Performance Benefits of Running RocksDB on Samsung NVMe SSDs

A Detailed Analysis
Executive Summary

The industry has been experiencing an exponential data explosion over the past decade, a trend that we see continuing in the coming years, moving us into something that we call a Data Deluge. As a result of this Data Deluge, huge amounts of unstructured data are ending up in massive data centers and cloud storage servers across the world. To store so much big data and process it at blinding speed, data center architects are pursuing storage solutions with very low latency and small, power-optimized form factors.

Hard disk drives (HDDs) cannot meet the necessary performance requirements for big data applications. Unfortunately, many solid state drives are also constrained by the maximum throughput limitation of a SATA interface. Recognizing the difficulty of the challenge, and to take advantage of the full speed of NAND flash memory, the industry is now progressing towards a more scalable enterprise interface known as NVMe.

With NVMe SSDs gaining considerable market traction, typical technical specifications on performance do not make it easy for customers to decide between SATA SSDs and NVMe SSDs. Very few benchmark studies have been done that compare real world applications, such as database workloads, running on NVMe SSDs versus SATA SSDs.

This whitepaper addresses that void. This initiative uses a persistent key-value store, RocksDB, on both NVMe SSDs (with and without multi-streaming) and SATA SSD to provide a comparative performance analysis for different workloads.

Using Samsung NVMe SSDs, the essential findings are:

- RocksDB is **3.8x faster** compared to SATA SSDs for 20% read/80% update workload.
- Average read latency decreased by **73%** for 80% read/20% update workload.
- 99th percentile read latency decreased by **78%** for 80% read/20% update workload.

Using Samsung NVMe multi-stream SSDs, the performance and endurance benefits are even better. The essential findings are:

- RocksDB is **4.4x faster** compared to SATA SSDs for 20% read/80% update workload.
- **SSD lifetime increased by 54%** (compared to NVMe SSDs without multi-streaming) with a 20% read/80% update workload.

Please note that these benchmark results were obtained without significant tweaking or optimization and, thus, are very achievable.
RocksDB Overview

RocksDB is an open-source, embeddable, persistent key-value store for fast storage. Developed by Facebook, it is based on the LevelDB database. RocksDB uses a Log Structured Database Engine for storage and features very flexible, tunable configuration settings that allow it to operate in a variety of production environments, including pure DRAM memory and Flash, using standard or distributed file systems such as HDFS.

In operation, inserts are fast because new data goes into an in-memory data-structure, `memTable`, and optionally into `logfile` – a persistent storage write-ahead commit log. This is later flushed to `sstfile` (in the L0 level described below) on non-volatile storage when the `memTable` is full. The entire database is stored in a set of `sstfiles` and arranged at multiple levels (the L0–L6 levels described below). Each `sstfile` contains data blocks that are sorted by key and stores metadata and indexing information.

To reduce read and space amplification, concurrent compactions at different levels remove duplicates and overwritten data. But, compaction increases the stored data's write amplification.

Write amplification in SSDs

The minimum SSD read granularity is a page. Though writes have a page granularity level, they may entail an additional erase cycle. Such erasures happen only at the erase block granularity level. As an SSD fills with data, a garbage collection process needs to copy valid page data from fragmented blocks before erasing the blocks to make them available for further writes. Here, fragmented means media blocks have areas with invalidated data intermixed with areas containing valid data. Copying valid page data leads to an increase in write amplification which effectively increases the time for actual data storage activities.

Introducing the NVMe SSD Multi-Stream Feature

Preventing garbage collection during write operations translates to highly efficient SSD processing usage. Samsung PM953S NVMe SSD multi-stream feature does exactly that. Multi-stream reduces valid data fragmentation within blocks by grouping data with the same lifetime within the same erase block.

In doing so, host applications provide data characteristic hints such as expected data-lifetime. An application could first select a stream ID to associate data in data-write requests, allowing the SSD to store data with the same stream ID in a same erase block. This can significantly decrease garbage collection activity, thereby reducing latency - translating to an IOPS throughput increase. Reducing the associated write amplification extends SSD NAND Flash media lifetime as well.

