

Quo Vadis MPI RMA? Towards a More Efficient Use of MPI One-Sided Communication

Joseph Schuchart*
schuchart@icl.utk.edu

Innovative Computing Laboratory (ICL), University of
Tennessee Knoxville (UTK)
Knoxville, TN, U.S.A.

José Gracia
gracia@hlrs.de

High-Performance Computing Center Stuttgart (HLRS)
Stuttgart, Germany

Christoph Niethammer
niethammer@hlrs.de

High-Performance Computing Center Stuttgart (HLRS)
Stuttgart, Germany

George Bosilca
bosilca@icl.utk.edu

Innovative Computing Laboratory (ICL), University of
Tennessee Knoxville (UTK)
Knoxville, TN, U.S.A.

ABSTRACT

The MPI standard has long included one-sided communication abstractions through the MPI Remote Memory Access (RMA) interface. Unfortunately, the MPI RMA chapter in the 4.0 version of the MPI standard still contains both well-known and lesser known short-comings for both implementations and users, which lead to potentially non-optimal usage patterns. In this paper, we identify a set of issues and propose ways for applications to better express anticipated usage of RMA routines, allowing the MPI implementation to better adapt to the application's needs. In order to increase the flexibility of the RMA interface, we add the capability to duplicate windows, allowing access to the same resources encapsulated by a window using different configurations. In the same vein, we introduce the concept of MPI memory handles, meant to provide life-time guarantees on memory attached to dynamic windows, removing the overhead currently present in using dynamically exposed memory. We will show that our extensions provide improved accumulate latencies, reduced overheads for multi-threaded flushes, and allow for zero overhead dynamic memory window usage.

KEYWORDS

MPI-RMA, Memory Handles, MPI, RDMA

ACM Reference Format:

Joseph Schuchart, Christoph Niethammer, José Gracia, and George Bosilca. 2021. Quo Vadis MPI RMA? Towards a More Efficient Use of MPI One-Sided Communication. In *Proceedings of EuroMPI 2021 (EuroMPI'21)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

*Also with High-Performance Computing Center Stuttgart (HLRS).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

EuroMPI'21, September 7th, 2021, Garching, Germany

© 2021 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM... \$15.00
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

Modern high-performance networks commonly provide the capability to directly access memory on a remote host for reading, writing, and atomic memory updates [2, 23]. The hardware is capable of transferring data without the involvement of the CPU at the target node once the upper software layers have properly set up the parameters for the transfer, e.g., registered memory with the network interface card (NIC) and exchanged the registration information with the peers involved. MPI implementations typically make use of these low-level network features to provide efficient transfer of large messages between peers communicating through point-to-point or collective operations [36].

The MPI RMA interface was introduced with MPI version 2.0 [24] and has seen a major overhaul in version 3.0 [25], including the addition of allocated and dynamic windows. The intention of this interface is to expose the network's low-level remote direct memory access (RDMA) capabilities to the user by providing procedures for *put*, *get*, and *accumulate* operations on *windows* that encapsulate memory for which registration information has been exchanged among the group of participating peers. By using MPI RMA, applications are able to decouple communication and synchronization, e.g., to perform bursts of communication before synchronizing through collective and point-to-point communication or by setting a signal flag at the target using accumulate operations.

In its current form, an MPI window is an object spanning across the processes in its group, i.e., its creation and destruction are collective operations. Active target synchronization involves collective operations. Passive target synchronization, on the other hand, only involves specific MPI calls at the origin of the operation (some MPI implementation may, however, depend on the target to call into MPI procedure to progress outstanding RMA operation). With the exception of dynamic windows, the memory accessed through these windows is static, requiring collective (re)allocations to increase the amount of memory accessible to RMA operations. Dynamic windows, on the other hand, allow for dynamically attaching and detaching memory segments, albeit at a significant penalty in performance, which will be discussed in Section 4.1.

Interest in RMA in the user community seems to be growing [4, 8]. However, the RMA chapter has seen little change during the work on the 4.0 release of the standard, despite there being several

known shortcomings that inhibit full and efficient usage of MPI RMA in applications and runtime systems [30, 32]. In this paper, we draw from our experience in using MPI-3 RMA in the context of the DASH project [15, 32] in general and the global task synchronization scheme built on top of it in particular [31]. We will discuss a number of issues found during this work and propose potential remedies that mostly consist in allowing the user to express the anticipated use of the RMA interface to the implementation through additional info keys (Section 2). In order to increase flexibility in applying these configurations, we propose an extension to the MPI RMA interface that allows users to duplicate windows, accessing the same memory using the same network resources but with configurations adjusted to the needs of different regions of the application (Section 3). In order to increase flexibility in the use of RDMA through the MPI RMA interface, we propose the addition of *MPI memory handles* that allow users to explicitly manage the registration information of memory attached to dynamic windows, alleviating the performance penalties that stem from the current design of dynamic windows (Section 4).

A brief discussion of additional improvements and changes that are beyond the scope of this work will be provided in Section 5. Section 6 discusses some of the implementation details for the proposed solutions before an evaluation of some of the proposals using micro benchmarks is presented in Section 7. Related work is discussed in Section 8 and conclusions are drawn in Section 9.

2 ADDITIONAL USER-PROVIDED INFORMATION

2.1 Thread-Scope Synchronization

Communication and synchronization in MPI RMA happens at the scope of processes, which encapsulate the memory made accessible to remote peers. Thus, a flush in passive target synchronization or a fence in active target synchronization ensure the completion of all operations previously issued by any thread in the current process to complete in the target process memory. Active target synchronization is collective either over the group of a window (`MPI_Win_fence`) or an otherwise provided group (post-start-complete-wait, PSCW). With passive target synchronization, flushes are local operations. Despite previous attempts at thread-specific communication endpoints [11], collective operations in MPI happen at the scope of processes. We will thus focus on the passive target synchronization.

In multi-threaded applications, individual threads may perform RMA operations independently. However, a thread calling into `MPI_Win_flush` potentially waits for the completion of operations issued by all threads of the same process to the same target, even though the completion of operations previously issued by the current thread might be sufficient for the application. This is depicted in Figure 1a, where the flush of Thread 1 is prolonged by operations issued by Thread 2. MPI implementations may use thread-specific network hardware resources (*endpoints* or *rails*) to reduce synchronization between threads when issuing RMA operations, e.g., by using one endpoint per thread or distributing operations across a fixed number of endpoints [17].

MPI supports flushing operations to a single target or to all targets in the group of the window, either with local or remote

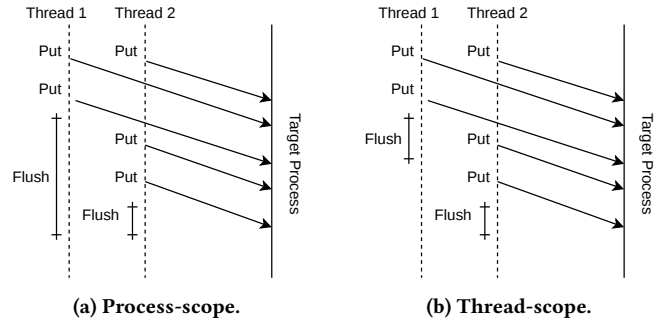


Figure 1: Flushes with process- and thread-scope. With process-scope flushes, Thread 1 potentially waits for the completion of all of its operations and the operations issued by Thread 2.

completion. We will focus on operations with remote completion, although the proposal also applies to flushes with local completion.

