App note

Graid Technology + ScaleFlux

Improve User Experience & SLAs by using Accelerated RAID and Hardware-Compression to control tail latency

Graid Technology Inc.

ScaleFlux

Table of Contents

1 Introduction

The benefits of RAID are well known. Aggregating the resources of multiple disks can improve throughput, add protection against one or more disk failures, and enable flexible capacity management. The RAID concept was introduced in a 1988 SIGMOD paper titled "A Case for Redundant Arrays of Inexpensive Disks (RAID)" by the legendary David A. Patterson, Garth Gibson, and Randy H. Katz [1]. At the time this paper was published, a typical disk provided throughput on the order of a MB/s with double digit millisecond latency. RAID quickly proved to be an indispensable tool to scale performance while improving reliability. Over three decades later, a single NVMe disk provides multiple GB/s of throughput with sub-millisecond latency (and 100,000 times more capacity). And yet the core tenet of RAID remains unchanged: make the array better than the sum of its disks.

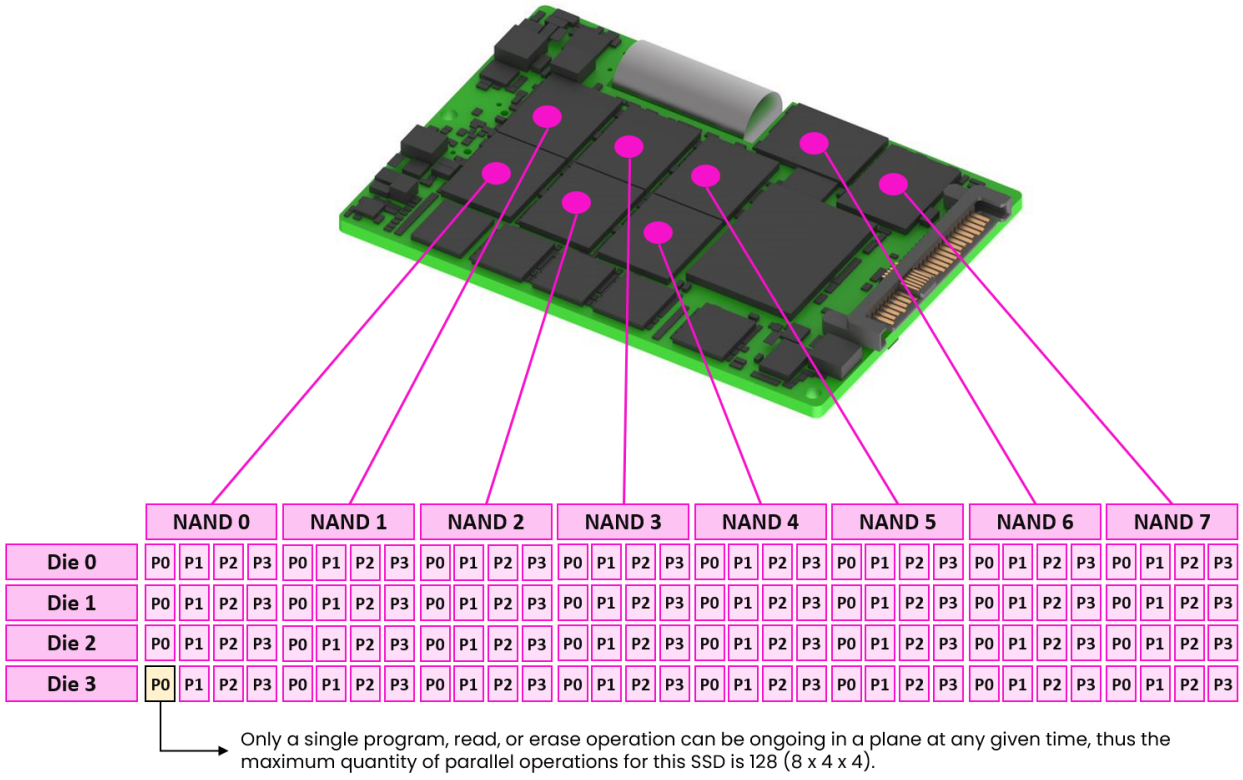
There are two key challenges faced by modern RAID solutions. First, IO performance is scaling much faster than traditional compute resources [x]. Second, latency has become the key performance metric in the datacenter. Specifically, read tail latency [x]. The SupremeRAID™ SR-1000 offers a solution to the RAID performance scaling challenge by adopting a heterogeneous architecture where compute-intensive RAID calculations are offloaded to a GPU. The massively parallel vector processing capability of a GPU aligns well with calculating RAID parity data for many simultaneous IO operations. The performance advantages of this approach are well-documented [x]. In this application note, we focus on the latency challenge and explore how Graid Technology solutions can be used as a tool to manage read tail latency.

