

Optimizing Space Amplification in RocksDB

Siying Dong, Mark Callaghan, Leonidas Galanis, Dhruba Borthakur,
Tony Savor, Michael Strum

Core Data Team, Facebook



Database Storage engine supports point / range lookup

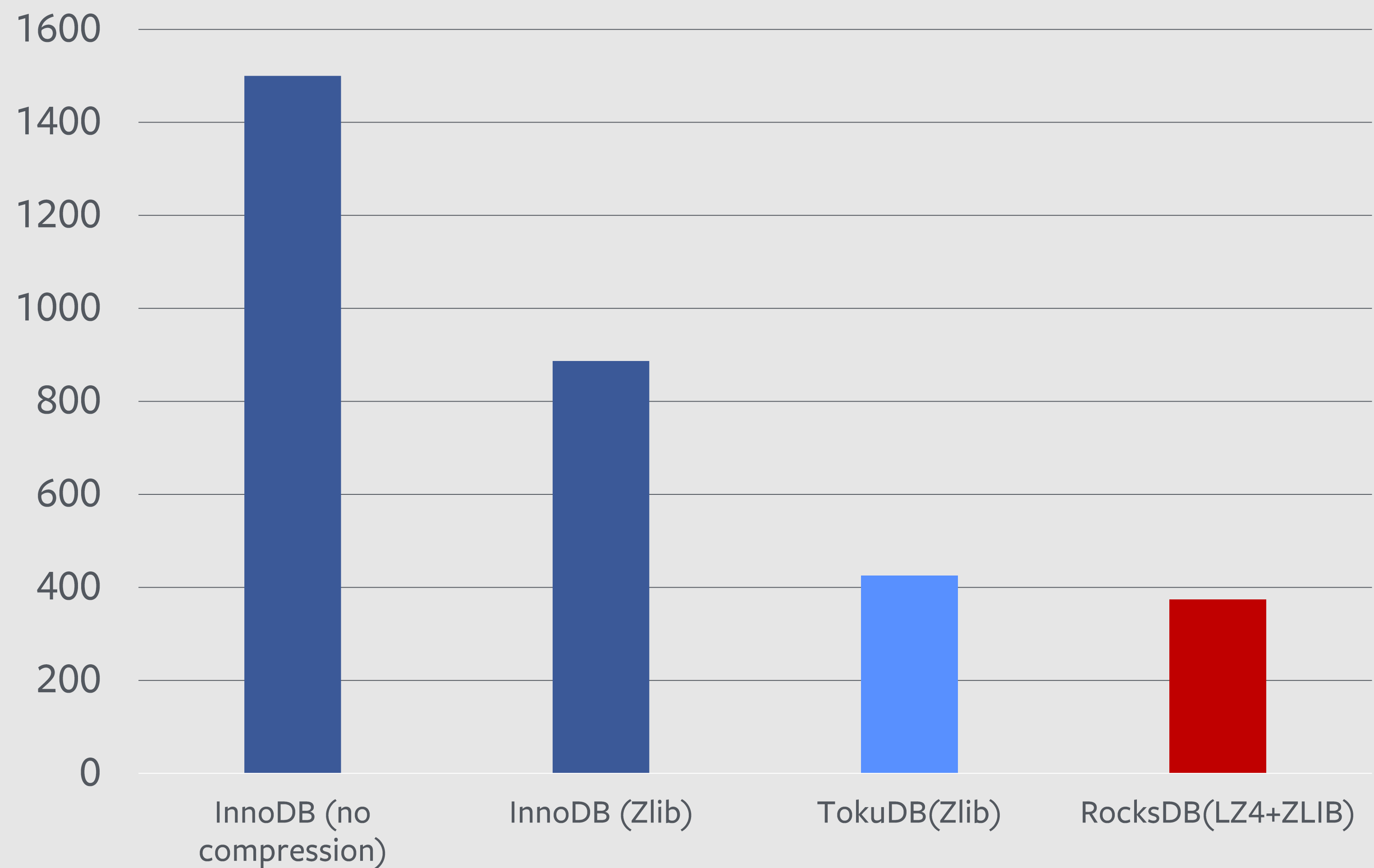
Usage of RocksDB

- MyRocks = MySQL + RocksDB
- MongoRocks = MongoDB + RocksDB
- Used in many other Facebook services
- Used in Yahoo, LinkedIn, Netflix, etc

How much can we improve
databases' space efficiency?

LinkBench DB Size Using MySQL

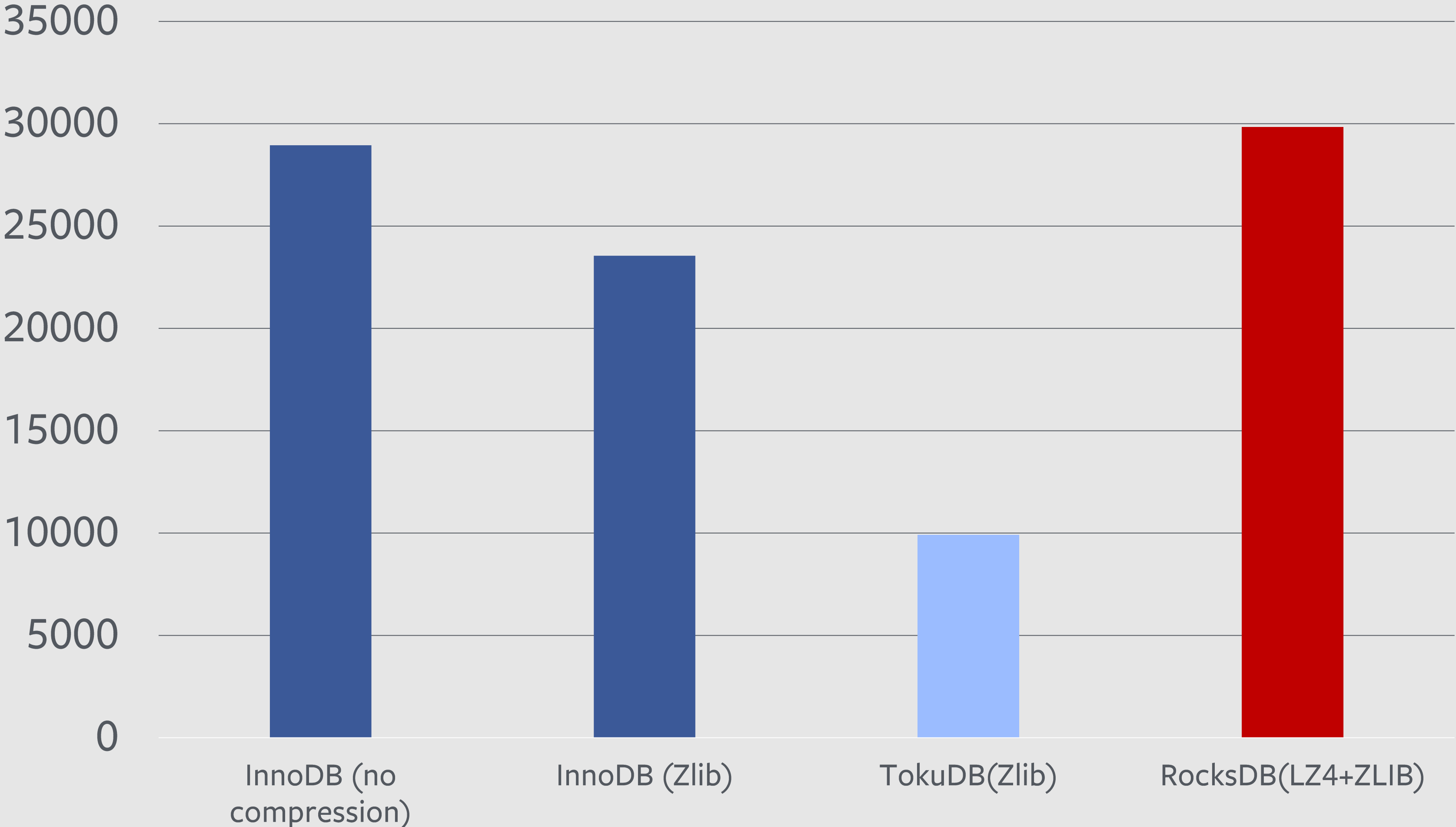
Size: GB



While preserving performance?

LinkBench Transaction/s

Transaction/s



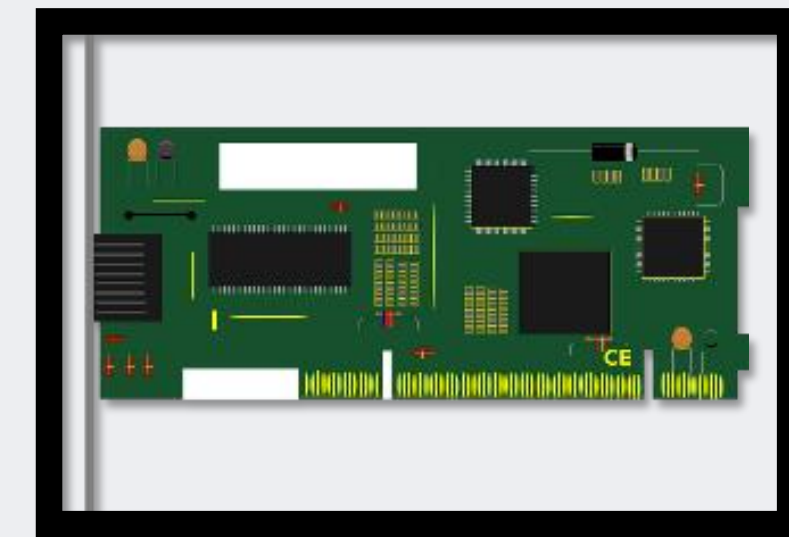
Why space efficiency?

