ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΕΠΙΣΤΗΜΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

ΠΑΡΟΥΣΙΑΣΗ / ΕΞΕΤΑΣΗ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Βαρδουλάκης Μιχαήλ

Μεταπτυχιακός Φοιτητής

Τμήμα Επιστήμης Υπολογιστών, Πανεπιστήμιο Κρήτης

Επόπτης Μεταπτυχιακής Εργασίας: Καθηγητής, Α. Μπίλας

Τρίτη, 2 Νοεμβρίου 2021, ώρα 17:00 μ.μ.

Join Zoom Meeting
https://zoom.us/j/96070560075

"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.

University of Crete

Computer Science Department

M.Sc. Thesis

Mixalis Vardoulakis

Master's Thesis Supervisor: Professor, A. Bilas

Tuesday, 2 November 2021, 17:00 p.m.

Join Zoom Meeting
https://zoom.us/j/96070560075

"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×, CPU overhead by up to 1.6×, and memory size needed for the write path by up to 2×, without increasing network bandwidth excessively, and by up to 1.3×. 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.