ANL/MCS-TM-00

MPICH Abstract Device Interface Version 3.4 Reference Manual

Draft of September 3, 2003

by

William Gropp
Ewing Lusk
David Ashton
Rob Ross
Rajeev Thakur
Brian Toonen
Mathematics and Computer Science Division
Argonne National Laboratory

MATHEMATICS AND
COMPUTER SCIENCE
DIVISION

Contents

1 Introduction **1**
1.1 Other Work . . . 1
1.2 MPI Overview . . . 1
1.3 A Layered Approach . . . 2

2 High-level Overview **3**

3 Basic Design Rationale **4**
3.1 Communication Types . . . 4
3.2 Additional Goals . . . 5
3.3 Other Relevant MPI Issues . . . 5
3.3.1 Structures Involved in Communication . . . 6
3.3.2 Communication Contexts and Groups . . . 7
3.4 Completing Point-to-Point Operations . . . 7
3.5 Supporting Collective Operations . . . 8
3.6 Data Segments . . . 9
3.7 Remote Memory Access . . . 10
3.7.1 RMA Aggregation . . . 10
3.7.2 Nonblocking RMA and Remote Completion . . . 11
3.8 Contexts for Collective, File, and Window Operations . . . 11
3.8.1 Context Id Generation . . . 12
3.9 Dynamic Processes . . . 12

4 ADI Layers **13**
4.1 Socket (TCP) communication . . . 13
4.2 Remote Put . . . 13
4.3 Shared Memory . . . 13

5 Summary **13**
5.1 Point to point communication . . . 14
5.2 Completion of Point to point communication . . . 14
5.3 Starting and Stopping . . . 15
5.4 RMA . . . 15
5.5 Dynamic Processes . . . 15
5.6 Device Hooks . . . 15

6 Integrating a New Device into the MPICH Build Tree **15**
6.1 Steps Needed in Implementing a Device . . . 16
6.2 Directory Structure . . . 17
6.3 Device Configuration and Setup . . . 17

A Data Structures **17**
A.1 Basic enumerations and types . . . 17
A.2 Major MPI Objects . . . 22
A.3 Threads . . . 32

B Basic Point-to-Point **36**

C Persistent Point-to-Point **43**

D Generalized Requests **45**

E Data Segment **46**

F Communication Buffer management **46**

G Process Topology **46**

H Progress Engine **48**

I Starting and stopping **50**

J Information about the device **54**

K RMA **55**
K.1 Memory Allocation for RMA 55

L Dynamic Processes **56**

M Collective Communication **57**

N Connections and Local Process Ids **57**

O Timers **58**

References **61**

Index **63**

1 Introduction

The goal of the MPICH-2 ADI-3 is to provide routines to support MPI-1 and MPI-2 operations. The target systems include clusters connected with either conventional networks or networks that support remote memory access such as Infiniband, large symmetric multiprocessors (SMPs), and experimental systems (particularly in support of research into MPI implementations).

1.1 Other Work

This section has not been written yet.

This section will briefly review other MPI implementation designs.

Other implementations of MPI include ones from IBM [**?**], Sun [6], MPI Software technology [**?**] and Critical Software [**?**]. The IBM, Sun, and MPI Software technology versions offer support for `MPI_THREAD_MULTIPLE`.

Several implementations have been developed that rely on remote memory operations, including IBM LAPI [1], DEC (now HP) memory channel [4], Infiniband [**?**], and BIP [**?**].

Several implementations of the MPI one-sided operations have been described; see [**?**], [**?**].

The major cluster implementations are MPICH [5] and LAM/MPI [2, 7].

Another implementation is MPI/Pro from MPI Software Technologies; [**?**] describes the implementation of one version of this product.

An example of the use of both lock-free shared-memory operations and threads as MPI "processes" is presented in [8].

1.2 MPI Overview

In order to better understand the design of the ADI, we first review MPI communication. MPI-1 defined both point-to-point and collective message-passing. In both of these, the sender and the receiver actively participates in the communication in the sense that communication does not occur until an MPI call is made to initiate it and the data is not guaranteed to be delivered until an MPI call is made by the destination process. In fact, because the destination in memory of the data is described by an MPI call made by the destination process, the data cannot be delivered to its final (user-specified) destination until the matching MPI call is made by the destination process. This feature has led many (but not all) MPI implementations to adopt a *polling* mode of communication where communication for MPI happens only within MPI calls, not asynchronously.[1] An alternative approach uses either one more separate threads or an interrupt to cause communication to happen outside of MPI calls made by the user[2].

MPI-2 introduced additional operations including remote memory access (RMA), dynamic process management, and parallel I/O. Remote memory access (also called one-sided communication) provides a way to express put, get, and accumulate operations into memory in a remote process. In MPI, these operations are all nonblocking; to ensure that these operations are locally complete, an additional MPI routine must be called. MPI provides two "flavors" of RMA completion: *active target* and *passive target*. In active target, the target process of an RMA operation (that is, the process that did *not* initiate the RMA operation) must call an MPI routine before the RMA completes. The simplest such routine is `MPI_Win_fence`; this is a collective call over all processes associated with the RMA window and is similar to a barrier. The other is the more esoteric "scalable completion" routines, which are called by a group of processes. In both of these cases, an MPI implementation may rely on the target processes calling an MPI routine, and thus these may (but are not required to) use a polling implementation. Unlike some other RMA APIs, MPI active target RMA operations may be applied to any memory belonging to the process.

[1] The only exception to this is cancelling of MPI Send operations; a strictly conforming implementation of this requires a guarantee that the remote process will respond without requiring an MPI call. Because of the extremely rare use of this feature, most polling-mode implementations do not support this case.

[2] MPICH-1 supported both modes but most implementations chose the polling mode. All ADI-2 implementation distributed with MPICH-1 used polling mode; however, some built by outside groups, such as the version for the Intel TFLOPS system, did *not* use polling mode.

The other form of RMA completion is handled by calls made only by the originating process. This is called passive target RMA. Because the target process is not required to make any MPI calls, this kind of RMA requires either very capable hardware that can handle all MPI RMA operations or the use of a non-polling agent at the target process, or a combination of these. Because these operations can be more difficult to implement efficiently, MPI allows an MPI implementation to require that passive target RMA operations be allowed only on memory allocated by `MPI_Alloc_mem`.

MPI-2 dynamic process management allows an MPI application both to create new processes and communicate with them and to connect two already running MPI programs together. The number of processes in an MPI program can thus change over the lifetime of the program, though the MPI routines to create or connect to processes are collective over a communicator, allowing an MPI implementation to ensure that these operations are handled in a scalable fashion.

Section 3 describes some of the design choices that these operations suggest. These choices are not the only ones possible, but we believe that they provide a consistent and efficient way to realize the communication defined by MPI.

1.3 A Layered Approach

The ADI described here is full-featured. This allows an implementor to take advantage of the opportunities for more efficient communication. However, to keep this flexibility from becoming a burden, the design of the ADI is also *layered*: the more advanced features can be emulated by the more basic features. The implementation of the ADI distributed with MPICH will include code to provide these more advanced features in terms of the more basic operations, allowing an implementor to quickly create a working implementation of the ADI and providing the opportunity to later enhance the performance by selectively replacing some of these emulations. Section 4 describes this in more detail.

The ADI described here resembles the communication routines of the MPI standard. The differences are

1. The MPI objects such as requests and communicators are not opaque objects; instead, the ADI uses pointers to structures with defined fields.

2. Checking for correct parameter values is not performed by the ADI routines; the implementation of the MPI routines can make these tests before calling the ADI routines.

3. Completion of MPI operations (e.g., `MPI_Wait`) is handled by a combination of a completion counter in requests and ADI calls to make progress, rather than through calls similar to the MPI wait and test calls. There is no `MPID_Wait` or `MPID_Test` operation.

4. RMA (remote memory access) operations are complemented by a special interface for low-latency operations, particularly in the passive target case.

5. Collective operations are built out of point to point operations (though provision is made to replace each collective operation with an optimized function). An enhancement is planned that allows the use of pipelined store and forward and scatter/gather (to collections of processes) communication.

6. Dynamic process operations use a similar interface to MPI, but the process of building the new intercommunicator is made more explicit through the use of *virtual connections*.

In addition, the ADI is *not* responsible for the construction and management of other MPI objects such as datatypes, attributes, error handlers, and reduction operations. However, hooks are provided to allow the MPI level to notify the ADI of changes in the other MPI objects.

Most implementors will choose to start with a simple interface based on only nonblocking `readv` and `writev`-like operations. This replaces the "channel" interface defined in ADI-2. An implementation of ADI-3 is provided with MPICH that is built on this simple interface; the channel interface and this implementation are described in [**?**].

adi-ov.eps

Figure 1: Simplified block diagram of the communication paths on a connection between two processes. The dashed lines separate the four communication types described below.

Another interface that is under development and that will be documented separately is the method interface. This defines a set of operations that are needed to implement communication on a single connection. This interface allows multiple communication methods to be used in a single MPI program, such as TCP, VIA, and shared memory.

2 High-level Overview

The ADI is relatively rich in functionality. Before diving into some of the details, we provide a brief overview of the ADI. Communication in the ADI takes place between processes. We call the object that describes the communication between two processes a *connection* and show a simplified block diagram of the objects used by the ADI in Figure 1. On each connection, we assume that the low-level communication is *nonblocking*, thus there are queues of pending send operations and ordered lists of receive operations. The lines show communication paths; note that data may be moved from user memory to user memory without passing through the send or receive agents. This allows the ADI to support so-called zero-copy[3] methods.

Because the ADI implementation is connection oriented, all details of the data transfer are hidden within the connection. This makes it relatively easy to support multiple communication methods between different processes; for example, the message format used with TCP communication need not look anything like the message format used by Infiniband communication within the same device. Similarly, flow control is handled on a connection basis, allowing the use of whatever method is appropriate for each communication method.

While the term "connection" is used, there is no requirement that an MPI program create all possible connections or use a connection-oriented low-level protocol. These are really *virtual connections*, with meaning only for the ADI and MPI layers.

It is important to bear in mind the difference between the operations that are sufficient to provide the MPI semantics and the operations that may be necessary to provide a higher-performance match

[3]This is zero copy as that term is understood by network designers. This really mean zero *extra* copies, and does not mean that the data is not copied from the source to the destination process.

between facilities provided by the hardware and software. The ADI endeavors to provide an effective compromise between a minimal set that can provide the full functionality and a richer (and larger) set that can exploit the capabilities of a wide range of systems. To bridge the gap between these, the MPICH ADI implementation provides some implementations of these higher-performance interfaces in terms of a smaller set of operations. (In ADI-2 used in MPICH-1, this smaller set was called the "channel interface".) A more complete discussion of this is presented in Section 4.

3 Basic Design Rationale

In this section, the basic operations that MPI requires of an ADI are described, as well as the design decisions that we have made based on these operations.

3.1 Communication Types

While all MPI interprocess communication (including MPI-1 and MPI-2) can be supported by a single, suitably powerful mechanism such as active messages, the MPI communication semantics described above suggest four separate types of communication operations. These are:

1. Two-party, point-to-point communication. This is the classic send-receive operation. This typically involves coordination between the sender and the receiver, handling such items as flow control, rendezvous messaging, and eager message delivery. Many low-latency implementations of this kind of communication rely on *polling* to advance the communication (make *progress* in MPI terms). Others may use a separate thread or an interrupt-driven mechanism (or a hybrid of both polling and non-polling). Because MPI semantics support nonblocking communication of arbitrary message sizes, any low-level support for communication in MPI must provide nonblocking, point-to-point communication. This is most easily accomplished if the low-level communications is also nonblocking.

2. Communication that has the property of local-completion. That is, a communication operation that must complete independently of any explicit action by the target (destination) process. This is required in MPI-1 for the implementation of `MPI_Cancel` (in the send case in most implementations) and is useful for `MPI_Abort`. In MPI-2, local completion is also needed for *passive target* remote memory access (RMA) operations. This kind of communication is typically implemented through the use of active messages or remote service requests (without polling).

3. Communication for active-target RMA. In MPI, active target RMA operations include remote put, get, and accumulate. These operations are completed by either an `MPI_Win_fence` call made by all processes in the MPI window object or by the combination of `MPI_Win_complete` and `MPI_Win_wait` at the origin (process that initiating a put, get, or accumulate) and target (process containing the memory accessed by a put, get, or accumulate) processes. This form is similar to the two-party, point-to-point communication because it can be implemented by using a pure polling interface. This communication form is separated from the point-to-point mode because the hardware in some systems allows some of the active-target RMA operations to be implemented directly by hardware or low-level software.

4. Communication for passive-target RMA. These operations must complete locally. If these operations must be implemented by communicating with an agent at the remote process, then some form of non-polling agent is required, such as an interrupt-driven active message or a separate communication thread. As in case three, this communication form is separated from case 2 (local completion for MPI-1) because the hardware in some systems allows passive target RMA operations to be implemented directly. We expect some systems to provide an extensive set of operations (e.g., direct access to memory on an SMP through a shared memory segment), others to provide more limited access (e.g., remote DMA through special network support), and others to be implemented on top of a non-polling communication layer such as active

```
(sketch of figure to be completed later)
1 can implement 3
2 can implement 4
1 can provide emulation of 2, but only approximately
```

Figure 2: Sample dependencies between communication approaches

messages. The ADI design is intended to provide a common set of entry points independent of the capabilities of the underlying system.

The ADI design makes no explicit choice between polling and non-polling implementations. Instead, it defines several kinds of polling *points* but allows a purely non-polling (interrupt-driven or separate communication thread) implementation as well.

The four types of communication, of course, can be implemented by a single, suitably powerful abstraction. However, achieving high performance (particularly low latency) requires abstractions that are close to the operations that are efficiently implemented in hardware. The emergence of remote access-style operations in networks [10, 9, **?**] encourages their use as primitives (communication types 3 and 4). Efficient handling of message-passing, particularly the need to manage the flow of point-to-point communications and scalable collective communications on top of more conventional two-sided communications such as TCP suggests communication type 1. Finally, the need to handle some MPI-1 operations that are infrequent and not performance sensitive, but must be handled reliably, suggests communication type 2. The top level ADI interface provides direct access to these four types of communication. Of course, the implementation of the ADI may implement some of these communication types in terms of others, such as reducing all four types to type two (active messages).

3.2 Additional Goals

To ensure that short messages have the lowest possible latency, the common cases should have direct paths to the low-level data transfer operations. In particular, the MPI and API layers should allow operations to to complete without requiring the creation of intermediate data structures. For example, sending a single word (particularly from a blocking `MPI_Send` operation) should not require creating and initializing internal data structures. However, to maintain a simple code base, we will strive to use common code. This suggests that a basic operation is a simple "attempt to send;" this allows the ADI to attempt to send a short data message and return success without creating, for example, a queue element that holds a pending communication operation. Only if the data cannot be communicated immediately will the ADI create an intermediate data structure to hold the state of the incomplete communication.

