A Real-Time, Low Latency, Key-Value Solution Combining Samsung Z-SSD™ and Levyx's Helium™ Data Store

January 2018

SAMSUNG

Levyx

A Real-Time, Low Latency, Key-Value Solution Combining Samsung Z-SSD™ and Levyx's Helium™ Data Store

Summary

In this whitepaper, we present a high performance key-value store solution that can be used in a wide variety of latency sensitive applications tied to the rapidly expanding era of hyper-speed "big data." By combining Levyx's high-performance Helium™ key-value store with Samsung's ultra-low latency Z-SSD™, we have demonstrated performance improvements of up to 10X over the conventional approaches of processing large-scale datasets.

Introduction

Over the past decade, many "big data" application requirements have evolved from those of a batch-oriented nature, to being increasingly time sensitive and now all information is demanded in real time. Many real-time operations and corresponding data-intensive analytics have fanned out from their Social Media origins to encompass a broad range of areas including Internet of Things (IoT), financial trading, manufacturing, e-commerce, cybersecurity, utility grids, life sciences, automotive and more. While existing big data architectures have been able to keep up with overall IO throughput requirements by simply scaling out (by adding more CPUs, nodes, servers, SSDs, etc.), they are increasingly challenged by workloads where very low latency is especially critical.

Samsung's Z-SSDs deliver a dramatic improvement in latency—making them strikingly efficient solutions for large-scale, real-time environments. In particular, opportunities centered on IoT are gaining momentum, and today, the number of intelligent interconnected devices is far outstripping the amount of associated human interaction. Managing latency in these environments is particularly challenging as machine-to-machine speed demands are beginning to drive latency and jitter requirements from social media-acceptable levels of 10-100s of milliseconds down into 10s of microseconds, and even lower.

The financial sector also requires low-latency solutions given that financial institutions are now predominantly automated and algorithmically-driven, where trades are executed in 100s of nanoseconds. The need for low-latency applied to large-scale data exists today in financial functions such as the back-testing of trading algorithms, compliance verification, risk analysis, and fraud detection. These capabilities must keep pace with the overall speed of transaction-intensive data mining and real-time analytics required in the world of finance.

Technology

Samsung, the leading supplier of flash-based technologies, has developed an extensive portfolio of flash storage products that effectively address the increasing density, throughput and performance requirements of today's big data applications.

Lightweight protocols like NVMe can leverage the parallelism of NAND technology to deliver one million IOPS and 3GB/s throughput from a single storage device. At the same time, network hardware can deliver 100Gbps throughput and dramatically reduce latency, while challenging assumptions about data locality and disaggregated storage.

Samsung has made major hardware advances to accommodate real-time, data-intensive applications to the extent that the conventional (non-optimized) software stack has become the single greatest source of latency and non-deterministic system performance. Levyx set out to resolve this and has re-engineered the data stack to fully extract the benefits of this new hardware, making the pairing of its Helium key-value store with Samsung's Z-SSDs a compelling, low-latency and high-performance flash-based solution.

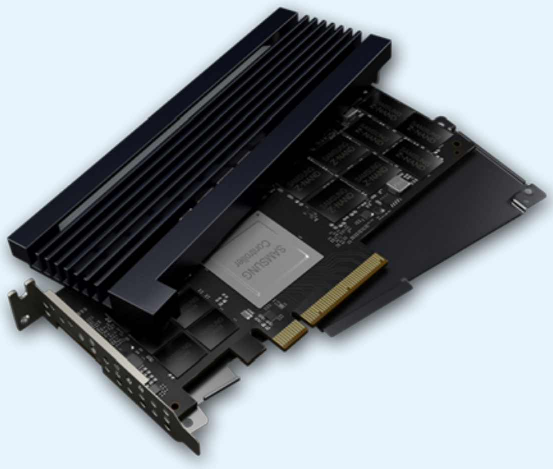
Samsung Z-SSD Ultra-low Latency Storage

One of Samsung's biggest breakthroughs over the past 18 months has been its development of a high performance, ultra-low latency solid state storage solution, the Z-SSD. This new generation of low-latency storage shares the fundamental structure of V-NAND, the industry-leading 3D flash production technology. Z-SSD also offers a unique circuit design and controller with which to maximize performance, offering four times faster latency and 1.6 times better sequential reading than the Samsung PM963 NVMe SSD.

The first low-latency solid state storage device based on Samsung's Z-NAND technology—the 800GB SZ985 provides an incredible 5.5 times lower latency than today's leading NVMe SSDs. Samsung's SZ985 has been designed with proven NAND technology for a high degree of reliability, exceptional scalability and an significantly improved total cost of ownership.

Z-SSD is positioned as a storage device that sits between DRAM and NVMe SSD. It delivers lower latency than any SSD or HDD in the market today. This is an optimal solution for systems designers looking to strike the ideal balance between performance and cost for their target applications.

Z-SSD™



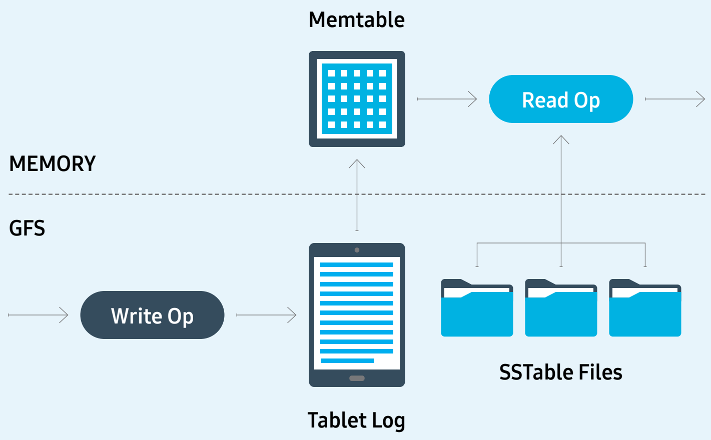
Samsung SZ985 Z-SSD	
Form Factor	HHHL
Interface	PCIe Gen3 x4
NAND	Z-NAND™ Technology
Port	Single
Data Transfer Rate (128KB data size, QD = 32)	
Seq. R/W (GB/s)	3.2 / 3.0
Data I/O Speed (4KB data size, sustained, QD = 32)	
Ran. R/W (IOPs)	750K / 170K
Latency (sustained random workload, QD = 1)	
Ran. R (Typical)	20μs
Ran. R (Best)	12μs
Ran. W (Typical)	16μs
DWPD	30
Capacity	800GB

Figure 1: Samsung SZ985 Technical Specifications

A Real-Time, Low Latency, Key-Value Solution Combining Samsung Z-SSD™ and Levyx's Helium™ Data Store

Helium, a Highly Efficient Key-Value Store from Levyx

Conventional big data architectures have yet to fully exploit the latest enterprise hardware advances. In fact, the vast majority of big data architectures and databases in use today are based on 10-year-old concepts built for commodity x86 hardware that was first available when the Google Big Table paper was published in 2006.



