intel + 阿里云

intel XEON

英特尔携手阿里云
构建数智基石
加速云上创新

云同行 AI加速

Contents
目录

案例篇 - 千行百业落地

产品篇

以数据为中心的硬件产品组合

软件及系统级优化

基础设施算力优化

基础设施存储优化

基础设施网络优化

操作系统和编排层优化

分析及 AI 性能优化

媒体服务应用优化

趋势篇

云服务迎挑战、频创新，为企业数字化转型打造坚实技术基座

在科技革命与产业变革浪潮持续推进的今天，运用前沿技术力量，积极拥抱数字化转型赋能创新发展，无疑已成为各领域企业客户应对变化的“必选项”，同时也推动着数字化投资的持续增长。数据显示，有超过半数的企业正计划进一步加大数字化投资力度，以保持业务竞争优势[1]。这些数字化投资，既包括使用人工智能（Artificial Intelligence, AI）和大数据（Big Data）实施业务创新、实现决策优化，也包括依托云服务（Cloud Service）和物联网（Internet of Things, IoT）来提升业务效率、推动降本增效。而这进一步表明采用先进技术加速创新与变革，已成为企业赢得未来竞争力的重要抓手。

数字化创新与变革的背后，离不开高效 IT 基础设施的支持。得益于在性能、敏捷性以及弹性可扩展等方面的优势，云服务正在事实上成为一系列数字化转型的基座；与此同时，产业应用场景的丰富也推动着云服务需求的高速增长。国际数据公司（IDC）发布的《中国公有云服务市场（2022 年下半年）跟踪》报告显示，2022 年下半年，中国公有云服务（IaaS/PaaS/SaaS）整体市场规模已达 188.4 亿美元[2]。越来越多的企业客户正通过阿里云等云服务提供商的开放平台，加速上云、用云的进程，并推动业务系统与云服务的融合。

更深、更广和更多维度的应用场景，以及与客户创新商业模式和业务流程的融合，也使云服务面临着更严峻和更多元化的挑战。例如，在科学计算、AI 等应用中，性能成为至关重要的因素；在直播、游戏等互联网应用中，接入能力、弹性可扩展能力则成为关键需求；而随着传统行业、政企行业的上云，云上数据安全防护也面临着前所未有的压力。这些差异化的需求无疑对云服务提供商也提出了进一步的要求，包括：

- **以算力、吞吐量为代表的性能要求**：用于满足强性能、高并行和大体量的云上工作负载高效执行所需；
- **差异化需求在不同架构上的一致性要求**：用于实现不同类型的云上工作负载在公有云、私有云、混合云等不同平台架构上的一致性体验；
- **数据共享带来的云上安全新要求**：用于消除客户在利用更丰富数据构建更优 AI、大数据等应用时的疑虑，更好实现数据价值及“数据可用不可见”；
- **实现绿色节能新要求**：用于在国家双碳政策下，达成云服务低碳化可持续发展要求所需的能效目标。

为应对这些新的要求，阿里云等云服务提供商正通过对千行百业云服务的应用模式和发展趋势的敏锐洞察，一方面

图 1-1-1 数字企业进化图（来源：埃森哲）

[1] 数据援引自由埃森哲发布的《2022企业数字化转型指数》：https://www.accenture.cn/cn-zh/insights/strategy/china-digital-transformation-index-2022
[2] 数据援引自由IDC发布的《中国公有云服务市场（2022下半年）跟踪》：https://www.idc.com/getdoc.jsp?containerId=prCHC50595423

借助新一代的软硬件产品与技术，来推进云基础设施优化；另一方面也积极推动云平台架构的不断变革以及云服务技术的持续创新，从而打造一系列灵活、高性能、高可用的解决方案，来为各行各业的数字化转型提供加速引擎。这些变革与创新也推动一系列新技术、新架构和新服务在云服务应用场景中的落地与实践，并展现出多彩的新趋势，包括：

■ 混合云与多云架构更广泛应用

对云服务能力的多样化诉求，往往无法通过单一类型的云平台来承载。结合容灾性、业务可行性以及成本议价等方面的因素，混合云和多云策略正成为企业近年来部署 IT 基础设施的普遍选择，如图 1-1-2 所示，数据表明，目前有 89% 的企业与机构正在打造其多云策略[3]。

