英特尔助力腾讯云深度优化云硬盘 CBS 产品，打造极速云存储体验

前言概述

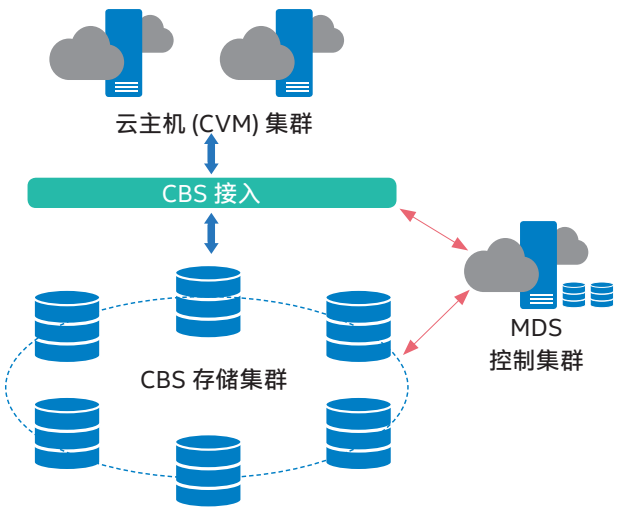
在更多企业核心系统“云化”的今天，云存储系统作为业务数据的重要载体，其性能表现正受到越来越多的关注。作为全球领先的云服务提供商之一，腾讯云通过先进的云硬盘 CBS (Cloud Block Storage) 产品为众多行业用户提供高效、可靠的持久性块存储服务，并在核心数据库、内容分发网络 (Content Delivery Network, CDN) 及电商系统等用户场景中获得了广泛的部署和使用。

为向用户提供性能更为卓越的企业级云存储服务，腾讯云与深度合作伙伴英特尔一起，以全新的存储引擎设计和英特尔® 傲腾™ 持久内存来重构和优化腾讯云的极速型固态盘 CBS产品。验证表明，新的产品方案能以更佳的带宽、更低的时延和更高的每秒读写次数 (Input/Output Per Second, IOPS)，为性能密集型用户业务场景打造极速云存储体验。

挑战：快速发展的云服务对云存储性能提出更高要求

无论是正兴的互联网、大数据或人工智能等领域，还是传统的金融、医疗和制造等行业，云服务都已逐渐成为企业下一代 IT 基础设施的标准之一；而作为企业未来业务数据的重要载体，包括云硬盘在内的云存储产品与解决方案的性能表现，也成为企业选择云服务的一个重要考量因素。

作为全球领先的云服务提供商之一，腾讯云一直以先进的云硬盘 CBS 产品为用户提供持久性块存储服务。典型的腾讯云 CBS 产品存储系统架构如图一所示，由 CBS 接入、MDS 控制集群以及 CBS 存储集群构成。当 CBS 接入收到 CVM 云主机集群的数据读写请求后，会根据 MDS 提供的集群路由信息，将读写请求转发至对应的 CBS 存储节点中。



图一 腾讯云 CBS 产品存储系统架构

依托于雄厚的技术积累以及持续不断的技术优化与演进，腾讯云 CBS 产品性能卓越，可用性、可靠性及可扩展性俱佳：

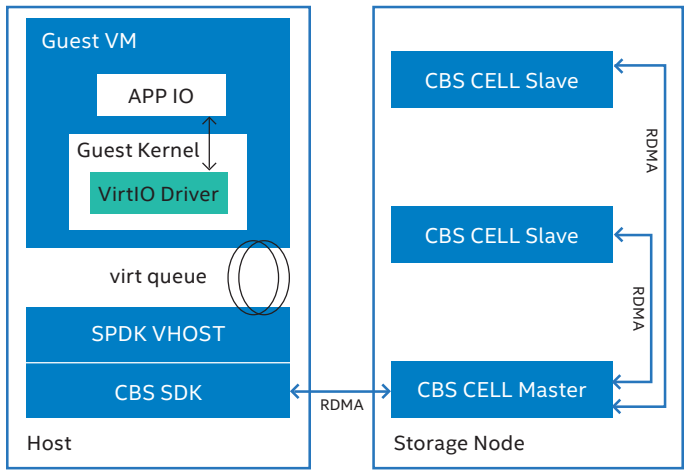
- **高性能：**基于英特尔高性能 NVMe 固态盘和腾讯云创新自研存储引擎的有效组合，CBS 产品目前已可为用户业务场景提供单盘最大 110W 的随机 IOPS 性能，以及最高 4Gbps 每秒的带宽能力；
- **高可用性：**通过高可用和容灾设计，CBS 产品能有效降低系统不可用概率，并可通过快照 (Snapshot) 方式备份用户数据，防止因篡改和误删导致数据丢失，保证在业务故障时能快速回退；
- **高可靠性：**通过三副本的分布式机制，CBS 产品可为用户提供高达 99.9999999% 的数据可靠性；而借助优异的数据复制机制，CBS 产品也能在副本出现故障时快速进行数据迁移恢复，保障用户业务不受影响；
- **高可扩展性：**CBS 产品允许用户根据业务需求自由配置存储容量，按需扩容。目前系统单磁盘容量最大可支持 32TB，单个云主机累计可挂载 640TB，使用户能够从容应对 TB/PB 级的大数据处理场景。

