MapReduce in MPI for Large-scale Graph Algorithms

Steven J. Plimpton and Karen D. Devine
Sandia National Laboratories
Albuquerque, NM
sjplimp@sandia.gov

Keywords: MapReduce, message-passing, MPI, graph algorithms, R-MAT matrices

Abstract

We describe a parallel library written with message-passing (MPI) calls that allows algorithms to be expressed in the MapReduce paradigm. This means the calling program does not need to include explicit parallel code, but instead provides "map" and "reduce" functions that operate independently on elements of a data set distributed across processors. The library performs needed data movement between processors. We describe how typical MapReduce funtionality can be implemented in an MPI context, and also in an out-of-core manner for data sets that do not fit within the aggregate memory of a parallel machine. Our motivation for creating this library was to enable graph algorithms to be written as MapReduce operations, allowing processing of Terabyte-scale data sets. We outline MapReduce versions of several such algorithms: vertex ranking via PageRank, triangle finding, connected component identification, Luby's algorithm for maximally independent sets, and single-source shortest-path calculation. To test the algorithms on arbitrarily large artificial graphs we generate randomized R-MAT matrices in parallel; a MapReduce version of this operation is also described. Performance and scalability results for the various algorithms are presented for varying size graphs on a distributed-memory cluster. For some cases, we compare the results with non-MapReduce algorithms, different machines, and different MapReduce software, namely Hadoop. Our open-source library is written in C++, is callable from C++, C, Fortran, or scripting languages such as Python, and can run on any parallel platform that supports MPI.

1 Introduction

MapReduce is the programming paradigm popularized by Google researchers Dean and Ghemawat [9]. Their motivation was to enable rapid development and deployment of analysis programs to operate on massive data sets residing on Google's large distributed clusters. They introduced a novel way of thinking about certain kinds of large-scale computations as a "map" operation followed by a "reduce" operation. The power of the paradigm is that when cast in this way, a nominally serial algorithm now becomes two highly parallel operations working on data local to each processor, sandwiched around an intermediate data-shuffling operation that requires inter-processor communication. The user need only write serial code for the application-specific map and reduce functions; the parallel data shuffle can be encapsulated in a library since its operation is independent of the application.

The Google implementation of MapReduce is a C++ library with communication between networked machines via remote procedure calls. It allows for fault tolerance when large numbers of machines are used, and can use disks as out-of-core memory to process petabyte-scale data sets. Tens of thousands of MapReduce programs have since been written by Google researchers and are a significant part of the daily compute tasks run by the company [10].

Similarly, the open-source Hadoop implementation of MapReduce [1], has become widely popular in the past few years for parallel analysis of large-scale data sets at Yahoo and other data-centric companies, as well as in university and laboratory research groups, due to its free availability. MapReduce programs in Hadoop are typically written in Java, although it also supports use of stand-alone map and reduce kernels, which can be written as shell scripts or in other languages.

More recently, MapReduce formulations of traditional number-crunching kinds of scientific computational tasks have been described, such as post-processing analysis of simulation data [20], graph algorithmics [8], and linear algebra operations [11]. The paper by Tu et al. [20] was particularly insightful to us, because it described how MapReduce could be implemented on top of the ubiquitous distributed-memory message-passing interface (MPI), and how the intermediate data-shuffle operation is conceptually identical to the familiar MPI_Alltoall operation. Their implementation of MapReduce was within a Python wrapper to simplify the writing of user programs. The paper motivated us to develop our own C++ library built on top of MPI for use in graph analytics, which we initially released as open-source software in mid-2009 [2]. We have since worked to optimize several of the library's underlying algorithms and to enable its operation in out-of-core mode on larger data sets. These algorithmic improvements are described in this paper and are part of the current downloadable version [2].

The MapReduce-MPI (MR-MPI) library described in this paper is a simple, lightweight implementation of basic MapReduce functionality, with the following features and limitations:

- *C++ library using MPI for inter-processor communication:* The user writes a (typically) simple main program which runs on each processor of a parallel machine, making calls to the MR-MPI library. For map and reduce operations, the library calls back to user-provided *map()* and *reduce()* functions. The use of C++ allows precise control over the memory and format of data allocated by each processor during a MapReduce. Library calls for performing a map, reduce, or data shuffle, are synchronous, meaning all the processors participate and finish the operation before proceeding. Similarly, the use of MPI within the library is the traditional mode of MPI_Send and MPI_Recv calls between pairs of processors using large aggregated messages to improve bandwidth performance and reduce latency costs. A recent paper [15] also outlines the MapReduce formalism from an MPI perspective, although they advocate a more asynchronous approach, using one-way communication of small messages.

- *Small, portable:* The entire MR-MPI library is a few thousand lines of standard C++ code. For parallel operation, the program is linked with MPI, a standard message passing library available on all distributed memory machines. For serial operation, a dummy MPI library (provided) can be substituted. As a library, it can be embedded in other codes [3] to enable them to perform MapReduce operations.

- *In-core or out-of-core operation:* Each MapReduce object that a processor defines allocates "pages" of memory, where the page size is determined by the user. Typical MapReduce operations can be performed using a few such pages (per processor). If the data set fits in a single page (per processor), the library performs its operations in-core. If the data set exceeds the page size, processors each write to temporary disk files (on local disks or a parallel file system) as needed and subsequently read from them. This out-of-core operation allows processing of data sets larger than the aggregate memory of all the processors, i.e. up to the available aggregate disk space.

- *Flexible programmability:* An advantage of writing a MapReduce program on top of MPI, is that the user program can invoke MPI calls directly, if desired. For example, one-line calls to MPI_Allreduce are often useful in determining the status of an iterative graph algorithm, as described in Section 4. The library interface also provides a user data pointer as an argument passed to all callback functions, so it is easy for the user program to store "state" on each processor, accessible during the map and reduce operations. For example, various flags can be stored that alter the operation of a map or reduce operation, as can richer data structures that accumulate results.

- *C++, C, and Python interfaces:* The C++ interface to the MR-MPI library allows user programs to instantiate and then invoke methods in one or more MapReduce objects. The C interface allows the library to be called from C or other high-level languages such as Fortran. The C interface also allows the library to be wrapped easily by Python via the Python "ctypes" module. The library can then be called from a Python script, allowing the user to write *map()* and *reduce()* callback functions in Python. If a machine supports running Python in parallel, a parallel MapReduce can also be run in this mode.

- *No fault tolerance:* Current MPI implementations do not enable easy detection of a dead processor or retrieval of the data it was working on. So like most MPI programs, a parallel program calling the MR-MPI library will hang or crash if a processor goes away. Unlike Hadoop, and its HDFS file system which provides data redundancy, the MR-MPI library reads and writes simple, flat files. It can use local per-processor disks, or a parallel file system, if available, but these typically provide no data redundancy.

The remainder of the paper is organized as follows. Sections 2 and 3 describe how in-core and out-of-core MapReduce primitives are formulated as MPI-based operations in the MR-MPI library. Section 4 briefly describes the formulation of several common graph algorithms as MapReduce operations. Section 5 gives performance results for these algorithms running on a parallel cluster for graphs ranging in size from 8 million to 2 billion edges; we highlight the performance and complexity trade-offs of a MapReduce approach versus other more special-purpose algorithms. The latter generally perform better but are harder to implement efficiently on distributed memory machines, due to the required explicit management of parallelism, particularly for large out-of-core data sets. Section 6 summarizes some lessons learned from the implementation and use of our library.

2 MapReduce in MPI

The basic datums stored and operated on by any MapReduce framweork are key/value (KV) pairs. In the MR-MPI library, individual keys or values can be of any data type or length, or combinations of multiple types (one integer, a string of characters, two integers and a double, etc); they are simply treated as byte strings by the library. A KV pair always has a key; a KV's value may be NULL. A related data type is the key/multivalue (KMV) pair, where all values associated with the same key are collected and stored contiguously as a multivalue, which is just a longer byte string with an associated vector of lengths, one integer length per value. In this section, we assume the datums operated on by each processor fit in local memory.

A typical MR-MPI program makes at least three calls to the MR-MPI library, to perform *map()*, *collate()*, and *reduce()* operations. In a *map()*, zero or more key/value pairs are generated by each processor. Often this is done using data read from files, but a *map()* may generate data itself or process existing KV pairs to create new ones. The KV pairs produced are stored locally by each processor; a *map()* thus requires no inter-processor communication. Users call the library with a count of tasks to perform and a pointer to a user function; the MR-MPI *map()* operation invokes the user function multiple times as a callback. Depending on which variant of *map()* is called, the user function may be passed a file name, a chunk of bytes from a large file, a task ID, or a KV pair. Options for assigning map tasks to processors are specified by the user and include assigning consecutive chunks of tasks to processors, striding the tasks across processors, or using a master-slave model that is useful when tasks have widely varying workloads.

The *collate()* operation (or data shuffle in Hadoop) identifies unique keys and collects all the values associated with those keys to create KMV pairs. This is done in two stages, the first of which requires communication, since KV pairs with the same key may be owned by any processor. Each processor hashes each of its keys to determine which processor will "own" it. The k-byte length key is hashed into a 32-bit value whose remainder modulo P processors is the owning processor rank. Alternatively, the user can provide a hash function which converts the key into a processor rank. Each processor then sends each of its KV pairs to the owning processor.

After receiving new KV pairs, the second stage is an on-processor computation, requiring no further communication. Each processor reorganizes its KV pairs into KMV pairs, one for each unique key it owns. This is done using a hash table, rather than a sort. A sort scales as $N \log_2(N)$, i.e. it requires $\log_2(N)$ passes through the N KV pairs. With hashing, the list of KMV pairs can be created in two passes. The first pass populates the hash table with needed count and length information; the second pass copies the key and value datums into the appropriate location in a new KMV data structure. Since the cost to lookup a key in a well-formed hash table is a constant-time $O(1)$ operation, the cost of the data reorganization is also $O(N)$. This methodology has the added benefit that the number of values in each KMV pair is known and can be passed to the user function during a reduce. For some reduce operations, this count is all the information a reduce requires; the values need not be looped over. The count is not available in Hadoop-style data shuffles, which sort the values; counting the values associated with the same key requires iterating over the values.

Note that the first portion of the *collate()* operation involves all-to-all communication (each processor sends and receives data from every other processor) using a distributed hash table. The communication can either be done via a MPI_Alltoall() library call, or by a custom routine that aggregates messages and invokes point-to-point MPI_Send() and MPI_IRecv() calls.

The *reduce()* operation processes KMV pairs and can produce new KV pairs for continued computation. Each processor operates only on the KMV pairs it owns; no communication is required. As with the *map()*, users call the library with a pointer to a user function. The MR-MPI *reduce()* operation invokes the user function, once for each KMV pair.

Several related MapReduce operations are provided by the library. For example, the *collate()* function described above calls two other functions: *aggregate()*, which performs the all-to-all communication, and *convert()*, which turns a list of KV pairs into KMV pairs. Both functions can be called directly. The *compress()* function allows on-processor operations to be performed; it is equivalent to a *convert()* with on-processor KVs as input, followed by a *reduce()*. The *clone()* function turns a list of KV pairs into KMV pairs, with one value per key. The *collapse()* function turns N KV pairs into one KMV pair, with the keys and values of the KV pairs becoming $2N$ values of a single multivalue assigned to a new key. The *gather()* function collects KV pairs from all processors to a subset of processors; it is useful for doing output from one or a few processors. Library calls for sorting datums by key or value or for sorting the values within each multivalue are also provided. These routines invoke the C-library *quicksort()* function to compute the sorted ordering of the KV pairs (or values within a multivalue), using a user-provided comparison function. The KV pairs (or values in a multivalue) are then copied into a new data structure in sorted order.

The interface to various low-level operations is provided so that a user's program can string them together in various ways to produce interesting MapReduce algorithms. For example, output from a *reduce()* can serve as input to a subsequent *map()* or *collate()*. KV pairs from multiple MapReduce objects can be combined to perform new sequences of *map()*, *collate()*, and *reduce()* operations.

