Tebis: Efficient Replica Index Construction for Persistent LSM-based Key-Value Stores

Michalis Vardoulakis

Thesis submitted in partial fulfillment of the requirements for the

Master of Science in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Prof. *Angelos Bilas*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).

University of Crete
Computer Science Department

Tebis: Efficient Replica Index Construction for Persistent LSM-based Key-Value Stores

Thesis submitted by
Michalis Vardoulakis
in partial fulfillment of the requirements for the
Master of Science in Computer Science and Engineering

THESIS APPROVAL

Author: _____
Michalis Vardoulakis

Committee approvals: _____
Angelos Bilas
Professor, Thesis Supervisor

_____
Kostas Magoutis
Associate Professor, Committee Member

_____
Polyvios Pratikakis
Assistant Professor, Committee Member

Departmental approval: _____
Polyvios Pratikakis
Assistant Professor, Director of Graduate Studies

Heraklion, October 2021

Tebis: Efficient Index Replication for Persistent LSM-based Key-Value Stores

Abstract

Log-Structured Merge tree (LSM tree) Key-Value (KV) stores have become a foundational layer in the storage stacks of datacenter and cloud services. Current approaches for achieving reliability and availability avoid replication at the KV store level and instead perform these operations at higher layers, e.g., the DB layer that runs on top of the KV store. The main reason for taking that approach is that past designs for replicated KV stores favor reducing network traffic and increasing I/O size. Therefore, they perform costly compactions to reorganize data in both the primary and backup nodes since they avoid sending the index over the network. Since all nodes in a rack-scale KV store function both as primary and backup nodes for different data shards (regions), this approach eventually hurts overall system performance.

In this paper, we design and implement *Tebis*, an efficient rack-scale LSM-based KV store that aims to significantly reduce the I/O amplification and CPU overhead in backup nodes and make replication in the KV store practical. We rely on two observations: (a) the increased use of RDMA in the datacenter, which reduces CPU overhead for communication, and (b) the use of KV separation that is becoming prevalent in modern KV stores. We use a primary-backup replication scheme that performs compactions only on the primary nodes and sends the pre-built index to the backup nodes of the region, avoiding all compactions in backups. Our approach includes an efficient mechanism to deal with pointer translation across nodes in the region index. Our results show that *Tebis* reduces in the backup nodes, I/O amplification by up to $3\times$, CPU overhead by up to $1.6\times$, and memory size needed for the write path by up to $2\times$, without increasing network bandwidth excessively, and by up to $1.3\times$. Overall, we show that our approach has benefits even when small KV pairs dominate in a workload (80%-90% of the total key-values). Finally, it enables KV stores to operate with larger growth factors (from 10 to 16) to reduce space amplification without sacrificing precious CPU cycles.

Αποδοτική Αντιγραφή Ευρετηρίων για Συστήματα Μόνιμης Αποθήκευσης Ζευγαριών Κλειδιού-Τιμής Βασισμένα σε LSM

Περίληψη

Τα συστήματα αποθήκευσης ζευγαριών κλειδιού-τιμής βασισμένα σε δένδρα Log-Structured Merge (LSM) έχουν γίνει ένα βασικό κομμάτι των λογισμικών αποθήκευσης δεδομένων σε κέντρα δεδομένων και υπηρεσίες υπολογιστικών νεφών. Τέτοια συστήματα πρέπει να αντιγράφουν τα δεδομένα τους, αλλά και μεταδεδομένα όπως το ευρετήριο, ώστε να να επιτύχουν να είναι αξιόπιστα και διαθέσιμα. Ως τώρα, τα συστήματα αποθήκευσης αποφεύγουν να δημιουργούν τα αντίγραφα των δεδομένων στο επίπεδο του συστήματος αποθήκευσης ζευγαριών κλειδιού-τιμής και προτιμούν να κάνουν αυτές τις διεργασίες σε υψηλότερα στρώματα, όπως για παράδειγμα στην βάση δεδομένων που τρέχει πάνω από το σύστημα αποθήκευσης ζευγαριών κλειδιού-τιμής. Παλαιότεροι σχεδιασμοί συστημάτων αποθήκευσης κλειδιού-τιμής προτιμούν να μειώσουν την κυκλοφορία στο δίκτυο και να αυξήσουν το μέγεθος των αιτημάτων εγγραφής δεδομένων στον δίσκο. Επομένως εκτελούν compactions για να αναδιοργανώσουν τα δεδομένα και στα κύρια και στα δευτερεύοντα αντίγραφα των δεδομένων, αφού αποφεύγουν να στείλουν το ευρετήριο χρησιμοποιώντας το δίκτυο. Καθώς όλοι οι κόμβοι σε ένα κατανεμημένο σύστημα αποθήκευσης ζευγαριών κλειδιού-τιμής λειτουργούν ταυτόχρονα ως κύριοι και ως δευτερεύοντες κόμβοι για διαφορετικά δεδομένα, μία τέτοια προσέγγιση βλάπτει την απόδοση ολόκληρου του συστήματος.

Σε αυτή την εργασία, σχεδιάζουμε και υλοποιούμε το *Tebis*, ένα αποδοτικό σύστημα αποθήκευσης ζευγαριών κλειδιού-τιμής βασισμένο σε δένδρο LSM με στόχο την δραστική μείωση του I/O amplification και του επεξεργαστικού κόστους για τα δευτερεύοντα αντίγραφα ώστε να γίνει πρακτική η αντιγραφή των δεδομένων στο επίπεδο του συστήματος αποθήκευσης ζευγαριών κλειδιού-τιμής. Βασιζόμαστε σε δύο παρατηρήσεις: (α) η αυξημένη χρήση του RDMA στα κέντρα δεδομένων, το οποίο μειώνει το επεξεργαστικό κόστος για επικοινωνία μεταξύ κόμβων και (β) την διαδεδομένη χρήση του διαχωρισμού ζευγαριών κλειδιού-τιμής σε σύγχρονα συστήματα αποθήκευσης ζευγαριών κλειδιού-τιμής. Χρησιμοποιούμε ένα πρωτόκολλο αντιγραφής δεδομένων primary-backup όπου μόνο ο κύριος κόμβος υπολογίζει το ευρετήριο και στη συνέχεια το στέλνει σε όλους τους δευτερεύοντες κόμβους, αποφεύγοντας έτσι όλα τα compactions στους δευτερεύοντες κόμβους. Η προσέγγιση μας περιλαμβάνει και έναν αποδοτικό μηχανισμό μετάφρασης των δεικτών του ευρετηρίου μεταξύ διαφορετικών κόμβων. Τα αποτελέσματα μας δείχνουν ότι το *Tebis* μειώνει το I/O amplification έως και 3 φορές, το επεξεργαστικό κόστος έως και 1,6 φορές, και την μνήμη που χρειάζεται για την εγγραφή δεδομένων έως και 2 φορές, αυξάνοντας τα δεδομένα του δικτύου έως το πολύ 1,3 φορές. Συνολικά, δείχνουμε ότι η μέθοδος μας έχει οφέλη ακόμα και σε περιπτώσεις όπου τα μικρά κλειδιά κυριαρχούν (80% - 90% επί του συνόλου κλειδιών-τιμών). Τέλος, η μέθοδος μας επιτρέπει σε συστήματα αποθήκευσης ζευγαριών κλειδιού-τιμής να λειτουργούν με μεγαλύτερους ρυθμούς αύξησης δεδομένων από επίπεδο σε επίπεδο (growth factor), όπως 10 έως 16, μειώνοντας την περιττή χρήση αποθηκευτικού χώρου λόγω των πολλαπλών επιπέδων (space amplification) χωρίς να επιφέρει επεξεργαστικό κόστος.

Acknowledgments

I was lucky enough to receive a great deal of support throughout my studies.

I would first like to thank my supervisor Prof. Angelos Bilas whose direction and guidance was invaluable in steering me in the right direction and whose door was always open when I needed to discuss any issues with my studies or my research.

Next, I would like to thank Giorgos Saloustros for our great collaboration on this project throughout both my undergraduate and post-graduate studies. Our discussions and our collaboration have helped me expand my knowledge in research and software engineering.

A big thank you to everyone at the Computer Architecture and VLSI Systems lab at ICS-FORTH for creating a fun environment to work in and for their valuable input on my work.

