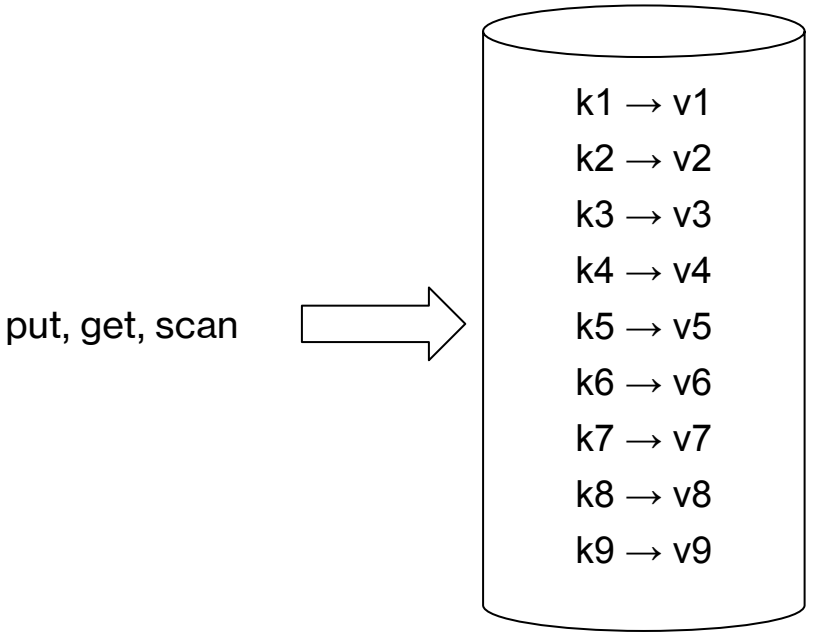
EvenDB: Optimizing Key-Value Storage for Spatial Locality

Eran Gilad, Edward Bortnikov, Anastasia Braginsky, Yonatan Gottesman, Eshcar Hillel (Yahoo Research), Idit Keidar (Technion), Nurit Moscovici (Outbrain), Rana Shahout (Technion)

