文章类型：论文|刊号（ISSN）：2737-4505(P) / 2737-4513(O)

虚拟机环境下高性能存储架构的设计与性能评估

李保成
西南财经大学天府学院
DOI:10.32629/etd.v6i11.17479

[摘　要] 本文旨在系统性地研究虚拟机环境下的高性能存储架构设计与性能评估。首先,深入剖析了虚拟机I/O路径中的性能瓶颈。其次,本文提出了一种融合多种优化策略的高性能存储架构设计方案,该方案涵盖了从硬件加速(如NVMe、SPDK)、软件栈优化(如vhost-user、io_uring)、到智能调度与资源管理(如QoS保障、NUMA感知)的多层次协同优化。在此基础上,本文设计并实施了一套全面的性能评估体系,通过构建基准测试平台,对所提出的架构在吞吐量、延迟、CPU开销及可扩展性等关键指标上进行了量化分析。实验结果表明,相较于传统的基于内核态virtio-blk的存储方案,本文提出的架构在高并发场景下能显著提升I/O性能(吞吐量提升最高达300%,延迟降低高达70%),同时有效降低了宿主机CPU开销。最后,本文总结了研究成果,并对未来在持久内存(PMem)、计算存储分离(CSD)以及AI驱动的智能存储调度等方向的发展趋势进行了展望。
[关键词] 虚拟机；高性能存储；I/O虚拟化；性能评估；SPDK；vhost-user；QoS
中图分类号：TP333　文献标识码：A

Design and Performance Evaluation of High-Performance Storage Architecture in Virtual Machine Environments

Baocheng Li
Tianfu College of Southwest University of Finance and Economics

[Abstract] This paper aims to systematically study the design and performance evaluation of high-performance storage architectures in virtual machine (VM) environments. First, it conducts an in-depth analysis of the performance bottlenecks in the VM I/O path. Subsequently, a high-performance storage architecture design integrating multiple optimization strategies is proposed, covering multi-level collaborative optimizations from hardware acceleration (e.g., NVMe, SPDK), software stack optimization (e.g., vhost-user, io_uring), to intelligent scheduling and resource management (e.g., QoS assurance, NUMA awareness). Based on this, a comprehensive performance evaluation system is designed and implemented. By constructing a benchmark testing platform, a quantitative analysis of the proposed architecture is performed on key metrics such as throughput, latency, CPU overhead, and scalability. Experimental results indicate that compared to the traditional kernel-based virtio-blk storage solution, the proposed architecture significantly improves I/O performance under high-concurrency scenarios (throughput increased by up to 300%, latency reduced by up to 70%), while effectively reducing the host CPU overhead. Finally, the research findings are summarized, and future development trends in directions such as persistent memory (PMem), computational storage disaggregation (CSD), and AI-driven intelligent storage scheduling are discussed.
[Key words] Virtual Machine; High-Performance Storage; I/O Virtualization; Performance Evaluation; SPDK; vhost-user; QoS

引言

数字化时代,企业级应用等业务对计算和存储性能需求激增。虚拟化技术虽提升数据中心资源利用率与运维灵活性,但资源抽象带来新挑战,尤其在I/O子系统方面。虚拟机监控器处理I/O请求时引入额外开销,形成“I/O墙”问题,对I/O密集型应用影响大。因此,如何在保证虚拟化优势的同时消除I/O性能瓶颈,设计高效存储架构,成学界和业界关注焦点。本研究将梳理痛点,提出创新高性能存储架构并验证,为构建下一代云原生基础设

文章类型：论文|刊号（ISSN）：2737-4505(P) / 2737-4513(O)

施提供支撑。为构建下一代云原生基础设施提供理论支撑和技术参考。值得注意的是,近年来以华为为代表的国产厂商在高性能存储领域持续投入,其推出的OceanStor Dorado系列全闪存存储服务器已广泛支持NVMe over Fabrics(NVMe-oF)、智能QoS调度及用户态I/O加速技术,并与KVM、OpenStack等虚拟化平台深度集成。这些工业实践为本文的研究提供了重要的现实参照和技术验证场景,也凸显了高性能虚拟化存储架构在国产化替代与信创生态中的战略价值[1]。

