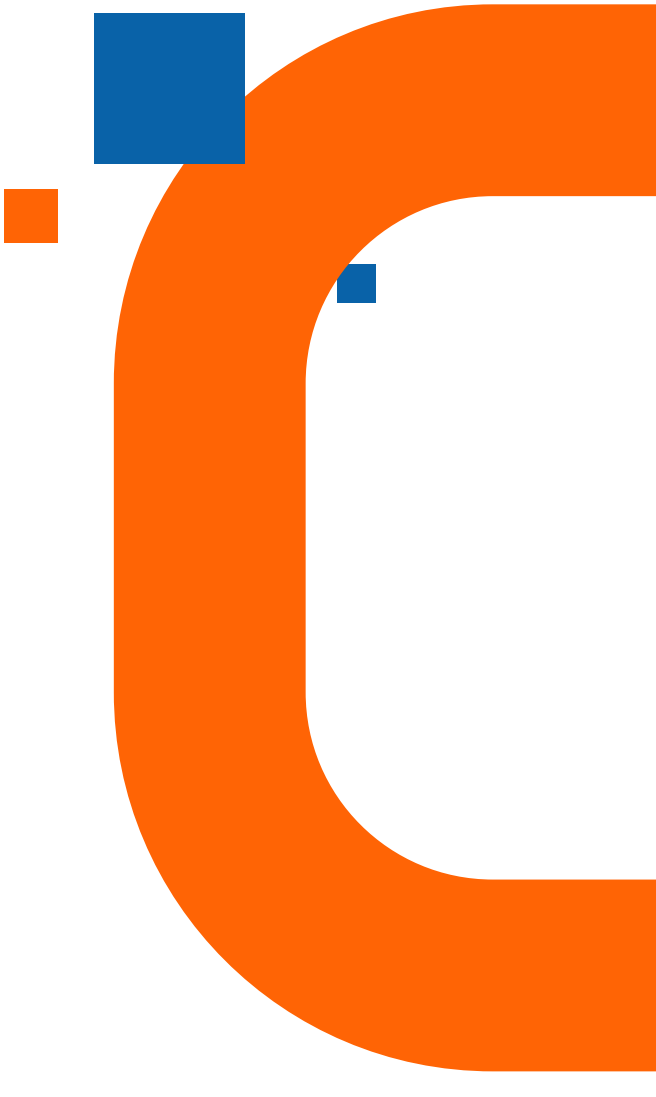
ExponTech
华瑞指数云

WiDE分布式引擎与英特尔®新一代平台

联合解决方案技术白皮书



WiDE™
无量数据引擎

华瑞指数云科技有限公司
英特尔（中国）有限公司

Digitize for a better future.

让未来更有数。

华瑞指数云WiDE分布式引擎与英特尔®新一代平台

联合解决方案技术白皮书

1 新一代分布式存储解决方案

1.1 执行综述

华瑞指数云自主研发的新一代分布式存储引擎，全称无量数据引擎（英文名WiDE，下文简称WiDE或者WiDE引擎），吸纳了学术界和工业界过去十余年来在分布式系统，存储技术，数据处理技术，系统编程模型上的最新进展，采用全新设计的分布式系统软件架构和分布式环境上的端到端无锁零拷贝编程模型从零开始自主研发，其关键设计目标和特征包括：

一、完全匹配云原生环境,支持私有云、公有云、混合多云，边缘环境上的部署，提供无缝一致的数据体验；

二、解决当前的分布式存储架构无法支持数据库，高性能数据分析等高端业务场景的问题，读写延迟降至微秒级别，单卷随机读写IO超百万IOPS；

三、能够充分发挥新一代存储介质，新一代网络技术的优势，充分利用这些新兴硬件的峰值性能；

四、全局元数据管理，从IO调度层面实现数据跨多池，跨多云和边缘的智能调度和自由流动，实现更好的系统可靠性，实现全场景数据的统一管理。

华瑞指数云的目标是基于完全自主研发的新一代分布式存储引擎WiDE作为统一数据基座，向上提供极速块设备接口(NVMe-oF, vhost),经典存储接口(对象存储S3、文件系统接口POSIX、NFS、SMB、块设备接口iSCSI和容器接口CSI)以及数据接口(大数据访问协议HDFS和OLAP SQL接口等)，从而实现在一个统一的数据平台上存储企业的全场景数据，并且通过存数一体化架构，提供统一数据管理和多样化的数据组织和分析能力，实现企业数据的全生命周期管理。无量数据引擎WiDE将推动软件定义存储SDS的整体架构换代，进入SDS 2.0时代。

WDS是基于无量数据引擎WiDE的极速分布式块存储产品，在配置全SSD介质的普通服务器集群环境上，可以提供单卷数百万IOPS和单路百微秒以内稳定时延，可以通过NVMe-oF、vhost、iSCSI和CSI等标准接口，与OpenStack、VMware、Kubernetes和物理机等常见的企业IT环境进行对接，满足企业关键应用极高的性能要求，同时在系统可靠性方面也有许多创新设计满足核心业务的高可靠需求，面向大型数据库、核心业务云化部署、HPDA、AI/ML等业务场景，在许多情况下都可以直接替换AFA全闪阵列，还具备可以水平扩展、软件定义、适配云原生环境、提供统一存储平台和统一数据管理等AFA全闪阵列所不具备的优势。

英特尔® 傲腾™ (Optane)技术是一项全新的非易失性内存技术。其高速和高密度特性消除了处理瓶颈，改进了诸如大数据、高性能计算 (HPC)、虚拟化、存储、云和游戏等要求严格的应用的性能。

英特尔® 傲腾™ 固态盘 P5800X 是一款性能出众的面向数据中心的高速固态盘产品，同时它还具备出色的耐用性，并支持 PCIe Gen4，非常适合进行热数据层缓存以及为各种类型的存储解决方案加速。凭借业界领先的低延迟、高质量服务 (QoS)、高速吞吐和高耐用性的组合，英特尔® 傲腾™ 固态硬盘 P5800X 可加速访问企业和云服务提供商所需的大型的、复杂的数据集，以运行其要求严苛的工作负载。重要的是，与其他固态硬盘技术不同，英特尔® 傲腾™ 固态硬盘可以同时进行读取和写入，而不会造成性能的降低。

华瑞指数云和英特尔实验室进行了一系列的联合测试，验证在英特尔®傲腾™固态盘以及全新架构的无量分布式存储引擎WiDE的联合作用下，新一代分布式存储系统所能达到的性能水平和系统能力。整个验证测试在英特尔实验室环境中进行，一共采用了6台标准服务器，其中3台用于组建分布式存储集群，3台作为计算节点发起对存储系统的功能和性能测试。

硬件配置

每台服务器配置（3台）：

CPU：Intel(R) Xeon(R) Gold 6240Y CPU @ 2.6Hz * 2

磁盘：英特尔®傲腾™固态盘P5800X 1.6TB * 4

网卡：英特尔®以太网网络适配器E810-2CQDA2（单口100GE）网卡 * 2

软件配置

在相同的硬件环境上以及标准的CentOS 7.9操作系统上，分别部署Ceph Nautilus版本和华瑞指数云基于WiDE引擎的WDS 2.0产品进行测试，其中的Ceph Nautilus版本已经经过了全面的系统调优，确保其能发挥Ceph软件栈的更优性能。

测试方法

针对Ceph Nautilus版本和华瑞指数云的WDS 2.0产品组成的分布式存储集群，分别使用FIO和Sysbench执行存储基准性能测试和Mysql数据库基准性能测试，比较两个不同的分布式存储软件在相同的硬件环境上的性能。

针对WDS 2.0产品，专门测试其对新一代硬件介质的峰值性能发挥能力以及分布式集群的线性扩展能力。

测试结果表明，华瑞指数云基于WiDE引擎的WDS 2.0产品可以在使用RDMA互联的分布式系统环境中，能够充分的发挥出英特尔®傲腾™固态盘以及网卡的峰值性能，在3台服务器节点，100G RDMA网络，一共12块英特尔®傲腾™固态盘P5800X的集群环境上，可以达到4k随机写324万IOPS（2副本）和244万IOPS(3副本)，4k随机读692万IOPS的性能，在达到如此惊人的IOPS的同时，还能保持稳定的500us以内的时延。测试集群的主要瓶颈是在网卡带宽上面，如果进一步扩充每个节点的网络带宽，将可以轻松达到3节点存储集群超1000万IOPS的性能。

与全面调优后的Ceph Nautilus版本相比，在性能上实现了数量级的惊人提升，无论是单卷性能还是集群性能均提升高达10到30倍以上，与此同时IO时延缩短为十分之一，从Ceph Nautilus版本的5ms左右大幅降低到WDS 2.0版本的500us左右。

在Mysql数据库的Sysbench测试中，验证访问16个表的每秒事务处理性能(TPS)，WDS 2.0版本可以在稳定的5ms以内时延，提供高达20000以上的TPS，而Ceph Nautilus版本没有任何办法可以提供5ms以内的稳定时延，将时延要求放宽到15ms以后，也只能提供数千TPS（对比数据结论详见章节3.2）。

测试结果也充分说明了WiDE分布式引擎已经把分布式软件栈的效率优化到了峰值，可以充分发挥所有新型硬件的峰值能力，甚至能将英特尔®傲腾™固态盘P5800X这样的目前市面上运行速度标杆式的SSD的能力发挥到峰值。随着硬件本身的性能提升或者硬件节点和数量的增加，端到端的分布式存储集群性能也可以近乎无损的线性增长，可以轻松达到上亿IOPS。

1.2 存储新介质及网络的技术发展

1.2.1 NVMe SSD 和英特尔® 3D XPoint

客户面临数字化转型的时候，都绕不开两个维度，即应用和数据。数据将成为改变传统的度量依据：流程的优化、产品的迭代、商业模式的创新，都将依据数据来驱动，企业需要基于数据来做实时的决策，实现高效管理。

随着海量数据分析和实时数据分析发展，越来越多的数据需要高效存储，同时还需要高效的进行数据读写处理，这给存储系统带来极大的压力。存储系统正在努力跟上新兴计算和网络技术的发展速度，以处理当前的数据密集型工作负载。存储介质反而成为了阻碍 IT创新和新生业务增长的瓶颈。尽管在过去一段时间，NAND闪存SSD被认为是提升存储性能的可行手段，但随着NAND闪存SSD技术的不断发展，从SLC到MLC、TLC，再到即将主流的QLC SSD，容量不断增长，但单位容量产生的IOPS越来越低，尤其是写性能的“踌躇不前”已经越来越不能满足应用的需求，尤其是在写操作占比较高的环境下。

采用3D XPoint技术的英特尔® 傲腾™ 产品被视为一种新的可行解决方案，已经在云计算领域开始大规模部署。Optane的特点在于，在每GB容量内，能提供比NAND闪存SSD高出一个甚至多个数量级的性能，尤其是写性能，且延迟显著降低。

相比于NAND闪存，采用3D XPoint技术的英特尔® 傲腾™ 产品速度可超出1000倍左右，同时延迟更低，寿命更持久。相比DRAM内存成本只有一半，存储密度却高出10倍，堪称NAND闪存发明以来最具突破性的一项存储技术。

1.2.2 RDMA网络

伴随着AI和高性能实时数据分析的热潮以及各种高性能新型存储介质的发展，数据需要以更快的速度在网络上传输，对网络通信时延提出了更高的要求，传统的TCP/IP协议栈在接收/发送报文时，内核需要做多次上下文切换，需要多次的数据拷贝，许多工作需要依赖主机CPU来完成，导致CPU持续高负载，网络协议栈处理需要十毫秒级的时延。在要求微秒级时延的高性能分布式系统中，网络协议栈的高时延成为最明显的瓶颈之一。

RDMA(Remote Direct Memory Access远程直接内存访问)是一种高性能无损网络技术，其内核旁路机制允许应用与网卡之间直接进行数据读写，不需要在内核态与用户态之间做上下文切换，其内存零拷贝机制，允许接收端直接从发送端的内存读取数据，不需要CPU的参与，极大的减少了 CPU 的负担。RDMA技术使网络协议栈时延下降到了10us以内，能充分发挥新兴的高性能存储介质的极低时延的优势，因此在最新的NVMe接口协议中，RDMA成为主流的网络通信协议栈，在AI、实时数据分析、全闪存储、数据库等追求极致性能的大潮中，RDMA替换TCP/IP已是大势所趋。

目前有三种主要的RDMA网络实现技术，分别是Infiniband、RoCE(RDMA over Converged Ether-net)、iWARP：

InfiniBand：设计之初就考虑了RDMA，从硬件级别保证可靠传输，提供更高的带宽和更低的时延，但是成本高，需要支持InfiniBand的专门网卡和专门交换机。

iWARP：基于TCP实现RDMA网络，相比RoCE、iWARP的大量TCP连接会占用大量的内存资源，对系统规格要求更高，可以使用普通的以太网交换机，但是需要支持iWARP的网卡。

RoCE：基于UDP实现RDMA，消耗的资源比iWARP少，可以使用普通的以太网交换机，但是需要支持RoCE的网卡，由于可以使用现有的以太网交换机，同时新一代网卡已经缺省支持RoCE, RoCE v2协议是目前实现高性能分布式存储系统的主流选择。

