NVMe SSD Write Performance

Introduction

*BittWare's Data Recorder reference design captures up to 200 Gb/s high-speed sensor data to NVMe solid-state drives (SSDs). This gives users an open architecture to build any data recorder system they require—but a common question we hear is "**How many drives are required for a given sustained data rate**?"*

We want to address this question as it's often assumed the calculation is a simple one: look at an SSD vendor's drive specification for write speed, and divide up the required maximum sustained bandwidth for the application, add a bit of margin, and you have your number of RAID0 (striped) drives required. The challenge is that write figure on the SSD spec sheet is never sustainable for the long record times of our high-performance customers.

Let's look at some real-world performance figures using common SSD drive types and dig into the reasons why sustained write isn't as simple as read speeds. We'll then conclude with some recommendations on drive arrays for given speeds using our Data Recorder reference design. The good news is our design can easily scale as required using our TeraBox servers and 250-SoC FPGA card.

SSD Drive Reputation

Solid-state drives with NVMe interfaces are known for being very fast. Indeed, they are faster than traditional hard disk drives (HDDs) for a certain workload. Specifically, SSDs are optimized for many more reads than writes and random access. Of course, for a data recorder, the focus is the opposite: **sequential writes**.

Another challenge is that NVMe SSDs often carry specifications on maximum "streaming write" bandwidth that is limited to a small portion of the drive's capacity when doing sustained writes. The majority of a long sustained write will significantly underperform this figure, and as drives get larger it gets worse. To see why, let's briefly look at how bits are stored on SSDs.

Larger SSDs Tend to Have Reduced Sustained Write Performance

Increasing SSD density means moving to newer technologies that happen to also make sustained writes slower. These SSD technology generations are called SLC, MLC, TLC, and finally QLC, in the order they appeared in the market, which is also in order from lowest density to highest density. The SSD designers are very aware they are making streaming writes slower over time. Thus they often reserve a portion of even the very newest, densest SSDs to operate in the old SLC mode. This allows write to stream at the old, faster speeds until that section of drive is filled up. After that, write streaming slows down to the speed supported by the rest of the drive.

SSD Drive Performance Degradation Thresholds

The first three thresholds are examined further in our Benchmarks section.

- Top performance is delivered only until the **SLC cache is filled up**. See Benchmarks section for details on when that happens for a range of drives.
- Then we see reduced performance until the drive hits a **capacity threshold** were it begins reducing the size of the SLC cache.
- Shrinking the SLC cache requires a **background copy** that further slows things down.

There's yet another reduction that occurs when a recorder begins to **overwrite old data** with newer data. We'll cover this in the next section.

NVMe SSD Write Performance

Overwriting Data Performance Degradation

Why overwriting data degrades performance

- SSDs must zero a page before writing to it.
- Thus, if you change just one byte, the SSD controller looks for a page it has zeroed in the past and then copies the whole page onto it, with that one byte changed.
- SSDs cannot zero a single page. That operation uses a block of pages.
- Every SSD is different, but, as an example, a Samsung 840 EVO has 2048 Kb pages. There are 256 such pages in a block.
- Zeroing a page is a relatively slow operation that occurs in the background when an NVMe Trim command tells the SSD that certain blocks are no longer in use.
- In a disk recorder, data is often captured into a ring data structure that wraps around when the disk is full.
- In this case, there is no opportunity to trim in advance. Also, even if an application does issue trim, the SSD recorder can be too busy to find spare time to perform the zeroing.
- Performance will significantly decline if every few page writes are followed by a pause while another block of pages is zeroed.

How to avoid this in the FPGA

- Your application could stop when a disk is full instead of wrapping around.
- Or you could issue trim commands and provide enough NVMe drives that each drive has a bit of performance to spare to zero blocks.

SSD Drive Benchmarks

Drive Benchmarking Methodology

- We use SPDK to benchmark NVMe drive performance
- Removes any filesystem overhead
- Maximizes DMA efficiency by aligning buffers to cache lines
- Empty drives are filled completely as fast as possible using a dedicated thread and ensuring the NVME DMA controller always has a write command to process

Testing Categories

Rather than benchmark specific drives, our goal was to choose examples from consumer and enterprise markets:

- Consumer Market: TLC with SLC
- Consumer Market: MLC with no SLC
- Enterprise Market: TLC with no SLC
- Enterprise Market: 3D XPoint

Consumer Market: TLC with SLC

Example drive tested: **Samsung 970 Plus, 1 TB**

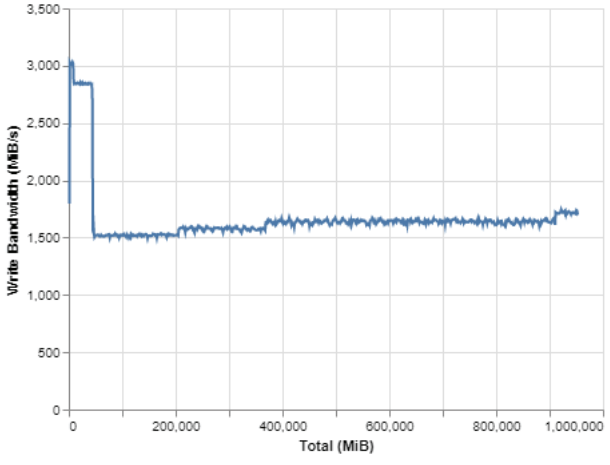
This is a typical TLC consumer-level drive with an SLC cache.

