SDS、硬件 RAID 与软件 RAID 哪个属于未来?

本白皮书将介绍 NVMe 存储快速发展的功能，及其与 SATA/SAS 技术一起提供的一系列高级配置。

确定改变的节奏：目前及以后的存储见解

确定改变的节奏：目前及以后的存储见解

最开始是软件定义网络 (SDN)，不久之后软件定义计算 (SDC)、软件定义存储(SDS) 甚至整个数据中心本身 (SDDC) 都开始虚拟化，作为超融合基础架构 (HCI) 中的元素。

所有这些连接的驱动者就是 PCIe – Peripheral Component Interconnect Express – 从集成 CPU、GPU、存储和联网发展为支持 I/O 虚拟化的高速总线。PCIe 还支持 Non-Volatile Memory Express (NVMe) 接口，可通过以多种不同尺寸封装的新型 SSD 进行最佳连接。

与 SAS 和 SATA 相比，NVMe SSD 在带宽、延时和功耗方面出现了一个巨大的飞跃。自 2011 年标准化以来，NVMe 存储接口与每一代新的 PCIe 相匹配，吞吐量成倍提升，PCIe 2.0 时是 5GT/s，PCIe 4.0 提升到了 16GT/s。当 PCIe 5.0 推出时，其吞吐量为 32GT/s，但要想成为主流，还有一段路要走。

当前的路线图已经制定，PCIe 6.0 规格提供 64Gt/s 的速度，PCIe 7.0 规格计划在 2025 年发布，目标是 128GT/s，同时保持与前代的兼容性。通过关联，NVMe 存储将吸取所有这些优点。

建立连接

目前，支持 PCIe 3.0 和 PCIe 4.0 的硬件很多，转换到 NVMe SSD 在表面上看起来很简单。例如，可以利用扩展卡 (AIC) 直接插入 PCIe 槽。具有 U.3“三连接器”接口的 NVMe 驱动器可与 SATA 及 SAS 存储一起直接用于服务器阵列。U.2 部署的驱动器槽位于正面，检修方便，没有 AIC 不易接触的缺点，因此是很多数据中心运营的切实选择。

或者，M.2 接口可连接 NVMe 驱动器并发挥其最大 PCIe x4 性能，前提是主板或 AIC 上有适当的 M 型连接器键。B 型键只能提供 SATA 3 或 PCIe x2 速度。您可能会看到有 B+M 键的 SSD，这种几乎可以肯定是 SATA 3 设备，能够兼容两种插槽类型。

因此，虽然这些物理形状允许 NVMe 存储设备安装在系统中，但真正重要的问题是，要完全重新思考如何最佳地管理 NVMe 生态系统中的存储设备。

例如，为获取最佳吞吐量，安装单个 U.2 或 M.2 NVMe SSD 需要每个设备使用四个 PCIe 通道 (PCIe x4)。旧系统存在的问题是，拿 CPU 来说，在安装占用 16 个 PCIe 通道的 GPU 后，24 个 PCIe 通道可能很快就用完了。随着最近的硬件配备多核，最多可支持 128 个 PCIe 通道，切换到此硬件的 PCIe 增加了通道数，这个问题得到了缓解。但即使如此，也必须将 PCIe 通道配置纳入基础架构升级规划，以确保最佳效果。

NVMe 存储需要不同的部署方法。虽然配置驱动器仍然采用熟悉的方式，比如软件 RAID 和硬件 RAID，但它们的利用率已经改进，支持 NVMe 存储通过 SATA 和 SAS SSD 提供的增益。

软件 RAID

NVMe 存储一项纯粹而简单的优点是，所有主要操作系统都有支持它的 NVMe 驱动程序。无论主机是 Windows、Linux、macOS 还是 Solaris，加入 NVMe SSD 后，设备都可以访问。VMware 的虚拟化环境支持 NVMe 驱动程序，适用于软件定义存储应用程序的选项非常多。

NVMe 存储设备的这种立即可用性是对所有主流操作系统中标配的软件 RAID 应用程序的补充。简单、高效的软件 RAID 功能适用于所有形式 – 从终端消费者、游戏发烧友和内容创作者，到成熟的企业部署 – 让人们便利地享有一套基本、强大的存储管理功能。