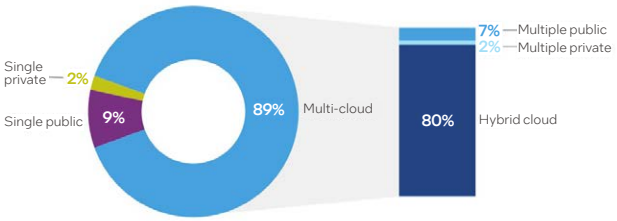


图 1-1-2 更多企业与机构正在打造多云策略[4]

通过有机融合公有云与企业本地的计算存储资源（包括私有云及本地 IT 基础设施等），并引入多云管理能力，混合云与多云策略能够发挥不同优势，帮助企业拓展云部署模式、丰富云服务类型，满足企业在新时期的数字化转型需求。

■ 云原生技术日益受到青睐

随着技术的不断完善和能力的持续提升，云原生正以全生命周期的技术链帮助企业革新 IT 架构，从优化云基础设施管理、云资源调度，到统一协同运维等能力，全方位帮助企业构建更加灵活的用云模式，提高云资源使用效率与服务敏捷性、降低总体拥有成本（Total Cost of Ownership, TCO）并提升服务质量，从而获得更多客户的青睐。

图 1-1-3 云原生技术日益普及

以云原生技术推动的容器云为例，其在公有云市场上的渗透率正在不断提升，有预测数据表明，到 2025 年，将有 50%-75% 的云应用会迁移到容器架构，同时私有云容器云平台市场的增速也将继续保持高位[5]。

■ 丰富的云应用驱动云优化策略

日趋多元化、动态化的云上负载，正驱动云服务提供商提供更具差异化的服务能力。以当前火热的 AI 为例，数据表明，未来数年中 AIaaS（AI as a Service，AI 即服务）市场的 CAGR（复合年均增长率）或达 40% 以上[6]。

为此，云服务商需要采用一系列的优化策略，来保证云服务的性能、敏捷性和可用性。例如，通过丰富的软件栈来实现异构计算、资源编排等能力，实现对云资源更有效地利用；或通过引入 DPU、IPU 等专用芯片产品，提升云平台承载高密度工作负载的性能表现。

■ 云上机密计算受到持续关注

借助云计算的敏捷性和灵活性来打破数据孤岛，进而提升数字化转型效能，是客户选择云服务时的关键诉求之一。但数据安全性一直是实现敏感数据价值挖掘的巨大障碍。

在云上部署的机密计算方案，可以有效帮助客户实现以上目标。数据表明，有 88% 来自数据敏感型行业的客户将采用机密计算，来对敏感型工作负载中的数据实施保护[7]。

[3、4] 数据援引自由Flexera发布的“2022 State of the Cloud Report”（2022 年云状态报告）：https://resources.flexera.com/web/pdf/Flexera-State-of-the-Cloud-Report-2022.pdf
[5] 数据援引自由艾瑞发布的“2020年中国容器云市场研究报告”：https://report.iresearch.cn/report/202012/3701.shtml
[6] 数据援引自由MarketsandMarkets发布的“AI as a Service Market by Offering (SaaS,PaaS,IaaS),Technology (Machine Learning, Natural Language Processing,Context Awareness,Computer Vision),Cloud Type (Public,Private, Hybrid),Organization Size,Vertical and Region -Global Forecast to 2028”：https://www.marketsandmarkets.com/Market-Reports/artificial-intelligence-ai-as-a-service-market-121842268.html
[7] 数据援引自由英特尔官网发布的《机密计算：催生云端新可能》：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/confidential-computing-possibilities-cloud.html

■ 液冷技术推进绿色云服务

作为云服务的算力基座，数据中心的能耗一直是颇受关注的焦点，这不仅关系到云服务的成本把控，更具有助力推进中国“双碳”政策实施的战略意义。近年来，云服务商一直致力于通过前沿的散热技术，如全液冷技术等降低 PUE（Power Usage Effectiveness，电能利用效率），提升数据中心绿色指数，加强可持续服务水平。

