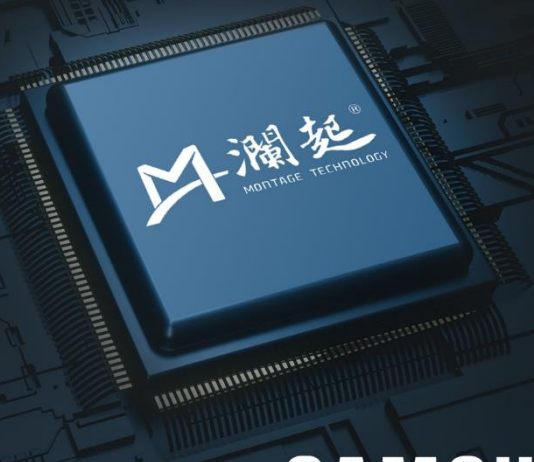

# 三星PCIe Gen5 NVMe SSD PM1743 + 澜起PCIe 5.0/CXL 2.0 Retimer 高性能信号完整性解决方案

# 白皮书




SAMSUNG

# 1. 概述

随着大数据、云计算、5G 等技术的发展，数据计算和高效存储的需求量呈指数级增长，更多的终端设备要求高带宽、强稳定的数据传输，PCIe 凭借其高速的传输速度成为服务器总线的主流解决方案。目前 PCIe 总线的传输速率已经从第一代的 2.5GT/s 演进到了第五代的 32GT/s。支持 PCIe Gen5 的设备能够更加契合当前数据中心及企业级应用的需求，同时各领域均已陆续推出 PCIe Gen5 相关产品，三星已于 2021 年发布了高性能的 PCIe Gen5 NVMe SSD:PM1743，澜起科技于 2023 年 1 月开始量产 PCIe 5.0/CXL 2.0 Retimer，不同厂商的 PCIe Gen5 企业级服务器也均陆续发布。可以预见，PCIe Gen5 将成为全球未来几年的市场潮流，属于 PCIe Gen5 的时代即将全面到来。

极高的信号传输速率使得 PCIe Gen5 能够更好的支持对吞吐量要求超高的高性能设备。但面对 PCIe 每一代成倍增长的信号传输速率，信号衰减变形问题也越来越明显，确保信号传输的稳定性和完整性已成为信号传输质量的瓶颈，限制了传输速率的进一步高速发展。

在数据中心存储应用领域，NVMe SSD 结合 PCIe 接口的方式，可以优化数据传输路径，显著提升数据传输带宽，并且大幅度降低读写延迟，因此已经成为企业级存储设备的首选。目前 PCIe Gen5 NVMe SSD 的主流通信系统链路构成中包括 CPU，主板 PCB，PCIe 扩展卡 PCB，线缆和 SSD 等，其总插入损耗显著超出 PCIe Gen5 标准中规定的端到端的允许信道损耗预算 36dB。为减小信道损耗，寻找可靠的链路预算压缩方案或者高性能的链路扩展解决方案已成为当务之急。市面上常见的高性能链路扩展解决方案主要包括高速 PCB 板材、信号中继器 Redriver 或信号增强器芯片 Retimer 等。其中 Retimer 作为一种协议感知设备，通过其良好的信号处理效果可以提升服务器和 SSD 之间的信号完整性，提高传输质量，并且它具有与高速 PCB 板材相比更低的成本，成为了系统信号完整性设计的最佳解决方案，为服务器 OEM 厂商提供了一种兼容的解决方案，便于服务器的设计和开发。

本白皮书主要介绍了三星公司联合澜起科技推出的基于三星 PCIe Gen5 NVMe SSD (PM1743)和澜起 PCIe 5.0/CXL 2.0 Retimer (M88RT51632)的高性能信号完整性解决方案，并在 PCIe Gen5 服务器上对该方案进行了链路稳定性和读写性能测试，以应对 PCIe Gen5 NVMe SSD 系统链路中遇到的信号完整性设计挑战。

# 2. 硬件设备

## 2.1 三星 PCIe Gen5 NVMe SSD PM1743

在 2021 年，三星就推出了采用其先进的 V-NAND 闪存技术以及最新的 PCIe Gen5 接口的高性能 SSD:PM1743，如图 1 所示。PM1743 在性能和能效方面都超越了上一代产品，其核心特性如下所示，凭借这些特性，PM1743 足以胜任大负荷企业级工作环境，将成为服务器和数据中心降本增效的首要选择。