基本软件 RAID 应用程序只能分别为性能和数据安全提供 RAID 0（带区）和 RAID 1（镜像）。**诚然，硬件 RAID 提供的 RAID 层级多于软件 RAID。**但即使如此，Mdraid（Linux 上的默认软件 RAID 应用程序）等应用程序也提供 RAID 4、5、6 和 10 – 可平衡性能与数据安全的组合。

当 SSD 还不能匹配个别硬盘的容量时，指定数量的硬盘的总存储容量要求也是配置 RAID 阵列时的主要考虑因素。而且，使用软件管理 RAID 存储环境中的数据分发和同位检查功能会影响执行这些例程的主机 CPU。算法操作的复杂性各异 – 例如，写入的计算密集性超过读取 – 如果数据吞吐量很大，并且 RAID 配置的冗余级别高，则这些任务可能会影响整体性能。

如果软件许可证按核心收费，让系统承担存储任务是否有意义？对硬件 RAID 的争论由来已久，但我们不再是 SATA/SAS 环境。在某种程度上，NVMe 中的延时和吞吐量增益及其对 PCIe 总线的直接访问可以弥补软件 RAID 中固有的性能不足。

通过设计改进

SATA 接口专为硬盘而设计，它与 SSD 的结合使用一直是一种妥协方案。SATA SSD 通过硬盘提升速度极为有效，但这只是快闪存储能够提供的一小部分功能。SATA 使用的高级主机控制器接口 (AHCI) 拥有所有传统的特点 – 围绕旋转型磁盘的物理限制创建了 120 多个命令 – 可通过系统升级兼容闪存，但终究是一个瓶颈。而 NVMe 可运行至少 13 个命令 – 10 个 admin，三个 I/O：read、write、flush。

关于命令队列，AHCI/SATA 技术只有一个，每个队列可以发送 32 个命令。而 NVMe 有 64,000 个 I/O 队列，每个队列最多 64,000 个命令，因此对 CPU 周期的使用特别低。

NVMe 存储使用的简化 PCIe 数据路径，结合其巨大的吞吐量和效率，可以在此领域中以不同的视角查看软件 RAID。软件 RAID 在此领域证明了效率，人们并没有在意它的局限性。的确，对很多人来说，作为传统意义上的唯一选择，硬件 RAID 必须发展，提供超出 NVMe 存储范围的功能。

硬件 RAID

硬件 RAID PCIe 卡具有专门的控制器芯片，可执行所有必要的计算功能，从目标存储硬件创建和管理 RAID 阵列。处理工作量全部卸给 RAID 卡，因此硬件 RAID 可以提供复杂性不同的广泛 RAID 级别，主机平台没有处理负担。

因为处理 RAID 算法未涉及昂贵的主机 CPU 资源，所以读写速度得到了优化，而且支持硬盘热插拔。使用软件 RAID 时，缺乏专门的处理资源会导致在高容量 SAS/SATA 环境中增加延时和吞吐量。与硬件 RAID 不同，硬盘更换通常需要在取下磁盘之前执行 RAID 管理程序，也通常需要重新引导。

PCIe 硬件 RAID 卡的成本、低延时、数据保护和缓存特性及其磁盘阵列扩展功能，为其在企业存储管理中赢得了一个位置。而且它也在不断发展。专用的仅 NVMe RAID 卡在市场上仍然是相对较新的产品，Broadcom、Marvell 和 Microchip 等供应商提供的三模 PCIe 第 4 代芯片 RAID (ROC) 卡支持 SATA、SAS 及 NVMe 的组合。

这些硬件 RAID 卡为 NVMe SSD 在混合存储环境中的共存提供简便的方式。按照基本的布线程序，U.2 背板可配置为使用 U.2 外形规格的 SATA/SAS 与 NVMe SSD 的组合。

U.3 标准的出现进一步推动了这一外形规格，通过适合真正三模背板的统一布线降低复杂性。但有一点要注意：U.3 物理硬盘接口相同，但插脚配置已更改。因此，U.3 硬盘可用于 U.2 背板，但 U.2 硬盘与 U.3 背板不兼容。

