Open. Together.
Unleash Stranded Flash Capacity - Disaggregated Storage Architecture, Trends and OCP Solutions

Manoj Wadekar, Storage Engineer, Facebook
Anjaneya “Reddy” Chagam, Sr Principal Engineer, Intel Corporation

Contributors: Vaidyanathan Krishnamoorthy (Intel), Jeff Smits (Intel), Siying Dong (FB)
Notices & Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. **No product can be absolutely secure.**

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit [http://www.intel.com/benchmarks](http://www.intel.com/benchmarks).

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit [http://www.intel.com/benchmarks](http://www.intel.com/benchmarks).

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

© 2019 Intel Corporation.
Intel, the Intel logo, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as property of others.*
Agenda

• Disaggregated Storage Architecture
• Test Configuration
• Benchmark Results
• Summary
Disaggregated Storage Architecture

**Compute Node**
- Application
- Operating System
- Hardware
- SSDs

**Head Node**
- Target
- Operating System
- Hardware

**Disaggregated**

- **Logical disaggregation**
  - Consumes physical or logical block devices
  - **Dynamic binding** based on workload requirements
  - Efficient, improved TCO

- **Physical disaggregation**
  - Static binding
  - Shared resources
  - Target can expose physical or logical devices

- **Local attached storage**
- **Static binding**
- **Stranded capacity**, IOPS
- Inefficient, increased TCO

PCIe

JBOF

Network

iSCSI, NVMe-oF, etc.
NVM Express (NVMe)
Standardized interface for non-volatile memory, http://nvmexpress.org

Performance: PCIe Gen3 1GB/s per lane (x4 = 4GB/s)
Low Latency: Direct CPU connection
No Host Bus Adapter: Lower power, lesser cost
Form Factor options: PCIe AIC, SFF-8639, M.2, SATA express, BGA
Remote Block Storage - Network Protocols

- Enables sharing of NVMe flash storage over network
- Can use traditional block protocols (e.g. iSCSI) or NVMe optimized protocols (e.g., NVMe/TCP)
- NVMe over Fabrics – supports multiple transports, extends NVMe efficiency over network
  - Poll and interrupt mode architecture
  - Kernel and user mode implementations
RocksDB Key-Value Store - Overview

• Type of NoSQL database that uses simple key/value pair mechanism to store data
• Alternative to limitations of traditional relational databases
  - Data structured and schema pre-defined
  - Mismatch with today’s workloads
  - Data growth in large and unstructured
  - Lots of random writes and reads
• NoSQL brings flexibility as application has complete control over what is stored inside the value
RocksDB Key-Value Store - Overview

• Key in a key-value pair must be unique
• Values identified via a key can be numbers, strings, images, videos etc.
• Common API operations: get(key) for reading data, put(key, value) for writing data and delete(key) for deleting keys
• **Phone Directory** database example:

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>(123) 456-7890</td>
</tr>
<tr>
<td>Kyle</td>
<td>(245) 675-8888</td>
</tr>
<tr>
<td>Richard</td>
<td>(787) 122-2212</td>
</tr>
</tbody>
</table>
Test Configuration - Hardware

**RocksDB TiogaPass Server**
- 2S Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz, 20 Cores, 40 Threads (27.5MB L3 Cache)
- 192 GB (12x 16GB, 1DPC) DDR4 2666
- Mellanox MT27710 ConnectX-4 Lx x8 PCIe NIC 25Gbps NUMA Node 0

**TiogaPass Head Node**
- 2S Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz, 20 Cores, 40 Threads (27.5MB L3 Cache)
- 192 GB (12x 16GB, 1DPC) DDR4 2666
- Mellanox MT27710 ConnectX-4 Lx x8 PCIe NIC 25Gbps NUMA Node 0

**Lightning JBOF**
- 15x INTEL® SSD DC P3500 (2.5” SFF) x4 PCIe 1.8TB
- All SSDs attached to Tioga Pass NUMA Node 0
## Test Configuration - Software

### Operating System

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Distro</strong></td>
<td>Fedora 27</td>
</tr>
<tr>
<td><strong>Kernel</strong></td>
<td>5.0.0-rc4</td>
</tr>
<tr>
<td><strong>Arch</strong></td>
<td>x86_64</td>
</tr>
</tbody>
</table>

**Tuning:**
- XFS filesystem, agcount=32, mount with discard
- CPU Profile: Performance
- NIC MTU: 9000
- Huge Pages: Turned off

