Non-Data-Communication Overheads in MPI: Analysis on Blue Gene/P

P. Balaji[1], A. Chan[2], W. Gropp[3], R. Thakur[1], and E. Lusk[1]

[1] Math. and Comp. Sci. Div., Argonne Nat. Lab., Argonne, IL 60439, USA
[2] Dept. of Astronomy and Astrophysics, Univ. of Chicago, Chicago, IL 60637
[3] Dept. of Computer Science, Univ. of Illinois, Urbana, IL, 61801, USA

Abstract. Modern HEC systems, such as Blue Gene/P, rely on achieving high-performance by using the parallelism of a massive number of low-frequency/low-power processing cores. This means that the local pre- and post-communication processing required by the MPI stack might not be very fast, owing to the slow processing cores. Similarly, small amounts of serialization within the MPI stack that were *acceptable* on small/medium systems can be brutal on massively parallel systems. In this paper, we study different non-data-communication overheads within the MPI implementation on the IBM Blue Gene/P system.

1 Introduction

As we move closer to the petaflop era, modern high-end computing (HEC) systems are undergoing a drastic change in their fundamental architectural model. With processor speeds no longer doubling every 18-24 months owing to the exponential increase in power consumption and heat dissipation, modern HEC systems tend to rely lesser on the performance of single processing units. Instead, they try to extract parallelism out of a massive number of low-frequency/low-power processing cores. IBM Blue Gene/L [1] was one of the early supercomputers to follow this architectural model, soon followed by other systems such as Blue Gene/P (BG/P) [5] and SiCortex [2].

While such an architecture provides the necessary ingredients for building petaflop and larger systems, the actual performance perceived by users heavily depends on the capabilities of the systems-software stack used, such as the MPI implementation. While the network itself is quite fast and scalable on these systems, the local pre- and post-data-communication processing required by the MPI stack might not be as fast, owing to the slow processing cores. For example, local processing tasks within MPI that were considered *quick* on a 3.6 GHz Intel processor, might form a significant fraction of the overall MPI processing time on the modestly fast 850 MHz cores of a BG/P. Similarly, small amounts of serialization within the MPI stack which were considered *acceptable* on a system with a few hundred processors, can be brutal when running on massively parallel systems with hundreds of thousands of cores.

In this paper, we study the non-data-communication overheads in MPI on BG/P. We identify various bottleneck possibilities within the MPI stack, with respect to the slow pre- and post-data-communication processing as well as serialization points, stress these overheads using different benchmarks, analyze the reasons behind such overheads and describe potential solutions for solving them.

2 BG/P Software and Hardware Stacks

BG/P has five different networks [6]. Two of them (10G and 1G Ethernet with JTAG interface) are used for file I/O and system management. The other three (3-D Torus, Global Collective and Global Interrupt) are used for MPI communication. The 3-D torus network is used for MPI point-to-point and multicast operations and connects all compute nodes to form a 3-D torus. Thus, each node has six nearest-neighbors. Each link provides a bandwidth of 425 MBps per direction (total 5.1 GBps). The global collective network is a one-to-all network for compute and I/O nodes used for MPI collective communication and I/O services. Each node has three links to this network (total 5.1 GBps bidirectional). The global interrupt network is an extremely low-latency network for global barriers and interrupts, e.g., the global barrier latency of a 72K-node partition is approximately 1.3μs. On BG/P, compute cores do not handle packets on the torus network; a DMA engine on each node offloads most of the network packet injecting and receiving work, enabling better overlap of computation and communication. However, the cores handle sending/receiving packets from the collective network.

BG/P is designed for multiple programming models. The Deep Computing Messaging Framework (DCMF) and the Component Collective Messaging Interface (CCMI) are used as general purpose libraries to support different programming models [9]. DCMF implements point-to-point and multisend protocols. The multisend protocol connects the abstract implementation of collective operations in CCMI to targeted communication networks.

IBM's MPI on BG/P is based on MPICH2 and is implemented on top of DCMF. Specifically, it borrows most of the upper-level code from MPICH2, including MPI-IO and the MPE profiler, while implementing BG/P specific details within a device implementation called `dcmfd`. The DCMF library provides low-level communication support. All advanced communication features such as allocation and handling of MPI requests, dealing with tags and unexpected messages, multi-request operations such as `MPI_Waitany` or `MPI_Waitall`, derived-datatype processing and thread synchronization are **not** handled by the DCMF library and have to be taken care of by the MPI implementation.