Thread safety is another goal of the MPICH2 implementation. There are really two separate situations. One is the use of multiple threads by the user's code, for example, in a `MPI_THREAD_MULTIPLE` mode. Another case is the use of a single thread by the user but multiple threads by the MPICH2 implementation. In addition, because thread safety introduces extra overhead to ensure that shared data structures are updated consistently, MPICH2 can be configured for both compile time and runtime specification of the level of support for user threads.

3.3 Other Relevant MPI Issues

MPI defines a number of objects such as requests, windows, and communicators. In many cases, these objects are natural choices for use within the ADI. Using these objects directly, rather than defining different objects for use by the ADI and translating between the MPI and ADI versions, avoids unnecessary overhead within the device. Let us look at several of these objects.

3.3.1 Structures Involved in Communication

MPI Requests. The use of nonblocking operations to implement the low-level communication requires an object to hold the current state of the communication and to record completion of the operation. The natural place to store this within the MPI request. Pending send operations must also be saved in a first-in-first-out queue (to maintain the message ordering guaranteed by MPI); this suggests that the queue contain MPI requests. On the receive side, the message-matching defined by MPI suggests saving the requests in an ordered list. Note the asymmetry between sends and receives. On the send side, requests are placed in a strict FIFO queue for each communication path (to maintain ordering of messages). On the receive side, requests for unmatched receives are kept in an ordered list, and this list, because of the wildcard source receive (`MPI_ANY_SOURCE`) is (logically at least) shared by all communication paths. Similarly, requests for unexpected messages (messages sent but for which no matching receive has yet been issued) are kept in an ordered list. Requests are also the logical place for the data structures relevant to packing and unpacking from complex datatypes to simpler layouts such as contiguous data buffers. This is discussed in more detail under MPI datatypes.

MPI Datatypes. MPI communication can be specified using datatypes that describe complex layouts in memory. An MPI implementation must convert these descriptions into data layouts that can be conveniently moved by the low-level communication layers. Such layers typically support only contiguous memory regions or Unix "io vectors" (`struct iovec`); MPI provides more general forms of data layouts. However, while the MPI datatypes are sufficient to express most forms of communication, there are no routines defined by the MPI standard to pack or unpack only a fraction of a datatype. For example, there is no MPI-defined method to pack as much as fits into a fixed sized buffer and return enough state so that a subsequent pack can pick up where the last left off. Such an operation is needed for any algorithm that packetizes data or that pipelines data transfers. Because this operation is needed both to handle packing noncontiguous data into temporary buffers needed by low-level communication routines (such as TCP `write` or `writev`) and by high-performance algorithms for collective communication, we have introduced a new data structure that is used to pack and unpack buffers described by MPI datatypes. This new structure is called a *segment* and is stored in a structure type named `MPID_Segment`. Segments are discussed in more detail in Section 3.6.

Thus, to handle the need to pack and unpack data, MPI requests also contain a segment. Combined with the datatype and the user-buffer, this gives enough information to move the data to and from the user buffer, even if an intermediate buffer is needed. Note that where possible, no intermediate buffer is used. For contiguous data, and for more general data formats that the underlying communication level supports, data can be moved without using any extra buffers. Segments are required to handle the general case.

Consequences. These considerations suggest that the MPI request (more specifically, the internally-defined structure to which an MPI request refers, which is `MPID_Request`) is the key object. The `MPID_Request` is used to store the progress of communication and order communication between processes. The ADI will use the request as its basic object.

The consequence of this is that the ADI point-to-point communication routines should usually return a request. The only exception is that blocking communication routines should not return a request if the communication is already complete. This allows the blocking communication routines to return completion without ever creating and managing a request. This suggests that the ADI interface for point-to-point communication look something like the following:

```
MPID_Send( buf, count, datatype, tag, <communicator info>, &request )
MPID_Isend( buf, count, datatype, tag, <communicator info>, &request )
MPID_Issend(...)
... similarly for other point-to-point functions
```

The exact form of the arguments that specify the communicator information will be discussed later. This allows a "blocking" send to return a request if the operation has not completed, giving the calling

routine more control over what steps to take to complete the request. It sets the request pointer to `NULL` if it was able to complete without creating a request.

The blocking receive case is similar to the blocking send case. If, when the receive routine is called, the data is available, no request should be created (the cost of creating a request isn't the real issue, it is the cost of initializing and managing the request).

Recall that in the receive case, there are two kinds of receives: posted (but unmatched by an incoming send) and unexpected (sent but unmatched by a receive). For thread-safety, operations on these two lists must be made atomically. These operations include

- check posted and return if found; else insert into unexpected queue
- check unexpected and return if found; else insert into posted queue

The MPID request is the appropriate list element to use in constructing these structures.

While the request is allocated by the ADI, `MPID_Datatype`s are allocated by the MPI implementation. In fact, most of the MPI objects, except for requests, will be allocated by the MPI implementation rather than within the ADI. This is an arbitrary choice; for greatest generality, the ADI could be responsible for allocating all MPI objects. However, we believe that most users of MPICH and ADI-3 will not need that flexibility, and managing the objects within the MPI layer instead of the ADI layer simplifies the implementation of the ADI. However, to make provision for any ADI-specific features that must be associated with an MPI object, the definition of the structure associated with each object includes

```
MPID_DEV_xxxx_DECL
```

where `xxxx` may be `COMM`, `DATATYPE`, etc. This provides a simple way to extend the objects defined by the MPI layer without forcing the ADI to provide a complete implementation. Whenever an object is created or destroyed, the MPI layer can call a *hook* routine with a name of the form

```
MPID_Dev_xxxx_create_hook( pointer to object, ... )
MPID_Dev_xxxx_destroy_hook( pointer to object, ... )
```

The other parameters will be defined as it becomes clear what is needed. For example, in the case of communicator creation, a pointer to the old communicator may be needed.

These routines are called after all other creation operations take place and before any of the destroy operations take place.

These calls have not yet been implemented

3.3.2 Communication Contexts and Groups

MPI Communicators describe both a collection of processes (an `MPI_Group`) and a unique communication context. As described later, in MPICH and ADI-3, the communication context is encoded as an integer. The target process of communication in MPI is described by a rank in a group associated with a communicator. While this suggests that MPI groups are fundamental data structures, in MPICH-2, groups are not used in the ADI at all. Instead, each communicator maintains an array of *connections* that are indexed by the rank. Using these connections, the MPI implementation provides complete support for the MPI group operations (e.g., `MPI_Group_union`). Explicit MPI Groups are not used by the ADI.

3.4 Completing Point-to-Point Operations

To complete some particular MPI communication (described by a request, such as in a call to `MPI_Wait`, it is necessary to have the ADI respond to *any* pending communication. Thus, it is not necessary to provide the ADI with a collection of request to test or wait on. Instead, we merely need to ask the ADI to try to make progress on communication and then check (using the completion counter in each request) whether any particular MPI requests have completed. In the absence of threads, a simple interface would look like

```
MPI_Waitsome( ... )
{
while( no completed requests found )
    for (i=0; i<count; i++) {
        if (any requests done, return those as complete)
    }
    MPID_Make_progress( TRUE );
}
```

However, if there are multiple threads, particularly if there are separate threads that can complete communication, then this code has a race condition, caused by the API for progress. Instead, a slightly more complex interface is needed to eliminate any race conditions. For example,

```
MPI_Waitsome( ... )
{
  while (1) {
    MPID_Progress_start( );   // Notes that we are about to
                              // check ready flags.  No completion
                              // counters will be set to complete
    for (i=0; i<count; i++) {
        if (any request done) { save info on request }
    }
    if (no requests done)
        MPID_Progress_wait();
    else {
        MPID_Progress_end();
        break;
    }
  }
}
```

The interface for test operations is similar, except that `MPID_Progress_wait` is replaced with `MPID_Progress_test`, and no outer loop is needed.

In a polling implementation, the "start" and "end" calls are no-ops and the "wait" and "test" calls are blocking and nonblocking polling calls respectively. In a nonpolling implementation, the "start" and "end" calls may set and clear a thread lock or access lock, and the "wait" and "test" calls may yield to a communication thread (in addition, the wait version could wait to be signaled through a condition variable). This interface allows us to use the same code for completing MPI nonblocking operations independent of the choice of polling, nonpolling, threaded, or nonthreaded implementations of the ADI.

3.5 Supporting Collective Operations

When implementing collective communication algorithms, the ability to both store and forward data is important for performance. For example, when complex MPI datatypes are used, it may be necessary when receiving data to first receive into a temporary buffer and then unpack that data into the user's buffer. Forwarding this same data on (for example, within a broadcast) in a separate MPI send operation then requires repacking the data from the user into a temporary buffer. To enable an ADI implementation to avoid this cost, and to make it easier to efficiently write the collective communication algorithms, the ADI will provide a variety of store and forward, scatter, and gather operations. Note that these operations can be emulated using only point-to-point; as described above, these can be built on top of simpler, point-to-point communication. These interfaces are still under design.

Consequences. Determining completion in the store and forward or multisend cases may involve more than one communication operation and possibly multiple communication methods. This argues

that the status of a communication be tracked with a completion counter rather than a simple flag, since multiple communication operations may be working with the same data buffer.

To best exploit the fact that the same data is both being received and sent, the ADI should be able to provide pointers to "good" memory for these operations. For example, such memory may be in a special, pinned page or within a sophisticated NIC.

The algorithms for efficient collective communication provide some information on the kinds of multi-party operations that are required. The ADI does not support the most general of these operations; the goal is to allow an MPI implementation to efficiently and correctly support the more common collective communication operations such as `MPI_Bcast`, `MPI_Scatter`, and `MPI_Allgather`.

The very first version of ADI-3 does not include these more general multi-party communication operations. It is the intent of ADI-3, however, to develop an efficient method for describing and implementing the operations needed for the MPI collective operations. We may also consider special support for collective argument checking, that is, checking that the parameters to a collective MPI routine are consistent amoung all of the involved processes, perhaps using a special header format or layered header format.

3.6 Data Segments

MPI datatypes are very general; a single instance of a datatype can describe an arbitrarily large amount of data that is not necessarily contiguous in memory. Further, MPI datatypes can be very concise; a vector datatype of a structure (that itself is not contiguous) describes a very complex memory layout with just a few words of memory. Further, MPI dataytpes may be described using only five different basic loop types [**?**, **?**]. Because few if any low-level communication layers support the full generality of MPI datatypes, it is sometimes necessary to pack and unpack data to intermediate buffers. In addition, it may be necessary to pack or unpack a single MPI datatype with multiple calls; this operation is not supported by the `MPI_Pack` and `MPI_Unpack` routines. For example, the code

```
MPI_Type_vector( 1000000, 1, 237, MPI_DOUBLE, &newtype );
MPI_Type_commit( &newtype );
MPI_Send( buf, 1, newtype, ... );
```

passes a single instance of an MPI datatype that describes 8 MB of data, even though it is completely described by the base address and four integers, one of which is the size of a `double`. Further, because this is a vector type, it cannot be represented efficiently with a `struct iovec`. If the underlying communication layer can only send a maximum of 64K at a time (for example, using a shared memory pool or a remote-memory communication area), it is necessary to incrementally pack this datatype, 64K at a time, into a temporary buffer. The segment routines provide this capability.

These routines are also needed for some implementations of the collective communication routines. Many of the better (and in some cases, the best) algorithms for the collective operations in MPI can be cast, at least for systems that are homogeneous is data representation, in terms of operations that view the data to be communicated as a contiguous range of bytes and that send different parts to different processes. To implement these algorithms for MPI requires handling MPI datatypes; even in a single collective operation, some MPI processes may provide a simple, contiguous data buffer while others specify a complex datatype. To implement these algorithms requires a method to extract parts of a data buffer, viewed as a range of bytes. This is a simple variation on the routines that can incrementally pack (and unpack) a data buffer described by an MPI datatype.

Question: We need to tie down the details of the segment creation and pack/unpack operations.

Issues with the segment routines:

1. Who allocates storage for the segment? That is, where does the contiguous buffer come from? In many cases, it would be nice if it was memory that was convenient for the ADI-3 device. In the case of shared memory or RMA-networks, using special memory saves a memory copy.

2. Who sets the amount of data to pack or unpack in each call? The specific concern here is to not stop in the middle of a natural data item, e.g., 3/8ths of the way through a double. The design

here allows the routines to make slight adjustments in the amount of data packed or unpacked in order to stay on "natural" boundaries.

3. How much do these routines need to know about MPI Datatypes? We'd like them to be generic so that they could be shared with other projects such as PVFS and HDF5.

3.7 Remote Memory Access

The MPI remote memory access model was deliberately designed to have very loose (but precise) synchronization requirements and to make minimal *demands* on the underlying hardware. For example, it is possible within the MPI model to support non-cache-coherent systems (such as the NEC vector supercomputers)[4]. However, the model also *allows* an MPI implementation to exploit special hardware capabilities.

The MPI specification is very careful in describing when a process's memory window is accessible to other processes (the *exposure epoch*) and when a process may be performing RMA operations (the *access epoch*). Understanding these is necessary in developing a correct MPI implementation. However, these concepts are deliberately made as general as possible to allow the greatest flexibility to an MPI implementation. In the discussion below, we will usually not refer to the access or exposure epochs. However, if there are questions as to what the terms "synchronization" or "completion" mean in the RMA context, consult the discussion of the RMA epochs in the MPI standard.

The MPI specification has a number of "as if" rules, such as "as if only one process accesses a memory window at a time". Naturally, if an implementation can perform an operation more efficiently without violating such "as if" rules, the implementation is free to do so. An example of this is passive target updates to disjoint regions in a memory window; the "as if" rule says that these must appear "as if executed sequentially," but an implementation can allow concurrent updates if they are known a priori to be disjoint. This suggests that operations be aggregated so that the range of affected bytes within the target window is known. Fortunately, the MPI specification allows such aggregation.

3.7.1 RMA Aggregation

One of the most misunderstood parts of the MPI RMA specification is the issue of when operations take place, particularly with the poorly named `MPI_Win_lock` and `MPI_Win_unlock` routines. MPI RMA allows the implementation considerable latitude in the timing of operations. An important case in point is aggregation (combining) of RMA operations. The approach of aggregating RMA operations has been developed in the BSP approach, and the MPI specification was designed to support this technique. For example, the following sequence of MPI calls

```
MPI_Win_lock( MPI_LOCK_SHARED, rank, 0, win );
MPI_Accumulate( &one, 1, MPI_INT, rank, 0, 1, MPI_INT, MPI_SUM, win );
MPI_Win_unlock( rank, win );
```

can be converted into a single, atomic update operation on some systems (particularly on those that have no direct access to remote shared memory, such as a system that only supports TCP communication).