凭借以上优势，腾讯云 CBS 产品在不同用户业务场景，如高负载 OLTP (On-line Transaction Processing，联机事务处理) 的金融交易系统、高吞吐的电商系统、面向人工智能的数据分析系统，以及高并发的 CDN 网络等中都具有不俗表现，并获得了用户的良好反馈。

但从 CBS 的产品架构中可以看到，基于分布式构建的存储集群，令来自网络的接入、传输时延等因素会对其整体性能产生影响，从而与本地化存储产生差异。这也是用户在核心数据库、CDN 网络等性能敏感场景中对采用 CBS 产品仍抱有迟疑的原因之一。而随着云服务逐渐成为企业业务系统的核心载体，更多更复杂的核心业务数据读写需求正驱动着腾讯云对极速型 CBS 产品开展进一步深度优化以提升性能，消除用户对 CBS 产品的顾虑。

针对 CBS 产品的架构、存储引擎以及硬件基础设施，腾讯云加入了对远程直接数据存取 (Remote Direct Memory Access，RDMA) 协议的支持，并与英特尔携手，开展了多方位的优化，包括：

- 加入轮询、算法优化、消除竞争以及消除锁等机制，优化 CBS 存储引擎；
- 引入由英特尔提供的 SPDK (Storage Performance Development Kit) 开发套件，优化 NVMe 固态盘的 IOPS 和时延性能。



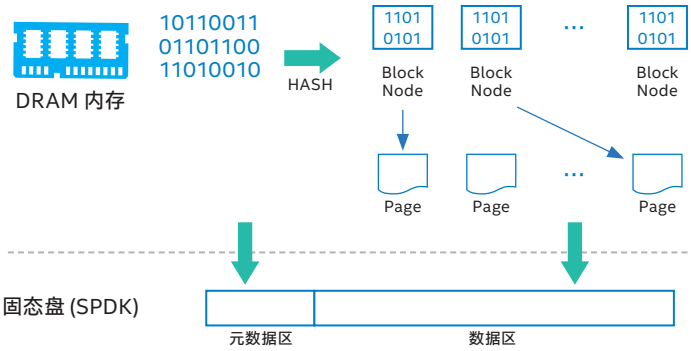
图二 CBS 极速云盘架构图

在进行上述架构、引擎和软件方案的优化后，双方发现固态盘产品本身的时延性能也成为进一步提升 CBS 产品整体性能的障碍。要应对这一问题，有效方法就是为方案寻找一种更具性能优势的存储介质。

为此，腾讯云与英特尔一起，借助英特尔® 傲腾™ 技术这一存储“黑科技”，以英特尔® 傲腾™ 持久内存作为新一代极速型 CBS 产品的存储核心，并重构数据落盘方案，来满足性能密集场景在时延上的更高要求。

解决方案：借力存储“黑科技”，为极速型 CBS 产品打造更佳性能

在腾讯云既有的极速型固态盘 CBS 产品设计中，数据的落盘过程如图三所示，来自计算集群的云主机数据首先通过 HASH 找到或分配到对应的块节点 (Block Node) 中，然后数据会被缓存到不同的 Page。接下来，系统需要进行执行两次写操作，一次将业务数据写入固态盘对应的数据区；另一次是将元数据 (Metadata) 以 LOG 方式追加 (wAppend) 写入固态盘中。

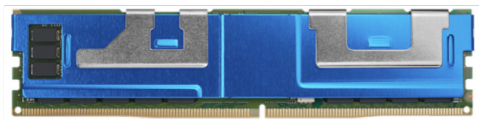


图三 腾讯云既有极速型 CBS 产品数据落盘过程

可以看到，这一过程需要对固态盘执行两次写操作。基于 NAND 闪存构建的固态盘写入时延通常为数十微秒，因此两次写入过程就会带来数十乃至近百微秒的时延。这一数字虽然看起来很小，但在端到端网络时延可达 1 毫秒 (1000 微秒) 的 5G 时代，其显然还是会制约 CBS 产品的整体性能。

