解决方案简介

超融合
英特尔® 傲腾™ 持久内存

基于 SMTX OS* 和英特尔® 傲腾™ 持久内存，打造高性能、低延迟超融合解决方案

英特尔® 傲腾™ 持久内存以其低延迟、持久化存储等特性为超融合基础架构 (HCI) 的创新带来更多可能，并助力 SmartX* 在高性能、低延迟超融合解决方案中加速冲刺。

作为超融合领域的专业领导厂商，SmartX* 深知硬件技术对超融合产品的重要性，并始终和英特尔在研发领域保持紧密合作。本次 SmartX 基于创新的英特尔® 傲腾™ 持久内存技术对 SMTX OS* 进行深度优化和架构升级，不仅为超融合产品性能带来质的飞跃，更是对解决性能敏感型核心业务向云化基础架构转型这一难题进行了深度的探索，有效加速了超融合基础架构在金融核心等领域的落地。

—— 徐文豪
SmartX CEO

执行概要

客户业务的迅猛发展和对计算资源、数据存储资源需求的激增使得传统计算与存储分离的 IT 基础架构面临重重挑战。超融合基础架构 (HCI) 的兴起与发展有助于企业更高效地池化及管理运算和存储资源，并从中获得更高的效率与投资回报。但在过去的几年里，将超融合基础架构 (HCI) 应用于企业核心业务和性能敏感型业务一直是有争议的话题。

作为国内领先的超融合产品提供商，SmartX 凭借其完全自主研发的超融合系统，致力打造高可靠、高性能、易扩展、生产就绪的超融合产品，帮助用户构建新一代的数据中心基础架构。其产品核心 SMTX OS* 超融合软件可以运行在几乎所有主流的 x86 服务器上，且已被广泛应用于金融、医疗和制造业等领域。

为了给用户提供更优质的 HCI 产品实现端到端的数据中心云化转型，SmartX 采用英特尔® 傲腾™ 持久内存搭配 NVMe NAND SSD 开发出了新一代全闪超融合解决方案。整体性能获得显著提升，能够更好地满足客户核心业务对 IO 吞吐、延迟等性能的严苛要求。

业务挑战

随着现代数据中心对系统架构效率和可扩展性诉求的不断提高，超融合基础架构 (HCI) 凭借着架构精简、易于扩展、稳定性高等诸多优势赢得了企业的广泛青睐。经过几年的发展，HCI 的需求在市场上高速增长，已在各行各业中迅速普及，应用场景由生产外围切入到如金融、医疗等行业的核心业务负载。据 Gartner 预测，到 2023 年，全球 HCI 设备预计将以 23% 的复合年增长率 (CAGR) 增长，市值预计将从 31 亿美元增长到 85 亿美元以上[1]。然而，企业为了不断提供更好的终端用户体验，对各级业务系统特别是核心业务的 IT 基础架构的要求也越来越高，性能问题一直是架构上的一大顾虑；超融合基础架构是否能顺利应用于核心环境、实现业务价值，也是企业用户最关心的课题之一。

SmartX 多年来始终致力于开发性能领先的超融合产品，并取得不错的成果。然而，在面对性能敏感型业务应用更为严苛的 IO 及延迟要求场景时，以往在超融合领域广泛使用的 NAND SSD 存储设备则成为瓶颈，逐渐成为超融合系统未来发展的阻碍。几乎所有高性能存储系统皆依赖缓存技术来加速存储性能，存储系统的能力直接反应在前端应用的响应速度上，超融合系统亦不例外。使用更快的存储部件作为缓存介质是各厂商一致的目标，

而曾经 NAND SSD 几乎是唯一选择。作为超融合领域的专业领导厂商，SmartX 不断尝试将超融合应用于核心业务场景，面对 NAND SSD 成为性能瓶颈的今天，如何突破当前超融合系统在 IO 及延迟方面的局限，是 SmartX 进一步提升超融合产品性能的关键。