The above discussion assumed that the KV or KMV pairs stored by a processor fit in its physical memory. In this case, MR-MPI performs only "in-core" processing and no disk files are written or read by any of the processors, aside from initial input data if it exists or final output data if it is generated. Note that the aggregate physical memory of large parallel machines can be multiple terabytes, which allows for large data sets to be processed in-core, assuming the KV and KMV pairs remain evenly distributed across processors throughout the sequence of MapReduce operations. The use of hashing to assign keys to processors typically provides for such load-balancing. In the next section, we discuss what happens when data sets do not fit in available memory.

3 Out-of-core Issues

Since the MapReduce paradigm was designed to enable processing of extremely large data sets, the MR-MPI library also allows for "out-of-core" processing, which is triggered when the KV or KMV pairs owned by one or more processors do not fit in local memory. When this occurs, a processor writes one or more temporary files to disk, containing KV or KMV pairs, and reads them back in when required. Depending on the parallel machine's hardware configuration, these files can reside on disks local to each processor, on the front end (typically a NSF-mounted file system for a parallel machine), or on a parallel file system/disk array.

When a user program creates a MapReduce object, a "pagesize" can be specified, which defaults to 64 Mbytes. As described below, each MR-MPI operation is constrained to use no more than a handful of these pages. The *pagesize* setting can be as small as 1 Mbyte or as large as desired, though the user should ensure the allocated pages fit in physical memory; otherwise a processor may allocate slow virtual memory. The *pagesize* is also the typical size of individual reads and writes to the temporary disk files; hence a reasonable *pagesize* ensures good I/O performance.

We now explain how the MapReduce operations described in the previous section work in out-of-core mode. The *map()* and *reduce()* operations are relatively simple. As a *map()* generates KV pairs via the user function, a page of memory fills up, one KV pair at a time. When the page is full, it is written to disk. If the source of data for the *map()* operation is an existing set of KV pairs, those datums are read, one page at a time, and a pointer to each KV pair is given to the user function. Similarly, a *reduce()* reads one page of KMV pairs from disk, passes a pointer to each pair, one at a time, to the user function that typically generates new KV pairs. The generated

pairs fill up a new page, which is written to disk when full, just as with the *map()* operation. Thus for both a *map()* and *reduce()*, out-of-core disk files are read and written sequentially, one page at a time, which requires at most two pages of memory.

A special case is when a single KMV pair is larger than a single page. This can happen, for example, in a connected component finding algorithm if the graph collapses into one or a few giant components. In this case, the set of values (graph vertices and edges) associated with a unique key (the component ID), may not fit in one page of memory. The individual values in the multivalue are then spread across as many pages as needed. The user function that processes the KMV pair is passed a flag indicating it received only a portion of the values. Once it has processed them, it can request a new set of values from the MR-MPI library, which reads in a new page from disk.

Performing an our-of-core *collate()* operation is more complex. Recall that the operation occurs in two stages. First, keys are hashed to "owning" processors and the KV datums are communicated to new processors in an all-to-all fashion. In out-of-core mode, this operation is performed on one page of KV data at a time. Each processor reads in a page of KV pairs, the communication pattern (which processors receive which datums in that page) is determined, and all-to-all communication is performed. Each processor allocates a two-page chunk of memory to receive incoming KV pairs. On average each processor should receive one page of KV pairs; the two-page allocation allows for some load imbalance. If the KV pairs are distributed unevenly so that this limit is exceeded, the all-to-all communication is performed in smaller, multiple passes until the full page contributed by every processor has been communicated. Performing an MPI_Alltoall(), or using the custom all-to-all routines provided by the MR-MPI library, requires allocation of auxiliary arrays that store processor indices, datum pointers, and datum lengths. For the case of many tiny KV datums, the total number of memory pages required to perform the all-to-all communication, including the source and target KV pages is at most seven.

When the communication stage of the *collate()* operation is complete, each processor has a set of KV pages, stored in an out-of-core KV file, that need to be reorganized into a set of KMV pages, likewise stored in a new out-of-core KMV file. In principle, creating one page of KMV pairs would require all KV pages be scanned to find all the keys that contribute values to that page. Doing this for each output page of KMV pairs could be prohibitively expensive. A related issue is that, as described in the previous section, a hash table is needed to match new keys with previously encountered keys. But there is no guarantee that the hash table itself will not grow arbitrarily large for data sets with many unique keys. We need an algorithm for generating KMV pairs that operates within a small number of memory pages and performs a minimal number of passes through the KV disk file. An algorithm that meets these goals is diagrammed in Figure 1; it reads the KV pages at most four times, and writes out new KV or KMV pages at most three times. It also uses a finite-size in-memory hash table, that stores unique keys as they are encountered while looping over the KV pages, as well as auxiliary information needed to construct the output pages of KMV pairs.

- (Pass 1) The KV pairs are read, one page at a time, and split into "partition" files, represented by KVp in the figure. Each partition file contains KV pairs whose unique keys (likely) fit in the hash table (HT). Initially, all KV pairs are assigned to the first partition file as the HT is populated. When the HT becomes full, e.g. at the point represented by the horizontal line shown on the leftmost vertical line as a downward scan is performed, the fraction of KV pairs read thus far is used to estimate the number of additional partition files needed. As subsequent KV pairs are read, they are assigned to the original partition file if the KV pair's key is already in the current HT. If not, a subset of bits in the hash value of the key is used to assign the KV pair to one of the new partition files. The number of needed partition files is estimated conservatively, rounding up to the next power-of-two, so that extra passes through

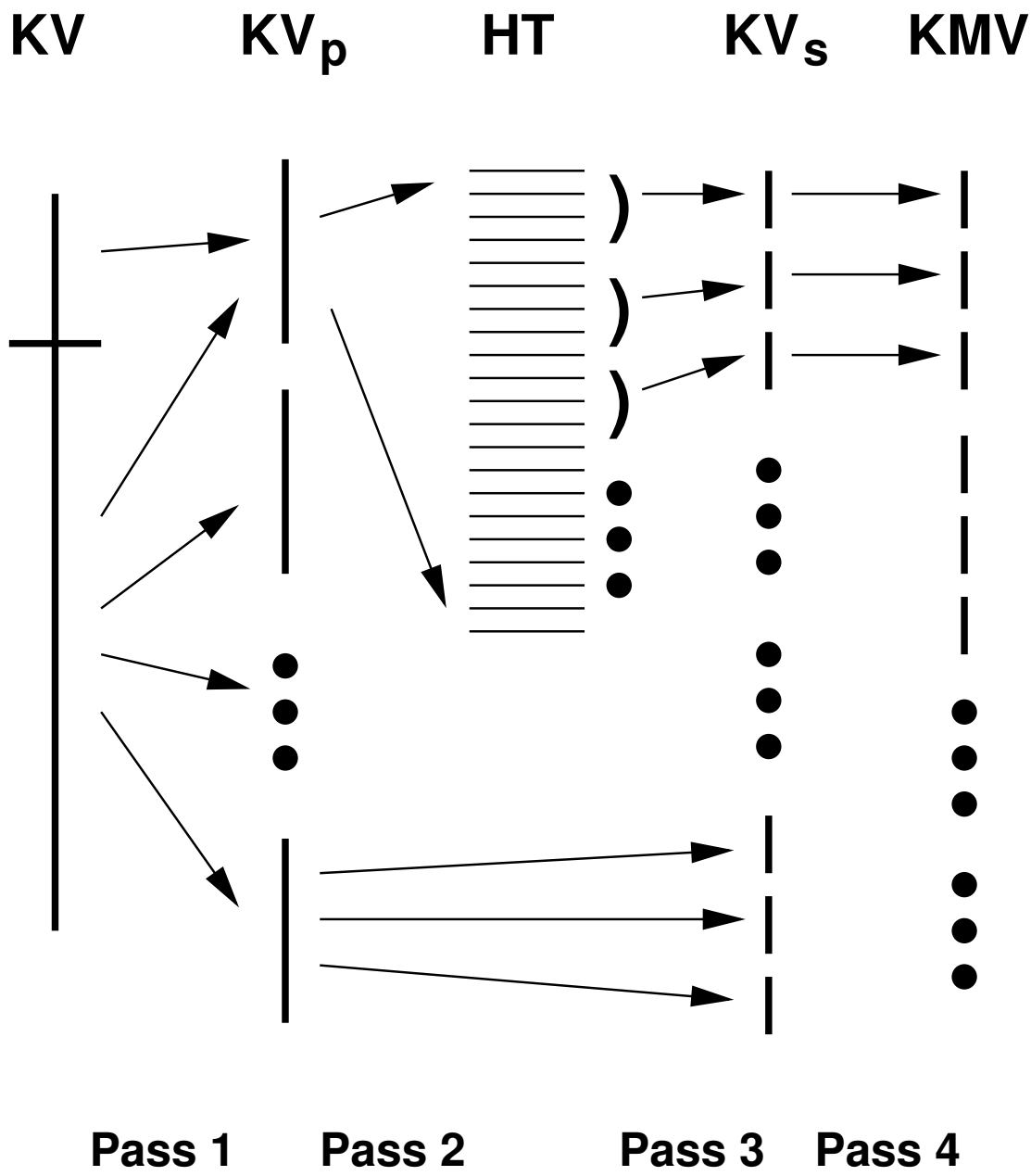

Figure 1: *Multi-pass algorithm for converting KV data to KMV data. The vertical lines represent out-of-core data sets. KV is key/value pairs; HT is an in-memory hash table; KMV is key/multivalue pairs.*

the KV pairs are almost never needed, and so that the bit masking can be done quickly. This first pass entails both a read and write of all the KV pairs.

- (Pass 2) The partition files are read, one at a time. The key for each KV pair is hashed into the HT, accumulating data about the number and size of values associated with each unique key. Before the next partition file is read, passes 3 and 4 are completed for the current partition. Eventually this pass entails a read of all KV pairs.

- (Pass 3) The unique keys in the HT for one partition are scanned to determine the total size of KMV for keys in the HT. The associated values create KMV pairs that span M memory pages. If $M > 1$, the associated partition file of KV pairs is read again, and each KV pair is assigned to one of M smaller "set" files, represented by KVs in the figure. Each set file contains the KV pairs that contribute to the KMV pairs that populate one page of KMV output. If a single KMV pair spans multiple pages because it contains a very large number of values, the corresponding KV pairs are still written to a single set file. Eventually, this pass reads all partition files, entailing both a read and write of all KV pairs.

- (Pass 4) A set file is read and the key/value data for each of its KV pairs is copied into the appropriate location in the page of KMV pairs, using information stored in the HT. When complete, the KMV page is written to disk. This pass eventually reads all set files, entailing a final read of all KV pairs. It also writes all KMV pairs to disk. If each KV pair has a unique key, the volume of KMV output is roughly the same as that of the KV input.

In summary, this data reorganization for the *collate()* operation is somewhat complex, but requires only a small, constant number of passes through the data. The KV datums are read from disk at most four times, and written to disk three times. The first two write passes reorganize KV pairs into partition and set files; the final write pass creates the KMV data set. Depending on the size and characteristics of the KV datums (e.g., the number of unique keys), some of these passes may not be needed. By contrast, a full out-of-core sort of the KV pairs, performed via a merge sort as discussed below is still an $O(N \log_2 N)$ operation. The number of read/write passes through the KV data files depends on the amount of KV data that can be held in memory, but can be a large number for big data sets.

The memory page requirements for the out-of-core *collate()* operation are as follows. Two contiguous pages of memory are used for the hash table. Intermediate passes 2 and 3 can require numerous partition or set files to be opened simultaneously and written to. To avoid small writes of individual KV datums, a buffer of at least 16K bytes is allocated for each file. The precise number of simultaneously open files is difficult to bound, but typically one or two additional pages of memory suffice for all needed buffers. Thus the total number of pages required for the data reorganization stage of the *collate()* operation is less than the seven used by the all-to-all communication stage.

