# Fine-Grained Multithreading Support for Hybrid Threaded MPI Programming

Pavan Balaji,[1] Darius Buntinas,[1] David Goodell,[1]
William Gropp,[2] and Rajeev Thakur[1]

[1]Mathematics and Computer Science Division,
Argonne National Laboratory, Argonne, IL 60439, USA

[2]Department of Computer Science,
University of Illinois, Urbana, IL, 61801, USA

**Abstract**

As high-end computing systems continue to grow in scale, recent advances in multi- and many-core architectures have pushed such growth toward more denser architectures, that is, more processing elements per physical node, rather than more physical nodes themselves. Although a large number of scientific applications have relied so far on an MPI-everywhere model for programming high-end parallel systems, this model may not be sufficient for future machines, given their physical constraints such as decreasing amounts of memory per processing element and shared caches. As a result, application and computer scientists are exploring alternative programming models that involve using MPI between address spaces and some other threaded model, such as OpenMP, Pthreads, or Intel TBB, within an address space. Such hybrid models require efficient support from an MPI implementation for MPI messages sent from multiple threads simultaneously. In this paper, we explore the issues involved in designing such an implementation. We present four approaches to building a fully thread-safe MPI implementation, with decreasing levels of critical-section granularity (from coarse-grain locks to fine-grain locks to lock-free operations) and correspondingly increasing levels of complexity. We present performance results that demonstrate the performance implications of the different approaches.

## 1 Introduction

High-end computing (HEC) systems have continued to grow in scale over the past few years. However, recent advances in multi- and many-core architectures have pushed such growth toward more denser architectures (more processing elements per physical node), rather than more physical nodes themselves. For example, more than 80% of the systems in the November 2008 ranking of the Top500 supercomputers belong to the multi/many-core processor family [20]. Even in the low-end server market, quad-core and hex-core processors are already available and are considered commodity processors today. With Intel's plans to support the 16-core Larrabee [16] processor by next year and SUN's plans to allow as many as 2048 threads within a single physical node in the near future [21], systems can be expected to continue to get denser.

The vast majority of parallel scientific applications running on HEC systems today rely on an MPI-everywhere model, where an MPI process is launched on each processing element. Each process explicitly communicates with other processes without sharing any part of the address space, regardless of whether it is on the same physical node. However, given the physical constraints of current and future generation parallel machines (including decreasing amounts of per-processing-element memory, shared caches, and per-process TLB space), many application and computer scientists are reconsidering this design and exploring alternative programming models that can be used with incremental additions to their existing programs. These models include using MPI between address spaces and relying on some threaded model, OpenMP [3], Pthreads [9], Intel Threading Building Blocks [15], within an address space. For example, the Sequoia benchmark suite [17] that was used recently for procurement of a 20 petaflops system at Lawrence Livermore National Laboratory contains many codes that use a hybrid MPI and threaded model.

The MPI-2 [11] standard already specifies a clear definition of interaction between MPI and all such models that internally rely on threads sharing the same address space. However, many MPI implementations either provide no support for such hybrid programming or rely on coarse-grained global locking to avoid multiple threads corrupting their internal stacks. This limitation is primarily because of the complexity associated with designing and implementing fine-grained locking support for threads [7]. However, as the number of processing elements within the same node continues to grow, the need for efficient threaded-MPI hybrid programming is becoming increasingly more important.

Thus, in this paper, we study the issues associated with fine-grained threading support in MPI. We propose four different approaches to building a fully thread-safe MPI implementation, with decreasing levels of critical-section granularity and correspondingly increasing levels of complexity. We also describe details of the implementation of our proposed schemes in MPICH2 [12], a popular implementation of the MPI-2 standard, as well as various experiments evaluating its performance in different scenarios. Our experimental results show that our proposed schemes can improve the performance of hybrid threaded-MPI programming significantly.

The rest of this paper is organized as follows. In Section 2 we discuss related work in the area of multithreading in MPI implementations. In Section 3 we briefly describe the thread-safety specification in MPI. In Section 4 we describe the four approaches to selecting granularity of critical sections. In Section 5 we present our experimental results. In Section 6 we present our conclusions and discuss future work.