In order to restrict the completion semantics of flush operations to operations previously issued by the calling thread, the user has to signal this intention to the implementation. While it would be feasible to introduce a new set of functions to accomplish this goal, the required extension of the API would force implementations to provide such functionality even if support for multi-threaded RMA was limited. It would also introduce the notion of thread-scope operations into an API that is otherwise oblivious of the existence of multiple threads of execution, with the exception of `MPI_Init_thread` used to signal their (anticipated) existence.

We thus propose the addition of an info key called `mpi_win_scope`, which specifies the scope of synchronization operations on a window. If the value of that key is set to process (the default), flushes behave as today with operations issued by all threads required to complete during a flush. However, if the scope is set to thread, the implementation is free to restrict the scope of a flush to the operations previously issued by the calling thread, as depicted in Figure 1b. Since the process scope is a superset of the thread scope, implementations ignoring this info key remain correct. Users can check the support for the thread scope by querying the value associated with that key from the info object attached to the window using `MPI_Win_get_info` [26, §12.2.7].

With the `mpi_win_scope` key set to thread on a window, implementations using thread-local endpoints only need to wait for the completion of operations on the endpoint assigned to the calling thread, potentially avoiding any synchronization between threads using RMA with passive target synchronization. The key has no effect on active target synchronization, since collective operations always happen at the process scope.

2.2 Operation Ordering

By default, MPI RMA guarantees the ordering of consecutive accumulate operations on the same memory location with the same data type and allows users to relax this constraint using the `accumulate_ordering` info key. However, in order to achieve ordering of put and get operations or to order accumulate operations to distinct memory location within a window, the application is required

```

int flag, one = 1;
MPI_Request req;
MPI_Rput(..., target, win, &req);
do {
    do_useful_work();
    MPI_Test(&req, &flag, MPI_STATUS_IGNORE);
} while (!flag);
/* Flush needed for remote completion */
MPI_Win_flush(target, win);
/* Signal that the data has been written by
 * incrementing a counter at the target */
MPI_Raccumulate(&one, ..., target, ..., win, &req);
do {
    do_useful_work();
    MPI_Test(&req, &flag, MPI_STATUS_IGNORE);
} while (!flag);

```

Listing 1: Using an atomic increment to signal the completion of a put, overlapping communication with useful work.

to wait for all outstanding operations to complete before issuing operations that are required to occur later in the sequence. This completion might entail at least the latency of a full round-trip in the network, depending on the number of previously issued operations.

Listing 1 provides an example in which an accumulate operation is used to signal the availability of data previously put into the target's memory. In order to hide the latency of both operations, the application tests on requests for both operations (Lines 6 and 15) and requires a flush in between (Line 9) to ensure remote completion before the signal is set.

Modern high-performance networks provide so-called *fence* operations, allowing users to request the hardware to order the completion of two operations Op_1 and Op_2 in the order in which they were issued, similar to a memory barrier in shared memory systems. We have previously proposed adding a function `MPI_Win_order` [30], which would translate either into a memory barrier or a fence in the network interface card. In multi-threaded applications, however, the default process-scope of MPI RMA would require a fallback to waiting for completion of all operations if thread-specific network resources are used. Using the thread-local scope proposed in Section 2.1 may provide a partial solution by constraining the scope of ordering to operations issued by individual threads. However, in cases where the implementation issues operations of a single thread to multiple hardware resources (e.g., for explicit load-balancing of communication) multiple streams of operations would again have to be synchronized by waiting for completion of prior operations.

The underlying problem, however, is that the ordering request may be injected into the operation stream Op_n, \dots, Op_m at any time. As a consequence, the MPI implementation has no prior knowledge of the ordering request at the time Op_n is issued and thus would have to either constrain itself to using configurations in which operations can safely be ordered or resort to waiting for completion to enforce ordering.

In order to provide the MPI implementation with *a priori* information about intended operation ordering we propose the concept of *ordered operation sequences*. The user thereby signals the request to order a set of operations *before* issuing the first operation included in that sequence. We propose a new info key that enables operation ordering on a given window, called `mpi_win_order`. While

```

int flag, one = 1;
MPI_Request req;
MPI_Put(..., target, win);
/* Signal that the data has been written by
 * incrementing a counter at the target */
MPI_Raccumulate(&one, ..., target, ..., win, &req);
do {
    do_useful_work();
    MPI_Test(&req, &flag, MPI_STATUS_IGNORE);
} while (!flag);

```

Listing 2: The example of Listing 1 with `mpi_win_order` set to true, avoiding the first flush by chaining two operations.

set to true, the sequence of operations issued to the same target on this window will complete at the target in the order in which the operations were issued. This provides sufficient information to the implementation to implement operation ordering without resorting to completion in the middle of the sequence, albeit potentially at the cost of using a single endpoint. However, in combination with the thread scope described previously ordering can be constrained to operations issued by the same thread.

Applications relying on ordering using this info key are required to check whether the operation ordering using the `mpi_win_order` info key is supported and fall back to flushes in between operations otherwise. Listing 2 provides a modified version of the example in Listing 1 with the `mpi_win_order` set to true, avoiding the intermittent flush and only testing for the accumulate operation request.

2.3 Hardware Accumulate Operations

MPI RMA accumulate operations are notoriously hard to implement efficiently: on the one hand, single-element operations such as `MPI_Fetch_and_op` and `MPI_Compare_and_swap` may benefit from the use of hardware atomic operations provided by the NIC due to the low latency of single-element operations implemented by the hardware. On the other hand, `MPI_Accumulate` allows users to issue operations on an arbitrary number of elements at the same time, which could potentially benefit from the higher bandwidth of operations performed by vector units on the host CPU [37]. The MPI standard requires implementations to provide element-wise atomicity of operations applied to the same memory location if using the same data type, regardless of whether they were issued through single-element operations or as part of a multi-element operation. In addition, not all possible operations are supported by the network hardware. For example, while networks commonly support addition and subtraction of integral values, support for integral value multiplication or operations on floating point values is often missing. Taken together, implementations once again cannot anticipate the operations that will be issued by the user and thus have to leave certain hardware features lay bare.

Implementation typically use two possible approaches: i) taking a lock at the target process before fetching the data, applying the operation, and writing the data back before releasing the lock; and ii) using active messages to transfer the data to the target and relying on the target CPU to perform the operation. Both approaches require the serialization of concurrent accumulate operations through some form of mutual exclusion device.

```

/**
 * Query whether the implementation employs hardware operations
 * intrinsic to the origin node to perform the operations listed
 * in ops on a maximum of max_count elements of type on the provided
 * window.
 * Flag will be set to 1 if intrinsic hardware operations at the
 * origin are used to perform these operations and 0 otherwise.
 */
int MPI_Win_op_intrinsic(const char *ops, MPI_Aint max_count,
                        MPI_Datatype type, MPI_Win win, int *flag);

```

Listing 3: Function to query the use of intrinsic hardware operations for a given set of operations on a number of elements of a certain data type.

An existing proposal to tackle this problem allows the user to specify how many elements will be used with which operations [35]. However, users would still have no information on whether accumulate operations will be executed in hardware or software. Some applications may rely on low-latency accumulate operations [6, 30] but the lack of transparency prevents them from picking an alternative algorithm if available. The proposed unidirectional signaling is hence not sufficient.

We propose the addition of a procedure (based on a previous suggestion in the MPI RMA working group [10]) that i) allows the application to query the implementation’s approach to performing a given accumulate operations for a given number of elements of a certain data type; and ii) to signal the anticipated usage pattern to the implementation. We borrow a concept from the C++ `std::atomic` wrapper type [20], which allows developers to query whether atomic modification of a given wrapped type is *lock-free* using the compile-time `is_lock_free` trait that signals whether a mutex or CPU-provided atomic operations are used to ensure atomicity. Contrary to `std::atomic`, MPI accumulate operations may apply to multiple elements at once and implementations may use a threshold for switching between hardware and software approaches.