基于 SMTX OS 和英特尔® 傲腾™ 持久内存的高性能、低延迟超融合解决方案

英特尔® 傲腾™ 持久内存简介

英特尔® 傲腾™ 持久内存是基于英特尔® 3D Xpoint™ 技术打造的创新存储产品，兼具易失性和非易失性，通过创建全新的存储层来填补内存和存储之间的性能差距（如图 1 所示）。它拥有两种模式：内存模式——这种模式下，可为系统提供大容量内存；App Direct 模式——该模式下，软件和应用可以直接访问英特尔® 傲腾™ 持久内存中的数据，有助于降低堆栈复杂性。英特尔® 傲腾™ 持久内存具备以下几大优势：

- **高性能低延迟：**英特尔® 傲腾™ 持久内存兼容 DDR4 内存封装，具备接近 DRAM 的传输速度；
- **非易失更安全：**在 App Direct 模式下，英特尔® 傲腾™ 持久内存可让系统在遭遇突然断电等意外情况后不仅不造成数据丢失还能更快恢复。同时，在这种模式下，可按字节 (Byte) 进行寻址，能有效解决写数据放大问题；
- **大容量：**英特尔® 傲腾™ 持久内存提供 128 GB、256 GB 和 512 GB 三种容量规格；
- **低成本：**在内存模式下，与 DRAM 相比，英特尔® 傲腾™ 持久内存的单位容量成本更低。

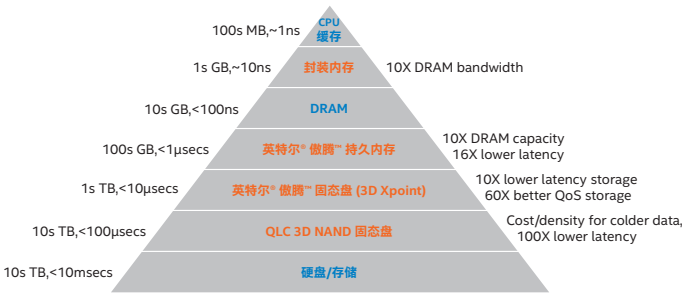

图 1. 英特尔® 傲腾™ 持久内存在速度金字塔中的位置

搭配英特尔® 傲腾™ 持久内存的超融合系统概述

英特尔® 傲腾™ 持久内存的**低延迟**、**非易失性**和**可按字节寻址**优势正是 SmartX 需要的高性能存储特性。因此，他们基于这款创新产品和 SMTX OS 打造了全新的全闪超融合解决方案（如图 2 所示）。英特尔® 傲腾™ 持久内存拥有更靠近 DRAM 的读写速度[2]，使得它与内存交换数据的过程中延迟更低，可大大缩短响应虚拟机请求的时间。

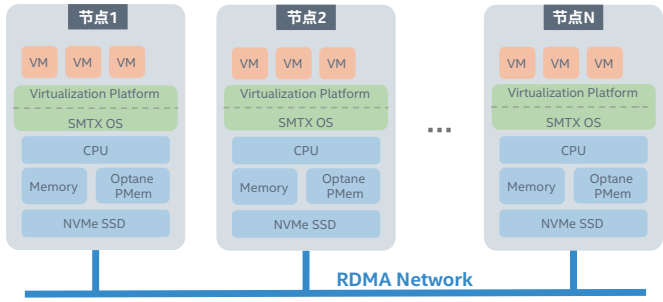

图 2. 采用英特尔® 傲腾™ 持久内存后的全闪超融合解决方案

结合 SmartX 针对英特尔® 傲腾™ 持久内存的一系列优化，新方案性能表现出色。据 SmartX 介绍，采用英特尔® 傲腾™ 持久内存的超融合一体机，三节点最小系统的 IOPS 仍可达约 120 万[3]，与以往仅使用 NAND SSD 相比有大幅提升。