Other MR-MPI library calls discussed in the previous section, such as *clone()*, *collapse()*, and *gather()*, can also operate in out-of-core mode, typically with one pass through the KV or KMV data and the use of one or two memory pages. Sorting KV pairs, by key or value, is an exception. An out-of-core merge sort is performed as follows. Two pages of KV datums are read from disk. Each is sorted using the C-library quicksort() function, as discussed in the previous section. The two pages are then scanned and merged into a new file which is written to disk as its associated memory page fills up. This process requires five memory pages, which includes vectors of KV datum pointers and lengths needed by the in-memory quicksort() operation. Once all pairs of pages have been merged, the sort continues in a recursive fashion, merging pairs of files into a new third file, without the need to perform additional in-memory sorts. Thus the overall memory requirement

is five pages. The number of read/write passes through the KV data set for the merge sort is $O(\log_2 M)$, where M is the number of pages of KV pairs. The number of passes could be reduced at the cost of allocating more in-memory pages.

4 Graph Algorithms in MapReduce

We begin with a MapReduce procedure for creating large, sparse, randomized graphs, since they are the input for the algorithms discussed below. R-MAT matrices [7] are recursively generated graphs with power-law degree distributions. They are commonly used to represent web and social networks. The user specifices six parameters that define the graph: the number of vertices N and edges M, and four parameters a, b, c, d that sum to 1.0 and are discussed below. The algorithm in Figure 2 generates M unique non-zero entries in a sparse $N \times N$ matrix A, where each entry A_{ij} represents an edge between graph vertices (V_i, V_j).

$M_{\text{remain}} = M$
while $M_{\text{remain}} > 0$:
    Map: Generate M_{remain}/P random edges (i, j) on each processor
        output Key = (i, j), Value = NULL
    Collate
    Reduce: Remove duplicate edges
        input Key = (i, j), MultiValue = one or more NULLs
        output Key = (i, j), Value = NULL
    $M_{\text{remain}} = M - N_{kv}$

Figure 2: *MapReduce algorithm for R-MAT graph generation on P processors.*

In the *map()* operation, each of P processors generates a $1/P$ fraction of the desired edges. A single random edge (i, j) is computed recursively as follows. Pick a random quadrant of the A matrix with relative probabilities a, b, c, and d. Treat the chosen quadrant as a sub-matrix and select a random quadrant within it, in the same manner. Repeat this process n times where $N = 2^n$. At the end of the recursion, the final "quadrant" is non-zero matrix element A_{ij}.

Though A is very sparse, the *map()* typically generates some small number of duplicate edges. The *collate()* and *reduce()* operations remove the duplicates. The entire map-collate-reduce sequence is repeated until the number of resulting key/value pairs $N_{kv} = M$. For reasonably sparse graphs this typically takes only a few iterations.

Note that the degree distribution of vertices in the graph depends on the choice of parameters a, b, c, d. If one of the four values is larger than the other three, a skewed distribution results. Variants of the above algorithm can be used when N is not a power-of-two, to generate graphs with weighted edges (assign a numeric value to the A_{ij} edge), graphs without self edges (require $i \neq j$), or graphs with undirected edges (require $i < j$). Or the general R-MAT matrix can be further processed by MapReduce operations to meet these requirements. For some algorithms below (e.g., triangle finding and maximal independent set generation), we post-process R-MAT matrices to create upper-triangular R-MAT matrices representing undirected graphs. For each edge (i, j) with $i > j$, we add edge (j, i), essentially symmetrizing the matrix; we then remove duplicate edges and self edges, leaving only edges with $i < j$.

By using more than one MapReduce object, we can improve the performance of the R-MAT generation algorithm, particularly when out-of-core operations are needed. In the enhanced algorithm (Figure 3), we emit newly created edges into MapReduce object E_{new} and *aggregate()* the edges to processors. We then add those edges to the MapReduce object E_{old}, containing all

previously generated unique edges; the edges of E_{old} are already aggregated to processors using the same mapping as E_{new}. We can then use the *compress()* function to reduce duplicates locally on each processor. The cost of the algorithm is reduced due to the smaller number of KV pairs communicated in the *aggregate()*, compared to the *collate()* operation of the original algorithm, particularly in later iterations of the algorithm where few edges are created.

$M_{\text{remain}} = M$
while $M_{\text{remain}} > 0$:
    Map E_{new}: Generate M_{remain}/P random edges (i, j) on each processor
        output Key = (i, j), Value = NULL
    Aggregate E_{new}
    Add E_{new} to E_{old}
    Compress E_{old}: Remove duplicate edges
        input Key = (i, j), MultiValue = one or more NULLs
        output Key = (i, j), Value = NULL
    $M_{\text{remain}} = M - N_{kv}$

Figure 3: *Enhanced MapReduce algorithm for R-MAT graph generation on P processors.*

4.1 PageRank

The PageRank algorithm assigns a relative numeric rank to each vertex in a graph. In an Internet context, it models the web as a directed graph $G(V, E)$, with each vertex $V_i \in V$ representing a web page and each edge $(V_i, V_j) \in E$ representing a hyperlink from V_i to V_j. The probability of moving from V_i to another vertex V_j is $\alpha/d_{out}(V_i) + (1 - \alpha)/|V|$, where α is a user-defined parameter (usually 0.8-0.9), $d_{out}(V_i)$ is the outdegree of vertex V_i, and $|V|$ is the cardinality of V. The first term represents the probability of following a given link on page V_i; the second represents the probability of moving to a random page. For pages with no outlinks, the first term is $\alpha/|V|$, indicating equal likelihood to move to any other page. Equivalently, the graph can be represented by a matrix A [16], with matrix entries $A_{ij} = \alpha/d_{out}(V_i)$ if vertex V_i links to V_j. To maintain the sparsity of A, terms for random jumps and zero out-degree vertices are not explicitly stored as part of the matrix. Rather, they are computed separately as adjustments to the PageRank vector. The kernel of the PageRank algorithm is thus a power-method iteration consisting of matrix-vector multiplications $A^T x = y$, where x is the PageRank vector from the previous iteration. A few additional dot product and norm computations per iteration are also required.

The algorithm in Figure 4 performs these iterations. The nonzeros of matrix A are initialized as described above; the PageRank vector x is initialized uniformly. A MapReduce object MT contains the indices of all-zero rows of A, corresponding to vertices with no outlinks. The PageRank algorithm is made more efficient by calling *aggregate()* once on the MapReduce objects representing the matrix A, the all-zero matrix rows MT, and the vector x before the PageRank iterations begin. Since these three MapReduce objects have the same key types, pre-aggregating the objects moves all KV pairs with a given key to a single processor. Thus, many of the PageRank operations become local operations, replacing *collate()* operations (which are equivalent to an *aggregate()* followed by *convert()*) with *convert()* calls. In Figure 4, steps where *convert()* operations can be substituted for *collate()* operations are marked with an asterisk. This strategy of pre-aggregating to improve data locality is useful in many MapReduce algorithms, but assumes that specific keys are always mapped to the same processor.

In Step 1 of a PageRank interation, MT is used in dot products to compute adjustments to the PageRank vector to represent the uniform probability of going from a leaf page to any other page.

Adjustments for random jumps from any page are also computed. In Step 2, the matrix-vector product $A^T x$ is computed. This is the most expensive part of the computation, requiring two passes over the nonzero structure of A. In the first pass, a *convert()* gathers all local row entries A_{ij} with their associated x_i entry, and a *reduce()* computes $A_{ij}x_i$. In the second pass, a *collate()* gathers, for each j, all contributions to the column sum $\sum_i A_{ij}x_i$, which is computed by a second *reduce()*. Steps 3 and 4 adjust and scale the product vector; the work is proportional to $|V|$. Step 5 computes the residual to determine whether the PageRank vector has converged; work is again proportional to $|V|$. In Step 6, the PageRank vector is overwritten prior to the next iteration. Throughout the algorithm, we exploit the ability to call MPI_Allreduce to compute global norms and residuals.

4.2 Triangle enumeration

A triangle in a graph is any triplet of vertices (V_i, V_j, V_k) where the edges $(V_i, V_j), (V_j, V_k), (V_i, V_k)$ exist. Figure 5 outlines a MapReduce algorithm that enumerates all triangles, assuming an input graph of undirected edges (V_i, V_j) where $V_i < V_j$ for every edge, i.e. an upper-triangular R-MAT matrix. This exposition follows the triangle-finding algorithm presented in [8].

The initial step is to store a copy of the graph edges as key/value (KV) pairs in an auxiliary MapReduce object G_0, for use later in the algorithm. The first *map()* operation converts edge keys to vertex keys with edge values. After a *collate()*, each vertex has a list of vertices it is connected to; the first *reduce()* can thus flag one vertex V_i in each edge with a degree count D_i. The second *collate()* and *reduce()* assign a degree count D_j to the other vertex in each edge. In the third *map()*, only the lower-degree vertex in each edge emits its edges KV pairs. The task of the third *reduce()* is to emit "angles" for each of these low-degree vertices. An "angle" is a root vertex V_i, with two edges to vertices V_1 and V_2, i.e. a triangle without the third edge (V_1, V_2). The *reduce()* emits a list of all angles of vertex V_i, by a double loop over the edges of V_i. Note that the aggregate volume of KV pairs emitted at this stage is minimized by having only the low-degree vertex in each edge generate angles.

In stage 4, KV pairs from the original graph G_0 are added to the current working set of KV pairs. The KV pairs in G_0 are edges that complete triangles for the angle KV pairs just generated. After the fourth *collate()*, a pair of vertices (V_i, V_j) is the key, and the multivalue is the list of all root vertices in angles that contain V_i and V_j. If the multivalue also contains a NULL, contributed by G_0, then there is a (V_i, V_j) edge in the graph. Thus all vertices in the multivalue are roots of angles that are complete triangles and can be emitted as a triplet key.

4.3 Connected component labeling

A connected component of a graph is a set of vertices where all pairs of vertices in the set are connected by a path of edges. A sparse graph may contain many such components. Figure 6 outlines a MapReduce algorithm that labels each vertex in a graph with a component ID. All vertices in the same component are labelled with the same ID, which is the ID of a vertex in the component. We assume an input graph of undirected edges (V_i, V_j). This exposition also follows the connected-component algorithm presented in [8], with the addition of logic that load-balances data across processors, when one or a few giant components exist in the graph.

The algorithm begins (before the iteration loop) by assigning each vertex to its own component or "zone," so that $Z_i = V_i$. Each iteration grows the zones, one layer of neighbors at a time. As zones collide due to shared edges, a winner is chosen (the smaller zone ID), and vertices in the losing zone are reassigned to the winning zone. When the iterations complete, each zone has become a fully connected component. The algorithm thus finds all connected components in the

graph simultaneously. The number of iterations required depends on the largest diameter of any component in the graph.

The first *map()* operation emits the vertices in each edge as keys, with the edge as a value. The current zone assignment of each vertex is added to the set of key/value (KV) pairs. The first *collate()* operation collects all the edges of a vertex and its zone assignment together in one multi-value. The first *reduce()* operation re-emits each edge, tagged by the zone assignment of one of its vertices.

Since each edge was emitted twice, the second *collate()* operation collects the zone assignments for its two vertices together. If the two zone IDs are different, the second *reduce()* operation chooses a winner (the smaller of the two IDs), and emits the loser ID as a key, with the winning ID as a value. If no zone ID changes are emitted, the algorithm is finished, and the iterations cease.

The third *map()* operation inverts the vertex/zone KV pairs to become zone/vertex pairs. The third *add()* operation adds the changing zone assignments to the set of KV pairs. The third *collate()* can then collect all the vertices of a zone and zero or more reassignments for the zone ID. Since a zone could collide with multiple other zones on the same iteration due to shared edges, the new zone ID becomes the minimum ID of any of the neighboring zones. If no zone reassignment value appears in the multi-value, the zone ID is unchanged. The final *reduce()* emits a KV pair for each vertex in the zone, with the vertex as a key and the new zone ID as a value.

Note that if a graph has only a few components, the third *collate()* operation, which collates by the zone ID, may generate a few very large key/multi-value (KMV) pairs. For example, if the entire graph is fully connected, in the last iteration, a single KMV pair contains all vertices in the graph and is assigned to one processor. This imbalance in memory and computational work can lead to poor parallel performance of the overall algorithm. To counter this effect, the various operations of stage 3 include extra logic. The idea is to partition zones whose vertex count exceeds a user-defined threshhold into P sub-zones, where P is the number of processors. The 64-bit integer that stores the zone ID also stores a bit flag indicating the zone has been partitioned and a set of bits that encode the processor ID.