1.3 分布式存储软件的技术发展

分布式存储相关的技术和产品在2005年前后问世，初期主要用于互联网等行业需要海量数据存储，高访问并发量，但是对数据的一致性要求并不苛刻等业务场景。随着云计算的普及以及数据量的急剧增加，在2015年左右分布式存储开始在企业级市场等到大量应用，尤其是在云计算环境中，存储资源需要池化和弹性扩容，分布式存储成为当然之选。由于分布式架构和软件定义带来的诸多优势，分布式存储也被认为是存储系统的未来趋势，将向传统的集中式存储阵列发起挑战。这个时期我们可以称之为软件定义存储的1.0时代，归结来看有如下核心特点：

- 基于标准硬件节点的Scale-out横向扩展，存储集群的容量和性能都可以随着扩容线性增长；
- 软硬件解耦，可以使用通用服务器作为存储服务器，解除硬件绑定；
- 分布式架构决定了存储系统可以按照资源池化管理，系统可以根据实际的容量性能需求进行部署，同时适应需求的变化进行快速的部署扩展。

然而，以Ceph为代表的第一代分布式存储软件诞生于18年以前，那时候还是一个较为初级的企业信息化阶段，硬件是以慢速的机械硬盘和高时延低带宽的TCP/IP网络为主，当时也远远没有办法预期到现在的数字化和智能化浪潮，这就导致其系统架构设计难以匹配现在的新兴硬件和新兴需求，在应用场景上先天不足，具有很大的局限性。第一代分布式存储虽然优点很明显，但是性能低、时延大、性能的稳定性和系统可靠性不足，无法匹配企业关键业务的需求，因而在过去的十多年来只能被应用于中低端业务场景，被贴上了“中低端存储”的标签。

与10多年前相比，现在的存储硬件、网络以及相关的技术方案已经发生了很多的变化。譬如在介质方面，存储已经实现了大规模的从机械硬盘向SSD固态硬盘的过渡，由此带来了硬件层面的超高IOPS和超低时延，新兴介质的性能相比磁介质，性能提升了1000倍，访问时延下降了数百倍，达到了10us级；网络带宽由过去的10M和100M已经提升到现在的10G和100G, 提升了1000倍，相应的，随着RDMA等新兴技术的出现，网络访问时延已经下降数百倍，达到了10~20us的水平。然而，目前市场上主流的第一代分布式存储IO栈还是采用的15~18年以前的系统架构，处理单个IO的时延会达到几百us之多，比今天的网络和介质要慢一个数量级，软件IO栈已经成为了整个数据存储系统的主要瓶颈。

为了适应高速IO处理的需求，过去十多年来学术界和工业界在本领域也提出了许多新想法，新技术，比如内存零拷贝、远程直接地址访问RDMA、无锁零中断IO处理、内核旁路(Bypass Kernel)、全用户态、RTC(Run to Completion)IO处理模型和异步轮询等。英特尔公司为了配合新的网络和存储介质的需要，主导推广了DPDK、SPDK开发套件和社区。实际上在商业厂家主导的集中式全闪阵列产品中，已经部分的应用了前述的这些技术，以达到发挥全闪介质和全闪阵列系统的极致性能的目的。而由于分布式系统整体的系统架构设计的复杂性，当前的分布式存储软件栈还没有全面的、成熟的应用上述技术，与集中式存储相比，还需要解决分布式系统的复制状态机、一致性保障、容错及故障恢复时的效率问题，因此在系统总体性能和可靠性上面还难以与集中式的全闪阵列产品相比。

在市场需求和技术发展的驱动下，以华瑞指数云无量数据引擎WiDE为代表的，采用全新系统架构和IO处理模型的新一代分布式存储系统软件应运而生了，它能够充分的压榨出硬件的性能，高效利用系统资源，在相同的硬件环境上仅仅通过软件技术的创新就能实现十多倍的性能提升，企业在硬件上的投资得以发挥最大效应。此外无量分布式存储引擎WiDE还针对云计算环境和企业数据管理的新要求，在系统架构上面做了许多重大创新，以一套核心数据引擎，支持许多种不同的存储协议和数据接口，同时在全局元数据管理能力的加持下，实现数据跨越多池做IO调度，跨越多云和边缘进行流动，进行统一数据管理。

2 解决方案参考架构

2.1 英特尔® 傲腾™ 固态盘P5800X介绍

英特尔®傲腾™固态盘P5800X是英特尔首款PCIe 4.0固态盘，也是目前市面上运行速度标杆性的数据中心固态盘。凭借业界领先的低延迟、高质量服务（QoS）、高速吞吐和高耐用性的组合，英特尔® 傲腾™ 固态硬盘 P5800X 可加速访问企业和云服务提供商所需的大型、复杂的数据集，以运行其要求严苛的工作负载。最重要的是，与其他固态硬盘技术不同，英特尔® 傲腾™ 固态硬盘可以同时进行读取和写入，而不会造成性能的降低。

英特尔®傲腾™固态盘P5800X采用2.5英寸U.2形态，容量可选400GB、800GB、1.6TB、3.2TB，支持PCIe 4.0，持续读写性能最高7.2GB/s、6.2GB/s(是上代的3倍)，4K随机读写性能最高都是150万IOPS，4K 70/30混合随机性能最高180万IOPS(是上代的3倍)，512Bytes随机读取性能高达460万IOPS，4K随机读取延迟低于6微秒、随机读写延迟低于25微秒(比上代提升40%)。

除了读写性能指标远超NAND闪存固态盘，英特尔®傲腾™固态盘P5800X的寿命也异常惊人，达到了不可思议的每天100次全盘写入，比上代的60次增加了67%。英特尔® 傲腾™ 固态盘P5800X系列填补了关键的存储性能空白，使这些固态硬盘成为高速缓存或热数据分层的理想选择。

2.2 英特尔®QAT介绍

英特尔®QAT是英特尔®针对网络安全和数据存储推出的一个硬件加速技术。英特尔®QAT专注数据安全和压缩加速，助力应用程序和平台的性能提升，在网络安全应用方面，英特尔® QAT支持多种对称数据加密（如 AES）、非对称公钥加密（如 RSA、椭圆曲线等）和数据完整性（SHA1/2/3 等）算法，加速数据的加解密和数字签名等操作。

英特尔®QAT还具备强大的压缩加速能力，提供了由QAT加速的同步压缩API，支持无状态并发压缩/解压模式、基于QAT异步API的流处理模式、线程安全压缩API，以及基于物理连续地址内存的零拷贝模式，能够将多个小数据压缩/解压请求整合到一个QAT硬件请求中，以达到提高吞吐量和降低CPU使用率的目的。

2.3 华瑞指数云无量分布式存储引擎WiDE介绍

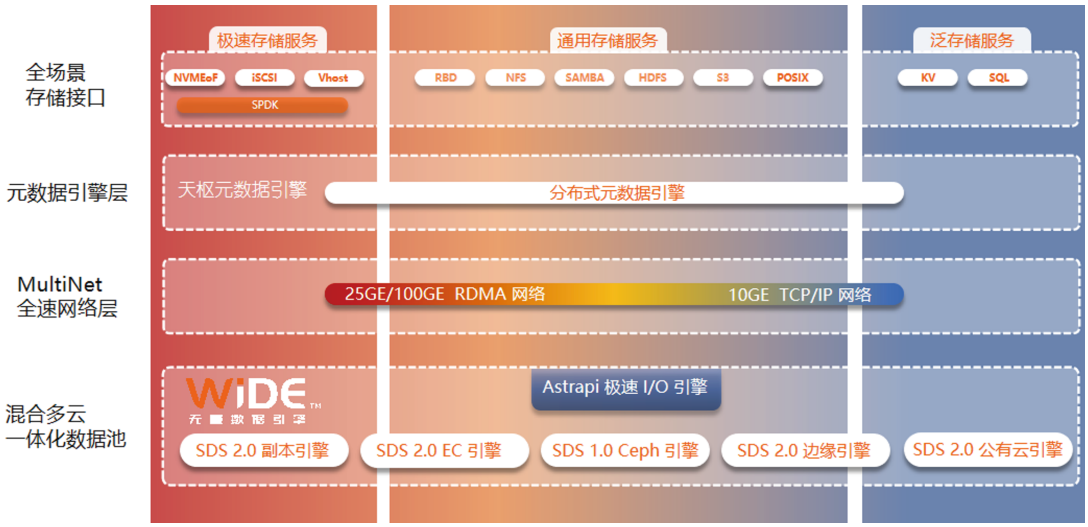


图1 华瑞指数云 SDS 2.0总体架构图

WiDE是华瑞指数云从零开始自主设计和研发的新一代分布式存储引擎，中文的名字为“无量数据引擎”，意思是可以存储无边无量的数据。就像引擎的名字一样WiDE引擎致力于解决全场景的存储需求，不但拥有全面的块、对象、文件等经典存储的常用协议，另外还会具备KV和SQL等泛存储的能力，真正做到一套存储引擎可以解决企业对存储的所有需求。华瑞指数云以WiDE作为统一的数据存储引擎，可以在上面不断研发各种分布式存储产品形态，比如WDS极速块存储，WFS极速文件存储等，并可组合成为一个可以存储全场景数据的统一存储平台。

图1是华瑞指数云以WiDE引擎为基础实现的SDS2.0总体架构，WiDE引擎实际上是提供了各种核心的基础数据存储能力：

【前瞻性的系统架构】

• 全局统一的元数据管理层（天枢元数据引擎）

天枢引擎将成为WiDE的大脑，实现元数据的全部统一管理，能够轻松地对跨越多云环境上的许多个存储池，实现统一调度和故障隔离管理，对于数据及其IO处理实现充分的抽象，除了基本的存储集群元数据管理功能之外，还将实现融合的全局数据管理，跨池跨云的智能IO调度和数据流动，全场景的数据生命周期管理。

• 全抽象的接口协议层USS（Unified Storage Service）

USS（Unified Storage Service）是整个WiDE引擎的核心协议适配层，实现了各种通用存储接口协议，包括NVMe-oF以及Vhost等新一代高性能块接口、文件协议（SMB、NFS、FUSE等）、对象存储S3协议、大数据HDFS协议、KV和SQL等数据访问协议等。灵活的USS协议层能够兼容适配企业各类业务场景使用不同的协议访问数据的需求，做到各种业务系统，各种计算与分析引擎都能够按熟悉的标准协议访问统一存储层的数据，并且能做到多种协议互通，访问同一份数据。

• 一体化多属性数据池

WiDE引擎支持多数据池架构，在统一的逻辑Namespace，统一的元数据管理架构，统一的协议层USS之下，可以支持多种存储池并存，包含高性能副本池，海量EC池，重删池以及纳管第三方CEPH池的能力，同时还支持搭建于公有云基础设施以及边缘节点上的数据池。通过多池架构可以满足企业各种业务场景对于数据存储和管理的多维度需求，例如：高性能、高可靠、大容量、低成本、高磁盘利用率等。

• 跨混合多云数据流动能力

WiDE支持数据IO跨多池的智能调度以及跨云的数据智能流动，能够从IO调度的层面实现数据跨边缘、数据中心、公有云之间的任意数据迁移和流动。借助这样的能力能够轻松实现数据的热温冷分级管理、边缘云整合、跨数据中心备份、跨地域容灾、公有云备份等通用能力。

【灵活、高速、稳定的网络架构】

• 线程级双栈网络

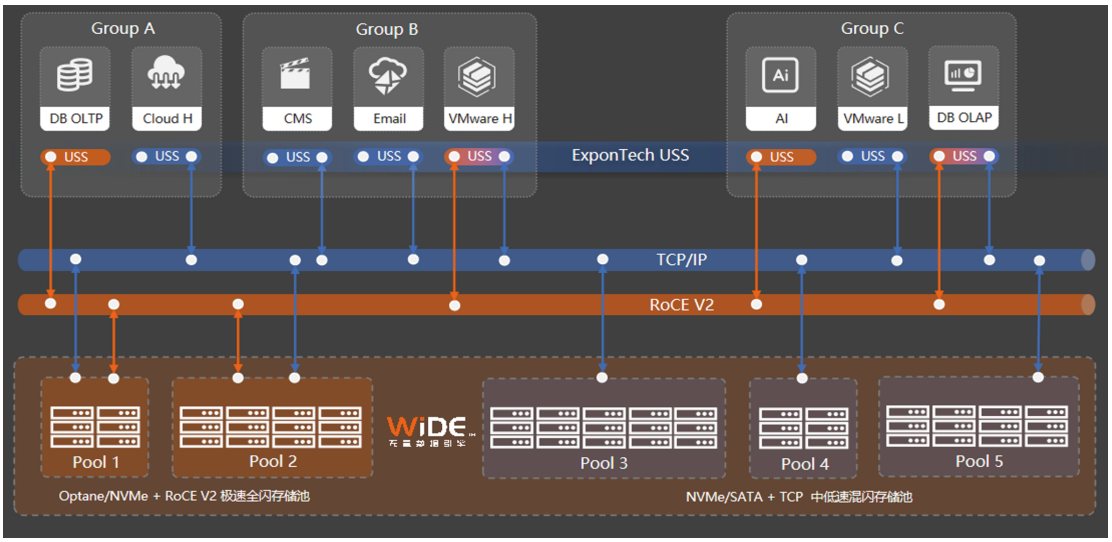


图2 线程级双栈网络支持

WiDE引擎支持线程级别双栈网络，同时支持TCP以及RoCEv2协议，在任意线程内部都可以实现高速RDMA网络和低速TCP/IP网络的任意选择，支持各种场景和业务类型下采用不同的网络协议访问存储，带来不同的性能体验，也会有不同的性价比收益。

