PERCONA

Structure overview: Btree / LSM

Emanuel Calvo

April 2017

Agenda

- Log Structured Merge Tree
- Balanced tree
- Technologies for both structures

LSM concepts

- Keep leveled SST files by merging them in the background.
- As SST are ordered, merging processes are efficient.

LSM components

- Memtable (in-memory tree ordered structure)
- SST (Sorted String Table)
- As data is ordered, sparse indexes are possible.

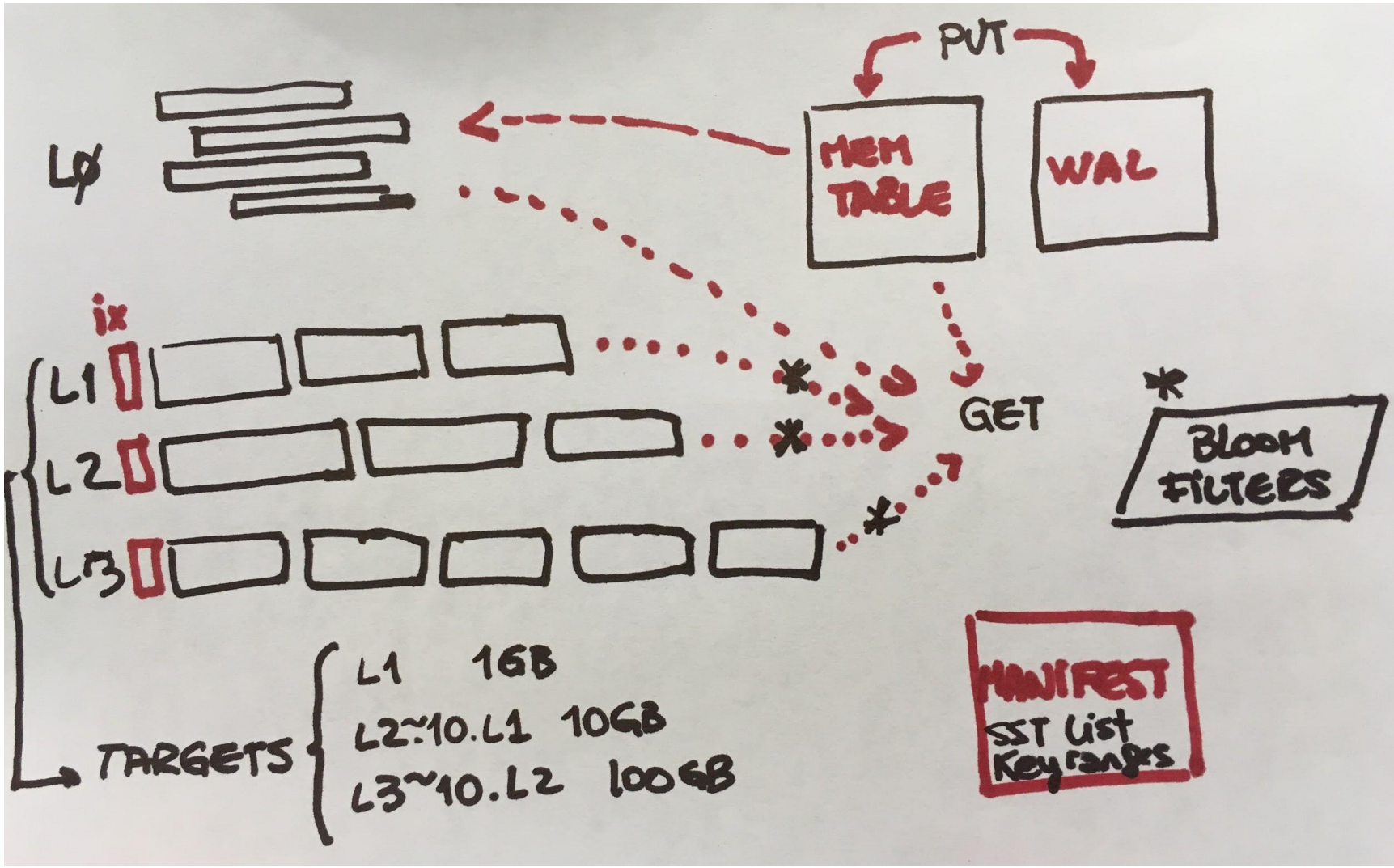
LSM Optimizations

- Bloom filters
  - Space-efficient probabilistic data structure
    - Possibly exists
    - Definitively not exists <- most expensive
    - (False negative cannot exist)

