Pragh: Locality-preserving Graph Traversal with Split Live Migration

Xiating Xie, Xingda Wei, Rong Chen, and Haibo Chen, ***Shanghai Jiao Tong University***

https://www.usenix.org/conference/atc19/presentation/xie

This paper is included in the Proceedings of the 2019 USENIX Annual Technical Conference.

July 10–12, 2019 • Renton, WA, USA

ISBN 978-1-939133-03-8

Open access to the Proceedings of the 2019 USENIX Annual Technical Conference is sponsored by USENIX.

Pragh: Locality-preserving Graph Traversal with Split Live Migration

Xiating Xie, Xingda Wei, Rong Chen, Haibo Chen
Shanghai Key Laboratory of Scalable Computing and Systems
Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University
Contacts: rongchen@sjtu.edu.cn

Abstract

Many real-world data like social, transportation, biology, and communication data can be efficiently modeled as a graph. Hence, graph traversal such as multi-hop or graph-walking queries has been key operations atop graph stores. However, since different graph traversals may touch different sets of data, it is hard or even impossible to have a one-size-fits-all graph partitioning algorithm that preserves access locality for various graph traversal workloads. Meanwhile, prior shard-based migration faces a dilemma such that coarse-grained migration may incur more migration overhead over increased locality benefits, while fine-grained migration usually requires excessive metadata and incurs non-trivial maintenance cost.

This paper proposes Pragh, an efficient locality-preserving live graph migration scheme for graph store in the form of key-value pairs. The key idea of Pragh is a split migration model which only migrates values physically while retains keys in the initial location. This allows fine-grained migration while avoiding the need to maintain excessive metadata. Pragh integrates an RDMA-friendly location cache from DrTM-KV to provide fully-localized accesses to migrated data and further makes a novel reuse of the cache replacement policy for lightweight monitoring. Pragh further supports evolving graphs through a check-and-forward mechanism to resolve the conflict between updates and migration of graph data. Evaluations on an 8-node RDMA-capable cluster using a representative graph traversal benchmark show that Pragh can increase the throughput by up to 19× and decrease the median latency by up to 94%, thanks to split live migration that eliminates 97% remote accesses. A port of split live migration to Wukong with up to 2.53× throughput improvement further confirms the effectiveness and generality of Pragh.

1 Introduction

Graph data ubiquitously exist in a wide range of application domains, including social networks, road maps, biological networks, communication networks, electronic payment, semantic webs, just to name a few examples [47]. Graph traversal (aka multi-hop or graph-walking) queries have been prevalent and important operations atop graph store to support emerging applications like fraud detection in e-commerce transaction [45], user profiling in social networking [11, 18, 6], query answering in knowledge base [52, 63], and urban monitoring in smart city [64].

With the increasing scale of data volume and the growing number of concurrent operations, running graph traversal workloads over distributed graph store becomes essential. Graph traversal workloads are severely sensitive to the access locality, while it is notoriously difficult to partition graph with good locality. For example, the difference of median latency for two-hop query (like friends-of-friends (FoF) [18]) over a Graph500 dataset (RMAT26) [12] is about 30× (0.75ms vs. 22.5ms) between a single machine and an 8-node cluster. Further, preserving locality is even more challenging where workloads and datasets may evolve, while it is common for many production applications [7, 16, 42, 33].

We argue that live migration of graph data is a necessary mechanism for preserving access locality in graph traversals, because existing alternatives have several limitations in many scenarios. First, locality-aware graph partitioning algorithms may improve the performance of a specific dataset and workload [27, 13]. However, *one partition scheme cannot fit all* [62]. Further, a proper graph partitioning scheme for a certain workload may be ineffective and even harmful to another graph traversal workload. Second, replicating data to multiple or all machines allows more (fast) localized read accesses, but also leads to excessive memory overhead as the increase of machines and heavy synchronization cost among replicas for write operations.

Hence, live migration becomes a compelling approach to preserve locality, which has been widely investigated in the database and distributed systems community over the last decade. Unfortunately, the unique characteristics of graph data and traversal operations significantly weaken the benefits of live migration using a shard-based approach, even which is adopted by almost all existing systems. For example,

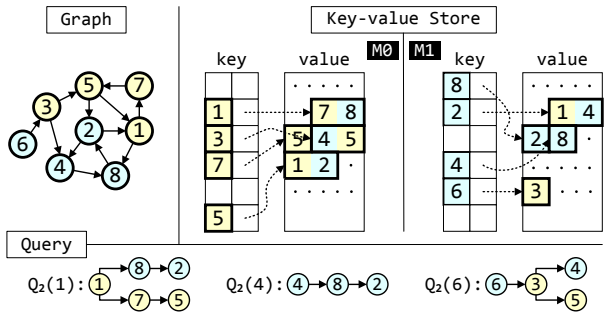

Fig. 1. A sample graph (G), key-value store over 2 machines, and three two-hop queries (Q_2).

using a typical shard-based live migration [25, 60] with an optimal migration plan on above two-hop query experiment will just decrease 29% (22.5ms vs. 15.9ms) median latency, still far from the performance of ideal setting (pure localized access). This is because the majority of the migrated data in a shard would likely have different location preferences. On the other hand, decreasing the size of the shard (fine-grained) would incur high memory and CPU overhead due to storing and maintaining excessive metadata (a location mapping for every shard).

In this paper, we present Pragh, an efficient locality-preserving live migration scheme for distributed in-memory graph store. The key idea of Pragh is a new migration scheme called *split live migration*, which separates the migration of keys and values. Only the value would be migrated physically, while the key would always be stationary at its initial location. This allows fine-grained migration (vertex granularity) while avoiding the need to maintain excessive metadata.

Pragh is made efficient and cost-effective with several key design choices. First, to migrate well-selected vertices (scattered over the entire store) efficiently, Pragh proposes a *unilateral migration protocol* such that the target machine can migrate vertices alone by carefully leveraging one-sided RDMA primitives, while the traversal workloads can concurrently execute on the store. Second, Pragh integrates split live migration with location-based caching [61] to provide *fully-localized* accesses to migrated data. This eliminates the restriction from the stationary key and unleashes the full power of split migration. Third, to support the evolving graph with live migration, Pragh designs a *check-and-forward mechanism* to resolve the conflict between updating and migrating data. Finally, fine-grained monitoring both local and remote accesses to every vertex may incur non-trivial memory and CPU overhead to traversal workloads. Pragh makes a novel reuse of the cache replacement policy to concentrate on tracking remote data accessed frequently. Pragh further provides two optional mechanisms (eager and deferred) for local access tracking to balance the accuracy and the timeliness of migration.

We have implemented Pragh by extending DrTM-KV [61], a state-of-the-art RDMA-enable key-value store, to store graph data and support split live migration. To demonstrate the effectiveness and efficiency of Pragh, we have conducted a set of experiments using a state-of-the-art graph traversal benchmark on an 8-node RDMA-capable cluster. The experimental results show that Pragh can increase the throughput by up to 19× and decrease the median latency by up to 94% through live migration, as the rate of remote accessing reduces from 86.2% to 2.0%. We have also integrated split live migration to Wukong [52], a state-of-the-art distributed graph store that leverages RDMA-based graph exploration (graph traversals in parallel) to provide highly concurrent and low-latency queries. An evaluation using original concurrent workload benchmark [52] shows that the throughput increases by up to 2.53× due to using split live migration.

Table 1: A detail analysis of shard-based live migration.

	Ideal	Shard-based	
		Before	After
Throughput (K ops/sec)	3,248	123	171
Median/50^{th} Latency (msec)	0.75	22.5	15.9
Tail/99^{th} Latency (msec)	4.2	76.6	59.2
Remote Access Rate (%)	0	86.2	64.4
Data Migration Rate (%)	-	-	85.6

2 Background and Motivation

2.1 Graph Store and Traversal Workload

The graph-structured store (aka graph store) becomes more and more prevalent in an increasing number of applications [47] for modeling the relationships among connected data. Due to fast lookup and good scalability, distributed key-value stores are widely used by existing graph systems [52, 64, 57, 22, 31, 24, 51, 63, 54] as the underlying storage layer to support graph traversal operations efficiently, which play a vital role for many emerging and crucial applications [45, 11, 18, 63, 52].

A natural way to build a graph model on top of the key-value store is to simply use the vertex as the key and the adjacency list as the value [51]. Further, separate key and value memory regions are used to support variable-sized key-value pairs [39, 61, 52]. Specifically, the key region is a fix-sized hash table, where each entry stores a key and an address (i.e., offset and size) of the value region. The value region stores variable-sized values consecutively. As shown in Fig. 1, a sample graph (G) is stored into a key-value store over two machines. Various graph traversal operations (like FoF, multi-hop query, and random walking) can be implemented by iteratively accessing key and value pairs. For example, the two-hop query on vertex 1 ($Q_2(1)$) will first retrieve neighbors of the start point (vertex 1) by hashing it as the key and accessing its value (vertex 7 and 8). The next hop will use the value in this hop as the keys (hash(7) and hash(8)) to retrieve their neighbors (vertex 2 and 5) recursively. The accesses over key and value may be either local or remote according to the partitioning scheme.

2.2 Poor Locality and Partitioning

For distributed in-memory stores, the locality of data accessing is quite important because accessing local memory is still more than 20× faster than accessing remote memory across networks, even using high-speed networks [22]. Unfortunately, the traversal on distributed graph data is notoriously slow due to poor data locality. Prior work shows that assigning vertices to N machines randomly will lead to the expected fraction of remote accesses reaching $1-\frac{1}{N}$ [27].

