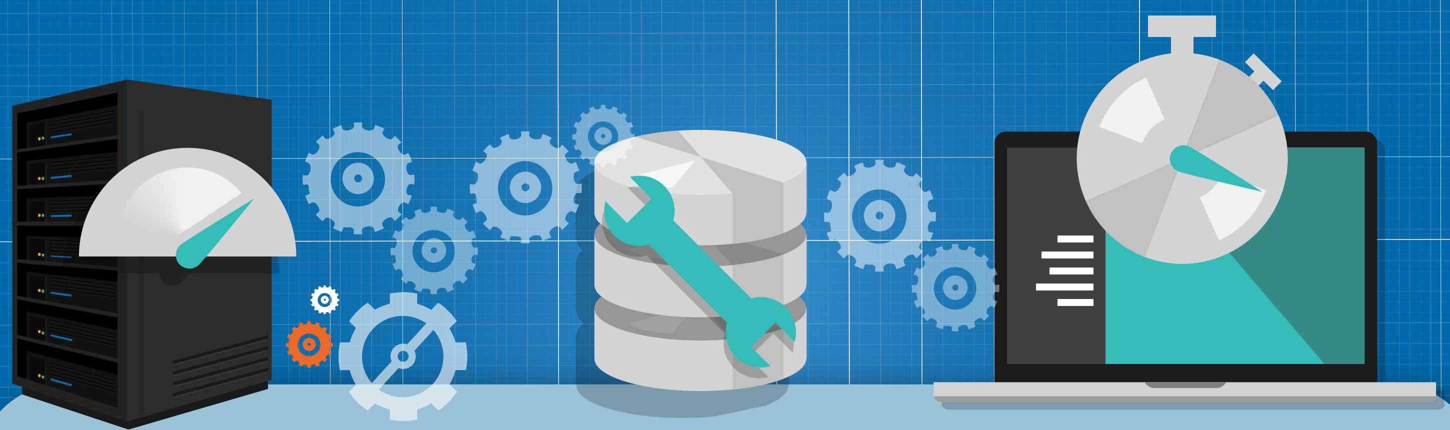
A Pliops Extreme Data Processor (XDP) solution boosted I/O performance compared to a Dell PowerEdge RAID Controller H755N front NVMe solution

Paired with NVMe drives, a Pliops XDP solution delivered more throughput than the Dell PERC H755N front NVMe solution and cut down on RAID volume rebuild times

RAID—or redundant array of independent disks—has long been a method of distributing data across storage to ensure data redundancy and bolster workload uptime. Advancements in drive technology have required RAID solutions to change too, with cards and software solutions on the market now compatible with NVMe™ drives.

In the Principled Technologies data center, we used the Flexible I/O tester (fio) tool to measure I/O performance (in a normal state and while undergoing a rebuild after a drive failure) of two RAID-capable solutions: Pliops Extreme Data Processor (XDP) and the Dell™ PowerEdge RAID Controller (PERC) H755N front NVMe (from the PERC 11 series) in a Dell PowerEdge™ R7525 server. With NVMe drives in RAID 5-like configuration, the Pliops XDP solution delivered more throughput before, during, and after a drive rebuild. This was the case without and with additional redundancy or a hot spare, and with the Pliops XDP solution, rebuild times decreased by half.

Getting the most from NVMe drives with Pliops XDP can help accelerate heavy I/O workloads and reduce rebuild times in case of failures to improve reliability and availability.



