解决方案简介

基于第四代英特尔® 至强® 可扩展处理器的 Curve 云原生高性能分布式存储系统

網易 NetEase

“存储等领域的基础软件对于数字化转型非常关键，伴随着数据的快速增长，用户需要一款性能更高、可用性和可靠性更好、自治能力更强的分布式存储系统，Curve 代表了网易数帆在存储软件方面的强大创新能力，持续助力软件定义基础设施生态的持续繁荣。在采用第四代英特尔® 至强® 可扩展处理器之后，我们进一步增强了 Curve 的性能表现，能够在更多高性能存储场景中扮演重要角色。”

—张晓龙
网易杭研基础平台总监

挑战

分布式存储系统已成为企业推进数字化转型的关键策略之一。相比传统架构，分布式存储系统具有更高的并发和更低的时延，可实现敏捷的扩展，帮助企业提升投资回报。但随着数据的快速增长，分布式存储系统在性能、TCO 等方面也遇到了多重挑战：

- **创新应用、海量数据带来严峻性能压力**

  用户增长与业务创新使得分布式存储系统必须具备更高的性能水平，以在特定的服务级别协议（SLA）标准下处理数据。这对存储系统的吞吐能力和时延造成了巨大压力，使得存储系统的性能压力远超传统业务。

- **存储集群规模增长可能导致 IO 性能抖动**

  在大规模的存储集群中，有可能出现单个节点或磁盘异常（慢盘坏盘、节点超载等），导致 IO 请求时延飙高，进而出现 IO 抖动。随着集群规模的增长，IO 抖动的问题可能变得更加严峻。

- **分布式存储系统的规模带来较高的 TCO 挑战**

  为了提升系统吞吐量并降低时延，面向关键业务的分布式存储系统通常需要较高的集群规模，这会直接带来高额的采购成本。此外，存储系统的服务器集群在空间占用、运维、供电散热等方面，也会带来较高的成本压力，因此提升单节点的性能密度成为重要发展方向。

解决方案概述

在人工智能（AI）、5G、边缘计算等数字化创新技术的驱动下，数据正在呈现爆发式增长的趋势。为挖掘海量数据背后的价值，高性能存储系统的需求急剧增加。在此背景下，传统存储系统逐渐暴露出性能不足、运维管理难、存储成本高、资源弹性不足、资源利用率低等问题，无法充分满足用户在 AI 训练、大数据、混合云等场景的存储需求。

为助力用户构建高性能存储系统、释放数据价值，网易主导自研了云原生高性能分布式存储系统 Curve，为企业提供一个高性能、低时延的存储底座，基于该存储底座，企业可以打造适用于不同应用场景的存储系统，如块存储、对象存储、云原生数据库等。网易与英特尔联合推出了第四代英特尔® 至强® 可扩展处理器的 Curve 方案，该方案借助高性能处理器与高性能数据库，实现了较高的存储性能表现。

Curve 云原生高性能分布式存储系统

Curve 是网易数帆主导自研的一款高性能、易运维、云原生的开源分布式存储系统，目前支持文件存储（CurveFS）和块存储（CurveBS）。Curve 可应用于主流的云原生基础设施平台，亦可作为云存储中间件使用 S3 兼容的对象存储作为数据存储引擎，为公有云用户提供高性价比的共享文件存储。

