Practical, Linear-time, Fully Distributed Algorithms for Irregular Gather and Scatter*

Jesper Larsson Träff
Faculty of Informatics, TU Wien
Vienna, Austria
traff@par.tuwien.ac.at

ABSTRACT

We present new, simple, fully distributed, practical algorithms with linear time communication cost for irregular gather and scatter operations in which processors contribute or consume possibly different amounts of data. In a homogeneous, linear cost transmission model with start-up latency α and cost per unit β, the new algorithms take time $3\lceil\log_2 p\rceil\alpha + \beta\sum_{i\neq r} m_i$ where p is the number of processors, m_i the amount of data for processor $i, 0 \leq i < p$, and processor $r, 0 \leq r < p$ a root processor determined by the algorithm. With a fixed, externally given root processor r, there is an additive time penalty of at most $\beta(M_{d'} - m_{r_{d'}} - \sum_{0\leq j<d'} M_j)$ for some $d' < \lceil\log_2 p\rceil$, where each M_j is the total amount of data in a tree of 2^j different processors with roots r_j as constructed by the algorithm. The worst-case time penalty is less than $\beta\sum_{i\neq r} m_i$. The algorithms have attractive properties for implementing the operations for MPI (the Message-Passing Interface). Standard algorithms using fixed trees take time either $\lceil\log_2 p\rceil(\alpha + \beta\sum_{i\neq r} m_i)$ in the worst case, or $(p-1)\alpha + \sum_{i\neq r}\beta m_i$. We have used the new algorithms to give prototype implementations for the `MPI_Gatherv` and `MPI_Scatterv` collectives of MPI, and present benchmark results from a small and a medium-large InfiniBand cluster. In order to structure the experimental evaluation we formulate new performance guidelines for irregular collectives that can be used to assess the performance in relation to the corresponding regular collectives. We show that the new algorithms can fulfill these performance expectations within a large margin, and that standard implementations do not.

KEYWORDS

MPI, Collective operations, Performance guidelines, Irregular collectives, Gather, Scatter

ACM Reference Format:
Jesper Larsson Träff. 2017. Practical, Linear-time, Fully Distributed Algorithms for Irregular Gather and Scatter. In *Proceedings of EuroMPI/USA '17, Chicago, IL, USA, September 25–28, 2017,* 10 pages.
https://doi.org/10.1145/3127024.3127025

*This work was in part supported by the Austrian FWF project "Verifying self-consistent MPI performance guidelines" (P25530). The computational results presented have in part been achieved using the Vienna Scientific Cluster (VSC).

EuroMPI/USA '17, September 25–28, 2017, Chicago, IL, USA

ACM ISBN 978-1-4503-4849-2/17/09…$15.00
https://doi.org/10.1145/3127024.3127025

1 INTRODUCTION

Gather and scatter operations are important collective operations for collecting and distributing data among processors in a parallel system with some chosen (and known) root processor, e.g., row-column gather-scatter in linear algebra algorithms. The problems come in two flavors, namely a *regular* (or homogeneous) variant in which all processors contribute or consume blocks of the same size, and an *irregular* (or inhomogeneous) variant in which the blocks may have different sizes. For the irregular variant, the root processor may or may not know the sizes of the blocks of data to be distributed to or collected from the non-root processors. While good algorithms and implementations exist for different types of systems for the regular problems, the irregular problems have been much less studied and often only trivial algorithms with less than optimal performance (for small to medium block sizes) are implemented. In this paper, we present new, simple algorithms for the irregular gather and scatter problems with many desirable properties for the practical implementation, and show experimentally with implementations for and in MPI (the Message-Passing Interface) [11] that they can perform much better and much more consistently than common algorithms and implementations.

Gather and scatter operations are included as collective operations in MPI in both variants [11, Chapter 5]. For the regular operations `MPI_Gather` and `MPI_Scatter`, usually fixed (binomial) trees are used (hierarchically) for short to medium sized blocks, while large blocks are sent or received directly from or to the root. Since the common block size is known, the MPI processes can consistently and without any extra communication decide which algorithm to use. Standard algorithms are surveyed by Chan *et al.* [5], and analyzed under a homogeneous, linear transmission cost model where they lead to optimal, linear bandwidth, and optimal number of communication rounds (binomial trees). Similar results for different communication networks were presented early by Saad and Schulz [12]. For the irregular `MPI_Gatherv` and `MPI_Scatterv` operations where only the root process has full information on the sizes of the blocks contributed by the other, non-root processes, the situation is different. Fixed (block oblivious) trees of logarithmic depth may lead to a large block being sent a logarithmic number of times, and letting the non-root processes send or receive directly from the root entails a linear number of communication start-ups which might be too expensive when the non-root blocks are small. Current MPI libraries, nevertheless, seem to use variations of these algorithms. Träff [14] gave algorithms specifically for MPI that rely on the global information on block sizes available at the root process and use sorting to construct good trees. These algorithms may therefore be too expensive when non-root blocks

are small. Variants of these algorithms were discussed and benchmarked by Dichev *et al.* [7]. Regular gather-scatter problems for heterogeneous multiprocessors where communication links may have different capabilities have been studied in several papers, e.g., [1, 9]. These algorithms also mostly rely on global knowledge (by one process) and sorting by the transmission times between processes to construct good communication schedules, but could be adopted to irregular gather-scatter problems. Boxer and Miller [3] study the regular gather-scatter problems on the coarse grained multiprocessor (CGM) and concentrate on the problem of finding good spanning trees for the machine in case. For hypercubes, compound scatter-gather computations are studied more precisely by Charles and Fraigniaud [6] who derive pipelined schedules for the regular gather and scatter problems. Simple algorithms for the regular problems in an asynchronous communication model that accounts for delays and permits overlap were presented in [13]. Bhatt *et al.* [2] study the irregular gather and scatter problems in tree networks, and derive (nearly) optimal schedules for arbitrary trees. This situation is somewhat orthogonal to the usual objective of finding both a good spanning tree and a corresponding schedule. The algorithms require full knowledge of the message sequences to be scattered and gathered.

In the following we present new, simple algorithms for the irregular gather and scatter problems with a number of desirable properties. For the analysis, we assume a homogeneous, fully connected network with 1-ported, bidirectional (telephone-like) communication. We let p denote the number of processors which are numbered consecutively from 0 to $p-1$. We assume that the cost of transmitting a message of m units between any two processors is linear and modeled as $\alpha+\beta m$, where α is a communication start-up latency, and β the transmission time per unit. A processor involved in communication can start the next transmission as soon as it has finished and selects from which other processor to receive the next message. In the irregular gather and scatter problems, each processor $i, 0 \leq i < p$ has a block of data of size m_i with $m_i \geq 0$ that it either wants to contribute to (gather) or consume from (scatter) some root processor $r, 0 \leq r < p$. The root r is usually a given processor, and this r is known to all other processors. At the root, blocks are stored in processor order, that is $m_0, m_1, m_2, \ldots, m_{p-1}$ (we assume that the root also has a block m_r for itself that does not have to be transmitted). Any consecutive sequence of blocks can be sent or received together as a single message. Our algorithms do not assume that the root knows the size of all p data blocks, although the `MPI_Gatherv` and `MPI_Scatterv` operations do make this assumption and require this to be the case.