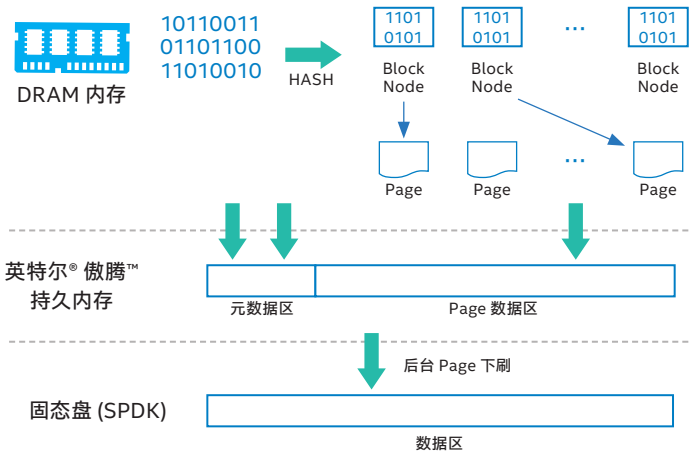
同时，NAND 固态盘数据写入需要以块为单位，且写入前需要做擦除操作等特性，一方面会带来写入效率的降低，另一方面也大幅降低了其使用寿命 (即所谓的“写入放大”问题)。此外，在 LOG 的回收过程中还存在相应的毛刺问题。

而基于英特尔® 傲腾™ 技术构建的英特尔® 傲腾™ 持久内存则可以帮助 CBS 产品有效应对以上问题。英特尔® 傲腾™ 技术通过一种全新的无晶体管存储架构，能在三维矩阵中堆叠存储网格，从而在提高存储密度、增强读写性能的同时，提供持久化的存储能力。持久内存可按字节寻址，可以像内存一样精准控制读写的位置和大小。



图四 英特尔® 傲腾™ 持久内存 200 系列

与传统 DRAM 内存相比，由英特尔® 傲腾™ 技术与其它英特尔先进存储控制技术、接口硬件，以及软件增强功能相结合构建的英特尔® 傲腾™ 持久内存具有两大显著优势：首先其存储密度更高、单位存储成本更低，可帮助用户更为经济地扩展云存储能力；其次，App Direct 模式下的英特尔® 傲腾™ 持久内存所具备的持久性特性，使之可以有效充当 CBS 产品的数据持久化存储载体。



图五 优化后腾讯云极速型 CBS 产品数据落盘过程

得益于英特尔® 傲腾™ 持久内存的创新特性，极速型 CBS 产品的数据落盘过程，如图五所示得以优化。首先来自计算集群的数据会通过 HASH 分配到对应的块节点并缓存到 Page 中，然后数据就马上会被持久化存储到英特尔® 傲腾™ 持久内存中，同时 Page/Block 的元数据也会原地更新到对应的数据区中。

除数据落盘过程实现优化之外，用户还可通过定制化的策略和算法，决定是否将英特尔® 傲腾™ 持久内存中的数据进一步下刷到固态盘中。例如，将需要频繁读写的“热数据”保留在持久内存中，而长时间不需访问的“冷数据”定期被后台转移至固态盘中，以有效降低 CBS 产品的总拥有成本 (Total Cost of Ownership, TCO)。

在提供先进存储硬件产品的基础上，英特尔® 持久内存开发工具包 (Persistent Memory Development Kit，PMDK) 为 CBS 产品提供了面向英特尔® 傲腾™ 持久内存的编程模型和环境。

以其中的 libpmem 库为例，作为 PMDK 中的底层库，其支持用内存映射方式访问持久内存，这一方式可将持久内存上的文件映射到应用程序的虚拟内存空间进行操作。通过规避内核参与和上下文切换带来的开销，使持久内存的性能可直接为应用程序提供助益。

同时，libpmem 库也可以检测处理器的特性而使用最为高效的持久化指令 (例如 CLWB、CLFHASHOPT 等) 将数据写入到持久内存中。CLWB 指令具有并发能力，同时可在刷新数据后仍然保证处理器缓存有效。除此之外，libpmem 还封装了 NTW (Non Temporal Write) 指令，该指令能利用写合并方式来绕过处理器缓存 (Cache)，直接将数据从 Store Buffer 中写入内存控制器的 WPQ 中，从而提高性能。

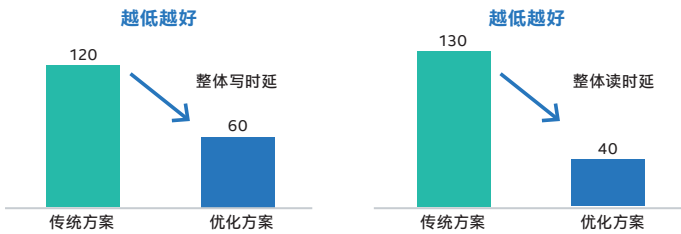
得益于以上特性，libpmem 库不仅能以丰富的接口帮助用户实现对整个写入流程更加细致和准确的控制，也通过使用内存映射 (Memory Mapping) 的访问方式，并结合 NTW 写入指令提升整个系统基于持久内存访问的写入性能，从而让英特尔® 傲腾™ 持久内存的各项特性在 CBS 新方案设计中发挥效能。

