A Comparative Analysis: Graid Technology SupremeRAID™ NVMe® & NVMe-oF™ RAID vs Xinnor xiRAID Software RAID

Executive Summary

Modern datacenter infrastructure must support data hungry applications and feed data to these applications at ever increasing speeds. NVM Express® (NVMe®) SSDs, which use the PCI Express® (PCIe®) physical interface, have become the frontrunner for raw storage device performance, both in terms of IO's per second (IOPS) and throughput, typically measured in Megabytes or Gigabytes per second (GB/s).

As enterprises refresh their servers and storage devices from the older PCIe 3 generation technology to PCIe 4 and more recently to PCIe 5, they need a clear path to provide storage performance and data protection for their highest-speed storage devices.

Providing RAID protection for high-performing NVMe SSDs while maintaining application performance can be a challenge due to the CPU consumption required to perform the RAID calculations.

GPUs are known for providing excellent acceleration to video graphics and related activities, in part, because these activities require many repetitive mathematical calculations to be performed and these GPUs are designed specifically to accomplish this purpose. RAID processing also involves extensive and repetitive mathematical calculations, especially for RAID5 and RAID6, so GPU acceleration is a natural fit.

Graid Technology has released two models of GPU-accelerated SupremeRAID™ controllers that deliver full performance of NVMe SSDs in PCIe Gen 3, 4, and 5 servers. These controllers combine two technologies, software defined storage (SDS) and GPU acceleration, to create an innovative RAID solution for NVMe storage. The GPU offloads the RAID computations to the hundreds of cores in the GPU, relieving the server CPU cores of the RAID computation burden. This reduces the load on the host CPUs and frees CPU cycles that can be directed to application workloads.

This report examines the advantages of GPU RAID and provides an analysis of test results from two competing NVMe RAID solutions:

- Graid Technology GPU RAID
- Xinnor xiRAID Software RAID

Advantages of GPU RAID

There are several advantages of offloading the NVMe RAID computations to a GPU, and to the Graid Technology SupremeRAID card in particular. These include:

- The RAID computations and processing are ***offloaded from the host CPU***
- The GPU provides ***hundreds of cores for RAID processing***
- The GPU RAID is ***not limited to the bandwidth of 16 PCIe lanes***
- The SupremeRAID card ***runs on Windows or Linux***.
- Not dependent on the ***number of host CPU cores***
- NVMe I/O does not run in ***user space***
- Not dependent on ***special host CPU instruction set extensions***
- Not dependent on the ***type of host CPU cores***
- No special adjustments to ***host interrupt coalescing***

The following sections explain these advantages in more detail.

Offloaded from the host CPU

RAID processing, especially RAID5 and RAID6, can be computationally taxing on host CPUs. The RAID and storage performance workload increases in the event of an SSD failure and while a rebuild takes place.

Typically, customers invest in host servers to run applications, oftentimes business-critical applications. Running the applications normally takes priority over running overhead work for infrastructure such as storage and RAID processing. By offloading the bulk of the RAID processing to the GPU, customers can ensure that the host CPU cycles can be focused on running their business applications.

Hundreds of cores for RAID processing

Because GPUs typically have hundreds, or even thousands, of processing cores, the RAID calculations and processing can be split across all these cores and run in parallel. This improves performance and provides a more consistent level of performance.

Not limited to the bandwidth of 16 PCIe lanes

Other ways to accelerate RAID performance include hardware RAID adapters based on ASICs or FPGAs. Typically, these solutions are delivered in the form of a PCIe add-in-card that supports either 8 or 16 PCIe lanes. It doesn’t take many 4-lane NVMe SSDs to saturate this type of solution.

The Graid Technology SupremeRAID solution is not limited to the amount of PCIe lanes in a single server slot but passes the I/O traffic through all available PCIe lanes to all the NVMe SSDs. This also means that SupremeRAID does not need special cabling to enable the GPU to connect to or manage the NVMe SSDs.

Runs on Windows or Linux

Graid Technology SupremeRAID solution is currently available for Microsoft® Windows® and Linux®.

Number of host CPU cores

Enterprises often deploy systems that require NVMe RAID, including servers and workstations, with different core counts depending on where these enterprises are in their server and workstation refresh cycles.