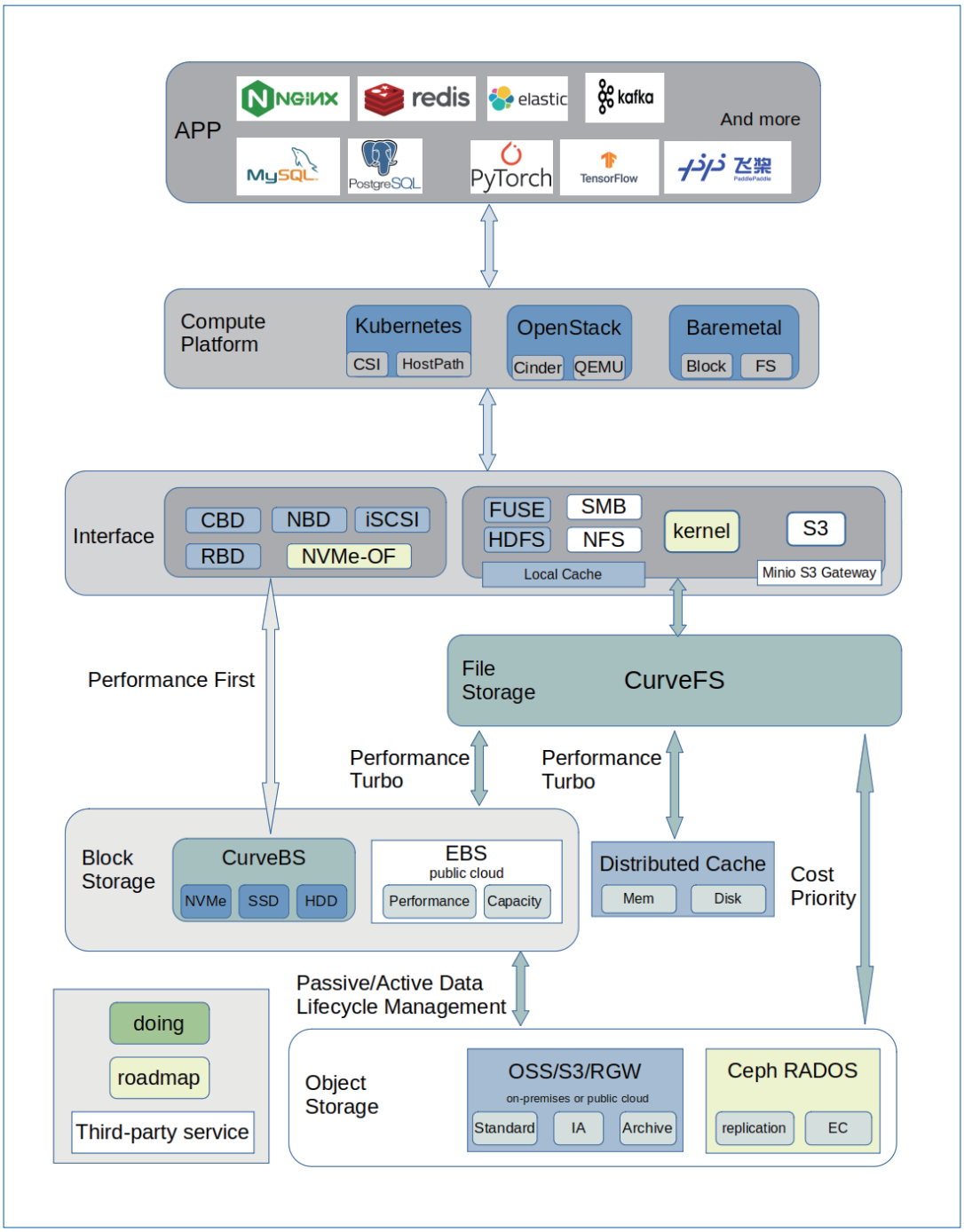

图 1. Curve 云原生高性能分布式存储系统架构

CurveBS 的核心应用场景主要包括：虚拟机/容器的性能型、混合型、容量型云盘或持久化卷，以及物理机的远程存储盘；高性能存算分离架构，以及基于 RDMA+SPDK 的高性能低时延架构，该架构是支撑 MySQL、kafka 等各类数据库、中间件的存算分离部署架构，有助于提升实例交付效率和资源利用率。CurveFS 的核心应用场景主要包括：AI 训练场景下的高性价比存储；大数据场景下的冷热数据自动化分层存储；公有云上高性价比的共享文件存储；以及混合云存储。

Curve 主要具有三大设计特点：高性能、高可用和自治。

高性能

Curve 参考了业界的存储系统，采用先进、高效的开源技术，设计了一个新架构实现高性能低时延的核心能力，采用高性能的 RPC 框架来保障网络数据流的高性能和低时延，基于 Raft协议实现多副本一致性下的低时延，并针对 Raft 协议的快照实现进一步优化。磁盘 IO 方面，Curve 通过更细粒度的地址空间的 hash 减少 IO 碰撞，增加 IO 并发度，并采用 chunkfilepool 减小 IO 放大倍数，从而有效发挥硬件的性能。

