Case Study: SupremeRAID™ and Solidigm D5-P5316 QLC NVMe

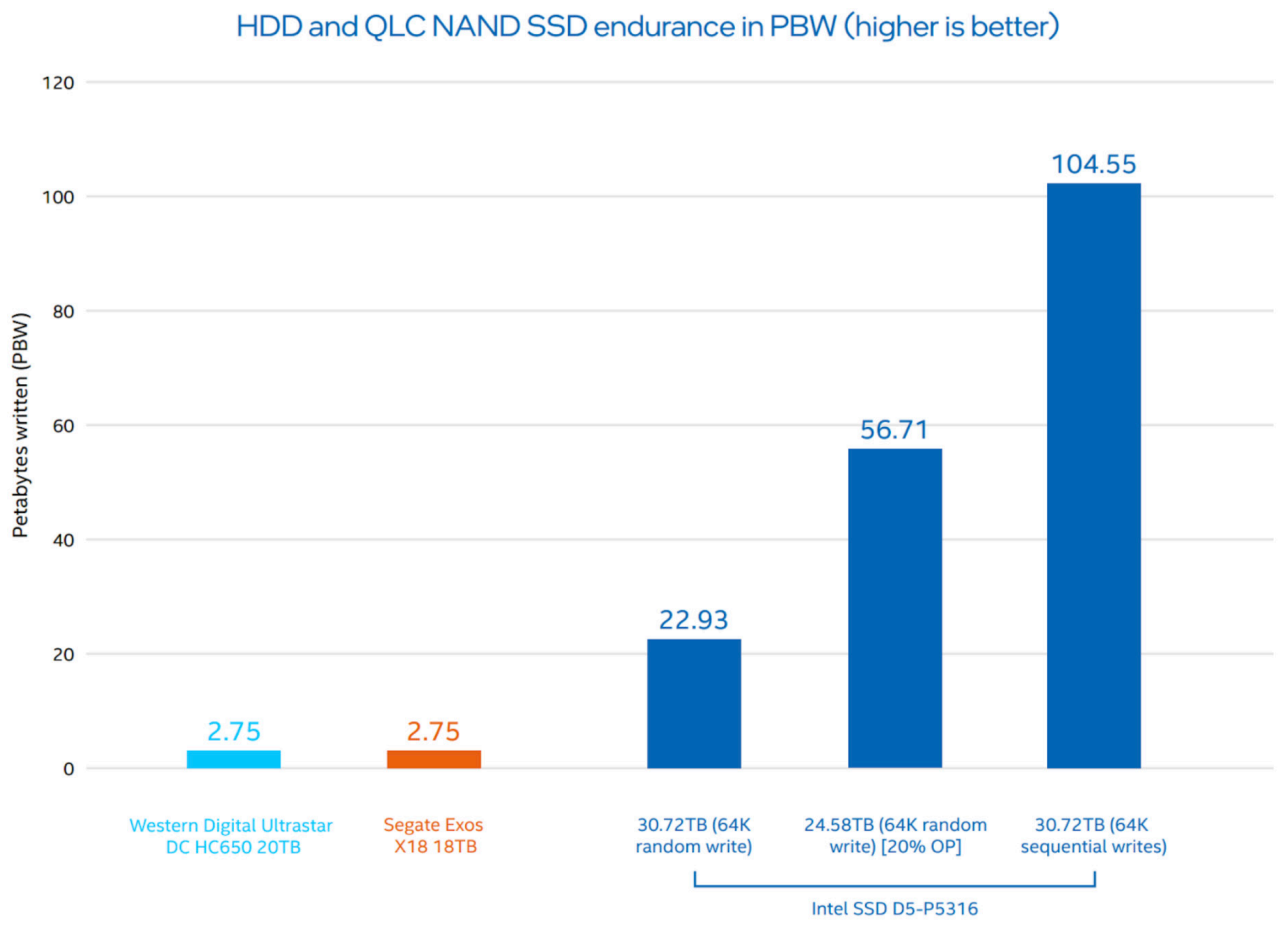
WHITEPAPER

Jonmichael Hands
VP Storage, Chia Network

Table of Contents

Why QLC

Storing more bits per cell is the most effective way to increase the amount of capacity and decrease cost of NAND based SSD. Early reports of QLC showed drives with severely limited performance and endurance, which hindered the adoption of QLC in the majority of applications. Solidigm has performed an amazing feat by getting PCIe 4.0 x4 performance of over 7GB/s reads, but an even more impressive 3.4GB/s of write bandwidth on a QLC drive. This comes with the fact that QLC paired with the Solidigm PCIe 4.0 controller works very well at the capacity points they have chosen, 15.36 and 30.72TB. At these large capacities, there is plenty of NAND to keep the controller busy, delivering high capacity and performance. Solidigm also solved the endurance problem by delivering 3000 P/E cycles (more than 2x QLC competitors); and notably, the 30.72 drive delivers a stunning 104.5 Petabytes Written (PBW) of endurance.



[QLC NAND Technology Is Ready for Mainstream Use in the Data Center]

But what about QLC in mainstream enterprise storage applications and storage arrays? Netapp performed a large-scale study of 2 million SSDs to look at endurance and reliability in the field, the findings were very interesting.

[Operational Characteristics of SSDs in Enterprise Storage Systems: A Large-Scale Field Study]

"The majority of SSDs in our data set consume PE cycles at a very slow rate. Our projections indicate that the vast majority of the population (~95%) could move toward QLC without wearing out prematurely."

Solidigm has done quite a bit of research into applications for QLC and the endurance and performance requirements, which you can read more about in their whitepaper titled [QLC NAND SSDs are Optimal for Modern Workloads]. QLC delivers phenomenal TCO, enterprise and cloud workloads would be wise to adopt.

