Western Digital.

White Paper

XINNOR xiRAID Power-of-Two (Po2) Performance Analysis with OpenFlex™ Data24 3200

Using one server and one OpenFlex Data24 3200 with Ultrastar® DC SN840-3.2 TB devices

Abstract

This investigation compares and contrasts performance differences between Power-of-Two (Po2) and Non-Power-of-Two (NPo2) RAID Sets using XINNOR's xiRAID, a third-party, software-based RAID solution. The number of data devices in a Po2 is 2N, where N is usually a small integer. While RAID Sets can contain an arbitrary number of devices, most solutions contain no more than 32. The number of devices in NPo2 configurations is not a Po2. The RAID configurations tested are parity solutions using an NPo2 R5-7+1 (7 data drives) and a Po2 R5-8+1 (8 data drives).

The testing methodology uses the standard Western Digital Platforms - Applications Engineering (PAE) test suites: Spec Sheet Sequential (SSS) and Spec Sheet Random (SSR) but are modified to demonstrate performance during the three RAID Life Cycles (Normal, Degraded, and Rebuild).

Three instances of all tests are performed, and their variability is analyzed with inter- and intra- Coefficient of Variation (CoV) analysis. These suites produce two Figures of Merit: the bandwidth (BW) and the CPU utilization. The study normalizes all data to support standard A/B comparisons. The SSR suite's final analysis (grand summary) shows that NPo2 has an 8.49% BW and a 1.52% CPU advantage for SSR. Po2 has a 6.99% BW and a 45.89% CPU advantage for SSS. Po2 SSS CPU efficiency is significant due to XINNOR's exploitation of Advance Vector Extensions (AVX).

November 2023

Table of Contents

Executive Summary

XINNOR, the company that provides xiRAID (a software-based RAID solution that exploits CPU Advanced Vector Extensions (AVX)), and Western Digital collaborated to demonstrate xiRAID with Western Digital's OpenFlex Data24 3200, a high-performance network storage JBOF (Just a Bunch of Flash) solution.

While Western Digital has tested xiRAID extensively, it has generally been with three R5-7+1 (7 data disks + 1 parity disk for eight total disks). This configuration naturally fits a 24-drive Data24 chassis. This document pivots to running a Power-of-Two (Po2) configuration for the data disks as it has proven efficient. Efficiency is increased by reading and writing full blocks to each data disk, which is not generally possible within NPo2s.

To simplify testing and analysis, this analysis compares a single Po2 (R5-8+1) configuration with a Non-Power-of-Two (NPo2) (R5-7+1) configuration. This analysis is not an endorsement of xiRAID or Xinnor by Western Digital, and no warranty of the product is either expressed or implied.

Configuration

The configuration required for this test is one server and one Data24 3200 using the Ultrastar DC SN840-3.2TB devices. Descriptions of each are provided with the links below.

- "OpenFlex Data24 3200 NVMe-oF Enclosure" on page 4
- "OpenFlex Data24 3200 Test Environment" on page 4

Workload

The following describes the workloads at a high level. In particular, when testing RAID, it is essential to perform the tests in all primary operational conditions (Normal, Degraded, and Rebuild), as this information is required for informed decisions.

- Spec Sheet Random (4K): SSR
  - Random Write: RW, Random Mixed (70%R): RM, Random Read: RR
- Spec Sheet Sequential (128K): SSS
  - Sequential Read: SR, Sequential Write: SW, Job-Split: SR-JS, Job-Split: SW-JS
- RAID Life Cycles
  - Normal, Degraded, and Rebuild

Process

Platforms - Applications Engineering (PAE) test using our Spec Sheet Sequential and Random (SSS and SSR, respectively) Suites with RAID Life Cycles (normal, degraded, and rebuild) with the two configurations identified above.

This approach demonstrates the performance across all significant operational conditions and provides critical information for stakeholders.

A key aspect of this analysis is that the results are normalized to performance per device. Doing this analysis corrects for the difference in drive counts (8 vs. 9) for the two configurations tested.

Results

The following table displays a macro view of the detailed results. Detailed results are presented in "Bandwidth Summary" on page 5 and "CPU Summary" on page 6.

