

应用案例

# Graid Technology + ScaleFlux

---

通过使用加速 RAID 和硬件压缩技术来控制尾部延迟，从而提升用户体验并更高效地满足服务水平协议 (SLA) 要求。



# 目录

|   |                            |    |
|---|----------------------------|----|
| 1 | <b>背景介绍</b>                |    |
|   | 背景介绍 .....                 | 3  |
| 2 | <b>理解固态硬盘（SSD）的延迟</b>      |    |
|   | 使用透明压缩技术降低延迟 .....         | 5  |
|   | 减少尾部延迟的其他方法 .....          | 6  |
|   | 使用 Graid 技术降低延迟 .....      | 7  |
| 3 | <b>评估 Graid 技术的延迟性能</b>    |    |
|   | 3.84 TB 的虚拟存储设备的性能表现 ..... | 9  |
| 4 | <b>结论</b>                  |    |
|   | 结论 .....                   | 10 |



## 1 背景介绍

RAID 的优势众所周知。通过聚合多个磁盘资源，可以提高吞吐量，增加对一个或多个磁盘故障的保护，并实现灵活的容量管理。RAID 的概念最早由 David A. Patterson、Garth Gibson 和 Randy H. Katz 于 1988 年在 SIGMOD 会议上发表的一篇题为《A Case for Redundant Arrays of Inexpensive Disks (RAID)》的论文中提出。在该论文发表时，典型磁盘的吞吐量约为每秒 1 MB，延迟为两位数毫秒级。RAID 很快被证明是提升性能和可靠性不可或缺的工具。

三十多年后的今天，单块 NVMe 磁盘的吞吐量已达数 GB/s，延迟低于毫秒级（容量更是提升了 10 万倍）。尽管如此，RAID 的核心理念依然未变：让磁盘阵列的整体性能优于磁盘的简单叠加。

然而，现代 RAID 解决方案面临两大主要挑战：

1. IO 性能的增长速度远快于传统计算资源的扩展速度；
2. 在数据中心，延迟已成为关键性能指标，尤其是读尾延迟（Read Tail Latency）。

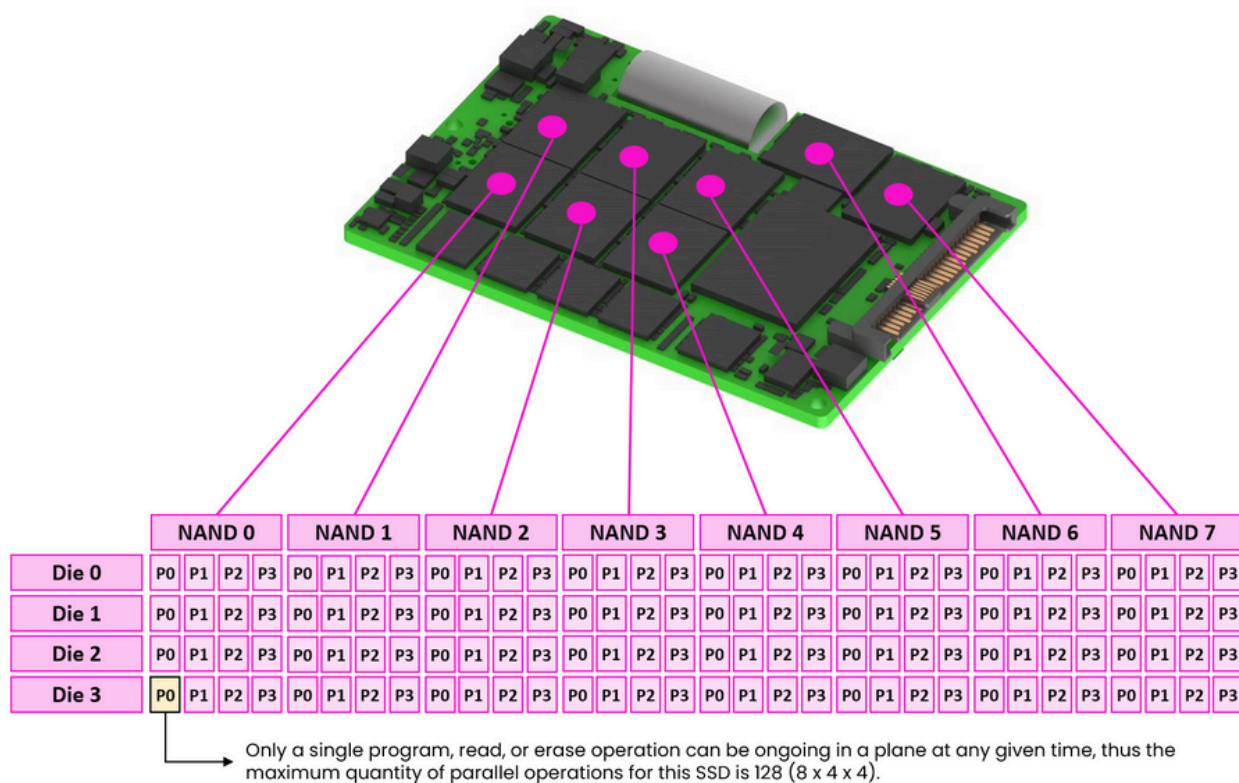
为解决 RAID 性能扩展的挑战，SupremeRAID™ SR-1000 采用了异构架构，将计算密集型的 RAID 运算卸载到 GPU 上完成。GPU 的强大并行计算能力非常适合同时处理大量 IO 操作的 RAID 校验数据计算。这种方法的性能优势已有充分文档支持。在本文中，我们将重点探讨延迟挑战，并研究如何通过 Graid Technology 的解决方案来管理读尾延迟。