Leveraging RocksDB’s SSTFile for Multi-Streaming
RocksDB uses a log-structured merge process. Its sstfile has multiple levels (L0 – LN), with each lower level (higher numbered) retaining increasingly colder data than its predecessor. This arrangement allows each sstfile level to have a unique assigned stream ID, differentiating consistently accessed/merged/deleted sstfile data (say, in Level L0) from colder sstfile data which is not accessed/merged/deleted as frequently (say, in Level L6).

**Benchmark Environment**

The selected server was a Dell PowerEdge R730xd with dual Intel Xeon E5-2640v3 running at 2.60 GHz and 64GB of memory. In total, the server provided 16 physical cores. With hyper-threading enabled, the total logical CPU count was 32. We kept a 1:4 *physical core to main memory* ratio (16 cores: 64GB memory).

For the SSD, we used the power-optimized, M.2 form factor Samsung PM953 and PM953S NVMe SSD which is targeted for data center and cloud service providers. That said, Samsung PM1725 NVMe SSD continues to be the flagship product within Samsung’s NVMe SSD portfolio. In terms of documented performance, the Samsung PM1725 is 4.5x faster with 4K random reads, 5.5x faster with 4K random writes than the Samsung PM953. Therefore, PM1725 performance is much higher.

<table>
<thead>
<tr>
<th>Processor/Memory Details</th>
<th>Operating System</th>
<th>SSD Details</th>
</tr>
</thead>
</table>
| **Processor Dual Socket:** Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz. | **Distro**: Ubuntu 14.04.1 LTS  
**Kernel**: 3.19.0-11-generic, patched for multi-stream support  
**Arch**: x86_64 | **SSD**:  
1 x Samsung NVMe PM953 M.2 480GB SSD  
1 x Samsung NVMe PM953S M.2 480GB SSD  
1 x professional grade commercial SATA 480 GB SSD |
| **Total Logical CPU**: 32  
**Total memory**: 64 GB | | |
The difference between PM953 and PM953S is the multi-stream feature. i.e., PM953S supports multi-stream feature while PM953 does not. For a baseline comparison, the selected SATA SSD is a professional-grade, commercially available, SSD.

<table>
<thead>
<tr>
<th>Software</th>
<th>Functionality</th>
<th>Version/Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>RocksDB</td>
<td>Persistent Embedded Key Value Store</td>
<td>3.11 (Modified to add multi-stream support)</td>
</tr>
<tr>
<td>YCSB</td>
<td>Yahoo Cloud Benchmark Tool</td>
<td>0.1.4</td>
</tr>
<tr>
<td>SSDB-Rocks</td>
<td>Provides an interface to RocksDB for YCSB</td>
<td>1.6.6</td>
</tr>
</tbody>
</table>

YCSB and SSDB have been set up to run on the same server. The dataset was initially created by performing 370 Million, 1K key-value record inserts. The same dataset was used for all benchmark runs. To get the SSD and the database into a stable steady-state, we performed preconditioning consisting of 20% read/80% update for 2 Billion operations. Benchmark workloads ran after this pre-conditioning.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of Operations</td>
<td>500 Million</td>
</tr>
<tr>
<td>No of YCSB Threads</td>
<td>32</td>
</tr>
</tbody>
</table>

With data centers already embracing super-fast SSDs, the world is ready to replace and, in some cases, complement rotational HDDs in many data-centers. Therefore, we didn't focus on providing an in-depth performance benefit analysis of migrating from HDD to SATA SSDs or NVMe SSDs.

We used metrics such as throughput, average read and update latency for comparing a commercially available SATA SSD, the Samsung PM953 NVMe SSD without multi-stream, and the same drive with multi-stream. Since actual NAND write operation data was unavailable on the SATA SSD, we only compared NVMe SSD Write Amplification Factors.
**Benchmark Results for 20% Read/80% Update**

The benchmark workload of 20% read and 80% update was performed for an operation count of 500 Million (i.e., 100 Million reads and 400 Million updates.)

![Figure 1. 20/80 Workload Throughput and WAF Performance Benefit](image)

The throughput increase with the Samsung PM953S NVMe SSD (with multi-stream) was 4.4x compared to the SATA SSD. (i.e., the benchmark runtime decreased by more than 77.2% when the Samsung NVMe SSD with multi-stream support was used.) The PM953S NVMe SSD multi-stream feature increased the throughput performance by 15.45% over the PM953 NVMe SSD (without multi-stream feature).