---

### RocksDB

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Version</strong></td>
<td>Master with commit 301da345aed32577da649ffdcea0f3b5e2fe979f</td>
</tr>
<tr>
<td><strong>Record Size</strong></td>
<td>Key - 16B, Value - 100B</td>
</tr>
<tr>
<td><strong>Database Size</strong></td>
<td>456 GB, 4 Billion keys</td>
</tr>
<tr>
<td><strong>RocksDB Instances</strong></td>
<td>Upto 9 (1 SSD per 3 instances)</td>
</tr>
<tr>
<td><strong>Read/Write Dataset</strong></td>
<td>5 million records</td>
</tr>
<tr>
<td></td>
<td>- Dataset size higher (&gt; 3:1 DRAM size)</td>
</tr>
<tr>
<td></td>
<td>- Compression Off</td>
</tr>
<tr>
<td><strong>Testing Tool</strong></td>
<td>db_bench</td>
</tr>
<tr>
<td><strong>Block Size</strong></td>
<td>8KB, Block Cache: 16GB</td>
</tr>
<tr>
<td><strong>Threads</strong></td>
<td>32 (for fill), 16 (for randrw &amp; randr), 1 (randw)</td>
</tr>
<tr>
<td><strong>Database &amp; Write-Ahead-Log</strong></td>
<td>co-located on the same drive</td>
</tr>
<tr>
<td><strong>Jemalloc memory allocator</strong></td>
<td>Direct IO for flush_and_compaction, reads</td>
</tr>
</tbody>
</table>

**NOTE:** see back up for detailed config
Test Methodology

Disaggregation Modes
1. Local NVMe SSD
2. iSCSI
3. NVMe/TCP

Scenarios
1. **Bulk Load** of 4 billion keys in sequential order (32 threads, compression off, Write-Ahead-Log disabled)
2. **Random Write** of 5 million keys (threads=1, Write-Ahead-Log enabled)
3. **Random Read** of 5 million keys (threads = 16)
4. **Multi-threaded Read & Single-threaded Write** of 5 million reads during updates (16 read threads, 1 write thread rate limited at 2MBps)

Test Execution
1. Drop page cache
2. Start system metrics collection
3. Run db bench (modified benchmark.sh)
4. Stop system metrics collection
Performance Comparison: Bulk Load

32 threads per RocksDB instance

**Avg Latency (μs) of 9 RocksDB Instances**

- iSCSI: 1.17
- NVMe TCP/IP: 1.17
- Local: 1.17

**Avg Throughput (Kops/s) of 9 RocksDB Instances**

- iSCSI: 855
- NVMe TCP/IP: 855
- Local: 855

Comparable Performance between local and network attached config (Sequential IO)
Performance Comparison: Random Read

16 threads per RocksDB instance

Minimal performance overhead with NVMe over TCP/IP
Performance Comparison: Random Write
1 thread per RocksDB instance

NVMe over TCP/IP performance is better compared to iSCSI
Performance Comparison: Multi-thread Read and Single-thread Write

16 read threads, 1 rate limited write thread (2MBps) per RocksDB instance

**Avg Latency (µs) of 9 RocksDB Instances**

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Latency (µs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSCSI</td>
<td>1,383</td>
</tr>
<tr>
<td>NVMe TCP/IP</td>
<td>250</td>
</tr>
<tr>
<td>Local</td>
<td>285</td>
</tr>
</tbody>
</table>

**Avg Throughput (Kops/s) of 9 RocksDB Instances**

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Throughput (Kops/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iSCSI</td>
<td>12</td>
</tr>
<tr>
<td>NVMe TCP/IP</td>
<td>64</td>
</tr>
<tr>
<td>Local</td>
<td>56</td>
</tr>
</tbody>
</table>

NVMe over TCP/IP scales better as number of clients increase.
Summary

• Locally attached SSDs result in stranded flash capacity and increased TCO.
• Disaggregating flash storage enables independent scaling of compute and storage resources for cloud workloads.
• NVMe over TCP/IP enables disaggregation of flash storage without requiring changes to networking infrastructure.
• RocksDB using NVMe over TCP/IP delivers scalability while delivering comparable performance to local storage.
Product/Facility Info

OCP TiogaPass 2S Server
https://www.opencompute.org/documents/facebook-k-2s-server-tioga-pass-specification

OCP Lightning NVMe JBOF
https://www.opencompute.org/documents/facebook-lightning-hardware-specification
Call to Action

• Take advantage of flash disaggregated architecture using OCP platforms
• Contribute to open source software to enable optimized flash disaggregation solutions
• Share production experience and best practices in OCP communities
BIOS Setup

Profiles
- CPU Power and Performance Policy: Performance
- Workload Configuration: Balanced
- Memory RAS Configuration: Maximum Performance
- Fan Profile: Performance

Enabled
- Hyper-Threading
- Enhanced Intel SpeedStep® Tech
- Intel® Turbo Boost Technology
- Uncore Frequency Scaling
- Performance P-Limit

Disabled
- Cluster on Die
- Early Snoop
- CPU C States
- Energy Efficient Turbo
Test Setup (Linux OS)

