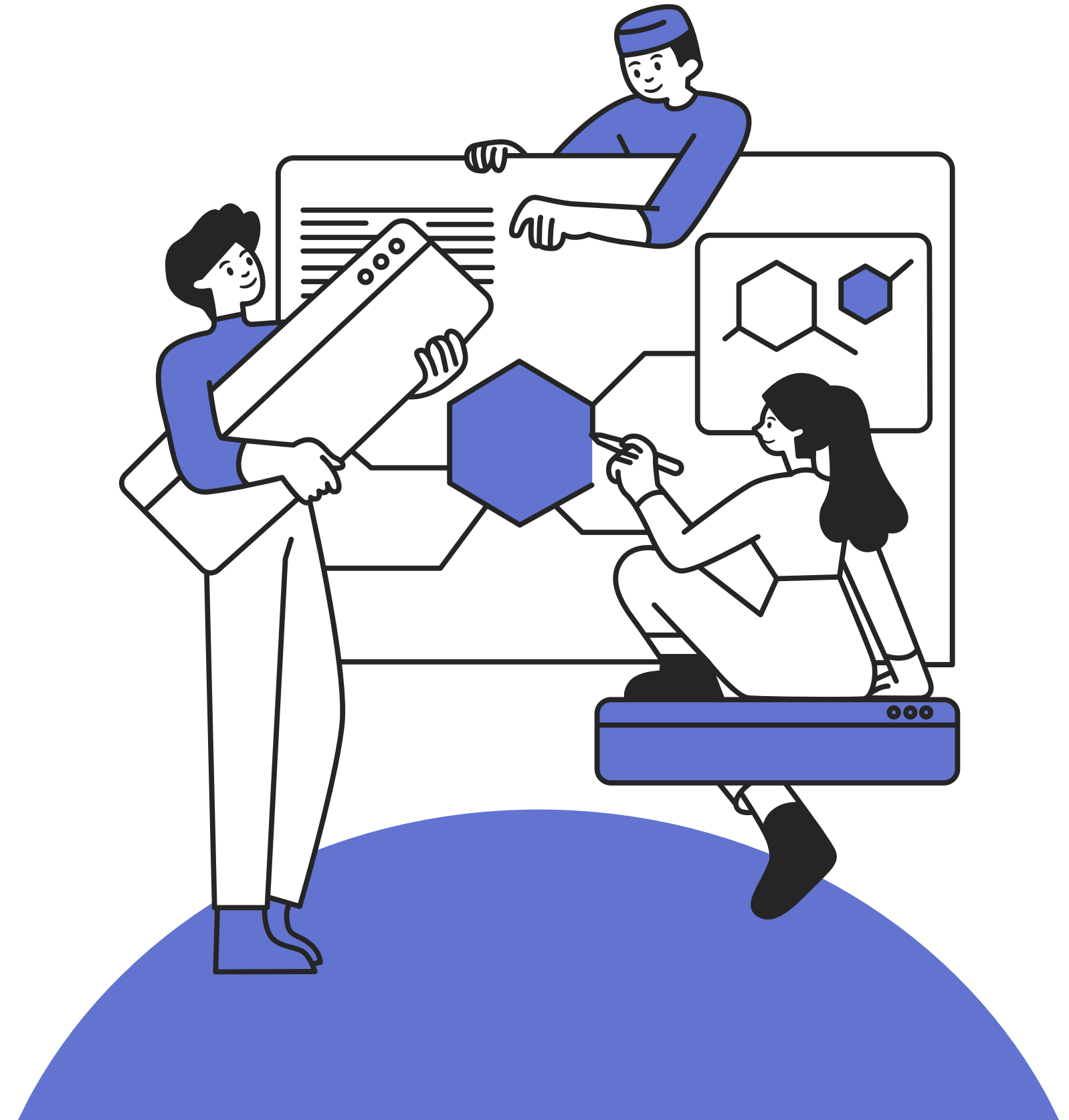


# KV Separation Mechanism

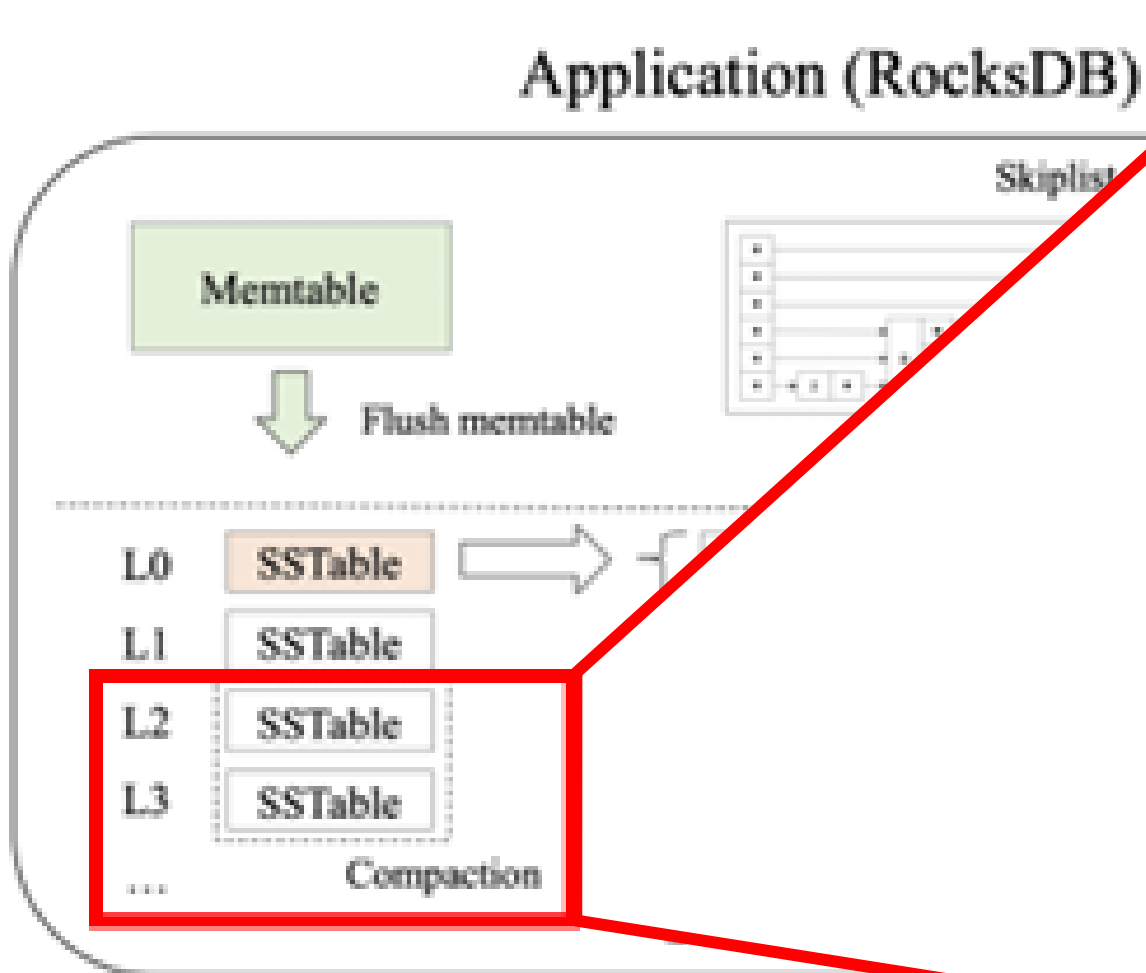
김은하, 유연주, 하지원



# 01

## What is BlobDB?

등장 배경: 기존 RocksDB의 한계



Compaction의 작동 방식: Read → Merge Sort → Write

문제점: Key만 정리하려 해도 붙어있는 Value까지 함께 읽고 다시 쓰는 과정이 반복된다.

$$WA \propto \frac{\text{Total Bytes Written to Disk}}{\text{Database Size}}$$

결과: WAF(Write Amplification Factor, 쓰기 증폭)가 급격히 증가하여 SSD 수명을 단축 시키고 시스템 성능을 저하 시킨다.