Last but not least, I want to thank my family, my friends and my girlfriend for supporting me through my studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Replicated persistent key-value (KV) stores are the heart of modern datacenter storage stacks [28, 24, 12, 10, 1]. These systems typically use an LSM tree [30] index structure because of its 1) fast data ingestion capability for small and variable size data items while maintaining good read and scan performance and 2) its low space overhead on the storage devices [13]. LSM trees organize their data in hierarchical levels where the first level (L_0) is kept in memory and the rest of the levels are on the storage device. While there are multiple ways to organize data across LSM tree levels [20, 30], in this work, we focus on leveled LSM-based KV stores that organize their levels in non-overlapping ranges. When a higher LSM tree level is full, its data is moved to a lower level through a *compaction* operation. Compactions incur high CPU overhead and increase I/O amplification in LSM-based KV stores.

To provide reliability and availability, state-of-the-art KV stores [10, 24] replicate their KV pairs in multiple, typically two or three [5], nodes. Current designs for replication optimize network traffic and favor sequential I/O to the storage devices both in the primary and backup nodes. Essentially, these systems perform costly compactions to reorganize data in both the primary and backup nodes to ensure: (a) minimal network traffic by moving user data across nodes, and (b) sequential device access by performing only large I/Os. However, this approach comes at a significant increase in device traffic (I/O amplification) and CPU utilization at the backups. Given that all nodes in a replicated KV store function both as primaries and backups at the same time for different regions, this approach hurts overall system performance. For this reason, in many cases, current approaches for reliability and availability avoid replication at the KV store level and instead perform these operations at higher layers, e.g. the DB layer that runs ontop of the KV store [10, 24].

Nowadays, state-of-the-art KV stores [10, 24] adopt the eager approach [33, 24, 10] which minimizes network traffic and recovery time at the expense of I/O amplification, CPU, and memory overhead at the secondaries. This approach is appropriate for systems designed for TCP/IP networks and Hard Disk Drives (HDDs).

In our work, we rely on two key observations: (a) the increased use of RDMA in the datacenter [35, 17], especially at the rack level [21, 29, 14, 22], reduces CPU overhead

for communication and (b) the use of KV separation that is becoming prevalent in modern KV stores [27, 25, 38, 31, 2, 15]. KV separation places KV pairs in a value log and keeps an LSM index where values point to locations in the value log. As a result, they only re-organize the keys (and pointers) in the multi-level structure. This technique introduces small and random read I/Os, which fast storage devices can handle, and reduces I/O amplification by up to 10x [3]. Additionally, recent works present hybrid KV placement [38, 25], a technique that extends KV separation and significantly improves garbage collection overhead for KV separation [9, 34], making it production ready.

We design and implement *Tebis*, an efficient rack-scale LSM-based KV store that significantly reduces I/O amplification and CPU overhead in secondary nodes and makes replication in the KV store practical. *Tebis* uses *Kreon* [8, 31] as its storage layer, an open-source persistent LSM-based KV store designed for fast storage devices (NVMe) and that uses KV separation to reduce I/O amplification. Moreover, *Tebis* uses one-sided RDMA communication for data replication, client-server, and server-server communication. One-sided operations allow one peer to directly read or write the memory of a remote peer without the remote one having to post an operation, hence bypassing the remote node CPU and consuming CPU cycles only in the originating node. *Tebis*'s main novelty lies in how it takes advantage of the design of its storage engine and RDMA networking to send a pre-build index from primaries to secondaries in order to eliminate compactions in the secondaries.

The three main design challenges in *Tebis* are the following. First, to efficiently replicate the data (value log) *Tebis* uses an efficient RDMA-based primary-backup communication protocol. This protocol does not require the involvement of the replica CPU in communication operations [36].

Second, since the index of the *primary* contains pointers to its value log, *Tebis* implements an efficient rewrite mechanism at the *backups*. Kreon performs all allocations in 2 MB segments and all logical structures in Kreon (level's indexes and value log) are represented as a list of segments on the device. During its log and index replication process, *Tebis* creates mappings between *primary* and *backup* segments. It later uses these mappings to efficiently rewrite pointers at the *backups*. This approach allows *Tebis* to operate at larger growth factors to save space without significant CPU overhead.

Finally, to reduce CPU overhead for client-server communication, *Tebis* implements an RDMA protocol with one-sided RDMA write operations. *Tebis*'s protocol supports variable size messages that are essential for KV stores. Since these messages are sent in a single round trip, *Tebis* is able to reduce the processing overhead at the server.

We evaluate *Tebis*'s performance using a modified version of the Yahoo Cloud Service Benchmark (YCSB) [11] that supports variable key-value sizes for all YCSB workloads, similar to Facebook's [7] production workloads. Our results show that our index replication method compared to a baseline implementation that performs compactions at the *backups* spends $10\times$ fewer CPU cycles per operation to replicate its index. Furthermore, it has $1.1 - 1.7\times$ higher throughput, reduces I/O amplification by $1.1 - 2.3\times$, and increases CPU efficiency by $1.2 - 1.6\times$. Overall, *Tebis* technique of sending and rewriting a pre-built index gives KV stores the ability to operate at larger growth factors and save space without spending precious CPU cycles [3, 13].

Chapter 2

Design

2.1 Overview

Tebis is a persistent rack-scale KV store that increases CPU efficiency in *backup* regions for data replication purposes. *Tebis* uses a primary-backup protocol [6] for replicating the data via RDMA writes without involving the CPU of the *backups* [36] for efficiency purposes. To reduce the overhead of keeping an up-to-date index at the *backups*, we design and implement the *Send Index* operation for systems that use KV-separation [27, 31, 2, 15] or hybrid KV placement [38, 25]. *Primary* servers, after performing a compaction from level L_i to L_{i+1}, send the resulting L'_{i+1} to their *backups* in order to eliminate compactions in *backup* regions. Because L'_{i+1} contains references to the primary's storage address space, *backups* use a lightweight rewrite mechanism to convert the *primary*'s L'_{i+1} into a valid index for their own storage address space. During the Send Index operation, the *backup* uses metadata (hundreds of KB) retrieved during the replication of the KV log to translate pointers of the *primary*'s KV log into its own storage space.

We design and implement an RDMA Write-based protocol for both its server-server and client-server communication. We build our protocol using one-sided RDMA write operations because they reduce the network processing CPU overhead at the server [23] due to the absence of network interrupts. Furthermore, *Tebis*, as a KV store, must support variable size messages. We design our protocol to complete all KV operations in a single round trip to reduce the messages processed per operation by the servers.

Tebis uses Kreon [8, 31] KV store for efficiently managing the index over its local devices. We modify Kreon to use direct I/O to write its KV log to further reduce CPU overhead for consecutive write page faults, since write I/Os are always large. Direct I/O also avoids polluting the buffer cache from compaction traffic.

Finally, *Tebis* partitions the key-value space into non-overlapping key ranges named *regions* and offers clients a CRUD API (Create, Read, Update, Delete) as well as range (scan) queries. *Tebis* consists of the three following entities, as shown in Figure 2.1:

1. *Zookeeper* [19], a highly available service that keeps the metadata of *Tebis* highly available and strongly consistent, and checks for the health of *Tebis region servers*.

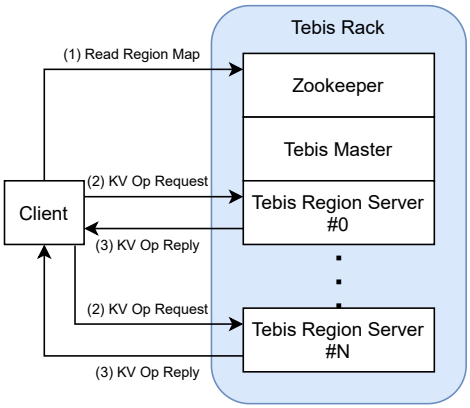


Figure 2.1: *Tebis* overview.

2. *Region servers*, which keep a subset of regions for which they either have the *primary* or the *backup* role.
3. *Tebis-Master*, which is responsible for assigning regions to *region servers* and orchestrating recovery operations after a failure.

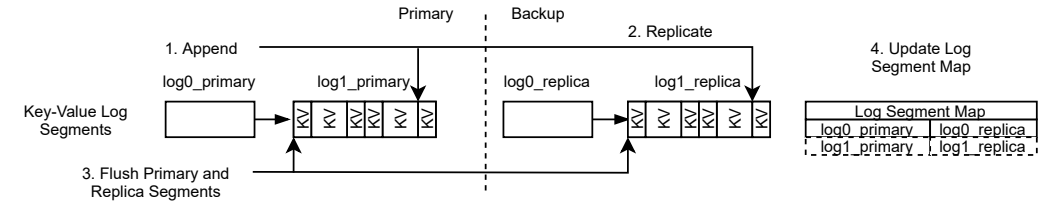
2.2 Primary-backup Value Log Replication

We design and implement a primary-backup replication protocol to remain available and avoid data loss in case of failures. Each *region server* stores a set of regions and has either the *primary* or *backup* role for any region in its set. The main design challenge that *Tebis* addresses is to replicate data and keep full indexes at the *backups* with low CPU overhead. Having an up-to-date index at each *backup* is necessary to provide a fast recovery time in case of a failure.