Thus, the information has to be query-able at runtime, for which we propose a new procedure called `MPI_Win_op_intrinsic` listed in Listing 3. For a given tuple describing the set of anticipated operations to be performed (`ops`) on the provided number elements (`max_count`) of a certain type on a specific window `win`, the implementation returns whether the operations will be performed with hardware operations intrinsic to the origin node, i.e., without relying on the participation of a CPU at the target.¹ The set of operations are described as a string containing a comma-delimited list of operations, using the second half of the name of predefined `MPI_Op` elements (e.g., “sum”), “replace” for `MPI_REPLACE`, and “cas” to denote `MPI_Compare_and_swap`.

The information obtained from a call to `MPI_Win_op_intrinsic` may then be used to set a new boolean info key called `mpi_assert_accumulate_intrinsic`. If set to true, the application asserts that it will only issue accumulate operations in configurations for which the implementations has signaled the use of intrinsic operations. The results of the application disregarding this assertion are undefined, leading to modifications that are potentially non-atomic.

¹We note that an accumulate operation using the network hardware technically relies on a processor at the target to perform the operation. However, the accumulate instruction is issued to the NIC at the origin and is thus intrinsic to the origin hardware.

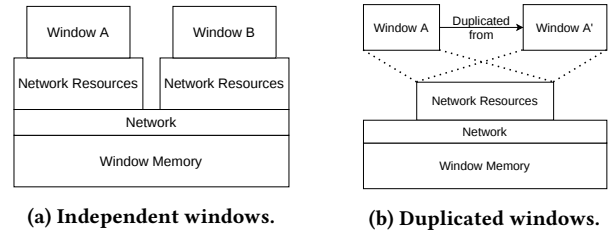


Figure 2: Two windows accessing the same window memory through the network.

With this bidirectional signaling mechanism we achieve two goals: a) providing transparency to the application on how the MPI implementation will handle a given configuration of accumulate operations; and b) allowing the application to announce their anticipated behavior to the implementation. This, in turn, will allow implementations to safely make use of hardware atomic operations if all anticipated operations used by the applications can be mapped onto the hardware and the number of elements is below the threshold controlling the switch in the trade-off between low latency and high bandwidth.

3 DUPLICATING MPI WINDOWS

In the previous section we have proposed three new info keys to help the user express the intended use of RMA operations on a window: restricting the scope of flushes to operations issued by the calling thread; ordering the operations issued on the window; and limiting the number of elements in accumulate operations to allow for the use of operations intrinsic to the origin hardware. However, changing the value of an info key overwrites the old value and thus makes it impossible to use different configurations concurrently, e.g., to request operation ordering on some threads but not on others or to use bandwidth-optimized accumulate operations on one part of the memory while using latency-optimized accumulate operations on another part of the window. Switching between these settings would require careful orchestration of info key values.

While it is legitimate to allocate memory in `MPI_Win_allocate` and pass that memory into a call to `MPI_Win_create` with different info key values, the resulting two windows are semantically independent with independent passive and active target synchronization semantics and no cross-window atomicity guarantees, as depicted in Figure 2a. In order to ease the task of managing different means of access to the same window memory and to keep windows with different info values in sync, we propose to add a *window duplication* function, as outlined in Listing 4. In contrast to two independently created windows, duplicated windows may share internal data structures and window memory while carrying potentially different access semantics. Duplicated windows can thus be considered as different handles to the same underlying memory and network resources, as depicted in Figure 2b.

This approach enables the use case described above in which different window settings are used in different parts of an application, while sharing the underlying resources. Some restrictions apply: as long as the duplicated windows use the same value for `mpi_assert_accumulate_intrinsic` accumulate operations are

```

/* Duplicate a window with different info key values.
 * Both the parent window and the new window access the same
 * memory regions with potentially different configurations. */
int MPIX_Win_dup_with_info(MPI_Win parentwin,
                           MPI_Info info,
                           MPI_Win *newwin);
    
```

Listing 4: Signature of window duplication function.

atomic across these windows. However, issuing accumulate operations on two windows having different values for the `mpi_assert_accumulate_intrinsic` is legal but the accumulate operations may not be atomic with respect to each other. It is up to the user to coordinate the correct use of this info key.

We propose a new function called `MPIX_Win_dup_with_info` that is used to duplicate the window with a new set of info keys. Its signature is shown in Listing 4. Info keys from the parent window will be duplicated into the new window, with the provided info keys overriding existing ones.

Since the original and duplicated windows are not logically separate, all synchronization operations applied to a window also apply to its duplicates, and vice versa. For example, the duplicated window may not be locked if the parent window has already been locked. In essence, window duplication is akin to assignment of an `MPI_Win` variable with the added ability to control certain info values.

The call to `MPIX_Win_dup_with_info` is a local operation and thus does not entail any synchronization with other processes in the group of the parent window. An MPI implementation may not be able to change certain info keys during this call and may thus reject the change by retaining the original or default value. Users should check whether the MPI implementation is able to support the requested configuration by querying the active info keys using `MPI_Win_get_info`.

4 DYNAMIC MEMORY HANDLES

In its current form, window creation is a collective operation in MPI. With the exception of dynamic windows, MPI windows and their memory are statically allocated or assigned, which allows the MPI implementation to exchange all relevant connection and registration information during window creation and enables the use of the network’s RDMA capabilities (Figure 3a). However, such static windows may be impractical if an application’s communication requirements changes over time, requiring repeated (collective) reallocation of windows. Moreover, applications may treat communication and memory allocation as orthogonal concerns such that the use of window memory would break through abstraction boundaries, requiring major restructuring efforts to integrate the allocation of memory through MPI windows.

In contrast, dynamic windows, as their name suggests, allow users to attach and detach memory dynamically in a local operation after the window has been created in a collective operation. Attaching the memory explicitly allows the MPI implementation to register the memory with the network device for later access using RDMA. Addressing in dynamic windows is done using absolute virtual addresses: after attaching memory to the dynamic window, the process distributes the virtual address that is then used as displacement at the origin of an RMA operation.

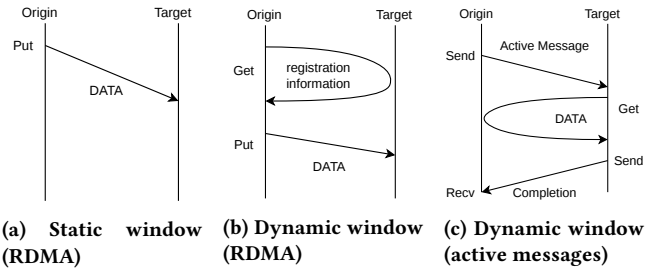


Figure 3: Possible implementations of put on RDMA-capable networks for static and dynamic windows.

Unfortunately, the use of virtual addresses is the greatest weakness of dynamic windows: in contrast to static windows, the origin of an RMA operation initially has no information on the underlying memory registration and thus has to either query this information from the target before issuing RDMA operations (Figure 3b) or fall-back to emulating remote memory accesses using active messages (AM, Figure 3c). Both approaches add considerable latency, especially to RMA operations on small amounts of data. Since the target may detach and reattach the same virtual base address, the registration information for the same virtual address may change at the target in between RMA operations at the origin. Thus, while caching techniques are possible, the origin has to at least verify the validity of the cached registration information on every RMA communication operation. The lack of *life-time guarantees* is thus the main reason for added overhead when using dynamic windows.