<EURO/SYS'20>

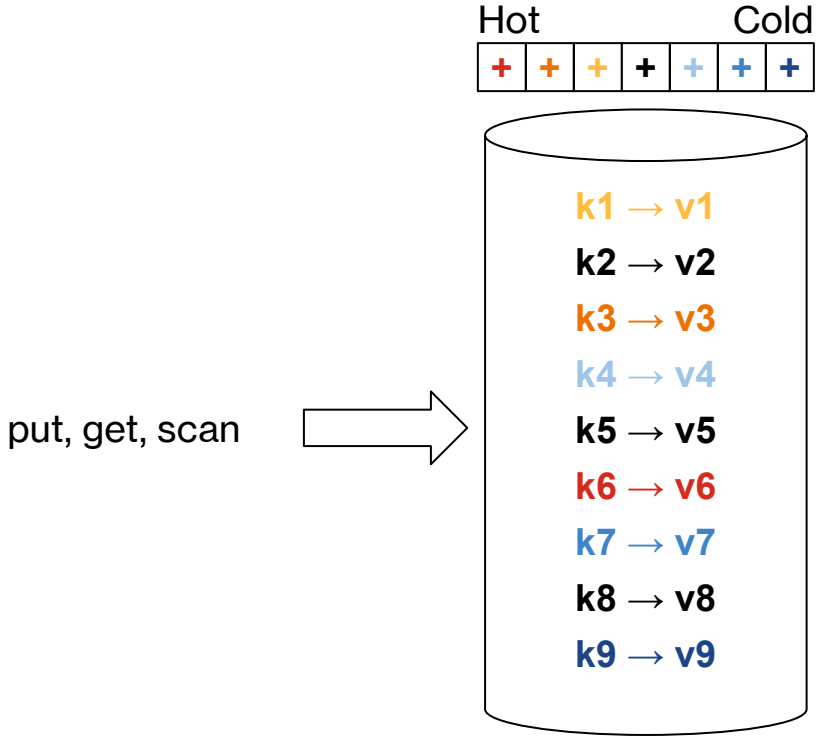
Key-value stores

- **key -> value mapping**



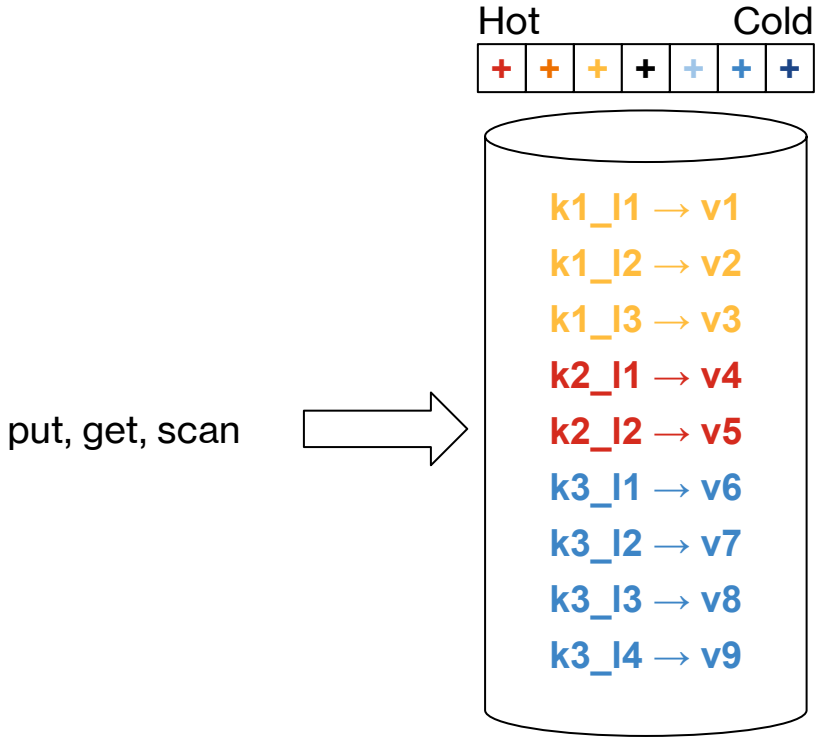
Key-value stores

- key -> value mapping
- **skewed workload: some *keys* are hotter**



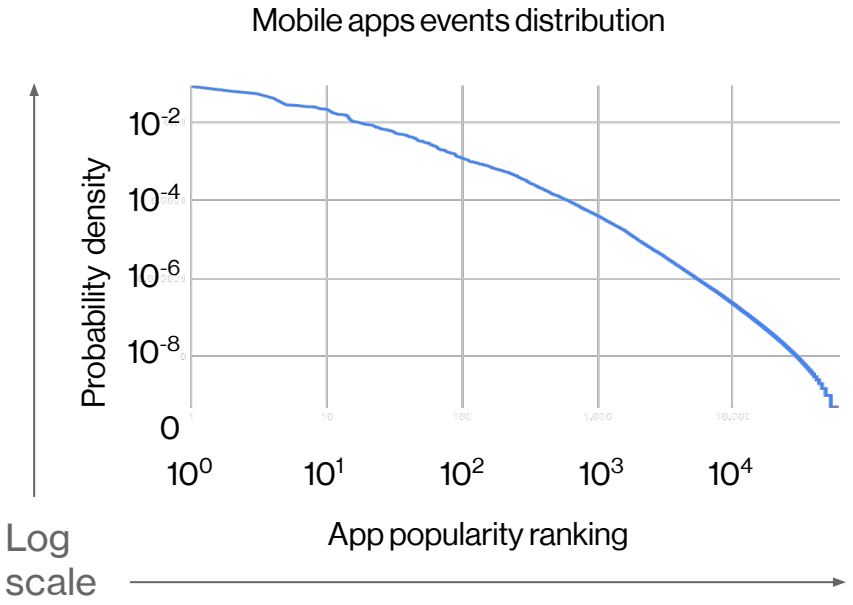
Key-value stores

- key -> value mapping
- skewed workload: some *keys* are hotter
- **spatial locality: some *ranges* are hotter**
  - e.g., complex keys

