

MPI Collectives for Multi-core Clusters: Optimized Performance of the Hybrid MPI+MPI Parallel Codes

Huan Zhou
HLRS, University of Stuttgart
Stuttgart, Germany
zhou@hlrs.de

Jose Gracia
HLRS, University of Stuttgart
Stuttgart, Germany
gracia@hlrs.de

Ralf Schneider
HLRS, University of Stuttgart
Stuttgart, Germany
schneider@hlrs.de

ABSTRACT

The advent of multi-/many-core processors in clusters advocates hybrid parallel programming, which combines Message Passing Interface (MPI) for inter-node parallelism with a shared memory model for on-node parallelism. Compared to the traditional hybrid approach of MPI plus OpenMP, a new, but promising hybrid approach of MPI plus MPI-3 shared-memory extensions (MPI+MPI) is gaining attraction. We describe an algorithmic approach for collective operations (with allgather and broadcast as concrete examples) in the context of hybrid MPI+MPI, so as to minimize memory consumption and memory copies. With this approach, only one memory copy is maintained and shared by on-node processes. This allows the removal of unnecessary on-node copies of replicated data that are required between MPI processes when the collectives are invoked in the context of pure MPI. We compare our approach of collectives for hybrid MPI+MPI and the traditional one for pure MPI, and also have a discussion on the synchronization that is required to guarantee data integrity. The performance of our approach has been validated on a Cray XC40 system (Cray MPI) and NEC cluster (OpenMPI), showing that it achieves comparable or better performance for allgather operations. We have further validated our approach with a standard computational kernel, namely distributed matrix multiplication, and a Bayesian Probabilistic Matrix Factorization code.

CCS CONCEPTS

- Computing methodologies → Parallel computing methodologies.

KEYWORDS

MPI, MPI shared memory model, collective communication, hybrid programming

ACM Reference Format:

Huan Zhou, Jose Gracia, and Ralf Schneider. 2019. MPI Collectives for Multi-core Clusters: Optimized Performance of the Hybrid MPI+MPI Parallel Codes. In *Proceedings of ICPP '19: International Conference on Parallel Processing (ICPP '19)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

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.

ICPP '19, August 5–8, 2019, Kyoto, JP

© 2019 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

MPI [19] has for decades retained its dominance in the circle of the parallel programming model. It is widely used to write portable and scalable parallel applications. Although MPI was originally designed for distributed-memory systems where single-core compute nodes were connected by a network, it is also constantly tuned and adapted for gaining acceptable performance on other types of systems, such as on the shared memory symmetric multiprocessor (SMP) and compound systems tightly coupled with SMP nodes. Nowadays, the multi-core technology is extensively incorporated into the current commodity supercomputers. Therefore, the increased number of on-node data transfers come along with the growing computational capability of a single chip. This will be certain to exacerbate memory bandwidth in the pure MPI-based applications and then limit per-core affordable problem sizes since the intra-node data transfers are performed conventionally using the shared memory as the transport layer. As a result, the intention of making full use of shared memory drives the programmers to seek for the hybrid method which mixes the MPI programming model (inter-node parallelism) with a shared memory programming approach (on-node parallelism) for allowing resources on a node to be shared.

Traditionally, the shared-memory models are OpenMP [4] or Pthread [3]. This hybrid approach associates MPI rank with system-level or user-level threads, which share memory inside the same address space. OpenMP, as the most popular higher-level threading abstraction, cooperates with MPI to benefit the parallel applications, especially at high core counts where the scalability of pure MPI code suffers a lot [5, 24]. Importantly, a naive approach of migrating the pure MPI codes to the hybrid MPI+OpenMP ones is to incrementally add OpenMP directives to the computationally intense parts of the MPI application codes. This approach can produce serial sections that are only executed by the master thread as to an MPI process. In this regard, such hybrid implementation may hardly outperform the pure MPI version, especially when the scaling of the MPI version is still good [13]. Besides, the limitations of the thread-based MPI libraries are fully identified in a paper [6]. Further, the MPI-3 standard introduces a promising process-based approach, namely the MPI Shared Memory (SHM) model [2, 11, 33], for supporting memory sharing between MPI processes within one node. This leads to an innovative hybrid approach to parallel programming with the combination of MPI and MPI SHM model [10]. The migration from the pure MPI codes to the hybrid MPI+MPI ones is straightforward since the work/data distribution across MPI processes (parallelism) stays unchanged. Even when the pure MPI applications are already good in scalability, this hybrid scheme is

expected to bring performance benefit due to the unchanged computationally parallelism and less communication overhead. There are so far very limited experiences in writing efficient hybrid MPI+MPI application codes, except for the study [10] that presents a hybrid MPI+MPI programming paradigm featuring point-to-point communication operations (e.g., halo exchanges). However, this paradigm is suboptimal since the halo zones are needed for storing copies of replicated data for on-node processes. Thus this does not comply with the hybrid programming scheme requiring only one copy of replicated data between processes that are eligible to share memory.

The MPI collectives (e.g., *MPI_Bcast*, *MPI_Allgather*, and *MPI_Allreduce*) are important as they are frequently invoked in a spectrum of scientific applications or kernels [21]. Their scalability is thus closely dependent on the performance of MPI collective communication operations, especially as the HPC systems continue to grow rapidly. Thus, accelerating the collectives is a key to driving and achieving large-scale scientific simulations. Therefore, efficiently extending our experiences to include the collective communication operations in the hybrid MPI+MPI context is in high demand and appealing as well. Basically, the collectives in the pure MPI version give a copy of the result to every on-node process, which should not be the case in the hybrid MPI+MPI version.

We (take allgather and broadcast as test cases) describe a hybrid implementation approach of collective operations in the hybrid MPI+MPI context, where only one copy of replicated data is maintained within one node. Meanwhile, we discuss the programmatic differences between the hybrid approach with the naive one in the pure MPI context, under the condition that they perform the same work distribution across cores (same parallelism in computation). Unlike the naive approach, we need to take extra consideration into the synchronization between on-node processes for hybrid approach to guarantee the data integrity. Further, we evaluate the time performance of the naive and hybrid approach (synchronization overhead inclusive) for micro and application-level benchmarks.

The rest of the paper is organized as follows. In Section 2 we give the related work and then introduce the background knowledge about the shared memory model and the bridge communicator in Section 3. The detailed implementation procedures of the hybrid collectives in the MPI+MPI context are depicted in Section 4. The experimental results and analyses based on the micro and application-level benchmarks are presented in Section 5. Section 6 discusses two critical issues and Section 7 concludes.

2 RELATED WORK

Traditionally collective algorithms have been well studied. For each of collectives, many algorithms exist, where the different algorithms cater to different message sizes and numbers of processes[28]. A lot of widely-used MPI implementations, such as MPICH [20], Intel MPI [12] and OpenMPI [22], choose the most appropriate algorithm to use at runtime.