## 2 Related Work

The issue of efficiently supporting multithreaded MPI communication has received only limited attention in the literature. In [7], we described and analyzed what the MPI Standard says about thread safety and what it implies for an implementation. We also presented an efficient multithreaded algorithm for generating new context ids, which is required for creating new communicators. Protopopov and Skjellum discuss a number of issues related to threads and MPI, including a design for a thread-safe version of MPICH-1 [14, 18]. Plachetka describes a mechanism for making a thread-unsafe PVM or MPI implementation

quasi-thread-safe by adding an interrupt mechanism and two functions to the implementation [13]. García et al. present MiMPI, a thread-safe implementation of MPI [6]. TOMPI [4] and TMPI [19] are thread-based MPI implementations, where each MPI rank is actually a thread. (Our paper focuses on conventional MPI implementations where each MPI rank is a process that itself may have multiple threads, all having the same rank.) USFMPI is a multithreaded implementation of MPI that internally uses a separate thread for communication [2]. A good discussion of the difficulty of programming with threads in general is given in [10].

## 3 Thread Safety in MPI

For performance reasons, MPI defines four "levels" of thread safety [11] and allows the user to indicate the level desired—the idea being that the implementation need not incur the cost for a higher level of thread safety than the user needs. The four levels of thread safety are as follows:

1. `MPI_THREAD_SINGLE` Each process has a single thread of execution.
2. `MPI_THREAD_FUNNELED` A process may be multithreaded, but only the thread that initialized MPI may make MPI calls.
3. `MPI_THREAD_SERIALIZED` A process may be multithreaded, but only one thread at a time may make MPI calls.
4. `MPI_THREAD_MULTIPLE` A process may be multithreaded, and multiple threads may simultaneously call MPI functions (with a few restrictions mentioned below).

MPI provides a function, `MPI_Init_thread`, by which the user can indicate the level of thread support desired, and the implementation will return the level supported. A portable program that does not call `MPI_Init_thread` should assume that only `MPI_THREAD_SINGLE` is supported. This paper focuses on the `MPI_THREAD_MULTIPLE` (fully multithreaded) case.

For `MPI_THREAD_MULTIPLE`, the MPI Standard specifies that when multiple threads make MPI calls concurrently, the outcome will be as if the calls executed sequentially in some (any) order. Also, blocking MPI calls will block only the calling thread and will not prevent other threads from running or executing MPI functions. (The example in Figure 1 must not deadlock for any ordering of thread execution.) MPI also says that it is the user's responsibility to prevent races when threads in the same application post conflicting MPI calls. For example, the user cannot call `MPI_Info_set` and `MPI_Info_free` on the same `info` object concurrently from two threads of the same process; the user must ensure that the `MPI_Info_free` is called only after `MPI_Info_set` returns on the other thread. Similarly, the user must ensure that collective operations on the same communicator, window, or file handle are correctly ordered among threads.

## 4 Choices of Critical-Section Granularity

To support multithreaded MPI communication, the implementation must protect certain data structures and portions of code from being accessed by multiple threads simultaneously

| *Process 0* | | *Process 1* | |
|---|---|---|---|
| Thread 0 | Thread 1 | Thread 0 | Thread 1 |
| MPI_Recv(src=1) | MPI_Send(dest=1) | MPI_Recv(src=0) | MPI_Send(dest=0) |

Figure 1: An implementation must ensure that this example never deadlocks for any ordering of thread execution.

in order to avoid race conditions. A portion of code so protected is called a *critical section* [5]. The granularity (size) of the critical section and the exact mechanism used to implement the critical section can have a significant impact on performance. In general, having smaller critical sections allows more concurrency among threads but incurs the cost of frequently acquiring and releasing the critical section. A critical section can be implemented either by using mutex locks or in a lock-free manner by using assembly-level atomic operations, such as compare-and-swap or fetch-and-add [8]. Mutex locks are comparatively expensive, whereas atomic operations are non-portable and can make the code more complex.

We describe four approaches to the selection of critical-section granularity in a thread-safe MPI implementation.

**Global** There is a single, global[1] critical section that is held for the duration of most MPI functions, except if the function is going to block on a network operation. In that case, the critical section is released before blocking and then reacquired after the network operation returns. A few MPI functions have no thread-safety implications and hence have no critical section (e.g., `MPI_Wtime`) [1, 7]. This is the simplest approach and is used in the past few releases of MPICH2.