We pair the SupremeRAID™ SR-1000 offering with ScaleFlux® CSD-3000 series NVMe SSDs. The CSD-3000 series features transparent compression in the datapath. This technology reduces the impact of write IOs on the NAND media, freeing the media to service more read IOs and thus lowering their latency. This technology pairs well with RAID solutions where host write activity is coupled with RAID-induced write activity (e.g., parity writes, rebuild).

2 Understanding SSD Latency

A single SSD is comprised of a controller and a collection of independent NAND Flash memory packages. The primary function of the controller is to turn this collection of NAND Flash memory packages into a monolithic storage media. In NVMe terminology, this storage media is made accessible to the host via one or more namespaces, which provide a continuous range of logical block addresses (or LBAs). The SSD controller is responsible for the dynamic mapping of logical block addresses to physical addresses in the media (through what's often referred to as the Flash Translation Layer, or FTL).

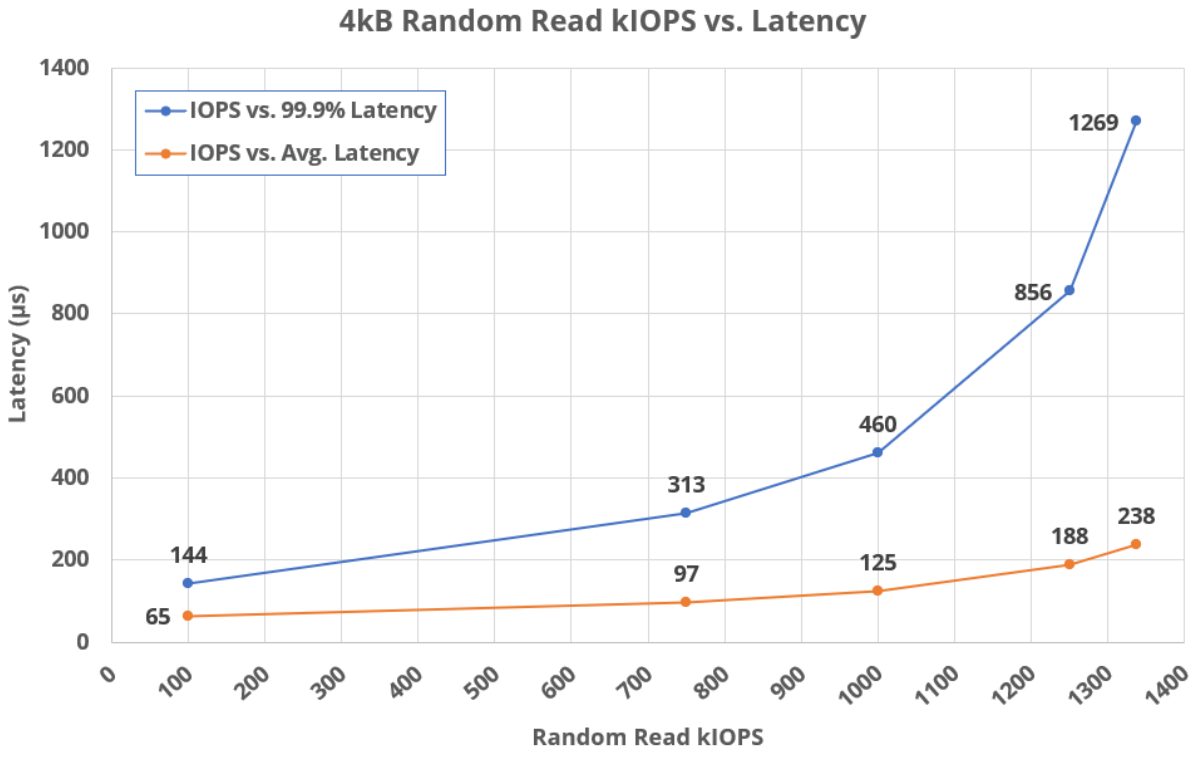
Each NAND Flash memory package contains one or more Flash Memory die inside. Each NAND Flash die is divided into a small number of planes, typically between two and eight. A key attribute of NAND Flash memory is that only a single program, read, or erase operation can be ongoing in a plane at any given time. Therefore, the quantity of NAND Flash die contained in an SSD (or more accurately the total number of planes) determines the maximum IO parallelism that can be achieved. Consider an SSD with eight NAND Flash memory packages, each containing four NAND Flash die with four planes per die:



If the above SSD offered 3.84TB of usable capacity, each plane holds about 30GB and the LBA space is effectively broken into 128 silos. To achieve the maximum parallelism of the SSD, the LBAs being accessed must be distributed across all planes; however, there is generally no way for the host to know which plane a particular LBA may belong to. Latency grows as the number of operations requiring access to the same plane increases.

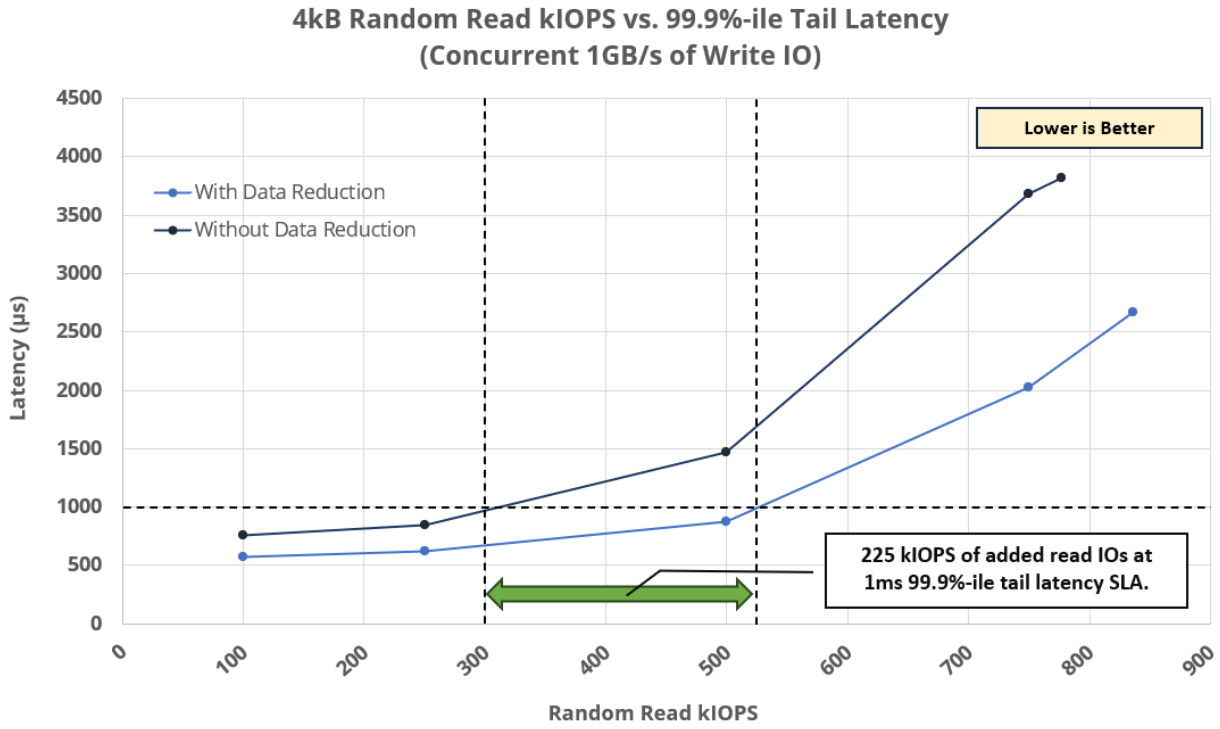
2.1 Latency Reduction with Transparent Compression

To illustrate the nature of SSD latency, the read latency response of a single 3.84TB CSD-3310 is measured with and without concurrent write activity. In the following chart, the average and 99.9th percentile latency is measured as the number of 4kB random read IOPS increases. In this case, there is no concurrent host write activity.



The key observation is that there are two operating regions: one where latency increases linearly with IOPS, and another where the relationship becomes exponential. In the above plot, the SSD operates in the linear region up to between 800k-900k IOPS. From 800k-900k IOPS to saturation at 1.33M IOPS, latency increases non-linearly. The increase in tail latency reflects the probabilistic nature that IOs contend for access to the same plane.

