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
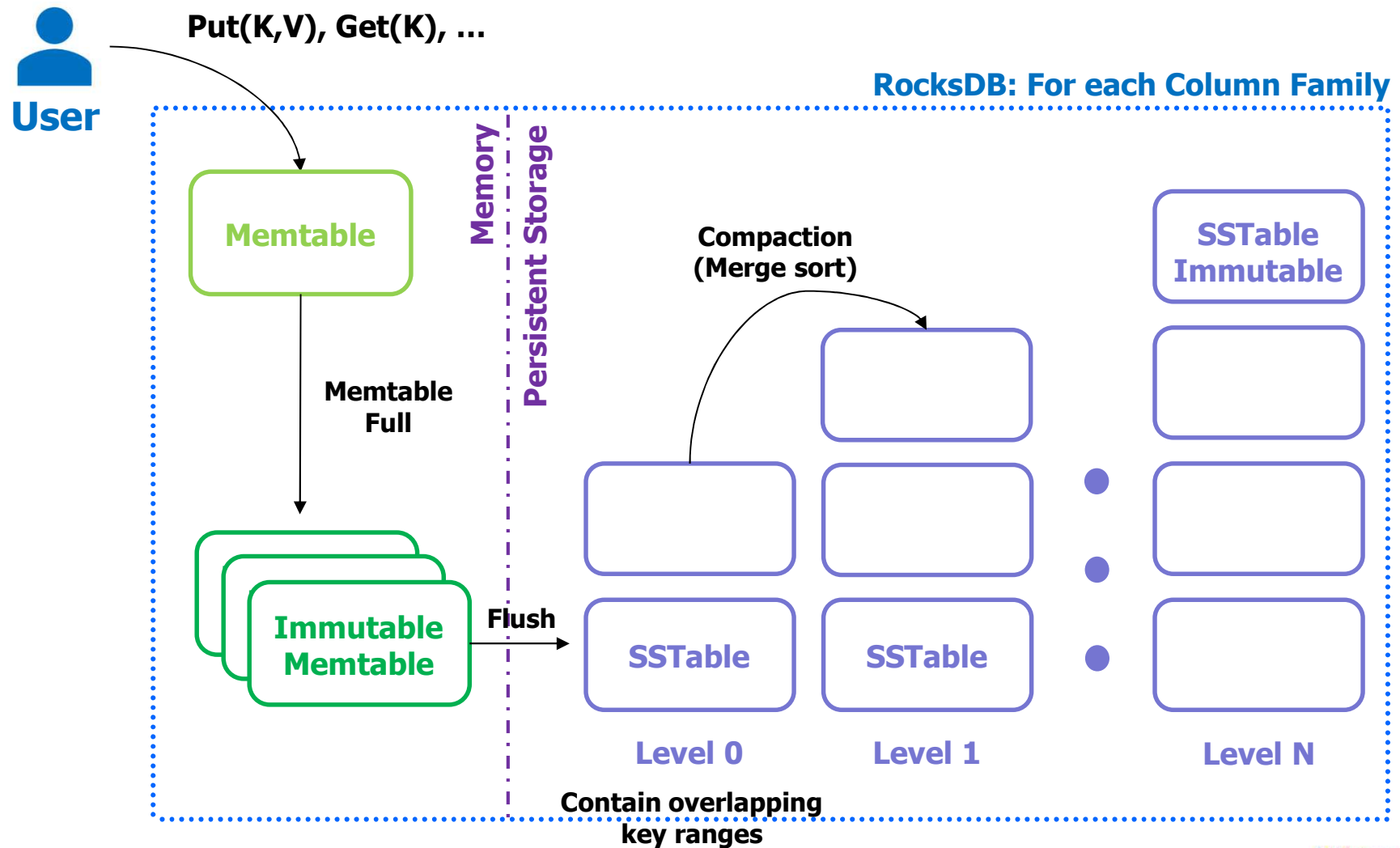
RocksDB Festival: Contents

■ Contents

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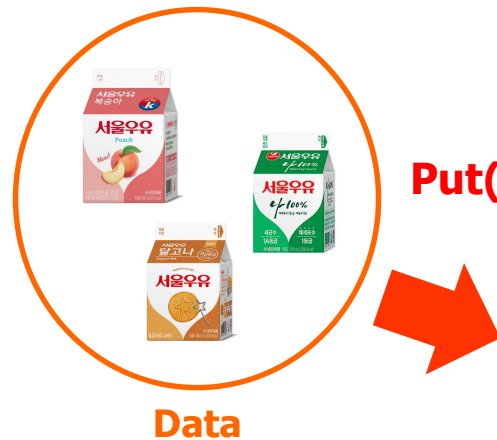- ✓ RocksDB Operation: Read
- ✓ RocksDB Operation: Compaction
- ✓ QnA

RocksDB

RocksDB Festival: Overall

- ## RocksDB Architecture

User

Put(K,V), Get(K), …

RocksDB: For each Column Family

Memory | **Persistent Storage**

Memtable

Memtable Full

Immutable Memtable

Flush

Compaction (Merge sort)

SSTable **SSTable**

SSTable Immutable

Level 0 **Level 1** **Level N**

Contain overlapping key ranges

SW스타랩

Dankook University
Embedded System

RocksDB Festival: Overall

■ RocksDB Architecture: Vending Machine

Put(K,V)

Data

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

SW스타랩

Dankook University
Embedded System

RocksDB Festival: Overall

- **RocksDB Architecture: Vending Machine**

Vending Machine Database	
Key	**Value**
초코우유1	700
초코우유2	700
딸기우유	700
흰 우유	500

Vending Machine Database	
Key	**Value**
초코우유1	700
초코우유2	700
딸기우유	700
흰 우유	500
초코우유1	800
커피우유	1000

Vending Machine Database	
Key	**Value**
초코우유1	800
초코우유2	700
딸기우유	700
흰 우유	500
커피우유	1000
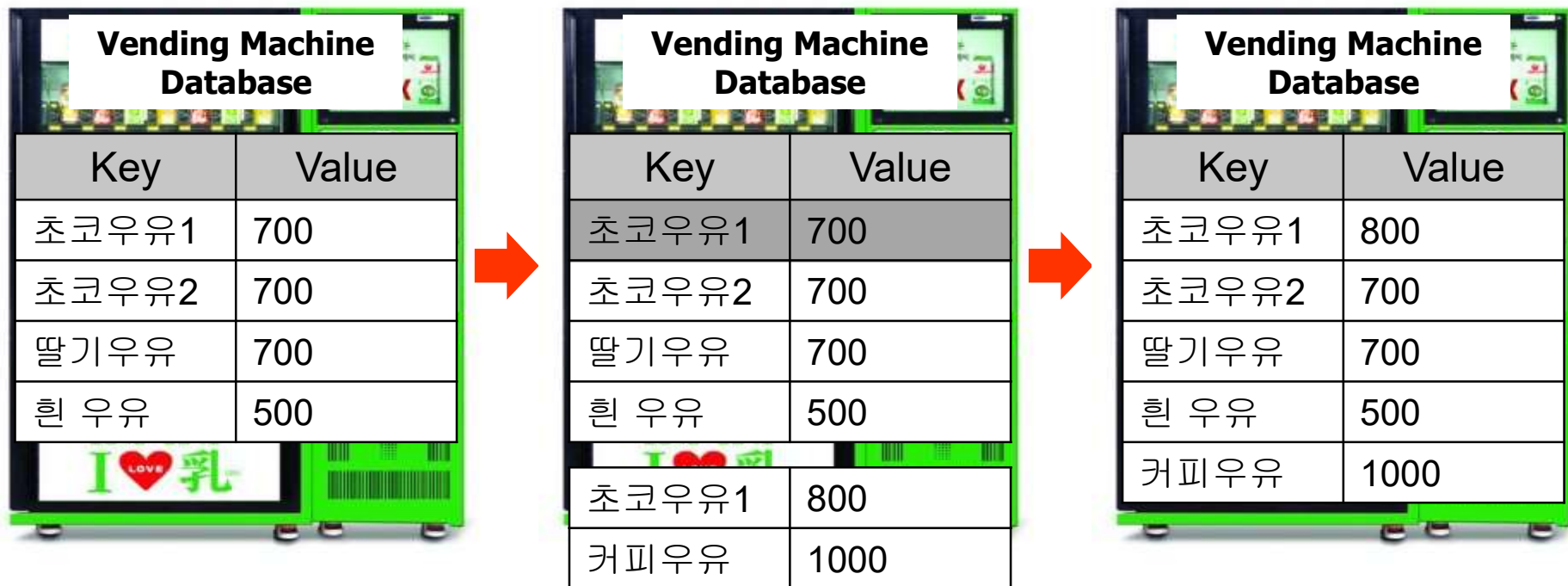
Compaction (Merge Sort)