我们将 SupremeRAID™ SR-1000 与 ScaleFlux® CSD3000 系列 NVMe SSD 结合使用。CSD3000 系列在数据路径中采用透明压缩技术，该技术降低了写 IO 对 NAND 介质的影响，从而释放介质去处理更多的读 IO，降低延迟。这项技术与 RAID 解决方案相得益彰，特别是在主机写操作与 RAID 引发的写操作（如校验写入、重建）结合的场景中。

## 2 理解固态硬盘（SSD）的延迟

一个 SSD 由一个控制器和多个独立的 NAND 闪存芯片包组成。控制器的主要功能是将这些 NAND 闪存芯片包转化为一个统一的存储介质。在 NVMe 术语中，这个存储介质通过一个或多个命名空间（namespace）提供给主机，命名空间提供一个连续的逻辑块地址（LBA）范围。SSD 控制器负责将逻辑块地址动态映射到介质中的物理地址，这一过程通常通过被称为闪存转换层（Flash Translation Layer, FTL）来完成。

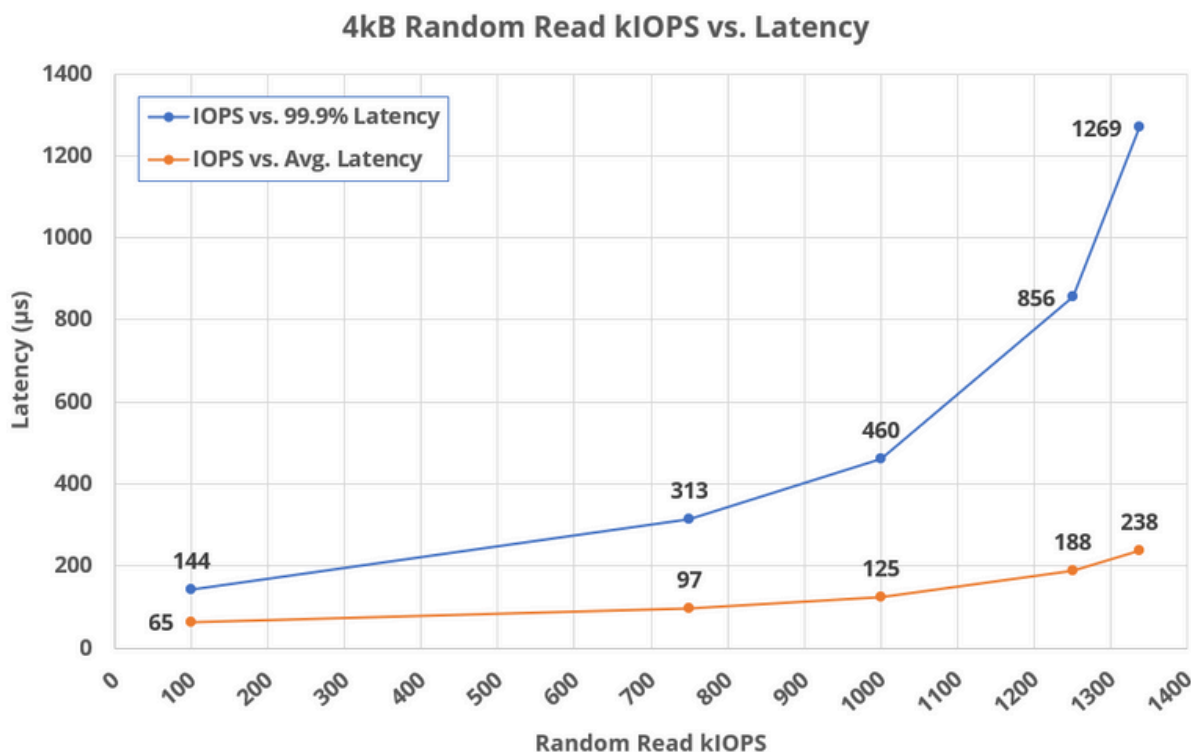
每个 NAND 闪存芯片包内部包含一个或多个闪存芯片（die）。每个 NAND 闪存芯片被划分为若干平面（plane），通常为 2 至 8 个平面。NAND 闪存的一项关键特性是，在任意时刻，一个平面只能执行一个编程、读取或擦除操作。因此，SSD 中包含的 NAND 闪存芯片数量（或者更准确地说，平面的总数）决定了能够实现的最大 IO 并行度。例如，考虑一个拥有 8 个 NAND 闪存芯片包的 SSD，每个包内包含 4 个 NAND 闪存芯片，每个芯片有 4 个平面。



如果上述 SSD 提供了 3.84TB 的可用容量，每个平面大约容纳 30GB，并且 LBA 空间有效地被分成了 128 个小块（silo）。为了实现 SSD 的最大并行性，数据访问（LBA 操作）应该尽可能分布在所有平面上；然而，通常主机无法知道某个特定的 LBA 属于哪个平面。当需要访问同一平面的操作数量增加时，延迟会随之增长。

## 2.1 使用透明压缩技术降低延迟

为了说明 SSD 延迟的特性，我们测量了 3.84TB CSD3000 在有无并发写操作的情况下的读延迟响应。下图展示了随着 4KB 随机读 IOPS 增加，平均延迟和 99.9% 百分位延迟的变化。在此测试中，没有并发的主机写操作。

