The Task

- To implement high-throughput main-memory index
- All completed updates must be durable using SSD
  - Recoverable after system crash
- Stores key-value pairs
  - Key (unique): up to 1024 byte
  - Value: up to 4096 byte
- Offers 5 key-value operations:
  - Insert, find, delete, compare & swap, and iterate (lexicographical ordered traversal)
- Should support high-concurrent accesses
  - All single-key operations must be atomic
- SSD is formatted using ext4 filesystem

Architecture Overview

- Key-value operations
- Workload
- Threads
- B+-Tree based Trie
- Cache hit
- Cache miss
- Disk Writer
- Garbage Collector
- SSD
- Logging

Main Structure: B+-Tree based Trie

- Trie uses B+-tree as a node
- Key is split into 8-byte chunks
  - Each chunk is used as a key for each level of B+-tree
- Skewed trees are merged into one tree
  - To speed up traversing common prefix

Concurrency Control

- Tree level readers-writer lock
- Only downward propagation
  - To avoid deadlock

Disk Writer

- All update operations send redo log to Disk Writer
  - Delete does not include value-related fields

Garbage Collector

- Reclaims invalid logs to gather free space
  - Victim selection: round-robin policy

The Task

- To implement high-throughput main-memory index
- All completed updates must be durable using SSD
  - Recoverable after system crash
- Stores key-value pairs
  - Key (unique): up to 1024 byte
  - Value: up to 4096 byte
- Offers 5 key-value operations:
  - Insert, find, delete, compare & swap, and iterate (lexicographical ordered traversal)
- Should support high-concurrent accesses
  - All single-key operations must be atomic
- SSD is formatted using ext4 filesystem

Main Structure: B+-Tree based Trie

- Trie uses B+-tree as a node
- Key is split into 8-byte chunks
  - Each chunk is used as a key for each level of B+-tree
- Skewed trees are merged into one tree
  - To speed up traversing common prefix

Concurrency Control

- Tree level readers-writer lock
- Only downward propagation
  - To avoid deadlock

Disk Writer

- All update operations send redo log to Disk Writer
  - Delete does not include value-related fields

Garbage Collector

- Reclaims invalid logs to gather free space
  - Victim selection: round-robin policy

Architecture Overview

- Hash Cache
  - Maps from key to key-value object for fast lookup

Hash Cache

- Maps from key to key-value object for fast lookup