To illustrate the performance impact of locality for graph store, we conducted a motivating experiment using two-hop queries (like FoF [18]) over Graph500 dataset [12] (RMAT26) on an 8-node RDMA-capable cluster. The graph is partitioned into 8 machines randomly (hash-based), and a set of vertices randomly sampled with a Zipf distribution (θ=0.99) is used to run two-hop queries which access fixed 100 friends of 100 friends. As shown in Tab. 1, the distributed setting using 8 machines only achieves less than 4% throughput (123K vs. 3,248K) and about 30× median (50^{th} percentile) latency (22.5ms vs. 0.75ms) of the ideal setting since the rate of remote accessing reaches up to 86.2%.[1]

Therefore, designing locality-aware graph partitioning algorithms has been an active area of research for a decade [27, 13], especially for graph analytics systems. However, *one partition scheme cannot fit all* [62]. It is hard or even impossible to handle dynamic workloads or evolving graphs only relying on static partition-based approaches. One example is shown in Fig. 1 such that $Q_2(1)$ and $Q_2(4)$ contends for the same vertices (8 and 2). The queries may arrive at different times, which causes *false contention*. Actually, prior work on production applications has shown that workloads change rapidly over time [7, 16, 42, 33].

2.3 Live Migration

Live migration (aka dynamic migration) is a compelling approach for handling dynamic workloads and has been widely investigated in the database and distributed systems communities [20, 21, 26, 25, 35, 60, 5]. Generally, a centralized coordinator will make the migration plan according to the statistics (e.g., access frequency) collected by the monitor on each machine. The migration threads on the source and/or target machines will implement the plan by migrating key-value pairs in a synchronous way (see Fig. 2(b)). Since the position of vertices may change after migration, additional metadata (POS) will be accessed to look up the latest positions of the key-value pairs before accessing them (see Fig. 2(a)). The metadata should be updated by the coordinator during live migration and usually is consistently cached at each machine to avoid remote lookup for every accesses.

[1]The ideal result is gained by running the benchmark on a single machine (fully local accessing). The throughput is further magnified 8× (the number of machines).

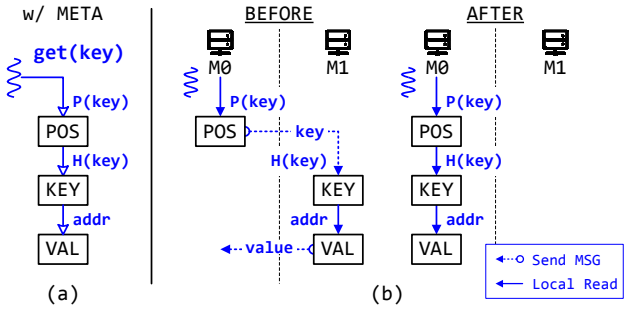

Fig. 2. (a) The sequence of an access on (kv-based) graph store and (b) a comparison of accesses before and after live migration.

Shard-based migration. A ubiquitous approach in live migration is to group the data into *shards* (*partitions*) by key ranges or key hashing [49, 53, 4, 35, 5]. Shards serve as the unit of migration for load balancing and locality-aware optimization. Prior work mainly focuses on relational workloads (e.g., TPC-C) or simple CRUD (Create, Read, Update, and Delete) workloads (e.g., YCSB [15]). Compared to traversing graph data, such workloads with datasets usually have high access locality (e.g., accessing 1% remote key in TPC-C). Consequently, leveraging shard-based migration on the graph store is ineffective and may be harmful, due to the following reasons:

First, migrating data at shard granularity will significantly weaken the benefits of data migration. Due to lacks of data locality, after migrating a shard, the majority of the migrated data in the shard would likely not be accessed by the workload at the target machine. Meanwhile, it will also increase the number of remote accesses at the source machine. Based on the above motivating experiment, we partition graph data into one hundred shards per machine (about 70K keys per shard), similar to prior work [11, 5]. All of the local and remote accesses to every shard are monitored and aggregated to make an optimal migration plan. As shown in Tab. 1, the rate of remote accessing only decrease from 86.2% to 64.4% even after migrating more than 85.6% of graph data (about 20GB). As a result, the throughput only increases 39% (123K vs. 171K) and the median latency also just decreases 29% (22.5ms vs. 15.9ms), still far from the performance of an ideal setting.

Second, though decreasing the size of shard could enhance the effectiveness of migration, it still faces the same drawbacks of static graph partitioning approaches when handling dynamic workloads, unless vertices (key-value pairs) serve as the unit of migration. For example, two irrelevant queries may contend the same shard even assigning two vertices to one shard by key ranges, like vertex 2 and 4 for $Q_2(1)$ and $Q_2(6)$ in Fig. 1. More importantly, the amount of metadata (POS) needed to manage the shards would incur extremely high memory pressure. For example, the metadata for the motivating experiment will consume about 3GB memory on each machine to support vertex granularity migration. Each machine has to cache the entire metadata since the workload

may access any vertex of the graph. Consequently, the size of metadata may exceed the size of graph data when the graph scales.

3 Approach and Overview

Our approach: split live migration. We propose a new migration approach, named *split live migration*, that enables live migration at the minimum level of granularity (i.e., key-value pair). A landmark difference compared to prior approaches is that *split live migration has no need of metadata at all*. This is the greatest advantage but also the biggest challenges for live migration.

The key principle of split migration is to separate the migration of keys and values. The key will always be *stationary* at its initial location, which can be found without metadata (e.g., key hashing). The value will be *migratory* on demand to improve locality or rebalance the load. Our design naturally tackles the issue of memory pressure by avoiding metadata due to the stationary key. Further, allowing fine-grained migration (even a single value) would maximize the effectiveness of data migration for graph store. However, there are still many challenges before making split live migration come true.

Opportunity: RDMA. Remote Direct Memory Access (RDMA) is a networking feature to provide cross-machine accesses with high speed, low latency, and kernel bypassing. The one-sided RDMA primitive (e.g., `READ`, `WRITE`, and `CAS`) allows one machine to directly access the memory of another machine without involving the host CPU. Much prior work has demonstrated the benefit of using RDMA for in-memory key-value stores [39, 22, 32, 61]. Generally, the `Get`/`Put` (read/write) operation first uses `RDMA READ`s to look up the location (address) of the value by hashing the given key, and then use `RDMA READ`/`WRITE` to retrieve/update the value (see the left part of Fig. 4(b)). We observe that *one-sided RDMA primitives decouple the accesses of keys and values, which make it easy and efficient to separate keys and values in physical.* It opens a new opportunity to *split* live migration.

Challenges and solutions. First, split live migration uses the key-value pair as the unit of migration, such that the key-value pairs which will be migrated are scattered over the entire graph store. Therefore, directly using existing protocols designed for shard-based migration may be inefficient. We propose a unilateral (target-only) migration protocol that the target machine can do it alone and efficiently by carefully leveraging one-sided RDMA primitives (§4.1).

Second, the basic split migration only migrates the values of key-value pairs, which can at most eliminate about half of the remote accesses. This is because the read access to the key of key-value pair (look up the location of the value) will still be remote. We address this challenge by integrating split migration with RDMA-friendly location-based caching [61] to provide *fully-localized* access to migrated data (§4.2).

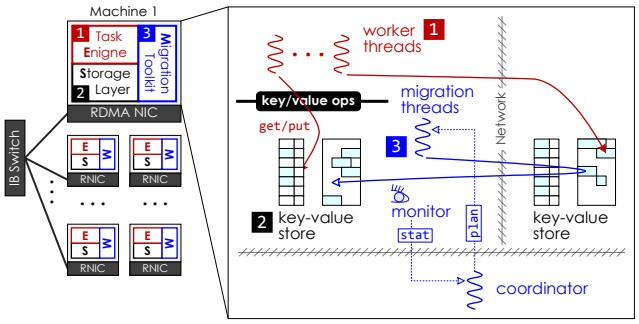

Fig. 3. The architecture of Pragh.

Third, the split of key and value after performing migration presents a new challenge to the support of evolving graphs, especially for the target-only protocol. We use a check-and-forward mechanism to resolve the conflict between data updating and data migrating tasks (§4.3).

Finally, to maximize the effectiveness of data migration, both local and remote accesses to every key-value pair should be tracked to generate an optimal migration plan. It may incur non-trivial memory and CPU overhead to traversal workloads. We design a lightweight, memory-saving monitor, which reuses the location cache to track frequent remote accesses and provides two optional mechanisms for local access tracking to balance the accuracy and the timeliness of live migration (§4.4).

Architecture. As shown in Fig. 3, Pragh is a distributed in-memory graph store with split live migration. It follows a decentralized architecture to deploy servers on a cluster of machines connected with a high-speed, low-latency RDMA network. Each server is composed of three components: task engines, a storage layer, and a migration toolkit. The task engine binds a worker thread on each core with a task queue to continuously execute operations (e.g., `Get` and `Put`) from clients or other servers. The storage layer adopts an RDMA-enabled key-value store over distributed hashtable to support a partitioned global address space. The migration toolkit enables a monitor to collect statistics of graph store and runs migration threads to perform live migration. Pragh scales by partitioning graph data randomly (hash-based) into multiple servers. Each server stores a partition of the graph, which is shared by all of the workers and migration threads on the same machine.

Execution flow. Pragh is designed to handle concurrent operations on graph data with low-latency and high-throughput. The key advantage of Pragh over previous systems is capable of physically migrating data to improve locality in a split way, which can promptly and significantly enhance performance for dynamic workloads.

