Τμήμα Μηχ. Η/Υ & Πληροφορικής **Εαρινό εξάμηνο 2022**
Πολυτεχνική Σχολή, Πανεπιστήμιο Ιωαννίνων **Διδάσκων: Σ. Αναστασιάδης**

ΜΥΥ601 Λειτουργικά Συστήματα

Ανακοίνωση: Τετάρτη 9 Μαρτίου, Παράδοση: Παρασκευή, 1 Απριλίου στις 21:00

Εργαστήριο 1:Υλοποίηση πολυνηματικής λειτουργίας σε μηχανή αποθήκευσης δεδομένων

1. Εισαγωγή

Σας δίνεται μια μηχανή αποθήκευσης (storage engine) υλοποιημένη στη γλώσσα C. Σας ζητείται να υλοποιήσετε την πολυνηματική λειτουργία των εντολών put και get που παρέχει η μηχανή αποθήκευσης. Η υλοποίησή σας θα επιτρέπει πολλαπλά νήματα να καλούν τις εντολές put και get ταυτόχρονα. Η μηχανή αποθήκευσης θα πρέπει να εκτελεί τις ταυτόχρονες λειτουργίες σωστά και να διατηρεί στατιστικά του χρόνου εκτέλεσης της κάθε λειτουργίας. Η υλοποίησή σας θα πρέπει να γίνει στην γλώσσα C και να βασίζεται στην βιβλιοθήκη Pthreads του Linux.

1.1 Μηχανή αποθήκευσης βασισμένη στο δέντρο LSM

Ο πηγαίος κώδικας που σας δίνεται υλοποιεί την μηχανή αποθήκευσης **Kiwi** που βασίζεται σε δέντρο log-structured merge (LSM-tree). Οι μηχανές αποθήκευσης είναι σημαντικό μέρος των σύγχρονων υποδομών νέφους καθώς είναι υπεύθυνες για την αποθήκευση και ανάκτηση δεδομένων στις τοπικές συσκευές μιας μηχανής. Ένα κατανεμημένο σύστημα αποθήκευσης χρησιμοποιεί χιλιάδες τέτοιες μηχανές για να πετύχει κλιμακώσιμη και αξιόπιστη λειτουργία (π.χ., Facebook, Google).

Το LSM-tree είναι η δομή δεδομένων στην οποία συχνά βασίζονται οι μηχανές αποθήκευσης. Η παρεχόμενη προγραμματιστική διεπαφή (API) περιλαμβάνει λειτουργίες put και get για ζεύγη κλειδιού-τιμής. Η λειτουργία put δέχεται ως παράμετρο ένα ζεύγος κλειδιού-τιμής που πρέπει να προστεθεί στην δομή. Η λειτουργία get δέχεται ως παράμετρο ένα κλειδί και αιτείται την αντίστοιχη τιμή εφόσον υπάρχει αποθηκευμένο στην δομή ζεύγος κλειδιού-τιμής με το συγκεκριμένο κλειδί, αλλιώς επιστρέφει σφάλμα αν το κλειδί δεν βρεθεί.

Για την γρήγορη αναζήτηση, η δομή διατηρεί τα αποθηκευμένα δεδομένα ταξινομημένα στην μνήμη ή τον δίσκο. Συνήθως, τα πιο πρόσφατα δεδομένα διατηρούνται στην μνήμη σε μια ταξινομημένη δομή που λέγεται memtable και συνήθως υλοποιείται με skip list. Η skip list είναι μια δομή δεδομένων που αποτελείται από πολλαπλές λίστες οργανωμένες σε διαφορετικά επίπεδα για να επιταχύνουν την αναζήτηση κάποιου στοιχείου. Όταν το memtable γεμίσει, μετακινείται στον δίσκο προκειμένου να αδειάσει η μνήμη και να μπορέσει να δεχτεί καινούρια δεδομένα. Κατά το διάστημα που το memtable παραμένει στην μνήμη, υπάρχει ένα αρχείο log που λαμβάνει τα εισερχόμενα δεδομένα και τα αποθηκεύει προσωρινά στον δίσκο για την γρήγορη ανάκτησή τους σε περίπτωση σφάλματος.

