Extreme Storage Performance with eXFlash DIMM and AMPS

lenovo

60East TECHNOLOGIES

For more information on AMPS, visit the 60East web site at http://www.crankuptheamps.com or contact 60East at info@crankuptheamps.com.

For more information on exFlash DIMMs, see the IBM/Lenovo product page at http://www-03.ibm.com/systems/x/options/storage/solidstate/exflashdimm/

AMPS, the Advanced Message Processing System from 60East Technologies, is the software behind some of the most demanding financial applications in the world. AMPS applications require predictable latency at high messaging volumes, while maintaining a persistent, queryable "flight recorder" and state of the world, or current value cache. This paper introduces AMPS, describes the demands that AMPS places on storage systems, and how adding flash storage on the memory channel improves performance and helps to meet those needs.

AMPS is engineered for high performance in real world conditions. Our customer applications regularly publish millions of transactional messages a second. AMPS provides a current value cache, the State of the World, that can hold the current state of those messages, and allows applications to run ad hoc queries on the State of the World and the transaction log, as well as play back messages from any point in time.

High Performance Storage

Storage matters. To provide this kind of performance, AMPS needs storage that runs at consistently high throughput with predictable low latency. AMPS records the state of the message with the metadata required to retrieve the message, such as a timestamp, in near real time for millions of messages a second. This requires the ability to sustainably write multiple gigabytes a second of data at predictable latencies.

What happens when the storage system isn't up to the demands of AMPS? AMPS publishers save messages locally until AMPS acknowledges that the message has been persisted. When the rate of message publication exceeds the throughput of the storage device, publishers need to retain messages longer. This requires more local storage. If transaction logging can't keep up with the rate of publication, the additional storage requirements, or "push back" on the publishers, can cause publishers to run out of storage.

One way to work around this is to allow publishers to discard messages before the messages have been acknowledged. This adds risk of lost or duplicate messages if the publisher or AMPS fails over, and masks the problem of slow storage rather than solving the problem.

As an example, a moderately-sized AMPS system logging a million 1KB messages per second requires more than a gigabyte per second of sustained write throughput, in addition to the throughput required to read the State of the World and the transaction log for queries. AMPS workloads mix heavy sequential writes with multiple sequential reads starting at arbitrary points in the transaction log. For most storage systems, this is a nightmare scenario – and the performance numbers show that.

Putting Storage to the Test

To evaluate storage performance, 60East engineers created a test using the AMPS high performance I/O engine. The test creates the same read/write pattern as the transaction logging that an AMPS instance performs under heavy load, using the same code path. However, the test allows our engineers to precisely adjust the read/write mix and time this single component. By eliminating other variables from the test, we can more accurately measure storage performance under AMPS workloads.

AMPS transaction logging is characterized by sequential reads and writes. For this test, we simulate 512 byte messages being written to storage. Our storage engine batches the messages into a 128 message chunk, submitting 64K bytes per operation.

When we add reads to the mix, the reads requested are also sequential access from the file: the same access pattern AMPS uses to replay transaction logs.

Hardware Specifications

We ran the tests on hardware with the following specifications:

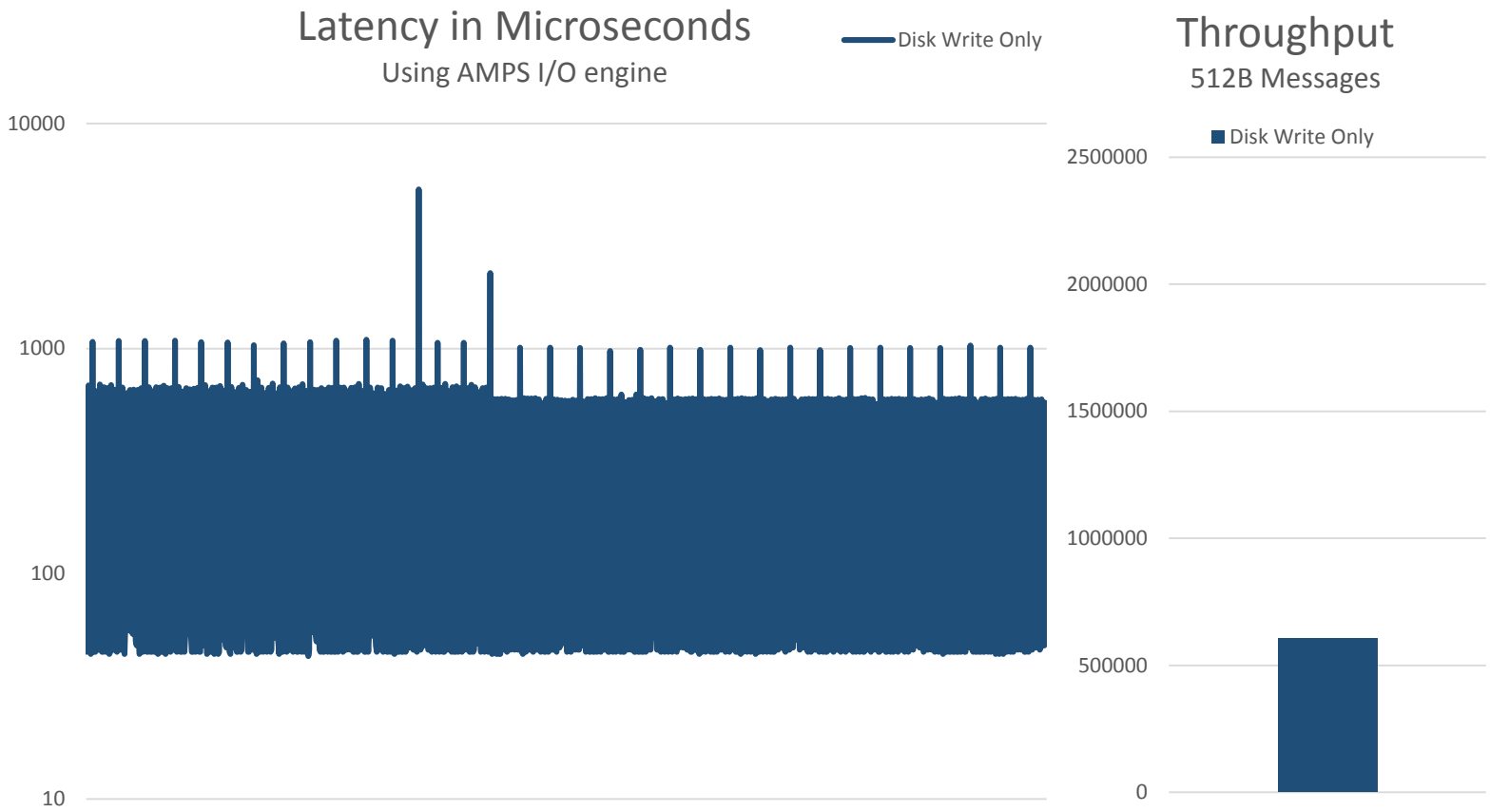
IBM x3650 M4	
Processor	2 x Intel® Xeon® E5-2690 @ 2.90GHz
Memory	64GB or 384GB
PCIe Storage	Leading vendor, 785GB capacity
eXFlash DIMMs	8 DIMMs, software RAID 0, 3TB total capacity
Disk	IBM 90Y8878 at 10KRPM, 300GB capacity

