To Push or To Pull: On Reducing Communication and Synchronization in Graph Computations

Maciej Besta¹, Michał Podstawski²³, Linus Groner¹, Edgar Solomonik⁴, Torsten Hoefler¹

¹Department of Computer Science, ETH Zurich;
²Perform Group Katowice; ³Katowice Institute of Information Technologies;
⁴Department of Computer Science, University of Illinois at Urbana-Champaign

maciej.besta@inf.ethz.ch, michal.podstawski@performgroup.com, gronerl@student.ethz.ch, solomona2@illinois.edu, htor@inf.ethz.ch

ABSTRACT

We reduce the cost of communication and synchronization in graph processing by analyzing the fastest way to process graphs: pushing the updates to a shared state or pulling the updates to a private state. We investigate the applicability of this push-pull dichotomy to various algorithms and its impact on complexity, performance, and the amount of used locks, atomics, and reads/writes. We consider 11 graph algorithms, 3 programming models, 2 graph abstractions, and various families of graphs. The conducted analysis illustrates surprising differences between push and pull variants of different algorithms in performance, speed of convergence, and code complexity; the insights are backed up by performance data from hardware counters. We use these findings to illustrate which variant is faster for each algorithm and to develop generic strategies that enable even higher speedups. Our insights can be used to accelerate graph processing engines or libraries on both massively-parallel shared-memory machines as well as distributed-memory systems.

CCS CONCEPTS

• Computer systems organization → Parallel architectures; Multicore architectures; Distributed architectures; • Theory of computation → Design and analysis of algorithms; Graph algorithms analysis; Shortest paths; Parallel algorithms; Shared memory algorithms; Massively parallel algorithms; • Computing methodologies → Parallel computing methodologies; Parallel algorithms; Shared memory algorithms.

This is an arXiv version of a paper published at ACM HPDC’17 under the same title

Code:
https://spcl.inf.ethz.ch/Research/Parallel_Programming/PushPull

1 INTRODUCTION

Graph processing underlies many computational problems in social network analysis, machine learning, computational science, and others [38]. Designing efficient parallel graph algorithms is challenging due to several properties of graph computations such as irregular communication patterns or little locality [57]. These properties lead to expensive synchronization and movements of large data amounts on shared- and distributed-memory (SM, DM) systems.

Direction optimization in breadth-first search (BFS) [4] is one of the mechanisms that are used to alleviate these issues. It combines the traditional top-down BFS (where vertices in the active frontier iterate over all unvisited neighbors) with a bottom-up scheme (where unvisited vertices search for a neighboring vertex in the active frontier [55]). Combining these two approaches accelerates BFS by ≈2.4x on real-world graphs such as citation networks [4].

We first illustrate that distinguishing between bottom-up and top-down BFS can be generalized to many other graph algorithms, where updates can be either pushed by a thread to the shared state (as in the top-down BFS), or pulled to a thread’s private state (as in the bottom-up BFS). As another example, consider a PageRank (PR) computation and assume a thread X is responsible for a vertex v. X can either push v’s rank to update v’s neighbors, or it can pull the ranks of v’s neighbors to update v. Despite many differences between PR and BFS (e.g., PR is not a traversal), PR can similarly be viewed in the push-pull dichotomy.

This notion sparks various questions. Can pushing and pulling be applied to any graph algorithm? How to design push and pull variants of various algorithms? Is pushing or pulling faster? When and why? Does it depend on the utilized programming model and abstraction? When and how can pushing or pulling be accelerated?

We seek to answer these and other questions and provide the first extensive analysis on the push-pull dichotomy in graph processing. Now, this dichotomy was identified for some algorithms [4, 60] and was used in several graph processing frameworks, such as Ligra [52] and Gemini [65]. Yet, none of these works analyzes the differences in formulations, complexity, and performance between the two approaches for various algorithms, environments, or models.

As a motivation, consider Figure 1 with the results of our push/pull variants of graph coloring [11]. They unveil consistent advantages of pushing. The figure also shows the speedup from a strategy GrS (“Greedy-Switch”) that (1) reduces the number of memory access with a traversal-based graph coloring, and (2) switches between push- or pull-based scheme and an optimized greedy variant.

Figure 1: (§ 6.1) Boman graph coloring [11] results and (§ 6.2) the analysis of the strategy Greedy-Switch (GrS); single node of a Cray XC30, 16 threads.

We provide the following contributions:
We first describe the necessary concepts.

We analyze pushing and pulling with PRAM and derive the differences in the amount of synchronization and communication in both variants of the considered algorithms.

We analyze performance of push- and pull-based algorithms for both SM and DM systems that represent fat-memory nodes and supercomputers. Various programming models are incorporated, including threading, Message Passing (MP), and Remote Memory Access (RMA) [25] for various classes of graphs. For detailed insights, we gather performance data (e.g., cache misses or issues branches and atomic instructions) using PAPI counters.

We incorporate strategies to reduce the amount of synchronization in pushing and memory accesses in pulling and illustrate that they accelerate various algorithms.

We provide performance insights that can be used to enhance graph processing engines or libraries.

Finally, we discuss whether the push-pull dichotomy is applicable in the algebraic formulation of graph algorithms.

2 MODELS, NOTATION, CONCEPTS

We first describe the necessary concepts.

2.1 Machine Model and Simulations

Parallel Random Access Machine (PRAM) [22] is a well-known model of a parallel computer. There are \( P \) processors that exchange data by accessing cells of a shared memory of size \( M \) cells. They proceed in tightly-synchronized steps: no processor executes an instruction \( i+1 \) before all processors complete an instruction \( i \). An instruction can be a local computation or a read/write from/to the memory. We use \( S \) and \( W \) to denote \textit{time} and \textit{work}: the longest execution path and the total instruction count. There are three PRAM variants with different rules for concurrent memory accesses to the same cell. EREW prevents any concurrent accesses. CREW allows for concurrent reads but only one write at a time. CRCW enables any concurrent combination of reads/writes and it comes with multiple flavors that differently treat concurrent writes. We use the Combining CRCW (CRCW-CB) [30]: the value stored is an associative and commutative combination of the written values.

Now, a simulation of one PRAM machine on another is a scheme that enables any instruction from the former to be executed on the latter. Simulation schemes are useful when one wants to port an algorithm developed for a stronger model that is more convenient for designing algorithms (e.g., CRCW) to a weaker one that models hardware more realistically (e.g., CREW). The used simulations are:

Simulating CRCW/CREW on CREW/EREW Any CRCW with \( M \) cells can be simulated on an MP-cell CREW/EREW with a slowdown of \( \Theta(\log n) \) and memory \( MP \) (similarly to simulating a CREW on an EREW) [30].

Limiting \( P \) (LP) A problem solvable on a \( P \)-processor PRAM in \( S \) time can be solved on a \( P' \)-processor PRAM \( (P' < P) \) in time \( S' = \left\lceil \frac{S P}{P'} \right\rceil \) for a fixed memory size \( M \).

2.2 Graph Model, Layout, and Notation

A tuple \( (V, E) \) models an undirected graph \( G: V \) is a set of vertices and \( E \subseteq V \times V \) is a set of edges; \( |V| = n \) and \( |E| = m \) \( d(v) \) and \( N(v) \) are the degree and the neighbors of a vertex \( v \). The (non-negative) weight of an edge \( (u, w) \) is \( \mathcal{W}_{(u,w)} \). We denote the maximum degrees for a given \( G \) as \( \hat{d}, \hat{d}_{in} \) (in-degree), and \( \hat{d}_{out} \) (out-degree). The average degree is denoted with a bar \( \bar{d} \). \( G \)'s diameter is \( D \).

The neighbors of each \( v \) form an array. The arrays of all the vertices form a contiguous array accessed by all the threads; we also store offsets into the array that determine the beginning of the array of each vertex. The whole representation takes \( n + 2m \) cells.

We partition \( G \) by vertices (1D decomposition) [16]. We denote the number of used threads/processes as \( P \). We name a thread (process) that owns a given vertex \( v \) as \( t[v] \). We focus on label-setting algorithms. In some of the considered schemes (e.g., PageRank) the number of iterations \( L \) is a user-specified parameter.

2.3 Atomic Operations

Atomic operations (atomics) appear to the system as if they occur instantaneously. They are used in lock-free graph computations to perform fine-grained updates [29, 44]. Here, we use CPU atomics that operate on integers. We now present the relevant operations:

- \textbf{Fetch-and-Add}(target, arg) (FAA): it increases \textit{*target} by arg and also returns \textit{*target}’s previous value.
- \textbf{Compare-and-Swap}(target, compare, value, result) (CAS): if \( \textit{*target} == \textit{compare} \) then \( \textit{*target} = \textit{value} \) and \( \textit{result} = \textit{true} \) are set, otherwise \( \textit{*target} \) is not changed and \( \textit{result} = \textit{false} \).

