Western Digital.

White Paper

SupremeRAID™ with OpenFlex™ Data24

Abstract

This white paper explores the performance benchmarks of SupremeRAID SR-1000 NVME-oF™ RAID card in conjunction with the Western Digital OpenFlex Data24 NVMe-oF Storage Platform.

July 2023

Table of Contents

Introduction

In Software-Composable Infrastructure (SCI), compute, storage, and networking resources are abstracted from their physical locations and are usually managed with software via a web-based interface. SCI makes data center resources as readily available as cloud services and is the foundation for private and hybrid cloud solutions. With the emergence of NVMe™ SSD and NVMe- oF™ technologies, SCI can disaggregate storage resources without sacrificing performance and latency. As NVMe SSD technology rapidly evolves, a significant performance bottleneck is introduced — RAID data protection.

RAID Configurations

In performing RAID computations, the user has historically had the following two options:

- O.S. Software RAID (e.g., MDADM on Linux®)
- Hardware RAID (e.g., a RAID Controller Card)

Software RAID

OS Software RAID provides an independent solution that can work with multiple media types (HDD or SSD) and protocols (SATA, SAS, NVMe). The challenge with O.S. Software RAID is generally poor performance with a high cost for CPU resources. Sequential bandwidth especially Read bandwidth, can achieve high-performance levels, but sequential writes require protection computations. Small block I/O patterns generally have even lower RAID performance levels to render this option generally usable. In summary, this option has the protocol independence needed on network-attached storage devices but lacks the required performance.

Hardware RAID

Hardware RAID was convenient because the SAS adapter card could provide it to the client who was in line with the storage housed in an external enclosure. In the HDD era, a simple ASIC on a RAID card was capable enough to handle all I/O – after all, even with SAS HDD, maximum performance was only around 200 IOPS and 150MB/s of throughput. However, a single NVMe SSD can now deliver around 1 M IOPS and 7Gb/s of throughput.

The hardware RAID Cards were slow to adapt from slower HDDs to higher performing NVMe SSDs. That transition has primarily occurred and can provide higher performance levels when using SSDs. The challenge with these RAID adapters is that they can only be used with their native physical protocols. They cannot be used with network-attached devices and don't scale performance fully or efficiently. In summary, these adapters can potentially have the needed local performance but do not offer protocol independence to work on network-attached devices, severely limiting their usefulness in modern Software-Composable Infrastructures or high-performance applications. These considerations also prevented their testing in these benchmarks.

In this paper, we discuss and benchmark a third option: Hardware-Accelerated Software RAID. This option provides protocol independence and the high performance needed for network-attached Flash storage.

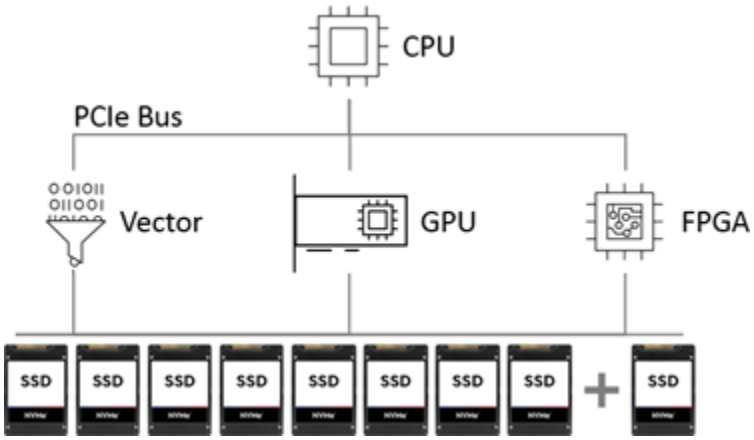
GPU-based Hardware Accelerated Software RAID

The challenge of implementing complex RAID levels such as 5 and 6 while maintaining high performance on NVMe drives is usually parity calculations. Hardware RAID parity calculations use a hardware engine within the ASIC, while software RAID can only use the CPU's instruction set, whose performance is often limited.

Offloading and parallelizing the CPU-intensive parity calculations onto a hardware accelerator often addresses this issue. There are a few potential hardware engines where these calculations can take place. The first option would be to utilize CPU extensions (e.g., Vector/SIMD) to offload and parallelize the parity calculations to improve RAID performance. A second option would be to offload and parallelize these calculations on dedicated hardware accelerators such as GPUs (DPUs) or FPGAs.

Graid Technology Inc. provides the GPU-based RAID solution tested in this project, the SupremeRAID™ SR-1000. The following figure provides a block diagram of its implementation.

Block Diagram of Advanced RAID Solutions (Vector, GPU, FPGA)



RAID Hardware Elements

While GPU-based solutions are promising, each server requires a GPU. Commercially available solutions of these technologies, at the time of writing were limited, but several Vector and FPGA solutions are available.

Solution Components

For this project, we chose the SupremeRAID SR-1000 NVMe-oF RAID card for performance benchmark in conjunction with the Western Digital OpenFlex Data24 NVMe-oF Storage Platform.

Hardware and Software Summary

The following table provides a list of components used in this test.