We will show in Section 4.1 that the difference in latency between allocated and dynamic windows is significant on all tested MPI implementations. We will then propose an extension to the MPI RMA interface to allow applications to explicitly exchange registration information and thus make life-time guarantees to the MPI implementation that enable it to use RDMA with zero overhead.

4.1 State of the Art

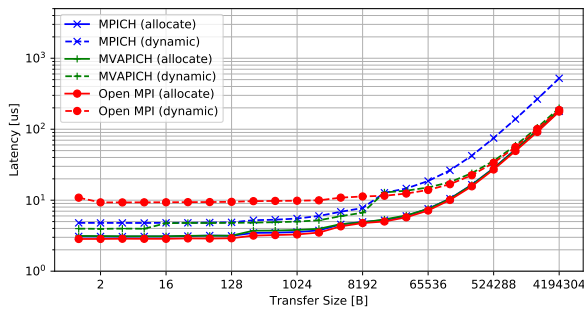
Table 1 lists the software used for comparison of dynamic and allocated windows in this section. All measurements were conducted on an HPE Apollo 6500 system *Hawk* installed at HLRS.² The nodes are equipped with dual-socket 64-core AMD EPYC 7742 processors and connected through Mellanox InfiniBand HDR200 in a 9D hyper-cube fabric.

4.1.1 Communication Latency. Figure 4 shows the latency of put operations measured using the OSU benchmark `osu_put_latency` benchmark on different MPI implementations. While the latencies on allocated windows are similar across the three implementations, the differences are significant for dynamic windows. Especially for smaller transfer sizes, the penalty of using dynamic windows over allocated windows ranges from a factor of 1.5× (MPICH, MVA-PICH) over 3× (Open MPI using UCX). As described earlier, this discrepancy between allocated and dynamic windows stems from the missing registration information, which either leads to a fall-back to AM-based emulation or requires fetching the registration information before issuing the actual operation.

²More details at <https://www.hlrs.de/systems/hpe-apollo-hawk/>

Table 1: Software configuration.

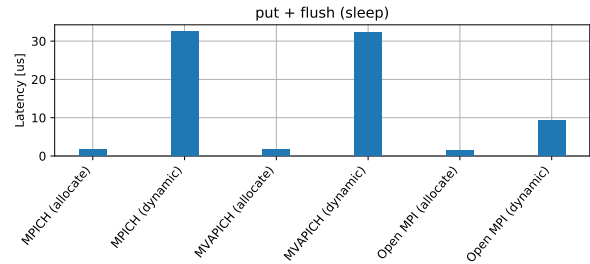
Software	Version	Configuration/Remarks
Open MPI	4.0.5	-with-ucx=...
MPICH	v4.0.a	-with-device=ch4:ucx
MVAPICH	2.3.5	-with-device=ch3:mrail -with-rdma=gen2
UCX	1.10.0	-enable-mt -with-xpmem -with-verbs -with-rdmacm
GCC	10.2.0	site installation
OSU Benchmarks	5.6.2	none

**Figure 4: Latency of put operations using allocated and dynamic windows.**

4.1.2 One-Sided Progress Behavior. In order to better understand the behavior of the implementations, we repeat a benchmark here that was used in [30] to determine the one-sided behavior of various accumulate MPI implementations. In this test, the target process is busy outside of MPI for a fixed amount of time before waiting in an MPI barrier for the origin to complete the execution of a number of RMA operations. For the results shown in Figure 5, the origin performs $n = 100\,000$ put operations, each followed by a flush, while the target is busy outside of MPI for $t = 3$ s. Thus, a latency of $\frac{t}{n} > 30\ \mu\text{s}$ indicates that the origin is not progressing until the target enters the MPI barrier. As can be seen in Figure 5, both MPICH and MVAPICH lack progress for dynamic windows, indicating an implementation relying on the target CPU to execute active messages (Figure 3c). Operations on dynamic windows using the UCX integration in Open MPI, on the other hand, progress, albeit at a significantly higher latency, as discussed in the previous section. We note that while AM-based emulation may yield sufficiently low-latencies in benchmarks such as the `osu_get_latency`, in practice it renders the performance of MPI RMA unpredictable as performance depends on the behavior of the target process, (partially) defeating the purpose of a one-sided programming interface.

4.2 Life-Time Control Through Memory Handles

Allowing users to provide MPI directly with registration information on memory to be accessed through windows provides both

**Figure 5: Average latency of 100,000 single-byte put and flush with the target sleeping for 3 s. Latencies above $30\ \mu\text{s}$ indicate no progress while the target process is not executing MPI calls.**

life-time information and avoid additional overhead during communication operations. We propose the following three additions to the MPI RMA interface (their signatures are detailed in Listing 5):

MPIX_Memhandle_create registers a memory region starting at base of size `size` with the provided dynamic window for later access through RMA operations. The function returns in `memhandle` a memory handle of size `memhandle_size`. The `memhandle` should be a buffer of at least `MPI_MAX_MEMHANDLE_SIZE` bytes. The memory handle contained in this buffer can be distributed to peer processes.

MPIX_Win_from_memhandle The received memory handle is passed to this function together with the same dynamic window. The function returns a new window object whose only allowed target is the provided target and the usual configuration of displacement unit, size, and info to control aspects of the newly created window.

MPIX_Memhandle_release Once all RMA operations have completed and the registered memory is not needed anymore (i.e., all peers have signaled completion) the memory handle can be released using this function. After a call to this function, no more RMA operations may be issued on windows created from this memory handle. The corresponding windows must be freed through a call to `MPIX_Win_free`.

Instead of sending the virtual address of an attached memory region to the peer the application now sends the registration information directly, which is an opaque data structure that is specific to the underlying implementation (which may differ from platform to platform for the same MPI implementation).

The call to `MPIX_Win_from_memhandle` is a local operation and the resulting window remains connected to its parent window. We restrict synchronization of such windows to passive target synchronization and require the lock and unlock to be applied on the parent dynamic window. These restrictions allow the MPI implementation to avoid allocating additional internal memory during the creation of the memory handle required to handle these synchronization operations. Thus, the only operations permitted on memory handle windows are put, get, and accumulate operations as well as flushes. We expect users to use shared locks and rely on other synchronization and signaling mechanisms such as collective operations, point-to-point operations, or accumulate operations. By allocating a separate window object, the implementation is

```

/* Maximum size of a memory handle, implementation specific */
#define MPI_MAX_MEMORY_HANDLE_SIZE <value>

/* Start exposure for memory and return handle to be sent
 * to peers. The memory handle is returned in memhandle
 * and has the actual size memhandle_size. The memhandle
 * argument should be a byte array of at least
 * MPI_MAX_MEMORY_HANDLE_SIZE elements. */
int MPIX_Memhandle_create(
    void *base, MPI_Aint size,
    MPI_Info info, MPI_Win parentwin,
    void *memhandle, int *memhandle_size);

/* Create a window from a memory handle. The data
 * in memhandle should have been filled in by a call to
 * MPIX_Memhandle_create and sent to a peer or used
 * locally. The parentwin argument is a previously allocated
 * dynamic window. A newly created window will be returned
 * in newwin. */
int MPIX_Win_from_memhandle(
    const void *memhandle,
    MPI_Aint size, int disp_unit,
    MPI_Info info, int target,
    MPI_Win parentwin, MPI_Win *newwin);

/* Release a memory handle, ending the associated memory's
 * exposure. The data in memhandle must have previously
 * been filled in by a call to MPIX_Memhandle_create.
 * It is erroneous to release a memory more than once. */
int MPIX_Memhandle_release(void *memhandle, MPI_Win parentwin);

```

Listing 5: The MPI Memory Handle interface.

not required to maintain and repeatedly traverse a list of attached memory handles but instead the resulting window identifies the remote memory region directly and allows for the implementation to stop tracking that remote memory region once the applications calls `MPIX_Win_free` on the memory handle window.