Τα αρχεία που αποθηκεύονται στο δίσκο είναι οργανωμένα σε πολλαπλά επίπεδα των οποίων το μέγεθος αυξάνει με γεωμετρική πρόοδο από τα χαμηλότερα προς τα υψηλότερα επίπεδα (0..6). Ένα αρχείο δίσκου στην δομή αυτή ονομάζεται Sorted String Table (sst) επειδή τα δεδομένα είναι ταξινομημένα με βάση τα κλειδιά που περιέχουν συμβολοσειρές. Όταν το memtable μετακινείται στον δίσκο, συγχωνεύεται με τα αρχεία (στο επίπεδο 0 ή 1) των οποίων τα κλειδιά επικαλύπτονται με τα κλειδιά του memtable. Επιπρόσθετα, ενεργοποιείται σύμπτυξη αρχείων (compaction) όταν το πλήθος των αρχείων στο επίπεδο 0 ξεπερνά ένα προκαθορισμένο κατώφλι (π.χ., 4), ή το συνολικό μέγεθος των αρχείων σε ένα επίπεδο ξεπερνάει ένα προκαθορισμένο κατώφλι για το επίπεδο αυτό. Με την σύμπτυξη, τα αρχεία ενός επίπέδου συγχωνεύονται σε ένα αρχείο που προστίθεται στο επόμενο επίπεδο.

1.2 Πολυνηματική υλοποίηση

Στην υλοποίηση που σας δίνεται, υπάρχουν ήδη δύο νήματα στην μηχανή αποθήκευσης. Το πρώτο νήμα καλεί μια ακολουθία λειτουργιών put ή get την μία μετά την άλλη όπως ορίζεται από

την γραμμή εντολών, ενώ το δεύτερο νήμα είναι υπεύθυνο για την σύμπτυξη αρχείων όταν ικανοποιούνται οι απαιτούμενες συνθήκες που αναφέρθηκαν παραπάνω.

Η λειτουργία put αλληλεπιδρά με το memtable προκειμένου να εισαγάγει ένα νέο στοιχείο στην δομή. Αν απαιτείται, το τρέχον memtable συγχωνεύεται με ένα αρχείο sst, που μπορεί να οδηγήσει σε περαιτέρω σύμπτυξη από το ένα επίπεδο στο επόμενο.

Η λειτουργία get κάνει αναζήτηση για το παρεχόμενο κλειδί πρώτα στο memtable και αν δεν βρεθεί εκεί στα αρχεία sst ξεκινώντας από το επίπεδο 0 και πηγαίνοντας σε επόμενα επίπεδα όπως απαιτείται. Είναι δυνατόν να έχει δρομολογηθεί η συγχώνευση ενός memtable, οπότε θα πρέπει και αυτό να υποστεί αναζήτηση πριν τα αρχεία sst. Τα αρχεία στα οποία θα γίνει αναζήτηση επιλέγονται με βάση το εύρος των κλειδιών που περιέχουν, υποθέτοντας ότι είναι περιττό να γίνει αναζήτηση σε ένα sst του οποίου το εύρος κλειδιών δεν περιέχει το κλειδί που αναζητείται. Η αναζήτηση σε ένα αρχείο sst γίνεται ταχύτερη με ένα Bloom filter που περιέχει τους κατακερματισμούς των κλειδιών που έχουν ήδη εισαχθεί στο συγκεκριμένο αρχείο.

Δεδομένης της πολυπλοκότητας του κώδικα, σας ζητείται να σχεδιάσετε την υλοποίηση σε βήματα. Μια τετριμμένη υλοποίηση θα χρησιμοποιούσε μία κλειδαριά για να εφαρμόσει αμοιβαίο αποκλεισμό στις λειτουργίες put και get. Μια βελτίωση θα επέτρεπε πολλαπλούς αναγνώστες ή ένα γραφέα να λειτουργεί κάθε φορά. Επιπλέον βήματα θα χρησιμοποιούσαν διαφορετικές κλειδαριές για το memtable και τα αρχεία sst, έτσι ώστε διαφορετικά νήματα να λειτουργούν ταυτόχρονα εφόσον δεν τροποποιούν το ίδιο μέρος του συστήματος το ίδιο χρονικό διάστημα. Η λύση θα πρέπει να λαμβάνει υπόψη ότι υπάρχει ήδη ένα νήμα που είναι υπεύθυνο για τις συγχωνεύσεις που τροποποιούν ασύγχρονα στο παρασκήνιο τα αρχεία sst.

Είστε υπεύθυνοι να τεκμηριώσετε τις αλλαγές που κάνετε στον κώδικα τόσο με σχόλια στον πηγαίο κώδικα που παραδίδετε, όσο και με λεπτομερή περιγραφή στην αναφορά που παραδίδετε. Κώδικας που δεν βγάζει νόημα και δεν τεκμηριώνεται επαρκώς δεν θα ληφθεί υπόψη στην βαθμολόγηση, ή μπορεί να την επηρεάσει αρνητικά.