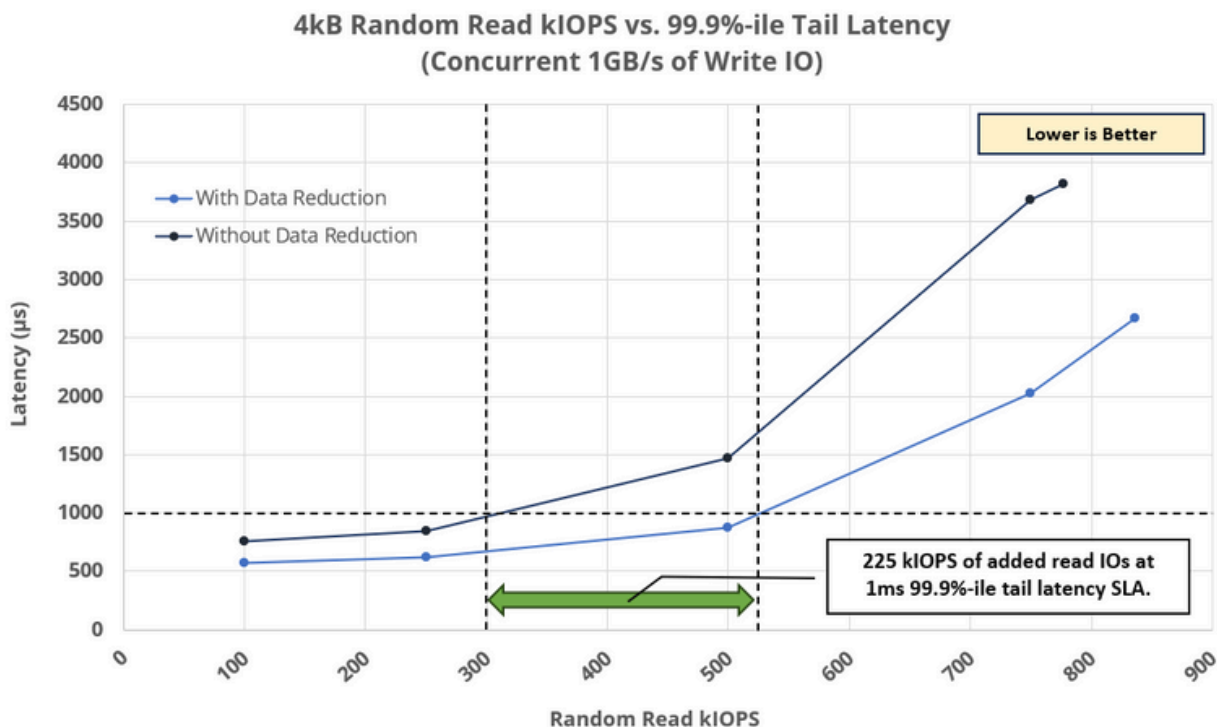


在图表中，可以看到两种不同的延迟增长模式：

- 线性增长区间：在 800k-900k IOPS 之间，延迟与 IOPS 增加成线性关系，延迟增长较为平稳。
- 非线性增长区间：当 IOPS 从 800k-900k 增加到 1.33M 时，延迟以更陡峭的非线性方式增长。

这反映了 I/O 请求竞争同一平面时，由于访问竞争，尾部延迟增加。

通常，读操作与写操作是并行进行的。与读操作相比，写操作大约需要 10 倍的时间来完成。尽管许多控制器采用了程序暂停（program suspend）等技术来缓解操作延迟的不匹配，写入和读取操作最终还是会竞争访问相同的存储介质，这种现象被称为写读干扰（write-to-read interference）。CSD3000 中的透明压缩技术有助于减少写读干扰，它通过去除数据冗余来减少 NAND 闪存中的写入操作。下图展示了当透明压缩技术应用于主机数据并实现约 2:1 压缩比时的效果。



透明压缩技术的效果十分显著。通过减少写读干扰，一个 1GB/s 的写入流可以将读 IOPS 从 300k 提升到 525k，同时保持相同的尾部延迟。

## 2.2 减少尾部延迟的其他方法

由于 SSD 延迟本质上受到介质中 NAND 闪存芯片数量的限制，一种显而易见的解决方案是简单地增加 NAND 闪存芯片的数量。当然，这也会增加 SSD 的容量，但这种“垂直”扩展是有限制的。数据中心级控制器通常有 16 条独立的数据总线（或通道），每条通道连接多个 NAND 闪存芯片封装。每个封装最多只能连接 8 个 NAND 闪存芯片。如果想要更多芯片，可以增加封装数量，但这样会增加通道的电容负载，影响管理效率。而且，在更小的 SSD 封装尺寸（如 E1.S）中，增加封装数量的空间可能有限。即使有更多的 NAND 闪存芯片，性能也可能受限于通道争用、控制器瓶颈或主机接口带宽不匹配。因此，在某些情况下，为了提高 SSD 性能，需要采取“水平”扩展，即增加更多的 SSD 来支持更大的工作负载。