/etc/sysctl.conf
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
net.core.netdev_max_backlog = 250000

/etc/security/limits.conf
* soft nofile 65536
* hard nofile 1048576
* soft nproc 65536
* hard nproc unlimited
* hard memlock unlimited

CPU Profile
echo performance> /sys/devices/system/cpu/cpu{0..n}/cpufreq/scaling_governor

Huge Page
echo never> /sys/kernel/mm/transparent_hugepage/defrag
echo never> /sys/kernel/mm/transparent_hugepage/enabled

Network
ifconfig <eth> mtu 9000
ifconfig <eth> txqueuelen 1000
Test Setup (RocksDB)

Options.error_if_exists: 0
Options.create_if_missing: 0
Options.paranoid_checks: 1
Options.env: 0x56126fe7b240
Options.info_log: 0x561270c35d90
Options.max_file_opening_threads: 16
Options.statistics: (nil)
Options.use fsync: 0
Options.statistics: (nil)
Options.max_log_file_size: 0
Options.max_manifest_file_size: 1073741824
Options.log_file_time_to_roll: 0
Options.allow_fallocate: 1
Options.allow_mmap_reads: 0
Options.allow_mmap_writes: 0
Options.use_direct_reads: 1
Options.create_missing_column_families: 0
Options.db_log_dir:
Options.wal_dir: /mnt/nvme2n1/wal
Options.table_cache_numshardbits: 6
Options.max_subcompactions: 4
Options.max_bgflushes: 7
Options.WAL_size_limit_MB: 0
Options.manifest_preallocation_size: 4194304
Options.is_fd_close_on_exec: 1
Options.advertise_random_on_open: 1
Options.db_write_buffer_size: 0
Options.write_buffer_manager: 0x561270c3de90
Options.access_hint_on_compaction_start: 1
Options.new_table_reader_for_compaction_inputs: 1
Options.random_access_max_buffer_size: 1048576
Options.use_adaptive_mutex: 0
Options.rate_limiter: 0x561270c3de90
Options.sst_file_manager.rate_bytes_per_sec: 0
Options.wal_recovery_mode: 2
Options.enable_thread_tracking: 0
Options.enable_pipelined_write: 1
Options.allow_concurrent_memtable_write: 1
Options.enable_write_thread_adaptive_yield: 1
Options.write_thread_max_yield_usec: 100
Options.write_thread_slow_yield_usec: 3
Options.row_cache: None
Options.wal_filter: None
Options.avoid_flush_during_recovery: 0
Options.allow_ingest_behind: 0
Options.recycle_log_files_num: 0
Options.two_write_queues: 0
Options.manual_wal_flush: 0
Options.max_background_jobs: 8
Options.max_background_compactions: 16
Options.avoid_flush_during_shutdown: 0
Options.writable_file_max_buffer_size: 1048576
Options.delayed_write_rate : 8388608
Options.max_total_wal_size: 17179869184
Options.stats_dump_period_ms: 21600000000
Options.stats_dump_period_sec: 600
Options.open_files: -1
Options.wal_bytes_per_sync: 8388608
Options.wal_bytes_per_sync: 8388608
Options.compression_readahead_size: 0

Compression algorithms supported:
kZSTDNotFinalCompression supported: 0
kZSTD supported: 0
kXpressCompression supported: 0
kLZ4HCCompression supported: 0
kLZ4Compression supported: 0
kBZip2Compression supported: 0
kZlibCompression supported: 1
kSnappyCompression supported: 0
Test Setup (RocksDB)

Fast CRC32 supported: Supported on x86
Options for column family [default]:
  Options.comparator: ledb.ByteWiseComparator
  Options.merge_operator: PutOperator
  Options.comparator_filter: None
  Options.comparator_filter_factory: None
  Options.memtable_factory: SkipListFactory
  Options.table_factory: BlockBasedTable
  table_factory options: flush_block_policy_factory:
    FlushBlockBySizePolicyFactory (0x561270c2cb20)
cache_index_and_filter_blocks: 1
  cache_index_and_filter_blocks_with_high_priority: 0
  pin_0_filter_and_index_blocks_in_cache: 1
  pin_top_level_index_and_filter: 0
  index_type: 0
  hash_index_allow_collision: 1
  checksum: 1
  no_block_cache: 0
  block_cache: 0x561270c2caa0
  block_cache_name: LRUCache
  block_cache_options:
    capacity: 34359738368
    num_shard_bits: 6
    strict_capacity_limit: 0
    memory_allocator: None
    high_pri_pool_ratio: 0.000
    block_cache_compressed: (nil)
  persistent_cache: (nil)
  block_size: 16384
  block_size_deviation: 10
  block赭heart_interval: 16
  index_block赭heart_interval: 1
  metadata_block_size: 4096
  partition_filters: 0
  use_delta_encoding: 1