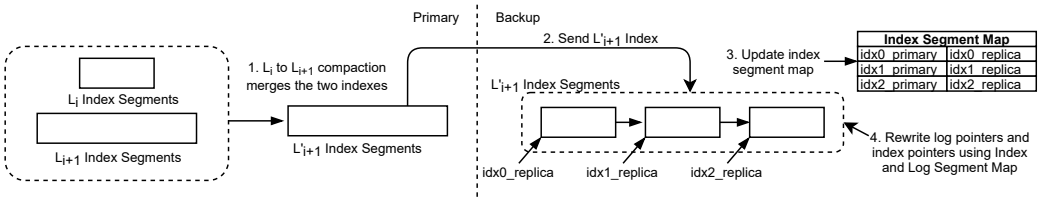
Tebis implements a primary-backup protocol over RDMA for replication [6, 36]. On initialization, the *primary* sends a message to each *backup* to request an RDMA buffer of the same size as Kreon's value log segment (2 MB). When a client issues an insert or update KV operation, the *primary* replicates this operation in its set of *backup* servers. The *primary* completes the client's operation in the following three steps:

1. Inserts the KV pair in Kreon, which returns the offset of the KV pair in the value log, as shown in step 1 in Figure 2.2a).
2. Appends (via RDMA write operation) the KV pair to the RDMA buffer of each replica at the corresponding offset, as shown in step 2 in Figure 2.2a).
3. Sends a reply to the client after receiving the completion event from all *backup*s for the above RDMA write operation.

The *backup*'s CPU is not involved in any of the above steps due to the use of RDMA write operations. When a client receives an acknowledgment it means that its operations has been replicated to all the memories of the replica set.



(a) Key-Value log replication process



(b) Key-Value index replication process

Figure 2.2: Replication in *Tebis*.

When the last log segment of the *primary* becomes full, the *primary* writes this log segment to persistent storage and sends a *flush* message to each *backup* requesting them to persist their RDMA buffer, as shown in step 3 in Figure 2.2a. *Backup* servers then copy their RDMA buffer to the last log segment of the corresponding Kreon region and write that log segment to their persistent storage. *Backup* servers also update their *log segment map*, as shown in step 4 in Figure 2.2a. The log segment map contains entries of the form <primary value log segment, replica value log segment>. Each *backup* server keeps this map and updates it after each flush message. *Backups* use this map to rewrite the *primary* index. We will discuss this index rewriting mechanism in more detail in Section 2.3.

Each *backup* keeps the log segment map per *backup* region in memory. The log segment map has a small memory footprint; for a 1 TB device capacity and two replicas, the value log will be 512 GB in the worst case. With the segment size set to 2 MB, the memory size of the log segment map across all regions will be at most 4 MB. In case of *primary* failure, the new *primary* informs its *backups* about the new mappings.

2.3 Efficient *Backup* Index Construction

Tebis instead of repeating the compaction process at each server to reduce network traffic, takes a radical approach. *Primary* executes the heavy, in terms of CPU, compaction process of L_i and L_{i+1} and sends the resulting L'_{i+1} to the *backups*. This approach has the following advantages. First, servers do not need to keep an L_0 level for their *backup* regions, reducing the memory budget for L_0 by $2\times$ when keeping one replica per region or by $3\times$ when keeping two replicas. Second, *backups* save device I/O and CPU since they do not perform compactions for their *backup* regions.

Essentially, this approach trades network I/O traffic for CPU, device I/O, and memory

at servers since network bulk transfers are CPU efficient due to RDMA. The main design challenge to address is sending the *primary* level index to the *backup* in a format that *backups* can rewrite with low CPU overhead. *Tebis* implements the rewriting process at the *backup* as follows.

When level L_i of a region in a *primary* is full, *Tebis* starts a compaction process to compact L_i with L_{i+1} into L'_{i+1}. The *primary* reads L_i and L_{i+1} and builds L'_{i+1} B+-tree index. L'_{i+1} is represented in the device as a list of segments (currently set to 2 MB) which contains either leaf or index nodes of the B+-tree. Leaf nodes contain pairs of the form <key prefix, pointer to value log> whereas index nodes contains pairs of the form <pivot, pointer to node>.

To transfer the L'_{i+1} index, the *primary* initially requests from each *backup* to register an RDMA buffer of segment size. *Tebis* only uses these buffers during the L'_{i+1} index transfer and deregisters and frees them once the compaction is completed.

On receiving a leaf segment, each *backup region server* parses it and performs the following steps. Leaf segments contain key prefixes that work across both indexes. The *backup* has to rewrite pointers to the value log before using them. *Tebis*'s storage engine performs all allocations in 2 MB aligned segments. As a result, the first high 22 bits of a device offset refer to the segment's start device offset. The remaining bits are an offset within that segment. To rewrite the value log pointers of the *primary*, the *backup* first calculates the segment start offset of each KV pair. Since all segments are aligned, it does this through a modulo operation with segment size. Then it issues a lookup in the log map and replaces the primary segment address with its local segment address.

For index segments, *Tebis* keeps in-memory an *index map* for the duration of L'_{i+1} compaction. *Backups* add entries to this map whenever they receive an index segment from the *primary*. This map contains entries using as the *primary*'s index segment as the key and the corresponding *backup*'s index segment as the value, as shown in Figure 2.2b. This mechanism translates pointers to index or leaf nodes within the segment the same way as it does for value log pointers in leaves. Finally, on compaction completion, the *primary* sends the root node offset of L_{i+1} to each *backup*, which each *backup* translates to its storage space.

2.4 Failure Detection

Tebis uses Zookeeper's ephemeral nodes to detect failures. An ephemeral node is a node in Zookeeper that gets automatically deleted when its creator stops responding to heartbeats of Zookeeper. Every *region server* creates an ephemeral node during its initialization. In case of a failure, the *Tebis-Master* gets notified about the failure and runs the corresponding recovery operation. In case of *Tebis-Master* failure, all *region servers* get notified about its failure through the ephemeral node mechanism. Then, they run an election process through Zookeeper and decide which node takes over as *Tebis-Master*.

2.5 Failure Recovery

Tebis uses Zookeeper, similar to other systems [1, 26], to store its *region map*. Each entry in the region map consists of the range of the region <start key, end key>, the *primary* server responsible for it, and the list of its *backup* servers. The region map is infrequently updated when a region is created, deleted after a failure, or during load-balancing operations. Therefore, in *Tebis* Zookeeper operations are not in the common path of data access.

The *Tebis-Master* reads the region map during its initialization and issues *open region* commands to each *region server* in the *Tebis* cluster, assigning them a *primary* or a *backup* role. After initialization, the role of the *Tebis-Master* is to orchestrate the recovery process in case of failures and to perform load balancing operations.

Clients read and cache the region map during their initialization. Before each KV operation, clients look up their local copy of the region map to determine the *primary region server* where they should send their request. Clients cache the region map since each region entry is 64 B, meaning just 640 KB are enough for a region map with 10,000 regions, and changes to it are infrequent. When a client issues a KV operation to a *region server* that is not currently responsible for the corresponding range, the *region server* instructs it to update their region map.

Tebis has to handle three distinct failure cases: 1) *backup* failure, 2) *primary* failure, and 3) *Tebis-Master* failure. Since each *Tebis region server* is part of multiple region groups, a single node failure results in numerous *primary* and *backup* failures, which the *Tebis-Master* handles concurrently. First, we discuss how we handle *backup* failures.

In case of a *backup* failure, the *Tebis-Master* replaces the crashed *region server* with another one that is not already part of the corresponding region's group. The *Tebis-Master* then instructs the rest of the *region servers* in the group to transfer the region data to the new member of the region group. The region experiencing the *backup* failure will remain available throughout the whole process since its *primary* is unaffected. However, during the reconstruction of the new *backup*, the region data are more susceptible to future failures, since there's one less *backup* copy.