The performance of competing software RAID solutions is often dependent on the number of CPU cores in the host system.

For example, an enterprise might deploy an application environment that requires NVMe RAID in a server with 8 cores, 16 cores, 32 cores, or more. A solution based on software RAID will likely perform differently depending on the number of CPU cores in the system.

Typically, I/O traffic to NVMe SSDs in Linux environments goes through kernel drivers. Some software RAID solutions attempt to run entirely in user space, rather than using kernel drivers to reduce the number of interrupts required. However, this type of solution is also dependent on the number of host CPU cores.

Intel ran several tests with NVMe drives in various configurations of server and NVMe-oF targets with various core counts, using the user space Storage Performance Development Kit (SPDK) NVMe drivers. The performance of these NVMe storage systems was dependent, in part, on the number of cores in the server hosting the NVMe SSDs. Another factor in the overall performance was the total number of NVMe SSDs in the configuration. These performance tests are available in the "Performance Reports" section of this page: https://spdk.io/doc/.

User space vs kernel space

For Linux solutions, the Graid SupremeRAID GPU solution treats Linux block devices as regular Linux block devices, accessed via kernel drivers, and maintains stable performance.

Some software RAID solutions have begun to experiment with SPDK to improve performance and CPU efficiency by running in user space rather than with kernel mode drivers. SPDK can provide performance improvements, but according to Xinnor, it does not currently support Linux block devices and SPDK support for Linux "ublk" targets is not currently considered production ready.

Special host CPU instruction set extensions

To reduce the time that complex RAID calculations can take, some software RAID solutions take advantage of instruction set extensions known as Advanced Vector Instructions (AVX). These extensions are known as single instruction multiple data (SIMD) instructions because they can perform the same instruction on multiple pieces of data simultaneously.

The multiple pieces of data can be 128 bits, 256 bits, or, in some cases, 512 bits in length, and can be placed in special registers in the CPU or at locations in memory.

Intel recently announced changes to its implementation of the AVX instructions in "The Converged Vector ISA: Intel® Advanced Vector Extensions 10 Technical Paper, version 1.0, July 2023." Intel comments in this technical paper that "the developer community has provided feedback that the current Intel AVX-512 enumeration method has become increasingly unwieldy over time."

The roadmap described in this technical paper highlights the somewhat confusing nature of these AVX instructions and which Intel processors and processor families support the various instructions, special registers, and different data lengths. This is true for currently available processors and for at least one or two future processor generations.

Intel dropped support for AVX-512 in its desktop processors beginning with its 12th generation ("Alder Lake") processors. In the "Intel® Architecture Instruction Set Extensions and Future Features Programming Reference, September 2023" they state: "Alder Lake Intel Hybrid Technology will not support Intel® AVX-512."

One security risk to using the AVX instructions is known as the "Gather Data Sampling (GDS)" or "Downfall" vulnerability, which was disclosed in August 2023 as CVE-2022-40982 and has a 6.5 (medium) severity score. In Intel's technical documentation for this vulnerability, they discuss the potential performance impacts of their vulnerability mitigation solution.

Software RAID solutions such as Xinnor xiRAID that depend on the AVX instructions may be negatively impacted by the upcoming Intel changes, security concerns, or may have difficulty providing consistent performance depending on the CPU model.

Type of host cores

In 2022, Intel introduced the concept of performance cores (p-cores) and efficient cores (e-cores) into desktop processors, similar in concept to smaller processors found in mobile phones. The e-cores of these processors do not support all the instruction set extensions that the p-cores support.

In August 2023, Intel announced a future series of server processors based on efficient cores that targets density-optimization and power efficiency. At this point, support for specific AVX instructions and data lengths is unclear for these new e-core server processors.

Host interrupt coalescing

In busy environments, host systems can be flooded with too many interrupts when processing data transfers from external devices. One way to alleviate this situation is to hold these interrupts for a short period of time and process groups of interrupts together as a single batched interrupt. This technique is known as *host interrupt coalescing* and has been used for some time with network adapters for network traffic.

