解决方案简介

2022 年 6 月

基于英特尔® 技术的浪潮云海® 超融合一体机 InCloud Rail 解决方案

“新一代超融合不仅是存储架构的融合，更是私有云方案的融合、异构算力的融合、云原生的融合、云数智一体化的融合。通过与英特尔合作推进英特尔® SPDK 在浪潮云海® 超融合一体机 InCloud Rail 中的应用，我们有效消除了分布式存储在内核上下文切换、CPU 中断等方面的性能瓶颈，充分发挥出存储介质的性能潜力，满足了金融和医疗等行业在不同应用场景中的性能需求，加速企业的智慧转型。”

— 张东

浪潮数据技术有限公司董事长

“超融合将存储、计算和网络融合到一个统一的系统中。凭借其简洁、灵活和可扩展性的特点，正在体现强大的方案优势，并获得市场的认可。英特尔与浪潮合作，致力推进超融合一体机的性能提升。为满足客户需求，加速市场智慧转型，持续贡献力量。”

— 李亚东

英特尔中国政企及全球 OEM 解决方案事业部总经理

概述

作为一种软硬一体化的基础设施架构，超融合具备易运维、易部署等典型优势，在多种行业与场景中得到了广泛应用。浪潮聚焦云原生、5G、大数据、云边端等应用场景，结合自身软硬件能力，近年来持续推动浪潮云海® 超融合一体机 InCloud Rail 的创新与演进。该一体机通过 InCloud DataCenter 云管理平台，支持跨云平台融合和异构虚拟化环境的集中管理。一体机同时搭载基础环境快速部署工具，部署速度远超传统解决方案。

在金融、医疗等行业的部署中，面向数据库等场景的超融合一体机面临着巨大的存储性能挑战，用户需要尽可能地提升数据吞吐能力并降低时延，以支撑关键型业务的高效运行。为了提升存储子系统的性能，浪潮使用了英特尔® 存储性能开发套件（SPDK），并通过 NVMe over Fabrics（NVMe-oF）协议进行加速。NVMe-oF 协议作为 iSCSI 协议的替代者，可以让主机以使用本机 NVMe 协议方式访问分布式存储，提供低延时、高吞吐的块存储设备，解决了内核态驱动读写 NVMe 磁盘时可能会出现的内核上下文切换和 CPU 中断问题。优化后的方案能够为客户虚拟机提供高性能的分布式存储，降低总体时延和减少时延差异，满足金融和医疗客户在数据库等多个场景中的需求。

挑战：金融与医疗行业面临存储性能瓶颈

由于用户规模庞大、业务与数据价值高等因素，金融和医疗机构一直是信息化和技术应用方面的先行者，业务发展对 IT 系统的依赖度非常高。近年来，随着金融与医疗机构

纷纷开启数字化转型，如何优化与重构 IT 基础设施，为上层应用和业务创新提供灵活可靠的 IT 服务平台，已经成为其数字化战略的发展重点。

在此背景下，越来越多的金融与医疗机构开始拥抱超融合，希望通过部署超融合一体机等方式，对硬件加以重构，以软件定义的方式打造灵活高效的 IT 基础设施，以便降低 IT 基础设施运维和扩展的压力，获得更大的业务弹性，从而将更多资源用于拓展业务。

在拥抱超融合的同时，金融与医疗机构也非常关注超融合能否提供强大的存储能力。伴随着用户增长与业务创新，金融与医疗机构内部场景化、移动化、碎片化的数据在快速增长，这给存储系统带来了巨大的压力。以金融数据库应用为例，互联网交易、数据风控、实时营销等应用需要数据库实现快速响应。而在大并发量的业务背景下，存储系统很容易成为性能瓶颈，影响业务的整体响应能力。

导致存储系统性能瓶颈的原因之一是存储介质。最近几年，存储介质得到了快速发展，如今 NVMe SSD 的 IOPS 已经远远高出 HDD 磁盘，时延从毫秒压缩到微秒，系统的性能瓶颈也由存储硬件本身逐渐转移到网络及处理器上。传统文件系统和调度器等方法由于无法充分发挥新存储介质的性能，成为存储系统的新瓶颈。这些瓶颈包括：

- 利用常规的 NVMe 内核驱动读写 NVMe 磁盘时，会遇到内核上下文切换和 CPU 中断问题。在高性能的全闪存储中，中断意味着时延的不确定，会导致较大时延和性能开销。
- 在传统的 I/O 模型中，应用程序提交读写请求后进入睡眠状态。待 I/O 完成后，中断会将其唤醒，中断开销成为了整个 I/O 时间中非常重要的一部分。