# 01

## What is BlobDB?

핵심 개념: Key-Value Separation (WiscKey)

기존 방식



SST = [ Key | Value ]

BlobDB 방식 (LSM-Tree (SST))



[ Key | Blob Pointer ]  
-> 작은 메타데이터만 저장

BlobDB 방식 (Blob File)

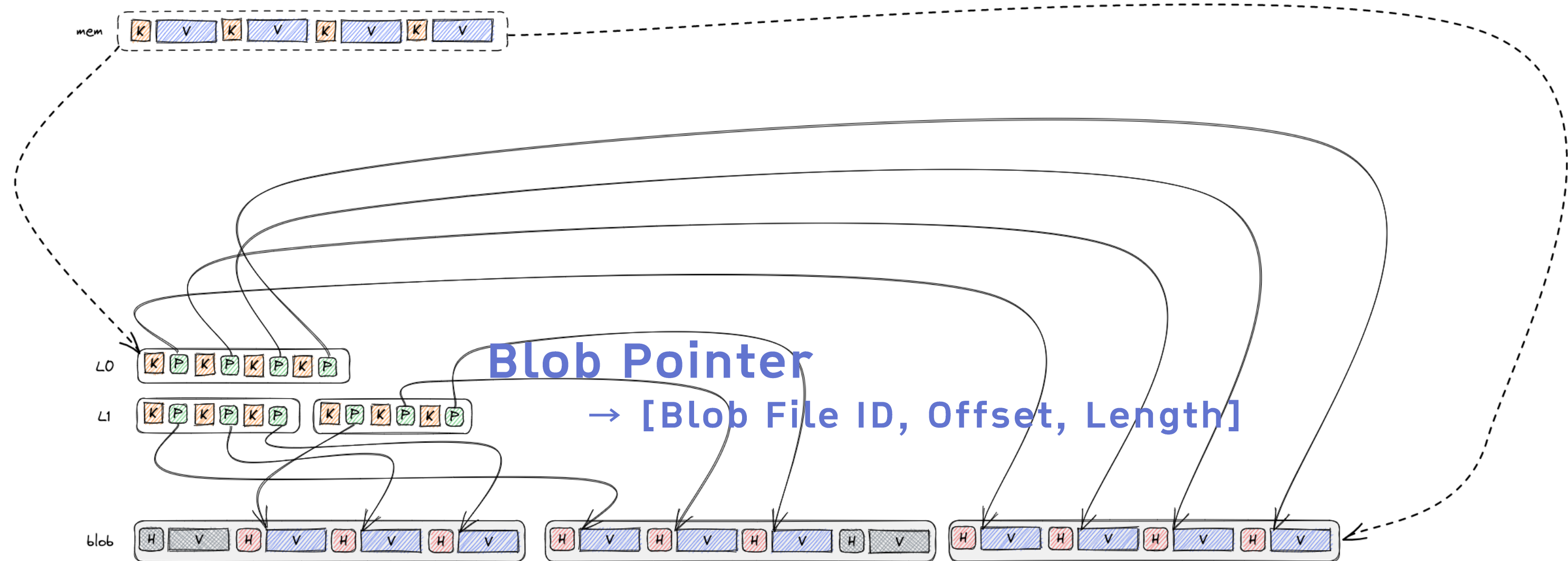


[ Real Large Value ]  
-> 실제 큰 데이터를 별도로 순차 저장

# 02

## How does it works in RocksDB?

### : 아키텍처 및 저장구조



**Blob Pointer**

→ [Blob File ID, Offset, Length]

**Blob File**

LSM tree with key-value separation. SST files (white background) contain keys (K) and blob pointers (P). Blob files (gray background) contain blob headers (H) and the actual values (V). Some blobs in the blob files may be unreferenced garbage.

source: <https://github.com/facebook/rocksdb/wiki/BlobDB>

# 02

## How does it works in RocksDB?

### : 데이터 쓰기 (Put) 프로세스

1. 임계값 확인
2. 로그 기록
3. Blob 파일 기록
4. 인덱스 생성
5. MemTable 저장

Write 요청 발생

↓

Value Size 체크

↓

if value < threshold

→ 기존 RocksDB 경로

else

→ BlobDB 경로

# 02

## How does it works in RocksDB?

### : 데이터 쓰기 (Put) 프로세스

1. 임계값 확인
2. 로그 기록
3. Blob 파일 기록
4. 인덱스 생성
5. MemTable 저장

데이터 유실 방지를 위해  
WAL(Write Ahead Log)에  
먼저 기록한다.