英特尔多元化产品与技术为云服务能力提升提供全量支持

得益于对云服务发展趋势的清晰洞察、在云服务市场的长期实践，以及与众多合作伙伴的协同努力，英特尔正围绕至强® 可扩展平台，借助其不断丰富的软、硬件产品体系及创新技术，为云服务商应对上述需求，拥抱新趋势提供更佳解决之道。

■ 面向数据中心的英特尔全栈产品组合不断丰富

云服务能力的不断提升需要从计算、存储到网络等多方面的全量基础设施支撑，其不仅需要一系列兼具高性能、可靠性、灵活性和安全性的产品与技术，作为其运营的必要物质和能力基础，还需要通过多种架构、软件优化策略，来充分释放硬件潜能，从而为客户所需的不同云负载提供强有力支持。英特尔面向数据中心，不断丰富的全栈产品组合，可为云服务平台提供有效的能力加成，具体包括：

- **更强算力输出**：工欲善其事，必先利其器，强劲的算力是高品质云服务的根本。英特尔以丰富的产品形态，为云平台打造了不同维度的算力引擎。从应对不同计算负载的通用处理器（英特尔® 至强® 可扩展平台），到擅长游戏、视频编解码的数据中心 GPU（英特尔® 数据中心 GPU Flex 系列、英特尔® 数据中心 GPU Max 系列）；从卸载处理器负载，加速平台数据处理效率的专用芯片（英特尔® IPU），到灵活实现服务验证的 FPGA（英特尔® FPGA 产品），不同的英特尔算力引擎，正为多元化的云服务提供强有力的算力输出。

- **更多架构优化**：在基础算力之外，英特尔还充分利用自身丰富的产品组合与软件工具，来开展架构优化。例如，对不同的云服务架构，从公有云、混合云到边缘云，英特尔都提供了端到端、向云而生的产品和解决方案，并以全栈软件开展技术生态链适配和优化。而对于云原生涉及的平台容器化、应用微服务化以及开发 / 运维一体化，英特尔也提供了云原生平台架构方案（包括硬件加速、软件组件实现、服务增强以及负载优化等），来予以强力支持。

- **更优应用加速**：为满足云上不同负载（例如 AI、科学计算、视频处理等）的高效处理需求，英特尔通过导入多种软硬件技术和工具来提供加速。例如，面向 AI 应用的高级矩阵扩展加速引擎（英特尔® AMX）、实现模型框架量化的软件工具（OpenVINO™ 工具套件）、面向数据处理与分析用的各类框架及软件工具（DAAL、BigDL 等），以及面向媒体服务的软件工具（英特尔® MediaSDK、SVT 等）。以 AI 应用为例，数据表明，目前已有 70% 的数据中心 AI 推理任务在英特尔® 至强® 可扩展平台上高效运行 [8]。

- **更大安全优势**：使用机密计算技术，客户可进一步利用敏感数据获得更多洞察，助力 AI 模型训练，应对大规模机器学习和应用等的挑战。英特尔® SGX 、英特尔® TDX 等技术为机密计算提供了良好数据安全保障，其中英特尔® SGX 是经过广泛验证的、数据中心可信执行环境（TEE）的技术实现方案，能大幅减少系统内的攻击风险。而英特尔® TDX 可将客户机操作系统和虚拟机应用都与云端主机、系统管理程序等隔离开来，使客户更方便进行大规模机密计算的部署和管理。

- **更丰富生态构建**：英特尔深知完备的生态建设对云服务能力提升的重要性，其积极与不同合作伙伴开展技术合作，让更多英特尔产品与技术优势在云服务的不同环节中得以体现。例如，英特尔作为各类开源项目的重要代码贡献者，一直在借助开源项目的成本优势和技术发展持续性，使云服务生态各个领域的伙伴都能从中获益。