Up to **50x** the random write throughput while running a fio workload (at 4KB block size)

Up to **21x** the throughput while undergoing a drive rebuild

Up to **2x** faster to complete a drive rebuild

About Pliops XDP

Pliops XDP is a low-profile PCIe card that makes NVMe RAID possible. The Pliops XDP includes these key features:[1]

- Performance
  - PCIe Gen 3 and Gen 4 NVMe, NVMe-oF, SAS, interface support
- Reliability
  - RAID 5-like Drive Fail Protection (DFP) protects against multiple single drive failures
- Capacity
  - Store up to 128TB of data on 64TB of protected physical capacity
  - Virtual Hot Capacity eliminates the need for a dedicated hot spare
  - Capacity expansion through in-line data compression
- Efficiency
  - Data serialization for increased SSD endurance

Pliops XDP supports drives from many vendors, including Samsung, WD, Micron, Intel, Kioxia, Hynix, Seagate, and more. Pliops reports that XDP is compatible with standard 1U and 2U servers, including those from Dell, HPE, Lenovo, Supermicro, Quanta, Wywinn, Inspur, Sugon, Fujitsu, and Hitachi. To learn more about Pliops XDP, visit https://pliops.com/.

Our test approach

For our testbed, we configured one Dell PowerEdge R7525 server that we used with both RAID-capable solutions we tested. To offer a glimpse into what each solution offers for data availability and bandwidth, we tested I/O performance in two states:

- Use case 1: I/O performance (normal state)
- Use case 2: I/O performance during a drive rebuild
  - Four 8TB NVMe SSDs in a RAID 5 configuration
  - Five 8TB NVMe SSDs in a RAID 5 configuration with additional redundancy (one extra drive for Pliops XDP Virtual Hot Capacity [VHC], or for a PERC hot spare)

To compare I/O performance, we used a tool called fio to create synthetic I/O workloads.

To learn more about our testbed as well as the step-by-step details of our tests, read the science behind the report. Please also note that while we are publishing this report in January 2023, we completed testing in May 2022.

Use case 1: Testing I/O performance (normal state)

First, we used fio to test I/O performance of both RAID-capable solutions in healthy environments to show what performance levels you might expect during normal operations at various block sizes. Smaller (4K and 8K) block sizes are typical of mail servers and many OLTP databases, while OLAP and data warehousing applications can also include I/O at 16K block sizes. Big data or analytics workloads can sometimes use larger block sizes (128K - 512K). While these workloads have different ranges of reads to writes (and we tested 100% random reads and writes), these workload/block size pairings present a general idea of which results might be relevant to your workloads. Figure 1 compares the random read bandwidth that the solutions achieved while in a healthy state (with no drive failures). Across block sizes, the Pliops XDP solution achieved significantly higher read bandwidth, delivering up to three times the read bandwidth of the Dell PERC H755N front NVMe solution (using a block size of 128).