These specifications are similar to the specifications that 60East customers use for production deployment of AMPS.

Spinning Disk

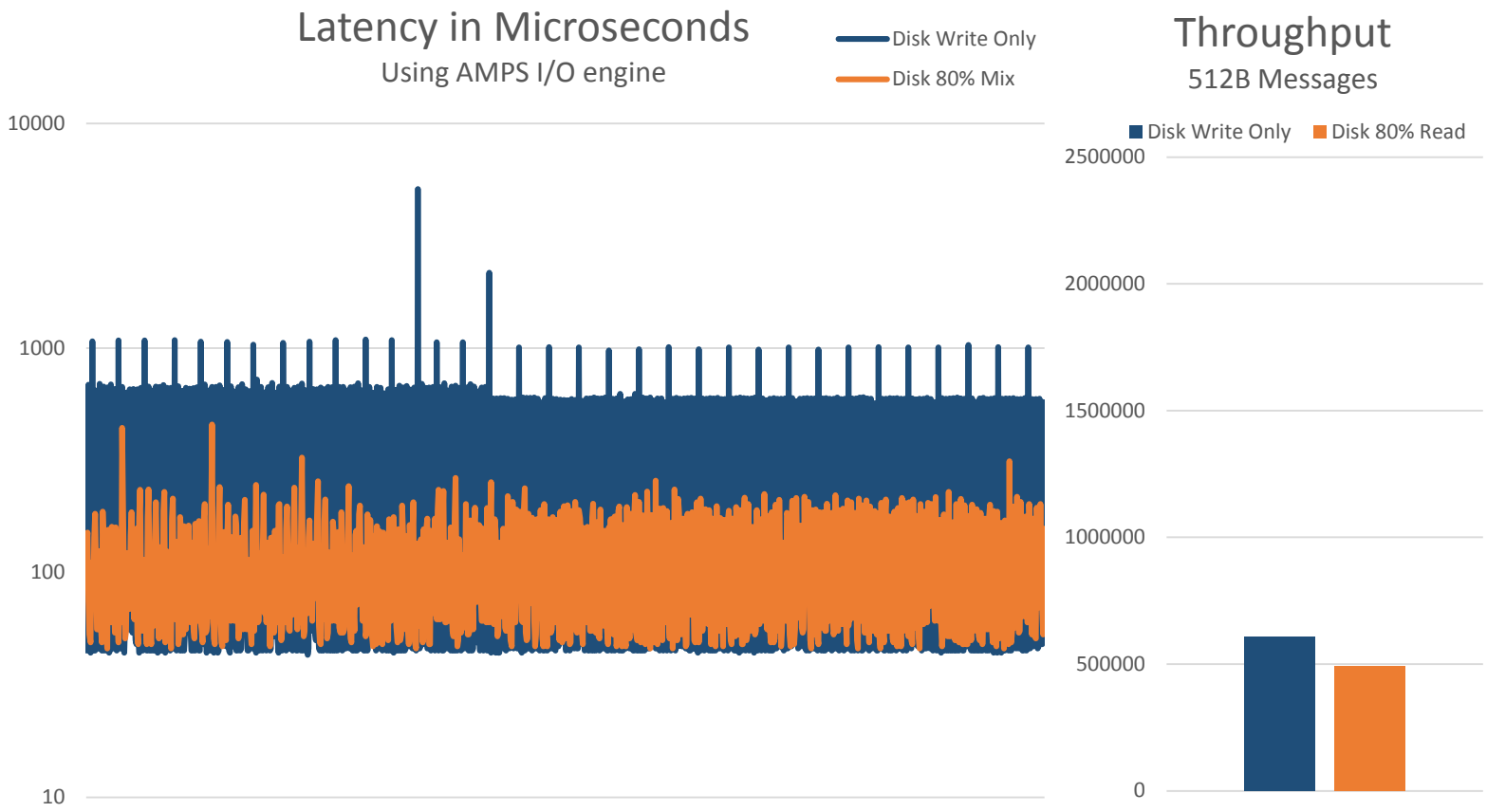
To get a baseline for storage performance, we ran our test using the drive in one of our performance lab machines. The access pattern for the AMPS transaction log, sequential writes and reads, matches the optimum access pattern for spinning disk. Most performance critical applications don't use this method any more, but comparing to spinning disk gives a good baseline comparison.