To allow low-latency implementation of single-element remote updates (e.g., put or accumulate), the ADI design allows the MPI implementation to perform aggregation of RMA operations without calling the ADI. This eliminates a layer of function calls in these simple cases[5]. The ADI provides a set of definitions that are used to decide the aggregation threshold, in terms of number of bytes and operations. The device can also specify that no aggregation is done at the MPI level, giving the ADI greater control at the cost of additional function calls. **How does the device specify this?**

[4] The ADI-3 design, however, does assume cache coherence with local memory; that is, the memory system on each node is cache coherent.

[5] In addition, the MPI functions could be inlined by a suitably sophisticated compiler, removing all function calls.

3.7.2 Nonblocking RMA and Remote Completion

The MPI RMA operations are nonblocking. Thus, there must be some way to indicate both local and remote completion. MPI provides users with three different mechanisms for marking completion in their code:

Fence. This is essentially a barrier synchronization, similar to the Cray SHMEM routine `shmembarrier`. To allow the ADI to exploit hardware and software features similar to those used by Cray `shmembarrier` in the Cray T3D and T3E, the ADI provides a similar routine (`MPID_Win_fence`), with the difference that it applies to MPI window objects on arbitrary groups of processes. An ADI implementation for a system that provides an efficient fence operation only on all processes can, of course, test for that case and execute different code when not all processes are involved in the MPI window object.

Passive. This is a kind of two-party synchronization since only the origin and target processes are involved, rather than all members of the group of the MPI window object. The ADI provides a simple completion counter variable that is zero on completion; this allows the flag to be a counter that contains the number of uncompleted operations or a simple boolean that indicates whether the operation is complete. The MPI calls that express this kind of synchronization are the misnamed `MPI_Win_lock` and `MPI_Win_unlock`. Note that calls by a process to `MPI_Win_lock` for its own rank (i.e., "lock my window") are different in behavior from calls to a remote process because the local process may use non-MPI operations to access the memory window (e.g., through simple references or assignments). ADI-3 assumes a cache-coherent memory system, which allows some important simplifications in handling these operations.

Scalable Multiparty. In this mode, not all processes in the window are involved (unlike the fence case), but both origin and target processes make calls to indicate when operations must complete. The MPI calls used to express this kind of synchronization are `MPI_Win_post`, `MPI_Win_start`, `MPI_Win_complete`, and `MPI_Win_wait`. This form can also be handled with counters that are updated by each partner process. Efficient handling of this approach remains a research issue, however.

3.8 Contexts for Collective, File, and Window Operations

MPI requires that different kinds of communication be non-interfering. That is, communication for collective operations such as `MPI_Bcast` and point-to-point operations such as `MPI_Send`, even on the same communicator, must not interfere with each other. With MPI-2, this is extended to operations on MPI File and Window objects, both of which may involve some communication by the implementation.

MPI also requires that communication within different communicators be noninterfering. Many implementations achieve this by using a hidden (to the MPI user) *context id*, which is simply an integer that is communicated along with the tag, source, etc. One approach that can be used to ensure that communication of different types be non-interfering is to use different communicators; in MPICH-1, each communicator was created with a second, hidden (from the user) communicator that was used for communication implementing collective operations.

However, this approach has a number of drawbacks, particularly in organizing the code. In addition, with files and windows as well, three hidden communicators (in addition to the one used for point-to-point communication) might be needed[6]. Finally, the implementation of intercommunicator collective operations, also introduced in MPI-2, adds additional complexity. In MPICH-2, we take a different route. Instead of creating hidden communicators, we allocate context ids in groups of four. The API routines for communication take an explicit context id offset as a parameter. These offsets have the following values. For intracommunicators:

0 –point to point communication,

[6] An alternative is to perform a shallow duplicate (without invoking attribute copy functions) of the communicators passed into the file and window creation routines.

1 –collective communication,

2 –communication for files,

3 –communication for window objects.

For intercommunicators:

0 –point to point,

1 –collective in group A,

2 –collective in group B,

3 –collective over both groups.

(Note that a single context value for "collective in local group" is adequate; however, having separate context values provides stronger separation and aids in debugging.)

3.8.1 Context Id Generation

Because libraries are encouraged to use it, the operation `MPI_Comm_dup` should be efficient. In particular, where possible, it should be a local operation (involving no communication). This is possible, at least for the first few dup's of a communicator, if each communicator caches some extra context ids when it is created. The MPI implementation will maintain a cache of context values in a communicator; communicators created by dup'ing that communicator will take a value from the cache. If the cache is empty, the MPI level will make a collective call to the ADI-3 routine that returns context ids. To simplify this, the ADI-3 routine returns a single value; the MPI level will multiply this by a fixed constant size to create a sequence of consecutive context ids. In addition, the MPI level will keep track of returned context ids (made available by freeing a communicator) so that the ADI can be told when a context id is available again.

In the current implementation, context id generation is provided by a utility routine above the level of the ADI. Because there are opportunties for implementations on hardware that provides remote memory operations, such as accumulate, future versions of the ADI may allow the ADI implementation to replace the default context id generation.

3.9 Dynamic Processes

There are two major goals here for the ADI: scale to large but not necessarily enormous numbers of processes and maintain modularity so that it is relatively easy to maintain and extend the implementation. A secondary goal is to structure the code so that initialization and rundown are efficient and do not spend a great deal of time setting up facilities that a program never uses.

MPI-2 adds dynamic process management. A consequence of this is that there are no absolute and global process ids. This observation suggests that all communication be considered locally in terms of (possibly virtual) connections to processes. Unlike ADI-2, there are no data structures that map a rank in a communicator into a "global rank" (i.e., rank in `MPI_COMM_WORLD`). Instead, communicators have arrays of virtual connections that are indexed by the rank.

There is a kind of local process id that is used by the MPI group operations. However, it is not critical item and, since the MPI group operations are implemented above the ADI, we do not discuss them here. The one exception is the routine to provide a mapping from any rank in a communicator to a value that uniquely specifies an MPI process *relative only to the calling process*. That is, we do not need a value that is globally correct, only one such that if two ranks in two different communicators *on the same process* refer to the same MPI process, then this mapping will give the same value on that process. Such values are called "local process ids".

4 ADI Layers

It is expected that many implementations of the ADI will be *layered*, building the four types of communication described in Section 3.1 in terms of simpler communication methods. This section outlines several possible implementations.

4.1 Socket (TCP) communication

Socket-based communication provides a two-party communication similar to the type 1 communication in Section 3.1. A simple implementation can map all four kinds of communication into simple `read` and `write` calls as follows (with some caveats):

1. Point to point communication is a close match. Some care is needed to handle flow control.
2. Local completion communication is more difficult. A simple (and not entirely correct) approach is to simply convert these operations (e.g., `MPID_Cancel_send`) into a message that is sent on the same socket as the type 1 point to point communication. All that this requires is that messages sent on the socket connecting two processes include a header that describes the type of message (e.g., MPI message envelope, MPI cancel message, MPI cancel acknowledgement, etc.). For those familar with the channel device in ADI-2, these are just the packet types.
3. Active target RMA. This can also be converted into simple messages sent on the same socket. Note that there are some complications in handling complex MPI datatypes.
4. Passive target RMA. Just like active target RMA. Note, however, that for correct behavior, the receive agent must be called at least occasionally, even if the target process makes no MPI calls.

The progress engine is implemented by using `select` or `poll`, and is a pure polling approach.

To correctly handle the local completion (type 2) and passive target (type 4) communication when using sockets, there must be an asynchronous communication agent. One easy way to do this is to take the progress engine and place it in a separate thread. The operating system then guarantees that the progress engine is called often enough to ensure that the locally completing communication operations make the necessary progress. However, to illustrate the advantages of the ADI design (in terms of four separate communication types), consider a different version where there are two sockets between communicating pairs of processes. The first socket is used in a polling mode for point-to-point and active target communication (types 1 and 3). This provides the low-latency associated with polling models. The second socket is used for the locally-completing communication (types 2 and 4), and uses a separate thread, process, or `SIGIO` to ensure local completion. This approach provides correctness with the MPI progress model without sacrificing the low latency of the polling approach for the more common operations. Note that such an implementation must still guard some data structure updates where different communication threads might update the same data structure.

Of course, other approaches are possible, including one that uses a mix of polling and nonpolling even on type 1 and type 3 communication.

4.2 Remote Put

Not yet written

4.3 Shared Memory

Not yet written

5 Summary

This section provides a short summary of the ADI routines. Complete details are provided in the ADI-3 reference manual.

5.1 Point to point communication

Each simple MPI communication operation has its counterpart here:

```
MPID_Send
MPID_Ssend
MPID_Rsend
MPID_Isend
MPID_Issend
MPID_Irsend
MPID_Recv
MPID_Irecv
MPID_Request_free
MPID_Iprobe
MPID_Probe
```

Note that these do not include the buffered send nor the two send-receive operations (`MPI_Sendrecv` and `MPI_Sendrecv_replace`). Send-receive operations are described in terms of separate send and receive operations; buffered send may use the optional `MPID_tBsend` or may simply use `MPID_Isend` with the sample implementation described in the MPI-1 standard.

Several MPI routines are represented by slightly different routines. These include

```
MPID_tBsend        - Used to optimize buffered sends (MPI_Bsend etc.)
MPID_Cancel_send   - Separate routines for cancelling sends and receives
MPID_Cancel_recv
```

In addition, the persistent communication routines have MPID equivalents:

```
MPID_Send_init
MPID_Ssend_init
MPID_Rsend_init
MPID_Recv_init
MPID_Startall
```

These are provided so that the device can effectively manage the persistent request, which may require allocating a device-specific request either when an init routine (e.g., `MPID_Send_init`) is called or when the request is started with `MPID_Startall`.

Note that there is no direct access to the message queues by the MPI layer. We recommend that the device implement the message queue interface defined for external tools and used by Totalview. This interface is described in [3] and the files `mpich2/src/mpi/debugger/` in MPICH2.

5.2 Completion of Point to point communication

Completion of nonblocking (or incomplete in the case of a request returned by `MPID_Send`, `MPID_Ssend`, `MPID_Rsend`, or `MPID_Recv`) operations is handled by calling the progress engine routines and checking the completion counter value in a request. The progress engine routines are

```
MPID_Progress_start
MPID_Progress_end
MPID_Progress_test
MPID_Progress_wait
```

This set of routines is required in order to provide an interface to the request's `completion counter` value that is thread-safe. In addition, there is a routine that indicates *polling points*; places in the MPI code where a polling implementation should check for incoming messages. This routine is nonblocking and is

```
MPID_Progress_poke
```

5.3 Starting and Stopping

The routines to start and stop the ADI match the MPI counterparts. Note however that there are no calls to implement "is initialized"; this is managed entirely at the MPI level.

```
MPID_Init
MPID_Finalize
```

5.4 RMA

The RMA routines haven't been set yet. However, they will include MPID versions of put, get, and accumulate, along with a few calls to handle the aggregated operations (e.g., lock/accumulate/unlock). In addition, we expect to separate the `MPI_Win_lock` and `MPI_Win_unlock` into two kinds of operations: non-local, where the operation is really a start/end RMA operation, and local, where it really is lock/unlock.

One question is whether there should be support at the MPI layer for caching data type descriptions (e.g., for complex MPI datatypes to be applied at the target process) and the corresponding ADI routines, or whether this should be handled entirely by the ADI.

The ADI routines that support RMA include

```
MPID_Win_put
MPID_Win_get
MPID_Win_accumulate
MPID_Win_do
MPID_Win_fence
MPID_Win_start
MPID_Win_end
MPID_Win_local_lock
MPID_Win_local_unlock
```

`MPID_Win_do` is used to send aggregated operations to the ADI. This provides a pointer to a description and an indication of whether this starts, ends, or continues a sequence of RMA operations. `MPID_Win_start` and `MPID_Win_end` are used for nonlocal uses of `MPI_Win_lock` and `MPI_Win_unlock`.

We may want `MPID_Win_do_put`, `MPID_Win_do_get`, and `MPID_Win_do_accumulate` instead of a single `MPID_Win_do`.

5.5 Dynamic Processes

The MPID routines are similar to the MPI routines. Undecided: does the MPI layer or the ADI layer setup the communicator? The ADI layer needs only provide the pointers to the connection structure; the MPI layer could build the communicators and access the context id values.

5.6 Device Hooks

To allow the device to track the creation and destruction of MPI objects (other than requests), the device may define hook routines. These have the form

```
MPID_Dev_xxx_create_hook
MPID_Dev_xxx_destroy_hook
```

where `xxx` is the object type, such as `comm`, `datatype`, `group`, etc. Many devices will define these as C preprocessor macros that expand to nothing, thus eliminating any function calls.

6 Integrating a New Device into the MPICH Build Tree

This section describes how to add a device or method into the MPICH build tree. This section is intended both to describe the process of adding a device and the rationale for the design of the device-dependent modules.

6.1 Steps Needed in Implementing a Device

Here are the steps that must be taken to implement a new device. Similar steps are required when adding to a device, such as creating a new channel-based device.

1. A new device is created by adding a new directory `src/mpid/<newdevice>`. Into this directory should go all of the files (and subdirectories) needed for the device. These include

   `setup_device:` This script, if present, is run *before* any configure in the device directory. It can setup any files or communicate any data that is needed by the device.

   `configure:` This should be a standard `autoconf` script (at least, it must accept the standard `configure` arguments. All device-specific tests, such as those for header files or libraries, should be made by the `configure` in the device directory. No changes of any kind should be made to the top-level `configure` (the one in the `mpich2` directory).

   `localdefs:` This file, if present, is sourced by the top-level `configure` after the device's `configure` is run. It should be used to communicate any variable values to the top-level configure. A common use it to add any libraries that may be required. For example, a file `localdefs.in` may be used that contains

   ```
   LIBS="$LIBS @EXTRA_LIBS@"
   ```

   The `configure` in the device directory includes `localdefs` in the `AC_OUTPUT` list, allowing the device's `configure` step to create the `localdefs` file.

2. Include files. Any include files that are needed by `mpiimpl.h` should be make available by including the directory path in the `CPPFLAGS` variable, set in the `setup_device` file mentioned above. Make sure that you append to this variable, as in

   ```
   CPPFLAGS="$CPPFLAGS -I$use_top_srcdir/src/mpid/mydevice"
   ```

   Note the use of the variable `use_top_srcdir`; the top-level `configure` sets this to the absolute path to the top-level source directory.

   Provide the files `mpidpre.h` and `mpidpost.h`. The implementation of all MPI routines include files in this order:

   mpi.h The standard `mpi.h` that all MPI users include

   mpidpre.h Any definitions needed *before* the provided definitions of the contents of the internal structures. This can included definitions that override parts of `mpiimpl.h`

   contents of mpiimpl.h The bulk of the internal definitions. This also includes information on the timers.

   mpidpost.h Any definitions needed by the device after the rest of the definitions in `mpiimpl.h`. In many cases, this file may be empty.