Similar to prior work [53, 60, 35], a *centralized coordinator* will make migration plan according to the statistics (e.g., access frequency) collected by the monitor on each server and migration policies. The details – how to make a proper policy and how to find an optimal plan – are beyond the

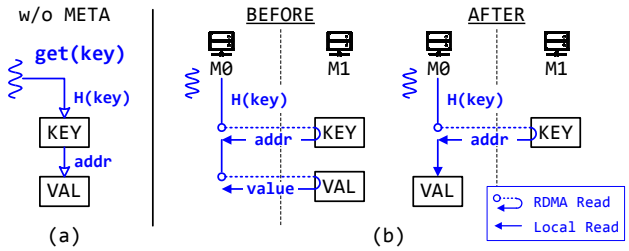

Fig. 4. (a) The sequence of an access on (kv-based) graph store without meta-data and (b) a comparison of accesses before and after split live migration.

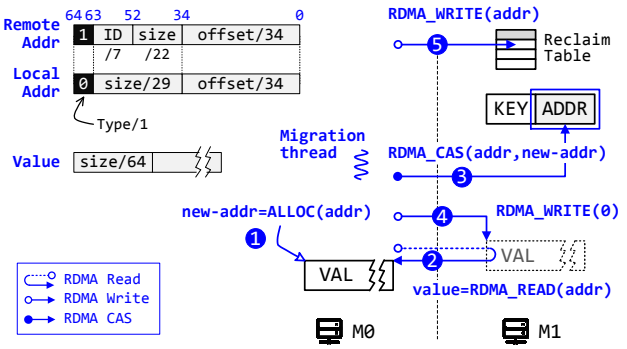

Fig. 5. The execution flow of basic split migration.

scope of this paper and are part of our future work. Currently, Pragh uses a simple threshold-based policy to generate migration plans. On each server, the monitor will track the accesses of worker threads to the key-value store in the background and report to the coordinator periodically (e.g., 10s) or instantly (e.g., when exceeding 100 times per second). The coordinator will compare the statistics from the applicant and the machine hosting the vertex at present, and approve the migration if the profit is more than a threshold (e.g., 50% more accesses per second). After that, the migration threads will migrate the key-value pairs according to the plan from the coordinator, while the worker threads will continue to execute queries by accessing the same key-value store concurrently. Note that the centralized coordinator is just used to collect a few statistics from servers and approve migrations by simply comparing the statistics. Further, the fine-grained approach commonly only needs to migrate much fewer vertices (e.g., 0.13% in §6.1). Hence, the coordinator may hardly become a bottleneck in a medium-sized cluster.

4 Split Live Migration

Pragh uses an RDMA-enabled key-value store over distributed hashtable to store graph data physically. For brevity, Pragh supposes that each vertex has a unique ID (vid) and use it as the key. The hash value of the key (`H(key)`) can be used to identify the host machine (`mid`) and the location in the key region (`off`). As shown in Fig. 4(a), to get neighbors of a given vertex, the worker thread first uses `H(key)` to look up the address of its value and then retrieves the value (a list of IDs of neighbors). For remote key-value pairs, `RDMA READ`s are used to access keys and values (see the left part of Fig. 4(b)), which are at least 20× slower than local reads. Hence, Pragh uses split live migration to eliminate such remote accesses.

4.1 Basic Split Migration

We start from the basic migration protocol, assuming that there only exist traversal workloads (i.e., `GET` operations) in the graph store. Since the key is always stationary in the split migration, Pragh will only move the value to the target machine. This could improve locality by avoiding remote accesses to the values (see the right part of Fig. 4(b)).

Address layout. To avoid the influence between accessing and migrating key-value pairs, the address (within the key) should be changed from local to remote in a lock-free way (e.g., compare-and-swap (CAS)). Therefore, both the local and remote location of value should use a 64-bit address uniformly, which can be modified atomically using both local and RDMA atomic instructions.[2]

Considering the machine ID should be added into the address, a simple layout may severely restrict the scope of address space. Pragh adopts a differentiated layout for local and remote addresses. As shown in the top left corner of Fig. 5, The most significant bit is used to present the type of address, local (0) or remote (1). For local addresses, the rest of the bits are used to store 29-bit value size and 34-bit offset within the value region. Thus, the size of a single value and value region on a single machine can reach 4GB and 128GB respectively (assuming 8-byte granularity and alignment). For remote addresses, the value offset still occupy 34 bits to present the entire remote value region, while the value size reduces to 22 bits for hosting 7-bit machine ID. Thus, the graph store can scale up to 128 machines, while the size of the maximum value that can be migrated is limited to 32MB. The observation is that the system will prefer to migrate the workloads rather than very large key-value pairs [52, 57]. Further, a large key-value pair can be split into multiple ones (vertex decomposition [52, 57]), and each one can be migrated separately.

Unilateral migration protocol. Similar to traditional shard-based migration systems, the split live migration also could be implemented by the collaboration of migration threads on source and target machines. However, the key-value pairs which will be migrated, are scattered over the entire graph store due to lacks of locality. It means that migrating multiple key-value pairs may incur a prolonged interruption to the concurrent graph accessing and/or lengthy migration delay since multiple addresses (within separated keys) should be modified by atomic operations (e.g., CAS).

Pragh proposes a unilateral (target-only) migration protocol based on one-sided RDMA primitives. It only uses the

[2]Note that RDMA primitives guarantee atomic 64-bit transfer [9], and `RDMA READ/WRITE` operations are also cache coherent with local accesses [22, 61].

```
   MIGRATE(key)
   1 retry:
   2   kmid = H(key).mid          ▸ e.g., key % machines
   3   addr = LOOKUP(kmid, key)
❶ 4   buf = ALLOC(addr.sz)
   5   new_addr = {1, local_mid, addr.sz, buf}
❷ 6   RDMA_READ(addr.mid, buf, addr.off, addr.sz)
❸ 7   if !RDMA_CAS(kmid, H(key).off, addr, new_addr)
   8   | goto retry                ▸ conflict w/ PUT
   9   zero = 0                    ▸ invalidate value
❹ 10  RDMA_WRITE(addr.mid, addr.off, zero, 8)
❺ 11  RDMA_WRITE(addr.mid, reclaim, addr, 8)  ▸ reclaim
```

Fig. 6. Pseudo-code of MIGRATE operation.[3]

migration thread on the target machine to migrate the key-value pair instantly, while the worker threads on every machine can still access the key-value pair concurrently. Fig. 5 illustrates three steps of the migration protocol (a detail pseudo-code is shown in Fig. 6). First, the migration thread on the target machine will allocate memory space in local value region (new_addr) to host migrated value (❶), according to the size in the original address. Second, the migration thread will retrieve the value using one RDMA READ from the original address to the new address (❷). Finally, one RDMA CAS is used to replace the original (local) address with the new (remote) address (❸).

Invalidation and reclaim. Unilateral migration protocol will incur two new problems. First, the memory of migrated value in the source machine should be invalidated. However, some worker threads may still have the original address of the migrated value and will access it in the future. To solve it, Pragh proposes a passive invalidation mechanism. The migration thread will invalidate the original memory of migrated value by zeroing (RDMA WRITE) the size within the value (❹). Before using the retrieved value, the worker thread should check whether the size within the value and address are equal. If not, the address should be regained. Note that the worker thread can safely read the value from the original memory before invalidation even it has just been migrated (❸).

Second, the memory of migrated value on the source machine should be reclaimed. However, it is hard or even impossible for the migration thread on the target machine to solely free the memory. Therefore, Pragh uses a lease-based mechanism to reclaim the memory of migrated values in the background by a garbage collection (GC) thread on the source machine.[4] The migration thread will actively write (RDMA WRITE) the original address to the reclaim table[5] of the source machine, at the end of live migration (❺). The GC thread on each machine will periodically check the reclaim table to free the expired memory, which has been migrated

[3]RDMA provides fences between different requests [38], and Pragh uses them before invalidating and reclaiming the memory (Line 10 and 11).
[4]Pragh uses the precision time protocol (PTP) [1] to implement lease.
[5]We implement the reclaim table like the circular buffer [22].

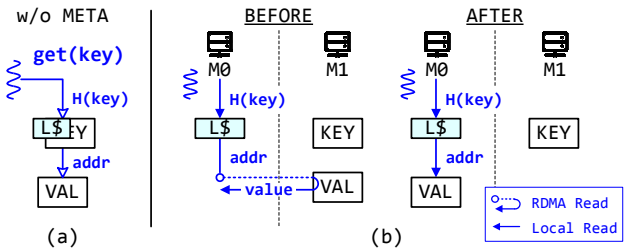

Fig. 7. (a) The sequence of an access on (kv-based) graph store with location cache and (b) a comparison of accesses before and after split live migration with location cache (for remote kv pair).

before a pre-agreed lease (e.g., 60s). All the worker threads comply with the convention that the value address obtained before a lease duration should not be used, since it may have been freed and reused. The performance impact could be trivial by using a long-term lease.

4.2 Fully-localized Split Migration

The basic split migration only avoids remote accesses to the values, which limits the effects of migration since only at most half of remote accesses can be eliminated.