filter_policy: rocksdb.BuiltinBloomFilter
  whole_key_filtering: 1
  verify_compression: 0
  read_amp_bytes_per_bit: 0
  format_version: 2
  enable_index_compression: 1
  block_align: 0
  Option.write_buffer_size: 134217728
  Options.max_write_buffer_number: 8
  Options.compression: NoCompression
  Options.bottommost_compression: Disabled
  Options.prefix_extractor: nullptr
  Options.memtable_insert_with_hint_prefix_extractor: nullptr
  Options.num_levels: 6
  Options.min_write_buffer_number_to_merge: 1
  Options.max_write_buffer_number_to_maintain: 0
  Options.bottommost_compression_opts.window_bits: -14
  Options.bottommost_compression_opts.level: 32767
  Options.bottommost_compression_opts.strategy: 0
  Options.bottommost_compression_opts.max_dict_bytes: 0
  Options.bottommost_compression_opts.zstd_max_train_bytes: 0
  Options.bottommost_compression_opts.enabled: false
  Options.compression_opts.window_bits: -14
  Options.compression_opts.level: 32767
  Options.compression_opts.strategy: 0
  Options.compression_opts.max_dict_bytes: 0
  Options.compression_opts.zstd_max_train_bytes: 0
  Options.compression_opts.enabled: false
  Options.level0_file_num_compaction_trigger: 4
  Options.level0_slowdown_writes_trigger: 20
  Options.level0_stop_writes_trigger: 20
  Options.target_file_size_base: 134217728
  Options.target_file_size_multiplier: 1
  Options.max_bytes_for_level_base: 1073741824
  Options.max_bytes_for_level_base: 8.000000
## Test Setup (RocksDB)

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[0]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[1]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[2]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[3]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[4]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[5]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_bytes_for_level_multiplier_addtl[6]</td>
<td>1</td>
</tr>
<tr>
<td>Options.max_sequential_skip_in_iterations</td>
<td>8</td>
</tr>
<tr>
<td>Options.max_compaction_bytes</td>
<td>3355443200</td>
</tr>
<tr>
<td>Options.arena_block_size</td>
<td>16777216</td>
</tr>
<tr>
<td>Options.soft_pending_compaction_bytes_limit</td>
<td>0</td>
</tr>
<tr>
<td>Options.hard_pending_compaction_bytes_limit</td>
<td>0</td>
</tr>
<tr>
<td>Options.rate_limit_delay_maxMilliseconds</td>
<td>1000000</td>
</tr>
<tr>
<td>Options.disable_auto_compactions</td>
<td>0</td>
</tr>
<tr>
<td>Options.compaction_style</td>
<td>kCompactionStyleLevel</td>
</tr>
<tr>
<td>Options.compaction_pri</td>
<td>kMinOverlappingRatio</td>
</tr>
<tr>
<td>Options.compaction_options_universal.size_ratio</td>
<td>1</td>
</tr>
<tr>
<td>Options.compaction_options_universal.min_merge_width</td>
<td>2</td>
</tr>
<tr>
<td>Options.compaction_options_universal.max_merge_width</td>
<td>4294967295</td>
</tr>
<tr>
<td>Options.compaction_options_universal.max_size_amplification_percent</td>
<td>200</td>
</tr>
<tr>
<td>Options.compaction_options_universal.compression_size_percent</td>
<td>-1</td>
</tr>
<tr>
<td>Options.compaction_options_universal.stop_style</td>
<td>kCompactionStopStyleTotalSize</td>
</tr>
<tr>
<td>Options.compaction_options_fifo.max_table_files_size</td>
<td>0</td>
</tr>
<tr>
<td>Options.compaction_options_fifo.allow_compaction</td>
<td>1</td>
</tr>
<tr>
<td>Options.compaction_options_fifo.ttl</td>
<td>0</td>
</tr>
<tr>
<td>Options.inplace_update_support</td>
<td>0</td>
</tr>
<tr>
<td>Options.inplace_update_num_locks</td>
<td>10000</td>
</tr>
<tr>
<td>Options.memtable_prefix_bloom_size_ratio</td>
<td>0.000000</td>
</tr>
<tr>
<td>Options.bloom_locality</td>
<td>0</td>
</tr>
<tr>
<td>Options.max_successive_merges</td>
<td>0</td>
</tr>
<tr>
<td>Options.optimize_filters_for_hits</td>
<td>1</td>
</tr>
<tr>
<td>Options.paranoid_file_checks</td>
<td>0</td>
</tr>
<tr>
<td>Options.force_consistency_checks</td>
<td>0</td>
</tr>
<tr>
<td>Options.report_bg_io_stats</td>
<td>0</td>
</tr>
<tr>
<td>Options.ttl</td>
<td>0</td>
</tr>
</tbody>
</table>