During the third *map()* operation, if the zone has been partitioned, the vertex is assigned to a random processor and the processor ID bits are added to the zone ID, as indicated by the Z_i^+ notation in Figure 6. Likewise, if Z_i has been partitioned, the third *add()* operation emits the zone ID change KV pair (Z_i,Z_{winner}) as (Z_i^+,Z_{winner}). In this case (Z_i,Z_{winner}), is emitted not once, but $P+1$ times, once for each processor, and once as if Z_i had not been partitioned (for a reason discussed below).

With this additional partitioning logic, the third *collate()* operation (which keys on zone IDs, some of which now include processor bits) collects only $1/P$ of the vertices in large zones onto each processor. But the multivalue on each processor contains all the zone-reassignments relevant to the unpartitioned zone. Thus, the third *reduce()* operation can change the zone ID (if necessary) in a consistent manner across all P multi-values that contain the zone's vertices. When the *reduce()* operation emits new zone assignments for each vertex, the zone retains its partitioned status; the partition bit is also explicitly set if the vertex count exceeds the threshhold for the first time.

Note that this logic does not guarantee that the partition bits of the zone IDs for all the vertices in a single zone are set consistently on a given iteration. For example, an unpartitioned zone with a small ID may consume a partitioned zone. The vertices from the partitioned zone retain their partitioned status, but the original vertices in the small zone may not set the partition bit of their zone IDs. On subsequent iterations, the third *add()* operation emits $P+1$ copies of new zone reassignments for both partitioned and unpartioned zone IDs, to ensure all vertices in the zone know the correct reassignment information.

4.4 Maximal independent set identification

An "independent" set of vertices from a graph is one where no pair of vertices in the set shares an edge. The set is "maximal" if no vertex can be added [1]. Finding a maximal independent set (MIS) is useful in several contexts, such as identifying large numbers of independent starting vertices for graph traversals. Luby's algorithm [17] is a well-known parallel method for finding a MIS. Figure 7 outlines a MapReduce version of Luby's algorithm, assuming an input graph of undirected edges (V_i, V_j) where $V_i < V_j$ for every edge, i.e. an upper-triangular R-MAT matrix.

Before the iterations begin, a random value is assigned to each vertex, which is used to compare pairs of vertices. The vertex ID itself could be used as the random value, but since the vertices eventually selected for the MIS depend on the randomization, this may introduce an unwanted bias. Instead a random number generator can be used to assign a consistent random value R_i to each vertex in each edge (via a *map()* operation), which is then carried along with the vertex ID through each stage of the algorithm. In the notation of Figure 7, each V_i is then really two quantities, a vertex ID (1 to N) and R_i. Alternatively, the R_i can be used to relabel the vertices in a random manner, assigning each a new, unique vertex ID from 1 to N. In this case, V_i is simply the new vertex ID, and the vertices in the final MIS could be remapped to recover their original IDs. These operations could be performed with a few extra MapReduce steps, outside the iteration loop.

The pre-iteration *clone()* operation converts the initial list of edge key/value (KV) pairs one by one to key/multivalue (KMV) pairs, with the edge as the key, and NULL as the value. Thus, the iterations can begin with a *reduce()* operation, without need for a *collate()* operation. A second empty MapReduce object is also initialized, which accumulates MIS vertices as the iterations proceed.

At each iteration, the current set of edges is examined to identify winning vertices. Vertices adjacent to winners are flagged as losers, and all edges adjacent to losers are flagged for removal at the start of the next iteration. Vertices with all losing edges are flagged as winners. At each iteration edges are removed from the graph, and winning vertices are added to the MIS. When the graph is empty, the MIS is complete.

The first *reduce()* operation flags the two vertices in each edge as a potential winner or loser. This is done by comparing the two random values for the vertices (or the randomized vertex IDs as explained above). The first *collate()* operation thus produces a multivalue for each vertex V_i which contains all the vertices it shares an edge with, and associated winner/loser flags. In the second *reduce()* operation, if V_i won the comparison with all its edge neighbors, then it is a "winner" vertex. All vertices connected to the winner are emitted with a loser flag. If V_i is not an overall winner, its edge vertices are emitted without the loser flag. The second *collate()* operation collects this new information for each vertex.

In the third *reduce()* operation, if V_i was flagged as a loser by any edge neighbor who was a winner, it becomes a losing vertex and emits all its edge vertices with a loser flag. Otherwise it emits all its edge vertices without the flag.

In the fourth *reduce()* operation, some vertices have all their edge vertices flagged as losers. These vertices are all winners, either the ones identified in the second *reduce()* operation, or additional vertices who would become singletons (having no edges) when edges containing a loser vertex are removed (on the next iteration). All of these winning vertices are emitted to the MapReduce object that stores the accumulating MIS.

All edges of each vertex are also emitted, retaining the loser flag if they have it; they are emitted with $E_{ij} = (V_i, V_j)$ as their key, with $V_i < V_j$. This is to ensure both copies of the edge come

[1] A maximal set is "maximum" if it is the largest maximal set. Finding a graph's maximum independent set is an NP-hard problem, which MapReduce is unlikely to help with.

together via the fourth *collate()* operation. The multivalue for each edge now has two values, either of which may be NULL or a loser flag.

At the first stage of the next iteration, any edge flagged by either value as a loser does not compare its two vertices; it is effectively deleted from the graph. The Luby iterations end when all edges have been removed from the graph. At this point, the second MapReduce object contains a complete MIS of vertices.

4.5 Single source shortest path calculation

The single-source shortest path (SSSP) algorithm computes, for a directed graph with weighted edges, the shortest weighted distance from a chosen source vertex to all other vertices in the graph. This operation is straightforward when global graph information is available. The source vertex is labeled with distance zero. Then in each iteration, edge adjacencies of labeled vertices are followed, and adjacent vertices are labeled with updated distances. When no distances change, the iterations are complete. Most importantly, only edges from modified vertices are visited in each iteration.

For distributed graphs, global information about edge adjacencies and labeled vertices is not available. Instead, following the Bellman-Ford-style algorithm of [6, 4, 12], every edge of the graph is visited in each iteration, although labels for an edge's vertices are updated only if the breadth-first traversal has reached those vertices. A MapReduce formulation of this algorithm, described in Figure 8, begins with a *map()* function loading edges (V_i, V_j) and source vertex V_s with distance 0. In each iteration, the edges (V_i, V_j) are collated with respect to V_i, and a *reduce()* operation updates each V_j's distance if V_i's distance is non-zero.

As with PageRank and R-MAT generation, significant savings in communication and execution time can be achieved by pre-aggregating certain MapReduce objects and storing data in more than one MapReduce object. The enhanced algorithm is shown in Figure 9. The MapReduce objects storing vertices and edges are aggregated to processors only once at the beginning of the computation. In each iteration of SSSP, only a small number of new candidate distances are generated. The enhanced algorithm uses MapReduce object U to store these updates. Only the updates in U are communicated; each update is aggregated to the same processor that stores the corresponding vertex and its edges. Thus, the total amount of communication needed is smaller than in the original algorithm, which aggregates all graph edges in each iteration. Updated vertex distances are stored in MapReduce object V.

5 Performance Results

In this section, we present performance results for the MapReduce graph algorithms of the preceeding section, implemented as small C++ programs calling our MR-MPI library. The benchmarks were run on a medium-sized Linux cluster of 2 GHz dual-core AMD Opteron processors connected via a Myrinet network. Most importantly for MR-MPI, each node of the cluster has one local disk, which is used for out-of-core operations in MR-MPI. To avoid contention for disk I/O, we ran all experiments with one MPI process per node. For comparisons with other implementations, we used either the same cluster or, where noted, Sandia's Cray XMT, a multi-threaded parallel computer with 500 MHz processors and a 3D-Torus network.

We ran each of the algorithms on three R-MAT graphs of different sizes, each on a varying number of processors. Details of the input data are shown in Table 1. The *small* problem (around 8M edges) can typically be run on a single processor without incurring out-of-core operations. The *medium* problem (around 134M edges) can be run on a single processor with out-of-core operations; larger

Data Set	# of vertices	# of edges	Maximum vertex degree
RMAT-20 (small)	$2^{20} \approx 1M$	$2^{23} \approx 8M$	$\approx 24K$
RMAT-24 (medium)	$2^{24} \approx 17M$	$2^{27} \approx 134M$	$\approx 147K$
RMAT-28 (large)	$2^{28} \approx 268M$	$2^{31} \approx 2B$	$\approx 880K$

Table 1: Characteristics of R-MAT input data for graph algorithm benchmarks.

processor configurations, however, do not necessarily require out-of-core operations. The *large* problem (around 2B edges) requires out-of-core operations on our 64-node cluster.

All data sets used R-MAT parameters $(a, b, c, d) = (0.57, 0.19, 0.19, 0.05)$ and generated 8 edges per vertex (on average). These values create a highly skewed degree distribution, as indicated by the maximumn vertex degree in Table 1.

The resulting timings give a sense of the inherent scalability of the MapReduce algorithms as graph size grows on a fixed number of processors, and of the parallel scalability for computing on a graph of fixed size on a growing number of processors. Where available, we compare the MapReduce algorithm with other parallel implementations, including more traditional distributed-memory algorithms and multi-threaded algorithms in the Multi-Threaded Graph Library (MTGL) [5] on the Cray XMT. We compute parallel efficiency on P processors as $(time_M \times M)/(time_P \times P)$ where M is the smallest number of processors on which the experiment was run, and $time_I$ is the execution time required on I processors.

5.1 R-MAT generation results

In Figure 10, we show the scalability of the R-MAT generation algorithm (Figure 3) for RMAT-20, RMAT-24 and RMAT-28. For RMAT-20 and RMAT-24, superlinear speed-up is shown. This speed-up is due to the decreased amount of file I/O needed with greater numbers of processors; with a larger total memory, a fixed-size problem requires less I/O with more processors. For RMAT-28, which requires significant out-of-core operations, parallel efficiency ranges from 52% to 97%. Comparisons between the enhanced algorithm (Figure 3) and the original algorithm (Figure 2) showed approximately 10% reduction in execution time for the enhanced algorithm.

5.2 PageRank results

In Figure 11, we show the performance of the MR-MPI PageRank algorithm (Figure 4) compared to a distributed-memory matrix-based implementation using the linear algebra toolkit Trilinos [14]. The matrix-based distributed-memory implementation of PageRank uses Trilinos Epetra matrix/vector classes to represent the graph and PageRank vector. Rows of matrix A and the associated entries of the PageRank vector x are uniquely assigned to processors; a random permutation of the input matrix effectively load balances the non-zero matrix entries across processors. Interprocessor communication gathers x values for matrix-vector multiplication and sums partial products into the y vector. Most communication is point-to-point communication, but some global communication is needed for computing residuals and norms of x and y.

Figure 11 shows the execution time per PageRank iteration for R-MAT matrices RMAT-20, RMAT-24 and RMAT-28. Converging the PageRank iterations to tolerance 0.002 requires five or six iterations. Several R-MAT matrices of each size were generated for the experiments; the average time over the matrices is reported here. The Trilinos implementations show near-perfect strong scaling for RMAT-20 and RMAT-24. The MR-MPI implementations also demonstrate good strong scaling. However, MR-MPI's execution time is at least an order of magnitude greater than the

Trilinos implementation. This result is due to two factors: (*i*) a higher volume of communication in the MapReduce implementation (where all edges are communicated in each iteration), compared to the Trilinos implementation (where only vertex-based data is communicated), and (*ii*) out-of-core operations in MR-MPI were performed because of the page-size restriction, while all Trilinos operations were performed in-core. The benefit of MR-MPI's out-of-core implementation is seen, however, with the RMAT-28 data set, which could be solved on smaller processor sets than the Trilinos implementation. For these experiments, the Trilinos implementation required 64 processors for the RMAT-28 data set, since it will not operate on matrices that exceed the aggregate memory of the processors.