# 02

## How does it works in RocksDB?

### : 데이터 쓰기 (Put) 프로세스

1. 임계값 확인
2. 로그 기록
3. Blob 파일 기록
4. 인덱스 생성
5. MemTable 저장

실제 Value를 활성화된  
Blob File의 끝에  
추가(Append) 한다

# 02

## How does it works in RocksDB?

### : 데이터 쓰기 (Put) 프로세스

1. 임계값 확인
2. 로그 기록
3. Blob 파일 기록
4. 인덱스 생성
5. MemTable 저장

해당 Value가 저장된  
위치 정보(Blob Pointer)  
생성

# 02

## How does it works in RocksDB?

### : 데이터 쓰기 (Put) 프로세스

1. 임계값 확인
2. 로그 기록
3. Blob 파일 기록
4. 인덱스 생성
5. MemTable 저장

Key와 Blob Pointer를  
짜지어 MemTable에 저장

이후 Flush → SST 파일

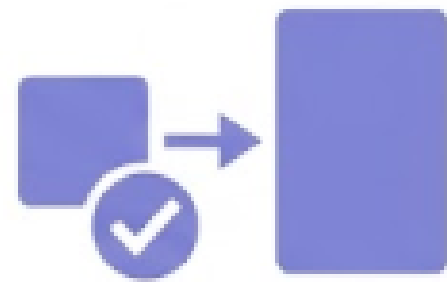
## 02

# How does it works in RocksDB?

: 데이터 읽기 (Get) 프로세스



LSM-Tree 탐색



포인터 확인



Blob 파일 접근



반환

# 02

## How does it works in RocksDB?

### : 삭제 (Delete), 업데이트 및 Garbage Collection (GC)

#### 삭제/업데이트

- 새로운 위치 포인터 쓰기
- 삭제 마커인 Tombstone 기록
- 기존 버전 Value는  
고아(Orphan)상태

#### Blob GC (Garbage Collection)

- 트리거: 특정 Blob 파일의 유효 데이터 비율이 낮아지면 GC 동작
- 재배치: 유효 데이터만 새 Blob 파일로 옮겨적기
- 메타데이터 갱신: Blob Pointer 업데이트
- 파일 삭제: 기존 Blob 파일 삭제

# 03

## BlobDB의 장단점

### Pros

- 쓰기 증폭(WAF) 최소화
- SSD 수명 연장
- 순차 쓰기 최적화

### Cons

- 읽기 성능 저하
- 공간 증폭(Space Amplification) 가능성
- 복잡성

→ Value 크기가 크고, 잦은 업데이트가 발생하는 워크로드에 최적화

# 04

## 실험 목적

BlobDB가 정말 compaction  
비용을 줄이는가?



Big value ->  
BlobDB 유리?

Value 크기에 따라 효과가  
달라지는가?



Small value ->  
BlobDB 의미 없음?

# 05

## 실험 설계

01

### Case 구조

	RocksDB	BlobDB
Big (32KB)	Case 1	Case 2
Small (256B)	Case 3	Case 4

02

- Key count : 100,000
- Workload : fillrandom -> overwrite
- Duration : 300s
- Compression : none
- LSM params : 동일
- Metrics : ops/sec, compaction read, stall