高性能存储用于 Journal & Cache

在本方案中，英特尔® 傲腾™ 持久内存采用 App Direct 模式，将其作为 Journal 和 Cache 盘，配合容量盘 NVMe SSD 一起使用（如图 3 所示）。

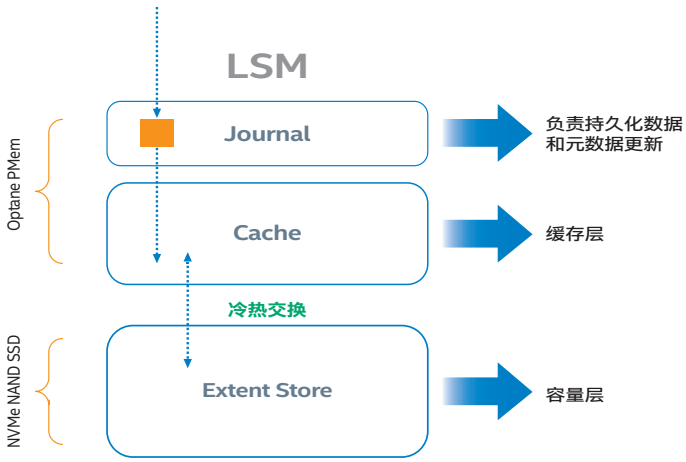

图 3. 采用英特尔® 傲腾™ 持久内存后的存储模块

在该方案中，英特尔® 傲腾™ 持久内存始终存放热数据（日志、数据和缓存），拥有比 NVMe SSD 更好的 IO 响应能力。在 APP Direct 模式下，英特尔® 傲腾™ 持久内存作为一种新型高速存储介质，其相对更大的单位容量（与 DRAM 相比）可保证更多经常被访问的热数据驻留在持久内存以获得出色的响应速度——这为数据加速提供了强有力的支撑。另外，英特尔® 傲腾™ 持久内存与 NVMe SSD 互相搭配，比以前的全闪方案更快，延迟得以进一步降低。在本方案中，系统配置少量持久内存便可持续为业务加速，对全闪架构性能的整体提升有很大的帮助作用。

按字节可寻址访问，Byte 对齐、避免写放大问题

在英特尔® 傲腾™ 持久内存的 App Direct 模式下，程序可通过以下两种模式访问持久内存设备：

- **Block Access 模式：** 应用通过标准的文件系统层（如：xfs, ext4 等）访问持久内存，如同访问普通存储设备一样；
- **Direct Access (DAX) 模式：** 应用访问持久内存如同访问内存一般，可直接访问并修改持久内存数据，并且支持以 Byte 为单位进行数据存取。

由于 DAX 模式在路径上绕过文件系统层与驱动层，可获得更低访问延迟，SmartX 在其底层的 SMTX ZBS 分布式块存储软件栈中专门开发了以 DAX 模式访问持久内存设备的适配程序，进一步缩短 IO 堆栈、降低延迟、提升性能。

基于 SSD 只能以块 (Block) 为单位进行访问的特性，SmartX 以往在传统 Journal 设计中采用 4 KiB 对齐方式，在某些应用场景下会造成严重的写放大问题。例如系统要求写入 128 bytes 大小的 IO，但由于 Journal 需要 4 KiB 对齐写入，因此必须将 128 bytes 数据进行扩充，用“0”填充至 4 KiB 大小，再写入介质中。因此，实际占用的容量是要求写入大小的 32 倍[4]。如此显著的写放大现象会造成 Journal 和 Cache 存储物理空间的大量浪费，增加了冷热数据交换的频率，从而增加数据交换延迟，对性能造成较大影响。

SmartX 通过在 SMTX ZBS 中以 DAX 方式访问持久内存设备，可像访问内存一样按字节 (Byte) 为单位对持久内存进行存取。SmartX 用这个特性重新设计 Journal 的写入机制，并以 64 bytes 对齐。此时往 Journal 写入 128 bytes 数据，则可拆分为 2 个 64 bytes 写入，避免了写放大的问题，有效提升性能。