高可用

Curve 被设计为核心组件均能容忍部分实例失败而不影响整个集群的可用性。无论是单台存储故障，还是系统扩容，Curve 的客户端 IO 都不会受到丝毫的影响，插拔硬盘、服务进程中断这些常见异常，IO 抖动也很小。此外，故障恢复过程对上层 IO 也不会造成明显影响。

自治

Curve 实现了一键部署、一键升级，运维只要很少的人工干预，并基于 Promethues 和 Grafana 等开源技术打造了全面的度量标准和告警体系。

采用第四代英特尔® 至强® 可扩展处理器提升性能表现

高性能存储系统依赖于高性能的硬件，尤其是 CPU 的性能更是堪称关键。为避免出现性能瓶颈，网易推荐采用搭载第四代英特尔® 至强® 可扩展处理器的服务器，以搭建 Curve 存储系统，满足多种场景的存储需求。

第四代英特尔® 至强® 可扩展处理器通过创新架构增加了每个时钟周期的指令，每个插槽多达 60 个核心，支持 8 通道 DDR5 内存，有效提升了内存带宽与速度，并通过 PCIe 5.0（80 个通道）实现了更高的 PCIe 带宽提升。第四代英特尔® 至强® 可扩展处理器提供了出色性能和安全性，可根据用户的业务需求进行扩展。借助内置的加速器，用户可以在 AI、分析、云和微服务、网络、数据库、存储等类型的工作负载中获得优化的性能。通过与强大的生态系统相结合，第四代英特尔® 至强® 可扩展处理器能够帮助用户构建更加高效、安全的基础设施。

第四代英特尔® 至强® 可扩展处理器内置了英特尔® In-Memory Analytics Accelerator(英特尔® IAA)、英特尔® 高级矩阵扩展（英特尔® AMX）、英特尔® Data Streaming Accelerator（英特尔® DSA）、英特尔® QAT 等高级硬件能力，能够加速 AI、数据分析、数据加解密等场景下的处理能力多种高级硬件特性，满足用户的多样化算力需求。

为了测试搭载第四代英特尔® 至强® 可扩展处理器的 Curve 块存储系统性能表现，网易进行了测试。测试使用 3 台 Server 节点和 1 台 Client 节点，配置如表 1 所示，并在同一配置上部署了 Ceph，进行了性能对比。

	Server	Client
处理器	2*英特尔® 至强® 金牌 6438Y+ 处理器	2*英特尔® 至强® 金牌 6430 处理器
内存	512 GB 总内存 (16x32 GB DDR5)	512 GB 总内存 (16x32 GB DDR5)
数据盘	4*2TB P4510 NVMe 固态盘（Curve/Ceph storage用）	4*2TB P4510 NVMe 固态盘
操作系统盘	1*480GB SATA 固态盘	1*480GB SATA 固态盘
操作系统	Debian GNU/Linux 11 (bullseye) Kernel: 5.10.0-26-amd64	

表 1. 节点硬件配置

测试工具采用了 fio 基准测试，对比了同等配置下 Curve 块存储与 Ceph 存储的 4K 随机性能。测试数据如图 2 所示，Curve 块存储 4K 随机写性能比 Ceph 高 1.07 倍，在混合读写场景（读写比7:3）下，性能比 Ceph 高 1.47 倍，随机读性能高 1.55 倍[1]。