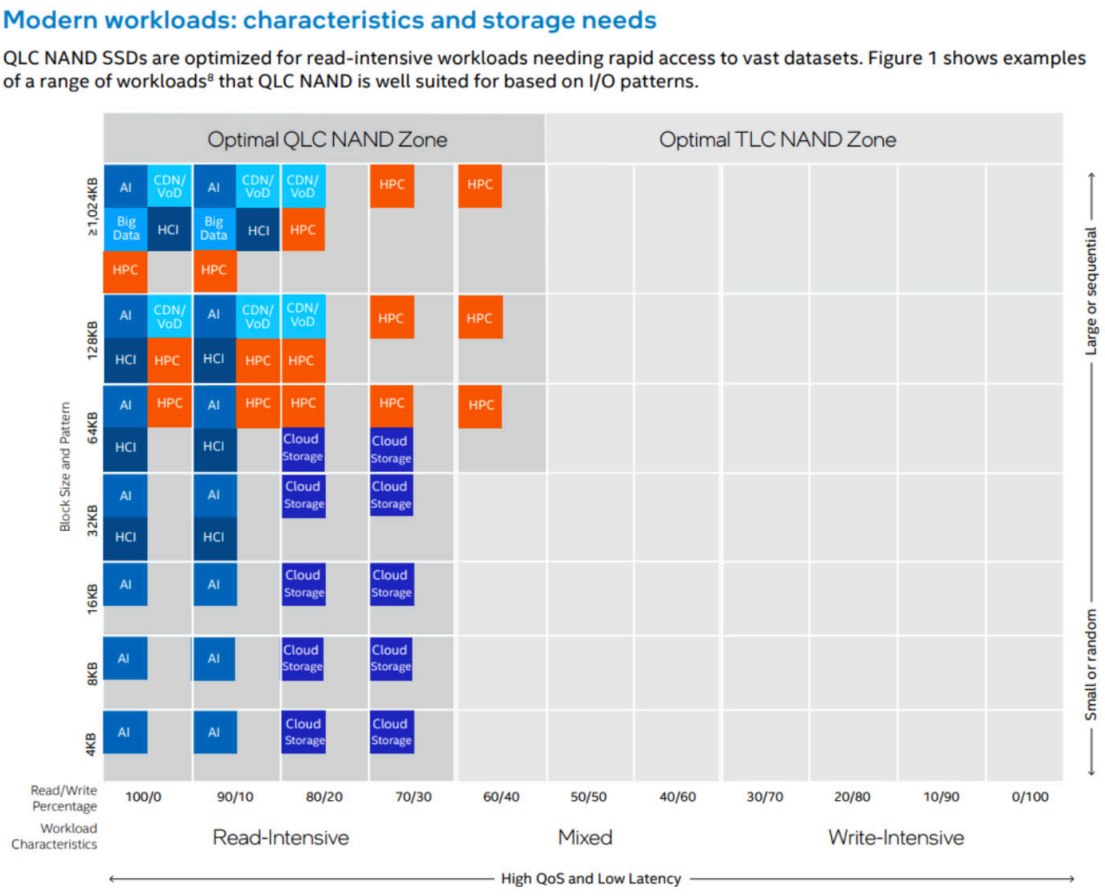


Figure 1. QLC NAND SSDs are optimized for read-intensive workloads.

Why SupremeRAID™

GRAID Technology Inc. developed a disruptive RAID technology based on special software coupled with GPU hardware technology to unlock the performance bottleneck of SSDs with RAID protection. SupremeRAID™ works by installing a virtual NVMe controller onto the operating system and integrating a PCIe device into the system equipped with a high-performance AI processor to handle all RAID operations of the virtual NVMe controller.

The solution offers many advantages:

World Record Performance
Take full advantage of NVMe performance —a single SupremeRAID™ card is capable of delivering up to 19 million IOPS and up to 110 GB/s of throughput, and redefines the performance standards for SSD RAID technology.

Plug & Play
Effortless installation, no cabling or motherboard re-layout required—SupremeRAID™ can be used for systems with Direct CPU connections (SSDs directly connected to the CPU via PCIe) without changing the hardware design as well as with PCIe architectures utilizing PCIe switches.

Flexible & Future Ready
Optimized for SCI (Software Composable Infrastructure), SupremeRAID™ delivers unmatched flexibility and automation capabilities with NVMe-OF support. It can also be used for external SSDs connected via NVMe-oF.

CPU Offload
Unlike traditional software RAID, SupremeRAID™ does not consume a large amount of CPU—offload your entire RAID computation to SupremeRAID™ to free-up CPU computing resources for 5G, AI and AIoT applications.

Highly Scalable
Easily add features like compression, encryption, or thin provisioning.

SupremeRAID™ delivers tremendous performance with comprehensive data protection and flexibility, not only resolving the performance bottleneck but also significantly reducing TCO.

Is Hardware RAID Dead?

Legacy Hardware RAID Cards

