WHITE PAPER

Intel® Storage Builders

Helium™ Key Value Store
Intel® SSD Data Center P4500 Series
Intel® Xeon® Scalable Processors
Software Defined Storage

Extraordinary Performance Running High Throughput, Low Latency NoSQL on Flash Memory

Combining Intel® SSD Data Center Series with Levyx® Software Enables In-Memory Like Performance Using Flash

"Pushing Flash into the memory hierarchy brings persistence to a sector of the market that until now has largely been about performance."

Reza Sadri, Levyx CEO

Executive Summary

Current technologies that access, process, and ultimately derive value from large-scale data sets are limited, as even the most prevalent techniques are not scalable, are highly inefficient in utilizing resources, or are quite simply not fast enough. Levyx has developed a simple, scalable I/O stack. Levyx's approach fundamentally rethinks the "traditional" data path and creates one that is designed for today's most advanced data center hardware architectures and the innovations on which they are founded, in order to fully maximize performance.

By benchmarking the performance of Levyx software on the Intel® Xeon® Scalable processors and Intel® SSD Data Center P4500 Series for PCIe* enterprise-class NVMe* storage devices, the data shows massive performance improvements for latency sensitive applications.

The Levyx High-Performance Data Path Paradigm

In the traditional server system I/O path, data must travel through main memory, the I/O subsystem, and into and out of the storage media (e.g., Flash). Flash and non-volatile memories (NVMs) are not optimized for conventional file systems, and OS kernels do not fully utilize the available bandwidth. As a result, block-oriented, unstructured access is highly inefficient. Levyx's software introduces a new I/O path that fundamentally disrupts the economics of data movement in real-time, bringing the benefits of affordable, high-speed data processing to an expanded variety of use cases and applications.

As shown in Figure 1, the Levyx data path is characterized by a single, persistent, high-capacity, high bandwidth, low-latency memory layer that is scalable with the number of cores in a system, cluster, or network, and with the bandwidth of the I/O system (Flash SSD or Storage-Class Memory). Under this schema, the benefits of Flash storage, especially NVMe*-based SSDs, are fully exploited, either in single-node or distributed system architectures. The result is a single, persistent memory layer that is object-oriented and allows for highly structured data access. This architecture, namely the Levyx Helium™ Data Store, is the basis for Levyx solutions designed for performance in high throughput, low latency NoSQL applications.

Levyx Helium Data Store provides a very fast and scalable platform for storing and retrieving data objects. It is specifically designed for applications that manipulate a large amount of data (billions of data items), and at the same time need to be very responsive. Some of the examples of Helium applications include:

- Real time analytics
- Analysis of network and server logs in data centers and cloud infrastructures
- Analysis of large sets of DNA sequences to find alignment and correlations
- Analysis of large, multibillion-node graphs

Helium is specifically designed to take advantage of new hardware trends, especially the emergence of CPUs with many cores, such as the Intel Xeon Scalable processors, and new nonvolatile storage technologies, like NVMe SSDs, including Intel SSD DC P4500 series for PCIe/NVMe devices. Levyx integrates high-capacity Flash and next-generation NVM fabrics into the system memory hierarchy.

In addition, Levyx employs rich and expressive key-value semantics instead of block interface, offering a native match to Big Data, HPC, and IoT application data demands. Its proprietary indexing methodologies are massively parallel, compressed, memory-resident, lock-free, and can track billions of ordered objects with microsecond lookup latencies. Since Levyx does object-level caching of data throughout the memory hierarchy, it can access petabytes of persistent storage at main-memory speeds and latencies. Finally, Levyx applies seamless and efficient translations of application-level parallelism to the physical processor, I/O bus, and Flash/NVM channel bandwidths. It uses advanced Intel instruction sets to build new, extremely efficient algorithms and data structures for processing many data items in parallel. In addition, its core algorithms inherently scale with the number of cores and number of I/O channels on the SSD side. See figure 1.