As the popularity of multi-core SMP clusters, much recent researches have been done in terms of SMP-aware (hierarchical) and shared memory collective algorithm. SMP-aware algorithms are discussed to distinguish between intra-node and inter-node communications [9, 34]. In [31], the authors, take regular all-to-all for

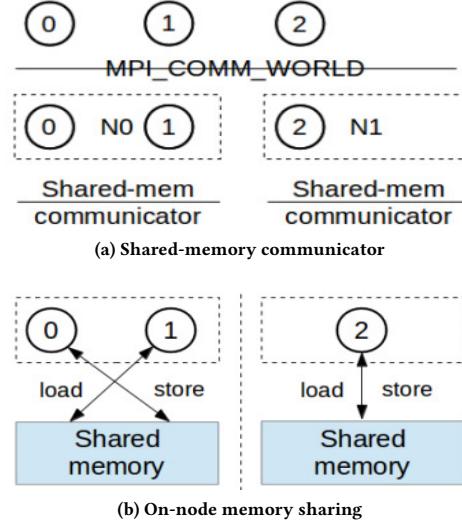


Figure 1: Shared-memory communicator splitting and shared-memory allocation.

example, give an insight into the hierarchical design of MPI collectives by solely using the MPI (3.0) functionalities for the hierarchical communicator splitting. The techniques for performing intra-node, shared-memory collectives are investigated in [6, 7, 15]. The hierarchical memory structure (multi-level caches) always appears along with the multi-cores. Several optimizations on the MPI collectives are studied based on this characteristic to take advantage of the shared cache as intra-node communication layer [8, 17]. In [18, 23], on top of SMP-aware, shared-memory is explored for intra-node communication to improve performance and RDMA is used for the inter-node communication. The above described SMP-aware collective algorithms are implemented based on a single leader per node. A multi-leader-based allgather algorithm is then proposed to reduce the contention on memory and the leader process [14].

[16] optimized the MPI+OpenMP hybrid applications by using idle OpenMP threads (refer to [27]) to parallelize the invoked MPI collectives.

3 BACKGROUND

In this section, we briefly describe the MPI-3 shared-memory facilities and the two-level (intra-node and across-node) of communicator splitting that is required for our work.

MPI SHM model: The MPI-3 SHM facility plays a significant role in supporting the hybrid MPI+MPI programming model since it allows for the direct load/store accesses to the shared data. Figure 1a splits out two disjoint, shared-memory sub-communicators out of *MPI_COMM_WORLD* on two nodes by calling *MPI_Comm_split_type* with the parameter of *MPI_COMM_TYPE_SHARED*. Each of the shared-memory communicators identifies a group of processes that can share data. Shown in Figure 1b, a region of shared-memory is allocated on each of these two shared-memory communicators via a call to *MPI_Win_allocate_shared*.

Furthermore, the function *MPI_Win_shared_query* is provided to query the base pointer to the memory on the target process. With

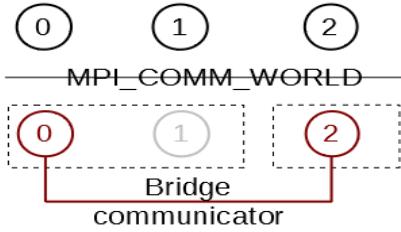


Figure 2: The creation of bridge communicator

the base pointer the locally-allocated memory can be accessed by the MPI processes in the group of the shared memory communicator with immediate load/store operations (see Fig. 1b).

Bridge communicator: Besides the above-mentioned shared-memory communicator, a bridge communicator is needed for the across-node data exchanges in the hybrid MPI+MPI codes. Shown in Figure 2, one process on each node (the lowest ranking process) is assigned as a *leader* to take responsibility for the data exchanges across nodes. All the leaders constitute a bridge communicator, which is initially brought up for the hierarchical algorithms of MPI collectives [14, 31]. Also [31] presents a paradigmatic implementation of splitting out the bridge communicator by using the facility provided by MPI-3 (*MPI_Comm_split*).

4 IMPLEMENTATION OF MPI COLLECTIVES IN THE HYBRID MPI+MPI VERSION

In the pure MPI version, the collectives require a copy of the received data to every process, which can, however, be omitted in the hybrid MPI+MPI version when the process advances the following computation by merely using the received data as an information without modifying it. This is true in most of the applications or kernels embracing the MPI collectives (refer to [21]). Therefore, it is reasonable to only maintain one global copy of the shared data, where the data to be sent is eligible to be shared by all on-node processes without the demand for intra-node memory copies. Not only is this implementation method expected to bring performance gain but it also can save the per-core memory resources and keep the per-core memory costs to constant as the number of the on-node processes increases.

Below we study the hybrid design of the widely-used collective operations – *MPI_Allgather* and *MPI_Bcast* by assuming that the SMP-style rank placement is employed. We then compare the implementation approaches of these two operations between in the pure MPI and in the hybrid MPI+MPI version and in the end perform a detailed benchmarking study to assess the benefits of using our approach over the naive one for the pure MPI version.

4.1 Allgather

The all-to-all gather problem is a typical MPI collective communication operation, where each process in a given communicator distributes personal data to all processes (including itself). Each process broadcasts data with the same size (*MPI_Allgather*) in the regular version, while the amount of data is not necessarily equal for the irregular version (*MPI_Allgatherv* [19]).

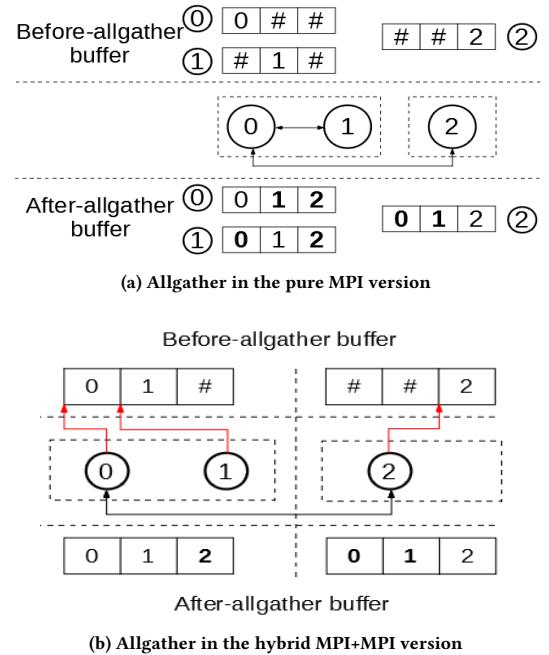


Figure 3: Illustration on the allgather approaches in the context of pure MPI and hybrid MPI+MPI. The notation $\#$ in the before-allgather buffer indicates an empty element. The received elements in the after-allgather buffer are stressed with bold font. The red arrow signifies a pointer while the black arrow denotes an inter-process communication.