解决方案：基于英特尔® 技术的浪潮云海® 超融合一体机 InCloud Rail

作为新一代超融合解决方案，浪潮云海® 超融合一体机 InCloud Rail 通过软件定义的计算、存储和网络技术实现了服务器的资源池化，使整个 IT 环境比单独的物理硬件具有更高的可用性、安全性和扩展性，能够有效满足企业对于降低成本、简化管理、提高安全性和扩展性的需求，助力企业向云计算迁移核心业务，构建企业云数据中心。

浪潮云海® 超融合一体机 InCloud Rail 实现了存储资源的池化和统一管理，并通过全闪存架构的软件定义存储 SmartONE 支持异构算力的融合。SmartONE 采用 etcd 组件作为集群管理，负责分布式存储的节点的关系、节点之间的元数据传递及一致性等。在资源管理方面，SmartONE 提供了 qemu、iSCSI 和 NVMe-oF 的服务入口，对接 VDI 卷管理入口，对外提供存储资

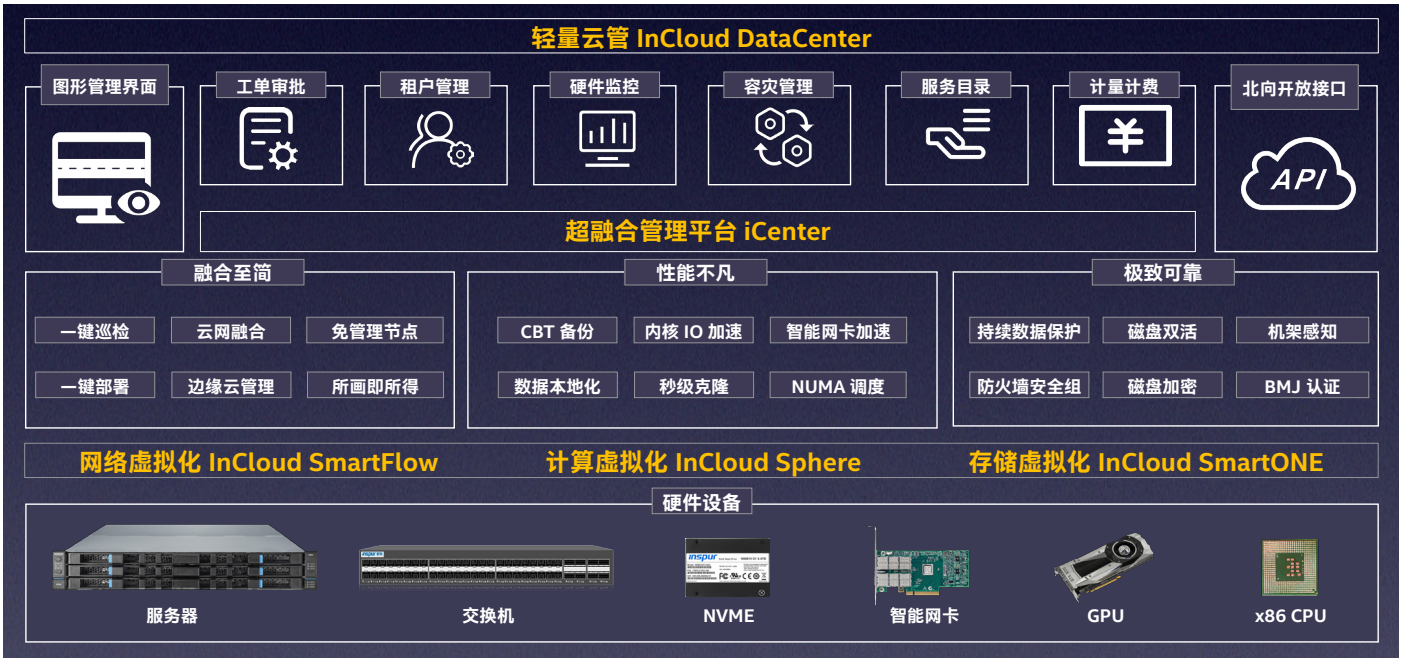

图 1. 浪潮云海® 超融合一体机 InCloud Rail 架构

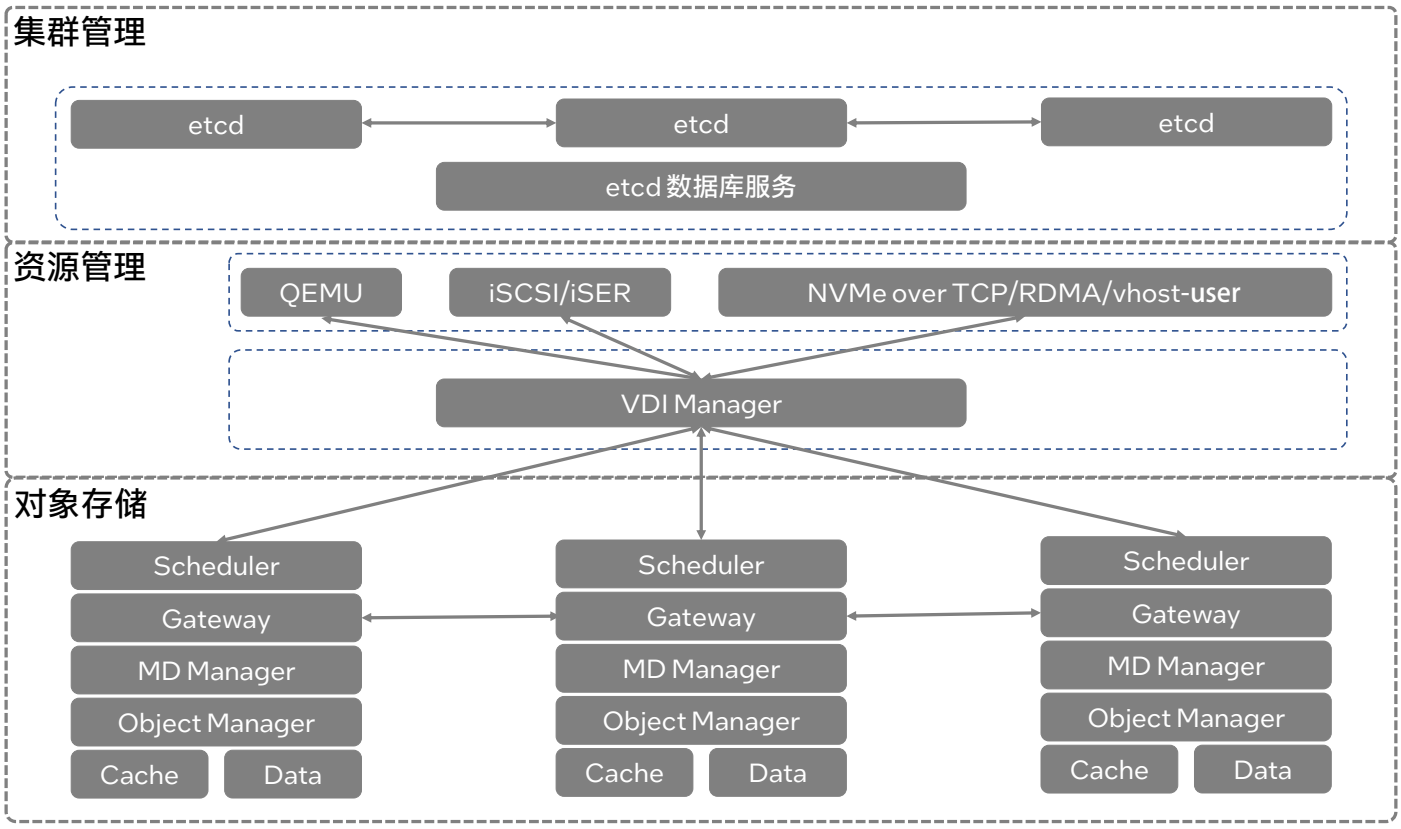

图 2. 浪潮 SmartONE 软件架构

源。在对象存储方面，SmartONE 主要接收 VDI 的 IO 请求，根据数据分布算法调度到相关节点，根据副本策略和 EC 规则调用 Gateway 分发到分布式存储的存储节点。

为了满足数据库等实时读写、随机访问超大规模数据集等场景对于存储性能的严苛要求，浪潮云海® 超融合一体机 InCloud Rail 除了探索使用新一代存储介质之外，还与英特尔合作，采用了英特尔® 至强® 可扩展处理器和英特尔® 以太网适配器 E810，并从存储引擎层面来化解性能瓶颈。