• 自研RDMA网络防拥塞实现

WiDE的高速网络主要使用RDMA（RoCEv2）协议，RoCEv2是基于无连接的UDP协议实现的，没有TCP实现的一套可靠传输机制，一旦出现丢包就会极大降低RDMA的传输效率。而实现不丢包的关键就是解决网络拥塞问题。一般的解决方案是通过网卡和交换机上配置PFC、ECN来实现流控，从而保证在高负载下性能的稳定。WiDE引擎为了充分发挥软件定义能力，自研了一套拥塞控制算法，在硬件不支持PFC、ECN等的条件下，也能实现网络防拥塞，保障性能平稳。

【极速且稳定的存储IO引擎】

• 自研极速无锁IO栈Astrapi

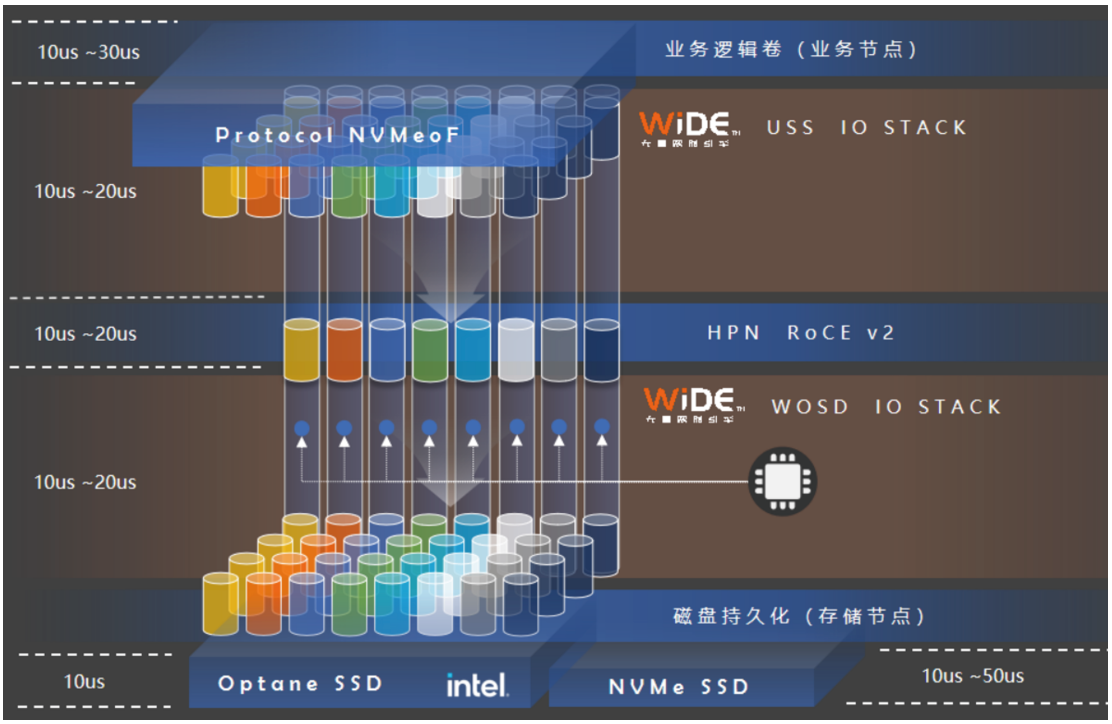


图3 Astrapi极速IO栈

WiDE使用了自主研发的全新架构的极速IO栈Astrapi，彻底改变了传统的系统软件的分层转发，依赖内核调用，基于多线程池和系统调用处理并发的IO处理模型，采用了在全用户态实现的无锁数据通道，基于RTC（Run to completion）IO处理模型，真正做到端到端没有上下文切换、无锁零中断、全用户态、内存零拷贝、极速网络IO处理等，采用绑核和异步轮询的方式来充分发挥CPU单核（虚拟线程核）的处理能力，整个软件栈的时延可以做到极低的20~40us，相比传统分布式IO软件栈400us的时延有了数量级的提升，能够真正充分发挥新一代硬件介质和网络技术的能力。

• 自研分布式复制协议实现

WiDE自主研发了一套支持强一致性的高性能分布式复制状态机，状态机基于DHT（一致性hash算法）和epoch来实现，多副本实现路径最短，一跳网络完成多副本分发，高效完成多副本强一致性；同时利用自研的复制状态机，能够最大限度保证故障场景高可靠性的同时兼顾高性能。WiDE自研的分布式复制状态机，整体的代码工程实现结合全自研的极速IO栈Astrapi及其编程模型来完成，相较于业界常用的基于Raft协议实现的复制状态机(比如ETCD的实现)，其整体的数据副本分发效率，故障场景和扩缩容时的数据迁移和恢复效率都要高效的多，在保证强一致性和高可靠性的同时，能够带来更好的可用性，更短的时延和非常优秀的性能体验。自研的分布式复制协议结合极速IO栈Astrapi，是WiDE引擎在分布式环境上能够实现端到端20~40us时延的关键环节。

• 自研单机磁盘管理存储引擎wstore

WiDE自主研发实现了单机磁盘管理引擎wstore，整个单机引擎同样结合极速IO栈Astrapi的架构和编程模型实现，实现全用户态无锁零拷贝并行下盘，实现可扩展的数据下盘路径，能够轻松对接SCSI协议的SATA接口和NVMe协议的NVMe接口，对于各种介质都可以充分发挥出最优的性能。同时还实现了磁盘亚健康检测能力，通过磁盘的I/O健康程度和错误码的识别，提前或者及时感知磁盘的故障和亚健康，保证业务的稳定运行。

2.4 华瑞指数云极速分布式存储产品WDS介绍

ExponTech WDS（WiDE Hyper-IO Disk Storage）是一款基于WiDE引擎开发出来的全自研的极速分布式块存储产品，是面向全新一代存储介质和新一代网络技术设计的软件架构，在设计目标是在系统性能、时延和可靠性方面大幅度超越当前的主流分布式块存储产品，达到可以与集中式AFA存储比肩的水平，能够承载企业核心关键业务，同时基于其分布式和软件定义的特征，可以在要求高性能和低时延的实时数据分析场景也发挥优势。

WDS凭借全新的分布式系统架构，打破了现在主流的分布式存储主要用于中低端业务场景的藩篱，可满足企业核心数据库、实时数据仓库、高性能虚拟化、HPDA（High Performa Data Analysis）、容器存储等应用对高性能、高可靠存储能力的需求，可广泛应用于金融、医疗、能源、交通、制造等行业，为企业关键应用提供持续卓越的数据存储服务能力。

WDS基于WiDE引擎开发，在继承了WiDE引擎的新一代分布式系统架构和能力的同时，实现了诸多关键的产品化特性：

【全场景极速块存储协议栈支持（iSCSI、NVMe-oF、vhost）】

WiDE引擎对外实现了一个非常重要的模块叫做统一存储服务层USS（Unified Storage Service）。借助USS的实现，WDS实现支持丰富的块协议栈，可以支持通用块设备和极速块设备协议，协议通过动态库形式提供，让网关通过动态库组合提供多协议的支持；所有协议都被抽象成USS的请求，和WiDE中多种存储池进行交互。当前在WDS产品中，依托于USS主要支持iSCSI、NVMe-oF、vhost三种协议，具体可以参考如下的逻辑图：

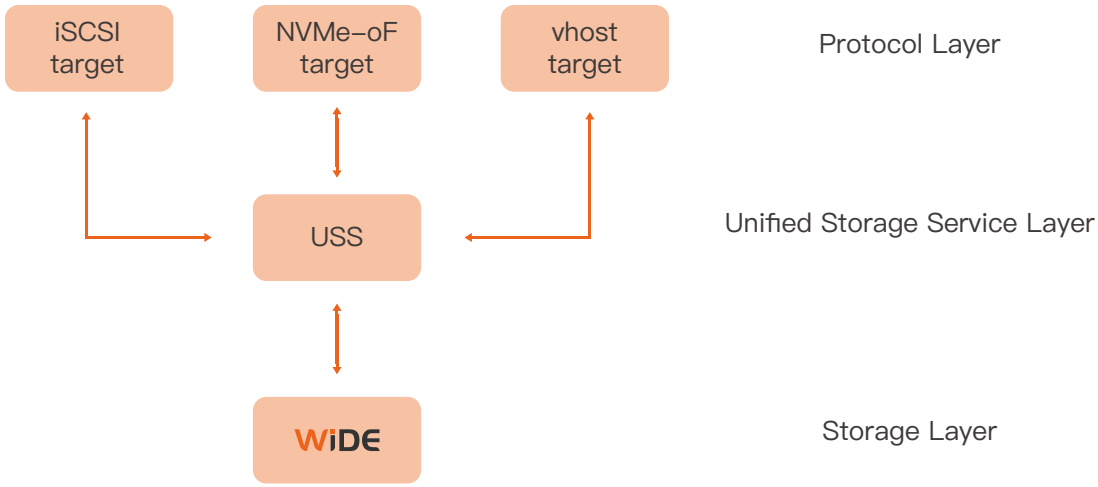


图4 基于通用协议层USS实现的多种块设备接口

【广泛的解决方案场景支持】

得益于USS支持的各种协议栈，WDS可以支持各种主流块存储应用场景，包括物理机、虚拟化、私有云和容器等：

- **支持VMware虚拟化场景：**

WDS支持通过iSCSI、NVMe-oF 和VMware进行对接，并能够提供非常优秀的性能，WDS已完整通过VMware的VAAI认证。

- **支持KVM（OpenStack）虚拟化场景：**

在OpenStack场景，WDS可以通过自己实现的Cinder driver，基于iSCSI和NVMe-oF对接OpenStack环境上的虚拟机。同时WDS还可以通过vhost方式对接KVM，在KVM虚拟化环境上提供实现极速的性能体验。

• **支持Kubernetes容器场景：**

在容器场景，WDS已经实现了CSI-iSCSI和CSI-NVMe-oF的对接，能够轻松让业务按照自己的需求来选择适配不同的协议栈接口。

• **支持本地虚拟块设备的直接挂载（非容器虚拟化场景）：**

WDS支持把USS部署于计算节点上，计算节点可以与存储节点超融合部署在一起，也可以采用存算分离方式独立部署，USS会在计算节点上虚拟出本地的块设备（支持NVMe设备或者SCSI设备）来直接与应用进行对接使用，该路径比较适用于非虚拟化环境上的数据库场景等，可以提供极速的性能体验。

【可组合式架构】

WDS利用WiDE提供的底层资源的充分利用能力实现了产品层面的可组合式架构，WiDE的极速IO栈可以实现对系统资源及其能提供的性能的精确度量，每一个进程和模块都可以量化成“资源+功能+性能指标”的组合，所以能够做到按容量、性能、可靠性、功能等许多维度来精确定义系统资源需求，并搭配相应的软件模块。

例如可以按照指定场景维度来定义硬件和软件的组合，按照用户场景的需求实现硬件和资源可量化配置，假设一个虚拟化场景需要单卷N万IOPS，集群需要M万IOPS，就可以按照这样的性能需求选择合适的网卡、磁盘介质、CPU和内存等硬件资源，然后按照性能需求计算出软件的线程和资源（CPU+内存）配置。另外也可以按照指定硬件维度来定义软件的组合，如果用户已经购买硬件，也可以通过可组合式架构能力，通过对软件资源的充分配置实现硬件资源的最大利用率，实现对硬件能力的可定制化利用。