In case of a *primary* failure, the *Tebis-Master* first promotes one of the existing *backup region servers* in that region group to the *primary* role, and updates the region map. The new *primary* already has a complete KV log and an index for levels L_i, where $i \geq 1$. The new *primary region server* replays the last few segments of its value log in order to construct an L_0 index in its memory before being able to server client requests. Now that a new *primary region server* exists for the group, the *Tebis-Master* handles this failure as if it were a *backup* failure. During the *primary* reconstruction process, *Tebis* cannot server client requests from the affected region.

When the *Tebis-Master* crashes, the rest of the *region servers* in the *Tebis* cluster will be notified through Zookeeper's ephemeral node mechanism, as discussed in Section 2.4. They will then use Zookeeper in order to elect a new *Tebis-Master*. During the *Tebis-Master* downtime, *Tebis* cannot handle any region failures, meaning that any region that has suffered a *primary* failure will remain unavailable until a new *Tebis-Master* is elected and initiates the recovery process for any regions that suffered a failure.

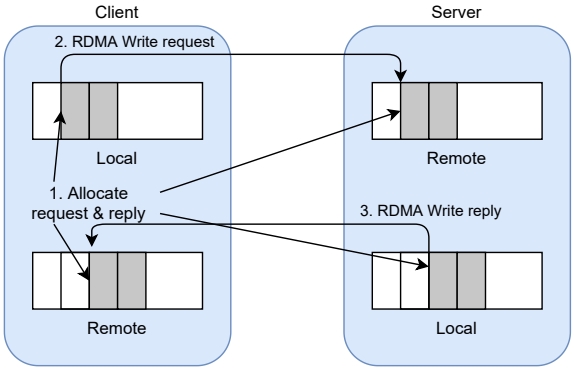


Figure 2.3: Allocation and request-reply flow of our RDMA Write-based communication protocol.

2.6 RDMA Write-based Communication Protocol

Tebis performs all client server communication via one-sided RDMA write operations [23] to avoid network interrupts and thus reduce the CPU overhead in the server's receive path [23, 22]. Furthermore, to avoid the overhead of registering and deregistering RDMA memory buffers per KV operation, the server and client allocate a pair of buffers during queue pair (QP) creation. Their size is dynamic within a range (256 KB currently) set by the client on QP creation. *region server* frees this buffer when a client disconnects or suffers a failure. The client manages these buffers to improve CPU efficiency in the server.

Clients allocate a pair of messages for each KV operation; one for their request and one for the server's reply. All buffers sizes are multiples of a size unit named *message segment size* (currently set to 128 bytes). Clients put in the header of each request the offset at their remote buffer where *region server* can write its reply. Upon completion of a request, the *region server* prepares the request's reply in the corresponding *local* circular buffer at the offset supplied by the client. Then it issues an RDMA write operation to the client's *remote* circular buffer at the exact offset. Figure 2.3 shows a visual representation of these steps. As a result, the *region server* avoids expensive synchronization operations between its workers to allocate space in the remote client buffers and update buffer state metadata (free or reserved). If the client allocates a reply of insufficient size, the *region server* sends part of the reply and informs the client to retrieve the rest of the data.

2.6.1 Receive Path

To detect incoming messages, in the absence of network interrupts, each *region server* has a *spinning thread* which spins on predefined memory locations in its corresponding clients' remote circular buffers, named *rendezvous points*. The spinning thread detects a new message by checking for a magic number in the last field of the message header, called the *receive field*, at the next rendezvous location. After it detects a new message header, it reads the payload size from the message header to determine the location of the message's tail. Upon successful arrival of the tail, it assigns the new client request to one

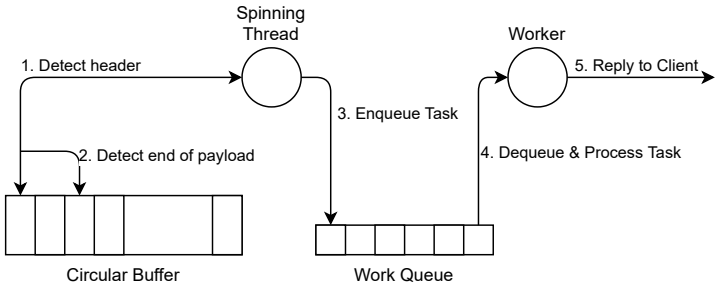


Figure 2.4: Message detection and task processing pipeline in *Tebis*. For simplicity, we only draw one circular buffer and a single worker.

of its workers and advances to the next rendezvous location of the circular buffer.

To support variable size messages *Tebis* adds appropriate padding so that their size is message segment aligned. This quantization has two benefits: 1) Possible rendezvous points are at the start of each segment, offset by the size of a message header minus the size of the header's receive field. Upon reception of a message the *region server* advances its rendezvous point by adding the current message size. 2) The *region server* does not have to zero the whole message upon each request completion; it only zeros the possible rendezvous points in the message segments where the request was written.

2.6.2 Reset Operation in Circular Buffers

There are two ways to reset the rendezvous point to the start of the circular buffer: 1) When the last message received in the circular buffer takes up its whole space, the server will pick the start of the circular buffer as the next rendezvous location, and 2) When the remaining space in the circular buffer is not enough for the client to send their next message, they will have to circle back to the start of the buffer. In this case, they will send a *reset rendezvous* message to inform the server that the next rendezvous location is now at the start of the circular buffer.

2.6.3 Task Scheduling

To limit the max number of threads, *Tebis* uses a configurable number of workers. Each worker has a private *task queue* to avoid CPU overhead associated with contention on shared data. In this queue, the spinning thread places new tasks, as shown in Figure 2.4. Workers poll their queue to retrieve a new request and sleep if no new task is retrieved within a set period of time (currently $100\ \mu s$). The primary goal of *Tebis*'s task scheduling policy is to limit the number of wake-up operations, since they include crossings between user and kernel space. The spinning thread assigns a new task to the current worker unless its task queue has more than a set amount of tasks already enqueued. In the latter case, the spinning thread selects a running worker with less than that set amount of queued tasks and assigns to it the new task. If all running workers already exceed the task queue limit, the spinning thread wakes up a sleeping worker and enqueues this task to their task queue.

Chapter 3

Evaluation Methodology

Our testbed consists of two servers where we run the KV store. The servers are identical and are equipped with an AMD EPYC 7551P processor running at 2 GHz with 32 cores and 128 GB of DDR3 DRAM. Each server uses as a storage device a 1.5 TB Samsung NVMe from the PM173X series and a Mellanox ConnectX 3 Pro RDMA network card with a bandwidth of 56 Gbps. We limit the buffer cache used by *Tebis*'s storage engine (Kreon) using *cgroups* to be a quarter of the dataset in all cases.

In our experiments, we run the YCSB benchmark [11] workloads Load A and Run A – Run D. Table 3.1 summarizes the operations run during each workload. We use a C++ version of YCSB [32] and we modify it to produce different values according to the KV pair size distribution we study. We run *Tebis* with a total of 32 regions across both servers. Each server serves as *primary* for the 16 and as *backup* for the other 16. Furthermore, each server has 2 spinning threads and 8 worker threads in all experiments. The remaining cores in the server are used for compactions.

In our evaluation, we also vary the KV pair sizes according to the KV sizes proposed by Facebook [7], as shown in Table 3.2. We first evaluate the following workloads where all KV pairs have the same size: Small (S), Medium (M), and Large (L).

Then, we evaluate workloads that use mixes of S, M, and L KV pairs. We use small-dominated (SD) KV size distribution proposed by Facebook [7], as well as two new mixed workloads: *MD* (medium dominated) and *LD* (large dominated). We summarize these KV size distributions in Table 3.2.

	Workload
Load A	100% inserts
Run A	50% reads, 50% updates
Run B	95% reads, 5% updates
Run C	100% reads
Run D	95% reads, 5% inserts

Table 3.1: Workloads evaluated with YCSB. All workloads use a Zipfian distribution except for Run D that use latest distribution.

	KV Size Mix S%-M%-L%	#KV Pairs	Cache per Server (GB)	Dataset Size (GB)
S	100-0-0	100M	0.38	3
M	0-100-0	100M	1.4	11.4
L	0-0-100	100M	11.9	95.2
SD	60-20-20	100M	2.8	23.2
MD	20-60-20	100M	3.3	26.5
LD	20-20-60	100M	7.5	60

Table 3.2: KV size distributions we use for our YCSB evaluation. Small KV pairs are 33 B, medium KV pairs are 123 B, and large KV pairs are 1023 B. We report the record count, cache size per server, and dataset size used with each KV size distribution.