Figure 3 illustrates the design difference of MPI collectives between in the pure MPI and hybrid MPI+MPI version. Refer to Sect. 6 for the case of the rank placement other than SMP-style. The two legends in this figure are drawn to show rather the communication path and status of on-node buffers than the switch of the allgather algorithms (recursive doubling or ring) in terms of different message sizes. Fig. 3a depicts the implementation algorithm of allgather in the pure MPI version and also marks the status of the buffer in each process, which maintains its buffer with the size of *MPI_COMM_WORLD* and works on it independently. We suppose that the allgather shown here is SMP-aware (cf. Sect. 2). Initially, process i assigns a valid value to the i -th element in its buffer as the local data, that is about to be sent to other processes. During this allgather operation, first the data is aggregated at the leader process, and then the leaders exchange data, and finally, the leaders broadcast data to their children. After this operation, the content sent from each process is placed in rank order in all processes' buffers and the copies of replicated data are observed from the after-allgather buffers of processes 0 and 1.

The implementation of allgather in the hybrid MPI+MPI version, shown in Fig. 3b, only demands one copy of buffer within one node, which is allocated as a shared-memory segment. The processes in the same node share this on-node buffer. Therefore, only the leaders (process 0 and 2) of the two nodes exchange all the valid elements when the all-to-all gather communication operation is required.

```

1  /* Hierarchical communicator splitting [31] */
2  comm = MPI_COMM_WORLD;
3  MPI_Comm_split_type(comm,
4      MPI_COMM_TYPE_SHARED, 0,
5      MPI_INFO_NULL, &sharedmemComm);
6  MPI_Comm_rank(sharedmemComm, &sharedmemRank);
7  leader = 0;
8  MPI_Comm_split(comm,
9      (sharedmemRank==leader)?0:MPI_UNDEFINED, 0,
10     &bridgeComm);
11 MPI_Comm_size(comm, &nprocs);
12 MPI_Comm_rank(comm, &rank);
13 msgSize = (sharedmemRank==leader)?msg*nprocs:0;
14 MPI_Win_allocate_shared(msgSize, sizeof(byte),
15     MPI_INFO_NULL, sharedmemComm,
16     &s_buf, &sharedmemWin);
17 r_buf = s_buf;
18 if (sharedmemRank != leader){
19     MPI_Win_shared_query(sharedmemWin, leader, &s_buf);
20 }
21 s_buf = s_buf + msg*rank;
22 s_buf[0..msg] = generate_random_character;
23 if (bridgeComm != MPI_COMM_NULL){// Leaders
24     if (bridgeCommSize > 1){// More than one node
25         MPI_Barrier(sharedmemComm);
26         MPI_Allgatherv(s_buf,...,r_buf, bridgeComm);
27         MPI_Barrier(sharedmemComm);
28     }
29     else{// Single node
30         MPI_Barrier(sharedmemComm);
31     }
32     else{// Children
33         if (bridgeCommSize > 1){
34             MPI_Barrier(sharedmemComm);
35             MPI_Barrier(sharedmemComm);
36         }
37         else
38             MPI_Barrier(sharedmemComm);
39     }
40 }
41 Each process accesses the updated r_buf;

```

Figure 4: Pseudo-code of the allgather implementation in the hybrid MPI+MPI version.

In order to enable the computational parallelism, a local pointer should be associated with each process to indicate an element of the on-node shared buffer. This element will be virtually viewed as the local data on which the individual process can work independently. In the example shown in Fig. 3b, process 0 and 1 point to the first and second element in the same buffer, respectively and process 2 points to the third element in a different buffer. Unlike the allgather in the pure MPI version, the aggregation and broadcast stages (intra-node message copies) are omitted in the hybrid MPI+MPI version. However, both of them face with the induced, irregular problem over sets of irregularly populated node [29]. This is due to that the sent message size varies from one node to another. Intuitively the irregular allgather variant (*MPI_Allgatherv*) is used for the across-node data transfers. This can also be replaced by other regular operations (e.g., broadcast) for performance improvement. In our approach, the irregular allgather variant is employed and evaluated in Section below.

In Figure 4 we list the allgather example written for the hybrid MPI+MPI version. First, the hierarchical (shared-memory and bridge) communicator splitting [31] is performed from line 2 to line 10. Then lines 11 through 20 allocate a contiguous region of

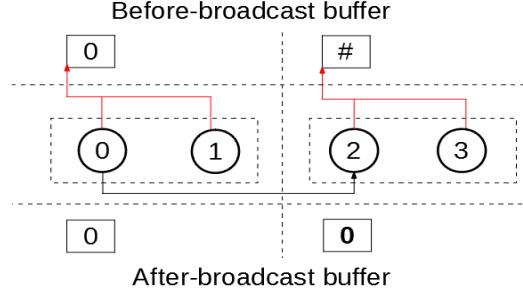


Figure 5: Illustration on the broadcast approach in the context of hybrid MPI+MPI. The buffer above denotes before-broadcast buffer, the buffer below denotes after-broadcast buffer. Refer to Fig. 3 for the the explanations of the notation **#**, bold font and arrows.

memory, that is shared among all processes in the given shared-memory communicator via a call to *MPI_Win_allocate_shared*. This function is invoked collectively with different memory sizes specified. Each of the on-node leaders asks for the contiguous memory that supposedly stores the on-node shared data. On contrary, other on-node processes (viewed as *children*) rely on the invocation of *MPI_Win_shared_query* to query the base pointer to the beginning address of the contiguous memory segment, which is originally allocated in the address space of the leader. Note that the hierarchical communicator splitting and the allocation of the shared-memory segment are one-offs, which can be amortized in the future by the repeatedly invoked *MPI_Allgatherv* operation. Next, with this base pointer, each process can compute out the address from which its local data starts and initializes its local data independently in lines 21 and 22. The local data of a process can also be viewed as its private data, on which only this process is eligible to write. The private data can also be shared (read) by other on-node processes. Finally, data exchange across nodes (on the bridge communicator) is performed from line 23 to line 39, where the omitted computation of the parameters of the sets of the received count and displacement for the irregular allgather is also a one-off. Besides the exchange, the on-node processes cooperate with each other by adding synchronization calls – barrier, which is specifically highlighted. A barrier operation is added before and after the irregular allgather operation, so as to guarantee data integrity. The leaders wait on the first barrier until all the data partitions are initialized by their children and the second barrier is invoked to make the children wait until the leaders finish data exchanges. In the extreme case of one node, only one barrier operation is performed on the shared-memory communicator to assure that the one-node buffer is ready.