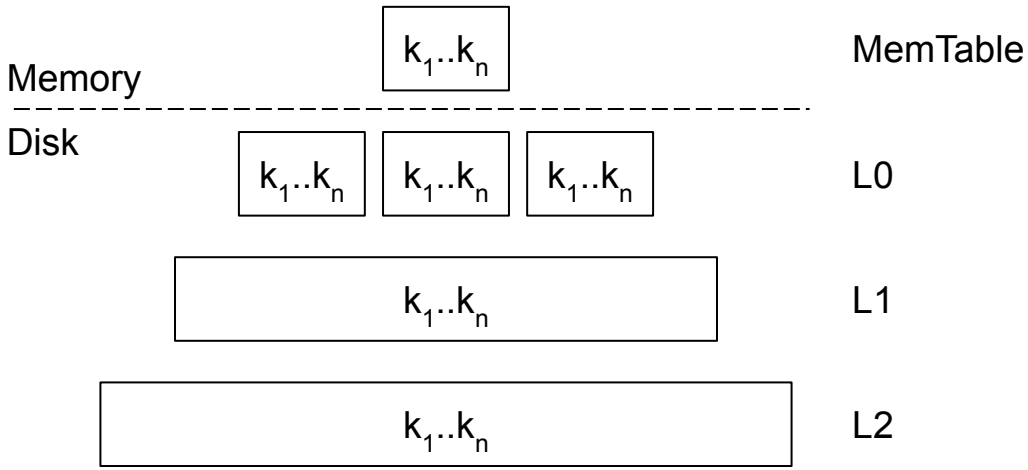


Key-value stores

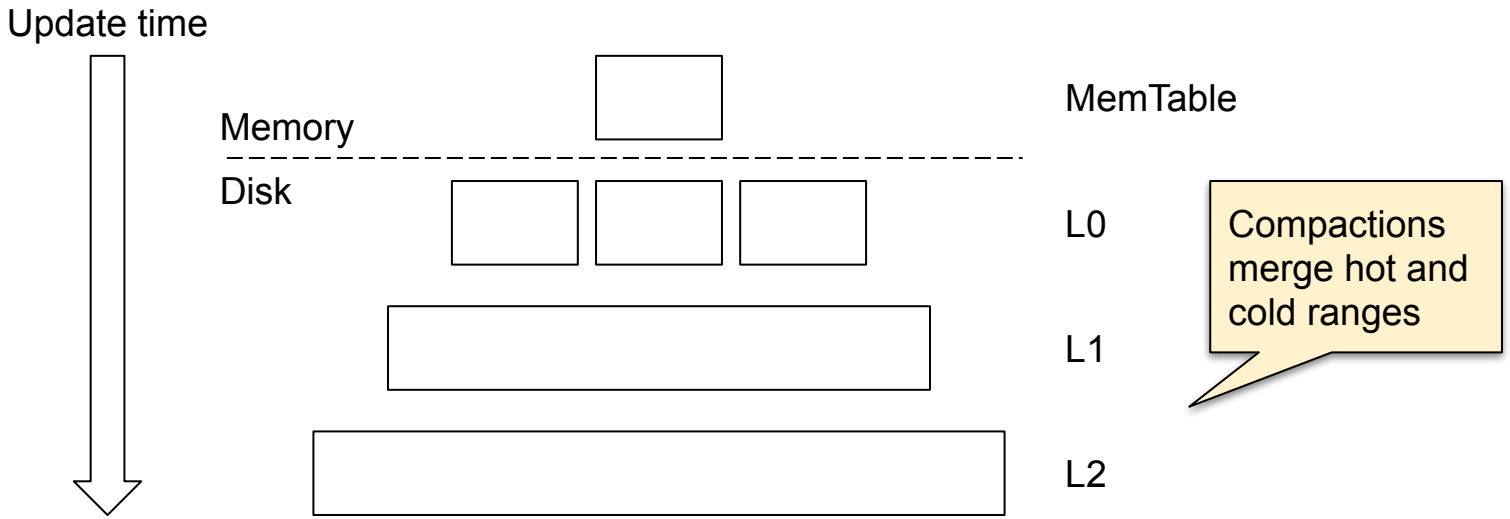
- key -> value mapping
- skewed workload: some *keys* are hotter
- spatial locality: some *ranges* are hotter
  - e.g., complex keys
- **Sample production trace:**
  - appname_timestamp
  - 1% of apps ⇒ 1% key prefixes ⇒ 94% of events