2.4 Communication & Synchronization

Unless stated otherwise, we associate communication with: intra- or inter-node reads and writes, messages, and collective operations other than barriers. Synchronization will indicate: any atomic operations, locks, and any form of barrier synchronization.

3 PUSH-PULL: APPLICABILITY

We first analyze what algorithms can be expressed in the push-pull (PP) dichotomy; we revisit existing schemes and discuss new cases.

3.1 PageRank (PR)

PR [15] is an iterative centrality algorithm that obtains the rank of each vertex \( v: r(v) = (1 - f)/|V| + \sum_{w \in N(v)} (f \cdot r(w)/d(w)); f \) is the \textit{damp factor} [15], PR is used to rank websites.

\textbf{Pushing and Pulling?} PR can be expressed in both [60]. In the former, \( f[v] \) updates all \( v \)’s neighbors with a value \( r(v)/d(v) \) (it pushes the value from \( v \) to \( N(v) \)). In the latter, \( f[v] \) updates \( v \) with values \( r(u)/d(u), u \in N(v) \) (it pulls the updates from \( N(v) \) to \( v \)).

3.2 Triangle Counting (TC)

In TC, one counts the number of triangles that each vertex \( v \in V \) is a part of; a triangle occurs if there exist edges \( \{v, w\}, \{w, u\}, \{u, v\} \), where \( u,w \in V \) and \( u, w \neq v, u \neq w \). TC is used in various statistics and machine learning schemes [48] and libraries such as igraph [19].

\textbf{Pushing and Pulling?} This algorithm is also expressible in both schemes. Consider a thread \( t[v] \) that counts the number of triangles associated with a vertex \( v \) (tpc(v)). It iterates over \( N(v) \) and, for each
\(u \in N(v)\), it iterates over \(N(u)\) and checks if \(\exists w \in V, v \neq w \neq u\) such that \(w \in N(u) \cap N(v)\); the final sums are divided by 2 at the end. If yes, then, in the push variant, it increments either one of \(tc(u)\) and \(tc(w)\) while in the pull scheme it increments \(tc(v)\).

### 3.3 Breadth-First Search (BFS)

The goal of BFS [18] is to visit each vertex in \(G\). The algorithm starts with a specified root vertex \(r\) and visits all its neighbors \(N(r)\). Then, it visits all the unvisited neighbors of the root’s neighbors, and continues to process each level of neighbors in one step. BFS represents graph traversals and is used the HPC benchmark Graph500 [44].

**Pushing and Pulling?** There exist both variants. The former is the traditional top-down BFS where \(t[v]\) (if \(v\) is in a frontier) checks each unvisited \(u \in N(v)\) and adds it to the next frontier \(F\) (it pushes the updates from \(v\) to \(N(v)\)). The latter is the bottom-up approach [4, 55]: in each iteration every unvisited vertex \(u\) is tested if it has a parent in \(F\) (the updates are pulled from \(N(u)\) to \(u\)).

### 3.4 Single Source Shortest Path (SSSP)

SSSP outputs the distance from a selected source vertex \(s\) to all other vertices. We consider \(\Delta\)-Stepping (SSSP-\(\Delta\)) [42] that combines the well-known Dijkstra’s and Bellman-Ford algorithms by trading work-optimality for more parallelism. It groups vertices into buckets and only vertices in one bucket can be processed in parallel. SSSP has applications in e.g., operations research.

**Pushing and Pulling?** Both are applicable when relaxing edges of each vertex \(v\) from the current bucket. In the former, \(v\) pushes relaxation requests to its neighbors in the buckets with unsettled vertices. In the latter, vertices in unsettled buckets look for their neighbors in the current bucket and perform (pull) relaxations. A similar scheme was used in the DM implementation of SSSP-\(\Delta\) [17].

### 3.5 Betweenness Centrality (BC)

BC measures the importance of a vertex \(v\) based on the number of shortest paths that lead through \(v\). Let \(\sigma_{s,t}(v)\) be the number of shortest paths between two vertices \(s, t\), and let \(\sigma_{s,t}(v)\) be the number of such paths that lead through \(v\). BC of \(v\) equals \(bc(v) = \sum_{s \neq t \in V} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}\).

Here, we consider Brandes’ algorithm [14, 46, 54]. Define the dependency of a source vertex \(s\) on \(v\): \(\delta_v(s) = \sum_{t \in V} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}\). Then, we have \(bc(v) = \sum_{s \in V} \delta_v(s)\). And one \(\delta_v(s)\) satisfies the following equation:

\[
\delta_v(s) = \sum_{w \in pred(s,v)} \frac{\delta_w(v)}{\sigma_{w,v}} (1 + \delta_v(w)); \quad \text{pred}(s, w) \text{ is a list of immediate predecessors of } w \text{ in the shortest paths from } s \text{ to } w.
\]

Brandes’ scheme uses this recurrence to compute \(bc(v)\) in two phases. First, BFS or SSSP traversals compute \(\text{pred}(s, v)\) and \(\sigma_{s,v}\), \(s \in V\), obtaining a tree \(T\) over \(G\). Next, \(T\) is traversed backwards (from the highest to the lowest distance) to compute \(\delta_v(s)\) and \(bc(v)\) based on the equations above. BC is a complexity central scheme used in biology, transportation, and terrorism prevention [2].

**Pushing and Pulling?** Both parts of Brandes BC can be expressed using push and pull. The first phase can compute shortest path information using either top-down or bottom-up BFS or push-and pull-based versions of SSSP. The second phase (backward accumulation) may also be cast as BFS from a starting frontier. In particular, one can either push partial centrality scores to predecessors or pull them from lists of successors [39].

### 3.6 Graph Coloring (GC)

GC assigns colors to vertices so that no two incident vertices share the same color and the number of colors is minimized. We consider Boman graph coloring (BGC) [11]. Here, each iteration has two phases. In phase 1, colors are assigned to vertices owned by each thread (i.e., to each partition \(P \in \mathcal{P}\) separately without considering other partitions (\(\mathcal{P}\) denotes a set of all partitions). The maximum number of available colors can be specified as a parameter \(C\). In phase 2, border vertices (i.e., vertices with at least one edge leading to another partition; they form a set \(B\)) are verified for conflicts. If there are any, the colors are reassigned. This may cause conflicts within partitions, which are resolved during the next iteration. More iterations \(L\) may improve a solution (fewer colors used).

GC has multiple applications in scheduling and pattern matching.

**Pushing and Pulling?** Both can be used in phase 2. For every border vertex \(v\), each \(u \in N(v) (t[u] \neq t[v])\) is analyzed. If \(v\) and \(u\) share the assigned color, then either \(u\)’s or \(v\)’s color is scheduled for a change (the update is pushed to or pulled from \(N(v)\)).

### 3.7 Minimum Spanning Tree (MST)

The goal of MST is to derive a spanning tree of \(G\) with the lowest sum of the included edge weights. The classical sequential algorithms: Prim [18] and Kruskal [18] lack parallelism. Therefore, we focus on the Boruvka [13] algorithm (more details on pushing and pulling in Prim and Kruskal are still provided in the technical report). In Boruvka, each vertex is first associated with its own supervertex. In each iteration, two incident supervertices are merged into one along an edge \(e_m\) of minimum weight. The algorithm proceeds until there is only one supervertex left. The selected minimum edges form the MST. MST algorithms are utilized in problems such as the design of broadcast trees [18].

**Pushing and Pulling in Boruvka?** First, selecting \(e_m\) adjacent to a given supervertex can be done by pushing (each supervertex overrides adjacent supervertices and their tentative minimal edges if it has a less expensive one) or by pulling (each supervertex picks its own \(e_m\)). Next, merging adjacent supervertices can also be done with pushing or pulling. Assume that each thread owns a number of supervertices. Now, it can either push the changes to the supervertices owned by other threads or pull the information on the adjacent supervertices and only modify its owned ones.