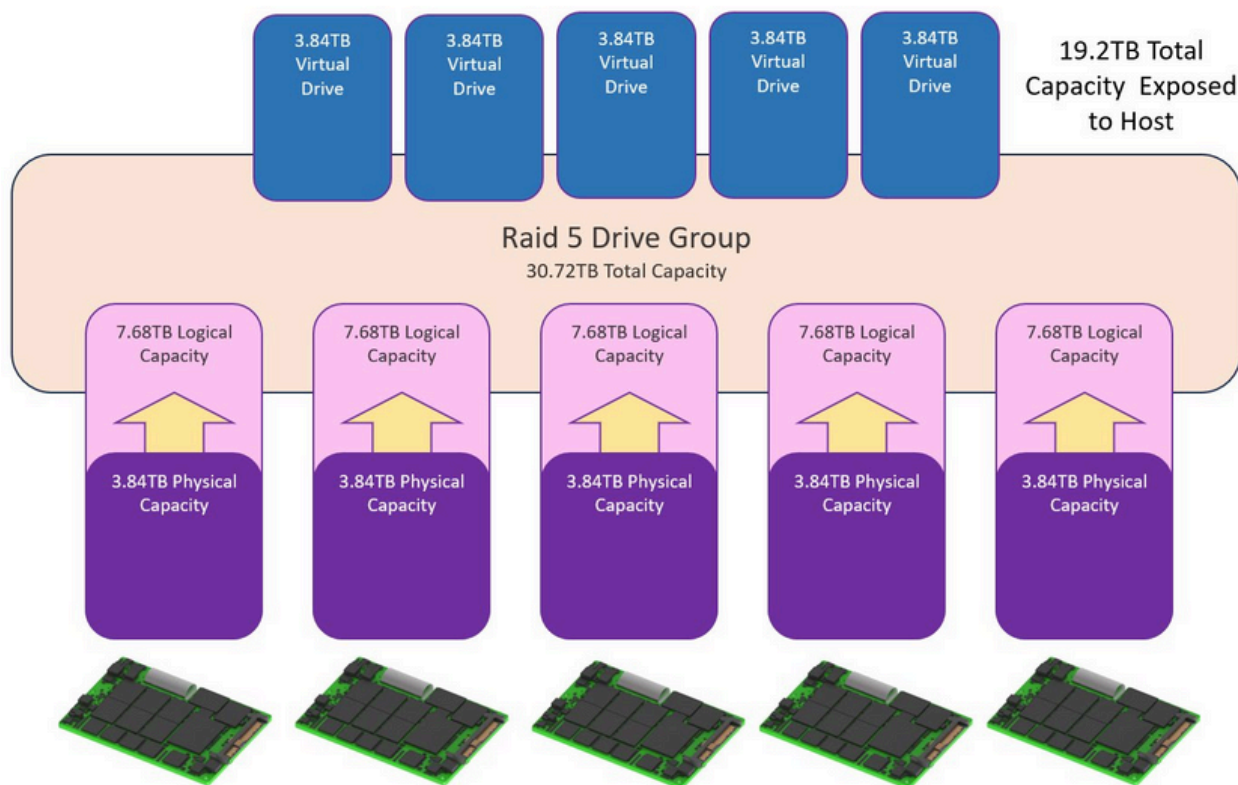
在需要多个磁盘来支持工作负载的环境中，必须有一个机制来管理数据在多个磁盘间的分布。这可以通过多种方式实现：通过软件的卷管理器（例如 LVM）、软件 RAID（例如 mdadm）、应用程序本身（例如 Aerospike 数据库）或硬件解决方案（例如传统 RAID 卡）。每种方法都有其优缺点。一些软件解决方案，如 mdadm，受限于 CPU 性能；而传统硬件解决方案则可能引入自身的瓶颈，无论是在 RAID 控制器 SoC 中，还是在主机 PCIe 带宽与 SSD 带宽之间的匹配不当。