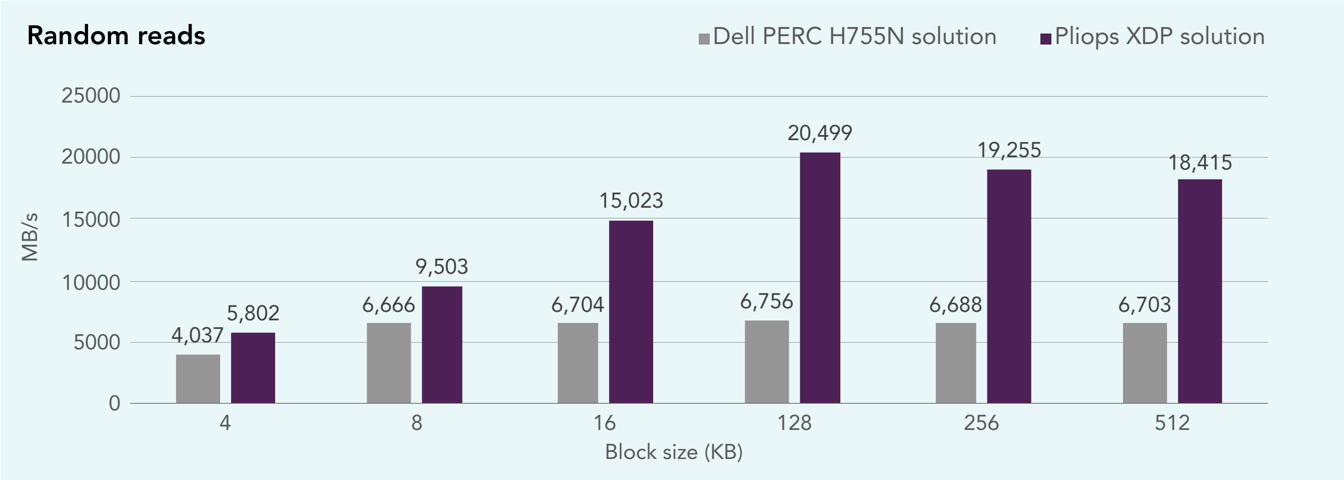


Figure 1: Random read bandwidth, in MB/s, at multiple block sizes for the Pliops XDP and Dell PERC H755N front NVMe solutions. Higher numbers are better. Source: Principled Technologies.

Random write performance differences were even more pronounced. As Figure 2 shows, the Pliops XDP solution achieved up to 50 times the random write bandwidth of the Dell PERC H755N front NVMe solution (at a 4K block size). Write bandwidth on the Pliops XDP solution was consistent across block sizes, while the Dell PERC H755N front NVMe solution gradually increased bandwidth achieved as block sizes increased. At the highest bandwidth mark for the Dell PERC H755N, the Pliops XDP solution still delivered over three times the bandwidth.

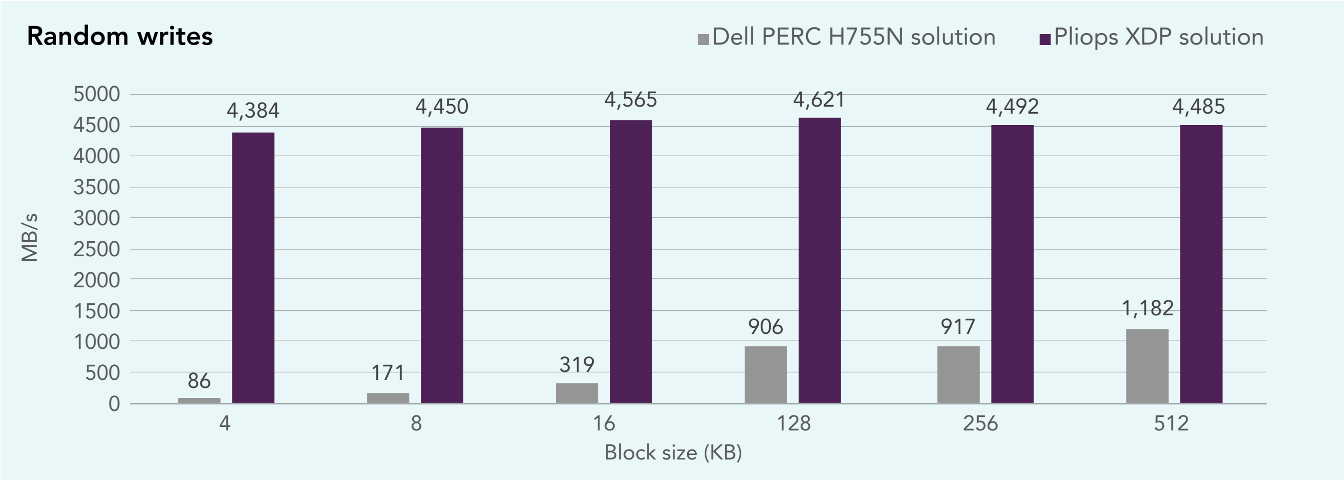


Figure 2: Random write bandwidth, in MB/s, at multiple block sizes for the Pliops XDP and Dell PERC H755N front NVMe solutions. Higher numbers are better. Source: Principled Technologies.