We examine the throughput (KOperations/s), efficiency (KCycles/operation), I/O amplification, and network amplification of *Tebis* for the three following setups: (1) without replication (No Replication), (2) with replication, using our mechanism for sending the index to the *backups* (Send Index), and (3) with replication, where the *backups* perform compactions to build their index (Build Index), which serves as a baseline. In Build Index, servers keeps additionally an L_0 level in memory for their *backup* regions, whereas in Send Index, they do not. For these two setups two be equal, and since we always use the same number of regions, in Build Index we configure each region L_0 size to be half of the L_0 size used in the other two setups.

We measure efficiency in cycles/op and define it as:

$$\textit{efficiency} = \frac{\frac{\textit{CPU\_utilization}}{100} \times \frac{cycles}{s} \times \textit{cores}}{\frac{\textit{average\_ops}}{s}} \; cycles/op,$$

where *CPU_utilization* is the average of CPU utilization among all processors, excluding idle and I/O wait time, as given by *mpstat*. As $cycles/s$ we use the per-core clock frequency. *average_ops*$/s$ is the throughput reported by YCSB, and $cores$ is the number of system cores including hyperthreads.

I/O amplification measures the excess device traffic generated due to compactions (for *primary* and *backup* regions) by *Tebis*, and we define it as:

$$\textit{IO\_amplification} = \frac{\textit{device\_traffic}}{\textit{dataset\_size}},$$

where *device_traffic* is the total number of bytes read from or written to the storage device and *dataset_size* is the total size of all key-value requests issued during the experiment.

Lastly, network amplification is a measure of the excess network traffic generated by *Tebis*, and we define it as:

$$\textit{network\_amplification} = \frac{\textit{network\_traffic}}{\textit{dataset\_size}},$$

where *network_traffic* is the total number of bytes sent by and received from the servers' network cards.

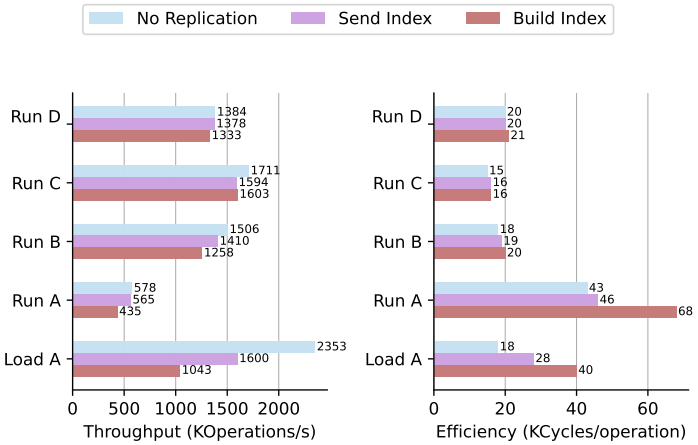


Figure 3.1: Performance and efficiency of *Tebis* for YCSB workloads Load A, Run A – Run D with the SD KV size distribution.

3.1 Experimental Evaluation

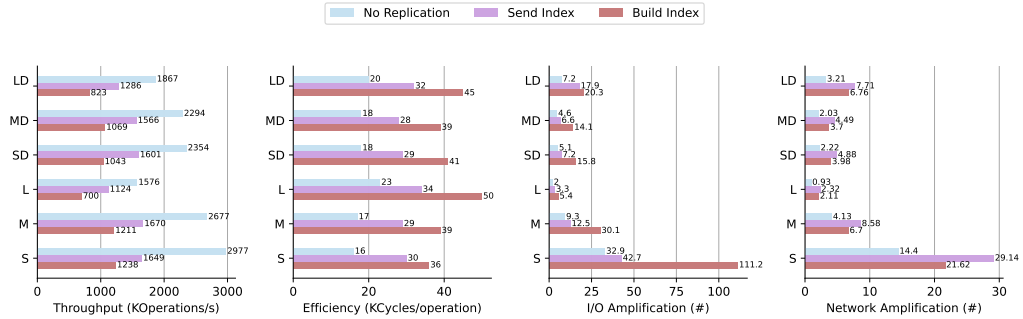
In our evaluation of *Tebis* we answer the following questions:

1. How does our *backup* index construction (Send Index) method compare to performing compactions in *backup* regions (Build Index) to construct the index?
2. Where does *Tebis* spend its CPU cycles? How many cycles does Send Index save compared to Build Index for index construction?
3. How does increasing the growth factor affect *Tebis*?
4. Does Send Index improve performance and efficiency in small-dominated workloads, where KV separation gains diminish?

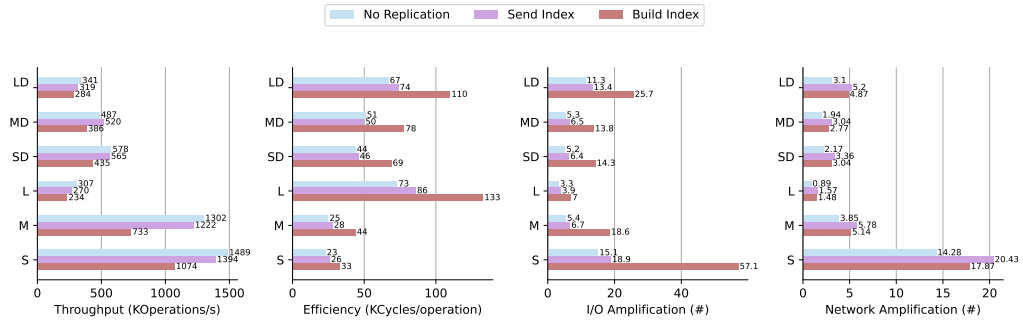
3.1.1 *Tebis* Performance and Efficiency

In Figure 3.1 we evaluate *Tebis* using YCSB workloads Load A and Run A – Run D for the SD [7] workload. Since replication doesn't impact read-dominated workloads, the performance in workloads Run B – Run D remains the same for all three deployments. We focus the rest of our evaluation on the insert and update heavy workloads Load A and Run A.

We run Load A and Run A workloads for all six KV size distributions and with growth factor 4 which minimizes I/O amplification (but not space amplification). We set the L_0 size to 64K keys for the No Replication and Send Index configurations and to 32K keys for the Build Index configuration, since Build Index has twice as many L_0 indexes. We measure throughput, efficiency, and I/O amplification for the three different deployments explained in Section 3. We summarize these results in Figure 3.2. We also report the tail latency in these workloads for the SD KV size distribution in Figure 3.3.



(a) Load A YCSB workload



(b) Run A YCSB workload

Figure 3.2: Throughput, efficiency, I/O amplification, and network amplification for the different key-value size distributions during the (a) YCSB Load A and (b) Run A workloads.

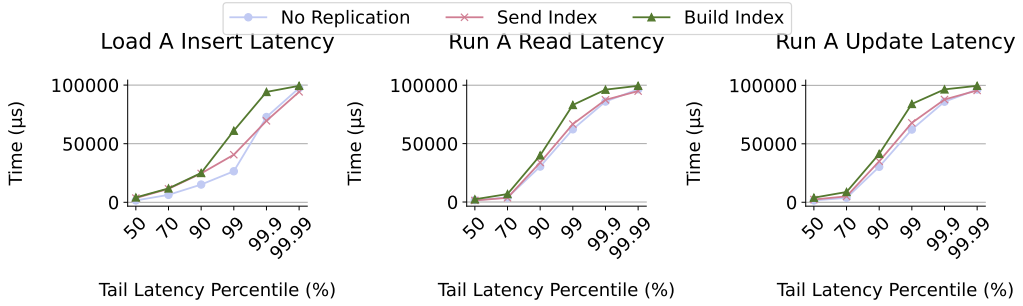


Figure 3.3: Tail latency for YCSB Load A and Run A workload operations using the SD key-value size distribution.

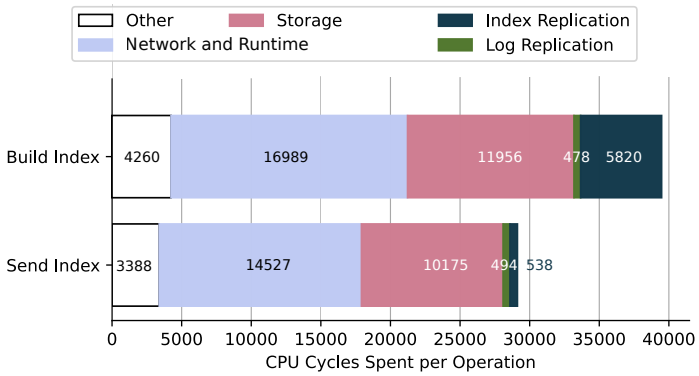


Figure 3.4: Breakdown of cycles spent per operation on network, storage, log replication and index replication.