- **卓越的性能：** PM1743 采用全新 PCIe Gen5 接口，顺序读写高达每秒 14GB/6GB，随机读写速度高达 2500K/280K IOPS，对比 PCIe Gen4 的产品，性能提升了 2 倍。
- **超高的能效：** 每瓦可达 657MB/s，比上一代产品提高了约 40%。这有望大幅降低服务器和数据中心的运营成本，同时也有助于减少碳足迹。
- **尖端的技术：** 采用三星先进的 V-NAND 闪存技术，为企业级服务器提供容量更大、性能更强，延迟更小的存储方案。
- **形态多样：** 拥有从 1.92TB 到 15.36TB 的各种容量，可提供传统的 2.5 英寸外形尺寸，以及厚度仅为 7.5 毫米的 EDSFF(E3.S)封装尺寸，一种专为新一代企业级服务器和数据中心设计的日益流行的固态硬盘外形尺寸，将企业级服务器的存储密度提升一倍。
- **强可用性：** 支持双端口运行，即使一个端口连接失败，通过将工作负载转移到第二个端口，也可以确保服务器的运行稳定性和业务的连续性，促进服务器和存储阵列的一致操作及其可用性。
- **高可靠性：** 利用增强遥测技术实现更有效地远程监控和分析，提供多租户功能并且性能不受影响，此外还通过加密和解密验证来确保更安全的存储，为用户提供了绝佳的可靠性。


| Samsung PM1743 | |
|---|---|
| 形态 | U.2/E3.S |
| 接口 | PCIe 5.0 x4 |
| NAND | V-NAND Technology |
| 端口模式 | Dual |
| 128KB 顺序读写性能 | |
| 顺序读/写带宽 (GB/s) | 14/6 |
| 4KB 随机读写性能 | |
| 随机读/写 (IOPS) | 2500k/280k |
| 4KB 随机读写时延 | |
| 随机读/写时延 (us) | 60/20 |
| 容量 (TB) | 2/4/8/16 |

图 1　三星 PM1743

# 2. 硬件设备

## 2.2 澜起 PCIe 5.0/CXL 2.0 Retimer M88RT51632

澜起科技的 PCIe 5.0/CXL 2.0 Retimer 芯片，采用先进的信号调理技术来提升信号完整性，增加高速信号的有效传输距离。该芯片符合 PCI-SIG 和 CXL 行业组织的相关技术规范，采用业界主流封装，传输速率高达 32 GT/s，在业界率先支持低于 5 ns 的超低传输时延。芯片支持 SRIS 和 Retimer 级联等复杂系统拓扑，是应对下一代服务器、企业存储、AI 加速系统中 PCIe/CXL 信号完整性挑战的理想解决方案。