Observation: location cache. Prior work [61, 23, 59] proposes location-based caching for RDMA-friendly key-value stores, which aims at avoiding remote accesses to the keys. Different to caching the content (value) of key-value pairs, the location cache (L$) only stores the location (address) of key-value pairs, which is very space-efficient and effective (see the left part of Fig. 7). We observe that *location cache is a perfect counterpart to split migration*. They focus on two different halves of the access to the remote key-value pair, and the candidates of them are also well matched, namely remote key-value pairs frequently accessed. Finally, a small cache has negligible memory overhead (e.g., 128MB) and lookup cost, yet it is sufficient to achieve fully-localized accesses for most workloads [46, 61].

Integration with location cache. Pragh extends the graph store with location cache (L$) and integrates it with split live migration to enable fully localized accesses after migration. Fig. 8 illustrates the pseudo-code of GET operation with the integration of location cache and split live migration. When accessing a remote key-value pair (Line 9), the worker thread will first check location cache (Line 21) and fill the cache (if missed) with the address of the value (Line 25) obtained by the remote access to the key (Line 24). Given the address, the worker thread will retrieve the value using one RDMA READ (Line 14).

If the worker threads access the key-value pair frequently enough, the value will be migrated to the local using the basic migration protocol. After that, the address stored in location cache will be updated by the new address, which points to the local value region. Therefore, the accesses to the key-value pair will be fully localized (Line 10-12), as shown in the right part of Fig. 7. In contrast, the local key-value pair could also

```
 GET(key, buf)
+1 retry:
 2   kmid = H(key).mid          ▸ e.g., key % machines
 3   addr = LOOKUP(kmid, key)
 4   if kmid == local_mid                  ▸ local key
+5   |  if addr.type == 0                  ▸ local value
 6   |  |  MEMCPY(buf, vals[addr.off], addr.sz)
+7   |  else                  ▸ remote value (migrated)
+8   |  |  RDMA_READ(addr.mid, buf, addr.off, addr.sz)
 9   else                                  ▸ remote key
+10  |  if addr.type == 1                   ▸ migrated
+11  |  |   && addr.mid == local_mid       ▸ local value
+12  |  |  MEMCPY(buf, vals[addr.off], addr.sz)
+13  |  else                               ▸ remote value
 14  |  |  RDMA_READ(addr.mid, buf, addr.off, addr.sz)
+15  if CHECK(addr, buf)
+16  |  if kmid != local_mid
+17  |  |  cache.DELETE(key)                ▸ invalidate
+18  |  goto retry

 LOOKUP(kmid, key)
 19  if (kmid == local_mid)                 ▸ local key
 20  |   return keys[H(key).off]
x21  if cache.FIND(key)
+22  |   && !EXPIRED(cache.GET(key).lease)
x23  |  return cache.GET(key).addr          ▸ cache hit
 24  RDMA_READ(kmid, addr, H(key).off, 8)
x25  cache.INSERT(key, addr)               ▸ fill cache
+26  cache.GET(key).lease = NOW()
 27  return addr
```

Fig. 8. Pseudo-code of GET operation with location cache. The code lines with "x" and "+" stand for additional instructions to integrate with location cache and split live migration, respectively.

be migrated to other machines, thus the type of address will be used to decide how to retrieve the value (Line 5-8).

Finally, the address stored in the location cache should also follow the convention of the invalidation and the reclaim mechanisms. First, if the retrieved value is invalid (Line 15), the worker thread has to delete the address in location cache for the remote key-value pair (Line 16-17), and needs to retry (Line 18). Second, the cached address must expire after a lease duration (e.g., 60s) from the last cache time (Line 22 and 26). Note that the duration of the (cache) lease should be equal or smaller than that of the (reclaim) lease (§4.1).

4.3 Full-fledged Split Migration

The basic migration protocol only considers traversal workloads (i.e., GET operations) concurrently execute in the graph store. Pragh extends it with a check-and-forward mechanism to support the evolving graph (i.e., PUT operations). For brevity, suppose that graph store has provided some mechanisms (e.g., snapshot read [52, 64]) to run traversal workloads over evolving graphs correctly.[6] Therefore, Pragh only tackles the conflict between split live migration and the change of graph. More specifically, Pragh only needs to consider the concurrent update to edges (i.e., change the value of a key-value pair).

[6]Pragh assumes the PUT operation will use atomic in-place updates on the key to ensure consistency, which is common in prior work [52, 64].

```
 PUT(key, val)
+1 retry:
 2   kmid = H(key).mid
 3   addr = LOOKUP(kmid, key)
+4   if addr.mid != local_mid              ▸ migrated
+5   |  SEND(addr.mid, key, val)          ▸ forward PUT op
+6   |  return false
 7   new_addr = WRITE_VALUE(addr, val)
 8   if !RDMA_CAS(kmid, H(key).off, addr, new_addr)
 9   |  goto retry          ▸ conflict w/ put or migrate
 10  zero = 0                          ▸ invalidate value
 11  MEMCPY(vals[addr.off], zero, 8)
 12  MEMCPY(reclaim, addr, 8)                ▸ reclaim
 13  return true
```

Fig. 9. Pseudo-code of PUT operation. The code lines with "+" stand for additional instructions to support split live migration.

We observe that both MIGRATE and PUT operations will change the address within the key atomically to mark the success of processing (Line 7 in Fig. 6 and Line 8 in Fig. 9).[7] Moreover, PUT operation will always be assigned to the machine hosting the key at first. So for key-value pairs migrated, a better choice is to forward the PUT operation to the machine hosting the value upon conflicts, which also ensures consistency and reclaims the memory. Consequently, Pragh adopts different strategies for MIGRATE and PUT operations when detecting the conflict over the address; MIGRATE operation will be retried (Line 8 in Fig. 6), while PUT operation will forward itself (Line 5 in Fig. 9), if it conflicts with some MIGRATE operation and then is retried (Line 9 in Fig. 9). Note that PUT operation will always update the address in the machine hosting the key using RDMA CAS, even though PUT operation is forwarded.

4.4 Lightweight Monitoring

To generate a proper migration plan, the coordinator should collect the statistics of both local and remote accesses to every key-value pair. A (much) higher remote access number from a certain machine to a key-value pair in the most recent interval (e.g., 10s) indicates that migrating the key-value pair to that machine may improve locality (fewer remote accesses). It has been a great challenge to track the accesses at the granularity of key-value pairs.[8] Even worse, the remote accesses using RDMA READ contributes much more extra burdens (both memory and CPU overhead) to the monitor, since each machine has to track the accesses to remote key-value pairs (except local key-value pairs).

Pragh designs a lightweight, memory-saving monitor for split live migration by tracking local and remote accesses separately. For remote accesses, worker threads may access

[7]Suppose that PUT operation will change the size or the offset of the address (or both), namely addr is not equal to new_addr.

[8]Relational database can leverage table schema to reduce the number of tuples should be tracked, by grouping co-accessed tuples into blocks [53, 25, 50, 60]. Unfortunately, graph store is generally schema-less.

any key-value pairs, while the monitor may (very likely) only care about remote key-value pairs *accessed frequently*. This observation also matches the intention of the location cache. Hence, Pragh reuses the cache to track (partial) remote accesses(remote key). The monitor relies on the replacement policy of cache to recognize the key-value pairs (worth tracking) freely. Note that the accesses for the values migrated to local will still be tracked through the cache.

For local accesses, reserving space for every key and tracking every access might be not worth, especially for a very large store. This is because only a small fraction of key-value pairs should be migrated for a while. For example, migrating less than 0.2% of key-value pairs is sufficient for the motivating experiment (§6.1). Therefore, Pragh allows to skip tracking local accesses to the key-value pairs and provides two optional mechanisms to balance the timeliness and the accuracy of split live migration. Note that the monitor on each machine will report to the coordinator when remote accesses to a key-value pair exceed a threshold.

Eager migration: The coordinator will eagerly approve the migration of the key-value pair. After migration, the machine hosting the key will track the (remote) accesses to the key-value pair using a separate table, and then may migrate it back in future if it accesses the key-value pair more frequently.[9]

Deferred migration: The coordinator will notify the machine hosting the key to track the (local) accesses to the key-value pair using a separate table. After a migration interval, the coordinator will decide whether to migrate the key-value pair according to the statistics from all of the machines.

4.5 Discussion

Even though the current design of split live migration highly relies on RDMA, we believe that it can still benefit graph traversal workloads without RDMA, including no need for metadata and vertex granularity migration. However, after migrating the value to local, the cost to retrieve the address would be almost the same as the cost to retrieve the value directly. Hence, location cache must be deployed even without RDMA. On the other hand, the lack of RDMA would also need to rethink the implementation of migration protocol. Our future work may extend Pragh to support commodity networks without RDMA.

5 Implementation

Fault tolerance. Pragh supposes distributed in-memory graph store has provided durability and/or availability by using specific mechanisms like checkpointing or replication. Pragh only needs to consider the interrupted migration tasks and the recovery of crashed machines, because split live migration only changes the location of key-value pairs rather than the content of key-value pairs.

Interrupted migration tasks: If the crashed machine is the source of migration, there is nothing to do since the key-value pair will be recovered on the crashed machine later. If the crashed machine is the target of migration, a corner case that the interruption occurs after replacing address (❸ in Fig. 5) but before reclaiming memory (❺ in Fig. 5) will cause a little memory leakage, which can be detected and reclaimed by scanning the entire value memory region in background.