Load Generating Servers	Quantity	Description
Platform	6	Lenovo® ThinkSystem™ SR650
Processor	2	Intel® 6154 200TDP 18-Core 3.0GHz
Memory	12	32GiB @ 2666MHz (384GiB)
Fabric	1	ConnectX®-5 100 Gb Ethernet HCA
RAID	1	SupremeRAID SR-1000
NVMe-oF Storage	**Quantity**	**Description**
Enclosure	1	Western Digital OpenFlex Data24 (FW4.0)
NVMe	24	Ultrastar® DC SN840 3.2TB (FW03)
Software	**Quantity**	**Description**
OS	1	RHEL 8.4
Kernel	1	4.18.0-305.25.1.el8_4.x86_64
MOFED	1	In-Box Mellanox 5.6

OpenFlex Data24 NVMe-oF Storage Platform

Western Digital's OpenFlex Data24 NVMe-oF Storage Platform is similar to a 2.5" SAS JBOD Enclosure. It provides 24 slots for NVMe drives and a maximum capacity of 368 TB[1] when using Western Digital Ultrastar DC SN840 15.36 TB devices. Unlike a SAS enclosure, the Data24's dual IO modules use Western Digital RapidFlex™ C1000 NVMe-oF Controllers. These controllers allow full access to all 24 NVMe drives over up to six ports of 100 Gb Ethernet.

The Data24 is a close replacement for the traditional SAS enclosures. However, the Data24 offers a significant benefit over these enclosures: the ability to integrate directly into Ethernet fabric, allowing for an Any-to-Any mapping of Object Storage Targets to Object Storage Servers.

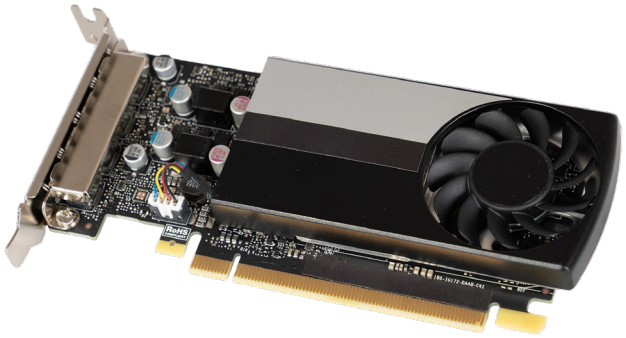
The OpenFlex Data24 design exposes the full performance of the NVMe SSDs to the network. With 24 Western Digital Ultrastar DC SN840 3.2 TB devices, the enclosure can achieve up to 71 GB/s of bandwidth and over 15 MIOPS at a 4K block size.

OpenFlex Data24 NVMe-oF Enclosure

SupremeRAID SR-1000 NVMe-oF RAID

The SupremeRAID SR-1000 PCIe 3.0 adapter delivers SSD performance in AI-accelerated compute, All Flash Array (AFA), and High Performance Computing (HPC) applications. Designed for both Linux and Windows® operating systems, it supports RAID levels 0/1/10/5/6/JBOD, while the core software license supports up to 32 native NVMe drives.

The SupremeRAID SR-1000 enables NVMe/NVMe-oF, SAS, and SATA performance while increasing scalability, improving flexibility, and lowering TCO. This solution eliminates the traditional RAID bottleneck in mass storage to deliver maximum SSD performance for high-intensity workloads. The following table displays Spec Sheet Data.



SupremeRAID™ SR-1000 for PCIe 3, 4, and 5 servers

SupremeRAID SR-1000 Spec Sheet Data[2]

Workload	SupremeRAID™ SR-1000	Number of Drives	Performance per Drive
4K Random Read	16.00 MIOPS	12	1.33 MIOPS
4K Random Write	0.75 MIOPS	12	0.06 MIOPS
512K Sequential Read	110.00 GB/s	20	5.50 GB/s
512K Sequential Write	11.00 GB/S	20	0.55 GB/S
4K Random Read in Rebuild	3.00 MIOPS	12	0.25 MIOPS

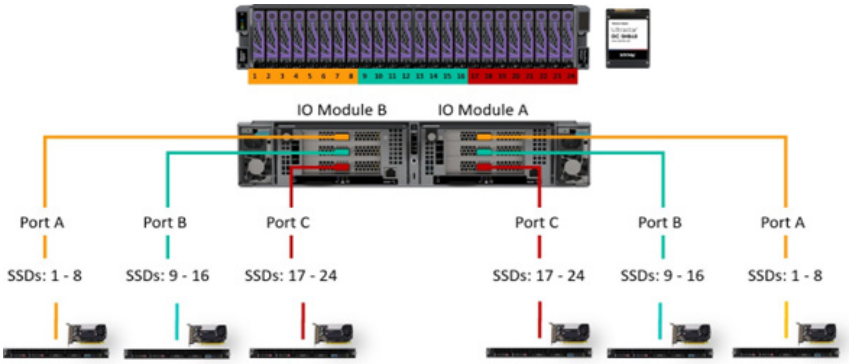
[1] One terabyte (TB) is equal to one trillion bytes. Actual user capacity may be less due to operating environment.
[2] Software: Linux Version: CentOS 8.5 | Hardware: CPU: Intel Xeon Gold 6338 CPU 32-Core with 2.0GHz x 2; Memory: SK Hynix HMA82GR7CJR8N-XN DIMM DDR4 3200 MHz 16GiB x 16; SSD: INTEL D7-P5510 SSDPF2KX038TZ 3.8TB x 20 | RAID Configuration: Random performance based on a drive group with 12 physical drives and 1 virtual drive; sequential performance based on a drive group with 20 physical drives and 1 virtual drive.

Benchmarking Infrastructure

The OpenFlex Data24

Available with one to three ports per IO Module. There are two IOM modules per Data24. The ratio of ports to IOM module will influence drive-to-port mapping options.

The configuration used for this benchmark is three ports per IO Module. In this configuration, a maximum of 8 physical drives are accessible per IO Module port. Each physical device can have up to 8 namespaces. Each device has a pair of ports, one port per IO Module displayed in the following figure.



