NVDIMM
The Savior of SSD Endurance in CEPH

David Tseng
Bigtera Inc. | Flash Memory Summit 2018
3 Replicas = 2100 MB/s
3 Replicas = 300 MB/s
Why CEPH?

**UNIFIED STORAGE**

- Object
- Block
- Filesystem

**SCALE-OUT**

- Pay as you grow
- No single point of failure
The Source of Write Amplifications
OSD
JOURNAL
WAF

Journal
Data
CEPH FileStore

2.98
4.33

© 2018 Bigtera | Flash Memory Summit
OSD
JOURNAL
WAF

NVDIMM

2.98

4.33

Journal
Data
CEPH FileStore
3 Replicas = 2100 MB/s

USER

CEPH

SSD

SSD

SSD
3 Replicas = 1200 MB/s
METADATA FOR EACH WRITE

Re-writing Same Key

PG INFO

OBJECT SNAPSHOT INFO

OBJECT INFO

PG LOG
METADATA STORAGE

EXT4
EXTENDED ATTRIBUTE
Limited length
Write amplification on Disk

KV
STORE
Compaction overhead
Write amplification on Disk

NVDIMM
STORAGE
Stores hot metadata
Minimal writes on Disk
OSD METADATA
Re-writing same KV pair
OSD METADATA
Re-writing same KV pair

More Merge
Less EXT4-J Lock Contention
OSD METADATA
Re-writing same KV pair

METADATA -> META CACHE -> KV -> EXT4

Store Metadata in NVDIMM
Flush when
(Idle & High cache pressure)
METADATA FOR EACH WRITE

New Log for Each Write
OSD METADATA
PGLog Traditional behavior

Insert new PGLog

KV
STORE
OSD METADATA
PGLog Traditional behavior

Remove old PG Logs  Insert new PGLog

KV STORE
OSD METADATA
PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG

NVDIMM Cache Entry

Head / Tail

Update PGLog Pos

Buffer

Insert new PGLog

Fixed Pre-allocated Space
OSD METADATA
PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG

NVDIMM Cache Entry

Head / Tail
Update PGLog Pos

Buffer
Insert new PGLog

Fixed Pre-allocated Space
OSD METADATA

PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG

NVDIMM Cache Entry

Head / Tail
Update PGLog Pos

Buffer
Insert new PGLog

Fixed
Pre-allocated
Space

16K Aligned
OSD METADATA
PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG

NVDIMM Cache Entry

Head / Tail

Buffer

16K Aligned

Fixed Pre-allocated Space
OSD METADATA
PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG

NVDIMM Cache Entry

Head / Tail
Buffer

Save to KV when
(Idle & High cache pressure)

Fixed
Pre-allocated
Space

KV STORE
OSD METADATA WAF

**Journal**

**Data**

<table>
<thead>
<tr>
<th></th>
<th>CEPH FileStore</th>
<th>VS Extreme M</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2.98</strong></td>
<td>4.33</td>
<td>2.21</td>
</tr>
</tbody>
</table>

© 2018 Bigtera | Flash Memory Summit
3 Replicas = 1200 MB/s
3 Replicas = 600 MB/s

USER

CEPH

SSD

SSD

SSD
EXT4 JOURNAL & METADATA
EXT4 WRITE AMPLIFICATION

Keep track of mtime

Re-writing Data

DATA

Re-writing Data

DATA

DATA BLOCKS

Update mtime

METADATA (INODE)

Journal Due to Metadata Update

JOURNAL

EXT4
EXT4 WRITE AMPLIFICATION

Do not touch mtime

Re-writing Data

OSD does _not_ require mtime information on filesystem

Re-writing Data

nomtime

DATA BLOCKS

METADATA (INODE)

JOURNAL

DATA

EXT4
EXT4 WAF

- CEPH FileStore: 4.33
- VS Extreme M: 2.21
- VS Extreme ME: 1.05

- Journal
- Data
3 Replicas = 600 MB/s

USER → 100 MB/s → CEPH → 200 MB/s → SSD

200 MB/s → SSD

200 MB/s → SSD

200 MB/s → SSD
3 Replicas = 300 MB/s

USER ➔ 100 MB/s ➔ CEPH ➔ 100 MB/s ➔ SSD ➔ 100 MB/s ➔ SSD ➔ 100 MB/s ➔ SSD

MISSION ACCOMPLISHED
**VirtualStor™ EXTREME**

**LOCK FREE QUEUES**
- Messenger
- OSD Workers
- IO Workers

**AVOID LOCK CONTENTION**
- Per thread data
- Smarter job scheduling
- Atomic data structures
- Parallel RW lock

**NVFAStore**
- AIO supported backend
- Specialized NVDIMM journal structure
- No locks in all data path

**CPU OPTIMIZATIONS**
- Memory spatial / temporal locality
  - Huge pages
  - Avoid memory copy
  - Less context switches
  - CPU Affinity
  - NUMA consideration

**OBJECT POOLS**
- Better reuse of frequently used objects
- Better memory spatial locality
  - Lock free allocation / free
  - Huge page to have better TLB cache hit
VIRTUALSTOR EXTREME

- Hardware Form Factor: 4 nodes in 2U chassis
- Spec Per Node:
  - 2x Intel Xeon Scalable Processor
  - 256 GB Memory
  - 2x 16 GB NVDIMM
  - 6x SATA SSD: [960 GB, 1.92 TB or 3.84 TB]
  - Network 2x 10GbE and 1x 1GbE IPMI
  - Add-on: Dual-Port 40GbE, Fiber Channel
BENCHMARK ENVIRONMENT

Brocade Turbolron 24X
10GbE Switch

Client Node* 4
Intel Gold 5118 x2
128GB RAM
Intel X710 10GbE
Ubuntu 16.04

Storage Nodes (2U4N)
SuperMicro 2029TP-HCOR
Intel Silver 4114 x2
192GB RAM
Viking NVDIMM 16GB x2
Intel 82599ES 10GbE
VirtualStor Extreme
Intel DC S4500 x6

Intel CeTune Benchmark
112GB Vdisk* 16 VM* 4 Nodes
7TB Dataset
1QD* 16 VM* 4 Nodes
64QD
SSD LATENCY
CEPH Mimic BlueStore VS VirtualStor Extreme

Mimic BlueStore

VirtualStor Extreme

Latency (ms)

0 200 400 600 800 1000 1200 1400 1600
times (s)

Latency (ms)

0 200 400 600 800 1000 1200 1400 1600
times (s)
CLIENT LATENCY
CEPH Mimic BlueStore VS VirtualStor Extreme

Mimic BlueStore

VirtualStor Extreme
Thank you for your attention.
Questions?