Hardware RAID is still used predominantly in the enterprise OEM space, where people buying a small number of servers need local data protection. Traditionally RAID was the tried-and-true method to protect data local to a server. The emergence of NVMe SSDs created a gap in hardware RAID when they were introduced in 2014: there was no RAID card support for many years, plus the RAID cards severely bottlenecked the performance. Only one to two drives were needed to saturate the PCIe 3.0 link on the RAID card, and RAID5 IOPS didn't scale. NVMe SSDs are known for the low latency on the PCIe bus; adding additional NVMe to SCSI translation and another controller was too much overhead.

Meanwhile it didn't require a lot of RAID card CPU cycles for HDD. But vendors have been forced to continue to add cores in the HW RAID controllers to keep up with SATA and SAS SSDs—an impossible race with PCIe 4.0, PCIe 5.0, and the ever-increasing performance of SSDs every year.

Software RAID

The DIY community has embraced open source solutions like ZFS, which offer impressive features for data durability, checksums, snapshots, and compression—but performance is abysmal, and cannot scale to modern NVMe SSDs. Alternative modern copy-on-write based file systems exist and boast data protection, checksums, snapshots as well as compression, such as btrfs. While btrfs has native RAID5 capability, it also comes with massive warnings that the software isn't mature. While improvements and innovation are occurring rapidly, there is still a massive performance bottleneck in software compared to the block device, suggesting more improvements are needed to keep vup with fast NVMe SSDs.

MDADM (Linux RAID) has come a long way by adding features such as partial parity log to protect against the RAID5 write hole while using SSDs with power loss protection, but the algorithms largely haven't changed and performance can vary greatly with devices.

