natural way of referring to individual processes is to specify their coordinates in the process grid. Usually a process does not have to know the PID numbers of *all* processes in the grid, because communication only takes place with processes in a certain neighborhood, and perhaps with a few distant processes. Non–scalable features like the explicit listing of all participating processes in the application could, therefore, be avoided, if there were a means to specify the information which is really required.

Additionally, the user–defined process structure can be exploited by the system in the assignment of processes to hardware processors, if this helps to improve the communication performance on a given machine. How this mapping is done in detail is outside the scope of this standardization proposal, since it is too much dependent on the particular hardware. The description of a "virtual process topology", however, only depends on the application, and is, therefore, machine–independent. The functions which are proposed in this paper only deal with this side of the mapping problem.

There are well–known techniques for mapping grid/torus structures to hardware topologies such as hypercubes or grids. But even for more complicated graph structures good heuristics often yield near–optimal results [14]. On the other hand, if there is no way for the user to specify the logical process arrangement as a "virtual topology", this information is most difficult to recover later. A random mapping is most likely the consequence. On some machines this will lead to unnecessary contention in the interconnect network. Some details about predicted and measured performance improvements which result from good process–to–processor mapping on modern wormhole–routing architectures can be found in [5, 4].

4.2 Virtual Topologies

In the most general case the connection structure of a set of processes can be represented by a graph. The nodes stand for the processes, and the edges connect processes which communicate with each other. MPI provides message–passing between any pair of processes in a group. There is no requirement for opening a channel explicitly. Therefore, a "missing link" in the user–defined process graph does not prevent the corresponding processes from exchanging messages. It only means that this connection is neglected in the virtual topology. As a consequence, an automatic mapping tool will not try to assign the processes to neighboring processors.

The graph definition function MPI_GRAPH presented in this paper does not include any weighting of the communication volume along the different edges. Processes are either neighbors or not. Experience with similar techniques in PARMACS [11, 12] show that this information is usually sufficient for a good mapping. Additionally, a more precise specification would be more difficult for the user to set up, and it would make the interface functions more complicated.

Discussion: One could also have two graph mapping functions, one of which uses communication weighting, whereas the other does not. The mapping function with weighting, MPI_MAP_WGRAPH, would then have an additional input argument `weights` with the same structure as the neighbor list `nglist`. On the higher abstraction level the function MPI_GRAPH would be complemented by MPI_WGRAPH in the same way.

In principle a function for the definition of such a graph structure would be sufficient for all applications. However, in many applications the graph structure is very regular, and the detailed set–up of the graph would be inconvenient for the user and resource–consuming during run time. A large fraction of all parallel applications use process topologies like rings, or two– or higher–dimensional grids or tori. These structures are completely defined by the number of dimensions and the numbers of processes in each coordinate direction. Also, the mapping of grids and tori is generally an easier problem then that of general graphs. Thus, it is desirable to do the mapping as grids and tori rather than obscuring the information in a graph.

Therefore, in addition to the general graph cartesian structures of arbitrary dimension are supported. The topology definition function MPI_CART can be used to describe both grids and tori by specifying per coordinate direction whether the process structure is periodic or not. Similar functions are contained in EXPRESS [15] and PARMACS. Note that an n–dimensional hypercube is an n–dimensional torus with 2 processes per coordinate direction. Thus, special support for hypercube structures is not necessary.

Based on the positions in the virtual topology a process can look up the PID numbers of other processes by calling MPI_INQPID. The processes are specified by their logical node numbers (graph) or rectangular process coordinates (grid/torus). Thus, it is possible to inquire for the PID numbers of only those (few) processes with which messages have to be exchanged.

Discussion: In the PT2PT discussion it is not yet decided how processes are addressed. This could, for example, be done by unique identifiers, or by their ranks in a group. The topology handling will be adapted to whatever decision will be taken in PT2PT.

4.3 Embedding in MPI

The support for virtual topologies has to be fully integrated in the framework of MPI. In particular, the virtual topology functions do not interfere with the functions for process group creation and with those for collective communication. The current design of the process topology functions is based on the following assumptions:

- MPI is free to choose any process in a group for any location in the topology. There are no special processes which have to be aligned with topology substructures. This assumption fits best to a "local SPMD model" in which all processes in the same group execute the same program.
- One important feature of MPI are the process groups. The treatment of virtual topologies has to take this concept into account. The functions presented in this paper are not only compatible with the process groups, they even take advantage of them for the definition of mapping overlays and dynamic remapping during execution time. Virtual process topologies help in the subdivision of process groups which in many applications has to be done along process coordinate lines. Furthermore, the definition of collective gather–type communication calls benefits from the natural ordering of processes in a logical topology.

The virtual topology handling follows the general MPI rule of functional layering. At the bottom it is based on the general process group creation function MPI_CREATE_GROUP plus a few topology specific mapping functions. In principle this is sufficient for defining topologies and use them in applications.

A number of higher–level functions provide an interface which is similar to what users of packages like Express and PARMACS are familiar with. This level can be written in a portable way in terms of the low–level functions. Implementations of the high–level functions which bypass the low level may be advantageous on some systems in terms of memory and cpu time usage, but this is not a requirement.

4.4 Overview of the proposed MPI functions

This chapter assumes the availability of the following MPI features which are defined in other chapters:

1. An existing process group (e.g., the ALL group) can be partitioned and/or renumbered using

   MPI_CREATE_GROUP(gid_old, gid_new, group_key, order_key) with

   IN gid_old group id of parent group,

   OUT gid_new group id of newly created subgroup of which the calling process is a member,

   IN group_key integer value used for the partitioning,

IN order_key integer value used for determining the ranks in the new group (see below),

This function is called by all processes in group *gid_old* loosely synchronously. It creates as many different subgroups as there are different values of *group_key*. The ranking in the newly created groups is determined by the *order_key* values, with the process ranks in the parent group for tie breaking.

2. There is an internal information cacheing facility, which can be used by MPI routines to store and retrieve information related to a process group. This facility was proposed in the Context Subcommittee. High–level topology functions use this possibility to store structure information for a process group, which is then later used by topology inquiry functions. Also, special shift functions which are based on the topological structure can take advantage of this information.

Discussion: If the information cacheing will not be part of MPI, passing the topology information for a group from one topology function to the other can be handled in several other ways which will be discussed later.

There has been a general decision in MPI to start all numbering with 0 instead of 1. Thus, coordinates in a cartesian structure are 0,1, ..., dim-1. In the same way, nodes in a general graph are numbered zero up.

4.4.1 Low–level topology functions

On the lowest level groups are created by the general function MPI_CREATE_GROUP. Using the input argument *newrank* the caller can change the process ranks in the new group as compared to the parent group. This feature can be used to efficiently map the processes in the group to the positions in the topology. For each class of topologies, therefore, there is a function which computes the rank of the calling process in the new group, based on the size and structure of the topology. Note that it is a valid implementation if the new ranks are identical to the old ranks. So, if for a given machine the placement of processes makes no difference, the implementation of the rank computation functions is trivial.

For cartesian structures the function is

newrank = MPI_MAP_CART(gid_old, ndims, dims, period) with

IN gid_old group id of parent group,

IN ndims number of dimensions of cartesian structure,

IN dims(*) number of processes in each coordinate direction,

IN period(*) periodicity specification in each coordinate direction.

The return value is the reordered rank of the calling process.

The corresponding function for general graph structures is

newrank = MPI_MAP_GRAPH(gid_old, nodes, index, nglist) with

IN gid_old group id of parent group,

IN nodes number of graph nodes,

IN index(*), nglist(*) definition of graph structure: For each graph node its set of neighbors is entered, i. e., the set of processes with which the node process is going to communicate most frequently. Since there is no weighting of node connections, it might be sensible to specify only those neighbors with which there is a considerable amount of communication.

For each node i $(0 \leq i \leq nodes - 1)$ the list of the neighbor nodes is stored at $nglist(j)$, $index(i) \leq j \leq index(i+1) - 1$. For the last process $index(i+1)$ points to the first free position in vector nglist. Note that the matrix is symmetric, i. e., if node i is a neighbor of node j, than node j is also a neighbor of node i. This property can be used for an internal checking of the input arguments.

Again, the function returns the reordered rank of calling process.

The definitions of the arguments `nodes`, `index`, and `nglist` are illustrated with the following simple example:

Assume there are four processes 0, 1, 2, 3 with the following adjacency matrix:

process	neighbors
0	1, 3
1	0
2	3
3	0, 2

Then, the input arguments are:

nodes =	4
index =	0, 2, 3, 4, 6
nglist =	1, 3, 0, 3, 0, 2

If the topology contains fewer processes than there are processes in the parent group, the newrank values for the remaining processes are set to MPI_NORANK. It is an error if the topology defines more processes than are contained in the parent group.

4.4.2 High–level topology functions

In principle, all that is needed to create an MPI process group with topological ordering are the two mapping functions and the general group creation routine. However, this is not what the applications programmer expects to see on his level of abstraction. Therefore, another layer of functions is defined on top of the low–level functions. This higher level can be portably implemented in terms of the level below. Of course MPI implementors are free to bypass the lower level for the sake of better performance or resource usage.

For cartesian and graph structures the functions MPI_CART and MPI_GRAPH create a process group with topological structure. They combine the call to the mapping function with the call to the group creation routine. The user can decide whether the ranking in the new group may be reordered (for a more efficient mapping) or not. In the latter case the rank of the calling process is copied from the parent group.

The option of not reordering the process ranks can be useful in the case of an overlay of different topologies. The user can then fully control which location in the first topology coincides with which location in the second one. The price which one obviously has to pay in this case is that for only one topology the process arrangement can be optimized.

As specified for the low–level functions it is an error if one attempts to define a topology with more locations than there are processes in the parent process group. If it does not use all available processes, the unused process will return MPI_NORANK as their rank in the new group.

MPI_CART is defined as

newrank = MPI_CART(gid_old, gid_cart, ndims, dims, period, reorder) with

IN gid_old group id of parent group,

OUT gid_cart group id of cartesian group (if calling process is part of it, else id of rest group),

IN ndims, dims(*), period(*) definition of cartesian structure, (for details see the definition of MPI_MAP_CART)

IN reorder ranking may be reordered or not.

The return value is the rank of the calling process in the cartesian group (or MPI_NORANK).

Apart from some local computations, MPI_CART is based on the two function calls

newrank = MPI_MAP_CART (gid_old, ndims, dims, period)

and

MPI_CREATE_GROUP (gid_old, gid_cart, 0, newrank) .

For graph structures the corresponding routine is

newrank = MPI_GRAPH(gid_old, gid_graph, nodes, index, nglist, reorder) with

IN gid_old group id of parent group,

OUT gid_graph group id of graph group (if calling process is part of it, else id of rest group),

IN nodes, index(*), nglist(*) definition of graph structure, (for details see the definition of MPI_MAP_GRAPH)

IN reorder ranking may be reordered or not.

In the context subcommittee a mechanism has been proposed for cacheing information which is associated with a process group. If that feature will be part of MPI, it can be used by the high–level topology functions to store the following information:

1. Type of topology (cartesian/graph)

2. For a cartesian topology:

- ndims (number of dimensions),
- dims(*) (numbers of processes per coordinate direction),
- period(*) (periodicity information),
- own_position(*) (own position in grid, could also be computed from rank and dims(*))

For a graph structure no information has to be stored, because the graph node number is the rank in the graph group, and the size of the graph is the number of processes in that group.