Here are the performance characteristics of the AMPS I/O engine with spinning disk, using writes only:



Notice that the latency chart uses a logarithmic scale: this is necessary because of the wide variance in numbers, and the fact that the numbers range from roughly 56µs to over 5ms. We use the same scale throughout this paper, to make it easy to compare the charts. Likewise, for the throughput chart, we use a consistent scale. The spinning disk can handle over 600,000 512B messages per second.

Few applications simply use sequential writes. To make the test reflect an active, real-world workload, we added in an 80% read mix, so that the AMPS engine was requesting 80% reads, 20% writes. The chart below shows the results.

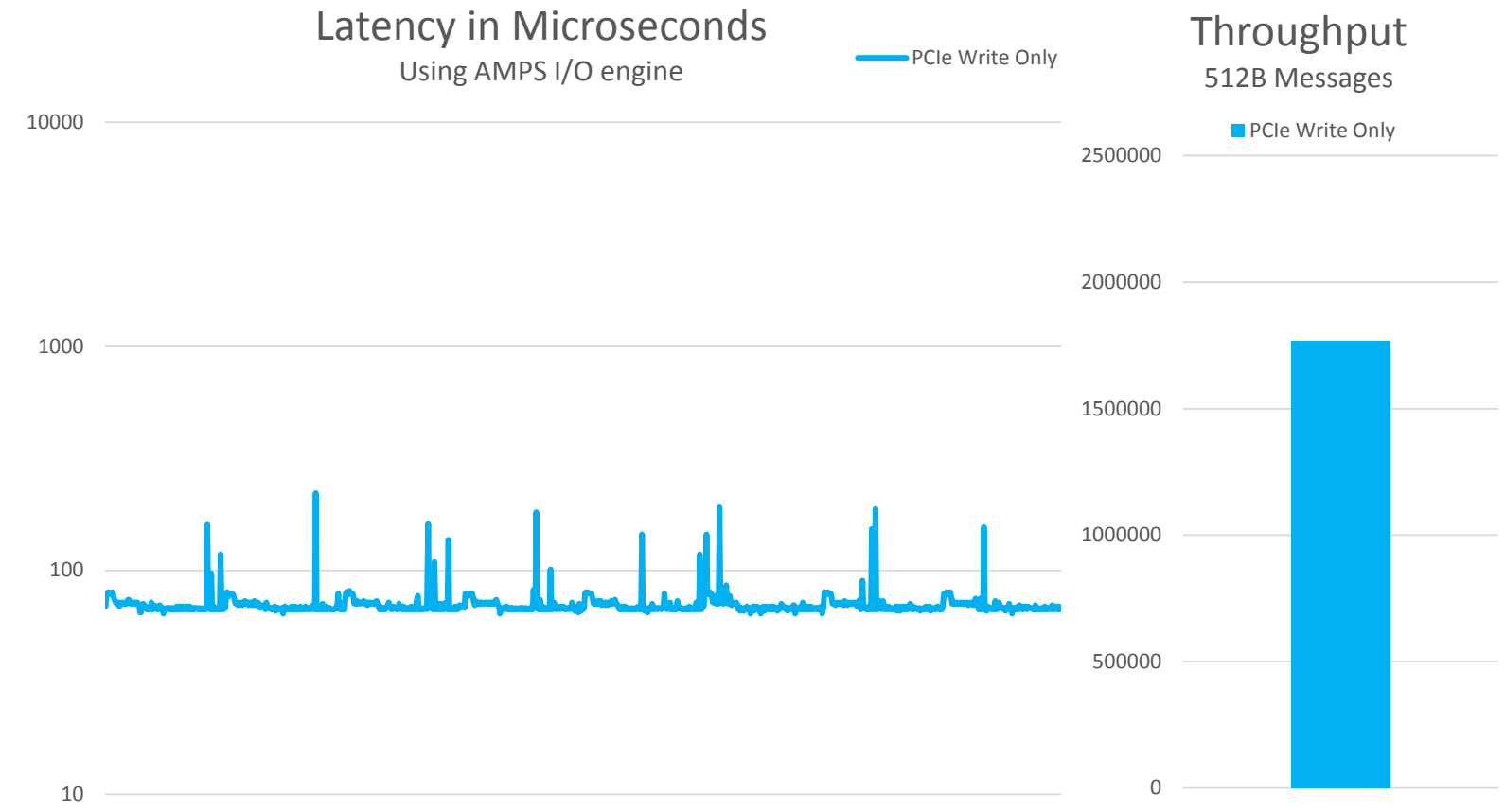


The latency drops somewhat when we add a heavy load of reads to the mix. However, this also reduces throughput to just under 500,000 messages per second. There's still a large variation in latency, though somewhat less than the write-only case.

Those are respectable numbers, and there's a good reason why systems still include spinning disk.

