白皮书

成本与性能也可兼得 — 英特尔® 傲腾™ 持久内存助力阿里云服务

阿里云

持久内存实例 re7p 基于阿里云第三代神龙架构以及第二代英特尔® 傲腾™ 持久内存 (BPS)，相比上代产品 re6p 的网络带宽/存储带宽实现了翻倍，算力提升 40% 以上；同时基于英特尔® 傲腾™ 持久内存推出的性能增强型本地盘实例产品 i4p 读写延时可以低至 170 纳秒，性能相对于传统 NVMe SSD 高两个数量级，此外，i4p 的单盘 IOPS 高达 140 万以上、单盘吞吐高达 9GB/s，重 IO 的应用在 i4p 上性能有了大幅提升，例如 RocksDB 数据库性能提升 2.5 倍、Click House 数据库性能提升 2 倍、NSQ 消息中间件业务恢复时间提升 2-3 倍。

—— 阿里云弹性计算产品专家
唐湘华

目录

前言

国际分析机构 Canalys 日前发布的 2021 年中国云计算市场报告[1] 显示，中国的云基础设施市场规模已达 274 亿美元，年增速超过 30%，是全球增速最快的市场之一。

随着整个社会进入智慧+ 时代，云计算成为数字经济的基础和中小企业数字化发展的助推器。尤其是新冠肺炎疫情暴发以来，远程办公、在线教育、网络会议等需求爆发式增长，进一步推动了云计算市场快速发展。

阿里云作为全球领先的云计算及人工智能科技公司，依托阿里巴巴强大的技术实力和业务场景，集中了国内外云计算领域的顶尖专家，致力于打造公共、开放的云计算服务平台，并将借助技术的创新，不断提升计算能力与规模效益，将云计算变成真正意义上的公共服务。

云服务器 ECS (Elastic Compute Service) 是阿里云提供的性能卓越、稳定可靠、弹性扩展的 IaaS (Infrastructure as a Service) 级别云计算服务。云服务器 ECS 免去了企业采购 IT 硬件的前期准备，让您像使用水、电、天然气等公共资源一样便捷、高效地使用服务器，实现计算资源的即开即用和弹性伸缩。阿里云 ECS 持续提供创新型云服务器，解决多种业务需求，助力企业业务发展。

挑战

随着国内企业数字化转型的推进，上云成为大多数企业的选择。数字化转型可以赋能业务的持续增长，也能帮助企业可持续发展，但发展快速的数字技术，也让企业的 IT 预算逐年增加，使企业面临现实的 “降本增效” 压力。

例如在企业云化的背景下，各个业务线越来越复杂，数据也呈几何式增长，对数据的存储能力、数据的实时响应速度、系统的稳定性的要求越来越高。在以往解决方案中，通常的做法是将热数据或活跃数据保存在内存中，以保证性能，但每 GB DRAM 的价格居高不下，造成 DRAM 的成本高昂；而且现有 DRAM DIMM 密度也限制了系统内存的物理容量，造成企业中的热数据很难全部存入到内存中。

