Key/Value SSD
Design Overview and Use Cases

Stanley Miao
Chief Engineer
Shannon Systems
Unstructured data dominates

80% Unstructured vs 20% Structured

- Database
- Tables
Use Cases

- Pika
- ceph
- CockroachDB
- MyRocks
- TiDB
- Apache Flink

- mongoDB
- levelDB
- RocksDB

SSD

KVSSD
The overhead cost of RocksDB

Application
RocksDB
Filesystem
Block Device
SSD
The overhead of RocksDB

1. Read/Write must be aligned to 4KB sectors
The overhead of RocksDB

1. block bitmap
2. map table: file address -> LBA.
3. the journal
The overhead of RocksDB

1. High write amplification caused by compaction.
2. Write Log
3. Map key -> file address.
The overhead of RocksDB

Application
RocksDB
Filesystem
Block Device
SSD

key -> file address
file address -> LBA
LBA -> PBA
The overhead of RocksDB

Application
RocksDB
Filesystem
Block Device
SSD

Log + Data
Journal + Data
Log Structured
Software Architecture Evolves

Application
RocksDB
Filesystem
Block Device
SSD

Application
KVSSD
KVSSD Software Stack

- Application
- User Space
- KV Library
- Kernel Space
- KVSSD Driver
- Hardware
- Open Channel SSD
KVSSD Driver Read/Write Flow

User Key

Index Table

Physical Address

Nand Flash

Nand Flash Page

Metadata | Key | Value

User Key/Value

Index Table

Physical Address

DRAM on SSD

Nand Flash

Nand Flash Page
• **KEY_SIZE**: 1B ~ 127B
• **VALUE_SIZE**: 1B ~ 16MB
• Persistency
• Atomic Write
Test Conditions: key_size = 16bytes, value_size = 4KB, threads = 10, WriteOption.sync=true
Test Conditions: key_size = 16bytes, value_size = 4KB, threads = 10, WriteOption.sync=true
Random Write Performance Comparison

Test Conditions: key_size = 16bytes, value_size = 4KB, threads = 10, WriteOption.sync=true
Use Cases: Pika's vire-benchmark result
github projects

https://www.github.com/shannon-sys

• pika
• mongodb
• kv_library
• benchmark tools
KV Library API

- Provide both C and C++ API
- Compatible with RocksDB API
- `kv_put/kv_get/kv_delete/kv_exist`
- Column Family
- Snapshot
- Iterator
- `Write_batch/Read_batch`
Shannon KV SSD products

- Interface: PCIe/u.2
- Capacity: 1TB/2TB/4TB
Questions?

Contact me:
Booth: 413
Email: stanley@shannon-sys.com
Wechat: stanleymiao