- **英特尔® 至强® 可扩展处理器：**该处理器专为数据中心现代化革新而设计，能够提高各种基础设施、企业应用及技术计算应用的运行效率，进而改善总体拥有成本（TCO），提升用户生产力。它拥有更高的单核性能，能够在计算、存储和网络应用中，为计算密集型工作负载提供卓越的性能和可扩展性。
- **英特尔® 以太网适配器 E810：**该网络适配器具备 100/25GbE 性能，支持单个或双端口连接，在 PCIe 4.0 x 16 插槽中提供了出色的性能，并支持应用程序设备队列（ADQ）、动态设备个性化（DDP）、RDMA iWARP 和 RoCEv2 等各种高级功能，能够有效提升工作负载的可预测性与吞吐率，同时降低应用时延。

浪潮采用英特尔® SPDK 优化存储性能

英特尔® SPDK 提供了一组工具、库和方案，用于编写高性能和可扩展的用户态存储应用程序。它通过使用多种关键技术来实现高性能和高扩展，诸如将一些驱动程序移至用户空间，避免了系统调用，并允许从应用程序进行零拷贝访问。它通过无锁化、消息机制和异步编程实现高性能应用框架，同时提供统一的用户态通用块设备来高效管理不同的存储后端设备。

使用英特尔® SPDK 之后，用户态的驱动通过轮询硬件而不是依赖中断来完成，可以有效降低总时延和减少时延差异，同时和内核驱动相比，在每个 CPU 内核的 IOPS 上具有更明显的性能优势。此外，英特尔® SPDK 具备 I/O 路径的无锁高性能模式，避免了所有在 I/O 关键路径中的锁，而是依靠消息传递在多个线程中共享资源，从而提高了并行性。

浪潮与英特尔合作，在 SmartONE 分布式存储的单机存储引擎、NVMe-oF 存储服务等模块中，采用了英特尔® SPDK 进行优化。

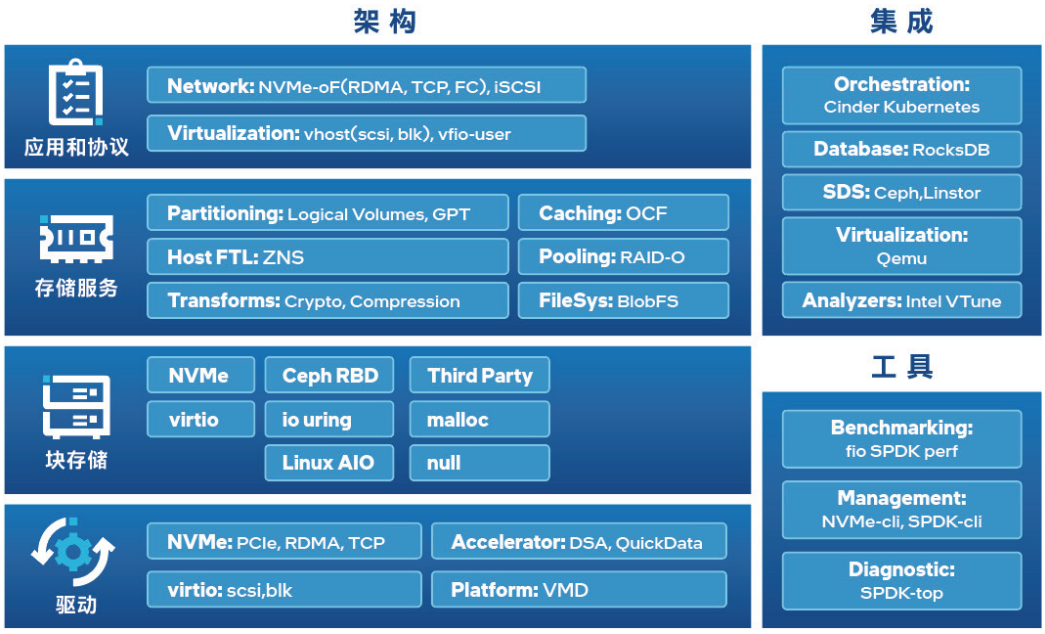

图 3. 英特尔® SPDK 架构

单机存储引擎

浪潮 SmartONE 分布式存储基于英特尔® SPDK 的 NVMe 驱动实现了高性能底座，支持和 NVMe 磁盘设备直接交互，同时采用无锁设计，并行处理 IO 命令。此外，NVMe 上层实现了专属 NVMe 的单机存储引擎系统，该引擎系统通过基于内存的元数据和日志管理系统，有效避免了传统文件系统的双写问题。