[1] 引自 Canalys《2021年中国云计算市场报告》

而保存在存储中的温数据，因为 NADN 介质的性能和耐用性不足，在一些近线存储的场景中，低延时的需求增加，超出了 NAND 固态盘的限制，无法满足需求。

由此可见用户不得不在扩展容量和降低成本之间做出选择，而在内存与存储间形成了一个新的需求缺口如图 1。

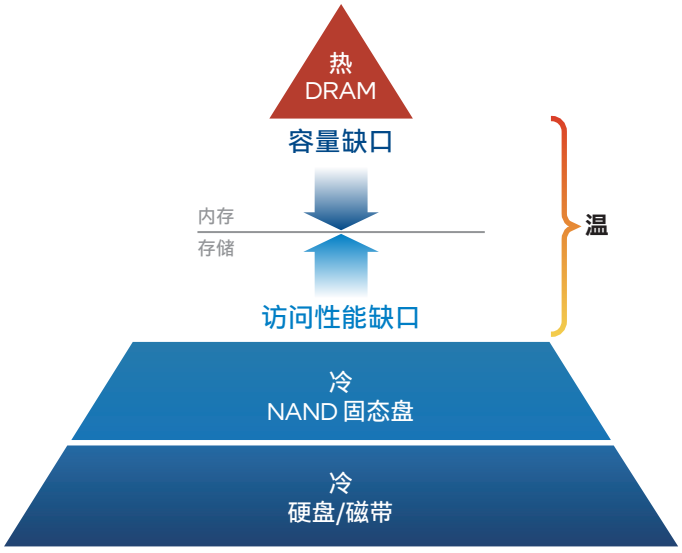


图1

因此整合云计算生态中的各项元素，满足企业业务和技术创新带来的快捷性需求，如何平衡成本与安全性问题成为企业上云的主要问题。作为阿里云 IaaS 层核心产品的云服务器 ECS 需要提供一套性能卓越、稳定可靠、极具性价比的解决方案。

为满足用户对数据性能及成本的需求，阿里云选择将持久化内存技术带到了云上，将阿里云神龙架构与英特尔® 傲腾™ 持久内存进行了深度融合，构建了云上持久内存型产品 (ecs.i4p/re7p)，为用户提供一种成本与性能都可兼得的解决方案。

破局

阿里云是全球第一家基于英特尔® 傲腾™ 持久内存推出云服务器的厂商，在 2021 年最新推出的持久内存实例 — re7p/r7p/i4p 都是基于阿里云第三代神龙架构以及第二代英特尔® 傲腾™ 持久内存 (BPS)，相比上代产品的网络与存储带宽实现了翻倍，算力提升 40% 以上。

同时阿里云还推出了基于持久内存的性能增强型本地盘实例产品 — i4p，同时结合阿里云自研的神龙架构，相对于传统的 NVMe 本地盘实例，i4p 实现了根本性的性能提升，读写延时可以低至 170 纳秒级别，单盘 IOPS 高达 140 万以上、单盘吞吐高达 9GB/s，对重 IO 应用来说，是极大的福音。

英特尔® 傲腾™ 持久内存 PMem

英特尔® 傲腾™ 持久内存 PMem 是一款颠覆传统的内存产品，基于 3D XPoint 介质，具有高速率、低延时、大容量、高性价比和持久保存数据和高级加密等优势。

英特尔® 傲腾™ 持久内存可帮助解决 DRAM 和 NAND 存储之间存在的性能和容量缺口。通过将具有突破意义的英特尔® 傲腾™ 技术与软件相结合，用户可以使用各种配置模式，实现诸多益处，比如显著扩展系统总内存容量；提高虚拟机密度，从而增加每台服务器的虚拟机数量和提升服务器整合率等。英特尔® 傲腾™ 持久内存提供了一种经济可靠的解决方案，可为数据中心和云应用提供满足新一轮数据需求所需的容量和性能。

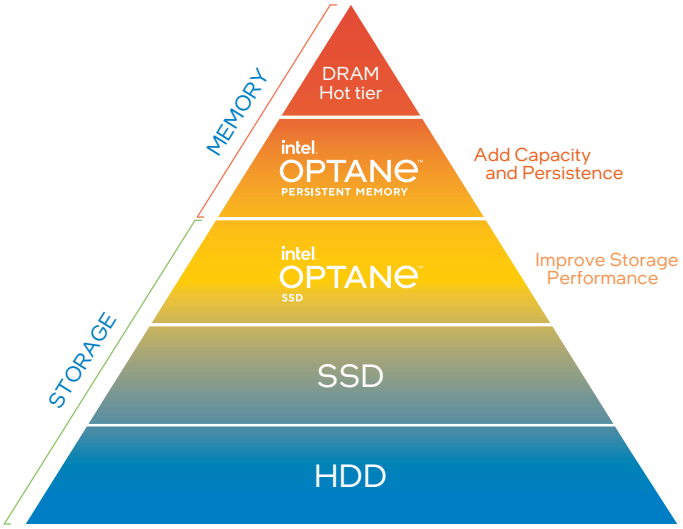


图2 在存储层级中，英特尔® 傲腾™ 持久内存位于 DRAM 下方

英特尔® 傲腾™ 持久内存具有字节寻址能力，传统的数据库主要依靠磁盘存储数据，但磁盘不能按字节寻址，所以数据库读取数据一般是按块读取（比如：4KiB）。持久内存具备字节直接访问 (DAX) 能力，应用直接访问持久内存介质没有内核参与、中断和上下文切换，使得持久内存的性能完全暴露给了应用，因此能够大大简化底层 I/O，加快查询速度。