Space Efficiency Is Important

HDD



SSD



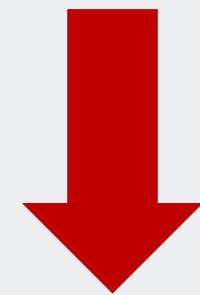
Random IOPS

Space Efficiency Is Important

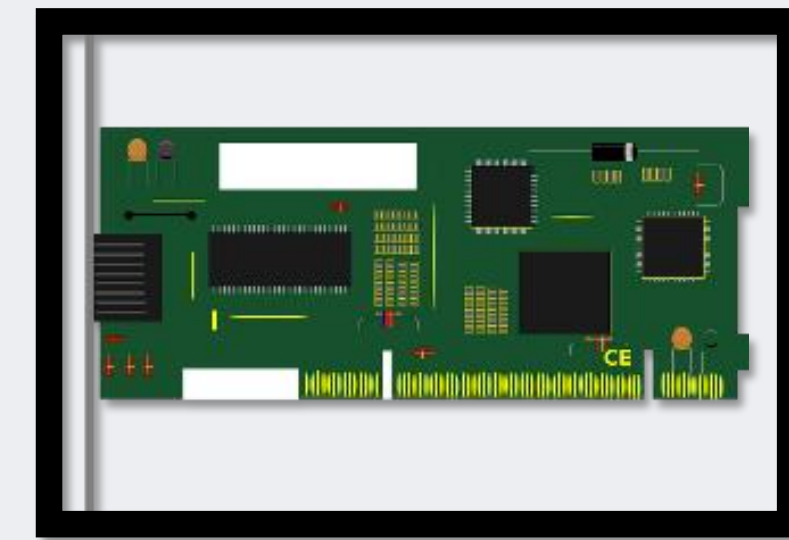
HDD



What we need



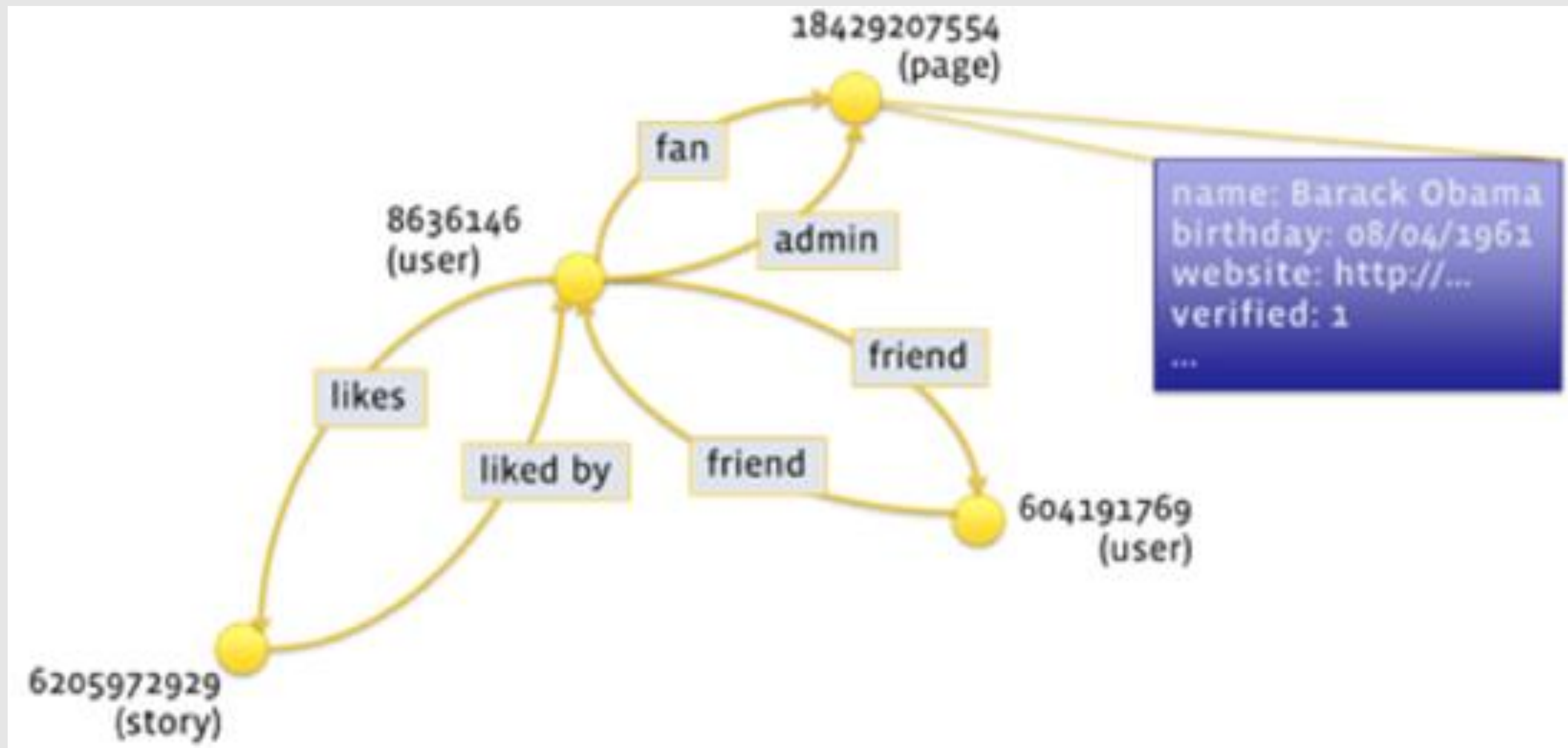
SSD



Random IOPS

Our Workload

LinkBench: benchmark that emulates Facebook social graph queries



Our Workloads for the Social Graph

- Data rows are in the order of magnitude of bytes
- Writes are random
- Reads are random
 - Half are point lookups
 - Half are range-queries
- Read/write ratio 2:1

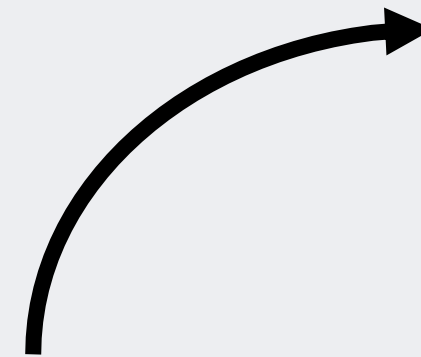
Opportunities to Improve Space Efficiency

Space inefficiency of the B-Tree

- Page fragmentation
- Burn SSD erase cycles fast
- CPU heavy to apply heavyweight compression