High-Level Performance Summary

Spec Sheet Suite	Figure of Merit	More Performant	Advantage in Percent	Comments
SSR 4K	BW	NPo2 (R5-7+1)	8.49%	NPo2 marginally outperforms on SSR in both Figures of Merit (BW and CPU).
	CPU	NPo2 (R5-7+1)	1.52%	
SSS 128K	BW	Po2 (R5-8+1)	6.99%	Po2 CPU efficiency, using the Advanced Vector Extensions (AVX), is significant with large block sequential workloads.
	CPU	Po2 (R5-8+1)	45.89%	

OpenFlex Data24 3200 NVMe-oF Storage Platform

Western Digital's OpenFlex Data24 3200 NVMe-oF Storage Platform is similar to a 2.5" SAS Enclosure. The Data24 3200 can be used as stand-alone storage but can also be a foundational block of a Software-Composable Infrastructure. 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 3200'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 six ports of 100 Gb Ethernet.

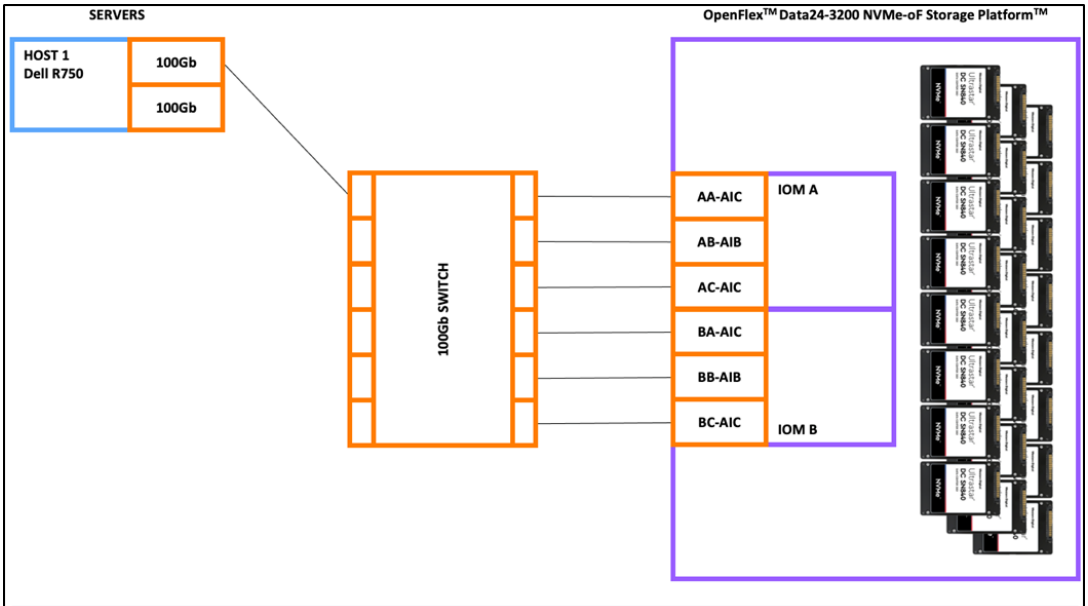
The Data24 is a close replacement for the traditional SAS enclosures. However, the Data24 3200 offers a significant benefit over these enclosures: the ability to integrate directly into Ethernet fabric, allowing for an Any-to-Any mapping of Object Storage Targets to Object Storage Servers.

The OpenFlex Data24 3200 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 71.4 GB/s of 128K bandwidth and over 16.7 MIOPS at a 4K block size.

OpenFlex Data24 3200 NVMe-oF Enclosure

Western Digital OpenFlex Data24 3200 Test Environment

- A block diagram of the tested configuration consisting of one server and one Data24 3200.
- NPo2 (R5-7+1) uses eight drives, while Po2 (R5-8+1) uses nine drives.
  - NPo2 uses eight drives from the first eight slots or "eight pack."
  - Po2 uses the first three drives of each of the three "eight packs" of the Data24.