# 06

## 결과 Big value (32KB) + RocksDB

```
Uptime(secs): 300.4 total, 300.3 interval
Flush(GB): cumulative 17.588, interval 17.588
AddFile(GB): cumulative 0.000, interval 0.000
AddFile(Total Files): cumulative 0, interval 0
AddFile(L0 Files): cumulative 0, interval 0
AddFile(Keys): cumulative 0, interval 0
Cumulative compaction: 43.51 GB write, 148.32 MB/s write, 37.29 GB read, 127.15 MB/s read, 385.2 seconds
Interval compaction: 43.51 GB write, 148.34 MB/s write, 37.29 GB read, 127.16 MB/s read, 385.2 seconds
Estimated pending compaction bytes: 9595639852
Write Stall (count): cf-l0-file-count-limit-delays-with-ongoing-compaction: 37, cf-l0-file-count-limit-stops-with-ongoing-compaction: 0, l0-file-count-limit-delays: 90, l0-file-count-limit-stops: 0, memtable-limit-delays: 123, memtable-limit-stops: 0, pending-compaction-bytes-delays: 0, pending-compaction-bytes-stops: 0, total-delays: 213, total-stops: 0
interval: 213 total count
Block cache AutoHyperClockCache@0x58ec989a77c0#3613 capacity: 32.00 MB seed: 867020383 usage: 4.00 KB table_size: 64 occupancy: 1 collections: 1 last_copies: 0 last_secs: 0.000103 secs_since: 300
Block cache entry stats(count,size,portion): Misc(1,0.00 KB,0%)

** File Read Latency Histogram By Level [default] **

** DB Stats **
Uptime(secs): 300.4 total, 300.3 interval
Cumulative writes: 583K writes, 583K keys, 583K commit groups, 1.0 writes per commit group, ingest: 17.84 GB, 60.82 MB/s
Cumulative WAL: 583K writes, 0 syncs, 583999.00 writes per sync, written: 17.84 GB, 60.82 MB/s
Cumulative stall: 00:03:39.932 H:M:S, 73.2 percent
Interval writes: 583K writes, 583K keys, 583K commit groups, 1.0 writes per commit group, ingest: 18268.35 MB, 60.83 MB/s
Interval WAL: 583K writes, 0 syncs, 583999.00 writes per sync, written: 17.84 GB, 60.83 MB/s
Interval stall: 00:03:39.932 H:M:S, 73.2 percent
Write Stall (count): write-buffer-manager-limit-stops: 0

user32222840@JeonJooOS:~/rocksdb$
```

## 결과

## Big value (32KB) + RocksDB

L0	23/22	1.67 GB	0.2	0.0	0.0	0.0	17.6	17.6	17.6	0.0	1.0	0.0	171.7	104.91
	62.35	231	0.454	583K	0	0.0	0.0							
L1	30/30	1.54 GB	0.0	29.0	15.9	13.1	20.7	20.7	7.6	0.0	1.3	169.4	120.9	175.40
	97.47	13	13.492	949K	271K	0.0	0.0							
L2	71/0	3.00 GB	1.2	8.3	4.6	3.7	5.2	5.2	1.5	1.5	1.1	80.7	50.8	104.93
	37.61	58	1.809	270K	100K	0.0	0.0							
Sum	124/52	6.21 GB	0.0	37.3	20.5	16.8	43.5	43.5	26.7	1.5	2.5	99.1	115.6	385.23
	197.42	302	1.276	1802K	372K	0.0	0.0							
Int	0/0	0.00 KB	0.0	37.3	20.5	16.8	43.5	43.5	26.7	1.5	2.5	99.1	115.6	385.23
	197.42	302	1.276	1802K	372K	0.0	0.0							
** Compaction Stats [default] **														
Priority	Files	Size	Score	Read(GB)	Rn(GB)	Rnp1(GB)	Write(GB)	WPreComp(GB)	Wnew(GB)	Moved(GB)	W-Amp	Rd(MB/s)	Wr(MB/s)	Comp(
sec)	CompMergeCPU(sec)	Comp(cnt)	Avg(sec)	KeyIn	KeyDrop	Rblob(GB)	Wblob(GB)							
-----														
Low	0/0	0.00 KB	0.0	37.3	20.5	16.8	25.9	25.9	9.1	0.0	0.0	136.2	94.7	280.33
	135.08	71	3.948	1219K	372K	0.0	0.0							
High	0/0	0.00 KB	0.0	0.0	0.0	0.0	17.6	17.6	17.6	0.0	0.0	0.0	171.7	104.91
	62.35	231	0.454	583K	0	0.0	0.0							
Blob file count: 0, total size: 0.0 GB, garbage size: 0.0 GB, space amp: 0.0														

