Evolution of Development Priorities in Key-value Stores Serving Large-scale Applications: The RocksDB Experience

wu

April 25, 2024

RocksDB is a key-value store targeting large-scale distributed systems and optimized for Solid State Drives (SSDs). This paper describes how our priorities in developing RocksDB have evolved. We describe how and why RocksDB's resource optimization target migrated from write amplification, to space amplification, to CPU utilization.

Lessons from running large-scale applications taught us that:

1. resource allocation needs to be managed across different RocksDB instances,
2. data format needs to remain backward and forward compatible to allow incremental software rollout,
3. appropriate support for database replication and backups are needed.

Lessons from failure handling taught us that:

1. data corruption errors needed to be detected earlier and at every layer of the system.

1 Introduction

Each RocksDB instance manages data on storage devices of just a single server node; it does not handle any inter-host operations, such as replication and load balancing, and it does not perform high-level operations, such as checkpoints

RocksDB and its various components are highly customizable, customizations can include the write-ahead log (WAL) treatment, the compression strategy, and the compaction strategy. RocksDB may be tuned for high write throughput or high read throughput, for space efficiency, or something in between.

Used in

- **Database**:
- **Stream processing**:
- **Logging/queuing services**:
- **Index service**:
- **Caching on SSD**:

Table 1: RocksDB use cases and their workload characteristics

	Read/Write	Read Types	Special Characteristics
Databases	Mixed	Get + Iterator	Transactions and backups
Stream Processing	Write-Heavy	Get or Iterator	Time window and checkpoints
Logging/Queues	Write-Heavy	Iterator	Support on HDD too
Index Services	Read-Heavy	Iterator	Bulk loading
Cache	Write-Heavy	Get	Can drop data

Table 2: System metrics for a typical use case from each application category

	CPU	Space Util	Flash Endurance	Read Bandwidth
Stream Processing	11%	48%	16%	1.6%
Logging/Queues	46%	45%	7%	1.0%
Index Services	47%	61%	5%	10.0%
Cache	3%	78%	74%	3.5%

2 Background

2.1 Embedded storage on flash based SSDs

The high performance of the SSD, in many cases, also shifted the performance bottleneck from device I/O to the network for both of latency and throughput. It became more attractive for applications to design their architecture to store data on local SSDs rather than use a remote data storage ser- vice. This increased the demand for a key-value store engines that are embedded in applications.

2.2 RocksDB architecture

Writes. Whenever data is written to RocksDB, it is added to an in-memory write buffer called **MemTable**, as well as an on-disk **Write Ahead Log (WAL)**. Memtable is implemented as a skiplist so keep the data ordered with $O(\log n)$ insert and search overhead. The WAL is used for recovery after a failure, but is not mandatory. Once the size of the MemTable reaches a configured size, then

1. the MemTable and WAL become immutable,
2. a new MemTable and WAL are allocated for subsequent writes,
3. the contents of the MemTable are flushed to a "Sorted String Table" (SSTable) data file on disk,
4. the flushed MemTable and associated WAL are discarded.

Each SSTable stores data in sorted order, divided into uniformly-sized blocks. Each SSTable also has an index block with one index entry per SSTable block for binary search.

Compaction. Levels higher than Level-0 are created by a process called **compaction**. The size of SSTables on a given level are limited by configuration parameters. When level-L's size target is exceeded, some SSTables in level-L are selected and merged with the overlapping SSTables in level-(L+1). This process gradually migrates written data from Level-0 to the last level. Compaction I/O is efficient as it can be parallelized and only involves bulk reads and writes of entire files.

Reads. In the read path, a key lookup occurs at each successive level until the key is found or it is determined that the key is not present in the last level.

RocksDB supports multiple different types of compaction:

- **Leveled Compaction**: levels are assigned exponentially increasing size
- **Tiered Compaction** (**Universal Compaction** in RocksDB): Similar to Cassandra or HBase. Multiple sorted runs are lazily compacted together, either when there are too many sorted runs, or the ratio between total DB size over the size of the largest sorted run exceeds a configurable threshold.

- **FIFO Compaction**: discards old files once the DB hits a size limit and only performs lightweight compaction. It targets in-memory caching applications.