**Brief Global** There is a single, global critical section, but it is held only when required. This approach permits concurrency between threads making MPI calls, except when common internal data structures are being accessed. However, it is more difficult to implement than Global, because determining where a critical section is needed, and where not, requires care.

**Per Object** There are separate critical sections for different objects and classes of objects. For example, there may be a separate critical section for communication to a particular process. This approach permits even more concurrency between threads making MPI calls, particularly if the underlying communication system supports concurrent communication to different processes. Correspondingly, it requires even more care in implementing.

**Lock Free** Instead of critical sections, lock-free (or wait-free) synchronization methods [8] are implemented by using atomic operations that exploit processor-specific features. This approach offers the potential for improved performance and greater concurrency. Complexity-wise, it is the hardest of the four.

[1] Global here means global to all threads of a process.

In this paper we implement and evaluate the first three approaches to selecting critical-section granularity. The lock-free approach is part of our future work as discussed in Section 6.

To manage building and experimenting with these four options in MPICH2, we have developed a set of abstractions built around named critical sections and related concepts. These are implemented as compile-time macros, ensuring that there is no extra overhead. Each section of code that requires atomic access to shared data structures is enclosed in a begin/end of a named critical section. In addition, the particular object (if relevant) is passed to the critical section. For example,

```
MPIU_THREAD_CS_BEGIN(COMM,vc)
... code to access a virtual communication channel vc
MPIU_THREAD_CS_END(COMM,vc)
```

In the Global mode, there is an "`ALLFUNC`" (all functions) critical section, and the other macros, such as the `COMM` one above, are defined to be empty so that there is no extra overhead. In the Brief-Global mode, the `ALLFUNC` critical section is defined to be empty, and others, such as the above `COMM` critical section, are defined to acquire and release a common, global mutex. The `vc` argument to the macro is ignored in that case. In the Per-Object mode, the situation is similar to that in Brief Global, except that instead of using a common, global mutex, the critical-section macro uses a mutex that is part of the object passed as the second argument of the macro.

## 5 Performance Evaluation

To assess the performance of each granularity option, we wrote a test that measures the message rate achieved by $n$ threads of a process sending to $n$ single-threaded receiving processes, as shown in Figure 2(a). The receiving processes prepost 128 receives using `MPI_Irecv`, send an acknowledgment to the sending threads, and then wait for the receives to complete. After receiving the acknowledgment, the threads of the sending process send 128 messages using `MPI_Send`. This process is repeated for 100,000 iterations. The acknowledgment message in each iteration ensures that the receives are posted before the sends arrive, so that there are no unexpected messages. The sending process calls `MPI_Init_thread`

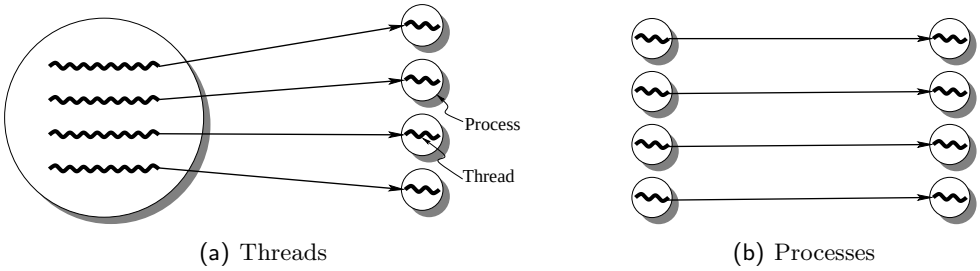


Figure 2: Illustration of test programs. Multiple threads or processes send messages to a different single-threaded receiving process.

with the thread level set to MPI_THREAD_MULTIPLE (even for runs with only one thread, in order to show the overhead of providing thread safety). The message rate is calculated as $n/avg_latency$, where $n$ is the number of sending threads or processes, and $avg_latency$ is $avg_looptime/(niters*128)$, where $avg_looptime$ is the execution time of the entire iteration loop averaged over the sending threads.

To provide a baseline message rate, we also measured the message rate achieved with separate processes (instead of threads) for sending. For this purpose, we used a modified version of the test that uses multiple single-threaded sending processes, as shown in Figure 2(b). The sending processes simply call MPI_Init, which sets the thread level to MPI_THREAD_SINGLE.