4.3 Example

An example for how we envision memory handles to be used is provided in Figure 6. After creating the window `win` from the memory handle received from Process A, Process B performs an arbitrary number of RMA operations on the corresponding target memory region before signaling completion back to Process A, which then releases the memory handle. We will show in Section 7.3 that latencies using memory handle windows are on par with allocated windows in our proof-of-concept implementation.

5 FURTHER IMPROVEMENTS

In this section we briefly discuss efforts that we consider beneficial for the future direction of the MPI RMA API but that are beyond the scope of this paper.

5.1 Completion Notification

Past work has focused on completion notification at the target and we support the approach presented in [34], which was based on a previous proposal [3]. However, we caution that instead of introducing new test/wait routines (`MPIX_Win_test_notify` and `MPIX_Win_wait_notify`) the notification should integrate with the existing request facilities. Due to the nature of progress in MPI, these functions have to ensure progress in order to drive non-RMA communication

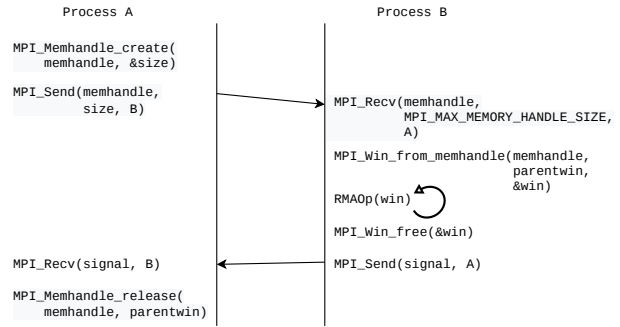


Figure 6: An example of using memory handles and the associated memory handle windows. RMAOp signifies any RMA operation on the window.

operations on whose completion at the origin the notification may depend, e.g., collectives or point-to-point operations.

Using a request-based notification mechanism (without relying persistent requests [3]) allows users to test or wait for completion all MPI-related communication, including RMA notifications. We envision a function such as `MPIX_Win_wait_notify(win, notify_id, request)` that returns a request, which can later be used with the regular request test and wait infrastructure in MPI. We leave an in-depth investigation into such an interface for future work.

5.2 Remote-Completing Request-Based Operations

MPI RMA provides variants of put, get, accumulate, and get-accumulate that return a request that can be used to test and wait for completion of that particular operation. However, the completion of a request returned by `MPIX_Rput`, `MPIX_Raccumulate`, and `MPIX_Rget_accumulate` only signals local completion, requiring a subsequent flush to achieve remote completion before signaling completion to the target. This flush, in turn, may wait for the completion of unrelated operations, potentially from other threads. In many cases, MPI implementations are able to implement request-based put with remote completion more efficiently than the application using a flush (e.g., by leveraging guarantees of the underlying transport library). We thus propose adding functions `MPIX_Rrput` (remote-completing request-base **put**), `MPIX_Rraccumulate`, and `MPIX_Rrget_accumulate` with similar signatures as their current request-based counter-parts.. The completion of requests provided by these procedures signal both local and remote completion, alleviating the need for an additional flush and thus potentially improving the efficiency of applications using request-based operations. An example of the case described above using a remote-completing request-based put is given in Listing 6. If a regular call to `MPIX_Rput` was used instead, a flush would be necessary in Line 8 before the call to `MPIX_Allreduce` was used to signal completion to all peers.

5.3 Deprecating Active Target Synchronization

We encourage efforts to engage with users of the active target synchronization interface to identify potential road-blocks in the transition to passive target synchronization, with the goal of phasing

```

1 int flag, one = 1;
2 MPI_Request req;
3 MPI_Rrput(..., target, win, &req);
4 do {
5     do_useful_work();
6     MPI_Test(&req, &flag, MPI_STATUS_IGNORE);
7 } while (!flag);
8 /* A call to MPI_Win_flush would be required if MPI_Rput was used */
9 MPI_Allreduce(&another_variable, ...);

```

Listing 6: The example of Listing 1 using a remote-completing request-based put and a collective operation for synchronization.

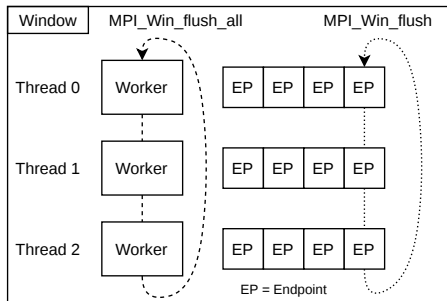


Figure 7: Flushes in the UCX one-sided communication module in Open MPI. Threads calling `MPI_Win_flush` iterate over the endpoints of all threads to ensure completion of all operations issued by the process.

out active target synchronization. By eventually removing active target synchronization, the RMA part of the standard would become cleaner and more concise, removing a significant portion of its complexity and providing easier access to the one-sided communication chapter. Moreover, we believe that the collective nature of active synchronization incurs excessive overhead and techniques mentioned earlier in this section may achieve similar goals more efficiently. However, such an effort has to be undertaken in collaboration with the community, which is beyond the scope of this paper. Previous surveys of MPI usage might help in identifying the relevant user groups [4, 21].

6 IMPLEMENTATION

6.1 Reference Implementation

The used reference implementation (the UCX one-sided communication module in Open MPI’s main development branch) employs thread-specific UCX worker objects for each window. Since UCX endpoints are specific to a worker, each thread also manages its own set of endpoints (connections with peers in the window) that are created upon the first access to that peer in the window and used to issue RMA operations. The worker and endpoints are stored in lists in the window, over which threads iterate during flushes on that window, as depicted in Figure 7. Access to that list and to each thread’s connection information are protected through mutexes to ensure thread-safety. While this scheme allows threads to issue operations on independent endpoints, it leads to significant synchronization overheads during a flush operation.

Our proof-of-concept implementations are based on this infrastructure and we note that other implementations may have different approaches, leading to different degrees of effectiveness of the proposed RMA extensions.

6.2 Thread-Scope Flushes

The aforementioned reference implementation allowed for an easy implementation of thread-scope flushes discussed in Section 2.1: instead of iterating over the list of workers or endpoints, a thread calling into a flush only operates on its local worker or endpoint.³ Thus, no access to workers or endpoints owned by other threads is necessary, greatly reducing the amount of both the amount of work and inter-thread synchronization required to achieve completion of operations issued by an individual thread.

6.3 Operation Ordering

When issuing RMA operations on windows for which the `mpi_win_order` info key discussed in Section 2.2 is set to true, the calling thread calls into `ucp_worker_fence` on the used UCX worker before issuing the actual operation. This function call ensures that the operation will complete at the target only after all previously issued operations have completed at the target.

The used UCX worker depends on the *scope* set for the window. Since `ucp_worker_fence` guarantees operation ordering for a specific worker only, all operations are funneled through the endpoints of a single worker when the *process scope* is enabled on the window (the default). While this may incur additional synchronization between threads, it allows for operation ordering without explicitly waiting for operations to complete using a full flush. With *thread scope* enabled, the calling thread invokes `ucp_worker_fence` on its local worker.