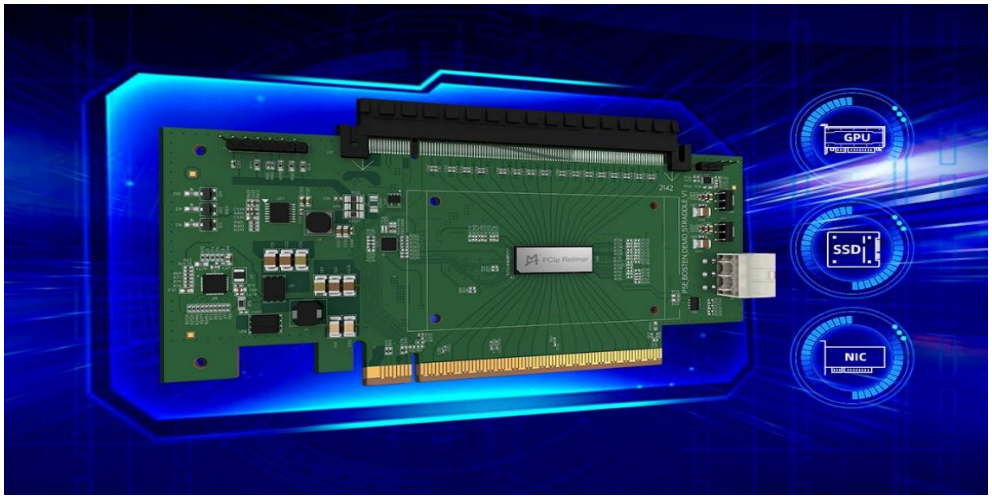



**图 2　澜起 M88RT51632**

## 2.3 基于 PM1743 和 M88RT51632 的信号完整性解决方案

三星公司和澜起科技联合推出了基于三星 PCIe Gen5 NVMe SSD (PM1743)和澜起 PCIe 5.0/CXL 2.0 Retimer (M88RT51632)的高性能信号完整性解决方案，如图 3 所示。通过使用搭载 M88RT51632 的 PCIe 扩展卡，确保了端到端的链路损耗预算均小于 36dB，符合规范要求；并进一步消除了串扰和反射等不良因素的影响，确保系统工作稳定可靠。

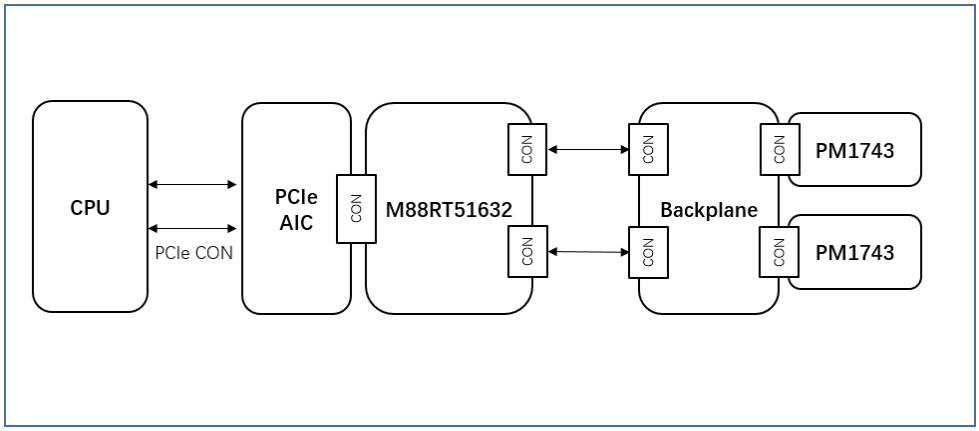



**图 3　三星-澜起 PCIe Gen5 NVMe SSD 解决方案**

# 3. 测试配置

本白皮书中，我们在 PCIe Gen5 服务器上对三星 PCIe Gen5 NVMe SSD PM1743 + 澜起 PCIe 5.0/CXL 2.0 Retimer M88RT51632 的信号完整性解决方案分别进行了以下场景的测试。

- 场景 1：链路稳定性测试
- 场景 2：读写性能测试

## 3.1 链路稳定性测试配置

在链路稳定性测试中，我们进行的测试主要包括基于 Intel CScripts 的测试和 Power Cycle 测试。Intel CScripts 测试主要在主板启动或调试期间进行，可以处理在调试固件时遇到的内存错误、故障转储和其他灾难性错误。Power Cycle 测试可以验证系统在多次开机或者重启后的稳定性和可靠性。Intel CScripts 的具体测试参数设置如表 1 所示。

表 1 Intel CScripts 测试设置

| 测试工具及内容 | 备注 | 参数配置 |
|---|---|---|
| 服务器 | N/A | CPU: Intel Sapphire Rapids SP<br>OS: CentOS 8.3.1-5 |
| Intel CScripts | N/A | 630113-EagleStream-CScripts-Rev2232-6000-Build2585 |
| IOMT 版本 | N/A | 632111/Windows_rev2p0p24 |
| UEFI Shell 版本 | N/A | 2.2 |
| Reset Tests | 热复位等 | 循环次数:10k |
| Power Management | L1 PM(D3hot), L1 ASPM | 循环次数:10k |
| New Feature | 动态链路宽度调节 | 循环次数:10k |
| Link Training | 反复建链，切速等 | 循环次数:10k |

## 3.2 读写性能测试配置

针对读写性能测试场景，我们分别在这一解决方案下对三星的 PCIe Gen5 NVMe SSD:PM1743 进行了裸盘和基于数据库的性能测试，并在相同的测试配置下，与不使用澜起 M88RT51632 的 PM1743 性能测试结果进行了对比，以观察性能差异。