In case the cacheing mechanism does not become part of MPI, there are two other possibilities: since the storage and retrieval of the above information is only done by the topology functions, it can internally be stored with other group information in an implementation–dependent way, which each vendor can do as is most convenient.

As an alternative, the information could be stored in an opaque object, a handle to which would have to be included in the argument lists of all related topology functions. However, this alternative would be much less elegant.

4.4.3 Topology inquiry functions

If a topology has been defined with one of the above high–level functions, the topology information can be looked up using inquiry functions. They all are local calls, and do not require any communication. In the following the assumption is made that the information cacheing is available. Otherwise the handle to the opaque information object has to be added to the argument lists (or replace the group identifier argument).

The function MPI_INQMAP returns the type of mapping which is assigned to a group.

map = MPI_INQMAP(gid).

The return value is one of the following:

- MPI_CARTTOP: cartesian topology,
- MPI_GRAPHTOP: graph topology,
- MPI_NOTOP: no topology.

For a process group with cartesian ordering which has been created by MPI_CART, MPI_INQCART looks up the following info:

dim = MPI_INQCART(gid, maxdim, size, period, pos) with

IN gid group id of group with cartesian structure,

IN maxdim length of vectors size(*), period(*), and pos(*) in the calling program,

OUT size(*) number of processes for each coordinate direction,

OUT period(*) periodicity yes/no for each coordinate direction,

OUT pos(*) position of calling process in cartesian structure.

The return value is the dimension of the cartesian structure.

For a process group with cartesian structure, the function MPI_INQPID translates the logical process coordinates to process ids as they are used by the point–to–point routines:

pid = MPI_INQPID(gid, location) with

IN gid group id of group with cartesian structure,

IN location coordinate vector of process.

In a sense, the inverse operation to MPI_INQPID is provided by the function:

MPI_INQLOC(gid, rank, location, maxdim) with

IN gid group id of group with cartesian structure,

IN rank rank of a process within group `gid`,

OUT location(*) coordinate vector of process,

IN maxdim length of vector `location` in the calling program (only for error checking).

MPI_INQLOC translates a rank in the group into a topology location.

Discussion: Depending on the decisions taken in the PT2PT subcommittee, MPI_INQPID might be changed to return the rank in the group rather than a pid.

4.4.4 Partitioning of cartesian structures

If a group has been created with MPI_CART, the function MPI_PARTC can be used to partition the group into subgroups which form hyperplanes of the original process structure. As an example, assume that MPI_CART has defined a two–dimensional grid structure on a process group. With a call to MPI_PARTC this group can then be split up into disjoint subgroups which represent rows or columns of the original group. The function syntax is:

gid_new = MPI_PARTC(gid, lostdim) with

IN gid group id of group with cartesian structure,

IN lostdim direction perpendicular to resulting hyperplanes.

The function returns the group identifier of the hyperplane group in which the calling process is contained.

MPI_PARTC uses the function MPI_CREATE_GROUP to partition the parent group, using the process coordinate in direction `lostdim` as the key value. After a group has been partitioned by MPI_PARTC, the inquiry functions MPI_INQCART and MPI_INQPID consistently refer to the grid sections if applied to a subgroup.

If the process topology is a cartesian structure, a collective shift operation most likely is used along a coordinate direction. The function MPI_SHIFTGR partitions a cartesian group into one–dimensional subgroups, which are all oriented in one coordinate direction. Every process is contained in exactly one subgroup:

gid_new = MPI_SHIFTGR(gid, direction) with

IN gid group id of group with cartesian structure,

IN direction coordinate direction of one–dimensional subgroups.

MPI_SHIFTGR uses the function MPI_CREATE_GROUP to partition the parent group. The return value is the group identifier of the one–dimensional subgroup to which the calling process belongs. The subgroup has a one–dimensional cartesian structure. It is either periodic (ring) or not (linear), depending on the periodicity of the parent group in the corresponding direction.

Discussion: Instead of creating one–dimensional subgroups and then use the standard collective communication (circular or end–off) shift functions, one could also define a topology based shift function. It would look very much like the standard shift, with an additional argument defining the coordinate direction for the shift. Depending on the periodicity information of the cartesian structure, the shift function could choose the circular or end–off version automatically.

In a portable program the number of available processors usually is not known at compile time. In the case of a cartesian topology the function MPI_MAKDIM helps in setting the numbers of processes per coodinate direction, depending on the number of processes in the group and optional constraints which can be specified by the user:

MPI_MAKDIM(gid, ndims, dims) with

IN gid group id of group which is to be structured,

IN ndims number of cartesian dimensions,

IN/OUT dims number of processes in each coordinate direction.

The processes of the group with ID number `gid` are divided as evenly as possible among `ndims` dimensions, subject to the constraint that all processes in the group are used. The caller may further constrain the operation of this routine by specifying elements of array `dims`. If `dims(i)` is set to a positive number, the routine will fix the number of processes in coordinate direction `i` to `dims(i)`. Processes will then be distributed among directions where `dims(i)` is set to 0, subject to the above constraints.

Note that an error will occur if `dims(i)` is positive and does not evenly divide the number of processes in the group. Negative input values of `dims(i)` will also cause an error. All elements of `dims(*)` must be set to 0 for completely automatic process distribution.

When the function returns, `dims(i)` will be set to the number of processes in coordinate direction `i`. For `dims(i)` not set by the user, `dims(i)` will be ordered (MAXIMUM, ..., MINIMUM). Array `dims(*)` is suitable for use as input to routine MPI_CART.

4.5 A simple example

The example in figure 4.1 shows how the grid definition and inquiry functions can be used in an application program. A partial differential equation, for example the Poisson equation, is to be solved on a rectangular domain. First the processes organize themselves in a two–dimensional structure. Each process then inquires for the PID numbers of its neighbors in the four directions. The numerical problem is solved by an iterative method, the details of which are hidden in the subroutine `relax`.

In each relaxation step each process computes new values for the solution grid function at all points owned by the process. Then the values at interprocess boundaries have to be exchanged with neighboring processes. For example, the exchange subroutine might contain a call like MPI_SENDC(...,neigh_pid(1),...) to send updated values to the left–hand neighbor (i-1,j). Note that there is only local communication in this exchange operation.

```
      integer ndims, mx_neigh
      parameter (ndims=2, mx_neigh=4)
      integer gid, gid_cart, dims(ndims), neigh_def(2), neigh_pid(mx_neigh)
      integer own_position(ndims), period(ndims), i, j, num_neigh, ret, rank
      real*8 u(0:101,0:101), f(0:101,0:101)
      data dims / ndims * 0 /
C     Create a grid structure in ALL group and inquire about own position
      gid = ALL
C     Set process grid size and periodicity
      ret = MPI_MAKDIM(gid, ndims, dims)
      if (ret .lt. 0) call error_handler(ret)
      period(1) = MPI_PER
      period(2) = MPI_PER
      rank = MPI_CART (gid, gid_cart, ndims, dims, period)
      ret = MPI_INQCART (gid_cart, ndims, dims, period, own_position)
C     Look up the PID numbers for the neighbors. Own process coordinates are (i,j).
C     Neighbors are (i-1,j), (i+1,j), (i,j-1), (i,j+1)
      i = own_position(1)
      j = own_position(2)
      neigh_def(1) = i-1
      neigh_def(2) = j
      neigh_pid(1) = MPI_INQPID (gid, neigh_def)
      neigh_def(1) = i+1
      neigh_def(2) = j
      neigh_pid(2) = MPI_INQPID (gid, neigh_def)
      neigh_def(1) = i
      neigh_def(2) = j-1
      neigh_pid(3) = MPI_INQPID (gid, neigh_def)
      neigh_def(1) = i
      neigh_def(2) = j+1
      neigh_pid(4) = MPI_INQPID (gid, neigh_def)
      num_neigh = mx_neigh
C     Initialize the grid functions and start the iteration
      call init (u, f)
      do 10 it=1,100
         call relax (u, f)
C     Exchange data with neighbor processes
         call exchange (u, neigh_pid)
10    continue
      call output (u)
      end
```

Figure 4.1: Set–up of process structure for two–dimensional parallel Poisson solver.

Chapter 5

Language Binding

Chapter 6

Formal Semantics

Chapter 7

Environmental Inquiry

Chapter 8

Profiling

8.1 Requirements

To meet the MPI profiling interface, an implementation of the MPI functions *must*

1. provide a mechanism through which all of the MPI defined functions may be accessed with a name shift. Thus all of the MPI functions (which normally start with the prefix "`MPI_`") should also be accessable with the prefix "`PMPI_`".

2. ensure that those MPI functions which are not replaced may still be linked into an executable image without causing name clashes.

3. document the implementation of different language bindings of the MPI interface if they are layered on top of each other, so that the profiler developer knows whether she must implement the profile interface for each binding, or can economise by implementing it only for the lowest level routines.

8.2 Discussion

The objective of the MPI profiling interface is to ensure that it is relatively easy for authors of profiling (and other similar) tools to interface their codes to MPI implementations on different machines.

Since MPI is a machine independent standard with many different implementations, it is unreasonable to expect that the authors of profiling tools for MPI will have access to the source code which implements MPI on any particular machine. It is therefore necessary to provide a mechanism by which the implementors of such tools can collect whatever performance information they wish *without* access to the underlying implementation.

We believe that having such an interface is important if MPI is to be attractive to end users, since the availability of many different tools will be a significant factor in attracting users to the MPI standard.

The profiling interface is just that, an interface. It says *nothing* about the way in which it is used. There is therefore no attempt to lay down what information is collected through the interface, or how the collected information is saved, filtered, or displayed.

While the initial impetus for the development of this interface arose from the desire to permit the implementation of profiling tools, it is clear that an interface like that specified may also prove useful for other purposes, such as "internetworking" multiple MPI implementations. Since all that is defined is an interface, there is no objection to its being used wherever it is useful.

As the issues being addressed here are intimately tied up with the way in which executable images are built, which may differ greatly on different machines, the examples given below should be treated solely as one way of implementing the objective of the MPI profiling interface. The actual requirements made of an implementation are those detailed in the Requirements section above, the whole of the rest of this chapter is only present as justification and discussion of the logic for those requirements.

The examples below show one way in which an implementation could be constructed to meet the requirements on a Unix system (there are doubtless others which would be equally valid).

8.3 Logic of the design

Provided that an MPI implementation meets the requirements above, it is possible for the implementor of the profiling system to intercept all of the MPI calls which are made by the user program. She can then collect whatever information she requires before calling the underlying MPI implementation (through its name shifted entry points) to achieve the desired effects.

8.4 Examples

8.4.1 Profiler implementation

Suppose that the profiler wishes to accumulate the total amount of data sent by the MPI_SEND function, along with the total elapsed time spent in the function. This could trivially be achieved thus :–

WARNING : syntax correction needed in examples when MPI syntax is defined.

```
static int totalBytes;
static int totalTime;
extern int now(void);   /* Some suitable timer function ! */

int MPI_SEND(const int tag, const int bytes, void * buffer, const int who)
{
   int tstart = now();
   int result = PMPI_SEND(tag,bytes,buffer,who);   /* Pass on all the arguments */

   totalBytes += bytes;                /* Accumulate byte count */
   totalTime  += now() - tstart;       /* and time              */

   return result;
}
```

8.4.2 MPI library implementation

On a Unix system, in which the MPI library is implemented in C, then there are various possible options, of which two of the most obvious are presented here. Which is better depends on whether the linker and compiler support weak symbols.