Mid-range Server (2006)	
Processor	2GHz
Cores	2
Storage	400GB HDD
RAM	8GB

Mid-range Server (2017)	
Processor	2.4GHz
Cores	44
Storage	16TB (4 X 4TB PCIe NVMe SSDS)
RAM	128GB

Figure 2: (from the Google Big Table Whitepaper, *circa* 2006)

At that time, the best approach to improve NoSQL database performance was to use a Log Structure Merge design to make the best possible use of HDDs, RAM and a limited number of cores per server. Besides scaling out, these architectures would cache in memory, maintain "write ahead" logs, and serialize random writes into sequential write logs. They would then perform asynchronous write compaction cycles on multi-level data structures to free up space and maintain the desired read performance.

Leading open source storage engines such as LevelDB and RocksDB are based on this approach and typically make use of single-threaded write paths as part of their design. However, the resulting compaction cycles and memory cache exhaustion could trigger database stalls and other disruptive behavior that can negatively impact performance. In general, conventional big data architectures have not fully exploited the latest hardware advances in networking and storage.

One new approach may change things dramatically. Levyx has developed a 'key-value store' software engine that unlocks the full potential of high performance and low-latency flash storage for the purpose of accelerating existing databases as well as enabling new classes of approaches to big data that require low-latency and reduced jitter. The new database storage software engine from Levyx— Helium—uses a "scale-in, then scale-out" approach to modern big data architecture to accommodate the latest evolutions in server design.

Commodity servers today are typically 20 times more powerful (cores, IO channels, memory) than their 2006 equivalents, and multi-channel flash solutions have largely displaced HDDs.

To help in sharply reducing latency, Levyx's Helium typically bypasses the OS file system and volume manager to directly interact with an NVMe block device. It also leverages multi-core/hyper-threading and multi-channel flash devices to create highly parallelized read and write paths. This allows Helium's performance to scale to accommodate increases in the size of IO channels and the number of cores. Latency jitter has been greatly reduced through single layer log-structured persistence and by maintaining a continuous background compaction/garbage collection process.

Improved latency is achieved by using more efficient look-ups and a patent-pending in-memory index that requires only 12 bytes of metadata per object. Since the index is sorted, point and range queries have proven to be extremely fast.

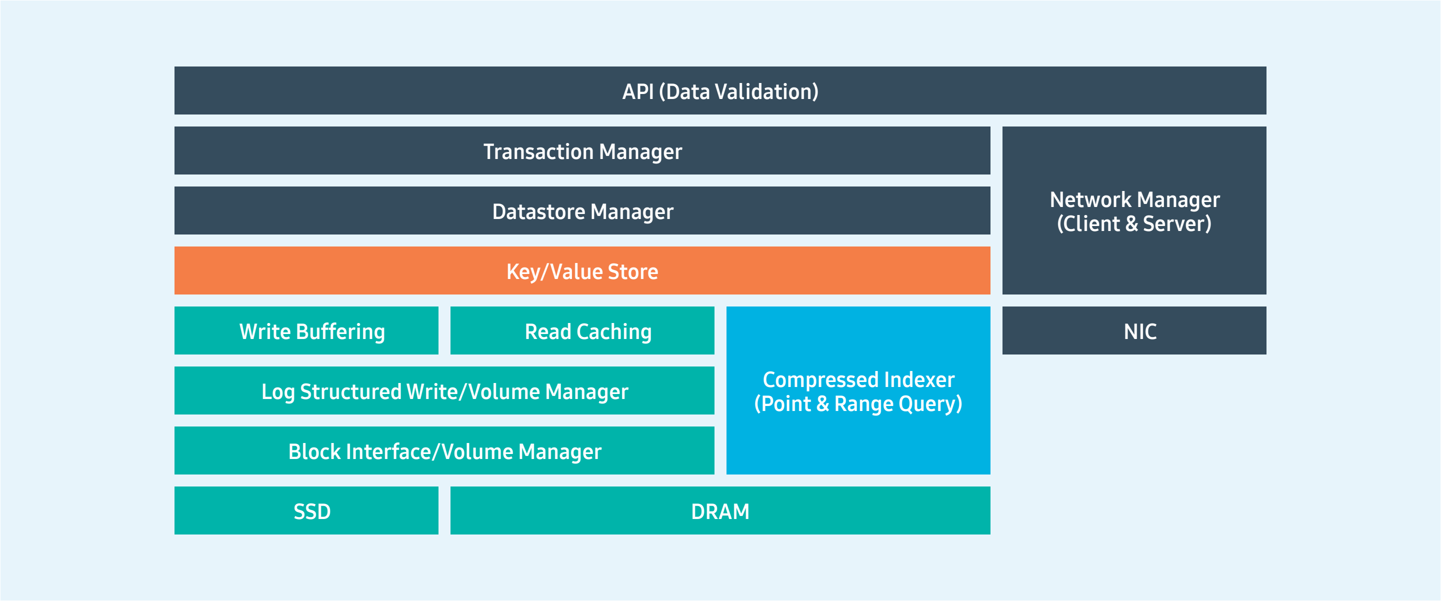

API (Data Validation)
Transaction Manager
Network Manager (Client & Server)
Datastore Manager
Key/Value Store
Write Buffering
Read Caching
Compressed Indexer (Point & Range Query)
NIC
Log Structured Write/Volume Manager
Block Interface/Volume Manager
SSD
DRAM

Figure 3: Helium™ Architecture

Helium can be widely deployed because it offers a simple-to-use API, in addition to supporting other popular key-value storage engine APIs such as RocksDB (HeRocks™ is Levyx's fully compatible version), LevelDB and Memcached. The Helium code has been ported to Linux (x86, Power and ARM), Windows and Mac, and runs in available user space in bare metal, container and cloud configurations. Besides NVMe/PCIe, Helium also can be configured to work with other storage-class memory including SAS/SATA, file systems, block device drivers, volume managers, FC and IB/RDMA.