Compared to Build Index, Send Index increases throughput by $1.1 - 1.7\times$ for all KV size distributions, increases CPU efficiency by $1.2 - 1.6\times$, and reduces I/O amplification by $1.1 - 3.0\times$. Also, it is crucial to notice that compared to No Replication, Build Index increases I/O amplification by $1.6 - 3.4\times$ while Send Index only increases I/O amplification by $1.4 - 1.5\times$, since eliminating compactions in *backup* regions means no additional read traffic for replication. Furthermore, *Tebis* increases CPU efficiency by replacing expensive I/O operations and key comparisons during compactions with a single traversal of the new index segments and hash table accesses to rewrite them.

Sending the *backup* region indexes over the network increases network traffic up to $1.2\times$. This trade-off favors *Tebis* since it pays a slight increase in network traffic for increased efficiency and decreased I/O amplification.

We also measure the tail latency for YCSB workloads Load A and Run A using the SD KV size distribution. As shown in Figure 3.3, Send Index improves the 99, 99.9, and 99.99% tail latencies from 1.1 to $1.5\times$ compared to Build Index for all Load A and Run A operations.

3.1.2 Cycles/Op Breakdown

We run YCSB workloads Load A and Run A and profile *Tebis* using *perf* with call graph tracking enabled. We profile *Tebis* while using Send Index and Build Index configurations. We use the call graph profiles generated to measure where CPU cycles are spent in *Tebis* for Send Index and Build Index. We count the CPU cycles spent on four major parts of our system:

- **Storage:** Cycles spent in KV store operations, excluding replication
- **Network and Runtime:** Cycles spent on detecting, scheduling, and processing client requests, excluding storage and replication

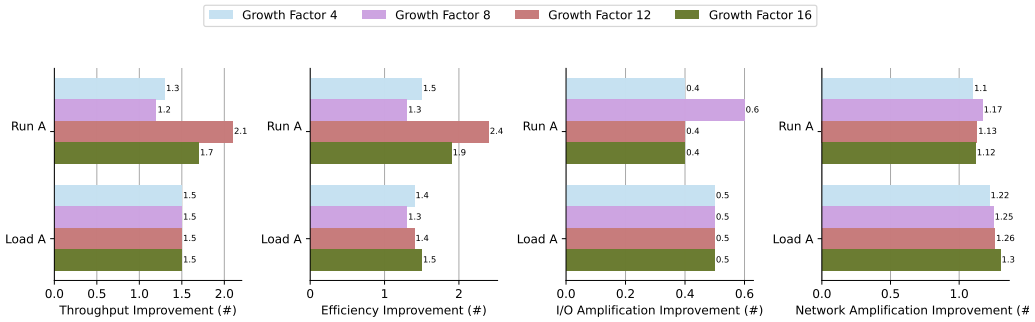


Figure 3.5: Send Index improvement over Build Index for Load A, Run A and different growth factors.

- **Log Replication:** Cycles spent on replicating KV pairs in a *backup*'s value log
- **Index Replication:** Cycles spent to construct indexes for *backup* regions. Send index spends these cycles to rewrite the index segments they receive from *primary region servers*. Build Index spends these cycles on compactions and iterating KV value log segments to insert them into Kreon's L_0 index
- **Other:** All cycles not spent in the above categories

Figure 3.4 summarizes the results of our profiling.

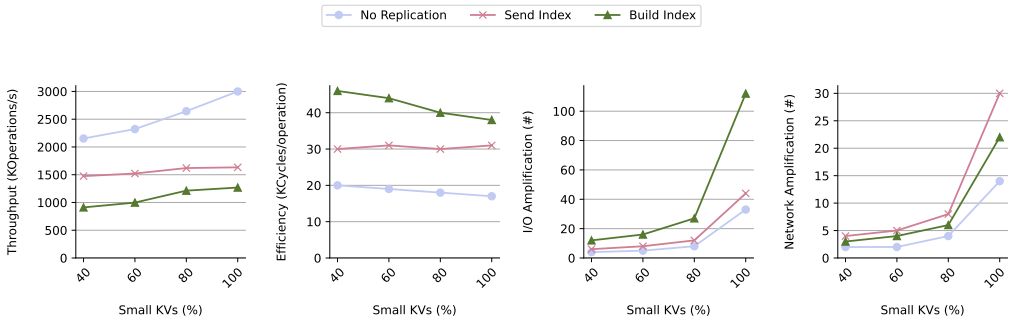
Tebis's Send Index technique requires 28% fewer cycles than performing compactions to construct *backup* indexes. This 28% cycles amount to roughly 12K cycles per operation. They are divided into: 5.5K fewer cycles for replicating *backup* indexes, 2K fewer cycles spent on storage, 2K fewer cycles spent on network and runtime processing, and 2.5K fewer cycles spent on other parts of the system. With the Send Index method, *Tebis region servers* spend $10\times$ fewer cycles on constructing *backup* indexes and $1.36\times$ fewer cycles overall when compared to using the Build Index method.

Sending the *primary* index to *backups* eliminates compactions for backup regions resulting in increased CPU efficiency during *backup* index construction. While *backup region servers* have to rewrite these index segments, the rewriting process only involves hash table lookups without requiring any read I/O, resulting in a more efficient *backup* index construction technique.

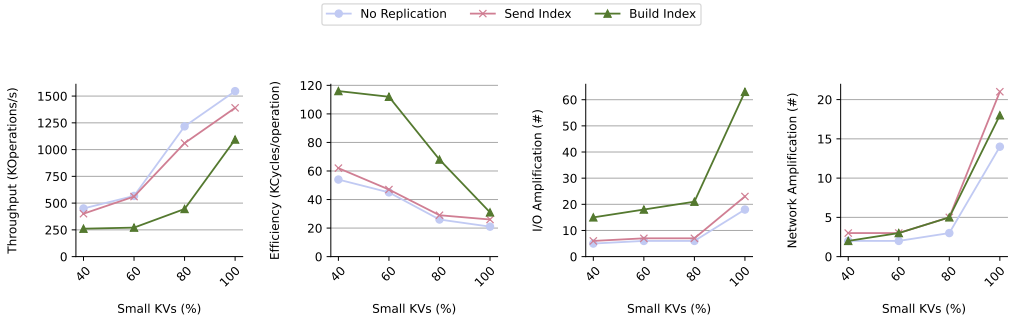
In comparison with Send Index, Build Index also spends 1.16x cycles on network and runtime processing. This is due to additional queuing effects which are a result of the increased load due to *backup* compactions.

3.1.3 Impact of Growth Factor

In Figure 3.5 we show that the gains in performance, efficiency, and I/O amplification during Load A remain constant when increasing the growth factor. However, during Run A, the gains of our Send Index approach compared to Build Index increase. Most notably, with growth factors 12 and 16, the performance improvement is 2.1 and $1.7\times$ respectively.



(a) Load A YCSB workload



(b) Run A YCSB workload

Figure 3.6: Throughput, efficiency, I/O amplification, and network amplification for increasing percentages of small KVs during (a) YCSB Load A and (b) Run A workloads.

Similarly, efficiency is improved by 2.4 and $1.9\times$, and I/O amplification is decreased by 60%.

KV stores intentionally increase growth factor [3, 13] (from 4 to 8-10) and pay the penalty of higher I/O amplification to reduce space. In the Build Index, this penalty is further amplified to two or three times according to the number of replicas per region. However, Send Index eliminates these redundant compactions and allows us to increase the growth factor and thus the space efficiency of LSM tree-based KV stores without sacrificing significantly performance or CPU efficiency.

3.1.4 Small KVs Impact

The KV separation [27, 31, 9] and hybrid placement techniques [25, 38] gains in I/O amplification decrease for $small \leq 33\ B$ KV pairs, which are important for internet-scale workloads [7]. This decrease is because the gains for KV separation of small KV pairs is around $2\times$ [38]. However, if we include also the garbage collection overheads, the gains further diminish making KV separation identical to putting KVs in-place as RocksDB [16] does.

In this experiment, we investigate the impact that small KV pairs percentage has on the efficiency of Send Index method. We set the growth factor to 12 and examine four workloads where we vary small KV pairs percentage to 40%, 60%, 80%, and 100%. In all four cases, we equally divide the remaining percentage between medium and large KV pairs.