Systems with weak symbols

If the compiler and linker support weak external symbols (e.g. Solaris 2.x, other system V.4 machines), then only a single library is required through the use of `#pragma weak` thus :–

```
#pragma weak MPI_EXAMPLE = PMPI_EXAMPLE

int PMPI_EXAMPLE(/* appropriate args */)
{
    /* Useful content */
}
```

The effect of this `#pragma` is to define the external symbol `MPI_EXAMPLE` as a weak definition, this means that the linker will not complain if there is another definition of the symbol (for instance in the profiling library), however if no other definition exists, then the linker will use the weak definition.

Systems without weak symbols

In the absence of weak symbols then one possible solution would be to use the C macro pre-processor thus :–

```
#ifdef PROFILELIB
#    ifdef __STDC__
#        define FUNCTION(name) P##name
#    else
#        define FUNCTION(name) P/**/name
#    endif
#else
#    define FUNCTION(name) name
#endif
```

Each of the user visible functions in the library would then be declared thus

```
int FUNCTION(MPI_EXAMPLE)(/* appropriate args */)
{
    /* Useful content */
}
```

The same source file can then be compiled to produce both versions of the library, depending on the state of the `PROFILELIB` macro symbol.

It is required that the standard MPI library be built in such a way that the inclusion of MPI functions can be achieved one at a time. This is a somewhat unpleasant requirement, since it may mean that each external function has to be compiled from a separate file. However this is necessary so that the author of the profiling library need only define those MPI functions which she wishes to intercept, references to any others being fulfilled by the normal MPI library. Therefore the link step can look something like this :–

```
% cc ... -lmyprof -lpmpi -lmpi
```

Here `libmyprof.a` contains the profiler functions which intercept some of the MPI functions. `libpmpi.a` contains the "name shifted" MPI functions, and `libmpi.a` contains the normal definitions of the MPI functions.

8.5 Multiple levels of interception

The scheme given here does not directly support the nesting of profiling functions, since it provides only a single alternative name for each MPI function. Consideration was given to an implementation which would allow multiple levels of call interception, however we were unable to construct an implementation of this which did not have the following disadvantages

- assuming a particular implementation language.
- imposing a run time cost even when no profiling was taking place.

Since one of the objectives of MPI is to permit efficient, low latency implementations, and it is not the business of a standard to require a particular implementation language, we decided to accept the scheme outlined above.

Note, however, that it is possible to use the scheme above to implement a multi-level system, since the function called by the user may call many different profiling functions before calling the underlying MPI function.

Unfortunately such an implementation may require more co-operation between the different profiling libraries than is required for the single level implementation detailed above.

Chapter 9

Contexts – Proposal I

Marc Snir

9.1 Contexts

A **communication context** (for short, **context**) is a mechanism for the modularization of MPI communication. Any MPI communication occurs within a context, and does not interfere with communication executed within another context. Furthermore, a context specifies a local name space for processes that communicate in this context. The processes that participate in a context are associated with a **rank**, which ranges from 0 to $n-1$, where n is the number of processes that participate in the context. This rank is used in interprocess communication as the local address of the process within that context. Thus, communication within a context is unaffected by communication outside that context.

A process may communicate simultaneously in several contexts. The context of a communication is explicitly stated as a parameter of the communication call.

A process that participates in a communication context accesses this context using a *context handle* (i.e., a handle to an opaque object that identifies a context). This handle can be used to

- Find information about this context, such as the number of processes that participate in the context, or the rank of the calling process within the context.
- Communicate with other processes that participate in the context; these processes are addressed using their context rank.
- Create new contexts.

Context handles cannot be transferred for one process to another; they can be used only on the process where they were created. Thus, "knowledge" about a context exists only locally, at the processes that participate in that context. Operations within a communication contexts (including the generation of new subcontexts) do not require communication with processes that do not participate in that context.

Follows examples of possible uses for contexts.

9.1.1 Loosely synchronous library call interface

Consider the case where a parallel application executes a "parallel call" to a library routine, i.e., where all processes transfer control to the library routine. If the library was developed separately, then one should beware of the possibility that the library code may receive by mistake messages send by the caller code, and vice-versa. The problem is solved by allocating a different context to the library, thus preventing unwanted interference.

9.1.2 Functional decomposition and modular code development

Often, a parallel application is developed by integrating several distinct functional modules, that is each developed separately. Each module is a parallel program that runs on a dedicated set of processes, and the computation consists of phases where modules compute separately, intermixed with global phases where all processes communicate. It is convenient to allow each module to use its own private process numbering scheme, for the intramodule computation. This is achieved by using a private module context for intramodule computation, and a global context for intermodule communication.

9.1.3 Collective communication

MPI supports collective communication within dynamically created groups of processes. Each such group can be represented by a distinct communication context. This provides a simple mechanism to ensure that communication that pertains to collective communication within one group is not confused with collective communication within another group, and avoids the introduction of two different mechanisms with similar functionality.

9.1.4 Lightweight gang scheduling

Consider an environment where processes are multithtreaded. Contexts can be used to provide a mechanism whereby all processes are time-shared between several parallel executions, and can context switch from one parallel execution to another, in a loosely synchronous manner. A thread is allocated on each process to each parallel execution, and a different context is used to identify each parallel execution. Thus, traffic from one execution cannot be confused with traffic from another execution. The blocking and unblocking of threads due to communication events provide a "lazy" context switching mechanism. This can be extended to the case where the parallel executions are spanning distinct process subsets. (MPI does not require multithreaded processes.)

9.2 Basic Context Operations

A global context **MPI_ALL** is predefined. All processes participate in this context when computation starts. MPI does not specify how processes are initially ranked within the context `MPI_ALL`. It is expected that the start-up procedure used to initiate an MPI program (at load-time or run-time) will provide information or control on this initial ranking (e.g., by specifying that processes are ranked according to their pid's, or according to the physical addresses of the executing processors, or according to a numbering scheme specified at load time).

Discussion: If we think of adding new processes at run-time, then `MPI_ALL` conveys the wrong impression, since it is just the initial set of processes.

The following operations are available for creating new contexts.

MPI_COPY_CONTEXT(newcontext, context)

Create a new context that includes all processes in the old context. The rank of the processes in the previous context is preserved. The call must be executed by all processes in the old context. It is a blocking call: No call returns until all processes have called the function. The parameters are

OUT newcontext handle to newly created context. The handle should not be associated with an object before the call.

IN context handle to old context

MPI_NEW_CONTEXT(newcontext, context, array_of_ranks, size)

OUT newcontext handle to newly created context at calling process. This handle should not be associated with an object before the call.

IN context handle to old context

IN array_of_ranks ranks in the old context of the processes that join the new context

IN size size of new context (integer)

A new context is created for the processes in the old context that are listed in the array. The processes are listed according to their rank in the old context. The rank of the processes in the new context is determined by their place in the list.

The call has to be executed by all processes listed in the array; all make the call with the same list of parameters. Processes in the old context that do not belong to the new context need not make the call. The call is blocking; no process returns from the call until all processes have executed the call.

MPI_SPLIT_CONTEXT(newcontext, context, key, index)

OUT newcontext handle to newly created context at calling process. This handle should not be associated with an object before the call.

IN context handle to old context

IN key integer

IN index integer

A new context is created for each distinct value of `key`; this context is shared by all processes that made the call with this key value. Within each new context the processes are ranked according to the order of the `index` values they provided; in case of ties, processes are ranked according to their rank in the old context.

This call is blocking: No call returns until all processes in the old context executed the call.

Particular uses of this function are:

(i) Reordering processes: All processes provide the same `key` value, and provide their index in the new order.

(ii) Splitting a context into subcontexts, while preserving the old relative order among processes: All processes provide the same `index` value, and provide a key identifying their new subcontext.

`MPI_COPY_CONTEXT` is a particular case of `MPI_SPLIT_CONTEXT`, when all processes provide the same key and index parameter.

MPI_RANK(rank, context)

OUT rank integer

IN context context handle

Return the rank of the calling process within the specified context.

MPI_SIZE(size, context)

OUT size integer

IN context context handle

Return the number of processes that belong to the specified context.

A context object is destroyed using the `MPI_FREE` function.

9.2.1 Usage note

Use of contexts for libraries: Each library may provide an initialization routine that is to be called by all processes, and that generate a context for the use of that library. A scheme for allowing each linked library to have its own initialization code would can be used for this purpose (assuming the library will not have several concurrent instantiations).

Use of contexts for functional decomposition: A harness program, running in the context `ALL` generates a subcontext for each module and then starts the submodule within the corresponding context.

Use of contexts for collective communication: A context is created for each group of processes where collective communication is to occur.

Use of contexts for context-switching among several parallel executions: A preamble code is used to generate a different context for each execution; this preamble code needs to use a mutual exclusion protocol to make sure each thread claims the right context.

Implementation note:

We outline here two possible implementations of contexts. They are by no means the only possible ones. In each implementation we assume that a context object is a pointer to a structure that describes the context. A component of this structure is a table of the processes that participate in the context, ordered by rank. We assume that the number of concurrently active contexts at each process is relatively small; say 16-32. In either implementation one might have disjoint message queues for each context, or have shared queues, with the right mechanisms for context matching and buffer allocation.

Proposal 1: Large context tags.

In this implementation we use large context tags (say 32 bits), so that matching of an incoming tag with a local process requires to perform a search in a hash table or another similar search structure (this occurs whenever a message is received). All messages sent within a context carry the same tag value. We use as context tag the pid of the lowest numbered process in the context (let's call it the context leader), concatenated with a counter that is incremented whenever this process allocates a new tag. This guarantee a unique tag for each group (special code needed for counter wraparound).

`MPI_COPY_CONTEXT` - A new tag is generated by incrementing the old tag by one; one can either copy the old context table, or create a new pointer to the old table. A global barrier synchronization is needed to make sure the call is blocking.

`MPI_SPLIT_CONTEXT` - A naive implementation is to have an all-to-all communication where each process gathers the `(key, index)` pairs of all processes in the old group. Each process can determine whether it is the leader of a new group and broadcast the new group tag to all members (using point to point communication in the old group, or using another all-to-all communication). Algorithmic minds will think of many possible optimizations.

`MPI_NEW_CONTEXT` - The lowest numbered process in the list broadcast the new context tag to all processes in the list (using point to point communication operations). A more robust implementation may entail the broadcast of the member list, for error checking.

`MPI_RANK, MPI_SIZE` - require local access to the context object

Proposal 2: Small context tags.

In that implementation the number of distinct context tag values is equal to the maximal number of contexts that can be active at the same node. Thus, the context tag of an incoming message can be used to index directly into a context table, avoiding the need for a search in a hash table. Each process has a unique context tag for incoming communication within this context. However, different context tag values may be used by different processes for the same context (this is necessary in order to densely populate the context tag range). The context table carries, for each member of the context, the context tag to be used when sending messages to it.

`MPI_COPY_CONTEXT` - A new tag is generated by each process for the new context, and broadcast to all other members of the context, using an all-to-all communication. A new context table is created, with these new tags.

`MPI_SPLIT_CONTEXT` - A naive implementation is to have an all-to-all communication where each process gathers the `(key, index, new_context_tag)` triples of all processes in the old group. Each process then creates a new context table for the processes that provided the same key value as it.