Performance Evaluation

Our comprehensive performance evaluation compares Levyx's Helium key-value store with RocksDB, a highly-prolific database engine originally developed and open-sourced by Facebook. We used the RocksDB tool: db_bench, as our benchmarking application with Levyx's HeRocks interface. HeRocks essentially becomes a fully RocksDB-compliant data engine, enhanced by Levyx. For our analysis, we compared the performance of Levyx's Helium key-value store to the LSM-based RocksDB using the Z-NAND SSD, and an enterprise-class NVMe SSD—the Samsung PM1725a.

We ran all of our tests on the same Dell PowerEdge R730XD server. The hardware and software specifications of the server are listed in Table 1. We compared RocksDB key-value store and Helium with HeRocks (a RocksDB API compatible wrapper) using db_bench.

Table 1:
Hardware/Software Test Specifications

Server	Dell PowerEdge R730XD
Processor	Per Processor : 22 Cores, 44 Threads Overall : 44 Cores, 88 Threads
Memory	256 GB
Storage	1 x Samsung SZ985 Z-SSD 800GB 1 x Samsung PM1725a NVMe SSD 1.6 TB
OS	CentOS 7.3.1611
Linux Kernel	3.10.0-514.26.2
RocksDB Version	5.2.1
Helium Version	2.12.0

Raw SSD performance was measured using the cloud harmony benchmark, with the results tabulated in Table 2.

Table 2:
Performance comparisons between PM1725a 2.5" NVMe SSD (Baseline) and Z-SSD

	PM1725a 2.5"		Z-SSD	
	Read	Write	Read	Write
Random IO Latency @ 4KB	83.1µs	22.2µs	15.3µs	10.5µs
Random IOPS @ 4KB	743K	140K	808K	151K
Sequential Throughput @128KB	3055MB/s	2150MB/s	3215MB/s	2846MB/s

Z-SSD's random read latency is one-fifth that of the PM1725a, which makes it well suited for latency sensitive applications like key-value stores.

Methodology

The dataset was created by running db_bench *filluniquerandom* workload, which creates a one billion key-value dataset. The db_bench options that were used in these tests are listed below.

```
./db_bench --num_levels=6 --key_size=20 --prefix_size=20 --keys_per_prefix=0 --value_size=400 --cache_size=17179869184 --cache_numshardbits=6
--compression_type=none --compression_ratio=1.0 --min_level_to_compress=-1 -disable_seek_compaction=1 --hard_rate_limit=2 --write_buffer_
size=134217728 --max_write_buffer_number=2 --level0_file_num_compaction_trigger=8 --target_file_size_base=134217728 --max_bytes_for_level_
base=1073741824 --disable_wal=0 --sync=0 --verify_checksum=1 --delete_obsolete_files_period_micros=314572800 --max_background_compactions=16
--max_background_flushes=1 --level0_slowdown_writes_trigger=16 --level0_stop_writes_trigger=24 --statistics=1 --stats_per_interval=1 --stats_
interval=10000000 --histogram=1 --open_files=500000 --memtablerep=prefix_hash --bloom_bits=10 --bloom_locality=1 --num=1000000000 --use_
existing_db=0 --db=PATH_TO_DB  --wal_dir=PATH_TO_WAL
```

The performance evaluation used two workloads that best approximated real-world scenarios.

- **read-while-writing** workload as specified on RocksDB website
- **read-random-write-random** with read and write set to 80% & 20% respectively

Each workload ran for 30 minutes.

Read-while-Writing

The *read-while-writing* workload was run with 32 threads (31 read threads and one write thread). Results are tabulated in Table 3.

Table 3: DB_Bench Read-while-Writing Workload Results

KV-Store	PM1725a		Z-SSD	
	Ops/Sec	Avg Read Latency	Ops/Sec	Avg Read Latency
HeRocks (Levyx)	350K	91.4μs	894K	35.8μs
RocksDB	67K	476.9μs	85K	374.3μs

Observe that HeRocks leverages Z-SSD to deliver a whopping 2.6x increase in Ops/Sec at 1/2.6th average read latency whenever Z-SSD is used as a drop-in replacement for the PM1725a, and potentially more of an increase when compared against other enterprise SSDs.

Further comparing performance when adding the benefits of Z-SSD, we see that the Helium key-value store leverages the low-latency attributes of Z-SSD to deliver a 10x improvement at only 1/10 the latency of RocksDB. Figure 4 shows that HeRocks consistently drove Z-SSD to deliver a read throughput of ~2.5GB/s with 32 threads, while the PM1725a hit just ~1.0GB/s. By scaling the number of db_bench threads (for the PM1725a) to 256, we attained similar Ops/sec as delivered by a Z-SSD at 32 threads but with 7.8x higher latency.