4.2 Broadcast

In a broadcast operation, a message is broadcast from the root to all processes (itself included) of the given group. Figure 5 shows the implementation algorithm of the broadcast operation (with process 0 as the root) in the hybrid MPI+MPI version. Unlike the pure MPI broadcast mechanism, where each process allocates a memory buffer to store the broadcast data, we maintain one shared memory segment for the broadcast data within each node. All processes

within a node independently access the broadcast data via a local pointer to the beginning of this shared memory segment. Here, performing the across-node broadcast operations over the process 0 and 2 is straightforward since the size of the broadcast message remains the same as the broadcast in the pure MPI version.

```

1   if (rank == root)
2     s_buf[0..msg] = generate_random_character;
3   /* Broadcast across nodes */
4   if (bridgeComm != MPI_COMM_NULL){// Leaders
5     if (bridgeCommSize > 1){// More than one node
6       MPI_Bcast(s_buf,..., root, bridgeComm);
7       MPI_Barrier(sharedmemComm);
8     }
9     else// Single node
10    MPI_Barrier(sharedmemComm);
11  }
12  else// Children
13    MPI_Barrier(sharedmemComm);
14
  Each process accesses the updated s_buf;

```

Figure 6: Pseudo-code of the broadcast implementation in the hybrid MPI+MPI version.

Figure 6 lists the pseudo code of the broadcast implementation in the hybrid MPI+MPI version. Fig. 4 can be referenced for the omitted operations, such as hierarchical communicator splitting and allocation of the shared-memory segment. However, unlike the allgather operation, only the root contributes to the broadcast data (see line 1-2). Therefore, the root other than other processes is eligible to modify the shared data. In this example, one barrier operation is needed after the broadcast operation to guarantee that the broadcast data is ready for all the on-node processes.

5 EVALUATION

In this section, we compare the performance of the proposed approach of MPI collectives for the hybrid MPI+MPI version with that of the naive approach for the pure MPI version, based on micro and application-level experiments. The micro experiments are conducted by measuring the allgather latency for different schemes across different message sizes and number of cores for different cluster configurations. This micro-benchmark was modified from the OSU benchmark¹ and averaged over 10000 executions.

The two clusters are listed below:

- (1) Cray XC40 (aka. Hazel Hen): Each node of Hazel Hen is a 2-socket system equipped with Intel Haswell E5-2680v3 processors. Each compute node has 24 cores running at 2.5 GHz with 128 GB of DDR4 main memory. The nodes are connected via the Cray Aries network which has a dragonfly topology. Cray MPI library is employed on this cluster.
- (2) NEC cluster (aka. Vulcan): The architectures of node and processors are the same as Hazel Hen. The nodes are connected via the InfiniBand network. OpenMPI library is employed on this cluster.

The application-level experiments consisting of SUMMA and BPMF, are conducted on the Cluster Hazel Hen (Cray MPI). The overhead of the incurred synchronization calls is always included

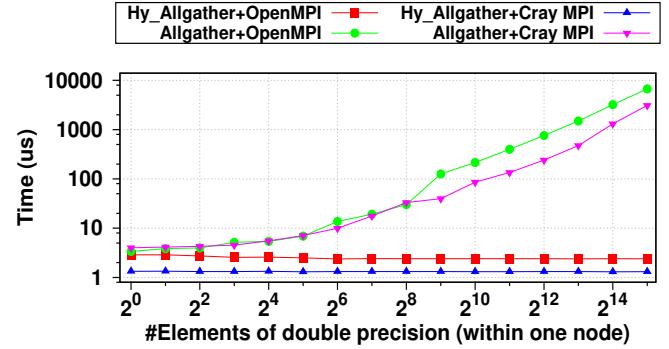


Figure 7: Comparison results between *Hy_Allgather* and *Allgather* within one full node. The number of the transferred elements is ranged from 1 to 32768.

(explained in Sect. 4) when we evaluate the time performance of our hybrid approach of MPI collectives. And the extra one-off activities are not evaluated in this section.

5.1 Allgather comparison

The abbreviations used for the allgather comparison in these micro experiments are as follows:

- *Hy_Allgather*: Our Allgather approach for the hybrid MPI+MPI version, which includes the synchronization calls.
- *Allgather*: The naive Allgather approach for the pure MPI version.

5.1.1 Extreme cases. We start with the investigation into the performance comparison between our hybrid and naive allgather approach under two extreme cases. In one scenario, we will have all MPI processes residing on the same node. In the other scenario, we allocate one MPI process per node, which means that this comparison will equal to a comparison between *MPI_Allgatherv* and *MPI_Allgather*. The count of the transferred elements of double precision floating point (8 B) is increased from 1 to 32768.

In Figure 7, we show the performance results under the first extreme case, which is also the best case for our approach since no inter-node data exchange is involved. The curves for OpenMPI and Cray MPI are of the similar tendencies. According to Fig. 4, it can be deduced that only an *MPI_Barrier* is performed for our approach and as we expected, the overhead of *Hy_Allgather* almost keeps constant. The overhead of *Allgather* goes up steadily as the message size grows and is always greater than that of *Hy_Allgather*.

Figure 8 shows the performance results under second extreme case, which is also the worst case for our approach since it loses the advantage offered by the on-node shared-memory mechanism. Figure 8a and 8b show the results running on the Vulcan with OpenMPI library and Hazel Hen with Cray MPI library, respectively. This extreme experiment is conducted across 4, 16 and 64 nodes. We can see that the time performance of *Hy_Allgather* is slightly inferior to that of *Allgather* because the implementation of *MPI_Allgatherv* is not as optimal as that of *MPI_Allgather*[29]. Further, it is obviously observed that the performance gap between *Hy_Allgather* and *Allgather* shrinks with 64 nodes.

