X-Engine

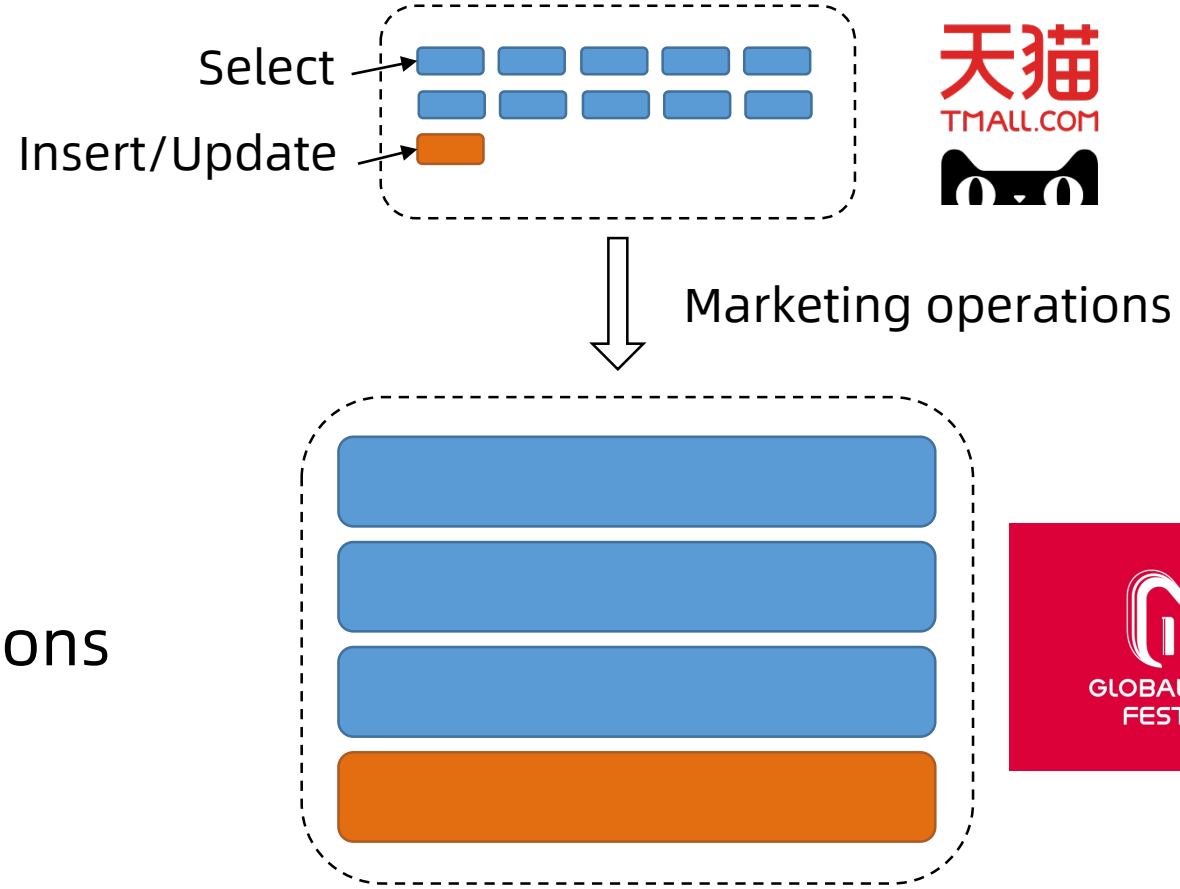
An Optimized Storage Engine for Large-scale E-commerce Transaction Processing

Authors: Gui Huang Xuntao Cheng Jianying Wang Yujie Wang Dengcheng He
Tieying Zhang Feifei Li Sheng Wang Wei Cao Qiang Li

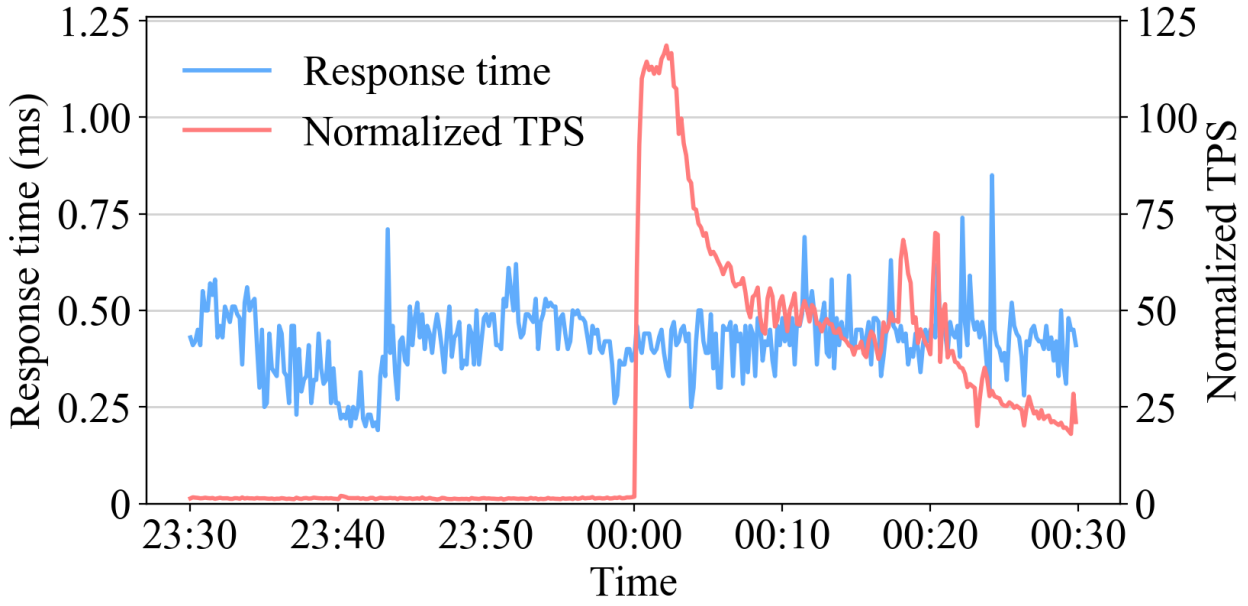
Presenter: Xuntao Cheng

DB for the e-commerce

- Storage cost
  - Business-critical data
  - Money burning SSDs
- Transactions
  - Mostly read-intensive
  - Ordinary days v.s. promotions