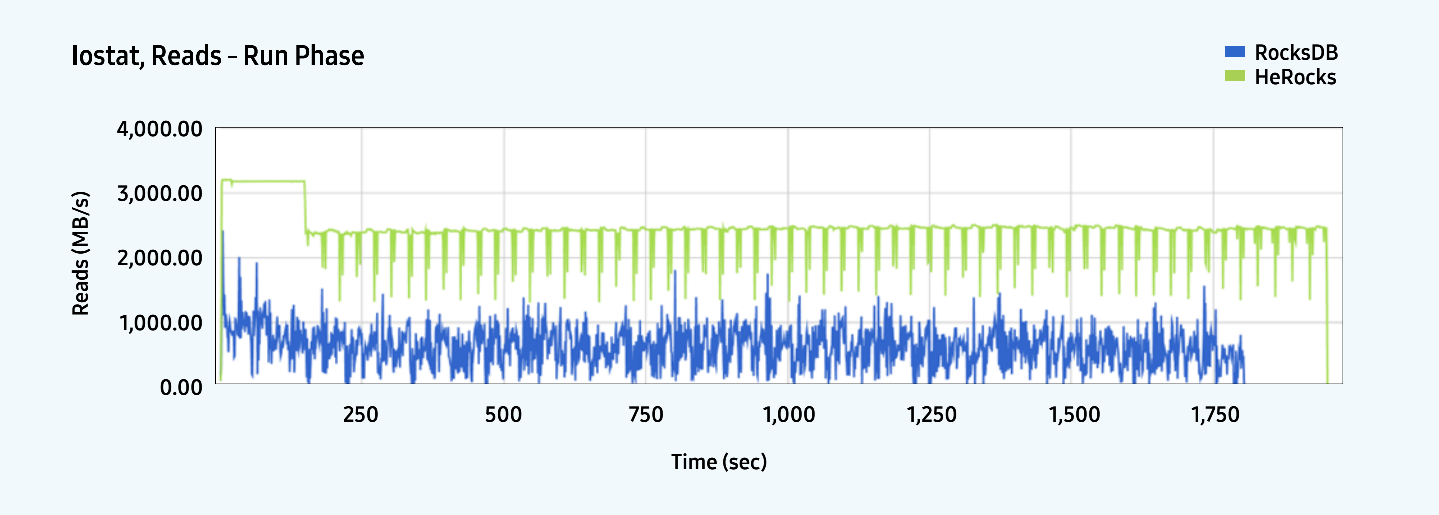


Figure 4: RocksDB vs HeRocks Disk Traffic

Helium's efficiency is again illustrated by CPU utilization measurements shown in Figure 5. This is due to the reduced number of lookups that Helium needs to obtain the requested key and its value.

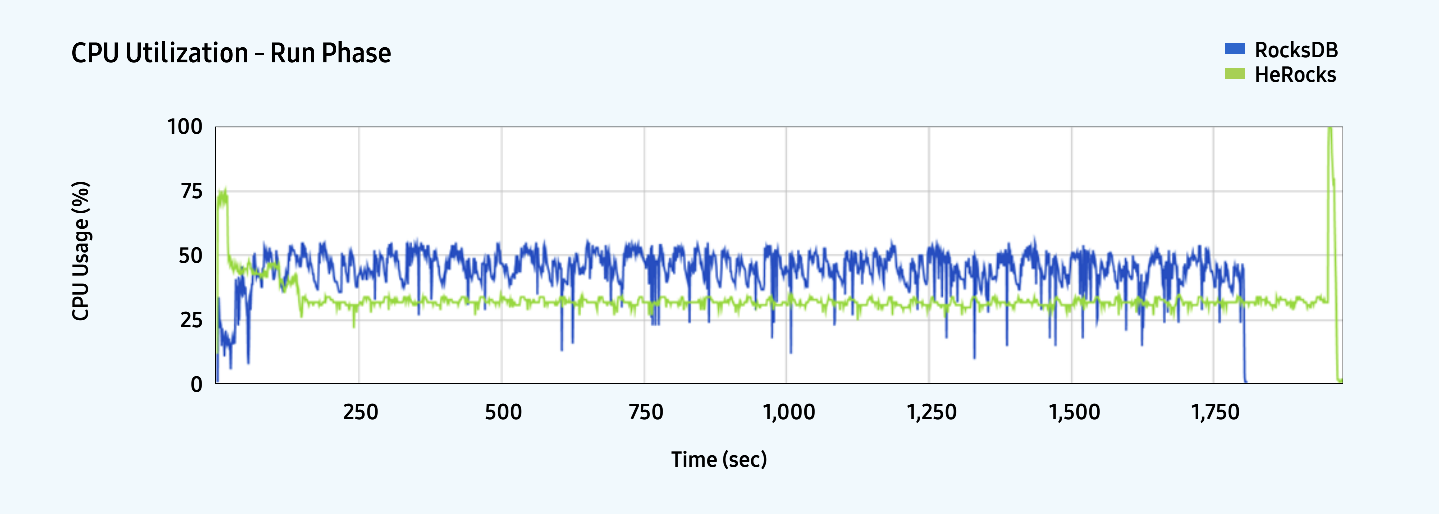


Figure 5: CPU Utilization - RocksDB vs HeRocks

Comparing the read latency distribution with the PM1725a (Figure 6 vs. Figure 7), we found that HeRocks contained two distinct clusters of bars. The 1st range (3-25 μs) can be attributed to cache hits coming from memory, and the 2nd (90-350 μs) attributed to cache misses serviced by the SSD. The IO-Path within Helium was shown to be extremely thin and free from complicated lookups. As a result HeRocks took advantage of fast completions from the memory, followed by a close-to-SSD raw latency response from the storage device for cache misses. RocksDB, on the other hand, had to first look through its SSTFiles at different log structured levels and only then hit the block cache inside the memory. Figure 7 illustrates this, showing that there were no completions in less than 25 microseconds.