System recovery: Pragh relies on the mechanism provided by graph store to detect machine failures, like Zookeeper [30]. It will notify surviving machines to assist the recovery of crashed machines, which needs to handle two kinds of key-value pairs. First, the key-value pairs hosted by a crashed machine will be reloaded by the substitute of the crashed machine. Before that, all surviving machines will flush addresses in location cache which point to the key-value pairs hosted by crashed machines (i.e., `H(key).mid`) whether they have been migrated or not, and reclaim the memory of values migrated from crashed machines. Second, the key-value pairs, hosted by a surviving machine but migrated to a crashed machine, will be reloaded by the surviving machine. Before that, all surviving machines will also flush addresses in location cache which point to the key-value pairs migrated to the crashed machines (i.e., `addr.mid`). The coordinator will record the latest target machines of values migrated persistently before approving the migration, which could help surviving machines reload vertices precisely. Moreover, all workloads running on surviving machines involving crash machines will be aborted and suspended until recovery is complete. Finally, the coordinator in Pragh is stateless and easy to recover. The coordinator failure will not influence the execution of worker threads and only pause launching new migration tasks and recovering crashed machines. The migration thread can continue to complete the outstanding migrations.

Optimizations. Pragh adopts a unilateral migration protocol (see §4.1) to migrate one key-value pair (vertex) at a time, which requires *at most* five one-sided RDMA operations: two `READ`s to lookup and retrieve the value, one `CAS` to change the address atomically, and two `WRITE`s to invalidate and reclaim the original memory. Though this approach can provide instant response to migration demands and fully bypass the CPU and kernel of source machine, the throughput of migration may be bottlenecked by the network due to too many RDMA operations with small payloads.

To remedy it, Pragh enables three optimizations to further accelerate split migration. First, in most cases, the migrated key-value pair is frequently accessed by the target machine; thus, its address (very likely) has been already cached in the location cache. It means that the migration thread can skip the first `RDMA READ` to look up the address. Second, Pragh will

[9] Pragh relies on the coordinator to prevent the "ping-pong" of migrations, which prefers not to migrate the vertex competed by multiple machines.

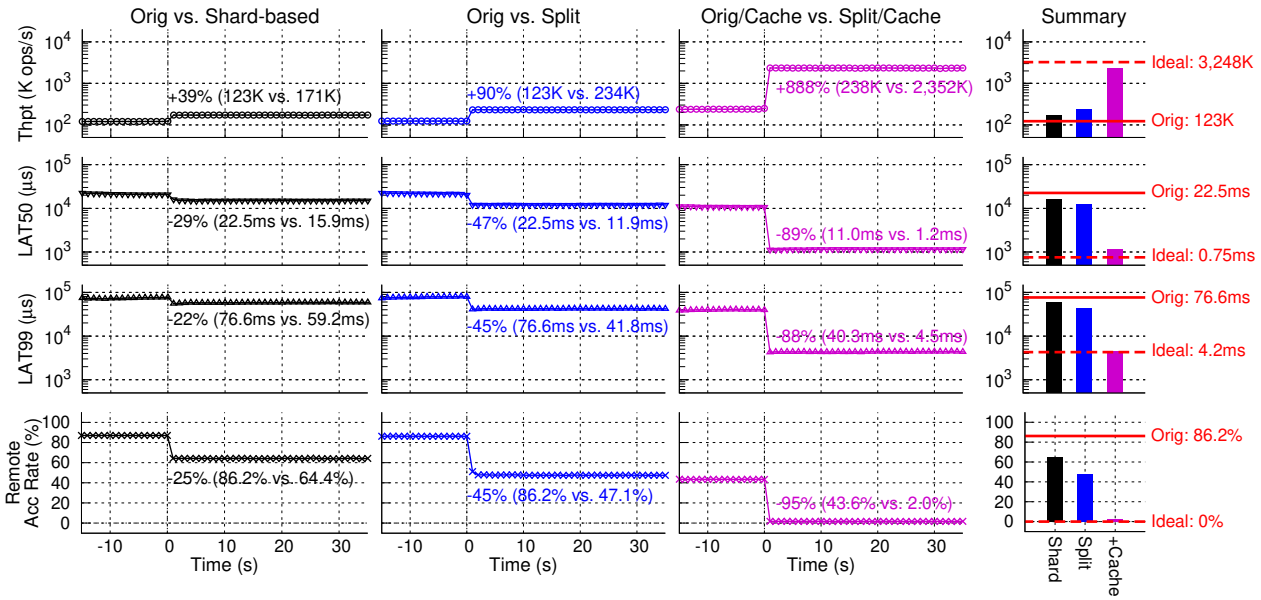

Fig. 10. *A comparison of migration schemes on the traversal benchmark with a skewed workload (a Zipf distribution with $\theta = 0.99$).*

migrate multiple key-value pairs concurrently in a pipelined fashion to better utilize network bandwidth. Each RDMA operation to migrate one key-value pair is implemented as one stage, and Pragh schedules these stages without waiting for the request completion. Finally, since the memory invalidation and reclaim are not on the critical path to migrate one key-value pair[10], Pragh enables passive ACK [59] to acknowledge the completion of such two `RDMA WRITE`s passively, which further reduces the network bandwidth. As a result, a single migration thread is sufficient to migrate more than one million vertices per second (§6).

Load balance. Though Pragh mainly focuses on using live migration to improve the locality of graph traversal workloads, it also can be used to rebalance load across machines, similar to prior work [53, 60, 35]. Basically, it all depends on the migration plan generated by the coordinator. Generally, the traversal workload will be sent to the machine hosting the initial vertex and run to completion. The remote key-value pairs will be retrieved by RDMA operations. Therefore, the coordinator should recognize such hotspots and generate proper plans to scatter them over all of the machines using live migration, like Pragh. Meanwhile, different goals also need different migration policies and statistics. It is orthogonal to the design of Pragh and beyond the scope of this paper.

6 Evaluations

Hardware configuration. All evaluations were conducted on a rack-scale cluster with 8 nodes. Each node has two 12-core Intel Xeon E5-2650 v4 processors and 128GB DRAM. Each node is equipped with two ConnectX-4 MCX455A 100Gbps InfiniBand NIC via PCIe 3.0 x16 connected to a Mellanox SB7890 100Gbps IB Switch, and an Intel X540 10GbE NIC connected to a Force10 S4810P 10GbE Switch. In all experiments, we reserve four cores on each CPU to generate requests to avoid the impact of networking between clients and servers as done in prior work [56, 58, 61, 14, 60, 52]. All experimental results are the average of five runs.

Traversal benchmark. Inspired by YCSB [15], we build a simple benchmark to evaluate the effectiveness of different migration approaches for graph traversal workloads. The traversal benchmark uses a synthetic graph provided by Graph500 [12]. In this paper, the graph with 2^{26} vertices and 2^{30} edges (RMAT26) is used as default dataset since we need to run the benchmark on a single machine to gain the performance of ideal setting (pure localized access). Note that the experimental results on larger graphs (e.g., RMAT29) are similar. The traversal benchmark consists of 95% two-hop queries (`GET`) and 5% edge updates/inserts (`PUT`), similar to YCSB-B (read-heavy) [15]. Note that the majority of many traversal workloads [11] are two-hop queries, and it is easy to compose other complicated queries like SPARQL query [52]. The starting vertices of two-hop queries are chosen according to a Zipf distribution with $\theta = 0.99$. The scope of starting vertices and the number of neighboring vertices retrieved could be configured. The default values are 2^{10} and 100, respectively. We will compare the performance impact with different settings in separate experiments.

Comparing targets. The following five results are provided in the evaluation of the traversal benchmark. ***Orig*** indicates the performance of running the benchmark over the graph data partitioned randomly and without data migration. ***Ideal*** is the result gained by running the benchmark on a single machine. Specifically, throughput is simply magnified by

[10]To ensure consistency, the (original) memory invalidation must be completed before the next `PUT` operation on (new) memory starts, which is easy to implement with the check-and-forward mechanism (§4.3).

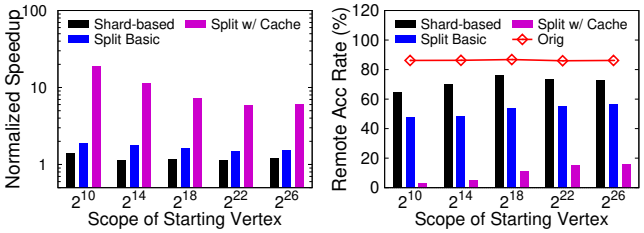

Fig. 11. *A comparison of migration benefits for different approaches with the increase of scopes of starting vertices.*

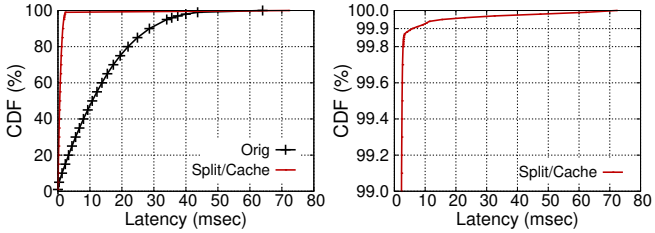

Fig. 12. *The CDF graph of latency for* Put *operations.*

the number of machines (i.e., 8×). ***Shard-based*** represents the performance of a shard-based migration approach, which deploys one hundred shards at each machine, similar to prior work [11, 5]. Note that we always generate optimal migration plans for shard-based migration by tracking every access but do not consider the tracking cost. ***Split/Cache*** and ***Split*** are the performance of Pragh using split live migration with and without location cache. The size of the location cache is 128MB. The migration plan is built by the statistics collected by our lightweight monitor. The default interval is set to 10 seconds.