虽然 U.3 的混合和匹配功能可能是一个值得追求的目标，但如何传播这种配置很可能变成另一个问题。

驱动器槽观察

当然，通用背板管理 (UBM) 标准的出现进一步促进了混合存储发展，并且同时兼容 U.2 和 U.3 设计。UBM 得到 20 多家领先的存储硬件供应商的联合拥护，可让主机和控制器设备发现背板功能，并支持检测和监控不同类型的设备（SATA、SAS 和 NVMe），即使是在单一驱动器槽中。UBM 也支持 SATA/SAS 扩展器和 PCIe 交换机，提供一系列切实的背板管理功能，进一步增强 U.2 和 U.3 系统架构。

三模 RAID 或 HBA（硬件总线适配器）卡将使用 x8 或 x16 PCIe 主机通道，并具有 PCIe 交换功能，可成倍扩大通道数，有效地增加带宽。比如，卡规格可能表示支持最多 32 台 NVMe 设备，但不等于支持 8 台完全 4 倍速度的 NVMe SSD，后者需要 32 个 PCIe 通道。在理论上，32 台 1 倍速度的物理 NVMe 驱动器可能适合，甚至在 PCIe 3.0 环境中，每台设备的运行速度为 1000MB/s – 比 SATA 的 600Mb/s 吞吐量快三分之二。即便如此，该配置也是 NVMe SSD 存储的次优用法，因为 NVMe SSD 存储的卓越性能通过 PCIe 通道并行化得到了巨大的提升。在混合使用环境中，三模控制器只能将 8 或 16 个通道用于 NVMe 存储，这又一次涉及是选择更少的驱动器还是选择更少吞吐量的问题。

无论是什么类型的驱动器，无缝集成在一个背板中都将启发构建能够在一个机箱内管理热 (NVMe)、温 (SAS/SATA) 和冷 (SATA/HDD) 存储要求的系统，拭目以待。

毕竟，在启用 NVMe 时分割 PCIe 通道分配以保持兼容旧存储设备是一种妥协，有其局限性和成本。为确保可靠性和更大的容量，很多操作只能关注驱动器刷新率，目前现有的 SAS/SATA 存储部署令人满意。U.2 存储还要坚持一段时间，但为了利用现有的 SAS/SATA 存储资产以及成本更低的专用控制器和扩展器，使用一种设备通道的配置可能司空见惯。同样，为最大提高性能和容量，专用 NVMe SSD 是最好的。

NVMe 存储推广的速度主要取决于工作负载的密集性及其对现有系统的增强作用。投资于实质性仅 NVMe 部署的云服务提供商已经开始收获效益，因为带宽的巨大增益意味着可以提供新的分层服务，以适应广泛的客户需求。