Data24 6x8 with 24 SSDs with 2 Namespaces Each

- The Western Digital Ultrastar DC SN840 devices are dual-ported NVMe drives. This architecture allows both paths to the device to be used, maximizing the performance potential of that device.
- Each of the six servers includes a SupremeRAID SR-1000 and Mellanox® CX5 RDMA network interface card (RNIC).
- This configuration allowed for a single path to a front-end IO Module port and the eight physical drives presented by that port. In this instance, each device has two namespaces. Each server pair that access a shared device is assigned one of the device's two namespaces.
- This configuration is considered non-HA; therefore, this benchmark used no redundant paths or multipathing.
- The server connects directly or via a switch. There is minimal performance impact with either implementation.

Benchmarking Methodology

Flexible IO (FIO) is the workload generator. The SupremeRAID SR-1000 solution uses the standard OpenFlex Data24 Spec Sheet Process.

Fundamentally, the process has two phases – the sequential process using 128K blocks (to measure bandwidth) and the random process using 4K blocks (to measure IOPS).

We ran three instances of the tests and averaged the results.

Also, we checked for excessive variability using the Coefficient of Variation (COV). Any extreme variability is investigated and resolved. Additional tests may be required if a clear cause exists, such as a test error, external interruption, etc.

Spec Sheet Sequential (SSS) Process 128K	
NVMe Format	
Conditioning	2x Sequential Fills
Figures of Merit: Steady State Performance	1. Sequential Write 2. Sequential Read 20m each

General Comments:

- The goal of this testing is reliable repeatable tests demonstrating the performance of the solution.
- All tests are run three times followed by a variability analysis. If too variable, we research the cause and possibly run additional tests.
- Conditioning establishes a well-known starting point and is a mandatory step in the process
- Conditioning is not a goal but a means to an end (reliable, repeatable tests).
- fio and fiodb are the workload generator and test harness, respectively.

Spec Sheet Random (SSR) Process 4K	
NVMe Format	
Conditioning	2x Sequential Fills
Figures of Interest: Peak Performance	1. Random Write 2. Random Mixed (70/30) 3. Random Read 4. Random Write QOS QD1 5. Random Read QOS QD1 5m each
Conditioning	1x Random Fill
Figures of Merit: Steady State Performance	1. Random Write 2. Random Mixed (70/30) 3. Random Read 4. Random Write QOS QD1 5. Random Read QOS QD1 5m each

Measured Performance – 4K Random IO

The first BASELINE tests (without simulated failures) were run with FIO and tested 4K Random Read, Random Mixed, and Random Write. The following figure shows the results.

BASELINE

Six servers with each connected to eight namespaces without RAID established a performance BASELINE. Aggregate results of the BASELINE show 15.3 million IOPS for 4K Random Reads, 12 million IOPS for 4K Random Mixed, and 6.26 million IOPS for Random Writes. These results are as expected and therefore serve as a good BASELINE for Random IO Tests. We compare all RAID Results to the BASELINE.

SupremeRAID RAID 5

We ran the same tests with the SupremeRAID SR-1000 Solution. We created a single eight namespace RAID 5 (7+1) set on each server, with the aggregate results showing 15.3 million IOPS for 4K Random Read, 6.17 million IOPS for 4K Random Mixed, and 2.6 million IOPS for 4K Random Writes.

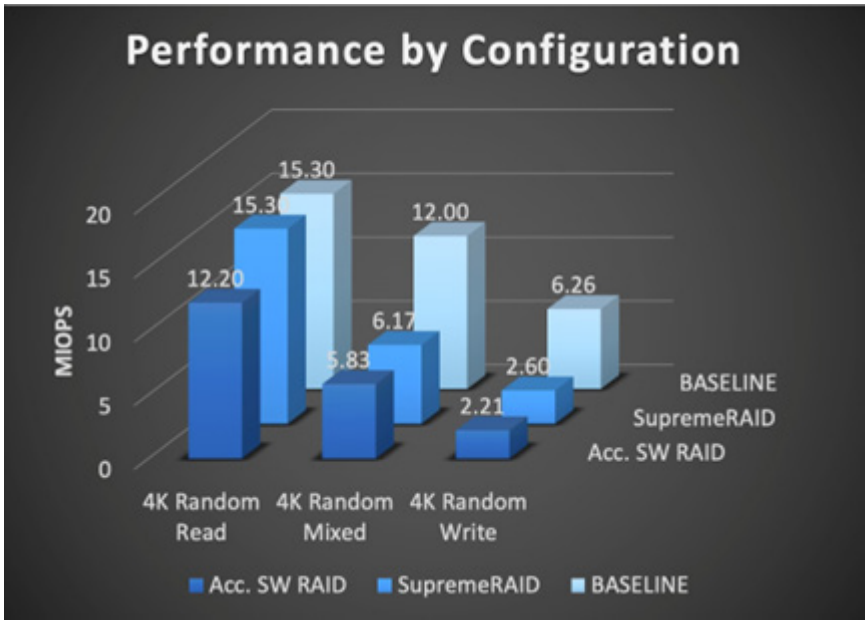
The Random Read IOPS matches the OpenFlex Data24 BASELINE results and demonstrates the SupremeRAID non-blocking architecture with a 4K Random Read workload while validating the test infrastructure. Random Mixed and Random Write workloads showed the expected performance drops associated with RAID 5. The read-modify-write (parity) calculations have an unavoidable compute cost and delay.

Advanced Software RAID Solution

We also tested another third-party advanced software-based RAID solution (that exploits advanced CPU instruction set features).

Again, we created a RAID 5 (7+1) set on the six servers, with the aggregate results showing 12.2 million IOPS for 4K Random Reads, 5.83 million IOPS for 4K Random Mixed, and 2.21 million IOPS for Random Writes.

In all instances, the advanced software RAID solution was less performant than the SupremeRAID GPU solution.