5.3 Triangle finding results

In Figure 12, we show the performance of the triangle finding algorithm (Figure 5). Execution times for this algorithm were too large to allow the problem to be easily run with RMAT-28 on our cluster. Parallel efficiencies for RMAT-20 ranged from 80% to 140%; for RMAT-24, they ranged from 50% to 120%.

5.4 Connected Components

In Figure 6, we show the performance of the connected component identification algorithm (Figure 6). A comparison is made with a hybrid "Giant Connected Component" implementation using both Trilinos and MR-MPI. Power law graphs often have one or two very large components and many very small components. The hybrid algorithm exploits this feature of the data by using inexpensive breadth-first search (BFS) from the vertex with highest degree to identify the largest components, followed by a more expensive algorithm to identify the small components in the remainder of the graph. In our hybrid implementation, we intially perform matrix-vector multiplications in Trilinos to perform a BFS, finding components that include (in total) 70% or more of the vertices. We then apply our MR-MPI algorithm to the remaining graph to identify the small components. This hybrid approach is quite effective, reducing the execution time to identify all components of RMAT-20 and RMAT-24 by 80-99%, as shown in Figure 13. The benefit of MR-MPI is seen, however, for RMAT-28, where the graph is too large to fit into memory, and thus Trilinos cannot perform the initial BFS for our hybrid algorithm.

5.5 Maximally independent set results

The execution times for the maximal independent set algorithm (Figure 7) are shown in Figure 14. Like the R-MAT generation results, superlinear speed-up of the algorithm occurs for RMAT-20 and RMAT-24, as more of the graph fits into processor memory and less file I/O is needed. For RMAT-28, the algorithm requires significant out-of-core operations. In this case, parallel efficiency is nearly perfect going from 8 to 64 processors.

5.6 Single-source shortest path results

The execution times for the single-source shortest path algorithms (Figures 8 and 9) are shown in Figure 15. The results show the benefit of using the enhanced algorithm, providing at least a 17% (and often greater) reduction in execution time due to reduced communication; only the updated distances are communicated throughout most of the enhanced algorithm. However, the execution times are still large compared to multi-threaded implementations; for example, Madduri et al. [18]

Implementation	Number of Processes	SSSP Execution Time
Hadoop	48	38,925 secs.
MR-MPI original (Figure 8)	48	13,505 secs.
MR-MPI enhanced (Figure 9)	48	8,031 secs.
XMT/C	32	37 secs.

Table 2: Execution times for SSSP with WebGraphB.

Data Set	R-MAT a	R-MAT b	R-MAT c	R-MAT d	# of vertices	# of edges	Maximum vertex degree
nice	0.45	0.15	0.15	0.25	2^{25}	2^{28}	1108
nasty	0.57	0.19	0.19	0.05	2^{25}	2^{28}	230,207

Table 3: Characteristics of R-MAT input data for PageRank and Connected Components scalability experiments.

report execution times of only 11 seconds on 40 processors for a multithreaded implementation on a Cray MTA-2 (the predecessor of the XMT) on graphs with one billion edges.

To compare our MR-MPI implementation with a wider set of algorithms, we performed experiments comparing MR-MPI, Hadoop, PBGL (Parallel Boost Graph Library) [13] and a multithreaded implementation on the Cray XMT using two web graphs: WebGraphA with 13.2M vertices and 31.9M edges, and WebGraphB with 187.6M vertices and 531.9M edges. In Figure 16, we show execution times for the SSSP algorithm using WebGraphA. Like our MR-MPI implementation, the Hadoop implementation is a Bellman-Ford-style [4, 12] algorithm. The XMT and PBGL implementations are based on delta-stepping [19], and do not require full iterations over the entire edge list to advance a breadth-first search. We observe that the MR-MPI implementation runs in less time than the Hadoop implementation, but requires significantly more time than the XMT and PBGL implementations. In experiments with WebGraphB, the benefit of the enhanced algorithm (Figure 9) is clearly shown, with a 40% reduction in execution time compared to the original SSSP algorithm (Figure 8). But the Bellman-Ford-style iterations are especially harmful to the MR-MPI and Hadoop implementations for WebGraphB, which required 110 iterations to complete; execution times for this data set are shown in Table 2.

5.7 Scalability to large numbers of processors

Finally, we demonstrate the scalability of our MR-MPI library to large numbers of processors. The library was used on Sandia's Redstorm and Thunderbird parallel computers. Redstorm is a large Cray XT3 with 2+ GHz dual/quad-core AMD Opteron processors and a custom interconnect providing 9.6 GB/s of interprocessor bandwidth. Thunderbird is a large Linux cluster with 3.6 GHz dual-core Intel EM64T processors connected by an Infiniband network. Because these systems do not have local disks for each processor, we selected a data set and page sizes that fit in memory, so out-of-core operations were not needed. For these experiments, we used an R-MAT data set with with 2^{25} vertices and 2^{28} edges, with parameters given in Table 3. We ran both the PageRank and Connected Components algorithms.

Figure 17, shows the performance of the various PageRank implementations on distributed memory and multi-threaded architectures. The MR-MPI and Trilinos implementations are described in Sections 4.1 and 5.2, respectively. In the multi-threaded MTGL implementation, rank propagates via adjacency list traversal in a compressed sparse-row data structure. To maintain its scalability, code must be written so that a single thread spawns the loop that processes all in-neighbors of a

given vertex; this detail enables the compiler to generate hotspot-free code.

The MR-MPI implementation demonstrated good scalability up to 1024 processors; however, as before, it required an order-of-magnitude more execution time than the matrix-based implementations on Redstorm. The distributed memory matrix-based implementations are competitive with the multi-threaded implementation in MTGL on the Cray XMT.

Similar results were obtained for the Connected Components algorithm, as shown in Figure 18. As with PageRank, the MR-MPI implementation showed good scalability up to 1024 processors, but required significantly more time than the hybrid algorithm using Trilinos and MR-MPI or the MTGL algorithm.

6 Lessons Learned

We conclude with several observations about performing MapReduce operations on distributed-memory parallel machines via MPI.

MapReduce achieves parallelism through randomizing the distribution of data across processors, which often intentionally ignores data locality. This translates into maximal data movement (during a data shuffle or *collate()* operation) with communication between all pairs of processors. But the benefit is often good load-balance, even for hard-to-balance irregular data sets. By contrast, more traditional distributed-memory parallel algorithms, e.g. for matrix operations, or grid-based or particle-based simulation codes, typically work hard to localize data and minimize communication. To do this they often require a lot of application-specific logic and parallel communication coding to create and maintain a data decomposition, generate ghost versions of nearby spatially-decomposed data, etc.

MapReduce algorithms can be hard to design, but are often relatively easy to write and debug. Thinking about a computational task from a MapReduce perspective is different from traditional distributed-memory parallel algorithm design. For example, with an MPI mindset, it often seems heretical to intentionally ignore data locality. However, writing small *map()* and *reduce()* functions is typically easy. And writing an algorithm that involves complex parallel operations without needing to write application-specific code to communicate data via MPI calls or to move data between out-of-core disk files and processor memory is often a pleasant surprise. Moreover, if the MapReduce algorithm is initially coded so that it runs correctly on one processor, it often works out-of-the-box on hundreds or thousands of processors, without the need for additional debugging.

Performing MapReduce operations on a fixed allocation of processors on a traditional MPI-based parallel machine is a somewhat different conceptual model than that of performing MapReduce operations on a cloud computer using (for example) Hadoop. In the former case, one can potentially control which processor owns which data at various stages of an algorithm. This is somewhat hidden from the user in a typical cloud-computing model, where data simply exists somewhere in the cloud and Hadoop ensures data moves where it is needed and is operated on by some processor. The cloud model is a nice data-centric abstraction which allows for fault tolerance both to data loss (via redundant storage) and to processor failure (via reassignment of work), neither of which is typically possible on current MPI-based parallel machines.

However, since the MPI implementation of MapReduce as described in this paper is processor-centric, one can sometimes fruitfully exploit the possibility for processors to maintain "state" over the course of multiple map and reduce operations. By controlling where data resides for maps and reduces (e.g. via a user-specified hash function), and by assuming that processor will always be available, more efficient operations with less data movement are sometimes possible. The discussion of enhanced graph algorithms in Section 4 illustrated this. To fully exploit this idea for large data

sets, mechanisms and data structures are needed for processors to store, retrieve, and efficiently find needed datums on local disk (e.g. static edges of a graph), so that archived state can be used efficiently during subsequent map and reduce operations.

Finally, though this paper focuses on graph algorithms expressed as MapReduce operations, there is nothing about the MR-MPI library itself that is graph specific. We hope the library can be generally useful on large-scale monolithic or cloud-style parallel machines that support MPI, for a variety of data-intensive or compute-intensive problems that are amenable to solution using a MapReduce paradigm.

7 Acknowledgements

We thank the following individuals for their contributions to this paper: Greg Bayer and Todd Plantenga (Sandia) for explaining Hadoop concepts to us, and for the Hadoop implementations and timings of Section 5; Jon Cohen (DoD) for fruitful discussions about his MapReduce graph algorithms [8]; Brian Barrett (Sandia) for the PBGL results of Section 5; Jon Berry (Sandia) for the MTGL results of Section 5, and for his overall support of this work and many useful discussions.

Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] Apache Hadoop WWW site: `http://hadoop.apache.org/`.

[2] MapReduce-MPI software WWW site: `http://www.sandia.gov/~sjplimp/download.html#mapreduce`.

[3] Titan Informatics Toolkit WWW site: `http://titan.sandia.gov/`.

[4] R. Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16(1):87–90, 1958.

[5] J. W. Berry, B. Hendrickson, S. Kahan, and P. Konecny. Software and algorithms for graph queries on multithreaded architectures. In *Proceedings of the 21st International Parallel and Distributed Processing Symposium*, March 2007.

[6] C. Bisciglia, A. Kimball, and S. Michels-Slettvet. Lecture 5: Graph algorithms and PageRank, 2007. `http://code.google.com/edu/submissions/mapreduce-minilecture/lec5-pager%ank.ppt`.

[7] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-MAT: A recursive model for graph mining. In *SIAM Data Mining*, 2004.

[8] J. Cohen. Graph twiddling in a mapreduce world. *Computing in Science and Engineering*, 11:29–41, 2009.

[9] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In *OSDI04: Sixth Symposium on Operating System Design and Implementation*, 2004.

[10] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[11] J. Ekanayake and G. Fox. High performance parallel computing with clouds and cloud technologies. In *First International Conference on Cloud Computing (CloudComp09)*, 2009.

[12] L. R. Ford and D. R. Fulkerson. *Flows in networks*. Princeton University Press, 1962.

[13] D. Gregor and A. Lumsdaine. The parallel BGL: A generic library for distributed graph computations. *Parallel Object-Oriented Scientific Computing*, July 2005.

[14] M. Heroux, R. Bartlett, V. H. R. Hoekstra, J. Hu, T. Kolda, R. Lehoucq, K. Long, R. Pawlowski, E. Phipps, A. Salinger, H. Thornquist, R. Tuminaro, J. Willenbring, and A. Williams. An Overview of Trilinos. Technical Report SAND2003-2927, Sandia National Laboratories, 2003.

[15] T. Hoefler, A. Lumsdaine, and J. Dongarra. Towards efficient mapreduce using mpi. In M. Ropo, J. Westerholm, and J. Dongarra, editors, *PVM/MPI*, volume 5759 of *Lecture Notes in Computer Science*, pages 240–249. Springer, 2009.

[16] A. N. Langville and C. D. Meyer. A survey of eigenvector methods for web information retrieval. *The SIAM Review*, 47(1):135–161, 2005.

[17] M. Luby. A simple parallel algorithm for the maximal independent set problem. *SIAM J. Comput.*, 15(4):1036–1055, 1986.

[18] K. Madduri, D. A. Bader, J. W. Berry, and J. R. Crobak. An experimental study of a parallel shortest path algorithm for solving large-scale graph instances. In *ALENEX*. SIAM, 2007.