# 06

## 결과 Big value (32KB) + BlobDB

```
Uptime(secs): 300.4 total, 300.4 interval
Flush(GB): cumulative 61.124, interval 61.124
AddFile(GB): cumulative 0.000, interval 0.000
AddFile(Total Files): cumulative 0, interval 0
AddFile(L0 Files): cumulative 0, interval 0
AddFile(Keys): cumulative 0, interval 0
Cumulative compaction: 61.66 GB write, 210.21 MB/s write, 0.59 GB read, 2.03 MB/s read, 361.0 seconds
Interval compaction: 61.66 GB write, 210.22 MB/s write, 0.59 GB read, 2.03 MB/s read, 361.0 seconds
Estimated pending compaction bytes: 0
Write Stall (count): cf-l0-file-count-limit-delays-with-ongoing-compaction: 0, cf-l0-file-count-limit-stops-with-ongoing-compaction: 0, l0-file-count-limit-delays: 0, l0-file-count-limit-stops: 0, memtable-limit-delays: 512, memtable-limit-stops: 0, pending-compaction-bytes-delays: 0, pending-compaction-bytes-stops: 0, total-delays: 512, total-stops: 0
interval: 512 total count
Block cache AutoHyperClockCache@0x61831a16c7c0#3737 capacity: 32.00 MB seed: 1435234329 usage: 4.00 KB table_size: 64 occupancy: 1 collections: 1 last_copies: 0 last_secs: 5.4e-05 secs_since: 300
Block cache entry stats(count,size,portion): Misc(1,0.00 KB,0%)

** File Read Latency Histogram By Level [default] **

** DB Stats **
Uptime(secs): 300.4 total, 300.4 interval
Cumulative writes: 2030K writes, 2030K keys, 2030K commit groups, 1.0 writes per commit group, ingest: 62.04 GB, 211.51 MB/s
Cumulative WAL: 2030K writes, 0 syncs, 2030999.00 writes per sync, written: 62.04 GB, 211.51 MB/s
Cumulative stall: 00:01:52.090 H:M:S, 37.3 percent
Interval writes: 2030K writes, 2030K keys, 2030K commit groups, 1.0 writes per commit group, ingest: 63532.64 MB, 211.53 MB/s
Interval WAL: 2030K writes, 0 syncs, 2030999.00 writes per sync, written: 62.04 GB, 211.53 MB/s
Interval stall: 00:01:52.090 H:M:S, 37.3 percent
Write Stall (count): write-buffer-manager-limit-stops: 0

user32222840@JeonJoo05:~/rocksdb$
```

## 결과

## Big value (32KB) + BlobDB

	Files	Size	Score	Read(GB)	Rn(GB)	Rnp1(GB)	Write(GB)	WPreComp(GB)	Wnew(GB)	Moved(GB)	W-Amp	Rd(MB/s)	Wr(MB/s)	Comp(sec)
L0	1/0	122.66 KB	0.2	0.0	0.0	0.0	0.1	0.1	0.1	0.0	1.0	0.0	224.2	279.18
	174.51	765	0.365	2028K	0	0.0	61.1							
L1	1/0	3.04 MB	0.0	0.6	0.1	0.5	0.5	0.5	0.0	0.0	8.9	7.4	6.7	81.82
	32.46	191	0.428	20M	1894K	0.0	0.0							
Sum	2/0	3.16 MB	0.0	0.6	0.1	0.5	0.6	0.6	0.1	0.0	1.0	1.7	174.9	361.00
	206.97	956	0.378	22M	1894K	0.0	61.1							
Int	0/0	0.00 KB	0.0	0.6	0.1	0.5	0.6	0.6	0.1	0.0	1.0	1.7	174.9	361.00
	206.97	956	0.378	22M	1894K	0.0	61.1							

\*\* Compaction Stats [default] \*\*

Priority	Files	Size	Score	Read(GB)	Rn(GB)	Rnp1(GB)	Write(GB)	WPreComp(GB)	Wnew(GB)	Moved(GB)	W-Amp	Rd(MB/s)	Wr(MB/s)	Comp(sec)
Low	0/0	0.00 KB	0.0	0.6	0.1	0.5	0.5	0.5	0.0	0.0	0.0	7.4	6.7	81.82
	32.46	191	0.428	20M	1894K	0.0	0.0							
High	0/0	0.00 KB	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	224.2	279.18
	174.51	765	0.365	2028K	0	0.0	61.1							