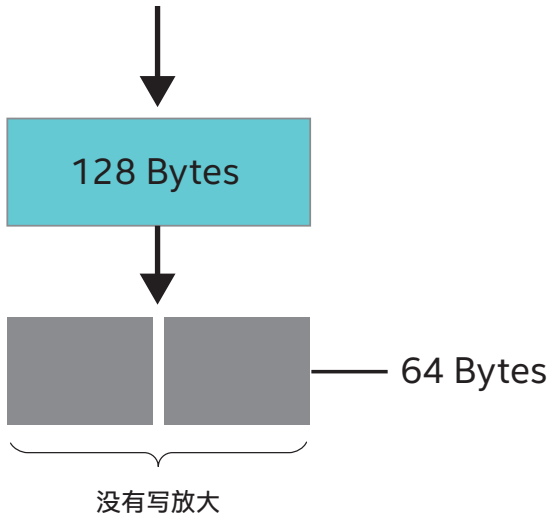

图 4. 采用英特尔® 傲腾™ 持久内存之后能够以 Byte 为单位访问

异步访问英特尔® 傲腾™ 持久内存，减少 CPU 资源使用，增加访问带宽

数据从 DRAM 内存写入英特尔® 傲腾™ 持久内存的过程需执行内存复制操作。这种串行操作在任务完成之前不会释放 CPU 资源，对 CPU 性能消耗比较大。因此，若默认将数据同步写入持久内存，CPU 资源将受到内存复制操作的影响而被大量占用，无法快速释放用以处理其他任务，通过测试，如果以同步的方式处理内存复制过程，单个 CPU 核心最大写入带宽只有 4 GB/s[5]，HCI 系统整体性能则相对较低。

在 HCI 系统中，宝贵的 CPU 资源必须更多的留给 VM 使用，因此需要严格控制 HCI 系统本身的 CPU 消耗。在有限的 CPU 资源下，为了充分发挥英特尔® 傲腾™ 持久内存的带宽性能，双方技术团队创新地引入了 IO/AT DMA 引擎实现了异步写入持久内存的机制，使得单个 CPU 核心的最大写入带宽提升至 10 GB/s[6]。这意味着在几乎不增加 CPU 开销的情况下，通过数据异步写入持久内存的优化方式，可获得 2.5 倍的性能提升[7]。英特尔® 傲腾™ 持久内存为 SmartX 团队带来的创新潜力可见一斑。

此外，SmartX 还针对英特尔® 傲腾™ 持久内存优化了 IO 栈以进一步降低 IO 损耗、并优化网络传输协议，采用 RDMA 存储网络来降低远程节点写入延迟，保证远程节点副本数据及时同步更新。

性能验证与成果

FIO 3.19 测试：带宽、读写速度和延迟性能优化明显

为了验证基于英特尔® 傲腾™ 持久内存和 SMTX OS 打造的全闪超融合解决方案的性能。英特尔和 SmartX 使用英特尔® 至强® 可扩展处理器 6240Y 分别在搭载和未搭载英特尔® 傲腾™ 持久内存的情况下，测试了该解决方案在 FIO 3.19 中的表现。

从图 5 可以看到，无论顺序还是随机读/写 4K 请求，搭载英特尔® 傲腾™ 持久内存的方案在 1/2/4/8/16/32/64/128 个工作负载下，其带宽 (MB/s) 均具有明显优势。其中，顺序读最高加速达到 2.86 倍，顺序写加速最高达到 2.90 倍；随机读加速最高达到 4.14 倍，随机写加速最高达到 2.79 倍[8]。

而在 P99[9] 延迟（线程数 = 16）测试中（如图 6 所示），搭载英特尔® 傲腾™ 持久内存的方案在 4K/8K/16K/32K/64K 大小下，顺序读、顺序写、随机读和随机写延迟均有明显降低，最高降幅分别为 78.46%、59.38%、76.75% 和 56.62%[10]。