The Value of Intel SSD DC P4500 Series for PCIe/NVMe

SSDs using the PCIe interface and NVMe protocol offer much faster response times and lower latency than traditional SATA and SAS SSDs. This family of data center SSDs from Intel, which includes the Intel SSD DC P4500 Series, consists of PCIe Gen3.0 x4 devices architected with the new high-performance NVMe protocol, delivering leading performance, low latency, and unsurpassed Quality of Service.

Matching the performance with world-class reliability and endurance, the DC P4500 Series offers a range of capacities—1 TB, 2 TB, and 4 TB—in both Add-In card and 2.5-inch form factors. With PCIe Gen3 support and NVMe queuing interface, the DC P4500 delivers excellent sequential read performance of up to 3.2 GB/s and sequential write speeds of up to 1.1 GB/s. Intel's PCIe storage devices also deliver very high random read IOPS of up to 490K and random write IOPS of 38K for 4K operations. Taking advantage of the direct path from the storage to the CPU by means of PCIe, the DC P4500 exhibits low latency of less than 82 µs for sequential access to the SSD.

Providing the foundation for a high-performance analytics engine running Levyx software, the DC P4500 has shown to accelerate data analytics on Levyx technologies, as illustrated by the following benchmarks.

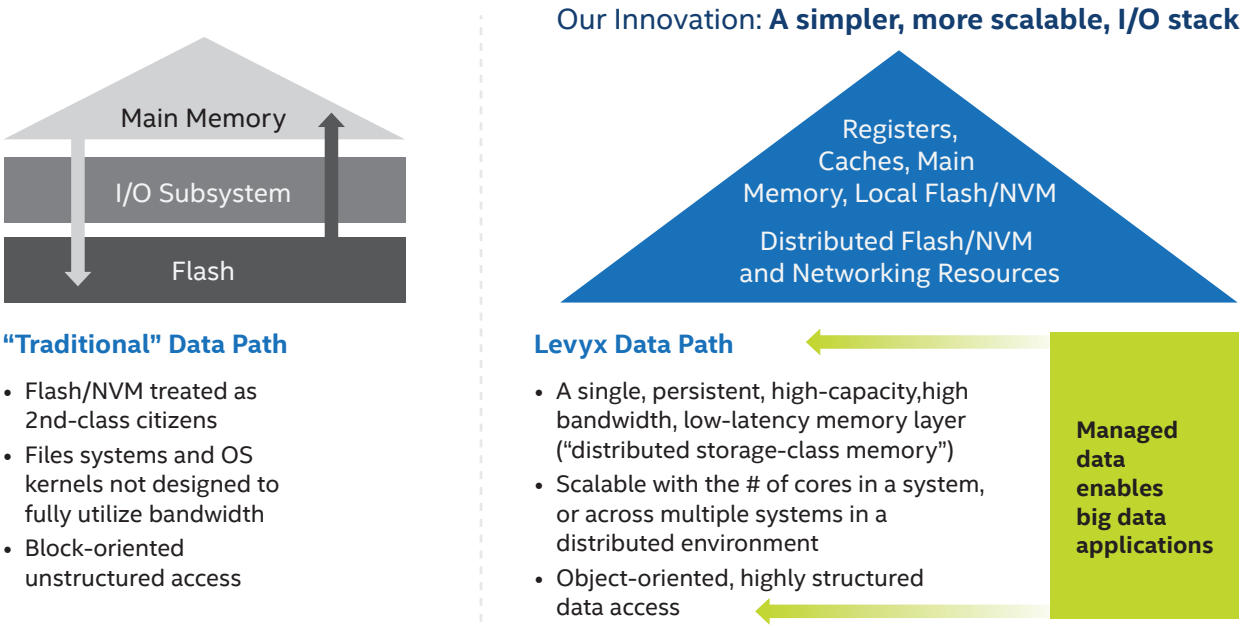

Figure 1. Levyx I/O stack

Benchmarking Levyx with the Intel SSD DC P4500 Series

Benchmarks measuring latency, throughput, and write amplification illustrate the high performance of this series of Intel® Storage Systems when combined with Levyx software.