The tsunami problem

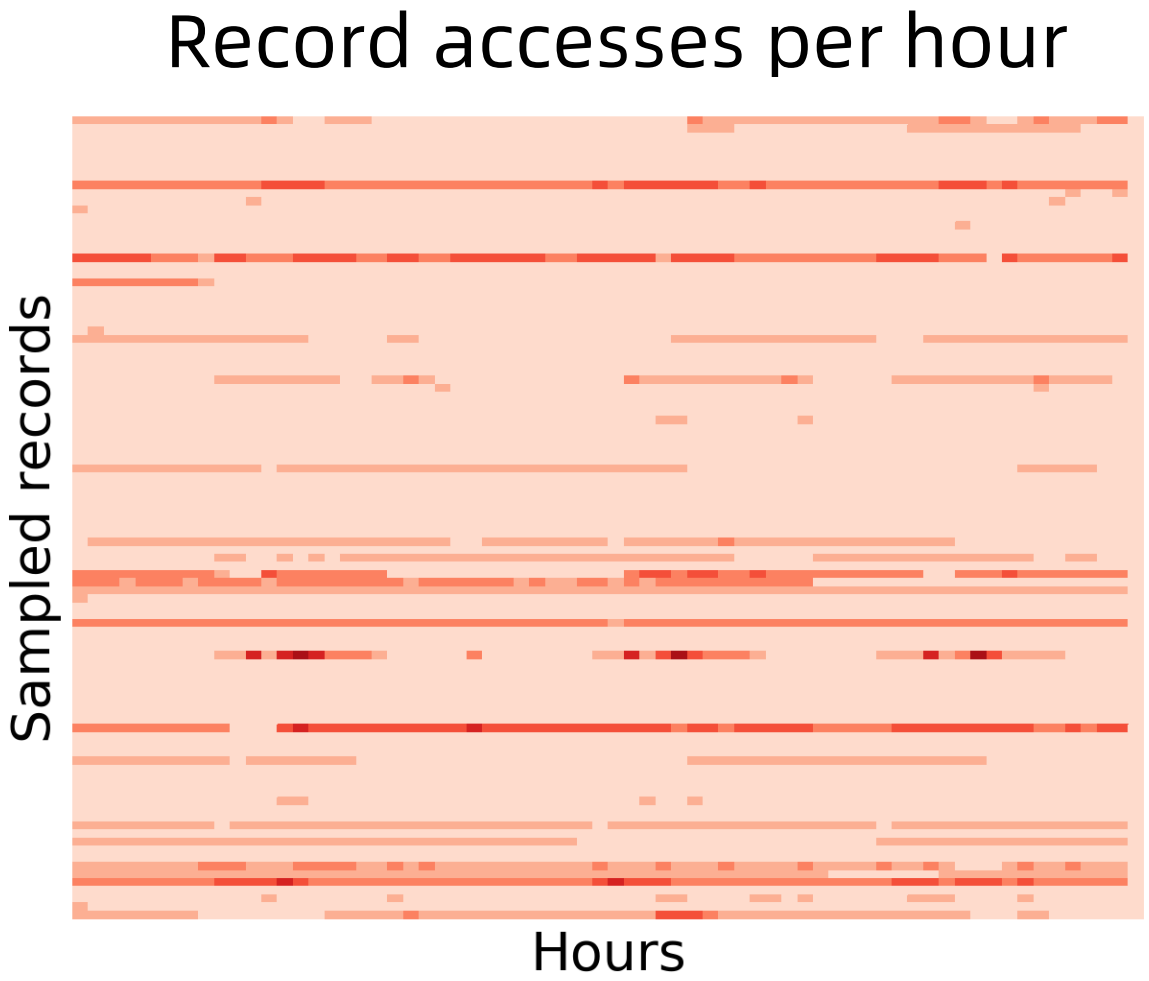


122 X spike

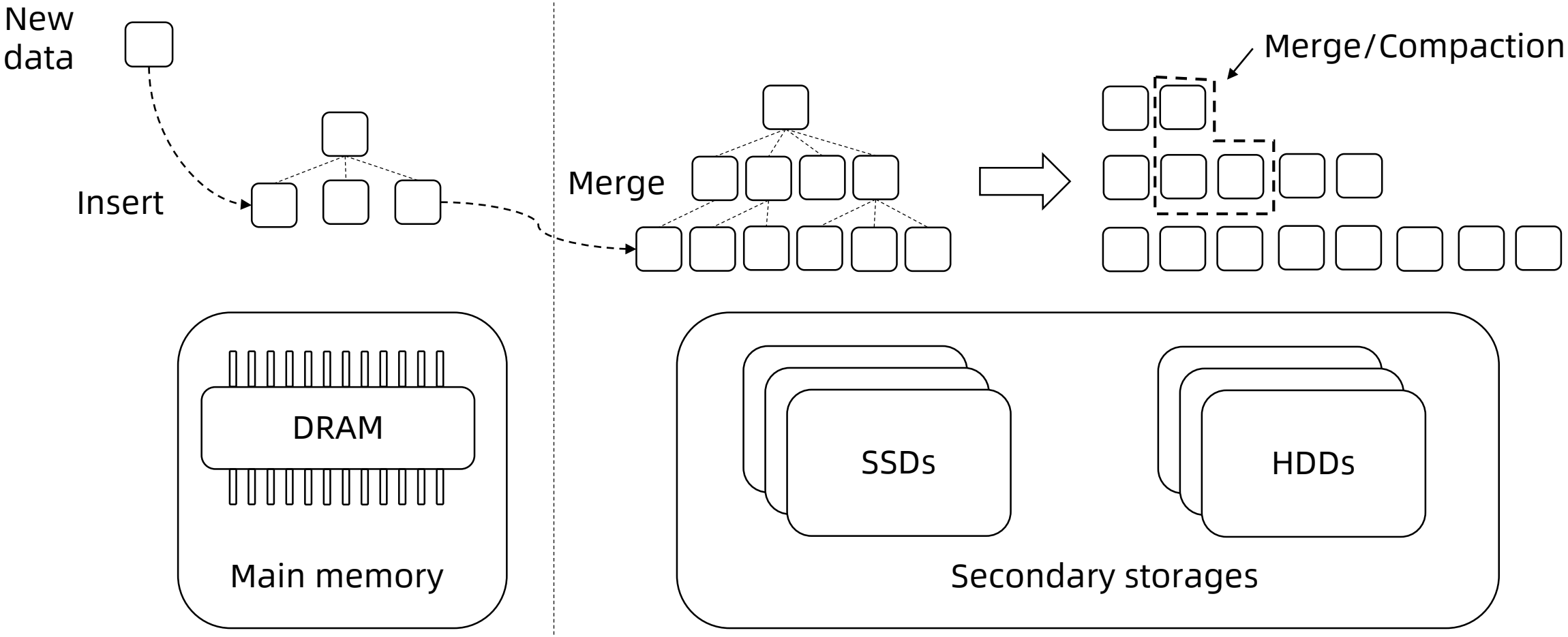
491 K sales transactions per second

11 Nov, 2018

Record temperatures

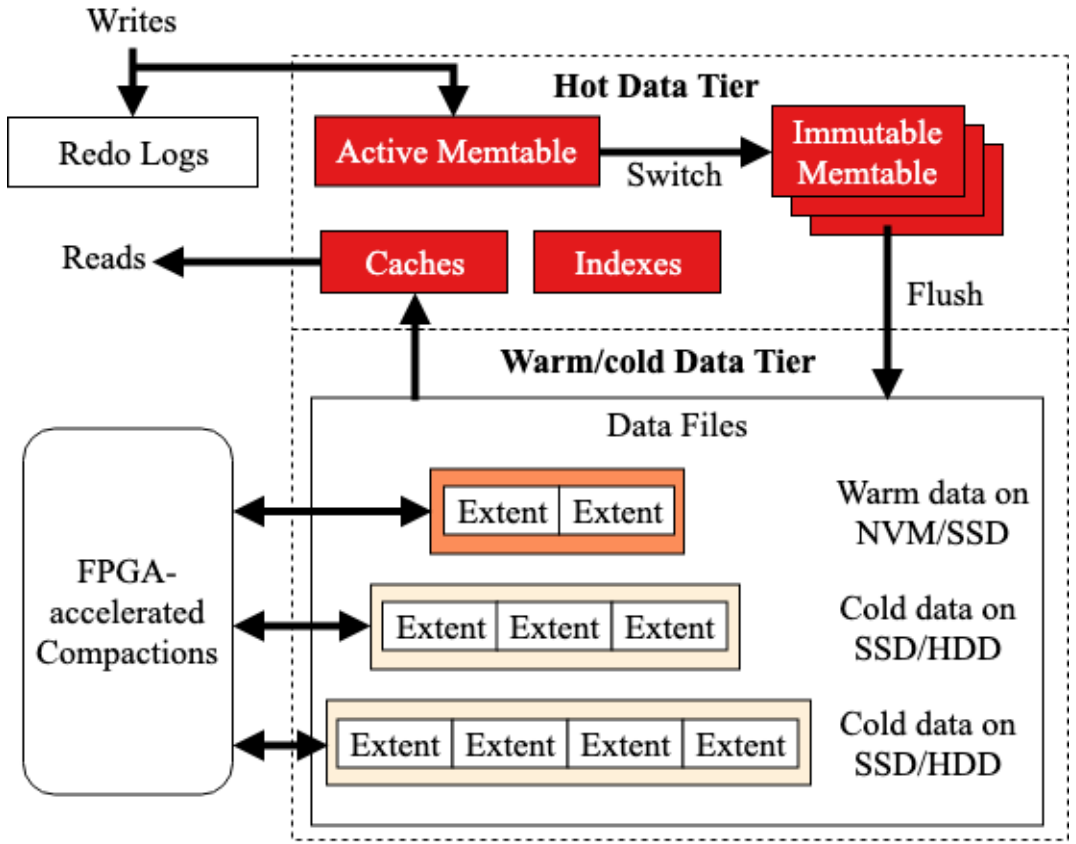