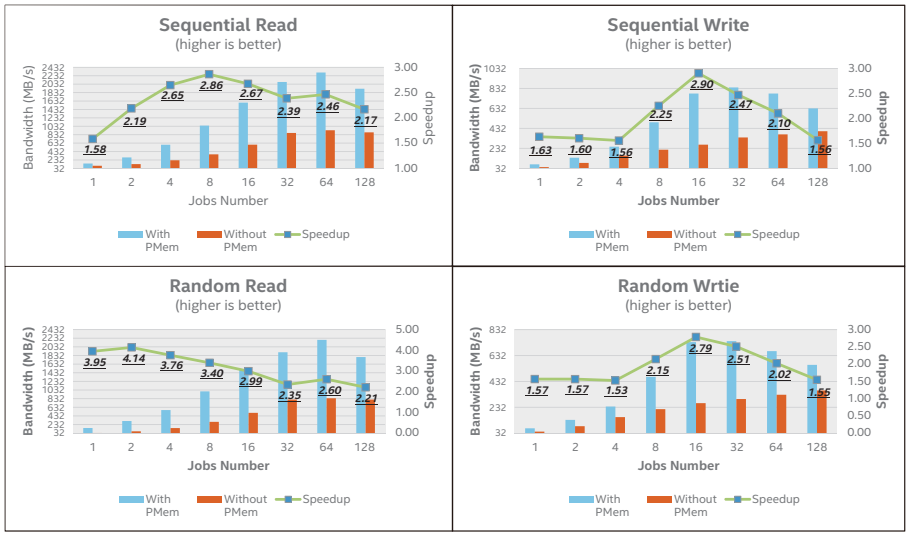

图 5. 在不同访问模式下 4K 请求读写对比测试结果

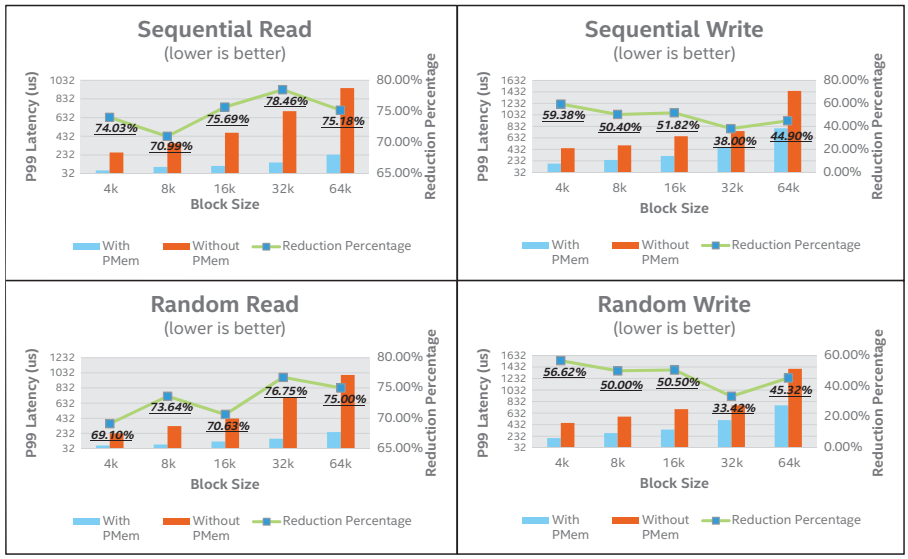

图 6. 不同访问模式下不同请求大小的 P99 延迟（线程数 = 16）对比测试结果

MySQL 8.0.20 测试：实现吞吐率与延迟的双重优化

为了进一步测试新方案在实际应用中的表现，我们测试了不同访问模式下 MySQL 8.0.20 在不同虚拟机中的吞吐率和延迟情况。如图 7 所示，单节点只运行 1 台虚拟机的时候，搭载英特尔® 傲腾™ 持久内存的方案在混合读写、只读、只写方面的吞吐率，分别比未搭载持久内存的方案高 3.20 倍、2.90 倍和 2.85 倍；在混合读写、只读、只写方面，使用英特尔® 傲腾™ 持久内存方案的 P95[11] 延迟均有明显降低，降幅分别为 72.16%、60.10% 和 73.13%[12]。