The testing was done with the system configuration listed in Table 1.

Servers	Dell PowerEdge T630*
Processors	2 socket Intel® Xeon® Platinum 8168 processors [96 cores with hyper-threading]
Memory	384 GB
OS	Linux* 3.10.0-327.28.2.el7.x86_64
Helium™ software	Version 3.0.0
SSD	DC P4500 2 TB capacity

Table 1. Test system configuration

	Number of Threads			
Percentile	**1**	**4**	**8**	**16**
100 Byte Payload Put (Write) Operation Latency (microsecond)				
99.000%	2	4	5	5
99.900%	4	7	8	11
99.990%	9	364	392	408
99.999%	464	704	509	575

	Number of Threads			
Percentile	**1**	**4**	**8**	**16**
100 Byte Payload Get (Read) Operation Latency (microsecond)				
99.000%	1	2	2	110
99.900%	2	6	31	237
99.990%	3	22	188	403
99.999%	4	39	320	708

Table 2. Latency (in microseconds)

Latency and Throughput

The Helium performance benchmarking application was used to generate key/value pairs and to measure throughput and latency of basic put/get/delete operations, using +/1 microsecond precision. For all the tests, the DC P4500 was preconditioned with two full-capacity write cycles. This eliminates any false new-drive reporting numbers, which is important to achieving results consistent with "real-world" implementations. (Never before written NAND SSDs are slightly faster the first time they are written.) Next, varying threads (e.g., 1, 4, 8, 16, ...) were used to generate load with 1 million operations per thread.

For the latency measurements, Helium was configured to work in asynchronous mode, and a 1 GB read cache and 1 GB write cache were used for intermediate data caching. Next, 1, 4, 8, or 16 threads were used to continuously issue put operations of 100 byte object payloads with 16 byte key size. For put operations, all data eventually landed on SSD. The exercises were repeated with the get operations. For get operations, all data originated from SSD.

Latency

The results for latency are summarized in Table 2. The leftmost column is the percentile of operations that completed in the specified time.

To put things into context, with Helium, for a single thread, 100 byte update, only one in 10,000 operations takes more than nine microseconds to complete, and only one in 100,000 operations takes more than 4 microseconds to complete. This makes Helium an ideal platform for latency-sensitive applications, such as Internet of Things (IoT), cyber-security and financial trading.

IOPs Performance

Figure 2 depicts measured I/O performance in millions of operations per second. For each data point, a total of one million operations per thread were performed and we scaled the number of threads from 1 through 24 with a write buffer of 1 GB and read cache of 4 GB.

The figure clearly illustrates an extraordinary level of performance. Currently, in order to achieve this level of performance (at millions of gets or puts per second), the only option is an in-memory system, which is costly, not scalable, power hungry, and space inefficient. For a solution with Helium and Intel SSD DC P4500 Series drives with the throughput demonstrated here, combined with the above latency measurement, the benchmarks clearly show how results similar to an in-memory based system can be achieved by using high-performance data paths, modern data indexing, and lock-free data structures.

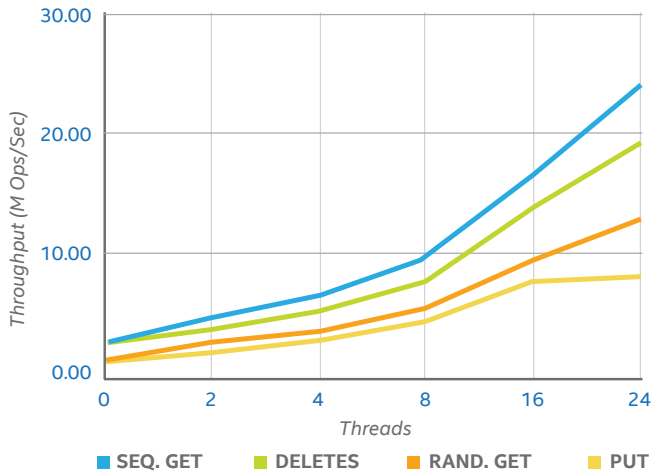

Figure 2. 100 Byte Payload, 1 GB write cache, 4 GB read cache

Read and Write Amplification

The Helium engine consumes 12 bytes to index any item (this is both on SSD and in memory). In addition, there are another 12 bytes of overhead per object for SSD. Therefore, there is a total of 24 bytes of overhead per value on SSD. The 24 bytes overhead is much lower compared to other NoSQL and key-value store, where just the in-memory overhead is upwards of 48 bytes and more so when serialized and written to SSD.