The advantage of this process is that multiple packets of data can be processed quickly under heavy load. The disadvantage is that latency can increase (gets worse) the longer the interrupts are held before processing, especially during periods of light traffic. Administrators must finely tune the settings to achieve the right balance of maximizing performance improvements under heavy load while minimizing latency under light load.

With the advent of high-speed NVMe SSDs, a similar phenomenon can occur in environments with many NVMe SSDs configured with software RAID. Host interrupt coalescing for software RAID can improve performance in some cases, but it requires tuning for the right settings to improve performance under heavy load while at the same time keeping latency low. For optimal performance, these interrupt coalescing parameters need to be adjusted as the workload or environment changes.

For example, if the timeout value associated with NVMe software RAID interrupt coalescing is increased appropriately for heavy load conditions, during times of light load conditions that same setting will cause performance to significantly degrade.

Another side effect of host interrupt coalescing for NVMe software RAID is that while performance will improve under heavy load, CPU utilization will also significantly increase, so much so that the software RAID processing can consume most of the available CPU cycles on most or all the CPU cores.

Performance results

Although each of the two companies provide various test results of their NVMe RAID solutions on their respective websites, none of these test configurations are similar enough to make reasonable comparisons. Graid Technology ran the following set of tests using the same hardware configuration, which Dennis Martin Consulting audited.

This set of tests uses the same PCIe gen 4 server, processors, memory, and NVMe SSDs, along with the same version of Ubuntu® Linux®.

These tests use the synthetic I/O benchmark known as fio. This is a well-known standard tool used by many in the industry. It is known as a synthetic benchmark because it is designed to run workloads by controlling specific technical I/O parameter settings such as block size, read and/or write, random and/or sequential I/O, and various numbers of I/O queue depths and jobs. This is different than an application benchmark where the test runs a real-world application, such as databases, video rendering, video streaming, backup applications, etc. These real-world benchmarks use a variety of these technical I/O parameters under the hood but are not specifically controlled by the test.

The specific test runs are identified by the I/O type (random or sequential, and read or write), then by the block size, then the number of jobs, and the I/O queue depth.

For example, the job titled “Rndread b4k j128 d32” has the following characteristics:

- Random read
- 4 KB block size reads
- 128 jobs
- I/O queue depth of 32 (32 I/O’s issued at the same time)

The environment for these tests included:

- Server: Supermicro SYS-220U-TNR
- CPU: 2x Intel Xeon Gold, 2.0 GHz, 32-core / 64-thread
- Memory: 16x SK Hynix DDR4 3200 MT/s, 16GB (HMA82GR7CJR8N-XN)
- NVMe SSDs: 22x Intel P5510, 3.84 TB (SSDPF2KX038TZ)

This server, and the NVMe SSDs, are PCI Express gen 4.

The operating system was Ubuntu Linux 22.04.1 LTS with kernel 5.15.0.57-generic.

The benchmark tool was fio 3.30.

The RAID versions were:

- Graid Technology SupremeRAID version 1.5
- Xinnor xiRAID version 4.0.1

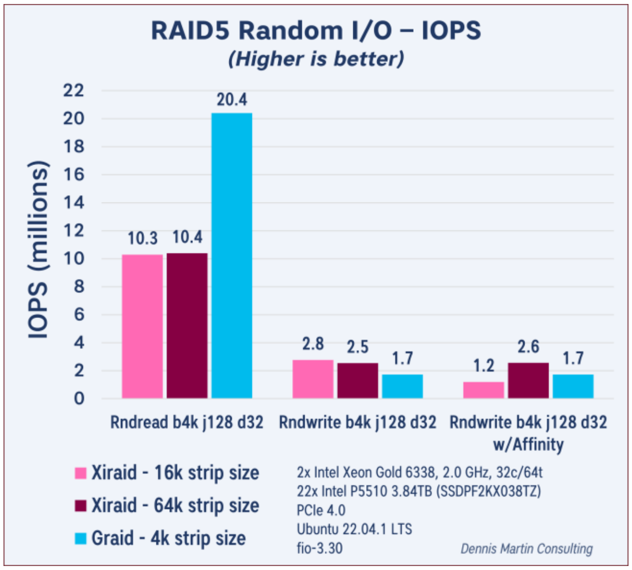
The strip size is the amount of data written to a single SSD as part of a single data stripe. The stripe width is the amount of data written to all the data SSDs simultaneously, not counting parity. In this set of tests, there are 22 SSDs, so the stripe width is (22-1) x the strip size.