We performed three sets of experiments to measure the impact of critical-section granularity. The first set does not perform any actual communication (all sends are to MPI_PROC_NULL), the second performs blocking communication, and the third performs non-blocking communication.

The tests were run on a single Linux machine with two 2.6 GHz, quad-core Intel Clovertown chips, (a total of 8 cores), with our development version of MPICH2 in which the ch3:sock (TCP) channel was modified to incorporate the thread-safety approaches described in this paper.

### 5.1 Performance with MPI_PROC_NULL

This test is intended to measure the threading overhead in the MPICH2 code in the absence of any network communication. For this purpose, we use MPI_PROC_NULL as the destination in MPI_Send and as a source in MPI_Irecv. In MPICH2, an MPI_Send to MPI_PROC_NULL is handled at a layer above the device-specific code and does not involve manipulation of any shared data structures.

Figure 3 shows the aggregate message rate of the sending threads or processes as a function of the number of threads or processes. In the multiple-process case, the message rate increases with the number of senders because there is no contention for critical sections. In the multithreaded case with Brief Global, the performance is almost identical to multiple processes because Brief Global acquires critical sections only as needed, and in this case no critical section is needed as there is no communication. With the Global mode, however, there is a considerable decline in message rate because, in this mode, a critical section is acquired on entry to an MPI function, which serializes the accesses by different threads.

Figure 4 shows the time the multithreaded process spent waiting for a mutex, averaged over the number of threads. This figure clearly shows that there is no mutex contention for Brief-Global granularity, while for Global granularity the time a thread spends waiting for a mutex increases with the number of threads.

The number of times a mutex is acquired was counted for each send and is shown in Table 1. The first data column shows the number of times a mutex is locked when sending to MPI_PROC_NULL. We see that for Global granularity the global mutex is only acquired once, while in Brief Global it is not acquired at all.

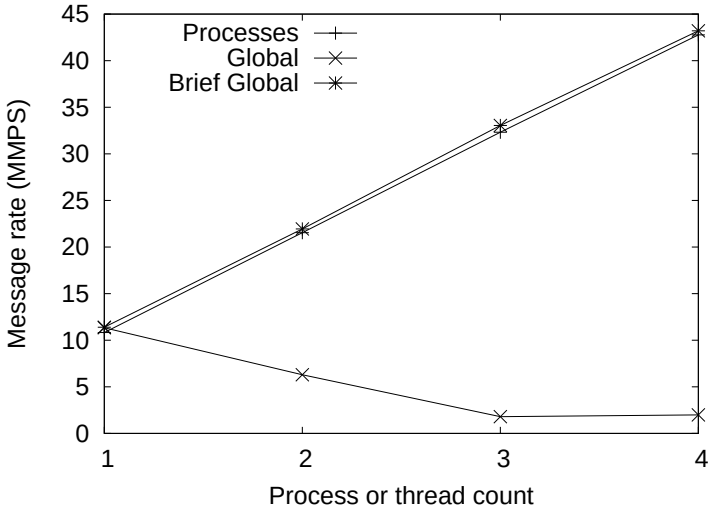


Figure 3: Message rate (in million messages per sec.) for a multithreaded process sending to `MPI_PROC_NULL` with *Global* and *Brief Global* granularities, compared to that with multiple processes.

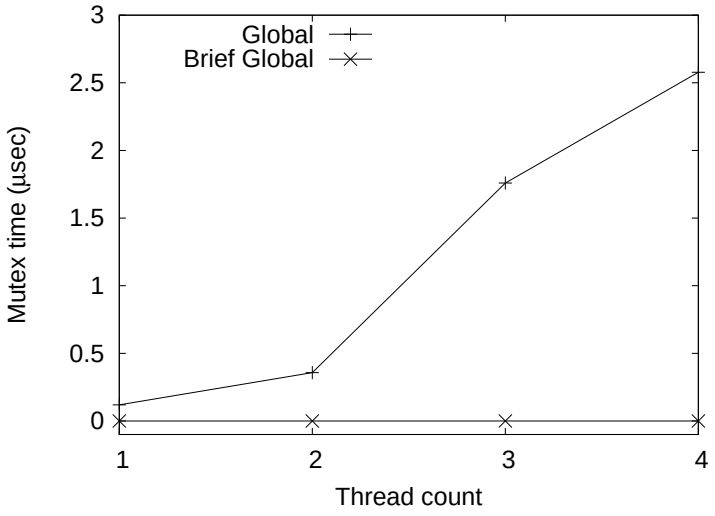


Figure 4: Per-thread mutex wait time for a multithreaded process sending to `MPI_PROC_NULL` with *Global* and *Brief Global* granularities, compared to that with multiple processes.