RocksDB Festival: Overall

■ RocksDB Architecture: Vending Machine

Vending Machine Database

Key	Value
초코우유1	700
초코우유2	700
딸기우유	700
흰 우유	500

→

Vending Machine Database

Key	Value
초코우유1	700
초코우유2	700
딸기우유	700
흰 우유	500

초코우유1	800
커피우유	1000

Management Ledger

Date	What
2021.07.12	커피 추가
2021.07.12	초코 변경
…	…
…	…
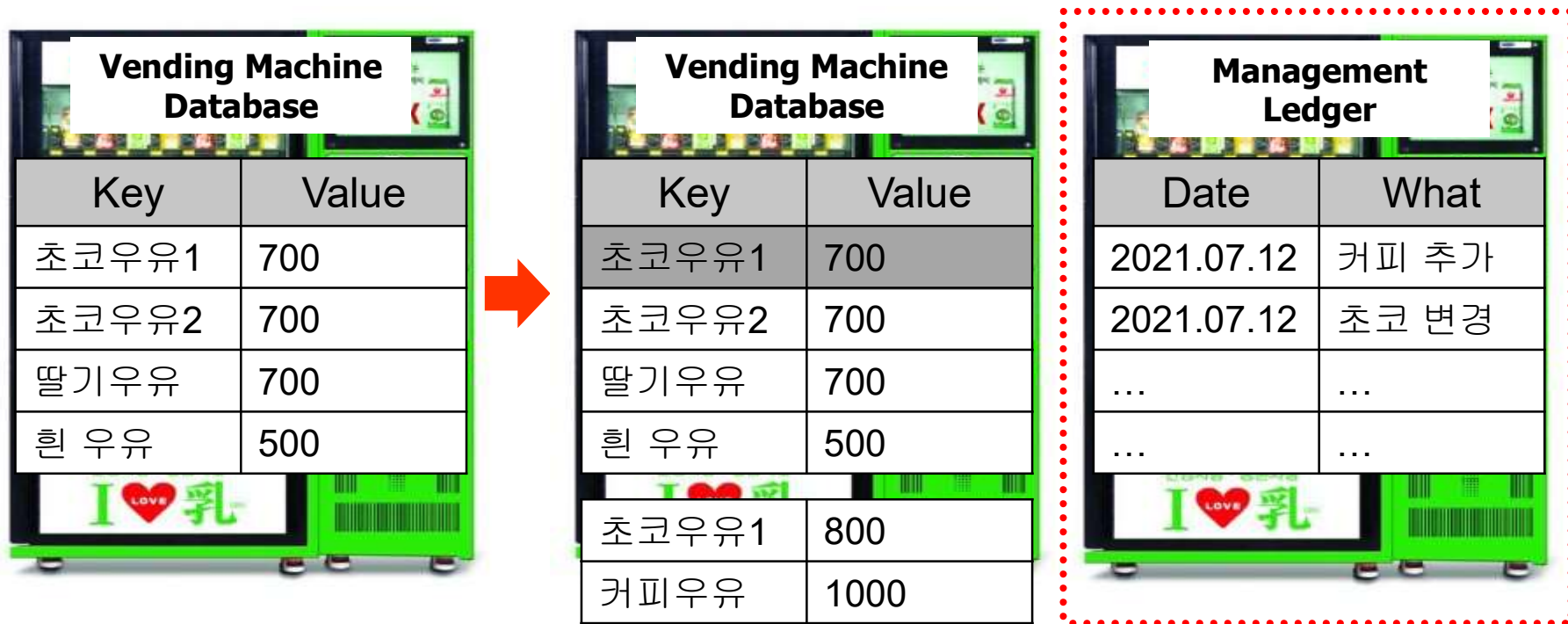
**WAL
(Write-Ahead-Log)**

RocksDB Festival: MemTable

■ **RocksDB Architecture: MemTable**

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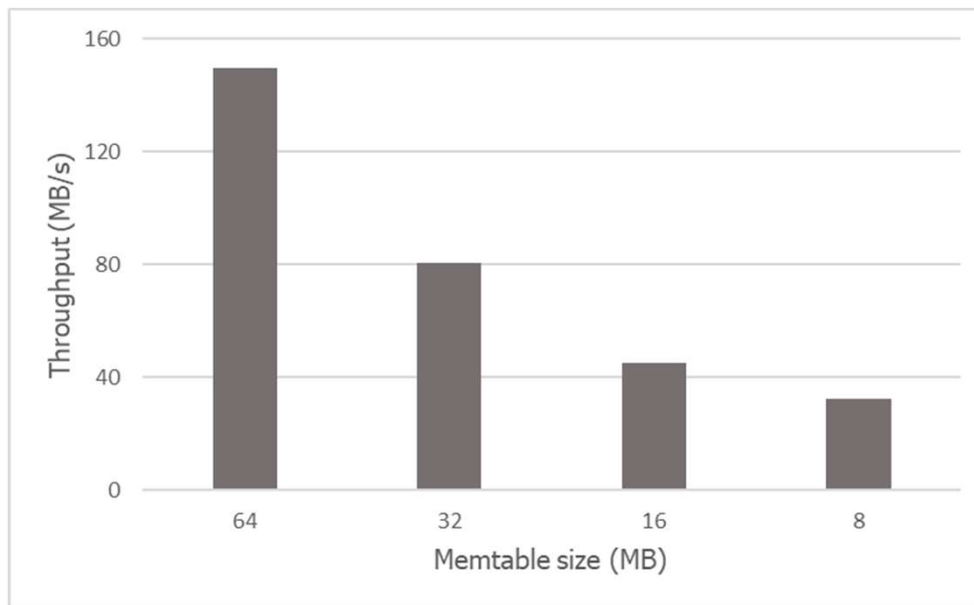- ✓ Immutable memtable flushed ➔ delete memtable (free)

K1	K10
K2	K20
K15	K6
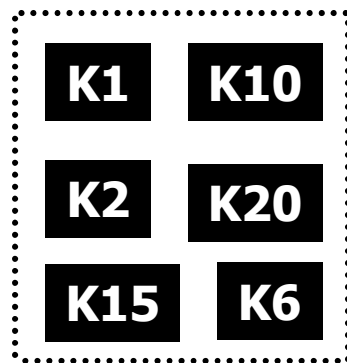
Input Data

MemTable

RocksDB Festival: MemTable

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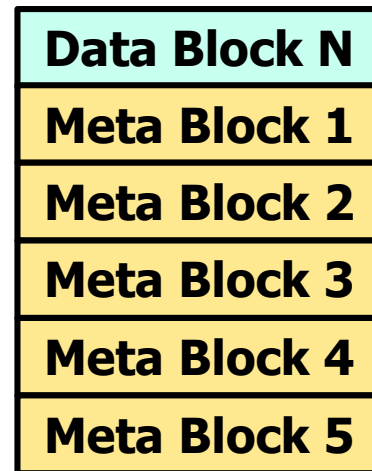 - ✓ Effect of MemTable Size: With RocksDB db_bench
  - ✓ Why ➔ Compaction Overhead