### 3.2.1 FIO 测试配置

在裸盘性能测试中，我们使用的测试工具为 FIO。FIO 是一个工作负载生成器，主要用于 SSD 基准性能测试和

# 3. 测试配置

压力测试，为了使 SSD 达到预期性能，FIO 的测试参数需要进行适当配置，本白皮书中的 FIO 测试方法参考 ODCC 企业级 SSD 测试标准，具体测试参数设置如表 2 所示。

表 2　FIO 参数设置

| 参数 | 备注 | 参数配置 |
| --- | --- | --- |
| **服务器** | N/A | CPU: AMD EPYC 9554 / OS: ubuntu22.04 |
| **FIO 版本** | N/A | 3.35 |
| **Numjobs** | 相同负载的任务个数 | 顺序读写 Numjobs = 1 / 随机读写 Numjobs =16 |
| **QueueDepth** | 队列深度 | 顺序读写 QueueDepth = 64 / 随机读写 QueueDepth = 64 |
| **Block Size** | 一次 I/O 的单位 | 顺序读写 Block Size = 32/64/128KB<br>随机读写 Block Size = 4/32/64KB |

### 3.2.2 MySQL 测试配置

本白皮书中选取了主流的关系型数据库 MySQL 进行测试。MySQL 具有高性能、高可靠性、跨平台灵活等优点，凭借其强大的联机事务处理能力成为全球使用广泛的开源数据库。

测试工具使用 sysbench，sysbench 是一个开源的、模块化的、跨平台的多线程性能测试工具，可以用来进行 CPU、内存、磁盘 I/O、线程、数据库的性能测试。它自带多种工作负载，可以进行不同场景下的压测。MySQL 在本白皮书中的具体测试参数设置如表 3 所示。

表 3　MySQL-SYSBENCH 参数设置

| | |
| --- | --- |
| **服务器** | CPU: AMD EPYC 9554 / OS: ubuntu22.04 |
| **数据库版本** | MySQL-5.7.36 |
| **测试工具** | sysbench-1.0.20 |
| **测试数据集** | 65%SSD 容量大小 |
| **测试工作负载** | oltp_read_write.lua:读写混合负载 |
| **测试线程数** | 不超过 CPU 核心数量 |
| **压测时长** | 3600 秒 |
| **测试结果** | TPS(transactions per second) / 平均时延 |

# 3. 测试配置

### 3.2.3 RocksDB 测试配置

在非关系型数据库的测试场景中，我们选取了 RocksDB 进行测试。RocksDB 依靠其快速低延迟的存储能力，可以适应不同的工作负载,并且满足作为数据库存储引擎、应用程序数据缓存及嵌入式工作负载等多种数据需求。

测试工具使用 YCSB，YCSB 是一款开源的 NoSQL 数据库性能测试工具，能够对云数据库进行测试。自带 6 种工作负载，可修改工作负载中的参数以满足不同的测试要求。RocksDB 在本白皮书中的具体测试参数设置如表 4 所示。

表 4　RocksDB-YCSB 参数设置

| | |
|---|---|
| **服务器** | CPU: AMD EPYC 9554/ OS: ubuntu22.04 |
| **数据库版本** | RocksDB-6.22 |
| **测试工具** | YCSB-0.18 |
| **测试数据集** | 65%SSD 容量大小 |
| **测试工作负载** | workloada:读写均衡负载 |
| **测试线程数** | 不超过 CPU 核心数量 |
| **压测时长** | 3600 秒 |
| **测试结果** | 吞吐量 / 平均时延 |

# 4. 测试结果

## 4.1 链路稳定性测试结果

本白皮书采用表 1 中的参数设置，分别对三星 PM1743 使用澜起 M88RT51632 与不使用时的情况进行了测试，对表 1 中的每项测试内容进行了一万次测试，同时也进行了一千次的 Power Cycle 测试。

测试结果如表 5 所示。在使用澜起 Retimer 的三星 PM1743 的 Power Cycle 测试过程中，未出现 SSD 意外断开连接或无法识别等问题，同时 PM1743 的带宽和速度相比其单独使用时均未降低。Intel CScripts 测试结果也表明，使用 Retimer 并未对 PM1743 的性能及功能产生额外影响，与单独使用 PM1743 的测试结果保持一致。这表明在模拟真实使用环境下，PM1743 + M88RT51632 的解决方案在提高主板上 PCIe 插槽和 CPU 间的信息交互质量的同时，仍能保障系统的稳定运行，提高系统信号的可靠性。