1 虚拟机I/O性能瓶颈分析

要设计高效的存储架构,必须首先理解性能瓶颈所在。典型的KVM虚拟机I/O路径(以virtio-blk为例)涉及多个环节,每个环节都可能成为性能瓶颈。

1.1虚拟化开销

这是最根本的瓶颈。当GuestOS发起一个I/O请求时,需要经历陷入、模拟、返回和处理四个阶段。具体而言,GuestOS通过virtio驱动将请求写入共享内存环(virtqueue),并触发一个hypercall通知QEMU；运行在Host上的QEMU进程被唤醒,从virtqueue中读取请求,并调用Host内核的块设备驱动来处理；Host内核完成I/O后,QEMU将结果写回virtqueue,并注入一个中断通知GuestOS;最后,GuestOS收到中断,从virtqueue中读取结果,完成本次I/O。这个过程涉及多次用户态-内核态切换、Guest-Host之间的上下文切换以及潜在的数据拷贝,这些操作消耗了大量的CPU周期,尤其在高I/O负载下,CPU开销会急剧增加,形成瓶颈。

1.2 I/O调度器冲突

现代操作系统内核通常包含多级I/O调度器。GuestOS内部有自己的I/O调度器(如CFQ、deadline),而HostOS同样有自己的调度器。这种双重调度不仅增加了延迟,还可能导致调度策略冲突。例如,Guest希望按顺序合并请求以优化磁盘寻道,但Host调度器可能为了公平性而打乱这些请求的顺序,导致物理设备无法发挥最佳性能。

1.3资源争用与隔离不足

在多租户的云环境中,多个虚拟机共享同一物理存储设备。如果缺乏有效的资源隔离和QoS保障机制,一个“吵闹的邻居”(NoisyNeighbor)可能会独占I/O带宽或产生大量I/O操作,导致其他虚拟机的I/O性能急剧下降,甚至服务不可用[2]。传统的cgroups虽然可以限制I/O带宽,但对于IOPS(每秒I/O操作数)和延迟的精细控制能力有限。

1.4 NUMA效应

在多路服务器(NUMA架构)上,CPU、内存和I/O设备分布在不同的节点上。如果虚拟机的vCPU、其使用的内存以及后端存储设备不在同一个NUMA节点上,跨节点访问会引入显著的延迟和带宽损耗。不合理的资源分配会放大这一问题。

2 高性能存储架构设计

针对上述瓶颈,本文提出一种分层解耦、软硬协同的高性能存储架构。该架构由三个核心层次构成：硬件加速层、用户态I/O引擎层和智能调度管理层。

2.1硬件加速层

本架构充分利用现代硬件特性,首先采用基于PCIe总线的NVMe协议SSD,相比传统的SATA/SASSSD,NVMe拥有更低的延迟、更高的IOPS和更优的并行性,是高性能存储的理想选择。NVMe原生支持多队列机制,能够与现代多核CPU架构良好匹配,充分发挥硬件并发能力[3]。此外,系统配置中广泛启用大页内存(HugePages),使用2MB或1GB的大页替代默认的4KB页,可以显著减少TLB缺失次数,提升内存访问效率。这对于需要频繁进行DMA操作的I/O路径尤为重要,因为每次DMA地址转换若发生TLB未命中,都会触发代价高昂的页表遍历,而大页内存有效缓解了这一问题。此外,随着国产高端存储服务器的发展,华为OceanStor系列存储设备在虚拟化环境中展现出卓越的I/O性能与可靠性。例如,华为OceanStor Dorado 8000采用自研SSD控制器与FlashLink技术,通过端到端NVMe协议栈优化、多队列并行处理及智能磨损均衡算法,显著提升了IOPS与延迟表现；同时,其内置的SmartQoS功能可实现基于租户的细粒度带宽与IOPS控制,与本文提出的QoS隔离机制高度契合。在实际部署中,将本文架构的后端存储对接至华为存储服务器(通过NVMe-oF或直连PCIe方式),可进一步释放硬件潜能,形成从虚拟机到国产高端存储的全栈高性能通路。

2.2用户态I/O引擎层

这是架构的核心,旨在绕过内核,实现极致的I/O性能。我们把存储后端驱动从内核态迁至用户态,选用Intel主导的SPDK框架。SPDK以轮询模式处理I/O事件,避免中断上下文切换开销,还实现真正零拷贝,让I/O数据在应用程序内存缓冲区和NVMe设备间直接传输,无需内核缓冲区中转。为将高性能延伸到虚拟机内,采用vhost-user作通信桥梁,它基于Unix域套接字,让用户态程序直接扮演vhost角色,使GuestOS的virtio-blk驱动可与用户态SPDK应用直接通信,绕过QEMU和Host内核块设备栈。对依赖Host文件系统的场景,则利用Linux5.x的io_uring接口,其高效异步I/O机制可减少系统调用,作为SPDK的补充。