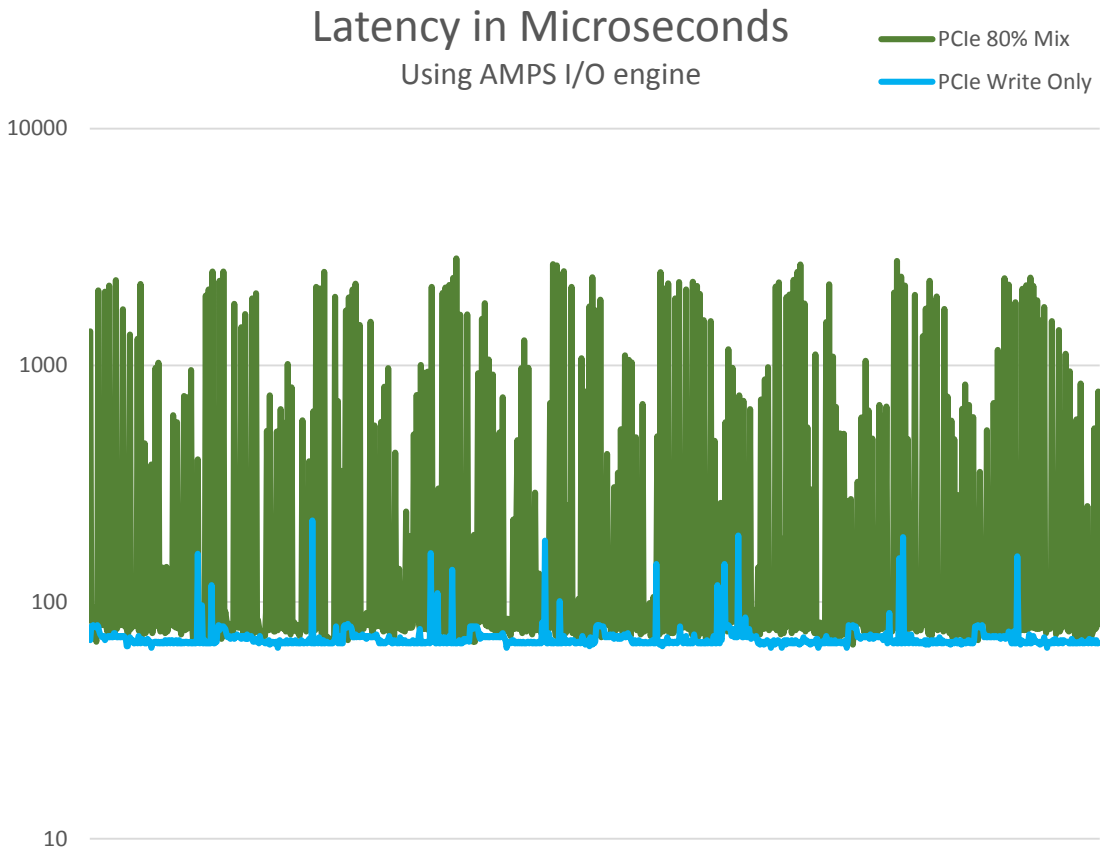
PCIe Flash

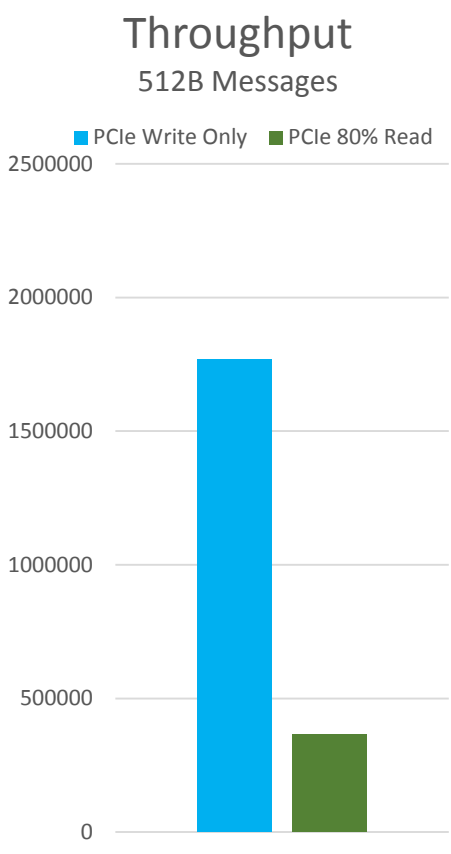
High-performance systems no longer rely on spinning disk for performance-critical storage. Instead, most current systems use PCIe-based flash storage. This storage doesn't rely on mechanical parts, and uses the PCIe bus for fast access to memory and CPU. Here are the results for PCIe write only.



The overall throughput numbers are much better, and the latency is lower compared to the spinning disk: this is clearly a better solution for high-performance applications. The spinning disk had a wide variation in latency times. PCIe is more consistent, although it still suffers from regular spikes in performance. With 100% writes, the PCIe solution can handle nearly 1.8 million messages a second, or three times the number of messages handled by the spinning disk.

As with the spinning disk, though, when we introduce reads, both latency and throughput suffer. We ran the test again, changing the mix of operations from 100% write to 80% read and 20% write.