而在单节点运行 4 台虚拟机时，如图 8 所示，搭载英特尔® 傲腾™ 持久内存方案在混合读写、只读、只写方面的吞吐率，分别比未搭载持久内存的方案提升了 1.39 倍、2.49 倍、2.98 倍；使用持久内存方案的 P95 延迟除了在混和读写场景外，在只读和只写的场景下皆有明显降低，降幅分别为 46.48% 和 50.28%[13]。

可以看到，与实际应用场景接近的 MySQL 数据库实测中，采用英特尔® 傲腾™ 持久内存后的全闪超融合解决方案在性能上得到很大程度的提升。新方案在优化超融合系统自身基础架构的同时，大幅改善了其上运行的业务系统的延迟，实现了端到端的 IO 加速。虚拟机端 IO 延迟的大幅度降低使得 SMTX OS 产品完全具备承载对延迟要求非常苛刻的核心业务的能力。

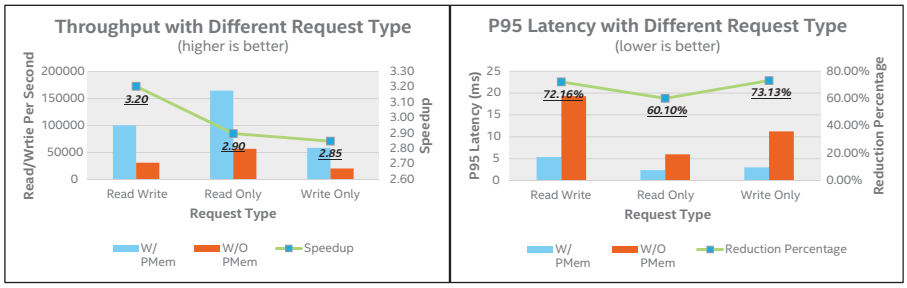

图 7. 不同访问模式下 MySQL Performance - 1VM on a Node

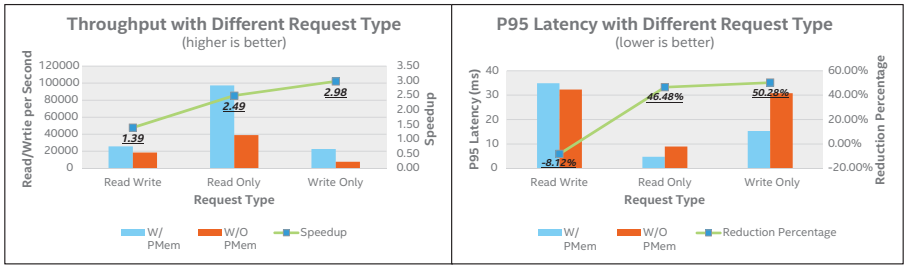

图 8. 不同访问模式下 MySQL Performance - 4VM on a Node

结论与未来展望

在本解决方案中，SmartX 针对英特尔® 傲腾™ 持久内存产品和 App Direct 模式进行了深度开发和重构设计，使得新一代全闪超融合产品得到如下优化：

- 使用英特尔® 傲腾™ 持久内存作为更高速的缓存介质来乘载 Journal 和 Cache，并优化 SMTX ZBS LSM 分层体系，让超融合分布式存储获得更好的性能提升；
- 扩大高速缓存容量，使得更多“热数据”可以常驻内存附近；
- 优化 Journal 存取数据的形式、并采用 Byte 级别的数据写入对齐，解决写放大问题；
- 减少系统 IO 堆栈、优化网络传输性能，降低超融合系统延迟。

