Western Digital.

Solution Brief

SupremeRAID™ SR-1010 with OpenFlex™ Data24 4000 Series

SupremeRAID™

Highlights

- Software RAID
- Hardware RAID
- GPU-based Hardware Accelerated Software RAID

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 Computations

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

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

Software RAID

Operating System (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 OS 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 do not 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.

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.

Graid Technology Inc.[1] provides the GPU-based RAID solution tested in this project, the SupremeRAID SR-1010.

[1] G-RAID is owned by and is a registered trademark of Western Digital. Graid Technology, Inc. is not affiliated with, related to, or endorsed by Western Digital. Western Digital does not endorse or authorize use of Graid by Graid Technology, Inc.

SupremeRAID SR-1010 with OpenFlex Data24 4000 Series

For more detail see the following resources:

- https://www.graidtech.com/supremeraid-for-ai-machine-learning/
- https://www.graidtech.com/storage-review-supremeraid-for-ai/
- https://www.graidtech.com/product/sr-1010/

OpenFlex Data24 4000 Series NVMe-oF Storage Platform

Western Digital's OpenFlex Data24 4000 Series 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 1.474PB[2] when using Ultrastar® DC SN655 61.44TB devices. Unlike a SAS enclosure, the Data24's dual IO modules use Western Digital RapidFlex™ C2000 NVMe-oF Controllers. These controllers allow full access to all 24 NVMe drives over up to twelve 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 Ultrastar DC SN655 15.36TB devices, the enclosure can achieve up to 135 GB/s of bandwidth and over 27 MIOPS at a 4K block size.

SupremeRAID SR-1010 NVMe-oF RAID

The SupremeRAID SR-1010 (for PCIe® 3, 4, and 5 servers) 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-1010 enables NVMe/NVMe-oF, SAS, and SATA performance while increasing scalability, improving flexibility, and lowering total cost of ownership (TCO). This solution eliminates the traditional RAID bottleneck in mass storage to deliver maximum SSD performance for high-intensity workloads.

SupremeRAID SR-1010 for PCIe 3, 4, and 5 servers

FIO Test Cases

FIO Spec Sheet Sequential Workloads consists of documenting 128KB Sequential Write and Sequential Read performance results under an 8 Drive Non-GRAID Baseline. The same tests are performed on a RAID 5 7+1 volume under Normal, Degraded, and Rebuild conditions. Each workload was run for 20 minutes to capture the average bandwidth in GBps during each GRAID volume state. Details of the test phases can be provided if needed.

FIO Spec Sheet Random Workloads consists of documenting 4KB Random Write, Random Mixed, and Random Read performance results under an 8 Drive Non-GRAID Baseline. The same tests are performed on a RAID 5 7+1 volume under Normal, Degraded, and Rebuild conditions. Each workload was run for 20 minutes to capture the average bandwidth in GBps during each GRAID volume state. Details of the test phases can be provided if needed.

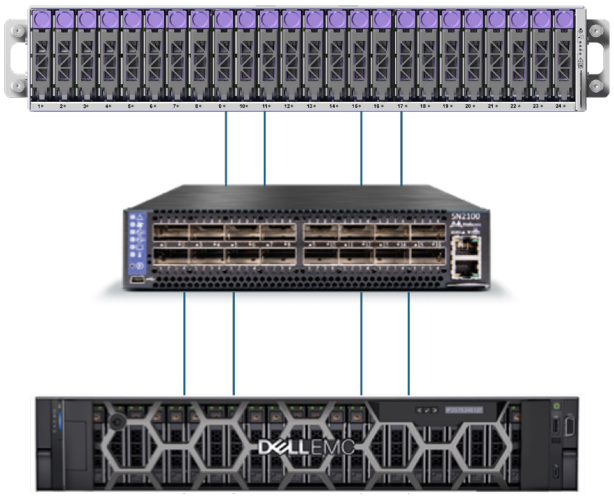
- A Normal condition consists of a completely optimal RAID 5 volume.
- A Degraded condition consists of a single drive being put into an offline Using the graidctl software.
- A Rebuild condition consists of the drive being put back into an online state by the graidctl command and having its data reconstructed from the rest of the drives in the RAID set.

[2] One terabyte(TB) is equal to one trillion bytes and one petabyte(PB) is equal to one quadrillion bytes. Actual user capacity may be less due to operating environment.

