CAS CS 561: Data Systems Architectures
Data-intensive Systems and Computing Lab
Department of Computer Science
College of Arts and Sciences, Boston University
http://bu-disc.github.io/CS561/

CS561 Spring 2024 - Research Project

Title: *Finding the optimal granularity of index*

Background:
The log-structured merge tree (LSM-tree) [1] is a key-value store widely adopted by many data-intensive applications because they offer fast ingestion and competitive reads. In order to further enhance the read performance, LSM-trees employ metadata such as Bloom filters and indexes and store useful blocks in the block cache. Since memory is a limited resource, the decision which blocks to cache is crucial for lookup performance. However, the state of the art treats all blocks equally. Furthermore, even though the benefit of each metadata varies per SST files based on the workload statistics, the metadata, such as bits-per-key of Bloom filters and granularity of index, is chosen beforehand and hard to modify at the run times.

Objective:

The objective of this project is to find out the optimal granularity of the index blocks in LSM-trees at the run time. The workflow for this project is as follows.

(a) Get familiarized with the query procedure using block cache in the vanilla implementation of the LSM-trees [2].
(b) Design the cost model of the lookup that can find the optimal granularity of index blocks based on the workload statistics.
(c) Design an algorithm that can dynamically change the granularity of the index block.
(d) Implement the proposed solutions on top of RocksDB, and analyze their performance with respect to the state-of-the-art.

Responsible Mentor: *Juhyoung Mun (jmun@bu.edu)*

References:
[1] Patrick E. O'Neil, Edward Cheng, Dieter Gawlick, Elizabeth J. O'Neil. The Log-Structured Merge-Tree (LSM-Tree). Acta Inf. 33(4), (1996)
[2] Facebook. RocksDB. http://github.com/facebook/rocksdb.