3. Testing codes for device-specific functions. Place these in a `test` subdirectory of the device. These tests should be performed through a `test` target in the device's `Makefile.sm`.

4. `mpiexec`. Decide whether you need a special `mpiexec` program. Many devices will be able to use any of the `mpiexec` programs in the various `src/pm` directories, such as `src/pm/mpd` and `src/pm/forker`. ("pm" is for process manager, and MPICH2 expects to install some process manager for all devices.) If you need a new `mpiexec`, it should be added to a new subdirectory within `src/pm`. There must be an installation target in the `Makefile.sm` for any new `mpiexec`. For example, if the `mpiexec` program requires the multi-purpose demon (`mpd`), ensure that the `mpd` is installed. For example, in the `src/pm/<my-process-manager>/Makefile.sm`, use the line

   ```
   install_BIN = mpd
   ```

to install the program `mpd` into the `bin` directory. The particular process manager to use is selected at configure time using the `--with-pm` option. Most devices should be able to use one of the process managers provided with MPICH2.

5. Device-specific documentation, such as environment variables and command-line arguments used only by a particular device. Place this information into the file `src/mpid/<yourdevice>devdoc.txt`. The `mpich` documentation generators will look for this file.

   All of these are included by the file `mpiimpl.h`.

6.2 Directory Structure

A device should be placed in a subdirectory of `mpich/src/mpid/`; for example, `mpich/src/mpid/mm` is the multi-method ADI delivered with `mpich`. The directory name is the same as the device name specified to the `mpich configure` with the `--with-device` option.

6.3 Device Configuration and Setup

Each device must have a `configure` script. This will be run by the `mpich configure` as part of the top-level configuration. Any other commands that a device needs for setup should be run using the `AC_OUTPUT_COMMANDS autoconf` macro. Autoconf version 2.13 or later, but before 2.50, should be used; we recommend using the macros defined in `mpich/confdb/aclocal.m4`, as they include fixes to `autoconf`[7]. Do not modify the `mpich configure` to support a device.

There must be a `echomaxprocname` target in the `Makefile` in the device's directory. This should look something like

```
echomaxprocname:
       @echo 128
```

This value will be used as the value for `MPI_MAX_PROCESSOR_NAME`, and must be an integer value. (The above will be replaced by an option to provide this value through the `localdefs` files.)

A Data Structures

This section contains descriptions of the data structures using in the ADI3 definition. These include both the major enumerated types and the structure definitions.

A.1 Basic enumerations and types

MPID_Object_kind — Object kind (communicator, window, or file)

Synopsis

```
typedef enum MPID_Object_kind {
  MPID_COMM       = 0x1,
  MPID_GROUP      = 0x2,
  MPID_DATATYPE   = 0x3,
  MPID_FILE       = 0x4,
  MPID_ERRHANDLER = 0x5,
  MPID_OP         = 0x6,
  MPID_INFO       = 0x7,
  MPID_WIN        = 0x8,
```

[7] There are serious bugs in the `AC_CHECK_HEADER` macro that are still present in `autoconf` 2.52, and version 2.57 is not backward compatible with earlier versions of `autoconf`, including 2.52.

```
  MPID_KEYVAL     = 0x9,
  MPID_ATTR       = 0xa,
  MPID_REQUEST    = 0xb
  } MPID_Object_kind;
```

Notes

This enum is used by keyvals and errhandlers to indicate the type of object for which MPI opaque types the data is valid. These are defined as bits to allow future expansion to the case where an object is value for multiple types (for example, we may want a universal error handler for errors return). This is also used to indicate the type of MPI object a MPI handle represents. It is an enum because only this applies only the the MPI objects.

Module

Attribute-DS

MPID_Lang_t — Known language bindings for MPI

Synopsis

```
typedef enum MPID_Lang_t { MPID_LANG_C
#ifdef HAVE_FORTRAN_BINDING
                         , MPID_LANG_FORTRAN
                         , MPID_LANG_FORTRAN90
#endif
#ifdef HAVE_CXX_BINDING
                         , MPID_LANG_CXX
#endif
} MPID_Lang_t;
```

A few operations in MPI need to know what language they were called from or created by. This type enumerates the possible languages so that the MPI implementation can choose the correct behavior. An example of this are the keyval attribute copy and delete functions.

Module

Attribute-DS

MPID_Request_kind — Kinds of MPI Requests

Synopsis

```
typedef enum MPID_Request_kind_t {
    MPID_REQUEST_SEND, MPID_REQUEST_RECV, MPID_PREQUEST_SEND,
    MPID_PREQUEST_RECV, MPID_UREQUEST } MPID_Request_kind_t;
```

Module

Request-DS

MPID_Comm_kind_t — Name the two types of communicators

Synopsis

```
typedef enum MPID_Comm_kind_t {
    MPID_INTRACOMM = 0,
    MPID_INTERCOMM = 1 } MPID_Comm_kind_t;
```

MPID_Errhandler_fn — MPID Structure to hold an error handler function

Synopsis

```
typedef union MPID_Errhandler_fn {
   void (*C_Comm_Handler_function) ( MPI_Comm *, int *, ... );
   void (*F77_Handler_function) ( MPI_Fint *, MPI_Fint *, ... );
   void (*C_Win_Handler_function) ( MPI_Win *, int *, ... );
   void (*C_File_Handler_function) ( MPI_File *, int *, ... );
} MPID_Errhandler_fn;
```

Notes

The MPI-1 Standard declared only the C version of this, implicitly assuming that `int` and `MPI_Fint` were the same.

Module

ErrHand-DS

Questions

What do we want to do about C++? Do we want a hook for a routine that can be called to throw an exception in C++, particularly if we give C++ access to this structure? Does the C++ handler need to be different (not part of the union)?
What is the interface for the Fortran version of the error handler?

MPID_Op_kind — Enumerates types of MPI_Op types

Synopsis

```
typedef enum MPID_Op_kind { MPID_OP_MAX=1, MPID_OP_MIN=2,
                            MPID_OP_SUM=3, MPID_OP_PROD=4,
                MPID_OP_LAND=5, MPID_OP_BAND=6, MPID_OP_LOR=7, MPID_OP_BOR=8,
                MPID_OP_LXOR=9, MPID_OP_BXOR=10, MPID_OP_MAXLOC=11,
                MPID_OP_MINLOC=12, MPID_OP_REPLACE=13,
                MPID_OP_USER_NONCOMMUTE=32, MPID_OP_USER=33 }
  MPID_Op_kind;
```

Notes

These are needed for implementing `MPI_Accumulate`, since only predefined operations are allowed for that operation.
A gap in the enum values was made allow additional predefined operations to be inserted. This might include future additions to MPI or experimental extensions (such as a Read-Modify-Write operation).

Module

Collective-DS

MPID_User_function — Definition of a user function for MPI_Op types.

Synopsis

```
typedef union MPID_User_function {
    void (*c_function) ( const void *, void *,
                        const int *, const MPI_Datatype * );
    void (*f77_function) ( const void *, void *,
                         const MPI_Fint *, const MPI_Fint * );
} MPID_User_function;
```

Notes

This includes a `const` to make clear which is the `in` argument and which the `inout` argument, and to indicate that the `count` and `datatype` arguments are unchanged (they are addresses in an attempt to allow interoperation with Fortran). It includes `restrict` to emphasize that no overlapping operations are allowed.
We need to include a Fortran version, since those arguments will have type `MPI_Fint *` instead. We also need to add a test to the test suite for this case; in fact, we need tests for each of the handle types to ensure that the transfered handle works correctly.
This is part of the collective module because user-defined operations are valid only for the collective computation routines and not for RMA accumulate.
Yes, the `restrict` is in the correct location. C compilers that support `restrict` should be able to generate code that is as good as a Fortran compiler would for these functions.
We should note on the manual pages for user-defined operations that `restrict` should be used when available, and that a cast may be required when passing such a function to `MPI_Op_create`.

Question

Should each of these function types have an associated typedef?
Should there be a C++ function here?

Module

Collective-DS

MPID_Copy_function — MPID Structure to hold an attribute copy function

Synopsis

```
typedef union MPID_Copy_function {
  int  (*C_CopyFunction)( int, int, void *, void *, void *, int * );
  void (*F77_CopyFunction)  ( MPI_Fint *, MPI_Fint *, MPI_Fint *, MPI_Fint *,
                              MPI_Fint *, MPI_Fint *, MPI_Fint * );
  void (*F90_CopyFunction)  ( MPI_Fint *, MPI_Fint *, MPI_Aint *, MPI_Aint *,
                              MPI_Aint *, MPI_Fint *, MPI_Fint * );
  /* The C++ function is the same as the C function */
} MPID_Copy_function;
```

Notes

The appropriate element of this union is selected by using the language field of the `keyval`. Because `MPI_Comm`, `MPI_Win`, and `MPI_Datatype` are all `int`s in MPICH2, we use a single C copy function rather than have separate ones for the Communicator, Window, and Datatype attributes. There are no corresponding typedefs for the Fortran functions. The F77 function corresponds to the Fortran 77 binding used in MPI-1 and the F90 function corresponds to the Fortran 90 binding used in MPI-2.

Module

Attribute-DS

MPID_Delete_function — MPID Structure to hold an attribute delete function

Synopsis

```
typedef union MPID_Delete_function {
  int  (*C_DeleteFunction)  ( int, int, void *, void * );
  void (*F77_DeleteFunction)( MPI_Fint *, MPI_Fint *, MPI_Fint *, MPI_Fint *,
                              MPI_Fint * );
  void (*F90_DeleteFunction)( MPI_Fint *, MPI_Fint *, MPI_Aint *, MPI_Aint *,
                              MPI_Fint * );
} MPID_Delete_function;
```

Notes

The appropriate element of this union is selected by using the language field of the `keyval`. Because `MPI_Comm`, `MPI_Win`, and `MPI_Datatype` are all `int`s in MPICH2, we use a single C delete function rather than have separate ones for the Communicator, Window, and Datatype attributes. There are no corresponding typedefs for the Fortran functions. The F77 function corresponds to the Fortran 77 binding used in MPI-1 and the F90 function corresponds to the Fortran 90 binding used in MPI-2.

Module

Attribute-DS

A.2 Major MPI Objects

MPID_Request — Description of the Request data structure

Synopsis

```
typedef struct MPID_Request {
    int          handle;
    volatile int ref_count;
#ifndef MPICH_SINGLE_THREADED
    MPID_Thread_lock_t mutex;
#endif
#ifndef MPICH_SINGLE_THREADED
    /* initialized flag/lock used to by recv queue and recv code for
       thread safety */
    MPID_Thread_lock_t initialized;
#endif
    MPID_Request_kind_t kind;
    /* completion counter */
    volatile int cc;
    /* pointer to the completion counter */
    /* This is necessary for the case when an operation is described by a
       list of requests */
    int volatile *cc_ptr;
    /* A comm is needed to find the proper error handler */
    MPID_Comm *comm;
    /* Status is needed for wait/test/recv */
    MPI_Status status;
    /* Persistent requests have their own "real" requests.  Receive requests
       have partnering send requests when src=dest. etc. */
    struct MPID_Request *partner_request;
    /* User-defined request support */
    MPI_Grequest_cancel_function *cancel_fn;
    MPI_Grequest_free_function *free_fn;
    MPI_Grequest_query_function *query_fn;
    void *grequest_extra_state;

    /* Other, device-specific information */
#ifdef MPID_DEV_REQUEST_DECL
```

```
    MPID_DEV_REQUEST_DECL
#endif
} MPID_Request;
```

Module

Request-DS

Notes

If it is necessary to remember the MPI datatype, this information is saved within the device-specific fields provided by `MPID_DEV_REQUEST_DECL`.
Requests come in many flavors, as stored in the `kind` field. It is expected that each kind of request will have its own structure type (e.g., `MPID_Request_send_t`) that extends the `MPID_Request`.

MPID_Comm — Description of the Communicator data structure

Synopsis

```
typedef struct MPID_Comm {
    int           handle;        /* value of MPI_Comm for this structure */
    volatile int  ref_count;
#if !defined(MPICH_SINGLE_THREADED)
    MPID_Thread_lock_t mutex;
#endif
    int16_t       context_id;    /* Assigned context id */
    int           remote_size;   /* Value of MPI_Comm_(remote)_size */
    int           rank;          /* Value of MPI_Comm_rank */
    MPID_VCRT     vcrt;          /* virtual connecton reference table */
    MPID_VCR *    vcr;           /* alias to the array of virtual connections
                                 in vcrt */
    MPID_VCRT     local_vcrt;    /* local virtual connecton reference table */
    MPID_VCR *    local_vcr;     /* alias to the array of local virtual
                                 connections in local vcrt */
    MPID_Attribute *attributes;    /* List of attributes */
    int           local_size;    /* Value of MPI_Comm_size for local group */
    MPID_Group   *local_group,   /* Groups in communicator. */
                 *remote_group;  /* The local and remote groups are the
                                    same for intra communicators */
    MPID_Comm_kind_t comm_kind;  /* MPID_INTRACOMM or MPID_INTERCOMM */
    char          name[MPI_MAX_OBJECT_NAME];  /* Required for MPI-2 */
    MPID_Errhandler *errhandler;  /* Pointer to the error handler structure */
    struct MPID_Comm    *local_comm; /* Defined only for intercomms, holds
                                        an intracomm for the local group */
    int           is_low_group;   /* For intercomms only, this boolean is
                                  set for all members of one of the
                                  two groups of processes and clear for
                                  the other.  It enables certain
                                  intercommunicator collective operations
                                  that wish to use half-duplex operations
```

```
                                to implement a full-duplex operation */
    struct MPID_Collops  *coll_fns; /* Pointer to a table of functions
                                            implementing the collective
                                            routines */
#ifdef MPID_HAS_HETERO
    int is_hetero;
#endif
  /* Other, device-specific information */
#ifdef MPID_DEV_COMM_DECL
    MPID_DEV_COMM_DECL
#endif
} MPID_Comm;
```

Notes

Note that the size and rank duplicate data in the groups that make up this communicator. These are used often enough that this optimization is valuable.
This definition provides only a 16-bit integer for context id's . This should be sufficient for most applications. However, extending this to a 32-bit (or longer) integer should be easy.
The virtual connection table is an explicit member of this structure. This contains the information used to contact a particular process, indexed by the rank relative to this communicator.
Groups are allocated lazily. That is, the group pointers may be null, created only when needed by a routine such as `MPI_Comm_group`. The local process ids needed to form the group are available within the virtual connection table. For intercommunicators, we may want to always have the groups. If not, we either need the `local_group` or we need a virtual connection table corresponding to the `local_group` (we may want this anyway to simplify the implementation of the intercommunicator collective routines).
The pointer to the structure `MPID_Collops` containing pointers to the collective routines allows an implementation to replace each routine on a routine-by-routine basis. By default, this pointer is null, as are the pointers within the structure. If either pointer is null, the implementation uses the generic provided implementation. This choice, rather than initializing the table with pointers to all of the collective routines, is made to reduce the space used in the communicators and to eliminate the need to include the implementation of all collective routines in all MPI executables, even if the routines are not used.
The macro `MPID_HAS_HETERO` may be defined by a device to indicate that the device supports MPI programs that must communicate between processes with different data representations (e.g., different sized integers or different byte orderings). If the device does need to define this value, it should be defined in the file `mpidpre.h`.