LSM-tree [O'Neil 1997]

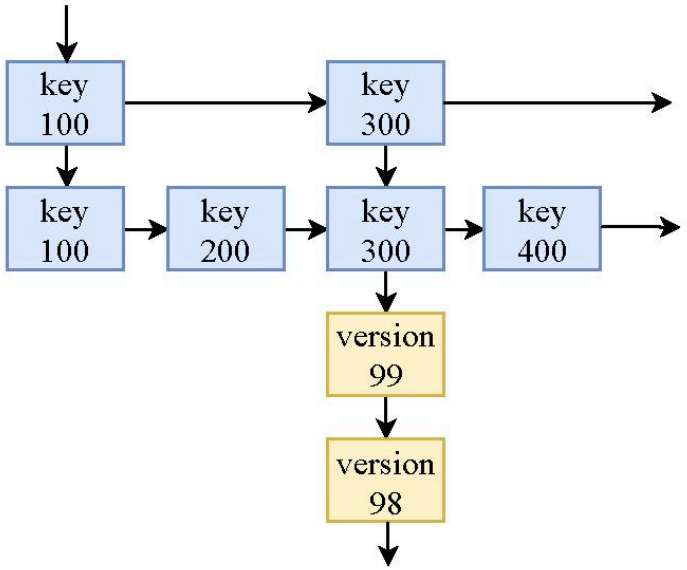


X-Engine architecture

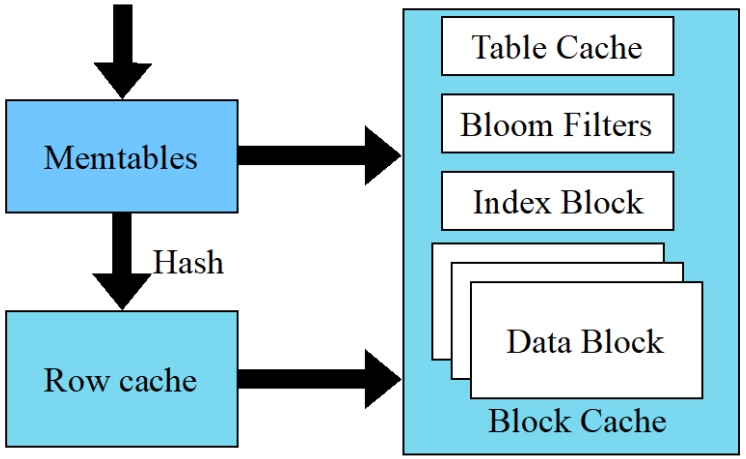
- Multi-version records with temperatures
- Logs first
- Specialized processors