修正预期

在 NVMe 强化的云提供商与较传统的数据中心这两个极端之间，是要求增加功能、提高效率、增强可扩展性的企业。

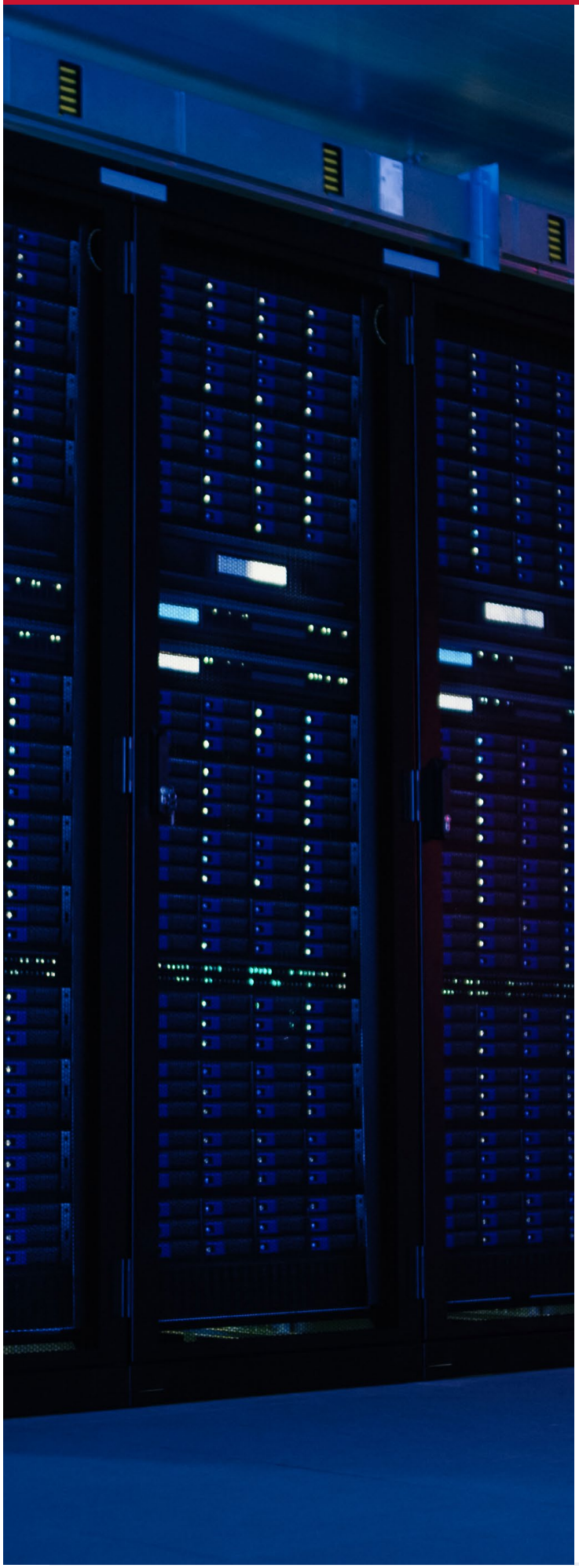
在 NVMe 强化的云提供商与较传统的数据中心这两个极端之间，是要求增加功能、提高效率、增强可扩展性的企业。

它采用 NVMe，但方针更有针对性 – 根据对成本、优点、集成和优化的研究逐渐采用。

系统的不足之处，如效率低下的应用程序会限制预期的延时和吞吐量增益，在 NVMe 存储用于缓存时会迅速显露出来。其他瓶颈会自行暴露，要想实现 PCIe/NVMe 生态系统的性能优点，必须解决这些瓶颈。

这不是对等的交换，而更像是自行车换动车。在这方面，SSD 规格也需要重新评估，因为服务水平协议可能坚持要求使用运营指标，而这些指标无法识别您可以使用 NVMe 存储执行的更多操作。

一个典型的示例是每天磁盘写入量 (DWPD) 数字，用于确定闪存在其保修期内的耐用性。闪存盘会遇到写入放大的问题，由于在内存单元中存储数据所采用的方法，此问题会增大 SSD 的磨损。本质上，存储单元不会直接存储数据，而必须先擦除后才可被重写，并且随着时间的过去，这一复杂的程序会造成存储性能的下降。使用超量配置（一种 SSD 容量的储备箱）来解决这些问题，并执行驱动器管理例程，例如垃圾收集。这是重新分配数据以释放存储块（然后为准备写入而擦除）的过程，是写入放大的主要原因。



进行分区

最近的 NVMe 2.0 规格增加了分区命名空间 (ZNS)，这为 NVMe SSD 读/写程序提供了新的方法。分区块管理接口位于主机与 NVMe SSD 之间。分区与磁盘分区有些类似，但它发生在主机应用程序层面。ZNS 可让 SSD 与主机通信，在 ZNS 写入时描述或揭示性能，例如提供最佳模式的详细信息和数据布局，并且擦除操作依序执行。

这种协作性互动将某些存储管理职能卸给主机应用程序，优点是可以减少对超量配置的需求，可能增加多达 20% 的存储容量。实施 ZNS 可改进 I/O 延时，减少 4x 至 5x 的驱动器放大率。此外，还可为不同的区域分配特定的工作负载或数据类型，以启用更可预测的性能模式。

分区名称空间的发展刚起步，但 ZNS 已经是 Linux kernel 5.9 的一项功能。此外，Microsoft、Alibaba 和 NetApp 支持的 ZNS 研究 – 着眼于大型超扩展运营 – 表明，在行业层面采用 ZNS 只是时间问题。

应用程序必须更新才可完全利用此功能集，因为它在不断发展，而且越来越多的 NVMe 驱动程序现在具有 ZNS，现有 NVMe SSD 实施 ZNS 在某些情况下只需要更新固件。