RocksDB Festival: SSTable

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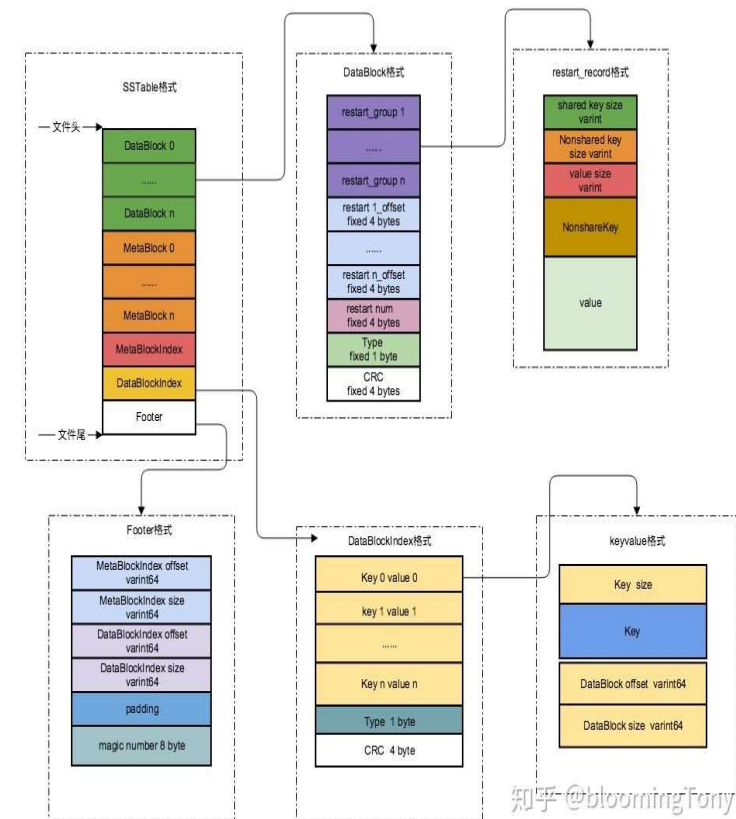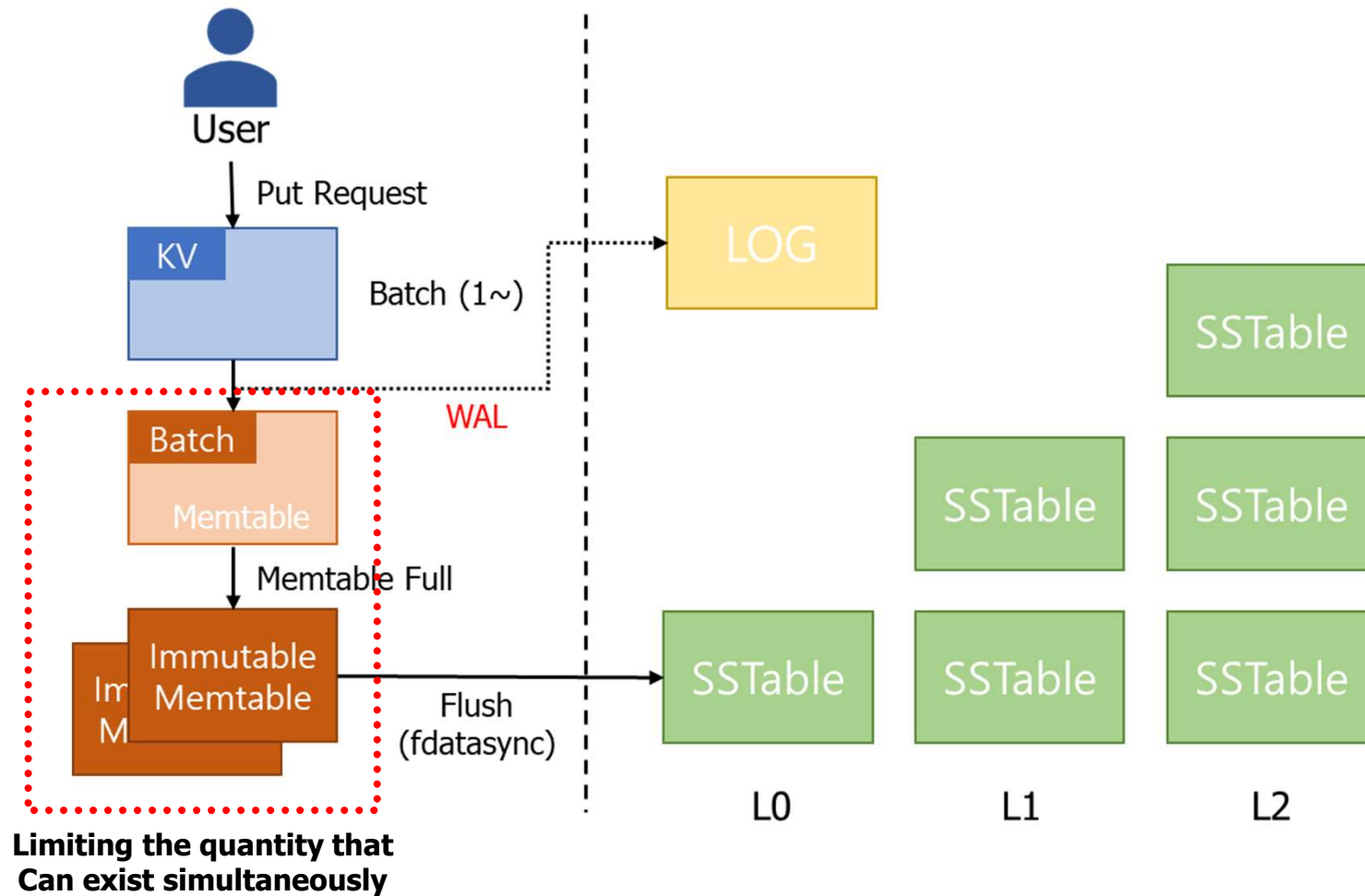 - ✓ Block, PlainTable, CuckooTable

Data Block 1
Data Block 2

...

Data Block N
Meta Block 1
Meta Block 2
Meta Block 3
Meta Block 4
Meta Block 5

...

Metaindex
Footer

Filter block

Index block

Compression block

Range deletion block

Stats block

知乎 @bloomingTony

Dankook University
Embedded System
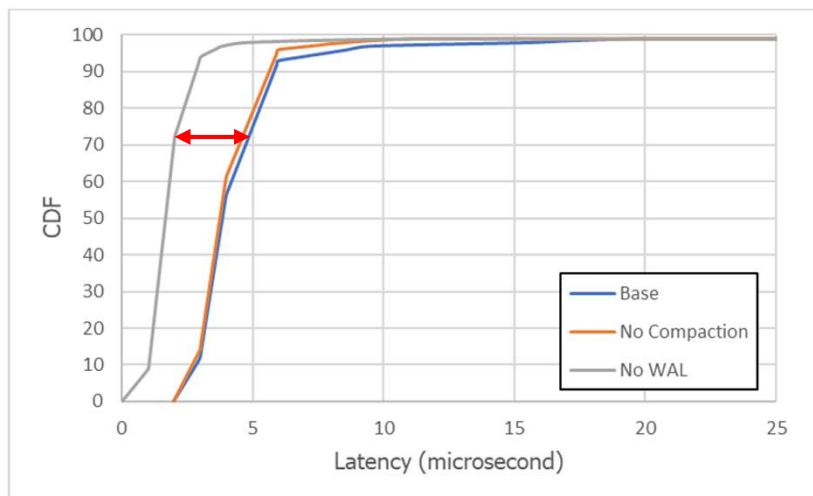
RocksDB Festival: Write and WAL