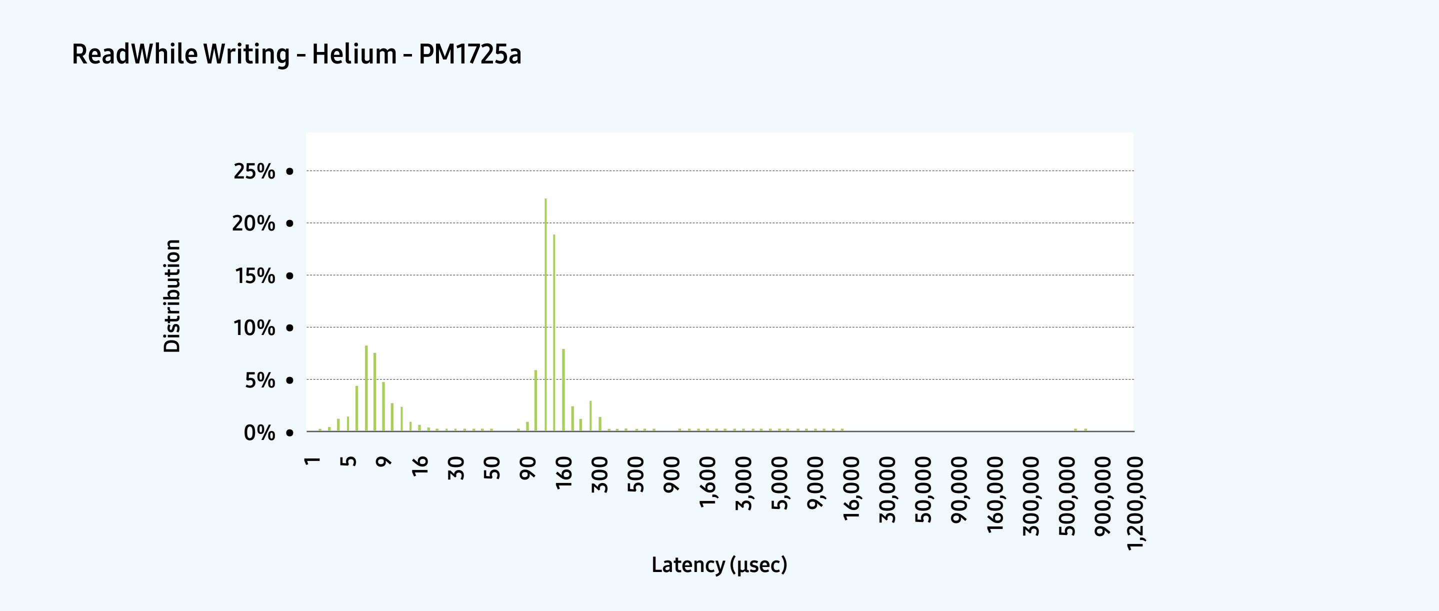


Figure 6: HeRocks Read Latency distribution when data resides in PM1725a

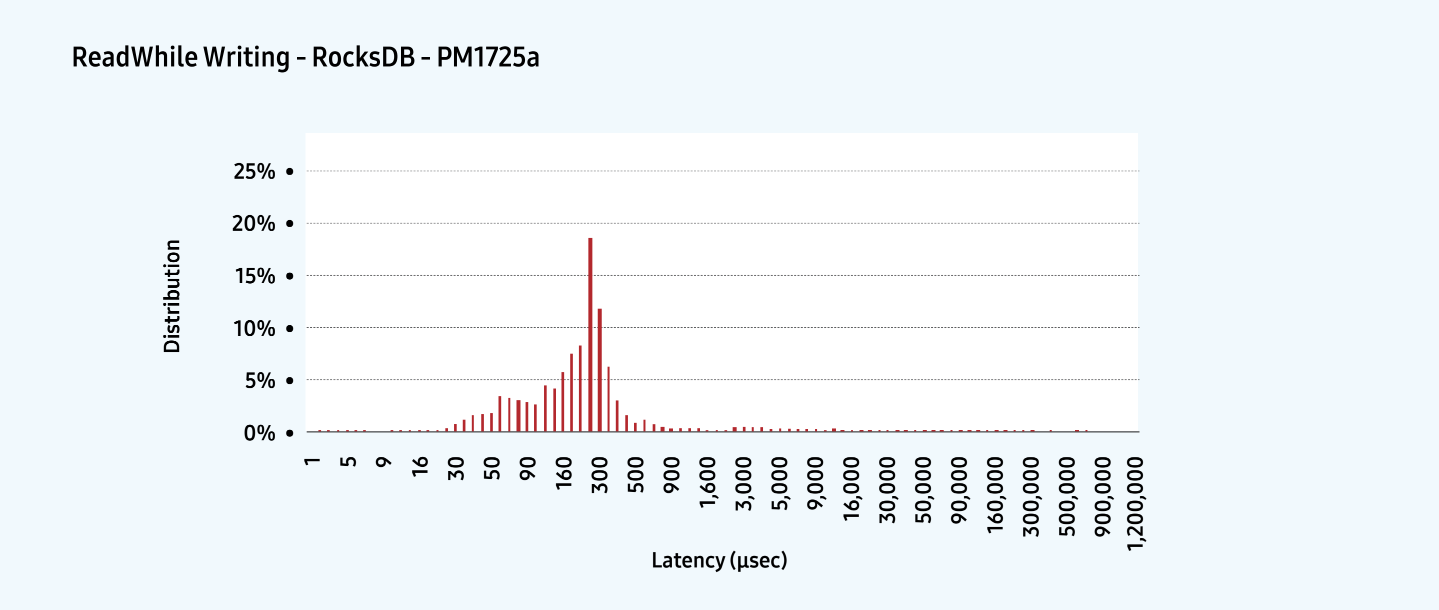


Figure 7: RocksDB Read Latency Distribution when data resides in PM1725a

The Helium architecture is designed to take full advantage of the ultra-low latency response time of the Z-SSD as seen in Figure 8 where the cluster of bars (data serviced from the SSD) clearly shifted to the left to lie within the range of 30 – 120 microseconds.