Table 1: Number of times a mutex is acquired per send operation for sending to `MPI_PROC_NULL` (Sec. 5.1), for blocking sends (Sec. 5.2), and for nonblocking sends (Sec. 5.3).

| | Communication Type | | |
|---|---|---|---|
| Granularity | `MPI_PROC_NULL` | Blocking | Nonblocking |
| Global | 1 | 1 | 1 |
| Brief Global | 0 | 1 | 8 |
| Per-Object | 0 | 1 | 8 |
| Per-Obj+TLS | 0 | 1 | 6 |
| Per-Obj+TLS+Atomic | 0 | 1 | 2 |

### 5.2 Performance with Blocking Sends

This test measures the performance when the communication path is exercised, which requires critical sections to be acquired. The test measures the message rate for zero-byte blocking sends. (Even for zero-byte sends, the implementation must send the message envelope to the destination because the receives could have been posted for a larger size.)

Figure 5 shows the results. Notice that because of the cost of communication, the overall message rate is considerably lower than with `MPI_PROC_NULL`. In this test, even Brief Global performs as poorly as Global because it acquires a large critical section during communication, which dominates the overall time. We then tried the Per-Object granularity, which demonstrated very good performance (comparable to multiple processes) because the granularity of critical sections in this case is per virtual channel (VC), rather than global. In MPICH2, a VC is a data structure that holds all the state and information required for a process to communicate with another process. Since each thread sends to a different process, they use separate VCs, and there is no contention for the critical section.

Figure 6 shows the mutex wait time for this test. As before, we see that with Global granularity, the mutex wait time increases with the number of threads, indicating that increasing the number of threads increases contention on the single mutex. We see a similar increase in mutex wait time with Brief-Global granularity as well because of the use of a single global mutex in both Global and Brief-Global cases. In the Per-Object case, very little time is spent waiting for mutexes because the threads are not contending for the same VC structures. The mutex wait time does increase very slightly in the Per-Object case, but it is most likely an artifact of the mechanism we used to time the mutex. From Table 1 we see that, in the blocking case, Global, Brief Global, and Per Object acquire the mutex the same number of times per send operation (once). In the Per-Object case, however, each thread locks a different mutex, resulting in a higher message rate (Figure 5) as there is no contention (Figure 6).

### 5.3 Performance with Nonblocking Sends

When performing a blocking send for short messages, MPICH2 does not need to allocate an `MPI_Request` object. For nonblocking sends, however, MPICH2 must allocate a request object to keep track of the progress of the communication operation. Requests are allocated

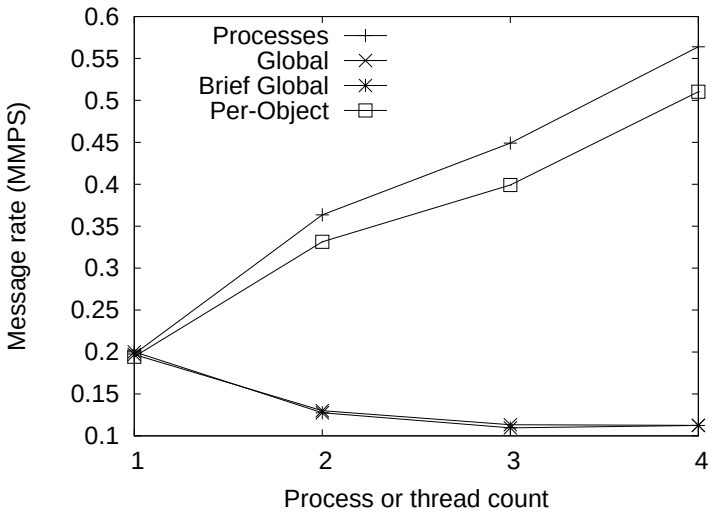


Figure 5: Message rates with blocking sends for *Global*, *Brief Global*, and *Per-Object* granularities.