RAID Initialization

RAID5 initialization first involves writing the parity data to the entire array. This is very bad for SSDs for two main reasons. First, it consumes unnecessary NAND endurance, writing a lot of data to the disk that is not user data. Second, and more importantly, modern SSDs use any blocks that have been unwritten to as spare area for garbage collection, reducing write amplification, increasing performance, and increasing endurance. A current software or hardware RAID5 starts off initialization by writing every single LBA in the array, which is catastrophically bad for modern SSDs, but especially QLC SSDs.

TRIM is arguably one of the most important things for SSDs, as a way for the host to tell the drive firmware which LBAs are no longer storing user data. Most drives support deterministic read zero after TRIM (meaning, if you read an LBA after it has been given a TRIM command, the drive always returns zeros).

SupremeRAID™ has an awesome trick that makes the parity calculation for RAID5 very easy when initialization. If A=0 and B=0, A XOR B = 0. (!) **No need to calculate the parity for the entire array since the GRAID software can just TRIM all the drives and ensure the parity and data are all zero.**

Old drives were not doing TRIM very efficiently, and this could take a long time, causing filesystem and operating system developers to recommend even turning off TRIM to improve performance in favor of a weekly task to TRIM all the unused LBAs. This is terrible on modern drives like the D5-P5316, which has new firmware algorithms to issue TRIM as a background operation on the drive and complete the command instantly on the host.

To illustrate how insanely high the TRIM bandwidth is on these drives, SupremeRAID™ software can initialize the entire RAID5 volume in about 1 second, compared to over 20 hours with Linux software RAID (as shown below).

Solution	Data	Parity	Initialization Time
GRAID	10	1 (RAID 5)	1 second
Mdadm	10	1 (RAID 5)	20.7 hours

ZFS Performance Comparison

I created a zpool of raidz1 (one parity) with 11 drives to compare to my 10+1 above.

```
sudo zpool create adprve raidz /dev/nvme0n1 /dev/nvme1n1 /dev/nvme2n1
/dev/nvme3n1 /dev/nvme4n1 /dev/nvme5n1 /dev/nvme6n1 /dev/nvme7n1
/dev/nvme8n1 /dev/nvme9n1 /dev/nvme10n1
```

```
zpool list

NAME     SIZE  ALLOC   FREE  CKPOINT  EXPANDSZ   FRAG    CAP  DEDUP
HEALTH  ALTROOT
adprve   154T  11.9T   142T        -         -     0%     7%  1.00x
ONLINE  -
```

The zfs array really does not like even a basic FIO script, with a 1TB test file the performance jumps all over the place, and the average performance is extremely low.

```
Sequential Read 1MB
   READ: bw=657MiB/s (689MB/s), 657MiB/s-657MiB/s (689MB/s-689MB/s),
io=38.5GiB (41.3GB), run=60001-60001msec

Sequential Write 1MB
  WRITE: bw=1165MiB/s (1222MB/s), 1165MiB/s-1165MiB/s (1222MB/s-1222MB/s),
io=68.3GiB (73.3GB), run=60001-60001msec

Random Read 64k
   READ: bw=80.1MiB/s (83.0MB/s), 80.1MiB/s-80.1MiB/s (83.0MB/s-83.0MB/s),
io=4804MiB (5037MB), run=60001-60001msec

Random Write 64k
  WRITE: bw=72.3MiB/s (75.8MB/s), 72.3MiB/s-72.3MiB/s (75.8MB/s-75.8MB/s),
io=4337MiB (4548MB), run=60001-60001msec
```

Synthetic tests aside, I do a basic file copy and see about 2.8GB/s peak read and write speeds during the file copy. A fraction of the performance that this array is capable of and less than a single disk! People may make an argument that this is “fine” performance for networked file storage - I disagree. ZFS makes users make an uncomfortable tradeoff between data integrity and performance. ZFS levels of paranoia on silent data corruption are just simply not needed when you have a proper enterprise SSD.

Data Protection: Using Enterprise Quality SSD

Modern enterprise SSDs have things like capacitors and firmware to keep data consistent on an unplanned power loss, error correction on all the internal memory, and many other features to deliver high reliability and low uncorrectable bit error rate.

Take a snapshot of the OCP NVMe SSD specification which is used by major hyperscalers like Meta and Microsoft for large-scale NVMe deployments.