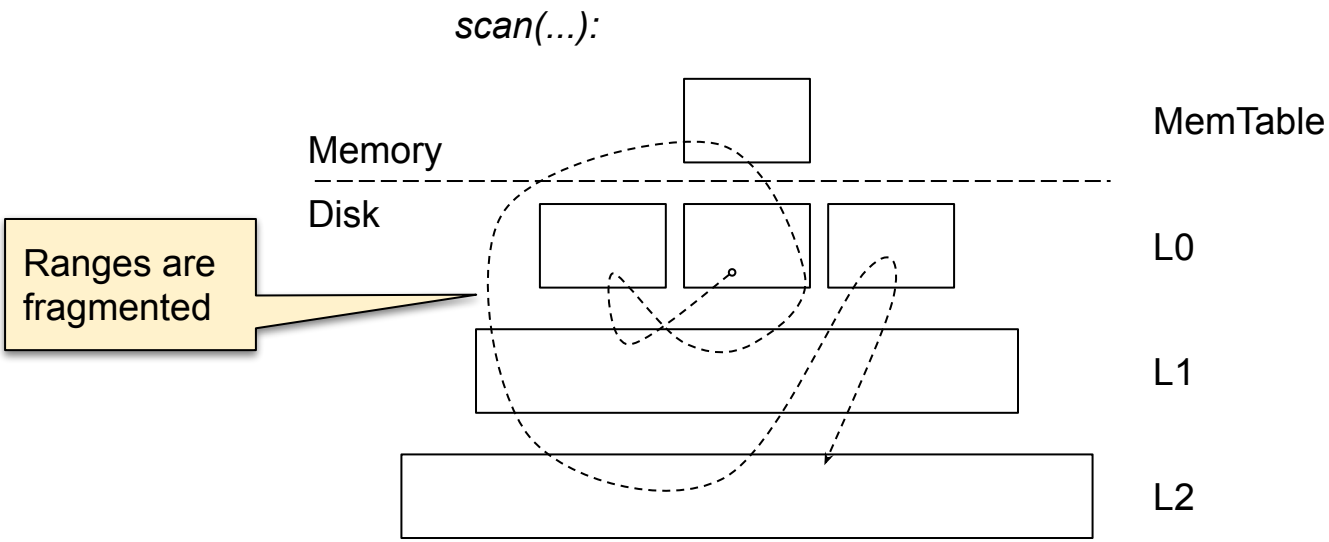
LSM-trees



LSM-trees are designed for temporal locality



LSM-trees are less suited for spatial locality



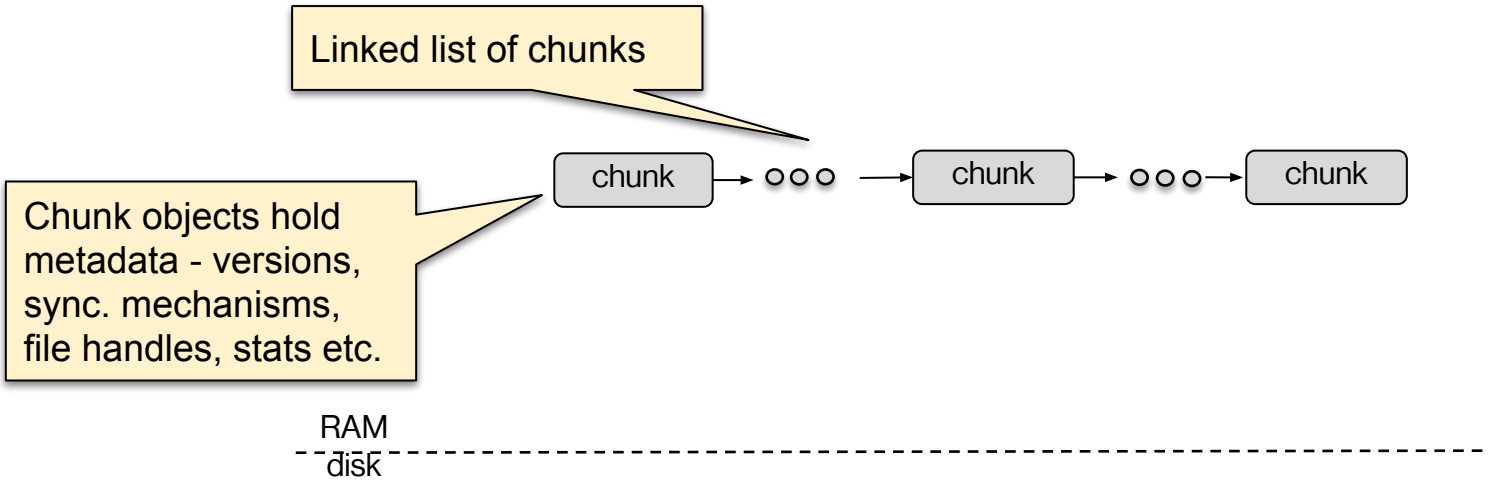
EvenDB

- Ordered key-value store
- Optimized for spatial locality
- Low write amplification
- Persistent, fast recovery
- Atomic operations, including scan

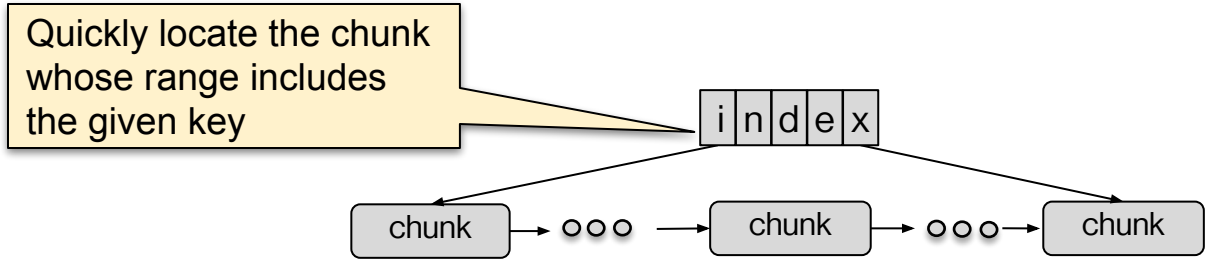
Chunk-based organization

- **Dynamically partitioned key space into *chunks***
  - Much smaller than shards
  - Much larger than blocks
- **Chunks are the basic unit for**
  - Disk I/O
  - Compaction
  - Memory caching
  - Concurrency control