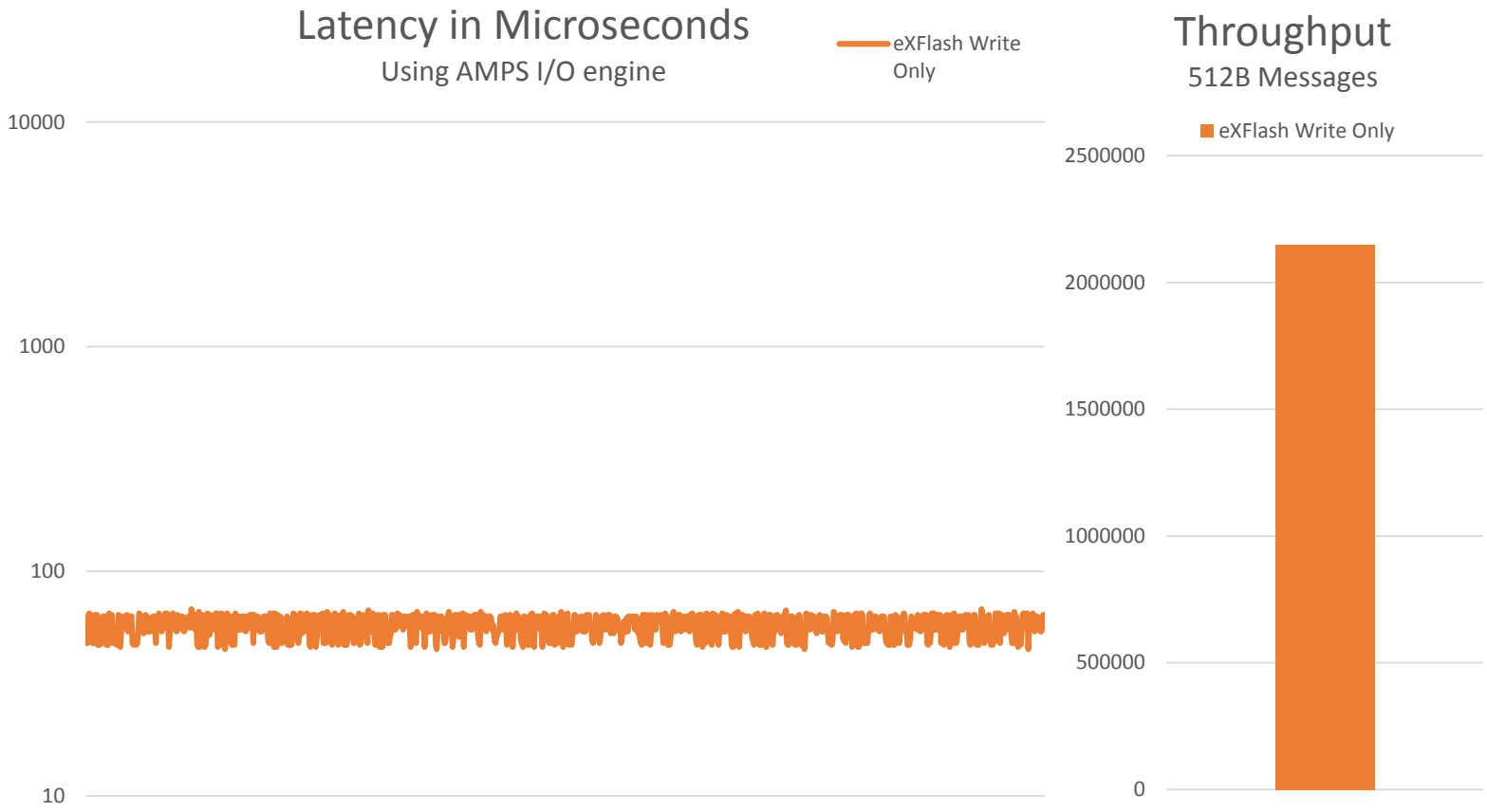


The performance of the PCIe flash storage system suffers dramatically when we add 80% reads to the workload. Latency not only becomes higher overall, but begins to vary drastically. Likewise, throughput degrades to the point where the PCIe storage system is processing fewer messages per second than the spinning disk. In our tests, at an 80% read mix, the PCIe storage system was only able to handle 365,000 512B messages.

PCIe offers good performance for write-heavy applications, but is not well suited for applications that include a heavy read mix in their I/O access patterns.

eXFlash DIMM

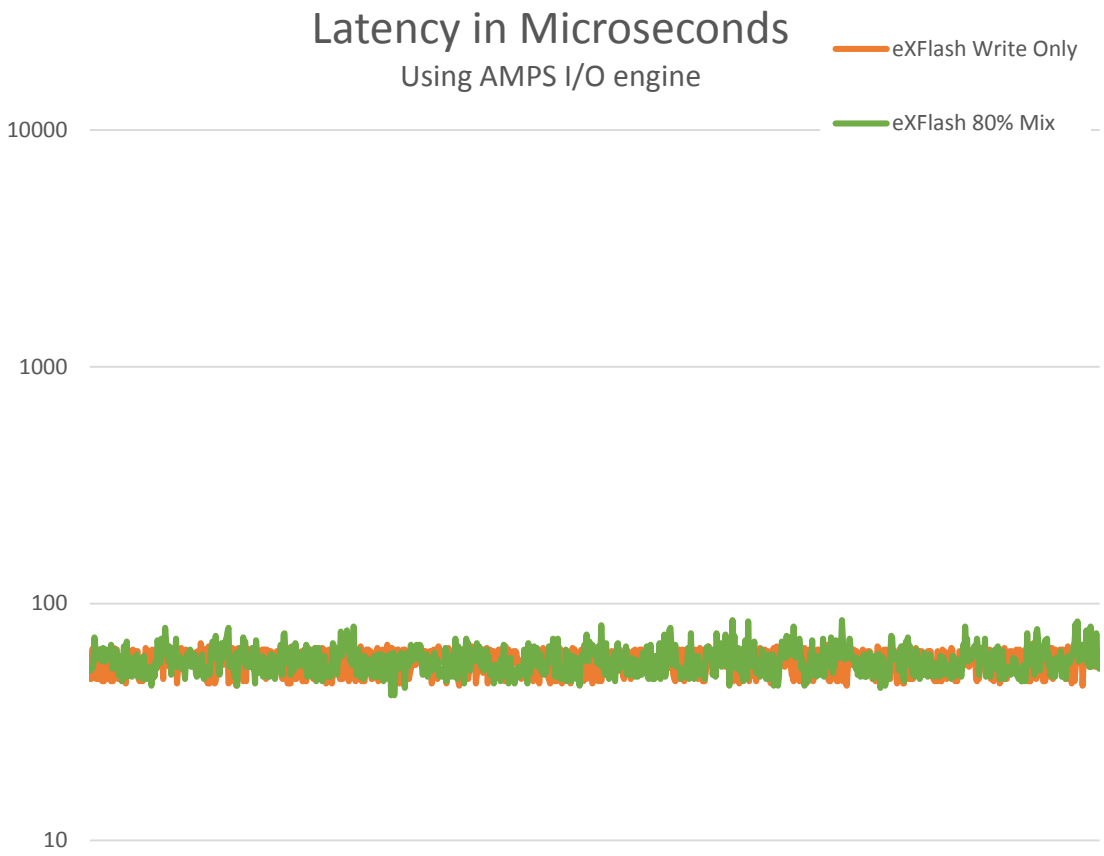
Finally, we ran the test on the new eXFlash DIMM from Lenovo. This is brand new technology that adds flash storage to the memory channel of the system. The memory channel is designed for predictable, high-speed data movement. As before, we first ran the test with a 100% write mix.