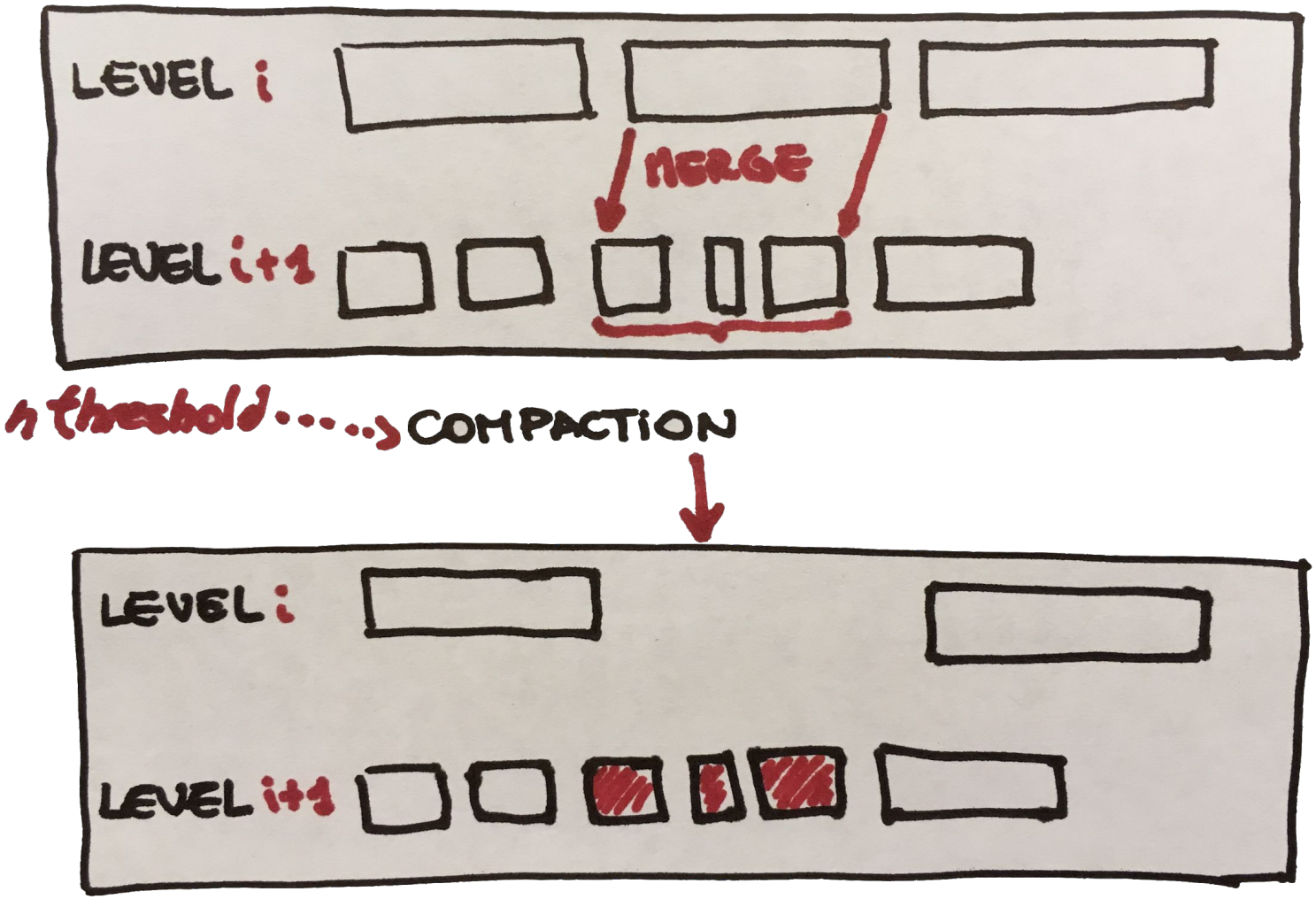
LSM Maintenance

- Memtable requires a WAL
- Compaction process can consume resources
  - Leveled compaction is more incremental
- Merging by selecting the last appearance of the key in the level.

LSM by RocksDB



LSM by RocksDB Merge/Compaction



LSM technologies

- WiredTiger
- MyRocks (inspired on LevelDB)
- LevelDB (on Cassandra and Hbase)