英特尔® 傲腾™ 持久内存有两种主要工作模式 — 内存模式和 App Direct 模式（以下简称 “AD 模式”）。

在内存模式下，CPU 内存控制器会把持久内存视作易失性的系统内存，而将 DRAM 内存用作持久内存的高速缓存。虽然该模式能够提供更大的内存容量，但在这种模式下数据访问请求会首先在 DRAM 内存上检查是否命中，如果命中，数据直接从内存中获取；如果没有命中，会再到持久内存上进行访问查询。

使用 AD 模式，支持行业标准 NVM 持久化内存编程模式的软件和应用，都能直接与英特尔® 傲腾™ 持久内存通信，且能充分利用其字节寻址能力，访问比文件系统更小的块文件，比如 64 字节、128 字节、256 字节的文件，从而显著降低延迟，满足高速业务场景的需求。利用持久内存容量大的特点，将大部分数据存储在持久内存中，而索引保存在 DRAM 中，这样就可以保证所有关键字的查询都在速度更高的 DRAM 内存中进行，也就能更好地让 DRAM 内存和持久内存实现互补，充分发挥各自的优势。

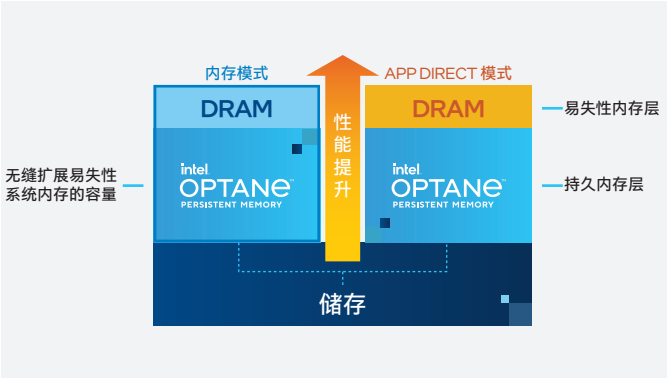


图3 英特尔® 傲腾™ 持久内存两种工作模式

内存增强型实例 re7p

阿里云服务器 ECS 内存型实例 r7p 和 re7p，基于阿里云自研的第三代神龙架构，同时采用第二代英特尔® 傲腾™ 持久内存 (BPS)，提供了超大 CPU 内存配比，可以到 1:20，即 1 个 vCPU 配 20G 内存，这 20G 内存，包含 4G 普通内存和 16G 持久内存。对于内存密集型应用，内存容量配比越大，性价比越高。

应用场景包括：内存型数据库 Redis；需要大容量 Page Cache 的应用，如：RocketMQ、Hadoop 集群、Spark 集群以及其他企业大内存需求应用；运行在此实例上可以大幅度降低单 GiB 内存的成本。

为测试使用第二代英特尔® 傲腾™ 持久内存的 re7p 性能，阿里云测试团队以部署内存型数据库 Redis 应用为例进行了性能测试。Redis 是一个支持持久化的内存数据库，不仅支持简单的 key-value 类型的数据，同时还提供 list、set、zset 和 hash 等数据结构的存储。Redis 以内存作为存储机制，能提供比固态盘更高的数据吞吐带宽和更低的数据处理延时，使得数据处理的速度得到大幅度提升，1ms 的响应延时在数据库圈内一骑绝尘。