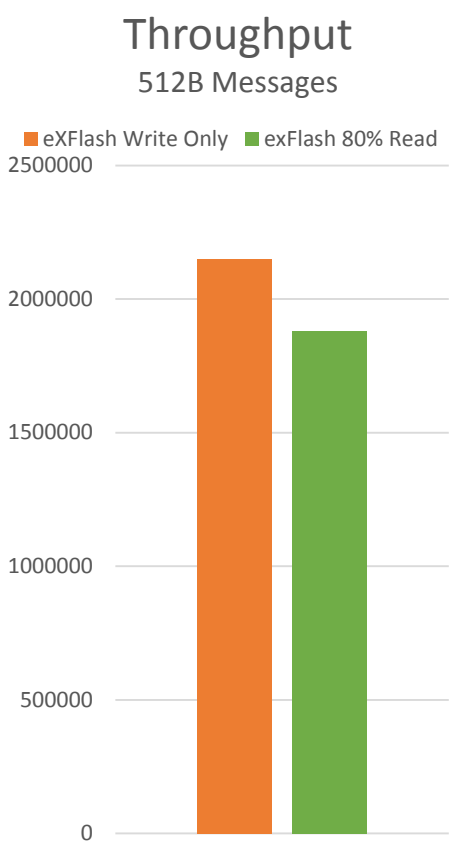


The overall latency and throughput numbers improve substantially as compared to PCIe storage. While the PCIe system has spikes of higher latency at regular intervals, the eXFlash DIMMs maintain consistent performance. The DIMM can handle approximately 2.15 million messages a second as compared with 1.84 in the PCIe system.

So far, so good. As we've seen throughout this paper, though, the question is whether the eXFlash DIMMs can handle a workload that is heavily weighted toward reads.

The eXFlash DIMMs perform amazingly well at an 80% read and 20% write mix. Overall latency is only slightly higher, with none of the wide variations in performance seen with the spinning disk and PCIe solutions. Throughput drops only slightly, to 1.87 million 512B messages per second – slightly *more* throughput than the PCIe card showed at a 100% write mix.



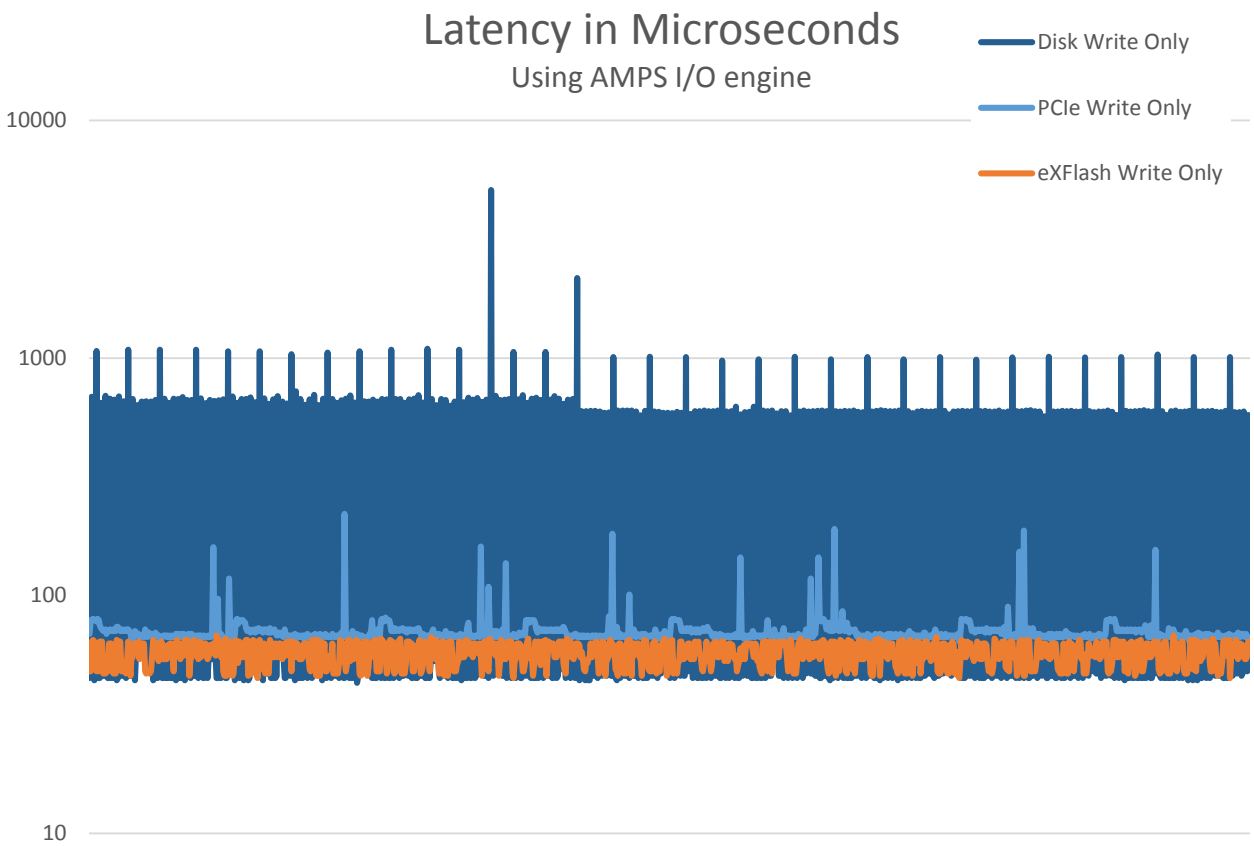


Summary

The numbers speak for themselves. In our testing, the eXFlash DIMMs outperform both PCIe and spinning disk on every dimension: higher throughput, lower latency, and rock solid predictability. The advantages are clear even when doing only sequential writes. When we add reads to the mix, the eXFlash DIMM technology leaves the other technologies behind.