¹<http://mvapich.cse.ohio-state.edu/benchmarks/>

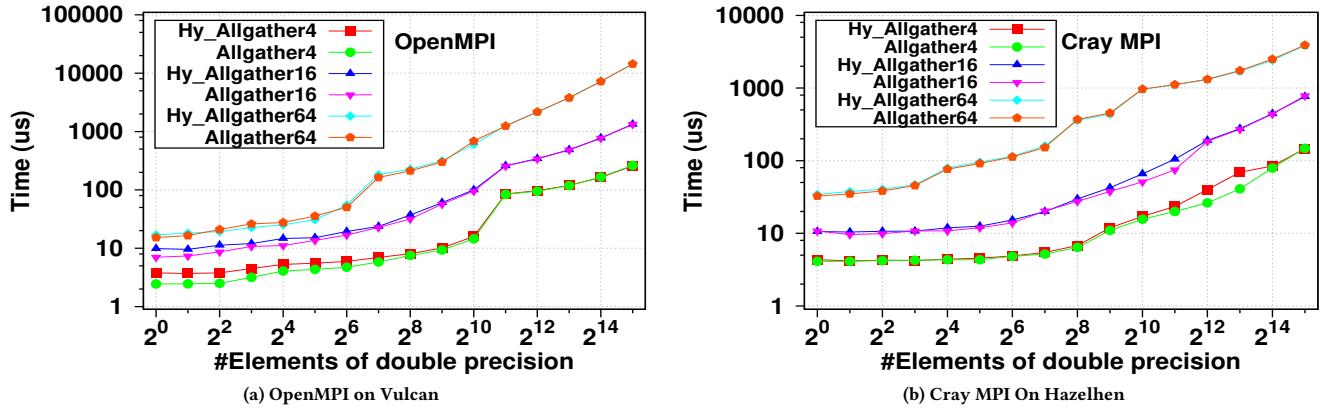


Figure 8: Comparison results between *Hy_Allgather* and *Allgather* with 4, 16 and 64 nodes and one MPI process per node. The number of transferred elements is ranged from 1 to 32768.

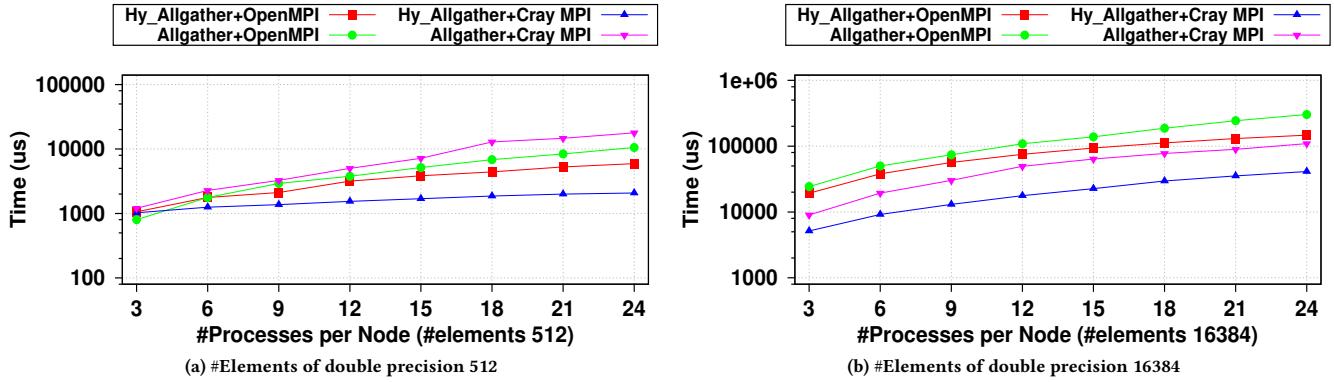


Figure 9: Comparison results between *Hy_Allgather* and *Allgather* across 64 nodes. The number of MPI processes per node ranges from 3 to 24.

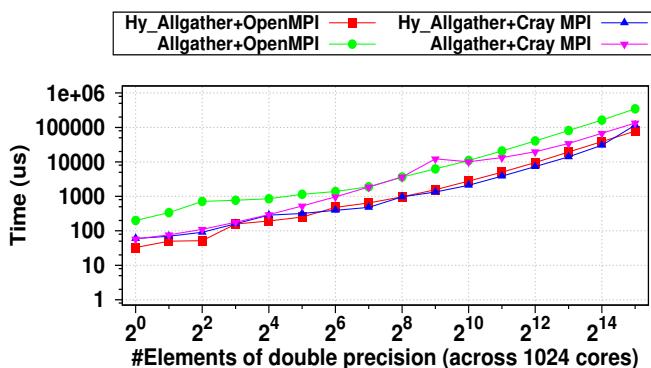


Figure 10: Comparison results between *Hy_Allgather* and *Allgather* on the irregularly populated nodes.

5.1.2 General case. In the plots shown in Sect. 5.1.1, we can identify the advantage of making use of on-node processes for our approach. Besides, our approach suffers from the scheme of one MPI process per node. Therefore, a further (more generalized) experiment is conducted to study how the advantage of *Hy_Allgather* in performance is affected by the number of processes per node. Figure 9 shows the results of this experiment with a fixed number of nodes of 64. We can observe the growths of the performance advantage of *Hy_Allgather* as the number of per-node-processes ranges from 3 to 24 for transferred element count of 512 (see Figure 9a) and 16384 (see Figure 9b) respectively. The *Hy_Allgather* starts to outperform *Allgather* when the number of processes per node is beyond 3 for OpenMPI with 512 elements and shows constant performance benefit in other situations. Moreover, like the case of 512 elements, the *Hy_Allgather* for OpenMPI with small message size (1 element) has outperformed the *Allgather* since the per-node-processes reached 3.

5.1.3 Irregular case. The experiments we discussed above cope with regular problems, where the number of MPI processes is identical across all nodes. However, *MPI_Allgatherv* suffers performance penalty from the irregular problem since its performance is determined by the maximum amount of data to be received by a node [29]. Therefore, Figure 10 compares the performance between *Hy_Allgather* and *Allgather* for an irregular problem, which allocates 24 MPI processes on 42 nodes and 16 MPI processes on one node. We can discern the advantage of our approach due to its constant lower latencies. In this section, only this irregular case is covered to preliminarily demonstrate the performance superiority of our approach even for the irregularly populated nodes.

5.2 Application-level benchmarks

In this section, we compare our approach of collectives in the hybrid MPI+MPI version to that in the pure MPI version for an application kernel of SUMMA and an application of BPMF. The software environment used for the study was CLE (Cray Linux Environment) with PrgEnv-gnu/5.2.82 and gcc version 6.2.0.

5.2.1 SUMMA. SUMMA is a scalable universal algorithm for implementing the dense matrix multiplication $C = A \times B$ [32]. In this application kernel, square matrices of size $N \times N$ are defined initially for simplicity [25]. The elements of the matrices are then decomposed in blocks of size $b \times b$ elements equally on the $\sqrt{P} \times \sqrt{P}$ cores, where b equals to N/\sqrt{P} . This application kernel consists of \sqrt{P} iterations, where each iteration triggers two broadcast communication operations of $b \times b$ data block on the row and column communicators. After obtaining the updated blocks, each process does its own computation. This benchmark is averaged over at least 20 executions.