表 5 链路稳定性测试结果

| 测试内容 | PM1743+M88RT51632 | PM1743 |
|---|---|---|
| SBR | 10k pass / 10k cycles | 10k pass / 10k cycles |
| Link Retrain | 10k pass / 10k cycles | 10k pass / 10k cycles |
| Link Disable/Enable | 10k pass / 10k cycles | 10k pass / 10k cycles |
| D3 Hot | 10k pass / 10k cycles | 10k pass / 10k cycles |
| TxEq Redo | 10k pass / 10k cycles | 10k pass / 10k cycles |
| Speed Change Retrain | 10k pass / 10k cycles | 10k pass / 10k cycles |
| Speed Change Max | 10k pass / 10k cycles | 10k pass / 10k cycles |
| Power Cycle | 1k pass / 1k cycles | 1k pass / 1k cycles |

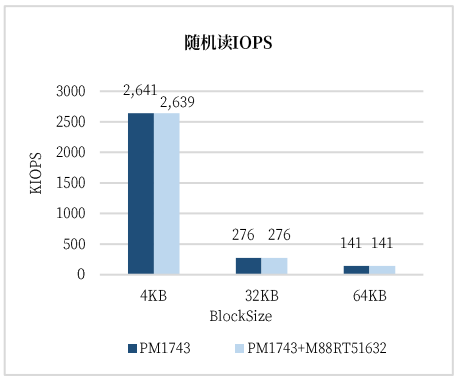
## 4.2 读写性能测试结果

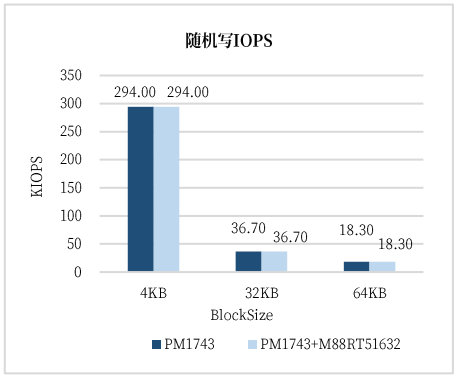
### 4.2.1 FIO 测试结果

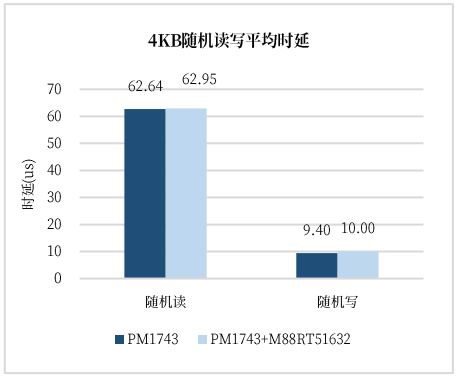
本白皮书中测试了三星 PM1743 配合澜起 M88RT51632 使用与单独使用三星 PM1743 在裸盘性能上的区别。主要进行了基于 FIO 的读写性能测试，测试项包括 32KB、64KB、128KB 的顺序读写和 4KB、32KB、64KB 的随机读写。

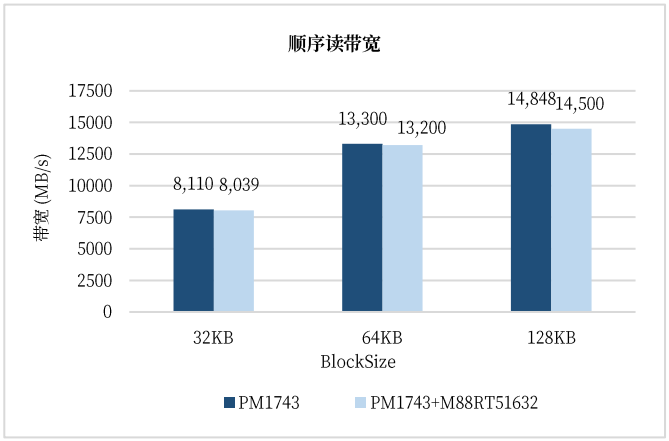
测试结果如图 4 所示。对比 PM1743 裸盘测试结果，使用 Retimer 后对不同 blocksize 的顺序带宽以及随机 IOPS 均无明显影响，顺序读写性能波动最高维持在 2%左右，随机读写 IOPS 差距最大为 0.08%。平均时延数据中 4KB 随机读几乎无差距，4KB 随机写为 6%。这表明使用 Retimer 后三星 PM1743 的系统读写性能与 PM1743 的基准性能结果保持了很好的一致性，该方案符合规格需求。