通过可组合式架构的能力，能真正实现存储系统资源的可量化的精确计算，最大限度使用硬件资源，充分发挥硬件性能的同时，可以实现降低能耗和减少资源浪费。

【完善的可视化的运维管理系统】

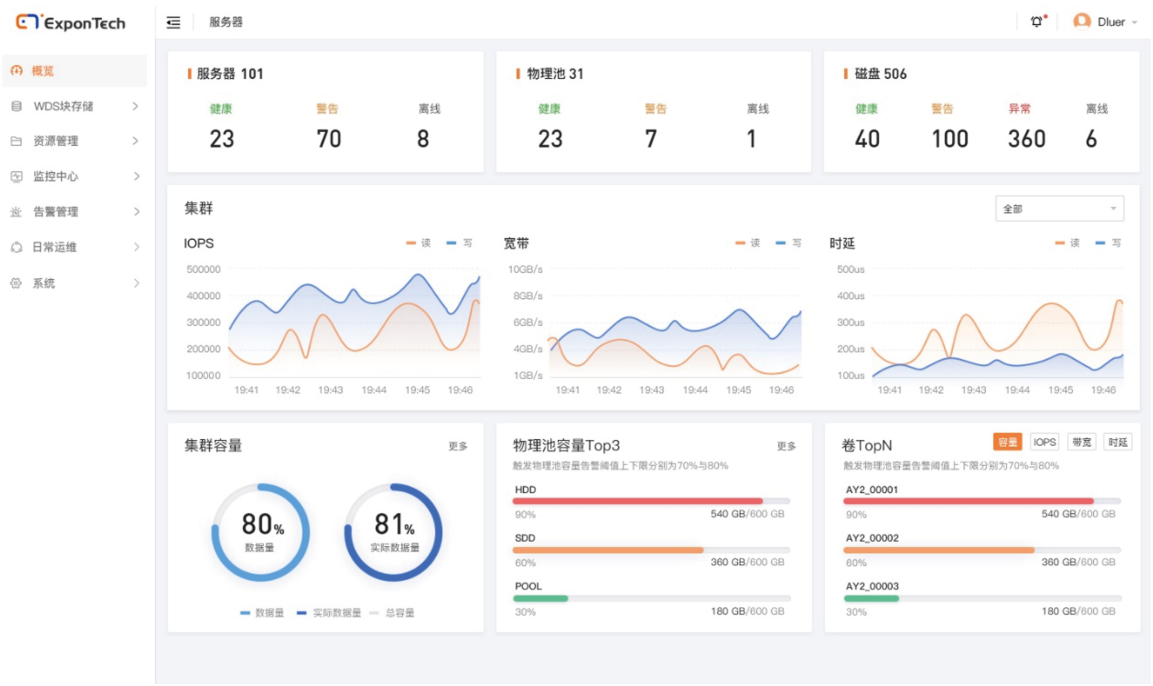


图5 运维管理平台

WDS实现了完整的运维管理系统和图形化的操作界面，实现了简便的安装部署、直观的系统资源管理、以及全功能的图形界面操作等；可运维性方面实现了实时的性能和容量监控、全场景的故障告警提示、深入的运维巡检等功能。

2.5 WiDE及WDS可靠性设计及最佳实践

可靠性和稳定性一直都是存储系统中重中之重的能力，WiDE自主研发的全新一代的分布式系统架构，在设计之初就重点思考和设计了一整套的提升可靠性和稳定性的系统能力。WiDE的可靠性系统能够让整个WiDE在极端硬件故障的场景下也可以保证业务的稳定运行；可以让系统在某些硬件组件亚健康的时候提前感知和预测，提前进行故障预判和处理，保证系统的稳定运行；可以提供各种容灾特性，在非人为等自然灾难发生时也可以实现数据不丢失，保障业务运行的连续性。

【支持多级Recovery QoS选择】

在分布式系统中由于有比较可靠的副本冗余机制，所以单盘，单服务器乃至单机柜的故障都不会给业务带来实质影响，但是发生此类故障之后系统会进入副本修复流程来重建数据副本，保证多个副本的一致，这个过程叫做Recovery，Recovery的过程中会有重建的数据流在集群内部产生，这样的重建数据流如果发生在业务高负载情况下，会或多或少给业务带来影响。

在WiDE和WDS的副本引擎的设计中独立设计了针对Recovery流程的QoS，能够通过多个QoS级别的设置，在业务高负载的时候，通过限制Recovery的流量来优先保证业务的稳定性，在业务低负载的时候放大Recovery的流量，以尽快完成数据重建，让业务的稳定性和Recovery能够形成充分的平衡。WiDE未来还会实现一整套端到端的QoS能力，从全场景、端到端来保证业务的稳定性。

【支持副本一致性校验】

在真实的生产环境中，当存储集群规模较大并且运行时间较长的时候会有一定概率出现某些磁盘的静默错误，这是一种出现概率小，但是一旦发生影响非常巨大的故障场景。WiDE和WDS的副本引擎实现了多副本间进行一致性校验的功能，通过定期的三副本比对和自动修复，能有效地降低了磁盘静默错误所带来的影响。后续还会与WiDE的端到端DIFF整合形成完整的多副本数据一致性保护能力，充分保证WiDE中数据的一致性。

【智能慢盘检测/SSD寿命监测】

慢盘是一种典型的亚健康问题，其表现就是硬盘IO的时延变长，但仍然可以提供数据访问。导致慢盘的原因最常见的是硬盘坏道、磁头异常、SSD寿命到期等硬件问题。存储系统中出现慢盘，会形成短板效应，导致业务性能急剧下降，严重时可能会导致服务不可用。

具体来说，慢盘表现为横向对比和纵向对比两个维度的性能下降。所谓横向对比是指，随着使用时间的推移，某些硬盘的性能明显低于系统内绝大多数其他硬盘的性能；而纵向对比则是指特定硬盘在使用过程中性能突降，可能在一段时间后恢复正常，也有可能不再可逆。

在WiDE和WDS中采用了智能慢盘检测算法，详细记录了盘上的IO模型（例如I/O大小、I/O时延、I/O类型、I/O数量），硬盘内部的SMART指标，IO路径上的错误记录，SSD的寿命衰减情况。根据历史的监控指标和集群的业务负载，利用机器学习算法，聚类算法，建立了当前环境下的动态阈值，不仅能及时发现绝对条件下的慢盘，也能发现相对条件下的慢盘，从而上报慢盘告警并及时隔离。

【智能网络亚健康】

在分布式系统中，网络亚健康是一个极具挑战性的问题。与断网这种非好即坏的明确故障场景不同，网络亚健康具有很强的模糊性，是否被判定为亚健康，很大程度上还取决于对当前业务的影响。在WiDE和WDS中实现了智能网络亚健康检测，根据组网模型和网络规模，智能的选择服务器节点的不同网络链路、不同网口自适应发送探测包，从链路、网口、网卡多个维度检测时延异常、错包和丢包等异常指标。

同时，对集群的IOPS、带宽和时延三个重要指标进行趋势分析，然后对网络指标和集群性能指标进行关联性分析，从而更加准确的检测网络亚健康问题。在检测到网络亚健康之后，运维平台会进行告警通知，并且根据亚健康的程度对链路、网口、网卡进行逐级隔离，并把业务切换到正常的链路、网口和网卡。

【支持容灾双活】

WDS在生产场景会应用于各种虚拟化或者数据库场景，对于可靠性和容灾的要求极高，WiDE和WDS在副本引擎中实现了短距离容灾双活的能力，能够在两个地域的副本之间形成强一致的副本同步，业务在极端故障场景仍然可以切换到另一个容灾区域，保证的连续性可以得到充分的保证。

【跨池数据流动】

多池架构以及跨池IO调度是WiDE引擎一个核心的技术能力，WiDE通过集中式的元数据管理天枢以及统一的协议层USS，真正实现了多池架构下的数据跨池流动和跨池管理，无论是块、文件、对象的逻辑管理元数据都不会局限于单池内部，带来了极其灵活的扩展能力，也带来了数据可靠性能力的大幅提升。

WiDE和WDS能够实现数据任意跨池，任何的冗余故障范围不会拘泥于一个单池，所以基于WiDE的产品可以采用构建多个池，而每个池的规模都比较小的方式来构建业务高可用集群，单池规模小会让整个系统的故障域很小，大大降低了极端故障的影响范围，而构建许多个小池又可以实现海量的数据容量，跨池的IO调度和数据流动又使得这无数个小池在逻辑上组成了一个大的数据池，使用起来非常方便。

另外通过跨多池的能力，可以轻松实现多池具备不同的属性，比如高性能全闪池、高性价比混闪池、海量EC池等，业务可以通过WiDE跨多池的能力轻松实现数据的分池分级，并可以在多池之间随意数据流动，充分发挥多池的优点和能力。

【分布式系统可观测性】

WiDE在框架中内置实现了全系统的可观测性和I/O的实时跟踪，实现了一整套的Trace系统，能够充分了解数据的均衡性和分布情况，及时识别不均衡的问题并及时修复调整。实现了完整I/O路径的时延和性能跟踪，可以观测并了解到任意一个影响时延的痛点，以最快的速度发现各种影响I/O时延稳定性的问题。

在生产环境由于网络或者硬件的各种物理不稳定原因，经常会有一些慢I/O（slow requests）让业务时延出现一些抖动的情况，通过WiDE的分布式可观测性系统能力可以及时发现隐含在系统中的深层次的故障点或隐患，及时修复保证业务的平稳。

【跨机柜跨故障域的可靠性】

在分布式系统中，很常见的一种场景是发生故障后，故障的影响范围快速扩散至整个系统，最终有可能导致所有业务都受到影响。因此，做好故障隔离是一种非常有效的技术手段。在大部分的分布式系统实现中，都是通过划分不同的故障域进行故障隔离的，WDS也提供了多个级别的故障域隔离，包括硬盘级、服务器级和机柜级，客户可以根据机房的条件和对可靠性的要求，灵活选择合适的故障域级别。更进一步，WiDE基于其多池架构以及跨池IO调度的能力，可以提供更加灵活的故障隔离手段和容灾手段。

3 英特尔实验室测试及结果

3.1 英特尔机房环境描述

硬件配置

极速性能服务器：

服务器个数：3台

主机名列表：Node84；Node85；Node86

CPU：Intel(R) Xeon(R) Platinum 8358（icelake） CPU @ 2.60GHz * 2

Mem：512GB

网卡：英特尔®以太网网络适配器E810-2CQDA2（单口100GE）网卡 * 2

OS：CentOS Linux release 7.9.2009 (Core)

普通性能服务器：

服务器个数：3台

主机名列表：Node61；Node62；Node64

CPU：Intel(R) Xeon(R) Gold 6240Y（cascadelake） CPU @2.60GHz * 2

Mem：256GB

网卡：英特尔®以太网网络适配器E810-2CQDA2（单口100GE）网卡 * 2

OS：CentOS Linux release 7.9.2009 (Core)