### 3.8 Push-Pull Insights

First, we present a generic difference between pushing and pulling. Recall that \(t[v]\) indicates the thread that owns \(v\). Define \(t \rightarrow v\) to be true if \(t\) modifies \(v\) during the execution of a given algorithm (\(t \rightarrow v \iff v\) modifies \(v\)). Then

\[
(\text{Algorithm uses pushing}) \Rightarrow (\exists t \in \{1, T\}, v \in V, t \rightarrow v \text{ and } t \neq t[v])
\]

\[
(\text{Algorithm uses pulling}) \Rightarrow (\forall t \in \{1, T\}, v \in V, t \rightarrow v \Rightarrow v = t[v])
\]

In pushing, any thread \(t\) may access and modify any vertex \(v \in V\) so that we may have \(t \neq t[v]\). In pulling, \(t\) can only modify its assigned vertices: \(t[v] = v\) for any \(v\) modified by \(t\). In §4, we show that this property determines that pulling requires less synchronization.
compared to pushing. However, pushing can often be done with less work, when only a subset of vertices needs to update its neighbors.

Second, our analysis shows that the push-pull dichotomy can be used in two algorithm classes: iterative schemes (PR, TC, GC, Boruvka MST) that derive some vertex properties and perhaps proceed in iterations until some convergence condition is met, and traversals (BFS, SSSP, Δ, BC).

## 4 THEORETICAL ANALYSIS

We now derive detailed specifications of push and pull algorithm variants and use them to investigate the differences between pushing and pulling. We (1) identify read and write conflicts, (2) conduct complexity analyses, and (3) investigate the amount of required atomics or locks. We focus on the CRCW-CB and CREW models.

### Algorithm Listings

Our schemes have multiple variants as many nested loops can be parallel; we indicate them with \[ \text{in par} \]. Unless specified otherwise, we only consider the loops without square brackets in complexity analyses. We mark the read/write conflicts in the parts of the code related to pushing or pulling with \( \text{in par} \), respectively. We indicate the data type in the modified memory cell to be either integer \( \text{in par} \) or float \( \text{in par} \). Finally, we use \( \text{in par} \) grey backgrounds to indicate pushing/pulling variants.

### Cost Derivations

We consider up to one processor per vertex, \( P \leq n \) (and \( P > \hat{d} \)). Thus, pulling avoids write-conflicts, as each thread accumulates updates for a given vertex. Still, pushing can update the same vertices multiple times at every iteration.

We formulate cost analyses of all algorithms via the primitives \( k\text{-RELAXATION} \) and \( k\text{-FILTER} \). \( k\text{-RELAXATION} \) corresponds to simultaneously propagating updates from/to \( k \) vertices to/from one of their neighbors for pushing/pulling. \( k\text{-FILTER} \) is used to extract the vertices updated in one or more \( k\text{-RELAXATIONS} \), and is non-trivial only when pushing updates. We let \( \hat{k} = \max(1,k/P) \) and quantify the cost of these primitives. When pushing, \( k\text{-RELAXATION} \) takes \( O(\hat{k}) \) time and \( O(\hat{k}) \) work. A \( k\text{-FILTER} \) invocation requires \( O(\log(P) + \hat{k}) \) time and \( O(\min(k,n)) \) work via a prefix sum.

When pushing, the cost of \( k\text{-RELAXATION} \) depends on the PRAM model. In the CRCW-CB model, \( k\text{-RELAXATION} \) takes \( O(\hat{k}) \) time and \( O(\hat{k}) \) work. In the CREW model, \( k\text{-RELAXATION} \) can be processed in \( O(\hat{k}\log(\hat{d})) \) time via binary-tree reductions. To update each vertex of degree \( d \) in the CREW model, we use a binary merge-tree with \( d \) leaves. Over all trees, at most \( k \) of \( m \) leaves contain actual updates.

We can avoid work for all nodes that are the roots of subtrees that do not contain updates, effectively computing a forest of incomplete binary trees with a total of \( k \) leaves and maximum height \( O(\log(\hat{d})) \). Each of \( P \) processors propagates \( k/P \) updates up the complete binary merge-tree associated with its vertices (requiring no setup time) in \( O(\hat{k}\log(\hat{d})) \) time with a total of \( O(\hat{k}(\log(\hat{d})) \) work.

### 4.1 PageRank

PR (Algorithm 1) performs \( O(L) \) steps of power iteration. For each step of power iteration, \( k\text{-RELAXATION} \) is called for \( i \in \{1, \ldots, \hat{d}\} \) with \( \sum_{i=1}^{\hat{d}} \hat{d} = m \). Thus the PRAM complexities of PR are (1) \( O(L(m/P+\hat{d})) \) time and \( O(L(m/P+\hat{d})) \) work using pulling, (2) \( O(L(m/P+\hat{d})) \) time and \( O(L(m/P+\hat{d})) \) work in pushing in CRCW-CB, and (3) \( O(L(m/P+\hat{d})) \) time and \( O(L(m/P+\hat{d})) \) work using pushing in CREW.

Conflicts Pushing/pulling entail \( O(L(m)) \) write/conflicts.

### 4.2 Triangle Counting

TC is shown in Algorithm 2; this is a simple parallelization of the well-known Seedteller strategy [49]. It employs \( k\text{-RELAXATION} \) for \( i \in \{1, \ldots, \hat{d}\} \) with \( \sum_{i=1}^{\hat{d}} \hat{d} = \hat{d} \). Thus the PRAM complexities of TC are (1) \( O(d(m/P+\hat{d})) \) time and \( O(md) \) work using pulling, (2) \( O(d(m/P+\hat{d})) \) time and \( O(md) \) work using pushing in CRCW-CB, and (3) \( O(d(m/P+\hat{d})) \) time and \( O(md) \) work using pushing in CREW. One can leverage more than \( n \) processors to lower the PRAM time-complexity of TC [53].

Conflicts Both variants generate \( O(md) \) read conflicts; pushing also has \( O(md) \) write conflicts.

### 4.3 Breadth-First Search

BFS is shown in Algorithm 3. We define a generalized version of BFS, where vertices enter the frontier only after a given number of neighbors have been in the frontier. The standard BFS is obtained by setting this number to 1, but to use BFS from within BC, we will employ a counter specific to each vertex. The BFS pseudo-code also employs a given accumulation operator to compute values for each
vertex as a function of values of its predecessors in the BFS tree. Our analysis assumes this operator is commutative and associative. The frontier \( F \) is represented as a single array while \( my_F \) is private for each process and contains vertices explored at each iteration. All \( my_F \)s are repeatedly merged into the next \( F \) (Line 8). We let \( f_i \) be the size of \( F \) in the \( i \)th iteration of the while loop.

```
1 /* Input: a graph \( G \), a vertex \( r \), the \( \Delta \) parameter. */
2 function \( \Delta \)-Stepping\((G, r, \Delta)\) { 
3   \( my_F[1..N] = [\infty,\ldots,\infty] \); \( R = R0 \); \( FCV \), such that for each \( e \in E \), \( \text{ready}[e]=0 \); 
4   while (\( f \neq 0 \)) { 
5     explore_my_F(); 
6     for \( w \in N(r) \) do { 
7       \( R[w] = R[r] \); 
8       \( \text{ready}[w] = 0 \); 
9     } 
10   } 
11   return \( my_F \); 
```

Algorithm 3: (§ 4.3) Push- and pull-based \( \Delta \)-Stepping.

The call to \( \text{explore}_{my_F} \) in pulling requires checking all edges, so it takes \( O(m/Pd) \) time and \( O(m) \) work. The call to \( \text{explore}_{my_F} \) in pushing needs \( O(d) \) consecutive \( f_i \)-RELAXATIONS, so it takes \( O(f_{di}) \) time where \( f_i = \max(1,f_{di}/P) \) and work \( O(f_{di}) \) in CRCW-CB (and \( O(\log(d)) \) more in CREW). Second, the merge of frontiers can be done via a \( \hat{\Delta} \)-RELAXATION and, in pushing, a \( \hat{\Delta} \)-FILTER. The \( \hat{\Delta} \)-FILTER is not required in pulling, since we check whether each vertex is in the frontier anyway. In pushing, the merge requires \( O(\log(P) + d_{fi}/P) \) time and \( O(\min(d_{fi},n)) \) work.

Thus, for a graph of diameter \( D \) (with \( D \) while-loop iterations) we derive the total cost using the fact that \( \sum_{i=1}^{D} f_i = n \), obtaining:

1. \( O(D(m/P + d)) \) time and \( O(Dm) \) work in pulling, (2) \( O(D(m/P + D + \log(P))) \) time and \( O(Dm) \) work in pushing in CRCW, and (3) a factor of \( O(D) \) more time and work in the CREW model. It is possible to achieve a lower time-complexity for BFS, especially if willing to sacrifice work-efficiency [23].

Conflicts There are \( O(m) \) write conflicts in pushing; pulling involves \( O(Dm) \) read conflicts.

Atoms/Locks Pushing requires \( O(m) \) CAS atoms.