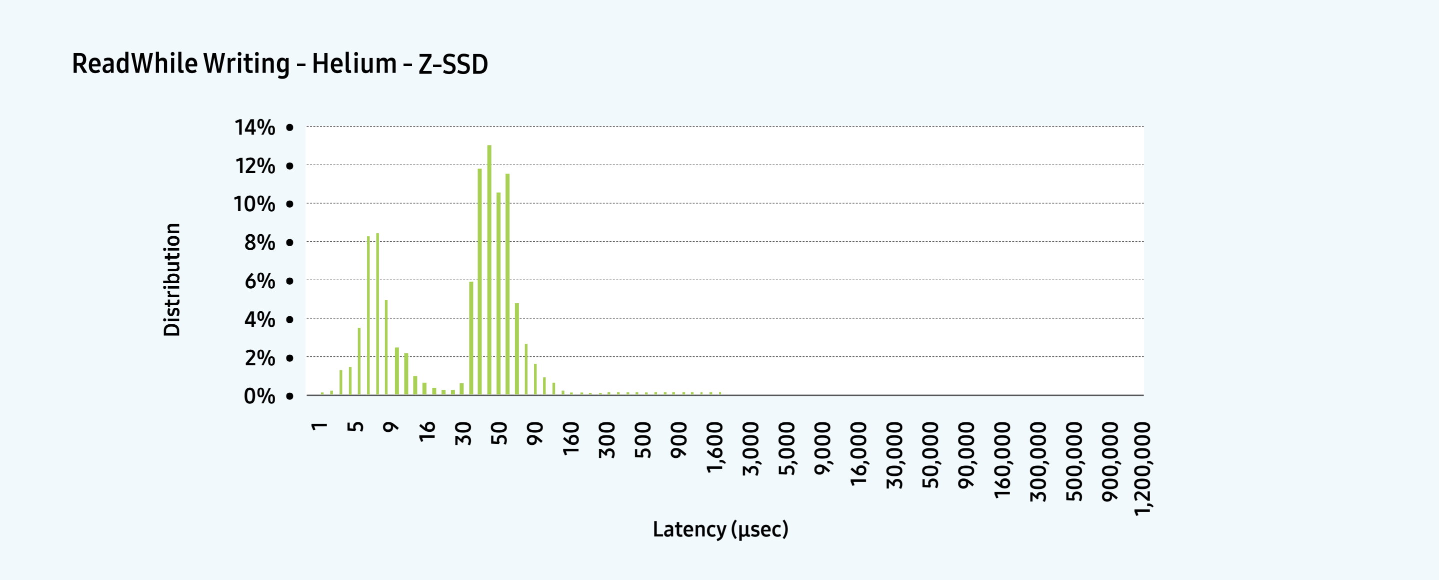


Figure 8: HeRocks Read Latency Distribution when data resides in Z-SSD

Switching to Z-SSD actually benefited both HeRocks and RocksDB. Figure 9 shows that the response from this drive peaked at 120 – 140 microsecond, far better than the 250 – 300 microsecond response time (Figure 7) logged when the PM1725a was used.

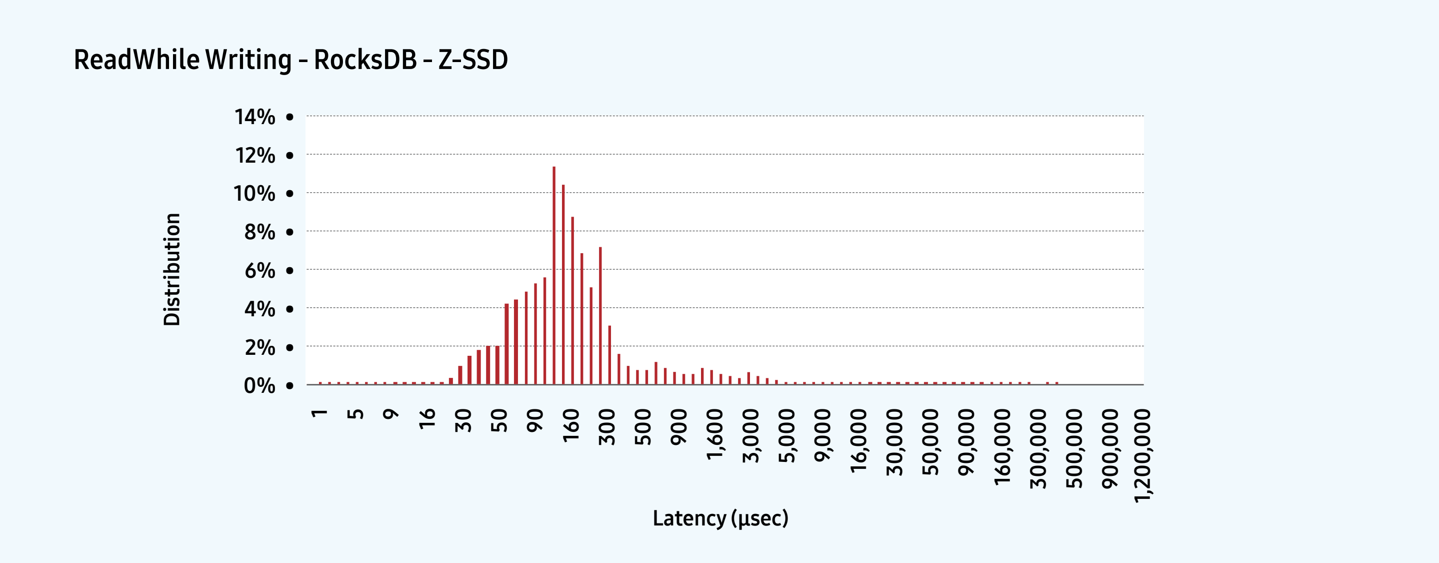


Figure 9: RocksDB Read Latency Distribution when data resides in Z-SSD

Read-Random Write-Random

The *Read-Random Write-Random* workload was run with 16 threads and an 80:20 read-to-write ratio. The results are tabulated in Table 4.

Table 4: DB_Bench Read-Random-Write-Random workload results

KV-Store	PM1725a		Z-SSD	
	Ops/Sec	Avg Read Latency	Ops/Sec	Avg Read Latency
HeRocks (Levyx)	277K	70.3μs	652K	28.7μs
RocksDB	143K	129.7μs	189K	91.4μs

Table 4 shows that HeRocks (with Z-SSD) delivers 3.5x better Ops/sec at one-third the average read latency over RocksDB under similar setup conditions. Comparisons to the popular PM1725a demonstrate the performance advantages of Helium and show how it is able to leverage the low-latency response time of Z-SSDs.

The 'read latency' distribution charts convey a similar story as discussed earlier. The latency when data is read from the disk falls within the 80 – 300 microsecond range (Figure 10) for HeRocks + the PM1725a vs. the 100 – 600 microsecond range (Figure 11) for RocksDB + the PM1725a.