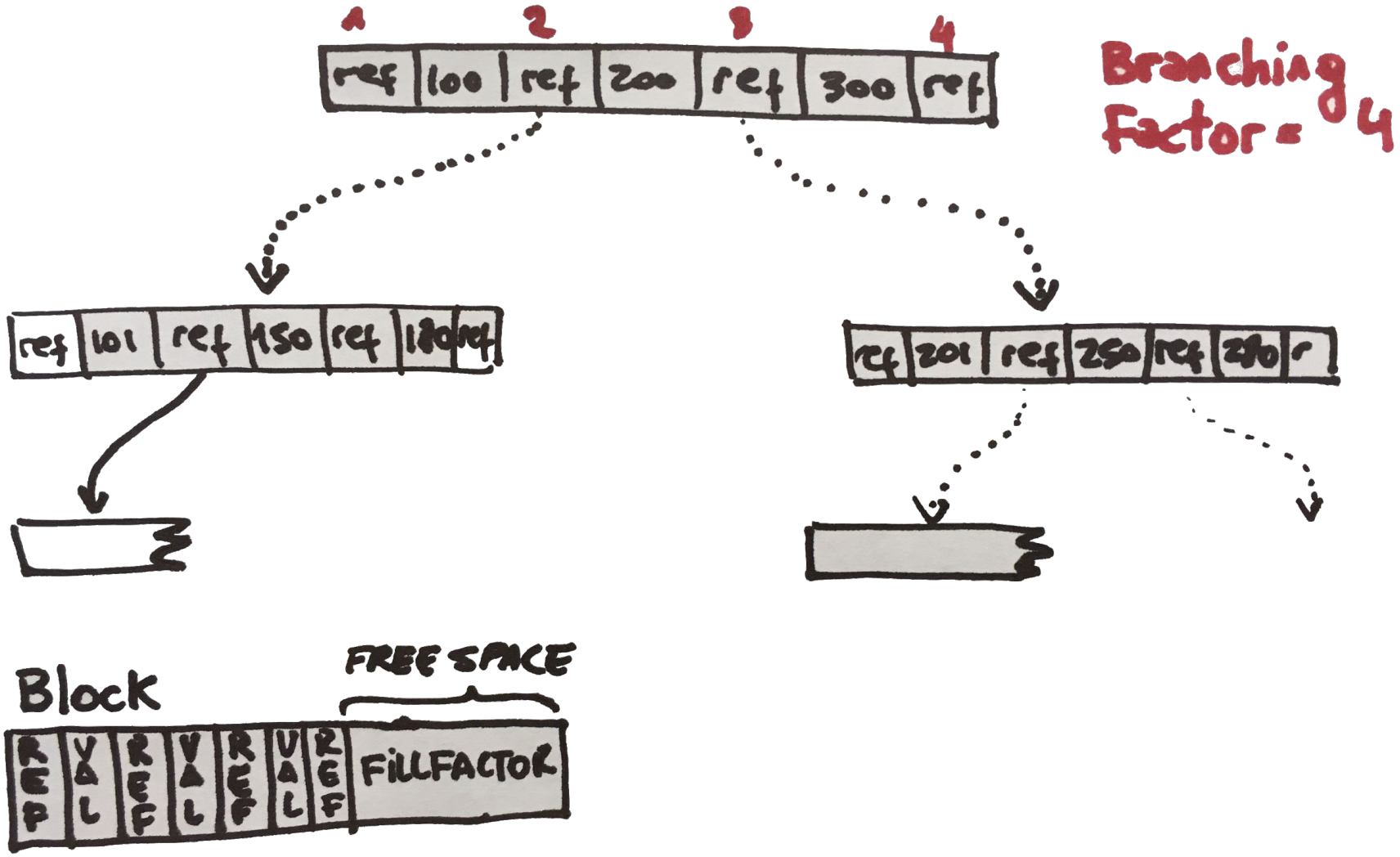
btree concept

- Writing in-place as closest to the underlying disk hardware, by fixed block size changes.

BTREE

- Branching factor
- Depth (O(log n))
- Write amplification is caused when splitting leaves + updating parent leaf.
  - Special mention on InnoDB as it uses clustered indexes. Postgres PK points to heap, forced to be updated.
- 4 level of 4kb block index with a few hundred branching factor can point up to ~256TB.
- Siblings levels in the same level point to the prev and next ones.

BTREE


1
2
3
4
ref
100
ref
200
ref
300
ref
Branching Factor = 4
ref
101
ref
150
ref
180
ref
ref
201
ref
250
ref
280
r
Block
FREE SPACE
REF
VAL
REF
VAL
REF
VAL
REF
FILLFACTOR

Columnar Store

- Analytics purposes
- Uses bitmaps for value pointing between column families.
- Less CPU cycles for row/column processing.

What's the deal

- LSM are efficient with low memory resources and when keys exists and are being modified recently (recently modified pattern)
- LSM are more efficient as the order the key-values before write to disk. Also merging is efficient.
- Also, LSM is good for writes.
- BTREE are good for reads, specially for range queries as it is based on blocks with contiguous data.

Technologies

- MongoDB(wiredtiger), LevelDB and MyRocks are the popular choices for LSM.
- PostgreSQL, MySQL and many other RDBMS are B+tree.
- Vertica , MariaDB Columnar Store, Cstore (Postgres, in-memory) and Redshift are columnar storage.

Most recent benchmarks between InnoDB and MyRocks

Linkbench Result

- 1.5B IDs, 32 query threads, 48 hour run, flash storage
- Space: 1172GB in InnoDB, 574GB in MyRocks (49%)
- QPS: 22227/s in InnoDB, 33094 in MyRocks
- Write KB/s: 152,422 in InnoDB, 66,932 in MyRocks (44%)

Sources

- The log-structured Merge-Tree
- LSMT in WiredTiger and WiredTiger
- BTREE vs LSM
- Columnar Storage

Don’t miss the next Percona Live 17 at Santa Clara!

www.percona.com/live