4.4 \( \Delta \)-Stepping SSSP

The algorithm works in epochs. In each epoch, a bucket \( b \) is initialized with vertices whose tentative distances are \( [(b-1)\Delta, b\Delta) \), and relaxations are computed until all vertices within distance \( b\Delta \) are found. This means that in epoch \( b \), edges are relaxed only from vertices whose final distances are within \( [(b-1)\Delta, b\Delta) \).

Algorithm 4: (§ 4.4) Push- and pull-based \( \Delta \)-Stepping SSSP.

Let \( L \) be the maximum weighted distance between any pair of vertices in the graph, and \( \ell_A \) be the number of iterations done in any epoch. If \( n_i \) vertices fall into the \( i \)th bucket, at the \( i \)th epoch \( O(LD) \) executions of \( n_i \)-RELAXATION will relax edges of vertices in the current bucket and up to \( \ell_A \) executions of \( n_i \)-FILTER will be used to update the set of vertices in the current bucket. So each edge will be relaxed \( O(LD) \) times. There are a total of \( L/\Delta \) epochs, so the complexity of \( \Delta \)-stepping is \( O((L/\Delta)mL + m/\Delta) \) time and \( O((L/\Delta)\ell_A) \) work using pulling, (2) \( O(\ell_AmL/P + (L/\Delta)\ell_A) \) time and \( O(\ell_AmL) \) work using pushing in CRCW-CB, (3) \( O(\log(d)) \) more than (2) using push in CREW. Pushing achieves a smaller cost, since we relax the edges leaving each node in only one of \( L/\Delta \) epochs. These results may be extrapolated to specific types of graphs considered in the original analysis [42].

Conflicts In pushing, there is a write conflict for each of \( O(\ell_AmL) \) edge relaxations. In pulling, there is a read conflict for each of \( O((L/\Delta)mL) \) edge relaxations.

Atoms/Locks In pushing, each edge relaxation can be performed via a CAS atomic (in total \( O(\ell_AmL) \) of these).

4.5 Betweenness Centrality

BC is illustrated in Algorithm 5. For each source vertex, we first compute a BFS to count the multiplicities of each shortest path and store all predecessors that are on some shortest path for each destination vertex. The list of predecessors is then used to define a shortest path tree. To calculate the partial centrality scores, this tree is traversed via BFS starting from the tree leaves. We use the \( \text{ready} \) array to ensure tree-nodes enter the frontier only once the partial centrality updates of all of their children are accumulated.

This algorithm (parallel Brandes) was described in detail [14, 39]. The approach is dominated by \( 2n \) BFS invocations, the cost of which is analyzed in § 4.3. For directed graphs, SSSP (e.g., \( \Delta \)-stepping) must be used to compute each shortest-path tree. Given the shortest-path tree the partial centrality scores can be computed via BFS in the same way as for undirected graphs. Computationally, the most significant difference of BC from SSSP and SSSP is, the presence of additional parallelism. Many source vertices can be processed
independently, so up to $O(n^2)$ processors can be used by running $n$ independent instances of BFS or SSSP.

Conflicts and Atomics/Locks The number of conflicts as well as atomic or lock matches that of BFS or SSSP and can vary by the factor of up to $O(n)$ (depending on the amount of additional parallelism). Yet, since the accumulation operator for the second BFS uses floating point numbers, locks are required instead of atomics. This can be alleviated by maintaining sets of successors instead of predecessors as proposed by Bader et al. [2], which we identify as another opportunity for using either pushing or pulling. We elaborate on it in the technical report.

More Parallelism Our analysis considered parallelism with $P < O(n)$. However, our pseudocodes specify additional potential sources of parallelism in many of the algorithms. Up to $m$ processors can be used in many cases (and even more for TC), but in this scenario, the distinction between pushing and pulling disappears.

4.7 Boruvka Minimum Spanning Tree

Push- and pull-based Boruvka is shown in Algorithm 7. Due to space constraints, it only displays pushing/pulling when selecting the minimum edge adjacent to each suprempt. The algorithm starts with $n$ supernodes and reduces their number by two at every iteration. The supernode connectivity graph can densify throughout the process with supernodes having degree $O(n)$. However, the supernodes will always contain no more than $m$ edges overall. Determining the minimum-weight edge for all supernodes requires $O(n^2/P)$ time and $O(m)$ work assuming each supernode is processed sequentially. Merging the vertices requires $O(log(n))$ time and $O(n)$ work via a tree contraction [24] (our implementation uses a more simplistic approach). Merging the edges connected to each vertex can be done via $O(n)$ invocations of a $k$-relaxation, where $k = O(n)$ at the first iteration and then the bound decreases geometrically. Over all $log(n)$ steps, the complexity of Boruvka is $O((n^2/P)^{log(d)})$ time and $O((n^2/P)^{log(d)})$ work using pushing, $(2)O((n^2/P)^{log(d)})$ time and $O((n^2/P)^{log(d)})$ work using pushing in CRCW-CB, $(3)O(log(d))$ more than $(2)$ using pushing in CREW.

Conflicts Pushing/pulling require $O((Lm)^2)$ write/read conflicts. Atomics/Locks In pushing and pulling the write conflicts can be resolved via CASes (a total of $O((Lm)^2)$ of these).

4.6 Boman Graph Coloring

We present BGC in Algorithm 6. The algorithm proceeds for $L$ iterations, a quantity that is sensitive to both the schedule of threads and the graph structure. To limit the memory consumption, we bounded the maximum count of colors to $C$. We use an opaque function init that partitions $G$ and thus initializes the set of border vertices $B$ and all the partitions $\mathcal{P} = \{P_1, P_2, \ldots\}$. The algorithm alternates between doing sequential graph coloring (seq_color_partition) and adjusting colors of border vertices. The adjustment of colors of bordering vertices corresponds to the invocation of a $\mathcal{D}$-relaxation, in the worst case $[2] = \Theta(n)$. Therefore, the complexity of BGC is $(1)O(L(m/P + d))$ time and $O(Lm)$ work using pulling, $(2)O(L(m/P + d))$ time and $O(Lm)$ work using pushing in CRCW-CB, $(3)O(log(d))$ more than $(2)$ using pushing in CREW.

Conflicts Pushing/pulling require $O((Lm)^2)$ write/read conflicts. Atomics/Locks The write conflicts in pushing can be handled via CAS atoms (in total $O(n^2)$ of them).

Algorithm 5: (§ 4.5) Push- and pull-based Betweenness Centrality.

Algorithm 6: (§ 4.6) Push- and pull-based Boman Graph Coloring.

Algorithm 7: (§ 4.7) Push- and pull-based Boruvka MST.
Directed Graphs} Pushing and pulling differ interestingly for directed graphs. Pushing entails iterating over all outgoing edges of a subset of the vertices, while pulling entails iterating over all incoming edges of all (or most) of the vertices. Thus, instead of $\hat{d}$ some cost bounds would depend on $\hat{d}_{\text{out}}$ and $\hat{d}_{\text{in}}$ for pushing and pulling, respectively; more details are in the technical report.

4.9 Discussion & Insights

We finally summarize the most important insights.

**Write/Read Conflicts** Pushing entails more write conflicts that must be resolved with locks or atomic hints (read conflicts must be resolved only under the EREW model). An exception is BC where the difference lies in the type of the data that causes conflicts (floats for pushing and integers for pulling as was remarked in the past work [39]). Moreover, traversals (BFS, BC (Part 2), SSSP) entail more read conflicts with pulling (e.g., $O(\Delta d)$ in the BFS based on pulling and none in the push-based BFS).

**Atoms/Locks** We now summarize how conflicts translate into used atoms or locks. In many algorithms, pulling atoms or locks completely (TC, PR, BFS, $\Delta$-Stepping, MST). In others (BC), it changes the type of conflicts from $\mathbb{L}$ to $\mathbb{L}'$, enabling the utilization of atoms and removing the need for locks [31].

**Communication/Synchronization** The above analyses show that pulling reduces synchronization compared to pushing (e.g., fewer atoms in TC). In contrast, pushing limits communication (e.g., the number of memory reads in BFS).

**Complexity** Pulling in traversals (BFS, BC, SSSP-$\Delta$) entails more time and work (e.g., see BFS). On the other hand, in schemes such as PR that update all vertices at every iteration, pulling avoid write conflicts. As a result, for PR and TC, pulling is faster than pushing in the PRAM CREW model by a logarithmic factor.

5 ACCELERATING PUSHING & PULLING

Our analysis in § 4 shows that most push- and pull-based algorithms entail excessive counts of atomic locks and reads/writes, respectively. We now describe strategies to reduce both.