Random Updates Using B-tree

Page Buffer



Row 1
Row 2'
Row 3
.....
Row N

Write +
Compress



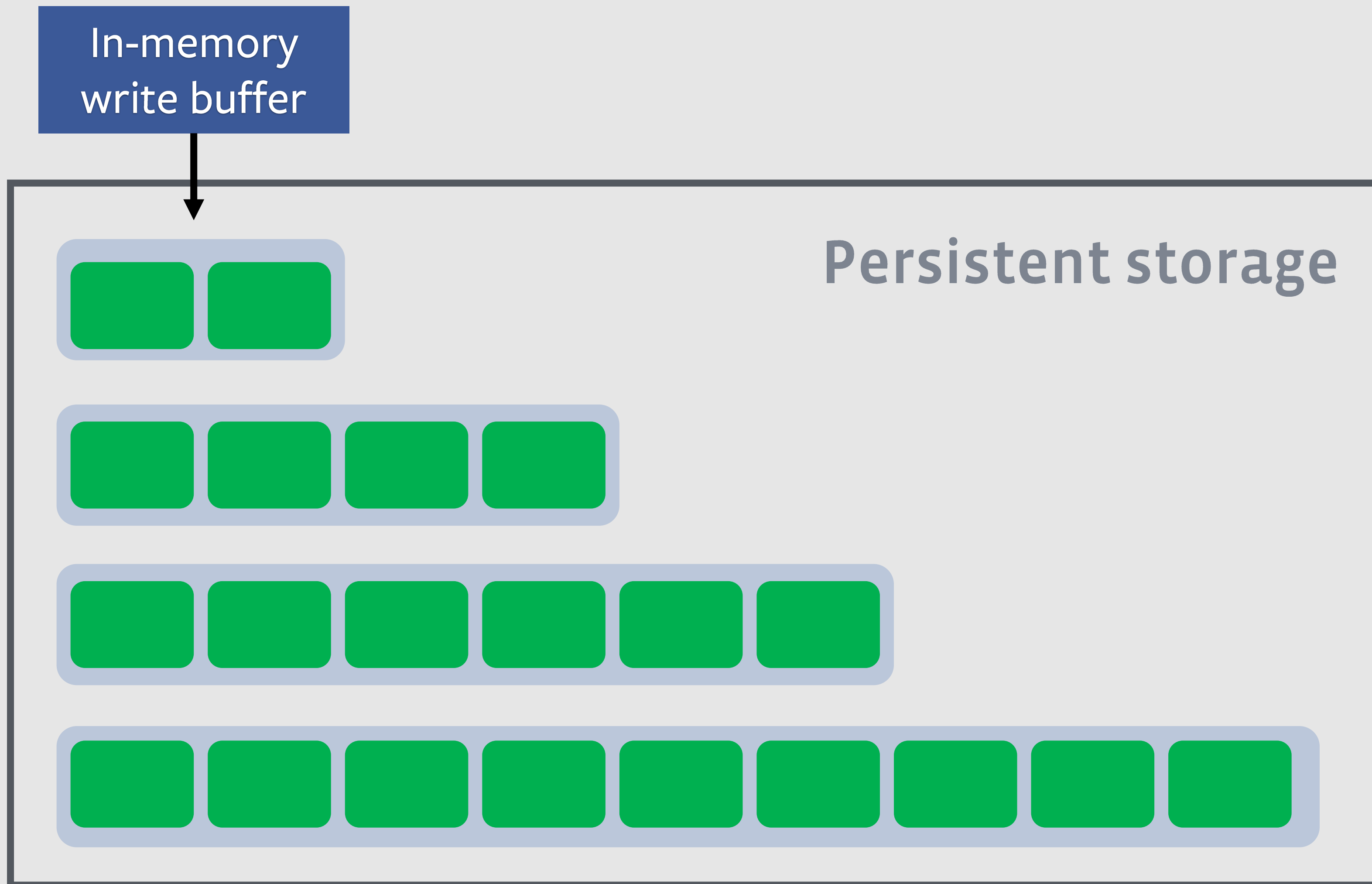
Disk

Row 1
Row 2
Row 3
.....
Row N

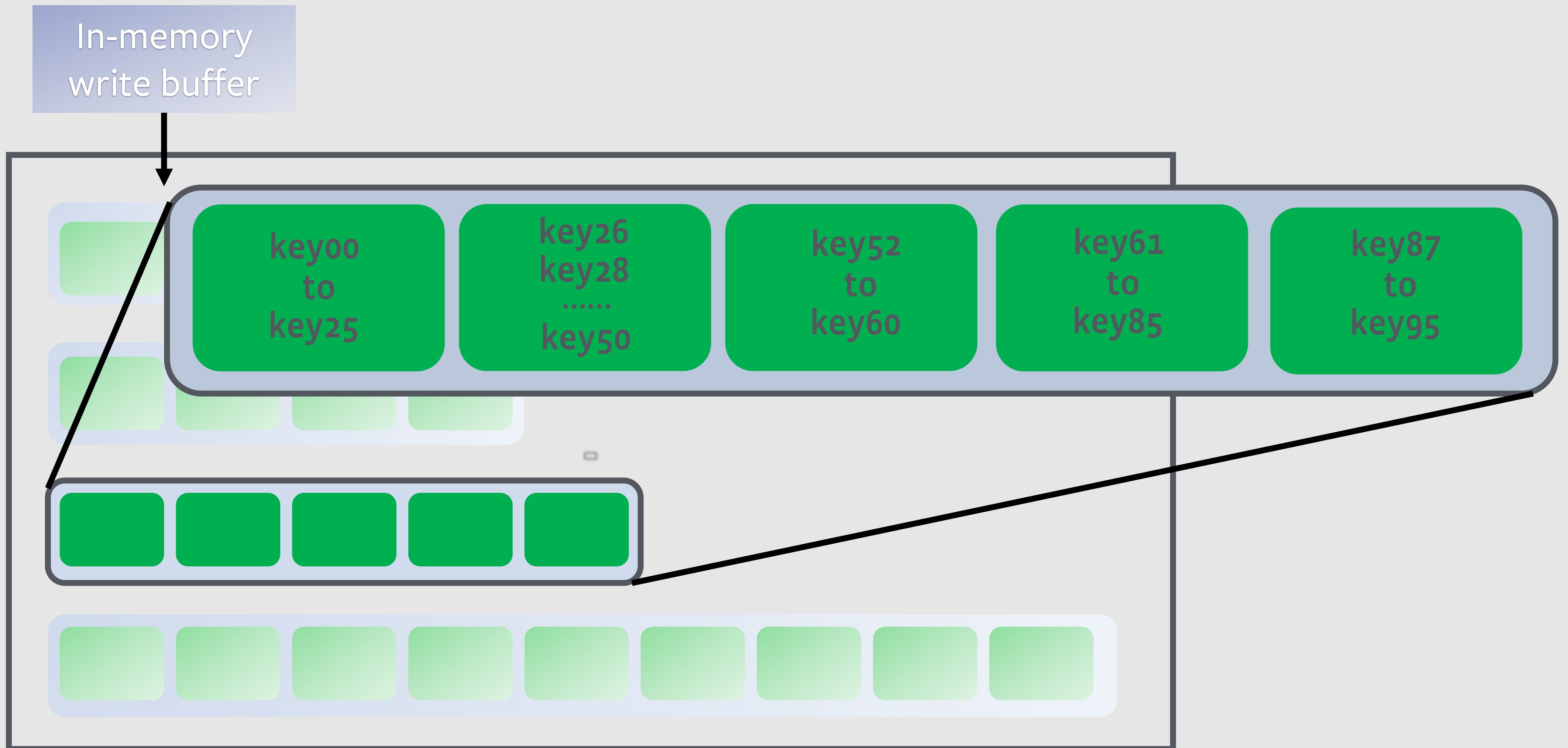
Row 1
Row 2'
Row 3
.....
Row N

Optimizing Space Efficiency in RocksDB

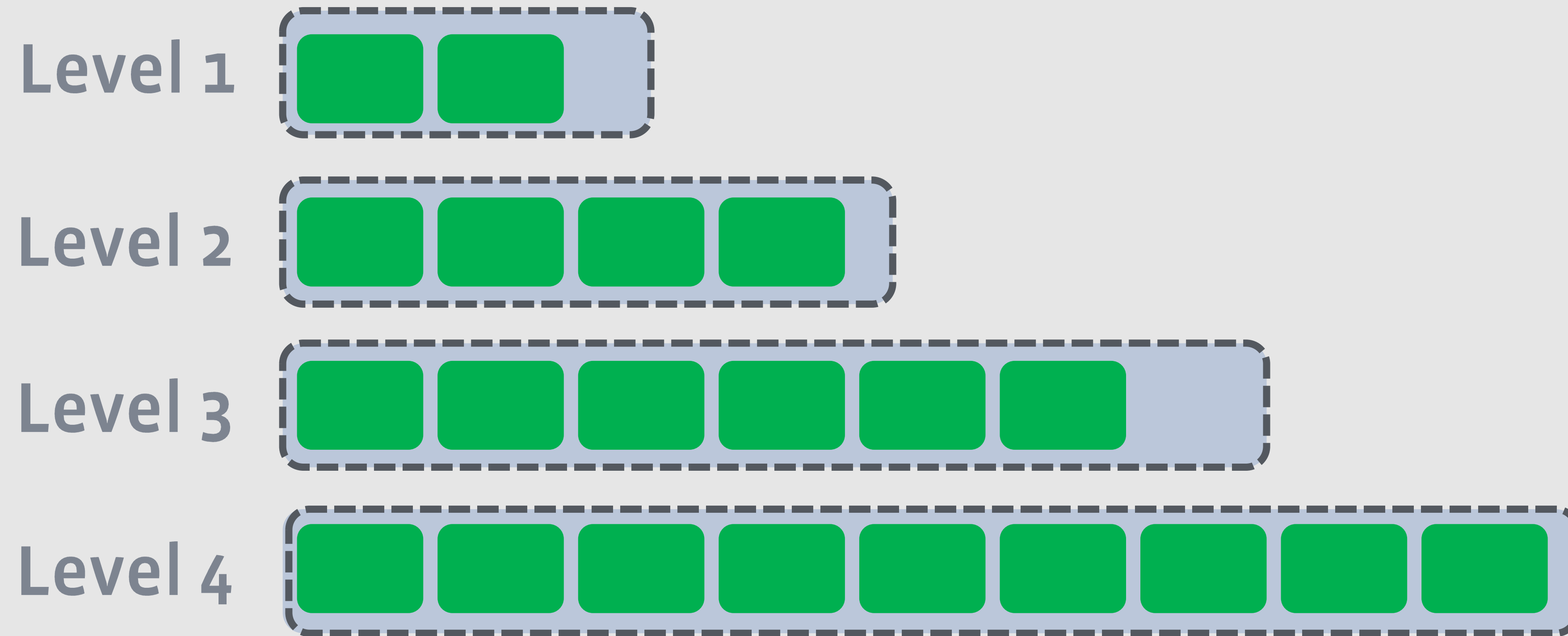
Log Structured Merge Tree



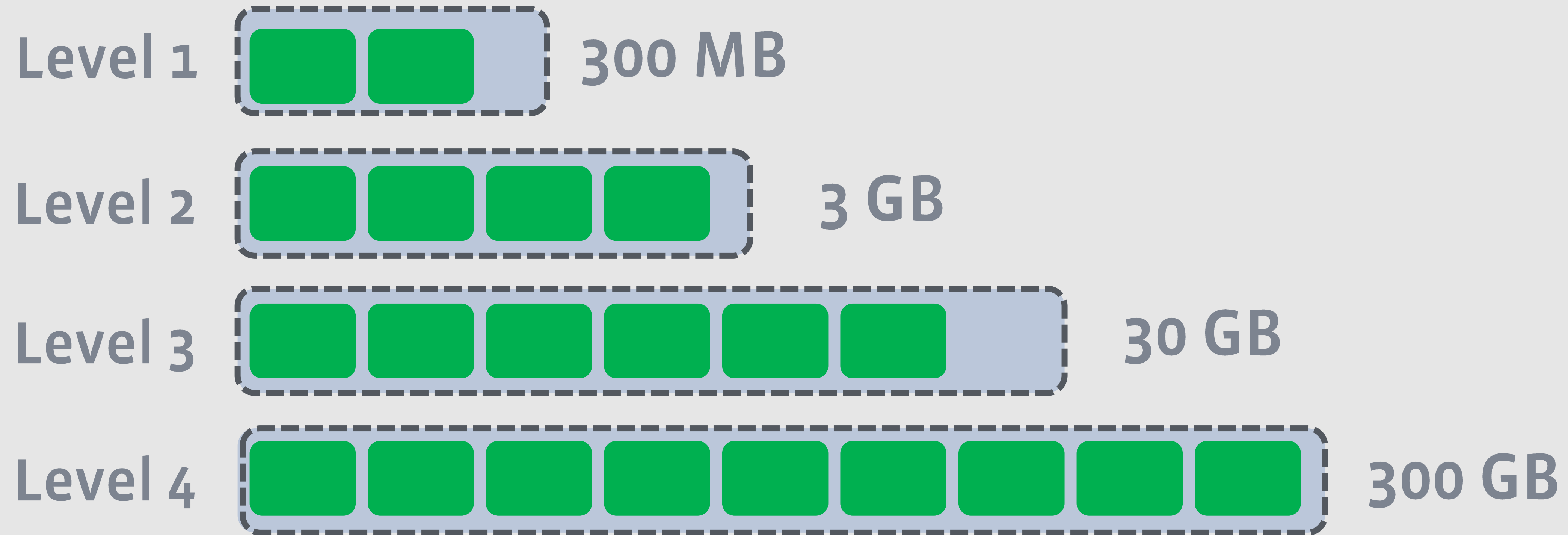