主存储介质：

英特尔®傲腾™固态盘P5800X 1.6T * 12

其中普通性能服务器将用于组建分布式存储集群，极速性能服务器将用于做计算结点，发起对分布式存储集群的各种压力测试。

网络拓扑：

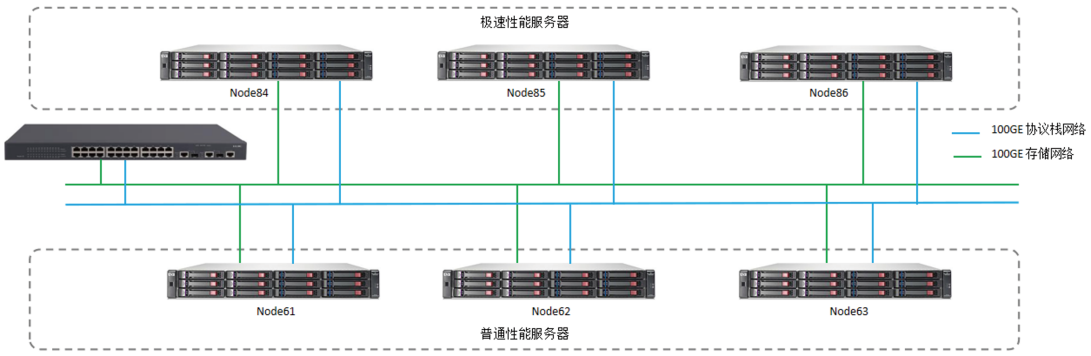


图6 测试环境网络拓扑图

3.2 【测试1】最优基线性能对比测试

3.2.1 硬件配置拓扑及测试方法

计算节点

3台极速性能服务器（CPU：Intel(R) Xeon(R) Platinum 8358（icelake）），详见“3.1 英特尔机房环境描述”内容。

存储节点

3台普通性能服务器（CPU：Intel(R) Xeon(R) Gold 6240Y（cascadelake）），详见“3.1 英特尔机房环境描述”内容。

存储节点主存

每台存储节点配置4块英特尔®傲腾™固态盘P5800X。

测试环境拓扑图

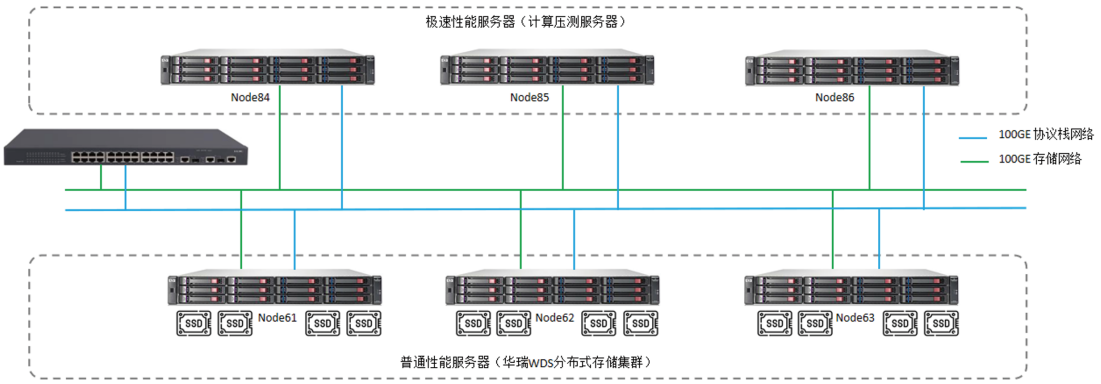


图7 基线性能测试环境拓扑图

测试方法

在相同的硬件环境上，分别测试不同的分布式存储软件（华瑞WDS产品 VS CEPH（N版本）），对比分析其基线性能。

3.2.2 WDS Benchmark测试

测试环境拓扑

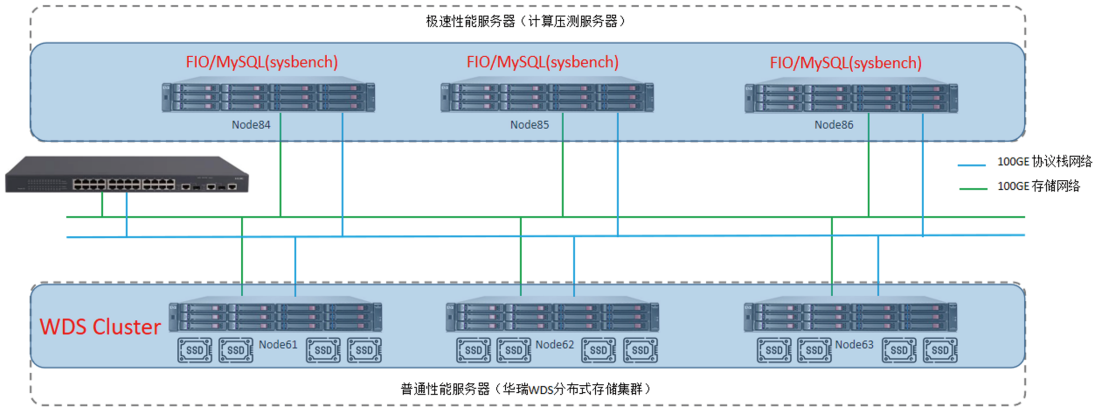


图 8 WDS Benchmark测试环境拓扑图

软件配置

- 基线压测工具
  FIO：FIO-3.7；
  MySQL（Sysbench）：SYSBENCH 1.0.17

- 分布式存储软件：华瑞指数云WDS
  版本：WDS 2.0.0, 配置为3副本同步写入
  网络协议：RoCEv2
  块存储协议：NVMe-oF(NVMe over RDMA)

测试方法

- FIO测试

  使用FIO工具使用单节点和三节点分别压测，测试出单卷性能以及集群性能的多个用例的基线性能。

- MySQL（Sysbench）测试
  在两个节点上搭建Mysql主备集群，在单节点上使用Sysbench进行压测，测试出纯写和读写两种场景的基线性能。

测试结果

(1) WDS FIO Benchmark数据

单卷时延	io模型		卷数量	iodepth	numjobs	iops	时延-us	带宽-MB
4k	随机写	randwrite	1	1	1	10100	89	40
	随机读	randread	1	1	1	13400	65	53
	读写混合	randwrite	1	1	1	3687	90	14
		randread	1	1	1	8601	64	34
8k	随机写	randwrite	1	1	1	9746	93	76
	随机读	randread	1	1	1	12500	70	98
	读写混合	randwrite	1	1	1	3457	94	27
		randread	1	1	1	8060	70	62

单卷性能	io模型		卷数量	iodepth	numjobs	iops	时延-us	带宽-MB
4k	随机写	randwrite	1	64	8	934000	544	3649
	随机读	randread	1	64	16	2593000	390	9890
	读写混合	randwrite	1	64	16	649000	672	2533
		randread	1	64	16	1513000	380	5910
8k	随机写	randwrite	1	64	4	481000	528	3760
	随机读	randread	1	64	8	1383000	357	10600
	读写混合	randwrite	1	64	8	381000	581	2979
		randread	1	64	8	890000	316	6953

集群性能	io模型		卷数量	iodepth	numjobs	iops	时延-us	带宽-MB
4k	随机写	randwrite	3	64	8	2439000	622	9527
	随机读	randread	3	64	32	6919000	438	26500
	读写混合	randwrite	3	64	16	1392000	520	5439
		randread	3	64	16	3248000	242	12400
8k	随机写	randwrite	3	64	4	1436000	531	10000
	随机读	randread	3	64	16	4047000	750	30900
	读写混合	randwrite	3	64	8	1004000	628	7843
		randread	3	64	8	2343000	334	17900

(2) WDS Sysbench数据（mysql数据库）

纯写场景

表数量	Threads	TPS(per S)	QPS(per S)	时延(ms) avg/95%
1	32	11584	69515	2.8/3.3
1	64	20733	124400	3.1/4
1	128	42531	261190	2.9/4.5
1	256	37236	223416	6.9/10.1
16	32	10287	61723	3.1/3.7
16	64	16725	100354	3.8/4.5
16	128	21831	130991	5.9/7.2
16	256	27181	163090	9.4/12.3

读写场景

表数量	Threads	TPS(per S)	QPS(per S)	时延(ms) avg/95%
1	32	6013	120260	5.3/6.8
1	64	15360	307206	4.2/5.4
1	128	26009	520182	4.9/5.9
1	256	24437	488752	10.5/13.5
16	32	5809	116191	5.5/6.8
16	64	13921	278436	4.6/5.9
16	128	21795	435914	5.9/6.6
16	256	23305	466117	10.9/13.7

3.2.3 Ceph Benchmark测试

测试环境拓扑

所有的硬件与3.2.2保持一致，不发生任何变化，只是安装的分布式存储软件由WDS改为CEPH，这样可以在完全相同的硬件条件下来比较WDS分布式存储与Ceph分布式存储的性能表现。

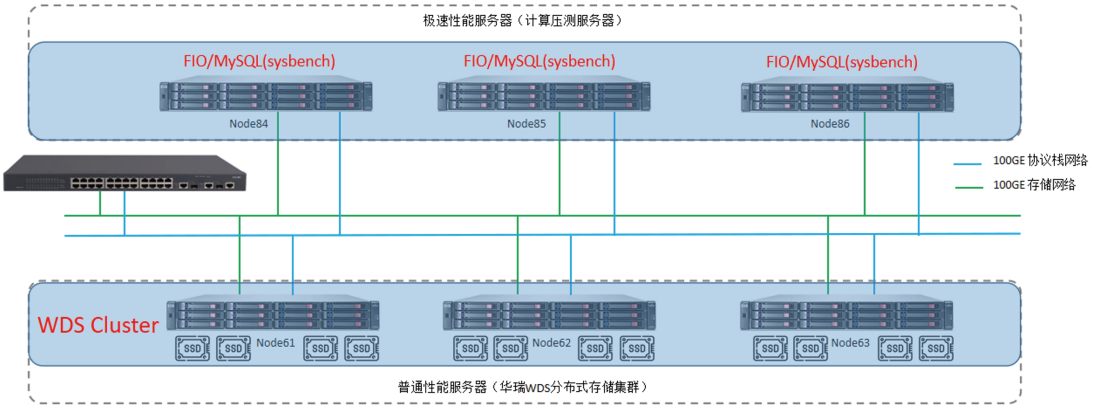


图9 Ceph Benchmark测试环境拓扑图

软件配置

- **基线压测工具**

  FIO：FIO–3.7

  MySQL（sysbench）：SYSBENCH 1.0.17

- **分布式存储软件：CEPH**

  版本：CEPH N版本，配置为3副本同步写入

  网络协议：RoCEv2

  块存储协议：KRBD

测试方法

- **FIO测试**

  使用FIO工具使用单节点和三节点分别压测，测试出单卷性和集群基线性能。

- **MySQL（Sysbench）测试**

  在两个节点上搭建Mysql主备集群，在单节点上使用Sysbench进行压测，测试出纯写和读写两种场景的基线性能。

测试数据

(1) CEPH benchmark数据

单卷时延	IO模型		卷数量	iodepth	numjobs	IOPS	时延-us	带宽-MB
4k	随机写	randwrite	1	1	1	689	1442	2.8
	随机读	randread	1	1	1	2310	422	9.4
	读写混合	randwrite	1	1	1	407	1564	1.6
		randread	1	1	1	952	365	3.9
8k	随机写	randwrite	1	1	1	630	1577	5.1
	随机读	randread	1	1	1	2196	445	18.8
	读写混合	randwrite	1	1	1	388	1607	3.1
		randread	1	1	1	908	395	7.3

单卷性能	IO模型		卷数量	iodepth	numjobs	IOPS	时延-us	带宽-MB
4k	随机写	randwrite	1	16	8	66800	1814	274
	随机读	randread	1	16	8	73100	1665	300
	读写混合	randwrite	1	16	8	21400	2052	87.8
		randread	1	16	8	49000	1545	209
8k	随机写	randwrite	1	16	8	67600	1790	554
	随机读	randread	1	16	8	72900	1667	598
	读写混合	randwrite	1	16	8	22800	1964	187
		randread	1	16	8	53100	1436	435

集群性能	IO模型		卷数量	iodepth	numjobs	IOPS	时延-us	带宽-MB
4k	随机写	randwrite	6	16	8	141000	5460	575
	随机读	randread	15	16	8	485000	3945	1987
	读写混合	randwrite	6	16	8	94900	4525	389
		randread	6	16	8	221000	1518	900
8k	随机写	randwrite	6	16	8	129000	5926	1059
	随机读	randread	15	16	8	476000	4024	3898
	读写混合	randwrite	6	16	8	91400	4803	748
		randread	6	16	8	213000	1531	1745

(2) CEPH Sysbench数据（mysql数据库）

纯写场景

表数量	Threads	TPS(per S)	QPS(per S)	时延(ms) avg/95%
1	64	6115	36694	10.5/14
1	128	10140	60841	12.6/17.6
1	256	15906	95439	16/22.7
1	512	20823	124940	24.6/35.6
16	64	6559	39359	9,8/13
16	128	6912	41472	16.5/18
16	256	8192	49155	31.2/33.1
16	512	11216	67301	45.6/56.8

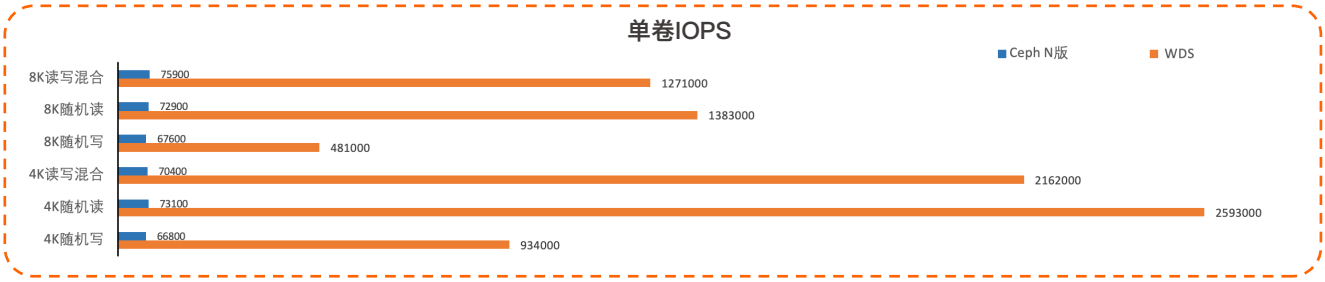
读写场景

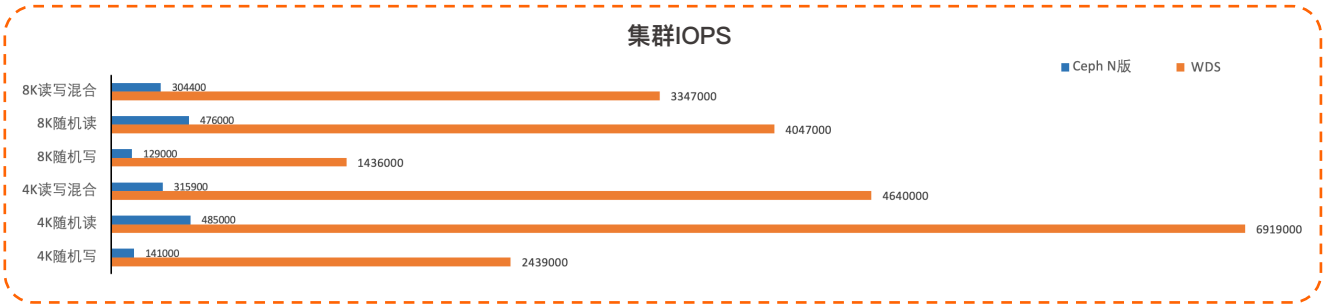
表数量	Threads	TPS(per S)	QPS(per S)	时延(ms) avg/95%
1	64	4464	89285	14.3/18
1	128	8994	179885	14.2/19
1	256	17676	353523	14.5/21.5
1	512	21498	429967	23.8/32.5
16	64	4584	91696	14/17.3
16	128	9966	199327	12.8/16.7
16	256	10313	206275	24.8/32.5
16	512	12069	241387	42.4/103