3 Experiments and Analysis

Here, we study the non-data-communication overheads in MPI on BG/P.

3.1 Basic MPI Stack Overhead

An MPI implementation can be no faster than the underlying communication system. On BG/P, this is DCMF. Our first measurements (Figure 1) compare the communication performance of MPI (on DCMF) with the communication performance of DCMF. For MPI, we used the OSU MPI suite [10] to evaluate

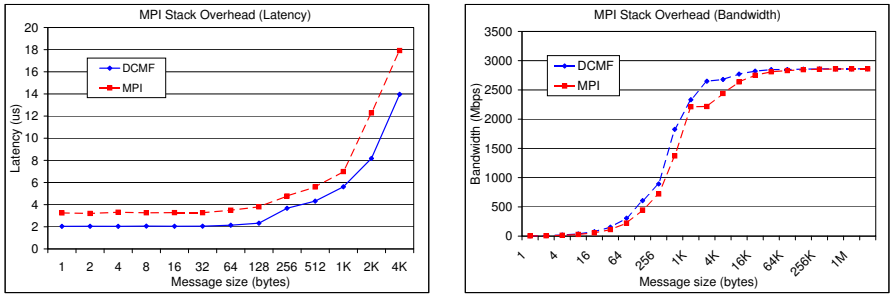

Fig. 1. MPI stack overhead

the performance. For DCMF, we used our own benchmarks on top of the DCMF API, that imitate the OSU MPI suite. The latency test uses blocking communication operations while the bandwidth test uses non-blocking communication operations for maximum performance in each case.

The difference in performance of the two stacks is the overhead introduced by the MPI implementation on BG/P. We observe that the MPI stack adds close to 1.1μs overhead for small messages; that is, close to 1000 cycles are spent for pre- and post-data-communication processing within the MPI stack. We also notice that for message sizes larger than 1KB, this overhead is much higher (closer to 4μs or 3400 cycles). This additional overhead is because the MPI stack uses a protocol switch from eager to rendezvous for message sizes larger than 1200 bytes. Though DCMF itself performs the actual rendezvous-based data communication, the MPI stack performs additional book-keeping in this mode which causes this additional overhead. In several cases, such redundant book-keeping is unnecessary and can be avoided.

3.2 Request Allocation and Queueing Overhead

MPI provides non-blocking communication routines that enable concurrent computation and communication where the hardware can support it. However, from the MPI implementation's perspective, such routines require managing `MPI_Request` handles that are needed to wait on completion for each non-blocking operation. These requests have to be allocated, initialized and queued/dequeued within the MPI implementation for each send or receive operation, thus adding overhead, especially on low-frequency cores.

In this experiment, we measure this overhead by running two versions of the typical ping-pong latency test—one using `MPI_Send` and `MPI_Recv` and the other using `MPI_Isend`, `MPI_Irecv`, and `MPI_Waitall`. The latter incurs the overhead of allocating, initializing, and queuing/dequeuing request handles. Figure 2 shows

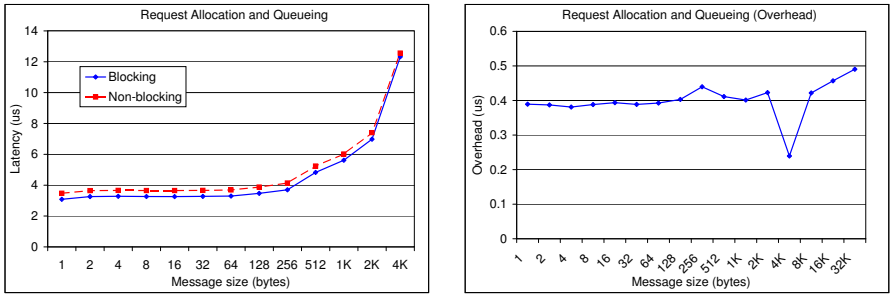

Fig. 2. Request allocation and queuing: (a) Overall performance; (b) Overhead.

that this overhead is roughly 0.4 μs or a little more than 300 clock cycles.[4] While this overhead is expected due to the number of request management operations, carefully redesigning them can potentially bring this down significantly.