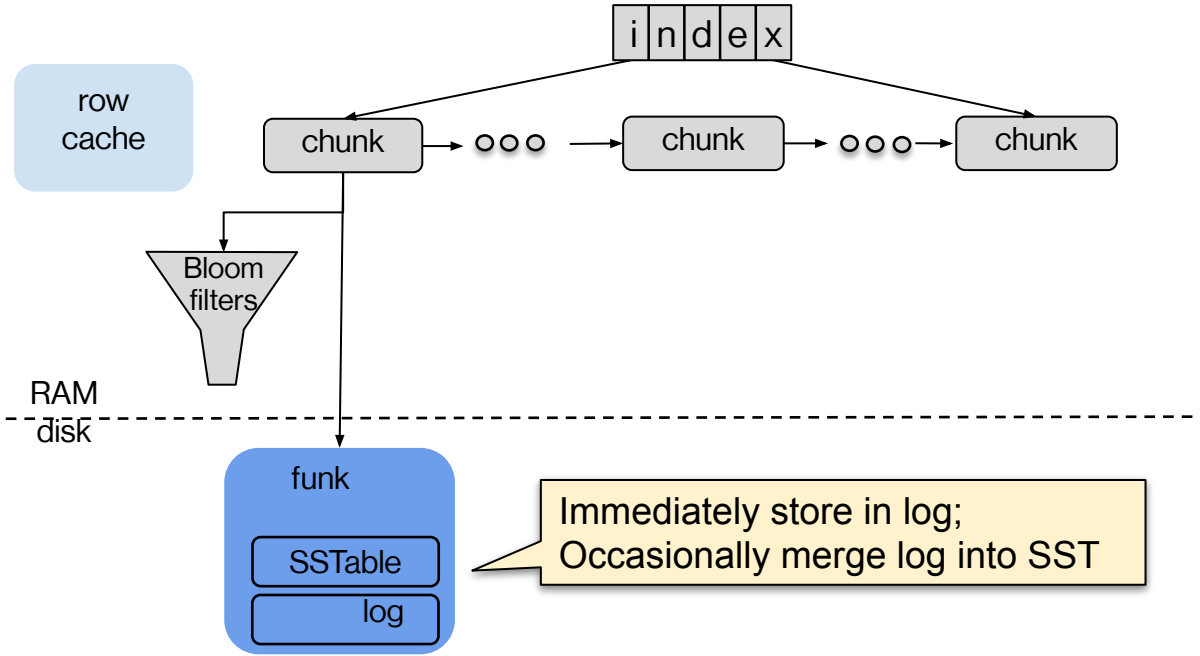
Chunks metadata



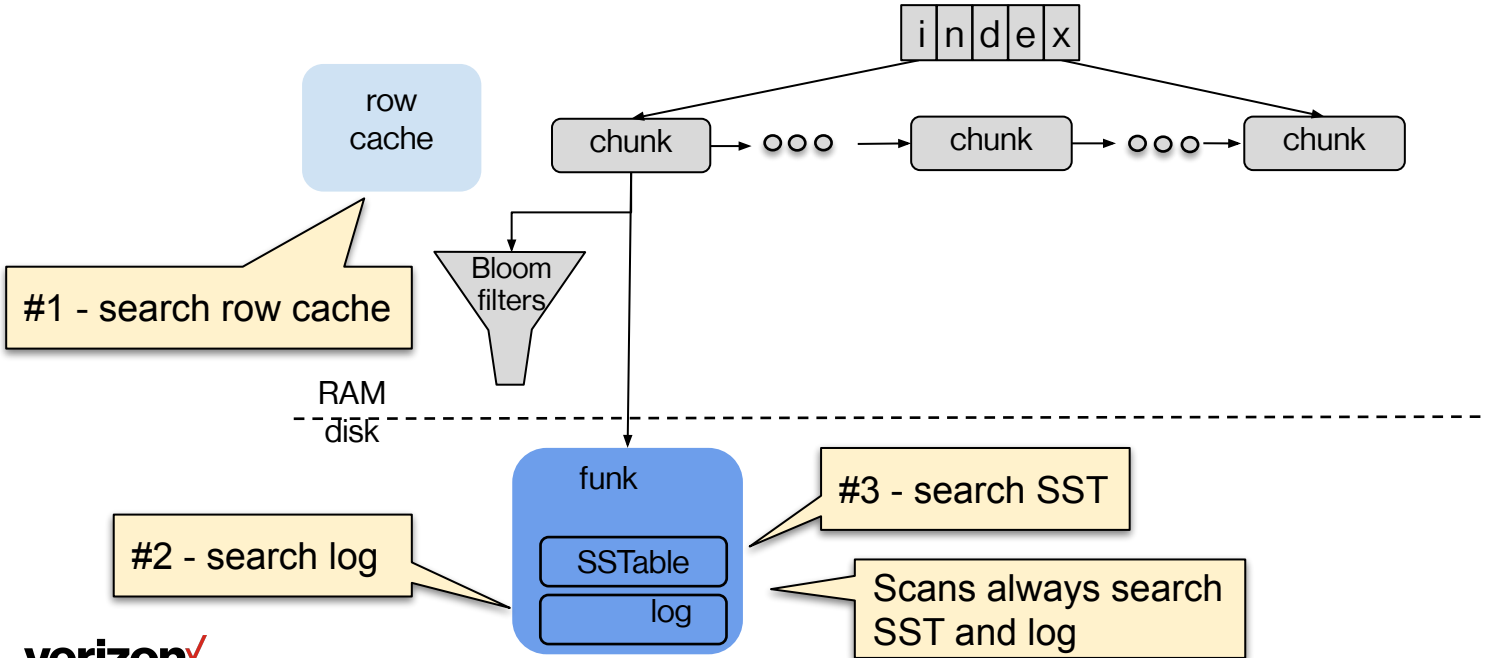
Chunks index



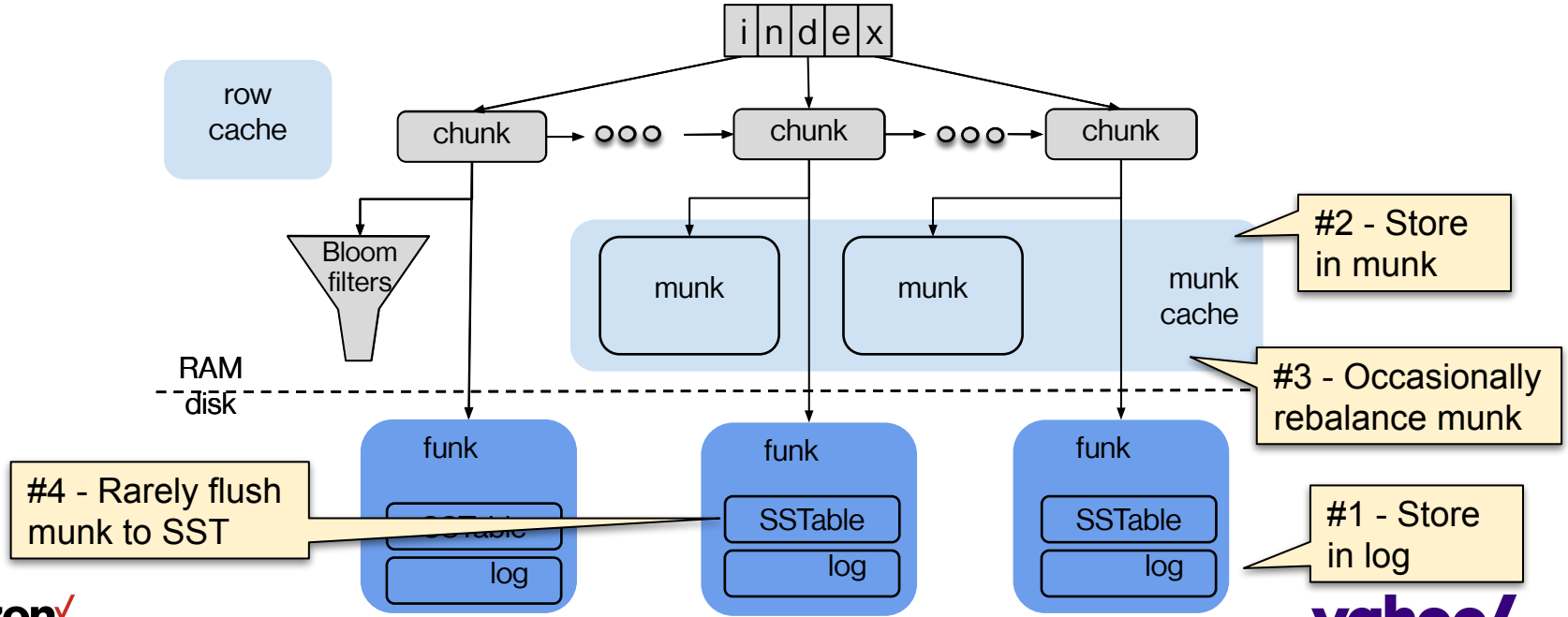
Disk storage - updates