Of course, there is typically write IO activity that is concurrent with read IO. A write operation to Flash Memory takes on the order of 10x more time to complete compared to a read operation. Many controllers have adopted program suspend and other techniques to ease the mismatch in operational latency, but ultimately writes and read compete for access to the same media in what is referred to as write-to-read interference. The transparent compression technology found in the CSD-3310 is an important tool in reducing this write-to-read interference. It removes redundancy in the data to reduce the quantity of write operations in the NAND Flash media. The following figure shows the effect of this transparent compression technology when applied to a host data that can be reduced by approximately a 2:1 ratio.



The results of transparent compression are dramatic. By reducing the write-to-read interference of a one GB/s write stream, the quantity of read IOPS that can be achieved with the same tail latency increases from 300k to 525k IOPS.

2.2 Other Methods to Reduce Tail Latency

Since SSD latency is fundamentally limited by the quantity of NAND Flash die present in the media, one obvious solution is to simply increase the number of NAND Flash die. Of course, this increases the capacity of the SSD and there are limits to this type of "vertical" scaling. A typical datacenter class controller has 16 independent data buses (or channels) on which to connect NAND Flash memories. One limitation comes from the quantity of NAND die that can fit into a single NAND Flash memory package. The upper limit today is eight die that can be connected to a channel by a single package. Additional packages can be added, but this increases capacitive loading on the channel to a point that is very difficult to manage. More packages are also difficult (or impossible) to fit in smaller SSD form factors, such as E1.S. Even with a very large number of dies, contention on the Flash channels, controller limitations/bottlenecks, or even the host interface itself may also constrain parallelism and therefore begin to drive latency higher. At some point, scaling SSD performance requires "horizontal" scaling – adding more SSDs to support the workload.

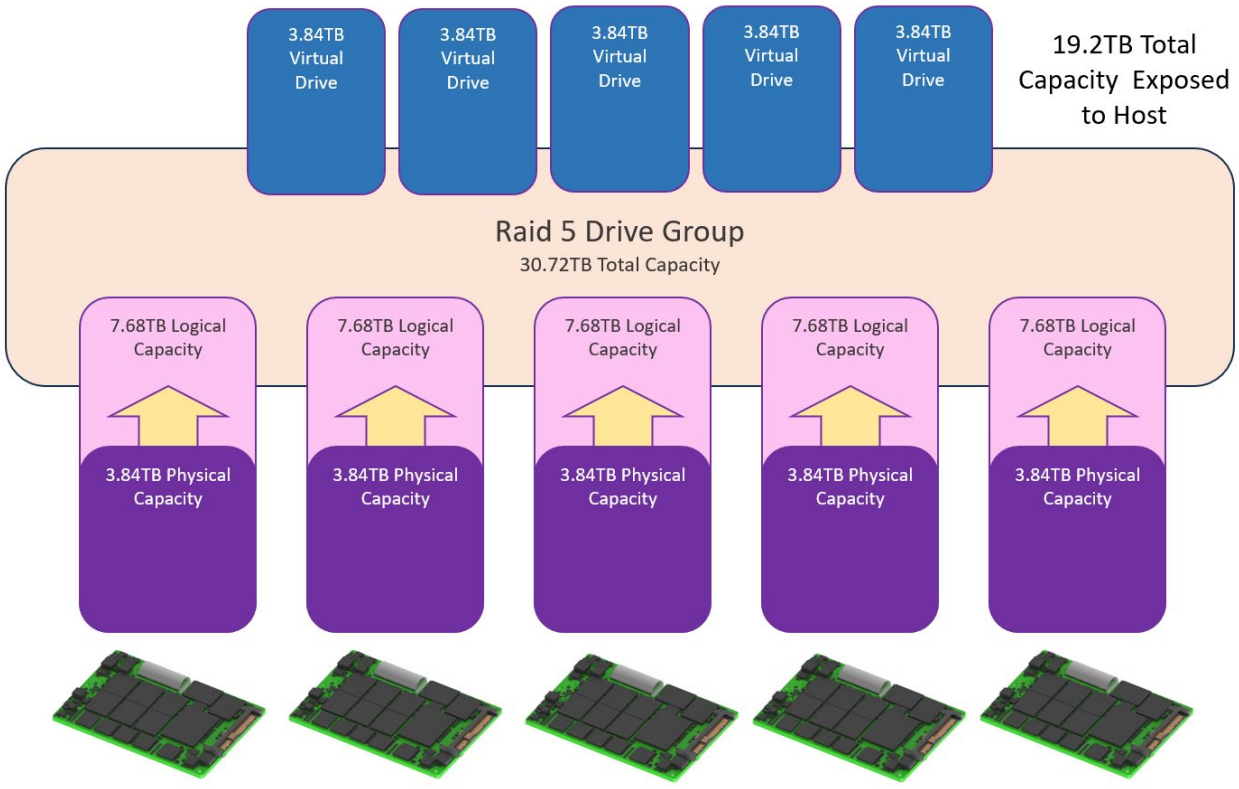
In an environment where multiple disks are required to support a workload, a mechanism is required to manage the distribution of data across multiple disks. This can be accomplished in a variety of ways: in software by a volume manager (e.g., LVM), by software RAID (e.g., mdadm), by the application itself (e.g., Aerospike Database), or by hardware solutions (e.g., traditional RAID cards). Each of these approaches has advantages and disadvantages. Some software solutions like mdadm are highly CPU limited. Traditional hardware solutions can introduce their own bottlenecks, whether in the RAID controller SoC, or simply in a mismatch in host PCIe bandwidth compared to SSD bandwidth.