## 2.3 使用 Graid 技术降低延迟

SupremeRAID 采用一种创新的架构，旨在高效地将多个高性能 NVMe SSD 聚合起来。它通过硬件卸载来实现高速度和低 CPU 利用率，同时避免限制 PCIe 连接性或在 SSD 与应用程序之间产生瓶颈。在 Graid Technology 的框架中，计算密集型的校验计算（即 RAID 校验）是通过硬件完成的，确保了 SR-1000 RAID 控制器不会在 SSD 和应用程序之间的读取数据路径中介入。这使得 SSD 的性能几乎保持原生 NVMe 水平，尤其是在低延迟方面。虽然 SupremeRAID 提供了磁盘故障保护和容量管理（例如隐藏多个磁盘参与一个逻辑卷的细节），但它能够在不影响延迟性能的情况下，继续保持高性能。这一能力使得 SupremeRAID 成为满足严格延迟 SLA 要求的理想选择。SupremeRAID 通过聚合 NVMe 设备，并保持这些设备的低延迟特性，能够提供水平性能扩展，以应对对低延迟的需求。这使其特别适合用于满足延迟敏感应用的性能扩展需求。

## 3 评估 Graid 技术的延迟性能

该测试使用了五个 3.84TB 容量的 CSD3000 NVMe SSD，将其集成到一个 RAID 5 池（或驱动器组）中。在导入 SupremeRAID 驱动器组之前，这些存储设备的容量被扩展至 7.68TB，扩展逻辑容量超过物理容量是透明压缩技术的一个特性，这种扩展是通过 NVMe 精简配置（thin provisioning）实现的，从而将写入数据的减少转化为额外的主机存储空间。在本次测试中，我们使用扩展的容量来回收 RAID 5 中用于校验数据的空间。具体方案如下图所示：





需要注意的是，系统中有 5 块 CSD3000，每个驱动器的容量为 3.84TB，因此 5 个驱动器的总物理存储容量是  $5 \times 3.84\text{TB} = 19.2\text{TB}$ 。每个驱动器的容量通过透明压缩扩展到 7.68TB。因此，SupremeRAID 系统看到的总容量是  $5 \times 7.68\text{TB}$ ，即 38.4TB。由于 RAID 5 会使用一个驱动器的容量来存储校验数据，驱动器组的可用容量减少了一个，变为 30.72TB。在 RAID 5 阵列上，5 个虚拟驱动器 被创建。每个虚拟驱动器的大小为 3.84TB，因此，总共有 19.2TB 的虚拟存储空间提供给主机。

为了补偿校验数据占用的空间，参与阵列的 CSD3000 SSD 必须达到最低 1.2:1 的数据压缩率（即写入数据减少 20%）。如果数据可以进一步压缩，还可以创建更多的虚拟驱动器来利用额外的容量。SupremeRAID 提供的灵活性使得存储容量管理更加灵活。通过透明压缩回收的额外存储空间，可以方便地用于创建更多虚拟驱动器或动态扩展存储池。graidcli 工具提供了简洁的汇总信息：

```
$ sudo graidctl list drive_group
```

✓List drive group successfully.