3.3 Overheads in Tag and Source Matching

MPI allows applications to classify different messages into different categories using tags. Each sent message carries a tag. Each receive request contains a tag and information about which source the message is expected from. When a message arrives, the receiver searches the queue of posted receive requests to find the one that matches the arrived message (both tag and source information) and places the incoming data in the buffer described by this request. Most current MPI implementations use a single queue for all receive requests, i.e., for all tags and all source ranks. This has a potential scalability problem when the length of this queue becomes large.

To demonstrate this problem, we designed an experiment that measures the overhead of receiving a message with increasing request-queue size. In this experiment, process P0 posts M receive requests for each of N peer processes with tag T0, and finally one request of tag T1 for P1. Once all the requests are posted (ensured through a low-level hardware barrier that does not use MPI), P1 sends a message with tag T1 to P0. P0 measures the time to receive this message not including the network communication time. That is, the time is only measured for the post-data-communication phase to receive the data after it has arrived in its local temporary buffer.

Figure 3 shows the time taken by the MPI stack to receive the data after it has arrived in the local buffer. Figures 3(a) and 3(b) show two different versions of the test—the first version keeps the number of peers to one ($N = 1$) but increases the number of requests per peer (M), while the second version keeps the number

[4] This overhead is more than the entire point-to-point MPI-level shared-memory communication latency on typical commodity Intel/AMD processors [7].

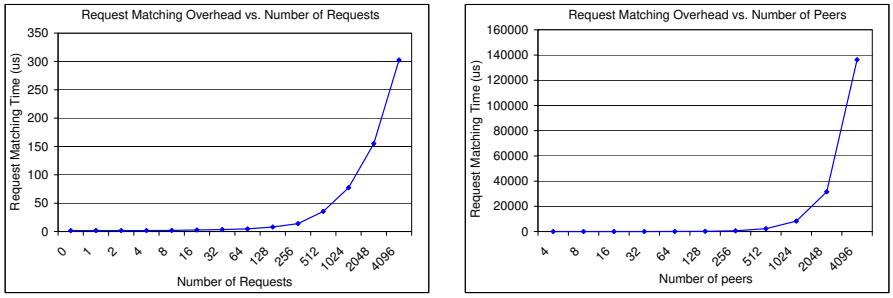

Fig. 3. Request matching overhead: (a) requests-per-peer, (b) number of peers.

of requests per peer to one ($M = 1$) but increases the number of peers (N). For both versions, the time taken increases rapidly with increasing number of total requests ($M \times N$). In fact, for 4096 peers, which is modest considering the size BG/P can scale to, we notice that even just *one request per peer* can result in a queue parsing time of about 140000μs.

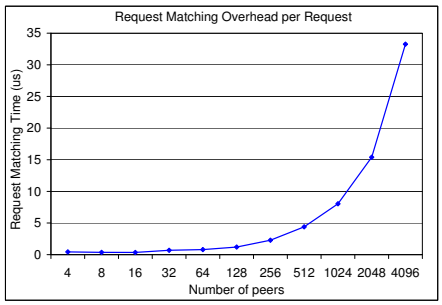

Fig. 4. Matching overhead per request

Another interesting observation in the graph is that the time increase with the number of peers is not linear. To demonstrate this, we present the average time taken per request in Figure 4—the average time per request increases as the number of requests increases! Note that parsing through the request queue should take linear time; thus the time per request should be constant, not increase. There are several reasons for such a counter-intuitive behavior; we believe the primary cause for this is the limited number of pre-allocated requests that are reused during the life-time of the application. If there are too many pending requests, the MPI implementation runs out of these pre-allocated requests and more requests are allocated dynamically.

3.4 Algorithmic Complexity of Multi-request Operations

MPI provides operations such as `MPI_Waitany`, `MPI_Waitsome` and `MPI_Waitall` that allow the user to provide multiple requests at once and wait for the completion of one or more of them. In this experiment, we measure the MPI stack's

capability to efficiently handle such requests. Specifically, the receiver posts several receive requests (`MPI_Irecv`) and once all the requests are posted (ensured through a low-level hardware barrier) the sender sends just one message that matches the first receive request. We measure the time taken to receive the message, not including the network communication time, and present it in Figure 5.