Our algorithms construct spanning trees of logarithmic depth, and need only the optimal $d = \lceil \log_2 p \rceil$ number of communication rounds for the tree construction, each round consisting of at most two communication steps. For the gathering or scattering of the data blocks, another at most $\lceil \log_2 p \rceil$ communication rounds are needed. Trees are constructed in a distributed manner, with each processor working only from gradually accumulated information, with no dependence on global information (e.g., from the root) on the sizes of other data blocks. The time for the root to gather or scatter all data blocks from or to the non-root processors is linear, namely $\lceil \log_2 p \rceil \alpha + \beta \sum_{0 \leq i < p, i \neq r} m_i$, with an additive time penalty of at most $\beta(M_{d'} - m_{r_{d'}} - \sum_{0 \leq j < d'} M_j)$ for some $d' < d$ where each M_j is the total amount of data in a tree of 2^j different processors as constructed by the algorithm for the case when the root is a fixed, externally given process (as is the case in `MPI_Gatherv` and `MPI_Scatterv`). The worst-case time penalty is (much) less than $\beta \sum_{i \neq r} m_i$. In contrast, for any fixed, block-size oblivious binomial tree it is easy to construct a worst case problem instance taking $\lceil \log_2 p \rceil (\alpha + \beta \sum_{0 \leq i < p, i \neq r} m_i)$ time steps, namely by choosing $m_i = 0$ for all processors except one being farthest away from the root. At all processors, blocks are always sent and received in order: Any receive operation receives a message consisting of blocks $m_k, m_{k+1}, \ldots, m_{k+l}$. No, potentially costly, local reordering of blocks in message buffers is therefore necessary.

We have implemented our algorithms[1] to support the `MPI_Gatherv` and `MPI_Scatterv` operations, and evaluated them with different block size distributions on a small InfiniBand cluster under three different MPI libraries, and a medium-large InfiniBand cluster under the vendor (Intel) MPI library. In order to structure the comparison against the native MPI library implementations we formulate expectations on the relative performance as new, self-consistent performance guidelines [10, 18]. We can show that the new algorithms can in many situations significantly outperform the native MPI library, and overall much better fulfill the formalized performance expectations.

2 PROBLEM AND ALGORITHM

We now present the algorithm for the irregular gather problem; the scatter algorithm is analogous. Each of the p processors has a data block of m_i units that it needs to contribute to some root process $r, 0 \leq r < p$. We organize the p processors in a $\lceil \log_2 p \rceil$-dimensional (incomplete), ordered hypercube which we use as a design vehicle, but communication can be between processors that are not adjacent in the hypercube. We let $H_d, 0 \leq d \leq \lceil \log_2 p \rceil$ denote a d-dimensional (incomplete) hypercube consisting of (at most) 2^d processors. We say that the hypercube H_d is *ordered* if the processors belonging to H_d form a consecutive range $[a2^d, \ldots, a2^d + 2^d - 1] = [a2^d, \ldots, (a+1)2^d - 1]$ for $a \in \{0, \ldots, \lceil p/2^d \rceil - 1\}$. The ordered hypercube H_{d+1} consisting of processors $[a2^{d+1}, \ldots, (a+1)2^{d+1} - 1]$ is built from two *adjacent*, ordered hypercubes H_d with processors $[2a2^d, \ldots, (2a+1)2^d - 1]$ and $[(2a+1)2^d, \ldots, (2a+2)2^d - 1]$. If p is not a power of two, the last H_d hypercube consists of the processors $[(\lceil p/2^d \rceil - 1)2^d, \ldots, p-1]$.

By an *ordered hypercube gather algorithm* for H_d we mean an algorithm for H_d in which a processor in one of the subcubes H_{d-1} which has gathered all data from the processors of this subcube sends all its data to a processor in the other subcube H_{d-1} which similarly has already gathered all data from that subcube. This processor will now have gathered all data in the hypercube H_d and will become the root processor of H_d. Note that this may require communicating along edges that do not belong to the hypercube, but of course do belong to the fully connected network.

[1] The prototype implementations used for evaluation are available at www.par.tuwien.ac.at/Downloads/TUWMPI/tuwgatherv.c

Lemma 2.1. *For any H_d, there exists an ordered hypercube gather algorithm that gathers the data to some root processor r in H_d in time $d\alpha + \beta \sum_{i \in H_d, i \neq r} m_i$.*

Proof. The claim follows by induction on d. For H_0 the sole processor $r \in H_0$ already has the data m_0 and there is no further cost. Let H'_{d-1} and H''_{d-1} be the two subcubes of H_d. By the induction hypothesis there is a processor r' of H'_{d-1} that has gathered all data of H'_{d-1} in $t' = (d-1)\alpha + \beta \sum_{i \in H'_{d-1}, i \neq r'} m_i$ time steps, and a processor r'' that has gathered all data of H''_{d-1} in $t'' = (d-1)\alpha + \beta \sum_{i \in H''_{d-1}, i \neq r''} m_i$ time steps. Of the two root processors r' and r'', the one with the smaller gather time (with ties broken in favor of the hypercube with the smallest amount of data) sends its data to the other root processor. Say, r' is the root with $t' \leq t''$. Processor r' sends a message of $\sum_{i \in H'_{d-1}} m_i$ units to root r'' which takes $\alpha + \beta \sum_{i \in H'_{d-1}} m_i$ time steps. Adding to the time t'' already taken by the slower r'' to gather the data from H''_{d-1} gives $(d-1)\alpha + \beta \sum_{i \in H''_{d-1}, i \neq r''} m_i + \alpha + \beta \sum_{i \in H'_{d-1}} m_i = d\alpha + \beta \sum_{i \in H_d, i \neq r} m_i$ as claimed. The root r'' of H''_{d-1} becomes the root r of H_d. □

Since roots with smaller gather times sends to roots with larger gather times, communication can readily take place with no delay for the sending gather root processor to become ready. Since subcubes are ordered, the data blocks received at a new root can easily be kept in consecutive order. Note that for the gather times of the two roots r' and r'', $t' \leq t''$ if and only if $\sum_{i \in H'_{d-1}, i \neq r'} m_i \leq \sum_{i \in H''_{d-1}, i \neq r''} m_i$, so that the shape of the constructed gather tree depends only on the block sizes and not on the relative magnitudes of α and β. For each H_d hypercube with root r, $\sum_{i \in H_{d-1}, i \neq r'} m_i$ is therefore an estimate of the time to construct H_d.

Lemma 2.2. *For any arbitrarily given root processor $r \in H_d$, there is an ordered hypercube algorithm that gathers all data in H_d to r in $d\alpha + \beta \sum_{i \in H_d, i \neq r} m_i$ time units with an additive time penalty of at most $\beta(M_{d'} - m_{r_{d'}} - \sum_{0 \leq j < d'} M_j)$ for some $d', d' < d$. The root processor gathers data from the roots in a sequence of ordered hypercubes $H_0, H_1, \ldots, H_{d-1}$, each with a total amount of data M_j, and d' is the last such hypercube for which waiting time is incurred.*

Proof. The construction of Lemma 2.1 is modified such that data are always sent to processor r if either $r' = r$ or $r'' = r$. The given root processor r will therefore receive blocks from $d-1$ linear gather time subcubes $H_0, H_1, \ldots, H_{d-1}$. The amount of data, and the time needed to gather the data in these d hypercubes is unrelated and may differ. Let $M_j = \sum_{i \in H_j} m_i$ be the amount of data in hypercube H_j with root processor r_j. If the time needed to gather the data in some $H_{d'}$ to $r_{d'}$, namely $\alpha d' + \beta(M_{d'} - m_{r_{d'}})$, is larger than the time needed to gather the data from the previous hypercubes $H_0, H_1, \ldots H_{d'-1}$ to r, the root processor is delayed until the data gather in $H_{d'}$ has completed. This delay is at most $\alpha d' + \beta(M_{d'} - m_{r_{d'}}) - (\alpha d' + \beta(\sum_{j<d'} M_j) = \beta(M_{d'} - m_{r_{d'}} - \sum_{j<d'} M_j)$. Let d' be the last hypercube in the sequence incurring such a delay. The total time to gather all data to the root r is therefore $d\alpha + \beta \sum_{i \in H_d, i \neq r} m_i$ plus the penalty of $\beta(M_{d'} - m_{r_{d'}} - \sum_{j<d'} M_j)$. □

The resulting construction is easy to implement, and better than first gathering to the linear time root determined by Lemma 2.1 and then sending to the externally given root r which would incur an extra communication round and sending the complete data $\sum_{i \in H_d} m_i$, effectively loosing half of the communication bandwidth (although still being linear). It is important to notice that we assume that for each hypercube H_d with root r there is no cost associated with the root block m_r. We return to this point in Section 3.