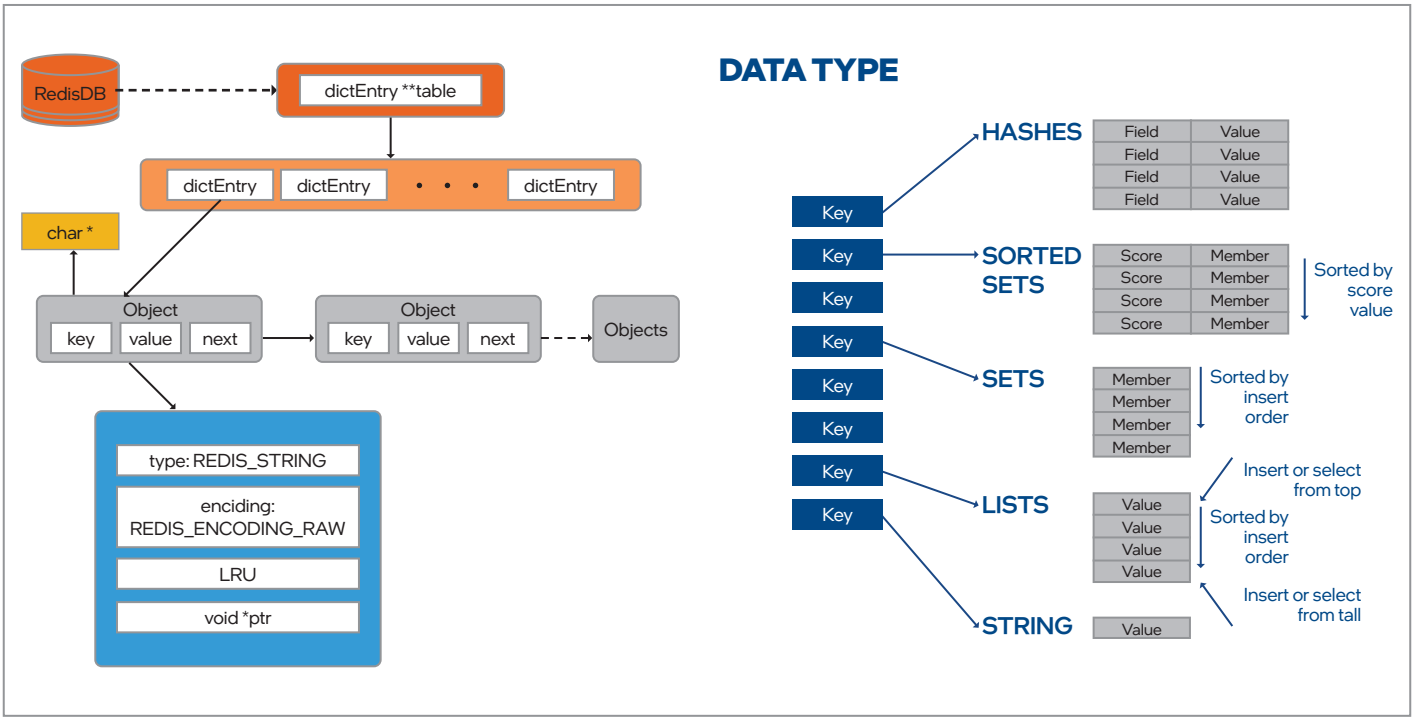


图4 Redis 架构框图

但在实际部署中，遇到了难题，Antirez 的 Redis 社区无法及时满足来自中国互联网客户的快速迭代需求。为此技术团队使用了 TieredMemDB[2]。TieredMemDB 是一个开源的 Redis 分支，它充分利用了 DRAM 和英特尔® 傲腾™ 技术的优势，与 Redis 完全兼容，并支持其所有结构和功能。

TieredMemDB 基于版本不小于 5.1 的内核提供的 KMEM DAX 特性，采用动态阈值算法来管理数据分布，大于或等于阈值的数据都会放到持久内存中，而小于阈值的数据将会放到 DRAM 中，定期监控 DRAM 和持久内存相关的分配器统计信息来不断调整动态阈值，使数据在 DRAM 和持久内存的分布始终符合预定的比例来获得最佳的性能。此外，通过对不同的客户实例设置不同的 DRAM 和持久内存的使用比例，来对客户进行必要的分层和 QoS 的管理。

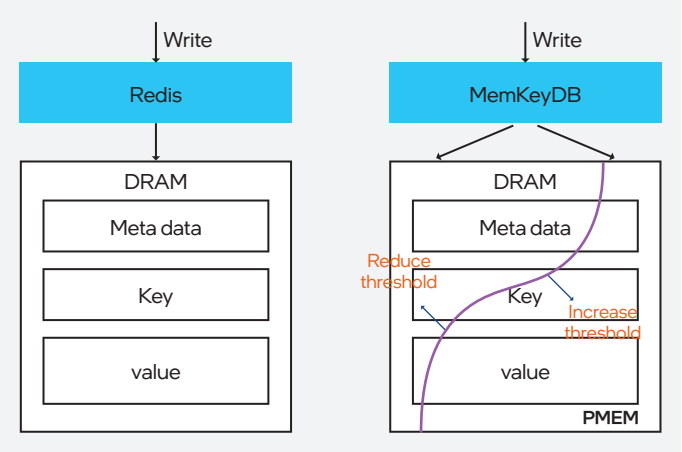


图5 动态阈值控制数据在 DRAM 和持久内存的比例

测试环境采用 re7p 与未采用持久内存的 r7 实例进行分组对比，在 r7.2xlarge 运行标准的 OpenSource Redis6.0.5 作为测试基线，re7p.2xlarge 上运行对应 Redis 版本的 TieredMemDB，测试结果比较如下：