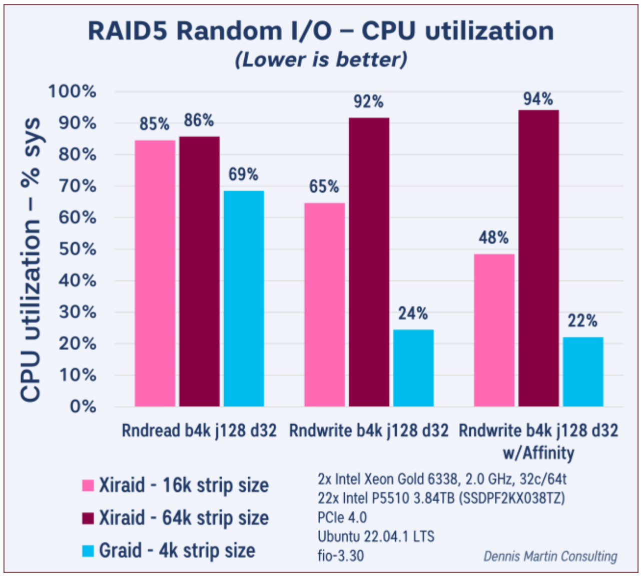
As the strip size increases, the stripe width also increases. If the stripe width exceeds the I/O block size for a write operation, then a read-modify-write must take place which leads to additional smaller I/O's, which can potentially lower performance and increase latency (the round-trip time).

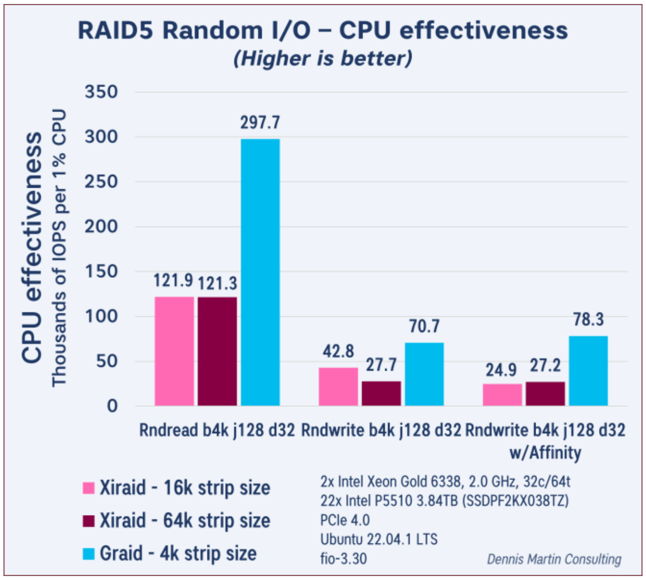
The performance results are divided into two sections, one for random I/O and one for sequential I/O. The key metric in these tests is "CPU Effectiveness" which is a measure of the amount of I/O performed in relation to the amount of host CPU consumed. This is expressed as I/O per 1% CPU.

RAID5 Random I/O

Random I/O is typically used for transactional applications where data can be placed at various (seemingly random) locations in the files, databases, volumes, etc.



