A (More) Write-Optimized LSM Tree

Ali Monfre
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

k = 3

\( d = 3 \)
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

\[ k = 3 \]
\[ d = 3 \]
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

k = 3
d = 3

n
n
n

3n
3n
3n
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

k = 3

n

n

n

3n

3n

3n
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

\( k = 3 \)

\( d = 3 \)
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

n

n

n

3n

3n

3n

k = 3

d = 3
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

k = 3

k = 3

k = 3

\[ d = 3 \]
Refresher on LSM trees

C0 – memory

C1 - disk

C2 - disk

\[ k = 3 \]

\[ d = 3 \]
Tradeoffs

• Writes are fast – memory buffer and merge structure
• Reads are slower – potentially have to seek every level of the tree each time

Can we make writes better without harming reads too much?
Updates

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

[...[(k1,v1)]]

[...[(key, val)]]

[...[(k1,v2)]]
Updates

C0 - memory

C1 - disk

C2 - disk

C3 (final level) - disk

Marked as “old”
Updates

C0 – memory

C1 – disk

C2 – disk

Marked as “old”

C3 (final level) – disk
Updates

C0 – memory

\[ (...(key, val)...) \]

C1 – disk

\[ (...(k1,v1)...), (...(k1,v2)...) \]

C2 – disk

\[ (...(k1,v1),(k1,v2)...), (...(k1,v3)...) \]

C3 (final level) – disk

\[ (...(k1,v3)...), Merged out at final level \]
Deletes

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

Deleted values are carried through until the final level
Reads

C0 – memory

C1 – disk

C2 – disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Reads

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Reads

- **C0** (memory)
- **C1** (disk)
- **C2** (disk)
- **C3** (final level, disk)

- Most recently updated in this level
- Most recently updated in this level
Reads

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Reads

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Reads

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Reads

C0 – memory

C1 – disk

C2 – disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Reads

C0 – memory

C1 - disk

C2 - disk

C3 (final level) – disk

Most recently updated in this level

Most recently updated in this level
Parallelization