4K Random IO Benchmarks

Throughput	Workload					
Solution	4K Random Read		4K Random Mixed		4K Random Write	
	MIOPS	Solution/Baseline	MIOPS	Solution/Baseline	MIOPS	Solution/Baseline
Baseline	15.30	100%	12.00	100%	6.26	100%
SupremeRAID RS	15.30	100%	6.17	51%	2.60	42%
ADVSW RS	12.20	80%	5.83	49%	2.21	35%

Measured Performance – 128K Sequential IO

Next, we ran the large block (128K) Spec Sheet Sequential Benchmark on the exact configuration previously described above. The results are shown below in the following figure.

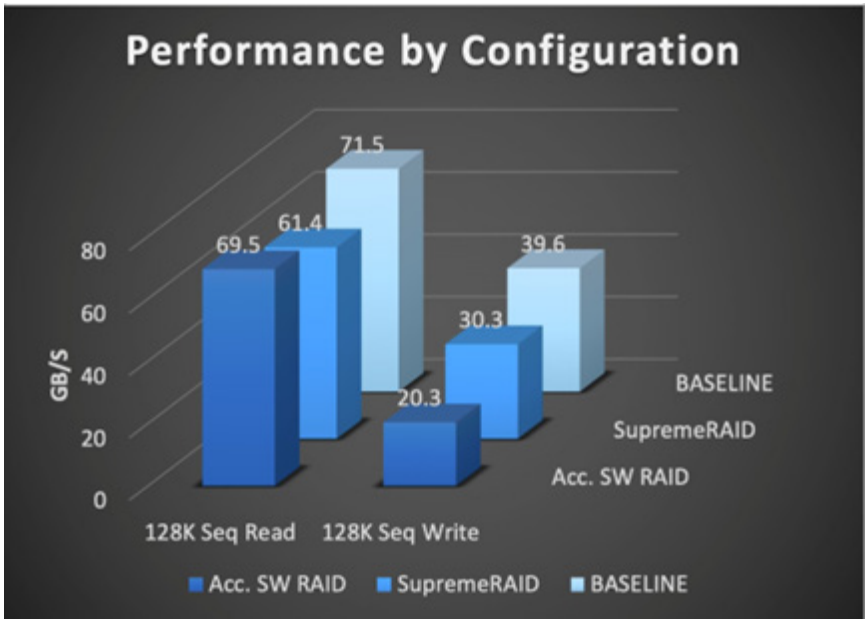
BASELINE

The Spec Sheet Sequential BASELINE achieved 71.5 GB/s for 128K Sequential Reads and 39.6 GB/s for 128K sequential writes. These results are as expected for the Data24 and therefore serve as a good BASELINE for Sequential IO Tests. We compare all RAID Results to the BASELINE.

SupremeRAID RAID 5

In this instance, the 128K Sequential Read results are 61.4 GB/s which is lower than the 71.5 GB/s demonstrated in the BASELINE, and these results are 12% slower when compared to the advanced software RAID results below. This slowdown isn't due to parity computation as there is none for Reads. The lower SupremeRAID Sequential Read Performance is because all data flows from the GPU to the SSDs in 4K blocks, which requires de-blocking and re-blocking all non-4K IO.

The 128K Sequential Write results are 30.3 GB/s, which outperform the advanced software RAID results (20.3 GB/s) by 49%, clearly demonstrating the advantages of offloading the compute (parity) calculations from the CPU to a software-enabled, GPU-based architecture.



Sequential IO Benchmarks

Throughput	Workload			
GB	128K Sequential Read		128K Sequential Write	
	GB	Solution/Baseline	GB	Solution/Baseline
Baseline	71.50	100%	39.60	100%
SupremeRAID RS	61.40	86%	30.30	77%
ADVSW RS	69.5	97%	20.30	51%

Performance and Efficiency

SupremeRAID NOMINAL vs. BASELINE

Below we compare the performance of the three SupremeRAID adapter operational states (NOMINAL, DEVICE DOWN, and DEVICE REBUILD) to the BASELINE solution – a Data24 FW4.0 using 24 SN840 3.2TB devices (each has two namespaces) in a 6x8n configuration (six servers with each using eight namespaces. All comparisons will be SupremeRAID STATE (nominal, device down, or device rebuild) to the BASELINE.

The first panel in the following figure shows the gross performance of the BASELINE for the 4K Random Writes (R.W.), Random Mixed (R.M.), and Random Reads (R.R.) at 6.26, 12.00, and 15.30 MIPS, respectively.

The second panel shows SupremeRAID NOMINAL performance for 4K R.W., R.M., and R.R. at 2.60, 6.12, and 15.30 MIOPS, respectively.

- SupremeRAID NOMINAL matches the BASELINE of 15.30 MIOPS, and this test shows that SupremeRAID NOMINAL is transparent for this workload.
- Otherwise, the BASELINE outperforms SupremeRAID NOMINAL by 58 and 49% for the 4K R.W. and R.M. workloads.

As measured by WORK/CPU%, SupremeRAID NOMINAL is 23% more efficient than the BASELINE for the 4K R.R. workload. Else, the BASELINE is ~22 and ~7% more efficient than SupremeRAID NOMINAL for the 4K R.W. and R.M. workloads, respectively.

The data also shows that the BASELINE outperforms SupremeRAID NOMINAL in Latency (i.e., RAID 5 increases the latency). Still, the COV (Coefficient of Variation) and CPU Percent (usr+sys) are better for SupremeRAID NOMINAL.

Observations:

- In many environments, RAID 5 can offset performance loss by protecting against a single device failing in a RAID Set.
- Without RAID 5, other data protection methods would have to be employed, which would likely be costlier than the RAID 5 solution, more complex, and could be more disruptive to production workloads.