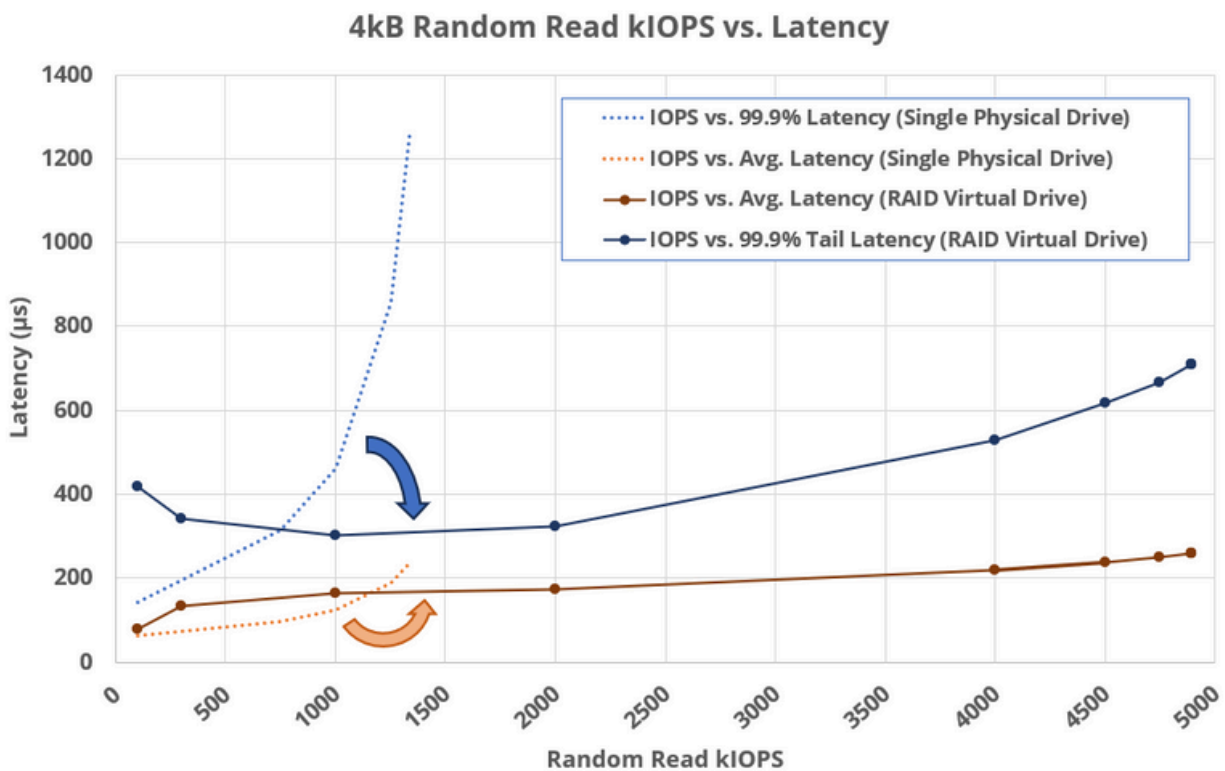
|         |       |        |          |        |        |                      | STATE |
|---------|-------|--------|----------|--------|--------|----------------------|-------|
| DG ID   | MODE  | VD NUM | CAPACITY | FREE   | USED   | CONTROLLER           |       |
| 00T1MAL | RAID5 | 5      | 28 TiB   | 10 TiB | 18 TiB | running: 0 prefer: 0 |       |