**Partition-Awareness (PA, in Pushing)** We first decrease the number of atoms by transforming the graph representation to limit memory conflicts. For this, we partition the adjacency array of each $v$ into two parts: local and remote. The former contains the neighbors $u \in N(v)$ that are owned by $f[v]$ and the latter groups the ones owned by other threads. All local and remote arrays form two contiguous arrays; offsets for each array are stored separately. This increases the representation size from $n + 2m$ to $2n + 2m$ but also enables detecting if a given vertex $v$ is owned by the executing thread (to be updated with a non-atomic) or if it is owned by a different thread (to be updated with an atomic). This strategy can be applied to PR, TC, and BGC. Consider PR as an example. Each iteration has two phases. First, each thread updates its own vertices with non-atoms. Second, threads use atoms to update vertices owned by other threads. Here, the exact number of atoms depends on the graph distribution and structure, and is bounded by 0 (if $V = \emptyset$) and $2m$ (if $V \neq \emptyset$). This former occurs if $G = (V, E)$ is bipartite (i.e., $V = U \cup W, U \cap W = \emptyset$) and each thread only owns vertices from either $U$ or $W$. The latter occurs if each thread owns all vertices in some $G$’s connected component. The number of non-atoms stays similar. We show this example in Algorithm 8. The overhead from a barrier (line 10) is outweighed by fewer write conflicts (none in line 8).

1 //The code below corresponds to lines 12-20 in Algorithm 8.
2 \begin{align*}
3 & \text{if } V \neq \emptyset, \text{\quad \text{(10) }}
4 & \text{then a conflict occurs and either } v \text{ or } u \text{ (depending on the selected strategy) is assigned a color } c_{i+1}, \text{ that was not used before.}
5 & \text{This scheme resembles a BFS traversal with multiple sources selected at the beginning and marked with a color } c_0, \text{ and a frontier constituted by vertices in } F. \text{ In pushing, the vertices in } F \text{ look for their uncolored neighbors and mark them with } c_i. \text{ In pulling, uncolored vertices look for colored neighbors that are in } F.
6
7 \text{Generic-Switch (GS, in Pushing/Pulling) Next, we use the idea of switching between pushing and pulling; we want to not only reduce communication, but also limit the iteration count. We refer to the strategy as Generic-Switch. As an example, consider the above-described BGC enhanced with Frontier-Exploit. Pushing itself results in the excessive number of iterations. This is because, when the number of vertices to be colored is low (our experiments indicate } c < 0.1n), \text{ threads often conflict with each other, requiring more iterations. Switching to pulling may prevent new iterations as no conflicts are generated. Yet, using pulling too early would entail excessive memory accesses (few vertices are colored). Thus, one must carefully select a switching moment or strategy, for example switch if the ratio of the number of the colored vertices to the generated conflicts (in a given iteration) exceeds a certain threshold.}
8
9 \text{Greedy-Switch (GrS, in Pushing/Pulling) Generic-Switch not always brings the desired speedups. For example, BGC with Frontier-Exploit may still need many iterations to color a small fraction of the remaining vertices due to many conflicts between threads that share vertices. In such cases, it is more advantageous to completely switch from a parallel variant (regardless of whether it does pushing or pulling) to an optimized greedy scheme.}

10\text{Generic-Switch (GS, in Pushing/Pulling)}
11\begin{algorithm}
12\text{PART 1: LOCAL UPDATES}
13for $v \in V_1$ do in par
14\text{PART 2: REMOTE UPDATES}
15for $u \in N(v)$ do in par
16\text{new_pr}[v] += $(f_pr[x])/d(u)$
17\text{new_pr}[u] += $(f_pr[x])/d(u)$
18\text{barrier(); //A lightweight barrier to synchronize all threads.}
19for $v \in V_2$ do in par
20for $u \in N(v)$ do in par
21\text{new_pr}[u] += $(f_pr[x])/d(u)$
22\end{algorithm}
Conflict-Removal (CR, Pushing/Pulling) The final strategy (see Algorithm 9) completely removes conflicts in both pushing and pulling. Consider BGC as an example. Instead of solving conflicts over border vertices (the set $B$) in each iteration, one can first use an optimized scheme (e.g., greedy sequential) to color them without any conflicts (thus, this scheme is advantageous if $|B|$ is small compared to $|V|$). The remaining vertices can then be colored in parallel; no conflicts occur either as every $v \in B$ is already colored.

| $L_1$ misses | $orc$ (PR) | $rca$ (PR) | $ljn$ (TC) | $rca$ (TC) | $orc$ (BGC) | $rca$ (BGC) | $pok$ (SSSP-Δ) | $rca$ (SSSP-Δ) |
|--------------|------------|------------|------------|------------|-------------|-------------|----------------|----------------|
| 335M         | 382M       | 572M       | 2.062M     | 10.560M    | 2.857M      | 10.815B     | 10.684B        | 9.290M 4.150M  |
| 243M         | 289M       | 466M       | 6.407M     | 13.077M    | 1.598M      | 7.666M      | 6.459M          | 1.399B 3.879B  |
| 183M         | 213M       | 348M       | 12.21k     | 27.4k      | 21.68k      | 6.644M      | 56.05M          | 8.345k 25.39M  |
| 136M         | 208M       | 664M       | 14.183B    | 36.492M    | 37.989M     | 43.128M     | 42.317M         | 1.049M 4.352M  |
| 118M         | 220M       | 324M       | 22025      | 51.238M    | 38.129M     | 42.918M     | 39.284M         | 1.179M 1.196M  |
| 234M         | 249M       | 535M       | 10564      | 25.246M    | 23.903M     | 32.984M     | 32.054M         | 1.219M 3.678M  |
| 214M         | 220M       | 535M       | 10564      | 25.246M    | 23.903M     | 32.984M     | 32.054M         | 1.219M 3.678M  |
| 214M         | 220M       | 535M       | 10564      | 25.246M    | 23.903M     | 32.984M     | 32.054M         | 1.219M 3.678M  |
| 214M         | 220M       | 535M       | 10564      | 25.246M    | 23.903M     | 32.984M     | 32.054M         | 1.219M 3.678M  |

Table 1: (§ 6.1) PAPI events for PR, BGC (average per iteration), and TC, SSSP-Δ (total count) for the SM setting (Daint, XC30, $T = 16$).

Algorithm 9: (§ 5) Example of Conflict-Removal with BGC.

6 PERFORMANCE ANALYSIS

Finally, we investigate the performance of push/pull variants and the described acceleration strategies. Due to a large amount of data we present and discuss in detail a small representative subset; the remainder is in the report (see the link on page 1).

Selected Benchmarks & Parameters We consider the push- and pull-based variants, strategies from § 5, strong- and weak-scaling, Hyper-Threading (HT), and static/dynamic OpenMP scheduling. Two types of synthetic graphs are used: power-law Kroecker [36] and Erdős-Rényi [21] graphs with $n = 2^{20}, \ldots, 2^{30}$ and $d \in \{2^1, \ldots, 2^{10}\}$. We also use real-world graphs (Table 2) of various sparsities: low $d$ and large $D$ (road networks), low $d$ and $D$ (purchase graphs), and large $d$ with low $D$ (communities). The graphs have up to 268M vertices and 4.28B edges.

| Type            | $|D|$ | $n$ | $m$ | $d$ |
|-----------------|------|----|-----|-----|
| R-MAT graphs    | rmat | 33M | 268M| 4.28B |
| Social networks | src  | 3.07M | 117M | 39  |
| pok             | 1.63M | 22.3M | 18.75 | 11 |
| Ground-truth [6] community | lnn | 3.99M | 34.6M | 6.87 | 17 |
| Purchase network | lnn | 262k | 900k | 3.43 | 32 |
| Road network    | rca  | 1.96M | 2.76M | 1.4 | 849 |

Table 2: (§ 6) The analyzed graphs with skewed degree distributions.

Used Programming Models We use threading to harness SM systems. For DM machines, we use Message Passing (MP, also denoted as Msg-Passing) and Remote Memory Access (RMA) [25]. In MP, processes communicate explicitly and synchronize implicitly with messages [8]. In RMA, processes communicate and synchronize explicitly by accessing remote memories with puts, gets, or atomics, and ensuring consistency with flushes [6, 9, 25].

Counted Events We incorporate the total of nine performance counters for detailed analyses of: cache misses (L1, L2, L3), reads and writes, conditional/unconditional branches, and data/instruction TLB misses. We also manually count issued atomics [50] and acquired locks [50]. Memory operations and cache/TLB misses are important as many graph algorithms are memory-bound [5]. Branches were also shown to impact performance in graph processing [28]. Finally, in distributed settings we count sent/received messages, issued collective operations, and remote reads/writes/atomics.

