MASSé: Media Aware Smart Storage Engine

Jack Zhang
Cloud & Enterprise Architect
yuan.zhang@intel.com
Agenda

• MASSé introductions, Tiered storage for Optane+QLC
• MASSé Evaluation and Proof
• What Comes Next

MASSé = Media Aware Smart Storage Engine
Feedbacks:

- “Why do I not see x number of times improvement over flash SSDs when dropped in an Intel® Optane™ SSD?”
- “Re-shaping writes into larger datasets and sequentially sending to a QLC SSD requires additional software investments, and implementations differ from application to application...is there a generic solution that supports this?”

Solution:

- Media Aware -- uniquely identifies and classifies heterogeneous SSDs by their media type, and builds inclusive data structures and algorithms, accordingly, helping to release maximum SSD capabilities to applications
- Smart -- intelligent module features such as data placements, IO re-shaping, key-value/virtual filesystem/virtual block APIs, workload pattern AI engine etc.
- Storage Engine -- replacement of filesystem and managing raw SSD blocks without modifying SSD firmware and kernel modules
Configurable Engine

Host/Application

node

Intel® Optane™ SSD DC P5800X

Intel® QLC 3D NAND SSD

SMART STORAGE ENGINE x5

SSE1

SSE2

SSE3

SSE4

SSE5

Best

Good

Better

Better

Better
Software Architecture

Key-Value, vFS, vBlock
Telemetry/Utility
NVMe Driver
GC Manager
KV database
Volume, Tier
Mapping Table
Workload AI Engine

Optane™ PMEM/SSD
NAND Flash SSD
Multi-Tier Architecture

DRAM

Intel Optane™ PM/SSD

QLC SSDs
Optane as write pad, QLC as capacity store
Data layout in QLC Flash

- Reserved Zone (n×LBand size)
- LBand Cluster 1 (default 512MB)
  - LBand Cluster 1 (default 512MB)
  - LBand Cluster 1 (default 512MB)

- Header
- Tail
  - Segment 1
  - Segment 2
  - Segment 3
  - Segment n
  - Segment m
  - LBand
MASSé Evaluation and Proof

1. MASSé vs RocksDB (media un-aware engine) performance comparison
2. MASSé performance with different SSD media
3. MASSé case study in real customer application, Bytedance TerarkdB
MASSé vs RocksDB

Test configurations:
CPU: Intel(R) Xeon(R) Gold 6142M CPU @ 2.60GHz, Memory: 384GB, Storage: Intel® Optane™ SSD P4800X 375GB, Intel® SSD DC P4510 8TB
Workloads: Index search.
db_bench, 64threads KV(23B, 100B), 1Billion kv pairs, readwhilewriting 50/50 r/w
For more complete information about performance and benchmark results, visit www.intel.com/benchmarks.
MASSé w/ different SSD media

Test configurations:
- CPU: Intel(R) Xeon(R) Gold 6142M CPU @ 2.60GHz
- Memory: 384GB
- Storage: QLC=Intel® SSD D5-P4326, TLC=Intel® SSD DC P4510 8TB “Optane” =Intel® Optane™ SSD DC P4800X 375GB
- db_bench: readwhilewriting, random 50% / 50%
- 64threads KV(16B, 4096B), 1Billion KV datasets

For more complete information about performance and benchmark results, visit www.intel.com/benchmarks
Replaces EXT4 FS

TerarkdB today

- RocksDB
  - Filesystem EXT4
    - NVMe Driver
- 3D XPOINT™ QLC

TerarkdB + MASSé

- RocksDB
  - vFile MASSé
  - Media Aware
- Filesystem EXT4
  - NVMe Driver
- Bypass FS EXT4
- 3D XPOINT™ QLC
  - Optane PM
Case Study TerarkDB: MASSé replacement of EXT4

Key=20B, Value=400B
readrandomwriterandom 70/30
100M entries, no read cache
3.2Billion Operations

OP/s

readrandomwriterandom 70/30

workloads

P4800X (EXT4)  P4510 (MASSé)  P4800X (MASSé)
ops        61606        99490        119782

194%

P4800X (EXT4)  P4510 (MASSé)  P4800X (MASSé)
Average 742.05  1839.25  3612.45
99% 5860.3  5860.3  6075.11
99.9% 5860.3  5860.3  6075.11
99.99% 5860.3  5860.3  6075.11

./db_bench --skvds=false (or true) --db=/mnt/Xdb (or /test)--benchmarks=readrandomwriterandom --threads=32 --readwritepercent=70 --num=100000000 --key_size=20 --value_size=400--options_file=../skvds_options --statistics=1 --histogram=1
What Comes Next

• Conclusions

1) MASSé is a high-performance and effective storage solution that releases the maximum power of heterogeneous SSD media. It is an inclusive design that reduces application burdens and encourages investments in new storage technologies.

2) By making the combination of Optane and QLC SSDs more effective, MASSé meets the growing demands of cloud and datacenter to improve performance while reducing cost

• Next steps

1) Design standard MASSé lib and userspace module, standardize vFile and vBlock interfaces

2) Design media aware RocksFS to replace RocksDB filesystems-- improve RocksDB performance especially with Optane, in general, RocksFS = abstract POSIX FS + MASSé

3) Opensource, MASSé revision 1.0 released at private https://github.com/TeamSKVDS/skvdsmaster

Please take a moment to rate this session.

Your feedback is important to us.