2.3 Using Graid Technology to Reduce Latency

The unique architecture of SupremeRAID offers a new path for efficiently aggregating high performance NVMe SSDs. It offers the speed and low CPU utilization of HW-offload without limiting PCIe connectivity or introducing a bottleneck between the SSD and applications. In the Graid Technology framework, only the compute-intensive parity calculations are handled in hardware such that the SR-1000 RAID controller itself is not in the read datapath between the SSDs and the application. This enables near-native NVMe SSD performance, and more specifically, low latency profiles. Of course, it does so while providing disk failure protection and capacity management (abstracting the fact that multiple disks are participating in a logical volume). The ability for SupremeRAID to aggregate NVMe devices while preserving their latency characteristics makes it uniquely suited as a tool to provide the horizontal performance scaling needed to meet latency SLA requirements.

3 Evaluating Graid Technology Latency Profiles

This evaluation considers a collection of five 3.84TB CSD 3310 NVMe SSDs integrated into a RAID 5 pool (or drive group). The individual latency profile of these SSDs was covered in Section 3.1. Before importing the CSD-3310 drives into a SupremeRAID drive group, the capacity of each drive was extended to 7.68TB. Extending the logical capacity above the physical capacity is a feature of transparent compression technology that turns the reduction in data written into additional host capacity through the standard NVMe thin-provisioning framework. In this test, we use the additional capacity to reclaim the capacity used by RAID 5 for parity data. The scheme is illustrated in the following diagram:



Note that there is 5 x 3.84TB or 19.2TB of physical storage. Because the capacity of the drives was extended (through NVMe thin provisioning), a total of 5 x 7.68TB or 38.4TB is seen by SupremeRAID. For RAID 5, the available capacity of the drive group drops by one drive's worth of capacity to 30.72TB (due to RAID5 parity overhead). 5 virtual drives of 3.84TB are then created providing a total of 19.2TB of storage to the host. The minimum aggregate data reduction rate of the CSD-3310s participating in the array must only be 1.2:1 (or 20% reduction in data written) to compensate for the capacity used for parity data. If the data is further compressible, additional virtual drives can be created to consume the additional capacity. The elasticity offered by SupremeRAID makes it very convenient to benefit from the capacity recovered by transparent compression.

The graidcli utility provides a succinct summary:

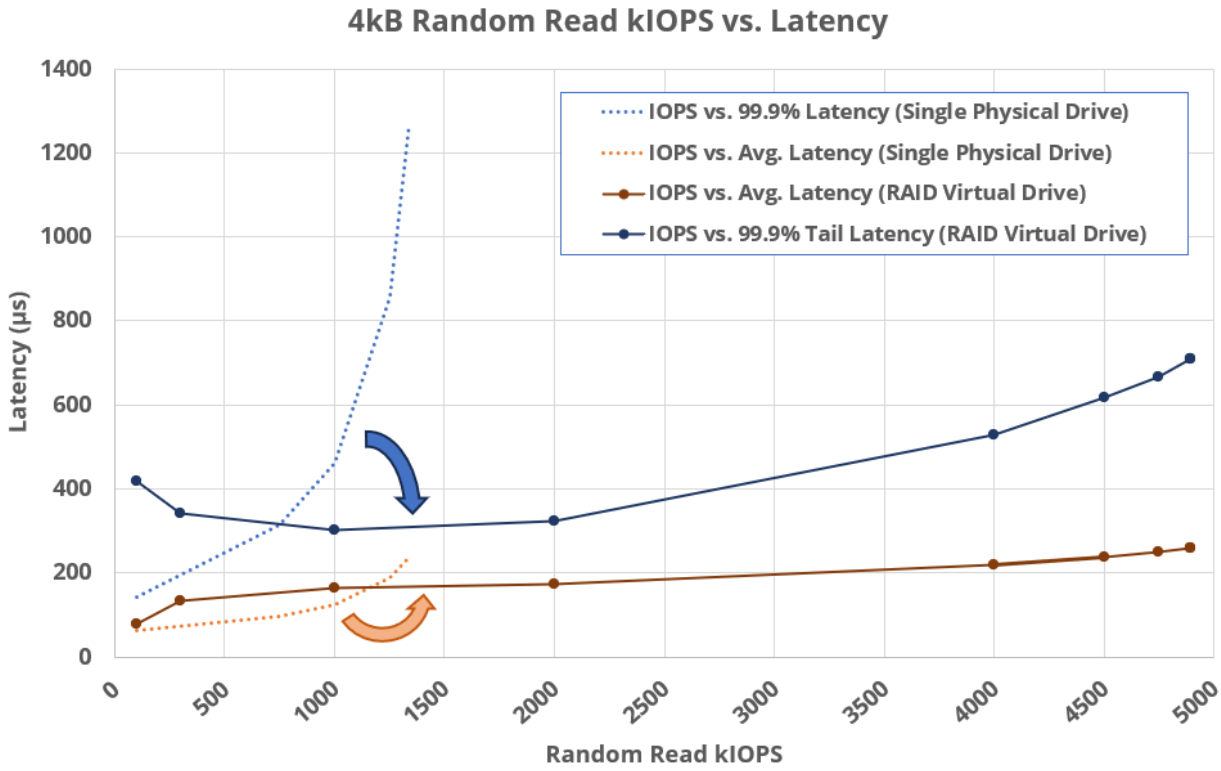
```
$ sudo graidctl list drive_group

✓List drive group successfully.
|————————|————————|————————|————————|————————|————————|————————————————————|————————————|
| DG ID  | MODE   | VD NUM | CAPACITY | FREE   | USED   | CONTROLLER                 | STATE      |
|————————|————————|————————|————————|————————|————————|————————————————————|————————————|
| 0      | RAID5  | 5      | 28 TiB   | 10 TiB | 18 TiB | running: 0 prefer: 0       | OPTIMAL    |
|————————|————————|————————|————————|————————|————————|————————————————————|————————————|
```

Note that the graidcli reports in TiB (1kB = 1024 bytes) and not TB (1kB = 1000 bytes).