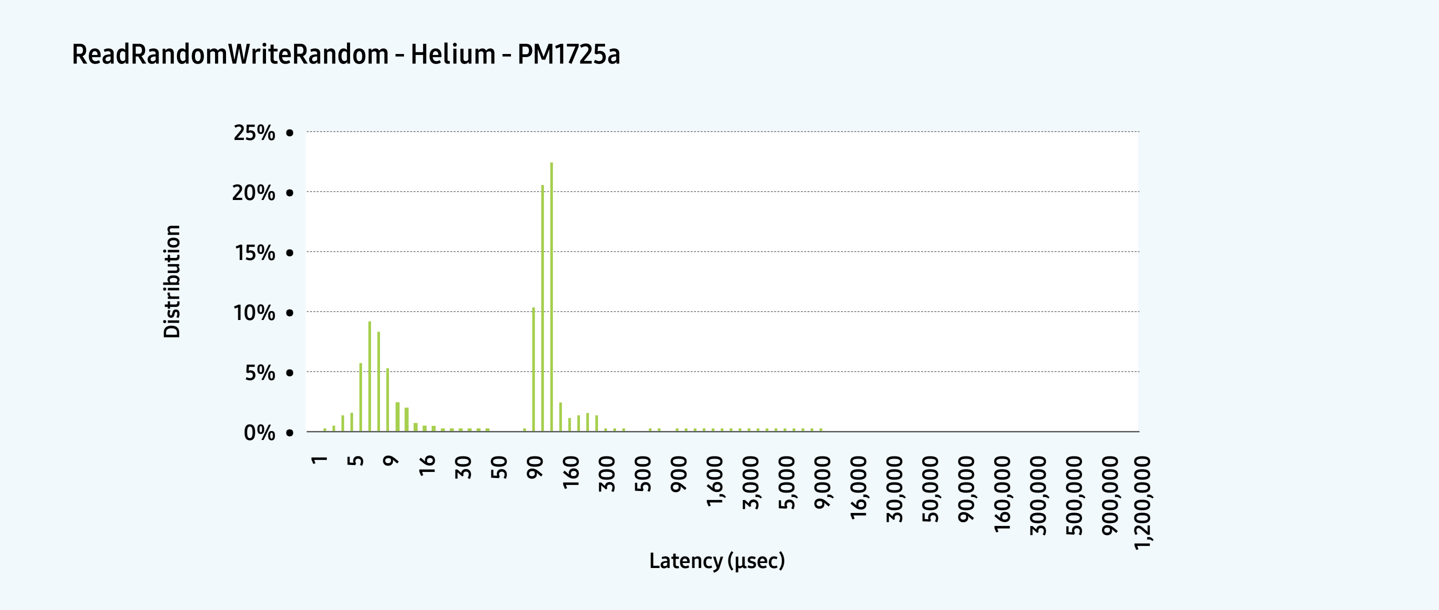


Figure 10: HeRocks Read Latency Distribution when data resides in PM1725a

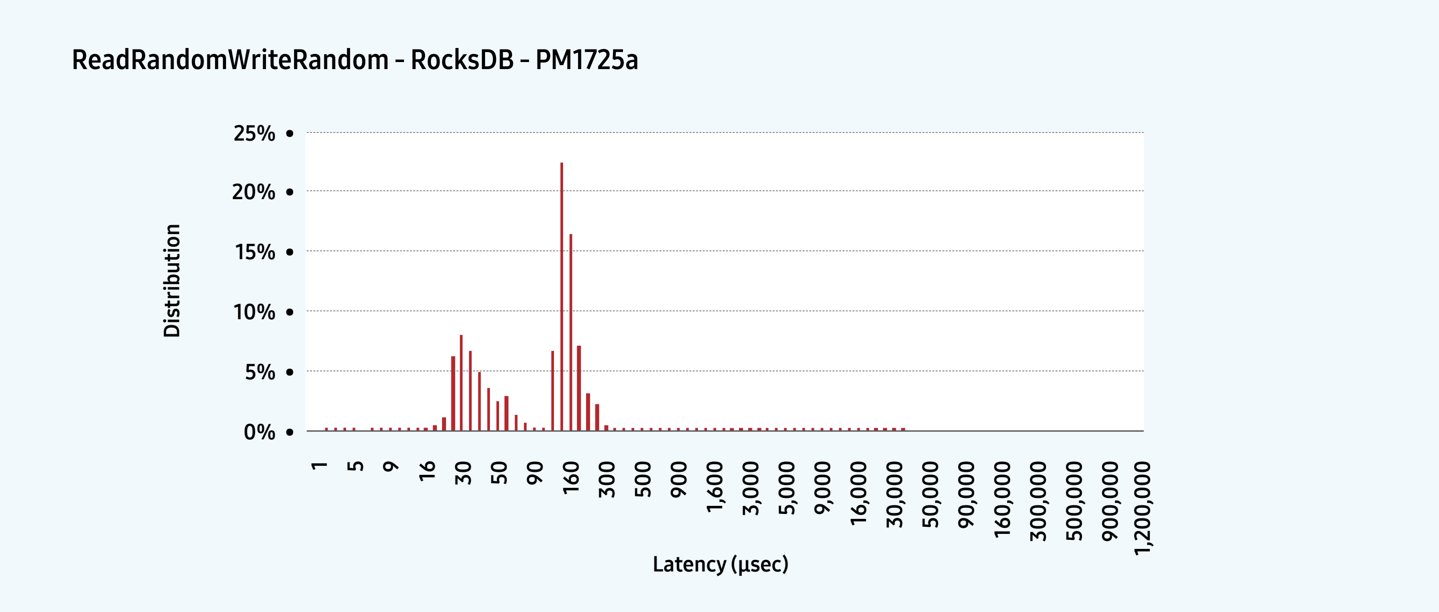


Figure 11: RocksDB Read Latency Distribution when data resides in PM1725a

Tail latencies are also a key performance metric. We can see from Figure 12 and Figure 13 that Helium and Z-SSD together deliver lower and shorter tail latencies.