6.1 Migration Benefits

To study the benefits of migration approaches, we run the traversal benchmark using different migration schemes and compare to the result of the original and ideal settings. As shown in Fig. 10, the original throughput and latency are about 26× slower than the ideal results (123K vs. 3,248K) since about 86.2% accesses to the key-value store are remote. Shard-based approach can only increase the throughput by 39% (123K vs. 171K) and decrease the median (50^{th} percentile) latency by 29% (22.5ms vs. 15.9ms) as it just removes about 25% remote accesses. Pragh can almost double the throughput and reduce the latency by half, thanks to the basic split migration, which removes nearly all remote accesses to the values. Using location cache can remove almost all of the remote accesses to the keys, as the cache hit rate is about 99%. Note that the performance of enabling basic split migration or location cache alone are similar, because both of them still need one `RDMA READ` to retrieve the remote key or value separately.

When combining two techniques, the throughput of Split/Cache can reach 2,352K queries per second (19× compare to Orig). It has achieved close to 72% of ideal performance. The remaining 2.0% of remote accesses is due to the competition on vertices shared by multiple queries running on different machines. Note that traditional migration scheme is hard to integrate with location cache, since they will migrate both keys and values physically and make location cache useless.

Migration time and network traffic. Both split migration and shard-based migration can complete migration in seconds since we optimize the data transmissions in both methods. For shard-based migration, we migrate the shards in block granularity to fully utilize network bandwidth. However, split migration is still faster even using more network round-trips for one key. This is because fine-grained migration migrates much less data than coarse-grained, shard-based migration. For this experiment, only 78,242 keys (0.13% of the total vertices) are migrated in split migration, where 782MB data is migrated in total. For comparison, 85.6% of shards (685 out of 800) are migrated in shard-based migration with a total size of 20GB key-value to transfer.

Scope of starting vertices. Generally, the query will start from a certain type of vertices (e.g., users or tweets in social networks), and the size of the subset of vertices may be various. Fig. 11 further presents the impact of using different scopes of starting vertices in the traversal benchmark from 2^{10} to 2^{26}. The speedup after migration decreases with the increase of scope steadily due to the increase of contention on key-value pairs accessed by multiple queries. It will also result in the rise of remote accesses (see Fig. 11(b)).

Impact on Put operations. To reveal the impact of check-and-forward mechanism in Pragh on the latency of Put operations, we use an update-heavy traversal benchmark, which consists of 50% two-hop queries (Get) and 50% edge updates/inserts (Put), similar to YCSB-A [15]. Fig. 12(a) shows the CDF graph of latency for Put operations with and without split live migration. After migration, the latency of 99.9% Put operations decreases significantly, thanks to the decline of waiting time in the queue. Moreover, as shown in Fig. 12(b), the check-and-forward mechanism will just impact the 99.9^{th} percentile latency, since about 0.11% Put operations updates migrated key-value pairs and is forwarded to another machine. Note that Pragh only migrates 0.13% of total key-value pairs. The increase of latency is mainly contributed by the extra cost for forwarding the operation, waiting in the queue, and re-executing the operation.

Uniform workload. We also evaluate the traversal benchmark with a uniform workload. Shard-based approach can hardly gain benefits and only increases the throughput by 8% (126K vs. 136K) after migration, as the remote access rate just drops from 85% to 82%. In contrast, the basic split migration eliminates over 43% of remote accesses and increases the throughput by 84% (126K vs. 232K). By using location cache, the throughput of Split/Cache can reach 1,521K queries per second (12× compared to Orig). The remote access rate reduces to 5%.

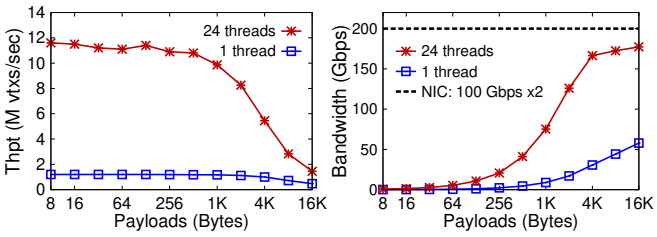

Fig. 13. *The throughput and bandwidth of unilateral migration using 1 and 24 threads.*

6.2 Migration Speed

To evaluate the capability of unilateral migration protocol, we conduct an experiment to migrate values from a remote machine to local with full speed. Fig. 13 shows the throughput of migration and network bandwidth consumed with the increase of payload (i.e., value) size. A single thread is enough to migrate values for millions of vertices per second with less than 4KB payloads. Using parallel migration with 24 threads can further increase the throughput of moving values to more than 10 million per second. Further, using multiple RDMA primitives to migrate a single value will not be limited by network. It should be noted that split live migration will only use the CPU of the target machine.

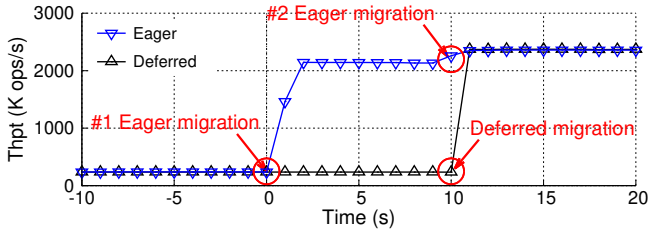

Fig. 14. *The throughput timeline for split live migration (w/ Cache) using eager or deferred mechanism.*

6.3 Eager Migration vs. Deferred Migration

Pragh provides two optional migration mechanisms, eager and deferred, to balance the accuracy and the timeliness of live migration. Fig. 14 compares these two mechanisms using the traversal benchmark. The monitor on each machine tracks remote accesses and reports the statistics to the coordinator periodically. After receiving statistics at 0 second, the coordinator adopts different mechanisms to notify migration threads. For eager migration, all of the migration threads will start migration directly, and the throughout reflects the benefits immediately, increasing from 239K to 2,142K. However, since the migration plan may not be optimal, the second migration happens at the next interval (after about 10s). The throughput further increases to 2,362K. For deferred migration, the coordinator will only ask monitors to track the local accesses on the potential key-value pairs for migration at 0 second, and do the migration with an optimal plan at the next interval. The throughput will directly increase from 239K to about 2,362K.

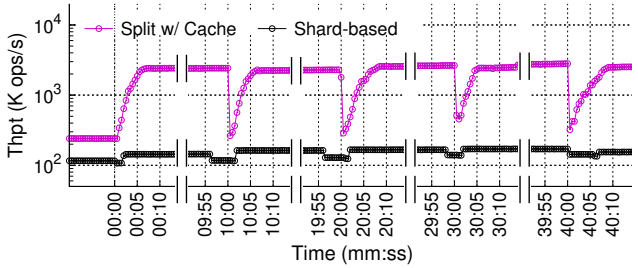

Fig. 15. *The throughput timeline for dynamic workloads using shard-based or split live migration.*

6.4 Dynamic Workloads

To study the effectiveness of split live migration in the face of dynamic workloads, we change workloads every 10 minutes by using non-overlapping scopes of starting vertices. As shown in Fig. 15, the performance notably drops every time the workloads change, because the location of vertex migrated for the current workload is very likely not suitable for the next workload. Shard-based migration can only provide very limited performance improvement as expected. Split migration with location cache can recover the performance after migration. Note that Pragh uses instant migration in this case, which is hard to implement in traditional migration approaches. When the monitor detects the frequency of accesses to some remote key-value pair exceeding a threshold (100 times per second), it will instantly report to the coordinator. Further, the migration on every machine can move values at any time, and there is no need to synchronize with other machines. Therefore, the performance is recovered gradually in about 5 seconds. Note that using a more aggressive policy could further reduce the time spent in recovery.

6.5 Application: RDF Graph and SPARQL Query

Wukong+M. To demonstrate the generality of Pragh, we have integrated *split live migration* with Wukong [52], called Wukong+M. Wukong is a state-of-the-art distributed graph store that leverages RDMA-based graph exploration to provide highly concurrent and low-latency SPARQL queries over large RDF graph datasets. RDF (Resource Description Framework) is a standard data model for the Semantic Web, recommended by W3C [2], which presents linked data as a set of $\langle subject, predicate, object \rangle$ triples forming a directed and labeled graph. SPARQL is the standard query language for RDF datasets, which can be supported by using graph exploration (i.e., graph traversals in parallel). We also implement an RDMA-friendly location cache on Wukong+M, similar to DrTM-KV [61].

Benchmark and workload. We use the Lehigh University Benchmark (LUBM) [3] which is widely used to evaluate the performance of RDF query systems [63, 36, 28, 52, 64, 37]. More specifically, we use LUBM-10240 dataset where each machine deploys about 32GB memory. We use the query set published in Atre et al. [8] and a mixed workload consist-

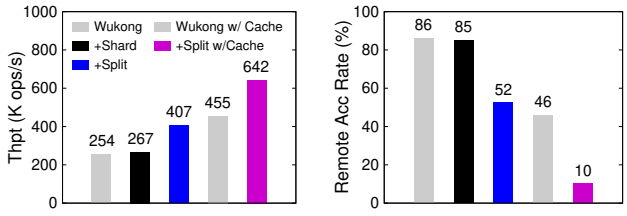

Fig. 16. *The comparison of (a) throughput and (b) remote access rate using a mixed workload for Wukong with various settings.*

ing of 6 classes as the same in the original paper [52]. The workload is skewed such that the starting vertices are chosen following a Zipf distribution ($\theta = 0.99$) over all vertices.