`MPI_NEW_CONTEXT` - Similar to `MPI_COPY_CONTEXT`.

`MPI_RANK, MPI_SIZE` - require local access to the context object

9.3 Advanced context operations

Additional functions are required to support a less static model, where processes may be created or deleted during execution. This requires contexts to grow or be merged. This section outlines possible mechanisms for such extension. The lack of interupt driven communication mechanisms in MPI restricts the functionality of such mechanisms: A new process can join an existing context only if all processes that participate in the old context execute a call to add this new process.

A predefined local context `MPI_ME` where only one process participates, is predefined for each process.

MPI_SPAWN(newcontext, oldcontext, array_of_environments, len)

Spawn new processes and create a new context that includes the processes in the old context, followed by the newly spawned processes. The array provides environment parameters for each newly generated process. The form of the array entries are implementation dependent – they may include information on the processor that is to run the new tasks, `argv, argc` arguments, etc.

OUT newcontext handle to new context

OUT oldcontext handle to old context

IN array_of_environments list of environments for new processes

IN len number of new processes (integer)

As particular cases, `MPI_SPAWN( newcontext, MPI_ME,...)` allows one process to spawn new processes, and create a new context that consists of the spawning process, followed by the spawned processes; `MPI_SPAWN( newcontext, MPI_ALL,...)` allows to add to create a new "universal" context that contains all previous processes and the newly spawned processes.

MPI_BCAST_CONTEXT(newcontext, context, root, array_of_ranks, count)

Behaves like `MPI_NEW_CONTEXT`, except that only one process in the old context, namely the process with rank `root`, has to compute the list of the ranks of the processes that participate in the new context. The new context does not necessarily include the process `root`. The call is blocking, and it has to be executed by all processes in the list and by the root process.

OUT newcontext handle to newly created context at calling process. This handle should not be associated with an object before the call.

IN context handle to old context

IN root index of new context creator in old context

IN array_of_ranks ranks in old contexts of members of new context; significant only at root

IN count size of new context; significant only at root

Implementation note:
The call can be implemented as a broadcast from the root to all new context participants, followed by the code of `MPI_NEW_CONTEXT` (the broadcast is executed using point-to-point communication).

Example: Suppose one has a server context `SERVER`, and a client context `CLIENT`. Processes in either contexts may not know about processes in the other context, but they all belong to context `MPI_ALL`, and all know about process `root` (the "nameserver"). The root process knows about each of the two contexts. A call to `MPI_BCAST_CONTEXT` can be used to create a new context that is the union of the `CLIENT` and `SERVER` contexts, so as to allow these to communicate.

Discussion:
Possible extension: the root has two lists; the list of processes that call to join the new context, and the list of processes that are to joint the next context. The remianing processes are returned a value indicating they were left out. This could be useful for master-slave code. The root is the master that creates new contexts for the parallel executuin of new tasks. All disponible processes execute the call, but only a subset is allocated.

MPI_MERGE_CONTEXT(newcontext, context1, context2, root)

This function allows to merge two contexts even when there is no context that encompasses the two merged contexts; it is only required that one process, namely `root`, be in both contexts to be merged. The call is blocking and must be executed by all processes that participate in the merged context. Each process provides its old context, and the index of process root within this old context (the index may be different in the two contexts that are merged). The root provides handles to both contexts that are to be merged. The processes in `context1` precede the processes in `context2` in the ranking of the new context.

OUT newcontext handle to newly created context at calling process. This handle should not be associated with an object before the call.

IN context1 handle to old context

IN context2 handle to second context; significant only at root

IN root index of root in old context

Example: a server (the "root") and a set of client processes participate in `context1`. The server spawns new "helpers", or has other available server processes join it. A call to `MPI_MERGE_CONTEXT` can be used to create a new context that make these new servers available to the parallel clinet.

Discussion:
Here, too, the function can be extended to allow the creation of a new context where only a subset of the processes in the old group participate
If `MPI_MERGE_CONTEXT` is available then the functionality of `MPI_SPAWN` can be reduced: It is suffcient to create a new context that consists of a spawning process (rather than spwning context) and the spawned processes. This new context can then be merged with a preexisting context.

Implementation note:
Implementation is similar to the implementation of `MPI_BCAST_CONTEXT`

9.4 Examples – Problems in Rick's list

Only the basic context functions are used in these examples.
Warning: late night work

Function that implements shuffle permutation in group/context

```
void function mpi_shuffle(inbuf, outbuf, datatype, count, context, size)

...

{
mpi_copy_context(newcontext, context);
mpi_rank(context, i);
dest = shuffle(i, size);
source = unshuffle(i, size);
mpi_irecvc (handle, inbuf, count, datatype, source, 0, nexcontext);
mpi_sendc (outbuf, count, datatype, dest, 0, newcontext);
mpi_wait(handle, ret_stat);
mpi_free(newcontext);
}
```

This works OK, if processes are single-threaded. If they are multithreaded, one has to make sure that other concurrent threads do not create new contexts. New context creation can be skipped if there are no pending messages in the context when mpi_shuffle is called.

Use of contexts for library development The code of a parallel library function should be written so that each message that is produced by some process is consumed by another process during the execution of the library code – i.e. the library should "clean its garbage".

A parallel library is invoked collectively by a group of processes. Any invokation of the library occurs within a context defined by the user. All processes that participate in such context invoke the parallel library, and matching invokations occur at all of them in the same order.

The library has to generate a new context, using the function `MPI_COPY_CONTEXT(newcontext, context)`, for each context `context` where from the library may be invoked. The library will then use the context `newcontext` for its communication whenever it is invoked within the context `context`. If the library uses collective communication within dynamically defined subgroups, then these subgroups will be created by splitting the group of processes defined by `newcontext`.

In the general case, the context of the invokation has to be passed as an explicit parameter when the library is invoked. The library code will generate a new context when it starts executing and will free this context when it terminates.

There are several special cases where dynamic context creation can be avoided.

Case 1: The library is always invoked within the context `MPI_ALL`. Then the new context can be "cached": A library collective initialization routine should be invoked by the user at the start of the program; this routine creates a copy of the context `MPI_ALL` and stores a handle to it in a C static variable (Fortran 77 COMMON). The context `MPI_ALL` need not be passed as a parameter when the library is invoked.

Case 2: The library is always invoked in a unique *library calling context* on each process. Then a library initialization routine should be invoked by the user on each process where the library may be invoked; the initialization routine is called after the user defined the library calling contexts and before the library is invoked. Each initialization call is a collective call within the library calling context. A copy of the calling context is created (by MPI_COPY_CONTEXT) and stored in a static library variable. Subsequent calls to the library need not pass the calling context as a parameter.

Case 3: The library may be invoked on each process within a fixed (small) number of library calling contexts. Copies of these contexts can be created before the library is invoked and "cached" in static variables. Subsequent invokations to the library need not create new copies, but only select the right preexisting copy.

This assumes one can test contexts (context handles) for equality. This should be said explicitely in the draft

Use of contexts for a host node computation model Let's assume two contexts:

- MPI_ALL with host being node zero
- MPI_NODE that does not include the host

We assume the load procedure ensures that host is process zero in ALL

```
host/node code                      translation
--------------                      -------------

I_am_the_host()                     (mpi_rank(MPI_ALL, rank);
                                    rank==0;)

form node group                     mpi_rank(MPI_ALL, task)
                                    mpi_split_context(MPI_NODE, MPI_ALL,
                                      (task==0),0);

Broadcast from host to nodes        mpi_bcast(MPI_ALL,0,...)

regular communications btwn         mpi_send(buffer,len,dest,tag,MPI_NODE)
nodes                               mpi_recv(buffer,len,source,tag,MPI_NODE)
     /* source, dest are ranging from 0 to #nodes-1 */

sum values from each node           mpi_reduce(inbuf,outbuf,len,MPI_ALL,
at host                                         0,MPI_ISUM)
       /* host (node zero) calls mpi_reduce with inbuf = 0 */
```

Chapter 10

Contexts – Proposal VII

Lyndon J Clarke & Rik J Littlefield

10.1 Introduction

This chapter is similar in basic principles to Proposal I and includes all of the functionality of that proposal as a subset — it extends in several ways and differs in some details. Certain features of other, now defunct, proposals discussed in the context subcommittee are included. In particular, this chapter proposes that:

1. Contexts and groups are not identical. A context is always associated with one group, but a group may have several contexts. Properties of groups are inherited by all of the associated contexts, for example process rank.

2. Context and group descriptors can be explicitly transferred to processes that are not members of the context or group.

3. In point-to-point messages, processes can be identified in any of three ways: by process, by rank in a shared context, or by ranks in separate sender and receiver contexts.

4. A "cache" facility is provided that allows modules to attach arbitrary information to both contexts and groups.

These extensions are somewhat independent of each other. The first reflects the observation that multiple modules often operate within each process group, so that context formation should be lighter weight than group formation. The second and third together provide expressive support for communication between modules within different groups of processes. The fourth allows modules to be significantly faster in common cases, without complicating their interface to the application.

Much of this proposal must be viewed as recommendations to other subcommittees of MPI, primarily the point-to-point communication subcommittee and the collective communications subcommittee. Concrete syntax is given in the style of the ANSI C host language, only for purposes of discussion.

10.2 Processes

This proposal views processes in the familiar way, as one thinks of processes in Unix or NX for example. Each process is a distinct space of instructions and data. Each process is allowed to compose multiple concurrent threads and MPI does not distinguish such threads.

Process Identifier

Each process is identified by a process-local *process handle*, which is a reference to a *process descriptor* of undefined size and opaque structure. In a static process model process handles can be obtained by mapping from a group (or context) and rank. In a future extension for dynamic processes, handles may be returned by process creation functions.

MPI provides a procedure which returns a handle for the calling process.

```
process = mpi_my_process()
```

Process Creation & Destruction

This proposal makes no statements regarding creation and destruction of processes.

MPI provides facilities for descriptor transmission allowing the user to explicitly transfer a process decriptor from one process to another. These facilities are described below.

10.3 Process Groups

This proposal views a process group as an ordered collection of (references to) distinct processes, the membership and ordering of which does not change over the lifetime of the group. The canonical representation of a group is a one-to-one map from the integers $(0, 1, \ldots, N-1)$ to handles of the N processes composing the group.

There may be structure associated with a process group defined by a process topology. This proposal makes no further statements regarding such structures.

Group Identifier

Each group is identified by a process-local *group handle*, which is a reference to a *group descriptor* of undefined size and opaque structure.

The initialization of MPI makes each process a member of the "initial" group. MPI provides a procedure that returns a handle to this group.

```
group = mpi_initial_group()
```

MPI provides facilities for descriptor transmission allowing the user to explicitly transfer a group descriptor from one process to another.

Group Creation and Deletion

MPI provides facilities which allow users to dynamically create and delete process groups. The procedures described here generate groups which are static in membership.

MPI provides a procedure which allows users to create one or more groups which are subsets of existing groups.

```
groupb = mpi_group_partition(groupa, key)
```

This procedure creates one or more new groups `groupb` which are distinct subsets of an existing group `groupa` according to the supplied values of `key`. This procedure is called by and synchronises all members of `groupa`.

MPI provides a procedure which allows users to create a group by permutation of an existing group.

```
groupb = mpi_group_permutation(groupa, rank)
```

This procedure creates one new group with the same membership as **groupa** with a permutation of process ranking, and returns the created group descriptor in **groupb**. It is called by and synchronises all members of **groupa**.