Accessing hot records



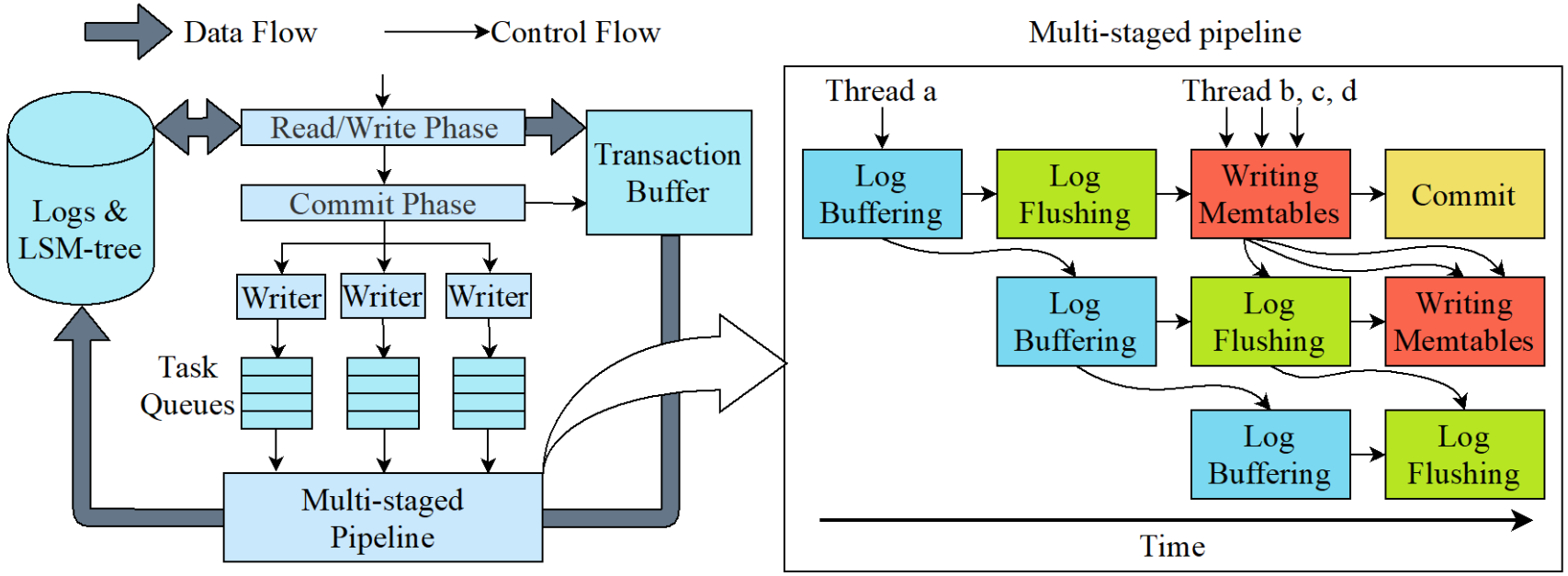
Multi-version memtable



Row/block caches

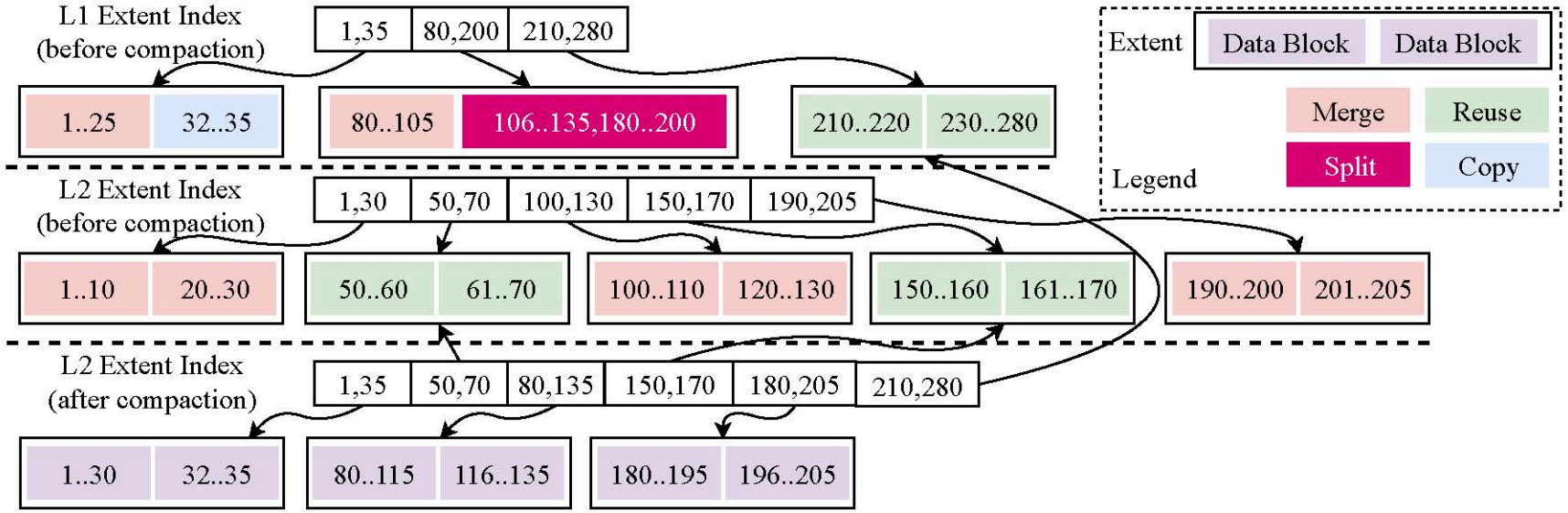
- Linked list for versions of newly inserted records
- Caches for flushed hot records