实例规格	re7p.2xlarge	r7.2xlarge
vCPU	8	8
内存 (GiB)	160	64
读 QPS	848445.52	767780.45
写 QPS	804766.93	780125.3
读延时 (ms) (p999)	1	1
写延时 (ms) (p999)	1	1
单 GB 内存价格 (元/月)	11.87	22.46

从上表可以看出持久内存实例 re7p 可以以更低的单 GB 内存价格实现与基于 OpenSource Redis 的 r7 实例读写 QPS 类似甚至更高的性能，同时满足服务级别协议 (SLA) (读写延时 <1ms)，此外用户还可以拥有更多的内存空间。

性能增强型本地盘实例 i4p

阿里云服务器 ECS 性能增强型本地盘实例 i4p，基于阿里云第三代神龙架构和第二代英特尔® 傲腾™ 持久内存，提供性能极高的本地盘。整机 IOPS 最高可达 300 万，单路访问延迟低至 30 微秒。

这款产品适用于磁盘类 KV 型数据库，例如：RocksDB、ClickHouse；OLTP、高性能关系型数据库进行 WAL 优化等；NoSQL 数据库，例如 Cassandra、MongoDB、HBase、Elasticsearch 等搜索场景，以及其他频繁将数据写入磁盘的 I/O 密集型应用，例如消息中间件、容器。

在 i4p 上运行 RocksDB，其读写性能比传统直接存本地盘吞吐高 2~2.8 倍，延时降低 23%；同时整体的数据可靠性也会更高：

a. SST 数据存放在 ESSD 中，数据三副本保护，再无数据丢失之忧；

b. WAL 日志放在 BPS 中，而基于 BPS 模拟的本地盘，其故障率远低于传统的 SSD 及机械硬盘，WAL 日志数据可靠性得到有效保障。

RocksDB 项目起源于 Facebook 的一个实验项目，旨在充分实现快存上存储数据的服务能力。RocksDB 借鉴了开源项目 LevelDB 的重要代码和 Apache HBase 项目的重要思想。最初的代码来源于开源项目 leveldb 1.5 分支。

RocksDB 基于 LSM-Tree 数据结构，可以在不同的生产环境 (纯内存、Flash、Hard Disks or HDFS) 中调优，支持不同的数据压缩算法。其主要设计点是在快存和高服务压力下性能表现优越，所以该 DB 需要充分挖掘 Flash 和 RAM 的读写速率。