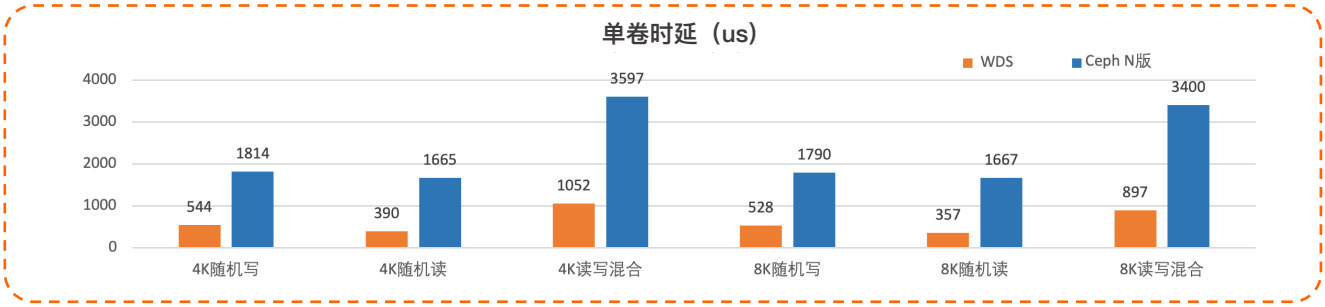
3.2.4 数据对比与结论

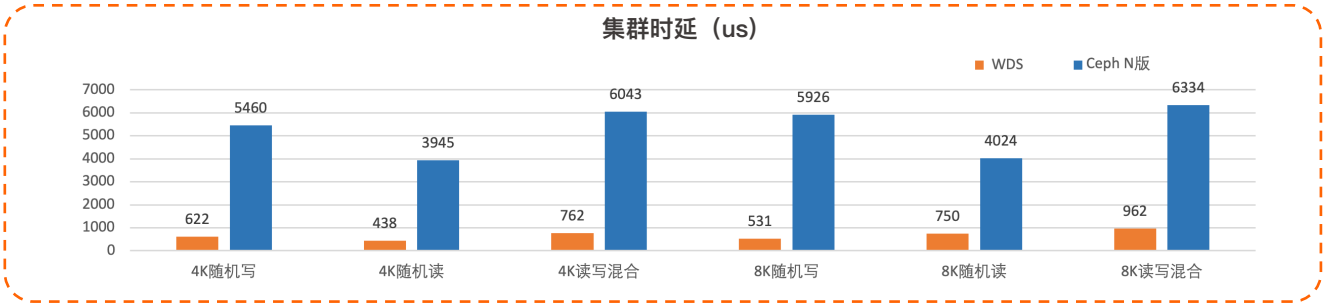
在相同的硬件环境上，分别部署了两套分布式存储软件：WDS2.0和Ceph N版本，测试FIO存储基线性能数据和SYSBENCH数据库基线性能数据，现在两套分布式存储软件的测试结果对比如下：

FIO存储基线性能对比



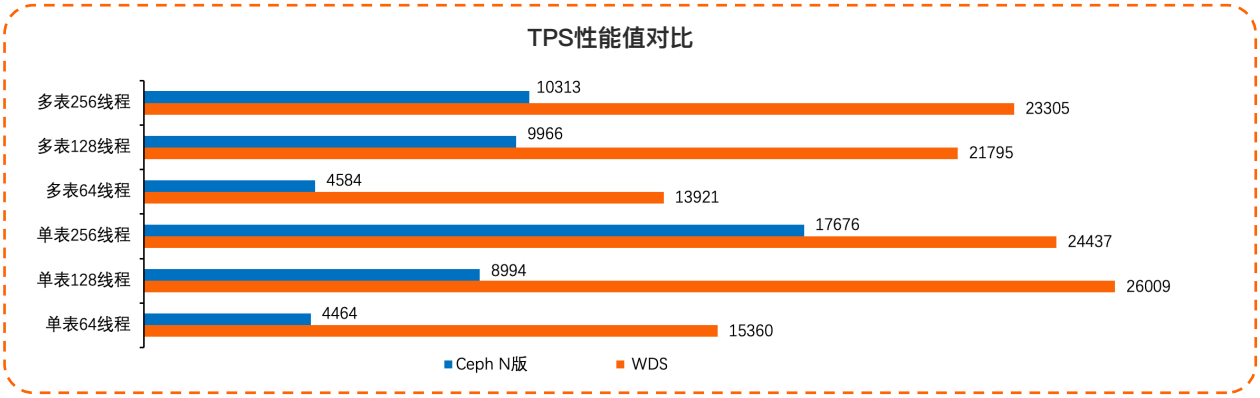


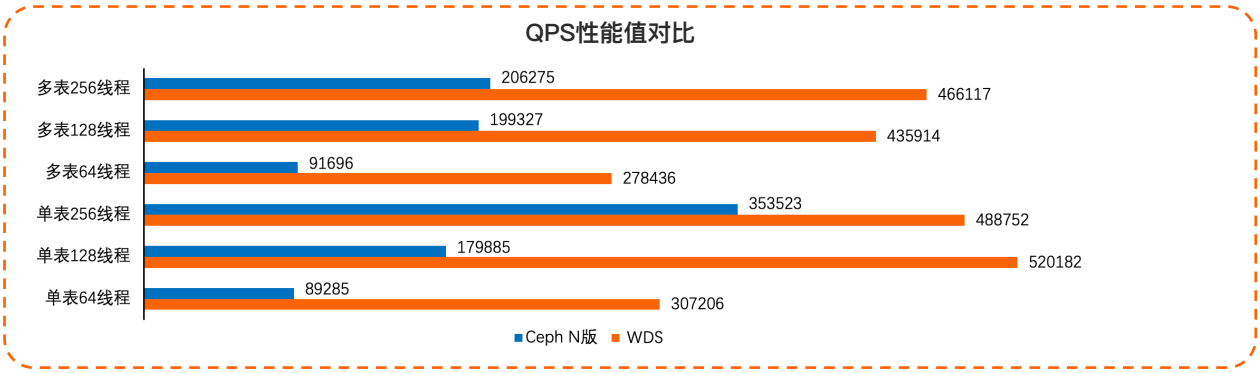


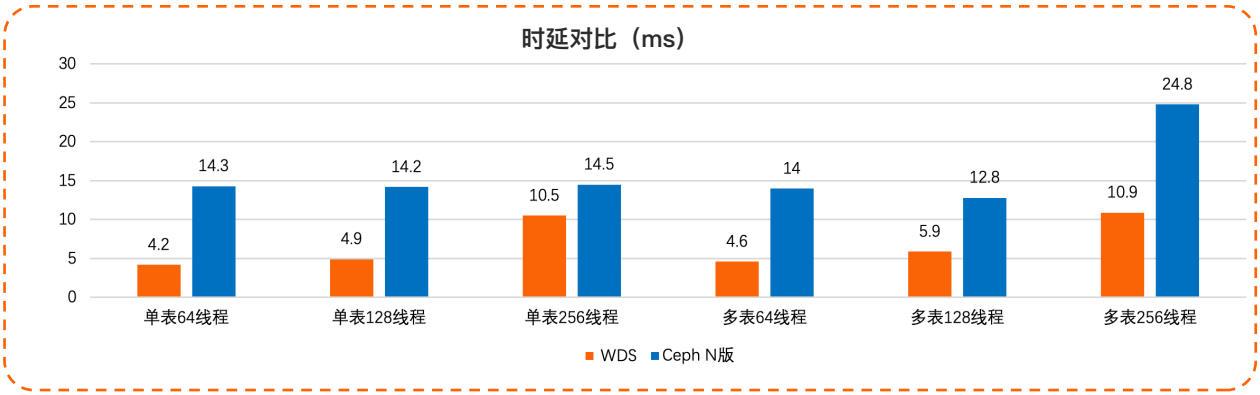


测试用例 (4K IO 模型)	单卷IOPS			集群IOPS			单卷时延(us)			集群时延(us)		
	WDS	Ceph N版	性能对比	WDS	Ceph N版	性能对比	WDS	Ceph N版	时延对比	WDS	Ceph N版	时延对比
随机写	934000	66800	14倍	2439000	141000	17.29倍	544	1814	30%	622	5460	11.3%
随机读	2593000	73100	35.4倍	6919000	485000	14.26倍	390	1665	23.4%	438	3945	11.1%
读写混合	2162000	70400	30.7倍	4640000	315900	14.68倍	1052	3597	29.2%	762	6043	12.6%
测试用例 (8K IO 模型)	单卷IOPS			集群IOPS			单卷时延(us)			集群时延(us)		
	WDS	Ceph N版	性能对比	WDS	Ceph N版	性能对比	WDS	Ceph N版	时延对比	WDS	Ceph N版	时延对比
随机写	481000	67600	7.11倍	1436000	129000	11.13倍	528	1790	29%	531	5926	8.9%
随机读	1383000	72900	18.9倍	4047000	476000	8.5倍	357	1667	21.4%	750	4024	18.6%
读写混合	1271000	75900	16.7倍	3347000	304400	11倍	897	3400	26.3%	962	6334	15.1%

SYSBENCH数据库基线性能对比







测试用例（单表）	TPS			QPS			时延(ms)	
	WDS	Ceph N版	性能对比	WDS	Ceph N版	性能对比	WDS	Ceph N版
单表64线程	15360	4464	3.44倍	307206	89285	3.44倍	4.2	14.3
单表128线程	26009	8994	2.89倍	520182	179885	2.89倍	4.9	14.2
单表256线程	24437	17676	1.38倍	488752	353523	1.38倍	10.5	14.5
测试用例（多表）	TPS			QPS			时延(ms)	
	WDS	Ceph N版	对比	WDS	Ceph N版	对比	WDS	Ceph N版
多表64线程	13921	4584	3.03倍	278436	91696	3.03倍	4.6	14
多表128线程	21795	9966	2.18倍	435914	199327	2.18倍	5.9	12.8
多表256线程	23305	10313	2.26倍	466117	206275	2.26倍	10.9	24.8

从实际测试数据对比来看，基于无量分布式存储引擎的WDS产品与系统调优后的Ceph Nautilus版本相比，实现了IOPS的数量级的惊人提升，无论是单卷性能还是集群性能均提升高达10到30倍以上，与此同时IO时延缩短为Ceph Nautilus版本的十分之一。

在Mysql数据库的Sysbench测试中，验证访问16个表的每秒事务处理性能(TPS),WDS可以在稳定的5ms以内时延，提供高达20000以上的TPS，而Ceph Nautilus版本没有任何办法可以提供5ms以内的稳定时延，将时延要求放宽到15ms以后，也只能提供数千TPS。

3.3【测试2】硬件瓶颈性能测试

3.3.1 硬件峰值性能测试(3块Optane P5800X)

测试目标

本测试的目的是验证WDS分布式存储软件使用单个进程就可以将一整块英特尔®傲腾™固态盘P5800X（不切分磁盘分区）的峰值能力发挥出来，也就是达到150万 IOPS以上（英特尔®傲腾™固态盘P5800X的4K随机读写性能极限）。而以Ceph为代表的当前主流分布式存储软件不太可能拥有使用单进程就能将一块英特尔®傲腾™固态盘P5800X发挥到150万 IOPS的能力。

计算节点

3台普通性能服务器（CPU：Intel(R) Xeon(R) Platinum 8358（icelake））（详见“3.1 英特尔®机房环境描述”内容）

存储节点

3台极速性能服务器（CPU：Intel(R) Xeon(R) Gold 6240Y（cascadelake））（详见“3.1 英特尔®机房环境描述”内容）

存储节点主存

每台存储节点配置1块英特尔®傲腾™固态盘P5800X

测试环境拓扑图

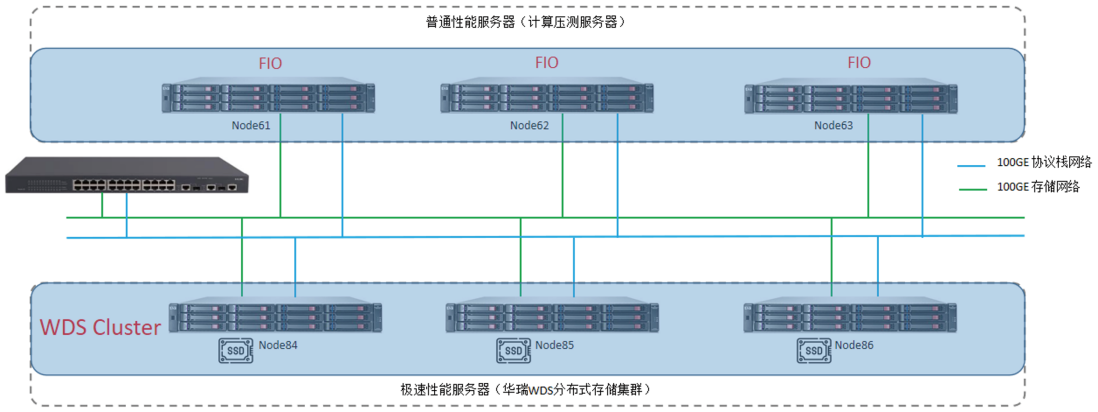


图10 硬件极限性能测试环境(3块傲腾™)拓扑图

软件配置

- 基线压测工具
  FIO：FIO–3.7

- 分布式存储软件：华瑞指数云WDS
  版本：WDS 2.0.0
  网络协议：RoCEv2
  块存储协议：NVMe–oF(NVMe over RDMA)