Optimizing the write path



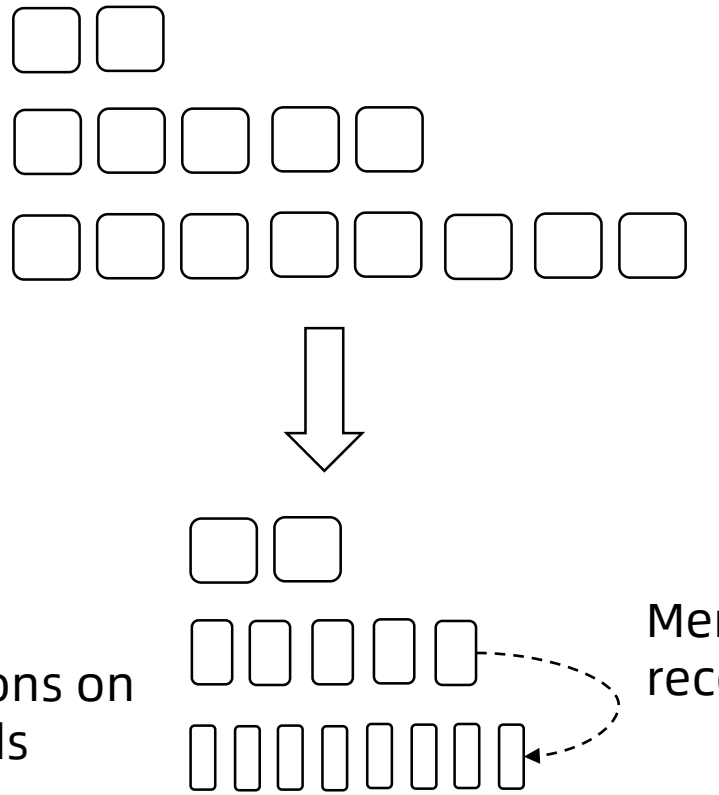
- Asynchronously buffering changes in transactions first
- Tuning thread-level parallelism for disk I/Os and memory writes

Slimming compactions



- Move pointers, not data
- Merge small blocks
  - If not possible, split them

Storage cost



Aggressive compressions on cold records

Merge cold records only

Dedicated compactions to reduce memory fragmentations

~50% space reduction

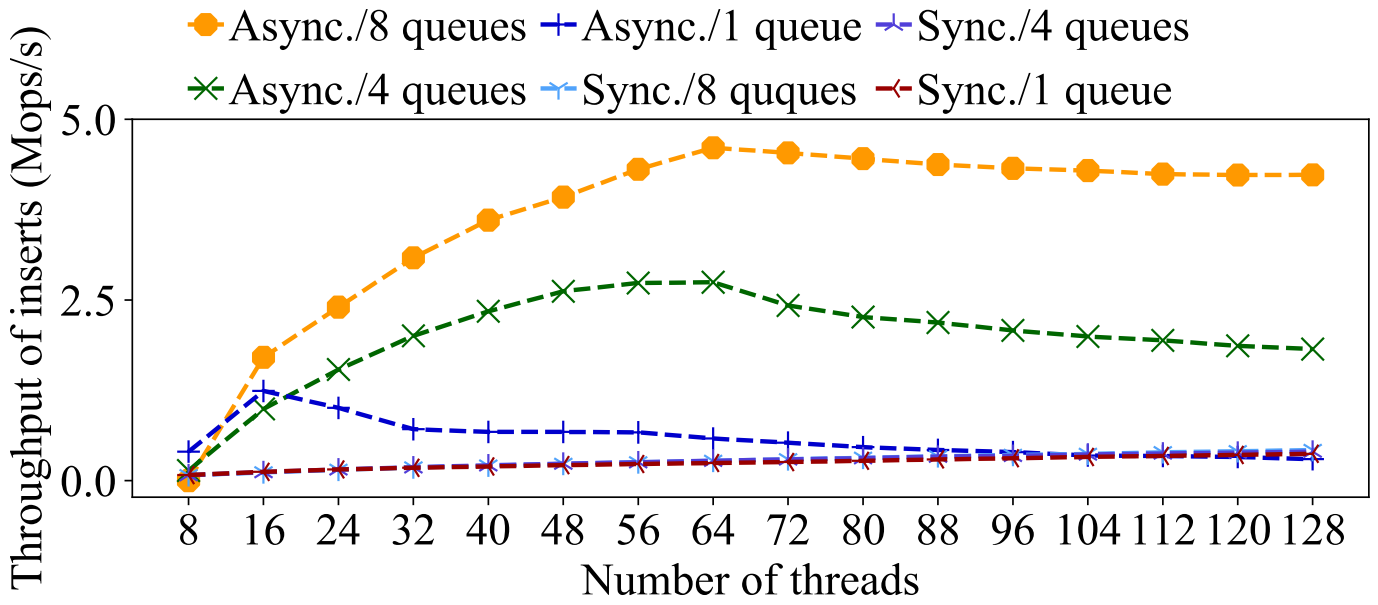
Summary of optimizations

- Optimizing the write path
  - Asynchronous writes in transactions
  - Multi-staged pipeline
  - Fast flush
- Reducing write/space amplifications
  - Small-size extents
  - Date reuse in compactions
  - FPGA-accelerated compactions
  - Incremental cache replacement
- Optimizing the read path
  - Caches (row, block)
  - Multi-version memtables
  - Multi-version metadata index

Experimental setup

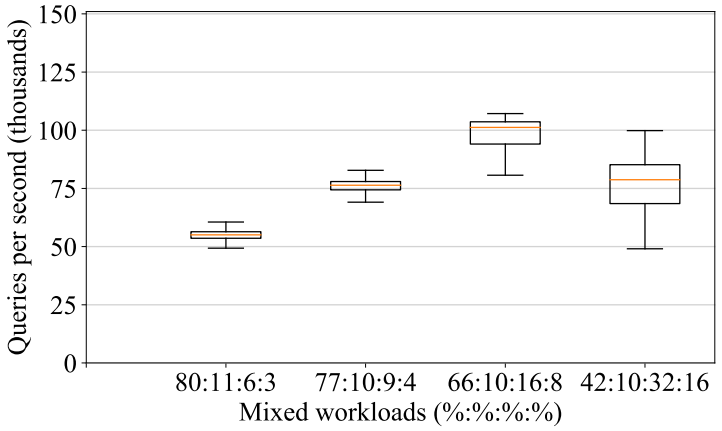
- Machines
  - Two 16-core Intel E5-2652 processors @ 2.3 GHz
  - 512 GB DDR4 main memory
  - A RAID of three 1TB SSDs
- Workloads
  - X-Bench: a self-developed stress-testing benchmark toolkit, capable of synthesizing e-commerce transactions
  - Dbbench for key-value tests
  - Sysbench for SQL tests