效果：创新硬件与优化设计为 CBS 产品带来全方位收益

与既有方案相比，基于英特尔® 傲腾™ 持久内存设计的 CBS 产品优化方案在落地实施后，带来了巨大的改变及收益，包括：

- **数据读写时延大幅缩短：** 一方面，相比 NAND 固态盘数十微秒的读写时延，英特尔® 傲腾™ 持久内存的读写时延可控制在 1 微秒以内；另一方面，借助 PMDK 提供的函数库与工具， 英特尔® 傲腾™ 持久内存可对整个写入流程实现更加细致和准确的控制，并有效提升系统的写入性能。
- **系统使用寿命有效提升：** 一方面，英特尔® 傲腾™ 持久内存可按字节寻址的特性有效解决了以往 NAND 固态盘的“写入放大”问题，从而避免因反复擦写造成的设备使用寿命降低；另一方面，英特尔® 傲腾™ 技术独有的存储结构也令英特尔® 傲腾™ 持久内存具有更长的使用期限；
- **增强存储空间使用效率：** 英特尔® 傲腾™ 技术允许单独访问和更新内存单元，所以英特尔® 傲腾™ 持久内存无需再执行垃圾收集，进而避免了以往 NAND 固态盘面临的回收毛刺问题，提升了存储空间的使用效率。

为验证新硬件与优化设计对 CBS 产品产生的效果，腾讯云与英特尔合作开展了多方位的验证测试。测试结果如图六所示，采用英特尔® 傲腾™ 持久内存构建的 CBS 产品方案与优化前相比，整体写时延从 120 微秒下降到 60 微秒，整体读时延从 130 微秒下降到 40 微秒，同时 IOPS 可高达 200W 以上，性能获得了有效提升[1]。



图六 新方案令 CBS 产品读写时延显著下降

展望：以先进产品与技术为用户创造更佳云存储体验

随着云计算、云存储技术的不断完善，云服务正在企业级业务系统中扮演越来越重要的角色，而用户也势必会对各类云服务的性能提出更多和更高的要求，这些技术与应用场景的互动有力推动着相关产品与技术的持续演进与优化。作为云服务行业的重要参与者和引领者，腾讯云与英特尔基于英特尔® 傲腾™ 持久内存开展的 CBS 产品优化及所取得的收益，正是这一趋势的显著体现。

面向未来，腾讯云与英特尔也将基于这一成功实践，在云计算、云存储等领域开展更广泛合作，运用更多先进产品和技术持续优化 CBS 等云服务产品。例如双方计划在基于英特尔® 傲腾™ 持久内存的方案设计中加入 RDMA 协议，从而有效降低处理器和内存开销。同时，随着全新第三代英特尔® 至强® 可扩展处理器的到来，其不仅能以更多的内核、更优化的架构和更大的内存容量为云服务产品带来更强性能助力，也能与新一代英特尔® 傲腾™ 持久内存形成良好的配合，为用户数据打造更佳云存储体验，使 CBS 等云存储产品成为未来企业级业务数据存储的可靠依托。

intel®

脚注和法律声明

[1] 性能测试结果基于「2021-04-27」进行的测试，且可能并未反映所有公开可用的安全更新。详情请参阅以下详细的测试配置信息。没有任何产品或组件是绝对安全的。
测试配置：处理器：双路英特尔® 至强® 铂金 8255C 处理器；内存1：384GB(32GB*12@2666 MHz)；内存 2：英特尔® 傲腾™ 持久内存 128G * 12；存储 1：英特尔® 固态盘 480GB；存储 2：英特尔® NVMe 固态盘 3.84TB*12；网络适配器：100GE*2；操作系统内核版本：5.4.110-1.el7.elrepo.x86_64
关于性能和基准测试程序结果的更多信息，请访问 www.intel.com/benchmarks。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。
描述的成本降低情景均旨在在特定情况和配置中举例说明特定英特尔产品如何影响未来成本并提供成本节约。情况均不同。英特尔不保证任何成本或成本降低。
英特尔技术特性和优势取决于系统配置，并可能需要支持的硬件、软件或服务得以激活。产品性能会基于系统配置有所变化。没有任何产品或组件是绝对安全的。更多信息请从原始设备制造商或零售商处获得，或请见 intel.com。