2.3智能调度与管理层

高性能不仅意味着峰值性能,更意味着在复杂多变的环境中保持稳定和公平。为此,在架构中集成了智能调度与管理机制。基于SPDK的框架,我们为每个虚拟机分配独立的I/O队列,并通过令牌桶或时间片轮转等算法,精确控制每个VM的IOPS、带宽上限和最低保障,从而有效防止“NoisyNeighbor”问题。在部署层面,我们强调NUMA感知调度：通过libvirt等管理工具,将VM的vCPU、其内存以及SPDK应用绑定到同一个NUMA节点上,同时确保SPDK使用的内存池也来自该节点,从而最大化本地内存访问,最小化跨NUMA节点的延迟[4]。最后,整个架构对外依然提供标准的virtio-blk或virtio-scsi接口,保证了GuestOS的兼容性和透明性,上层应用无需任何修改即可享受底层的性能提升,实现了高性能与易用性的统一。

3 性能评估

3.1实验环境与配置

硬件平台：双路IntelXeonSilver4210CPU(2.2GHz,10corespersocket),128GBRAM,2块IntelOptaneP4800XNVMeSSD。

软件环境：HostOS:Ubuntu22.04LTS(Kernel5.15)

Hypervisor:QEMU6.2+KVM

GuestOS:Ubuntu22.04LTS

对比方案A(Baseline)：传统virtio-blk+QEMU+HostKernelBlockLayer。

对比方案B(Proposed)：virtio-blk+vhost-user+SPDK(用户态NVMe后端)。

测试工具：FIO(FlexibleI/OTester),用于生成各种I/O负载(顺序/随机,读/写,不同队列深度和块大小)。

3.2评估指标

吞吐量(Throughput)：单位时间内完成的I/O数据量(MB/s)。

IOPS：每秒完成的I/O操作数。

延迟(Latency)：包括平均延迟、99%尾部延迟(p99latency),这是衡量服务质量的关键指标。

CPU开销(CPUUtilization)：完成相同I/O负载所消耗的HostCPU百分比。

3.3实验结果与分析

3.3.1实验一：吞吐量与IOPS对比

在4K随机写、队列深度(iodepth)为64的场景下：Baseline方案的IOPS约为85K,吞吐量约为332MB/s。Proposed方案的IOPS达到了惊人的250K,吞吐量约为976MB/s。性能提升接近300%,这充分证明了绕过内核和QEMU带来的巨大收益。

3.3.2实验二：延迟分析

在相同的4K随机写负载下：Baseline方案的平均延迟为750μs,p99延迟高达2.1ms。Proposed方案的平均延迟降至220μs,p99延迟仅为630μs。延迟的大幅降低,特别是尾部延迟的改善,对于延迟敏感型应用至关重要。

3.3.3实验三：CPU开销

在达到各自最大IOPS时：Baseline方案消耗了约70%的单个物理核心CPU资源。Proposed方案仅消耗了约25%的单个物理核心CPU资源。这意味着在相同的硬件上,可以承载更多的虚拟机或为其他计算任务释放宝贵的CPU资源。

3.3.4实验四：QoS隔离效果

在同一台宿主机上运行两个虚拟机(VM1和VM2)。对VM1施加极限的4K随机写负载,同时监控VM2在执行轻量级4K随机读时的延迟。在Baseline方案下,VM2的p99延迟从正常的300μs飙升至超过5ms。在启用了QoS的Proposed方案下,即使VM1满载,VM2的p99延迟依然稳定在400μs以内,隔离效果显著。

4 结语

本文针对虚拟机I/O性能瓶颈,设计出融合硬件加速、用户态I/O引擎与智能调度的高性能存储架构。该架构把存储后端迁至用户态,借vhost-user协议打通与虚拟机直通链路,绕过传统性能瓶颈。性能评估显示,在吞吐量等关键指标上提升显著,且QoS机制保障多租户服务质量与资源隔离,为高性能云基建提供可行路径。不过,该领域仍有研究空间,如集成持久内存、探索计算存储管理调度、AI驱动智能调度及推动技术标准化等。未来工作中,将进一步探索与国产存储生态的深度融合,特别是将本架构适配于华为OceanStor等企业级存储服务器平台,验证其在真实金融、政务等高可靠场景下的性能与稳定性,推动高性能虚拟化存储技术的自主可控与产业化落地。

[参考文献]

[1]卢明.基于HCI环境下虚拟机密度与存储性能的平衡策略研究[C]//重庆市大数据和人工智能产业协会,重庆建筑编辑部,重庆市建筑协会.智慧建筑与智能经济建设学术研讨会论文集(一).中节能大地(杭州)环境修复有限公司,2025:796-799.

[2]顾武雄.管理虚拟机存储区策略[J].网络安全和信息化,2021,(04):71-72.

[3]赖策,祝元仲.虚拟机基于不同存储池模式下的磁盘性能测试分析[J].轻工科技,2020,36(07):90-91.

[4]谢慧,张澎,王鲁达.虚拟机网络存储性能优化技术研究[J].河南科技,2017,(21):41-42.