Use case 2: Testing I/O performance during a drive rebuild

Four 8TB NVMe SSD drives in a RAID 5-like configuration

We tested performance (using a 70/30 random read/write mix) with four NVMe drives in a RAID 5-like configuration with no VHC or hot spare. We set rebuild priority to the highest setting for each controller to reflect a use case where data redundancy is more important than I/O performance.

Table 1 presents the bandwidth and rebuild times during fio tests for both the Pliops XDP solution and the Dell PERC H755N front NVMe solution with a four-drive RAID 5-like configuration with high-priority rebuild settings.

Before the fault, the Pliops XDP solution provided 13 times the bandwidth of the Dell PERC H755N front NVMe solution. While in its faulted, pre-rebuild state, the Pliops XDP solution achieved 18 times the MB/s that the Dell PERC H755N front NVMe solution did.

As Figure 3 shows, during the rebuild, the Pliops XDP solution achieved 17 times the throughput. After the rebuild, the Pliops XDP solution continued to provide more bandwidth than the Dell PERC H755N front NVMe solution, delivering 11 times the throughput after returning to a healthy state.

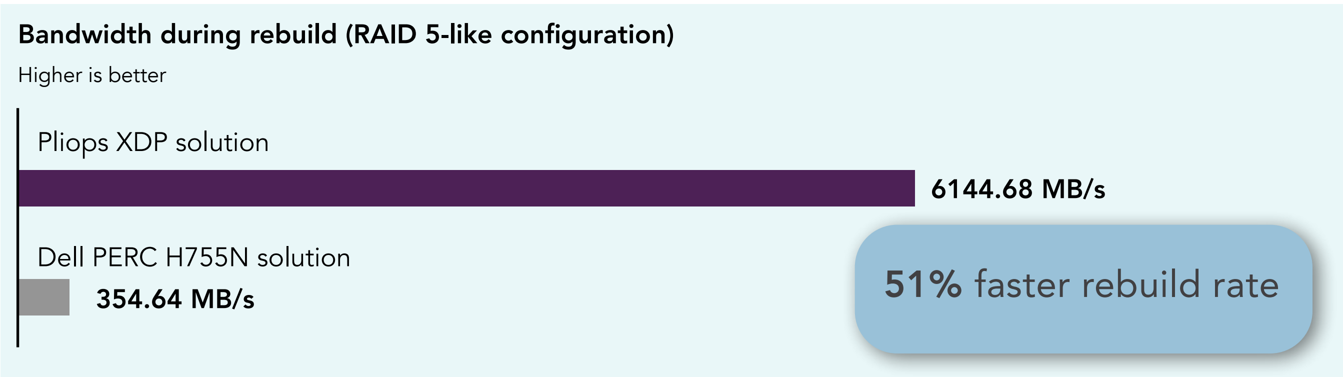


Figure 3: Bandwidth, in MB/s, during a drive rebuild (four-drive RAID 5 array with high-priority rebuild settings). Higher numbers are better. Source: Principled Technologies.

Using Pliops XDP, the rebuild took 51 percent less time than it did on the Dell PERC H755N front NVMe solution. Because rebuild times depend on the amount of data processed, and thus on the size of the array, we also present the rebuild time normalized by the amount of usable data in the array (minutes per terabyte). This metric could be used to estimate rebuild time for other arrays. The Pliops XDP solution rebuilt at a rate of over 39 minutes per TB, while the Dell PERC H755N front NVMe solution took 82 minutes per TB.

Table 2: Bandwidth test results using fio for both RAID solutions with a four-drive RAID 5 array with high-priority rebuild settings. Source: Principled Technologies.