As shown in Figure 3.6, Send Index has from 1.2 to $2.3\times$ better throughput and efficiency than Build Index across all workloads. I/O amplification for Build Index increases from 7.4 to $9.3\times$. From the above, we conclude that the Send Index method has significant benefits even for workloads that consist of 80%-90% small KV pairs.

Chapter 4

Related Work

In this section we group related work in the following categories: (a) LSM tree compaction offload techniques, (b) Log and index replication techniques, and (c) efficient RDMA protocols for KV stores:

Compaction offload: Acazoo [18] splits its data into shards and keeps replicas for each shard using the ZAB [19] replication protocol. Acazoo offloads compaction tasks from a shard's primary to one of its replicas. Then, on compaction completion, it reconfigures the system through an election to make the server with the newly compacted data the primary.

Hailstorm [4] is a rack-scale persistent KV store. It uses a distributed filesystem to provide a global namespace and scatters SSTs across the rack (in a deterministic way). Thus it can scale its I/O subsystem similar to HBase/HDFS [1]. Unlike HBase, it schedules compaction tasks to other servers through the global namespace offered by the distributed filesystem substrate.

Unlike these systems, *Tebis* can efficiently keep both the primary and backup indexes up to date through the send index operation by using RDMA to perform efficient bulk network transfers.

Log and index replication techniques: Rose [33] is a distributed replication engine that targets racks with hard disk drives and TCP/IP networking where device I/O is the bottleneck. In particular, it replicates data using a log and builds the replica index by applying mutations in an LSM tree index. The LSM tree removes random reads for updates and always performs large I/Os. *Tebis* shares the idea of Rose to use the LSM tree to build an index at the replica. However, it adapts its design for racks that use fast storage devices and fast RDMA networks where the CPU is the bottleneck. It does this by sending and rewriting the index and removing redundant compactions at the *backups*.

Tailwind [36] is a replication protocol that uses RDMA writes for data movement, whereas for control operations, it uses conventional RPCs. The primary server transfers log records to buffers at the backup server by using one-sided RDMA writes. Backup servers are entirely passive; they flush their RDMA buffers to storage periodically when the primary requests it. They have implemented and evaluated their protocol on RAMCloud, a scale-out in-memory KV store. Tailwind improves throughput and latency compared to RAMCloud. *Tebis* adopts Tailwind's replication protocol for its value log but

further proposes a method to keep a backup index efficiently.

Active-Memory [39] is a primary-backup replication protocol that, like *Tebis*, identifies the CPU and not the network traffic as the main bottleneck in modern data serving systems. Active-Memory takes advantage of RDMA in order to have the primary perform transaction updates directly in the memory of their backups. The primary takes advantage of the in-order delivery guarantee of RDMA connected QPs to write an undo log entry in each backup before altering its memory, to make sure that the backups recoverable in case of a primary failure. While Active-Memory focuses on transactional in-memory databases, *Tebis* is a persistent rack-scale LSM KV store where the main performance bottleneck is the compactions required to construct the multi-level LSM tree index. Because of this, *Tebis* focuses on taking advantage of excess network bandwidth in order to eliminate compactions when building the backup LSM tree indexes.

Efficient RDMA protocols for KV stores: Kalia *et al.* [23] analyze different RDMA operations and show that one-sided RDMA write operations provide the best throughput and latency metrics. *Tebis* uses one-sided RDMA write operations to build its protocol.

A second parameter is whether the KV store supports fixed or variable size KVs. For instance, HERD [22], a hash-based KV store, uses *RDMA writes* to send requests to the server, and *RDMA send* messages to send a reply back to the client. Send messages require a fixed maximum size for KVs. *Tebis* uses only RDMA writes and appropriate buffer management to support arbitrary KV sizes. HERD uses unreliable connections for RDMA writes, and an unreliable datagram connection for RDMA sends. Note that they decide to use RDMA send messages and unreliable datagram connections because RDMA write performance does not scale with the number of outbound connections in their implementation. In addition, they show that unreliable and reliable connections provide almost the same performance. *Tebis* uses reliable connections to reduce protocol complexity and examines their relative overhead in persistent KV stores. We have not detected scalability problems yet.

Other in-memory KV stores [29, 14, 37] use one-sided RDMA reads to offload read requests to the clients.

For instance, Pilaf [29] argues that read requests dominate in datacenters and therefore focuses on using RDMA to increase the CPU efficiency of those requests. In more detail, Pilaf clients use RDMA read operations to read a pointer to a KV from its hash table and then to read the KV. RDMA read is a one-sided operation, meaning that Pilaf nodes do not take part in the communication required to complete a client's read request.

FaRM [14] makes use of RDMA read to offload read requests to the clients, similarly to Pilaf. FaRM stores KV pairs in a hash table and clients can use RDMA read to retrieve them. FaRM design a new hashing scheme that tries to minimize the number of RDMA read operations required for a client to retrieve a KV.

In contrast to Pilaf and FaRM, *Tebis* does not use RDMA reads since lookups in LSM tree-based systems are complex. Typically, lookups consist of multiple accesses to the devices to fetch data. These data accesses must also be synchronized with compactions.

While FaRM clients use RDMA read to read data from its hash table, they send insert requests to the server using RDMA write. Clients write these requests in a circular buffer

and use another RDMA write to advance the circular buffer's tail. The sender keeps a copy of the buffer's head pointer, which the receiver updates in order to make space available to the sender. In *Tebis* we also use RDMA write to send insert requests. However, we quantize our circular buffer into fixed size segments so that the sender and the receiver can both determine the next message location, without having to resort to updating the receiver's pointers from the sender and vice-versa. These differences mean that *Tebis* uses a single RDMA Write for each insert request, since it doesn't need to update any pointers on the receiver's side, while FaRM makes better use of its communication buffer space, since it doesn't have to pad messages.

Chapter 5

Conclusions & Future Work

In this paper, we design *Tebis*, a replicated persistent LSM-based KV store that targets racks with fast storage devices and fast network (RDMA). *Tebis* implements an RDMA write-based client-server protocol and replicates its data using an efficient RDMA write-based primary-backup protocol. *Tebis* identifies the CPU instead of the network as the bottleneck and because of it takes a novel approach to keep an up-to-date index at the *backups* and avoid rebuilding it in case of a failure. Instead of performing compactions at the *backup* servers (Build Index) *primary* sends a pre-built index after each level compaction (Send Index), trading a slight increase in network traffic for increased CPU efficiency and decreased I/O amplification. *Tebis* implements an efficient index rewrite mechanism at the *backups*, which is used to translate the *primary* index's pointers into valid *backup* index pointers. Compared to Build Index, we find that Send Index increases throughput by up to $1.7\times$, CPU efficiency by up to $1.6\times$, decreases I/O amplification by up to $3.0\times$, and decreases tail latency by up to $1.5\times$ during YCSB Load A and Run A workloads.

Our approach enables KV stores to operate with larger growth factors in order to save space (6% space saved by increasing the growth factor from 10 to 16), without inducing significant CPU overhead. Furthermore, we show that Send Index provides significant benefits even in workloads where small KVs account for as much as 90% of the total operations. We believe that the Send Index technique can be adopted by state-of-the-art replicated LSM-based KV stores to increase their CPU and space efficiency.

Lastly we have identified the following two areas for future work. First, offloading or parallelizing compactions within the rack in order to better distribute their CPU overhead and limit their effect on the tail latency observed by the client.

Second, making use of backup copies to serve read and scan requests. In our current design, all client operations are served from the primary copy of a region. However, in cases where one the servers that has a backup copy has a lower load than the primary server, it could be beneficial for the backup server to serve read and scan requests in order to alleviate the pressure on the primary server. Since *Tebis* has a complete LSM tree index for all the out of memory levels, serving read and scan requests from storage is straight forward. A more in-depth approach is required if one is to ensure that a more recent copy is not in the in-memory L_0 tree.

Bibliography

[1] Apache. Hbase. `https://hbase.apache.org/`, 2018.

[2] Aurelius. Titandb, June 2012.

[3] Nikos Batsaras, Giorgos Saloustros, Anastasios Papagiannis, Panagiota Fatourou, and Angelos Bilas. Vat: Asymptotic cost analysis for multi-level key-value stores, 2020.

[4] Laurent Bindschaedler, Ashvin Goel, and Willy Zwaenepoel. Hailstorm: Disaggregated compute and storage for distributed lsm-based databases. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '20, page 301–316, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Dhruba Borthakur et al. Hdfs architecture guide. *Hadoop apache project*, 53(1-13):2, 2008.