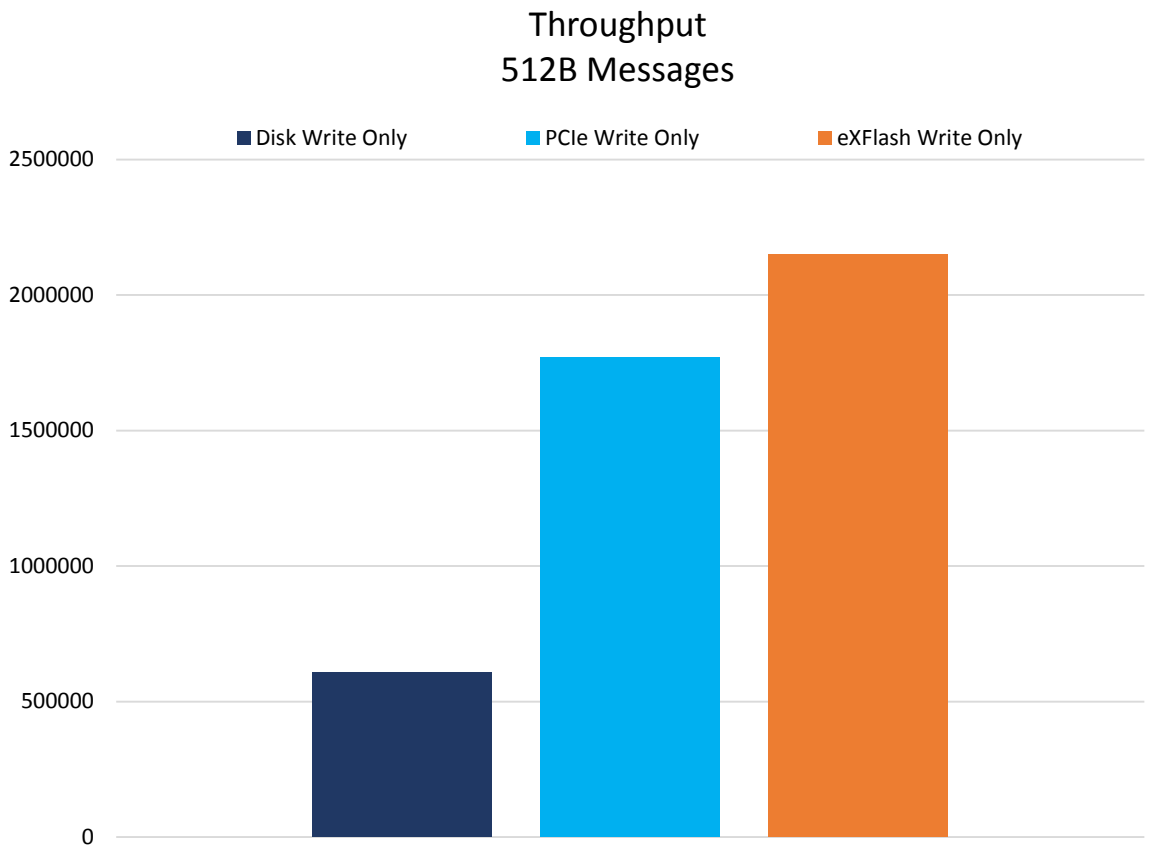
The expanded charts below show all of the tests plotted together, on the same scale as the individual charts above.

In write-only workloads, eXFlash DIMMs show lower latency and less variation for write only workloads.



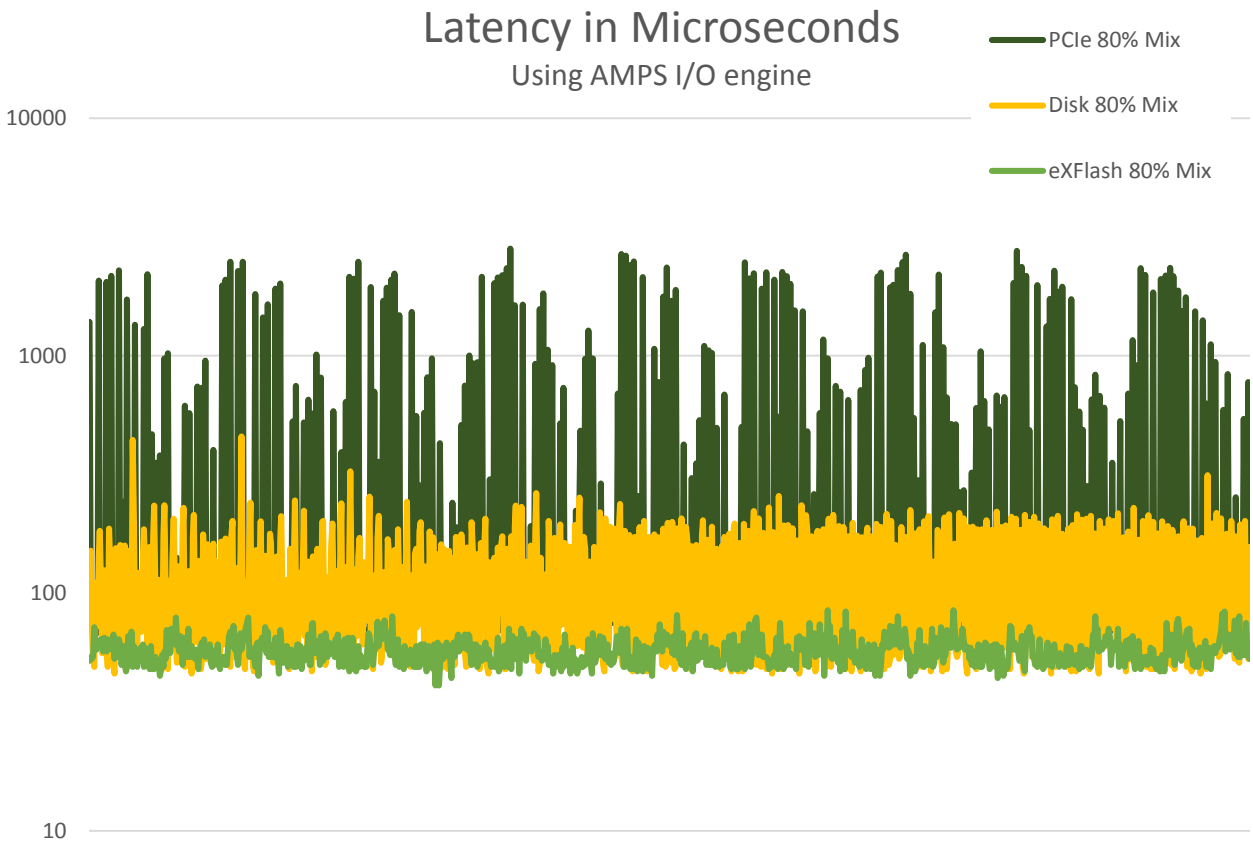
As you can see, latency is dramatically different for exFlash DIMMs overall. The latency is lower and more consistent than any other solution in the write only case.