- **RocksDB Operation: Write and WAL**

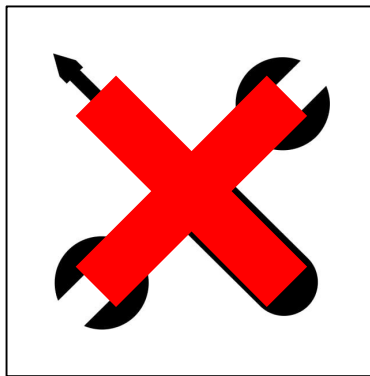# RocksDB Festival: Write and WAL

- **RocksDB Operation: Write and WAL**
  - ✓ 0000~~.log ➔ log file name
  - ✓ 0000~~.sst ➔ SSTable file name
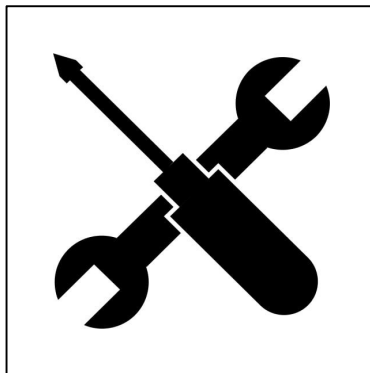
`000033.sst    000035.log`

**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**
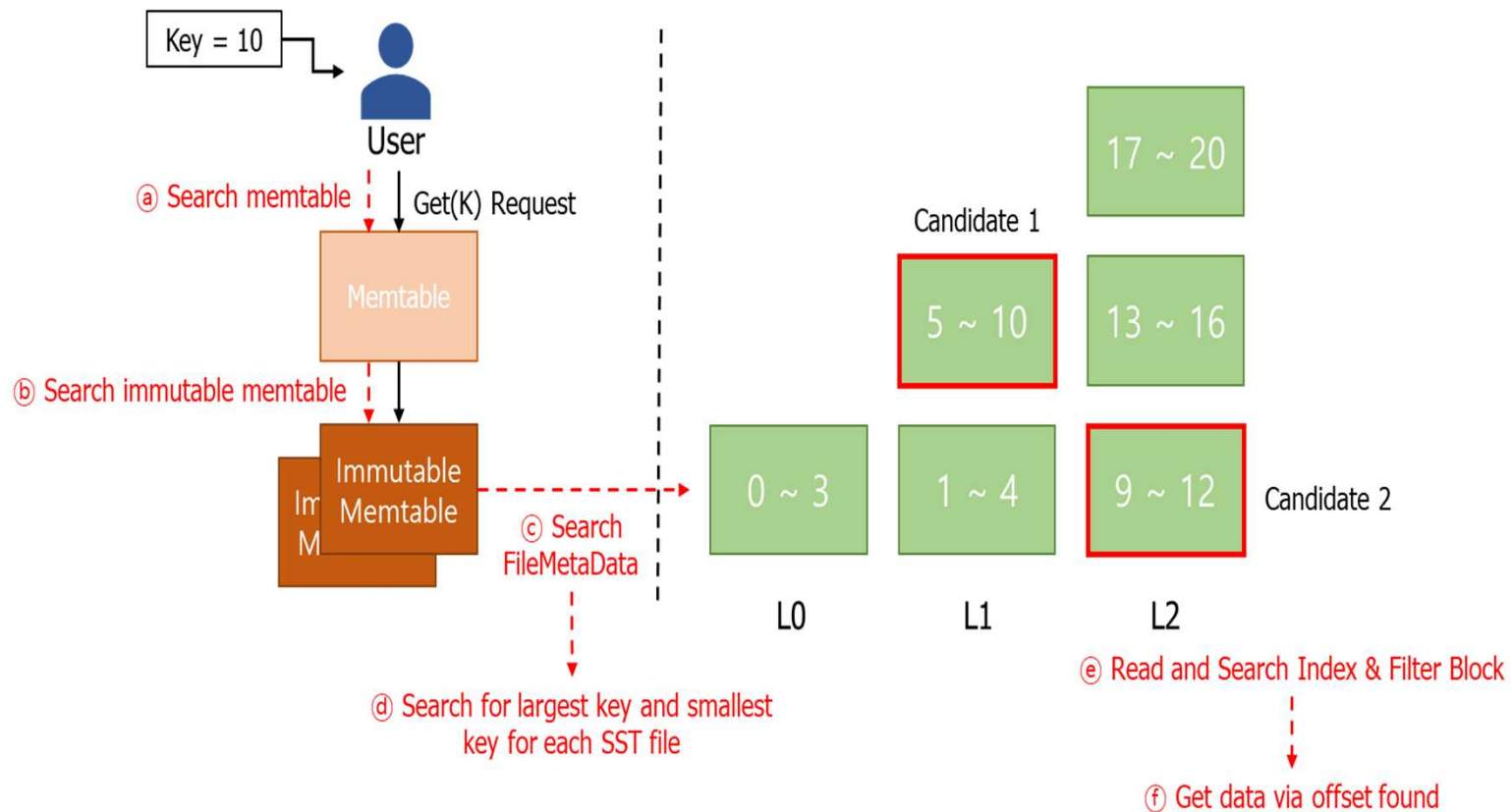
**Do proceed with
WAL operation**

?

RocksDB Festival: Read
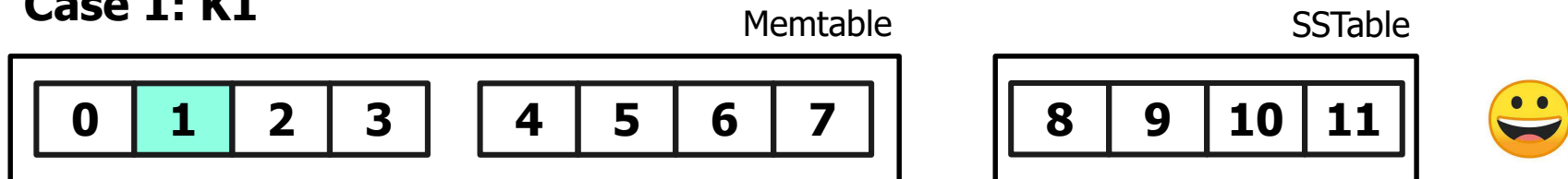
- ## RocksDB Operation: Read

RocksDB Festival: Read

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

Case 1: K1

Memtable   SSTable

| 0 | 1 | 2 | 3 | | 4 | 5 | 6 | 7 | | 8 | 9 | 10 | 11 | 😃

Case 2: K1

Memtable   SSTable

| 2 | 4 | 5 | 10 | | 4 | 7 | 8 | 23 | | 1 | 5 | 11 | 13 | 😭

Case 3: K1

| Memtable: 2 ~ 23 | | SSTable: 1~13 | 😄😄😄

Memtable   SSTable

| 2 | 4 | 5 | 10 | | 4 | 7 | 8 | 23 | | 1 | 5 | 11 | 13 |

RocksDB Festival: Compaction

- **RocksDB Operation: Compaction**

RocksDB Festival: Compaction

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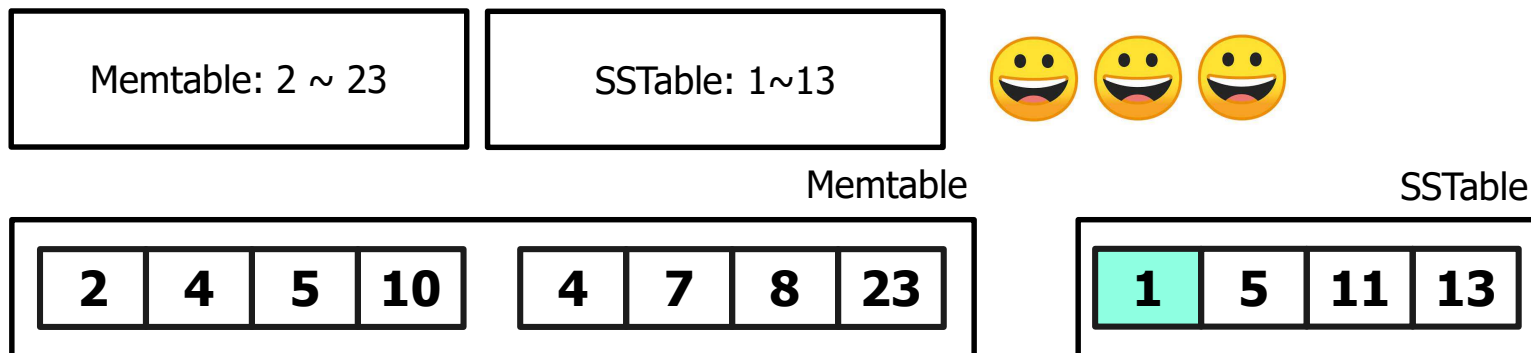 - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification

SSTable

Level 0 | 0 | 5 | 10 | 20 |

SSTable   SSTable

Level 1 | 5 | 7 | 10 | 11 | | 13 | 15 | 19 | 22 |

SSTable   SSTable   SSTable

Level 2 | 14 | 19 | 20 | 21 | | 23 | 25 | 30 | 32 | | 45 | 50 | 51 | 86 |
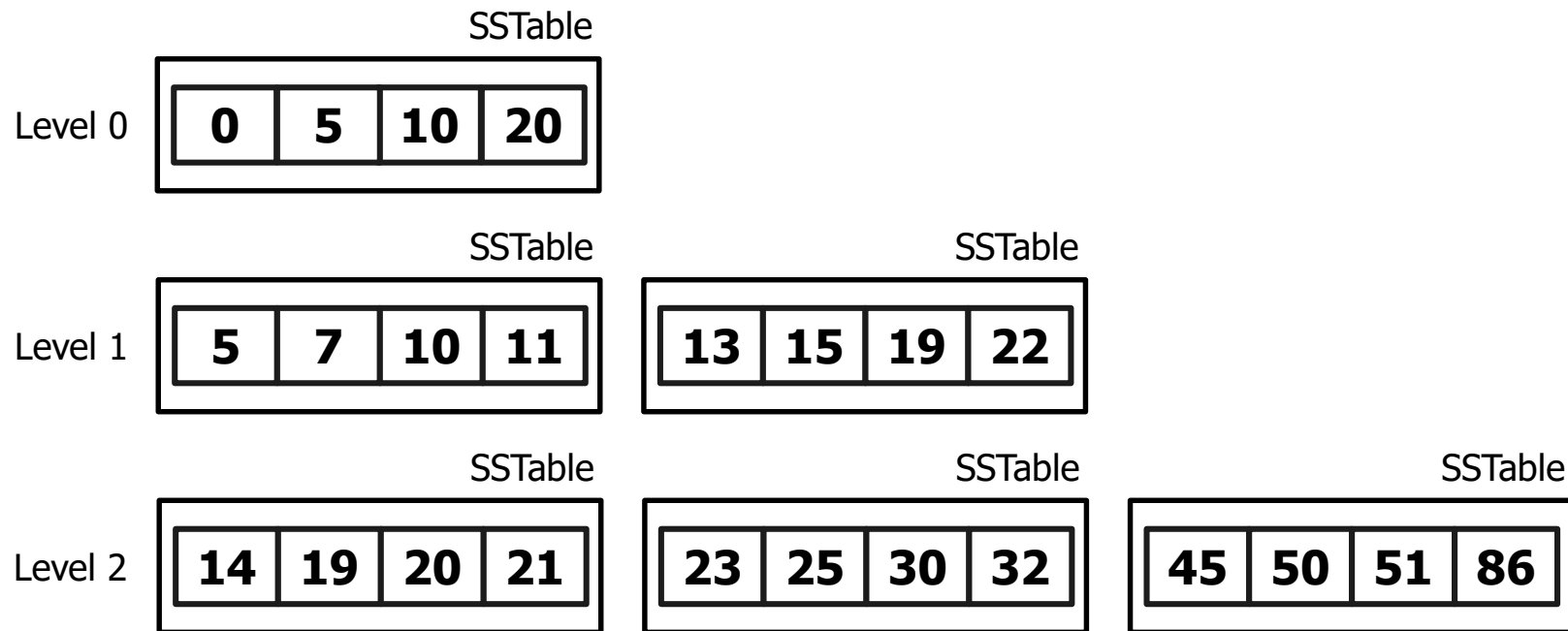
RocksDB Festival: Compaction