需要注意的是，graidcli 报告的容量单位是 TiB（1kB = 1024 字节），而非 TB（1kB = 1000 字节）。

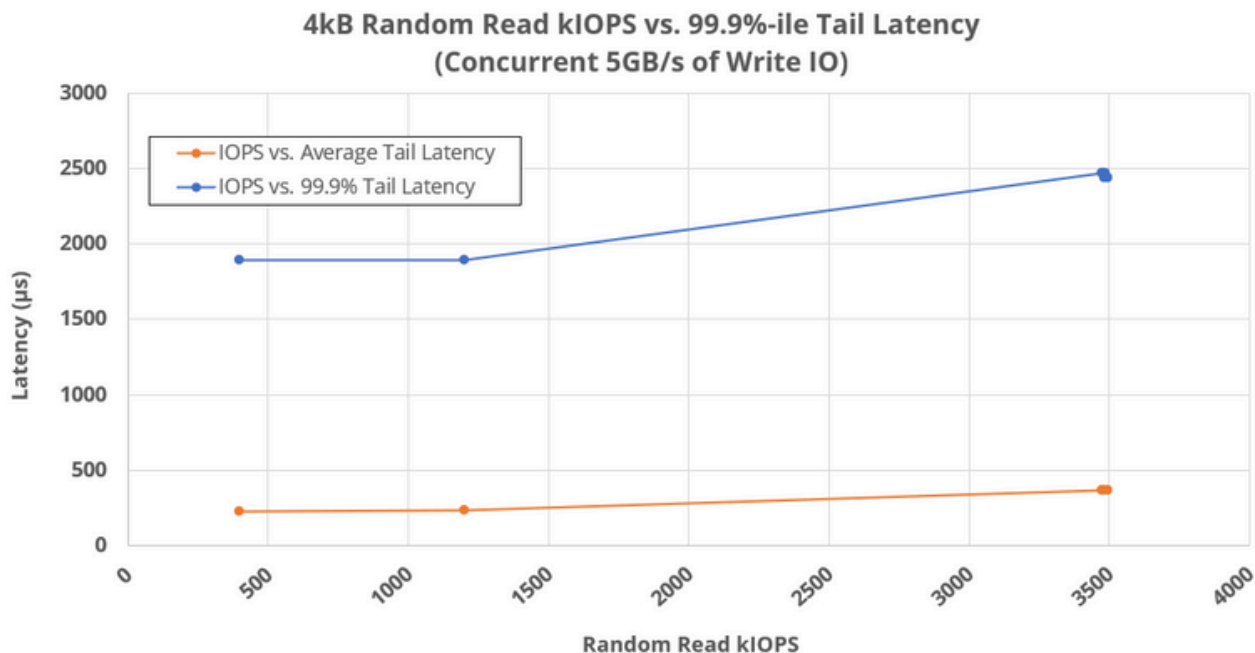


### 3.1 3.84 TB 的虚拟存储设备的性能表现

单个 3.84TB 虚拟驱动器可以利用整个 RAID 池的资源，使其性能远高于单独一个驱动器。下图展示了虚拟驱动器与单个驱动器的 4kB 随机读取性能对比：



单个驱动器的最大 IOPS 为 1.3M，而虚拟驱动器的 IOPS 能够扩展到接近 6M，同时保持低于 1 毫秒的读尾延迟。这显示了 SupremeRAID 阵列能够在 RAID 阵列中的所有驱动器之间水平扩展性能。



当给虚拟驱动器增加 5GB/s 的高写入负载时，尽管负载非常高，随机读取 IOPS 的延迟响应依然保持平稳，直到 3.5M IOPS 饱和为止。

## 4 结论

将单个 3.84TB CSD3000 驱动器与 SupremeRAID 3.84TB 虚拟驱动器进行了对比。结果表明，SupremeRAID 阵列通过池化所有 SSD 的资源，使得每个 SSD 的资源可以被单个逻辑卷充分利用，从而提高性能。

通过池化多个 NVMe SSD 并创建能充分利用底层存储并行性的逻辑卷，SupremeRAID 提供了一种关键的工具来有效管理尾部延迟。这个工具有两个主要用途：

- 创建虚拟卷：这些虚拟卷的性能超越任何物理卷，能够在数百万 IOPS 的性能水平下提供一致且低延迟的响应。
- 避免浪费 SSD 性能：通过多个虚拟卷的配置，能够充分利用驱动器池提供的全部性能，从而避免 SSD 性能的浪费。这种方式特别适合突发性工作负载，因为在这种情况下，单个虚拟驱动器可以在任何容量点上提供数百万 IOPS 和低延迟。