[8] 数据基于英特尔对截至 2021 年 12 月运行 AI 推理工作负载的全球数据中心服务器装机容量的市场建模。

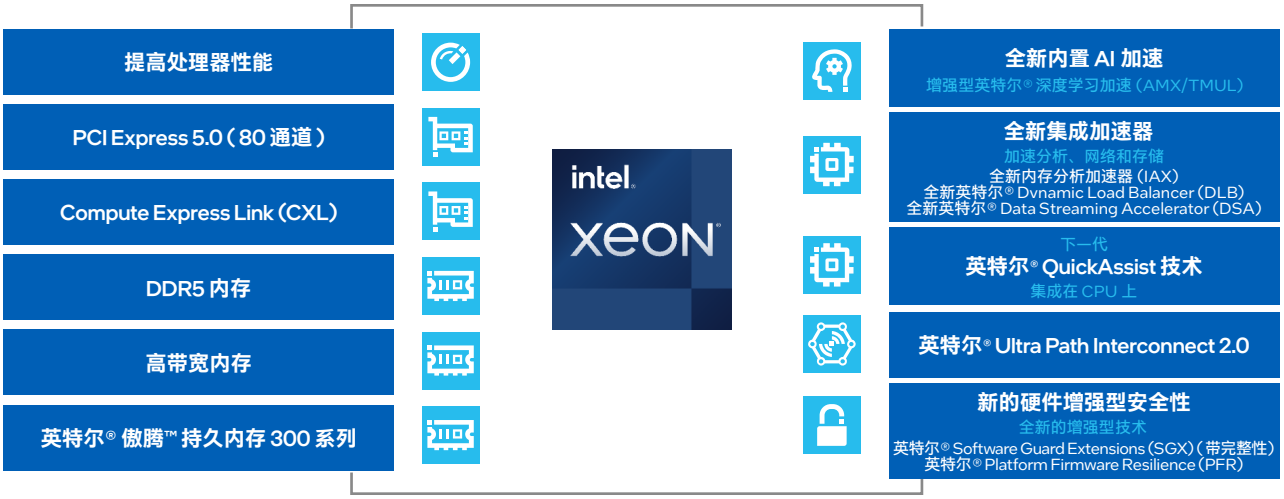


图 1-1-4 第四代英特尔® 至强® 可扩展处理器代际性能提升显著[10]

■ 英特尔® 至强® 可扩展平台持续演进

云服务的繁荣推动数据中心芯片市场规模迅速扩大，数据表明，其市场容量在五年后或达 1,100 亿美元[9]。而得益于英特尔® 至强® 可扩展处理器在运行大型云上工作负载时的出色性能，英特尔正通过持续的处理器演进策略以及加速器整合方案，来有效提升其代际性能，从而应对不断变化和快速扩展的云服务市场需求。

作为至强® 可扩展平台演进的重要标志，目前正在快速推向市场的是第四代英特尔® 至强® 可扩展处理器，其旨在为 AI、数据分析、科学计算、安全以及网络等工作负载提供更强劲的性能，内置有英特尔® 高级矩阵扩展（英特尔® AMX）、英特尔® 动态负载均衡器（英特尔® DLB）等一系列加速引擎，可从容应对云服务的复杂算力和应用优化需求，目前全球前十大云服务提供商都在推进这一处理器的部署。

图 1-1-5 第四代英特尔® 至强® 可扩展处理器内置七大加速器

[9] 数据援引自公开媒体报道《英特尔披露面向2025的至强产品路线图；四大要点回顾数据中心投资者网络研讨会》：https://www.eet-china.com/mp/a207058.html

[10] 如欲了解第四代英特尔® 至强® 可扩展处理器的更详细信息，请访问：https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors.html

面向未来需求，英特尔也给出了明确的至强 ® 可扩展处理器演进路线，包括：

图 1-1-6 至强® 可扩展平台未来演进路线

- **2023 年第四季度：第五代英特尔® 至强® 可扩展处理器（代号 “Emerald Rapids”）。**作为下一款性能核（P-core）产品，其将拥有更高的内核性能，可在相同功率范围内实现更高的每瓦性能，同时通过内置加速器为具体的工作负载进行优化；

- **2024 年上半年：英特尔® 至强® 处理器（代号 “Sierra Forest”）。**作为第一款能效核（E-core）处理器，其将通过优化的每瓦性能、高内核密度，以及高吞吐量性能，对能效进行优化；