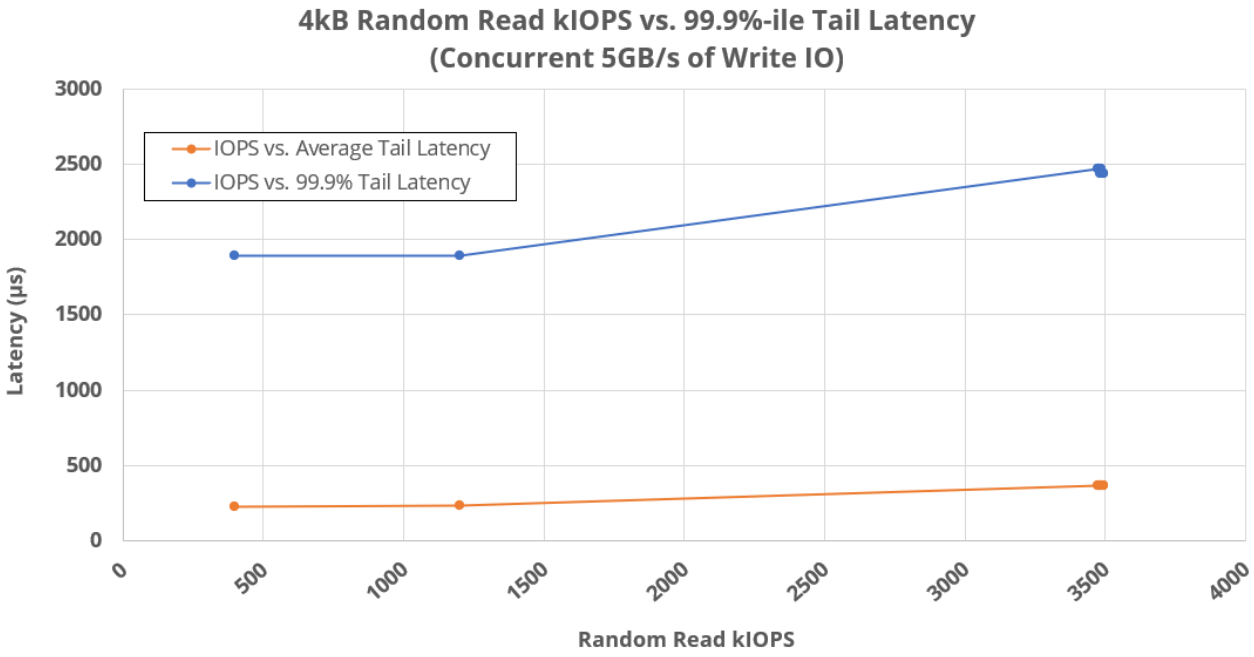
3.1 3.84 TB Virtual Drive Performance

The single 3.84TB virtual drive can use the resources of the full RAID pool, which allows it to exhibit much higher performance than a single drive in isolation. The following figure shows the 4kB random read performance of the virtual drive compared to the single drive case:



Where a single drive saturates at 1.3M IOPS, the virtual drive is able scale out to nearly 6M IOPS while maintaining sub 1ms read tail latency. This shows the SupremeRAID array's ability to scale performance horizontally across all drives in the RAID array.

Adding a very high write workload of 5GB/s to the virtual drive.



Under a very high write workload, the latency response as the number of random read IOPS remains very flat until saturation is reached at 3.5M IOPS.

4 Conclusion

In the above section, the performance of a single 3.84TB CSD-3310 drive tested in isolation was compared to a SupremeRAID 3.84TB virtual drive supported by a RAID 5 drive group consisting of five 3.84TB CSD-3310s. The results show that the SupremeRAID array enabled all the resources of each SSD to be leveraged by a single logical volume.

The ability to pool NVMe SSDs and provision logical volumes that can utilize the full parallelism provided by the underlying storage can be used as a critical tool to manage tail latency. This tool can be used to:

1. Create virtual volumes that outperform any physical volume and provide consistent, low tail latency at multi-million IOPS performance levels.
2. Avoid stranding SSD performance by many virtual volumes to use the full performance provided by the drive pool. This is particularly suited to bursty workloads where a single virtual drive can deliver many millions of IOPS and low latency at any capacity point.

The extended capacity feature offered by the 3.84TB CSD-3310s was used to increase their logical capacity to 7.68TB. This provided a larger capacity pool for the SupremeRAID drive pool from which to create virtual drives. The elastic nature of the drive pools enables more physical drives to be added when additional physical capacity is needed, or additional virtual drives to be created when there is capacity that can be recovered by the CSD-3310's transparent compression capability.

About Graid Technology

> *SupremeRAID™: The Revolutionary Next-Generation NVMe RAID Controller*

Our mission is to deliver our customers the next generation of IT storage infrastructure for NVMe and NVMeoF SSDs, without sacrificing the performance they need. SupremeRAIDTM is a revolutionary GPU-based RAID that delivers the resiliency, speed, ease of use, flexibility and TCO the market demands for the future of high- performance workloads.

Graid Technology Inc.

www.graidtech.com

About ScaleFlux

> *The Better SSD delivered to your door.*

ScaleFlux helps customers harness data growth as a competitive advantage by rethinking the data pipeline for the modern data center. The company is passionate about bringing innovative technologies to market by building products that accelerate value creation from data while reducing complexity. With its initial products, ScaleFlux offers a simple solution to enterprise customers struggling with data-hungry workloads – a better SSD built with computational storage.

ScaleFlux

sales@scaleflux.com

www.scaleflux.com