MPI provides a procedure which allows users to create a group by explicit definition of its membership as a list of process handles.

```
group = mpi_group_definition(listofprocess)
```

This procedure creates one new group **group** with membership and ordering described by the process handle list **listofprocess**. It is called by and synchronises all processes identified in **listofprocess**.

MPI provides a procedure which allows users to delete user created groups.

```
mpi_group_deletion(group)
```

This procedure deletes an existing group **group**. It is called by and synchronises all members of **group**.

MPI may provide additional procedures which allow users to construct process groups with a process group topology.

Group Attributes

MPI provides a procedure which accepts a valid group handle and returns the rank of the calling process within the identified group.

```
rank = mpi_group_rank(group)
```

MPI provides a procedure which accepts a valid group handle and returns the number of members, or *size*, of the identified group.

```
size = mpi_group_size(group)
```

MPI provides a procedure which accepts a valid group handle and process order number, or *rank*, and returns the valid process handle to which the supplied rank maps within the identified group.

```
process = mpi_group_process(group, rank)
```

MPI may provide additional procedures which allow users to determine the process group topology attributes.

MPI provides a group descriptor cache facility which allows the user to attach attributes to group descriptors.

10.4 Communication Contexts

This proposal views a communication context as the combination of a process group and a protection mechanism that avoids collision between messages sent to different contexts. The context inherits process ranking from its associated group, referred to as a *frame*. Each process group may be used as a frame for multiple contexts.

Context Identifier

Each context is identified by a process-local *context handle*, which is a reference to a *context descriptor* of undefined size and opaque structure.

The creation of a process group allocates a *base context* which inherits the created group as a frame and can be thought of as an attribute of the created group. MPI provides a procedure which accepts a valid group handle and returns a handle to the base context within the identified group.

```
context = mpi_base_context(group)
```

MPI provides facilities for descriptor transmission allowing the user to explicitly transfer a context descriptor from one process to another.

Context Creation and Deletion

MPI provides facilities which allows user to dynamically create and delete contexts in addition to the base context associated with a process group. Contexts created in this fashion can be thought of as copies of the base context of the process group.

MPI provides a procedure which allows users to create contexts. This procedure accepts the handle of a group of which the calling process is a member, and returns a handle to the new context.

```
context = mpi_context_creation(group).
```

This procedure must be called loosely synchronously by all members of `group`. The procedure may not actually synchronize the member processes — it is suggested that this is a lightweight procedure that can be implemented so as to not require interprocess communication.

MPI provides a procedure which allows users to delete user created contexts. The procedure accepts a context handle that was created by the calling process and deletes the identified context.

```
mpi_context_deletion(context)
```

This procedure has the same synchronization behavior as context creation.

Context Attributes

MPI provides a procedure which allows users to determine the process group that is the frame of a context.

```
group = mpi_context_frame(context)
```

MPI provides a group descriptor cache facility which allows user to attach attributes to group descriptors.

10.5 Descriptor Facilities

This section describes the descriptor transmission and user cache facilities.

Transmission Facility

MPI provides a mechanism whereby the user can transmit a valid descriptor in a message such that the received descriptor handle is valid. This can be integrated with the capability to transmit typed messages, and it is suggested that a notional data type should be introduced for this purpose, e.g. `MPI_DSCR_TYPE`. There are other reasonable approaches to providing this facility.

The descriptor is translated as necessary to be meaningful on the destination process, storage is allocated for it, and a handle to that storage is returned. Decorations are not transmitted.

Handles are guaranteed to be unique within each process — if processes A and B independently send to process C a descriptor for an object D, then process C will get two copies of the same handle. As with all transfers of descriptors, the receiving process is responsible for releasing the descriptor and its handle when it is no longer needed or becomes stale. MPI provides a procedure which frees a descriptor.

```
mpi_free_dscr(handle)
```

A descriptor registry service which allows descriptors to be identified by name would be a useful additional feature. This service can be implemented at the user level using the point-to-point chapter of MPI and the descriptor transmission facilities. These services can be deferred in this session of MPI.

Cache Facility

MPI provides a "cache" facility that allows an application to attach arbitrary pieces of information, called *decorations*, to context and group descriptors. Decorations are local to the process and are not included if the descriptor is sent to another process. This facility is intended to support optimizations such as saving persistent communication handles and recording topology-based decisions by adaptive algorithms.

MPI provides the following services related to cacheing:

Generate key: Generate cache key.

```
keyval = mpi_GetDecorationKey()
```

Store decoration: Store decoration in cache by key.

```
mpi_SetDecoration(handle, keyval, decoration_val,
decoration_destructor_routine)
```

Retrieve decoration: Retrieve decoration from cache by key.

```
mpi_TestDecoration(handle,keyval,decoration)
```

Delete decoration: Delete decoration from cache by key.

```
mpi_DeleteDecoration(handle,keyval)
```

Each decoration consists of a pointer or a value of the same size as a pointer, and would typically be a reference to a larger block of storage managed by the module. As an example, a global operation using cacheing to be more efficient for aall contexts of a group after the first call might look like this:

```
   static int gop_key_assigned = 0; /* 0 only on first entry */ static
MPI_KEY_TYPE gop_key; /* key for this module's stuff */

   efficient_global_op (context, ...)  int context_handle; { struct
gop_stuff_type *gop_stuff; /* whatever we need */ int group_handle =
mpi_context_frame(context_handle);

     if (!gop_key_assigned) /* get a key on first call ever */ {
gop_key_assigned = 1; if ( ! (gop_key = MPI_GetDecorationKey()) ) {
MPI_abort ("Insufficient keys available"); } } if (MPI_TestDecoration
(group_handle,gop_key,&gop_stuff)) { /* This module has executed in
this group before.  We will use the cached information */ } else { /*
This is a group that we have not yet cached anything in.  We will now
do so.  */ gop_stuff = /* malloc a gop_stuff_type */

       /* ... fill in *gop_stuff with whatever we want ... */

       MPI_SetDecoration (group_handle, gop_key, gop_stuff,
gop_stuff_destructor); } /* ... use contents of *gop_stuff to do the
global op ... */ }

    gop_stuff_destructor (gop_stuff) /* called by MPI on group delete
*/ struct gop_stuff_type *gop_stuff; { /* ... free storage pointed to
by gop_stuff ... */ }
```

The cache facility could also be provided for process descriptors, but it is less clear how such provision would be useful. It is suggested that the cache store, retrieve and delete decoration procedures should fail when applied to a process descriptor handle.

10.6 Point-to-Point Communication

This proposal recommends three forms for MPI point-to-point message addressing and selection: null context; closed context; open context. It is further recommended that messages communicated in each form are distinguished such that a `Send` operation of form X cannot match with a `Receive` operation of form Y, requiring that form is embedded into the message envelope.

The three forms are described, followed by considerations of uniform integration of these forms in the point-to-point communication chapter of MPI.

Null Context Form

The *null context* form contains no message context. Message selection and addressing are expressed by

`(process, tag)`

where: `process` is a process handle; `tag` is a message tag.

`Send` supplies the `process` of the receiver. `Receive` supplies the `process` of the sender.

`Receive` can wildcard on `process` by supplying the wildcard descriptor handle value `MPI_WILDCARD`. In this case the receiver may have obtained the process descriptor of the sender, and the null descriptor handle `MPI_NULL` is returned in the relevant point-to-point enquiry procedure.

Closed Context Form

The *closed context* form permits communication between members of the same context. Message selection and addressing are expressed by

`(context, rank, tag)`

where: `context` is a context handle; `rank` is a process rank in the frame of `context`; `tag` is a message tag. The calling process must be a member of the frame of `context`.

`Send` supplies the `context` of the receiver (and sender), and the `rank` of the receiver. `Receive` supplies the `context` of the sender (and receiver), and the rank of the sender. The `(context, rank)` pair in `Send` (`Receive`) is sufficient to determine the process identifier of the receiver (sender).

`Receive` cannot wildcard on `context`. `Receive` can wildcard on `rank` by supplying the wildcard integer `MPI_DONTCARE`. This proposal makes no statement about the provision for wildcard on `tag`.

Open Context Form

The *open context* form permits communication between members of any two contexts. Message selection and addressing are expressed by

`(lcontext, rcontext, rank, tag)`

where: `lcontext` is a context handle; `rcontext` is a context handle; `rank` is a process rank in the frame of `rcontext`; `tag` is a message tag. The calling process must be a member of the frame of `lcontext` and need not be a member of the frame of `rcontext`.

`Send` supplies the context of the sender in `lcontext`, the context of the receiver in `rcontext`, and the `rank` of the receiver in the frame of `rcontext`. `Receive` supplies the context of the receiver in `lcontext`, the context of the sender in `rcontext`, and the `rank` of the sender in the frame of `rcontext`. The `(rcontext, rank)` pair in `Send` (`Receive`) is sufficient to determine the process identifier of the receiver (sender).

`Receive` cannot wildcard on `lcontext`. `Receive` can wildcard on `rcontext` by supplying the wildcard descriptor handle value `MPI_WILDCARD`, in which case it must also wildcard on `rank` since the process descriptor of the sender cannot be determined. In this case the receiver may not have obtained the context descriptor of the sender, and the null descriptor handle `MPI_NULL` is returned in the relevant point-to-point enquiry procedure. `Receive` can wildcard on `rank` by supplying the wildard intger value `MPI_DONTCARE`.

Uniform Integration

The three forms of addressing and selection described have different syntactic frameworks. We can consider integrating these forms into the point-to-point chapter of MPI by defining a further orthogonal axis (as in the multi-level proposal of Gropp & Lusk) which deals with form. This is at the expense of multiplying the number of `Send` and `Receive` procedures by a factor of three, and some further but trivial work with details of the current point-to-point chapter which uniformly assumes a single addressing and selection form.

There are various approaches to unification of the syntactic frameworks which may simplify integration. Two options are now described, each based on retention and extension of the framework of the closed and open contexts forms.

The framework of the open context form could be adopted and extended. The null context form is expressed as `(MPI_NULL, MPI_NULL, process, tag)`, which is a little clumsy. The closed context form is expressed as `(MPI_NULL, context, rank, tag)`, which is marginally inconvenient. The open context form is expressed as `(lcontext, rcontext, rank, tag)`, which is of course natural.

The framework of the closed context form could be adopted and extended. The null context form is expressed as `(MPI_NULL, process, tag)`, which is marginally inconvenient, and requires that descriptor handles are expressed as intgers. The closed context form is expressed as `(context, rank, tag)`, which is of course natural. Expression of the open context form requires a little more work. We can use the `context` field as "shorthand notation" for the `(lcontext, rcontext)` pair at the expense of introducing some trickery. We define a "duplet descriptor" which is formally composed of two references to contexts, and provide a procedure which constructs such a descriptor given two context descriptors. Both `Send` and `Receive` accept a duplet descriptor in `context`, are able to distinguish the duplet descriptor from a singlet descriptor, and treat the duplet as shorthand notation. It is conjectured that using this framework is the best choice for MPI.

10.7 Collective Communication

Symmetric collective communication operations are compliant with the closed context form described above. This proposal recommends that such operations accept a context descriptor which identifies the context (thus frame) in which they are to operate.