Module

Communicator-DS

Question

For fault tolerance, do we want to have a standard field for communicator health? For example, ok, failure detected, all (live) members of failed communicator have acked.

MPID_Group — Description of the Group data structure

Synopsis

```
typedef struct MPID_Group {
    int          handle;
    volatile int ref_count;
    int          size;           /* Size of a group */
    int          rank;           /* rank of this process relative to this
                                 group */
    int          idx_of_first_lpid;
    MPID_Group_pmap_t *lrank_to_lpid; /* Array mapping a local rank to local
                                     process number */
    /* We may want some additional data for the RMA syncrhonization calls */
  /* Other, device-specific information */
#ifdef MPID_DEV_GROUP_DECL
    MPID_DEV_GROUP_DECL
#endif
} MPID_Group;
```

The processes in the group of `MPI_COMM_WORLD` have lpid values 0 to `size`-1, where `size` is the size of `MPI_COMM_WORLD`. Processes created by `MPI_Comm_spawn` or `MPI_Comm_spawn_multiple` or added by `MPI_Comm_attach` or `MPI_Comm_connect` are numbered greater than `size - 1` (on the calling process). See the discussion of LocalPID values.
Note that when dynamic process creation is used, the pids are *not* unique across the universe of connected MPI processes. This is ok, as long as pids are interpreted *only* on the process that owns them.
Only for MPI-1 are the lpid's equal to the *global* pids. The local pids can be thought of as a reference not to the remote process itself, but how the remote process can be reached from this process. We may want to have a structure `MPID_Lpid_t` that contains information on the remote process, such as (for TCP) the hostname, ip address (it may be different if multiple interfaces are supported; we may even want plural ip addresses for stripping communication), and port (or ports). For shared memory connected processes, it might have the address of a remote queue. The lpid number is an index into a table of `MPID_Lpid_t`'s that contain this (device- and method-specific) information.

Module

Group-DS

MPID_Win — Description of the Window Object data structure.

Synopsis

```
typedef struct MPID_Win {
    int           handle;             /* value of MPI_Win for this structure */
    volatile int  ref_count;
    int fence_cnt;     /* 0 = no fence has been called;
                          1 = fence has been called */
    MPID_Errhandler *errhandler;  /* Pointer to the error handler structure */
    void *base;
    MPI_Aint    size;
    int          disp_unit;      /* Displacement unit of *local* window */
    MPID_Attribute *attributes;
    MPID_Group *start_group_ptr; /* group passed in MPI_Win_start */
```

```
    int start_assert;            /* assert passed to MPI_Win_start */
    MPI_Comm    comm;         /* communicator of window (dup) */
    volatile int my_counter;  /* completion counter for operations
                                 targeting this window */
    void **base_addrs;     /* array of base addresses of the windows of
                              all processes */
    int *disp_units;      /* array of displacement units of all windows */
    int **all_counters;    /* array of addresses of the completion
                                 counters of all processes */
#ifdef USE_THREADED_WINDOW_CODE
    /* These were causing compilation errors.  We need to figure out how to
       integrate threads into MPICH2 before including these fields. */
#ifdef HAVE_PTHREAD_H
    pthread_t wait_thread_id; /* id of thread handling MPI_Win_wait */
    pthread_t passive_target_thread_id; /* thread for passive target RMA */
#elif defined(HAVE_WINTHREADS)
    HANDLE wait_thread_id;
    HANDLE passive_target_thread_id;
#endif
#endif
    /* These are COPIES of the values so that addresses to them
       can be returned as attributes.  They are initialized by the
       MPI_Win_get_attr function */
    int  copyDispUnit;
    MPI_Aint copySize;

    char          name[MPI_MAX_OBJECT_NAME];
  /* Other, device-specific information */
#ifdef MPID_DEV_WIN_DECL
    MPID_DEV_WIN_DECL
#endif
} MPID_Win;
```

Module

Win-DS

Notes

The following 3 keyvals are defined for attributes on all MPI Window objects:

```
MPI_WIN_SIZE
MPI_WIN_BASE
MPI_WIN_DISP_UNIT
```

These correspond to the values in `length`, `start_address`, and `disp_unit`.
The communicator in the window is the same communicator that the user provided to `MPI_Win_create` (not a dup). However, each intracommunicator has a special context id that may be used if MPI communication is used by the implementation to implement the RMA operations. There is no separate window group; the group of the communicator should be used.

Question

Should a MPID_Win be defined after MPID_Segment in case the device wants to store a queue of pending put/get operations, described with MPID_Segment (or MPID_Request)s?

MPID_Op — MPI_Op structure

Synopsis

```
typedef struct MPID_Op {
     int                handle;      /* value of MPI_Op for this structure */
     volatile int       ref_count;
     MPID_Op_kind       kind;
     MPID_Lang_t        language;
     MPID_User_function function;
  } MPID_Op;
```

Notes

All of the predefined functions are commutative. Only user functions may be noncummutative, so there are two separate op types for commutative and non-commutative user-defined operations. Operations do not require reference counts because there are no nonblocking operations that accept user-defined operations. Thus, there is no way that a valid program can free an MPI_Op while it is in use.

Module

Collective-DS

MPID_Datatype — Description of the MPID Datatype structure

Synopsis

```
typedef struct MPID_Datatype {
    int           handle;            /* value of MPI_Datatype for structure */
    volatile int  ref_count;
    int           is_contig;     /* True if data is contiguous (even with
                                    a (count,datatype) pair) */
    int           n_contig_blocks; /* number of contiguous blocks in one instance of this type */
    int           size;          /* Q: maybe this should be in the dataloop? */
    MPI_Aint      extent;        /* MPI-2 allows a type to be created by
                                    resizing (the extent of) an existing
                                    type */
    MPI_Aint      ub, lb,        /* MPI-1 upper and lower bounds */
                  true_ub, true_lb; /* MPI-2 true upper and lower bounds */
    int           alignsize;     /* size of datatype to align (affects pad) */
    /* The remaining fields are required but less frequently used, and
       are placed after the more commonly used fields */
    int loopsize; /* size of loops for this datatype in bytes; derived value */
```

```
    struct MPID_Dataloop *loopinfo; /* Original loopinfo, used when
                                     * creating and when getting contents */
    int           has_sticky_ub;   /* The MPI_UB and MPI_LB are sticky */
    int           has_sticky_lb;
    int           is_permanent;  /* True if datatype is a predefined type */
    int           is_committed;  /* */

    int           eltype;      /* type of elementary datatype. Needed
                                 to implement MPI_Accumulate */

    int           loopinfo_depth; /* Depth of dataloop stack needed
                                     to process this datatype.  This
                                     information is used to ensure that
                                     no datatype is constructed that
                                     cannot be processed (see MPID_Segment) */

    struct MPID_Attribute   *attributes;    /* MPI-2 adds datatype attributes */

    int32_t       cache_id;      /* These are used to track which processes */
    /* MPID_Lpidmask mask; */         /* have cached values of this datatype */

    char          name[MPI_MAX_OBJECT_NAME];  /* Required for MPI-2 */

    /* The following is needed to efficiently implement MPI_Get_elements */
    int           n_elements;   /* Number of basic elements in this datatype */
    MPI_Aint      element_size; /* Size of each element or -1 if elements are
                                   not all the same size */
    MPID_Datatype_contents *contents;
    /* int (*free_fn)( struct MPID_Datatype * ); */ /* Function to free this datatype */
    /* Other, device-specific information */
#ifdef MPID_DEV_DATATYPE_DECL
    MPID_DEV_DATATYPE_DECL
#endif
} MPID_Datatype;
```

Notes

The `ref_count` is needed for nonblocking operations such as

```
MPI_Type_struct( ... , &newtype );
MPI_Irecv( buf, 1000, newtype, ..., &request );
MPI_Type_free( &newtype );
...
MPI_Wait( &request, &status );
```

Module

Datatype-DS

Notes

Alternatives

The following alternatives for the layout of this structure were considered. Most were not chosen because any benefit in performance or memory efficiency was outweighed by the added complexity of the implementation.
A number of fields contain only boolean inforation (`is_contig`, `has_sticky_ub`, `has_sticky_lb`, `is_permanent`, `is_committed`). These could be combined and stored in a single bit vector.
`MPI_Type_dup` could be implemented with a shallow copy, where most of the data fields, particularly the `opt_dataloop` field, would not be copied into the new object created by `MPI_Type_dup`; instead, the new object could point to the data fields in the old object. However, this requires more code to make sure that fields are found in the correct objects and that deleting the old object doesn`t invalidate the dup`ed datatype.
A related optimization would point to the `opt_dataloop` and `dataloop` fields in other datatypes. This has the same problems as the shallow copy implementation.
In addition to the separate `dataloop` and `opt_dataloop` fields, we could in addition have a separate `hetero_dataloop` optimized for heterogeneous communication for systems with different data representations.
Earlier versions of the ADI used a single API to change the `ref_count`, with each MPI object type having a separate routine. Since reference count changes are always up or down one, and since all MPI objects are defined to have the `ref_count` field in the same place, the current ADI3 API uses two routines, `MPIU_Object_add_ref` and `MPIU_Object_release_ref`, to increment and decrement the reference count.

MPID_Info — Structure of an MPID info

Synopsis

```
typedef struct MPID_Info {
    int                handle;
    struct MPID_Info   *next;
    char               *key;
    char               *value;
} MPID_Info;
```

Notes

There is no reference count because `MPI_Info` values, unlike other MPI objects, may be changed after they are passed to a routine without changing the routine's behavior. In other words, any routine that uses an `MPI_Info` object must make a copy or otherwise act on any info value that it needs.
A linked list is used because the typical `MPI_Info` list will be short and a simple linked list is easy to implement and to maintain. Similarly, a single structure rather than separate header and element structures are defined for simplicity. No separate thread lock is provided because info routines are not performance critical; they use the `common_lock` in the `MPIR_Process` structure when they need a thread lock.
This particular form of linked list (in particular, with this particular choice of the first two members) is used because it allows us to use the same routines to manage this list as are used to manage the list of free objects (in the file `src/util/mem/handlemem.c`). In particular, if lock-free routines for updating a linked list are provided, they can be used for managing the `MPID_Info` structure as well.

The MPI standard requires that keys can be no less that 32 characters and no more than 255 characters. There is no mandated limit on the size of values.

Module

Info-DS

MPID_Errhandler — Description of the error handler structure

Synopsis

```
typedef struct MPID_Errhandler {
  int                handle;
  volatile int       ref_count;
  MPID_Lang_t        language;
  MPID_Object_kind   kind;
  MPID_Errhandler_fn errfn;
  /* Other, device-specific information */
#ifdef MPID_DEV_ERRHANDLER_DECL
    MPID_DEV_ERRHANDLER_DECL
#endif
} MPID_Errhandler;
```

Notes

Device-specific information may indicate whether the error handler is active; this can help prevent infinite recursion in error handlers caused by user-error without requiring the user to be as careful. We might want to make this part of the interface so that the `MPI_xxx_call_errhandler` routines would check.
It is useful to have a way to indicate that the errhandler is no longer valid, to help catch the case where the user has freed the errhandler but is still using a copy of the `MPI_Errhandler` value. We may want to define the `id` value for deleted errhandlers.

Module

ErrHand-DS

MPID_Keyval — Structure of an MPID keyval

Synopsis

```
typedef struct MPID_Keyval {
    int                 handle;
    volatile int        ref_count;
    MPID_Lang_t         language;
    MPID_Object_kind    kind;
    void                *extra_state;
    MPID_Copy_function  copyfn;
```

```
    MPID_Delete_function delfn;
  /* other, device-specific information */
#ifdef MPID_DEV_KEYVAL_DECL
    MPID_DEV_KEYVAL_DECL
#endif
} MPID_Keyval;
```

Module

Attribute-DS

MPID_Attribute — Structure of an MPID attribute

Synopsis

```
typedef struct MPID_Attribute {
    int          handle;
    volatile int ref_count;
    MPID_Keyval  *keyval;           /* Keyval structure for this attribute */
    struct MPID_Attribute *next;    /* Pointer to next in the list */
    long        pre_sentinal;       /* Used to detect user errors in accessing
                                    the value */
    void *      value;              /* Stored value */
    long        post_sentinal;      /* Like pre_sentinal */
    /* other, device-specific information */
#ifdef MPID_DEV_ATTR_DECL
    MPID_DEV_ATTR_DECL
#endif
} MPID_Attribute;
```

Notes

Attributes don't have `ref_count`s because they don't have reference count semantics. That is, there are no shallow copies or duplicates of an attibute. An attribute is copied when the communicator that it is attached to is duplicated. Subsequent operations, such as `MPI_Comm_attr_free`, can change the attribute list for one of the communicators but not the other, making it impractical to keep the same list. (We could defer making the copy until the list is changed, but even then, there would be no reference count on the individual attributes.)
A pointer to the keyval, rather than the (integer) keyval itself is used since there is no need within the attribute structure to make it any harder to find the keyval structure.
The attribute value is a `void *`. If `sizeof(MPI_Fint)` > `sizeof(void*)`, then this must be changed (no such system has been encountered yet). For the Fortran 77 routines in the case where `sizeof(MPI_Fint)` < `sizeof(void*)`, the high end of the `void *` value is used. That is, we cast it to `MPI_Fint *` and use that value.

Module

Attribute-DS

A.3 Threads

Thread support in MPI requires both a thread-safe device and thread-safe handling of MPI objects that may be shared among threads. These routines provide for atomic updates to reference counts on MPI objects. Other routines provide thread locks on a communicator basis, as well as a process-wide thread lock that may be used when such a coarse-grain lock is sufficient.

MPID_MAX_THREAD_LEVEL — Indicates the maximum level of thread support provided at compile time.

Values

Any of the `MPI_THREAD_xxx` values (these are preprocessor-time constants)

Notes

The macro `MPID_MAX_THREAD_LEVEL` defines the maximum level of thread support provided, and may be used at compile time to remove thread locks and other code needed only in a multithreaded environment.
A typical use is

```
#if MPID_MAX_THREAD_LEVEL >= MPI_THREAD_MULTIPLE
   lock((r)->lock_ptr);
   (r)->ref_count++;
   unlock((r)->lock_ptr);
#else
   (r)->ref_count ++;
#fi
```

Module

Environment-DS

MPIU_Object_add_ref — Increment the reference count for an MPI object

Synopsis

```
.vb
MPIU_Object_add_ref( MPIU_Object *ptr )
.ve
```

Input Parameter

ptr Pointer to the object.

