RocksDB Festival

RocksDB Architecture and Operation

Supported by IITP, StarLab.

July 12, 2021

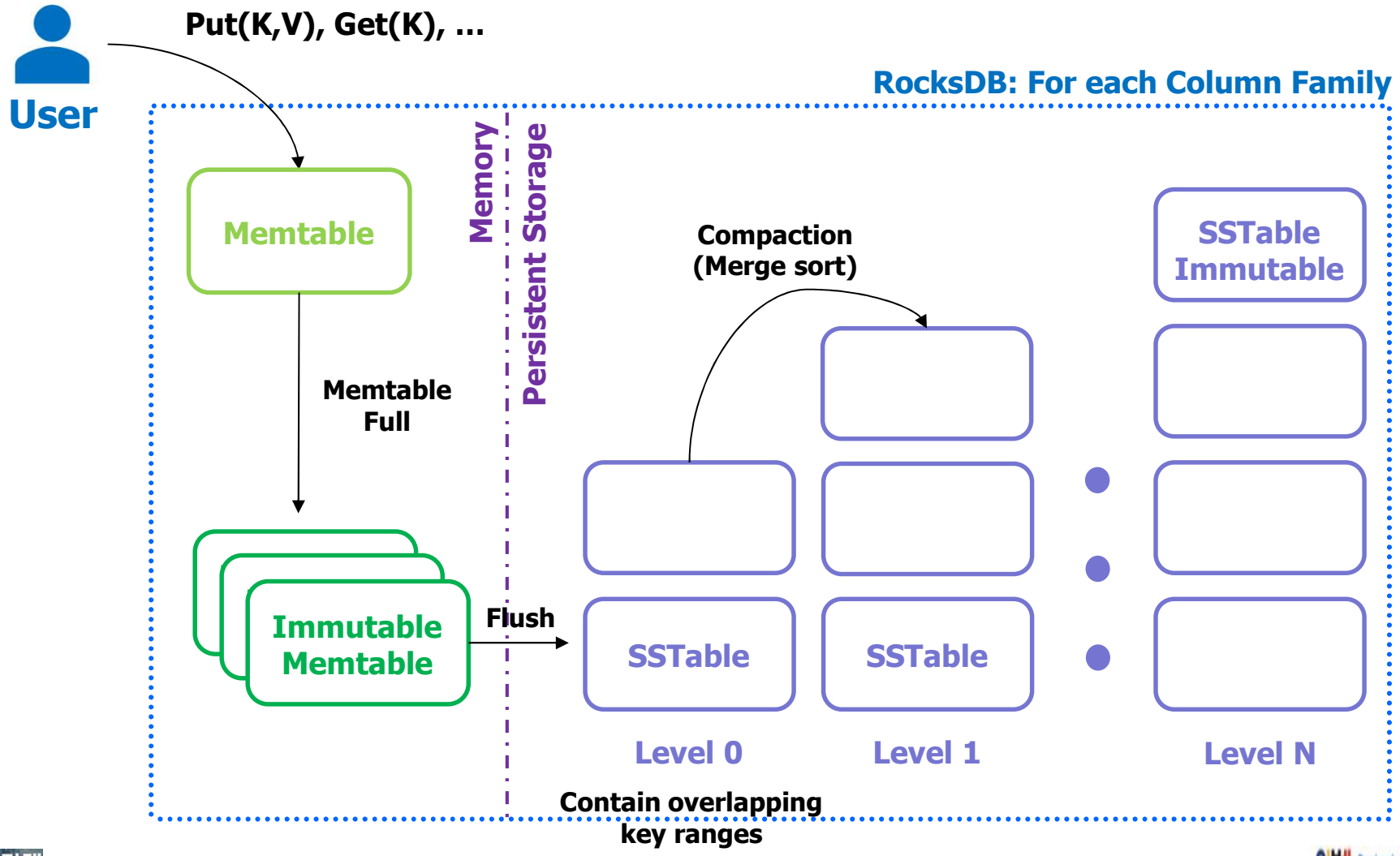
Hojin Shin, Jongmoo Choi
choijm@dankook.ac.kr
http://embedded.dankook.ac.kr/~choijm

RocksDB Festival: Contents

- Contents
  - ✓ RocksDB Architecture: Overall
  - ✓ RocksDB Architecture: MemTable
  - ✓ RocksDB Architecture: SSTable
  - ✓ RocksDB Operation: Write and WAL
  - ✓ RocksDB Operation: Read
  - ✓ RocksDB Operation: Compaction
  - ✓ QnA

RocksDB Festival: Overall

- RocksDB Architecture



RocksDB Festival: Overall

- RocksDB Architecture: Vending Machine

Put(K,V)

Get(K)
??