SERVERS	SWITCH & INTERFACES	Data24-3200 & STORAGE	LEGEND
DELL PowerEdge R750	MELLANOX SN3800	Data24-3200	BLUE = SERVERS
CPU: 2 x INTEL XEON GOLD 6354 @ 3.00 GHZ	ONYX OS	FW: 1.0.0	ORANGE = NETWORK
DRAM: 512 GiB	FW: 3.9.2110	WDC SN840 3.2 TB X 24	PURPLE = Data24-3200
OS: RedHat 8.7	CONNECTX-6 Dx	FW: R2210010	
KERNEL: 4.18.0-425.3.1.el8.x86_64	FW: 22.37.1014	PN: WUS4C6432DSP3X1	

OpenFlex Data24 3200 Test Environment

[1] One gigabyte (GB) is equal to one billion bytes and one terabyte is equal to one trillion bytes. Actual user capacity may be less due to operating environment.

RAID Overview

RAID has many configurations, but the two dominant configurations are:

- Mirrored where the amount of device capacity is at least two times the capacity needed for non-RAID solutions.
  - Mirrored solutions' major benefit is little if any falloff in performance with a failed device.
  - However, a failed device can fail one-half of the mirrored pair, resulting in no protection for another failure in the working half of the pair.
- Parity-based solutions add one or more parity devices (often just one) to a set of devices to create a RAID Set. So, if D data devices are needed, you can add P parity devices where P is one or more for a total of D+P devices.
  - With one added parity device and seven data devices, the nomenclature would be R5-7+1.
  - Adding additional parity devices so that there are two or more parity devices creates a RAID6 Raid Set. With 7 data devices and two parity devices, the nomenclature would be R6-7+2. This configuration could sustain two device failures without data loss.
  - Parity RAID solutions can protect from a device failure, but performance is less during the Degraded and Rebuild conditions.

PAE has focused on testing RAID parity solutions because this is the lowest cost solution if it can meet the business objectives (Parity RAID generally has lower write workload performance).

- This study used two RAID 5 (parity RAID) configurations, R5-7+1 and R5-8+1.

Bandwidth Summary

- The following table shows all the tests run against the test configuration for both the Po2 (8+1) and NPo2 (7+1) and the performance analysis based on the bandwidth of the various tests.
- The top of this table is for SSR, and the bottom is for SSS. The green shading indicates that the Po2 is the most performant, and the red shading indicates that the NPo2 performs better.
- The first of the two rightmost columns show the relative performance of each of the RAID Life Cycles, with the NPo2 being the baseline and the Po2 being the challenger. The rightmost column shows the equally weighted average of the three RAID Life Cycles. Of course, RAID systems are designed for higher reliability, and the vast majority of time will be spent in normal operation. The table provides enough information to support a more detailed analysis by adjusting the weighting of the RAID Life Cycle as appropriate.
- The NPo2 is more performant in all tests for the SSR workloads and provides an overall 8.49% advantage.
- For the SSS workloads, the NPo2 is more performant only in the rebuild state by 6.17%, while Po2 is more performant across all of the RAID Life Cycles by 6.99%.

Note: The following Bandwidth Summary table reduces a large spreadsheet to a reasonably small table. All details are available.

	Bandwidth							
	Life_Cycle	Workload	NPo2BWraw	Po2BWraw	NPo2BWnorm	Po2BWnorm	Avg per SS and Life Cycle	Sum per SS Workloads
			BW_GB_s	BW_GB_s	BW_GB_s	BW_GB_s		
Spec Sheet Random (SSR)	Normal	Random Write	3.13	3.46	0.39	0.38		
	Normal	Random Mixed	6.87	7.26	0.86	0.81		
	Normal	Random Read	11.44	11.43	1.43	1.27	-6.31%	
	Degraded	Random Write	2.44	2.69	0.30	0.30		
	Degraded	Random Mixed	4.64	4.81	0.58	0.53		
	Degraded	Random Read	6.54	6.43	0.82	0.71	-7.45%	
	Rebuild	Random Write	1.98	2.02	0.25	0.22		
	Rebuild	Random Mixed	3.58	3.59	0.45	0.40		
	Rebuild	Random Read	5.32	5.09	0.67	0.57	-11.73%	-8.49%
Spec Sheet Sequential (SSS)	Normal	sread	11.80	11.80	1.48	1.31		
	Normal	swrite	8.02	10.58	1.00	1.18	3.08%	
	Degraded	sread	9.05	11.80	1.13	1.31		
	Degraded	swrite	7.45	11.08	0.93	1.23	24.06%	
	Rebuild	sread	9.15	7.88	1.14	0.88		
	Rebuild	swrite	7.10	8.87	0.89	0.99	-6.17%	6.99%