DATA24-FW4.0-SN840-3.2T SSR (8, 9, 10) BASELINE					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	6.26	0.95	0.32%	132.60%	47,210
Measure_RM_4KB_QD256 Random Fill	12.00	0.45	0.27%	215.22%	55,757
Measure_RR_4KB_QD256 Random Fill	15.30	0.39	0.54%	310.20%	49,323

SupremeRAID R5 7+1 DATA24-FW4.0-SN840-3.2T SSR (8, 9, 10) NOMINAL					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	2.60	6.50	0.22%	70.38%	36,942
Measure_RM_4KB_QD256 Random Fill	6.12	3.31	0.19%	117.72%	51,988
Measure_RR_4KB_QD256 Random Fill	15.30	1.10	0.12%	251.40%	60,859

RATIO SupremeRAID/BASELINE					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	0.42	6.84	0.69	0.53	0.78
Measure_RM_4KB_QD256 Random Fill	0.51	7.36	0.69	0.55	0.93
Measure_RR_4KB_QD256 Random Fill	1.00	2.82	0.22	0.81	1.23

PERCENT CHANGE: ((SupremeRAID/BASELINE)-1)*100					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	-58%	584%	-31%	-47%	-22%
Measure_RM_4KB_QD256 Random Fill	-49%	636%	-31%	-45%	-7%
Measure_RR_4KB_QD256 Random Fill	0.00%	182%	-78%	-19%	23%

UUIDS: 1ef987fa, aa4d1feb	FIO RESULTS	BASELINE BETTER	Supreme RAID BETTER	SIMILAR

SupremeRAID DEVICE DOWN vs. BASELINE

In this test, we remove a device from RAID Set, and then the Spec Sheet Random benchmark is run. Common SupremeRAID control commands remove the device from the RAID Set to simulate a failed device as seen in the following table.

We continue to compare the SupremeRAID performance to the BASELINE performance. The reasons for this are:

- The performance through the RAID Life Cycle (Nominal, Device Down, and Device rebuild) can vary with respect to the BASELINE.
- It is sensible to compare to a well-known BASELINE – this would likely be a customer's current solution, i.e., the BASELINE.

The last panel shows that SupremeRAID DEVICE DOWN performance for 4K R.W., R.M., and R.R. is 62, 62, and 44 percent lower than the BASELINE. All three of these are significant impacts from the BASELINE.

- But we must remember that this solution has eliminated a single point of failure.
- The cost and complexity of alternatives can be high and time-consuming. SupremeRAID DEVICE DOWN has:
- Higher latency but a lower COV, i.e., it is more stable.
- Fifty percent lower CPU (usr+sys) for all three tests.

Efficiency per CPU%, shown in the rightmost column, is computed by dividing the Work in IOPS by the CPU% required to generate this workload. This calculation provides the number of IOPS per CPU% shown in the rightmost column of the first two panels.

- The third panel is the ratio of SupremeRAID and BASELINE for each workload, while the fourth panel converts the last panel to percent difference.
- SupremeRAID CPU efficiency is 16 percent more than the BASELINE for the 4K R.R. workload.
- BASELINE CPU efficiency is 23 and 16 percent better for the 4K R.W. and R.M. workloads.

DATA24-FW4.0-AN840-3.2T SSR (8, 9, 10) BASELINE					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	6.26	0.95	0.32%	132.60%	47,210
Measure_RM_4KB_QD256 Random Fill	12.00	0.45	0.27%	215.22%	55,757
Measure_RR_4KB_QD256 Random Fill	15.30	0.39	0.54%	310.20%	49,323

SupremeRAID R5 7+1 DATA24-FW4.0-AN840-3.2T SSR (8, 9, 10) DEVICE DOWN					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	2.38	7.08	0.22%	65.10%	36,559
Measure_RM_4KB_QD256 Random Fill	4.58	4.27	0.19%	97.32%	47,061
Measure_RR_4KB_QD256 Random Fill	8.60	1.96	0.17%	150.66%	57,082

RATIO SupremeRAID/BASELINE					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	0.38	7.45	0.69	0.49	0.77
Measure_RM_4KB_QD256 Random Fill	0.38	9.49	0.71	0.45	0.84
Measure_RR_4KB_QD256 Random Fill	0.56	5.03	0.31	0.49	1.16

PERCENT CHANGE: ((SupremeRAID/BASELINE)-1)*100					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	-62%	645%	-31%	-51%	-23%
Measure_RM_4KB_QD256 Random Fill	-62%	849%	-29%	-55%	-16%
Measure_RR_4KB_QD256 Random Fill	-44%	403%	-69%	-51%	16%

UUIDS: 1ef987fa, c81717b8	FIO RESULTS	BASELINE BETTER	Supreme RAID BETTER	SIMILAR

SupremeRAID DEVICE REBUILD vs. BASELINE

In this section the failed device from the previous test is added back in. Then, the rebuild process starts, and the standard Spec Sheet Random Benchmark is run as seen in the following table.

The standard workloads take 20 minutes for each of the three IO Types. The rebuild operates at its highest performance and completes in 85 minutes (25 minutes longer than the standard workloads).

As previously shown, the MIOPS column shows the performance of the three workloads in Panels 1 and 2 for the BASELINE and SupremeRAID, respectively.

At a high level, there is not much performance difference between SupremeRAID DEVICE DOWN and SupremeRAID DEVICE REBUILD in terms of throughput or efficiency.