In order to measure the amount of write and read amplification (that is the ratio of data stored to overhead), one million 100-byte put requests of values were sent to SSD, and the amount of data that was written to SSD was simultaneously measured (which includes the payload and the overhead, including keys and metadata) using statistics from the block device. As illustrated in Table 3, for 100-byte value and 16-byte key sizes, the write amplification is less than 1.2. The same pattern is seen for read amplification (the read amplification is less than 1 for lower thread counts due to caching). Low write/read amplification directly translates to a much higher utilization of the Intel SSD DC P4500 bandwidth for reading and writing real data instead of overhead data.

	Number of Threads			
Measurement	**1**	**4**	**8**	**16**
Write Amplification 100 Byte Value	1.1102	1.1052	1.1043	1.1039
Read Amplification 100 Byte Value	0.5528	0.5520	0.5855	1.1036

Table 3. Read and Write Amplification

Comparing HeRocks™ and RocksDB

Improving upon the open-source version of the popular RocksDB, Levyx has developed HeRocks, which is essentially a clone of RocksDB using Helium as its internal data engine. In this test, HeRocks is compared to conventional RocksDB. Benchmarking was completed using db_bench, which is the default performance test tool of RocksDB. Two different benchmarks were run with 64 threads, 100 byte value sizes, and 1 million operations per thread – (i) fillseq which performs writes utilizing multiple threads, and (ii) readwhilewriting which is read operation heavy (64 threads) with one thread writing at the same time. The DC P4500 SSD is used to store data in both cases (formatted as xfs file system).

Program	**Throughput (ops/sec)**	**99% Latency**
RocksDB	286K	373 usec
HeRocks	1.1M	34 usec

Table 4. Throughput and latency, fillseq (writes)

Program	**RocksDB**	**HeRocks**	**HeRock Gain**
Throughput (ops/sec)	286K	1.1M	>4X
99% Latency	373 usec	34 usec	10X

Program	**Throughput (ops/sec)**	**99% Latency**
RocksDB	778K	332 usec
HeRocks	4.3M	216 usec

Table 5. Throughput and latency, readwhilewriting (64 thread read with 1 thread writes)

As evidenced in Tables 4 and 5, the Levyx-aided version of RocksDB outperforms the conventional version by up to 5X in terms of throughput. Latency is also reduced by up to 10X. Both throughput and latency measurements are indicators of how well Levyx software pairs with Intel SSD DC P4500 Series .

Conclusion

Levyx software and Intel SSD DC P4500 Series are highly compatible for processing high throughput and low latency workloads that have real time requirements. The results highlighted by the benchmark testing clearly show how performance similar to an in-memory based system can be achieved. The combination of these advanced technologies can have an enormous impact on the future of big data processing.

We invite you to test our software here: **http://www.levyx.com/helium**. If you are running data-intensive workloads and not getting the most out of your infrastructure, please contact use at **info@levyx.com**.

To find out more about us and our products, visit **www.levyx.com**.

To learn more about Intel® storage technologies, visit **www.intel.com/storage** or Intel® Solid State Drives at **intel.com/ssd**.

The Solutions Library on the Intel Builders home page can help you find reference architectures, white papers, and solution briefs that can help you build and enhance your data infrastructure. **https://builders.intel.com/solutionslibrary**

You can also follow Intel® Builders on Twitter* by using **#IntelBuilders**

0218/YMB/HBD/PDF Please Recycle 336312-001US