How fast can we achieve

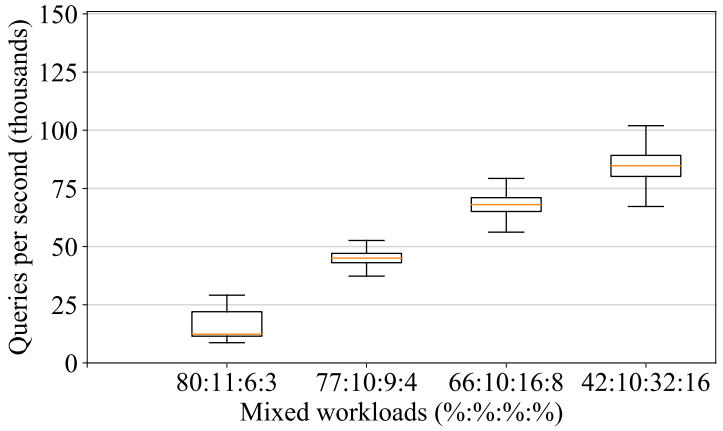


- 11 times faster than synchronous writes
- CPU efficiency ↑

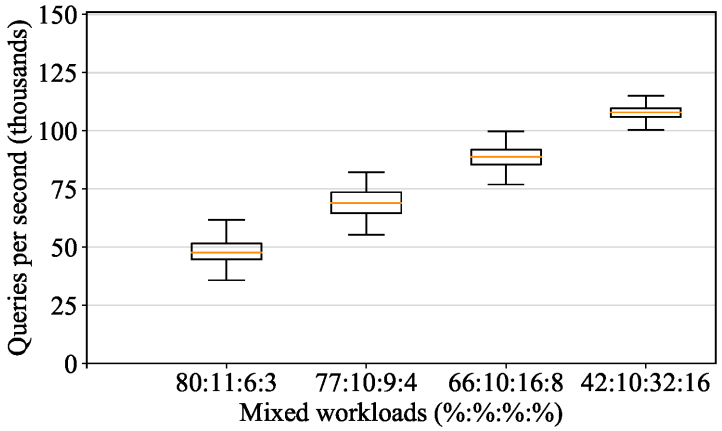
E-commerce transactions



InnoDB



RocksDB



X-Engine

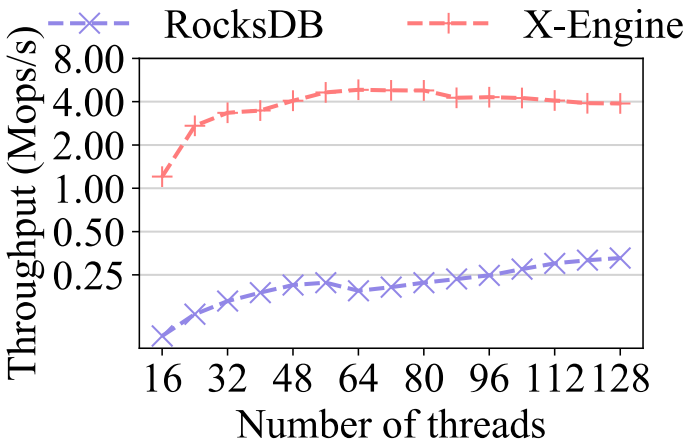
Left: non-promotional workload -> Right: promotional workload

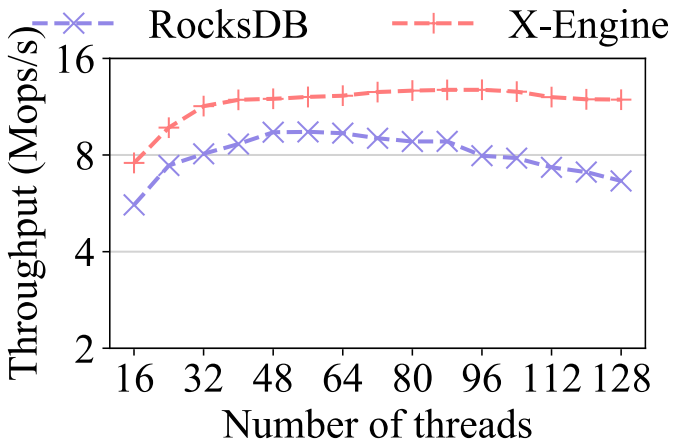
Plug X-Engine into MySQL, and compare it with other MySQL alternatives:

- Similar performance with InnoDB in non-promotional workload.
- Outperforms InnoDB in promotional workload.