6.3 End to End Data Protection

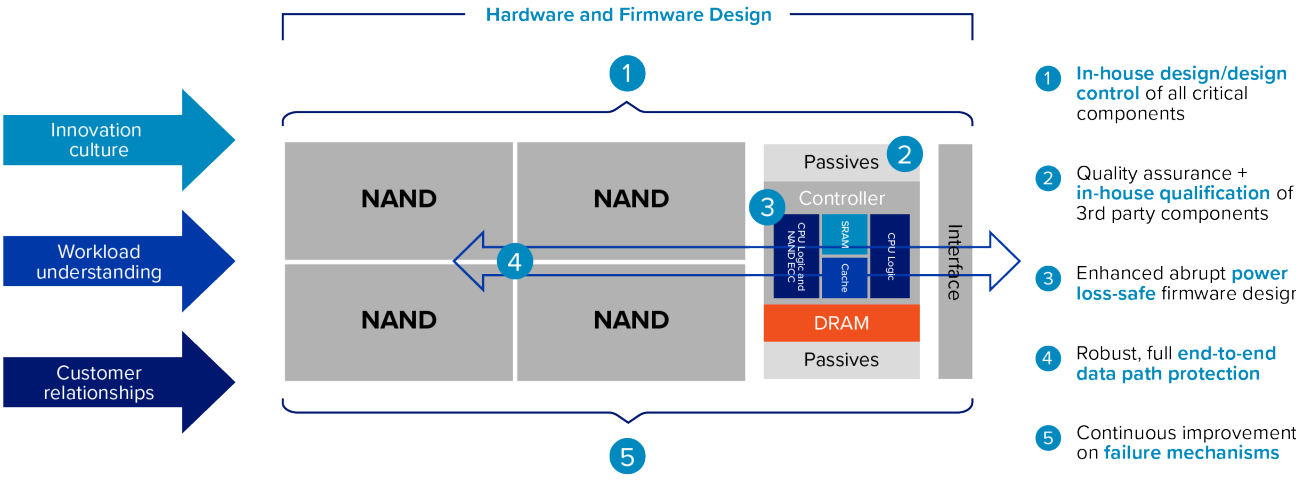
Requirement ID	Description
E2E-1	All user data and metadata shall be protected using overlapping protection mechanisms throughout the entire read and write path in the device including all storage elements (registers, caches, SRAM, DRAM, NAND, etc.).
E2E-2	At least one bit of correction and 2 bits of detection is required for all memories. This shall be for all memories regardless of function.
E2E-3	The entire DRAM addressable space shall be protected with at least one-bit correction and 2 bits of detection scheme (SECDED). This includes but not limited to the following: • Flash translation layer (FTL). • Mapping tables (including metadata related to deallocated LBAs). • Journal entries. • Firmware scratch pad. • System variables. • Firmware code.
E2E-4	Silent data corruption shall not be tolerated under any circumstances.
E2E-5	The device shall include a mechanism to protect against returning the data from the wrong logical block address (LBA), including previous copies from same LBA, to the host. It is acceptable if the device stores additional/modified information to provide protection against returning wrong data to the host. Device shall perform host LBA integrity checking on all transfers to and from the media.

The thing to note…
Silent data corruption shall not be tolerated under any circumstances.

A vendor will get instantly disqualified from a qualification at a hyperscaler if ANY silent data corruption is found. Enterprises and cloud service providers take this very seriously and drive vendors have a zero tolerance for silent data corruption.

Solidigm explains the features of their enterprise controllers and SSDs in a recent webcast titled, Go Under the Hood with the Industry's Most Advanced PCIe 4.0 SSDs.

Designed with industry-leading **Reliability and Quality**



Reliability is something that the Solidigm brand is built on, and that customers trust. Why is this important? You can have high reliability, data protection, AND high-performance by using a solution like the SupremeRAID™ card. Silent data corruption on drives like the D5-P5316 is simply not tolerated and handled by the drive firmware. There is no reason to kill your performance by using a solution like ZFS that was made during the time that block devices could not be trusted.