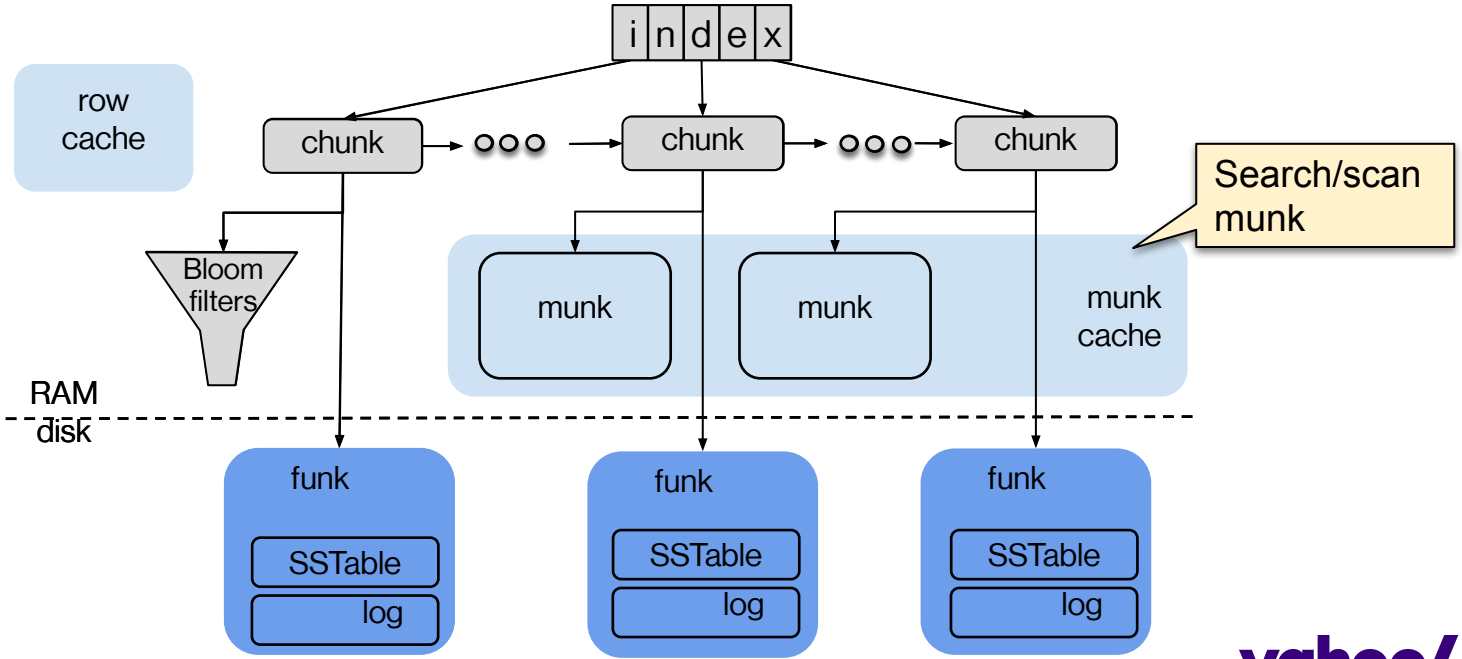
Disk storage - lookups



Memory cache - updates



Memory cache - lookups

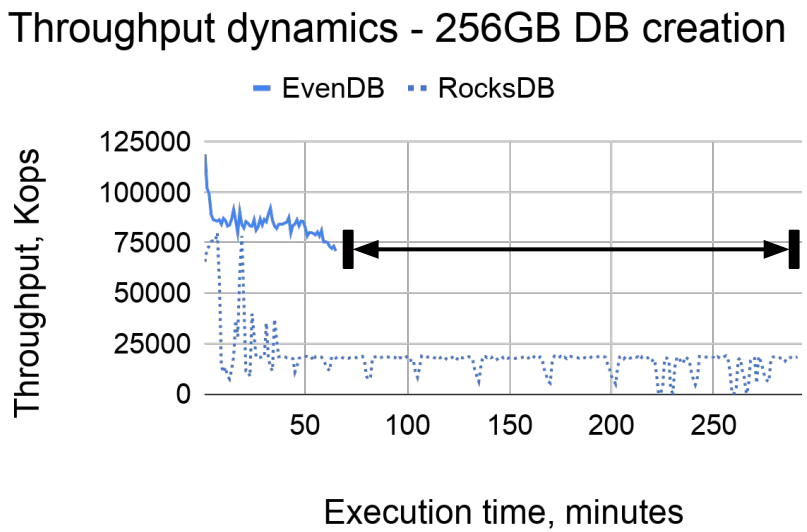


Evaluation

- **3 benchmark suites**
  - Traces from internal production system, 256GB DB - some presented next
  - Standard and extended YCSB benchmarks - results in paper