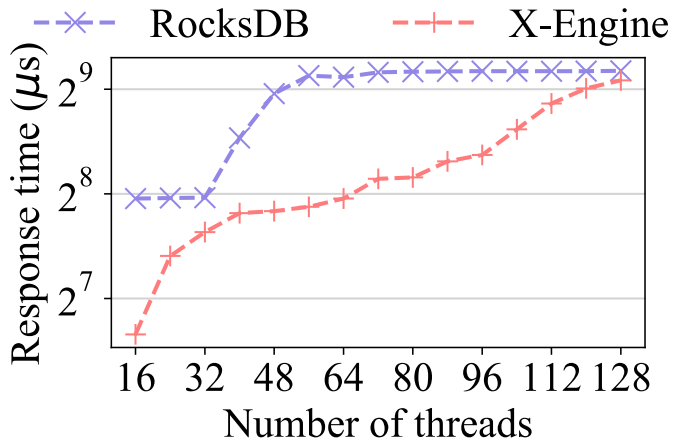
Peak in-memory performance

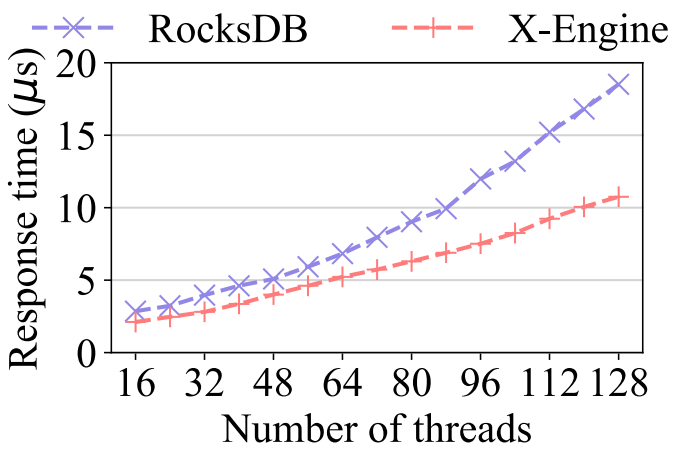
Put





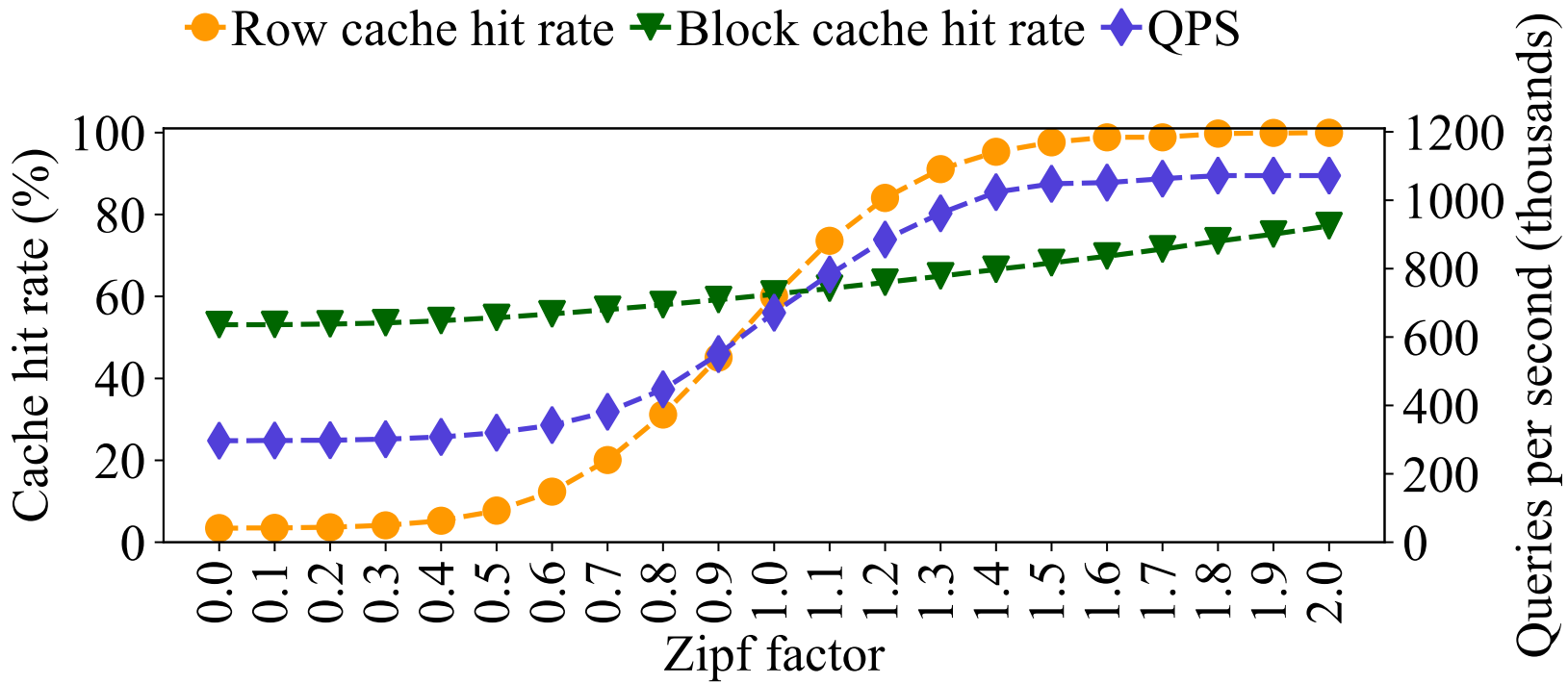
Get





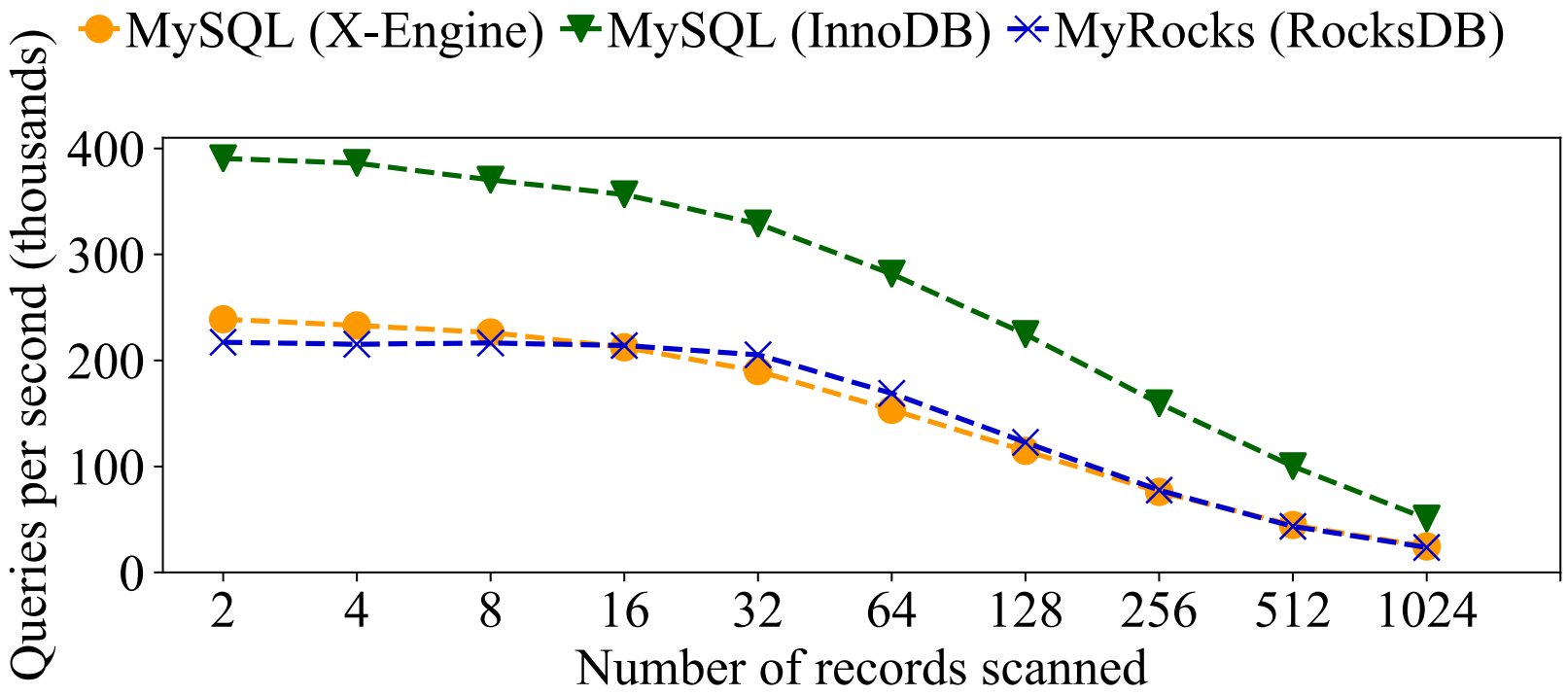
X-Engine has outstanding memory-only performance.