RocksDB 支持高效的 point lookup 和 range scan 操作，需支持配置各种参数在高压力的随机读、随机写或者二者流量都很大时性能调优。

[2] https://github.com/TieredMemDB/TieredMemDB

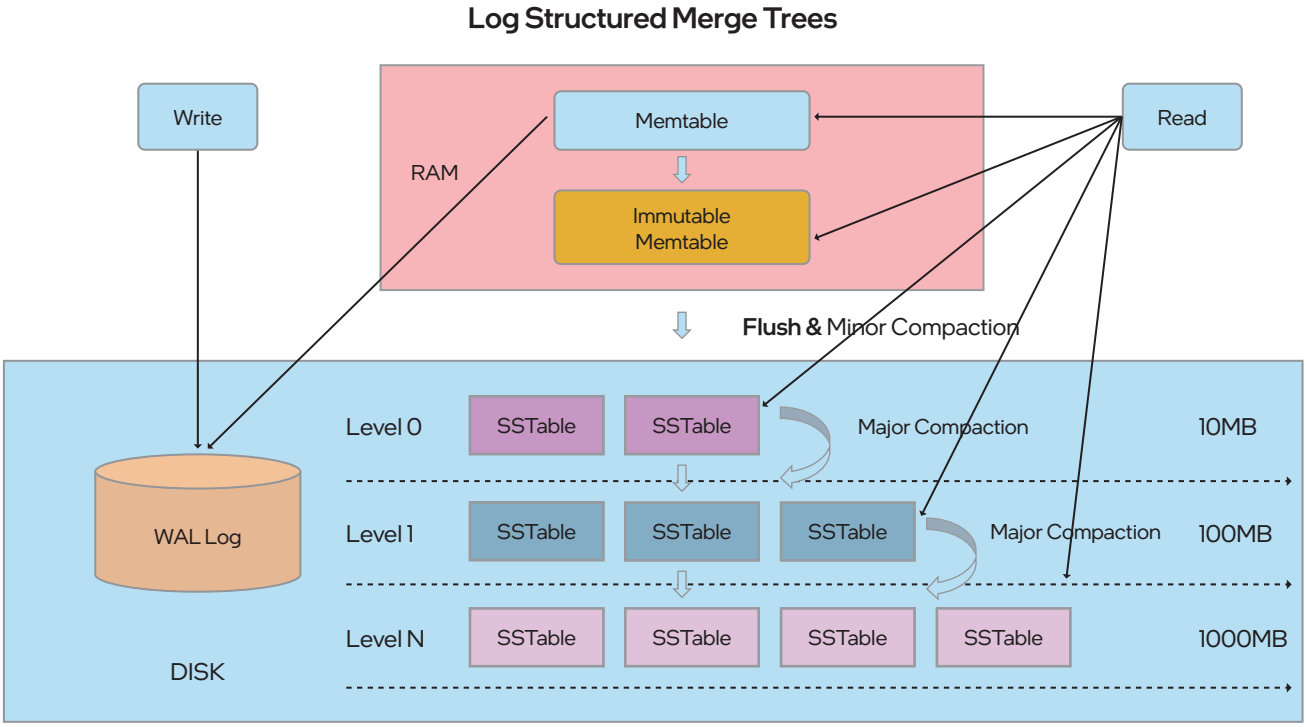


图6 LSM-Tree 持久内存优化的方案

基于英特尔® 傲腾™ 持久内存的优化

• 写优化

RocksDB 的写操作，先写 WAL，再写 Memtable，在出现系统崩溃的时候，WAL 日志可以用于完整的恢复 Memtable 中的数据，以保证数据库能恢复到原来的状态。在默认配置的情况下，RocksDB 通过在每次写操作后对 WAL 调用 flush 来保证一致性。在使用持久内存的场景下，可用使用 mmap 函数来映射 WAL 文件到应用的虚拟内存空间进行读写，以及使用 NT-Store 指令来加速写的效率。如图 7 所示：

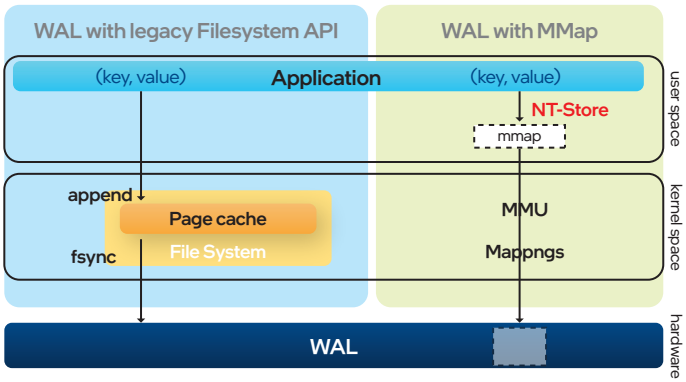


图7 WAL 写优化

• 读优化

RocksDB 的 Persistent Cache 是为了在低延迟的设备上提高读性能而设计的。PMem 作为一款低延时设备，其空间远大于 DRAM 并且其数据是压缩的。因此，采用持久内存存放 Persistent Cache 可以减少 block cache 对 DRAM 的消耗，从而提高 RocksDB 整体读流程的 cache 的命中率。