Blob file count: 315, total size: 25.5 GB, garbage size: 22.4 GB, space amp: 8.0

# 06

## 결과 Small value (256B) + RocksDB

```
Uptime(secs): 300.0 total, 300.0 interval
Flush(GB): cumulative 3.797, interval 3.797
AddFile(GB): cumulative 0.000, interval 0.000
AddFile(Total Files): cumulative 0, interval 0
AddFile(L0 Files): cumulative 0, interval 0
AddFile(Keys): cumulative 0, interval 0
Cumulative compaction: 4.87 GB write, 16.61 MB/s write, 4.84 GB read, 16.52 MB/s read, 73.9 seconds
Interval compaction: 4.87 GB write, 16.61 MB/s write, 4.84 GB read, 16.52 MB/s read, 73.9 seconds
Estimated pending compaction bytes: 127458799
Write Stall (count): cf-l0-file-count-limit-delays-with-ongoing-compaction: 0, cf-l0-file-count-limit-stops-with-ongoing-compaction:
0, l0-file-count-limit-delays: 0, l0-file-count-limit-stops: 0, memtable-limit-delays: 0, memtable-limit-stops: 0, pending-compaction-bytes-delays: 0, pending-compaction-bytes-stops: 0, total-delays: 0, total-stops: 0
Block cache AutoHyperClockCache@0x628b9dbca7c0#3857 capacity: 32.00 MB seed: 1938088603 usage: 4.00 KB table_size: 64 occupancy: 1 collections: 1 last_copies: 0 last_secs: 5.1e-05 secs_since: 300
Block cache entry stats(count,size,portion): Misc(1,0.00 KB,0%)

** File Read Latency Histogram By Level [default] **

** DB Stats **
Uptime(secs): 300.0 total, 300.0 interval
Cumulative writes: 36M writes, 36M keys, 36M commit groups, 1.0 writes per commit group, ingest: 9.71 GB, 33.15 MB/s
Cumulative WAL: 36M writes, 0 syncs, 36215999.00 writes per sync, written: 9.71 GB, 33.15 MB/s
Cumulative stall: 00:00:0.000 H:M:S, 0.0 percent
Interval writes: 36M writes, 36M keys, 36M commit groups, 1.0 writes per commit group, ingest: 9947.02 MB, 33.16 MB/s
Interval WAL: 36M writes, 0 syncs, 36215999.00 writes per sync, written: 9.71 GB, 33.16 MB/s
Interval stall: 00:00:0.000 H:M:S, 0.0 percent
Write Stall (count): write-buffer-manager-limit-stops: 0

user32222840@JeonJoo0S:~/rocksdb$
```

# 06

## 결과 Small value (256B) + RocksDB

```
** Compaction Stats [default] **
Level   Files   Size      Score Read(GB)  Rn(GB) Rnp1(GB) Write(GB) WPreComp(GB) Wnew(GB) Moved(GB) W-Amp Rd(MB/s) Wr(MB/s) Comp(sec)
) CompMergeCPU(sec) Comp(cnt) Avg(sec) KeyIn KeyDrop Rblob(GB) Wblob(GB)
-----
L0      4/4     94.86 MB    0.0      0.0      0.0      0.0      3.8      3.8      3.8      0.0      1.0      0.0      83.0      46.82
      34.40      164    0.286      35M      0      0.0      0.0
L1      1/1     26.69 MB    0.0      4.8      3.8      1.0      1.1      1.1      0.0      0.0      0.3     183.2     40.4      27.05
      21.39      41    0.660      18M     14M      0.0      0.0
Sum     5/5    121.55 MB    0.0      4.8      3.8      1.0      4.9      4.9      3.8      0.0      1.3      67.1     67.4      73.88
      55.78     205    0.360      54M     14M      0.0      0.0
Int     0/0      0.00 KB    0.0      4.8      3.8      1.0      4.9      4.9      3.8      0.0      1.3      67.1     67.4      73.88
      55.78     205    0.360      54M     14M      0.0      0.0

** Compaction Stats [default] **
Priority Files   Size      Score Read(GB)  Rn(GB) Rnp1(GB) Write(GB) WPreComp(GB) Wnew(GB) Moved(GB) W-Amp Rd(MB/s) Wr(MB/s) Comp(
sec) CompMergeCPU(sec) Comp(cnt) Avg(sec) KeyIn KeyDrop Rblob(GB) Wblob(GB)
-----
Low     0/0      0.00 KB    0.0      4.8      3.8      1.0      1.1      1.1      0.0      0.0      0.0     183.2     40.4      27.05
      21.39      41    0.660      18M     14M      0.0      0.0
High    0/0      0.00 KB    0.0      0.0      0.0      0.0      3.8      3.8      3.8      0.0      0.0      0.0      83.0      46.82
      34.40     164    0.286      35M      0      0.0      0.0

Blob file count: 0, total size: 0.0 GB, garbage size: 0.0 GB, space amp: 0.0
```