- Datasheet says streaming writes at "max 3.3 GB/sec"
- We measured 2.7-2.8 GB/sec until the SLC cache fills
- It then dropped to roughly 1.6 GB/sec
- The drive managed to keep writing at that same speed when we started overwriting stale data

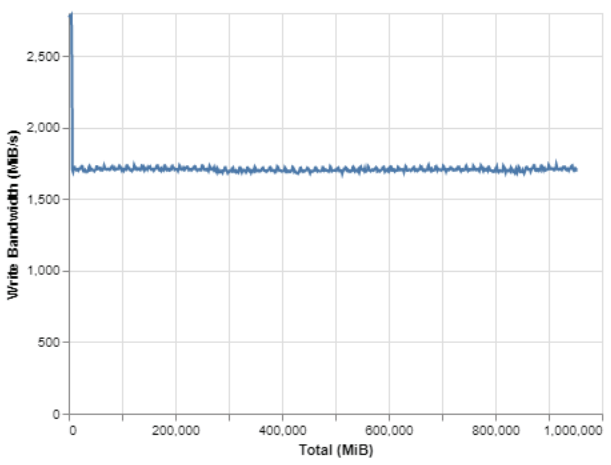
Achieving 100G Recorder Performance:

RAID 0 configuration would require at least **8** drives.

Empty



Rewrite full drive



NVMe SSD Write Performance

Consumer Market: MLC with no SLC

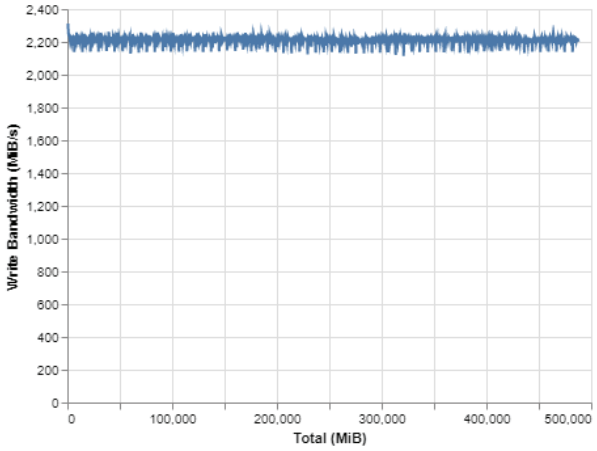
Example drive tested: **Samsung 970 Pro, 512 GB**

This is an MLC drive targeting the consumer market without an SLC cache.

- Datasheet says streaming writes at "max 2.3 GB/sec"
- There is no SLC cache in this drive. We measured roughly 2.2 GB/sec with no drop off
- The drive managed to keep writing at that same speed when we started overwriting stale data

Achieving 100G Recorder Performance:

RAID 0 configuration would require at least **6** drives



Enterprise Market: TLC with no SLC

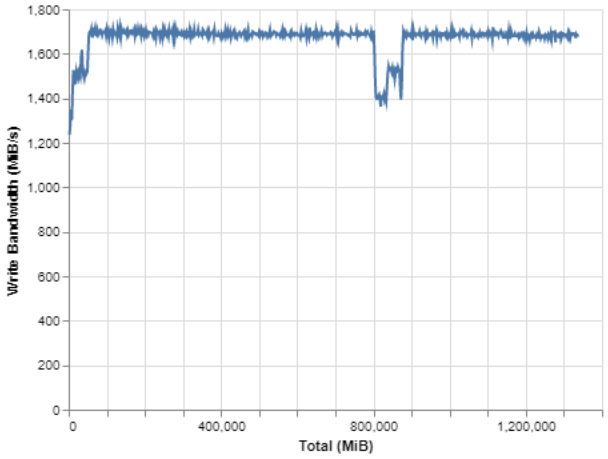
Example drive tested: **Samsung 1725b, 1.6 TBM**

This is an MLC drive targeting the consumer market without an SLC cache.