MPI does plan to describe symmetric collective communication operations. It is not possible to determine whether this proposal is sufficient to allow implementation of the collective communication chapter of MPI in terms of the point-to-point chapter of MPI without loss of generality, since the collective operations are not yet defined.

Asymmetric collective communication operations, especially those in which sender(s) and receiver(s) are distinct processes, are compliant with the open context form described above. This proposal recommends that such operations accept a pair of context descriptors (a duplet descriptor) which identify the contexts (thus frames) in which they are to operate.

MPI does not plan to describe asymmetric collective communication operations. Such operations are expressive when writing programs beyond the SPMD model, which are composed of communicative functionally distinct process groups. These services can be deferred in this session of `mpi`.

10.8 Conclusion

This chapter presented a proposal for communication contexts and process groups with MPI. In the proposal process groups are created dynamically and are static in membership. Associated with each process group are one or more communication contexts which inherit process ranking.

The recommendations for point-to-point communication are powerful. The proposal provides process addressed communication which occurs within an extended context. The proposal also contains closed context communication addressed in terms of context and rank which protects messages belonging to one context from those belonging to other contexts. The proposal also contains open context communications adressed in terms of sender context, receiver context, and process rank which provides expressive power for intercommunication between modules within different groups.

The proposal is extensible to a number of features which might be included in future sessions of MPI, for example: dynamic processes; dynamic groups; multiple group collective communications.

Chapter 11

Contexts – Proposal III

A. Skjellum *et al.*

11.1 Introduction

This chapter takes a slightly different approach to contexts and groups, than does Proposal VII. It is of roughly equal conceptual "power" as Proposal VII, with some differences. As appropriate, this chapter borrows directly from Proposal VII, by Clarke and Littlefield.

1. Contexts are supported to discriminate between messages in the system. A context is a conceptual extension of the tag space into a system-defined part (not wildcardable), and a totally user-defined part (the traditional 32-bit tag).
2. A context is a lower-level concept than a group, so that contexts not associated with groups are permitted. This permits the user to develop codes that build on the server model, or which build up groups dynamically (not otherwise supported by MPI1).
3. Groups are used to describe cooperative communication in the system. Groups have one or more context of communication associated with them. When created, a group is given a context of communication.
4. Context and group descriptors can be explicitly transferred to processes that are not members of the context or group.
5. In point-to-point messages, processes can be identified in either of two ways: by opaque process identifier, or by rank in a group; in either case, communication scope is within a given context.
6. The cache facility, allowing groups to add additional information (described in Proposal VII) is embraced by this Proposal, with reservations as noted. The possible need to omit this cacheing feature from MPI1 should not invalidate the remainder of Proposal VII from further consideration (severability).

11.2 Processes

This Proposal views processes in the familiar way, as one thinks of processes in Unix or NX for example. Each process is a distinct space of instructions and data. Each process is allowed to compose multiple concurrent threads and MPI does not distinguish such threads. MPI shall be thread-aware, but not thread supporting, so we make every attempt to make thread safe programs possible in defining what follows.

Process Identifier

Each process is identified by an opaque process identifier, which is associable to a *process descriptor* of undefined size and opaque structure, through MPI accessor calls. In a static process model process identifiers can be obtained by mapping from a group and rank. In a future extension for dynamic processes, identifiers may be returned by process creation functions.

MPI provides a procedure that returns an identifier for the calling process.

```
my_process = mpi_my_process()
```

This identifier can be converted to a transmittable form by MPI converter functions, though opaque, it is conceptually a pointer.

Process Creation & Destruction

This proposal makes no statements regarding creation and destruction of processes.

MPI provides facilities for identifier transmission allowing the user explicitly to transfer a process identifier from one process to another. These facilities are described below. MPI also provides means to transfer underlying information about the opaque process descriptor underlying.

11.3 Process Groups

This proposal views a process group as an ordered collection of distinct processes (via process identifiers), the membership and ordering of which does not change over the lifetime of the group. The canonical representation of a group is a one-to-one map from the integers $(0, 1, \ldots, N-1)$ to identifiers of the N processes composing the group.

There may be structure associated with a process group defined by a process topology. This proposal makes no further statements regarding such structures.

There may be non-enumerative ways to construct and manipulate special groups, or for special machine architectures (*e.g.*, cohorts). This proposal makes no further statements about such special groups, other than the desirability of avoiding group-name enumeration, when possible.

Group Identifier

Each group is identified by an opaque *group identifier*, which is a associable to a *group descriptor* of undefined size and opaque structure, through MPI accessor functions.

The initialization of MPI makes each process a member of the “initial” group. MPI provides a procedure that returns an identifier to this group.

```
group_ident = mpi_initial_group()
```

MPI provides facilities for descriptor transmission allowing the user explicitly transfer a group descriptor from one process to another.

Group Creation and Deletion

MPI provides facilities which allow users dynamically to create and delete process groups. The procedures described here generate groups which are static in membership.

MPI provides a procedure that allows users to create one or more groups which are subsets of existing groups.

```
new_group = mpi_group_partition(old_group, key)
```

This procedure creates one or more new groups `new_group` which are distinct subsets of an existing group `old_group` according to the supplied values of `key`. This procedure is called by and synchronises all members of `new_group`. No overlapping is permitted, so that exactly one `new_group` is achieved in each `old_group` process. The new groups have new contexts of communication. The number of new contexts depends on the number of different key values asserted.

MPI provides a procedure which allows users to create a group by permutation of an existing group.

```
    new_group = mpi_group_permutation(old_group, rank)
```

This procedure creates one new group with the same membership as `old_group` with a permutation of process ranking, and returns the created group descriptor in `new_group`. It is called by and synchronises all members of `gha`. The new group has a new context of communication.

MPI provides a procedure that allows users to create a group by explicit definition of its membership as a list of process identifiers.

```
    new_group = mpi_group_definition(array_of_process_ids,length,
context_to_use)
```

This procedure creates one new group `new_group` with membership and ordering described by the process handle list `array_of_process_ids`, of length `length`. If `context_to_use` is specified as the special context `MPI_GET_CONTEXT`, then the system allocates the new context. Else, the system trusts the user to have allocated the context legally (see below). This procedure must be called by and synchronises all processes identified in the list. A further approach to new group definition is as follows

```
    new_group = mpi_group_def_by_leader(leader_id, in_length,
in_array_of_process_ids, context_to_use, out_array_of_process_ids,
out_length)
```

This weaker form requires all future group members to have identified only a leader (specified by `leader_id`). The leader knows the length a names of all participants. This is a synchronization of all participants. Same semantics for `context_to_use` as above.

MPI provides a way formally to duplicate a group, in order to obtain a separate context of communication. This can be achieved by using other operations, but this procedure allows optimization in some implementations (no explicit group copy, for instance).

```
new_group = mpi_group_duplicate(old_group)
```

The only difference between the old and new groups is that there is a new context of communication.

MPI provides a procedure which allows users to delete user created groups.

```
    mpi_group_deletion(group)
```

This procedure deletes an existing group `group`. It is called by and synchronises all members of `group`.

MPI may provide additional procedures which allow users to construct process groups with a process group topology.

Group Attributes/Accessors

MPI provides a procedure that accepts a valid group identifier and returns the rank of the calling process within the identified group.

```
    rank = mpi_group_rank(group)
```

MPI provides a procedure that accepts a valid group identifier and returns the number of members, or *size*, of the identified group.

```
size = mpi_group_size(group)
```

MPI provides a procedure that accepts a valid group identifier *rank-in-group*, and returns the valid process identifier to which the supplied rank maps within the identified group.

```
process_id = mpi_group_process(group, rank)
```

MPI may provide additional procedures which allow users to determine the process group topology attributes.

MPI provides a group descriptor cache facility thath allows the user to attach attributes to group descriptors. See Proposal VII for details.

11.4 Communication Contexts

This proposal views a communication context as a partition of the tag space, which is a protection mechanism that avoids collision between messages sent between processes. Process groups have one or more contexts in *MPI*. Unlike Proposal VII, more contexts are obtained for a group using the above-discussed group creation and replication functions. Replication may only be formal for good implementations.

Context Identifier

Each context is identified by an opaque process identifier. It is conceptually an integer assigned by the system to partition a large tag space into a user-defined and system-controlled subspaces. This stategy provides the minimal level of isolation needed to build large libraries, and is close to practice.

Context Creation and Deletion

MPI provides facilities that allow user dynamically to allocate and free contexts. When contexts are used with groups, these calls are not needed. For more advanced users (such as building your own dynamic groups), these calls will be used. Above, where `context_to_use` appears as an argument, the following call would have been used to secure such a context in advance.

```
    mpi_context_creation(number_of_contexts_wanted,array_of_contexts,
number_of_contexts_provided)
```

This call is called by any process, with no synchronization to other processes.

MPI provides a procedure that allows users to delete user- created contexts. The procedure accepts a context identifier array, containing zero or more contexts created previously in the system.

```
mpi_context_deletion(context_array,length)
```

No synchronization occurs here. The user can do erroneous things by freeing contexts that are still in use.

For general applications, it may be nice to have a name service for contexts (necessary for building dynamic groups and servers, for yourself). Herewith:

```
    mpi_associate_contexts_with_name(string_name,context_array,length)
mpi_disassociate_contexts_with_name(string_name)
mpi_get_contexts_by_name(string_name,max_length,out_length,
context_array
```

As with context generation, the above calls assume a simple reactive, global server, or shared name space mechanism (both achieveable easily in practice).

11.5 Descriptor Facilities

This section describes the descriptor transmission and user cache facilities.

Conversion Facility

MPI provides a mechanism whereby the user can convert a valid descriptor (*e.g.*, a group descriptor) idenfied through an identifier (*e.g.*, a group identifier) for use in a message such that the received descriptor can be reconstructed on the remote end. This can be integrated with message transmission as the user sees fit, without additional complication to the send/receive semantics of MPI. An example follows:

```
    error = mpi_group_group_transmit(group, group_buffer, max_length,
act_length)
```

If the buffer is not long enough to hold the information, an error occurs. A network independent format can be assumed in the `group_buffer`. Cached "attributes" are not transmitted (see below).

Cache Facility

MPI provides a "cache" facility that allows an application to attach arbitrary pieces of information, called *attributes*, to context and group descriptors. Attributes are local to the process and are not included if the descriptor is sent to another process. This facility is intended to support optimizations such as saving persistent communication handles and recording topology-based decisions by adaptive algorithms.

MPI provides the following services related to cacheing. We call our attributes 'attributes'; Proposal VII calls them (equivalently) decorations (no big difference, except naming, is anticipated).

Generate key: Generate cache key.

```
keyval = mpi_get_attribute_key()
```

Store attribute: Store attribute in cache by key.

```
mpi_set_attribute(handle, keyval, attribute_val,
attribute_destructor_routine)
```

Retrieve attribute: Retrieve attribute from cache by key.

```
mpi_Test_Attribute(handle,keyval,attribute)
```

Delete attribute: Delete attribute from cache by key.

```
    mpi_delete_attribute(handle,keyval)
```

Each attribute consists of a pointer or a value of the same size as a pointer, and would typically be a reference to a larger block of storage managed by the module. Our example will appear in a later draft, because we have semantic differences from some of the ancillary aspects of the example of Proposal VII.

The cache facility could also be provided for process identifiers, but it is less clear how such provision would be useful. It is suggested that the cache store, retrieve and delete attribute procedures should fail when applied to a process identifiers.