Figure 11 compares the performance between the SUMMA implementation using the naive broadcast approach in the pure MPI version (*Ori_SUMMA*) and that using our broadcast approach in the hybrid MPI+MPI version (*Hy_SUMMA*) with the per-core matrix sizes of 8×8 (see Fig. 11a), 64×64 (see Fig. 11b), 128×128 (see Fig. 11c) and 256×256 (see Fig. 11d). The time performance ratios of *Ori_SUMMA* to *Hy_SUMMA* are also provided in Fig. 11. Each matrix element is double precision float point data type. According to Fig. 6, a barrier synchronization across the processes in the row or column communicator needs to be added after each of the two above mentioned broadcast operations in *Hy_SUMMA*. The observed ratios that are consistently over one, demonstrate the advantage of *Hy_SUMMA* in time performance (compared to *Ori_SUMMA*) with the increasing core count for varying matrix sizes. For small sizes, e.g., 8×8 , SUMMA is improved by as much as five times, when all processes are arranged in the same node. This is attributed to the fact that the *Hy_SUMMA* allows parallel computation without any data movement in between.

5.2.2 BPMF. The BPMF (Bayesian Probabilistic Matrix Factorization) application [1, 26], based on machine learning, can be used to predict compound-on-target activity in chemogenomics. The BPMF application works with the external library - Eigen. Eigen is a high-level C++ library of template headers for linear algebra, matrix and vector operations, geometrical transformations, numerical solvers and related algorithms. The BPMF version we used is an MPI code.

We did 3 runs of the application and averaged the execution times. The variance of timings was small. Within the application, the number of iterations to be sampled is set to be 20. Each iteration consists of two distinct sampling regions on movies and users. Both regions end with the all-to-all gather communication routines. Refer to the link ² for more information about the BPMF code.

In this experiment we constantly use the chembl_20 dataset as input data. Below we measure the (strong scaling) performance of the BPMF application versions implemented with the naive approach for allgather (*Ori_BPMF*) and our approach for it (*Hy_BPMF*). Specifically, Figure 12 illustrates the time performance ratio of *Ori_BPMF* to *Hy_BPMF*. According to Fig. 4, a barrier synchronization across the on-node processes needs to be added before and after the all-to-all gather communication operations in *Hy_BPMF*. It is clearly observed that the ratios are always above 1, which signifies that our allgather approach applied in *Hy_BPMF* can bring performance benefit for this BPMF application. Besides, the ratio curve is on a slow rise as the increasing number of cores. It can be deduced that the profit from the adoption of our allgather approach tends to grow steadily as the process population increases. Such incremental benefit can bring better scalability. The ratio slightly increases (by 3.9%) with a total of 1024 cores.

Additionally, each of the two barrier calls works on the processes in a shared-memory communicator (refer to Fig.1a), the size of which will not exceed 24 (aka. the number of cores per node) on Hazel Hen. Therefore, the extra overhead brought by the synchronization is independent of the number of MPI processes used, which further encourages the adoption of our approach on large-scale computers.

6 DISCUSSION

In this section, two critical issues related to the synchronization method and rank placement in our study are further addressed.

Explicit synchronization: Synchronization and communication between processes are more decoupled in hybrid MPI+MPI, compared to that in the pure MPI. In hybrid MPI+MPI, synchronization operations are explicitly added to guarantee the data integrity and support a deterministic computation. The synchronization mechanism comes in two flavors:

- Heavy-weight means: MPI Barrier across the on-node processes. This is used in our paper and is also a common synchronization mechanism to collectively synchronize among a set of processes.
- Light-weight means: pairs of MPI point-to-point communications, whereby one process can wait for another process to reach a given point in the program. This is cheaper compared to Barrier and is basically used for the non-collective (regular or irregular) communication patterns.

Rank placement: MPI processes are not necessarily mapped to the physical processors in the SMP-style way, under which our approach for allgather is proposed. In order to validate our approach for other rank placements than SMP-style, the MPI derived datatype can be employed [31]. However, the procedures of packing and unpacking always come with performance penalty. Another way is to precompute the node-sorted global rank array [31], with which