Vending Machine Database

Key	Value
초코우유1	700
초코우유2	700
딸기우유	700
흰 우유	500

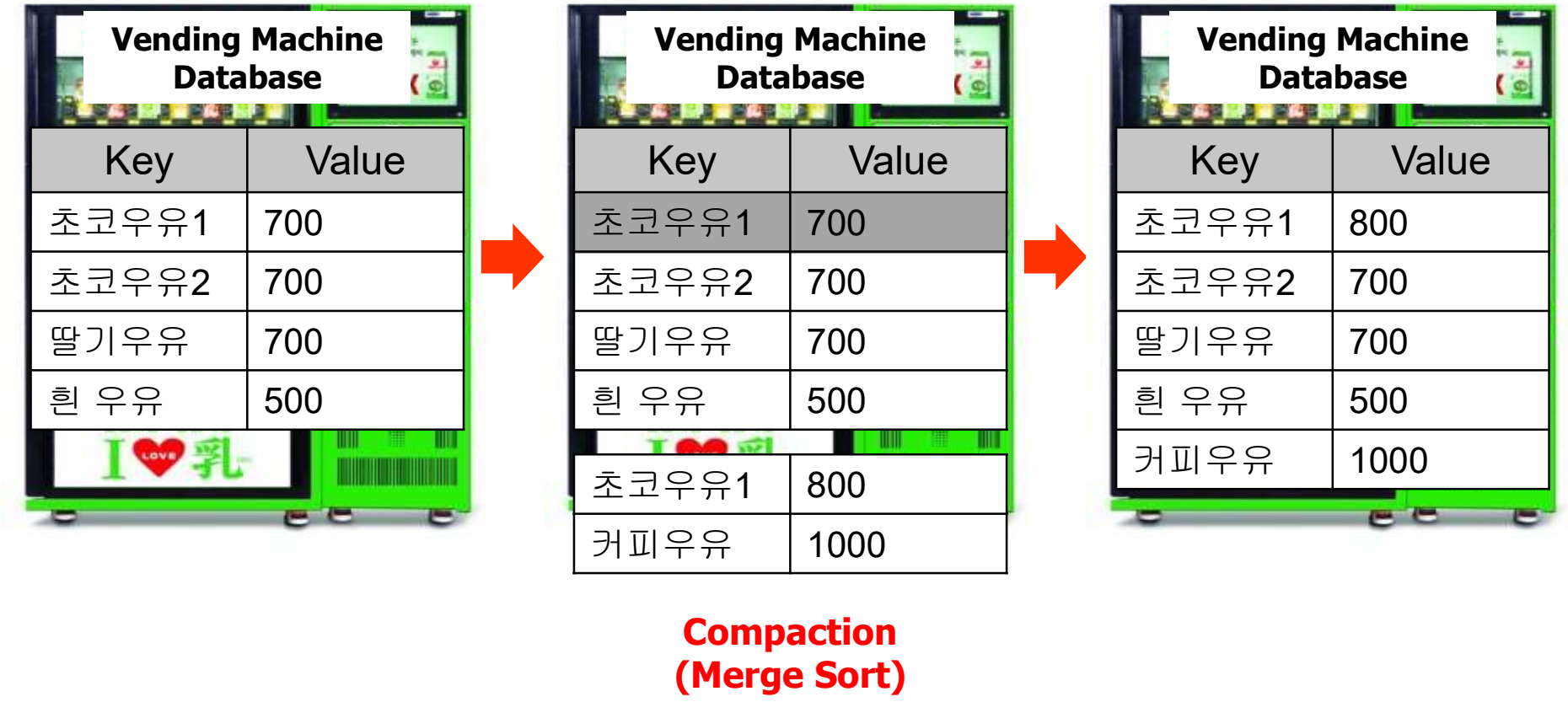
Candidate 1

Key	Value
700	초코, 딸기
500	흰 우유
…	…
…	…

Candidate 2

RocksDB Festival: Overall

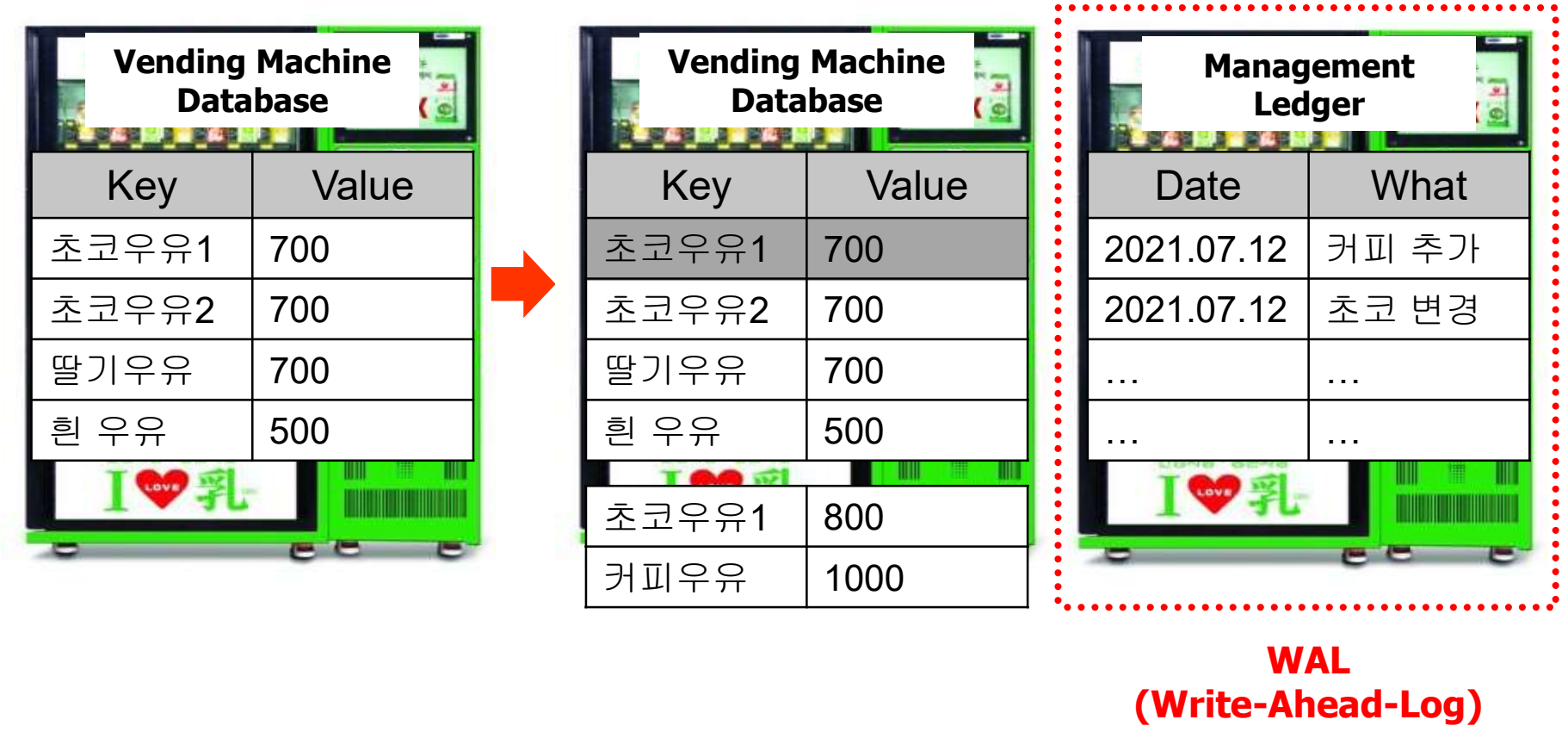
- RocksDB Architecture: Vending Machine



Compaction (Merge Sort)

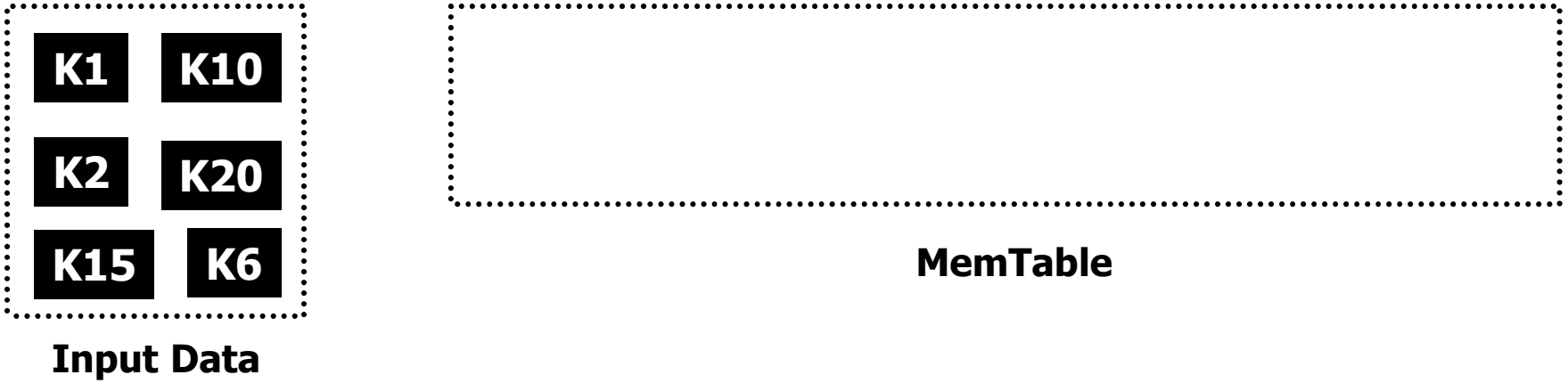
RocksDB Festival: Overall

- RocksDB Architecture: Vending Machine



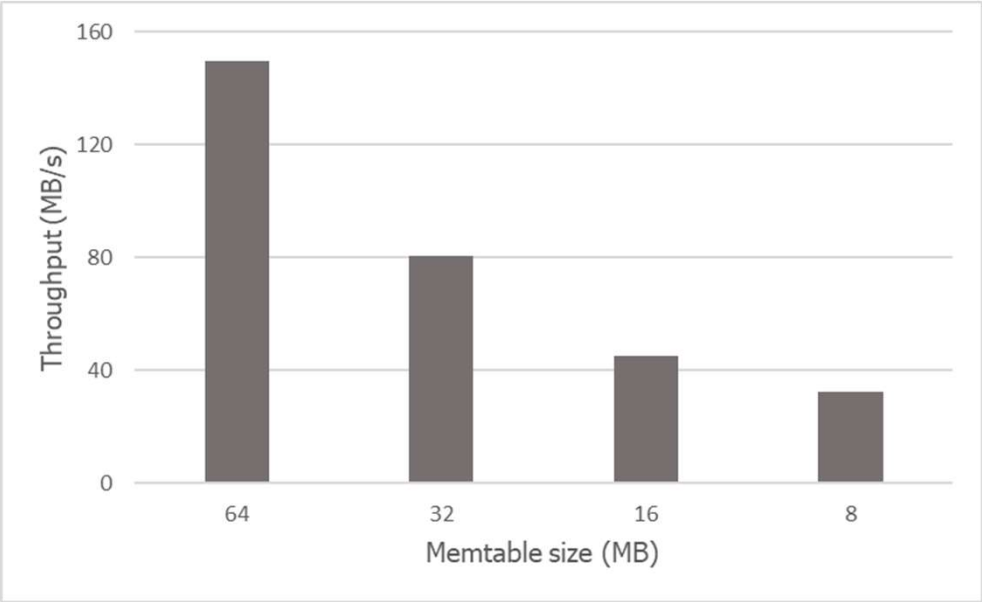
RocksDB Festival: MemTable

- RocksDB Architecture: MemTable
  - ✓ In-memory data structure
  - ✓ MemTable full ➔ Immutable memtable (Read-only)
  - ✓ Type before being flushed to SSTable
  - ✓ Immutable memtable flushed ➔ delete memtable (free)