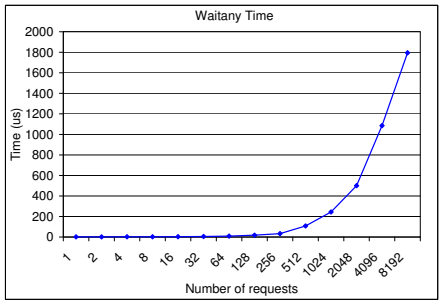

Fig. 5. `MPI_Waitany` Time

We notice that the time taken by `MPI_Waitany` increases linearly with the number of requests passed to it. We expect this time to be constant since the incoming message matches the first request itself. The reason for this behavior is the algorithmic complexity of the `MPI_Waitany` implementation. While `MPI_Waitany` would have a worst-case complexity of `O(N)`, where `N` is the number of requests, its best-case complexity should be constant (when the first request is already complete when the call is made). However, the current implementation performs this in two steps. In the first step, it gathers the internal request handles for each request (takes `O(N)` time) and in the second step does the actual check for whether any of the requests have completed. Thus, overall, even in the best case, where the completion is constant time, acquiring of internal request handlers can increase the time taken linearly with the number of requests.

3.5 Overheads in Derived Datatype Processing

MPI allows non-contiguous messages to be sent and received using derived datatypes to describe the message. Implementing these efficiently can be challenging and has been a topic of significant research [8, 11, 3]. Depending on how densely the message buffers are aligned, most MPI implementations pack sparse datatypes into contiguous temporary buffers before performing the actual communication. This stresses both the processing power and the memory/cache bandwidth of the system. To explore the efficiency of derived datatype communication on BG/P, we looked only at the simple case of a single stride (vector) type with a stride of two. Thus, every other data item is skipped, but the total amount of data packed and communicated is kept uniform across the different datatypes (equal number of bytes). The results are shown in Figure 6.

These results show a significant gap in performance between sending a contiguous messages and a non-contiguous message (with the same number of bytes). The situation is particularly serious for a vector of individual bytes (`MPI_CHAR`). It is also interesting to look at the behavior for shorter messages (Figure 6(b)). This shows, roughly, a 2 μs gap in performance between contiguous send and a send of short, integer or double precision data with a stride of two.

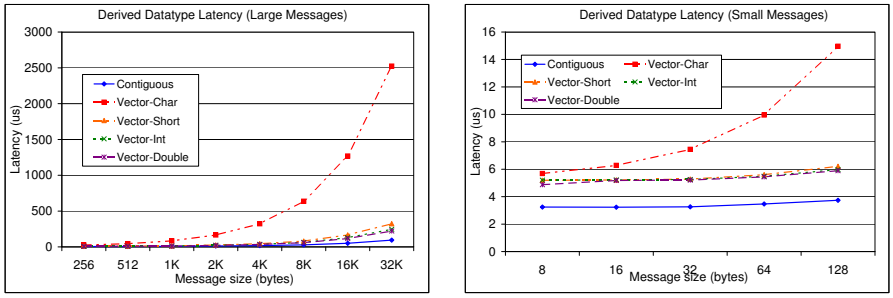

Fig. 6. Derived datatype latency: (a) long messages and (b) short messages

3.6 Buffer Alignment Overhead

For operations that involve touching the data that is being communicated (such as datatype packing), the alignment of the buffers that are being processed can play a role in overall performance if the hardware is optimized for specific buffer alignments (such as word or double-word alignments), which is common in most hardware today.

In this experiment (Figure 7), we measure the communication latency of a vector of integers (4 bytes) with a stride of 2 (that is, every alternate integer is packed and communicated). We perform the test for different alignment of these integers—"0" refers to perfect alignment to a double-word boundary, "1" refers to an misalignment of 1-byte. We notice that as long as the integers are within the same double-word (0-4 byte misalignment) the performance is better as compared to when the integers span two different double-words (5-7 byte misalignment), the performance difference being about 10%. This difference is expected as integers crossing the double-word boundary require both the double-words to be fetched before any operation can be performed on them.