²<https://github.com/ExaScience/bpmf/>

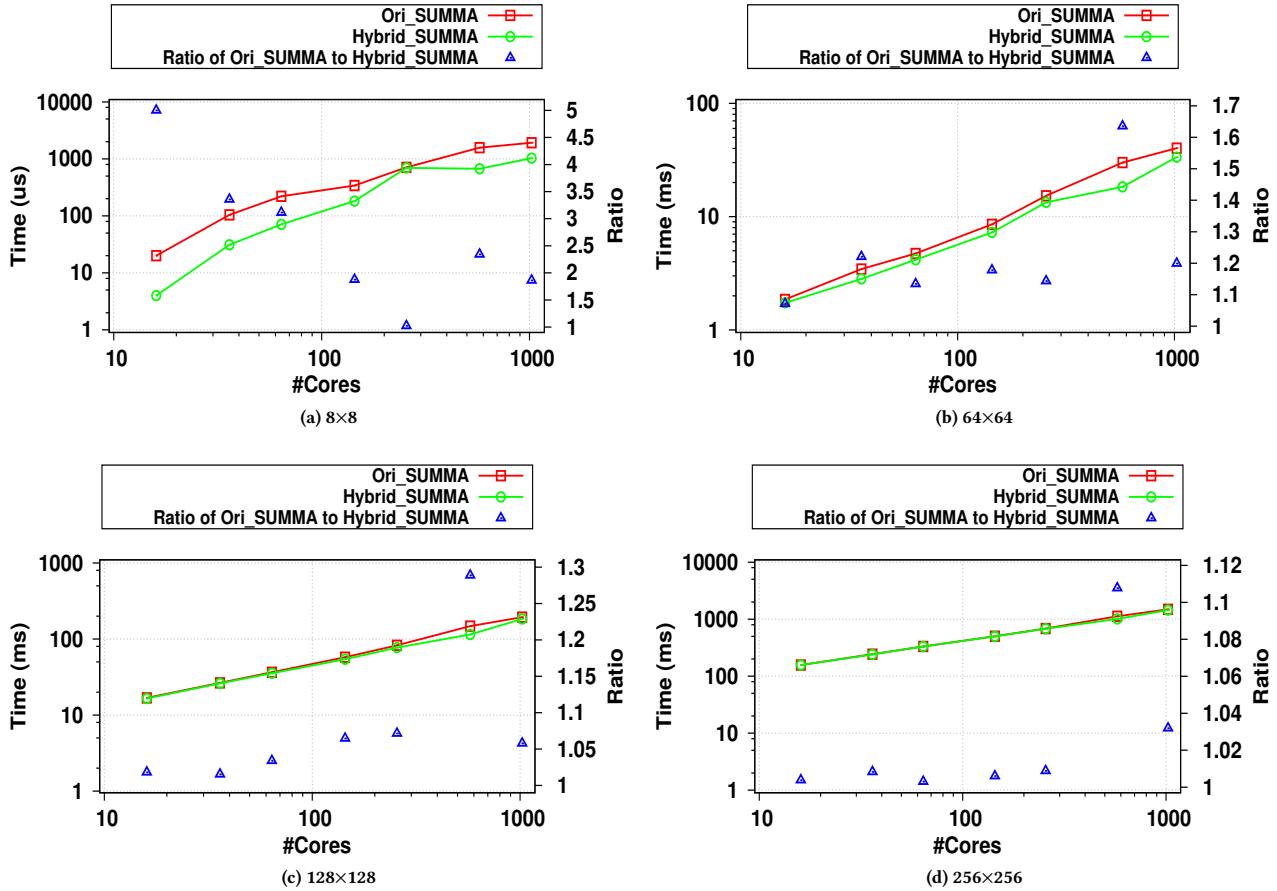


Figure 11: Performance comparison with respect to execution time (left y-axes) of the original SUMMA benchmark (red) and the version using hybrid collectives (green). Note, that the total problem size increases with number of cores and execution time is expected to scale as $\sqrt{\#cores}$. The ratio of execution times of the two versions (blue) is shown on the right y-axes. The panels correspond to different per-core matrix sizes as indicated.

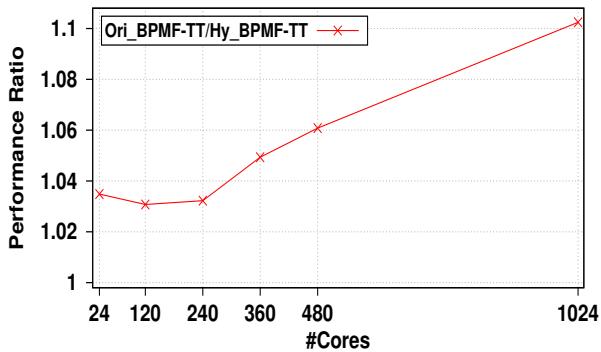


Figure 12: Performance ratio of *Ori_BPMF* to *Hy_BPMF* as the number of the MPI processes (cores) is increased from 24 to 1024, for BPMF application. TT means the TotalTime of the tested 20 iterations in BPMF application.

we can deduce the corresponding place of its block in the receive buffer in terms of any given global rank.

7 CONCLUSION

The prevailing application of multi-core technology encourages economically memory usage and thus the hybrid approach combining message passing (across-node) and shared memory (on-node). MPI-3 Shared Memory (SHM) model emerges as an innovative means of generating an efficient process-based hybrid MPI+MPI context on the multicore environment. The MPI collective operation, especially all-to-all, however, is not a scalable communication pattern. Therefore, adapting the implementation of MPI collectives to the hybrid MPI+MPI context is a key to obtaining a memory-efficient as well as high performance scientific applications. In this paper, we have proposed an efficient approach implementing the MPI collectives (such as *MPI_Bcast* and *MPI_Allgather*) in the hybrid MPI+MPI context. This can fix the memory consumption issue at the large-scale systems by only maintaining one copy of replicated data shared

by all processes within one node. Besides, the occurrence of inter-process data transfers is greatly reduced by removing the on-node memory copies. Inserting synchronization calls appropriately is necessary to guarantee the determinacy of the shared data. Our approach of allgather was on par with or outperformed that in the pure MPI context on Hazel Hen and Vulcan. We did not discuss the time performance comparison of allgather for the message sizes larger than 256 kB, where a pipeline method could, however, be applied [30]. Further, we observed consistent performance gains up to five times for the SUMMA application kernel. The total time of BPMF application using our allgather approach fell by up to 10%.

In our approach of MPI collectives for the hybrid MPI+MPI version, the issue concerning the idle cores could also happen when the leaders perform the inter-node data transfers and their children wait on a barrier operation. But this problem is lighter than that in the MPI+OpenMP applications, where the OpenMP threads are only active during the parallel computation phase and idle during the MPI calls [27]. Besides, from the programming perspective, it is straightforward to let the on-node MPI processes overlap with the network traffic by working on their own data regions. Here, the synchronization may be accelerated by using shared flags [8] to signal other processes when required.

The hybrid MPI+MPI programming is promising but has not gained the same attention as the hybrid MPI+OpenMP programming method, since so far quite a few experiences on the writing of efficient hybrid MPI+MPI application codes are explored. This paper can help to liberate the HPC programmers' mind from the traditional multi-threaded MPI approaches when it comes to multi-core environments, by highlighting an efficient approach for implementing MPI collectives in the hybrid MPI+MPI context. In addition, more experiences (e.g., p2p communications) are expected to popularize the implementation of the hybrid MPI+MPI application codes.

8 ACKNOWLEDGMENTS

The authors would like to thank Tom Vander Aa for offering the original code of BPMF application (aka. *Ori_BPMF*). Part of this work was supported by European Community's Horizon 2020 POP project under grant agreements n. 676553 and n. 824080.