Notes

In an unthreaded implementation, this function will usually be implemented as a single-statement macro. In an `MPI_THREAD_MULTIPLE` implementation, this routine must implement an atomic increment operation, using, for example, a lock on datatypes or special assembly code such as

```
try-again:
   load-link          refcount-address to r2
   add                1 to r2
   store-conditional  r2 to refcount-address
   if failed branch to try-again:
```

on RISC architectures or

```
lock
inc                   refcount-address or
```

on IA32; "lock" is a special opcode prefix that forces atomicity. This is not a separate instruction; however, the GNU assembler expects opcode prefixes on a separate line.

Module

MPID_CORE

Question

This accesses the `ref_count` member of all MPID objects. Currently, that member is typed as `volatile int`. However, for a purely polling, thread-funnelled application, the `volatile` is unnecessary. Should MPID objects use a `typedef` for the `ref_count` that can be defined as `volatile` only when needed? For now, the answer is no; there isn't enough to be gained in that case.

MPIU_Object_release_ref — Decrement the reference count for an MPI object

Synopsis

```
.vb
MPIU_Object_release_ref( MPIU_Object *ptr, int *inuse_ptr )
.ve
```

Input Parameter

objptr Pointer to the object.

Output Parameter

inuse_ptr Pointer to the value of the reference count after decrementing. This value is either zero or non-zero. See below for details.

Notes

In an unthreaded implementation, this function will usually be implemented as a single-statement macro. In an `MPI_THREAD_MULTIPLE` implementation, this routine must implement an atomic decrement operation, using, for example, a lock on datatypes or special assembly code such as

```
try-again:
   load-link          refcount-address to r2
   sub                1 to r2
   store-conditional  r2 to refcount-address
   if failed branch to try-again:
   store              r2 to newval_ptr
```

on RISC architectures or

```
   lock
   dec                   refcount-address
   if zf store 0 to newval_ptr else store 1 to newval_ptr
```

on IA32; "lock" is a special opcode prefix that forces atomicity. This is not a separate instruction; however, the GNU assembler expects opcode prefixes on a separate line. `zf` is the zero flag; this is set if the result of the operation is zero. Implementing a full decrement-and-fetch would require more code and the compare and swap instruction.
Once the reference count is decremented to zero, it is an error to change it. A correct MPI program will never do that, but an incorrect one (particularly a multithreaded program with a race condition) might.
The following code is *invalid*:

```
MPID_Object_release_ref( datatype_ptr );
if (datatype_ptr->ref_count == 0) MPID_Datatype_free( datatype_ptr );
```

In a multi-threaded implementation, the value of `datatype_ptr->ref_count` may have been changed by another thread, resulting in both threads calling `MPID_Datatype_free`. Instead, use

```
MPID_Object_release_ref( datatype_ptr, &inUse );
if (!inuse)
    MPID_Datatype_free( datatype_ptr );
```

Module

MPID_CORE

MPID_Comm_thread_lock — Acquire a thread lock for a communicator

Synopsis

```
.vb
void MPID_Comm_thread_lock( MPID_Comm *comm )
.ve
```

Input Parameter

comm Communicator to lock

Notes

This routine acquires a lock among threads in the same MPI process that may use this communicator. In all MPI thread modes except for `MPI_THREAD_MULTIPLE`, this can be a no-op. In an MPI implementation that does not provide `MPI_THREAD_MULTIPLE`, this may be a macro.
It is invalid for a thread that has acquired the lock to attempt to acquire it again. The lock must be released by `MPID_Comm_thread_unlock`.
Note that there is also a common per-process lock (`common_lock`). That lock should be used instead of a lock on lock on `MPI_COMM_WORLD` when a lock across all threads is required.
A high-quality implementation may wish to provide fair access to the lock.
In general, the MPICH implementation tries to avoid using locks because they can cause problems such as livelock and deadlock, particularly when an error occurs. However, the semantics of MPI collective routines make it difficult to avoid using locks. Further, good programming practice by MPI programmers should be to avoid having multiple threads using the same communicator.

Module

Communicator

See Also

`MPID_Comm_thread_unlock`

Questions

Do we also need versions of this for datatypes and window objects? For example, communicators, datatypes, and window objects all have attributes; do we need a thread lock for each type? Should we instead have an MPI Object, on which some common operations, such as thread lock, reference count, and name are implemented? Note that there is a common lock for operations that are infrequently performed and do not require fine-grain locking.

MPID_Comm_thread_unlock — Release a thread lock for a communicator

Synopsis

```
.vb
void MPID_Comm_thread_unlock( MPID_Comm *comm )
.ve
```

Input Parameter

comm Communicator to unlock

Module

Communicator

See Also

MPID_Comm_thread_lock

B Basic Point-to-Point

This section provides definitions for the point-to-point communication functions. These roughly parallel their MPI counterparts, with the exception that all (even MPID_Send and MPID_Recv are non-blocking, and the optional MPID_tBsend routine that may be used to improve the performance of buffered sends.

MPID_Send — MPID entry point for MPI_Send

Synopsis

```
int MPID_Send( const void *buf, int count, MPI_Datatype datatype,
               int dest, int tag, MPID_Comm *comm, int context_offset,
               MPID_Request **request )
```

Notes

The only difference between this and MPI_Send is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, a context id offset is provided in addition to the communicator, and a request may be returned. The context offset is added to the context of the communicator to get the context it used by the message. A request is returned only if the ADI implementation was unable to complete the send of the message. In that case, the usual MPI_Wait logic should be used to complete the request. This approach is used to allow a simple implementation of the ADI. The ADI is free to always complete the message and never return a request.

Module

Communication

MPID_Ssend — MPID entry point for MPI_Ssend

Synopsis

```
int MPID_Ssend( const void *buf, int count, MPI_Datatype datatype,
               int dest, int tag, MPID_Comm *comm, int context_offset,
               MPID_Request **request )
```

Notes

The only difference between this and MPI_Ssend is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, a context id offset is provided in addition to the communicator, and a request may be returned. The context offset is added to the context of the communicator to get the

context it used by the message. A request is returned only if the ADI implementation was unable to complete the send of the message. In that case, the usual `MPI_Wait` logic should be used to complete the request. This approach is used to allow a simple implementation of the ADI. The ADI is free to always complete the message and never return a request.

Module

Communication

MPID_Rsend — MPID entry point for MPI_Rsend

Synopsis

```
int MPID_Rsend( const void *buf, int count, MPI_Datatype datatype,
               int dest, int tag, MPID_Comm *comm, int context_offset,
               MPID_Request **request )
```

Notes

The only difference between this and `MPI_Rsend` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, a context id offset is provided in addition to the communicator, and a request may be returned. The context offset is added to the context of the communicator to get the context it used by the message. A request is returned only if the ADI implementation was unable to complete the send of the message. In that case, the usual `MPI_Wait` logic should be used to complete the request. This approach is used to allow a simple implementation of the ADI. The ADI is free to always complete the message and never return a request.

Module

Communication

MPID_Isend — MPID entry point for MPI_Isend

Synopsis

```
int MPID_Isend( const void *buf, int count, MPI_Datatype datatype,
               int dest, int tag, MPID_Comm *comm, int context_offset,
               MPID_Request **request )
```

Notes

The only difference between this and `MPI_Isend` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Issend — MPID entry point for MPI_Issend

Synopsis

```
int MPID_Issend( const void *buf, int count, MPI_Datatype datatype,
                int dest, int tag, MPID_Comm *comm, int context_offset,
                MPID_Request **request )
```

Notes

The only difference between this and `MPI_Issend` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Irsend — MPID entry point for MPI_Irsend

Synopsis

```
int MPID_Irsend( const void *buf, int count, MPI_Datatype datatype,
                int dest, int tag, MPID_Comm *comm, int context_offset,
                MPID_Request **request )
```

Notes

The only difference between this and `MPI_Irsend` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_tBsend — Attempt a send and return if it would block

Synopsis

```
int MPID_tBsend( const void *buf, int count, MPI_Datatype datatype,
                 int dest, int tag, MPID_Comm *comm, int context_offset )
```

Notes

This has the semantics of `MPI_Bsend`, except that it returns the internal error code `MPID_WOULD_BLOCK` if the message can't be sent immediately (t is for "try").
The reason that this interface is chosen over a query to check whether a message *can* be sent is that the query approach is not thread-safe. Since the decision on whether a message can be sent without blocking depends (among other things) on the state of flow control managed by the device, this approach also gives the device the greatest freedom in implementing flow control. In particular, if another MPI process can change the flow control parameters, then even in a single-threaded implementation, it would not be safe to return, for example, a message size that could be sent with `MPI_Bsend`.
This routine allows an MPI implementation to optimize `MPI_Bsend` for the case when the message can be delivered without blocking the calling process. An ADI implementation is free to have this routine always return `MPID_WOULD_BLOCK`, but is encouraged not to.
To allow the MPI implementation to avoid trying this routine when it is not implemented by the ADI, the C preprocessor constant `MPID_HAS_TBSEND` should be defined if this routine has a nontrivial implementation.
This is an optional routine. The MPI code for `MPI_Bsend` will attempt to call this routine only if the device defines `MPID_HAS_TBSEND`.

Module

Communication

MPID_Recv — MPID entry point for MPI_Recv

Synopsis

```
int MPID_Recv( void *buf, int count, MPI_Datatype datatype,
               int source, int tag, MPID_Comm *comm, int context_offset,
               MPI_Status *status, MPID_Request **request )
```

Notes

The only difference between this and `MPI_Recv` is that the basic error checks (e.g., valid communicator, datatype, source, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, a context id offset is provided in addition to the communicator, and a request may be returned. The context offset is added to the context of the communicator to get the context it used by the message. As in `MPID_Send`, the request is returned only if the operation did not complete. Conversely, the status object is populated with valid information only if the operation completed.

Module

Communication

MPID_Irecv — MPID entry point for MPI_Irecv

Synopsis

```
int MPID_Irecv( void *buf, int count, MPI_Datatype datatype,
               int source, int tag, MPID_Comm *comm, int context_offset,
               MPID_Request **request )
```

Notes

The only difference between this and `MPI_Irecv` is that the basic error checks (e.g., valid communicator, datatype, source, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Request_release — Release a request

Synopsis

```
void MPID_Request_release(MPID_Request *)
```

Input Parameter

request request to release

Notes

This routine is called to release a reference to request object. If the reference count of the request object has reached zero, the object will be deallocated.

Module

Request

MPID_Cancel_send — Cancel the indicated send request

Synopsis

```
int MPID_Cancel_send(MPID_Request *)
```

Input Parameter

request Send request to cancel

Return Value

MPI error code.

Notes

Cancel is a tricky operation, particularly for sends. Read the discussion in the MPI-1 and MPI-2 documents carefully. This call only requests that the request be cancelled; a subsequent wait or test must first succeed (i.e., the request completion counter must be zeroed).

Module

Request

MPID_Cancel_recv — Cancel the indicated recv request

Synopsis

```
int MPID_Cancel_recv(MPID_Request *)
```

Input Parameter

request Receive request to cancel

Return Value

MPI error code.

Notes

This cancels a pending receive request. In many cases, this is implemented by simply removing the request from a pending receive request queue. However, some ADI implementations may maintain these queues in special places, such as within a NIC (Network Interface Card). This call only requests that the request be cancelled; a subsequent wait or test must first succeed (i.e., the request completion counter must be zeroed).

Module

Request

MPID_Iprobe — Look for a matching request in the receive queue but do not remove or return it

Synopsis

```
int MPID_Iprobe(int, int, MPID_Comm *, int, int *, MPI_Status *)
```

Input Parameters

source rank to match (or `MPI_ANY_SOURCE`)
tag Tag to match (or `MPI_ANY_TAG`)
comm communicator to match.
context_offset
context id offset of communicator to match

Output Parameter

flag true if a matching request was found, false otherwise.
status `MPI_Status` set as defined by `MPI_Iprobe` (only valid when return flag is true).

Return Value

Error Code.

Notes

Note that the values returned in `status` will be valid for a subsequent MPI receive operation only if no other thread attempts to receive the same message. (See the discussion of probe in Section 8.7.2 (Clarifications) of the MPI-2 standard.)
Providing the `context_offset` is necessary at this level to support the way in which the MPICH implementation uses context ids in the implementation of other operations. The communicator is present to allow the device to use message-queues attached to particular communicators or connections between processes.
Devices that rely solely on polling to make progress should call MPID_Progress_poke() (or some equivalent function) if a matching request could not be found. This insures that progress continues to be made even if the application is calling MPI_Iprobe() from within a loop not containing calls to any other MPI functions.

Module

Request

MPID_Probe — Block until a matching request is found and return information about it

Synopsis

```
int MPID_Probe(int, int, MPID_Comm *, int, MPI_Status *)
```

Input Parameters

source rank to match (or `MPI_ANY_SOURCE`)
tag Tag to match (or `MPI_ANY_TAG`)
comm communicator to match.
context_offset
context id offset of communicator to match

Output Parameter

status `MPI_Status` set as defined by `MPI_Probe`

Return Value

Error code.

Notes

Note that the values returned in `status` will be valid for a subsequent MPI receive operation only if no other thread attempts to receive the same message. (See the discussion of probe in Section 8.7.2 Clarifications of the MPI-2 standard.)
Providing the `context_offset` is necessary at this level to support the way in which the MPICH implementation uses context ids in the implementation of other operations. The communicator is present to allow the device to use message-queues attached to particular communicators or connections between processes.

Module

Request

C Persistent Point-to-Point

This section provides the definitions for the persistent communication routines. Note that there is no `MPID_Start` routine; we currently believe that `MPID_Startall` is all that is required, as it is easy to implement `MPI_Start` with `MPID_Startall`, and `MPID_Startall` is required to allow the device the option of scheduling communication for all of the requests passed to it.

MPID_Send_init — MPID entry point for MPI_Send_init

Synopsis

```
int MPID_Send_init( const void *buf, int count, MPI_Datatype datatype,
                    int dest, int tag, MPID_Comm *comm, int context_offset,
                    MPID_Request **request )
```

Notes

The only difference between this and `MPI_Send_init` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Ssend_init — MPID entry point for MPI_Ssend_init

Synopsis

```
int MPID_Ssend_init( const void *buf, int count, MPI_Datatype datatype,
                     int dest, int tag, MPID_Comm *comm, int context_offset,
                     MPID_Request **request )
```

Notes

The only difference between this and `MPI_Ssend_init` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Rsend_init — MPID entry point for MPI_Rsend_init

Synopsis

```
int MPID_Rsend_init( const void *buf, int count, MPI_Datatype datatype,
                     int dest, int tag, MPID_Comm *comm, int context_offset,
                     MPID_Request **request )
```

Notes

The only difference between this and `MPI_Rsend_init` is that the basic error checks (e.g., valid communicator, datatype, dest, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Recv_init — MPID entry point for MPI_Recv_init

Synopsis

```
int MPID_Recv_init( void *buf, int count, MPI_Datatype datatype,
                    int source, int tag, MPID_Comm *comm, int context_offset,
                    MPID_Request **request )
```

Notes

The only difference between this and `MPI_Recv_init` is that the basic error checks (e.g., valid communicator, datatype, source, and tag) have been made, the MPI opaque objects have been replaced by MPID objects, and a context id offset is provided in addition to the communicator. This offset is added to the context of the communicator to get the context it used by the message.