DATA24-FW4.0-AN840-3.2T SSR (8, 9, 10) BASELINE					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	6.26	0.95	0.32%	132.60%	47,210
Measure_RM_4KB_QD256 Random Fill	12.00	0.45	0.27%	215.22%	55,757
Measure_RR_4KB_QD256 Random Fill	15.30	0.39	0.54%	310.20%	49,323

SupremeRAID R5 7+1 DATA24-FW4.0-AN840-3.2T SSR (8, 9, 10) DEVICE REBUILD					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	1.84	9.20	0.30%	50.22%	36,639
Measure_RM_4KB_QD256 Random Fill	3.61	5.51	0.25%	76.92%	46,932
Measure_RR_4KB_QD256 Random Fill	7.11	2.37	0.24%	123.30%	57,664

RATIO SupremeRAID/BASELINE					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	0.29	9.68	0.94	0.38	0.78
Measure_RM_4KB_QD256 Random Fill	0.30	12.24	0.93	0.36	0.84
Measure_RR_4KB_QD256 Random Fill	0.46	6.08	0.44	0.40	1.17

PERCENT CHANGE: ((SupremeRAID/BASELINE)-1)*100					Efficiency (Work/CPU%)
DESCRIPTION	**MIOPS**	**LAT-US**	**COV**	**USR+SYS**	**IOPS/CPU%**
Measure_RW_4KB_QD256 Random Fill	-71%	868%	-6%	-62%	-22%
Measure_RM_4KB_QD256 Random Fill	-70%	1124%	-7%	-64%	-16%
Measure_RR_4KB_QD256 Random Fill	-54%	508%	-56%	-60%	17%

UUIDS: 1ef987fa, c81717b8	FIO RESULTS	BASELINE BETTER	Supreme RAID BETTER	SIMILAR

Those considering a RAID 5 solution must be able to meet their Service Levels with the lowest SupremeRAID performance or augment RAID 5 with workload shedding or deferring, fail-over, etc.

There are two essential Business Continuity and Disaster Recovery (BCDR) objectives:

- Recovery Time Objective (RTO) and
- Recovery Point Objective (RPO)

RAID 5 essentially addresses and eliminates RTO and RPO assuming just one failed device. Of course, there are many elements to a comprehensive BCDR, but a well-planned and well-sized RAID solution can manage single instances of device failure.

Performance by Configuration Summary

Next, we ran the large block (128K) Spec Sheet Sequential Benchmark on the exact configuration previously described above. The results are shown in the following figure.

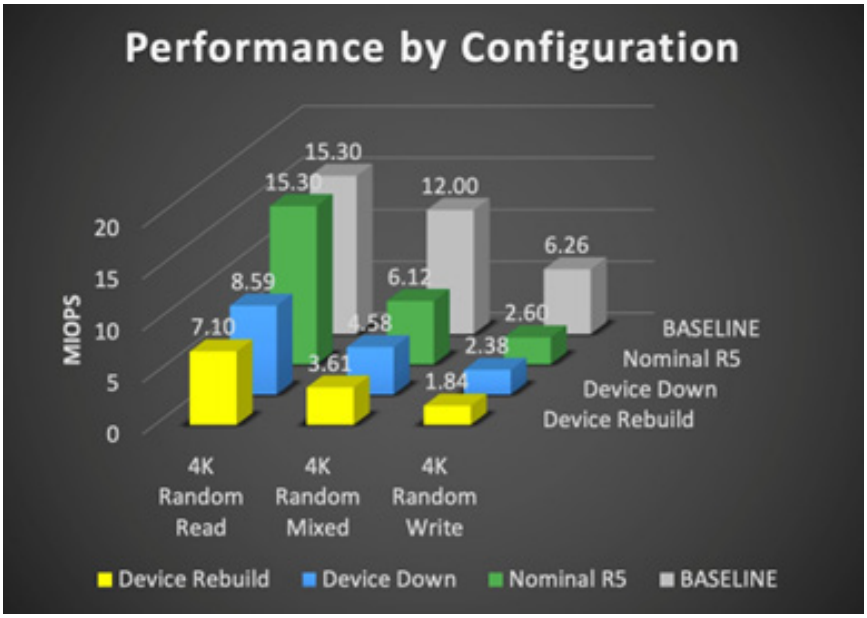
The Performance by Configuration Summary and Efficiency by Configuration Summary figures in the following section provide 3D and table data that summarizes SupremeRAID SR-1000 adapter R5 Performance on our standard Spec Sheet Random (SSR) benchmark. These figures show:

The BASELINE is a Data24 with 24 SN840 3.2TB devices (each with two namespaces) tested with the SSR benchmark. A SupremeRAID R5 7+1 implementation on the Data24 described above for:

- SupremeRAID NOMINAL
- SupremeRAID DEVICE DOWN
- SupremeRAID DEVICE REBUILD

The slopes of the chart elements are as expected – monotonically decreasing from:

- Left to right
- Back to front
- Left rear to right front



Performance by Configuration Summary

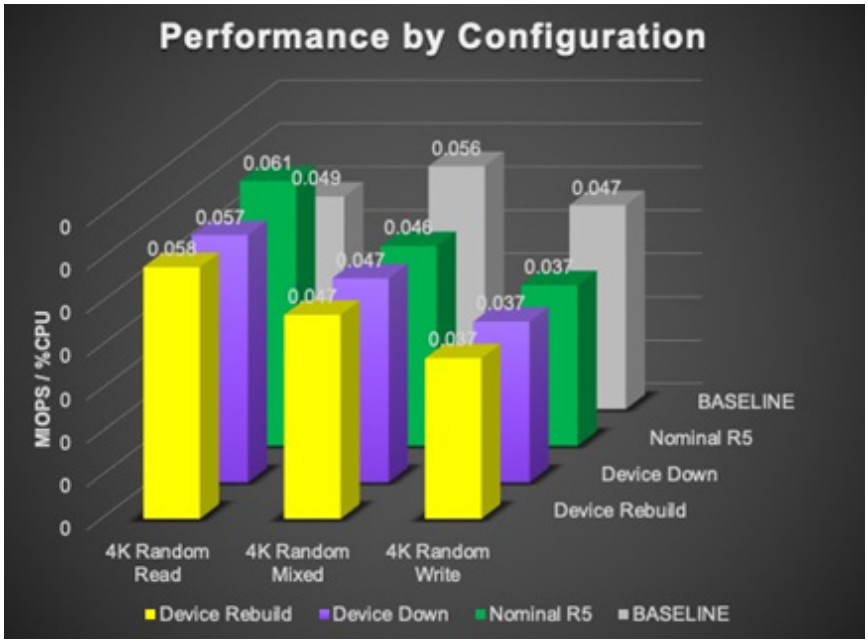
Throughput	WORKLOAD					
SupremeRAID / BASELINE	**4k Random Read**		**4K Random Mixed**		**4K Random Write**	
SOLUTION	MIOPS	SOLUTION/BASELINE	MIOPS	SOLUTION/BASELINE	MIOPS	SOLUTION/BASELINE
BASELINE	15.30	100%	12.00	100%	6.26	100%
SupremeRAID R5	15.30	100%	6.12	51%	2.60	42%
Device Down	8.59	56%	4.58	38%	2.38	38%
Device Rebuild	7.10	46%	3.61	30%	1.84	29%

Efficiency by Configuration Summary

The Random Read performance for the BASELINE and the SupremeRAID NOMINAL is similar at ~15.30 MIOPS. But, because of the move of Parity Calculations from Server CPUs to the GPU, the SupremeRAID Solution is about 24% more efficient than the BASELINE.

This finding is unique in this study. It is achieved by:

- Moving the CPU cycles to the GPU and
- The efficient SupremeRAID 4K read pipeline.
- SupremeRAID efficiency is similar for each workload (Random Reads, Random Mixed, and Random Writes) across the three life cycles (Nominal, Device Down, and Device Rebuild).



Performance by Configuration Summary

EFFICIENCY	WORKLOAD					
SupremeRAID / BASELINE	**4k Random Read**		**4K Random Mixed**		**4K Random Write**	
SOLUTION	MIOPS/ CPU%	SOLUTION/ BASELINE	MIOPS/ CPU%	SOLUTION/ BASELINE	MIOPS/ CPU%	SOLUTION/ BASELINE
BASELINE	0.049	100%	0.056	100%	0.047	100%
NOMINAL R5	0.061	124%	0.046	82%	0.037	79%
Device Down R5	0.057	116%	0.047	84%	0.037	79%
Device Rebuild R5	0.058	118%	0.047	84%	0.037	79%

Conclusion

An NVMe-oF storage enclosure such as the OpenFlex Data24 allows for a broader degree of performance, flexibility, and cost savings not found with traditional hardware or OS-based software RAID.

This GPU architecture outperformed the Advanced Software RAID solution in all areas except large block sequential reads in these tests.

Consider the following:

- SupremeRAID SR-1000 adapter is essentially a plug-and-play solution using a commercially available GPU.
- SupremeRAID allows competitive pricing as the silicon architecture is not proprietary for this use.
- The ability to separate the data path from the logic path adds value and flexibility.
- A GPU upgrade or a GPU firmware upgrade could provide new features and performance improvements, possibly with low operational impacts
- Traditionally, the data path has presented itself as the bottleneck via an AISC-based RAID controller or CPU computation. Direct IO between the CPU and GPU is efficient and allows the GPU's massive computational capability to manage RAID calculations in the data path.

GPU release cycles are regular, and it is fair to anticipate that performance should improve as GPU architectures are enhanced (along with server motherboard architecture – such as PCIe Gen 4). This regular product cycle, in turn, allows the consumer to balance performance requirements against the capabilities of the GPU – essentially driving a tighter cost versus performance model.