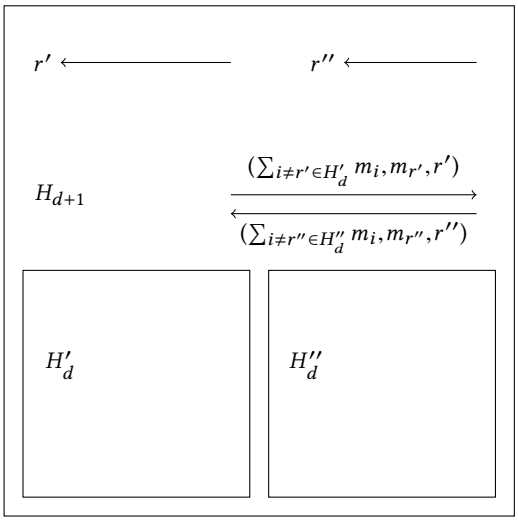

Figure 1: An iteration of the algorithm of Lemma 2.3 showing the communication necessary to join two adjacent, ordered hypercubes H'_d and H''_d into the larger hypercube H_{d+1}. The fixed roots first exchange information on the gather times, the size of the root data blocks, and the identity of the gather roots in the respective subcubes. In the next step, the gather roots r' and r'' receive this information from their fixed roots, so that they can consistently determine which will be the gather root for H_{d+1}.

The communication structure of an ordered hypercube gather algorithm is a binomial tree with a particular numbering of the tree roots determined by the p data block sizes $m_i, 0 \leq i < p$. An example is shown in Figure 2. This tree can be constructed efficiently as shown by the next lemma which is illustrated in Figure 1.

Lemma 2.3. *For any H_d, the gather communication tree can be constructed in d communication rounds, each comprising at most two send and receive operations.*

Proof. The communication tree is constructed iteratively, maintaining the following invariant. Each H_d has a predetermined, fixed root that can readily be computed by any processor, and a gather root r which will gather the data from H_d as per Lemma 2.1. The fixed root and the gather root are not necessarily distinct processors. Both the fixed and the gather root processors know that they have this role and which processor has the other role, and each knows the total amount of data in H_d. When the hypercube H_{d+1} is formed from H'_d and H''_d, the fixed root of H'_d knows which processor is the fixed root of H''_d and vice versa. The gather roots do not know the gather root of the other subcube.

For all H_0 subcubes the invariant holds with fixed and gather root being the sole processor in each H_0. To maintain the invariant for

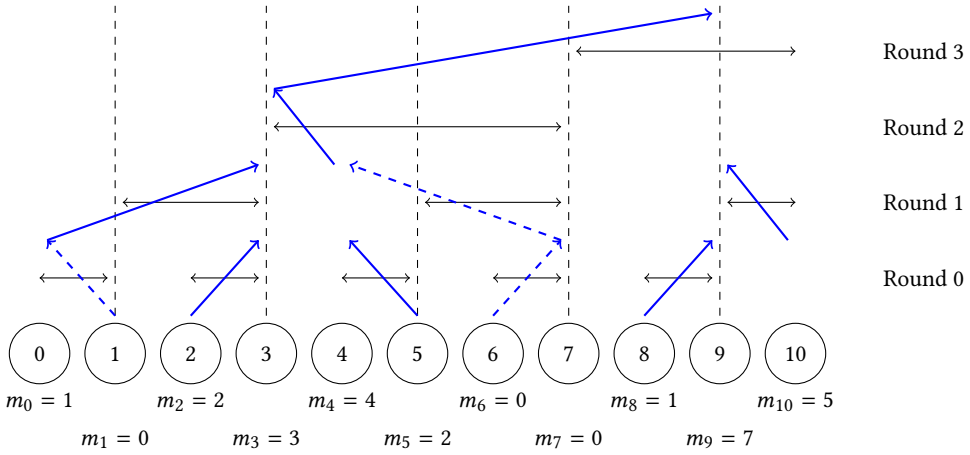

Figure 2: A linear-time, ordered gather tree for $p = 11$ processors and root 9 with the indicated block sizes m_i as constructed by the algorithm of Lemma 2.3. Thick (blue), upward arrows are the gather tree edges with dotted arrows indicating data sizes of zero with no actual communication. Thin (black), horizontal double arrows indicate the exchange between fixed roots as needed to construct the ordered gather tree.

H_{d+1}, the two fixed roots of the H_d subcubes exchange information on their estimated gather time, the size of the root data blocks, and the identity of the gather root processors. Both fixed roots can now determine which gather root will be the gather root of H_{d+1}, namely the gather root of the subcube with the largest gather time estimate $\sum_{i \in H_d, i \neq r} m_i$ (with ties broken arbitrarily, but consistently). The first time a fixed root of some H_d by the exchange determines that it will become a gather root of H_{d+1}, this new gather root knows that it is a gather root. To maintain the invariant for the following iterations, if the gather root of H_d does not know whether it will be the gather root of H_{d+1}, it receives information on the gather root in H_{d+1} from its fixed root in H_d which per invariant knows the identity of the gather root in H_d. By exchanging both the gather times $\sum_{i \in H_d, i \neq r} m_i$ and the sizes of the root data blocks m_r, gather roots can compute the amount of data to be received in each communication round.

The construction takes $\lceil \log_2 p \rceil$ iterations in each of which pairwise exchanges between the fixed roots of adjacent hypercubes take place. After this, at most one transmission between each fixed root and its corresponding gather root is necessary, except for the first communication round where no such transmission is needed. Thus at most $2\lceil \log_2 p \rceil - 1$ dependent communication operations are required. All information exchanged is of constant size, consisting of the gather time estimate, the size of the root data block, and the identity of the gather root. □

As fixed root for a subcube H_d consisting of processors $[a2^d, \ldots, (a+1)2^d - 1]$ we can choose, e.g., the last processor $i = (a+1)2^d - 1$. For the fixed root of this H_d to find the fixed root of its adjacent subcube, the d'th bit of i has to be flipped. If the d'th bit is a 1, the processor will survive as the fixed root of H_{d+1}. If the number of processors p is not a power of two, only the last processor $p-1$ has to be specially treated. If a fixed root in iteration d, by flipping bit d, determines that its partner fixed root is larger than $p-1$ it instead chooses $p-1$ as fixed root in its adjacent, incomplete hypercube. In this iteration, processor $p-1$ must be prepared to act as fixed root. In iteration d, if processor $p-1$ has bit d set, it knows that some lower numbered fixed root has chosen $p-1$ as fixed root, and this adjacent fixed root is $(\lceil p/2^d \rceil - 1)2^d - 1$. Otherwise, if bit d is not set, processor $p-1$ has no role in iteration d.

Together, these remarks and the three lemmas give the main result.