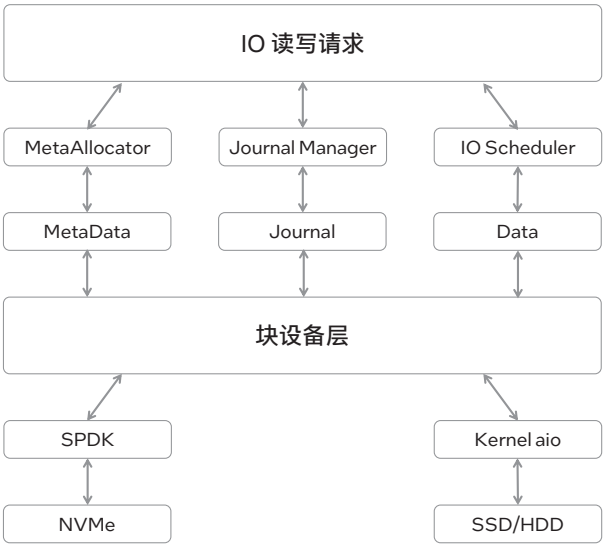

图 4. 专属 NVMe 的单机存储引擎系统

基于英特尔® 技术的新一代单机存储引擎实现了性能的显著提升。测试数据显示，浪潮 SmartONE 分布式存储的单机存储引擎几乎达到了 NVMe 物理硬盘支持的性能上限。

NVMe-oF 分布式存储服务

NVMe-oF 分布式存储服务为超融合平台提供两种块服务解决方案，其一是采用 vhost-user 技术方案，可以为虚拟机提供更短的 IO 路径；其二是作为存储服务，为服务器提供 NVMe-oF 的 TCP/RDMA 的块服务。SmartONE 利用 SPDK vhost-user 技术，直接消除 Guest 虚拟机通过 PCIE 方式访问 NVMe 设备，避免内核更新 PCI 配置空间；同时直接在用户态捕获 QEMU 虚拟 IO，以零拷贝方式将数据传输到存储系统中。

SmartONE 支持以 NVMe-oF 存储协议方式提供块存储服务，它可以提供 TCP 和 RDMA 两种形式的外部访问。NVMe-oF 存储协议作为 iSCSI 协议的替代者，可以让主机以使用本机 NVMe 协议的方式访问分布式存储，提供低延时、高吞吐的块存储设备。

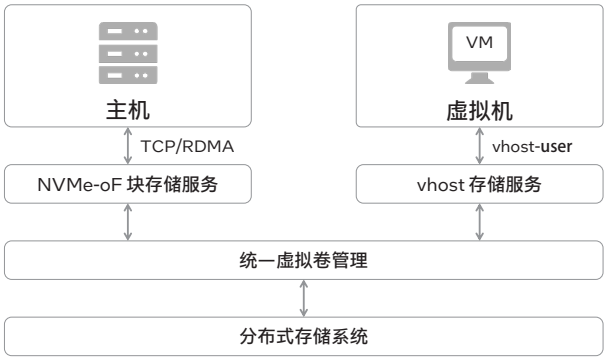

图 5. SmartONE NVMe-oF 分布式存储服务

为了验证 SmartONE NVMe-oF 分布式存储服务的性能表现，浪潮分别对比了 iSCSI 存储协议在浪潮 M5 系列服务器[1] 和浪潮 M6 系列服务器[2] 上的性能表现，以及 NVMe-oF 存储协议在浪潮 M6 系列服务器上的性能表现（NVMe 副本模式，3 副本）。

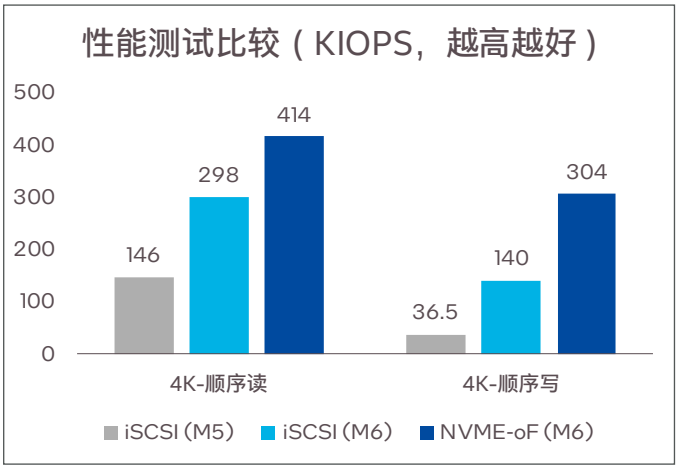

图 6. iSCSI 与 NVME-oF 存储协议在 M5/M6 平台上的 4K 读写性能比较[3]