[19] U. Meyer and P. Sanders. Delta-stepping: A parallel single source shortest path algorithm. In *In ESA '98: Proceedings of the 6th Annual European Symposium on Algorithms*, pages 393–404. Springer-Verlag, 1998.

[20] T. Tu, C. A. Rendleman, D. W. Borhani, R. O. Dror, J. Gullingsrud, M. O. Jensen, J. L. Klepeis, P. Maragakis, P. Miller, K. A. Stafford, and D. E. Shaw. A scalable parallel framework for analyzing terascale molecular dynamics simulation trajectories. In *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pages 1–12, Piscataway, NJ, USA, 2008. IEEE Press.

Input:
∗ Pre-aggregated MapReduce object A containing $|V| \times |V|$ matrix: Key $= i$, Value $= [j, A_{ij}]$
∗ Pre-aggregated MapReduce object MT containing indices of all-zero rows of A: Key $= i$, Value $= 0$
∗ Pre-aggregated MapReduce object x containing initial PageRank vector: Key $= i$, Value $= 1/|V|$
While $(residual > tolerance)$
1 Compute contribution for random jumps and zero-outdegree vertices
Add: x to MT
∗ Convert MT: Vertex is Key
Reduce MT:
Input: Key $= i$; MultiValue $= x_i, 0$ if $i \in MT$; MultiValue $= x_i$ if $i \notin MT$
Compute: Local contributions c_{local}
Output: If $nvalues = 2$, Key $= i$; Value $= 0$
MPI_Allreduce c_{local} with MPI_SUM to compute global contribution c_{global}
2 Compute $y = A^T x$
Copy x to y
Add A to y
∗ Convert y: Key is row index of A
Reduce y:
Input: Key $= i$; Multivalue $= i, [j, A_{ij}]$ for nonzeros A_{ij} in A
Output: Key $= j$; Value $= A_{ij} y_i$
Collate y: Key is column index of A
Reduce y:
Input: Key $= j$; Multivalue $= A_{ij} y_i$ for all non-zero entries A_{ij} of column j
Output: Key $= j$; Value $= \sum_i A_{ij} y_i$
3 Add contribution c_{global} to y; compute local max norm n_{local} of y
Map y:
Input: Key $= i$, Value $= y_i$
Compute: $y_i = y_i + c_{global}$; if $y_i > n_{local}$, $n_{local} = y_i$
Output: Key $= i$, Value $= y_i$
MPI_Allreduce n_{local} with MPI_MAX to compute global max norm n_{global}
4 Scale y by n_{global}
Map y:
Input: Key $= i$, Value $= y_i$
Output: Key $= i$, Value $= y_i / n_{global}$
5 Compute residual
Add y to x
∗ Convert x: Key is index i
Reduce x:
Input: Key $= i$; Multivalue $= x_i, y_i$
Compute: if $|x_i - y_i| > resid$, $resid = |x_i - y_i|$
MPI_Allreduce with MPI_MAX to compute $residual$
6 $x = y$;
Output: PageRank vector x

Figure 4: *MapReduce algorithm for PageRank vertex ranking; pre-aggregating the MapReduce objects allows many communication intensive collate() operations to be replaced by strictly local convert() operations (as marked with asterisks).*

1 Copy: G_0 = copy of edge KV pairs from input graph
1 Map: Convert edges to vertices
input Key = (V_i, V_j), Value = NULL
output Key = V_i, Value = V_j
output Key = V_j, Value = V_i
1 Collate
1 Reduce: Add first degree to one vertex in edge
input Key = V_i, MultiValue = $(V_j, V_k, ...)$
for each V in MultiValue:
if $V_i < V$: output Key = (V_i, V), Value = $(D_i, 0)$
else: output Key = (V, V_i), Value = $(0, D_i)$
2 Collate
2 Reduce: Add second degree to other vertex in edge
input Key = (V_i, V_j), MultiValue = $((D_i, 0), (0, D_j))$
output Key = (V_i, V_j), Value = (D_i, D_j) with $V_i < V_j$
3 Map: Low degree vertex emits edges
if $D_i < D_j$: output Key = V_i, Value = V_j
else if $D_j < D_i$: output Key = V_j, Value = V_i
else: output Key = V_i, Value = V_j
3 Collate
3 Reduce: Emit angles of each vertex
input Key = V_i, MultiValue = $(V_j, V_k, ...)$
for each V_1 in MultiValue:
for each V_2 beyond V_1 in MultiValue:
if $V_1 < V_2$: output Key = (V_1, V_2), Value = V_i
else: output Key = (V_2, V_1), Value = V_i
4 Add: Add G_0 edge KV pairs to angle KV pairs
4 Collate
4 Reduce: Emit triangles
input Key = (V_i, V_j), MultiValue = $(V_k, V_l, NULL, V_m, ...)$
if NULL exists in MultiValue:
for each non-NULL V in MultiValue:
output Key = (V_i, V_j, V), Value = NULL

Figure 5: *MapReduce algorithm for triangle enumeration.*

Iterate:

1 Map: Convert edges to vertices
- input Key = $E_{ij} = (V_i, V_j)$, Value = NULL
- output Key = V_i, Value = E_{ij}
- output Key = V_j, Value = E_{ij}

1 Add: Zone assignment of each vertex
- output Key = V_i, Value = Z_i

1 Collate: Vertex as key

1 Reduce: Emit edges of each vertex with zone of vertex
- input Key = V_i, MultiValue = $EEEE...Z$
- for each E in MultiValue:
  - output Key = E_{ij}, Value = Z_i

2 Collate: Edge as key

2 Reduce: Emit zone re-assignments
- input Key = E_{ij}, MultiValue = Z_iZ_j
- Z_{winner} = min(Z_i,Z_j); Z_{loser} = max(Z_i,Z_j)
- if Z_i and Z_j are different:
  - output Key = Z_{loser}, Value = Z_{winner}

2 Exit: if no output by Reduce 2

3 Map: Invert vertex/zone pairs
- input Key = V_i, Value = Z_i
- if Z_i is not partitioned:
  - output Key = Z_i, Value = V_i
- else:
  - output Key = Z_i^+, Value = V_i for a random processor

3 Add: Changed zones (Z_i,Z_{winner})
- if Z_i is not partitioned:
  - output Key = Z_i, Value = Z_{winner}
- else:
  - output Key = Z_i^+, Value = Z_{winner} for every processor
  - output Key = Z_i, Value = Z_{winner}

3 Collate: Zone ID as key

3 Reduce: Emit new zone assignment of each vertex
- input Key = Z_i or Z_i^+, MultiValue = $VVVV...ZZZ...$
- Z_{new} = min(Z_i or Z_i^+,Z,Z,Z,...)
- partition Z_{new} if number of V > threshhold
- for each V in MultiValue:
  - output Key = V_i, Value = Z_{new}

Figure 6: *MapReduce algorithm for connected component labeling.*

Map: assign random values to each vertex V_i and V_j of edge E_{ij}.
Clone: convert edge KV pairs directly to KMV pairs
Create: empty MapReduce object MRv for maximal independent set vertices

while $N_{edges} > 0$:

1 Reduce: Determine WINNER/LOSER vertex of each edge
- input Key = E_{ij} with $V_i < V_j$, Multivalue = NULL or LOSER
- if either value is LOSER, emit nothing
- V_w = WINNER vertex, V_l = LOSER vertex
- output Key = V_w, Value = V_l WINNER
- output Key = V_l, Value = V_w LOSER

1 Collate: Vertex as key

2 Reduce: Find WINNER vertices
- input Key = V_i, MultiValue = $VVVV$...
- if all V are WINNERs:
  - for each V_j in Multivalue:
    - output Key = V_j, Value = V_i LOSER
- else:
  - for each V_j in Multivalue:
    - output Key = V_j, Value = V_i

2 Collate: Vertex as key

3 Reduce: Find LOSER vertices
- input Key = V_i, MultiValue = $VVVV$...
- if any V is LOSER:
  - for each V_j in Multivalue:
    - output Key = V_j, Value = V_i LOSER
- else:
  - for each V_j in Multivalue:
    - output Key = V_j, Value = V_i

3 Collate: Vertex as key

4 Reduce: Emit WINNER vertices and LOSER edges
- input Key = V_i, MultiValue = $VVVV$...
- if all V are LOSERs:
  - output to MRv, Key = V_i, Value = NULL
- for each V_j in Multivalue:
  - if V_j is LOSER:
    - output Key = E_{ij} with $V_i < V_j$, Value = LOSER
  - else:
    - output Key = E_{ij} with $V_i < V_j$, Value = NULL

4 Collate: Edge as key

Figure 7: *MapReduce algorithm for finding a maximal independent set of graph vertices via Luby's algorithm.*

Input:
    Source vertex V_s
    MapReduce object V containing graph vertices V_i with distances d_i from V_s (initially ∞):
        Key = V_i; Value = ∞
    MapReduce object E containing graph edges (V_i, V_j) with weights w_{ij}:
        Key = V_i; Value = $[V_j, w_{ij}]$
Map V: Set distance for source vertex V_s to zero
    Input: Key = V_i; Value = ∞
    Output: Key = V_i; Value = 0 if $V_i = V_s$, otherwise Value = ∞
while not *done*:
    done = 1;
    Add: E to V
    Collate V: Key is vertex V_i
    Reduce V: Loop over all edges to perform breadth-first search from labeled vertices
        Input: Key = V_i; Multivalue = edges $[V_j, w_{ij}]$, candidate distances d_i
        Compute: Find smallest candidate distance d_{min} for V_i
            If $d_{min} \neq d_i$, *done* = 0;
        Output: Key = V_i; Value = d_{min} (updated distance for V_i)
            and if $d_{min} \neq d_i$, Key = V_j; Value = $d_{min} + w_{ij}$ for each edge (V_i, V_j)
    MPI_Allreduce on *done* with MPI_MIN
Output: MapReduce object V with Key = V_i, Value = d_i.

Figure 8: *MapReduce algorithm for single-source shortest path (SSSP) calculation.*

Input:
    Source vertex V_s.
    Pre-aggregated MapReduce object V containing graph vertices V_i with distances d_i from V_s:
        Key = V_i; Value = ∞.
    Pre-aggregated MapReduce object E containing graph edges (V_i, V_j) with weights w_{ij}:
        Key = V_i; Value = $[V_j, w_{ij}]$
U = new MapReduce object to store updates to distances.
Map U: Add V_s with distance $d_s = 0$ to U.
    Input: V_s
    Output: Key = V_s; Value = 0.
while not *done*:
    done = 1;
    Aggregate U.
    Add U to V. Empty U.
    Compress V: Pick best candidate distance for each vertex.
        Input: Key = V_i; Multivalue = candidate distances d_i.
        Compute: Find smallest candidate distance d_{min} for V_i.
            If $d_{min} \neq d_i$, *done* = 0;
        Output to V: Key = V_i; Value = d_{min}.
        Output to U: if $d_{min} \neq d_i$, Key = V_i; Value = d_{min}.
    MPI_Allreduce on *done* with MPI_MIN.
    if not *done*
        Add: U to E. Empty U.
        Compress E: Generate candidate distances for neighbors of changed vertices.
            Input: Key = V_i; Multivalue = $[V_j, w_{ij}]$ for edges (V_i, V_j)
                and distance d_i if updated in previous step.
            Output to U: If updated distance d_i exists,
                Key = V_j; Value = $d_i + w_{ij}$ for each edge (V_i, V_j).
Output: MapReduce object V with Key = V_i, Value = d_i.