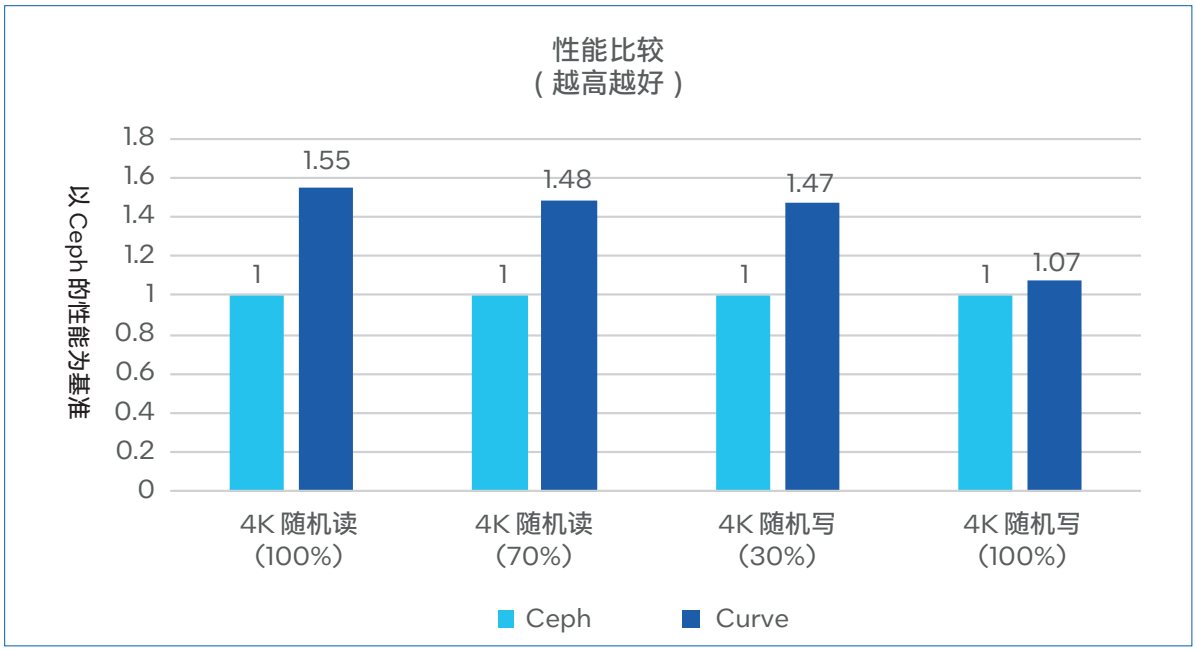

图 2. Ceph 与 Curve 性能比较

在 Curve 块存储中，网络传输中服务端处理可能会占用内存带宽与 CPU 资源，影响系统性能。通过在 Curve 块存储中启用远程直接数据存取（Remote Direct Memory Access, RDMA）技术，有助于进一步提升存储性能：RDMA 技术可以让计算机直接访问远程计算机的内存，而无需在本地和远程计算机之间进行数据复制，从而提高网络通信的性能和效率。

网易数帆测试了启用 RDMA 后 Curve 的性能，测试数据如图 3 所示，启用 RDMA 后 Curve 性能最高提升到 1.86 倍，同时时延也有大幅改进[2]。