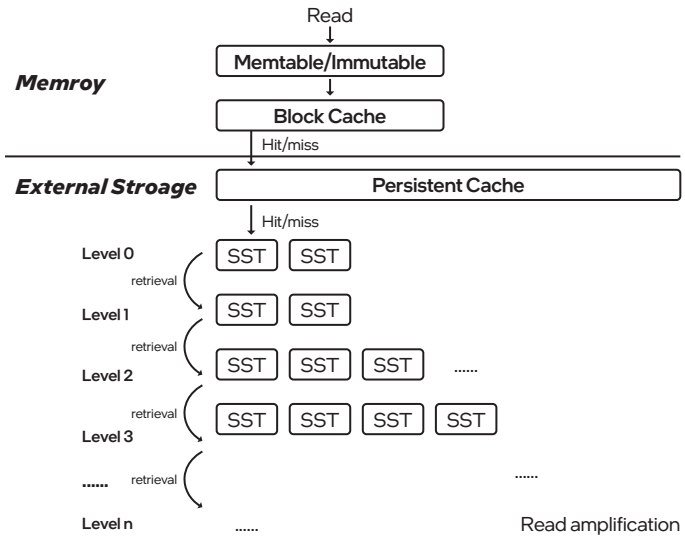


图8 读优化

测试

为了验证高端本地盘 i4p 的性能，阿里云的测试团队将部署 RocksDB 的 i3 实例（无持久内存）与 i4p 实例，进行读写性能测试。

• 测试环境

实例规格	i3	i4p
物理核数	26	32
内存容量	384	256
BPS 容量	0	1T
ESSD/SSD 容量	1.8T	1.8T

• 测试方案

- i3 作为基准线，SST 和 WAL 文件都指向 NVMe SSD，读利用 DRAM 作为 block cache。
- i4p 中，WAL/Persistent cache 路径指向持久内存，SST 数据文件指向 NVMe SSD。WAL 利用 libpmem 优化写，读利用持久内存存放比 block cache 更大容量的 Persistent cache 而获得收益。
- Persistent cache 设置为 128GB，Block cache 设置为 10GB。
- 持久内存与 NVMe SSD 公平比较，打开 RocksDB 的 io-direct 选项避免读写 page cache。

• 版本编译

源码：https://github.com/pmem/pmem-rocksdb
make db_bench ROCKSDB_ON_DCPMM=1 DEBUG_LEVEL=0 -j

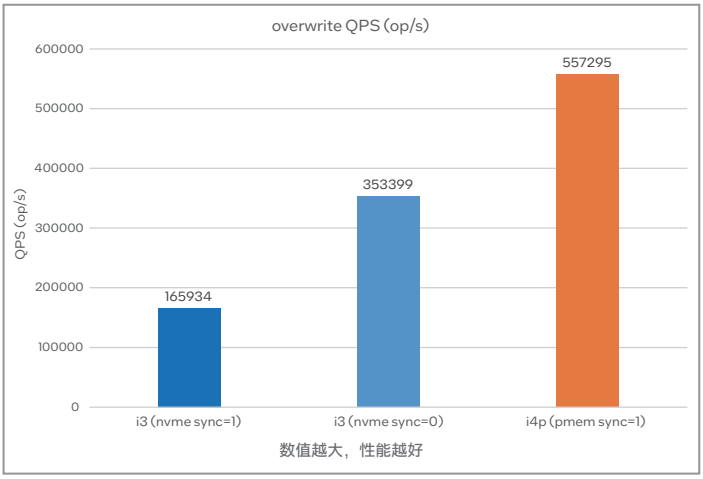
测试步骤：（1）生成 100GB 的 Dataset
（2）利用 db_bench 写操作

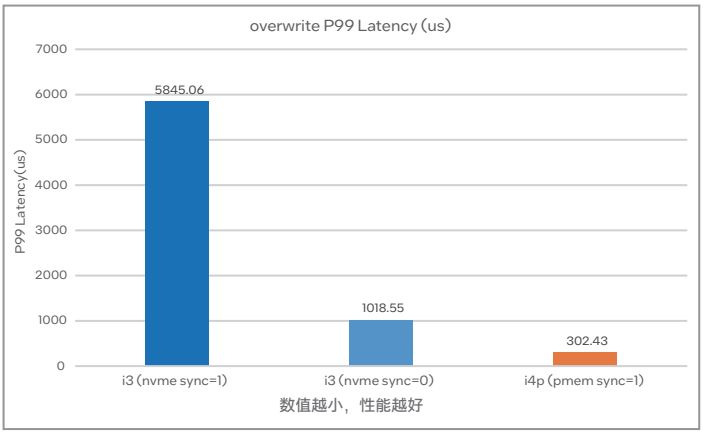
主要参数：

Dataset:	100GB
Key size:	16B
Value size:	128B
Threads:	96
Writes/Reads:	500000

• 写性能测试结果

在 i4p 上，将 WAL 落在持久内存上（sync 打开，获得数据强一致性）并利用 mmap 优化后的写性能比 WAL 落在 i3 上的 NVMe SSD 写性能要高。当 sync 关闭后（断电时，数据可能存在丢失风险），虽然在 i3 的 NVMe SSD 上写的性能要增加一倍，但是还是比 i4p 低不少。