One level in the LSM-tree



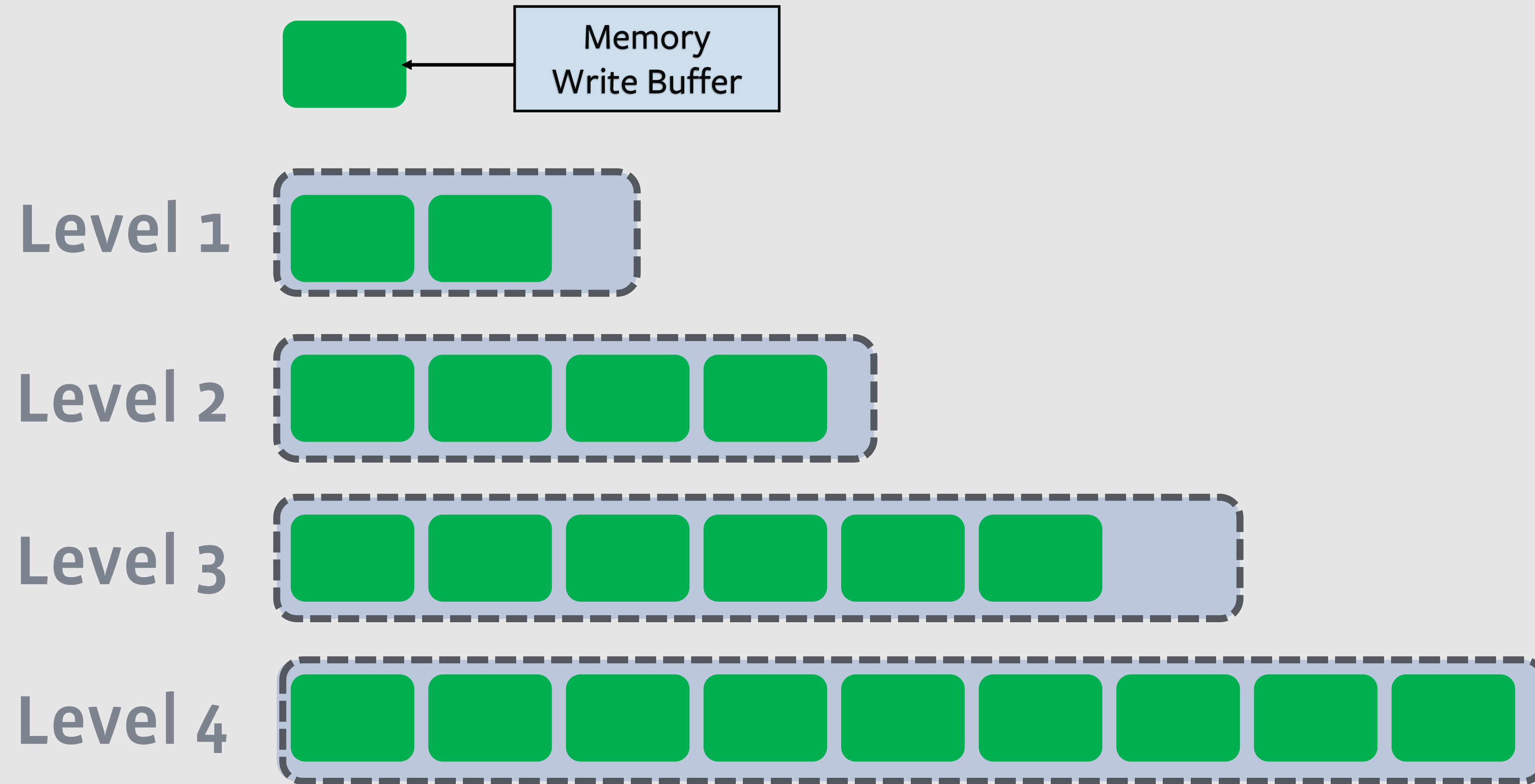
Leveled compaction



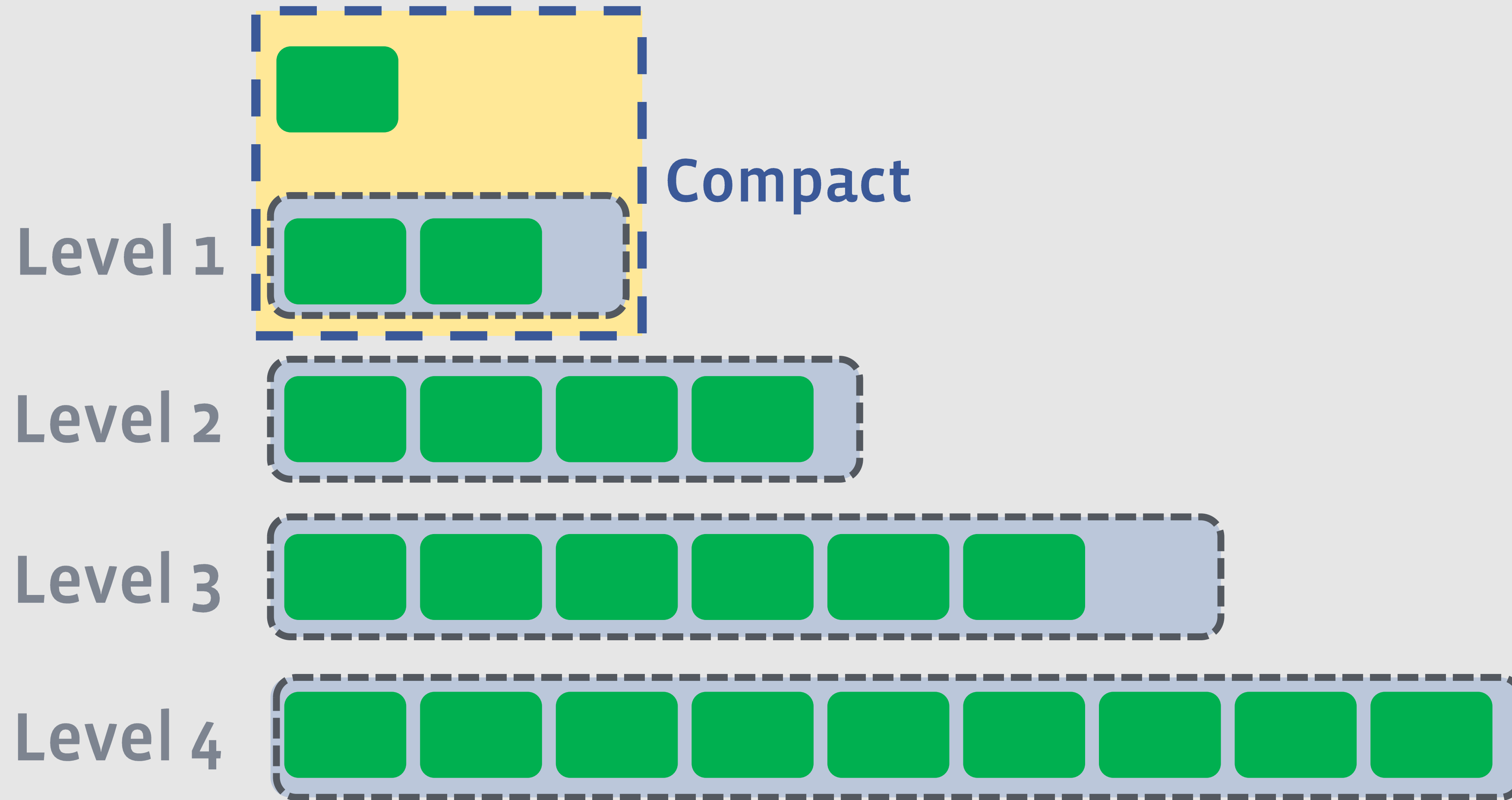
Leveled compaction targets



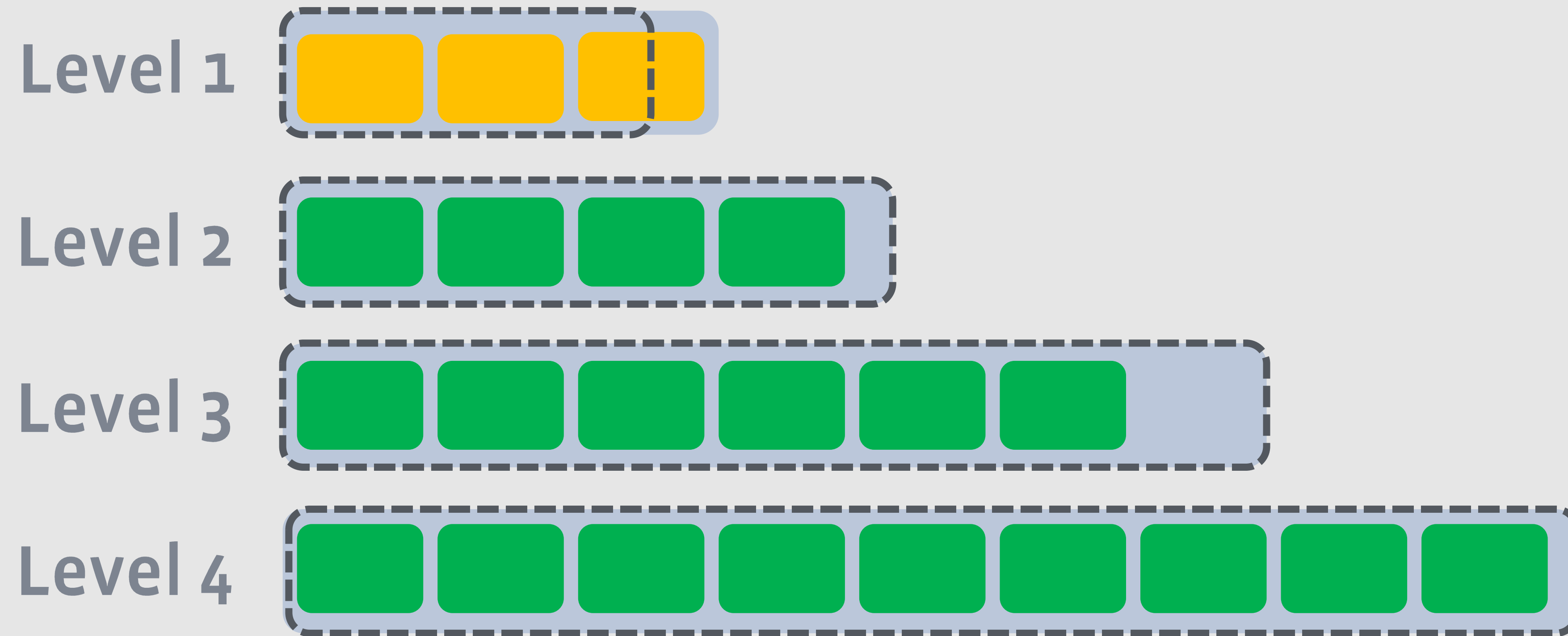
Leveled compaction



Leveled compaction

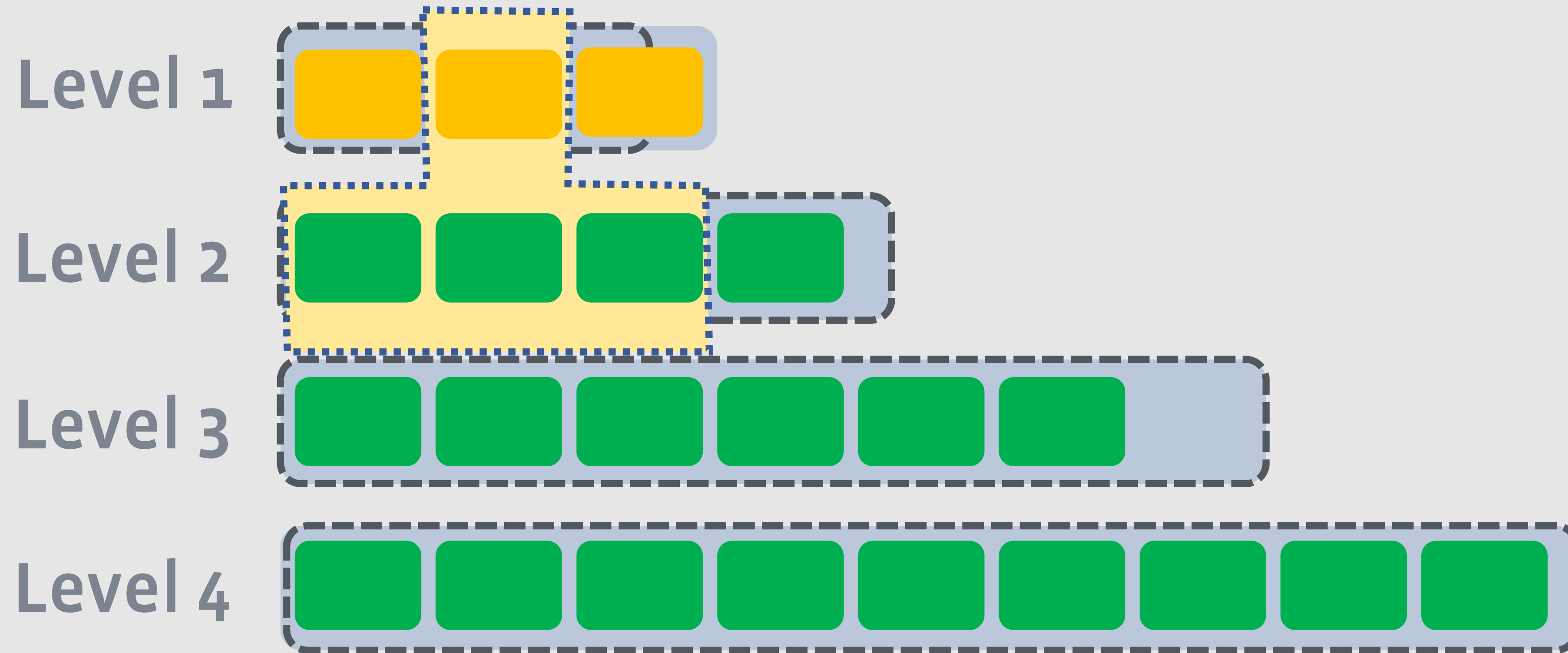


Leveled compaction

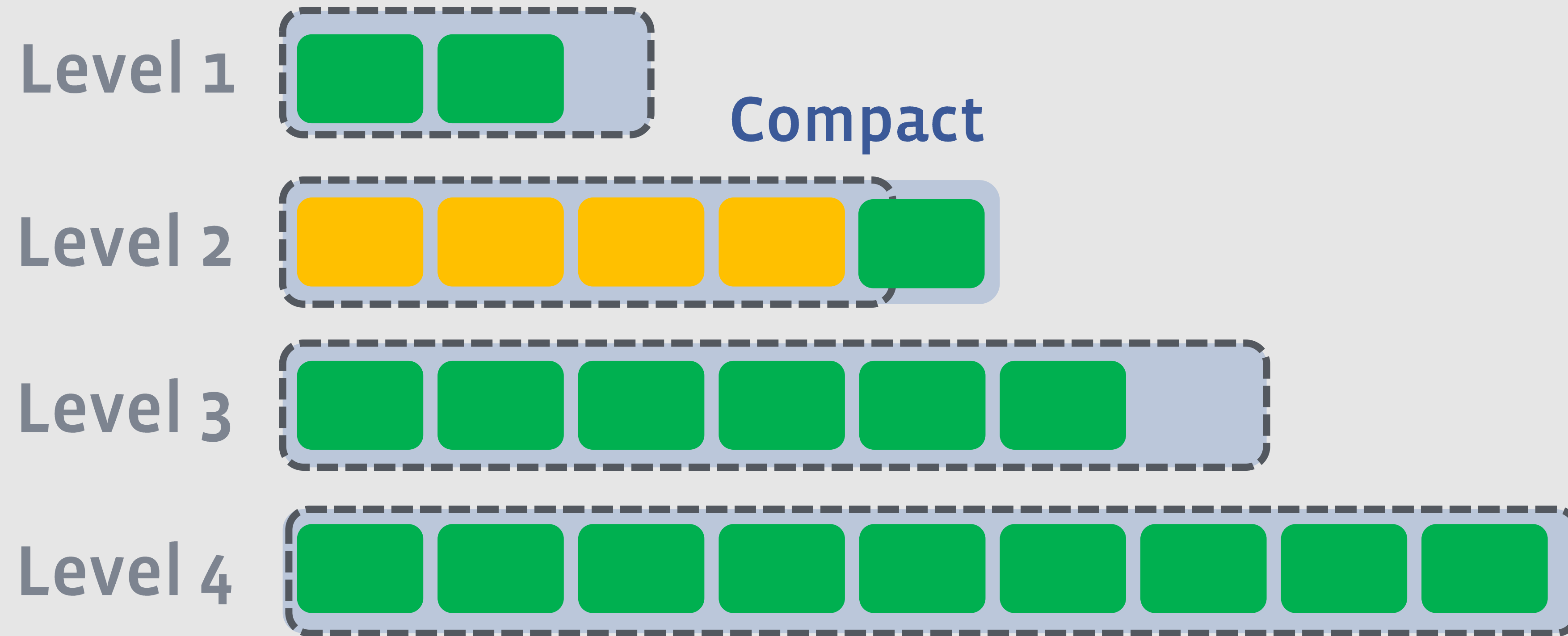


Leveled compaction

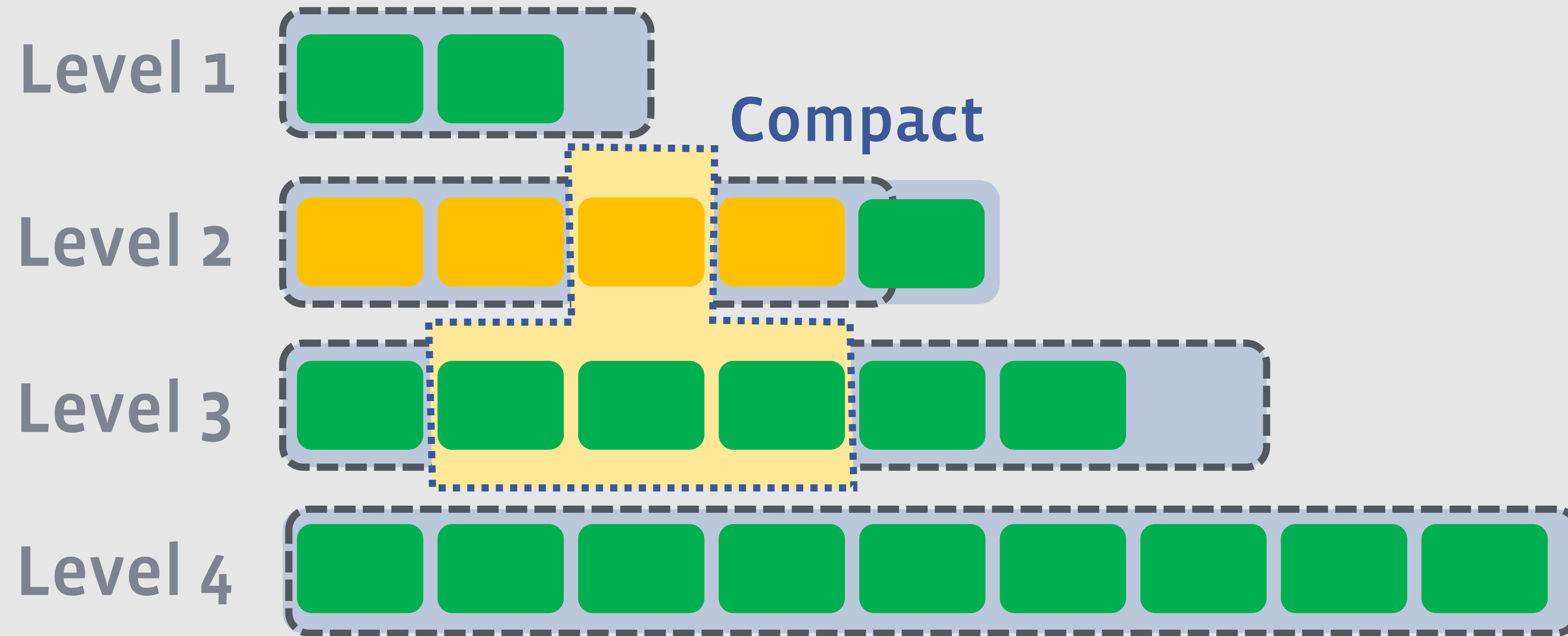