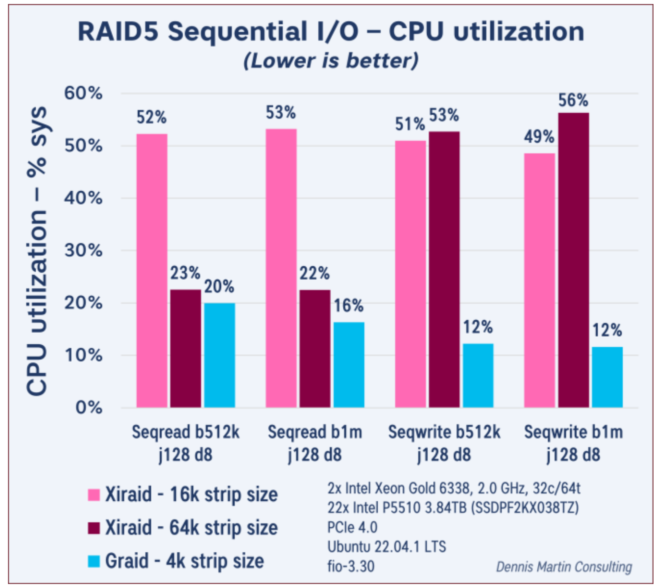


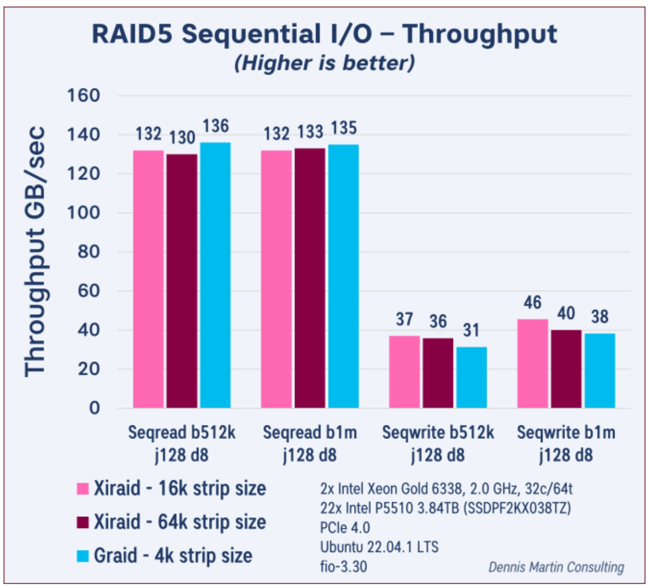


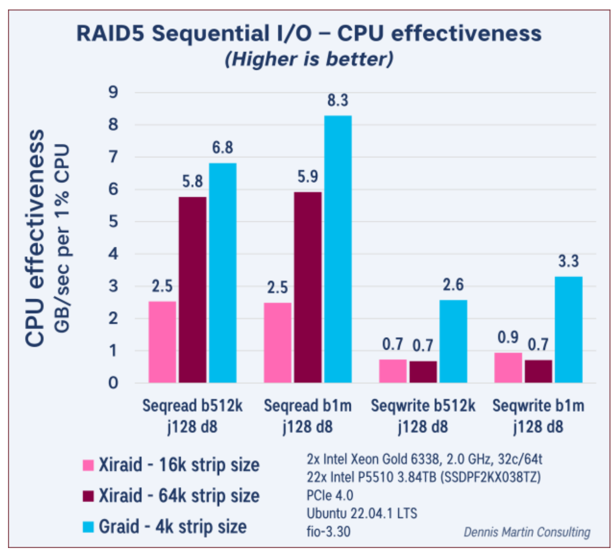
The Graid SupremeRAID solution has much better random I/O CPU effectiveness. This means that more CPU cycles are available for application workload processing for the given storage workload.

RAID5 Sequential I/O

Sequential I/O reads and writes are used by applications such as backup applications, video streaming, database imports and exports, and other applications that read or write entire files, databases, volumes, etc., in sequence, or at least large portions of the data sequentially.







Although the overall sequential throughput results are similar for the two solutions, the Graid SupremeRAID solution has much better sequential I/O CPU effectiveness, especially for sequential writes. As we observed in the random I/O tests, this means that more CPU cycles are available for application workload processing for the given storage workload.

Conclusion

The Graid Technology SupremeRAID solution offers several advantages over other NVMe software RAID solutions, including Xinnor xiRAID, offering better CPU effectiveness, freeing up host CPU resources to focus on application workloads rather than storage RAID overhead.

The SupremeRAID solution is not dependent on the number or type of host CPU cores, instruction set inconsistencies, and does not require fine tuning of host interrupts. It is available for Windows or Linux environments and provides the advantages of GPU offloading. For Linux environments, Graid SupremeRAID treats NVMe block devices in the normal way that production environments expect.

This is a solution well worth investigating for customers that are deploying NVMe SSDs into production, either in a host server or in an NVMe-oF target environment.