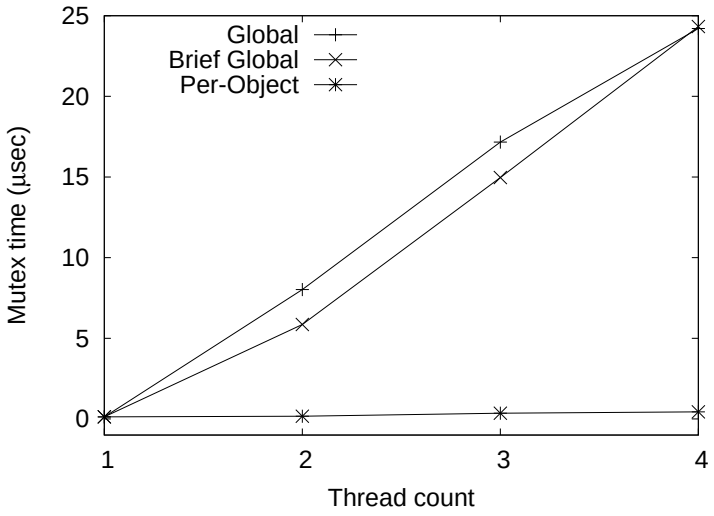


Figure 6: Per-thread mutex wait time with blocking sends for *Global*, *Brief Global*, and *Per-Object* granularities.

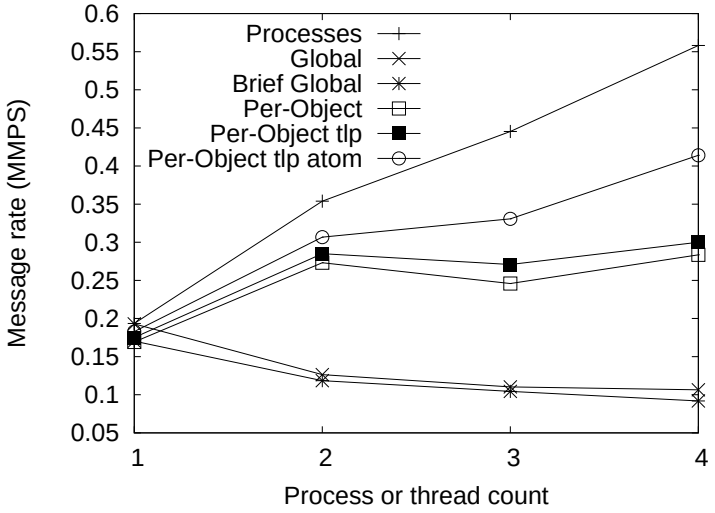


Figure 7: Message rates with nonblocking sends. *Per-Object tlp* is the thread-local request-pool optimization and *Per-Object tlp atom* updates reference counts using atomic assembly instructions.

from a common pool of free requests, which must be protected by a critical section. When a request is completed, it is freed and returned to the common pool. As a result, the common request pool becomes a source of critical-section contention.

Each request object also uses a reference count to determine when the operation is complete and when it is safe to free the object. Since any thread can cause progress on communication, any thread can increment or decrement the reference count. A critical section is therefore needed, which can become another source of contention. All this makes it more difficult to minimize threading overhead in nonblocking sends than blocking sends.

We modified the test program to use nonblocking sends and measured the message rates. Figure 7 shows the results. Notice that the performance of Per-Object granularity is considerably affected by the contention on the request pool, and the message rate does not increase beyond more than two threads.

To reduce the contention on the common request pool, we experimented with providing a local free pool for each thread. These thread-local pools are initially empty. When a thread needs to allocate a request and its local pool is empty, it will get it from the common pool. But when a request is freed, it is returned to the thread's local pool. The next time the thread needs a request, it will allocate it from its local pool and avoid acquiring the critical section for the common request pool. The graph labeled "Per-Object tlp" in Figure 7 shows that by adding the thread-local request pool, the message rate improves, but only slightly. The contention for the reference-count updates still has a negative impact on the message rate.