#+CAPTION Write amplification, overhead and read I/O for three compaction types

Compaction	Leveled	Tiered	FIFO
Write Amplification	16.07	4.8	2.14
Max Space Overhead	9.8%	94.4%	N/A
Avg Space Overhead	9.5%	45.5%	N/A
# I/O per Get() with bloom filter	0.99	1.03	1.16
# I/O per Get() without bloom filter	1.7	3.39	528
# I/O per iterator seek	1.84	4.80	967

3 Evolution of resource optimization targets

3.1 Write amplification

Write amplification emerges at two levels:

1. SSDs themselves introduct write amplification: by their observation between 1.1 and 3.

2. Storage and database software also generae write amplification; this can sometimes be as high as 100 (e.g., when an entire 4KB/8KB/16KB page is written out for changes of less than 100 bytes)

   Level Compaction in RocksDB usually exhibits write amplification between 10 and 30, which is several times better than when using B-trees in many cases.

3.2 Space amplification

We observed that for most applications, space utilization was far more important than write amplification, given that neither flash write cycles nor write overhead were constraining.

In fact the number of IOPS utilized in practice was low compared to what the SSD could provide. As a result, we shifted our resource optimization target to disk space.

We developed **Dynamic Leveled Compaction**, where the size of each level in the tree is automatically adjusted based on the actual size of the last level.

	# keys (millions)	200	400	600	800	1000
Dynamgic Leveled	Fully compacted size (GB)	12.0	24.0	36.0	48.0	60.1
	Steady DB size (GB)	13.5	26.9	40.4	54.2	67.5
	Space overhead (%)	12.4	11.8	12.2	12.7	12.4
LevelDB-style Compaction	Fully Compacted size (GB)	12.0	24.0	36.4	48.3	60.3
	Steady DB size (GB)	15.1	26.9	42.5	57.9	73.8
	Space overhead (%)	25.6	12.2	16.9	19.7	22.4

3.3 CPU utilization

1. prefix bloom filter
2. applying the bloom filter before index lookups
3. bloom filter improvements

3.4 Adapting to newer technologies

Disaggregated (remote) storage appears to be a much more interesting optimization target and is a current priority. Faster networks currently allow many more I/Os to be served remotely, so the performance of running RocksDB with remote storage has become viable for an increasing number of applications.

3.5 Main Data Structure Revisited

WiscKey/ForrestDB

4 Lessons on serving large-scale systems

4.1 Resource management

The fact that a host may run many RocksDB instances has implications on resource management. Given that the instances share the host's resources, the resources need to be managed both globally (per host) and locally (per instance) to ensure they are used fairly and efficiently. When running in single process mode, having global resource limits is im- portant, including for

1. memory for write buffer and block cache
2. compaction I/O bandwidth
3. compaction threads
4. total disk usage
5. file deletion rate

4.2 WAL treatment

For example, if copies of the same data exist in multiple replicas, and one replica becomes corrupted or inaccessible, then the storage system uses valid replica(s) from other unaffected hosts to rebuild the replica of the failed host. For such systems, RocksDB WAL writes are less critical. Further, distributed systems often have their own replication logs (e.g., Paxos logs), in which case RocksDB WAL are not needed at all.

4.3 Rate-limited file deletions

Rate-limited file deletions RocksDB typically interacts with the underlying storage device via a file system. These file systems are flash-SSD-aware; e.g., XFS, with realtime discard, may issue a **TRIM** command to the SSD whenever a file is deleted. TRIM commands are commonly believed to improve performance and flash endurance. However, it may also cause performance issue. In addition to updating the address mapping, the SSD firmware also needs to write these changes to FTL(Flash Translation Layer)'s journal in flash, which in turn may trigger SSD's internal garbage collection. To avoid TRIM activity spikes and associated increases in I/O latency, we introduced rate limiting for file deletion to prevent multiple files from being deleted simultaneously.

4.4 Data format compatibility

It is important that the data on disk remain both backward and forward compatible across the different software versions.

4.5 Managing configurations

4.6 Replication and backup support

Bootstraping a new replica by copying all the data from an existing one can be done in two ways:

1. read all keys from a source replica and then written to the destination replica (**logical copying**). On the source side, RocksDB supports data scanning operations by offering the ability to minimize the impact on concurrent online queries; e.g., by providing the option to not cache the result of these operations

2. Copying SSTables and other files directly (**physical copying**). RocksDB assist physical copying by identifying existing database files at a current point in time, and preventing them from being deleted or mutated.

5 Lessons on failure handling

5.1 Frequency of silent corruptions

CPU/memory corruption does happen rarely and it is difficult to accurately quantify.