Module

Communication

MPID_Startall — MPID entry point for MPI_Startall

Synopsis

```
int MPID_Startall(int count, MPID_Request *requests[])
```

Notes

The only difference between this and `MPI_Startall` is that the basic error checks (e.g., count) have been made, and the MPI opaque objects have been replaced by pointers to MPID objects.

Rationale

This allows the device to schedule communication involving multiple requests, whereas an implementation built on just `MPID_Start` would force the ADI to initiate the communication in the order encountered.

D Generalized Requests

This section provides the interface to the generalized (user-defined) requests.

MPID_Request_create — Create and return a bare request

Synopsis

```
MPID_Request * MPID_Request_create(void)
```

Return value

A pointer to a new request object.

Notes

This routine is intended for use by `MPI_Grequest_start` only. Note that once a request is created with this routine, any progress engine must assume that an outside function can complete a request with `MPID_Request_set_completed`.
The request object returned by this routine should be initialized such that ref_count is one and handle contains a valid handle referring to the object.

E Data Segment

F Communication Buffer management

The routines in this section fill two roles. In the implementation of routines such as `MPID_Isend` where the only device-specific routines are the few in `MPID_CORE`, it is necessary to have the ability to pack and unpack partial messages. Consider the case where the user calls `MPI_Isend( buf, 1, my_huge_indexed_type, ... )` where the total amount of data being sent is 100 MB. Since `MPID_CORE` can only move contiguous data, we need to convert the described data into a sequence of contiguous memory segments of some reasonable size (we dont want to have to allocate a 100 MB temporary buffer). Thus, to implement this operation, we need a routine that can be called repeatedly to get the next m bytes from the users described data buffer. Further, we need to be able to pause and remember where we are in processing a datatype. The MPI unpack routine does not have this flexibility.

The other place where these routines are needed is in the implementation of efficient versions of the MPI collective communication routines. These algorithms often need to look at the message as a range of bytes from which segments are extracted and moved. The implementation of the MPI collective routines provided with MPICH will use these routines express the algorithms.

These routines also provide a way to specify a data area that may be used in store-and-forward algorithms, without requiring copies to and from an intermediate device buffer.

Note that `MPID_Requests` include a segment descriptoin within them.

G Process Topology

This section describes a routine that allows the device to communicate information on the process topology to the MPI process topology routines. This routine is designed for hierarchically-organized machines. Additional routines for other topologies (e.g., a scalable mesh topology) will be considered as needed.

MPID_Topo_cluster_info — Return information on the hierarchy of interconnections

Synopsis

```
int MPID_Topo_cluster_info( MPID_Comm *comm,
                            int *levels, int my_cluster[], int my_rank[] )
```

Input Parameter

comm Communicator to study. May be `NULL`, in which case `MPI_COMM_WORLD` is the effective communicator.

Output Parameters

levels The number of levels in the hierarchy. To simplify the use of this routine, the maximum value is `MPID_TOPO_CLUSTER_MAX_LEVELS` (typically 8 or less).
my_cluster For each level, the id of the cluster that the calling process belongs to.
my_rank For each level, the rank of the calling process in its cluster

Notes

This routine returns a description of the system in terms of nested clusters of processes. Levels are numbered from zero. At each level, each process may belong to no more than cluster; if a process is in any cluster at level i, it must be in some cluster at level i-1.
The communicator argument allows this routine to be used in the dynamic process case (i.e., with communicators that are created after `MPI_Init` and that involve processes that are not part of `MPI_COMM_WORLD`).
For non-hierarchical systems, this routine simply returns a single level containing all processes.

Sample Outputs

For a single, switch-connected cluster or a uniform-memory-access (UMA) symmetric multiprocessor (SMP), the return values could be

```
level       my_cluster         my_rank
0           0                  rank in comm_world
```

This is also a valid response for *any* device.
For a switch-connected cluster of 2 processor SMPs

```
level       my_cluster         my_rank
0           0                  rank in comm_world
1           0 to p/2           0 or 1
```

where the value each process on the same SMP has the same value for `my_cluster[1]` and a different value for `my_rank[1]`.
For two SMPs connected by a network,

```
level       my_cluster         my_rank
0           0                  rank in comm_world
1           0 or 1             0 to # on SMP
```

An example with more than 2 levels is a collection of clusters, each with SMP nodes.

Limitations

This approach does not provide a representations for topologies that are not hierarchical. For example, a mesh interconnect is a single-level cluster in this view.

Module

Topology

H Progress Engine

The progress engine provides a way for the MPI implementation and the ADI implementation to coordinate progress on communication. The design makes no assumptions about whether the ADI uses polling or a separate communication thread or interrupts or any other mechanism to make progress.

Here are several implementation sketches. First, a polling implementation for a single-threaded implementation:

```
MPID_Progress_start - no-op
MPID_Progress_end - no-op
MPID_Progress_test - select with no wait
MPID_Progress_wait - select with infinite wait
MPID_Progress_poke - select with no wait
```

A non-polling, single-threaded implementation that used a separate thread for communication could use (this isn't correct yet)

```
MPID_Progress_start - Set flag indicating checks in progress
MPID_Progress_end - Clear flag
MPID_Progress_test - yield to communication thread
MPID_Progress_wait - if no completions since flag set,
                     wait on condition variable.  Otherwise, return.
MPID_Progress_poke - either no-op or yield to communication thread
```

See the generalized request functions in Section D for some additional requirements on the progress engine.

MPID_Progress_start — Begin a block of operations that check the completion counters in requests.

Synopsis

```
void MPID_Progress_start(void)
```

Notes

This routine is used to inform the progress engine that a block of code will examine the completion counter of some `MPID_Request` objects and then call one of `MPID_Progress_end`, `MPID_Progress_wait`, or `MPID_Progress_test`.
This routine is needed to properly implement blocking tests when multithreaded progress engines are used. In a single-threaded implementation of the ADI, this may be defined as an empty macro.

Module

Communication

MPID_Progress_end — End a block of operations begun with MPID_Progress_start

Synopsis

```
void MPID_Progress_end(void)
```

Notes

This instructs the progress engine to end the block begun with `MPID_Progress_start`. The progress engine is not required to check for any pending communication.
The purpose of this call is to release any locks initiated by `MPID_Progess_start`. It is typically used when checks of the relevant request completion counters found a completed request. In a single threaded ADI implementation, this may be defined as an empty macro.

Module

Communication

MPID_Progress_test — Check for communication since `MPID_Progress_start`

Synopsis

```
int MPID_Progress_test(void)
```

Return value

An mpi error code.

Notes

Like `MPID_Progress_end` and `MPID_Progress_wait`, this completes the block begun with `MPID_Progress_start`. Unlike `MPID_Progress_wait`, it is a nonblocking call. It returns the number of communication events, which is only indicates the maximum number of separate requests that were completed. The only restriction is that if the completion status of any request changed between `MPID_Progress_start` and `MPID_Progress_test`, the return value must be at least one.
This function used to return TRUE if one or more requests have completed, FALSE otherwise. This functionality was not used so we removed it.

Module

Communication

MPID_Progress_wait — Wait for some communication since `MPID_Progress_start`

Synopsis

```
int MPID_Progress_wait(void)
```

Return value

An mpi error code.

Notes

This instructs the progress engine to wait until some communication event happens since `MPID_Progress_start` was called. This call blocks the calling thread (only, not the process). Before returning, it releases the block begun with `MPID_Progress_start`.

Module

Communication

MPID_Progress_poke — Allow a progress engine to check for pending communication

Synopsis

```
int MPID_Progress_poke(void)
```

Return value

An mpi error code.

Notes

This routine provides a way to invoke the progress engine in a polling implementation of the ADI. This routine must be nonblocking.
A multithreaded implementation is free to define this as an empty macro.

Module

Communication

I Starting and stopping

MPID_Init — Initialize the device

Synopsis

```
int MPID_Init( int *argc_p, char ***argv_p,
               int requested, int *provided,
               /* MPID_Comm **parent_comm,*/ int *has_args, int *has_env )
```

Input Parameters

argc_p Pointer to the argument count
argv_p Pointer to the argument list
requested Requested level of thread support. Values are the same as for the `required` argument to `MPI_Init_thread`, except that we define an enum for these values.

Output Parameter

provided Provided level of thread support. May be less than the requested level of support.
parent_comm `MPID_Comm` of parent. This is null for all MPI-1 uses and for processes that are *not* started with `MPI_Comm_spawn` or `MPI_Comm_spawn_multiple`.
has_args Set to true if `argc_p` and `argv_p` contain the command line arguments. See below.
has_env Set to true if the environment of the process has been set as the user expects. See below.

Return value

Returns `0` on success and an MPI error code on failure. Failure can happen when, for example, the device is unable to start or contact the number of processes specified by the `mpiexec` command.

Notes

Null arguments for `argc_p` and `argv_p` *must* be valid (see MPI-2, section 4.2)
Multi-method devices should initialize each method within this call. They can use environment variables and/or command-line arguments to decide which methods to initialize (but note that they must not *depend* on using command-line arguments).
This call also initializes all MPID data needed by the device. This includes the `MPID_Request`s and any other data structures used by the device.
The arguments `has_args` and `has_env` indicate whether the process was started with command-line arguments or environment variables. In some cases, only the root process is started with these values; in others, the startup environment ensures that each process receives the command-line arguments and environment variables that the user expects. While the MPI standard makes no requirements that command line arguments or environment variables are provided to all processes, most users expect a common environment. These variables allow an MPI implementation (that is based on ADI-3) to provide both of these by making use of MPI communication after `MPID_Init` is called but before `MPI_Init` returns to the user, if the process management environment does not provide this service.
This routine is used to implement both `MPI_Init` and `MPI_Init_thread`.
Setting the environment requires a `setenv` function. Some systems may not have this. In that case, the documentation must make clear that the environment may not be propagated to the generated processes.
The `parent_comm` argument may not be the right interface.

Module

MPID_CORE

Questions

The values for `has_args` and `has_env` are boolean. They could be more specific. For example, the value could indicate the rank in `MPI_COMM_WORLD` of a process that has the values; the value `MPI_ANY_SOURCE` (or a `-1`) could indicate that the value is available on all processes (including this one). We may want this since otherwise the processes may need to determine whether any process

needs the command line. Another option would be to use positive values in the same way that the `color` argument is used in `MPI_Comm_split`; a negative value indicates the member of the processes with that color that has the values of the command line arguments (or environment). This allows for non-SPMD programs.

Do we require that the startup environment (e.g., whatever `mpiexec` is using to start processes) is responsible for delivering the command line arguments and environment variables that the user expects? That is, if the user is running an SPMD program, and expects each process to get the same command line argument, who is responsible for this? The `has_args` and `has_env` values are intended to allow the ADI to handle this while taking advantage of any support that the process manager framework may provide.

Alternately, how do we find out from the process management environment whether it took care of the environment or the command line arguments? Do we need a `PMI_Env_query` function that can answer these questions dynamically (in case a different process manager is used through the same interface)?

Can we fix the Fortran command-line arguments? That is, can we arrange for `iargc` and `getarg` (and the POSIX equivalents) to return the correct values? See, for example, the Absoft implementations of `getarg`. We could also contact PGI about the Portland Group compilers, and of course the `g77` source code is available. Does each process have the same values for the environment variables when this routine returns?

If we don't require that all processes get the same argument list, we need to find out if they did anyway so that `MPI_Init_thread` can fixup the list for the user. This argues for another return value that flags how much of the environment the `MPID_Init` routine set up so that the `MPI_Init_thread` call can provide the rest. The reason for this is that, even though the MPI standard does not require it, a user-friendly implementation should, in the SPMD mode, give each process the same environment and argument lists unless the user explicitly directed otherwise.

How does this interface to PMI? Do we need to know anything? Should this call have an info argument to support PMI?

The following questions involve how environment variables and command line arguments are used to control the behavior of the implementation. Many of these values must be determined at the time that `MPID_Init` is called. These all should be considered in the context of the parameter routines described in the MPICH2 Design Document.

Are there recommended environment variable names? For example, in ADI-2, there are many debugging options that are part of the common device. In MPI-2, we can't require command line arguments, so any such options must also have environment variables. E.g., `MPICH_ADI_DEBUG` or `MPICH_ADI_DB`.

Names that are explicitly prohibited? For example, do we want to reserve any names that `MPI_Init_thread` (as opposed to `MPID_Init`) might use?

How does information on command-line arguments and environment variables recognized by the device get added to the documentation?

What about control for other impact on the environment? For example, what signals should the device catch (e.g., `SIGFPE`? `SIGTRAP`?)? Which of these should be optional (e.g., ignore or leave signal alone) or selectable (e.g., port to listen on)? For example, catching `SIGTRAP` causes problems for `gdb`, so we'd like to be able to leave `SIGTRAP` unchanged in some cases.

Another environment variable should control whether fault-tolerance is desired. If fault-tolerance is selected, then some collective operations will need to use different algorithms and most fatal errors detected by the MPI implementation should abort only the affected process, not all processes.

MPID_Finalize — Perform the device-specific termination of an MPI job

Synopsis

```
int MPID_Finalize(void)
```

Return Value

`MPI_SUCCESS` or a valid MPI error code. Normally, this routine will return `MPI_SUCCESS`. Only in extrordinary circumstances can this routine fail; for example, if some process stops responding during the finalize step. In this case, `MPID_Finalize` should return an MPI error code indicating the reason that it failed.

Notes

Module

MPID_CORE

Questions

Need to check the MPI-2 requirements on `MPI_Finalize` with respect to things like which process must remain after `MPID_Finalize` is called.

MPID_Abort — Abort at least the processes in the specified communicator.

Synopsis

```
int MPID_Abort( MPID_Comm *comm, int mpi_errno, int exit_code )
```

Input Parameters

comm	Communicator of processes to abort
mpi_errno	MPI error code containing the reason for the abort
exit_code	Exit code to return to the calling environment. See notes.

Return value

`MPI_SUCCESS` or an MPI error code. Normally, this routine should not return, since the calling process must be a member of the communicator. However, under some circumstances, the `MPID_Abort` might fail; in this case, returning an error indication is appropriate.