To alleviate the reference-count contention, we modified MPICH2 to use atomic assembly instructions for updating reference counts (instead of using a mutex). The graph labeled "Per-Object tlp atom" in Figure 7 shows that the message rate improves even further with this optimization, and increases with the number of threads. It is still less than

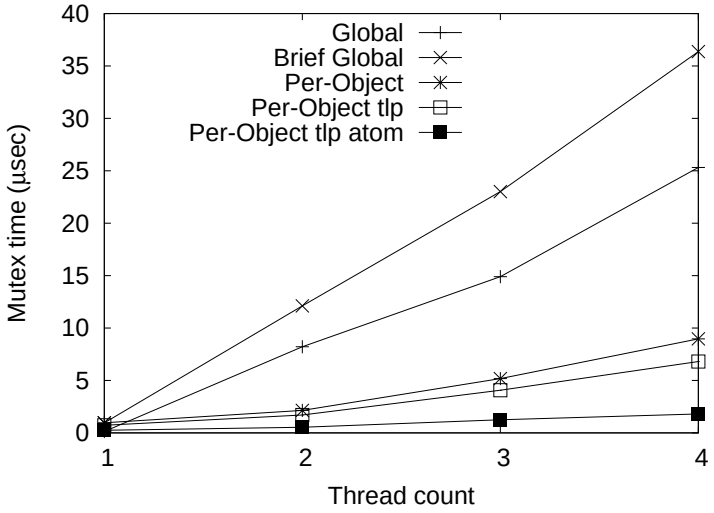


Figure 8: Per-thread mutex wait times with nonblocking sends. *Per-Object tlp* is the thread-local request-pool optimization and *Per-Object tlp atom* updates reference counts using atomic assembly instructions.

in the multiple-process case, but some performance degradation is to be expected with multithreading because critical sections cannot be completely avoided.

Figure 8 shows the mutex wait times for each of the granularities. Again, we see the mutex wait time of Global granularity increasing with thread count. Interestingly, we also see the mutex wait time of Brief Global increase faster than Global. This higher wait time is because of the smaller critical section in Brief Global, which required the mutex to be acquired eight times to send each message as shown in Table 1. Specifically, the mutex is acquired when a request object is allocated or a reference count is updated, following which the mutex is immediately released. The mutex must then be reacquired when entering the progress engine to actually send the message. As mentioned previously, while requests do not need to be allocated when performing blocking sends, they do need to be allocated for nonblocking sends. Therefore, we did not see this overhead in the previous tests. Even though the overall size of the critical section is decreased compared to Global granularity, the thread must contend for the mutex multiple times for each send, increasing the overall mutex wait time.

With Per-Object granularity, a mutex is allocated the same number of times for each nonblocking send as in Brief Global (Table 1), but the mutex wait time is much lower than Global and Brief Global (Figure 8). The lower wait time is because Per Object uses separate mutexes to lock different data structures, many of which are accessed by only one thread, and hence there is no contention. Using the thread-local request-pool optimization with Per Object reduces both the number of times a mutex needs to be locked as well as the overall mutex wait time. Combining this optimization with reference-count updates using atomic assembly instructions further decreases the mutex count and mutex wait time.

## 6 Conclusions and Future Work

We have studied the problem of improving the multithreaded performance of MPI implementations and presented several approaches to reducing the critical-section granularity, which can impact performance significantly. Such optimizations, however, require careful implementation.

While it is clear that atomic use and update of the communication engine is essential, it is equally important to ensure that all shared data structures, including MPI datatypes, requests, and communicators, are updated in a thread-safe way. For example, the reference-count updates used in most (if not all) MPI implementations must be thread atomic. This is not just a theoretical requirement: In some early experiments, we did not atomically update the reference counts, assuming that the very small race condition would not affect the results; but, by doing so, we regularly encountered failures in our communication-intensive tests. This experience suggests that the quasi-thread-safe approach proposed by Plachetka [13], in which only the access to the communication engine is serialized, is not sufficient.

We plan to implement Lock-Free granularity in MPICH2 in the future. As part of this work, we are implementing a portable library of atomic operations (such as compare-and-swap, test-and-set, and fetch-and-add). The atomic operations are implemented separately for different architectures by using assembly-language instructions. By using these atomic operations, we can replace many of the critical sections with lock-free code in a portable manner.

The abstractions we have employed to control critical-section granularity are similar to what is required for transactional memory. We plan to use these abstractions to explore the use of transactional memory.