测试方法

使用FIO工具，仅使用单客户端节点进行压测，针对单个卷到多个卷分别进行测试，直到将测试环境中配置的英特尔®傲腾™固态盘P5800X的达到性能上限。

测试数据

单卷性能	IO模型		卷数量	iodepth	numjobs	带宽–MB	IOPS	时延–us
4k	随机写	randwrite	1	32	12	3345	856000	443.86
	随机读	randread	1	32	32	7563	1936000	517.14
	读写混合	randwrite	1	32	16	1908	488000	434.91
		randread	1	32	16	4451	1139000	255.01

集群性能	IO模型		卷数量	iodepth	numjobs	带宽–MB	IOPS	时延–us
4k	随机写	randwrite	3	32	8	6255	1601000	475.62
	随机读	randread	3	16	32	17500	4592000	327.29
	读写混合	randwrite	3	32	12	4090	1047000	525.1
		randread	3	32	12	9543	2443000	240.2

3.3.2 性能扩展性测试(6块Optane P5800X)

测试目标

在3.3.1的测试能够达到英特尔®傲腾™固态盘存储介质的性能上限之后，进一步进行测试，在原测试环境中的三个物理节点中各自增加一块英特尔®傲腾™固态盘P5800X，扩大了存储介质的性能上限，然后继续使用FIO工具，针对单个卷到多个卷分别进行加压测试，直到吞吐量达到网卡带宽的上限，以此来观察整个分布式存储集群的IOPS，时延等性能表现。

计算节点

与3.3.1测试相同。

存储节点

与3.3.1测试相同。

存储节点主存

每台存储节点配置2块英特尔®傲腾™固态盘P5800X。

测试环境拓扑图

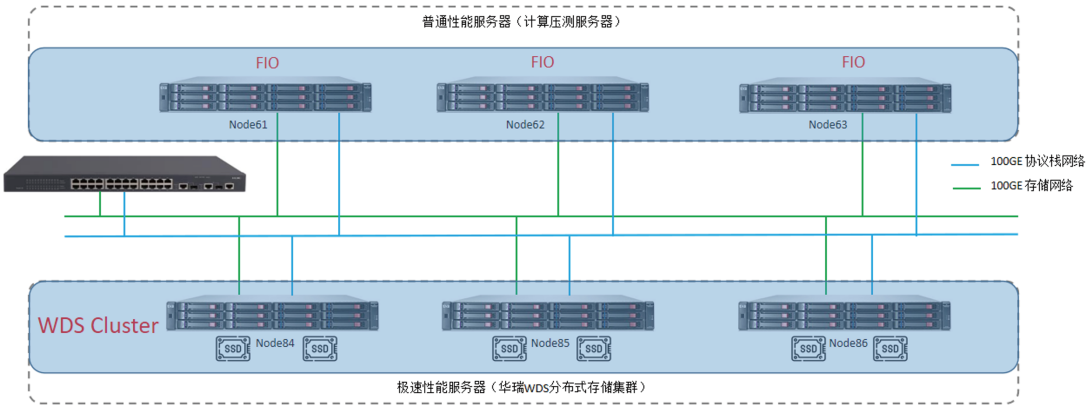


图10 性能扩展测试环境(6块Optane P5800X)拓扑图

软件配置

- 基线压测工具
  FIO：FIO–3.7
- 分布式存储软件：华瑞指数云WDS
  版本：WDS 2.0.0
  网络协议：RoCEv2
  块存储协议：NVMe–oF(NVMe over RDMA)

测试数据

集群性能	IO模型		卷数量	iodepth	numjobs	带宽-MB	IOPS	时延-us
4k	随机写	randwrite	3	16	32	9990	2619000	579.31
	随机读	randread	3	32	32	21000	5756000	522.09
	读写混合	randwrite	3	32	16	5535	1417000	450.36
		randread	3	32	16	12600	3306000	264.24

3.3.3 测试结果分析

第一组测试：最小集群(3节点共3块Optane P5800X)单卷测试

测试结果表明，单卷的4k随机读可以达到190万 IOPS，已经将网卡的带宽发挥到接近理论峰值。由于单卷是挂载在单台物理节点上进行测试，单台物理机的网卡带宽成为瓶颈，在PCIe-3.0平台上，100GE网卡的理论上限大约在200万IOPS。

第二组测试：最小集群(3节点共3块Optane P5800X)多卷测试

测试结果表明，WiDE引擎的最小化的软件部署可以将英特尔®傲腾™固态盘P5800X的性能发挥到接近峰值。实际测试中，三个计算节点并发三个卷，整体的集群性能达到了459万IOPS，集群中总共三块Optane P5800X，每一块都跑到了其理论性能上限150w IOPS，WiDE引擎软件将硬件介质的性能峰值都压榨出来了。

下图是针对其中一块英特尔®傲腾™固态盘的iostat数据截图:

```
nvme2n1   0.00   0.00 1560044.50   0.00  6093.92   0.00   8.00   347.20   0.19   0.19   0.00   0.00 113.55
nvme2n1   0.00   0.00 1559925.50   0.00  6093.46   0.00   8.00   340.15   0.19   0.19   0.00   0.00 112.45
nvme2n1   0.00   0.00 1560046.00   0.00  6093.93   0.00   8.00   352.99   0.19   0.19   0.00   0.00 114.90
nvme2n1   0.00   0.00 1560045.50   0.00  6093.92   0.00   8.00   355.91   0.19   0.19   0.00   0.00 116.25
nvme2n1   0.00   0.00 1560019.50   0.00  6093.83   0.00   8.00   335.93   0.18   0.18   0.00   0.00 112.80
nvme2n1   0.00   0.00 1559999.00   0.00  6093.74   0.00   8.00   352.68   0.19   0.19   0.00   0.00 113.80
```

图11 硬件介质iostat数据截图

第三组测试：性能扩展性(3节点共6块Optane P5800X)测试

在测试2的三个物理节点中各自增加一块英特尔®傲腾™固态盘P5800X，扩大了存储介质的性能上限，继续使用三个计算节点并发三个卷来进行压测。测试结果表明，整个集群的性能上限将随着磁盘介质的增加而线性增长，直到增长到另一个物理瓶颈---网卡的带宽上限。

本次测试环境中，共有6块英特尔®傲腾™固态盘P5800X，性能上限是900万IOPS，三台计算节点的网卡理论性能上限是600万 IOPS，而集群的实测IOPS达到了575万IOPS，接近网卡带宽上限，已经把三块100GE网卡在PCIe-3.0平台上的带宽发挥到了接近峰值。

测试结论

磁盘个数	网卡个数	集群IOPS	硬件瓶颈点	瓶颈带宽
3 Optane P5800X	1*100GE网卡	193万	100GE网卡（PCIE-3.0 X8）	7.2 GB/s
3 Optane P5800X	3*100GE网卡	459万	英特尔®傲腾™固态盘P5800X	150万 IOPS
6 Optane P5800X	3*100GE网卡	575万	300GE网卡（PCIE-3.0 X8）	21.9 GB/s

注：PCIe-3.0 X8的总线带宽7.8GB/s折算4K IOPS理论值为200万，单块英特尔® 傲腾™固态盘P5800X的标称上限性能为150万IOPS

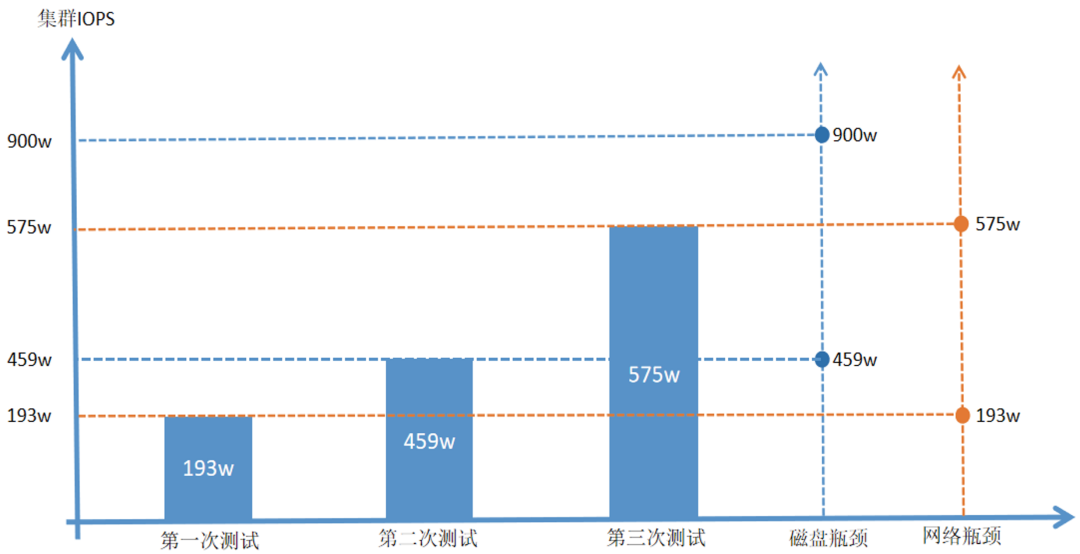


图12 集群性能与硬件上限性能对比图

如上图，整个测试随着网卡和磁盘的不断增加，整个集群的能力可以线性增长。随着卷的增长物理磁盘介质的性能成为瓶颈，增加更多的介质之后，集群性能继续线性增长，直到网卡带宽成为瓶颈，证明了WiDE分布式引擎可以充分发挥所有新型硬件的峰值性能能力，如果硬件线性增长，端到端的集群性能也可以线性增长。

另外，测试2中的测试结果也充分说明了WiDE分布式引擎已经把分布式软件栈的效率优化到了接近峰值，甚至能将英特尔®傲腾™固态盘P5800X这样的目前市面上运行速度最快的SSD的能力发挥到峰值。而以Ceph为代表的当前主流分布式存储软件，受限于其系统架构和IO处理模型的低效，即使用再多的CPU也不可能将超低时延超高性能的英特尔®傲腾™固态盘介质的能力完全发挥出来。

3.4【测试3】英特尔®QAT压缩卡与WDS整合测试

3.4.1 测试目标

使用英特尔®QAT卡部署在计算节点和存储节点的存储集群内部互联网络，希望可以发挥出来其硬件压缩的能力，传输过程中会进行压缩与非压缩的测试对比，希望可以通过英特尔®QAT卡硬件加速的数据压缩能力，在有限的网络（25GE）上发挥出超过网络带宽的系统性能。

3.4.2 参考架构

计算节点

1台极速性能服务器（CPU：Intel(R) Xeon(R) Gold 6240Y（cascadelake）），配置详见“3.1 英特尔机房环境描述”内容。

存储节点

2台极速性能服务器（CPU：Intel(R) Xeon(R) Gold 6240Y（cascadelake）），配置详见“3.1 英特尔机房环境描述”内容。

存储节点主存

每台存储节点配置4块英特尔®傲腾™固态盘P5800X。

QAT卡部署

计算节点和两台存储节点分别部署一块英特尔®QAT卡。

网络配置

为了更好的表现压缩卡的效果，我们将100GE网络限速到25GE，模拟25GE网络带来的效果。

测试环境拓扑图

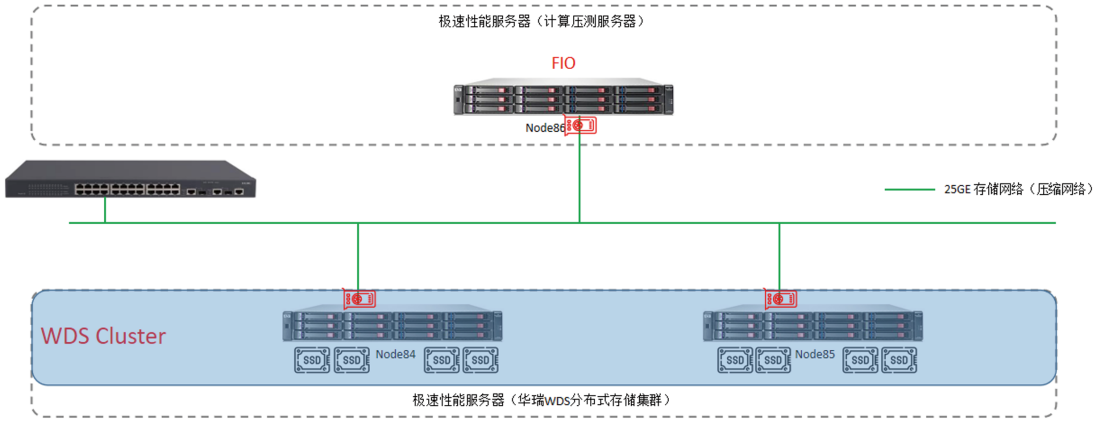


图13 英特尔®QAT压缩卡与WDS整合测试环境拓扑图

软件配置

- **基线压测工具**

  FIO：FIO–3.7

- **分布式存储软件：华瑞指数云WDS**

  版本：WDS 2.0.0

  网络协议：RoCEv2

  块存储协议：NVMe–oF(NVMe over RDMA)