6.4 Hardware Accumulate Operations

Open MPI already provides support for an info key on windows, called `acc_single_intrinsic`, that allows users to signal to the implementation that only single-element accumulate operations with support for intrinsic operations will be used. It’s effectiveness for ensuring low-latency accumulate operation on supported hardware has been shown in [30]. Due to space limitations, we refrain from any further discussion of both the implementation and its effectiveness as the proposal presented in Section 2.3 provides the same guarantees as the existing `acc_single_intrinsic` info key.

6.5 Memory Handle Windows

In contrast to collectively allocated windows, memory handle windows only provide access to a single memory region at a specific target process. It is thus sufficient to store the information for that target in the window.⁴ Our implementation employs the parent window UCX worker and endpoint information described in Section 6.1 for communication and only stores the memory handle’s

³The proof-of-concept implementation for the thread-scope and operation ordering info keys can be found at <https://github.com/devreal/ompi/tree/mpi-win-dup-with-info>.

⁴The proof-of-concept implementation for memory handle windows can be found at <https://github.com/devreal/ompi/tree/osc-win-memhandle-parentwin>.

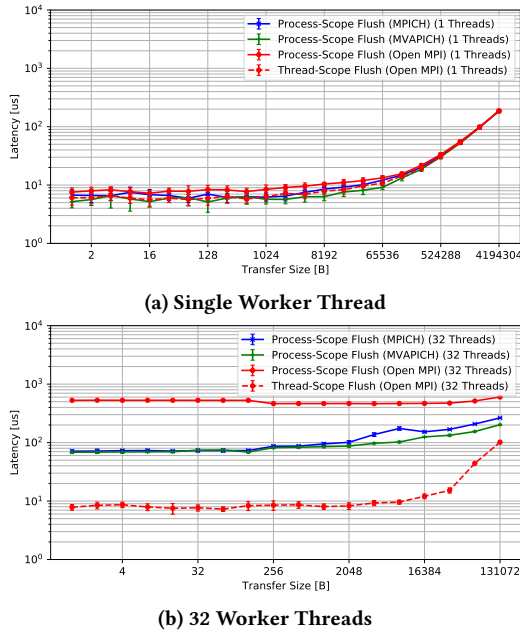


Figure 8: Latency of multi-threaded put and flush with process-scope and thread-scope flushes.

registration information in the window, allowing for fast creation (and destruction) of memory handle windows.

7 EVALUATION

7.1 Thread-Scope Flushes

We use the RMA-MT benchmark to measure latencies of RMA operations in a multi-threaded context [13]. The existing benchmark covers both active and passive target synchronization, with multiple worker threads issuing RMA operations and the main thread performing the ensuing RMA synchronization. While this pattern may be useful for fork-join thread models such as OpenMP work-sharing loops, it is inadequate for task-based applications that typically do not exhibit synchronization points. We have thus extended the benchmark to include a variant in which threads perform RMA operations followed by flushes, a pattern that is commonly found in applications using puts and flushes to ensure remote completion.

The latencies of a put followed by a flush when selecting either thread- or process-scope using the `mpi_win_scope` info key (Section 2.1) with Open MPI as well as using MPICH and MVAPICH are shown in Figure 8. The slightly higher latency of process-scope flushes in the case of a single worker thread shown in Figure 8a for Open MPI can be explained by the fact that the single worker thread has to perform a flush on its endpoint and the main thread’s endpoint, which is not required when using thread-scope flushes. Overall, however, the latencies of the different implementations are mostly similar. By contrast, for 32 worker threads shown in Figure 8b, the use of thread-scope flushes leads to an order of magnitude lower latencies for small transfer sizes compared to MPICH and MVAPICH and close to two orders of magnitude for Open MPI. For larger transfer sizes a factor of two is achieved.

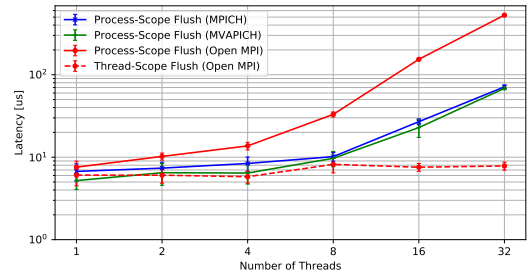


Figure 9: Latency of single-byte put and flush with process-scope and thread-scope when scaling the number of threads.

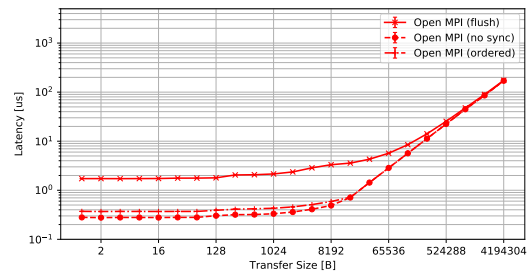


Figure 10: Latency of put with flush as well as put without intermediate synchronization but ordering using the `mpi_win_order` info key.

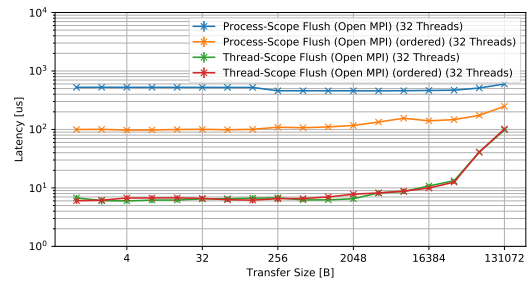


Figure 11: Latencies of put and flush using 32 worker threads with and without operation ordering enabled.

The thread-scaling behavior for single-byte transfer sizes is shown in Figure 9. The benefit of using thread-scope flushes (where appropriate) becomes clear, as it reduces the amount of work each thread has to perform inside a flush, reducing the inter-thread synchronization to a minimum.

7.2 Operation Ordering

We have modified the `osu_put_latency` benchmark to include an option suppressing intermediate flushes, i.e., puts are issued in a loop and synchronization happens during `MPI_Win_unlock` after the specified number of operations have been started. This allows us to better observe the overhead of operation ordering. Figure 10 shows the latency for regular put and flush as well as the variant without intermediate synchronization, both with and without operation ordering enabled using the `mpi_win_order` info key discussed in

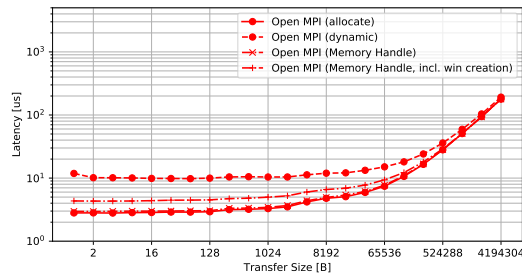


Figure 12: Latency of put using allocated windows and memory handle windows.

Section 2.2. While some additional latency can be observed due to the requested ordering, the latency of ordered puts is still significantly lower than the if flushes were used to enforce ordering of RMA operations. While this is by no means surprising (flushes likely incur at least the latency of a full network round trip), it underscores that the MPI RMA interface should provide means for ordering operations beyond waiting for completion.

We use the same RMA-MT benchmark to compare the impact of operation ordering using the `mpi_win_order` info key discussed in Section 2.2 on the latency of put operations. With 32 worker threads (Figure 11), enabling operation ordering with process-scope flushes reduces latencies since, as described in Section 6.3, all operations are funneled through a single endpoint.

7.3 Memory Handles

Similar to the previous results, the latency of puts measured using the `osu_put_latency` using allocated, dynamic, and memory handle windows are shown in Figure 12. The difference between allocated windows and windows created from memory handles is negligible. Compared to the latencies of today’s dynamic windows discussed in Section 4, the benefits of combining the flexibility of dynamic windows with the use of direct RDMA without the additional overhead of querying registration information become clear.