## Acknowledgments

This work was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357 and Award DE-FG02-08ER25835. We thank Sameer Kumar and others in the MPI group at IBM Research and IBM Rochester for discussions about efficient support for thread safety in MPI.

## References

[1] Analysis of thread safety needs of MPI routines. `http://www.mcs.anl.gov/research/projects/mpich2/design/threadlist.htm`.

[2] Sadik G. Caglar, Gregory D. Benson, Qing Huang, and Cho-Wai Chu. USFMPI: A multi-threaded implementation of MPI for Linux clusters. In *Proceedings of the IASTED Conference on Parallel and Distributed Computing and Systems*, 2003.

[3] Barbara M. Chapman and Federico Massaioli. OpenMP. *Parallel Computing*, 31(10-12):957 – 959, 2005.

[4] Erik D. Demaine. A threads-only MPI implementation for the development of parallel programs. In *Proceedings of the 11th International Symposium on High Performance Computing Systems*, pages 153–163, July 1997.

[5] E. W. Dijkstra. Solution of a problem in concurrent programming control. *Communications of the ACM*, 8(9):569, September 1965.

[6] Francisco García, Alejandro Calderón, and Jesús Carretero. MiMPI: A multithread-safe implementation of MPI. In *Recent Advances in Parallel Virtual Machine and Message Passing Interface, 6th European PVM/MPI Users' Group Meeting*, pages 207–214. Lecture Notes in Computer Science 1697, Springer, September 1999.

[7] William Gropp and Rajeev Thakur. Thread safety in an MPI implementation: Requirements and analysis. *Parallel Computing*, 33(9):595–604, September 2007.

[8] Maurice Herlihy. Wait-free synchronization. *ACM Transactions on Programming Languages and Systems*, 11(1):124–149, January 1991.

[9] IEEE/ANSI Std. 1003.1. Portable Operating System Interface (POSIX)–Part 1: System Application Program Interface (API) [C Language], 1996 edition.

[10] Edward A. Lee. The problem with threads. *Computer*, 39(5):33–42, May 2006.

[11] Message Passing Interface Forum. MPI-2: Extensions to the message-passing interface, July 1997. `http://www.mpi-forum.org/docs/docs.html`.

[12] MPICH2. `http://www.mcs.anl.gov/mpi/mpich2`.

[13] Tomas Plachetka. (Quasi-) thread-safe PVM and (quasi-) thread-safe MPI without active polling. In *Recent Advances in Parallel Virtual Machine and Message Passing Interface, 9th European PVM/MPI Users' Group Meeting*, pages 296–305. Lecture Notes in Computer Science 2474, Springer, September 2002.

[14] Boris V. Protopopov and Anthony Skjellum. A multithreaded message passing interface (MPI) architecture: Performance and program issues. *Journal of Parallel and Distributed Computing*, 61(4):449–466, April 2001.

[15] James Reinders. *Intel Threading Building Blocks*. O'Reilly & Associates, Inc., Sebastopol, CA, USA, 2007.

[16] Larry Seiler, Doug Carmean, Eric Sprangle, Tom Forsyth, Michael Abrash, Pradeep Dubey, Stephen Junkins, Adam Lake, Jeremy Sugerman, Robert Cavin, Roger Espasa, Ed Grochowski, Toni Juan, and Pat Hanrahan. Larrabee: A many-core x86 architecture for visual computing. In *SIGGRAPH '08: ACM SIGGRAPH 2008 papers*, pages 1–15, New York, NY, USA, 2008. ACM.

[17] Sequoia benchmark codes. `https://asc.llnl.gov/sequoia/benchmarks/`.

[18] Anthony Skjellum, Boris Protopopov, and Shane Hebert. A thread taxonomy for MPI. In *Proceedings of the 2nd MPI Developers Conference*, pages 50–57, June 1996.

[19] Hong Tang and Tao Yang. Optimizing threaded MPI execution on SMP clusters. In *Proceedings of the 15th ACM International Conference on Supercomputing*, pages 381–392, June 2001.

[20] Top500 supercomputer sites. `http://www.top500.org/lists/2008/11`, November 2008.

[21] Ashlee Vance. Sun's Niagara 3 will have 16-cores and 16 threads per core. *The Register*, June 2008. `http://www.theregister.co.uk/2008/06/23/sun_niagara_k2`.

The submitted manuscript has been created in part by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.