Theorem 2.4. *For any number of processors p in a homogeneous, fully connected, linear time cost communication network, the irregular gather problem with root r and block size m_i for processor $i, 0 \leq i < p$ can be solved in time at most $3\lceil \log_2 p \rceil \alpha + \beta \sum_{0 \leq i < p, i \neq r} m_i$ with an additive time penalty of at most $\beta(M_{d'} - m_{r'_d} - \sum_{0 \leq j < d'} M_j)$ for some $d' < \lceil \log_2 p \rceil$, each M_j being the total amount of data in a tree of 2^j distinct processors with local root r_j.*

The linear time gather trees can likewise be used for the irregular scatter problem. Also, both tree construction and communication algorithms can be extended to k-ported communication systems, which reduces the number of communication rounds needed from $\lceil \log_2 p \rceil$ to $\lceil \log_{k+1} p \rceil$. It is perhaps worth pointing out that the constructions also provide ordered communication trees for the regular gather and scatter (as well as for reduction-to-root) operations with the optimal $\lceil \log_2 p \rceil$ number of communication rounds. If all processors know the common root r, tree construction can be done without any actual, extra communication.

3 MPI IMPLEMENTATIONS

We have implemented our irregular gather and scatter algorithms in MPI (available at www.par.tuwien.ac.at/Downloads/TUWMPI/tuwgatherv.c) with the same interface as the corresponding MPI operations, and can thus readily compare our `TUW_Gatherv` and `TUW_Scatterv` to other MPI implementations. We use the algorithm of Lemma 2.3 to construct communication trees which we (for gather) represent as a sequence of receive operations followed

by a send operation at each process. Where appropriate, we use non-blocking sends and receives to better absorb delays by some MPI processes finishing late. For non-root processes, intermediate buffers gather (scatter) data from (to) the processes' children. Since the sizes of all received (sent) data are known by construction, and since all children send (receive) rank ordered data blocks, it is easy to keep blocks stored in intermediate buffers in rank order. Our implementations copy each subtree root's own block to/from the intermediate buffer, incurring a cost proportional to m_r not accounted for in Lemma 2.1. This explicit copy could be avoided by using an indexed derived datatype to handle communication of the root block. Alternatively, Lemma 2.1 can be modified to account for an explicit copy of the root blocks. All processes in the `MPI_Gatherv` and `MPI_Scatterv` operations supply an MPI datatype describing the types and structure of their blocks. Since all data blocks must eventually match the datatype supplied by the root process, it is possible to receive and send all intermediate data blocks with a correct MPI derived datatype. To this end, the signature datatype described in [16] can be used. Since blocks can be described by different types with different counts by different processes, it is important that the "smallest common block" of the signature type is used for communication. At the root, data blocks are to be stored as described by the list of displacements and block sizes supplied in the root process' `MPI_Gatherv` or `MPI_Scatterv` call. This can be accomplished by constructing a corresponding indexed derived datatype for each of the children describing where the data blocks go. No explicit, intermediate buffering at the root is therefore necessary, and in that sense zero-copy implementation of the algorithms are possible. If the root displacements describe a contiguous segment of blocks in rank order (as may be the case in applications), no such datatype is necessary, and the blocks can be received directly into their correct positions in the root receive buffer. Our prototype implementation works under this assumption.

Despite the linear gather time guaranteed by the algorithm, sending large data blocks multiple times through the gather tree incurs unnecessary, repeated transmission costs. Practical performance might be better if such large blocks would be sent directly to the root process. It would be possible to implement graceful degradation behavior [14] by introducing a gather subtree threshold beyond which a subtree in the gather tree shall send its data directly to the root. This is not entirely trivial, and we have not yet implemented this potential improvement.

4 A PADDING PERFORMANCE GUIDELINE FOR IRREGULAR COLLECTIVES

Self-consistent MPI performance guidelines formalize expectations on the performance of given MPI operations by relating them to the performance of other MPI operations implementing the same functionality [18]. If a performance guideline is violated, it gives a constructive hint to the application programmer and the MPI library implementer how the given operation can be improved in the given context. Performance guidelines thus provide sanity checks for MPI library implementations, and can be helpful in structuring experiments [4, 10].

In order to use regular collectives correctly, the application programmer must know that all processes supply the same data sizes and each process must know this data size. The irregular collectives have a weaker precondition: It suffices that each process by itself knows its data size with the only requirement that processes that pairwise exchange data must know and supply the same sizes. If an irregular collective is used in a situation where a regular one could have been used instead, we would expect the regular collective to perform better, or at least not worse in that situation. This is captured in the performance guideline for `MPI_Gatherv` below.

$$\texttt{MPI\_Gather}(m) \quad \leq \quad \texttt{MPI\_Gatherv}(m) \tag{1}$$

Here m is the total amount of data to be gathered at the root process, and the guideline states that in a situation where `MPI_Gather` can be used ($m_i = m/p$), this should perform at least as well as using instead `MPI_Gatherv`, all other things (e.g., root process, communicator) being equal in the two sides of the equation. If the guideline is violated, which can be tested experimentally, there is something wrong with the MPI library, and the user would do better by using `MPI_Gatherv` instead of `MPI_Gather`. There are reasons to expect that the guideline is not violated. The `MPI_Gather` operation is more specific, does not take long argument lists of counts and displacements, and good, tree-based algorithms exist and may have been implemented for this operation.

A common way of dealing with slightly irregular problems is to transform them into regular ones by padding all buffers up to some common size and solving the problem by a corresponding regular collective operation. The argument for having the specialized, irregular collectives in the MPI specification is that a library can possibly do better than (or at least as good as) this manual solution. Thus, we would like to expect that `MPI_Gatherv` performs no worse than first agreeing on the common buffer size and then doing the regular collective on this, possibly larger common size. This is expressed in the second irregular performance guideline.

$$\begin{aligned}\texttt{MPI\_Gatherv}(m) \quad \leq \quad & \texttt{MPI\_Allreduce}(1) + \\ & \texttt{MPI\_Gather}(m')\end{aligned} \tag{2}$$

where $m' = p \max_{0 \leq i < p} m_i$ is the total amount of data to be gathered by the regular `MPI_Gather` as computed by the `MPI_Allreduce` operation. Again, if experiments show this guideline violated, there is an immediate hint for the application programmer on how to do better: Use padding. Here we assume that the application programmer can organize the padded buffers such that no copying back and forth between buffers is necessary; this may not always be possible, so the guideline should not be interpreted too strictly but allow some extra slack on the right-hand side upper bound. Nevertheless, it constrains what should be expected by a good implementation of `MPI_Gatherv`.

The second guideline is particularly interesting for the regular case where $m_i = m/p$. Here it says that the overhead for `MPI_Gatherv` compared to `MPI_Gather` should not be more than a single, constant-sized `MPI_Allreduce` operation. This may be difficult for MPI libraries to satisfy, but indeed, if it is not, the usefulness of `MPI_Gatherv` may be questionable.

The two performance guidelines give less trivial performance expectations against which to test our new algorithms instead of only comparing to the `MPI_Gatherv` and `MPI_Gather` implementations in some given MPI library.