Compact



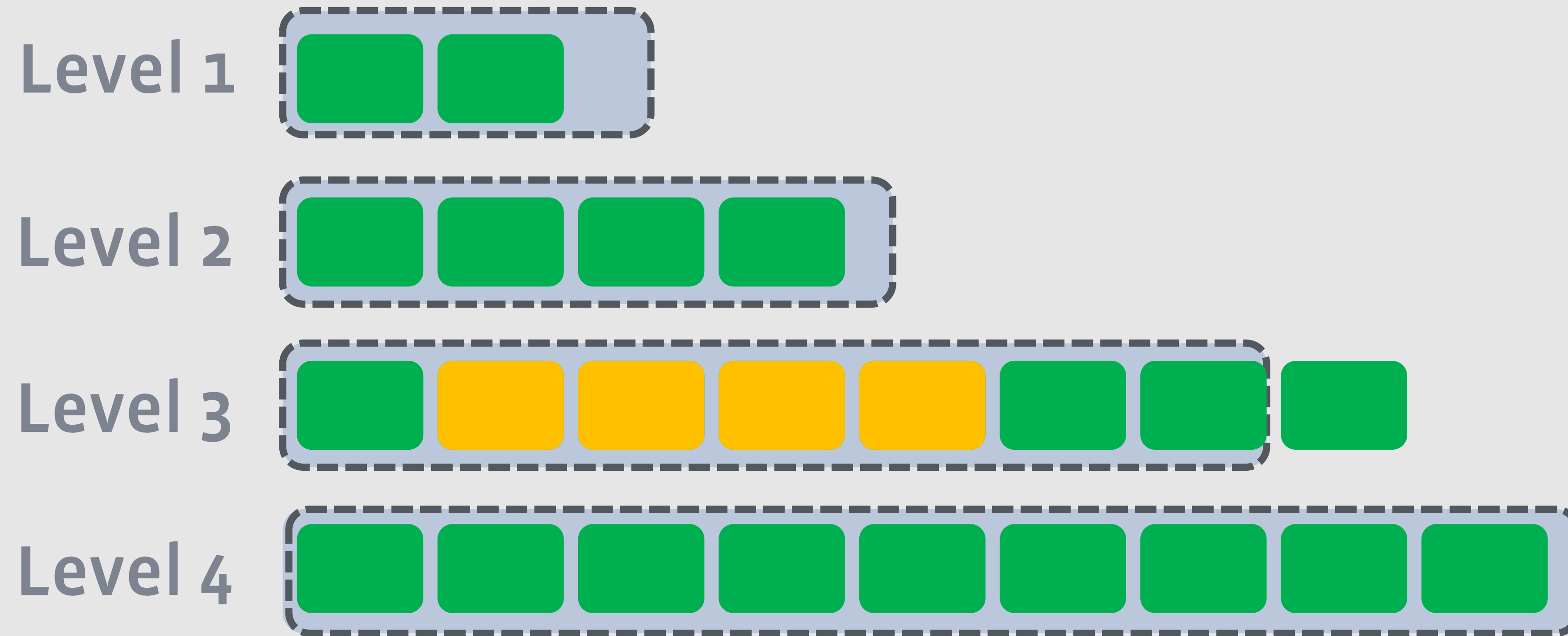
Leveled compaction



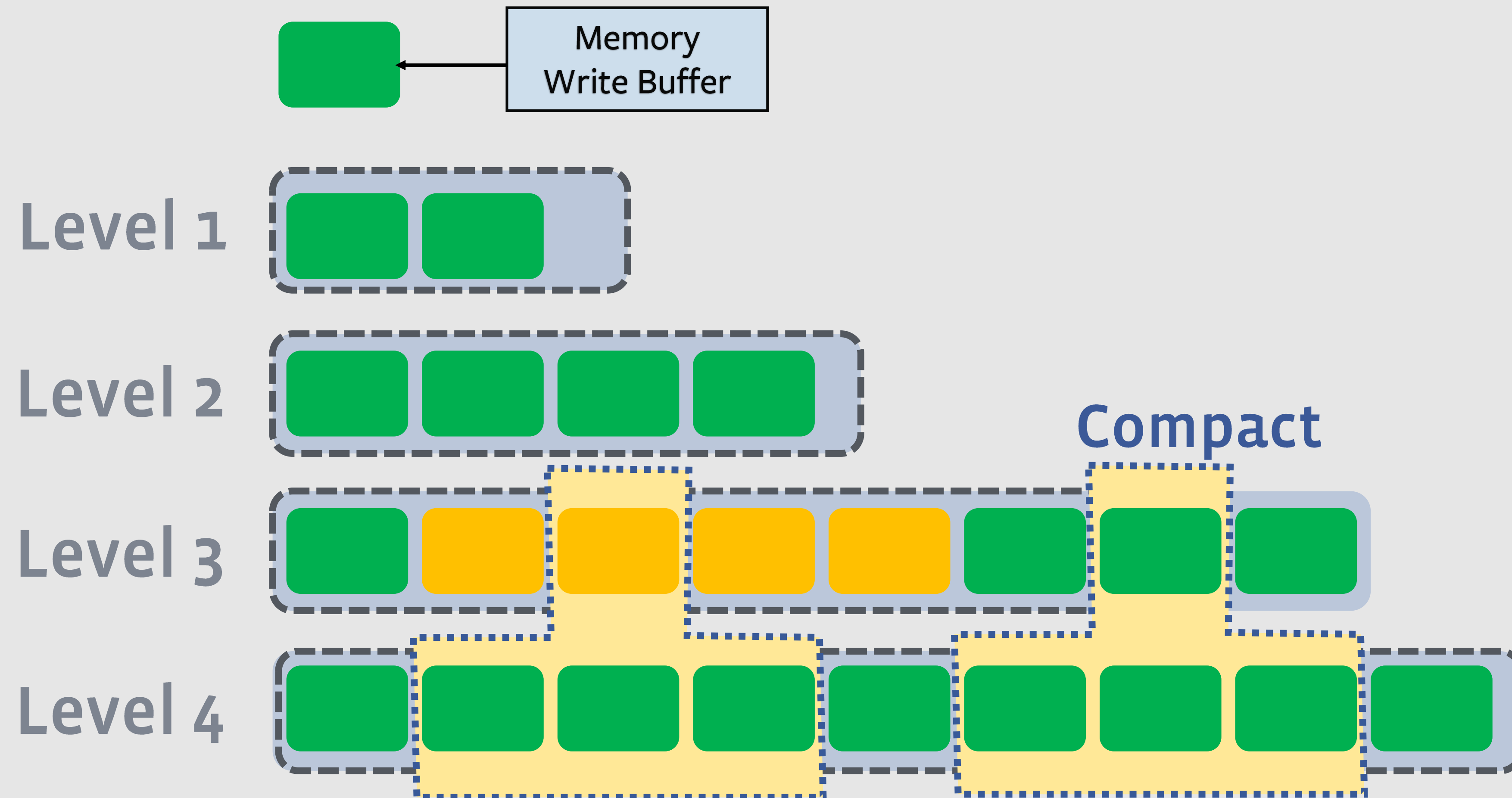
Leveled compaction



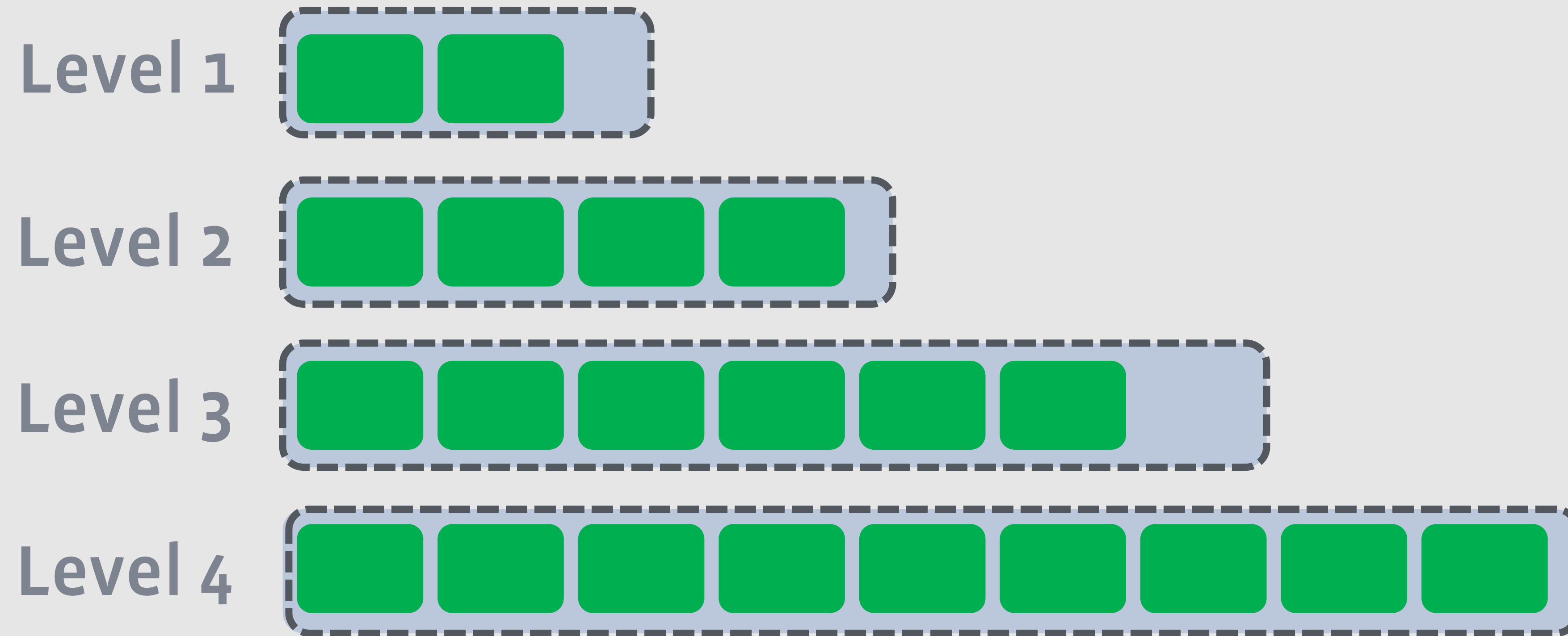
Leveled compaction



Leveled compaction

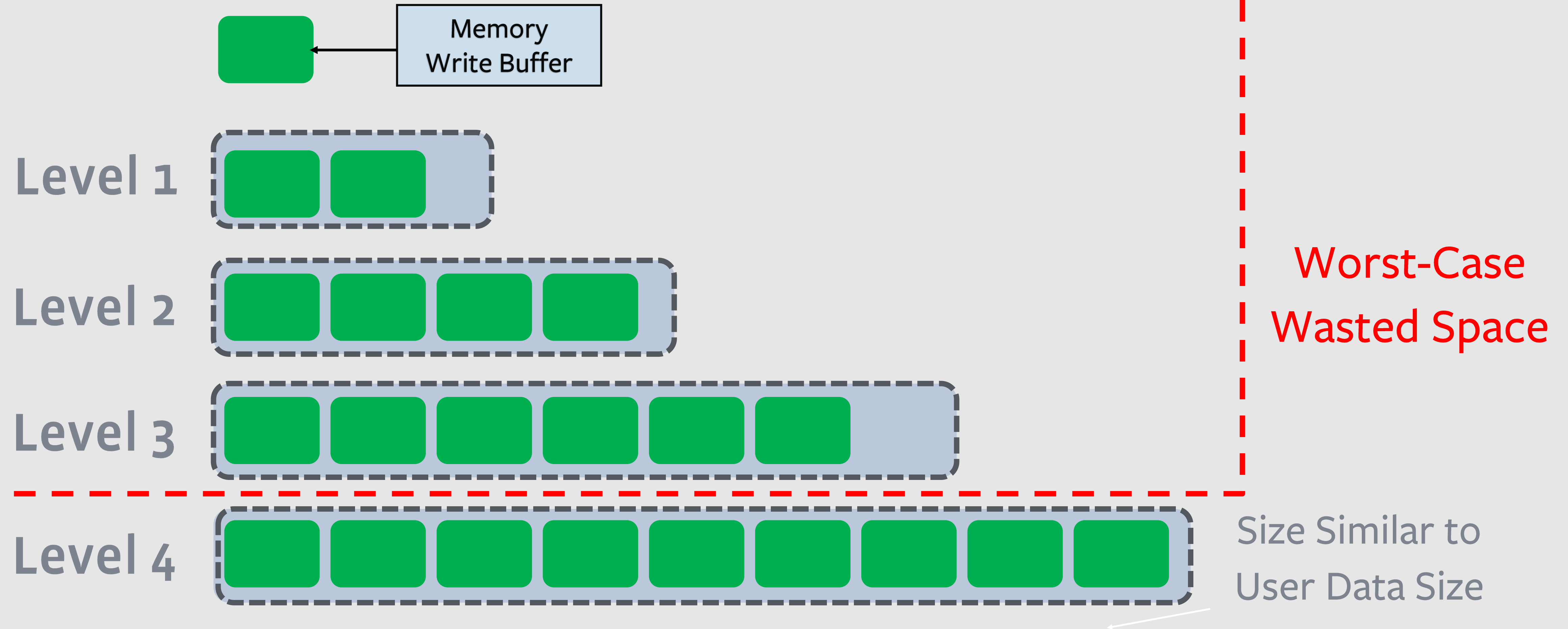


Leveled compaction

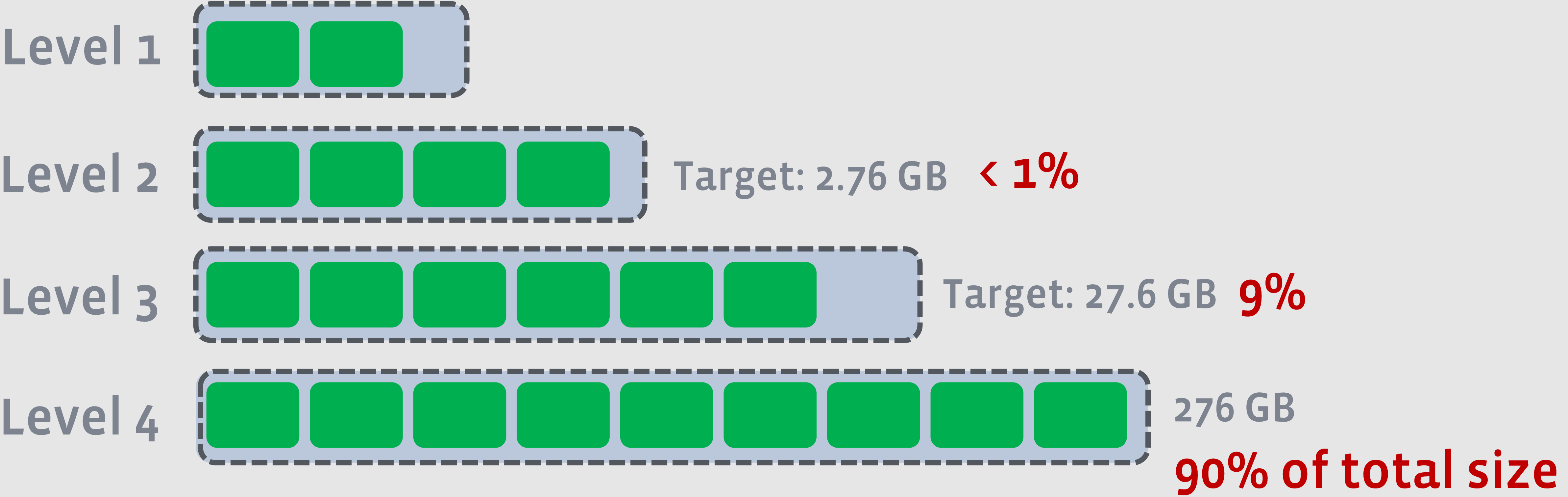


Space Efficiency Techniques

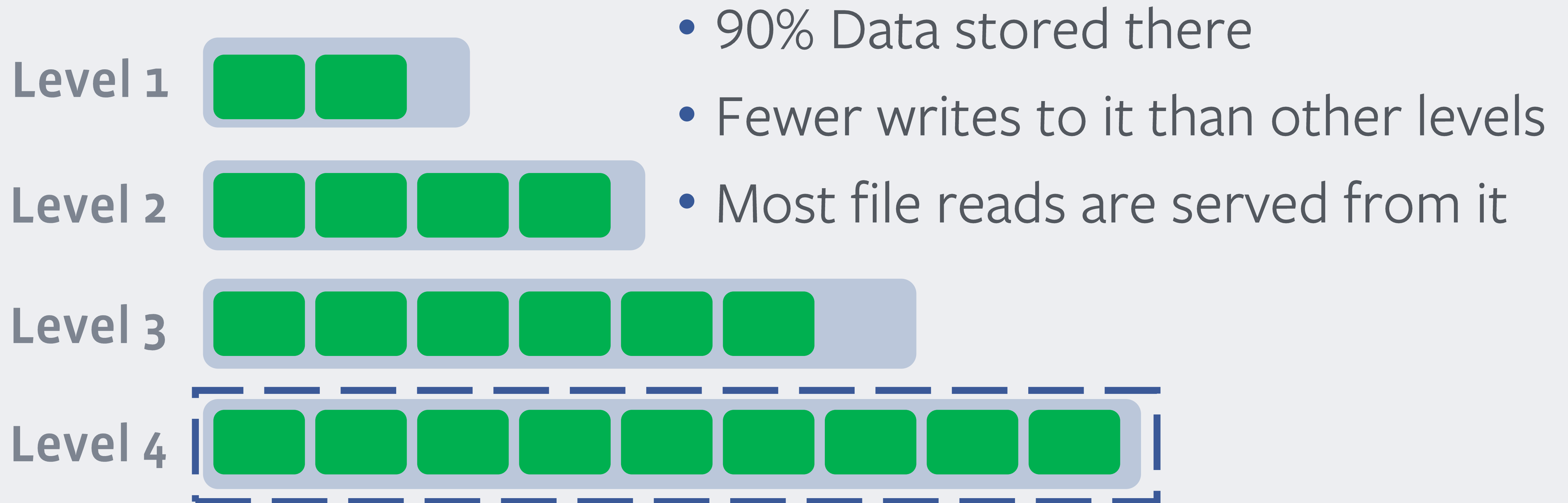