使用 3.84TB CSD3000 提供的扩展容量功能，将每个驱动器的逻辑容量扩展至 7.68TB。这为 SupremeRAID 驱动器池提供了一个更大的容量池，可以从中创建更多的虚拟驱动器。驱动器池的弹性特性使得在需要更多物理容量时，可以添加更多物理驱动器；或者当 CSD3000 的透明压缩技术能够回收更多容量时，可以创建更多的虚拟驱动器。这个特性使得存储池能够灵活地扩展和优化。

# 关于 Graid Technology

我们的使命是为客户提供下一代 NVMe 和 NVMeoF SSD 存储基础设施，同时不牺牲他们所需的性能。SupremeRAID™ 是一款革命性的基于 GPU 的 RAID 技术，能够为高性能工作负载的未来提供市场所需的可靠性、速度、易用性、灵活性和总拥有成本（TCO）。



[www.graidtech.com](http://www.graidtech.com)



Graid Technology Inc.

# 关于 ScaleFlux

ScaleFlux 锐钲是大规模部署计算存储的领导者，旨在帮助其客户利用数据增长作为竞争优势，提供企业级计算存储芯片解决方案，其硬件计算加速引擎极大优化了 NVMe SSD，提升了存储的能力。有效加速应用程序并优化数据中心、企业和边缘网络的基础设施资源。让客户在处理数据库、分析、物联网和 5G 等工作负载时获得更大的竞争优势。



ScaleFlux | 锐钲



[sales@scaleflux.com](mailto:sales@scaleflux.com)



[www.scaleflux.cn](http://www.scaleflux.cn)