Latencies with added window creation and destruction are included in Figure 12, adding approximately $1 \mu\text{s}$ and still being significantly lower than the latencies for existing dynamic windows while employing the network’s RDMA capabilities. We believe that such an overhead is sufficiently low to allow applications to rapidly create and destroy memory handle windows for use with RMA operations.

8 RELATED WORK

Several improvements to MPI’s ability of handling multi-threaded communication has been proposed over the years, ranging from thread-safe probes [18] over thread-specific endpoints [11] to partitioned communication [16] and (most recently) the use of continuations [29, 33]. Work on improved implementation support for multi-threaded MPI in general [28] and RMA in particular [17] has also been described. The thread-scope flushes proposed in this work provide additional information to the implementation to better leverage some of that earlier work.

Several abstractions for one-sided communication provide collective allocation capabilities, including OpenSHMEM [7, 27] and

GASNet [5], but lack local allocation of exposed memory. Lower-level PGAS abstractions such as GASPI [1] and LCI [9] provide dynamic local allocation of exposed memory. The memory handle windows proposed in this work aim at closing the gap to these low-level abstractions and increase the flexibility of MPI RMA.

OpenSHMEM has introduced so-called contexts to provide isolation between threads, at the cost of significant extension of the API [12]. The proposed duplicate windows with thread-scope setting is an attempt to achieve a similar to goal, without significantly extending the RMA interface.

The proposed memory handle windows may be useful for applications to work around limitations of hardware tag matching engines [22] by reducing the number of exchanged messages, e.g., by organizing multiple data transfers through MPI RMA and using matched messages for signaling purposes only. The proposed info key for ordering RMA operations might enable implementations to utilize triggered operations on network hardware, which already have proven useful in the implementation of collective operations [19] and in the implementation of fence operations in OpenSHMEM [14].

9 CONCLUSIONS

We have identified several short-comings of the RMA part of the current MPI standard versions that potentially cause low performance due to high costs of synchronization and a lack of usage of available hardware resources in RMA operations. By allowing users to provide additional information on the anticipated usage of windows, implementations can adapt to the application’s usage patterns, enabling improved performance, e.g., by constraining the scope of flushes, reducing the number of flushes by enforcing the ordering of operations, and by constraining the number of elements in accumulate operations. By introducing the duplication of windows, we provide a way for users to maintain differently configured handles to the same window resources, facilitating easy switching between configurations in different parts of an application. Additionally, we propose to add the notion of memory handles to the MPI RMA interface, enabling bare-metal performance of dynamic windows by allowing users to manage memory registration information and provide life-time guarantees of memory segments, which eliminates costly querying at the target before performing RMA operations. Our benchmarks show that the proposed additions to the RMA chapter can greatly reduce the synchronization overhead, allowing applications to make better use of the hardware capabilities through the MPI RMA interface.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union’s Horizon 2020 research and innovation programme under the ChEES project, grant agreement No. 823844. This material is based upon work supported by the National Science Foundation under Grant No. #1664142 and the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the US Department of Energy Office of Science and the National Nuclear Security Administration.