Notes

In a fault-tolerant MPI implementation, this operation should abort *only* the processes in the specified communicator. Any communicator that shares processes with the aborted communicator becomes invalid. For more details, see (paper not yet written on fault-tolerant MPI).
In particular, if the communicator is `MPI_COMM_SELF`, only the calling process should be aborted.
The `exit_code` is the exit code that this particular process will attempt to provide to the `mpiexec` or other program invocation environment. See `mpiexec` for a discussion of how exit codes from many processes may be combined.
An external agent that is aborting processes can invoke this with either `MPI_COMM_WORLD` or `MPI_COMM_SELF`. For example, if the process manager wishes to abort a group of processes, it should cause `MPID_Abort` to be invoked with `MPI_COMM_SELF` on each process in the group.

Question

An alternative design is to provide an `MPID_Group` instead of a communicator. This would allow a process manager to ask the ADI to kill an entire group of processes without needing a communicator. However, the implementation of `MPID_Abort` will either do this by communicating with other processes or by requesting the process manager to kill the processes. That brings up this question: should `MPID_Abort` use `PMI` to kill processes? Should it be required to notify the process manager? What about persistent resources (such as SYSV segments or forked processes)?
This suggests that for any persistent resource, an exit handler be defined. These would be executed by `MPID_Abort` or `MPID_Finalize`. See the implementation of `MPI_Finalize` for an example of exit callbacks. In addition, code that registered persistent resources could use persistent storage (i.e., a file) to record that information, allowing cleanup utilities (such as `mpiexec`) to remove any resources left after the process exits.
`MPI_Finalize` requires that attributes on `MPI_COMM_SELF` be deleted before anything else happens; this allows libraries to attach end-of-job actions to `MPI_Finalize`. It is valuable to have a similar capability on `MPI_Abort`, with the caveat that `MPI_Abort` may not guarantee that the run-on-abort routines were called. This provides a consistent way for the MPICH implementation to handle freeing any persistent resources. However, such callbacks must be limited since communication may not be possible once `MPI_Abort` is called. Further, any callbacks must guarantee that they have finite termination.
One possible extension would be to allow *users* to add actions to be run when `MPI_Abort` is called, perhaps through a special attribute value applied to `MPI_COMM_SELF`. Note that is is incorrect to call the delete functions for the normal attributes on `MPI_COMM_SELF` because MPI only specifies that those are run on `MPI_Finalize` (i.e., normal termination).

Module

MPID_CORE

J Information about the device

MPID_Get_processor_name — Return the name of the current processor

Synopsis

```
int MPID_Get_processor_name( char *name, int *resultlen)
```

Output Parameters

name	A unique specifier for the actual (as opposed to virtual) node. This must be an array of size at least `MPI_MAX_PROCESSOR_NAME`.
resultlen	Length (in characters) of the name

Notes

The name returned should identify a particular piece of hardware; the exact format is implementation defined. This name may or may not be the same as might be returned by `gethostname`, `uname`, or `sysinfo`.
This routine is essentially an MPID version of `MPI_Get_processor_name` . It must be part of the device because not all environments support calls to return the processor name.

K RMA

(not yet defined)

K.1 Memory Allocation for RMA

MPID_Mem_alloc — Allocate memory suitable for passive target RMA operations

Synopsis

```
void *MPID_Mem_alloc( size_t size, MPID_Info *info )
```

Input Parameter

size	Number of types to allocate.
info	Info object

Return value

Pointer to the allocated memory. If the memory is not available, returns null.

Notes

This routine is used to implement `MPI_Alloc_mem`. It is for that reason that there is no communicator argument.
This memory may *only* be freed with `MPID_Mem_free`.
This is a *local*, not a collective operation. It functions more like a good form of `malloc` than collective shared-memory allocators such as the `shmalloc` found on SGI systems.
Implementations of this routine may wish to use `MPID_Memory_register`. However, this routine has slighly different requirements, so a separate entry point is provided.

Question

Since this takes an info object, should there be an error routine in the case that the info object contains an error?

Module

Win

MPID_Mem_free — Frees memory allocated with `MPID_Mem_alloc`

Synopsis

```
int MPID_Mem_free( void *ptr )
```

Input Parameter

ptr Pointer to memory allocated by `MPID_Mem_alloc`.

Return value

`MPI_SUCCESS` if memory was successfully freed; an MPI error code otherwise.

Notes

The return value is provided because it may not be easy to validate the value of `ptr` without attempting to free the memory.

Module

Win

MPID_Mem_was_alloced — Return true if this memory was allocated with `MPID_Mem_alloc`

Synopsis

```
int MPID_Mem_was_alloced( void *ptr )
```

Input Parameters

ptr Address of memory
size Size of reqion in bytes.

Return value

True if the memory was allocated with `MPID_Mem_alloc`, false otherwise.

Notes

This routine may be needed by `MPI_Win_create` to ensure that the memory for passive target RMA operations was allocated with `MPI_Mem_alloc`. This may be used, for example, for ensuring that memory used with passive target operations was allocated with `MPID_Mem_alloc`.

Module

Win

L Dynamic Processes

(not yet designed)

M Collective Communication

(not yet designed)

N Connections and Local Process Ids

These routines are used to manage connections. MPI Communicators contain references to these tables; group operations use the local process id (see `MPID_VCR_Get_lpid`) to identify processes.

MPID_VCRT_Create — Create a virtual connection reference table

Synopsis

```
int MPID_VCRT_Create(int size, MPID_VCRT *vcrt_ptr)
```

MPID_VCRT_Add_ref — Add a reference to a VCRT

Synopsis

```
int MPID_VCRT_Add_ref(MPID_VCRT vcrt)
```

MPID_VCRT_Release — Release a reference to a VCRT

Synopsis

```
int MPID_VCRT_Release(MPID_VCRT vcrt)
```

MPID_VCRT_Get_ptr —

Synopsis

```
int MPID_VCRT_Get_ptr(MPID_VCRT vcrt, MPID_VCR **vc_pptr)
```

MPID_VCR_Dup —

Synopsis

```
int MPID_VCR_Dup(MPID_VCR orig_vcr, MPID_VCR * new_vcr)
```

MPID_VCR_Get_lpid — Get the local process id that corresponds to a virtual connection reference.

Synopsis

```
int MPID_VCR_Get_lpid(MPID_VCR vcr, int * lpid_ptr)
```

Notes

The local process ids are described elsewhere. Basically, they are a nonnegative number by which this process can refer to other processes to which it is connected. These are local process ids because different processes may use different ids to identify the same target process

O Timers

Timer support is defined so that the device may either provide or use the timers. We expect most devices to use one of the timers defined in `src/mpi/timer`

MPID_Wtime — Return a time stamp

Synopsis

```
void MPID_Wtime( MPID_Time_t * timeval)
```

Output Parameter

timeval A pointer to an `MPID_Wtime_t` variable.

Notes

This routine returns an *opaque* time value. This difference between two time values returned by `MPID_Wtime` can be converted into an elapsed time in seconds with the routine `MPID_Wtime_diff`. This routine is defined this way to simplify its implementation as a macro. For example, the for Intel x86 and gcc,

```
#define MPID_Wtime(timeval) \
    __asm__ __volatile__  ( "cpuid ; rdtsc ; mov %%edx,%1 ; mov %%eax,%0" \
                            : "=m" (*((char *) (var_ptr))), \
                              "=m" (*(((char *) (var_ptr))+4)) \
                            :: "eax", "ebx", "ecx", "edx" ); \
```

For some purposes, it is important that the timer calls change the timing of the code as little as possible. This form of a timer routine provides for a very fast timer that has minimal impact on the rest of the code.
From a semantic standpoint, this format emphasizes that any particular timer value has no meaning; only the difference between two values is meaningful.

Module

Timer

Question

MPI-2 allows `MPI_Wtime` to be a macro. We should make that easy; this version does not accomplish that.

MPID_Wtick — Provide the resolution of the `MPID_Wtime` timer

Synopsis

```
double MPID_Wtick( void )
```

Return value

Resolution of the timer in seconds. In many cases, this is the time between ticks of the clock that `MPID_Wtime` returns. In other words, the minimum significant difference that can be computed by `MPID_Wtime_diff`.
Note that in some cases, the resolution may be estimated. No application should expect either the same estimate in different runs or the same value on different processes.

Module

Timer

MPID_Wtime_diff — Compute the difference between two time stamps

Synopsis

```
void MPID_Wtime_diff( MPID_Time_t *t1, MPID_Time_t *t2, double *diff )
```

Input Parameters

t1, t2 Two time values set by `MPID_Wtime` on this process.

Output Parameter

diff The different in time between t2 and t1, measured in seconds.

Note

If `t1` is null, then `t2` is assumed to be differences accumulated with `MPID_Wtime_acc`, and the output value gives the number of seconds that were accumulated.

Question

Instead of handling a null value of `t1`, should we have a separate routing `MPID_Wtime_todouble` that converts a single timestamp to a double value?

Module

Timer

MPID_Wtime_init — Initialize the timer

Synopsis

```
void MPID_Wtime_init(void)
```

Note

This routine should perform any steps needed to initialize the timer. In addition, it should set the value of the attribute `MPI_WTIME_IS_GLOBAL` if the timer is known to be the same for all processes in `MPI_COMM_WORLD` (the value is zero by default).
If any operations need to be performed when the MPI program calls `MPI_Finalize` this routine should register a handler with `MPI_Finalize` (see the MPICH Design Document).

Module

Timer

MPID_Wtime_todouble — Converts an MPID timestamp to a double

Synopsis

```
void MPID_Wtime_todouble( MPID_Time_t *timeval, double *seconds )
```

Input Parameter

timeval `MPID_Time_t` time stamp

Output Parameter

seconds Time in seconds from an arbitrary (but fixed) time in the past

Notes

This routine may be used to change a timestamp into a number of seconds, suitable for `MPI_Wtime`.

MPID_Wtime_acc — Accumulate time values

Synopsis

```
void MPID_Wtime_acc( MPID_Time_t *t1, MPID_Time_t *t2, MPID_Time_t *t3 )
```

Input Parameters

t1,t2,t3 Three time values. `t3` is updated with the difference between `t2` and `t1`: `*t3 += (t2 - t1)`.

Notes

This routine is used to accumulate the time spent with a block of code without first converting the time stamps into a particular arithmetic type such as a `double`. For example, if the `MPID_Wtime` routine accesses a cycle counter, this routine (or macro) can perform the accumulation using integer arithmetic.
To convert a time value accumulated with this routine, use `MPID_Wtime_diff` with a `t1` of zero.

Module

Timer

References

[1] M. Banikazemi, R. K. Govindaraju, R. Blackmore, and D. K. Panda. Implementing efficient MPI on LAPI for the IBM-SP: Experiences and performance evaluation. In *International Parallel Processing Symposium (IPPS '99)*, pages 183–190, April 1999.

[2] Greg Burns, Raja Daoud, and James Vaigl. LAM: An open cluster environment for MPI. In John W. Ross, editor, *Proceedings of Supercomputing Symposium '94*, pages 379–386. University of Toronto, 1994.

[3] James Cownie and William Gropp. A standard interface for debugger access to message queue information in MPI. In Jack Dongarra, Emilio Luque, and Tomàs Margalef, editors, *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, volume 1697 of *Lecture Notes in Computer Science*, pages 51–58. Springer Verlag, 1999.

[4] Marco Fillo and Richard B. Gillett. Architecture and implementation of MEMORY CHANNEL2. *DIGITAL Technical Journal*, 9(1), 1997. `http://www.digital.com/info/DTJP03/DTJP03HM.HTM`.

[5] William Gropp, Ewing Lusk, Nathan Doss, and Anthony Skjellum. A high-performance, portable implementation of the MPI Message-Passing Interface standard. *Parallel Computing*, 22(6):789–828, 1996.

[6] Ferhat Hatay, Rolf van deVaart, and Josh Simons. Sun HPC ClusterTools™ software: Ubiquitous parallel computing, 2001.

[7] LAM-MPI. World Wide Web. `www.lam.org`.

[8] Hong Tang, Kai Shen, and Tao Yang. Program transformation and runtime support for threaded MPI execution on shared memory machines. *ACM Transactions on Programming Languages and Systems*, 0(0):999–1025, January 2000.

[9] VI Architecture. Was `http://www.viarch.org` but has disappeared.

[10] Thorsten von Eicken, Anindya Basu, Vineet Buch, and Werner Vogels and. U-Net: A user-level network interface for parallel and distributed computing. In *Proceedings of the 15th ACM Symposium on Operating Systems Principles (SOSP)*, pages 40–53. ACM Press, December 1995.

Index

MPID_Abort, 53
access epoch, 10
active target, 1
MPID_Attribute, 31

MPID_Cancel_recv, 41
MPID_Cancel_send, 40
MPID_Comm, 23
MPID_Comm_kind_t, 19
MPID_Comm_thread_lock, 34
MPID_Comm_thread_unlock, 35
connection, 3
context id, 11
MPID_Copy_function, 21

MPID_Datatype, 27
MPID_Delete_function, 21

epoch
    access, 10
    exposure, 10
MPID_Errhandler, 30
MPID_Errhandler_fn, 19
exposure epoch, 10

MPID_Finalize, 52

MPID_Get_processor_name, 54
MPID_Group, 24

hook routine, 7

MPID_Info, 29
MPID_Init, 50
MPID_Iprobe, 41
MPID_Irecv, 40
MPID_Irsend, 38
MPID_Isend, 37
MPID_Issend, 38

MPID_Keyval, 30

MPID_Lang_t, 18

MPID_MAX_THREAD_LEVEL, 32
MPID_Mem_alloc, 55
MPID_Mem_free, 55
MPID_Mem_was_alloced, 56

MPID_Object_kind, 17
MPIU_Object_add_ref, 32
MPIU_Object_release_ref, 33
MPID_Op, 27
MPID_Op_kind, 19

passive target, 1
polling, 1
    points, 5, 14
MPID_Probe, 42
progress, 4
MPID_Progress_end, 48
MPID_Progress_poke, 50
MPID_Progress_start, 48
MPID_Progress_test, 49
MPID_Progress_wait, 49

MPID_Recv, 39
MPID_Recv_init, 45
MPID_Request, 22
MPID_Request_create, 45
MPID_Request_kind, 18
MPID_Request_release, 40
MPID_Rsend, 37
MPID_Rsend_init, 44

segment, 6
MPID_Segment, 6
MPID_Send, 36
MPID_Send_init, 43
MPID_Ssend, 36
MPID_Ssend_init, 44
MPID_Startall, 45

MPID_tBsend, 38
thread safety
    issues, 7, 8
MPID_Topo_cluster_info, 46

MPID_User_function, 20

MPID_VCR_Dup, 57
MPID_VCR_Get_lpid, 58
MPID_VCRT_Add_ref, 57
MPID_VCRT_Create, 57
MPID_VCRT_Get_ptr, 57
MPID_VCRT_Release, 57
virtual connections, 2

MPID_Win, 25
MPID_Wtick, 59
MPID_Wtime, 58
MPID_Wtime_acc, 61
MPID_Wtime_diff, 59
MPID_Wtime_init, 60
MPID_Wtime_todouble, 60