SupremeRAID SR-1010 with OpenFlex Data24 4000 Series

Topology

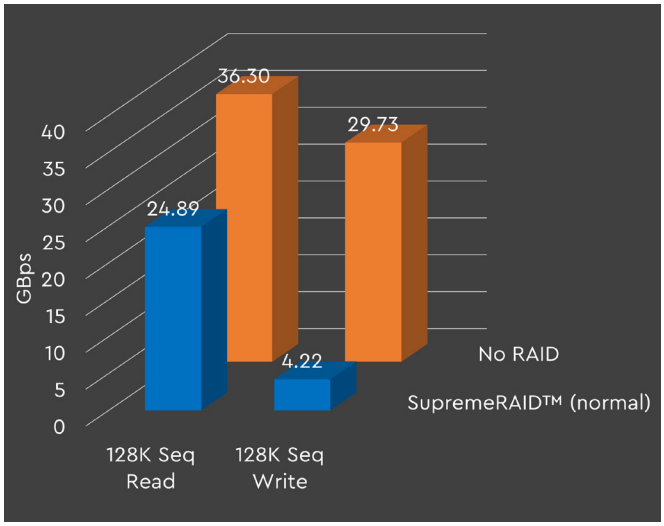
- The OpenFlex Data24 4240 and the Dell® R750 Server are attached to a Mellanox® SN2100 switch.
- The Dell R750 has a pair of Mellanox ConnectX-6 RNICs connected to the switch with 2 cables from each RNIC running at 100 Gbps.
- The OpenFlex Data24 4240 is connected to the switch from Ports 1 and 4 on each IOM also running at 100 Gbps.
- NVMe Native Multipathing is configured to round robin across all 4 paths.
- The protocol in use is RoCE v2.



Results

In the following tables, BASELINE performance is defined as eight drives under test in the OpenFlex Data24 4000 Series with no RAID in place. The SOULTION/BASELINE column percentages are derived from the solution value divided by the baseline value.

Normal OpenFlex Data24 4200 with SupremeRAID SSS vs OpenFlex Data24 4200 with normal No RAID:

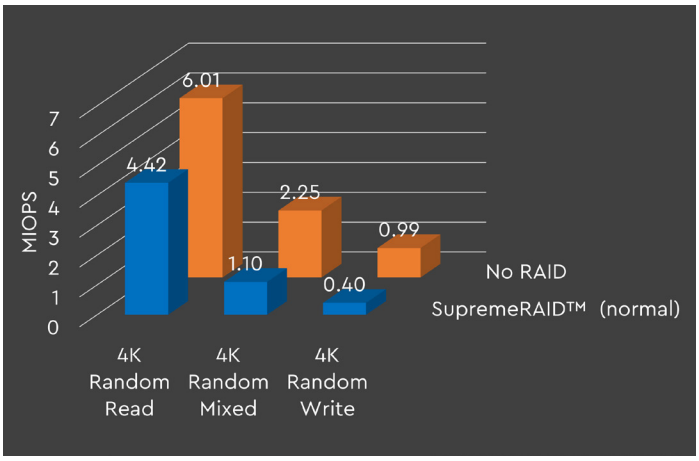


OpenFlex Data24 4200 No RAID vs SupremeRAID SR1010 v1.6.1 SSS Performance by Configuration:

GBps	SupremeRAID (normal)	SOLUTION/BASELINE	No RAID
128K Seq Read	24.89	68.5%	36.3
128K Seq Write	4.22	14.2%	29.73

SupremeRAID SR-1010 with OpenFlex Data24 4000 Series

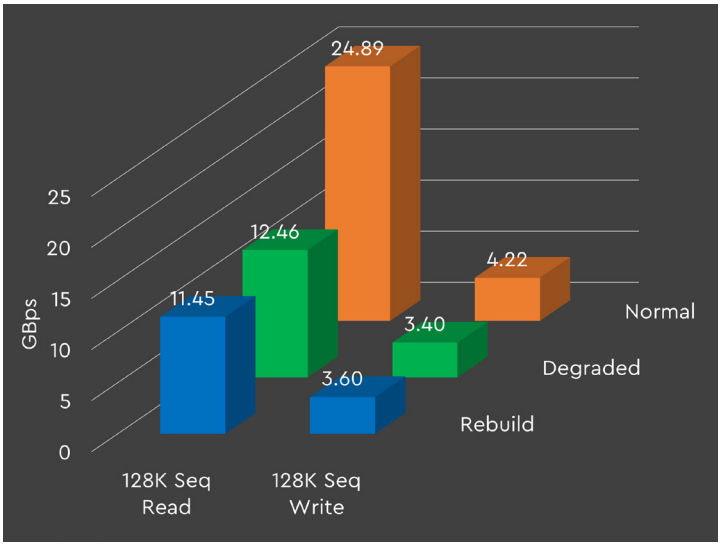
OpenFlex Data24 4200 No RAID vs SupremeRAID SR1010 v1.6.1 SSR Performance by Configuration:



Normal OpenFlex Data24 4200 with SupremeRAID SSR vs normal OpenFlex Data24 4200 with No RAID:

MIOPS	SupremeRAID (normal)	SOLUTION/BASELINE	No RAID
4K Random Read	4.42	73.5%	6.01
4K Random Mixed	1.1	48.9%	2.25
4K Random Write	0.4	40.40%	0.99

OpenFlex Data24 4200 SupremeRAID SR1010 v1.6.1 SSS Normal/Degraded/Rebuild/Performance by Configuration:

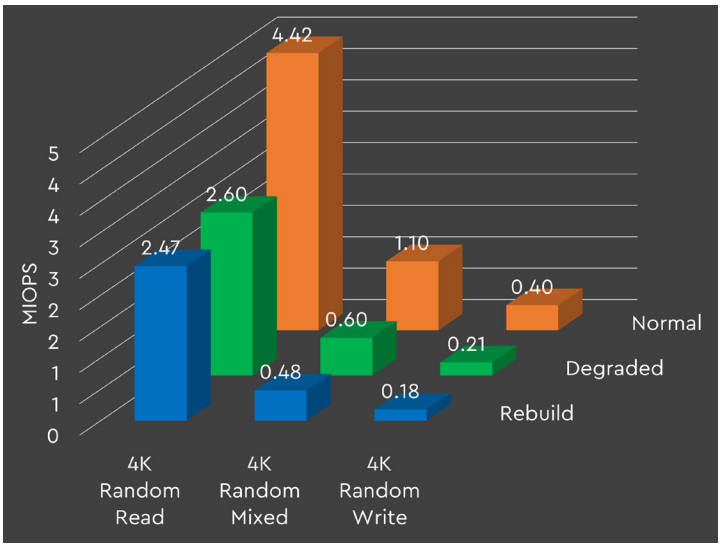


OpenFlex Data24 4200 with SupremeRAID SSS Normal/Degraded/Rebuild:

MIOPS	Rebuild	SOLUTION/BASELINE	Degraded	SOLUTION/BASELINE	Normal
128K Seq Read	11.45	46%	12.46	50.06%	24.89
128K Seq Write	3.36	85.71%	3.4	80.57%	4.22

SupremeRAID SR-1010 with OpenFlex Data24 4000 Series

OpenFlex™ Data24 4200 SupremeRAID SR1010 v1.6.1 SSS Normal/Degraded/Rebuild Performance by Configuration:



OpenFlex Data24 4200 with SupremeRAID SSR Normal/Degraded/Rebuild:

MIOPS	Rebuild	SOLUTION/BASELINE	Degraded	SOLUTION/BASELINE	Normal
4K Random Read	2.47	55.88%	2.6	58.82%	4.42
4K Random Mixed	0.48	43.63%	0.6	54.54%	1.1
4K Random Write	0.18	45%	0.21	52.50%	0.4

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.

Consider the following:

- SupremeRAID SR-1010 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.

SupremeRAID SR-1010 with OpenFlex Data24 4000 Series

Infrastructure Overview

Storage	Western Digital OpenFlex Data24 4000 Series OpenFlex Data24 4240 is using a Gen4 PCIe architecture and a total of 12 front end Ethernet AIC/RFX ports, each capable of 100Gb/s (12.5GB/s)[3]. Network Storage Protocol: RDMA with RoCE v2
Drives	8 x Ultrastar DC SN655 NVMe 15.36TB Dual ported enterprise SSDs. These drives offer high-capacity, cost-optimized, read-intensive performance for data-intensive applications. Firmware version RC610008 Offering 122.88TB raw capacity in this 8 Drive configuration. Single Namespace per drive
Server Hardware	Dell PowerEdge R750, BIOS 1.13.2 Dual Intel® Xeon® Gold 6354 CPU 18 Cores @ 3.00GHz 256GiB Memory 2 x Mellanox ConnectX-6 RNICs (PCIe x16 slots) 1 RNIC on NUMA 0 1 RNIC on NUMA 1 Model: MCX613106A-VDA/T, FW: 20.33.1048 GRAID SR-1010, version 1.6.1
Ethernet Switch	Mellanox/NVIDIA® SN2100 with 3.10.4504 FW 8 x 100Gb Direct Attach Copper (DAC) cables Server, Switch, and Data24 are configured with RDMA with RoCE v2 and appropriate lossless settings
Server Software	RHEL 9.4: 5.14.0-427.42.1.el9_4.x86_64 Mellanox OFED: doca-host-2.7.0-209000_24.04_rhel94.x86_64

[3]This gives a theoretical performance threshold of 150Gb/s per chassis.

Western Digital.

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

March 2025 / Revision 01