Performance. As shown in Fig. 16, Wukong+M (+Split) with location cache (w/ Cache) can outperform all other counterparts by up to 2.53×, thanks to split live migration that eliminates about 88% remote accesses (from 86% to 10%). Shard-based live migration (+Shard) only improves the mixed query throughput by about 5%, since it is hard to balance requirements for keys in each shard. The basic split migration (+Split) outperforms shard-based migration by 1.52× (407K ops/s vs. 267K ops/s) due to allowing fine-grained migration. After enabling location cache (+Split w/ Cache), the throughput further increases by 1.58× (642K ops/s vs. 407K ops/s).

7 Related Work

Live migration on relational stores. There have been many efforts to provide live migration features for distributed relational databases, considering different low-level architectures, such as shared-storage [20, 21, 26, 48, 19, 10] or partitioned database [53, 25, 60, 35]. They mainly focus on migrating shards efficiently across machines for balancing load and reducing latency. There are two main types of approaches: pre-copy based [20, 21, 60] and post-copy based [26, 25, 35].

To the best of our knowledge, almost all such systems adopt shard-based mechanisms (e.g., range or hash partitioning [17, 43]) and the changes of the ownership of shard are necessary when migration. Hence, they must maintain the state of shards explicitly by using internal global data structures or external location services [55, 4, 5]. Differently, split live migration fixes the (logical) location of data to avoid the maintenance overhead, which makes it different from all of the previous approaches.

The inherent drawback of one-off sharding has driven a few recent efforts to support dynamic sharding [25], auto sharding [4] and application-specific sharding [5] techniques. However, when shards still serve as the unit of migration, it is hard to balance the effectiveness (granularity) and efficiency (CPU and memory) for large-scale graph data with dynamic workloads due to lacks of locality.

Live migration on graph stores. The increasing importance of graph data models has stimulated a few recent designs of vertex migration or graph re-partitioning techniques targeting graph systems [44, 62, 34, 41, 65], since it is hard or even impossible to handle dynamic workloads or evolving graphs only relying on static partition-based approaches [27, 13]. The most related work is Mizan [34], a distributed graph processing system that leverages fine-grained vertex migration to improve *load balance* for iterative analytics workloads (e.g., PageRank and DMST [34]) over a *static* graph. Further, vertex migration can only happen when all worker threads reach a synchronization barrier (*stop-the-world*), and all selected vertices in one machine can only be migrated to a pairwise machine (*non-flexible*). By contrast, Pragh uses live migration to preserve *locality* for *concurrent* and *dynamic* traversal operations over *evolving* graphs. Thus, it makes many fine-grained migrations *on demand*, and vertices can be migrated to any machines *flexibly*.

Most graph re-partitioning approaches [44, 62, 65] need to maintain global metadata to map vertices to partitions, and use multiple phases to iteratively migrate vertices for reducing the communication cost. Therefore, these design choices make them slow to react to changes of workloads and other real-time events. Pragh can provide instant response to migration demands using lightweight monitoring and unilateral migration protocol.

Further, data replication has been used to improve the locality of traversal workloads over graph stores [29, 62, 40] by duplicating vertices on multiple machines. However, it will consume more memory and complicate the design of graph store in the face of evolving graphs. It should be noted that data replication is orthogonal to live migration, and integrating split live migration with fine-grained vertex replication [27, 13] is part of our future work.

8 Conclusion

This paper presents Pragh, an efficient locality-preserving live migration scheme for graph store. The key idea of Pragh is *split live migration*, which allows fine-grained migration while avoiding the need to maintain excessive metadata. Several key designs like the unilateral migration protocol, the integration of location-based caching, and the check-and-forward mechanism for evolving graphs made Pragh fast and full-fledged. Evaluations using both a graph traversal benchmark and SPARQL workloads confirmed the effectiveness and generality of Pragh.

Acknowledgments

We sincerely thank our shepherd Dushyanth Narayanan and the anonymous reviewers for their insightful suggestions. This work was supported in part by the National Natural Science Foundation of China (No. 61772335, 61572314, 61732010), the National Youth Top-notch Talent Support Program of China, and a research grant from Alibaba Group through Alibaba Innovative Research (AIR) Program. Corresponding author: Rong Chen (rongchen@sjtu.edu.cn).

References

[1] IEEE 1588 Precision Time Protocol (PTP) Version 2. http://sourceforge.net/p/ptpd/wiki/Home/.

[2] Semantic Web. https://www.w3.org/standards/semanticweb/.

[3] SWAT Projects - the Lehigh University Benchmark (LUBM). http://swat.cse.lehigh.edu/projects/lubm/.

[4] A. Adya, D. Myers, J. Howell, J. Elson, C. Meek, V. Khemani, S. Fulger, P. Gu, L. Bhuvanagiri, J. Hunter, et al. Slicer: Auto-sharding for datacenter applications. In *OSDI*, pages 739–753, 2016.

[5] M. Annamalai, K. Ravichandran, H. Srinivas, I. Zinkovsky, L. Pan, T. Savor, D. Nagle, and M. Stumm. Sharding the shards: managing datastore locality at scale with akkio. In *13th USENIX Symposium on Operating Systems Design and Implementation*, OSDI '18, pages 445–460, 2018.

[6] T. G. Armstrong, V. Ponnekanti, D. Borthakur, and M. Callaghan. Linkbench: A database benchmark based on the facebook social graph. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 1185–1196, New York, NY, USA, 2013. ACM.

[7] B. Atikoglu, Y. Xu, E. Frachtenberg, S. Jiang, and M. Paleczny. Workload analysis of a large-scale key-value store. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 53–64, New York, NY, USA, 2012. ACM.

[8] M. Atre, V. Chaoji, M. J. Zaki, and J. A. Hendler. Matrix "bit" loaded: A scalable lightweight join query processor for rdf data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 41–50, New York, NY, USA, 2010. ACM.

[9] D. Barak. VERBS Programming Tutorial. *OpenSHMEM*, 2014.

[10] S. Barker, Y. Chi, H. J. Moon, H. Hacigümüş, and P. Shenoy. "cut me some slack": Latency-aware live migration for databases. In *Proceedings of the 15th International Conference on Extending Database Technology*, EDBT '12, pages 432–443, New York, NY, USA, 2012. ACM.

[11] N. Bronson, Z. Amsden, G. Cabrera, P. Chakka, P. Dimov, H. Ding, J. Ferris, A. Giardullo, S. Kulkarni, H. C. Li, et al. Tao: Facebook's distributed data store for the social graph. In *USENIX Annual Technical Conference*, pages 49–60, 2013.

[12] D. Chakrabarti, Y. Zhan, and C. Faloutsos. R-mat: A recursive model for graph mining. In *Proceedings of the 2004 SIAM International Conference on Data Mining*, pages 442–446. SIAM, 2004.

[13] R. Chen, J. Shi, Y. Chen, and H. Chen. Powerlyra: Differentiated graph computation and partitioning on skewed graphs. In *Proceedings of the Tenth European Conference on Computer Systems*, EuroSys '15, pages 1:1–1:15, New York, NY, USA, 2015. ACM.

[14] Y. Chen, X. Wei, J. Shi, R. Chen, and H. Chen. Fast and general distributed transactions using rdma and htm. In *Proceedings of the Eleventh European Conference on Computer Systems*, EuroSys'16, pages 26:1–26:17, New York, NY, USA, 2016. ACM.

[15] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears. Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC'10, pages 143–154. ACM, 2010.

[16] C. Curino, E. Jones, Y. Zhang, and S. Madden. Schism: A workload-driven approach to database replication and partitioning. *Proc. VLDB Endow.*, 3(1-2):48–57, Sept. 2010.

[17] C. Curino, E. P. Jones, S. Madden, and H. Balakrishnan. Workload-aware database monitoring and consolidation. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 313–324, New York, NY, USA, 2011. ACM.

[18] M. Curtiss, I. Becker, T. Bosman, S. Doroshenko, L. Grijincu, T. Jackson, S. Kunnatur, S. Lassen, P. Pronin, S. Sankar, G. Shen, G. Woss, C. Yang, and N. Zhang. Unicorn: A system for searching the social graph. *Proc. VLDB Endow.*, 6(11):1150–1161, Aug. 2013.

[19] S. Das, D. Agrawal, and A. El Abbadi. Elastras: An elastic, scalable, and self-managing transactional database for the cloud. *ACM Trans. Database Syst.*, 38(1):5:1–5:45, Apr. 2013.

[20] S. Das, S. Nishimura, D. Agrawal, and A. El Abbadi. Live Database Migration for Elasticity in a Multitenant Database for Cloud Platforms. *CS, UCSB, Santa Barbara, CA, USA, Tech. Rep*, 9:2010, 2010.

[21] S. Das, S. Nishimura, D. Agrawal, and A. El Abbadi. Albatross: Lightweight elasticity in shared storage databases for the cloud using live data migration. *Proc. VLDB Endow.*, 4(8):494–505, May 2011.

[22] A. Dragojević, D. Narayanan, O. Hodson, and M. Castro. FaRM: Fast remote memory. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, NSDI'14, pages 401–414. USENIX Association, 2014.

[23] A. Dragojević, D. Narayanan, E. B. Nightingale, M. Renzelmann, A. Shamis, A. Badam, and M. Castro. No compromises: Distributed transactions with consistency, availability, and performance. In *Proceedings of the 25th Symposium on Operating Systems Principles*, SOSP'15, pages 54–70, New York, NY, USA, 2015. ACM.