	Pliops XDP solution			Dell PERC H755N front NVMe solution		
Stage	Bandwidth (MB/s)			Bandwidth (MB/s)		
	Read	Write	Total	Read	Write	Total
Before fault	7,077.10	3,032.54	10,109.63	520.86*	223.86*	744.72*
Faulted (pre-rebuild)	4,941.46	2,117.55	7,059.01	268.50	115.00	383.50
During rebuild	4,301.21	1,843.47	6,144.68	247.95	106.67	354.62
After rebuild	7,642.94	3,276.22	10,919.16	651.23	278.13	929.36
Rebuild time	5 hrs 18 min			11 hrs		
Rebuild rate	39.7 min/TB			82 min/TB		

*PERC H755N performance prior to triggering the rebuild was lower than expected due to the controller card performing a background sync. We would expect performance to be comparable to the after-rebuild performance, as shown in the low-priority rebuild test where a background sync did not occur.

Five 8TB NVMe SSD drives in a RAID 5-like configuration with additional redundancy

We destroyed the array and created a new one providing additional redundancy. For the Pliops XDP solution we created a five-drive RAID 5-like array with VHC, and for the Dell PERC H755N front NVMe solution, we created a four-drive RAID 5 array with an additional drive as a hot spare. Table 2 presents the bandwidth results and rebuild times during fio tests for both the Pliops XDP solution and the Dell PERC H755N solution with a four-drive RAID 5-like configuration and hot spare with high-priority rebuild settings. Before the fault, the Pliops XDP solution provided 16 times the bandwidth of the Dell PERC H755N front NVMe solution.

As Figure 4 shows, during the drive rebuild, the Pliops XDP solution achieved 21 times the MB/s that the Dell PERC H755N front NVMe solution did. Post-rebuild, the Pliops XDP solution offered 12 times the throughput of the Dell PERC H755N front NVMe solution.

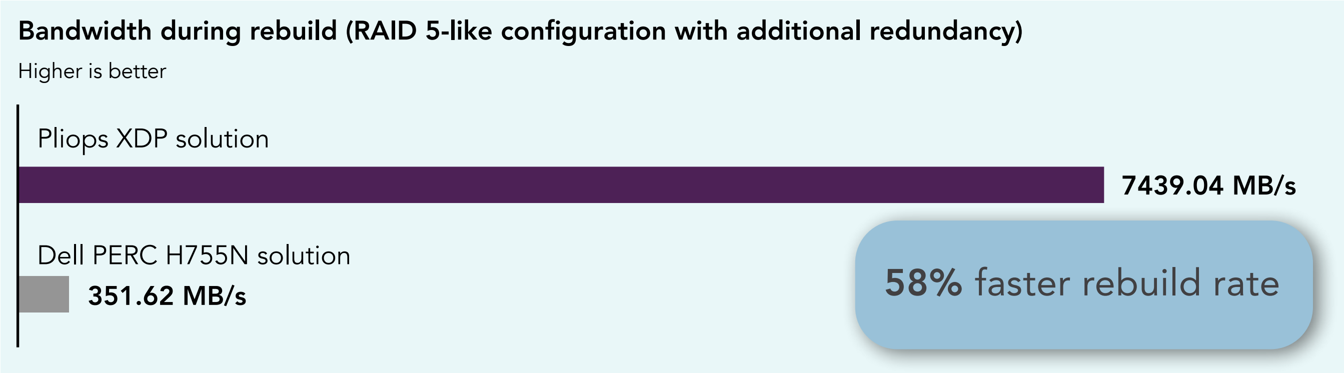


Figure 4: Bandwidth, in MB/s, during a drive rebuild (RAID 5 array with additional redundancy with high-priority rebuild settings). Higher numbers are better. Source: Principled Technologies.

Using additional redundancy resulted in the fastest rebuild times for both solutions, dropping to just over 4 hours for the Pliops XDP solution and 10 hours for the Dell PERC H755N front NVMe solution—a 58 percent reduction in time for the Pliops XDP solution. The Pliops XDP solution rebuilt at a rate of 31 minutes per TB, while the Dell PERC H755N had a rebuild rate of 75 minutes per TB—more than twice as long to return to a normal, healthy state.