Wasted Space in LSM-tree



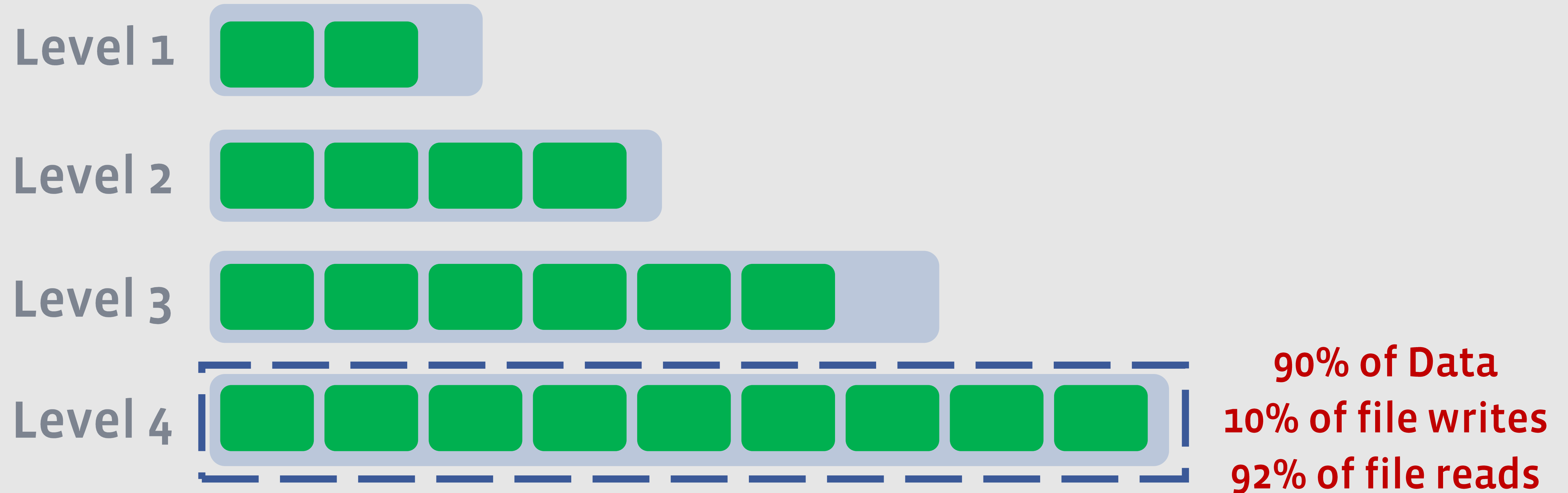
Wasted Space No More Than 10%



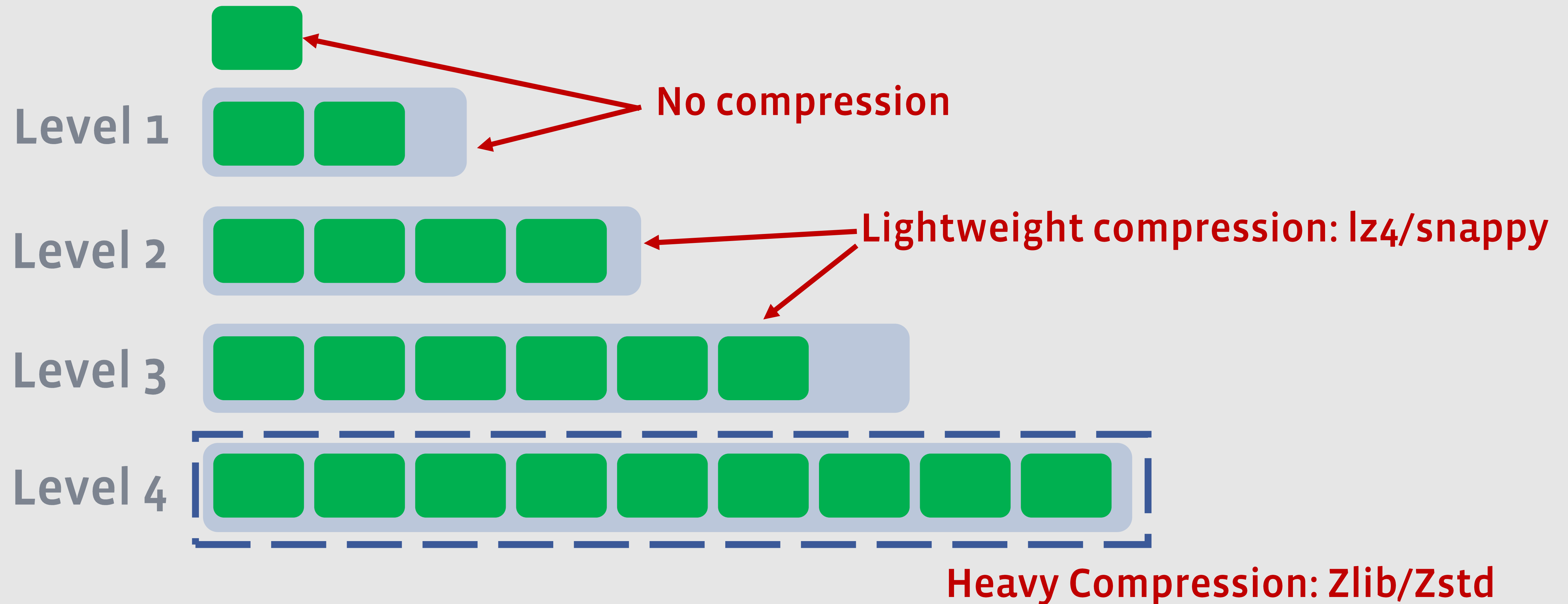
Data In The Last Level



Typical Access Stats for the Last Level



Tiered Compression



Space Efficiency Techniques

- Control wasted space
- Data compression
 - Tiered compression
 - Dictionary compression
- Other Techniques: in the paper