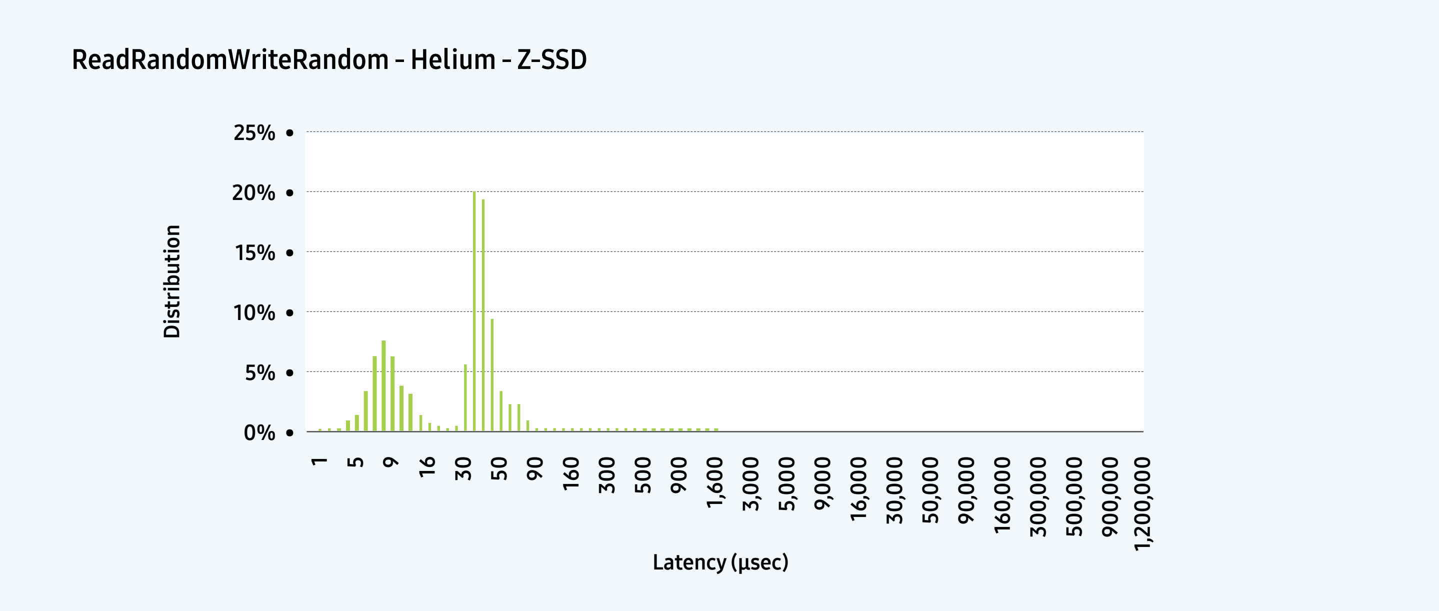


Figure 12: HeRocks Read Latency Distribution when data resides in Z-SSD

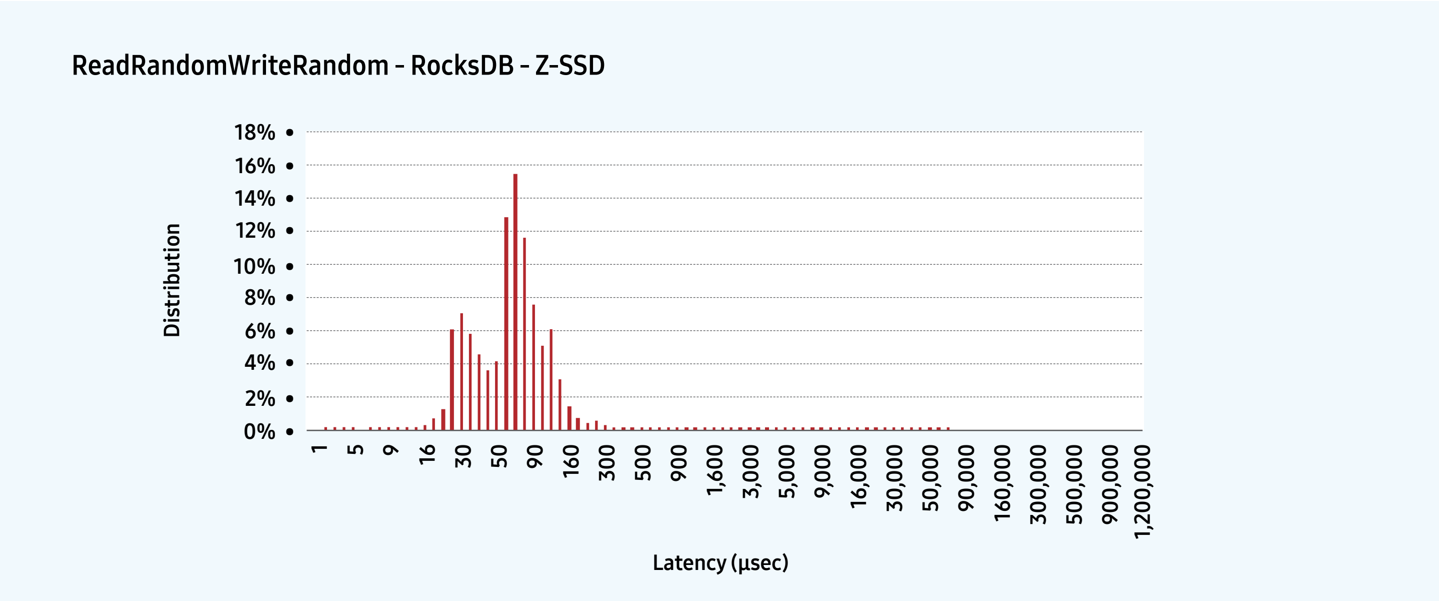


Figure 13: RocksDB Read Latency Distribution when data resides in Z-SSD

Conclusion

Storage systems that deliver near-real-time responses are required for most of today's IoT, edge and financial service applications. Clearly, legacy software stacks are unable to leverage the superior performance of devices like Samsung's Z-SSD. We have demonstrated a solution that combines Levyx's Helium key-value database engine and Samsung state-of-the-art NAND technology to dramatically improve performance. The Z-SSD enables enterprise system designers to create highly innovative systems that optimize cost and latency for real-time applications.

Results from *Read-While-Writing* and *Read-Random-Write-Random* workloads demonstrate the combined value of Z-SSD and Levyx's Helium for real-time applications. The 10X improvement at 1/10 the latency for read-heavy applications and over 3X improvement for random access patterns show that the solution tested here can be used for a wide range of big data applications and bring significant value to the data center customers running them.

A Real-Time, Low Latency, Key-Value Solution Combining Samsung Z-SSD™ and Levyx's Helium™ Data Store

About Levyx

Levyx has developed next-generation database storage and query offload engines that fully exploit the latest commodity hardware technologies including multi-core servers, internal and external flash systems, and IO offload engines. The result is unprecedented performance and latency reductions for IO intensive workloads, such as financial service backtesting and streaming analytics.

Levyx's new software-based engines allow for first time, persistent computing to be possible on "big data" platforms such as Apache Spark thru the use of SSDs instead of volatile memory-only designs. Levyx is now delivering the world's fastest key value store (Helium) and world's first Distributed Storage/Analytics Offload Engines running on Flash (Xenon).

About Samsung

Samsung inspires the world and shapes the future with transformative ideas and technologies. The company is redefining the worlds of TVs, smartphones, wearable devices, tablets, digital appliances, network systems, and memory, system LSI, foundry and LED solutions.

Embodying this vision, Samsung has been leading efforts to enable the creation of smaller-footprint servers that will decrease data center operational expenses, with Samsung low-latency Z-SSDs, high-density DRAM, and extremely fast NVMe SSD storage drives.

For more information, contact:

Levyx

info@levyx.com

SAMSUNG

msl-inquiry@ssi.samsung.com