REFERENCES

- [1] Thomas Alrutz, Jan Backhaus, Thomas Brandes, Vanessa End, Thomas Gerhold, Alfred Geiger, Daniel Grünewald, Vincent Heuveline, Jens Jägersküpper, Andreas Knüpfer, Olaf Krzikalla, Edmund Kügeler, Carsten Lojewski, Guy Lonsdale, Ralph Müller-Pfefferkorn, Wolfgang Nagel, Lena Oden, Franz-Josef Pfreundt, Mirko Rahn, Michael Sattler, Mareike Schmidtobreck, Annika Schiller, Christian Simmendinger, Thomas Soddemann, Godehard Sutmann, Henning Weber, and Jan-Philipp Weiss. 2013. *GASPI – A Partitioned Global Address Space Programming Interface*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35893-7_18
- [2] Bob Alverson, Edwin Froese, Larry Kaplan, and Duncan Roweth. 2012. *Cray XC Series Network*. Technical Report. Cray Inc. www.cray.com/sites/default/files/resources/CrayXCNetwork.pdf
- [3] Roberto Belli and Torsten Hoefler. 2015. Notified Access: Extending Remote Memory Access Programming Models for Producer-Consumer Synchronization. In *Proceedings of the 2015 IEEE International Parallel and Distributed Processing Symposium (IPDPS '15)*. IEEE Computer Society, 871–881. <https://doi.org/10.1109/IPDPS.2015.30>
- [4] David E. Bernholdt, Swen Boehm, George Bosilca, Manjunath Grentla Venkata, Ryan E. Grant, Thomas Naughton, Howard P. Pritchard, Martin Schulz, and Geoffroy R. Vallee. 2018. A Survey of MPI Usage in the US Exascale Computing Project. *Concurrency Computation: Practice and Experience* (09-2018 2018). <https://doi.org/10.1002/cpe.4851>
- [5] Dan Bonachea and Paul H. Hargrove. 2018. GASNet-EX: A High-Performance, Portable Communication Library for Exascale. (10 2018). <https://doi.org/10.25344/S4QP4W>
- [6] Benjamin Brock, Aydın Buluç, and Katherine Yelick. 2019. BCL: A Cross-Platform Distributed Data Structures Library. In *Proceedings of the 48th International Conference on Parallel Processing (Kyoto, Japan) (ICPP 2019)*. Association for Computing Machinery, New York, NY, USA, Article 102, 10 pages. <https://doi.org/10.1145/3337821.3337912>
- [7] Barbara Chapman, Tony Curtis, Swaroop Pophale, Stephen Poole, Jeff Kuehn, Chuck Koelbel, and Lauren Smith. 2010. Introducing OpenSHMEM: SHMEM for the PGAS Community. In *Proceedings of the Fourth Conference on Partitioned Global Address Space Programming Model (PGAS '10)*. Association for Computing Machinery. <https://doi.org/10.1145/2020373.2020375>
- [8] Gregor Daiß, Parsa Amini, John Biddiscombe, Patrick Diehl, Juhan Frank, Kevin Huck, Hartmut Kaiser, Dominic Marcelllo, David Pfander, and Dirk Pfüger. 2019. From Piz Daint to the Stars: Simulation of Stellar Mergers Using High-Level Abstractions. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19)*. <https://doi.org/10.1145/3295500.3356221>
- [9] H. Dang, R. Dathathri, G. Gill, A. Brooks, N. Dryden, A. Lenharth, L. Hoang, K. Pingali, and M. Snir. 2018. A Lightweight Communication Runtime for Distributed Graph Analytics. In *2018 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. <https://doi.org/10.1109/IPDPS.2018.00107>
- [10] James Dinan. 2018. *Query hardware acceleration for accumulates*. <https://github.com/mpiwg-rma/rma-issues/issues/6>
- [11] James Dinan, Pavan Balaji, David Godell, Douglas Miller, Marc Snir, and Rajeev Thakur. 2013. Enabling MPI Interoperability through Flexible Communication Endpoints. In *Proceedings of the 20th European MPI Users' Group Meeting (EuroMPI '13)*. Association for Computing Machinery. <https://doi.org/10.1145/2488551.2488553>
- [12] James Dinan and Mario Flajslik. 2014. Contexts: A Mechanism for High Throughput Communication in OpenSHMEM. In *Proceedings of the 8th International Conference on Partitioned Global Address Space Programming Models (PGAS '14)*. ACM, 10:1–10:9. <https://doi.org/10.1145/2676870.2676872>
- [13] Matthew G. F. Dosanjh, Taylor Groves, Ryan E. Grant, Ron Brightwell, and Patrick G. Bridges. 2016. RMA-MT: A Benchmark Suite for Assessing MPI Multi-threaded RMA Performance. In *2016 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. <https://doi.org/10.1109/CCGrid.2016.84>
- [14] M. Flajslik and J. Dinan. 2015. On the Fence: An Offload Approach to Ordering One-Sided Communication. In *2015 9th International Conference on Partitioned Global Address Space Programming Models*. 1–12. <https://doi.org/10.1109/PGAS.2015.9>
- [15] K. Fuerlinger, T. Fuchs, and R. Kowalewski. 2016. DASH: A C++ PGAS Library for Distributed Data Structures and Parallel Algorithms. In *2016 IEEE 18th International Conference on High Performance Computing and Communications*. <https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0140>
- [16] Ryan E. Grant, Matthew G. F. Dosanjh, Michael J. Levenhagen, Ron Brightwell, and Anthony Skjellum. 2019. Finepoints: Partitioned Multithreaded MPI Communication. In *High Performance Computing, Michèle Weiland, Guido Juckeland, Carsten Trinitis, and Ponnuswamy Sadayappan (Eds.)*. Springer International Publishing, Cham, 330–350.
- [17] Nathan Hjelm, Matthew G. F. Dosanjh, Ryan E. Grant, Taylor Groves, Patrick Bridges, and Dorian Arnold. 2018. Improving MPI Multi-threaded RMA Communication Performance. In *Proceedings of the 47th International Conference on Parallel Processing (ICPP 2018)*. ACM, 58:1–58:11. <https://doi.org/10.1145/3225058.3225114>
- [18] Torsten Hoefler, Greg Bronevetsky, Brian Barrett, Bronis R. de Supinski, and Andrew Lumsdaine. 2010. Efficient MPI Support for Advanced Hybrid Programming Models. In *Recent Advances in the Message Passing Interface*, Rainer Keller, Edgar Gabriel, Michael Resch, and Jack Dongarra (Eds.). Springer Berlin Heidelberg.
- [19] Nusrat Sharmin Islam, Gengbin Zheng, Sayantan Sur, Akhil Langer, and Maria Garzaran. 2019. Minimizing the Usage of Hardware Counters for Collective Communication Using Triggered Operations. In *Proceedings of the 26th European MPI Users' Group Meeting (Zürich, Switzerland) (EuroMPI '19)*. Association for Computing Machinery. <https://doi.org/10.1145/3343211.3343222>
- [20] ISO/IEC TS 19571:2016 2011. *ISO/IEC TS 19571:2016: Programming Languages – C++*. Standard. International Organization for Standardization, Geneva, CH. <https://www.iso.org/standard/50372.html>
- [21] Ignacio Laguna, Ryan Marshall, Kathryn Mohror, Martin Ruefenacht, Anthony Skjellum, and Nawrin Sultana. 2019. A Large-Scale Study of MPI Usage in Open-Source HPC Applications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (Denver, Colorado) (SC '19)*. Association for Computing Machinery. <https://doi.org/10.1145/3295500.3356176>
- [22] Scott Larson Nicoll Levy and Kurt Brian Ferreira. 2019. Evaluating Tradeoffs Between MPI Message Matching Offload Hardware Capacity and Performance. (7 2019). <https://doi.org/10.1145/3343211.3343223>
- [23] Mellanox Technologies. 2013. *Connect-IB: Architecture for Scalable High Performance Computing*. Technical Report. http://www.mellanox.com/related-docs/applications/SB_Connect-IB.pdf
- [24] MPI v2.0 2003. *MPI-2: Extensions to the Message-Passing Interface*. Technical Report. <https://www.mpi-forum.org/docs/mpi-2.0/mpi2-report.pdf> Last accessed April 23, 2021.
- [25] MPI v3.0 2012. *MPI: A Message-Passing Interface Standard, Version 3.0*. Technical Report. <https://www.mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf> Last accessed April 23, 2021.
- [26] MPI v4.0 2021. *MPI: A Message-Passing Interface Standard, Version 4.0*. Technical Report. <https://www.mpi-forum.org/docs/mpi-4.0/mpi40-report.pdf>
- [27] Open Source Software Solutions, Inc. 2020. *OpenSHMEM Application Programming Interface Version 1.5*. Open Source Software Solutions, Inc. http://openshmem.org/site/sites/default/site_files/OpenSHMEM-1.5.pdf
- [28] T. Patinyasakdikul, D. Eberius, G. Bosilca, and N. Hjelm. 2019. Give MPI Threading a Fair Chance: A Study of Multithreaded MPI Designs. In *2019 IEEE International Conference on Cluster Computing (CLUSTER)*. 1–11. <https://doi.org/10.1109/CLUSTER.2019.8891015>
- [29] Joachim Protze, Marc-André Hermans, Ali Demiralp, Matthias S. Müller, and Torsten Kuhnle. 2020. MPI Detach – Asynchronous Local Completion. In *27th European MPI Users' Group Meeting (EuroMPI/USA '20)*. Association for Computing Machinery. <https://doi.org/10.1145/3416315.3416323>
- [30] J. Schuchart, A. Bouteiller, and G. Bosilca. 2019. Using MPI-3 RMA for Active Messages. In *2019 IEEE/ACM Workshop on Exascale MPI (ExaMPI)*. 47–56.
- [31] Joseph Schuchart and José Gracia. 2019. Global Task Data-Dependencies in PGAS Applications. In *High Performance Computing, Michèle Weiland, Guido Juckeland, Carsten Trinitis, and Ponnuswamy Sadayappan (Eds.)*. Springer International Publishing.
- [32] Joseph Schuchart, Roger Kowalewski, and Karl Fuerlinger. 2018. Recent Experiences in Using MPI-3 RMA in the DASH PGAS Runtime. In *Proceedings of Workshops of HPC Asia (HPC Asia '18)*. ACM. <https://doi.org/10.1145/3176364.3176367>
- [33] Joseph Schuchart, Christoph Niethammer, and José Gracia. 2020. Fibers Are Not (P)Threads: The Case for Loose Coupling of Asynchronous Programming Models and MPI Through Continuations. In *27th European MPI Users' Group Meeting (Austin, TX, USA) (EuroMPI/USA '20)*. Association for Computing Machinery. <https://doi.org/10.1145/3416315.3416320>
- [34] Marc Sergeant, Célia Tassadit Aitkaci, Pierre Lemariner, and Guillaume Papauré. 2019. Efficient Notifications for MPI One-sided Applications. In *Proceedings of the 26th European MPI Users' Group Meeting (EuroMPI '19)*. ACM, Article 5, 5:1–5:10 pages. <https://doi.org/10.1145/3343211.3343216>
- [35] Min Si. 2018. *New info hints to enable network hardware atomics in RMA atomics*. <https://github.com/mpiwg-rma/rma-issues/issues/8>
- [36] Sayantan Sur, Hyun-Wook Jin, Lei Chai, and Dhableswar K. Panda. 2006. RDMA Read Based Rendezvous Protocol for MPI over InfiniBand: Design Alternatives and Benefits. In *Proceedings of the Eleventh ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (New York, New York, USA) (PPoPP '06)*. Association for Computing Machinery, New York, NY, USA, 32–39. <https://doi.org/10.1145/1122971.1122978>
- [37] Dong Zhong, Qinglei Cao, George Bosilca, and Jack Dongarra. 2020. Using Advanced Vector Extensions AVX-512 for MPI Reductions. In *27th European MPI Users' Group Meeting (Austin, TX, USA) (EuroMPI/USA '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3416315.3416316>