Experimental Setup and Architectures We use the following systems to cover various types of machines:

- **Cray XC nodes** from the CSCS supercomputing systems. We use XC50 and XC40 nodes from the Piz Daint machine. An XC50 node contains a 12-core Intel Xeon E5-2690 CPU with 64 GiB RAM. Each XC40 node contains an 18-core Intel Xeon E5-2695 CPU with 64 GiB RAM. We also show results for XC30 nodes (with an 8-core Intel E5-2670 Sandy Bridge CPU and 32 GiB RAM) from a past Daint version. Finally, we also provide results for XC40 nodes from a past Piz Dora system (referred to as XC40’); they contained 12-core Intel Haswells E5-2690 and 64 GiB RAM. All nodes are HT-enabled. The interconnection [7] in all the cases is based on Cray’s Aries and it implements the Dragonfly topology [33]. This machines represents massively parallel HPC systems.

- **Trivium V70.05** is a server with Intel Core i7-4770 (with four 3.4 GHz Haswell 2-way multi-threaded cores). Each core has 32 KB of L1 and 256 KB of L2 cache. The CPU has 8 MB of shared L3 cache and 8 GB of RAM. This option represents commodity machines.

Infrastructure and Implementation Details We use the PAPI library (v5.4.1.0) to access performance counters. We spawn one MPI process per core (or per one HT resource if applicable). We use Cray-mpiich (v7.2.2.2) for MP and the foMPI library (v0.2.1) [25] for RMA. We also use OpenMP 4.0 and TBB from the Intel Programming Environment 6.0.3. We compile the code (with the -O3 flag) with g++ v4.9.2 (on Trivium) and Cray GNU 5.2.40 g++ (on CSCS systems). The information refers to the current Daint system; others are covered in the technical report.
6.1 Shared-Memory Analysis

We first analyze the differences in the SM setting. The representative PAPI data for selected schemes is in Table 1. For each scheme, we discuss in more detail the results for graphs with: large $\bar{d}$ and low $D$, and low $\bar{d}$ and large $D$.

**PageRank** PR results can be found in Table 3. In graphs with both high $\bar{d}$ (orc, ljn, poc) and low $\bar{d}$ (rca, am), pulling outperforms pushing by $\approx 3\%$ and $\approx 19\%$, respectively. The former requires no atomics, but its speedup is moderate as it also generates more cache misses and branches as it accesses various neighbors, requiring more random memory reads.

| $G$ | PageRank [ms] | Triangle Counting [s] |
|-----|---------------|-----------------------|
|     | orc pok ljn am rca | orc pok ljn am rca |
| Pushing | 572 129 264 6.68 | 11.78k 139.9 803.5 |
| Pulling | 557 103 240 5.42 | 11.37k 135.3 769.9 |

Table 3: (§ 6.1) Time per iteration for PageRank [ms] and the total time to compute for Triangle Counting [s] (SM setting, Daint, XC30, $T = 16$).

**Triangle Counting** We now proceed to TC (Table 3). Large amounts of time are due to the high computational complexity (§ 4.2); this is especially visible in graphs with high $\bar{d}$. Here, pulling always outperforms pushing (by $\approx 4\%$ for orc and $\approx 2\%$ for rca). This is due to atomics but also more cache misses caused by atomics.

**Graph Coloring** The BGC results are presented in Figure 1. Pulling is always faster than pushing (by $\approx 10\%$ for orc and $\approx 9\%$ for rca for iteration 1). More detailed measurements indicate that the number of locks acquired is the same in both variants, but pushing always entails fewer cache/TLB misses and issued reads and writes.

**\Delta-Stepping** The outcomes for orc and am can be found in Figure 2. Both push and pull variants use locks. Yet, a higher number of memory accesses issued in most iterations in the pull-based scheme limits performance. As expected, the difference decreases after several iterations because the frontier grows (with pushing), requiring more memory accesses. This is especially visible in graphs with high $\bar{d}$ where pulling outperforms pushing (e.g., iteration 6 for orc). Moreover, illustrate in Figure 2c that the larger $\Delta$ is, the smaller the difference between pushing and pulling becomes.

**Breadth-First Search** The results are similar to SSSP-$\Delta$; pushing outperforms pulling in most cases. This is most visible for rca (high $D$, low $\bar{d}$) due to many memory accesses.

6.2 Acceleration Strategies

We now evaluate the acceleration strategies (§ 5).

**Partition-Awareness (PA)** We start with adding PA to PR (Table 6a). In graphs with higher $\bar{d}$ (orc, ljn, poc), pushing+PA outperforms pulling (by $\approx 24\%$). This is because PA decreases atomics (by 7%) and comes with fewer cache misses ($\approx 30\%$ for L1, $\approx 34\%$ for L2, and $\approx 69\%$ for L3) than pulling. In sparser graphs (rca, am), surprisingly pushing+PA is the slowest ($\approx 205\%$ than pushing). This is because fewer atomics issued in pushing+PA are still dominated by more branches ($\approx 23\%$), reads ($\approx 44\%$), and cache misses ($\approx 53\%$ for L3). We conjecture that in graphs with high $\bar{d}$, PA enhances pushing as the latter entails more atomics that dominate the performance. This is visible as both variants reduce the number of cache misses if adjacency lists are long and use better cache prefetchers. Then, for low $\bar{d}$, adjacency lists are short on average, giving more cache misses in pushing+PA and pushing, making pulling the fastest. The worst performance of pushing+PA is due to the synchronization overheads (it splits each iteration into two phases separated by a barrier) that are no longer compensated with more effective cache utilization.

**Frontier-Exploit (FE), Generic/Greedy-Switch (GS/GrS)** We now apply these strategies to BGC, ensuring the same number of colors for each coloring. All three strategies entail very similar ($< 1\%$ of difference) times to compute each iteration. Here, we select GrS and compare it to simple pushing/pulling; see Figure 1. Faster iterations are due to fewer memory accesses as predicted in § 5. Next, we show that the strategies differ in the number of iterations, see Table 6b. The largest iteration count (especially visible for orc/ljn) is due to FE. As predicted, this is because of conflicts. Both switching strategies reduce the iteration count.

6.3 Distributed-Memory Analysis

We also conduct a distributed-memory analysis.

6.3.1 PageRank. First, we use RMA for push- and pull-based PR. The former uses remote atomics ($\text{MPI}_{-}\text{Atomic}$) to modify ranks. The latter reads the ranks with remote gets ($\text{MPI}_{-}\text{Get}$). Next, we design PR with MP. Here, we use the collective $\text{MPI}_{-}\text{Alltoallv}$ [43] to exchange the information on the rank updates among processes. This variant is unusual as it combines pushing and pulling: each
communicated updates must first be placed in designated send buffers. Yet, the used MPI_Accumulate is implemented with costly underlying locking protocol. Next, pulling suffers from communication overheads as it fetches both the degree and the rank of each neighbor of each vertex.

**Memory Consumption** RMA variants only use $O(1)$ storage (per process) in addition to the adjacency list. Contrarily, PR with MP may require up to $O((nd)/P)$ storage (per process) for send and receive buffers.

**6.3.2 Triangle Counting.** Similarly to PR, we develop push- and pull-based TC with RMA and with MP. In pushing, we increase remote counters with an FAA. The MP-based TC uses messages to instruct which counters are augmented. To reduce communication costs, updates are buffered until a given size is reached.

**Performance** The results are in Figure 3. RMA variants always outperform MP; pulling is always faster than pushing ($<1\%$ for orc and $\approx25\%$ for ljn for $P = 48$). This is different from PR as the counters in TC are integer and the utilized RMA library offers fast path codes of remote atomic FAAs that access 64-bit integers. The MP variant is the slowest because of the communication and buffering overheads.

**Memory Consumption** Both RMA schemes fetch $N(v)$ of each analyzed vertex $v$ to check for potential triangles. This is done with multiple MPI_Gets, with two extremes: a single get that fetches all the neighbors, or one get per neighbor. The former requires the largest amount of additional memory ($O(\hat{d})$ storage per process) but least communication overheads. The latter is the opposite.

**6.4 Further Analyses**

We now show that the relative differences between pushing and pulling do not change significantly when varying the used machine. We verify that PR comes with the most relevant difference; see Table 4. Results vary most in denser graphs (orc, pok, ljn); for example pushing outperforms pulling on Trivium while the opposite is true on Dora. Contrarily, the results are similar for rca and am. Thus, the overheads from branches, reads, and cache misses (that are the highest in graphs with lowest $\hat{d}$) dominate performance.

