Tutorial on LSM-trees

Database System Technology - CSC443H1

Niv Dayan - Feb 2, 2023

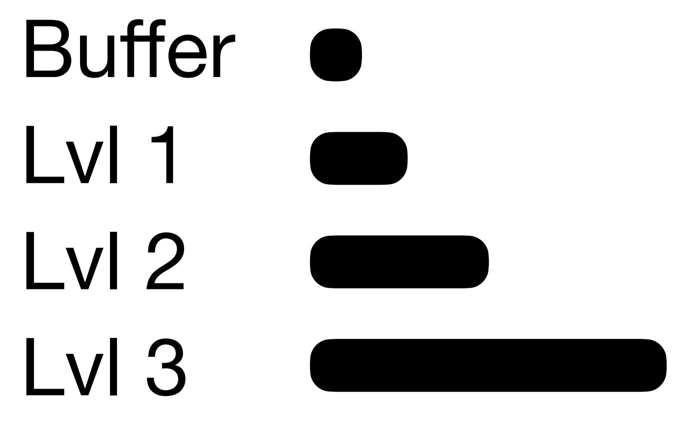
We will start at 3:10 pm

Question 1

We would like to choose KV-store engine between BerkeleyDB vs. RocksDB for a workload consisting of 10% random gets and 90% random puts. BerkeleyDB uses a B-tree. RocksDB uses a leveled LSM-tree with size ratio T=10 and a buffer size P=64MB. Assume N=2^40 and B=2^7. All internal nodes fit in memory. We're using a disk drive. What's your choice?

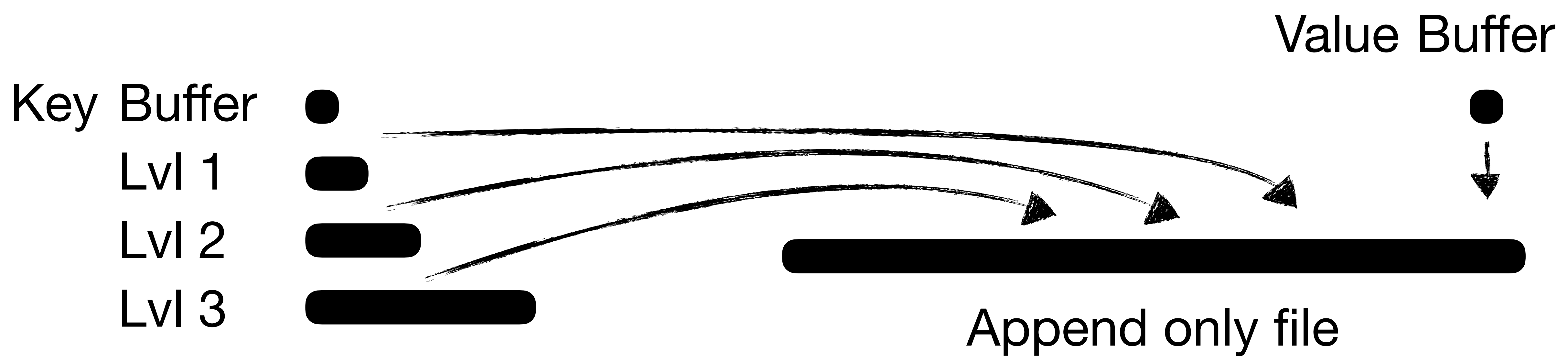
Question 2

A friend tells you they switched from using a B-tree to a basic LSM-tree (with size ratio 2), yet write-amplification actually increased. There are $N=2^{40}$ entries and $B=2^5$. Explain why this happened. Identify three tuning options for reducing write-amplification and the trade-off of each one of them.