Figure 9: *Enhanced MapReduce algorithm for single-source shortest path (SSSP) calculation. Pre-aggregating the vertices and edges, and later aggregating only updates to vertex distances, reduces communication and execution time.*

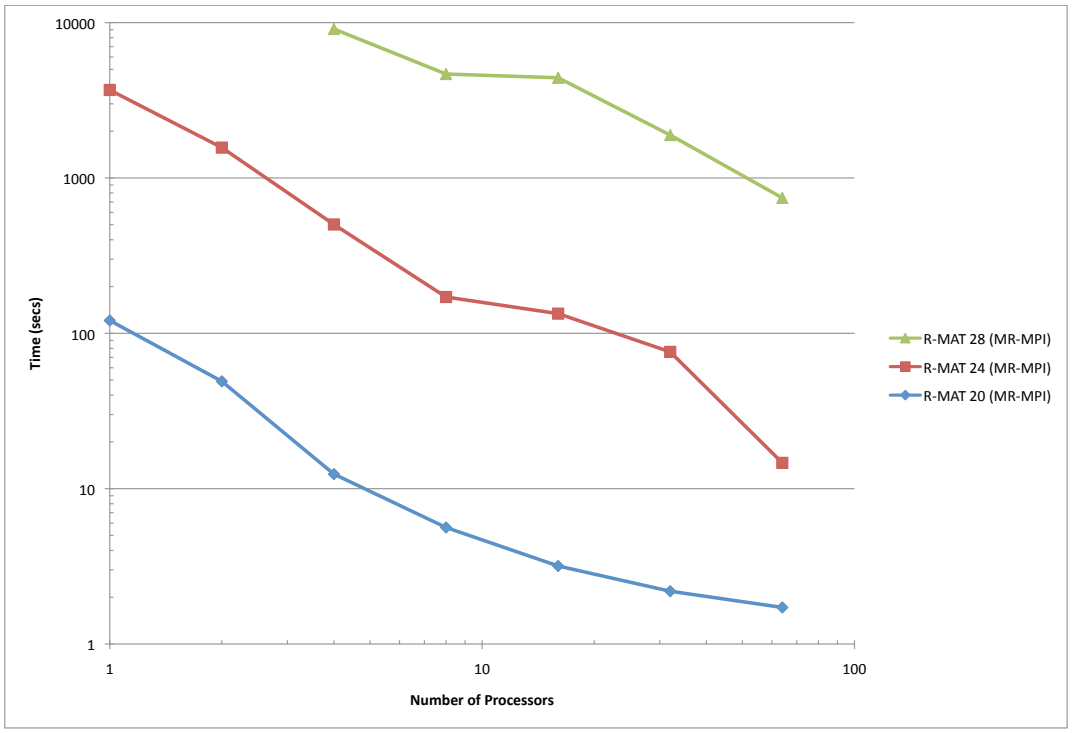

Figure 10: Performance of the MR-MPI R-MAT generation algorithm (Figure 3)

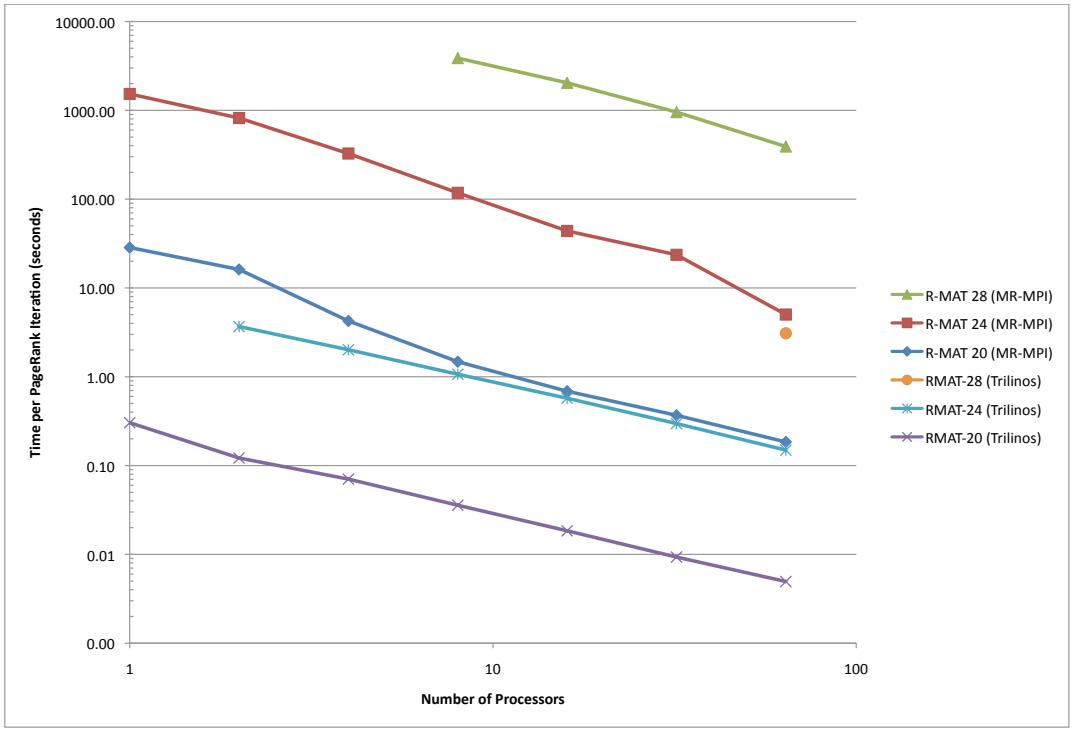

Figure 11: Comparison of PageRank implementations using MR-MPI (Figure 4) and Trilinos' matrix/vector classes on R-MAT data sets.

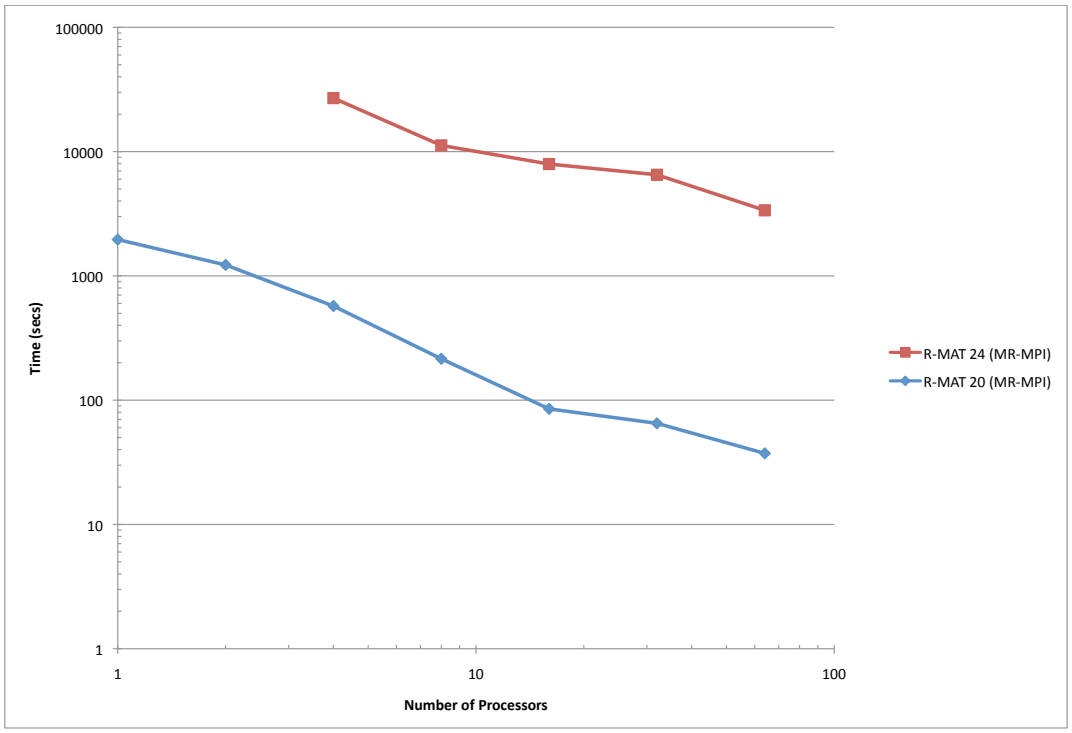

Figure 12: Performance of the MR-MPI triangle-finding algorithm (Figure 5).

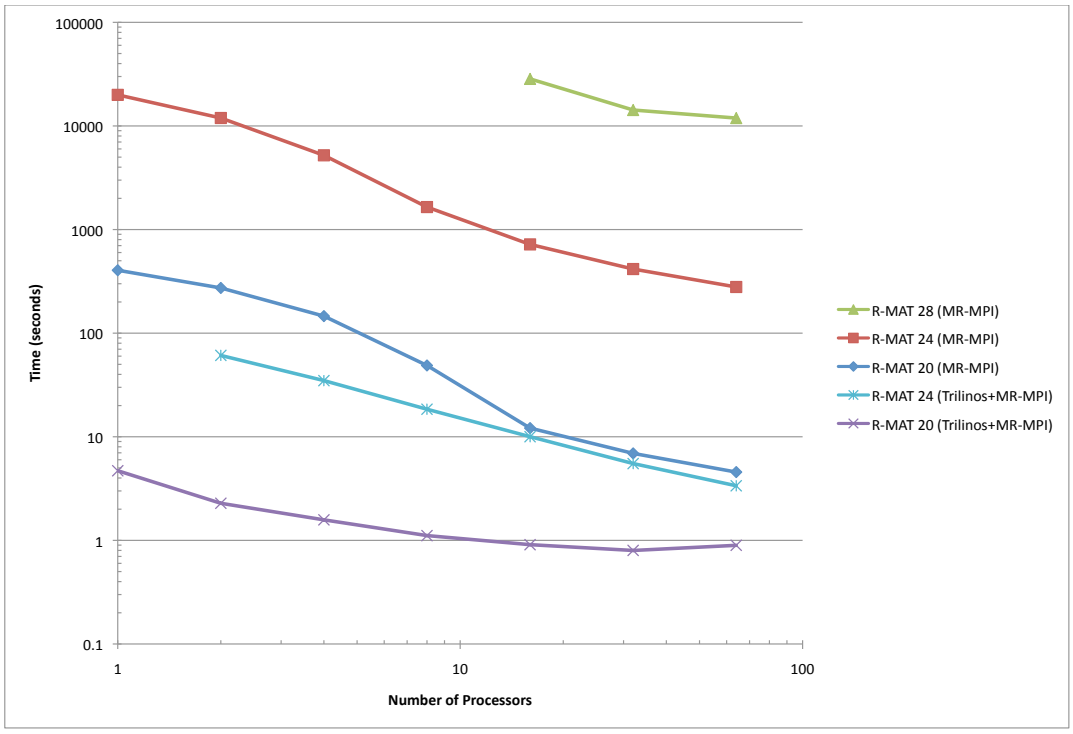

Figure 13: Performance of the MR-MPI connected components algorithm (Figure 6) compared with a hybrid "Giant Connected Component" algorithm based on Trilinos.

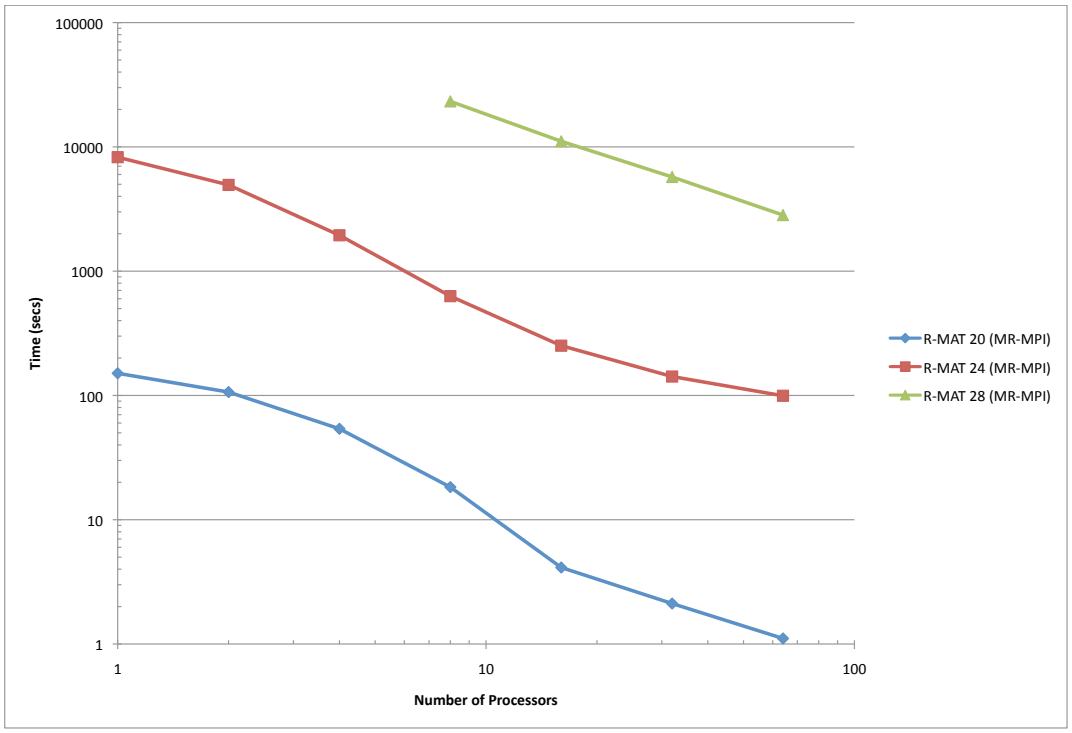

Figure 14: Performance of the MR-MPI maximal independent set algorithm (Figure 7).