- Datasheet says streaming writes at "2.0 GB/sec"
- Most of the time this drive delivered 1.7 GB/sec; however, it occasionally dropped to 1.3 GB/sec
- This is a TLC drive targeting the enterprise market apparently with no SLC cache; it was sold by Dell as an "NVMe Mixed Use Express Flash;" the "enterprise" label means a longer lifetime specification

Achieving 100G Recorder Performance:

RAID 0 configuration would require at least **10** drives



Enterprise Market: 3D XPoint

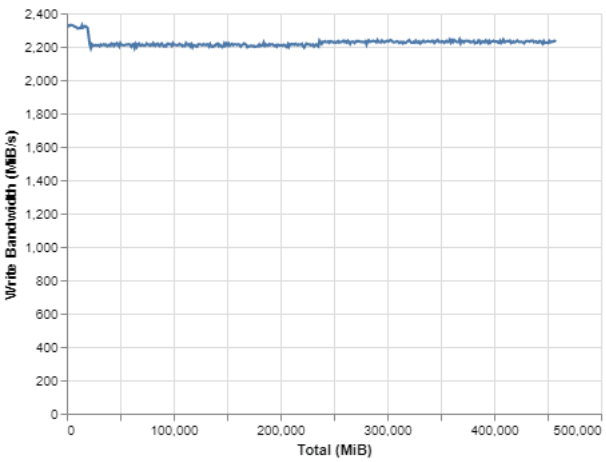
Example drive tested: **Intel Optane, 480 GB**

Optane drives use 3D XPoint memory technology and have significantly better endurance than traditional SSD drives. However, they are not available in the same high capacities as traditional NVMe drive technologies.

- Datasheet specifies 2.2 GB/s sequential write bandwidth
- Consistently delivered 2.2 GB/s or better

Achieving 100G Recorder Performance:

RAID 0 configuration would require at least **6** drives



NVMe SSD Write Performance

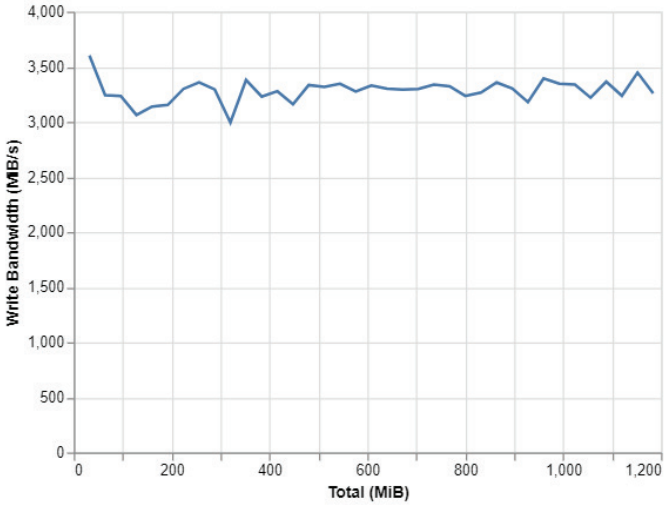
Enterprise Market: Gen4 3D TLC

Example drive tested: **Intel SSD D7-P5510 3.84TB**

Drives using PCIe Gen4 have substantially improved performance. To take advantage of the optimized data-moving engines we used large buffer sizes of 256KB or more.

- Datasheet specifies 3.3 GB/s sequential write bandwidth
- Consistently delivered 3.2 GB/s or better

Achieving 100G Recorder Performance:
RAID 0 configuration would require at least **5** drives.



Benchmark Summary

The use of multi-cell architectures to get larger drive sizes comes at a cost of sustained writes for long periods. The biggest impact of this comes on consumer SSDs. The best real-world streaming write speed that matches the claimed maximum figure comes from Intel Optane drives; however these are also the most limited in drive size.

For those specifying systems for sustained writes, the key is to obtain or perform real-world benchmarks on the target drives. Our recommendations above for sustained writes at 100 Gb/s are based on such real-world test data.

Even so, there are further considerations in writing to SSDs that should be considered, which we will cover next.

Drive Lifetime

A final consideration in write performance is lifetime of the drives. Writing to an SSD wears it out—and specifically data recorder applications can potentially stress the lifetime of the drives.

- SSD lifetime is specified as Drive Writes Per Day (DWPD) for the length of the warranty period.
  - There is a wide variance in SSD lifetime but, to establish an expectation, note that a drive rated 1 DWPD can be overwritten 1,800 times.

An "enterprise" SSD includes extra flash cells (over provisioning) to allow for longer lifetimes (larger DWPD).

- A 100 GbE disk recorder can generate up to 12.5 GB/sec of packet data.
  - That does not include metadata (PcapNg packet headers)
- If your packet recorder uses twelve expensive 30 TB enterprise SSDs, that means your array can hold about 8 hours of 100 GbE traffic. The array will wear out in just over 1.5 years if DWPD is 1.

However, this example is illustrative, not realistic.

- No 100 GbE link will run saturated with traffic 24/7.
- Commercial cybersecurity installations want to hold 7-10 days of traffic, not just 8 hours.

Conclusion

While SSDs bring impressive performance over traditional hard-disk drives, the improvements are application-specific and, in some cases, performance is reduced.

For applications with sustained writes, such as those targeted by our Data Capture and Recorder projects, there are a number of factors to consider beyond the maximum sustained write specification.

*Get in touch with BittWare to **learn more about our Storage acceleration and Sensor Processing products and solutions**. Or, visit the BittWare website at www.bittware.com.*

Rev 2022.09.20 | September 2022

BittWare
a molex company