1.3 Στατιστικά απόδοσης

Η υλοποίηση που σας δίνεται περιλαμβάνει κάποιες πολύ βασικές μετρήσεις της απόδοσης των λειτουργιών read (get) και write (put/add). Σας ζητείται να εξασφαλίσετε ότι λαμβάνονται σωστά οι μετρήσεις της απόδοσης κατά την πολυνηματική υλοποίηση που θα παραδώσετε. Σωστή πολυνηματική λειτουργία σημαίνει ότι οι μεταβλητές που καταγράφουν την απόδοση ενημερώνονται ατομικά από τα διαφορετικά νήματα χρησιμοποιώντας κλειδαριές για αμοιβαίο αποκλεισμό.

Η αναφορά σας θα πρέπει να περιλαμβάνει δοκιμές απόδοσης σε διαφορετικά σενάρια που κρίνετε σκόπιμα. Παραδείγματα είναι φορτία εργασίας που περιέχουν μόνο put ή μόνο get, καθώς και μίγμα put και get σύμφωνα με συγκεκριμένα ποσοστά για κάθε λειτουργία. Μπορείτε να παρουσιάσετε τα αποτελέσματα είτε απλά με screenshot ή καλύτερα με γραφικές παραστάσεις (π.χ., από το excel).

2. Προετοιμασία

Κατεβάστε την εικονική μηχανή MYY601Lab1.zip (1.2GB) και αποσυμπιέστε την στο τοπικό σκληρό δίσκο σας. Η εικονική μηχανή τρέχει την έκδοση 10 (“Buster”) του Debian Linux 64-bit. Κατεβάστε και εγκαταστήστε τον VMware Player v15 στο μηχάνημά σας. Η μηχανή είναι ρυθμισμένη να χρησιμοποιεί 1 επεξεργαστικό πυρήνα και 1GB RAM, αλλά αυτό αλλάζει εύκολα από τις ρυθμίσεις της εικονικής μηχανής. Για να εκμεταλλευτείτε το παραθυρικό περιβάλλον, εξασφαλίστε ότι τρέχετε την εικονική μηχανή σε κατάσταση πλήρους οθόνης πατώντας το κατάλληλο εικονίδιο πάνω αριστερά στον VMware Player. Μπορείτε να κάνετε login ως απλός χρήστης με το όνομα myy601 και το ίδιο ως κωδικό. Αν χρειάζεται να εγκαταστήσετε επιπλέον πακέτα στο σύστημα Linux της εικονικής μηχανής, μπορείτε να κάνετε login ως root (με κωδικό myy601) και να τρέξετε **apt install <package>**. Η εικονική μηχανή έχει εγκατεστημένη την γλώσσα προγραμματισμού C και τον φυλλομετρητή Firefox. Για διευκόλυνσή σας, μπορείτε να κάνετε copy-paste κείμενο και αρχεία μεταξύ του λειτουργικού συστήματος της εικονικής

μηχανής και του λειτουργικού συστήματος του μηχανήματός σας, π.χ., για να μεταφέρετε τον κώδικα που έχετε γράψει κατά την υποβολή με turnin.

Φρεσκάρετε τις γνώσεις σας στη γλώσσα C και εξασφαλίστε ότι κατανοείτε τις κλήσεις συναρτήσεων της βιβλιοθήκης Pthreads για την δημιουργία και τον συγχρονισμό των νημάτων. Επιπλέον, εξοικειωθείτε με τον εκσφαλματωτή **gdb** προκειμένου να κάνετε βηματική εκτέλεση και να δείτε τα τρέχοντα περιεχόμενα διαφόρων μεταβλητών (Cheatsheet & Manual). Θα χρειαστείτε να χρησιμοποιήσετε βασικά εργαλεία της εικονικής μηχανής για να ανοίξετε τα αρχεία του πηγαίου κώδικα (π.χ., nedit, xemacs). Ανοίξτε ένα τερματικό και μετακινηθείτε στον κατάλογο **~/kiwi/kiwi-source** που περιέχει τον πηγαίο κώδικα της μηχανής αποθήκευσης και ένα απλό benchmark (**kiwi-bench**). Δημιουργήστε τα εκτελέσιμα εκτελώντας **make all** και μετακινηθείτε στον κατάλογο **bench** που περιέχει το benchmark. Μπορείτε να εισάγετε ένα πλήθος (π.χ., 100.000) στοιχείων στην μηχανή με την εντολή **./kiwi-bench write 100000** και στην συνέχεια να τα διαβάσετε με την εντολή **./kiwi-bench read 100000**.