测试方法

使用FIO工具使用单节点压测，将带宽压满，观察压缩比，分析实际带来的带宽提升效果。

3.4.3 测试数据

IO模型	块大小	不开QAT压缩	开QAT压缩	压缩比
randwrite	128k	IOPS=11.6k, BW=1456MiB/s	IOPS=15.4k, BW=1930MiB/s	1.32
	512k	IOPS=2912, BW=1456MiB/s	IOPS=4323, BW=2162MiB/s	1.48
	1024k	IOPS=1456, BW=1456MiB/s	IOPS=2241, BW=2242MiB/s	1.53
randread	128k	IOPS=23.3k, BW=2913MiB/s	IOPS=28.6k, BW=3575MiB/s	1.22
	512k	IOPS=5826, BW=2913MiB/s	IOPS=7153, BW=3577MiB/s	1.22
	1024k	IOPS=2913, BW=2913MiB/s	IOPS=3577, BW=3577MiB/s	1.22
write	128k	IOPS=11.7k, BW=1457MiB/s	IOPS=16.5k, BW=2064MiB/s	1.41
	512k	IOPS=2913, BW=1457MiB/s	IOPS=4331, BW=2166MiB/s	1.48
	1024k	IOPS=1456, BW=1457MiB/s	IOPS=2211, BW=2211MiB/s	1.51
read	128k	IOPS=23.3k, BW=2913MiB/s	IOPS=28.5k, BW=3564MiB/s	1.22
	512k	IOPS=5826, BW=2913MiB/s	IOPS=7134, BW=3567MiB/s	1.22
	1024k	IOPS=2913, BW=2913MiB/s	IOPS=3567, BW=3568MiB/s	1.22

3.4.4 数据结果分析

我们集成了英特尔®QAT压缩卡进行测试，使用比较随机的数据进行测试，依然可以看出压缩的效果，借助于英特尔®QAT压缩卡的能力，25GE网卡可以达到总带宽3GB/S，实现业务带宽大于网卡物理带宽的效果，实测数据表明带宽有20%~50%的提升，压缩比能够达到20%~40%。本次测试中使用的是随机数据，如果一些有特征的业务数据，预期将会有更加好的效果。

在真实的业务场景WDS及WiDE借助于英特尔®QAT的方案可以大大提升互联带宽，并使用比较低带宽的网卡实现高带宽的效果。

3.5 总体测试结果分析

通过在实际物理环境上的一系列测试，充分说明了WiDE引擎以及基于WiDE引擎的新一代分布式存储系统WDS采用的全新设计的分布式系统架构以及极速IO栈具备极为高效的IO处理能力，可以提供极高性能和极低时延，同时还具备优异的并发能力和接近无损的水平扩展能力。

在测试1中我们测试了WDS的基准存储性能以及运行数据库的性能，测试结果表明，其在3台服务器节点，100G RDMA网络，一共12块英特尔®傲腾™固态盘P5800X的集群环境上，4k随机写可以达到324万IOPS（2副本）和244万IOPS(3副本)，4k随机读达到692万IOPS，在达到惊人的IOPS的同时，还能保持稳定的500us以内的时延。

与系统调优后的Ceph Nautilus版本相比，实现了IOPS性能的数量级的惊人提升，无论是单卷性能还是集群性能均提升高达10到30倍以上，与此同时IO时延缩短了90%，从Ceph Nautilus版本的5ms左右大幅降低到WDS的500us左右。

在Mysql数据库的Sysbench测试中，验证访问16个表的每秒事务处理性能(TPS),WDS 2.0版本可以在稳定的5ms以内时延，提供高达20000以上的TPS，而Ceph Nautilus版本没有任何办法可以提供5ms以内的稳定时延，将时延要求放宽到15ms以后，也只能提供数千TPS。

在测试2中，我们验证了WiDE引擎和WDS产品的峰值性能和水平扩展性，测试结果充分说明了WiDE分布式引擎已经把分布式软件栈的效率优化到了接近峰值，可以充分发挥所有新型硬件的峰值能力，甚至能将英特尔®傲腾™固态盘P5800X这样的目前市面上运行速度标杆性的SSD的能力发挥到峰值。随着硬件本身的性能提升或者硬件节点和数量的增加，端到端的分布式存储集群性能也可以近乎无损的线性增长，可以轻松达到上亿IOPS。

我们对WDS产品的性能水平扩展能力做了一些估测：
如果在物理节点上使用双口100GE网卡，性能表现可达到：

性能能力	IOPS
单卷峰值性能（4k）	4,000,000+ IOPS
3节点峰值性能（4k）	12,000,000+ IOPS

如果由三个物理节点线性扩容到更多的物理节点，可以实现性能的同比例飞跃：

性能能力	IOPS
3节点峰值性能（4k）	12,000,000+ IOPS
6节点峰值性能（4k）	24,000,000+ IOPS
12节点峰值性能（4k）	48,000,000+ IOPS
16节点峰值性能（4k）	64,000,000+ IOPS
24节点峰值性能（4k）	96,000,000+ IOPS

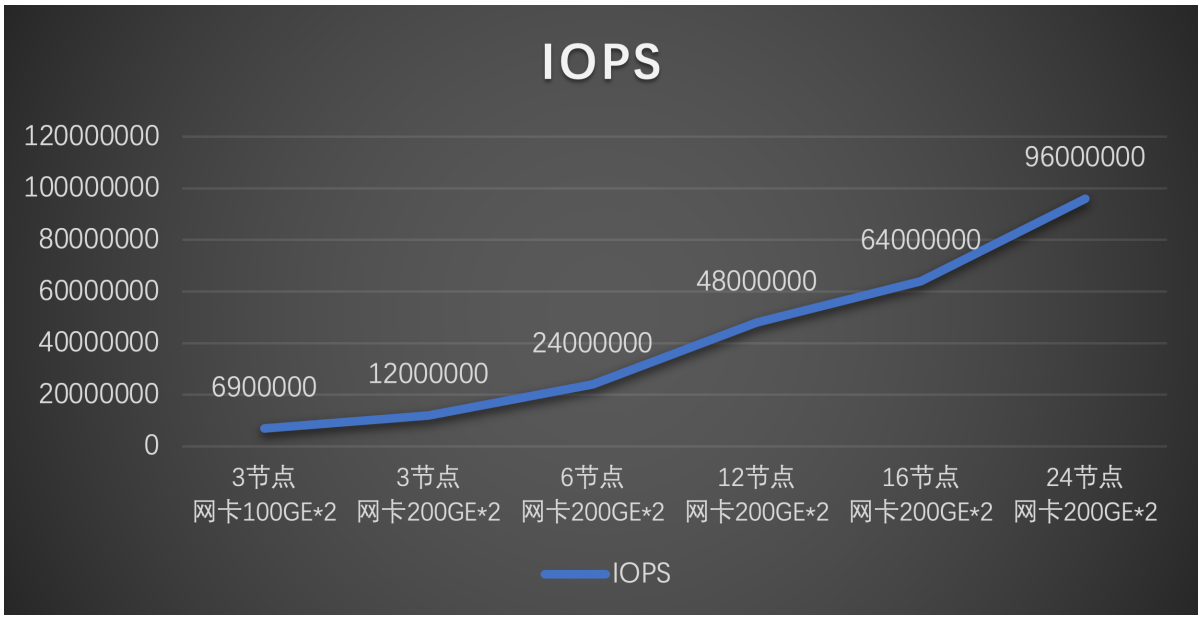


图14 性能水平扩展

在测试3中，通过集成英特尔®QAT卡硬件加速的数据压缩能力，WDS可以使用比较低带宽的网卡实现高带宽的效果，带宽有20%~50%的提升，让整个分布式存储系统的性能得到进一步的扩展。

4 结论

在机械硬盘时代，单块机械磁盘的性能很低，IOPS一般在120–150之间，因此存储系统的性能瓶颈主要体现在磁盘数量上，存储控制器很少成为制约IO处理能力的关键因素。随着介质进化到全面使用NVMe SSD，傲腾等新技术，单盘的性能比机械硬盘提升数百倍乃至上千倍，可以达到数十万/百万IOPS，这时传统集中式存储架构的控制器数量就容易变成系统性能瓶颈，并且控制器最大扩展数量也非常有限，会出现系统扩容更多容量时，性能几乎不会有任何提升，昂贵的SSD介质的能力无法充分发挥出来的情况，导致资源的极大浪费。此外，集中式存储结构与云计算时代要求的水平扩展，资源池化等特征也天然不适配。

分布式存储架构由于是水平扩展，每一个节点都相当于既是控制器同时也提供存储容量，因此系统的总体性能能够随着集群的节点数和容量线性扩展，有望成为充分释放新一代介质的潜力和新一代硬件技术红利的最佳选择。然而，当前的分布式存储引擎和相关产品，只能提供毫秒级的时延，并且时延还不太稳定，主要只能覆盖非结构化海量数据场景以及对性能和稳定性要求不高的一般结构化数据场景，无法对传统存储阵列形成更大范围的替代，还远远不足以作为全闪介质时代和全面云化、数字化时代的主力架构。

华瑞指数云WiDE这样的新一代分布式存储引擎实现了分布式系统架构的IO处理效率的10倍级的提升，时延缩短到十分之一，适应分布式云环境的数据的统一管理和智能IO调度，性能和容量的无级水平扩展，并且与英特尔第三代至强处理器相结合能够充分发挥更优的性能，实测的集群性能可以轻松达到几千万IOPS以及微秒级时延，将大大的扩大分布式存储架构的适用场景和应用范围，并且由于其软件定义的设计，只需要使用标准商用服务器硬件，单TB成本远低于传统高端存储，多池多故障域的数据流动和分布式架构容错设计，大幅度的提升了系统的可靠性，成为全闪介质时代以及云原生时代实现高效的数据存储的最佳选择。原来企业客户不得不采购高端全闪的一些业务场景，现在也可以采用基于WiDE引擎的分布式存储，实现与云计算平台架构的适配，实现关键业务真正可以在云原生环境上稳定高效运行。WiDE引擎还可以提供多池架构下的IO调度和数据流动，这样在一个平台上既可以存储海量非结构数据，也能存储要求高性能高可靠的结构化数据，还可以做高性能的数据分析，真正实现数据原生于一个数据平台上，只保留一份数据却可以被各类应用以各种接口访问，避免各种数据孤岛和数据复制拷贝带来的问题。WiDE分布式存储引擎及相关产品的出现，将直接推动SDS和分布式存储进入2.0时代，引领存储系统架构的变革，并逐步成为企业新一代数据架构的基石。

关于华瑞指数云（ExponTech）

华瑞指数云（ExponTech）是基于新一代分布式架构的数据基础设施整体解决方案提供商，是软件定义存储2.0 (SDS2.0)以及存数一体化架构的业界首创者和技术领导者，致力于帮助企业和组织建设“以数据为中心”的新型IT基础架构，自主研发下一代分布式数据存储，数据管理，新一代大数据平台，混合多云数据平台等面向数据和数字化经济的基础软件产品，系统和解决方案，实现在一个统一的数据平台上存储，管理和分析企业的全场景数据，构建数字化的基石，充分发挥数据的价值。

ExponTech的员工90%以上为技术人员，核心团队来自华为以及腾讯、阿里、IBM、微软、EMC等全球500强企业，60%的员工拥有研究生及以上学历。团队骨干曾经成功主持研究和开发了中国第一代商业化分布式存储，私有云产品和公有云平台，受到中国以及欧洲的领先企业和政府机构客户广泛使用。

关于英特尔

英特尔是半导体行业和计算创新领域的全球领先厂商，创造改变世界的科技，造福地球上每一个人。英特尔创始于1968年，拥有50余年推动技术创新和耕耘产业生态的成功经验。如今，英特尔正转型为一家以数据为中心的公司，致力于做可信赖的性能领导者，释放数据无限潜能。英特尔与合作伙伴一起，推动人工智能、5G、智能边缘等转折性技术的创新和应用突破，驱动智能互联世界，帮助解决人类面临的重大挑战。

英特尔1985年进入中国，员工超10,300人，在中国有22个办公地点。中国是英特尔全球战略之重，拥有除美国总部外最全面的业务部署，覆盖前沿研究、产品技术开发、精尖制造、产业生态合作、市场营销、客户服务、风险投资和企业社会责任等。

英特尔扎根中国36年，与中国产业伙伴的合作久经考验、风雨同舟。作为在中国扎根最久的跨国公司之一，英特尔在中国的战略从未变过，始终如一，就是“做正确的事”，就是与中国同行远行。

ExponTech

华瑞指数云科技有限公司

北京：朝阳区广顺北大街33号院

深圳：南山区尚美国际大厦A座4011

西安：高新区丈八五路10号科技资源统筹中心A座2层

成都：武侯区人民南路四段27号商鼎国际1–1–17–2

洛阳：伊川高新四路华瑞科技园

长沙：开元东路192号世景国际B栋2030

WiDE分布式存储引擎
和英特尔®傲腾™固态盘
联合解决方案技术白皮书

联系电话：400–100–5719

邮箱：support@expontech.com

网站：www.expontech.com