SmartX 致力于打造出色的超融合基础架构，并与英特尔携手提供了英特尔® 傲腾™ 持久内存在超融合部署上的最佳实践。测试数据表明，通过引入持久内存技术，能够显著提高超融合系统的性能，帮助消除企业对超融合产品性能的疑虑、解决关键应用场景实际部署的难题，实现利用新一代全闪超融合系统替换更多裸金属服务器和全闪阵列的可能。凭借着性能、架构等诸多优势，更进一步为企业打造新一代云化数据中心。英特尔将继续携手 SmartX 探索更多前沿技术在超融合系统中的应用，致力引领下一阶段超融合创新。

关于 SmartX

SmartX（北京志凌海纳科技有限公司）是中国领先的超融合产品与企业云解决方案提供商，通过构建强大而易用的新一代 IT 基础设施为企业数字化转型提供坚实基础。目前，SmartX 产品主要服务于金融、医疗、大型制造业、商业连锁、地产等领域客户，其中头部行业客户包括交通银行、泰康保险集团、国泰君安证券、海尔、京东方、恒大地产等国内客户，以及韩国 SBS 电视台、Cafe24 等海外客户。与此同时，通过打造新一代面向企业云的存储和计算引擎，SmartX 将赋能企业构建完整的虚拟化和云原生基础设施，为企业 IT 转型奠定基石。

[1] 数据来自 Gartner 调查报告“Competitive Landscape: Hyperconverged Infrastructure, China”，2019 年 9 月 20 日。

[2] 数据来自英特尔官网，DRAM < 0.1 Microsecend Latency，Intel Optane DC persistent Memory < 1 Microsecend Latency。https://www.intel.cn/content/www/cn/zh/architecture-and-technology/optane-dc-persistent-memory-case-gallery.html

[3] 数据来自 SmartX 白皮书《基于 SMTX OS 与英特尔傲腾持久内存的高性能、低延迟超融合解决方案》。

[4] 4KiB = 4 * 1024 Byte，倍数计算方法为：4 * 1024 / 128 = 32。

[5,6,7] 数据来自 SmartX 白皮书《基于 SMTX OS 与英特尔傲腾持久内存的高性能、低延迟超融合解决方案》。

[8,10,12,13] 由英特尔和 SmartX 于 2020 年 8 月 11 日测试。**基准配置**（无英特尔® 傲腾™ 持久内存）：节点数：3，路数：2，处理器：英特尔® 至强® 金牌 6240Y 处理器（18 核，36 线程），启用超线程，启用睿频加速，DDR 内存配置：384 GB（12 插槽/32GB/2666），存储：1 x 英特尔® S3700 200GB（操作系统盘）+ 4 x 英特尔® P4610 1.6TB，网卡：1 x 英特尔® XC710 10Gb + 1 x Mellanox connectx-5 100Gb，操作系统：Smartx customized based on CentOS 7.4，内核：3.10.0-1127.el7.x86_64，工作负载和版本：SMTX OS v4.5，MySQL 8.0.20, Sysbench 1.0.19 和 FIO 3.19

新配置（有英特尔® 傲腾™ 持久内存）：节点数：3，路数：2，处理器：英特尔® 至强® 金牌 6240Y 处理器（18 核，36 线程），启用超线程，启用睿频加速，DDR 内存配置：384 GB（12 插槽/32GB/2666），存储：1 x 英特尔® S3700 200GB（操作系统盘）+ 4 x 英特尔® P4610 1.6TB，英特尔® 傲腾™ 持久内存：1TB（8 插槽/128GB/2666 MHz），网卡：1 x 英特尔® XC710 10Gb + 1 x Mellanox connectx-5 100Gb，操作系统：Smartx customized based on CentOS 7.4，内核：3.10.0-1127.el7.x86_64，工作负载和版本：SMTX OS v4.5，MySQL 8.0.20, Sysbench 1.0.19 和 FIO 3.19

[9] 10 秒内最慢 1% 的 IO 请求延迟。

[11] 10 秒内最慢 5% 的 IO 请求延迟。