Row and block caches



Row cache is very impactful for highly skewed point queries, which are common in e-commerce workload.

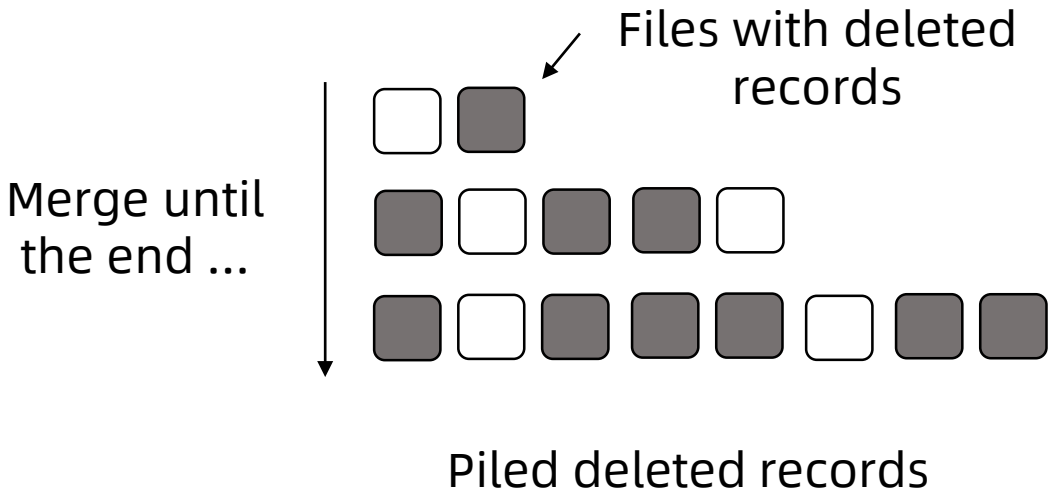
Range lookups



Range scans are drawbacks in LSM-tree systems. However, they are minor in e-commerce workloads.

Challenges

- Delayed compactions
- Write amplification
- Identification of cold records
- Benchmarking



Q & A

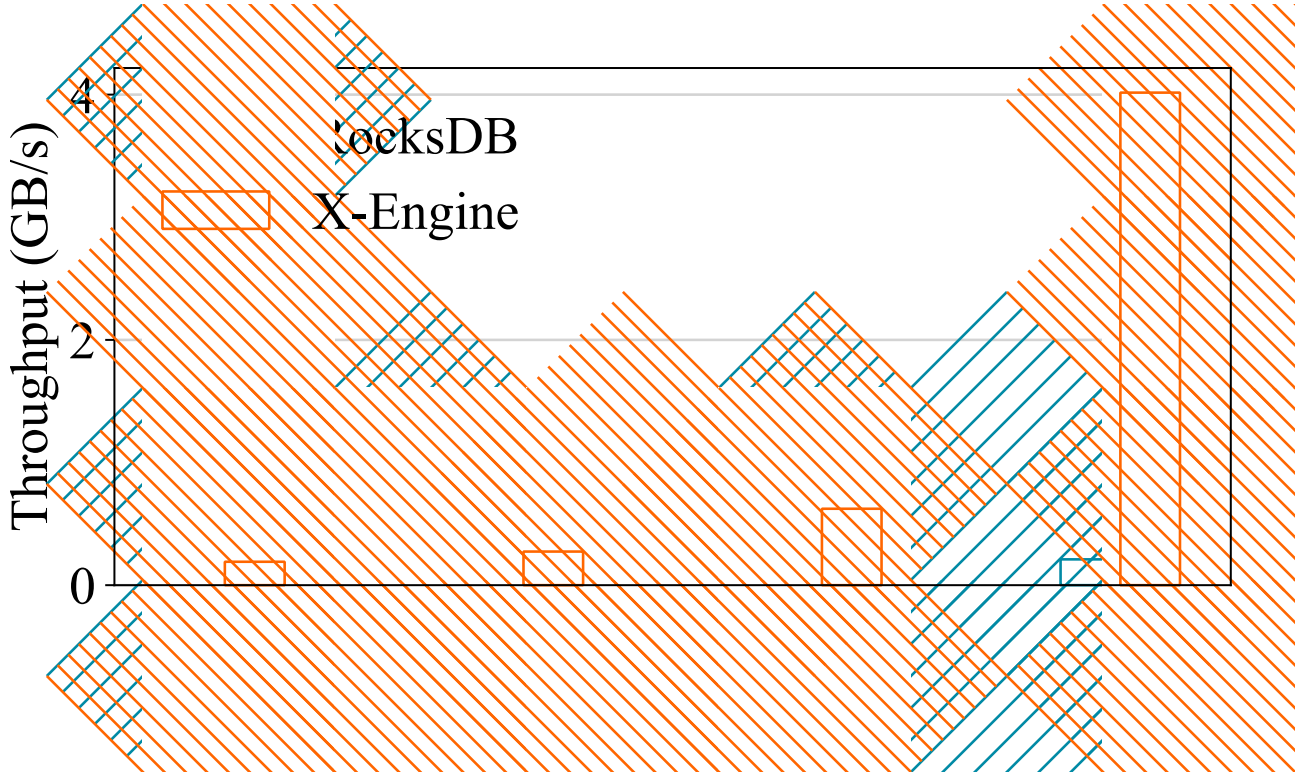
微信群

技术交流钉钉群

微信公众号

Backup slides

Data reuse in compactions



Small-size extents unleash more opportunities for data reuse during compaction.