GRAID Technology Inc. is also constantly making updates to their software, like the updated consistency check feature. I tested it on my 10+2 RAID6 array of the P5316 and it completes in just under 6 hours to do a full scan of the 154TB usable volume.

```
sudo graidctl describe consistency_check
✔Describe consistency check successfully.
Schedule Mode:      off
Schedule Base:
Excluded DGs:       []
Policy:             stop_on_error
Next Schedule:
Current Task:       1 DG(s)
                    -DG0: Checking (progress: 12.46%)

                          Start Time: 2022-07-13 16:20:43 +0000 UTC
                          End Time:
```

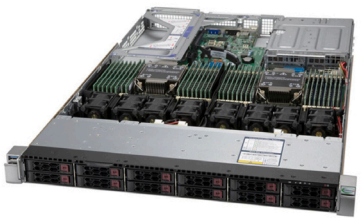
Blast Radius

It was thought that “big drives” don’t work well with RAID 5 because of the “blast radius” or the amount of storage taken offline in the event of a failure. For HDDs this was true, because RAID rebuilds can take days to weeks. Blast radius is a good way to describe failure domains, like failure of a rack takes down all the servers in the rack, and failure of a node takes down all the storage in the node. It is used in the calculation of durability, which is the actual metric that is important in RAID. RAID is meant to protect user data. Durability is the probability of not losing user data, generally expressed in “nines” over an annual period. We can actually use a handy equation to calculate durability if we know the failure rate of an individual drive and the mean time to recovery, which we measure, of rebuild time. The SupremeRAID™ solution offers a full one “nines” better durability at RAID6 than Linux software RAID.

Solution	Data	Parity	Rebuild Time Hours	AFR per device	Durability
GRAID	10	1 (RAID 5)	4.7	0.44%	99.97%
GRAID	10	2 (RAID 6)	4.7	0.44%	7 9’s
Linux sw	10	1 (RAID 5)	20.7	0.44%	99.968%
Linux sw	10	2 (RAID 6)	20.7	0.44%	6 9’s

1U Server Density Leadership

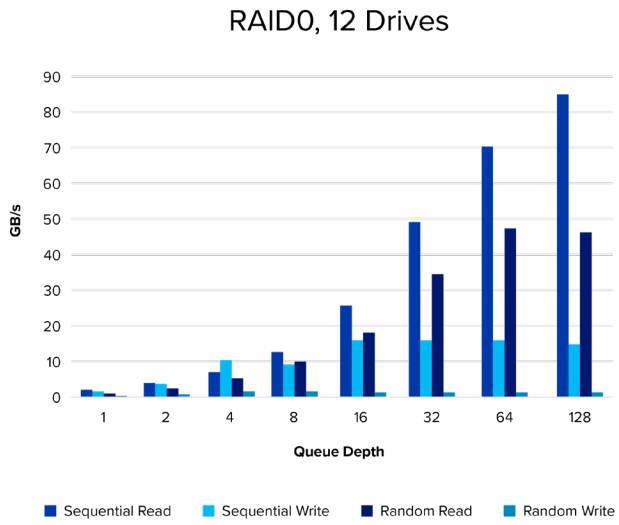
With U.2 we are able to reach 184TB raw with 15.36TB and 368TB raw with 30.72TB capacity in a standard Supermicro 1U server!