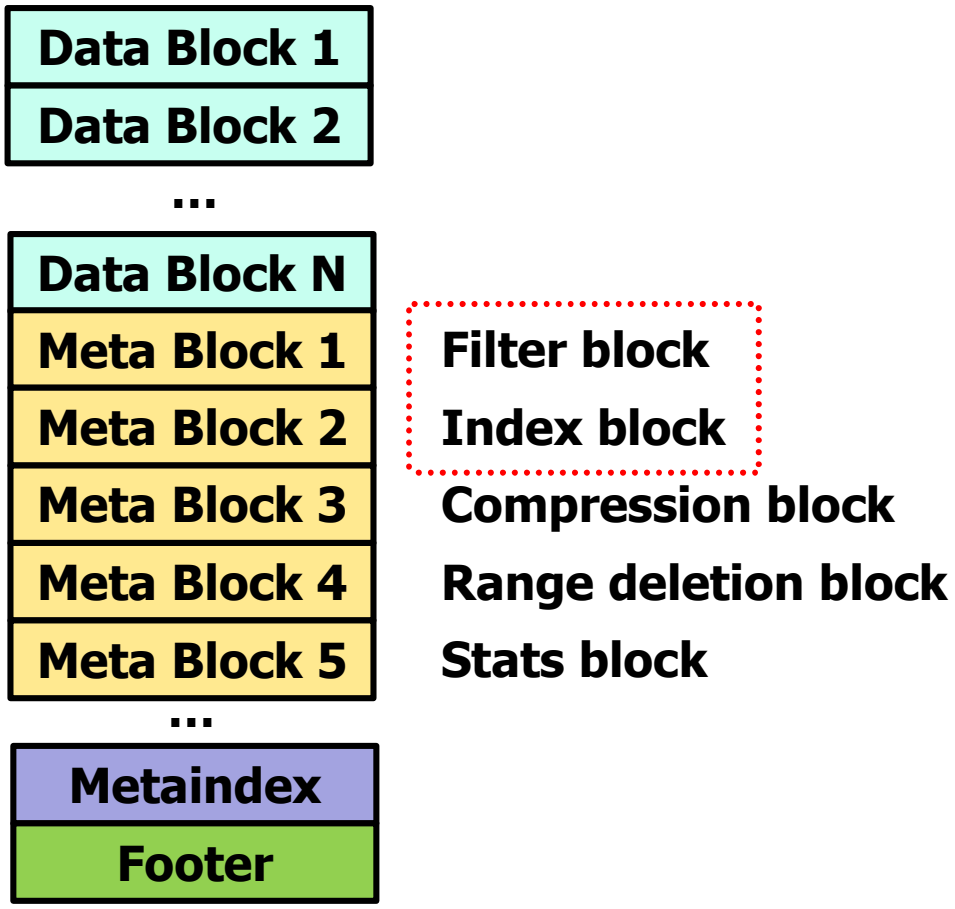
RocksDB Festival: MemTable

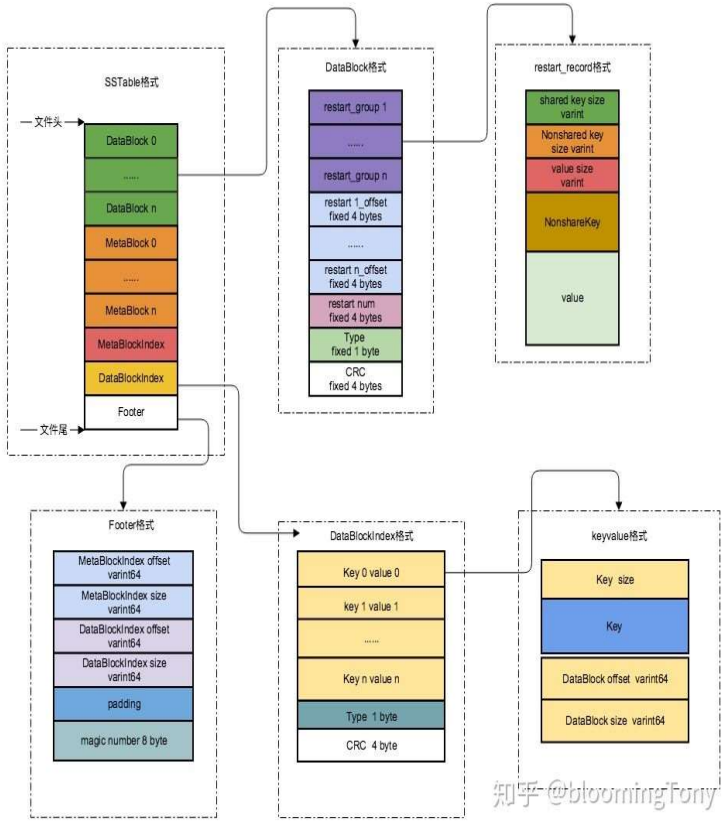
- RocksDB Architecture: MemTable
  - ✓ Effect of MemTable Size: With RocksDB db_bench
  - ✓ Why ➔ Compaction Overhead



RocksDB Festival: SSTable

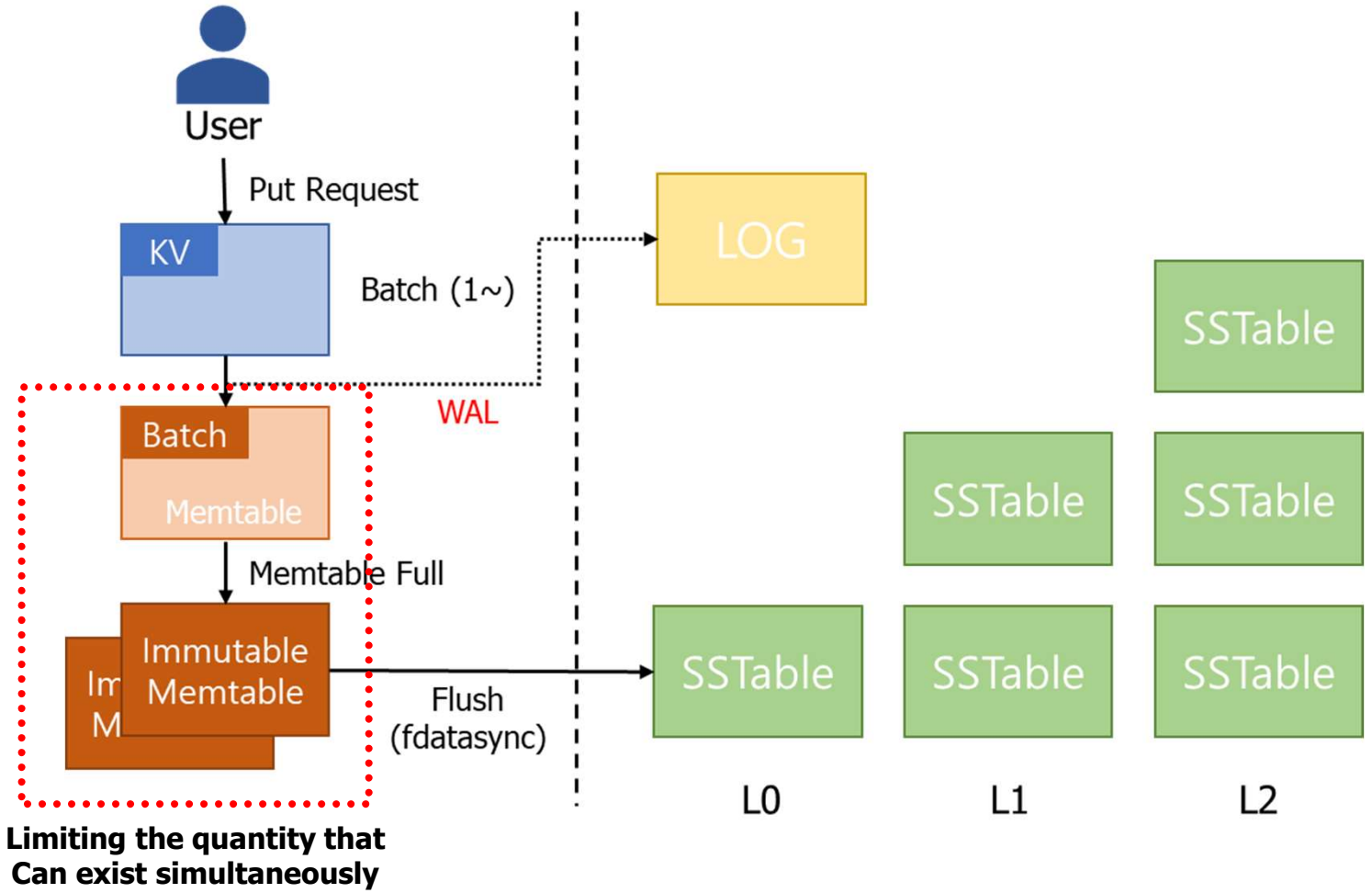
- RocksDB Architecture: SSTable
  - ✓ Sorted String Table
  - ✓ Block, PlainTable, CuckooTable





RocksDB Festival: Write and WAL

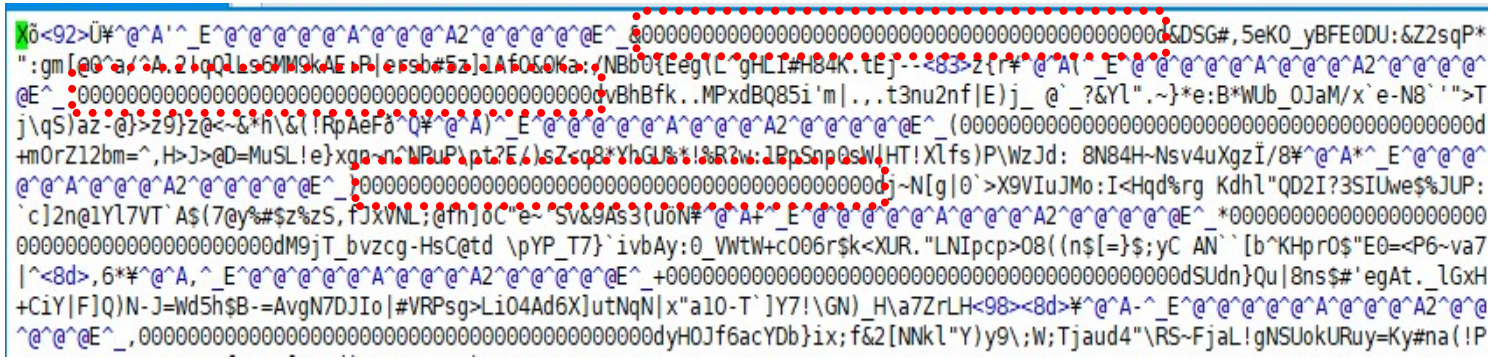
- RocksDB Operation: Write and WAL

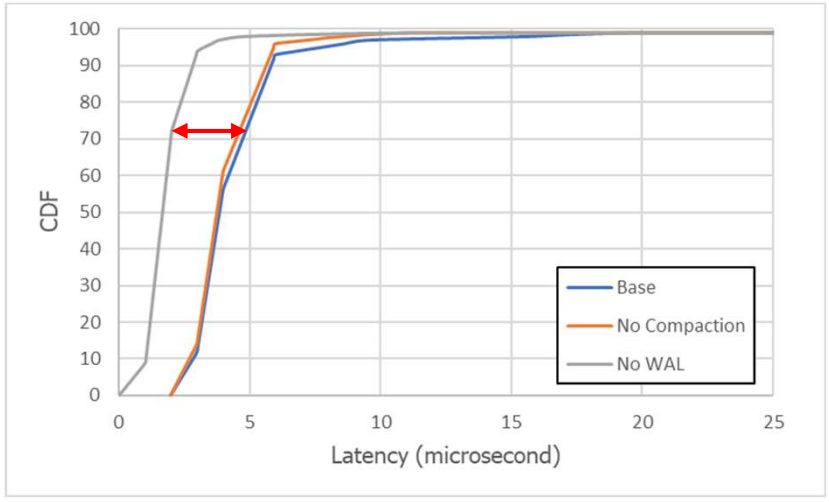


RocksDB Festival: Write and WAL

- RocksDB Operation: Write and WAL
  - ✓ 0000~~.log ➔ log file name
  - ✓ 0000~~.sst ➔ SSTable file name

```
000033.sst  000035.log
```