[6] Navin Budhiraja, Keith Marzullo, Fred B. Schneider, and Sam Toueg. Distributed systems (2nd ed.). chapter The Primary-backup Approach, pages 199–216. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 1993.

[7] Zhichao Cao, Siying Dong, Sagar Vemuri, and David H.C. Du. Characterizing, modeling, and benchmarking rocksdb key-value workloads at facebook. In *18th USENIX Conference on File and Storage Technologies*, FAST '16, pages 209–223, Santa Clara, CA, February 2020. USENIX Association.

[8] CARV-ICS. Kreon. `https://github.com/CARV-ICS-FORTH/kreon`, 2021.

[9] Helen H. W. Chan, Yongkun Li, Patrick P. C. Lee, and Yinlong Xu. Hashkv: Enabling efficient updates in kv storage via hashing. In *Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference*, USENIX ATC '18, pages 1007–1019, Berkeley, CA, USA, 2018. USENIX Association.

[10] Kristina Chodorow. *MongoDB: The Definitive Guide*. O'Reilly Media, second edition, 5 2013.

[11] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. Benchmarking cloud serving systems with ycsb. In *Proceedings of the 1st ACM Symposium on Cloud Computing*, SoCC '10, pages 143–154, New York, NY, USA, 2010. ACM.

[12] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon's highly available key-value store. *ACM SIGOPS operating systems review*, 41(6):205–220, 2007.

[13] Siying Dong, Mark Callaghan, Leonidas Galanis, Dhruba Borthakur, Tony Savor, and Michael Strum. Optimizing space amplification in rocksdb. In *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. www.cidrdb.org, 2017.

[14] Aleksandar Dragojević, Dushyanth Narayanan, Orion Hodson, and Miguel Castro. Farm: Fast remote memory. In *Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation*, NSDI'14, pages 401–414, Berkeley, CA, USA, 2014. USENIX Association.

[15] Facebook. Blobdb. `http://rocksdb.org/`, 2018. Accessed: November 15, 2021.

[16] Facebook. Rocksdb. `http://rocksdb.org/`, 2018.

[17] Yixiao Gao, Qiang Li, Lingbo Tang, Yongqing Xi, Pengcheng Zhang, Wenwen Peng, Bo Li, Yaohui Wu, Shaozong Liu, Lei Yan, Fei Feng, Yan Zhuang, Fan Liu, Pan Liu, Xingkui Liu, Zhongjie Wu, Junping Wu, Zheng Cao, Chen Tian, Jinbo Wu, Jiaji Zhu, Haiyong Wang, Dennis Cai, and Jiesheng Wu. When cloud storage meets RDMA. In *18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21)*, pages 519–533. USENIX Association, April 2021.

[18] Panagiotis Garefalakis, Panagiotis Papadopoulos, and Kostas Magoutis. Acazoo: A distributed key-value store based on replicated lsm-trees. In *2014 IEEE 33rd International Symposium on Reliable Distributed Systems*, pages 211–220, 2014.

[19] Patrick Hunt, Mahadev Konar, Flavio P. Junqueira, and Benjamin Reed. Zookeeper: Wait-free coordination for internet-scale systems. In *Proceedings of the 2010 USENIX Conference on USENIX Annual Technical Conference*, USENIXATC'10, pages 11–11, Berkeley, CA, USA, 2010. USENIX Association.

[20] H. V. Jagadish, P. P. S. Narayan, S. Seshadri, S. Sudarshan, and Rama Kanneganti. Incremental organization for data recording and warehousing. In *Proceedings of the 23rd International Conference on Very Large Data Bases*, VLDB '97, pages 16–25, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[21] Jithin Jose, Hari Subramoni, Miao Luo, Minjia Zhang, Jian Huang, Md. Wasi-ur Rahman, Nusrat S. Islam, Xiangyong Ouyang, Hao Wang, Sayantan Sur, and Dhabaleswar K. Panda. Memcached design on high performance rdma capable interconnects. In *Proceedings of the 2011 International Conference on Parallel Processing*, pages 743–752, 2011.

[22] Anuj Kalia, Michael Kaminsky, and David G. Andersen. Using rdma efficiently for key-value services. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, pages 295–306, New York, NY, USA, 2014. ACM.

[23] Anuj Kalia, Michael Kaminsky, and David G. Andersen. Design guidelines for high performance rdma systems. In *Proceedings of the 2016 USENIX Conference on Usenix Annual Technical Conference*, pages 437–450, 2016.

[24] Avinash Lakshman and Prashant Malik. Cassandra: A decentralized structured storage system. *SIGOPS Oper. Syst. Rev.*, 44(2):35–40, April 2010.

[25] Yongkun Li, Zhen Liu, Patrick P. C. Lee, Jiayu Wu, Yinlong Xu, Yi Wu, Liu Tang, Qi Liu, and Qiu Cui. Differentiated key-value storage management for balanced i/o performance. In *2021 USENIX Annual Technical Conference (USENIX ATC '21)*, pages 673–687. USENIX Association, July 2021.

[26] Todd Lipcon, David Alves, Dan Burkert, Jean-Daniel Cryans, Adar Dembo, Mike Percy, Silvius Rus, Dave Wang, Matteo Bertozzi, Colin Patrick McCabe, et al. Kudu: Storage for fast analytics on fast data. *Cloudera, inc*, 28, 2015.

[27] Lanyue Lu, Thanumalayan Sankaranarayana Pillai, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. Wisckey: Separating keys from values in ssd-conscious storage. In *14th USENIX Conference on File and Storage Technologies (FAST 16)*, pages 133–148, Santa Clara, CA, February 2016. USENIX Association.

[28] Yoshinori Matsunobu, Siying Dong, and Herman Lee. Myrocks: Lsm-tree database storage engine serving facebook's social graph. *Proc. VLDB Endow.*, 13(12):3217–3230, August 2020.

[29] Christopher Mitchell, Yifeng Geng, and Jinyang 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, Berkeley, CA, USA, 2013. USENIX Association.

[30] Patrick O'Neil, Edward Cheng, Dieter Gawlick, and Elizabeth O'Neil. The log-structured merge-tree (lsm-tree). *Acta Inf.*, 33(4):351–385, June 1996.

[31] Anastasios Papagiannis, Giorgos Saloustros, Pilar González-Férez, and Angelos Bilas. An efficient memory-mapped key-value store for flash storage. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '18, pages 490–502, New York, NY, USA, 2018. ACM.

[32] Jinglei Ren. Ycsb-c. `https://github.com/basicthinker/YCSB-C`, 2016.

[33] Russell Sears, Mark Callaghan, and Eric Brewer. rose: Compressed, log-structured replication. *Proc. VLDB Endow.*, 1(1):526–537, August 2008.

[34] Chen Shen, Youyou Lu, Fei Li, Weidong Liu, and Jiwu Shu. Novkv: Efficient garbage collection for key-value separated lsm-stores. October 2020.

[35] Arjun Singhvi, Aditya Akella, Maggie Anderson, Rob Cauble, Harshad Deshmukh, Dan Gibson, Milo M. K. Martin, Amanda Strominger, Thomas F. Wenisch, and Amin Vahdat. Cliquemap: Productionizing an rma-based distributed caching system. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, SIGCOMM '21, page 93–105, New York, NY, USA, 2021. Association for Computing Machinery.

[36] Yacine Taleb, Ryan Stutsman, Gabriel Antoniu, and Toni Cortes. Tailwind: Fast and atomic rdma-based replication. In *Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference*, USENIX ATC '18, pages 851–863, Berkeley, CA, USA, 2018. USENIX Association.

[37] Xingda Wei, Jiaxin Shi, Yanzhe Chen, Rong Chen, and Haibo Chen. Fast in-memory transaction processing using rdma and htm. In *Proceedings of the 25th Symposium on Operating Systems Principles*, pages 87–104, 2015.

[38] Giorgos Xanthakis, Giorgos Saloustros, Nikos Batsaras, Papagiannis Anastasios, and Angelos Bilas. Parallax: Hybrib key-value placement in lsm-based key-value stores. In *Proceedings of the ACM Symposium on Cloud Computing*, SoCC '21, New York, NY, USA, 2021. ACM.

[39] Erfan Zamanian, Xiangyao Yu, Michael Stonebraker, and Tim Kraska. Rethinking database high availability with rdma networks. *Proc. VLDB Endow.*, 12(11):1637–1650, July 2019.