3.7 Unexpected Message Overhead

MPI does not require any synchronization between the sender and receiver processes before the sender can send its data out. So, a sender can send multiple messages which are not immediately requested for by the receiver. When the receiver tries to receive the message it needs, all the previously sent messages are considered *unexpected*, and are queued within the MPI stack for later requests to handle. Consider the sender first sending multiple messages of tag T0 and finally one message of tag T1. If the receiver is first looking for the message with tag T1, it considers all the previous messages of tag T0 as *unexpected* and queues them in the unexpected queue. Such queueing and dequeuing of requests (and potentially copying data corresponding to the requests) can add overhead.

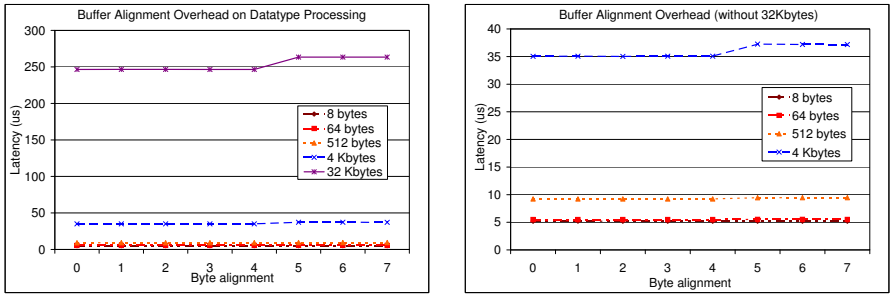

Fig. 7. Buffer alignment overhead

To illustrate this, we designed an experiment that is a symmetric-opposite of the tag-matching test described in Section 3.3. Specifically, in the tag-matching test, we queue multiple receive requests and receive one message that matches the last queued request. In the unexpected message test, we receive multiple messages, but post only one receive request for the last received message. Specifically, process P0 first receives M messages of tag T0 from each of N peer processes and finally receives one extra message of tag T1 from P1. The time taken to receive the final message (tag T1) is measured, not including the network communication time, and shown in Figure 8 as two cases: (a) when there is only one peer, but the number of unexpected messages per peer increases (x-axis), and (b) the number of unexpected messages per peer is one, but the number of peers increases. We see that the time taken to receive the last message increases linearly with the number of unexpected messages.

3.8 Overhead of Thread Communication

To support flexible hybrid programming model such as OpenMP plus MPI, MPI allows applications to perform independent communication calls from each thread by requesting for `MPI_THREAD_MULTIPLE` level of thread concurrency from the MPI implementation. In this case, the MPI implementation has to perform appropriate locks within shared regions of the stack to protect conflicts caused due to concurrent communication by all threads. Obviously, such locking has two drawbacks: (i) they add overhead and (ii) they can serialize communication.

We performed two tests to measure the overhead and serialization caused by such locking. In the first test, we use four processes on the different cores which send 0-byte messages to `MPI_PROC_NULL` (these messages incur all the overhead of the MPI stack, except that they are never sent out over the network, thus imitating an infinitely fast network). In the second test, we use four threads with `MPI_THREAD_MULTIPLE` thread concurrency to send 0-byte messages

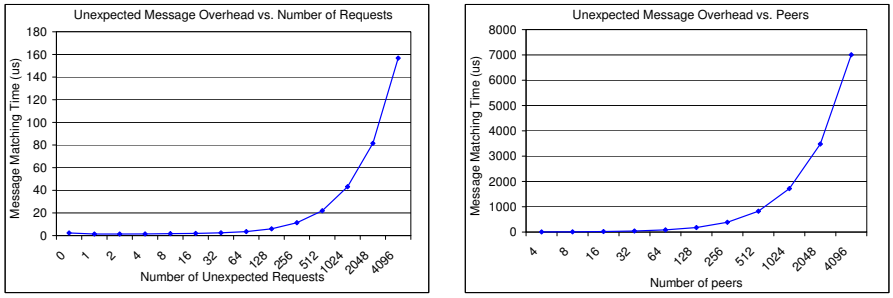

Fig. 8. Unexpected message overhead: (a) Increasing number of messages per peer, with only one peer; (b) Increasing number of peers, with only one message per peer.

to `MPI_PROC_NULL`. In the threads case, we expect the locks to add overheads and serialization, so the performance to be lesser than in the processes case.