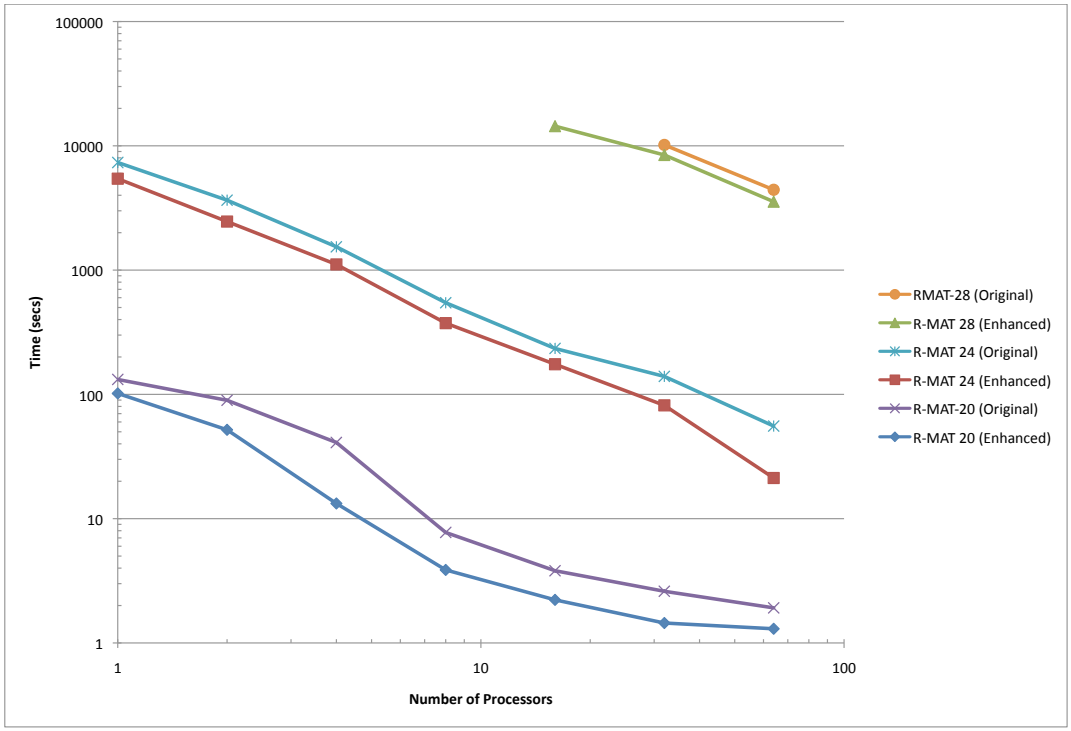

Figure 15: Execution times for SSSP using MR-MPI with R-MAT matrices. Both the original algorithm (Figure 8) and the enhanced algorithm (Figure 9) are included.

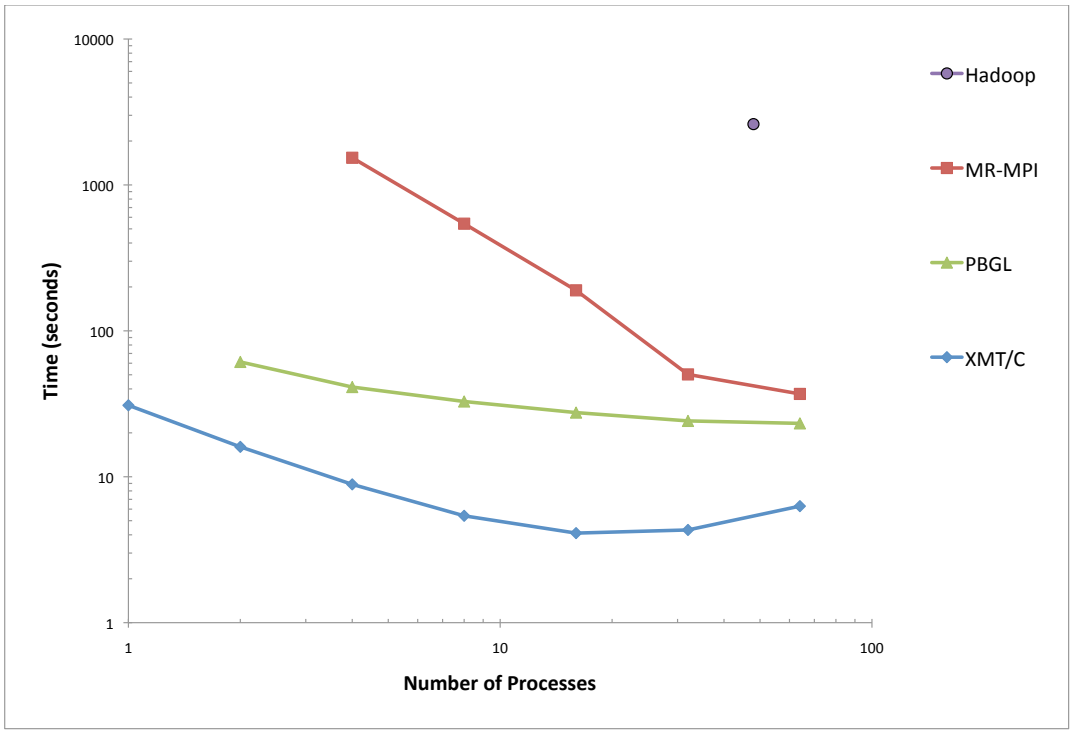

Figure 16: SSSP using MR-MPI for WebGraphA with 13.2M vertices and 31.9M edges. Runtimes using Hadoop and PBGL are also shown.

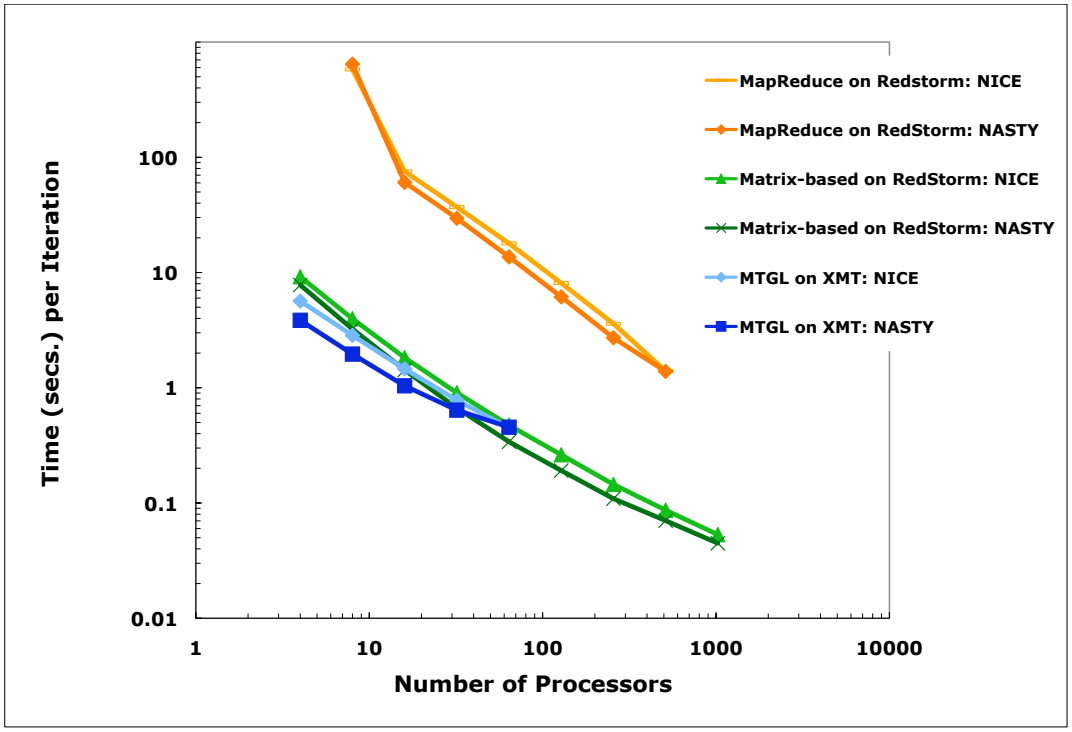

Figure 17: Scalability comparison of PageRank using MapReduce (MR-MPI), matrix-based (Trilinos), and multi-threaded (MTGL) implementation on the R-MAT data sets in Table 3.

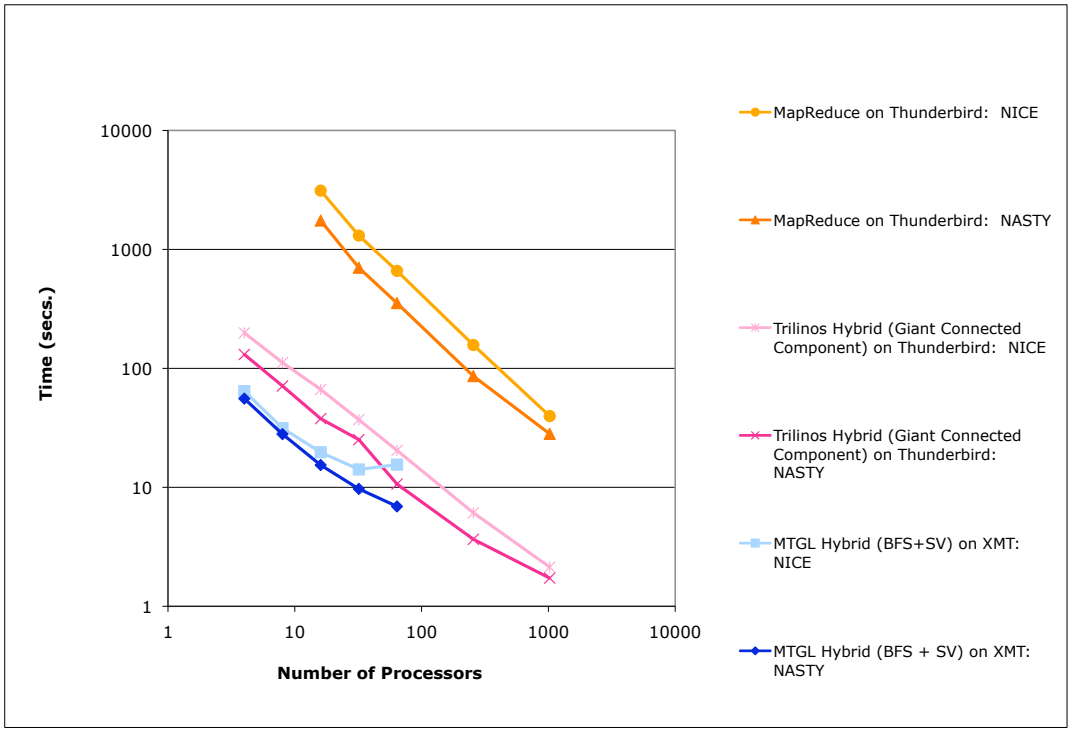

Figure 18: Scalability comparison of Connected Components algorithms using MapReduce (MR-MPI), matrix-based/MapReduce hybrid (Trilinos/MR-MPI), and MTGL implementations on the R-MAT data sets in Table 3.