对于想着正确规格的系统架构师，现在需要重写解释 DWPD 真正含义的规则手册。在实施 ZNS 后，写入放大率大幅降低，等于驱动器耐用度大幅提升。您还需要多少驱动器？随着超量配置的大幅减少，驱动器容量大幅增加。展望数据管理的未来，使用 NVMe SSD 和 ZNS 接口真的可以实现事半功倍。

软件定义存储

NVMe 真正采用混合路径，从 M.2 驱动器和 PCIe 附加卡到 U.2 或 U.3 存储。新兴的企业和数据中心 SSD 外形 (EDSFF) 是另一种存储形式，专为在长短（L 和 S）配置中使用双宽度（E.1 和 E.3）驱动器的 NVMe 生态系统而设计。E.1L 驱动器以 1U 机箱支持高存储密度，采用更加灵活的 E.1S 大小，具有适合可扩展性的热效率优势。E.3 驱动器封装为替换 U.2 2.5 英寸 SSD 驱动器，适合较传统的 2U 服务器和驱动器阵列机箱，设计为每个驱动器容纳更多闪存芯片，以增加存储密度。

当然，NVMe 作为一个共同标准，其驱动程序在所有主流操作系统上受支持，因此所述任何选项的实施都比较简便。如何选择取决于存储特性以及最符合工作负载和冗余要求的配置。这可能涉及集成边缘服务器上的 NVMe 存储与 SAS/SATA 硬件，进一步减少密集型运算。硬盘驱动器甚至带式备份也可能成为存储基础架构的一部分。企业存储管理中不缺少专有平台，协调这些不同的存储系统可能使复杂性迅速上升。这是软件定义存储 (SDS) 发挥作用的地方，提供统一混合存储设备操作和优化其利用率的方式。

在软件定义存储领域，可用的存储资源从存储硬件提取并虚拟化。使用行业标准协议，甚至可以通过 SDS 虚拟化访问专有硬件，单片存储设备可不受约束地加入也可能使用商品服务器构建了新型低成本可扩展存储器的大型存储池。这种不受约束的特性可避免在更换、升级或扩展存储硬件时发生中断。

在所有可用的存储器并入虚拟池后，需要做出配置决定，能帮助这些分配的功能非常多，其中包括自动化。在 SDS

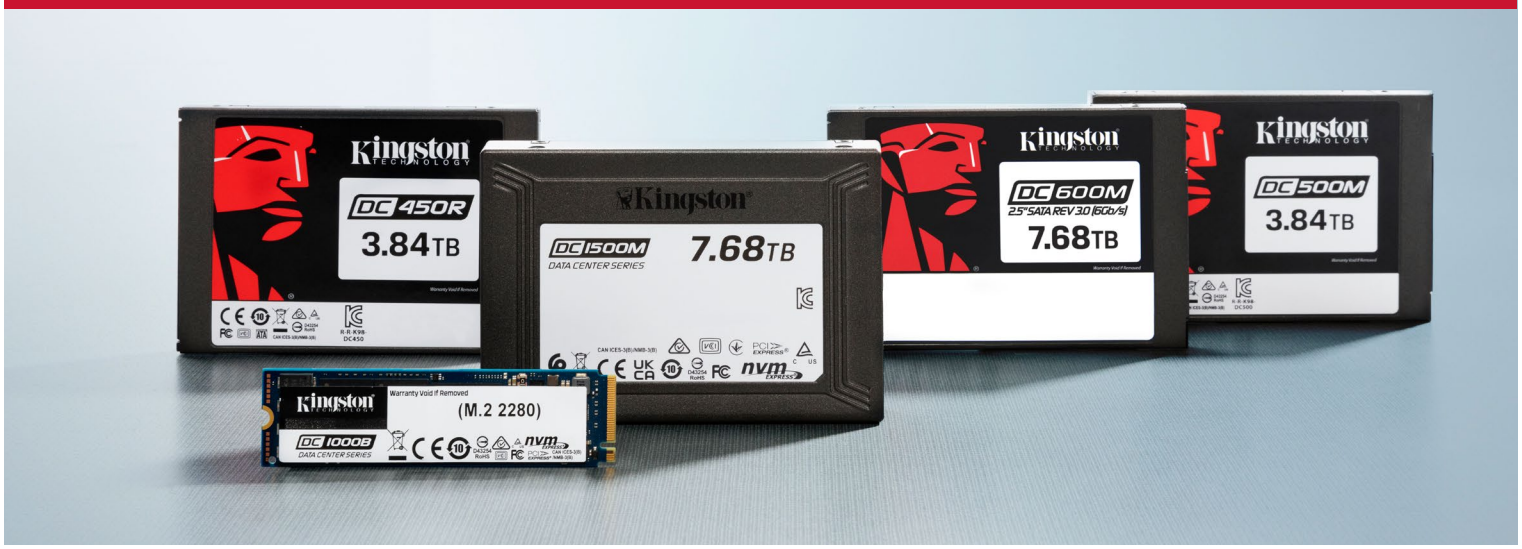
仪表板中，根据不同池中存储器的硬件配置文件来识别热、温和冷存储。通过使用脚本，可以执行任务来分配和分发最符合这些存储库的数据加载。