[24] A. Dubey, G. D. Hill, R. Escriva, and E. G. Sirer. Weaver: A high-performance, transactional graph database based on refinable timestamps. *Proc. VLDB Endow.*, 9(11):852–863, July 2016.

[25] A. J. Elmore, V. Arora, R. Taft, A. Pavlo, D. Agrawal, and A. El Abbadi. Squall: Fine-grained live reconfiguration for partitioned main memory databases. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 299–313, New York, NY, USA, 2015. ACM.

[26] A. J. Elmore, S. Das, D. Agrawal, and A. El Abbadi. Zephyr: Live migration in shared nothing databases for elastic cloud platforms. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, pages 301–312, New York, NY, USA, 2011. ACM.

[27] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. In *Presented as part of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*, pages 17–30, Hollywood, CA, 2012. USENIX.

[28] S. Gurajada, S. Seufert, I. Miliaraki, and M. Theobald. Triad: A distributed shared-nothing rdf engine based on asynchronous message passing. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 289–300, New York, NY, USA, 2014. ACM.

[29] R. Harbi, I. Abdelaziz, P. Kalnis, N. Mamoulis, Y. Ebrahim, and M. Sahli. Accelerating sparql queries by exploiting hash-based locality and adaptive partitioning. *The VLDB Journal*, 25(3):355–380, June 2016.

[30] P. Hunt, M. Konar, F. P. Junqueira, and B. Reed. Zookeeper: Wait-free coordination for internet-scale systems. In *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference*, USENIX ATC'10, pages 11–11. USENIX Association, 2010.

[31] B. Iordanov. Hypergraphdb: A generalized graph database. In *Proceedings of the 2010 International Conference on Web-age Information Management*, WAIM'10, pages 25–36, Berlin, Heidelberg, 2010. Springer-Verlag.

[32] A. Kalia, M. Kaminsky, and D. G. Andersen. Using rdma efficiently for key-value services. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM'14, pages 295–306. ACM, 2014.

[33] A. Khandelwal, R. Agarwal, and I. Stoica. Blowfish: Dynamic storage-performance tradeoff in data stores. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, pages 485–500, Santa Clara, CA, Mar. 2016. USENIX Association.

[34] Z. Khayyat, K. Awara, A. Alonazi, H. Jamjoom, D. Williams, and P. Kalnis. Mizan: A system for dynamic load balancing in large-scale graph processing. In *Proceedings of the 8th ACM European Conference on Computer Systems*, EuroSys '13, pages 169–182, New York, NY, USA, 2013. ACM.

[35] C. Kulkarni, A. Kesavan, T. Zhang, R. Ricci, and R. Stutsman. Rocksteady: Fast migration for low-latency in-memory storage. In *Proceedings of the 26th Symposium on Operating Systems Principles*, SOSP '17, pages 390–405, New York, NY, USA, 2017. ACM.

[36] K. Lee and L. Liu. Scaling queries over big rdf graphs with semantic hash partitioning. *Proceedings of the VLDB Endowment*, 6(14):1894–1905, 2013.

[37] C. Mayer, R. Mayer, J. Grunert, K. Rothermel, and M. A. Tariq. Q-graph: preserving query locality in multi-query graph processing. In *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, page 6. ACM, 2018.

[38] Mellnox. RDMA Aware Networks Programming User Manual, Rev 1.7. http://www.mellanox.com/related-docs/prod_software/RDMA_Aware_Programming_user_manual.pdf.

[39] C. Mitchell, Y. Geng, and J. Li. Using one-sided rdma reads to build a fast, cpu-efficient key-value store. In *Proceedings of the 2013 USENIX Conference on Annual Technical Conference*, USENIX ATC'13, pages 103–114. USENIX Association, 2013.

[40] J. Mondal and A. Deshpande. Managing large dynamic graphs efficiently. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 145–156, New York, NY, USA, 2012. ACM.

[41] D. Nicoara, S. Kamali, K. Daudjee, and L. Chen. Hermes: Dynamic partitioning for distributed social network graph databases. In *EDBT*, pages 25–36, 2015.

[42] A. Pavlo, C. Curino, and S. Zdonik. Skew-aware automatic database partitioning in shared-nothing, parallel oltp systems. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 61–72, New York, NY, USA, 2012. ACM.

[43] A. Pavlo, C. Curino, and S. Zdonik. Skew-aware automatic database partitioning in shared-nothing, parallel oltp systems. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 61–72, New York, NY, USA, 2012. ACM.

[44] J. M. Pujol, V. Erramilli, G. Siganos, X. Yang, N. Laoutaris, P. Chhabra, and P. Rodriguez. The little engine(s) that could: Scaling online social networks. In *Proceedings of the ACM SIGCOMM 2010 Conference*, SIGCOMM '10, pages 375–386, New York, NY, USA, 2010. ACM.

[45] X. Qiu, W. Cen, Z. Qian, Y. Peng, Y. Zhang, X. Lin, and J. Zhou. Real-time constrained cycle detection in large dynamic graphs. *Proc. VLDB Endow.*, 11(12):1876–1888, Aug. 2018.

[46] L. Rietveld, R. Hoekstra, S. Schlobach, and C. Guéret. Structural properties as proxy for semantic relevance in rdf graph sampling. In *International Semantic Web Conference*, pages 81–96. Springer, 2014.

[47] S. Sahu, A. Mhedhbi, S. Salihoglu, J. Lin, and M. T. Özsu. The ubiquity of large graphs and surprising challenges of graph processing. *Proc. VLDB Endow.*, 11(4):420–431, Dec. 2017.

[48] O. Schiller, N. Cipriani, and B. Mitschang. Prorea: Live database migration for multi-tenant rdbms with snapshot isolation. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, pages 53–64, New York, NY, USA, 2013. ACM.

[49] M. Serafini, E. Mansour, A. Aboulnaga, K. Salem, T. Rafiq, and U. F. Minhas. Accordion: elastic scalability for database systems supporting distributed transactions. *Proceedings of the VLDB Endowment*, 7(12):1035–1046, 2014.

[50] M. Serafini, R. Taft, A. J. Elmore, A. Pavlo, A. Aboulnaga, and M. Stonebraker. Clay: Fine-grained adaptive partitioning for general database schemas. *Proc. VLDB Endow.*, 10(4):445–456, Nov. 2016.

[51] B. Shao, H. Wang, and Y. Li. Trinity: A distributed graph engine on a memory cloud. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 505–516, New York, NY, USA, 2013. ACM.

[52] J. Shi, Y. Yao, R. Chen, H. Chen, and F. Li. Fast and concurrent rdf queries with rdma-based distributed graph exploration. In *Proc. OSDI*, 2016.

[53] R. Taft, E. Mansour, M. Serafini, J. Duggan, A. J. Elmore, A. Aboulnaga, A. Pavlo, and M. Stonebraker. E-store: Fine-grained elastic partitioning for distributed transaction processing systems. *Proc. VLDB Endow.*, 8(3):245–256, Nov. 2014.

[54] Titan. Titan Data Model. http://s3.thinkaurelius.com/docs/titan/current/data-model.html, 2018.

[55] N. Tran, M. K. Aguilera, and M. Balakrishnan. Online migration for geo-distributed storage systems. In *USENIX Annual Technical Conference*, 2011.

[56] S. Tu, W. Zheng, E. Kohler, B. Liskov, and S. Madden. Speedy transactions in multicore in-memory databases. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, SOSP'13, pages 18–32. ACM, 2013.

[57] S. Wang, C. Lou, R. Chen, and H. Chen. Fast and concurrent rdf queries using rdma-assisted gpu graph exploration. In *Proc. USENIX ATC*, 2018.

[58] Z. Wang, H. Qian, J. Li, and H. Chen. Using restricted transactional memory to build a scalable in-memory database. In *Proceedings of the Ninth European Conference on Computer Systems*, EuroSys'14, pages 26:1–26:15, New York, NY, USA, 2014. ACM.

[59] X. Wei, Z. Dong, R. Chen, and H. Chen. Deconstructing rdma-enabled distributed transactions: Hybrid is better! In *13th USENIX Symposium on Operating Systems Design and Implementation*, OSDI '18, pages 233–251, 2018.

[60] X. Wei, S. Shen, R. Chen, and H. Chen. Replication-driven live reconfiguration for fast distributed transaction processing. In *Proceedings of the 2017 USENIX Annual Technical Conference*, USENIX ATC'17, pages 335–347, Santa Clara, CA, 2017. USENIX Association.

[61] X. Wei, J. Shi, Y. Chen, R. Chen, and H. Chen. Fast in-memory transaction processing using rdma and htm. In *Proceedings of the 25th Symposium on Operating Systems Principles*, SOSP '15, pages 87–104, New York, NY, USA, 2015. ACM.

[62] S. Yang, X. Yan, B. Zong, and A. Khan. Towards effective partition management for large graphs. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 517–528, New York, NY, USA, 2012. ACM.

[63] K. Zeng, J. Yang, H. Wang, B. Shao, and Z. Wang. A distributed graph engine for web scale rdf data. In *Proceedings of the 39th international conference on Very Large Data Bases*, PVLDB'13, pages 265–276. VLDB Endowment, 2013.

[64] Y. Zhang, R. Chen, and H. Chen. Sub-millisecond stateful stream querying over fast-evolving linked data. In *Proc. SOSP*, 2017.

[65] A. Zheng, A. Labrinidis, and P. K. Chrysanthis. Planar: Parallel lightweight architecture-aware adaptive graph repartitioning. In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*, pages 121–132. IEEE, 2016.