Table 1: Results for NEC MPI, $p = 35 \times 16 = 560$. Running times are in microseconds (μs). Numbers in red, bold font show violations of the performance guidelines, Guideline (1) in the `MPI_Gather` column, Guideline (2) in column `MPI_Gatherv`

Problem	m	m'	`MPI_Gather`		Guideline (2)		`MPI_Gatherv`		`TUW_Gatherv`	
			avg	min	avg	min	avg	min	avg	min
Same	560	560	337.83	18.12	59.43	36.95	1755.53	**1194.95**	50.20	20.98
	5600	5600	49.69	33.86	76.92	57.94	1753.57	**1235.96**	45.06	37.91
	56000	56000	183.12	169.99	207.12	186.92	1810.53	**1347.06**	138.28	119.92
	560000	560000	1307.84	1293.18	1336.58	1312.02	2644.22	**2297.16**	824.08	802.04
	5600000	5600000	16874.53	**9708.17**	9793.19	9739.16	9243.09	7848.02	7972.50	7951.97
Random	844	1120	31.67	20.98	55.82	37.91	1707.73	**1415.01**	62.38	28.85
	5989	11200	68.23	57.94	93.61	77.01	1720.36	**948.91**	88.62	42.92
	57615	112000	302.17	290.16	324.11	305.18	1851.97	**1305.82**	185.11	144.96
	571327	1119440	2243.32	2229.93	2258.11	2242.09	2702.19	**2511.02**	1054.04	827.07
	5546939	11189360	21490.23	18246.89	31686.29	18262.15	7940.00	7556.92	15855.80	9074.93
Spikes	1036	2800	38.74	29.09	59.58	43.87	5262.01	**1353.03**	58.94	28.85
	6391	28000	111.11	98.94	134.16	118.97	1726.01	**1243.11**	76.38	40.05
	57945	280000	684.50	667.10	703.41	684.98	1759.52	**1068.12**	163.28	113.96
	570446	2800000	4839.81	4822.02	4874.30	4847.05	2029.88	1711.85	1088.44	1050.00
	6100438	28000000	46745.77	44775.96	50961.33	44814.11	8909.78	8166.07	10682.46	8350.13
Decreasing	842	1680	33.10	25.99	53.16	41.96	1690.14	**845.91**	40.90	25.03
	5900	11760	70.28	61.99	92.52	79.87	1731.86	**952.96**	94.30	61.04
	56400	112560	315.14	299.93	341.26	318.05	1986.37	**1213.07**	212.21	174.05
	561320	1120560	2244.04	2228.98	2267.95	2247.10	2900.91	**2758.03**	1223.44	1196.86
	5610320	11200560	22649.90	18422.13	18473.74	18440.96	7727.55	7624.86	12091.68	11695.86
Alternating	560	560	5793.04	20.03	48.54	34.81	4837.82	**1188.04**	32.95	20.98
	5600	8400	58.68	49.83	86.03	67.00	1720.53	**1029.01**	64.41	41.01
	56000	84000	244.91	231.03	269.17	246.05	1860.01	**1120.09**	153.70	139.00
	560000	840000	1649.08	1631.98	1666.12	1648.90	11386.20	**2473.12**	980.67	967.03
	5600000	8400000	29599.52	13819.93	22029.07	13828.04	7947.92	7896.18	14881.42	9907.96
Two blocks	2	560	28.39	18.12	50.70	36.95	5.74	1.91	28.09	18.12
	20	5600	49.09	41.01	72.24	58.89	6.89	2.86	28.25	18.12
	200	56000	185.02	171.90	209.54	189.07	7.31	2.86	29.83	19.07
	2000	560000	1306.06	1291.04	1331.58	1313.92	12.96	8.11	33.93	23.13
	20000	5600000	16682.08	9497.17	9559.33	9510.04	39.24	36.00	63.39	51.02

5 EXPERIMENTS

Finally, we present a preliminary evaluation of the `TUW_Gatherv` implementation. We evaluate by comparing to `MPI_Gather` and `MPI_Gatherv` of common MPI libraries guided by the performance guidelines explained in Section 4.

We test our algorithm on gather problems of varying degrees of irregularity. Let $b, b > 0$ be a chosen, average block size (in some unit, here `MPI_INT`). We have p MPI processes, and use as fixed gather root $r = \lfloor p/2 \rfloor$. Our problems are as follows with names indicating how block sizes are chosen for the processes.

Same: For process i, $m_i = b$.

Random: Each m_i is chosen uniformly at random in the range $[1, 2b]$.

Spikes: Each m_i is either $\rho b, \rho > 1$ or 1, chosen randomly with probability $1/\rho$ for each process i.

Decreasing: For process i, $m_i = \lfloor \frac{2b(p-i)}{p} \rfloor + 1$

Alternating: For even numbered processes, $m_i = b + \lfloor b/2 \rfloor$, for odd numbered processes $m_i = b - \lfloor b/2 \rfloor$.

Two blocks: All $m_i = 0$, except $m_0 = b$ and $m_{p-1} = b$.

These problem types, except for the last, specifically always have $m_i > 0$. This choice ensures that an implementation cannot take advantage of not having to send empty blocks. We perform a series of (weak scaling) experiments with $b = 1, 10, \ldots, 10\,0000$; the total problem size in each case is $m = \sum_{0 \leq i < p} m_i$ and increasing linearly with p (except for the two blocks problems). For comparison with `MPI_Gather` and for the padding performance Guideline (2), the padded block size is $\max_{0 \leq i < p} m_i$ and the total size $m' = p \max_{0 \leq i < p} m_i$. For the **spikes** problems, we have taken $\rho = 5$.

In our experiments we perform 75 time measurements of each of the collective operations with 10 initial, not timed, warmup calls, and compute average and minimum times (the fastest completion time seen over the 75 repetitions). For the average times we have not done any outlier removal. Before each measurement, MPI processes are synchronized with the native `MPI_Barrier` operation, and the running time of a measurement is the time of the slowest process, which for the gather operations is usually the root process.

Table 2: Results for MVAPICH, $p = 35 \times 16 = 560$. Running times are in microseconds (μs). Numbers in red, bold font show violations of the performance guidelines, Guideline (1) in the `MPI_Gather` column, Guideline (2) in column `MPI_Gatherv`

Problem	m	m'	`MPI_Gather`		Guideline (2)		`MPI_Gatherv`		`TUW_Gatherv`	
			avg	min	avg	min	avg	min	avg	min
Same	560	560	56.69	25.03	295.59	77.01	1083.19	**887.87**	53.32	38.15
	5600	5600	250.77	84.88	284.61	114.92	1229.84	**1013.99**	77.34	48.88
	56000	56000	317.90	302.08	366.31	323.06	1403.33	**1122.95**	193.72	155.93
	560000	560000	3736.72	**3597.02**	3852.69	3761.05	3760.52	3576.99	1149.36	1132.01
	5600000	5600000	31109.11	**30703.07**	30893.50	30778.17	7758.92	7701.16	15844.17	15782.12
Random	844	1120	84.15	76.06	105.57	93.94	2357.83	**1780.03**	60.69	49.83
	5989	11200	141.02	136.14	161.29	150.92	2413.02	**1760.01**	71.50	56.98
	57615	112000	573.50	535.01	608.74	573.87	2597.00	**1867.06**	168.84	157.12
	571327	1119440	5006.39	4858.97	7115.21	5001.07	3812.06	3521.92	1856.02	849.96
	5546939	11189360	54121.13	53570.03	54998.99	52770.14	8087.99	7569.07	13339.92	13207.91
Spikes	1036	2800	98.12	81.06	117.19	102.04	2353.78	**1659.87**	57.43	47.92
	6391	28000	204.39	198.84	226.27	215.05	2351.88	**1694.92**	70.14	56.98
	57945	280000	1200.79	1181.13	1225.05	1204.01	2432.12	**1807.93**	157.08	144.96
	570446	2800000	5283.16	4540.92	5573.90	5078.08	2965.85	2594.95	1153.44	1137.97
	6100438	28000000	130602.19	129426.96	131404.61	129746.20	8642.77	8414.03	13026.98	12981.18
Decreasing	842	1680	85.32	72.96	110.13	96.80	2349.97	**1697.78**	53.90	44.82
	5900	11760	144.53	133.99	165.36	154.97	2398.28	**1858.00**	86.67	75.10
	56400	112560	580.63	550.03	607.42	577.93	2753.80	**1947.16**	235.09	220.06
	561320	1120560	5091.47	4953.86	5138.67	5054.95	4221.21	3954.17	4238.99	2449.99
	5610320	11200560	54855.09	54377.08	55570.82	54203.99	7769.08	7728.82	19093.30	19005.78
Alternating	560	560	84.05	57.94	107.30	91.79	2298.24	**1638.17**	52.58	42.92
	5600	8400	127.73	115.87	152.87	139.95	2327.04	**1632.93**	69.52	56.98
	56000	84000	441.42	401.02	469.97	445.84	2538.67	**2073.05**	172.13	157.12
	560000	840000	4525.85	4410.98	4591.90	4480.84	3777.33	3542.90	1132.91	1110.79
	5600000	8400000	52944.41	46571.02	46649.89	46545.03	11048.42	7767.92	15145.07	15044.93
Two blocks	2	560	81.62	73.19	279.81	92.03	7.70	1.91	40.42	30.99
	20	5600	613.77	108.00	197.26	126.12	7.39	1.91	40.66	30.99
	200	56000	787.44	301.84	477.54	332.83	8.76	1.91	82.04	30.99
	2000	560000	3760.01	3598.93	3855.75	3764.87	11.48	2.86	45.08	33.86
	20000	5600000	32092.41	31884.91	31959.97	31888.96	41.83	36.00	74.30	62.94