通过其虚拟存储层，SDS 提供灵活性和扩展性；管理适合企业需求及客户可变需求的存储环境的创建和部署，从提取和配置虚拟机 (VM) 到镜像和复制。

关于 NVMe SSD，SDS 平台可使用一项称为 NVMe 传递的功能通过 PCIe 总线直接访问存储。例如，VMware 有其自己用于 ESXi/vSAN SDS 平台的 NVMe 存储驱动程序，可使用一项称为 VMDirectPath I/O 的功能直接向虚拟机分配 NVMe 存储。根据主机 CPU 配置，每个虚拟机最多支持 16 台传递设备。

整体来说，启用 NVMe 传递可最大程度减小主机的干扰，改进性能并简化为虚拟机实例及其他服务配置 NVMe SSD。为此，第三方软件或硬件 RAID 控制器是否支持 NVMe RAID 功能对 SDS 不算什么问题，因为它可以直接配置 NVMe 软件 RAID。

虽然 SDS 有潜力成为数据管理的“万能灵药”，但其成本和初始配置复杂性可能让一些要求比较简单的企业望而却步。但就像存储本身一样，这些成本可缩放，可使用不同的版本适用较小的硬件部署。



掌控改变节奏

存储在不断发展，但变化很少会在一朝一夕之间完成，因为现有资产很可能纳入淘汰战略计划中。因此，存储会随着硬盘和 SATA SSD 等技术持续发展。这些技术各有其市场，将继续在存储阵列中提供有用的服务。例如，Kingston 的 DC600M 2.5 英寸混合用途企业 SATA SSD，容量翻了一番，达到 7.68TB。

全世界的数据中心仍然以硬件 RAID 和主机总线适配器为主，供应商持续创新，以满足不断扩展的 IT 行业的要求。

通过与 Broadcom 及 Microchip 合作，Kingston SSD 进行严格试验，确保符合当今数据驱动型技术的严格要求。

测试计划采用这些领先供应商的存储适配器，涉及严苛的工作负载和挑战性的配置，旨在确保 Kingston 企业 SSD 可以提供合格的性能、耐用性及可靠性。毋庸置疑，Kingston 的 U.2 DC1500M PCIe NVMe Gen3x4 企业 SSD 经历了这一切。它的容量高达 7.68TB，再结合 1 DWPD，完全有资格进入最新一代服务器和存储阵列。

即使软件定义存储设备可以围绕商品硬件而构建，SSD 的选择也至为关键。消费级 SSDs 在成本上看似吸引人，但与专为耐用性和持续高带宽负载而构建的企业 SSD 相比，可能是一种假实惠。从超融合基础架构提供服务需要性能可预测性，从而有效地管理工作负载并满足客户期望。Kingston 企业 SSD 能够与 VMWare 存储应用程序结合使用，确保即使在虚拟的软件定义存储世界也能实现真实世界的目标。

存储配置在不断变化，但为适应不同商业模式，变化节奏不一样。在每个层面，从旧接口连续性到 NVMe 创新都有改善。如果感觉升级很困难，Kingston 的“咨询专家”服务可提供帮助。它免费帮助您做出适合您的企业和预算的重要决定。因此，只要您在此过程中遇到任何问题，Kingston 都会在您左右。

#KingstonIsWithYou