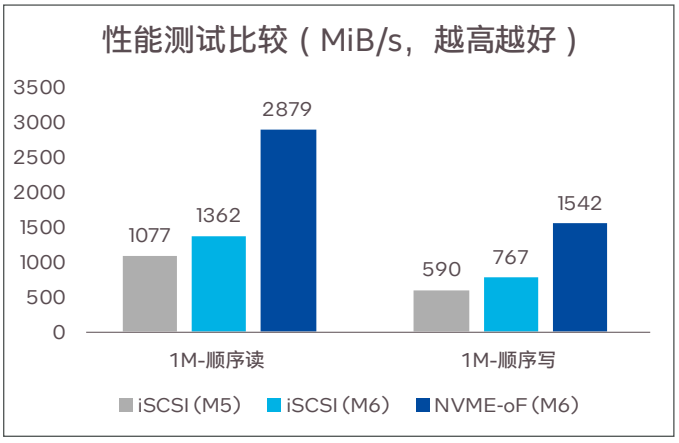

图 7. iSCSI 与 NVMe-oF 存储协议在 M5/M6 平台上的 1M 读写带宽比较[4]

测试数据显示，NVMe-oF 存储协议的 4K 读写性能和 1M 读写带宽均有显著提升，能够出色地满足金融、医疗关键型应用对于存储性能的严苛要求。

收益：消除性能瓶颈，打造高性能存储

得益于英特尔® 软硬件产品与技术的应用，浪潮 SmartONE 分布式存储系统能够充分发挥 NVMe 固态盘的性能潜力，应对在 IOPS 和时延方面有着较高要求的数据库应用场景。

- 显著提升存储系统的性能，并降低数据时延，能够满足有着苛刻要求的数据库应用场景，同时有效控制分布式存储系统的总体拥有成本（TCO）；
- 充分发挥英特尔® 硬件的性能优势，并从新一代英特尔® 硬件的创新中获益；
- 集成于超融合一体机中，能够通过一体机所预置的虚拟化平台、云管理平台、快速部署工具以及相关的工作流程，显著简化部署、管理和运维工作的复杂性。

展望

面向云数智一体的应用环境，浪潮将继续强化包括存储能力在内的超融合系统的创新，采用新一代英特尔® 至强® 可扩展处理器、英特尔® 傲腾™ 持久内存、英特尔® 以太网适配器等硬件产品，以及领先的软件方案，在高性能、高可靠、易运维、易扩展四个层面提升核心竞争力，满足全行业、全场景的需求。

浪潮还计划选择英特尔® 傲腾™ 持久内存作为单机存储引擎缓存层的存储介质。英特尔® 傲腾™ 持久内存是一项变革性的内存技术，提供了融合高速、高性价比、大容量、持久数据保护、高级加密等优势于一体的内存选项。其结合英特尔® PMDK 使用，可以支持应用直接访问持久内存设备，而不需要经过文件系统的页高速缓存系统、系统调用和驱动，从而能够降低 I/O 过程的开销，显著缩短数据时延。

未来，浪潮还将坚持以用户为中心，紧密结合前沿技术发展趋势与市场需求，持续创新引领、迭代优化，打造集融合至简、性能不凡、强大可靠、开放生态特性于一身的超融合一体机，持续为用户业务上云、数字化转型升级贡献力量。

[1] 浪潮 M5 系列服务器基于第二代英特尔® 至强® 可扩展处理器。更多信息请访问：https://www.inspur.com/lcjtww/2509917/2509945/2510106/index.html。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。

[2] 浪潮 M6 系列服务器基于第三代英特尔® 至强® 可扩展处理器。更多信息请访问：https://www.inspur.com/lcjtww/2509917/m6/index.html。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。

[3,4] 数据援引自浪潮与 2022 年 6 月开展的测试。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。

关于浪潮

浪潮集团是中国领先的云计算、大数据服务商，拥有浪潮信息、浪潮软件、浪潮国际三家上市公司。主要业务涉及云计算、大数据、工业互联网、新一代通信及若干应用场景。已为全球一百二十多个国家和地区提供 IT 产品和服务。浪潮是中国最早的 IT 品牌之一，一直秉承创新的理念，数次在中国信息产业发展的重要历史阶段，通过提供领先技术，提升竞争实力，成为新一代信息技术领军企业，全面服务经济社会的数字化转型和高质量发展。

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®