# 4. 测试结果



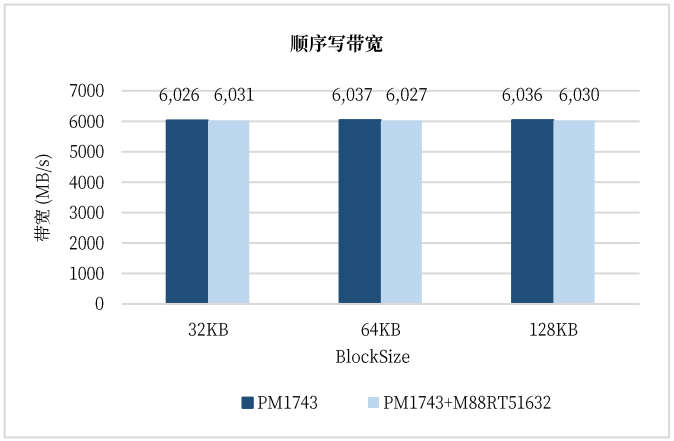
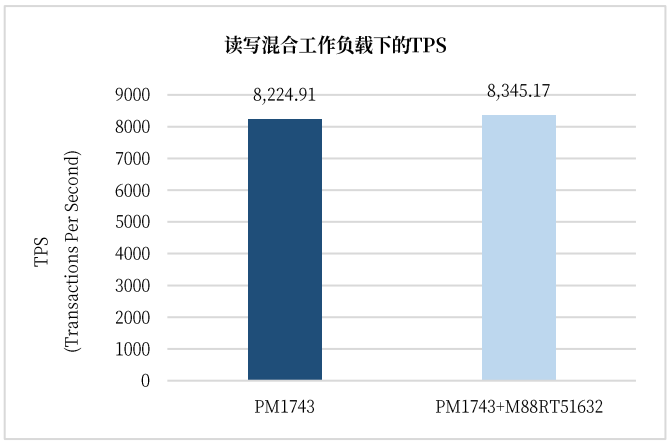




**图 4　FIO 测试结果**

### 4.2.2 MySQL 性能测试结果

在主流关系型数据库 MySQL 的测试中，我们通过 sysbench 分别对使用 M88RT51632 的 PM1743 和不使用的 PM1743 进行了测试，为更接近真实工作负载，本白皮书中使用 sysbench 提供的 oltp_read_write.lua（读写混合工作负载）模拟测试环境进行测试，选取测试结果中的 TPS(transactions per second)和平均时延作为衡量指标。

测试结果如图 5 所示。与未使用 Retimer 的 PM1743 的 MySQL 测试结果相比较，TPS 与平均时延的变化幅度均维持在 1.5%左右，这表明该方案系统链路稳定，NVMe SSD 与 Retimer 一起使用对 MySQL 数据库应用的性能未产生损耗。