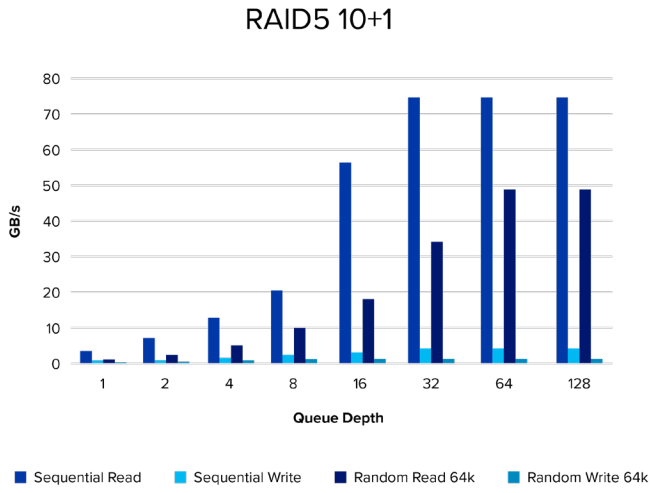
1U Performance Leadership

RAID 0 with 12 drives

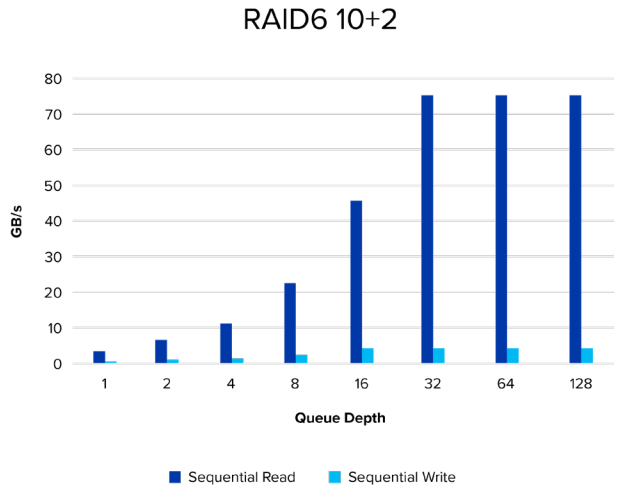
We see RAID 0 scaling almost perfectly on reads, 84GB/s is 7GB/s per drive. The RAID 0 write performance hits 14GB/s, which is very good, but lower than the theoretical capability on these devices. During the testing the team identified potential software improvements on these QLC devices to achieve maximum performance, which you can read about in the “future testing” section.



RAID5 10+1 drives



RAID6 10+2 drives



Future Testing

It is becoming more common on high capacity SSDs to use a high block size on the flash translation layer, as some vendors call “indirection unit”, which Intel describes as Achieving Optimal Performance & Endurance on Coarse-grained Indirection Unit SSDs. This allows the SSD vendor to significantly reduce the amount of DRAM on the SSD and reduce the cost per gigabyte of the drives, allowing for even better TCO. Good performance can be achieved by optimizing the workload to have higher block size or more sequential pattern for writes; reads are not impacted. The Solidigm and GRAID Technology Inc. teams have identified some potential software improvements in the future to fully take advantage of SSDs like the D5-P5316 write performance and workload patterns.

Conclusion

Can we have data protection, high-performance, AND high-capacity? Yes. No longer are we forced to choose between reliability and performance. As demonstrated in this case study, the SupremeRAID™ solution delivers both.

With massive capacities of up to 368TB raw in a 1U server or 737TB raw in a standard 2U server leveraging high capacity QLC based NVMe drives, SupremeRAID™ delivers outstanding RAID5 performance, great rebuild times and high durability with a full “9” extra durability over Linux software RAID. SupremeRAID™ has the needed flexibility to operate with many different filesystems for applications that require specific features, so you aren’t locked into a single solution. We observed how SupremeRAID™ can scale performance well above and beyond the competition in software RAID while completely offloading the CPU. SupremeRAID™ is going to be a really good fit for storage use cases in standard enterprise applications.

System Configuration

Server

- Ultra SuperServer SYS-120U-TNR
- 12 NVMe U.2 in 1U Server
- SupremeRAID™ SR-1010
- Solidigm D5-P5316 15.36TB, x11
- RAID5 10+1

Software

- Ubuntu 20.04, 5.4.0-121-generic kernel
- https://github.com/facebookincubator/FioSynth
- https://github.com/facebookincubator/FioSynth/blob/main/wkldsuites/RAIDPeakWklds modified for 64k random read / write instead of 4k
- graidctl 1.2.2-84.ge4ceb0b3

Copyright © 2021–2022 GRAID Technology Inc. All Rights Reserved. SupremeRAID™ is among the trademarks of GRAID Technology Inc. and/or its affiliates in the United States, certain other countries, and/or the EU. For more information, please visit www.graidtech.com. GRAID Technology Inc. reserves the right to make changes without further notice to any products or data described herein. Information provided by GRAID Technology Inc. is believed to be accurate. However, GRAID Technology Inc. does not assume any liability arising from the use of any application or product described herein, neither does it convey any license under its patent rights nor the rights of others.