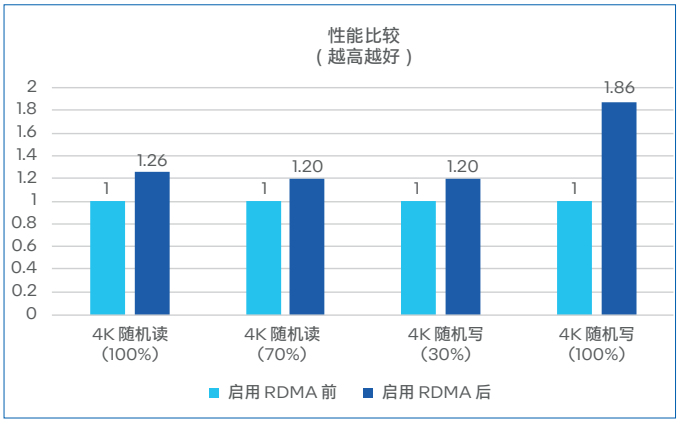

图 3. 启用 RDMA 前后的 Curve 性能比较

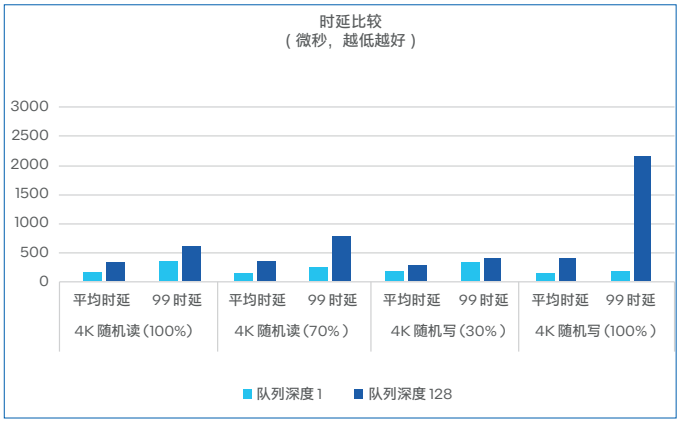

图 4. 启用 RDMA 后不同队列深度的读写时延

[1,2] 数据援引自网易数帆截止 2023 年 11 月的内部测试结果。测试配置 – Server：2*英特尔® 至强® 金牌6438Y+ 处理器，512 GB 总内存 (16x32 GB DDR5)，4*2TB P4510 NVMe 固态盘（Curve/Ceph storage用），1*480GB SATA 固态盘，Debian GNU/Linux 11 (bullseye)，Kernel 5.10.0-26-amd64；Client：2*英特尔® 至强® 金牌6430 处理器，512 GB 总内存 (16x32 GB DDR5)，4*2TB P4510 NVMe 固态盘，1*480GB SATA 固态盘，Debian GNU/Linux 11 (bullseye)，Kernel 5.10.0-26-amd64。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。

收益

基于第四代英特尔® 至强® 可扩展处理器的 Curve 高性能分布式存储方案在吞吐、时延、TCO 等方面具备出色优势，能够为金融、互联网、交通、能源等各行业提供高效的云原生存储解决方案。

- **应对超高性能场景的存储需求，实现快速响应：**得益于第四代英特尔® 至强® 可扩展处理器的强大性能，以及 Curve 在性能方面的优化，该方案实现了卓越的性能表现，能够有效应对超大并发量业务在吞吐、时延方面的挑战，提升业务服务能力。
- **降低大规模集群下的性能抖动：**Curve 采用了优化数据均衡工具、Backport 下线主动发现机制、优化 Peering 耗时、恢复速度控制、实现异步恢复等方式，并配置了数量众多的监控告警，以便及时发现问题节点、进行故障处理，从而降低存储异常导致 IO 抖动的问题。
- **有效控制 TCO，提高投资回报**：基于第四代英特尔® 至强® 可扩展处理器的 Curve 提升了单节点的性能密度，有助于降低集群构建、扩展带来的成本压力。
- **提升可用性**：Curve 采用多种数据高可用机制，网易的应用实践显示，系统上线以来，未出现数据不一致和丢数据的情况，没有发生过重大故障，实现了卓越的数据可靠性和服务可用性。

展望

在现有成果的基础上，Curve 将会持续优化性能表现，包括支持云原生数据库、为 K8s 集群提供存储服务，并积极挖掘第四代英特尔® 至强® 可扩展处理器、英特尔® 以太网适配器等新一代硬件的性能潜力，探索 io_uring、QUIC、英特尔® SPDK、RAMA 等技术的应用，将其反馈给开源社区，为客户提供更高性能的存储系统。

英特尔将与网易继续深度合作，通过联合技术创新、产品验证、场景化落地等方式，持续释放第四代英特尔® 至强® 可扩展处理器在性能和可扩展性等方面的优势，加速软硬件的融合，高效地利用计算、存储、网络等资源，充分发挥云原生高性能分布式存储系统的优势，助力企业数字化创新。

关于网易

网易（NASDAQ: NTES；HKEX：9999）作为全球领先的数字内容与互联网技术公司，长期投入引擎、人工智能、区块链、数字孪生、云游戏等技术领域，致力于实现“网聚人的力量，以科技创新缔造美好生活”的使命愿景。网易不仅是全球领先的在线游戏开发与发行公司，还拥有中国领先的在线音乐平台、智能学习平台、自营品质电商品牌、资讯传媒平台、电子邮件及企业服务，覆盖全中国超过 10 亿用户，影响全球 200 多个国家和地区。

关于英特尔

英特尔（NASDAQ: INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。如需了解英特尔创新的更多信息，请访问英特尔中国新闻中心 newsroom.intel.cn 以及官方网站 intel.cn。

intel®

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见 www.Intel.com/PerformanceIndex

性能测试结果基于配置信息中显示的日期进行测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

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。