REFERENCES

- [1] Tom Vander Aa, Imen Chakroun, and Tom Haber. 2016. Distributed Bayesian Probabilistic Matrix Factorization. In *CLUSTER*. IEEE Computer Society, 346–349.
- [2] Mikhail Brinckiy and Mark Lubin. 2017. *An Introduction to MPI-3 Shared Memory Programming*. Technical Report.
- [3] David R. Butenhof. 1997. *Programming with POSIX Threads*. Addison-Wesley Professional.
- [4] Leonardo Dagum and Ramesh Menon. 1998. Open MP: an Industry-Standard API for Shared-Memory Programming. *IEEE Comput Sci Eng* 5 (1998), 46–55.
- [5] Nikolaos Drosinos and Nectarios Koziris. 2004. Performance Comparison of Pure MPI vs Hybrid MPI+OpenMP Parallelization Models on SMP Clusters. In *Parallel & Distributed Processing, IPDPS*. IEEE Computer Society, 15.
- [6] Andrew Friedley, Greg Bronevetsky, Torsten Hoefer, and Andrew Lumsdaine. 2013. Hybrid MPI: efficient message passing for multi-core systems. In *Supercomputing, SC*, William Gropp and Satoshi Matsuoka (Eds.). ACM, 18:1–18:11.
- [7] Andrew Friedley, Torsten Hoefer, Greg Bronevetsky, Andrew Lumsdaine, and Ching-Chen Ma. 2013. Ownership passing: efficient distributed memory programming on multi-core systems. In *Principles and Practice of Parallel Programming, PPOPP*, Alex Nicolau, Xiaowei Shen, Saman P. Amarasinghe, and Richard W. Vuduc (Eds.). ACM, 177–186.
- [8] Richard L. Graham and Galen M. Shipman. 2008. MPI Support for Multi-core Architectures: Optimized Shared Memory Collectives. In *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting, PVM/MPI*. Springer, 84–93.
- [9] Khalid Hasanov. 2015. *Hierarchical approach to optimization of MPI collective communication algorithms*. Ph.D. Dissertation. University College Dublin, Ireland.
- [10] Torsten Hoefer, James Dinan, Darius Buntinas, Pavan Balaji, Brian Barrett, Ron Brightwell, William Gropp, Vivek Kale, and Rajeev Thakur. 2013. MPI+ MPI: a new hybrid approach to parallel programming with MPI plus shared memory. *Computing* 95, 12 (2013), 1121–1136.
- [11] Torsten Hoefer, James Dinan, Darius Buntinas, Pavan Balaji, Brian W. Barrett, Ron Brightwell, William Gropp, Vivek Kale, and Rajeev Thakur. 2012. Leveraging MPI's One-Sided Communication Interface for Shared-Memory Programming. In *European MPI Users' Group Meeting, EuroMPI (Lecture Notes in Computer Science)*, Jesper Larsson Träff, Siegfried Benkner, and Jack J. Dongarra (Eds.), Vol. 7490. Springer, 132–141.
- [12] IntelMPI. 2018. Intel MPI. Available at: <https://software.intel.com/en-us/mpi-library> (accessed 20th March 2019).
- [13] INTERTWinE. 2017. *Best Practice Guide to Hybrid MPI + OpenMP Programming*. Technical Report.
- [14] Krishna Kandalla, Hari Subramoni, Gopal Santhanaraman, Matthew Koop, and Dhabaleswar K Panda. 2009. Designing multi-leader-based allgather algorithms for multi-core clusters. In *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*. IEEE, 1–8.
- [15] Shigang Li, Torsten Hoefer, and Marc Snir. 2013. NUMA-aware shared-memory collective communication for MPI. In *High-Performance Parallel and Distributed Computing, HPDC*, Manish Parashar, Jon B. Weissman, Dick H. J. Epema, and Renato J. O. Figueiredo (Eds.). ACM, 85–96.
- [16] Aurèle Mahéo, Patrick Carribault, Marc Péache, and William Jalby. 2014. Optimizing Collective Operations in Hybrid Applications. In *European MPI Users' Group Meeting, EuroMPI/ASIA*, Jack Dongarra, Yutaka Ishikawa, and Atsushi Hori (Eds.). ACM, 121.
- [17] Amith R. Mamidala, Rahul Kumar, Debraj De, and Dhabaleswar K. Panda. 2008. MPI Collectives on Modern Multicore Clusters: Performance Optimizations and Communication Characteristics. In *Cluster, Cloud, and Grid Computing, CCGRID*. IEEE Computer Society, 130–137.
- [18] Amith R. Mamidala, Abhinav Vishnu, and Dhabaleswar K. Panda. 2006. Efficient shared memory and RDMA based design for MPI_Allgather over infiniband. In *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting, PVM/MPI*. Springer, 66–75.
- [19] Message Passing Interface Forum. 2015. MPI: A Message-Passing Interface Standard, Version 3.1. Available at: <http://mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf> (accessed 20th March 2019).
- [20] MPICH. 2019. MPICH. Available at: <https://www.mpich.org/> (accessed 20th March 2019).
- [21] NASA. 2016. NASA Parallel Benchmarks. Available at: <https://www.nas.nasa.gov/publications/npb.html> (accessed 20th March 2019).
- [22] OpenMPI. 2019. Open MPI: Open Source High Performance Computing. Available at: <https://www.open-mpi.org/> (accessed 20th March 2019).
- [23] Ying Qian and Ahmad Afshahi. 2007. RDMA-based and SMP-aware Multi-port All-Gather on Multi-rail QsNet^{II} SMP Clusters. In *International Conference on Parallel Processing, ICPP*. IEEE, 48.
- [24] Rolf Rabenseifner, Georg Hager, and Gabriele Jost. 2009. Hybrid MPI/OpenMP Parallel Programming on Clusters of Multi-Core SMP Nodes. In *Parallel, Distributed and Network-Based Processing, PDP*, Didier El Baz, François Spies, and Tom Gross (Eds.). IEEE Computer Society, 427–436.
- [25] Juan-Antonio Rico-Gallego, Juan Carlos Diaz Martin, and Alexey L. Lastovetsky. 2016. Extending τ -Lop to model concurrent MPI communications in multicore clusters. *Future Generation Comp. Syst.* 61 (2016), 66–82.
- [26] Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian Probabilistic Matrix Factorization using Markov chain Monte Carlo. In *Proceedings of the International Conference on Machine Learning*, Vol. 25.
- [27] Min Si, Antonio J. Peña, Pavan Balaji, Masamichi Takagi, and Yutaka Ishikawa. 2014. MT-MPI: multithreaded MPI for many-core environments. In *International Conference on Supercomputing, ICS*, Arndt Bode, Michael Gerndt, Per Stenström, Lawrence Rauchwerger, Barton P. Miller, and Martin Schulz (Eds.). ACM, 125–134.
- [28] Rajeev Thakur, Rolf Rabenseifner, and William Gropp. 2005. Optimization of collective communication operations in MPICH. *The International Journal of High Performance Computing Applications* 19, 1 (2005), 49–66.
- [29] Jesper Larsson Träff. 2009. Relationships between regular and irregular collective communication operations on clustered multiprocessors. *Parallel Processing Letters* 19, 01 (2009), 85–96.
- [30] Jesper Larsson Träff, Andreas Ripke, Christian Siebert, Pavan Balaji, Rajeev Thakur, and William Gropp. 2008. A simple, pipelined algorithm for large, irregular all-gather problems. In *European Parallel Virtual Machine/Message Passing Interface Users' Group Meeting, PVM/MPI*. Springer, 84–93.

- [31] Jesper Larsson Träff and Antoine Rougier. 2014. MPI collectives and datatypes for hierarchical all-to-all communication. In *European MPI Users' Group Meeting, EuroMPI/ASIA*. ACM, 27.
- [32] Robert A. Van de Geijn and Jerrell Watts. 1997. SUMMA: scalable universal matrix multiplication algorithm. *Concurrency - Practice and Experience* 9, 4 (1997), 255–274.
- [33] Huan Zhou, Kamran Idrees, and José Gracia. 2015. Leveraging MPI-3 Shared-Memory Extensions for Efficient PGAS Runtime Systems. In *European Conference on Parallel and Distributed Computing, Euro-Par (Lecture Notes in Computer Science)*, Jesper Larsson Träff, Sascha Hunold, and Francesco Versaci (Eds.), Vol. 9233. Springer, 373–384.
- [34] Hao Zhu, David Goodell, William Gropp, and Rajeev Thakur. 2009. Hierarchical collectives in mpich2. In *European Parallel Virtual Machine/Message Passing Interface UsersâŽ Group Meeting, PVM/MPI*. Springer, 325–326.