Our first test system is a small InfiniBand cluster with 36 nodes each consisting of two 8-core AMD Opteron 6134 processors running at 2.3GHz. The interconnect is a QDR InfiniBand MT26428. We have tried the implementations with three different MPI libraries, namely NEC MPI-1.3.1, MVAPICH2-2.2 and OpenMPI-2.0.1 using the gcc 4.9.2 compiler with -O3 optimization. We present the results in tabular form, see Table 1, Table 2 and Table 3. Running times are in microseconds (μs).

Average and minimum, best observed time differ considerably (which may be due to outliers), and comparison based on averages may not be well-founded. Nevertheless, the results show the three library implementations of the `MPI_Gather` and `MPI_Gatherv` operations to (surprisingly) differ considerably in quality. For `MPI_Gather`, this can best be seen for the **same** problem type, where the NEC MPI minimum time is about 9000 μs and MVAPICH at 31000 μs with MVAPICH at 13000 μs for the largest problem instance. Also for the smaller problem sizes, the differences can be considerable. For all three libraries, it is also clear that `MPI_Gatherv` is implemented with a trivial algorithm compared to `MPI_Gather`; this makes `MPI_Gatherv` an expensive operation for small problem sizes. On the other hand, the algorithms used for `MPI_Gather` are not well chosen for large problem instances, where for all three libraries, the simple, direct to root implementations used for `MPI_Gatherv` perform better. The trivial performance Guideline (1) is violated in such cases.

For the smaller instances of the irregular problem types, all libraries fail Guideline (2) with their `MPI_Gatherv` implementations by large factors, whereas `TUW_Gatherv`, except for the **two blocks** problems, easily fulfill the padding guideline, often by a considerable factor; the new `TUW_Gatherv` implementation is faster than the library implementations often by factors of 5 to more than 20. There are even cases where `TUW_Gatherv` is faster than the library `MPI_Gather` implementations (seen for the **same** block size problem type).

Our second system is a medium-large InfiniBand/Intel cluster consisting of 2000 Dual Intel Xeon E5-2650v2 8-core processors running at 2.6GHz, interconnected with an InfiniBand QDR-80

Table 3: Results for Open MPI, $p = 35 \times 16 = 560$. Running times are in microseconds (μs). Numbers in red, bold font show violations of the performance guidelines, Guideline (1) in the `MPI_Gather` column, Guideline (2) in column `MPI_Gatherv`

Problem	m	m'	`MPI_Gather`		Guideline (2)		`MPI_Gatherv`		`TUW_Gatherv`	
			avg	min	avg	min	avg	min	avg	min
Same	560	560	173.80	45.00	239.85	123.00	967.65	**170.00**	219.27	67.00
	5600	5600	156.08	74.00	242.00	153.00	969.65	**176.00**	124.20	79.00
	56000	56000	926.56	**764.00**	1065.35	872.00	430.83	191.00	749.55	535.00
	560000	560000	3619.15	**3378.00**	3758.16	3463.00	1788.52	1006.00	2390.60	1899.00
	5600000	5600000	14500.19	13844.00	14405.19	14053.00	14210.47	13993.00	16043.45	**15939.00**
Random	843	1120	106.56	61.00	174.33	125.00	1139.69	115.00	121.85	70.00
	5892	11200	155.07	90.00	207.55	160.00	1073.25	134.00	132.68	89.00
	57435	112000	1119.33	1071.00	1238.20	1184.00	485.49	193.00	560.53	509.00
	542098	1116640	7284.08	7059.00	7526.65	7228.00	2214.36	1581.00	1890.25	1842.00
	5687094	11173120	20778.19	19741.00	20860.49	20379.00	14068.03	13896.00	12936.69	12858.00
Spikes	984	2800	108.67	62.00	187.49	137.00	1125.93	**144.00**	125.40	70.00
	6146	28000	209.33	127.00	256.87	196.00	1056.89	**214.00**	132.23	77.00
	54951	280000	1885.92	1829.00	2004.80	1932.00	1142.00	285.00	594.28	444.00
	525455	2800000	10698.63	10047.00	10799.77	10433.00	3012.12	2909.00	1764.33	1707.00
	5000460	28000000	41352.32	40729.00	42246.07	41248.00	10434.21	10021.00	10070.91	9774.00
Decreasing	842	1680	106.77	56.00	181.56	134.00	1107.64	126.00	123.28	69.00
	5900	11760	143.09	96.00	421.40	161.00	1258.08	**518.00**	162.55	106.00
	56400	112560	1106.97	1058.00	1233.16	1158.00	451.28	195.00	638.89	589.00
	561320	1120560	7348.33	7125.00	7994.61	7220.00	2585.16	2077.00	2586.47	2427.00
	5610320	11200560	20940.69	20003.00	20914.79	20360.00	13757.08	13538.00	26159.17	21454.00
Alternating	560	560	137.84	58.00	245.55	124.00	1209.64	**136.00**	120.92	63.00
	5600	8400	142.93	94.00	201.03	155.00	1204.63	**282.00**	115.28	71.00
	56000	84000	935.68	886.00	1028.65	987.00	329.56	178.00	539.87	476.00
	560000	840000	7471.88	7397.00	7592.03	7491.00	2289.97	1825.00	2048.68	1718.00
	5600000	8400000	18110.28	17509.00	18244.43	17995.00	13781.65	13633.00	14882.91	14797.00
Two blocks	2	560	99.95	53.00	175.83	127.00	16.81	2.00	109.59	59.00
	20	5600	120.04	62.00	189.44	148.00	24.79	2.00	115.88	65.00
	200	56000	1195.01	1144.00	1310.89	1246.00	24.93	2.00	112.56	56.00
	2000	560000	3583.89	3179.00	3566.36	3305.00	16.37	4.00	188.45	71.00
	20000	5600000	14093.35	13816.00	14272.21	14026.00	98.35	56.00	155.19	117.00

network[2]. The MPI library is Intel MPI 2017.1 and the compiler is Intel MPI 2017.1 with optimization level -O3. The benchmark was executed first with the default environment for this machine, which pins the MPI processes to the 16 cores per node; also the choice of `MPI_Gather` and `MPI_Gatherv` implementations was left to the environment. Results can be found in Table 4 for $p = 6400$, respectively. The most conspicuous observation about this MPI library is the poor quality of both `MPI_Gather` and `MPI_Gatherv`, rendering our `TUW_Gatherv` implementation several orders of magnitude faster for small problems. The `TUW_Gatherv` implementation therefore satisfies Guideline (2) by a large margin, and `MPI_Gather` fails the trivial Guideline (1) compared to `TUW_Gatherv` by a very large factor.