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

SSTable

Level 0

0	5	10	20

SSTable   SSTable

Level 1

5	7	10	11

13	15	19	22

SSTable   SSTable   SSTable

Level 2

14	19	20	21

23	25	30	32

45	50	51	86

RocksDB Festival: Compaction

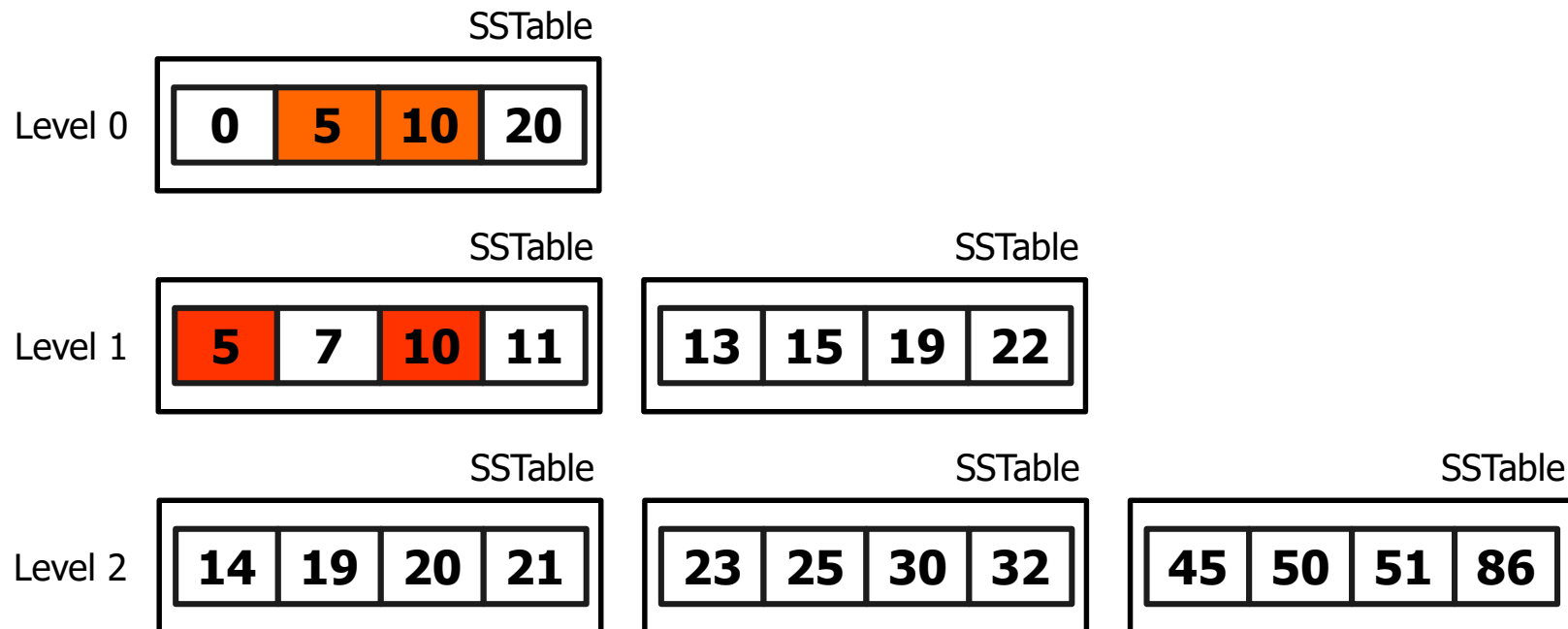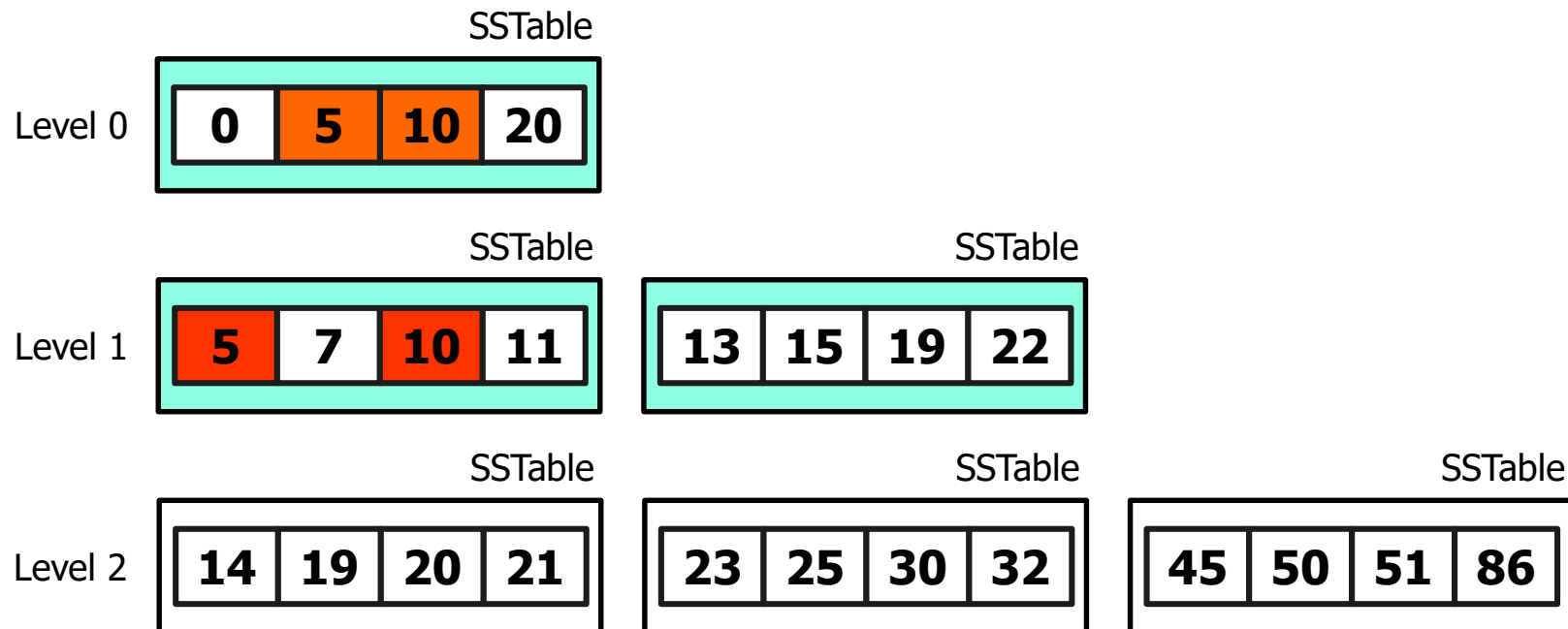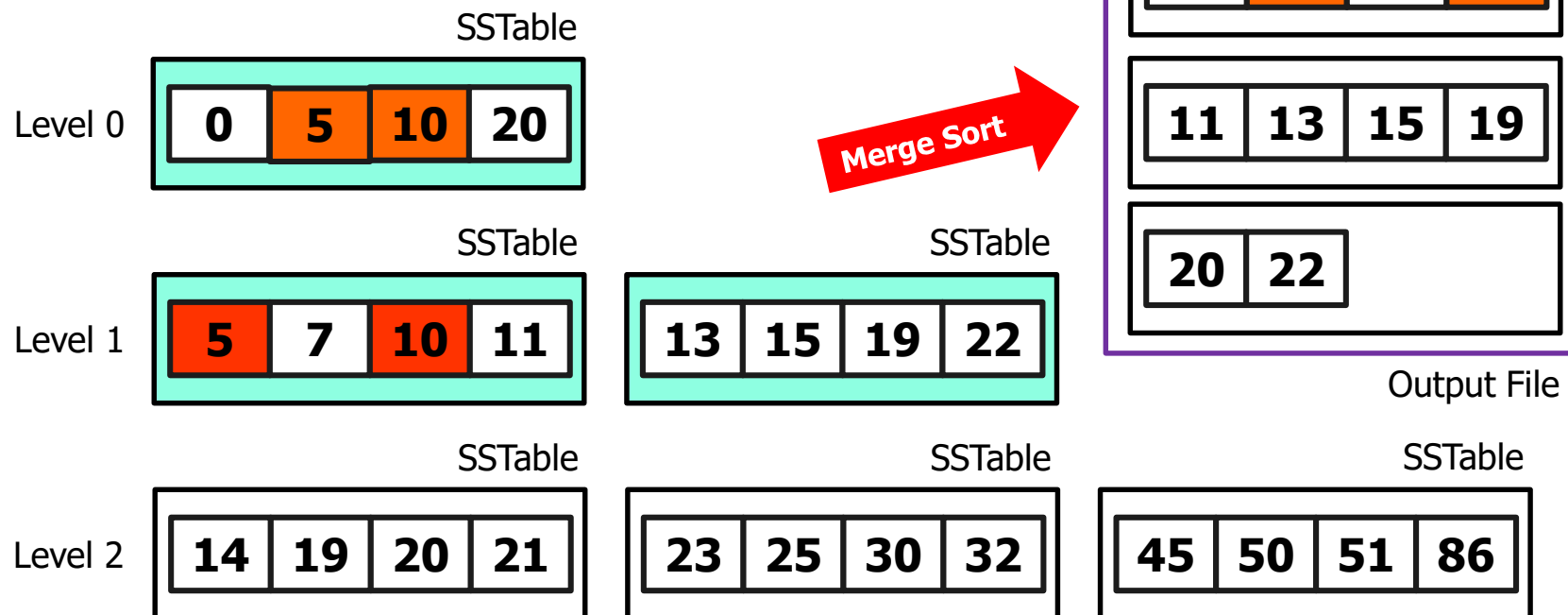- **RocksDB Operation: Compaction**
  - ✓ Leveled Compaction, Universal Compaction, FIFO compaction
  - ✓ Write Amplification, Space Amplification

RocksDB Festival: Compaction

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

SSTable

Level 0

| 0 | 5 | 10 | 20 |

SSTable

SSTable

Level 1

| 5 | 7 | 10 | 11 |

| 13 | 15 | 19 | 22 |

SSTable

SSTable

SSTable

Level 2

| 14 | 19 | 20 | 21 |

| 23 | 25 | 30 | 32 |

| 45 | 50 | 51 | 86 |

Merge Sort

| 0 | 5 | 7 | 10 |

| 11 | 13 | 15 | 19 |

| 20 | 22 |

Output File

RocksDB Festival: Compaction

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

Level 0

	SSTable			SSTable			SSTable
Level 1	**0** **5** **7** **10**			**11** **13** **15** **19**			**20** **22**

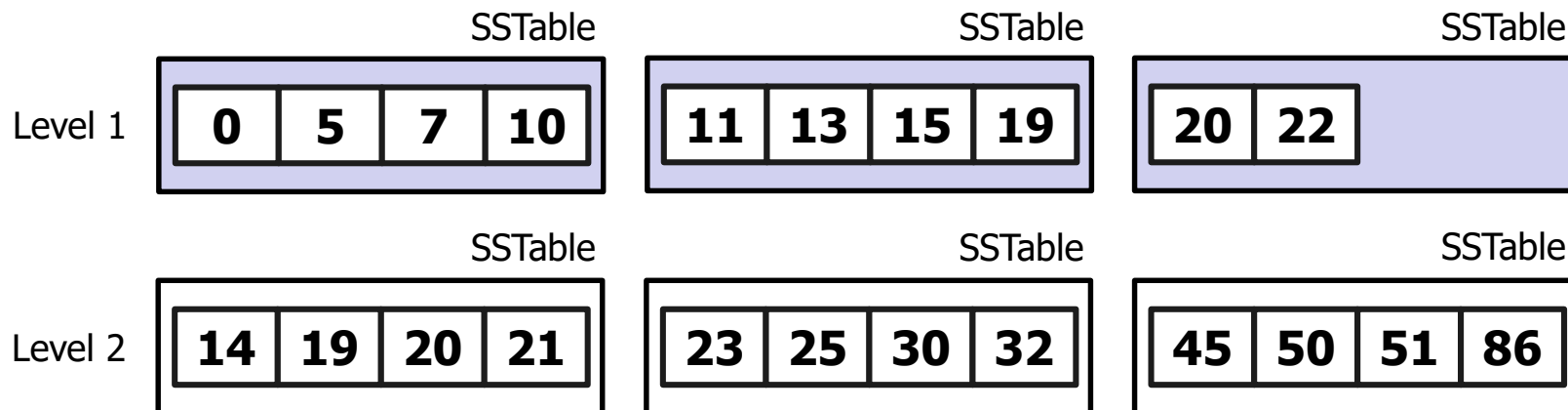|	SSTable			SSTable			SSTable
Level 2	**14** **19** **20** **21**			**23** **25** **30** **32**			**45** **50** **51** **86**