Implementations should use AVL trees, or similar efficient data structures to provide relatively efficient access to attributes.

11.6 Point-to-Point Communication

This proposal recommends two forms for MPI point-to-point message addressing and selection: by-group notation, by process-ID notation; always, one is working in a context, as this is the fundamental management tactic of MPI for messages. As a group always has a context at creation, and an ALL group is anticipated, this should prove fine for a static process model. We disagree significantly from Proposal VII in what follows.

The two forms are described, followed by considerations of uniform integration of these forms in the point-to-point communication chapter of MPI.

Group-Rank Form

The *group-rank* form permits communication between members of the same context and group. Message selection and addressing are expressed by

```
(group, rank, tag)
```

where: `group` is a group identifier; `rank` is a process rank in that group; `tag` is a message tag. The calling process must be a member of the frame of `context`.

`Send` determines the context using information in the group identifier. It does all necessary mappings to the process identifier space. `Receive` cannot wildcard on context, so a valid matching receive must refer to the same group information. `Receive` can wildcard on `rank` by supplying the wildcard integer `MPI_DONTCARE`. This proposal makes the following statement about the provision for wildcard on `tag`. Two integer-form wildcard is needed for layering: `care_bits`, `dont_care_bits`. Tags are matched if and only if

$$(received\_tag AND NOT dont\_care_bits) XOR care\_bits == 0 \tag{11.1}$$

This general format can be used to partition the tag space for virtual topologies or other user-defined needs, and is quite important to the standard's flexibility.

Process-Identifier Form

Communication takes place using the following parameters:

```
(context, process_identifier, tag)
```

where: `context` is a context identifier, `process_identifier` is a process identifier, `tag` is a message tag. The calling process must be a member of the same context as the recipient. There is no reference to groups here. Contexts can have been shared by using a least common ancestor prior to this call, or by the above-mentioned context naming service.

There is never wildcarding on context. Wildcarding on `process_identifier` is through `MPI_DONTCARE`. Tag wildcarding is through the integer pair described above.

Uniform Integration

The two forms of addressing and selection described have different syntactic frameworks. We can consider integrating these forms into the point-to-point chapter of MPI by defining a further orthogonal axis (as in the multi-level proposal of Gropp & Lusk) which deals with form. This is at the expense of multiplying the number of `Send` and `Receive` procedures by a factor of two, and some further but trivial work with details of the current point-to-point chapter which uniformly assumes a single addressing and selection form. No further details, other than naming that disambiguates the rank-group form from the process-id-context form is really needed, and the naming would seem uncontroversial.

11.7 Collective Communication

Symmetric collective communication operations are compliant with the group-rank form described above. This proposal recommends that such operations accept a group-identifier (which contains context and other information) needed to operate correctly. We recommend that the tag argument be included in collective calls where this could help with debugging.

MPI does plan to describe symmetric collective communication operations. It is impossible to determine whether this proposal is sufficient to allow implementation of the collective communication chapter of MPI in terms of the point-to-point chapter of MPI without loss of generality, since the collective operations are not yet defined.

Asymmetric collective communication operations, especially those in which sender(s) and receiver(s) are distinct processes, should be made compliant with the group-rank form described above.

MPI1 should forego non-blocking collective operations, but ask vendors to support thread models in lieu of such operations.

11.8 Conclusion

This proposal is substantially different than either Proposal VII or I. Contexts are integer tag partitions here, and are fundamentally lower-level objects than groups. Groups have contexts in this proposal, but not all operations require the presence of group-scope. To avoid invidious comparisons here, more substantial comparisons of all three proposals are deferred the following appendix.

Chapter 12

Summary of context subcommittee proposals

The three proposals in the context subcommittee share common features, and have differences both in concept and detail. Two of these proposals contain features which are "separable" and could equally appear as components of one or more other proposals. This summary identifies feature of proposals as: Common Features; Separable Features; Concept Differences; Detail Differences.

Hopefully the summary will: (a) help us to discuss the important differnces between the proposals and make agreements on how we should proceed with respect to those issues; (b) help us to isolate the separable points and make separate agreements on those issues.

I hope that the summary is both accurate and complete. Please make corrections and additions if you discover such. I apologise in advance for my errors, which are surely inevitable.

12.1 Common Features

12.1.1 Process group management

In each proposal groups are created dynamically and have static membership. In each proposal a group can be created as a partition of an existing group and as a permutation of an existing group. In each proposal there is a defined group containing all (or perhaps all initial) processes. Each proposal allows (or suggests) that a group can be created as an explicit list of processes.

12.1.2 Provision for point-to-point communication within group

In each proposal point-to-point communcation of scope closed within a group can be expressed in terms of a reference to a group coupled with a process rank within the group.

12.1.3 Provision for collective communication within group

In each proposal collective communication of scope closed within a group can be expressed in terms of a reference to a group.

12.1.4 Opacity of group and process description

In each proposal the description of groups and processes is opaque. Groups and processes are referred to by a handle like object.

12.1.5 Fields of point-to-point communication

In each proposal point-to-point communication accepts three fields, inclusive of message tag, in addressing and selection.

12.2 Separable Features

12.2.1 Tag usage in point-to-point communication

Proposal III describes tag selection for Receive in a two-integer form. Proposals I and VII say nothing about tag usage.

This feature can be placed in all Proposals I, III and VII.

12.2.2 Tag usage in collective communication

Proposal III suggests that tag should be used as an argument to collective communication where this will assist debugging. Proposals I and VII say nothing about tag usage.

This feature can be placed in all Proposals I, III and VII.

12.2.3 Context or Group cache

Proposal VII describes a "cache" facility associated with contexts and groups. Proposal III describes a similar "cache" facility associated with groups.

This feature can be placed in all Proposals I, III and VII.

12.2.4 Opaque object (descriptor) transmission

Proposal VII suggests that opaque object transmission can be provided by integration with transmission of typed data. Proposal III suggests that opaque transmission is provided by a mechanism for flattening a descriptor into a memory buffer. These are details of different ways of providing the feature.

This feature can be placed in Proposals III and VII. This feature cannot be placed in Proposal I.

12.2.5 Context registry

Proposal III describes a context name registry service. Proposal VII indicates that such a service would be useful.

This feature can be placed in Proposals III and VII. This feature cannot be placed in Proposal I.

12.3 Concept Differences

12.3.1 Concept of CONTEXT and GROUP

In Proposal I CONTEXT and GROUP are identical concepts and are not distinguished.

In Proposal III CONTEXT is a lower degree concept than GROUP. The GROUP concept inherits aspects of the CONTEXT concept.

In Proposal VII CONTEXT is a higher concept than GROUP. The CONTEXT concept inherits aspects of the GROUP concept.

12.3.2 Scope of point-to-point communication

In Proposal I the scope of point-to-point communication is limited to the group. Processes which are members of distinct groups can only communicate through a common ancestor group.

In Proposals III and VII the scope of point-to-point communication is not limited. Processes which are members of distinct groups can communicate without reference to a common ancestor group.

12.3.3 Transmission of group or context

In Proposal I the CONTEXT cannot be transmitted from one process to another.

In Proposals VII and III both CONTEXT and GROUP can be transmitted from one process to another. In Proposal VII PROCESS can alo be transmitted (Proposal III suggests such but makes no specific provision, presumably a small oversight?)

12.4 Detail differences

12.4.1 Manifestation of context

In Proposals I and VII context is an opaque object.

In Proposal III context is an integer.

12.4.2 Deletion of group

In Proposals VII and III groups can be deleted.

In Proposal I there is no provision for group deletion (possibly a small oversight?).

12.4.3 Duplication of group

In Proposals I and III there is explicit provision for duplication of an existing group to form a new (distinct, homomorphic) group.

In Proposal VII there is no such provision as similar funtionality is provided by the context (although the provision for group partition, permutation and definition can be used to create a snapshot copy of a group).

12.4.4 Global shared variables

Proposals I and VII do not require global shared variables.

Proposal III requires a global shared variable (which can be implemented as such or of course in the traditional approach as a global service process.)

12.4.5 Process identifier addressed communication

Proposal I does not make provision for process identifier addressed communication.

Proposal III makes provision for process identifier addressed communication within multiple distinct tag spaces.

Proposal VII makes provision for process identifier addressed communication within a single distinct tag space.

12.4.6 Inter-group communication

Proposal I does not provide inter-group communication as it limits the scope of point-to-point communication to be closed within a group.

Proposal VII provides inter-group communication in a triplet addressing form: sender (receiver) group, receiver (sender) group, sender (receiver) rank.

Proposal III provides inter-group communication as process identifier addressed communication.

Bibliography

[1] V. Bala and S. Kipnis. Process groups: a mechanism for the coordination of and communication among processes in the Venus collective communication library. Technical report, IBM T. J. Watson Research Center, October 1992. Preprint.

[2] V. Bala, S. Kipnis, L. Rudolph, and Marc Snir. Designing efficient, scalable, and portable collective communication libraries. Technical report, IBM T. J. Watson Research Center, October 1992. Preprint.

[3] R. Butler and E. Lusk. User's guide to the P4 programming system. Technical Report TM-ANL–92/17, Argonne National Laboratory, 1992.

[4] S. Chittor and R. J. Enbody. Performance evaluation of mesh–connected wormhole–routed networks for interprocessor communication in multicomputers. In *Proceedings of the 1990 Supercomputing Conference*, pages 647–656, 1990.

[5] S. Chittor and R. J. Enbody. Predicting the effect of mapping on the communication performance of large multicomputers. In *Proceedings of the 1991 International Conference on Parallel Processing, vol. II (Software)*, pages II–1 – II–4, 1991.

[6] J. Dongarra, R. Hempel, A. Hey, and D. Walker. A proposal for a user–level, message–passing interface in a distributed memory environment. Technical Report TM-12231, ORNL, October 1992.

[7] Edinburgh Parallel Computing Centre, University of Edinburgh. *CHIMP Concepts*, June 1991.

[8] Edinburgh Parallel Computing Centre, University of Edinburgh. *CHIMP Version 1.0 Interface*, May 1992.

[9] G. Geist and V. Sunderam. Network based concurrent computing on the PVM system. Technical Report TM-11760, Oak Ridge National Laboratory, June 1991.

[10] G. A. Geist, M. T. Heath, B. W. Peyton, and P. H. Worley. A user's guide to PICL: a portable instrumented communication library. Technical Report TM-11616, Oak Ridge National Laboratory, October 1990.

[11] R. Hempel. The ANL/GMD macros (PARMACS) in fortran for portable parallel programming using the message passing programming model – users' guide and reference manual. Technical report, GMD, Postfach 1316, D-5205 Sankt Augustin 1, Germany, November 1991.

[12] R. Hempel, H.-C. Hoppe, and A. Supalov. Parmacs–6.0 library interface specification. Technical report, GMD, Postfach 1316, D-5205 Sankt Augustin 1, Germany, December 1992.

[13] nCUBE Corporation. *nCUBE 2 Programmers Guide, r2.0*, December 1990.

[14] H. Mühlenbein O. Krämer. Mapping strategies in message–based multiprocessor systems. *Parallel Computing*, 9:213–225, 1989.

[15] Parasoft Corporation. *Express Version 1.0: A Communication Environment for Parallel Computers*, 1988.

[16] Paul Pierce. The NX/2 operating system. In *Proceedings of the Third Conference on Hypercube Concurrent Computers and Applications*, pages 384–390. ACM Press, 1988.