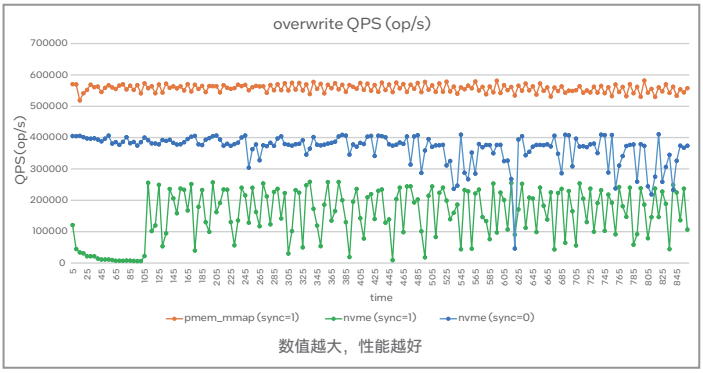


图9 i3 与 i4p 写性能比较

• 读性能测试结果

从实时 QPS 可以看到，i4p 的持久内存用于 persistent cache 有一个比较长的加载的过程，cache 填充完毕后，QPS 主要维持在 110 万左右，i3 的 NVMe SSD 上 10G 的 block cache 加载过程很短，加载后 QPS 维持在 30w 左右。

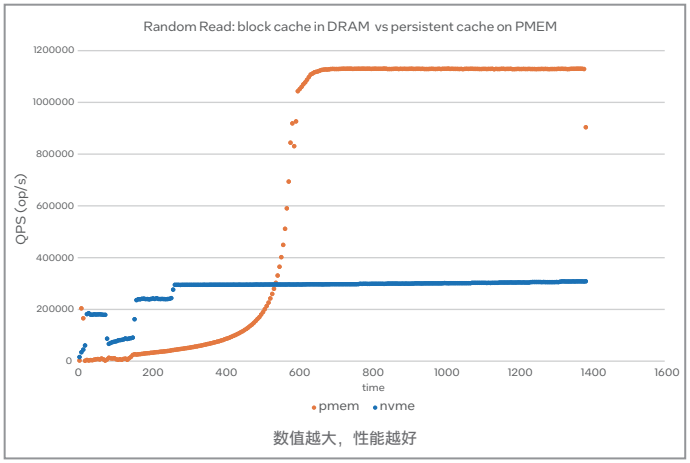


图10 i3 与 i4p 读性能比较

综上所述，利用英特尔® 傲腾™ 持久内存可以加速 RocksDB 的读写性能。基于 WAL 的优化可以将写性能提升 3.4X，基于Persistent Cache 的优化可以提升 3.7X 的读性能。

效果

基于英特尔® 傲腾™ 持久内存和阿里云第三代神龙架构的实例 re7p，一经推出后，带来了巨大的改变与收益：

- 性能得到一定提升：i4p 与同类产品相比，读写时延可控制在 1 微秒以内，可以有效提升系统的读写性能，降低延时。
- 总体拥有成本大大减少：对于现金流不足、IT 部门预算有限的企业，阿里云 ECS re7p/i4p，提供了更大的内存容量、更便宜的内存介质。刚好可以满足用户的降本诉求，为他们带来的切实的降本增效。

展望

随着云计算已经发展成为支持数字经济发展、承载各个行业实现数字化转型的重要基础设施。用户对云服务性能的要求会越来越高，这些技术与应用场景的呼应必将有力推动产品的不断创新和发展。

作为国内云服务行业的引领者，未来，阿里云将会与英特尔紧密合作，对基于英特尔® 傲腾™ 持久内存的 re7p/i4p 实例产品，进行持续的迭代更新，发挥自身技术与规模的红利，保持阿里云 ECS 持久内存实例行业领先的竞争力，为更多的企业用户提供稳定、高效、极具性价比的云服务。