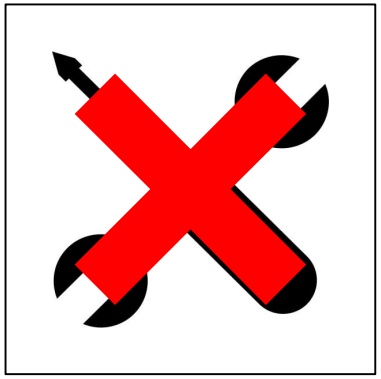




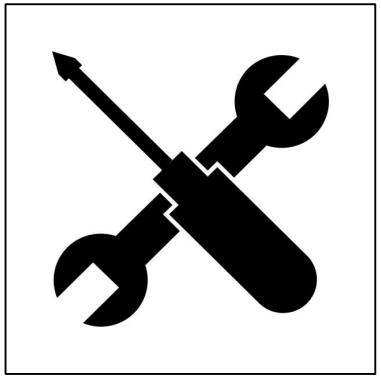
WAL overhead (Write Latency)

RocksDB Festival: Write and WAL

- RocksDB Operation: Write and WAL
  - ✓ If system failures can be predicted in advance
  - ✓ Control WAL overhead



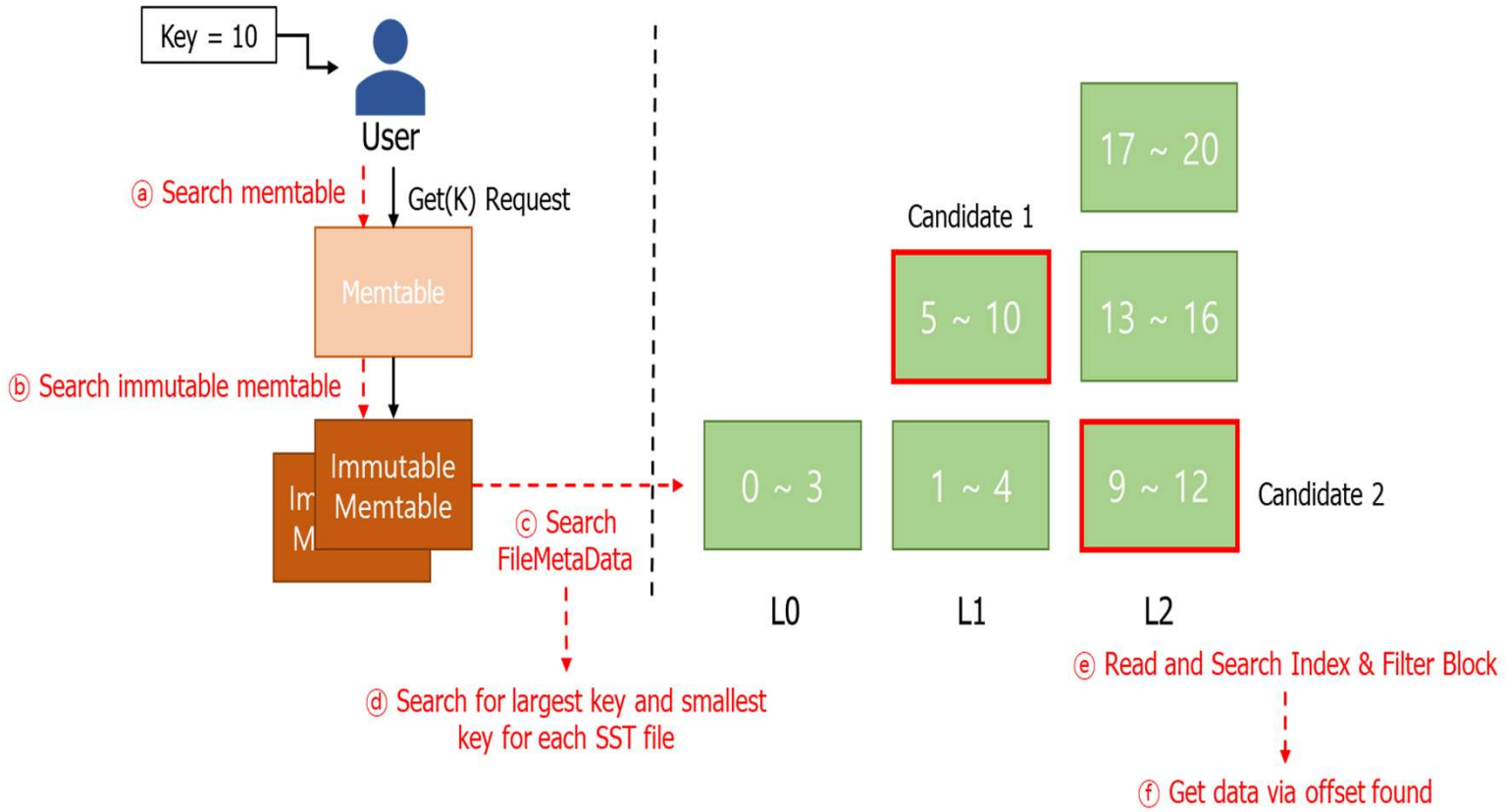
Do not proceed with WAL operation



Do proceed with WAL operation

RocksDB Festival: Read

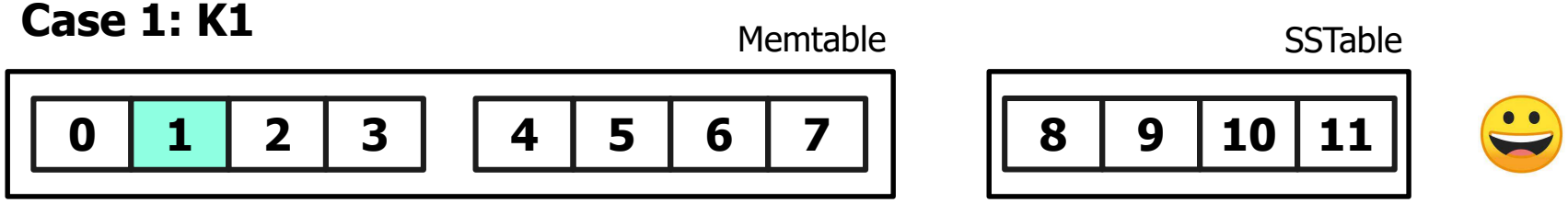
- RocksDB Operation: Read