Table 3: Bandwidth test results using fio for both RAID solutions with additional redundancy for the RAID 5 arrays with high priority rebuild settings. Source: Principled Technologies.

	Pliops XDP solution			Dell PERC H755N front NVMe solution		
Stage	Bandwidth (MB/s)			Bandwidth (MB/s)		
	Read	Write	Total	Read	Write	Total
Before fault	8,103.94	3,473.67	11,577.61	505.21*	217.38*	722.59*
During rebuild	5,206.91	2,232.13	7,439.04	245.87	105.75	351.62
After rebuild	7,887.18	3,379.75	11,266.92	644.61	278.33	922.94
Rebuild time	4 hrs 11 min			10 hrs 5 min		
Rebuild rate	31.5 min/TB			75 min/TB		

*PERC H755N performance prior to triggering the rebuild was lower than expected due to the controller card performing a background sync. We would expect performance to be comparable to the after-rebuild performance, as shown in the low-priority rebuild test where a background sync did not occur.

Table 4: Summary of performance wins that Pliops XDP offered in our tests.

	Pliops XDP offered:
Normal performance (random read)	Up to 3x the throughput
Normal performance (random write)	Up to 50x the throughput
Performance during rebuild (RAID 5-like state)	17x the throughput
Rebuild duration (RAID 5-like state)	51% less time / 2x faster
Performance during rebuild (with VHC or hot spare)	21x the throughput
Rebuild duration (with VHC or hot spare)	58% less time / 2x faster

Conclusion

Vendors are innovating to provide NVMe RAID with increasingly strong I/O performance that helps administrators take better advantage of the greater I/O and bandwidth NVMe drives can offer, paired with the data resiliency of RAID. In our tests, the Pliops XDP solution accelerated our fio workload to deliver up to 50 times the write throughput of a Dell PERC H755N front NVMe solution in an otherwise identically configured server. When we simulated a drive failure, the Pliops XDP solution delivered up to 21 times the throughput while undergoing a drive rebuild. Plus, the Pliops XDP solution consistently took less than half the time to complete a drive rebuild after failure, taking as much as 58 percent less time compared to the Dell PERC H755N front NVMe solution. By completing RAID volume rebuilds in less time using Pliops XDP, organizations can reduce the time that resources are diverted from workloads and improve high availability to meet SLAs.

1. Pliops, "Pliops Extreme Data Processor Data Sheet," accessed May 22, 2022, https://pliops.com/resource/pliops-extreme-data-processor-data-sheet/.

We concluded our hands-on testing on May 17, 2022. During testing, we determined the appropriate hardware and software configurations and applied updates as they became available. The results in this report reflect configurations that we finalized on May 4, 2022 or earlier. Unavoidably, these configurations may not represent the latest versions available when this report appears.

System configuration information

Table 5: Detailed information on the system we tested.

Server configuration information	Dell PowerEdge R7525
BIOS name and version	Dell 2.6.6
Non-default BIOS settings	Hyperthreading disabled
Operating system name and version/build number	Red Hat Enterprise Linux® 8.5, 4.18.0-348.7.1.el8_5.x86_64
Date of last OS updates/patches applied	4 May 2022
Power management policy	Performance Per Watt
Processor	
Number of processors	2
Vendor and model	AMD EPYC™ 7262
Core count (per processor)	8
CPU threads per core (hyperthreading)	1
Core frequency (GHz)	3.20
Stepping	0
Memory module(s)	
Total memory in system (GB)	128
Number of memory modules	16
Vendor and model	Micron® 9ASF1G72PZ-3G2E2t
Size (GB)	8
Type	PC4-25600
Speed (MT/s)	3,200
Speed running in the server (MT/s)	3,200
Storage controller	
Vendor and model	Dell PERC H755N front NVMe
Cache size	8GB
Firmware version	52.16.1-4158
Driver version	megaraid_sas 4.18.0-348.7.1.el8_5.x86_64