Clearly, multi-streaming has a profound effect on the Write Amplification factor. It reduced extra data writes by 93%, thereby increasing the lifetime of the SSD by 54%.
The average read latency was reduced by 79% with a multi-stream-enabled PM953S NVMe SSD and 76% with an NVMe SSD when compared with a SATA SSD. The average update latency was reduced by 51% with a multi-stream-enabled PM953S NVMe SSD and 42% with a PM953 NVMe SSD when compared with a SATA SSD.

The 99th percentile read and update latency was reduced by 70% and 54% with a multi-stream-enabled PM953S NVMe SSD when compared with a SATA SSD.

**Benchmark Results for 50% Read/50% Update**

The benchmark workload of 50% read and 50% update were performed for an operation count of 500 Million (i.e., 250 Million reads and 250 Million updates.)

The throughput increase with the Samsung PM953S NVMe SSD (with multistreaming) was 4x compared to the SATA SSD (i.e., the benchmark runtime decreased by more than 75% when using the Samsung PM953S NVMe SSD with multi-stream support. The PM953S NVMe SSD multi-stream feature increased throughput performance by 4.6% over the PM953 NVMe SSD (without multistreaming).

With a multi-stream-enabled SSD, extra writes decreased by 91%, thereby extending the SSD’s lifetime by 39%.
Figure 3. 50/50 Workload Throughput and WAF Performance Benefit

Compared with the SATA SSD, the average read latency decreased 76% with the multi-stream-enabled PM953S NVMe SSD and 75% with the PM953 NVMe SSD. The average update latency decreased 37%.
with multi-stream-enabled PM953S NVMe SSD and 30% with PM953 NVMe SSD in comparison to a SATA SSD.

The 99th percentile read and update latency was reduced by 95% and 47% respectively with multi-stream-enabled PM953S NVMe SSD when compared with a SATA SSD.

**Benchmark Results for 80% Read/20% Update**

The benchmark workload of 80% read and 20% update was performed for an operation count of 500 million (i.e., 400 million reads and 100 million updates.)

![Figure 5. 80/20 Workload Throughput and WAF Performance Benefit](image)

The Samsung PM953S NVMe SSD (with multi-stream) throughput increases 3.74x compared to SATA SSD throughput (i.e., the benchmark runtime decreased more than 73% using the Samsung PM953S NVMe SSD with multi-stream support).

In addition, the Samsung NVMe multi-stream-enabled SSD reduced extra writes by 90%, thereby extending the lifetime of the SSD by 27%.
Figure 6. 80/20 Workload Average Read and Update Latency Performance Benefit

When compared with the SATA SSD, the average read latency of the multi-stream-enabled PM953S NVMe and PM953 NVMe SSDs decreased by 74% and 73%, respectively. In comparison with an SATA SSD, the average update latency for the multi-stream-enabled PM953S NVMe SSD decreased by 52%, and by 51% for PM953 NVMe SSD.

The 99th percentile read and update latencies were reduced by 64% and 53% respectively, with a multi-stream-enabled PM953S NVMe SSD when compared to a SATA SSD.

**Conclusion**

These benchmark results were obtained without significant tweaking or optimization and, therefore are very achievable. I/O stack optimization opportunities still exist, which could enable higher performance, inching towards the maximum NVMe SSD hardware capability. Also, this whitepaper did not explore the power reduction benefits of the power-optimized Samsung PM953 (M.2 form factor).

Benchmark results show that using Samsung PM953S NVMe multi-Stream SSD over SATA SSD in this application configuration provides:
- 4.4x overall IOPS throughput improvement
- 54% SSD lifetime increase (compared to PM953 NVMe SSDs)
- 79% read latency reduction

This throughput increase and runtime reduction clearly indicate that, to deliver better performance, real world applications such as RocksDB can exploit faster NVMe devices that provide storage intelligence like the T10 standardized multi-stream feature. Using such devices simultaneously sharply extend an SSD’s lifetime.