Figure 9 shows the performance of the two tests described above. The difference between the one-process and one-thread cases is that the one-thread case requests for the `MPI_THREAD_MULTIPLE` level of thread concurrency, while the one-process case requests for no concurrency, so there are no locks. As expected, in the process case, since there are no locks, we notice a linear increase in performance with increasing number of cores used. In the threads case, however, we observe two issues: (a) the performance of one thread is significantly lower than the performance of one process and (b) the performance of threads does not increase at all as we increase the number of cores used.

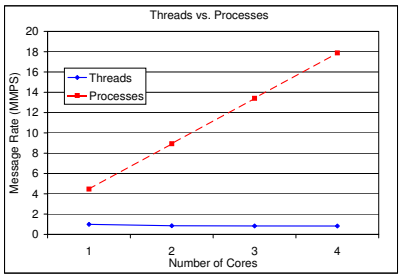

Fig. 9. Threads vs. Processes

The first observation (difference in one-process and one-thread performance) points out the overhead in maintaining locks. Note that there is no contention on the locks in this case as there is only one thread accessing them. The second observation (constant performance with increasing cores) reflects the inefficiency in the concurrency model used by the MPI implementation. Specifically, most MPI implementations perform a global lock for each MPI operation thus allowing only one thread to perform communication at any given time. This results in virtually *zero* effective concurrency in the communication of the different threads. Addressing this issue is the subject of a separate paper [4].

4 Conclusions and Future Work

In this paper, we studied the non-data-communication overheads within MPI implementations and demonstrated their impact on the IBM BlueGene/P system. We identified several bottlenecks in the MPI stack including request handling, tag matching and unexpected messages, multi-request operations (such as `MPI_Waitany`), derived-datatype processing, buffer alignment overheads and thread synchronization, that are aggravated by the low processing capabilities of the individual processing cores on the system as well as scalability issues triggered by the massive scale of the machine. Together with demonstrating and analyzing these issues, we also described potential solutions for solving these issues in future implementations.

Acknowledgments

This work was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract DE-AC02-06CH11357.

References

1. http://www.research.ibm.com/journal/rd/492/gara.pdf.
2. http://www.sicortex.com/products/sc5832.
3. P. Balaji, D. Buntinas, S. Balay, B. Smith, R. Thakur, and W. Gropp. Nonuniformly Communicating Noncontiguous Data: A Case Study with PETSc and MPI. In *IPDPS*, 2007.
4. P. Balaji, D. Buntinas, D. Goodell, W. Gropp, and R. Thakur. Toward Efficient Support for Multithreaded MPI Communication. Technical report, Argonne National Laboratory, 2008.
5. Overview of the IBM Blue Gene/P project. `http://www.research.ibm.com/journal/rd/521/team.pdf`.
6. IBM System Blue Gene Solution: Blue Gene/P Application Development. `http://www.redbooks.ibm.com/redbooks/pdfs/sg247287.pdf`.
7. D. Buntinas, G. Mercier, and W. Gropp. Implementation and Shared-Memory Evaluation of MPICH2 over the Nemesis Communication Subsystem. In *Euro PVM/MPI*, 2006.
8. W. Gropp, E. Lusk, and D. Swider. Improving the Performance of MPI Derived Datatypes. In *MPIDC*, 1999.
9. S. Kumar, G. Dozsa, G. Almasi, D. Chen, M. Giampapa, P. Heidelberger, M. Blocksome, A. Faraj, J. Parker, J. Ratterman, B. Smith, and C. Archer. The Deep Computing Messaging Framework: Generalized Scalable Message Passing on the Blue Gene/P Supercomputer. In *ICS*, 2008.
10. D. K. Panda. OSU Micro-benchmark Suite. http://mvapich.cse.ohio-state.edu/benchmarks.
11. R. Ross, N. Miller, and W. Gropp. Implementing Fast and Reusable Datatype Processing. In *Euro PVM/MPI*, 2003.