Contribution ID: **312** Type: **not specified**

A zone-aware cache system for distributed databases

Wednesday, 14 September 2022 10:05 (20 minutes)

Modern analytical distributed database platform requires massive data from remote filesystems(e.g. HDFS). A cache layer is necessary to eliminate the network bottleneck by caching data items with smaller granularity (e.g. 64KB ~ 128KB).

There are three major challenges to implementing such a cache system:
1. Predictable latency (latency spike is not acceptable)
2. Good enough user-side throughput (low end-to-end write amplification)
3. High-density storage per server with reasonable cost (SSD Cache is required)

Our previous solution is to use a traditional storage engine TerarkDB (which is a fork of RocksDB with much better throughput, and with a lot of optimization) on the EXT4 filesystem. But the result still can't meet our expectations:
1. A lot of latency spikes (we don't want to magically tuning it again and again under different workloads)
2. Too much write amplification so we cannot get enough write throughput.
2. Cannot provide predictable space cost (space amplification) and cannot make use of QLC SSD (random write), thus high-density storage cost is not acceptable

To solve these problems, we re-designed our Cache system by following the ZNS principles:
1. In-memory metadata and record-level indexing (thanks to large item size), so we have no read amplification.
2. Append only IO model with limited active write point, so we can make use of ZNS devices
3. User-controlled GC, so we would be able to use almost all space of the disk (a few reserved zones for data migration is enough), this is not possible on lsm-tree and conventional drives
4. Emergency data sacrifice (a cache system can usually tolerance some data loss), so we can make sure the device space is always fully utilized

Under benchmarks, we've got: 1) Much lower storage cost (QLC SSD & fully utilize disk space); 2) Stable latency (user-controlled GC & record-level indexing); 3) 5X+ better write throughput (append-only IO model);

Further works: we still haven't tested it under ZNS QLC SSD yet but expect to have a stable performance.

I agree to abide by the anti-harassment policy

Yes

Primary author: GUO, Kuankuan

Presenter: GUO, Kuankuan

Session Classification: Zoned Storage Devices (SMR HDDs & ZNS SSDs) MC

Track Classification: LPC Microconference: Zoned Storage Devices (SMR HDDs & ZNS SSDs) MC