**6.5 Push-Pull Insights**

We finally summarize the most important insights on the push-pull performance for the considered systems.
We now discuss various aspects of push and pull variants.

Various graph algorithms can be expressed with linear algebra (LA) operations such as matrix-vector (MV) multiplication. It enables a concise specification by abstracting from details such as scheduling vertices for processing in the next iteration [32]. We now illustrate that it is possible to frame LA-based graph algorithms in push and pull variants.

**Brief Recap** A crucial notion is the adjacency matrix of $G$ (denoted as $A$) that encodes $G$’s structure. The element in row $i$ and column $j$ of $A$ equals 1 iff there is an edge from vertex $j$ to vertex $i$, and equals 0 otherwise. For simplicity, we focus on unweighted graphs, but our conclusions apply to the weighted case.

The graph algorithms that we consider can be cast as matrix-vector multiplications (MVs) $A \odot x^{(k)}$, where $x^{(k)}$ is the algorithm state in iteration $k$ and $\odot$ is matrix-vector multiplication operator over an appropriate semiring. The adjacency matrix $A$ is generally sparse, while $x^{(k)}$ may or may not be sparse depending on the computation. For example, in PR, each $x^{(k)}$ is dense, while in BFS, the sparsity of $x^{(k)}$ depends on the number of vertices in the $k^{th}$ frontier. We refer to the case when the vector is dense as SpMV, and when the vector is sparse, SpMSpV. The dichotomy between push and pull algorithm variants is mirrored by the dichotomy between the Compressed Sparse Column (CSC) and Compressed Sparse Row (CSR) representations of $A$.

A CSR representation stores each row of $A$ contiguously. The $i^{th}$ row of $A$ contains all vertices with an edge to vertex $i$. Consequently, performing an SpMV in the CSR layout involves iterating over each row and multiplying each nonzero element in the row by appropriate entries of the vector. Thus, each entry of the output can be computed independently by a thread. This scheme is equivalent to pulling updates for each vertex. For SpMV, CSR (pulling) works extremely well, but for SpMSpV, it is not clear how to efficiently exploit the sparsity of the vector $x^{(k)}$.

A CSC representation stores each column of $A$ contiguously. The $i^{th}$ row of $A$ contains all vertices with an edge from vertex $i$. Consequently, performing an SpMV in the CSC layout involves iterating over each column and multiplying each nonzero element in the column by the same entry of the vector, while accumulating to different elements of the output vector. Here, atomics or a reduction tree are necessary to combine updates to each output vector element. This scheme is equivalent to pushing updates from each vertex, as each thread is naturally assigned a different column of $A$ and nonzero entries of $x^{(k)}$. For SpMSpV, CSC (pushing) facilitates exploiting the sparsity of the vector by simply ignoring columns of $A$ that match up to zeros in $x^{(k)}$.

### 7.2 Push-Pull: Programming Models

Push/pull differences depend on the programming model:

- **Threading/RMA** The difference lies in the used atomics. An example is TC: no atomics (pulling) and FAA (pushing).

- **MP (Point-to-Point Messages)** In iterative algorithms with fixed communication patterns (e.g., TC) pushing gives more speedup as pulling increases the message count. In traversals, pushing-pulling switching offers highest performance [4, 17].

- **MP (Collectives)** In collectives such as MPI_Alltoallv, all processes both push and pull the data, eliminating the distinction between these two.

### 7.3 Push-Pull: Code Complexity

Push and pull variants considered in this work come with similar code complexity. Still, pull schemes can be more challenging in achieving high performance. Consider the inner loop in PR where a thread iterates over $N(v)$ of a given $v$. In pushing, updates are conducted simply with atomics. Contrarily, in pulling, one must also fetch the degrees of neighbors. This is similar for other pull variants and poses more challenges in making the code fast.

### 7.4 Push-Pull: Gather-Apply-Scatter

Finally, we discuss the relationship between the push-pull dichotomy and the well-know Gather-Apply-Scatter (GAS) abstraction [27]. In GAS, one develops a graph algorithm by specifying the gather, apply, and scatter functions. They run in parallel for each vertex.
v and respectively: bring some data from v’s neighbors, use it to modify v’s value, and write the result to a data structure. We now describe two algorithms designed with GAS (SSSP and GC) [27] and show how to develop them with pushing or pulling.

SSSP Here, each vertex v is processed in parallel by selecting v’s incident edge e that offers a path to the selected root s with the lowest distance. If it is lower than the current distance from v to s, the value is updated accordingly and N(s) are scheduled for processing in the next iteration. Now, push or pull can be applied when v updates its distance to s. In the former, a neighboring vertex that performed a relaxation in the previous iteration updates its neighbors (pushes the changes) with new distances. In the latter, each vertex scheduled for updates iterates over its neighbors (pulls the updates) to perform a relaxation by itself.

GC Every vertex v collects the set of colors on N(v) to compute a new unique color. Next, the new colors are then scheduled for the color recomputation in the next iteration. This algorithm is a special case of BGC: each vertex constitutes a separate partition (i.e., ∀e∈V∀uc∈N(v)Δ[u] ≠ Δ[v]). Thus, the same approach can be incorporated.

8 RELATED WORK

Push and Pull Algorithm Variants Several graph algorithms that approach the pushing and pulling distinction have been proposed. The bottom-up (pull) BFS was described by Suzumura et al. [55] while Beamer et al. [4] introduced a direction-optimizing BFS that switches between top-down (push) and bottom-up (pull) variants. Madduri et al. [39] proposed several improvements to BC, one of which inverts the direction of modifications in the backward traversal to eliminate critical sections. Whang et al. [60] described pulling and pushing in PR. Finally, Chakaravarthy et al. [17] inverts the direction of message exchanges in the distributed ∆-Stepping algorithm. All these schemes are solutions to single problems. We embrace and generalize them in the push-pull analysis.

Pushing/Pulling in Graph Frameworks Various graph processing frameworks were introduced, for example PBGL [29], Pregel [40], GraphBLAS [41], Galois [34], HAMA [51], PowerGraph [27], GraphLab [37], and Spark [62]. Some use pushing and pulling in certain ways, by: sending and receiving messages (Pregel), using the GAS abstraction (PowerGraph), switching between sparse and dense graph structures (Ligra [52]), switching the direction of updates in a distributed environment (Gemini [65]), using pushing and pulling in 3D task-partitioning [63], or pushing and pulling to/from disk [59]. Yet, none of them comes with an analysis on the push-pull dichotomy, focusing on the framework design. Finally, Doekemeijer et al. [20] list graph processing frameworks that have push- or pull-based communication. Our theoretical analysis and performance observations can serve to help better understand and improve graph processing frameworks.

Accelerating Strategies The Grace framework [45] partitions the graph similarly to Partition-Awareness, but its goal is to reduce caching overheads instead of atomics in pushing. Ligra uses a scheme similar to Generic-Switch as it switches between sparse and dense graph representations [52]. Finally, Salihoglu et al. [47] enhance Pregel-based systems with various schemes. Among others, similarly to Greedy-Switch, they propose to switch from a Pregel-based distributed scheme to a sequential algorithm variant.

Pushing/Pulling outside Graph Processing Borokhovich et al. [12] analyzed gossip algorithms in network coding for information spreading using push, pull, and exchange communication schemes. Swamy et al. [56] designed an asymptotically optimal push-pull method for multicasting over a random network. Intel TBB uses a push-pull protocol in its flow graphs, biasing communication to prevent polling and to reduce unnecessary retries [58]. An analysis of push and pull in software engineering has also been conducted [64]. None of these works addresses graph processing.

9 CONCLUSION

Graph processing has become an important part of various CS research and industry fields, including HPC, systems, networking, and architecture. Its challenges, described by Lumsdaine et al. almost 10 years ago [38], have still not been resolved and accelerating graph computations remains an important goal that must be attained for the ability to process the enormous amounts of data produced today.

In this work, we accelerate graph algorithms by deriving the most advantageous direction of graph updates out of the two options: pushing the updates from the private to the shared state, or pulling the updates in the opposite direction. We illustrate in a detailed analysis that the Push-Pull (PP) dichotomy, namely using either pushing or pulling, can be applied to various algorithms such as triangle counting, minimum spanning tree computations, or graph coloring. We provide detailed specifications, complexity analyses, and performance data from hardware counters on which variant serves best each algorithm and why pushing and pulling differ. These insights can be used to improve various graph processing engines.