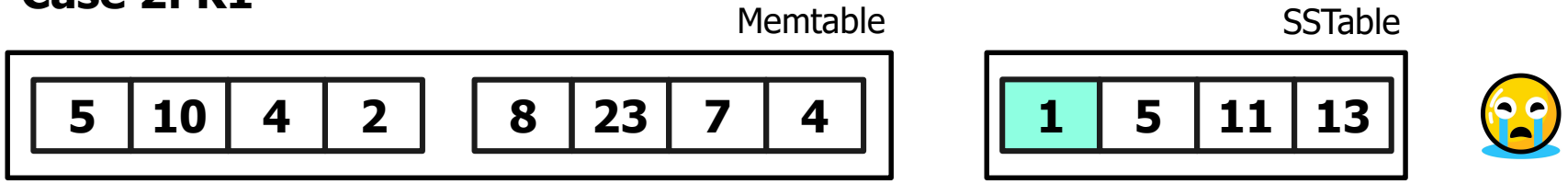
RocksDB Festival: Read

- RocksDB Operation: Read
  - ✓ Index block, Bloom filter block

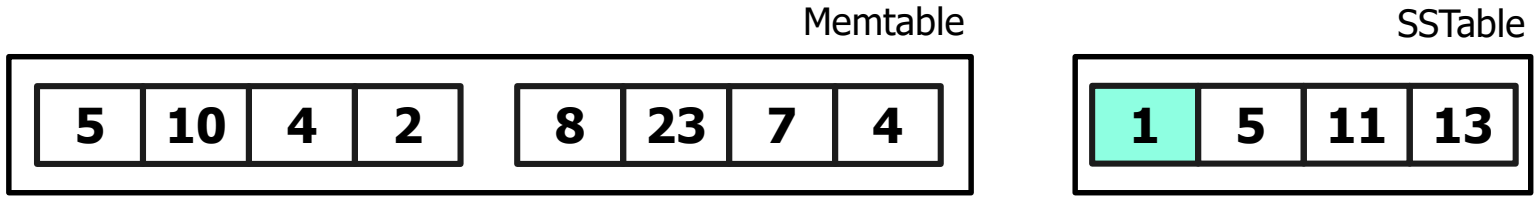
Case 1: K1



Case 2: K1

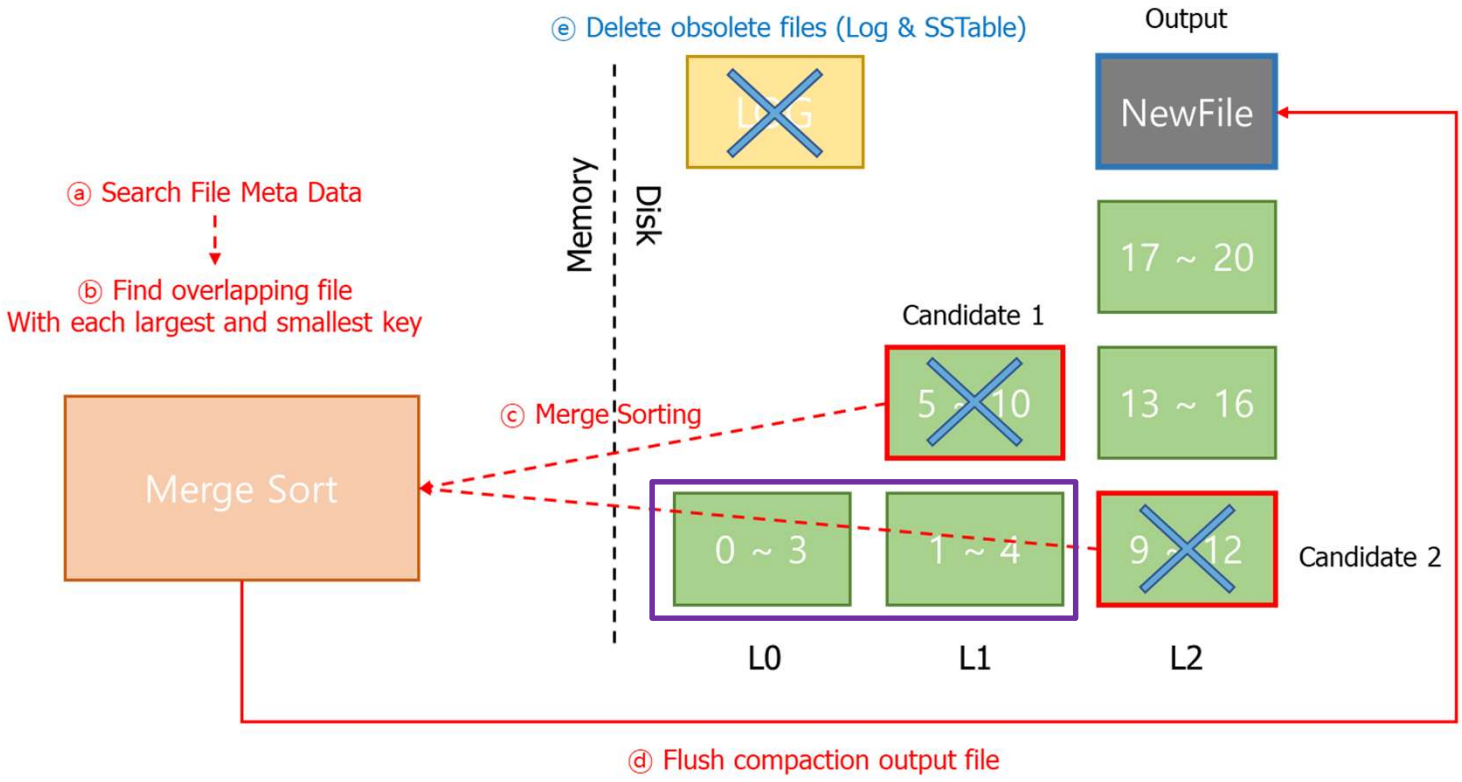


Case 3: K1



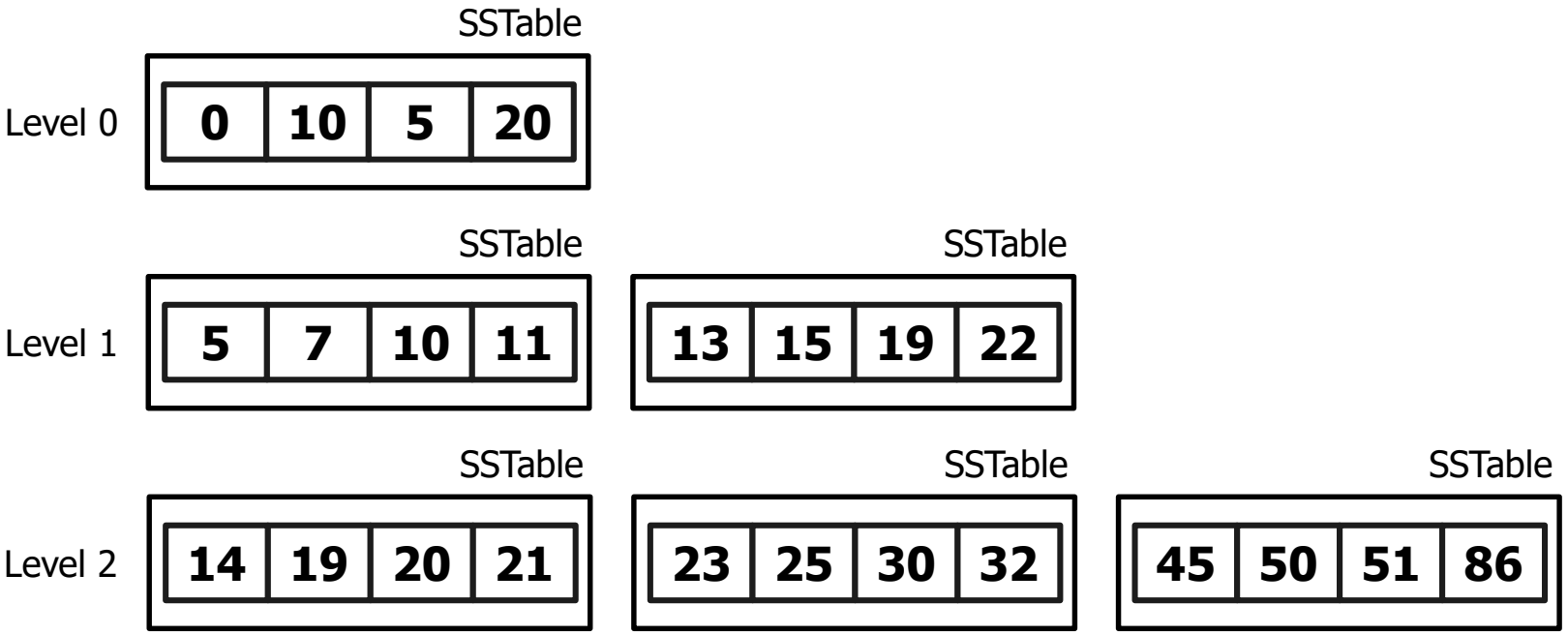
RocksDB Festival: Compaction

- RocksDB Operation: Compaction



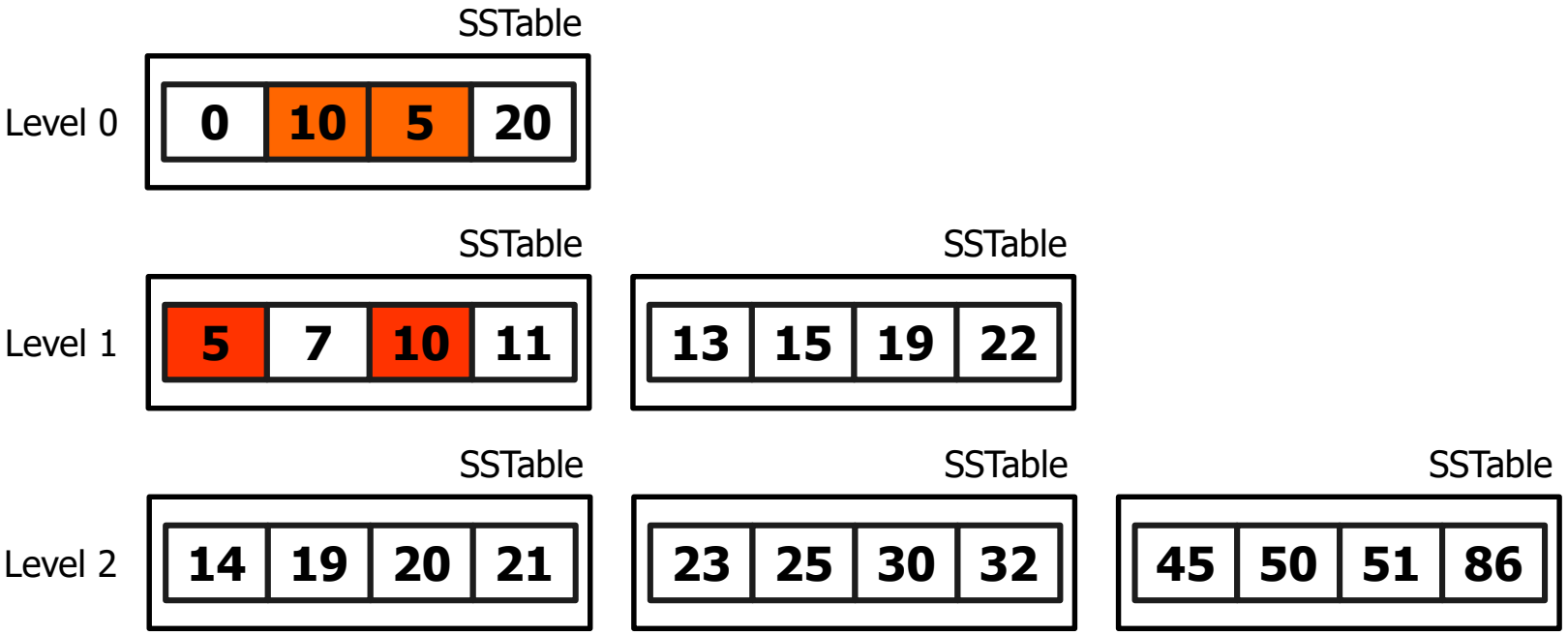
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