- How to balance CPU and Memory: in the paper

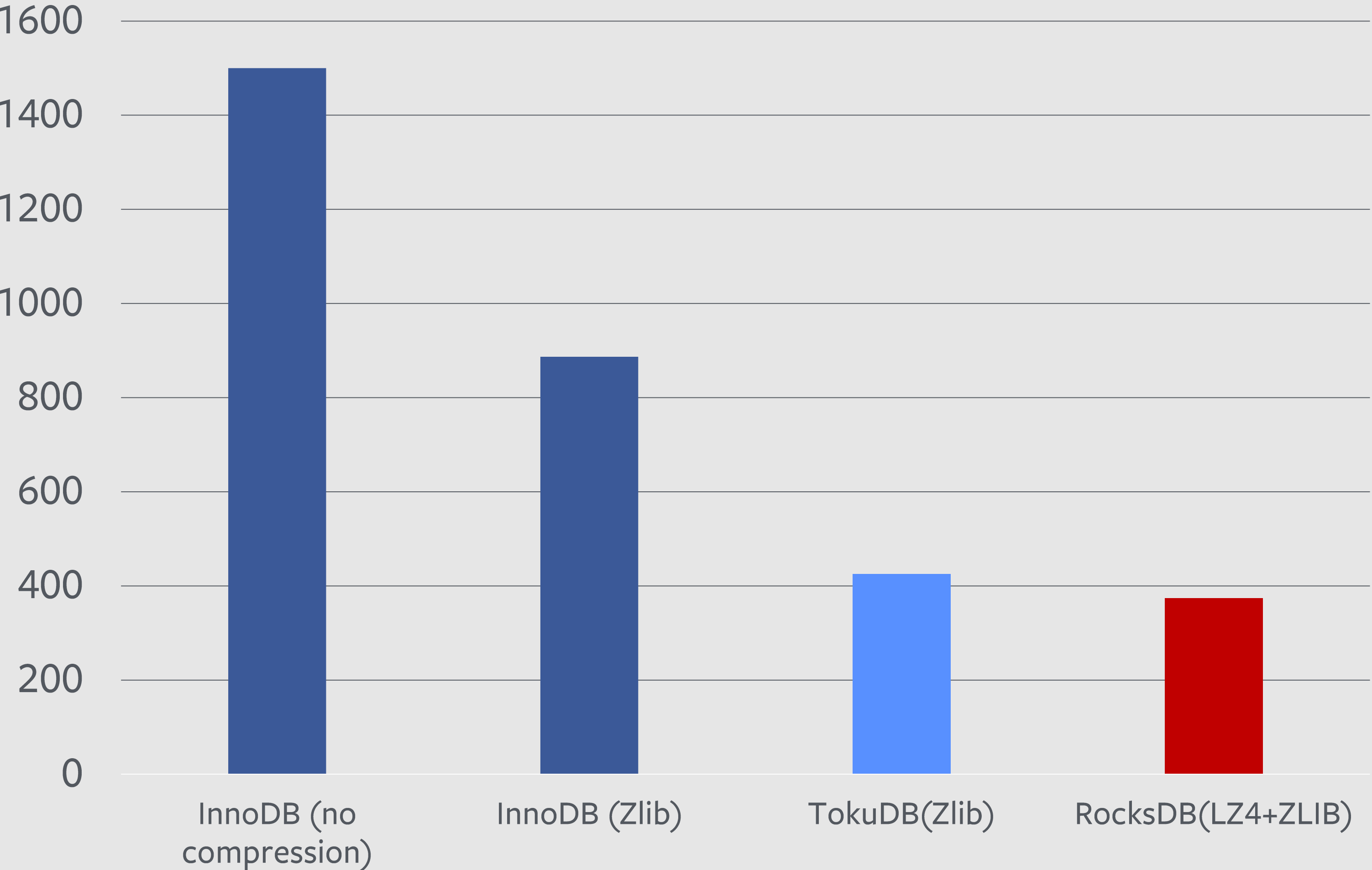
Evaluation

MyRocks = MySQL + RocksDB Engine

OLTP SQL DBMS Using LSM-tree at a very large scale

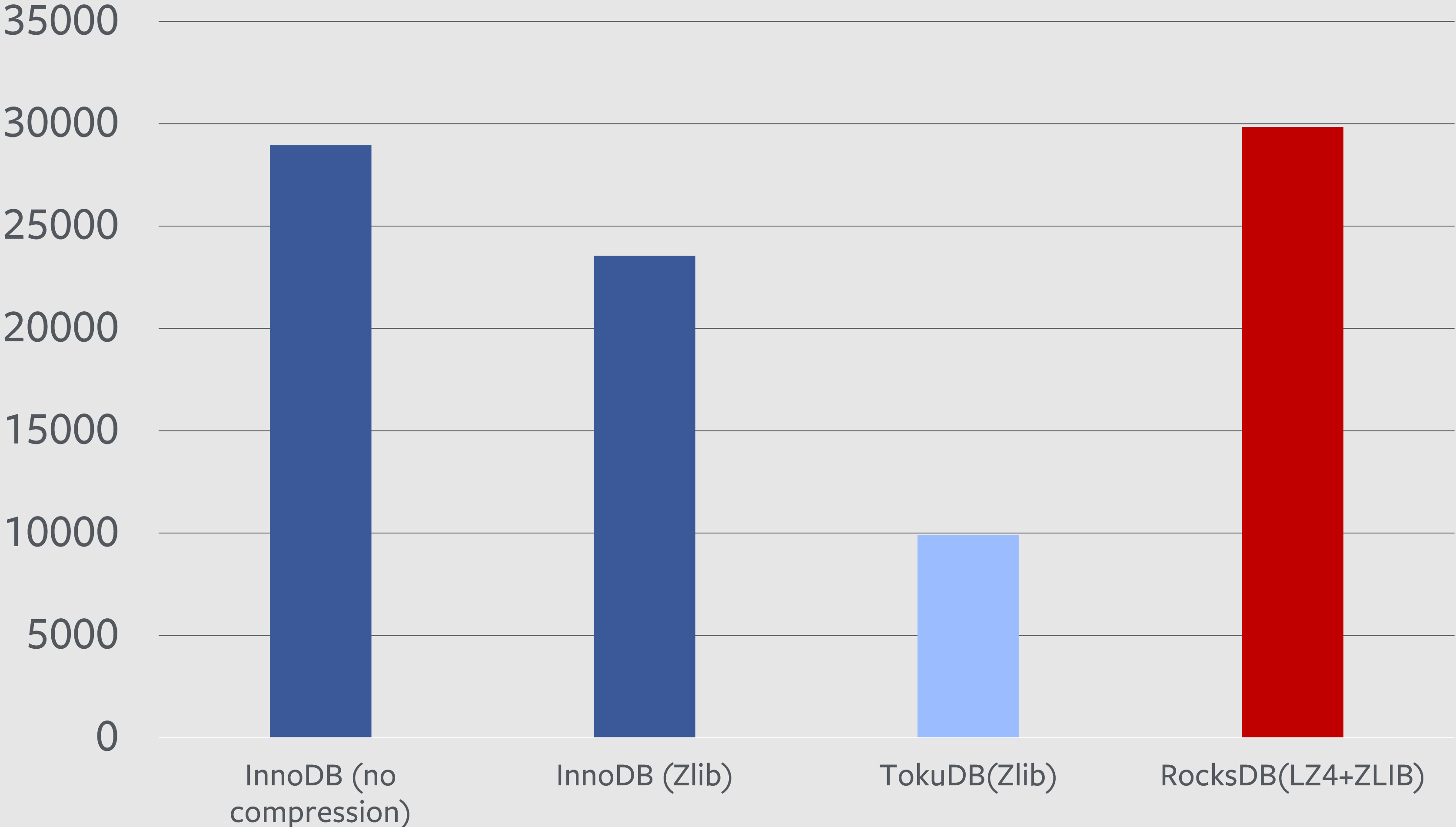
Big Space Saving

Size: GB



Without Sacrificing Performance

Transaction/s




Performance vs. Efficiency

Why efficiency?



Performance



Resource
Efficiency

Why efficiency?



Performance
meet the need



Resource Efficiency
as good as possible

Take-away

- We are willing to trade performance for efficiency as long as required service levels are met
- Space efficiency is the main bottleneck for SSD
- RocksDB is a widely used storage engine that can improve space efficiency
- RocksDB uses LSM-tree and applies several techniques to improve space amplification

facebook