Furthermore, we identify that pushing usually suffers from excessive amounts of atomics/locks while pulling entails more memory reads/writes. We use generic strategies to limit the amount of both, accelerating the processing of road networks, citation graphs, social networks, and others.

Our analysis illustrates that the decision on using either pushing or pulling is not limited to merely applying updates in PageRank or sending messages in BFS, but is related to a wide class of algorithms, strategies, graph abstractions, and programming models. Our PP dichotomy can easily be generalized to other concepts related to graph processing, for example vectorization [10].

ACKNOWLEDGMENTS

We thank Hussein Harake, Colin McMurtrie, and the whole CSCS team granting access to the Greina, Piz Dora, and Daint machines, and for their excellent technical support.

REFERENCES

[1] Baruch Awerbuch and Yossi Shiloach. 1987. New connectivity and MSF algorithms for shuffle-exchange network and PRAM. IEEE Trans. on Comp. 36, 10 (1987), 1258–1263.
[2] David A Bader et al. 2007. Approximating betweenness centrality. In Algorithms and Models for the Web-Graph. Springer, 124–137.
[5] David A. Bader and Guojing Cong. 2004. Fast shared-memory algorithms for computing the minimum spanning forest of sparse graphs. In *Par. and Dist. Proc. (IPDPS)*, IEEE, 39.

[4] Scott Beamer, Krste Asanović, and David Patterson. 2013. Direction-optimizing breadth-first search. *Scientific Programming* 21, 3–4 (2013), 137–148.

[3] Venkatesan T. Chakaravarthy et al. 2014. Scalable single source shortest path algorithms for distributed networks. In *Proc. of Intl. Conf. on World Wide Web (WWW)* 2014. 107–117.

[2] Edgar Solomonik, Maciej Besta, Flavio Vella, and Torsten Hoefler. 2017. Multicore triangle computations without reductions. In *Proc. of ACM Symp. on Par. in Alg. and Arch.* 303–314.

[1] Jure Leskovec et al. 2010. Kronecker graphs: An approach to modeling networks. *J. of Machine Learning Research* 11, Feb (2010), 985–1042.

Yucheng Low et al. 2010. GraphLab: A new framework for parallel machine learning. *preprint* arXiv:1006.4990 (2010).

[13] Andrew Lumsdaine, Douglas Gregor, Bruce Hendrickson, and Jonathan W. Berry. 2007. *Challenges in Parallel Graph Processing*. *Proc. Par. Let.* 17, 1 (2007), 5–20.

[14] Kamesh Madduri et al. 2009. A faster parallel algorithm and efficient multi-threaded implementations for evaluating betweenness centrality on massive datasets. In *Par. & Dist. Proc. (IPDPS)*, IEEE Intl. Symp. on -18.

[15] Grzegorz Malewicz et al. 2010. Pregel: a system for large-scale graph processing. In *ACM SIGMOD Intl. Conf. on Manag. of Data (SIGMOD)* '10. 135–146. https://doi.org/10.1145/1807167.1807184

[16] Tim Mattsson et al. 2014. Standards for graph algorithm primitives. *arXiv preprint arXiv:1408.0393* (2014).

[17] Ulrich Meyer and Peter Sanders. 2003. A stepping: a parallelizable shortest path algorithm. *Journal of Algorithms* 49, 1 (2003), 114–152.

[18] MPI Forum. 2012. MPI: A Message-Passing Interface Standard. Version 3. (2012).

[19] Richard C. Murphy et al. 2010. Introducing the graph 500. *Cray User’s Group (CUOG)* (2010).

[20] Vijayan Prabhakaran et al. 2012. Managing Large Graphs on Multi-Cores with Graph Awareness. In *USENIX Annual Technical Conference*, Vol. 12.

[21] Dimitrios Parouzas and Keshav Pingali. 2013. Betweenness centrality: algorithms and implementations. In *ACM SIGPLAN Notices*, Vol. 48. ACM, 55–46.

[22] Semih Salihoglu and Jennifer Widom. 2014. Optimizing graph algorithms on Pregel-like systems. *Proceedings of the VLDB Endowment* 7, 7 (2014), 577–588.

[23] Nadjat Satish et al. 2014. Navigating the maze of graph analytics frameworks using massive graph datasets. In *ACM SIGMOD Intl. Conf. on Man. Data*. 979–990. https://doi.org/10.1145/2588556.2588623

[24] Thomas Schank. 2007. Algorithmic aspects of triangle-based network analysis. Ph.D. Dissertation. University Karlsruhe.

[25] Hermann Schwarz, Maciej Besta, and Torsten Hoefler. 2015. Evaluating the cost of atomic operations on modern architectures. In *2015 International Conference on Parallel Architecture and Compilation (PACT)*, IEEE. 445–456.

[26] Sangwon Seo et al. 2010. HAMA: An Efficient Matrix Computation with the MapReduce Framework. *In Inf. Conf. on Cloud Comp. Tech. and Science (CLOUD-COM)* '10. 721–726. https://doi.org/10.1109/CloudCom.2010.17

[27] Julian Shun and Guy E Blelloch. 2013. Ligra: a lightweight graph processing framework for shared memory. In *ACM SIGPLAN Notices*, Vol. 48. 135–146.

[28] J. Shun and K. Tangwongsan. 2015. Multicore triangle computations without tuning. In *2015 IEEE 31st Intl. Conf. on Data Engineering*. 149–160. https://doi.org/10.1109/ICDE.2015.7113280

[29] Edgar Solomonik, Maciej Besta, Flavio Vella, and Torsten Hoefler. 2017. Scaling betweenness centrality using communication-efficient sparse matrix multiplication. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–14.

[30] Toyotaro Suzumura et al. 2011. Performance characteristics of Graph500 on large-scale distributed environment. *In Workload Char. (ISWC)*, IEEE Intl. Symp. on. 149–158.

[31] Vasuki Narasimha Swamy et al. 2013. An Asymmetrically Optimal Push-Pull Method for Multicasting Over a Random Network. *Inf. Theory, IEEE Tran. on* 59, 1 (2013), 5075–5087.

[32] Adrian Tate, Amir Kamil, Anshu Dubey, Armin Größlinger, Brad Chamberlain, Brice Goglin, Carter Edwards, Chris J Newburn, David Padua, Didem Unat, et al. 2014. Programming abstractions for data locality. (2014).

[33] Michael Voss. Understanding the internals of ibb-graph: Balancing Push and Pull. (????).

[34] Zhigang Wang et al. 2016. Hybrid Pull/Push for I/O-Efficient Distributed and Iterative Graph Computing. In *ACM Intl. Conf. on Man. of Data*. 479–494.

[35] Joyce Jiyoung Whang et al. 2015. Scalable Data-Driven PageRank: Algorithms, System Issues, and Lessons Learned. In *Euro-Par*. Par. Proc. 438–450.

[36] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.

[37] Matei Zaharia et al. 2012. Resilient Distributed Datasets: A Fault-tolerant Abstraction for In-memory Cluster Computing. In *Proc. of the USENIX Conf. on Par. in Alg. and Arch.* 499–512.

[38] Jure Leskovec et al. 2010. Kronecker graphs: An approach to modeling networks. *J. of Machine Learning Research* 11, Feb (2010), 985–1042.

[39] Angela Leggieri and Andrew Lumsdaine. 2005. The parallel BGL: A generic library for distributed graph computations. *Par. Obj.-Or. Scientific Comp. (POOSC)* (2005), 2.

[40] Tim J. Harris. 1994. A survey of PRAM simulation techniques. *ACM Comp. Surv.* 26, 2 (1994), 187–206.

[41] Intel. Inc. 2015. 64 and IA-32 Architectures Software Developer’s Manual. (2015).

[42] Jeremy Keppner and John Gilbert. 2011. Graph algorithms in the language of linear algebra. In *Par. Proc. Let.* 22. 163–177.

[43] John Kim et al. 2008. Technology-Driven, Highly-Scalable Dogenalty: On *Ann. Intl. Symp. on Comp. Arch.* (ISCA ’08). 77–88. https://doi.org/10.1109/ISCA.2008.19

[44] Subhodip Kundu et al. 2007. Optimistic parallelism requires abstractions. In *ACM SIGPLAN Conf. on Proc. Lang. Des. and Impl. (PLDI)* ’07. 211–222. https://doi.org/10.1145/1256734.1250579

[45] Charles E. Leiserson and Tao B. Schardl. 2010. A work-efficient parallel breadth-first search algorithm (or how to cope with the nondeterminism of reducers). In *Proc. of ACM Symp. on Par. in Alg. and Arch.* 303–314.