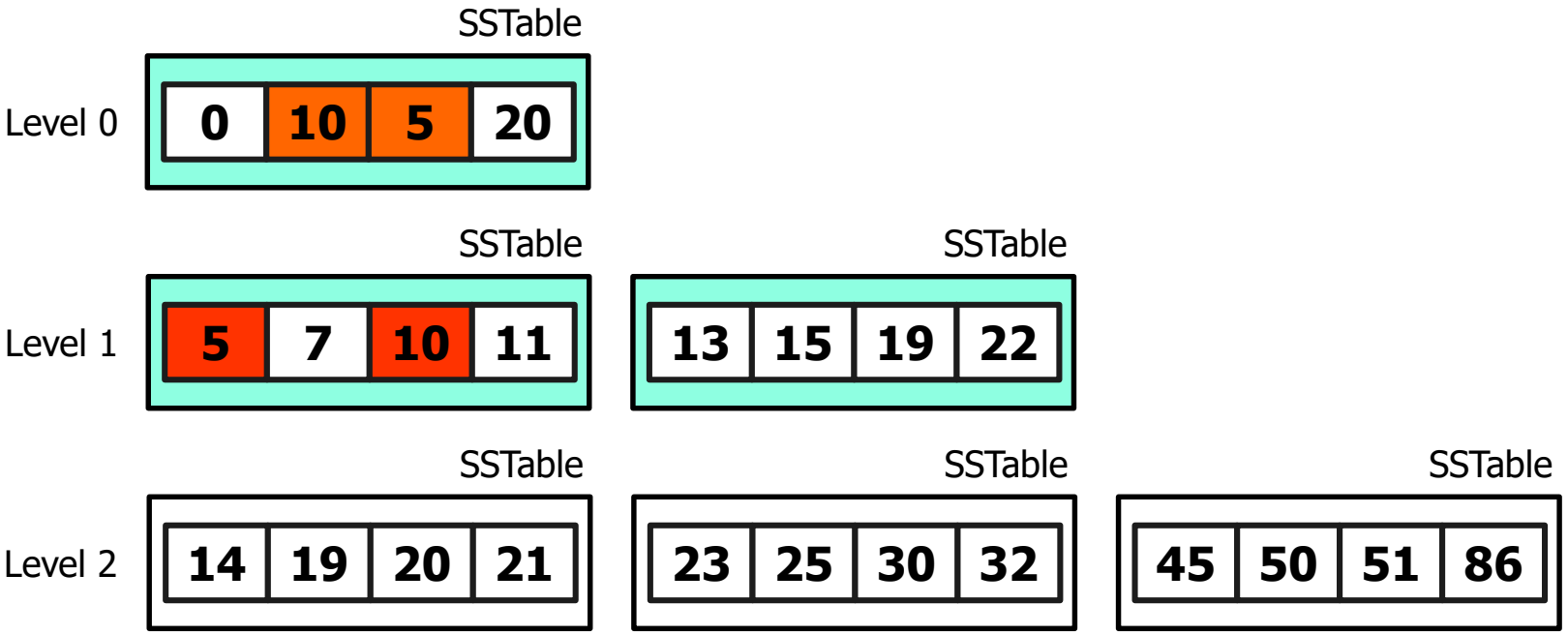
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



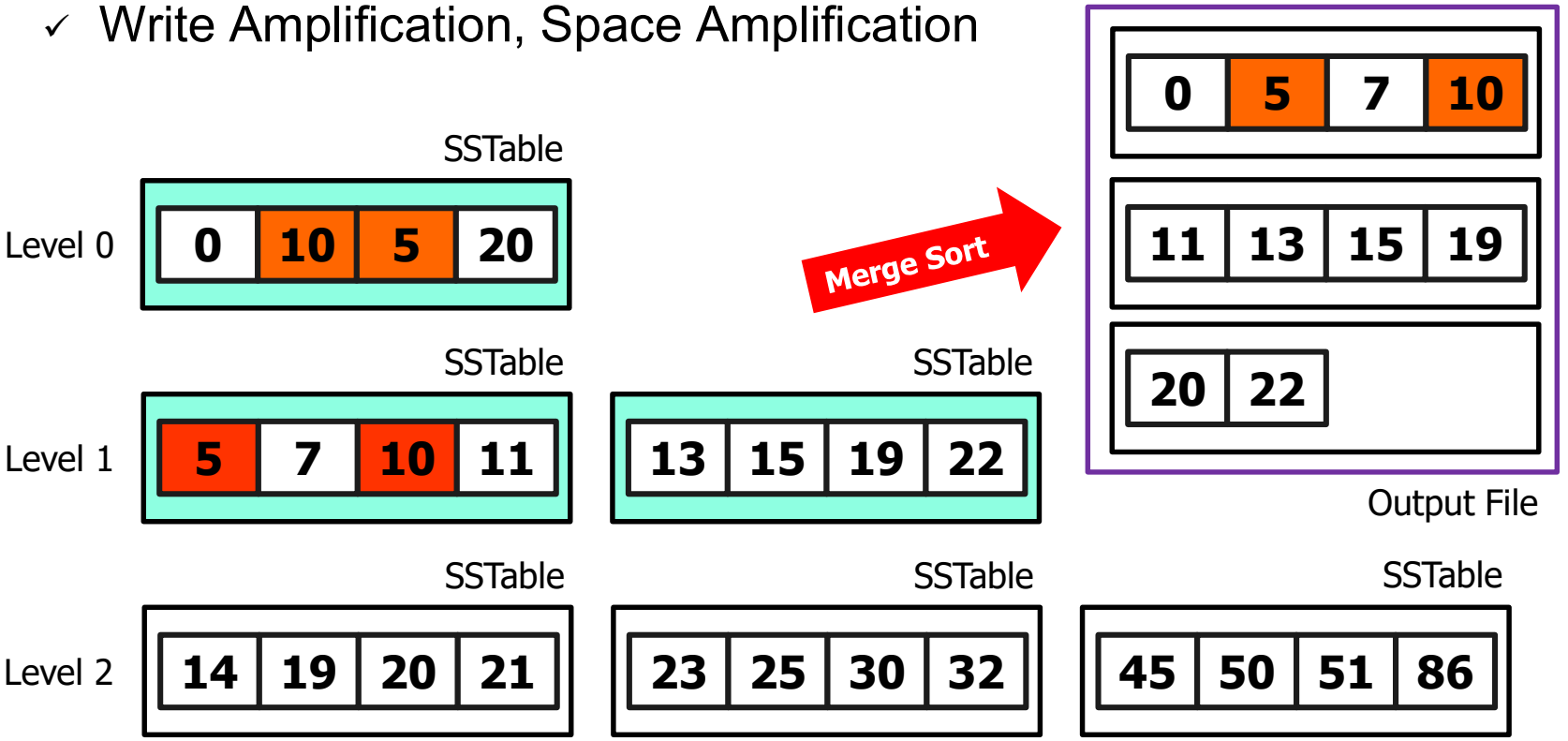
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



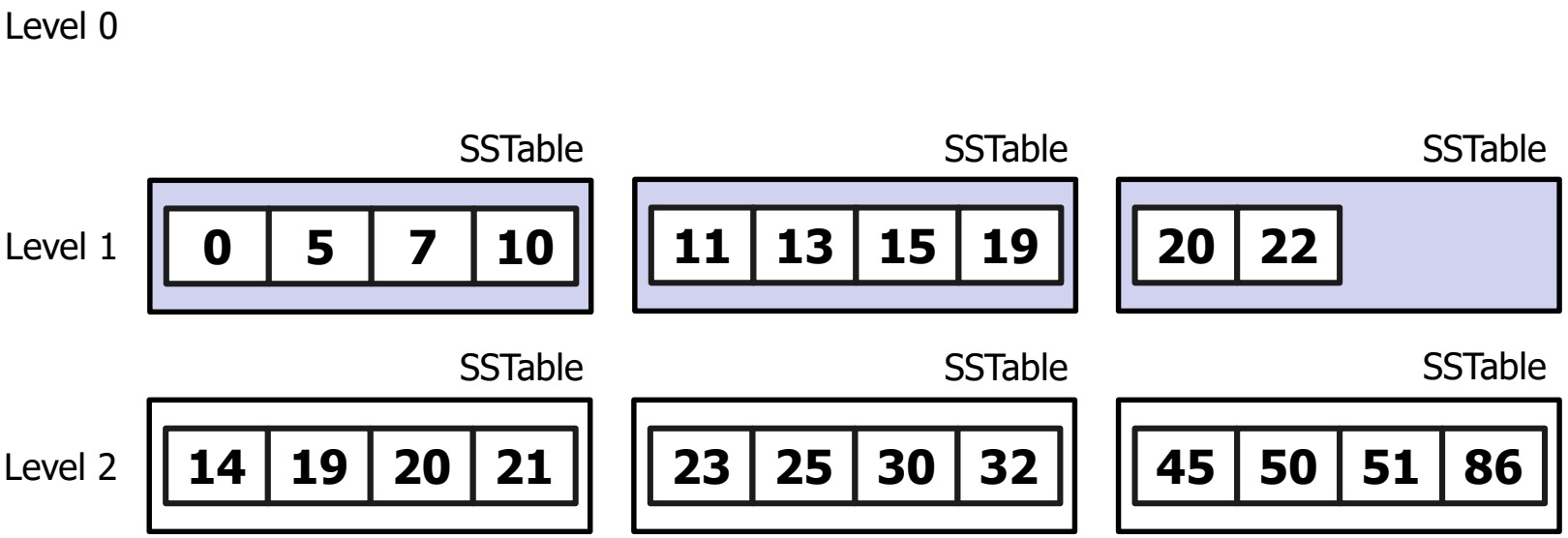
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