One reason for the poor performance on the larger cluster is that the `MPI_Gather` and `MPI_Gatherv` implementations used by default are ill chosen. The Intel MPI 2017.1 library indeed contains different algorithms and implementations that can be chosen by environment variables (`I_MPI_ADJUST`); it can also be controlled for which message and process ranges particular implementations shall be used. For `MPI_Gather`, the library lists four different algorithms (no specific references are given), namely 1) binomial, 2) topology aware binomial, 3) a socalled Shumilin's algorithm, and 4) binomial with segmentation. For `MPI_Gatherv`, three choices are possible, namely 1) linear, 2) topology aware linear, and 3) k-nomial (with radix $k = 2$). For completeness, we ran the benchmark with all these explicit choices as well, each for the full problem size range. These results can be found in the report [17] accompanying this paper. The best, and most promising implementation choices are 1) binomial for `MPI_Gather`, and 3) k-nomial for `MPI_Gatherv`. The results are shown together in Table 5, and performance guidelines violations are marked in red as in the previous tables. The results show that while a binomial tree algorithm can work well also for `MPI_Gatherv` for almost regular problems, there is a penalty as the problems get more irregular, which the `TUW_Gatherv` implementation pays to a much lesser extent, still often being faster than the native `MPI_Gather` implementation. The `TUW_Gatherv` implementation in all cases outperforms the k-nomial `MPI_Gatherv` for the Intel MPI

[2]This is the so-called Vienna Scientific Cluster, see vsc.ac.at. The author thanks for access to this machine.

Table 4: Results for Intel MPI, default algorithms settings, $p = 400 \times 16 = 6400$. Running times are in microseconds (μs). Numbers in red, bold font show violations of the performance guidelines, Guideline (1) in the `MPI_Gather` column, Guideline (2) in column `MPI_Gatherv`

Problem	m	m'	`MPI_Gather`		Guideline (2)		`MPI_Gatherv`		`TUW_Gatherv`	
			avg	min	avg	min	avg	min	avg	min
Same	6400	6400	143224.73	103827.95	134339.83	116875.89	198096.70	**147197.96**	483.80	50.07
	64000	64000	127234.13	96213.10	119072.93	96263.17	157288.30	**127314.09**	1119.38	158.07
	640000	640000	135538.09	92452.05	115266.33	98417.04	184500.61	**150245.90**	1208.13	964.16
	6400000	6400000	174751.44	111608.03	168121.69	127385.85	229787.77	**162585.97**	8883.16	8399.01
	64000000	64000000	478717.77	351228.00	448441.26	345794.92	521594.60	400845.05	82694.19	81154.11
Random	9606	12800	112400.91	83522.08	119423.66	97676.99	165242.37	**143667.94**	1149.91	80.11
	67404	128000	113700.98	92371.94	114324.38	100589.04	159724.97	**135301.83**	746.22	169.04
	640316	1280000	130619.07	97826.96	117179.47	100124.84	196255.20	**154004.10**	1431.84	959.16
	6365598	12800000	266579.81	133107.19	256535.67	194566.01	240661.89	174778.94	9395.70	8400.92
	64366469	128000000	383281.22	240486.86	382698.89	234140.87	568810.39	449437.14	85171.69	84529.88
Spikes	11576	32000	115877.39	83566.90	107814.19	95278.98	155785.61	**128391.03**	1546.82	67.00
	68483	320000	112930.08	88580.85	106333.22	95693.83	158825.65	**129721.16**	912.51	162.12
	641627	3200000	186607.47	99061.97	179326.41	102573.87	170716.86	**143610.00**	1132.60	957.97
	6465108	32000000	363413.76	282375.81	352513.17	273236.04	228685.47	170839.07	9048.14	8200.88
	64505110	320000000	592158.24	477687.12	589359.86	469671.96	271884.27	195724.96	139265.59	138741.97
Decreasing	9602	19200	112524.48	86571.93	114034.30	97607.14	166033.34	**143018.01**	372.60	56.98
	67220	134400	113499.48	88942.05	111749.48	99705.93	169566.57	**137952.09**	297.82	197.89
	643400	1286400	130733.47	100178.96	122006.55	96143.96	204257.36	**170412.06**	1309.55	1060.96
	6404400	12806400	256993.15	163550.14	249910.52	174806.12	236896.72	**177741.05**	9904.83	9175.06
	64013600	128006400	358370.11	231245.99	381372.81	230093.00	531949.64	438292.98	87494.51	86869.00
Alternating	6400	6400	103188.01	73747.87	103039.74	90380.19	155132.28	**132302.05**	200.55	47.92
	64000	96000	108217.47	81782.10	104217.22	80809.83	144754.90	**114871.03**	285.87	151.87
	640000	960000	128934.07	87894.20	108839.05	91517.93	189208.25	**128816.84**	1136.50	955.10
	6400000	9600000	209340.04	124116.18	210117.61	134249.93	251695.84	**152398.11**	9099.22	8328.91
	64000000	96000000	515877.56	367428.06	516283.90	398251.06	605431.46	**453334.81**	81952.00	81115.96
Two blocks	2	6400	102869.81	77906.85	99906.61	88593.96	17.60	15.97	365.01	30.99
	20	64000	105361.90	76925.99	104937.77	83439.11	17.81	15.97	56.81	30.04
	200	640000	107798.00	88968.99	105016.54	91776.85	17.78	16.93	146.50	29.80
	2000	6400000	184417.88	108457.09	179452.45	118996.14	21.39	16.93	165.49	36.00
	20000	64000000	435541.00	252403.02	416582.60	335694.07	38.32	30.04	83.75	63.90

2017.1 library by a factor of two to three (noteworthy also for the "Two blocks" distribution where a linear algorithm performs best). The violations of performance Guideline (2) by `MPI_Gatherv` are surprising, and likely due to overhead in determining block sizes to be used in the k-nomial tree. The `TUW_Gatherv` implementation has no guideline violations, even when compared to the good Intel MPI 2017.1 library implementations.

6 CONCLUSION

This paper described new, simple algorithms for performing irregular gather and scatter operations as found in MPI in linear communication time, a considerable improvement over both fixed, data oblivious logarithmic depth trees and direct communication with the root. An experimental evaluation shows that the resulting implementation can, especially for overall small problem instances be considerably faster than current MPI library `MPI_Gatherv` implementations by large factors. Our prototype implementations can readily be incorporated into existing MPI libraries. The algorithms were derived under the assumption of homogeneous, linear communication costs, which has often been pointed out to be inadequate for, e.g. hierarchically structured systems [8]. It would be possible to use the algorithms in a hierarchical fashion and still have overall linear-time performance.

The tree construction technique of Lemma 2.3 can be applied to other problems as well, for instance to construct good, problem dependent trees for sparse reduction operations [15].

REFERENCES

[1] Zina Ben-Miled, José A. B. Fortes, Rudolf Eigenmann, and Valerie E. Taylor. On the implementation of broadcast, scatter and gather in a heterogeneous architecture. In *Thirty-First Annual Hawaii International Conference on System Sciences (HICSS)*, pages 216–225, 1998.