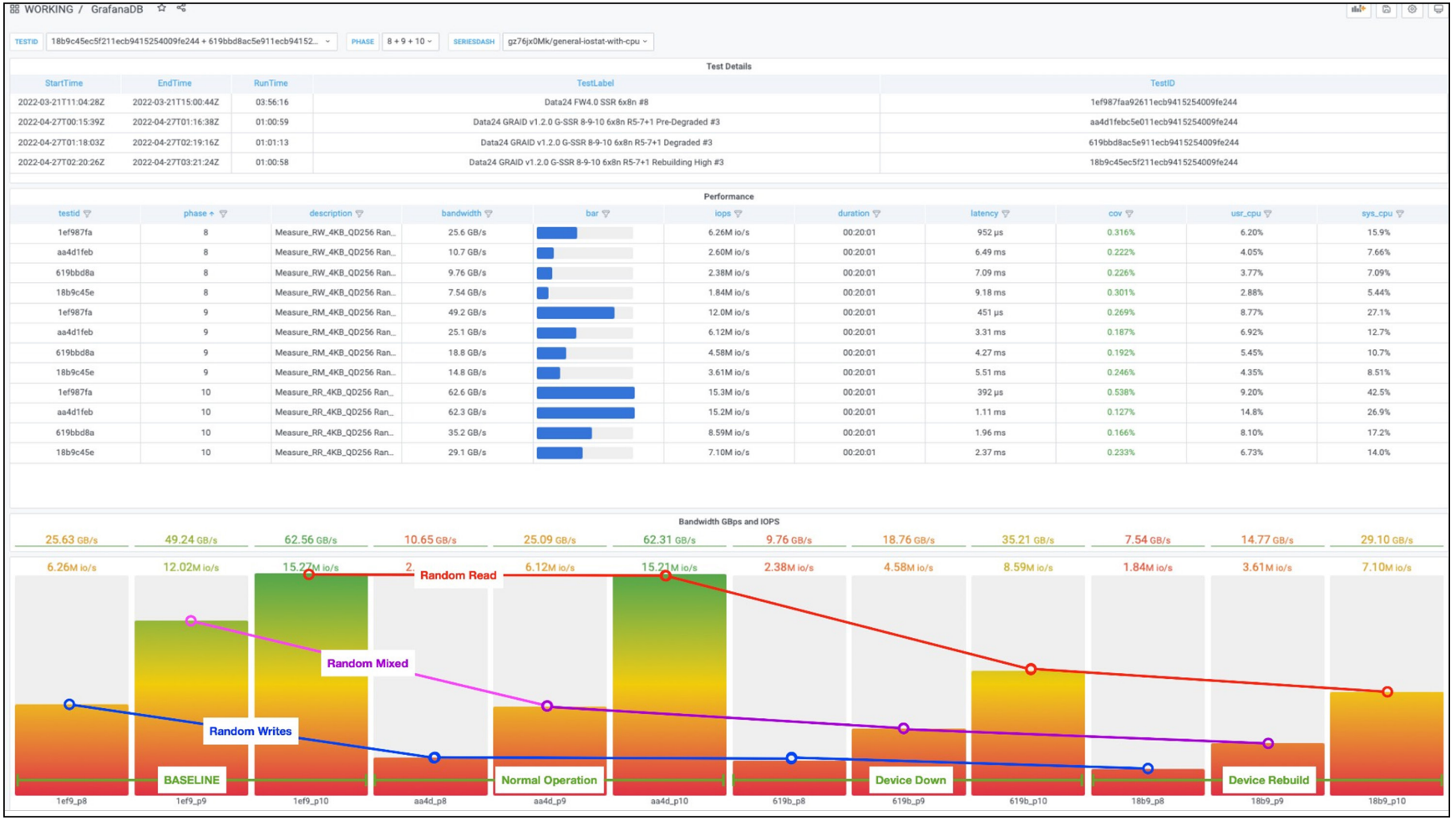
There are potential benefits to be realized in the server architecture when using this solution. Tradition hardware RAID is unlikely to meet the performance potential of NVMe devices. Such cards scale poorly and require additional cables for device connectivity. RAID Add-In-Cards (AIC) can add complexity and cost, use extra PCIe slots, and disrupt airflow. A GPU-based RAID solution may reduce or eliminate these issues. Additionally, CPU cycles are freed to be assigned elsewhere or, if not required, allowing for lower specification (lower cost) CPU to be considered.

The critical issue in this solution is making RAID 5, which has always been the most desired RAID configuration (add one device to eliminate a single point of failure), sufficiently performant for use across most general storage needs. The SupremeRAID does this in an elegantly simple implementation requiring no significant changes to the environment.

The SupremeRAID R5 Life Cycle Chart with BASELINE figure shows the absolute and relative performance for the various workloads of the RAID Life Cycle (Initialization, Nominal, Device Down, and Device Rebuild). Potential consumers should understand this information to assess the applicability of Graid Technology, Inc.'s solution.

The SupremeRAID R5 Life Cycle Chart with RAID Set and Constituent Devices figure provides a unique view of both the RAID Set used by the customer and the underlying Constituent Devices that make up the RAID Set.

Appendix 1: SupremeRAID R5 Life Cycle Chart with BASELINE



SupremeRAID R5 Life Cycle Chart with BASELINE

Appendix 2: SupremeRAID R5 Life Cycle Chart with RAID Set and Constituent Devices

SupremeRAID R5 Life Cycle Chart with RAID Set and Constituent Devices

Appendix 3: Document Details

Contributors

Name	Company	Title
John Gatch	Western Digital	Technologist, Platforms Field Engineering
Calvin Falldorf	Western Digital	Principal Engineer, Platforms Field Engineering
Niall Macleod	Western Digital	Director, Platforms Field Engineering
Barrett Edwards	Western Digital	Sr. Director, Platforms Field Engineering

References

Document Title	Version	Date
SupremeRAID White Paper 2022-07-25-V0.91	V0.91	July 25, 2022
SupremeRAID White Paper 2022-07-25-V0.93	V0.93	July 28, 2022
SupremeRAID White Paper 2022-07-25-V1.0	V1.0	August 3, 2022
SupremeRAID White Paper 2022-08-16-V1.1	V1.1	August 16, 2022
SupremeRAID White Paper 2023-07-6-V1.2	V1.2	July 6, 2023

Version History

Contributor	Version	Date
Niall MacLeod	V0.1	July 1, 2022
Calvin Falldorf and John Gatch	V0.3	July 8, 2022
John Gatch	V0.5	July 15, 2022
Niall MacLeod, Calvin Falldorf, and John Gatch	V0.7	July 22, 2022
Scot Rives, Western Digital Legal, first review	V0.91	July 25, 2022
John Gatch, Updated with Scot Rives updates	V0.93	July 28, 2022
John Gatch, Updated with Scot Rives updates	V1.0	August 3, 2022
John Gatch, updated title per Niall MacLeod	V1.1	August 16, 2022
Scot Rives, Aaron Fennimore, and Scott Hamilton updates	V1.2	July 6, 2023

Document Feedback

For feedback, questions, and suggestions for improvements to this document, send an email to the Data Center Systems (DCS) Technical Marketing Engineering (TME) team distribution list at pdl-dcs-tm@wdc.com.

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
www.westerndigital.com