3. Εργασία

Η εργασία σας ζητά να υλοποιήσετε τις λειτουργίες put και get με δυνατότητα εκτέλεσης σε διαφορετικά νήματα.

i. Πρώτα χρειάζεστε μια βασική κατανόηση του μονοπατιού που ακολουθούν οι λειτουργίες put (**db_add**) και get (**db_get**). Δεν χρειάζεται να καταλάβετε όλες τις λεπτομέρειες του κώδικα αλλά να αποκτήσετε μια βασική ιδέα για τον τρόπο με τον οποίο μια λειτουργία εξυπηρετείται είτε από το memtable ή από τα αρχεία sst. Είναι χρήσιμο να πλοηγηθείτε στον κώδικα με την χρήση την τεκμηρίωσης που διατίθεται online στον επόμενο σύνδεσμο (Τεκμηρίωση κώδικα).
   a. Στην επιλογή Data Structures θα βρείτε τον ορισμό σημαντικών δομών, όπως οι **_db**, **_memtable**, **_skiplist**, **_sst**. Ειδικότερα στην δομή **_sst** υπάρχουν ήδη μεταβλητές συγχρονισμού τύπου **pthread_mutex** ή **pthread_cond** που χρειάζεται να έχετε υπόψη σας όταν βελτιώνετε τον ταυτοχρονισμό του συστήματος.
   b. Στην επιλογή Files μπορείτε να βρείτε σημαντικά αρχεία τύπου κεφαλίδας που ορίζουν την διεπαφή σημαντικών ενοτήτων του κώδικα. Ειδικότερα, θα πρέπει να εξοικειωθείτε με το αρχείο **db.h** και τις συναρτήσεις που θα τροποποιήσετε (π.χ., **db_add**, **db_get**), και τις αντίστοιχες συναρτήσεις στο **memtable.h**, **skiplist.h**, και **sst.h**.
   c. Το αρχείο **config.h** περιέχει σημαντικούς ορισμούς των macros που καθορίζουν την συμπεριφορά του συστήματος (π.χ., **BACKGROUND_MERGE**).

ii. Εκτελέσετε το benchmark **kiwi-bench** με διαφορετικά πλήθη εισαγόμενων και αναζητούμενων στοιχείων και να παρακολουθήσετε τα μηνύματα στο τερματικό. Τρέξετε τον κώδικα στο gdb (**gdb kiwi-bench**) για να παρακολουθήσετε τις συναρτήσεις που εκτελούνται κατά τις διαφορετικές εκτελέσεις.

iii. Μπορείτε να ξεκινήσετε πολλαπλά νήματα στο benchmark (με κατάλληλη επέκταση του **bench.c**) και να καλέσετε λειτουργίες get που δεν δημιουργούν συγκρούσεις στις δομές. Αυτό όμως δεν μπορείτε να το κάνετε εύκολα με τις λειτουργίες put επειδή θα οδηγήσει σε segmentation fault και πιθανόν να αφήσει την δομή σε ασυνεπή μορφή. Εάν διαπιστώσετε πρόβλημα στα αποθηκευμένα δεδομένα, θα πρέπει να σβήσετε τον κατάλογο **testdb** με **make clean** ή **rm -rf testdb** στον κατάλογο **bench**.

iv. Προσδιορίστε το επίπεδο ταυτοχρονισμού που θέλετε να πετύχετε. Αυτό εξαρτάται από τα μέρη της δομής που εμπλέκονται στις διάφορες λειτουργίες και το αν διαβάζουν μόνο τα δεδομένα ή επιπλέον τροποποιούν τα δεδομένα. Μια απλή προσέγγιση είναι να επιτρέψετε να τρέχει μία μόνο λειτουργία τη φορά και να το εξασφαλίσετε αυτό με αμοιβαίο αποκλεισμό. Θα πρέπει να βελτιώσετε αυτή την τετριμμένη λύση προσθέτοντας επιπλέον ταυτοχρονισμό με την χρήση πολλαπλών κρίσιμων περιοχών και πειραματισμό με συγχρονισμό reader-writer στα διάφορα επίπεδα του LSM-tree (π.χ., skip list, sst, συγχώνευση, σύμπτυξη, κλπ).

v. Μπορείτε να επεκτείνετε τα ορίσματα της γραμμής εντολής του αρχείου **bench.c** για να προσδιορίσετε το μίγμα λειτουργιών put και get καθώς και το ποσοστό από τον κάθε τύπο (π.χ., 50-50, 10-90, κλπ).

vi. Θα πρέπει να διατηρείτε στατιστικά της απόδοσης των λειτουργιών (π.χ., χρόνο απόκρισης, ρυθμοαπόδοση) με τον απαραίτητο συγχρονισμό για τον αμοιβαίο αποκλεισμό μεταξύ των διαφορετικών νημάτων που ενημερώνουν τις αντίστοιχες μεταβλητές. Στο τέλος, μπορείτε να τυπώνετε τα στατιστικά που μετρήσατε.