# 06

## 결과 Small value (256B) + BlobDB

```
Uptime(secs): 300.0 total, 300.0 interval
Flush(GB): cumulative 3.403, interval 3.403
AddFile(GB): cumulative 0.000, interval 0.000
AddFile(Total Files): cumulative 0, interval 0
AddFile(L0 Files): cumulative 0, interval 0
AddFile(Keys): cumulative 0, interval 0
Cumulative compaction: 4.34 GB write, 14.82 MB/s write, 4.25 GB read, 14.49 MB/s read, 73.6 seconds
Interval compaction: 4.34 GB write, 14.82 MB/s write, 4.25 GB read, 14.49 MB/s read, 73.6 seconds
Estimated pending compaction bytes: 0
Write Stall (count): cf-l0-file-count-limit-delays-with-ongoing-compaction: 0, cf-l0-file-count-limit-stops-with-ongoing-compaction: 0, l0-file-count-limit-delays: 0, l0-file-count-limit-stops: 0, memtable-limit-delays: 0, memtable-limit-stops: 0, pending-compaction-bytes-delays: 0, pending-compaction-bytes-stops: 0, total-delays: 0, total-stops: 0
Block cache AutoHyperClockCache@0x557a6d2447c0#3973 capacity: 32.00 MB seed: 320770534 usage: 4.00 KB table_size: 64 occupancy: 1 collections: 1 last_copies: 0 last_secs: 5.7e-05 secs_since: 300
Block cache entry stats(count,size,portion): Misc(1,0.00 KB,0%)

** File Read Latency Histogram By Level [default] **

** DB Stats **
Uptime(secs): 300.0 total, 300.0 interval
Cumulative writes: 32M writes, 32M keys, 32M commit groups, 1.0 writes per commit group, ingest: 8.71 GB, 29.72 MB/s
Cumulative WAL: 32M writes, 0 syncs, 32459999.00 writes per sync, written: 8.71 GB, 29.72 MB/s
Cumulative stall: 00:00:0.000 H:M:S, 0.0 percent
Interval writes: 32M writes, 32M keys, 32M commit groups, 1.0 writes per commit group, ingest: 8915.40 MB, 29.72 MB/s
Interval WAL: 32M writes, 0 syncs, 32459999.00 writes per sync, written: 8.71 GB, 29.72 MB/s
Interval stall: 00:00:0.000 H:M:S, 0.0 percent
Write Stall (count): write-buffer-manager-limit-stops: 0

user32222840@JeonJoo0S:~/rocksdb$
```

## 결과 Small value (256B) + BlobDB

```
Blob file count: 0, total size: 0.0 GB, garbage size: 0.0 GB, space amp: 0.0
```

# 06 결과

Case	Value	구조	Compaction read	Stall	Blob file
1	32KB	RocksDB	37.29 GB	73.2%	X
2	32KB	BlobDB	0.59 GB	37.3%	315개
3	256B	RocksDB	4.84 GB	0%	X
4	256B	BlobDB	4.25 GB	0%	X

## Big value (32KB)

- Compaction read : 37.29 → 0.59 GB (대폭 감소)
- Stall : 73.2% → 37.3% (절반 수준 감소)
- Blob file 생성 0

## Small value (256B)

- Compaction read : 4.84 vs 4.25 (거의 차이X)
- Stall : 둘 다 0%
- Blob file 생성 X

=> Big value에서만 BlobDB 효과가 있다!

# THANK YOU

---