[17] A. Skjellum and A. Leung. Zipcode: a portable multicomputer communication library atop the reactive kernel. In D. W. Walker and Q. F. Stout, editors, *Proceedings of the Fifth Distributed Memory Concurrent Computing Conference*, pages 767–776. IEEE Press, 1990.

[18] A. Skjellum, S. Smith, C. Still, A. Leung, and M. Morari. The Zipcode message passing system. Technical report, Lawrence Livermore National Laboratory, September 1992.

[19] V. Sunderam. PVM: a framework for parallel distributed computing. *Concurrency: Practice and Experience*, 2(4):315–339, 1990.

[20] D. Walker. Standards for message passing in a distributed memory environment. Technical Report TM-12147, Oak Ridge National Laboratory, August 1992.

Appendix A

Process Topology Routines

In this appendix Fortran specifications for the proposed topology functions are given.

- **MPI_CART** Map a cartesian structure (grid/torus)
- **MPI_MAKDIM** Divide a process group among n cartesian dimensions
- **MPI_PARTC** Partition a cartesian structure
- **MPI_GRAPH** Map a graph
- **MPI_INQMAP** Check mapping type
- **MPI_INQCART** Find size of cartesian structure and own position
- **MPI_INQGR** Find own node number in graph
- **MPI_INQPID** Look up PID numbers

NAME

MPI_CART Arrange the processes of a process group in a cartesian structure (torus/grid) with `n` dimensions.

SYNOPSIS

integer function MPI_CART (gid, ndim, dims, period)
integer gid
integer ndim
integer dims(*)
integer period(*)

INPUT ARGUMENTS

gid	the ID number of the process group to be structured.
ndim	the number of dimensions of the torus.
dims	number of processes in each coordinate direction.
period	periodicity specification in each coordinate direction.

DESCRIPTION

The processes of the group with ID number `gid` are arranged in a cartesian structure (grid/torus). By assigning neighboring positions in the logical process torus to neighboring processors in the hardware the system can optimize the locality of inter–process communications. The number of dimensions is specified by `ndim`. Array `period` specifies the periodicity in each coordinate direction. If `period(i)` is set to **MPI_PER** then dimension `i` is treated as periodic. If `period(i)` is set to **MPI_NPER** then dimension `i` is treated as non-periodic. If all `period(i)` are set to **MPI_PER**, then the topology is a torus.

Note that a n-dimension hypercube is a torus with 2 processes per coordinate direction and MPI_PER. Thus, ndim=n, dims(1:n)=2 and period(1:n)=MPI_PER.

RETURN VALUE

If the calling process is assigned a position in the grid, **MPI_CART** returns the total number of processes in the grid. If the calling process is not in the grid, the return value is set to 0. A value of -1 is returned if an error occurs.

NAME

MPI_MAKDIM Divide the processes of a process group among n dimensions of a Cartesian grid.

SYNOPSIS

integer function MPI_MAKDIM (gid, ndim, dims)
integer gid
integer ndim
integer dims(∗)

INPUT ARGUMENTS

gid	the ID number of the process group to be structured.
ndim	the number of dimensions of the torus.
dims	number of processes in each coordinate direction.

OUTPUT ARGUMENTS

dims	number of processes in each coordinate direction.

DESCRIPTION

The processes of the group with ID number `gid` are divided as evenly as possible among `ndim` dimensions, subject to the constraint that all processes in the group are used. The caller may further constrain the operation of this routine by specifying elements of array `dims`. If `dims(i)` is set to a positive number, the routine will fix the number of processes in coordinate direction `i` to `dims(i)`. Processes will then be distributed among directions where `dims(i)` is set to 0, subject to the above constraints.

Note that an error will occur if `dims(i)` is positive and does not evenly divide the number of processes in the group. Negative input values of `dims(i)` will also cause an error. All elements of `dims(*)` must be set to 0 for completely automatic process distribution.

When the function returns, `dims(i)` will be set to the number of processes in coordinate direction `i`. For `dims(i)` not set by the user, `dims(i)` will be ordered (MAXIMUM, ..., MINIMUM). Array `dim(*)` is suitable for use as input to routine MPI_CART.

RETURN VALUE

If the routine completes without error, the return value is set to 0. A value of -1 is returned if an error occurs.

NAME

MPI_PARTC Partition a process group with cartesian structure in subgroups.

SYNOPSIS

integer function MPI_PARTC (gid, lostdim)
integer gid
integer lostdim

INPUT ARGUMENTS

gid	the ID number of the process group to be partitioned.
lostdim	the coordinate direction by which the process map will be reduced.

DESCRIPTION

It is assumed that the processes of the group with ID number `gid` are arranged in a cartesian structure of size (n1,n2,...) by a previous call to MPI_CART. Then, MPI_PARTC creates a number of subgroups with cartesian structures, the dimension of which is one lower than its parent's. The direction of splitting of the parent structure is specified by `lostdim`. If, for example, `lostdim = 2`, then `n2` subgroups with size (n1,n3,n4,...) are formed.

RETURN VALUE

Upon successful completion **MPI_PARTC** returns the group ID number of the subgroup to which the calling process belongs. A value of -1 is returned if an error occurs.

NAME

MPI_GRAPH Arrange the processes of a process group in a general graph structure.

SYNOPSIS

integer function MPI_GRAPH (gid, nodes, index, nglist)
integer gid
integer nodes
integer index($*$)
integer nglist($*$)

INPUT ARGUMENTS

gid	the ID number of the process group to be structured.
nodes	the number of nodes in the graph.
index	index vector into vector nglist (see below).
nglist	for each graph node: list of neighboring nodes (see below).

DESCRIPTION

The processes of the group with ID number `gid` are arranged in a graph structure. By assigning neighboring graph nodes to neighboring processors in the hardware the system can optimize the locality of inter–process communications. The number of nodes is specified by `nodes`.

Input to MPI_GRAPH is for each node its set of neighbors, i. e., the set of processes with which the node process is going to communicate most frequently. Since there is no weighting of node connections, it might be sensible to specify only those neighbors with which there is a considerable amount of communication.

For each node i $(1 \leq i \leq nodes)$ the list of the neighbor nodes is stored at $nglist(j),\ index(i) \leq j \leq index(i+1)-1$. For the last process $index(i+1)$ points to the first free position in vector nglist. Note that the matrix is symmetric, i. e., if node i is a neighbor of node j, than node j is also a neighbor of node i. This property can be used for an internal checking of the input arguments.

RETURN VALUE

If the calling process is assigned a position in the graph, **MPI_GRAPH** returns its node number. If it is not in the graph, a value of 0 is returned. A value of -1 is returned if an error occurs.

NAME

MPI_INQMAP Find out mapping type of a process group.

SYNOPSIS

integer function MPI_INQMAP (gid)

INPUT ARGUMENTS

gid the ID number of the process group.

DESCRIPTION

The function MPI_INQMAP tells the calling process the type of mapping which has been applied to the process group with the ID number `gid`, of which the calling process is a member.

RETURN VALUE

Upon successful completion **MPI_INQMAP** returns

MPI_NOMAP:	if no mapping has been assigned,
MPI_CARTMAP:	in case of a cartesian mapping, and
MPI_GRAPHMAP:	in case of a graph mapping.

A value of -1 is returned if an error occurs.

NAME

MPI_INQCART Look up cartesian structure and own position.

SYNOPSIS

integer function MPI_INQCART (gid, maxdim, size, period, pos)
integer gid
integer maxdim
integer size(∗)
integer period(∗)
integer pos(∗)

INPUT ARGUMENTS

gid	the ID number of the process group with a cartesian structure.
maxdim	the dimension of vectors `size` and `pos` in the calling program.

OUTPUT ARGUMENTS

size	vector returning the size of the grid in each coordinate direction.
period	periodicity characteristics in each coordinate direction.
pos	vector returning the position of the calling process in the grid.

DESCRIPTION

After the processes of the group `gid` have been assigned a cartesian structure (function MPI_CART), the function MPI_INQCART tells the calling process the parameters of that structure and its own logical position.

If, for example, a three–dimensional grid mapping has been applied, then `size` returns the three values `nx,ny,nz`, and `period(i)`, i=1,2,3, is set to **MPI_NPER**. The vector `pos` contains the logical process coordinates `i,j,k` with $1 \leq i \leq nx$, $1 \leq j \leq ny$, and $1 \leq k \leq nz$. If the calling process is not assigned a position in the grid structure, the coordinates are set to zeroes. If the vectors `size`, `period` and `pos` are too short, MPI_INQCART reports an error.

RETURN VALUE

Upon successful completion **MPI_INQCART** returns the number of dimensions of the cartesian structure, `ndim`. A value of -1 is returned if an error occurs.

NAME

MPI_INQGR Look up size of graph mapping and position of calling process.

SYNOPSIS

integer function MPI_INQGR (gid, nodes)

INPUT ARGUMENTS

gid the ID number of the process group with a graph structure.

OUTPUT ARGUMENTS

nodes the total number of nodes in the graph structure.

DESCRIPTION

After the processes of the group `gid` have been structured as a graph (function MPI_GRAPH), the function MPI_INQGR tells the calling process the size of the graph and its own node number. Nodes are numbered from 1 to `nodes` (cf. function MPI_GRAPH).

RETURN VALUE

Upon successful completion **MPI_INQGR** returns the node number `i` of the calling process, $1 \leq i \leq nodes$. If the calling process is not assigned a node in the graph, zero is returned. A value of -1 is returned if an error occurs.

NAME

MPI_INQPID Look up PID numbers of selected processes.

SYNOPSIS

integer function MPI_INQPID (gid, nproc, coord, plist)
integer gid
integer nproc
integer coord(∗)
integer plist(∗)

INPUT ARGUMENTS

gid	the ID number of the process group with a cartesian or graph structure.
nproc	the number of processes for which the PID numbers are to be looked up (see below).
coord	the logical process coordinates of the specified processes (see below).

OUTPUT ARGUMENTS

plist	the list of PID numbers of the specified processes.

DESCRIPTION

Usually a process only communicates with a relatively small number of other processes in a large process structure (a cartesian structure or a graph). Given the logical coordinates of those other processes, MPI_INQPID looks up the corresponding PID numbers. Those PID numbers can then be used for the message–passing.

The processes of which the PID numbers are required are specified by the input arguments `nproc` and `coord`. Two mechanisms are available:

1. Separately specified processes: `nproc` is set to the number of processes $1 \leq nproc$. Vector `coord` contains the logical coordinates for each process. In periodic directions the coordinates are interpreted modulo the number of processes in the corresponding direction (torus topology). As an example, assume the processes are mapped as a two–dimensional torus, and the PID numbers of three processes are to be looked up. Then, `nproc` is set to 3, and the first 6 entries of `coord` are set to $i_1, j_1, i_2, j_2, i_3, j_3$.

   Similarly, if the mapping was done with MPI_GRAPH, the entries of vector `coord` are set to the logical node numbers of the specified graph processes.

2. No individual selection: in this case `nproc` is set to a negative number $nproc \leq -1$. Vector `plist` returns the PID numbers of the *first* `|nproc|` processes. The vector `coord` is not used in this case. Note that this feature does not scale to very high process numbers. In small applications, however, it can be used to get a global list of all PID numbers in the map.

RETURN VALUE

Upon successful completion **MPI_INQPID** returns 0. A value of -1 is returned if an error occurs.