Throughput numbers on a pure write workload also favor the exFlash DIMM, as shown in the following chart.



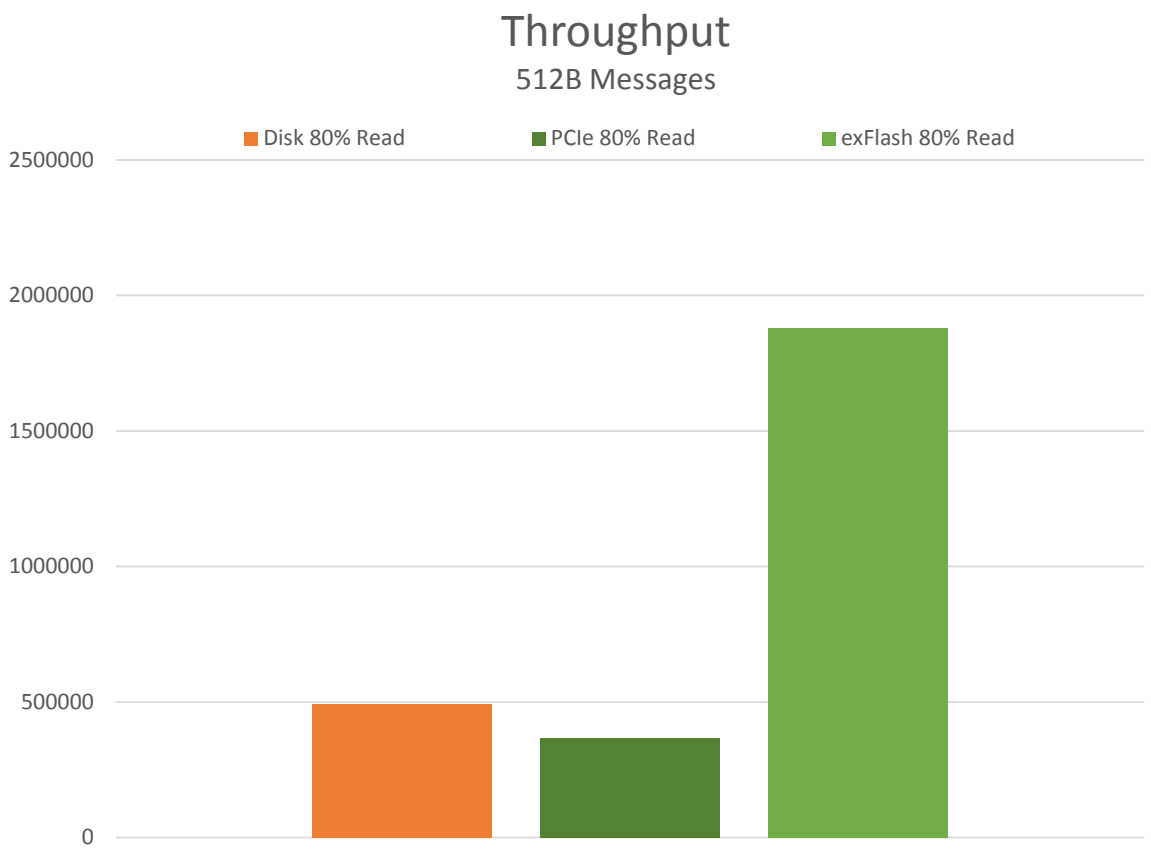
exFlash DIMMs provides higher throughput at a lower and more consistent latency than either the spinning disk or the leading PCIe flash storage solution in pure write workloads.

Charting the 80% read, 20% write mix side by side, the results are similar.



At the 80% read, 20% write mix, the PCIe solution has much wider variability in results than the spinning disk, while the eXFlash DIMM continues to have both lower and more consistent latency.

The throughput comparisons across technologies are even more impressive.



Both spinning disk and PCIe flash storage suffer a dramatic drop in throughput at the 80% read mix. eXFlash DIMMs, on the other hand, have only a slight drop in throughput.

Bottom Line

Our testing shows that there is no other technology on the market that shows the level of performance and throughput achieved with eXFlash DIMMs. eXFlash DIMMs are the only technology in the market that can keep up with AMPS I/O and deliver the uncompromising performance our customers demand.

At 60East, we specialize in high performance messaging systems. We've worked closely to test the technology behind eXFlash DIMMs while that technology was in development. We tell our customers that if they want to get the most out of AMPS, storage on the memory channel and eXFlash DIMMs need to be included in their planning.