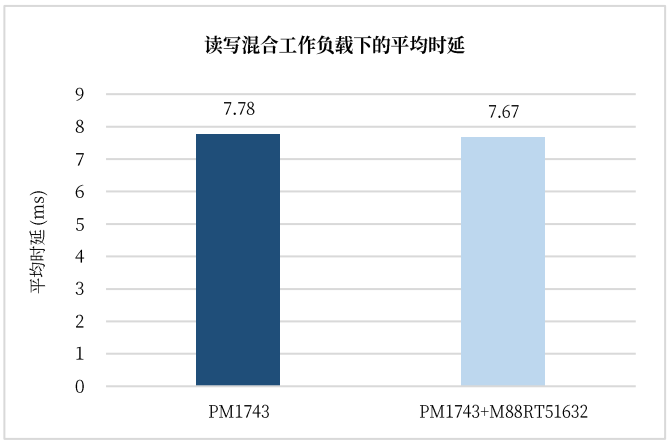



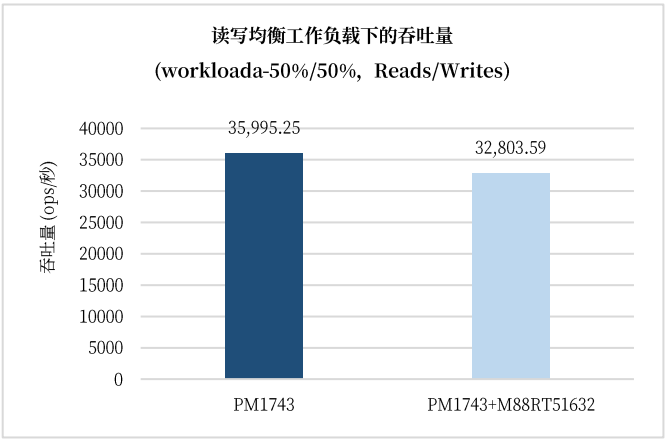

**图 5　MySQL 测试结果**

# 4. 测试结果

### 4.2.3 RocksDB 性能测试结果

在非关系型数据库 RocksDB 的测试中，我们在和 MySQL 测试相同的服务器配置下，分别对单独使用 PM1743 和搭载澜起 M88RT51632 的方案进行了测试，选取测试结果中的吞吐量和平均时延作为衡量指标。

测试结果如图 6 所示。在 RocksDB 的吞吐量和平均时延中，数据均出现了小幅度的变化，对吞吐量的数据而言，直连 PCIe 插槽的 PM1743 与使用 Retimer 的 PM1743 相比出现了约 9%的波动，同时，读写操作的平均时延均有约 7%的变化。针对 RocksDB 整体测试结果而言，该方案的性能及时延数据变化维持在 10%之内，符合预期。



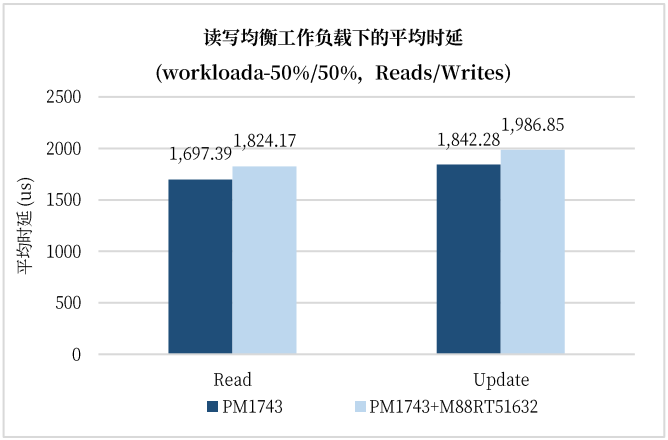




**图 6　RocksDB 测试结果**

# 5. 总结

本白皮书中我们对三星和澜起科技联合推出的基于三星 PCIe Gen5 NVMe SSD (PM1743)和澜起 PCIe 5.0/CXL 2.0 Retimer (M88RT51632)的高性能信号完整性解决方案进行了验证，在 PCIe Gen5 服务器平台分别对该方案进行了链路稳定性测试和读写性能测试。链路稳定性测试结果显示，该方案能够充分保证系统信号的稳定性和完整性。读写性能测试结果显示，不管是在裸盘状态下还是在搭载应用的状态下，该方案的表现都可以达到预期效果，与直连 PCIe 扩展卡的 NVMe SSD 的基准性能相比，使用搭载 PCIe Gen5 Retimer 的 PCIe 扩展卡，NVMe SSD 性能波动在预期范围之内。该方案能够满足当前的高带宽、低延迟要求，充分保障了系统链路的稳定性和系统信号完整性，增强了 NVMe SSD 在读写过程中的数据传输质量，增加了高速信号的有效传输距离。

面对云计算、云服务厂商、数据中心等对存储资料高速读取、交互的需求，PCIe Gen5 NVMe SSD 凭借其更高的 IO 速率和更低的能耗，已逐渐成为各大厂商在 PCIe Gen5 时代搭建基础部署的存储设备首选。PCIe Gen5 在给具有高吞吐量要求的设备提供强有力支持的同时，也进一步缩减了信号传输衰减距离，三星 PCIe Gen5 NVMe SSD + 澜起 PCIe 5.0/CXL 2.0 Retimer 的解决方案保障了主板上部分较远的 PCIe 插槽和 CPU 间的信息交互质量，同时兼顾了服务器 OEM、ODM 和终端用户对系统容量、读写带宽和灵活拓扑的需求。随着 PCIe Gen5 生态系统的进一步完善，信号完整性方案的需求会极大提升，该方案将持续演进，获得更广泛的应用。