RocksDB Festival: Compaction

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

Level 0

	SSTable				SSTable				SSTable	
Level 1	0	5	7	10		11	13	15	19	20	22

	SSTable				SSTable				SSTable	
Level 2	14	19	20	21		23	25	30	32	45	50	51	86

RocksDB Festival: Compaction

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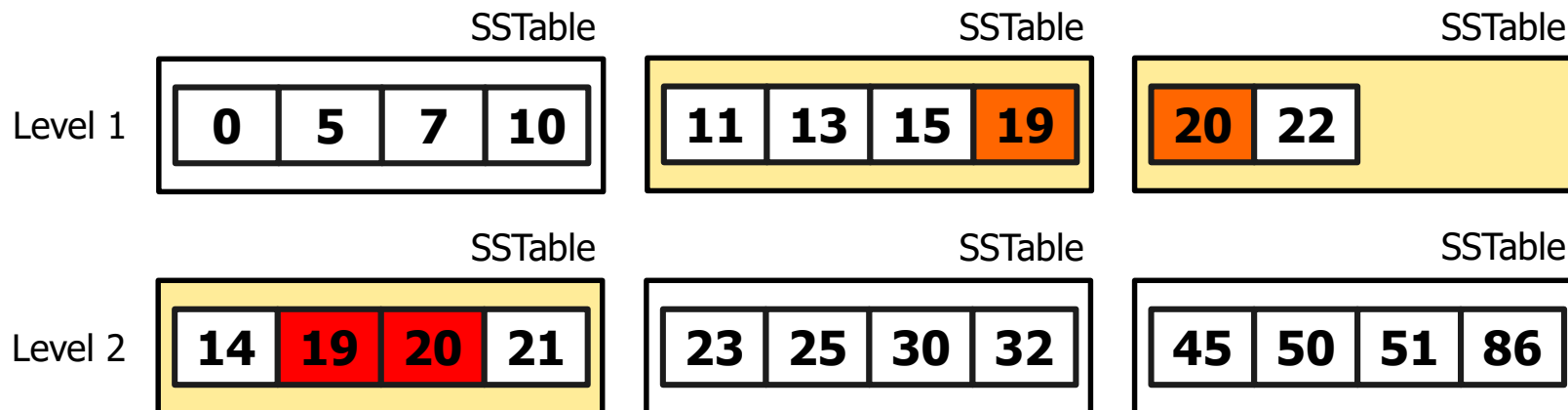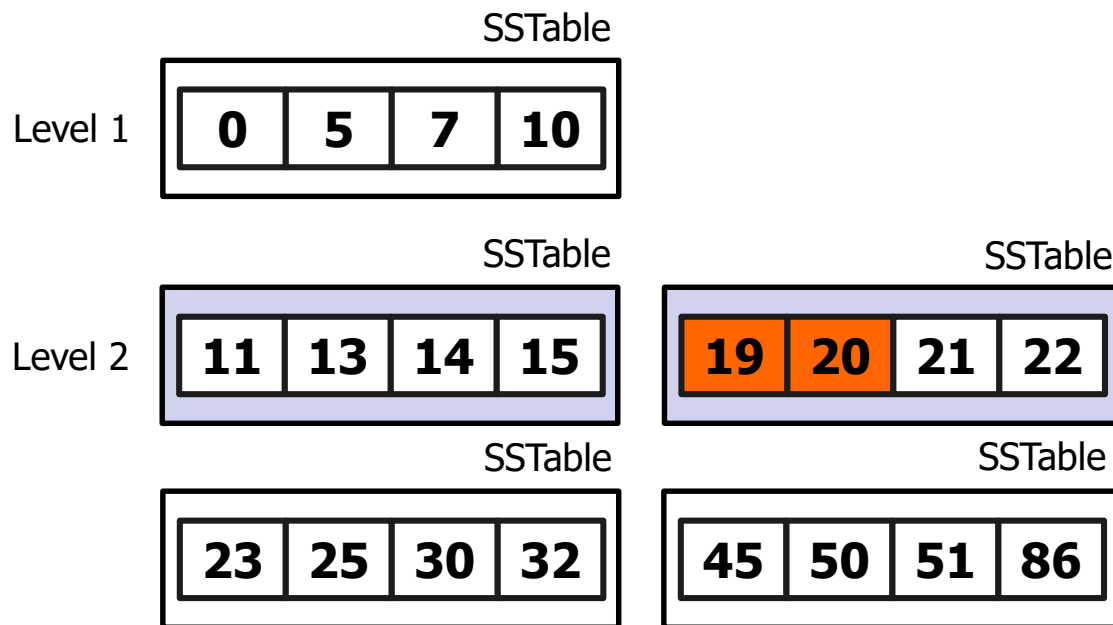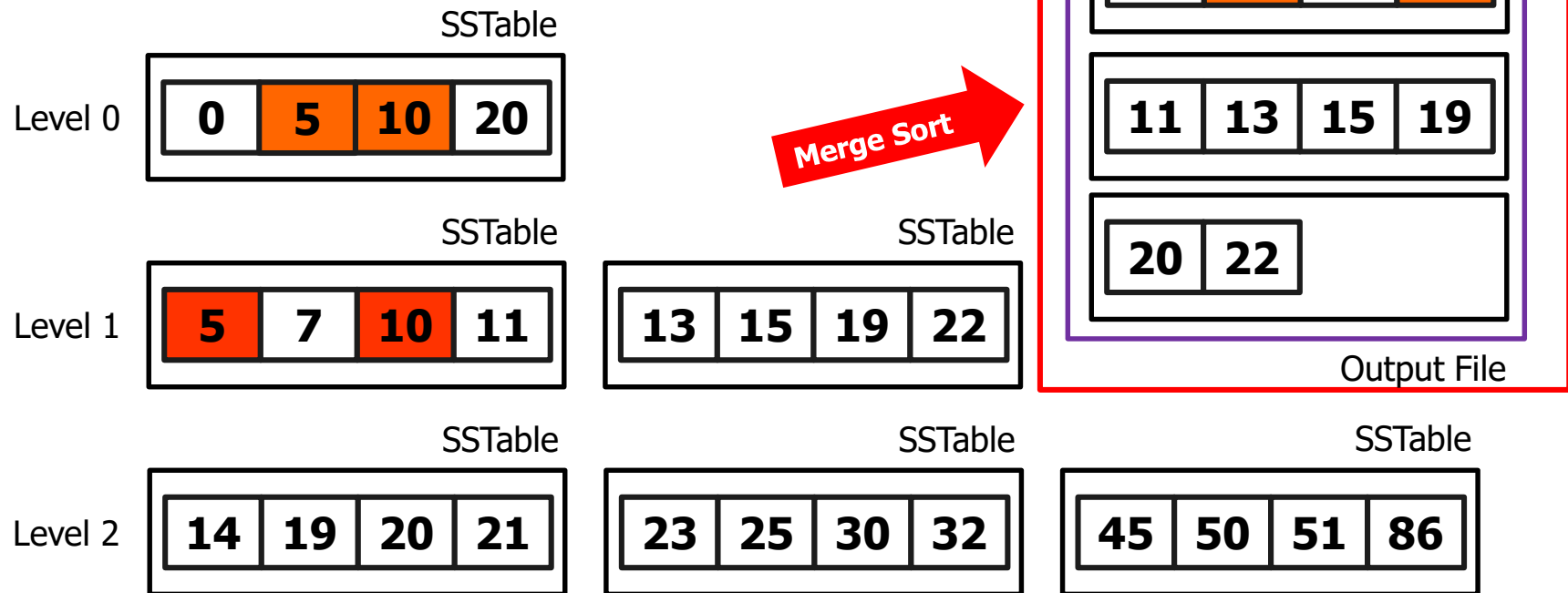 - ✓ Write Amplification, Space Amplification