[2] Sandeep N. Bhatt, Geppino Pucci, Abhiram Ranade, and Arnold L. Rosenberg. Scattering and gathering messages in networks of processors. *IEEE Transactions on Computers*, 42(8):938–949, 1993.

[3] Laurence Boxer and Russ Miller. Coarse grained gather and scatter operations with applications. *Journal of Parallel and Distributed Computing*, 64(11):1297–1310, 2004.

[4] Alexandra Carpen-Amarie, Sascha Hunold, and Jesper Larsson Träff. On expected and observed communication performance with MPI derived datatypes.

Table 5: Results for Intel MPI, $p = 400 \times 16 = 6400$, overall best settings with binomial (1) for `MPI_Gather` and k-nomial (3) for `MPI_Gatherv`. Running times are in microseconds (μs). Numbers in red, bold font show violations of the performance guidelines, Guideline (1) in the `MPI_Gather` column, Guideline (2) in column `MPI_Gatherv`

Problem	m	m'	`MPI_Gather`		Guideline (2)		`MPI_Gatherv`		`TUW_Gatherv`	
			avg	min	avg	min	avg	min	avg	min
Same	6400	6400	286.35	35.05	317.71	64.13	995.37	**123.98**	680.45	53.17
	64000	64000	299.46	182.15	517.45	246.05	902.68	**259.16**	819.50	157.83
	640000	640000	1259.86	1109.84	1247.51	1140.12	2281.50	**1932.86**	1619.93	962.02
	6400000	6400000	13873.83	13394.12	14012.66	13395.07	19681.53	**19338.13**	9631.22	8454.08
	64000000	64000000	147270.64	144801.14	147243.89	144757.03	193182.23	191246.99	83206.43	81207.99
Random	9648	12800	171.47	50.07	257.15	80.11	754.59	**131.13**	750.79	61.04
	67611	128000	438.32	299.93	451.96	334.02	1150.92	330.92	463.82	166.18
	644099	1280000	2375.90	2069.95	2211.75	2100.94	2226.74	1948.12	1330.13	972.99
	6348958	12800000	27995.71	27446.99	28207.40	27549.98	20149.21	19526.00	10030.93	9416.10
	64822758	127993600	295035.76	290700.91	294802.43	290601.02	194914.27	193742.99	87632.58	86949.83
Spikes	11668	32000	284.55	77.96	214.27	108.00	448.54	**144.96**	909.52	62.94
	70149	320000	821.23	600.10	897.93	660.18	626.34	305.18	580.06	161.89
	627655	3200000	5825.37	5667.21	5983.96	5713.94	2524.01	1981.02	1662.80	938.89
	6590083	32000000	70683.11	69520.00	72239.96	69606.07	20666.20	20349.98	9497.62	8433.82
	63605128	320000000	720784.90	710308.07	724077.12	710829.02	195790.83	193390.13	86017.73	84006.07
Decreasing	9602	19200	79.90	45.06	332.72	87.02	1077.95	**123.02**	290.08	55.07
	67220	134400	437.02	314.95	580.22	342.85	899.55	332.83	713.06	201.94
	643400	1286400	2177.98	2104.04	2323.18	2146.96	2821.24	**2244.00**	1633.08	1086.00
	6404400	12806400	27142.07	26774.88	27319.14	26860.00	22353.53	21329.88	9713.58	9071.11
	64013600	128006400	291315.60	286717.89	290913.77	286838.05	205636.62	202884.91	87562.04	86633.92
Alternating	6400	6400	49.76	37.91	212.76	64.85	341.39	**118.97**	806.74	51.02
	64000	96000	252.95	231.98	305.06	262.98	889.07	**271.08**	784.63	154.97
	640000	960000	1693.25	1624.11	2019.42	1688.00	2490.22	**1931.91**	1352.45	967.03
	6400000	9600000	17999.47	17704.01	18028.99	17511.13	20032.64	**19195.08**	8487.01	8227.83
	64000000	96000000	217804.71	214792.01	217947.53	214887.14	193477.21	191688.06	82237.05	80953.12
Two blocks	2	6400	51.97	30.99	82.22	61.99	627.13	**93.94**	673.90	29.80
	20	64000	606.29	180.96	357.92	218.87	532.54	103.95	592.47	39.82
	200	640000	1170.90	1126.05	1219.57	1170.87	430.17	101.09	471.95	29.80
	2000	6400000	10994.00	10734.80	11114.82	10782.96	433.90	107.05	581.97	38.86
	20000	64000000	141559.13	139204.03	142086.64	139410.97	607.52	234.13	581.15	63.90

Submitted, 2017.

[5] Ernie Chan, Marcel Heimlich, Avi Purkayastha, and Robert A. van de Geijn. Collective communication: theory, practice, and experience. *Concurrency and Computation: Practice and Experience*, 19(13):1749–1783, 2007.

[6] Henri-Pierre Charles and Pierre Fraigniaud. Scheduling a scattering-gathering sequence on hypercubes. *Parallel Processing Letters*, 3:29–42, 1993.

[7] Kiril Dichev, Vladimir Rychkov, and Alexey L. Lastovetsky. Two algorithms of irregular scatter/gather operations for heterogeneous platforms. In *Recent Advances in the Message Passing Interface; 17th European MPI Users' Group Meeting (EuroMPI)*, pages 289–293, 2010.

[8] William Gropp, Luke N. Olson, and Philipp Samfass. Modeling MPI communication performance on SMP nodes: Is it time to retire the ping pong test. In *Proceedings of the 23rd European MPI Users' Group Meeting, EuroMPI*, pages 41–50, 2016.

[9] Jun-ichi Hatta and Susumu Shibusawa. Scheduling algorithms for efficient gather operations in distributed heterogeneous systems. In *Proceedings of the 2000 International Workshop on Parallel Processing (ICPPW)*, pages 173–180, 2000.

[10] Sascha Hunold, Alexandra Carpen-Amarie, Felix Donatus Lübbe, and Jesper Larsson Träff. Automatic verification of self-consistent MPI performance guidelines. In *Euro-Par Parallel Processing*, volume 9833 of *Lecture Notes in Computer Science*, pages 433–446, 2016.

[11] MPI Forum. *MPI: A Message-Passing Interface Standard. Version 3.0*, September 21st 2012. www.mpi-forum.org.

[12] Youcef Saad and Martin H. Schultz. Data communication in parallel architectures. *Parallel Computing*, 11(2):131–150, 1989.

[13] Susumu Shibusawa, Hiroyuki Makino, Shigeki Nimiya, and Jun-ichi Hatta. Scatter and gather operations on an asynchronous communication model. In *Proceedings of the 2000 ACM Symposium on Applied Computing (SAC)*, pages 685–691, 2000.

[14] Jesper Larsson Träff. Hierarchical gather/scatter algorithms with graceful degradation. In *18th International Parallel and Distributed Processing Symposium (IPDPS)*, page 80, 2004.

[15] Jesper Larsson Träff. Transparent neutral element elimination in MPI reduction operations. In *Recent Advances in Message Passing Interface. 17th European MPI Users' Group Meeting*, volume 6305 of *Lecture Notes in Computer Science*, pages 275–284. Springer, 2010.

[16] Jesper Larsson Träff. A library for advanced datatype programming. In *23rd European MPI Users' Group Meeting (EuroMPI)*, pages 98–107. ACM, 2016.

[17] Jesper Larsson Träff. Practical, linear-time, fully distributed algorithms for irregular gather and scatter. arXiv:1702.05967, 2017.

[18] Jesper Larsson Träff, William D. Gropp, and Rajeev Thakur. Self-consistent MPI performance guidelines. *IEEE Transactions on Parallel and Distributed Systems*, 21(5):698–709, 2010.