Server configuration information	Dell PowerEdge R7525
Local storage (type A)	
Purpose	OS
Number of drives	1
Drive vendor and model	Dell Express Flash CD5 960G SFF
Drive size (GB)	960
Firmware	1.2.0
Drive information (speed, interface, type)	NVM Express 1.3
Local storage (type B)	
Purpose	Drives under test
Number of drives	6
Drive vendor and model	Intel® SSDPE2KX080T8
Drive size (GB)	8,000
Firmware	VDV10131
Drive information (speed, interface, type)	NVMe-MI1.0
Network adapter	
Vendor and model	Broadcom® Gigabit Ethernet BCM5720
Number and type of ports	2 x 1GbE
Driver version	tg3, 4.18.0-348.7.1.el8_5.x86_64
Cooling fans	
Vendor and model	Dell System Board Fa4
Number of cooling fans	6
Power supplies	
Vendor and model	Dell Delta 800W PSU
Number of power supplies	2
Wattage of each (W)	800

How we tested

Configuring the server

We applied all Dell firmware updates to the server as of 1 March 2022. We set all BIOS settings to their defaults, but disabled hyperthreading for the processor.

During our tests, we had only a single RAID controller card in the server at a time: the Extreme Data Processor (XDP) card was absent during tests of the Dell PERC H755N front NVMe controller, and vice versa.

The motherboard has four NVMe ports for storage drives, two ports for each processor, and each card used the ports differently. The Dell PERC H755N front NVMe controller came factory configured with a single NVMe cable connecting the controller to the motherboard at a port associated with processor 0. The Dell PERC H755N front NVMe controller serviced six drives: five 8TB Intel NVMe drives for the test volumes, and one 960GB Dell drive for the OS. When configuring the Pliops XDP card, we installed it in riser 1 (associated with processor 0), which provided it with one x16 Gen4 connection. We attached three NVMe cables from the drive enclosure to NVMe ports, with each cable servicing two drives. Two cables connected to NVMe ports associated with processor 1, and one to an NVMe port associated with processor 0.

Installing and configuring the OS

Note: We followed the instructions in XDP Pliops Extreme Data Processor User Guide, version 1.7.6, for preparing the OS for the XDP.

We installed Red Hat Enterprise Linux (RHEL) 8.5 and applied all OEM patches available as of 4 May 2022. At the request of the client, we downgraded the Linux kernel to version 4.18.0-348.7.1.el8_5. We added the iommu=soft to the kernel's boot command line, and rebooted.

We installed the additional RPMs: nvme-cli, libpciaccess, libaio, pciutils, dmsetup, device-mapper, sysstat, and tuned.

We enabled `tuned` and set its profile to throughput-performance.

We expanded the Pliops software archive v1.20.3.1_redhat8.5.tar.gz.

We allocated 26,000 2MB hugepages assigned to NUMA node 0, using the `pliops-hugepages.service` method.

We installed Dell's PERC command-line interface perccli, version 007.1910.0000.0000, to operate on the Dell PERC H755N front NVMe card.

Finally, we installed RHEL fio, version 3.19.

Summary of the test method

See the Pliops XDP and Dell PERC H755N front NVMe sections below for specific details; the general steps we performed for each device were as follows.

Four-drive RAID 5 array (with no VHC or hot-spare drive)

1. Create the RAID 5 array. Disable autorebuild.
2. Determine the usable size of the array and precondition the array using `fio-tester` with the workloads specified in the accompanying CSV file, replacing its fio-size parameters with the array's size.
3. From this same fio-tester run, gather performance metrics for approximately 735 FIO workloads, as specified in the client-supplied CSV file. These metrics are the read and write bandwidths.
4. (Step 5 test) We performed the faulted drive test for high rebuild priority using the bash script `Case1_Step5_HighPriority_IOstat.sh`. We plotted and analyzed the iostat output to determine the average read and write bandwidths for each phase of the test (optimal, three-drive degraded, rebuilding, restored).
5. (Step 3 test) We performed the faulted drive test for low rebuild priority using the bash script `Case1_Step5_HighPriority_IOstat.sh`. We plotted and analyzed the iostat output to determine the average read and write bandwidths for each phase of the test (optimal, three-drive degraded, rebuilding, restored).