- **State-of-the-art LSM: RocksDB**

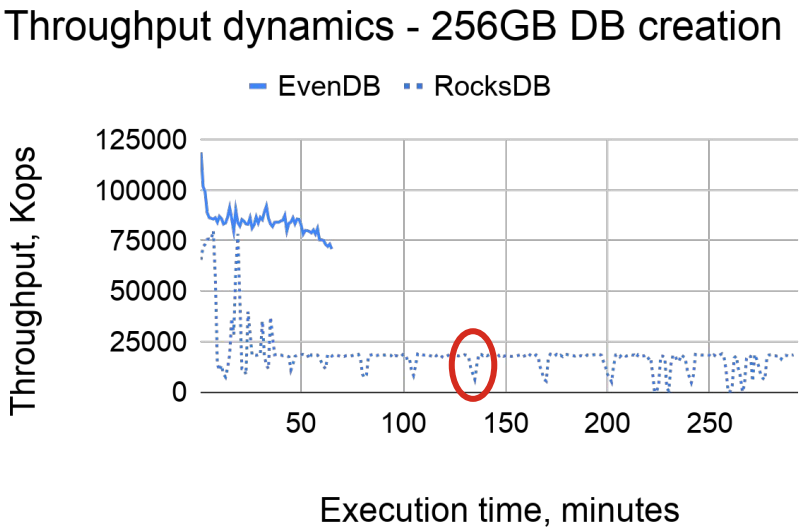
Real dataset ingestion

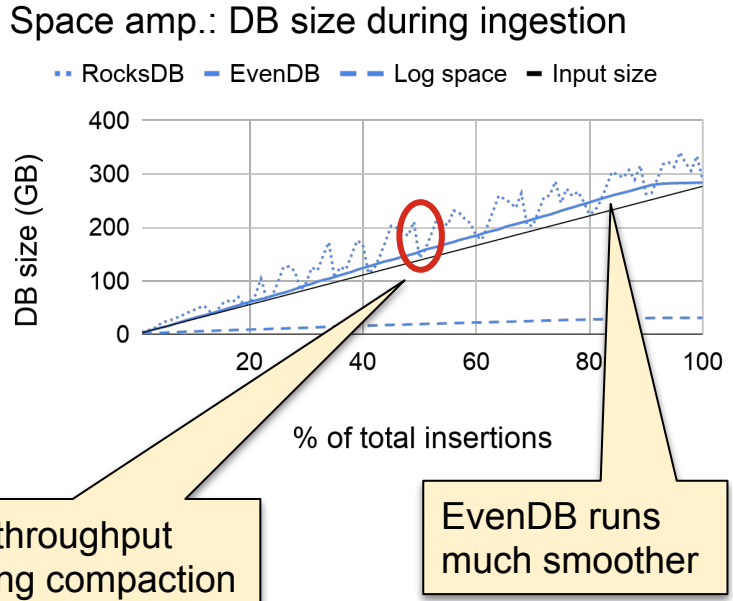


EvenDB 4.4x faster,

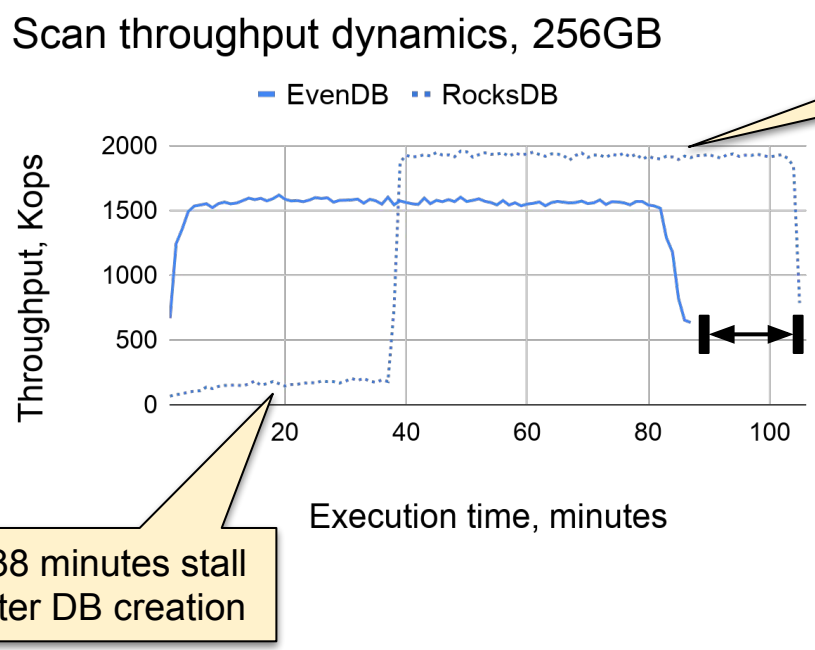
write amp. 4x lower (better)

Compactions impact





Real dataset scans



EvenDB 1.2x faster than RocksDB

Summary

- **EvenDB introduces a novel key-value store architecture**
- **Chunk arrangement better suited for spatially-local workloads than LSM:**
  - Lower write amplification
  - Single level of storage
  - Memory serves reads and writes
- **EvenDB outperforms RocksDB when:**
  - Workload is spatially-local or most working set fits in RAM
  - In par otherwise
  - Demonstrated in real workload and synthetic YCSB benchmarks