Η βαθμολόγηση θα γίνει με βάση τις λειτουργίες που έχετε υλοποιήσει, την ποιότητα του κώδικα που έχετε προσθέσει και την κατανοητή τεκμηρίωσή του στην αναφορά που παραδίδετε.

4. Τι θα παραδώσετε

Μπορείτε να προετοιμάσετε την λύση ατομικά ή σε ομάδες μέχρι τριών ατόμων. Η υποβολή θα γίνει από ένα μέλος της ομάδας μόνο. Υποβολή μετά την λήξη της προθεσμίας οδηγεί σε αφαίρεση 10% του βαθμού ανά ημέρα μέχρι 50%. Για παράδειγμα, αν στείλετε την λύση σας 1 ώρα μετά την προθεσμία, ο μέγιστος βαθμός που μπορείτε να λάβετε είναι 9 στα 10. Αν στείλετε την λύση μία εβδομάδα μετά την προθεσμία, ο μέγιστος βαθμός σας πέφτει σε 5 από τα 10. Υποβάλετε την λύση σας εγκαίρως με την εντολή

/usr/local/bin/turnin lab1_22@myy601 Report.pdf kiwi-source.zip

Θα βαθμολογηθείτε σύμφωνα με την περιγραφή που περιέχεται στο αρχείο **Report.pdf** και την συνοδευόμενη υλοποίηση στο αρχείο **kiwi-source.zip**. Η αναφορά σας περιλαμβάνει τα ονόματα των μελών της ομάδας και τον αριθμό μητρώου σας. Επιπλέον περιγράφει τι έχετε υλοποιήσει και πώς το κάνατε περιγράφοντας μία-μία τις γραμμές κώδικα που προσθέσατε ή τροποποιήσατε. Γραμμές κώδικα που αλλάξατε ή προσθέσατε αλλά δεν περιλαμβάνονται στην αναφορά δεν μετράνε στον τελικό βαθμό ή μπορεί να μετρήσουν αρνητικά αν δείχνουν βασικές ελλείψεις στις προγραμματιστικές γνώσεις σας.

Ενθαρρύνεστε να προσθέσετε σχόλια στον πηγαίο κώδικα που γράψατε προκειμένου να διευκολύνετε τον εξεταστή σας να κατανοήσει την εργασία σας. Θα πρέπει επιπλέον να συμπεριλάβετε στην αναφορά σας την έξοδο της τελικής εντολής **make** καθώς και την έξοδο από την εκτέλεση με διάφορες παραμέτρους. Στην έξοδο εκτέλεσης συμπεριλάβετε μόνο την εντολή που χρησιμοποιήσατε για να τρέξετε το **kiwi-bench** και τις μετρήσεις απόδοσης χωρίς τις λεπτομερείς πληροφορίες αποσφαλμάτωσης που παράγει η εκτέλεση, εκτός και αν υπάρχει κάτι συγκεκριμένο που θέλετε να δείξετε. **Αν οι έξοδοι εκτέλεσης που περιλαμβάνονται στην αναφορά δεν αντιστοιχούν στον κώδικα που παραδίδετε, θα μηδενιστεί ολόκληρη η άσκηση για ευνόητους λόγους.**

Το αρχείο **kiwi-source.zip** που υποβάλετε είναι ένα συμπιεσμένο αρχείο με τα τροποποιημένα αρχεία. Για κάθε αρχείο που αλλάξατε θα στείλετε μόνο το τροποποιημένο και όχι το αρχικό. Θα πρέπει να εκτελέσετε **make clean** στον κατάλογο **kiwi-source** πριν δημιουργήσετε το αρχείο zip για να μην συμπεριλάβετε αρχεία εκτελέσιμου κώδικα (.o, .a, κλπ). Ο κώδικας που παραδίδετε θα πρέπει να μπορεί να μεταγλωττιστεί με τα εργαλεία που περιλαμβάνονται στην εικονική μηχανή που σας δίνεται.

Βιβλιογραφία

[1] https://github.com/google/leveldb