Four-drive RAID 5 array with additional redundancy (with VHC or hot-spare drive)

1. Create the RAID 5 array. Disable autorebuild.
2. Determine the usable size of the array and precondition the array using fio-tester with the workloads specified in the accompanying CSV file, replacing its fio-size parameters with the array's size.
3. From this same fio-tester run, we also gather performance metrics for approximately 735 FIO workloads, as specified in the client-supplied CSV file. These metrics are the read and write bandwidths.
4. (Step 5 test) We performed the faulted drive test for high rebuild priority using the bash script `Case2_Step5_HighPriority_IOstat.sh`. We plotted and analyzed the iostat output to determine the average read and write bandwidths for each phase of the test (optimal, rebuilding, restored).
5. (Step 3 test) We performed the faulted drive test for low rebuild priority using the bash script `Case2_Step3_LowPriority_IOstat.sh`. We plotted and analyzed the iostat output to determine the average read and write bandwidths for each phase of the test (optimal, rebuilding, restored).

Testing the Pliops XDP solution

We first updated the firmware of the XDP card using xdp-installer, and rebooted the system. The OS presented the XDP card as device `nvme6n1`.

We created the RAID 5 array for use case 1:

```
sudo ./xdp-installer install --level=5 --resources="/dev/nvme1n1 /dev/nvme3n1 /dev/nvme4n1 /dev/
nvme5n1" --vhc=0 --clean
```

We created the array for use case 2:

```
sudo ./xdp-installer install --level=5 --resources="/dev/nvme0n1 /dev/nvme1n1 /dev/nvme3n1 /dev/
nvme4n1 /dev/nvme5n1" --vhc=1 --clean
```

We performed the tests for the two arrays as described in the section Summary of the test method.

Testing Dell PERC H755N front NVMe controller

For the Dell PERC H755N font NVMe controller, we followed the same procedure and used the same tools and inputs for the preconditioning and performance workloads tests, only changing the size of the usable array. The Dell PERC command-line interface perccli tool does not have a command to fault a drive. Instead, we used the tool to take one drive offline, and alter the array to a degraded state.

We created the RAID 5 array for use case 1 and disabled auto-rebuild with `percli`:

```
sudo perccli /c0 add vd r5 name=test01 drives=252:2,3,6,7
sudo perccli /c0 set autorebuild=off
```

We created the array for use case 2 with `percli`:

```
sudo perccli /c0 add vd r5 name=test02 drives=252:2,3,6,7 spares=252:1
sudo perccli /c0 set autorebuild=on
```

We set the following device parameters to improve performance by running the following script with the array's block device name (e.g., "sda" or "sdb") as an argument:

```
#!/bin/bash
DEVICE="$1"
if [[ -b "/dev/$DEVICE" ]]; then
  echo "Using /dev/$DEVICE"
else
  echo "Device /dev/$DEVICE cannot be used as a block file."
  exit 1
fi

echo 1024 > /sys/block/$DEVICE/device/queue_depth
echo    0 > /sys/block/$DEVICE/queue/nomerges
echo 2048 > /sys/block/$DEVICE/queue/nr_requests
echo    0 > /sys/block/$DEVICE/queue/rq_affinity
echo none > /sys/block/$DEVICE/queue/scheduler
```

We set the rebuild priority for the array:

```
perccli /c0 set rebuildrate=0       # low priority
perccli /c0 set rebuildrate=100     # high priority
```

We took one drive offline:

```
sudo perccli /c0/e252/s2 set offline
```

We performed the tests for the two arrays as described in the section Summary of the test method.

This project was commissioned by Pliops.