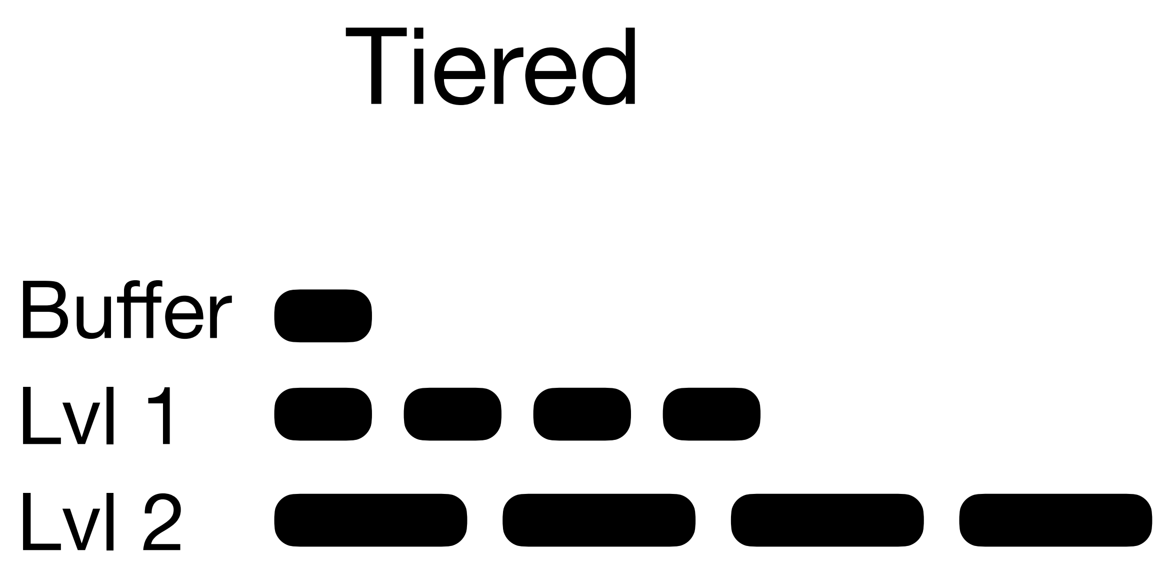
Question 3

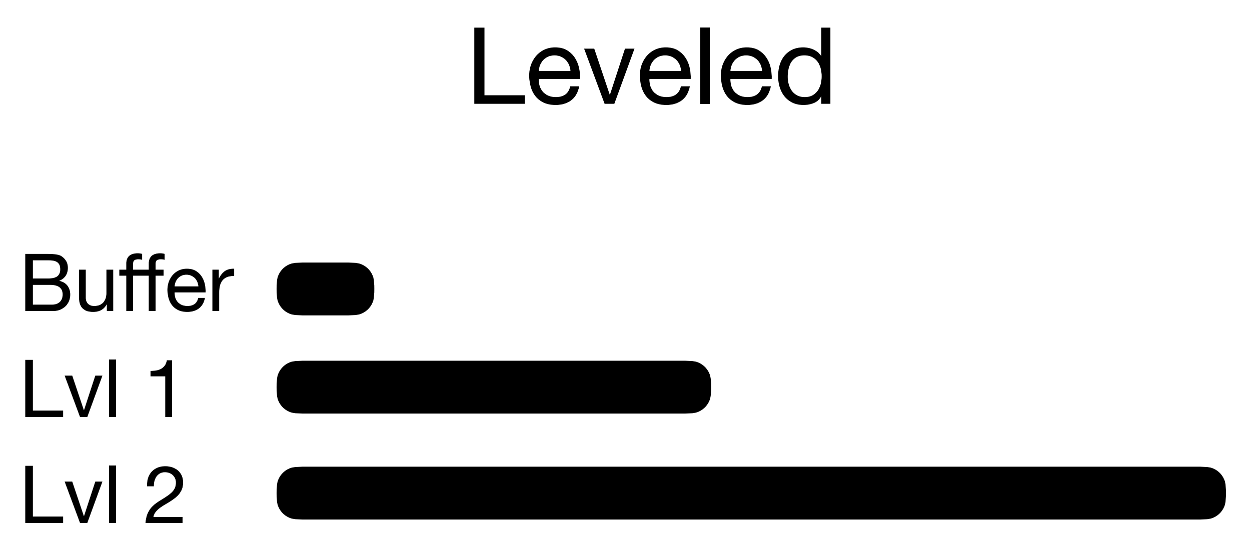
In a clustered LSM-trees, the values are stored within the LSM-tree alongside their keys. Another approach is an unclustered LSM-tree, which stores values in an append-only file and indexes them using key-pointer pairs from within the LSM-tree. What's the impact of clustered vs. unclustered LSM-trees on put/get/scan performance. Propose adjusted cost models assuming a basic LSM-tree (size ratio T=2). Hint: let K be the number of key-pointer pairs fitting into a page, and let B be the number of key-value pairs fitting in a page. Based on your models, identify cases where each of these approaches shines. Assume a 1 page buffer.



Question 4

An LSM-tree can store multiple versions for a given entry, where only the most recent is considered up-to-date and the rest are obsolete. The obsolete entries are eventually discarded during compaction, but meanwhile they consume space. This phenomenon is known as space-amplification, defined as: (physical space taken up) / (logical data size). Quantify space-worst-case amplification for a leveled vs. tiered LSM-tree with size ratio T between any two adjacent levels. Assume all levels are totally full.


Tiered
Buffer
Lvl 1
Lvl 2


Leveled
Buffer
Lvl 1
Lvl 2