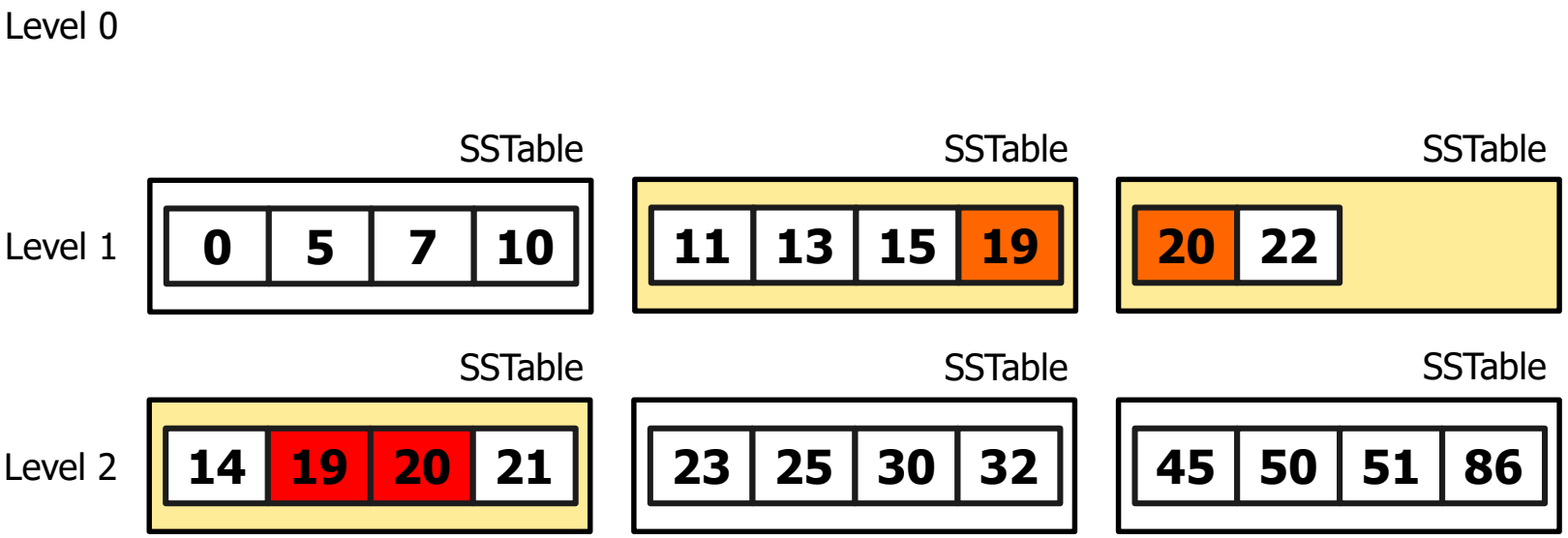
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



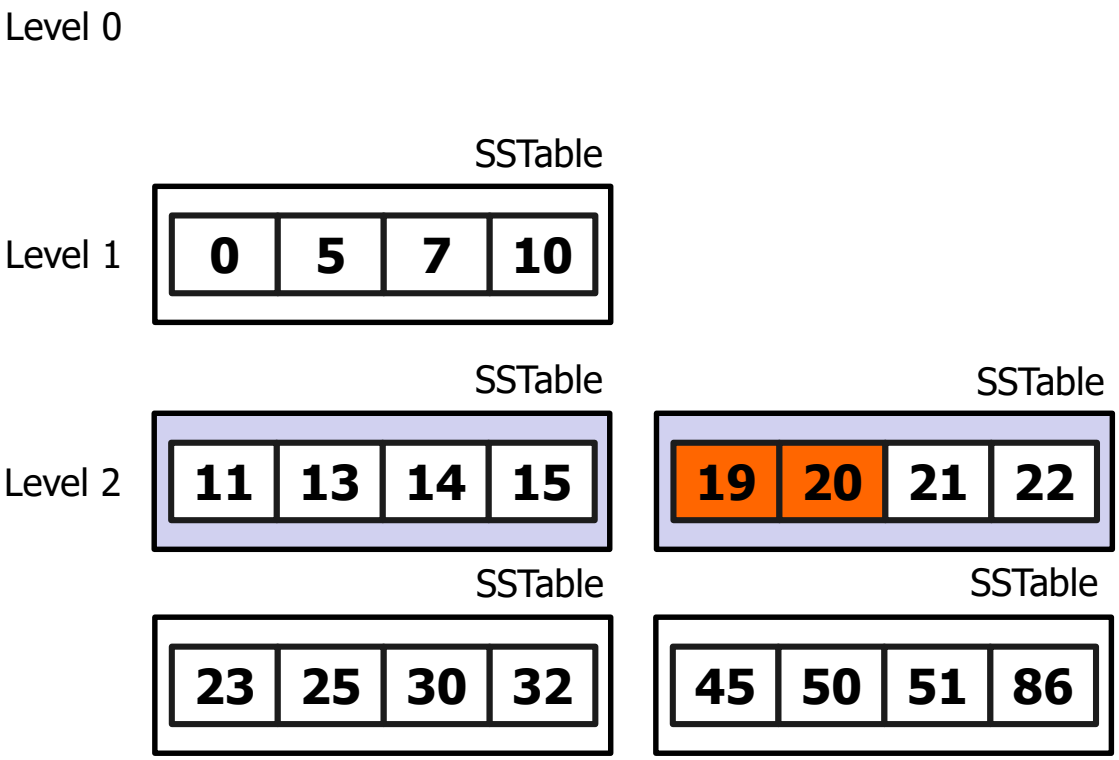
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



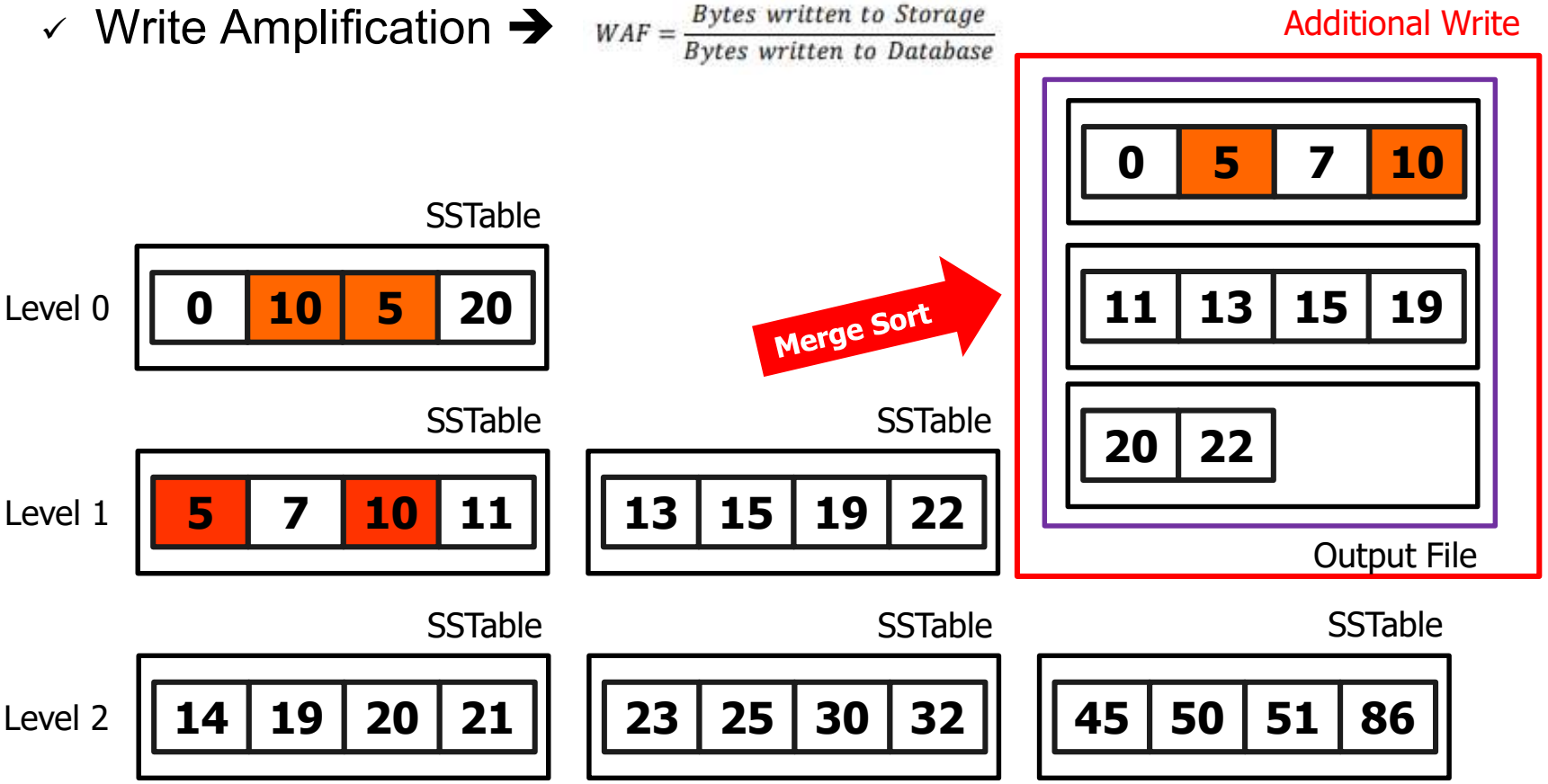
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification



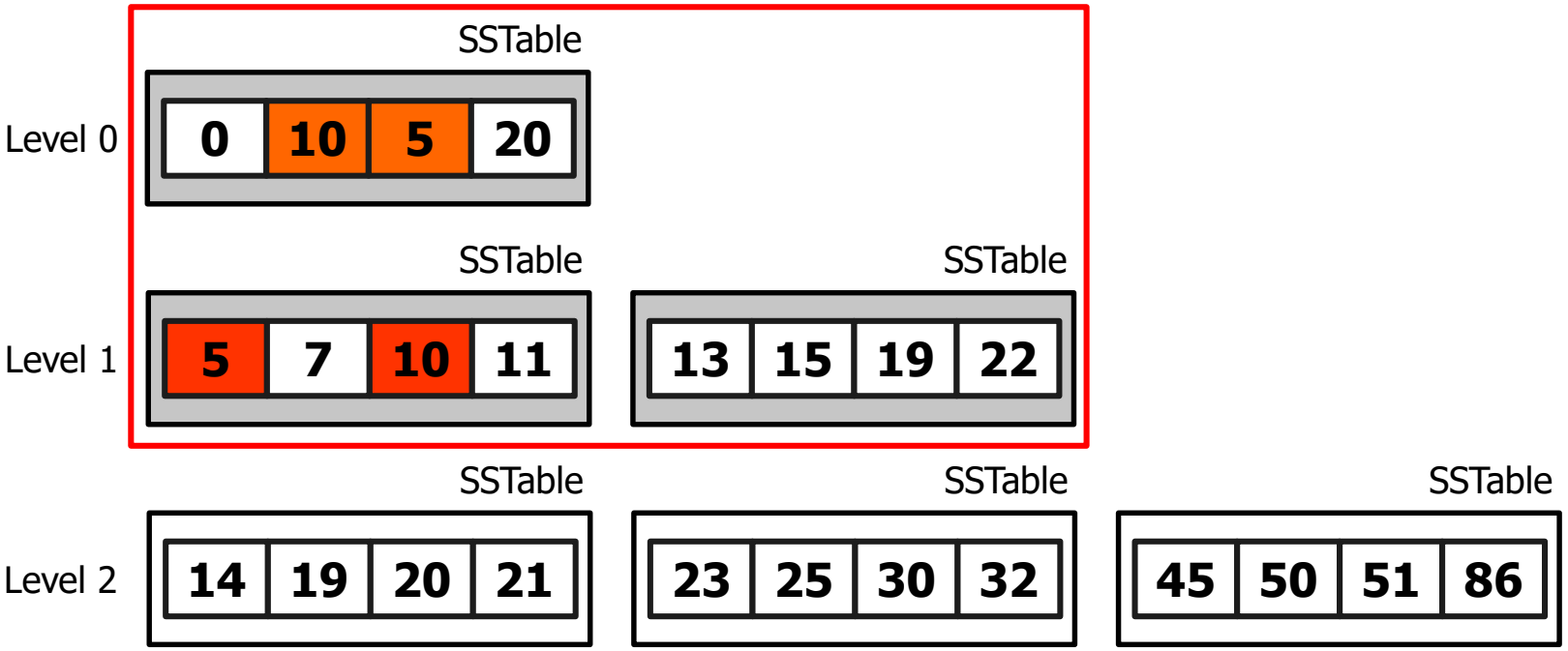
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Write Amplification ➔ $WAF = \frac{Bytes\ written\ to\ Storage}{Bytes\ written\ to\ Database}$



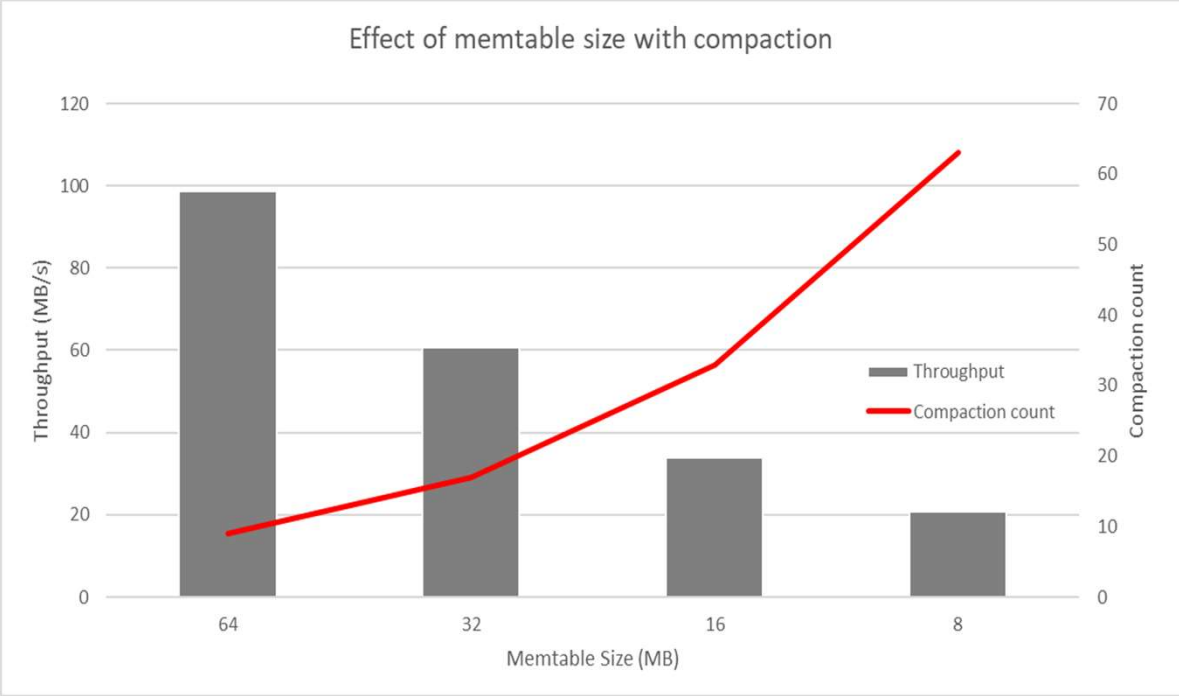
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Space Amplification



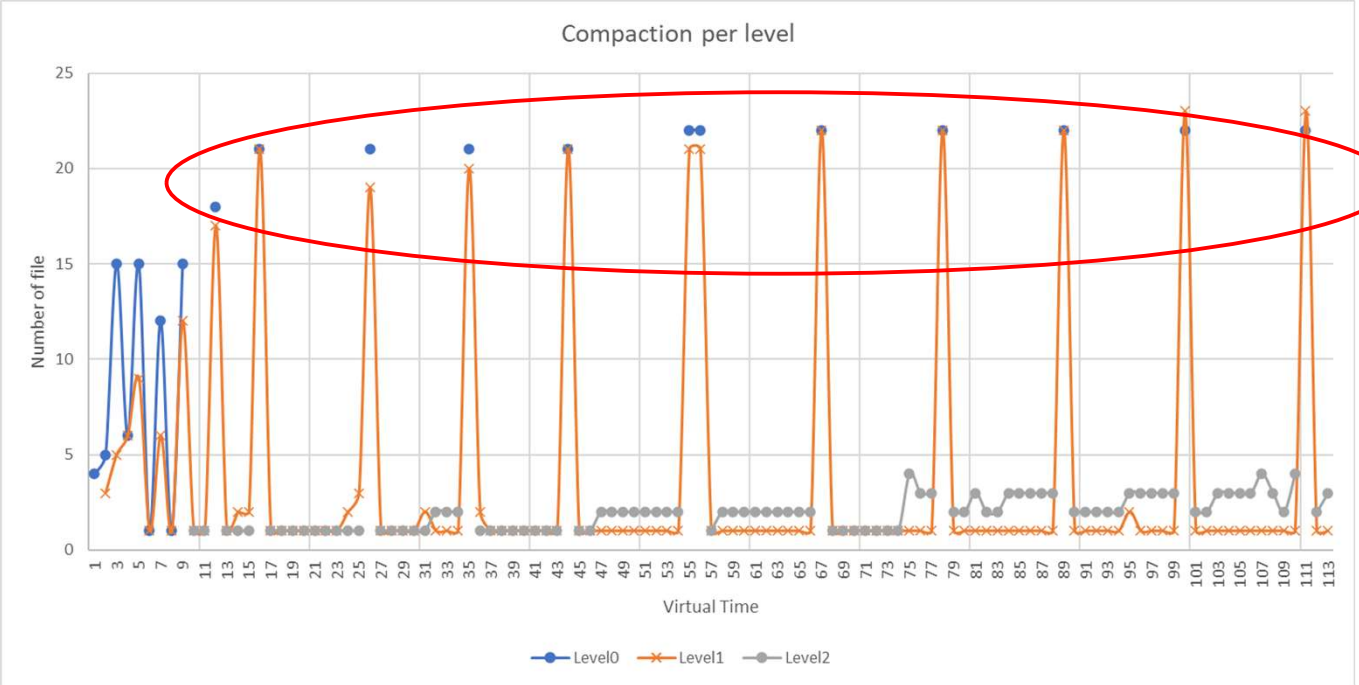
RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Compaction Overhead



RocksDB Festival: Compaction

- RocksDB Operation: Compaction
  - ✓ Compaction Trigger



QnA