Bandwidth Summary

CPU Summary

- The general comments for Bandwidth Summary still apply to this data. The shading is still the same, but because CPU usage is a time measure, lower (faster) is better, and the table accommodates this relationship.
- While NPo2 is still the most performant solution with an overall advantage of 1.52%, it achieves this because of its normal operation efficiency.
- Po2 is the most performant solution in all categories of the SSS by a 45.89% margin.
  - This is impressive and likely accomplishes this by exploiting AVX for the larger block sizes. The 4K workload may not be able to exploit AVX as there are 32x more IOs that have to be handled, and each IO may require server CPU resources.

Note: The following CPU Summary table reduces a large spreadsheet to a reasonably small table. All details are available.

	CPU							
	Life_Cycle	Workload	NPo2CPUraw	Po2CPUraw	NPo2CPUnorm	Po2CPUnorm	Avg per SS and Life Cycle	Sum per SS Workloads
			CPU PCT (1-iowait-devbusy)	CPU PCT (1-iowait-devbusy)	NPo2CPUraw / NPo2BWraw	Po2CPUraw / Po2BWraw		
Spec Sheet Random (SSR)	Normal	Random Write	37.09	53.84	11.84	15.55		
	Normal	Random Mixed	50.65	56.97	7.37	7.85		
	Normal	Random Read	55.70	52.47	4.87	4.59	-7.92%	
	Degraded	Random Write	49.93	54.13	20.50	20.15		
	Degraded	Random Mixed	54.27	55.49	11.69	11.53		
	Degraded	Random Read	51.82	50.67	7.93	7.88	1.25%	
	Rebuild	Random Write	46.52	46.53	23.48	23.01		
	Rebuild	Random Mixed	47.45	45.42	13.25	12.66		
	Rebuild	Random Read	41.84	40.17	7.86	7.89	2.12%	-1.52%
Spec Sheet Sequential (SSS)	Normal	sread	11.03	9.59	0.93	0.81		
	Normal	swrite	21.71	15.81	2.71	1.49	48.09%	
	Degraded	sread	11.56	13.23	1.28	1.12		
	Degraded	swrite	21.54	15.68	2.89	1.42	59.14%	
	Rebuild	sread	9.51	8.59	1.04	1.09		
	Rebuild	swrite	18.98	14.33	2.68	1.62	30.45%	45.89%

CPU Summary

General RAID Observations

Choosing a RAID solution should not be taken lightly; it requires knowledge and effort to tailor it. You need to know in detail your workload and business requirements.

This study is an example and a starting point for RAID solution design and analysis.

In particular:

- Parity RAID is often chosen because it has the lowest general cost, as you can protect against losing data from a single device failure by adding one more device to a set of devices.
- The RAID configuration must be chosen with respect to the workload requirements and the performance reduction when a RAID Set is degraded or rebuilding. In some cases, Parity RAID will not provide the required performance, and a Mirrored RAID solution is required.

xiRAID R5 is a well-designed and reliable solution, but as indicated in the prior charts, there are considerations in terms of performance, and the business requirements must be met during Normal, Degraded, and Rebuild RAID Life Cycles.

xiRAID performance:

- xiRAID RAID5 performance for the Normal Life Cycle for read workloads approaches No-RAID performance.
- CPU consumption for sequential workloads is excellent as xiRAID exploits AVX.

The classic Disaster Recovery questions (data unavailability or data loss is usually a disaster) are: What is the RTO (Recovery Time Objective) and RPO (Recovery Point Objective)? RTO and RPO are defined by the business, statutory and regulatory requirements, due diligence, company image, etc.

RAID can help avoid disasters, but you probably still need a good backup and recovery strategy.

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