Level 0

SSTable

Level 1

0	5	7	10

SSTable   SSTable

Level 2

11	13	14	15

19	20	21	22

SSTable   SSTable

23	25	30	32

| 45	50	51	86

RocksDB Festival: Compaction

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

Additional Write

SSTable

Level 0

| 0 | 5 | 10 | 20 |

Merge Sort

| 0 | 5 | 7 | 10 |

| 11 | 13 | 15 | 19 |

| 20 | 22 |

Output File

SSTable   SSTable

Level 1

| 5 | 7 | 10 | 11 | | 13 | 15 | 19 | 22 |

SSTable   SSTable   SSTable

Level 2

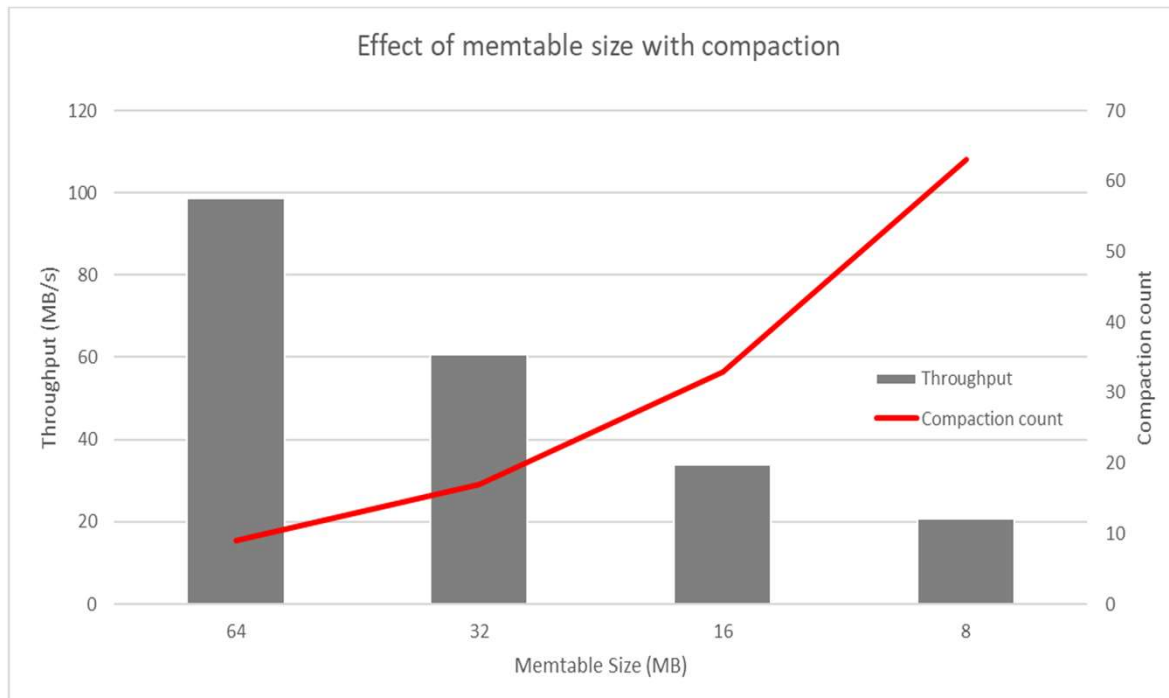| 14 | 19 | 20 | 21 | | 23 | 25 | 30 | 32 | | 45 | 50 | 51 | 86 |

RocksDB Festival: Compaction

- **RocksDB Operation: Compaction**
  - ✓ Space Amplification
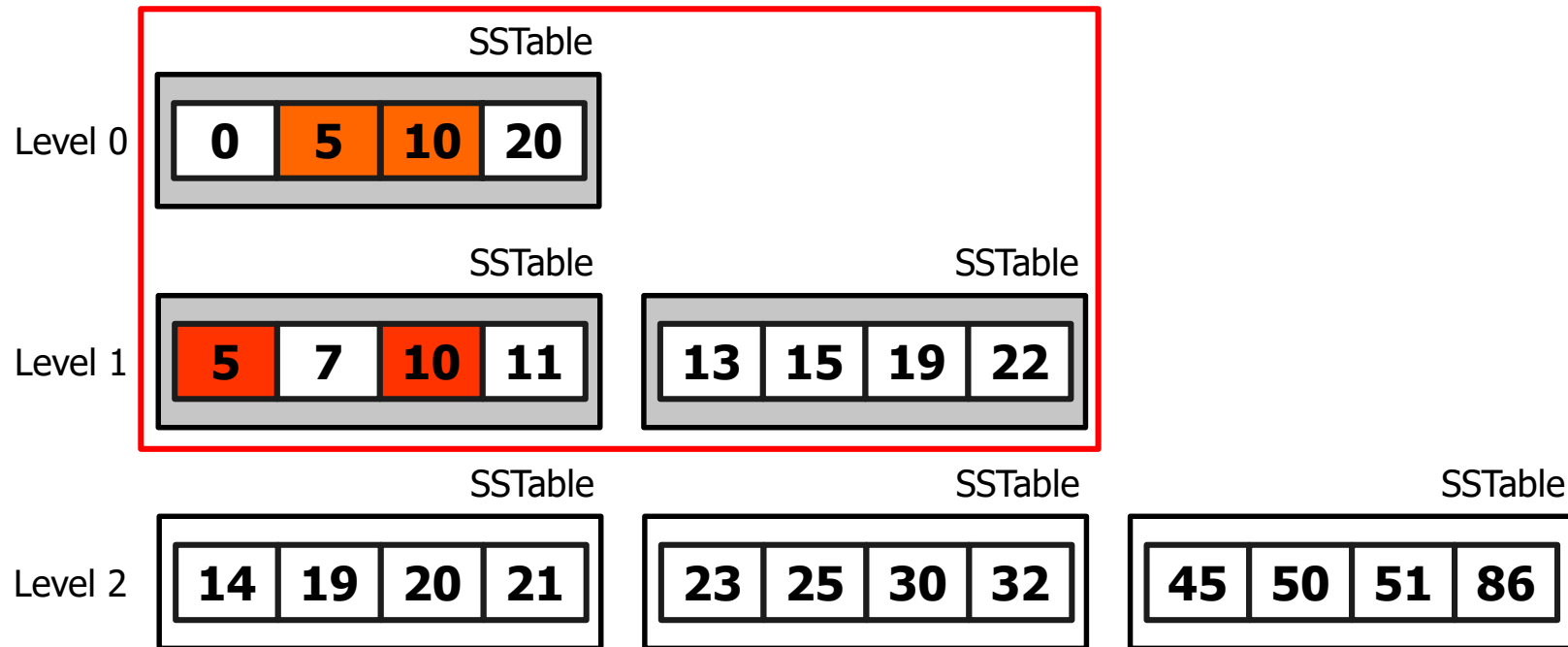
RocksDB Festival: Compaction

■ **RocksDB Operation: Compaction**
  ✓ Compaction Overhead

Effect of memtable size with compaction

RocksDB Festival: Compaction

- **RocksDB Operation: Compaction**
  - ✓ Compaction Trigger

RocksDB Festival

- **Next Week**
  - ✓ 각 팀 발표 준비
    - 팀 구성
    - **RocksDB db_bench** 설치 및 실험 결과 (관심 분야)
    - 선택한 주제와 선택한 이유
    - 이번 스터디 동안 기대하는 결과

QnA