- **研发跟进：英特尔® 至强® 处理器（代号 “Granite Rapids”）。**在这一产品中，英特尔将通过一种名为多路合并阵列（MCR）的新型 DIMM，使内存接口在 DDR5 的基础上，实现每秒 8,800 兆次的传输速度；

- **面向未来：计划在 2025 年上市的 “Clearwater Forest”。**其将采用领先的英特尔 18A 制程工艺，实现更强、更优的能效表现。

一系列新技术、新架构和新服务，以及英特尔软硬件产品与技术的运用，正在阿里云为千行百业提供高品质云服务的过程中，实施深入探索、获得充分验证和成功落地，包括：

- 持续创新的软硬件产品有效提升基础设施性能，帮助阿里云平台构建出强劲且持续提升的服务能力；

- 丰富的产品组合与软件工具，对阿里云上不同的云服务应用，包括 AI、科学计算以及视频服务等都提供了卓有成效的优化和加速；

- 对云原生涉及的平台容器化、应用微服务化以及开发 / 运维一体化，英特尔都能以愈加丰富的产品与技术方案来予以支持；

- 与阿里云等共同探索绿色计算技术，利用全浸没式液冷等技术，帮助数据中心实现更优的 PUE 表现，提升节能减碳水平。

来自众多客户的业务实践表明，云服务效能的提升和云上业务的创新，与云平台所配备的软硬件产品与技术支持息息相关。而作为重要的合作伙伴，英特尔为阿里云服务提供了以高性能处理器为核心，从硬件到软件的一系列产品与技术支持，也让双方的携手合作，成为阿里云高品质云服务的鲜明注脚与有力保证。

本白皮书不仅将介绍英特尔推出的丰富软硬件产品与技术组合，还将展示阿里云和英特尔基于各自优势，强强联手，携手开展多维度云服务技术优化，实现业务创新的一系列方案与实践，为行业客户更好地探索云中价值，实现数字化转型提供参考和助力。

案例篇

云服务
基础能力提升

基础设施优化

搭载第四代至强® 可扩展处理器并开启多项加速器，阿里云 ECS g8i 实例通用与场景化性能双提升

为满足各行业不同工作负载对云计算在算力和安全性等方面的更高要求，阿里云持续推动云平台迭代创新，采用第四代英特尔® 至强® 可扩展处理器，依托 CIPU + 飞天技术架构，构建第八代企业级弹性计算实例规格族 ECS (Elastic Compute Service) g8i，新实例在性能、稳定性和安全性等方面全面提升，并率先实现机密虚拟机能力在云上的落地，同时显著降低用云成本和门槛。

第四代至强® 可扩展处理器加持，g8i 通用性能强劲

第四代至强® 可扩展处理器贯彻以结果为导向、工作负载至上的策略，通过集成高性能核、更多内核数量、业内高需求的数据中心工作负载的相关加速器，以及业界领先的 DDR5、CXL1.1、PCIe 5.0，助力客户解决在 AI、分析、网络、安全、存储和科学计算等领域面临的重大计算挑战。

g8i 实例采用 CIPU + 飞天技术架构，搭载第四代至强® 可扩展处理器，网络性能及存储 I/O 均实现大幅演进。g8i 还标配阿里云自研 eRDMA 大规模加速能力，标志着 eRDMA 能力的全面商业化。阿里云 CIPU 所独有的 eRDMA 可让网络时延低至 8 微秒 [11]，且可依托 RDMA 协议栈的高性能、低开销特性，释放更多 CPU 负载，使其更专注于业务处理。

这些独具的优势，在第四代至强® 可扩展处理器的支持下，使得 g8i 更加如虎添翼，全核睿频 p0n 达到 3.2GHz，性能相比上一代实例最大提升 60%，在计算、网络、存储、安全等方面均有炸裂表现。[12]

CPU 型号	第四代英特尔® 至强® 可扩展处理器	整体性能提升 **60%**
最大 vCPU	192	密度提升 **50%**
网络 PPS	3,000万	提升达 **25%**
物理网络	2 x 100G	提升达 **100%**
存储 IOPS	100万	提升达 **56%**
存储延迟	低至百微秒	全面搭载 NVMe，支持共享盘

表 1 g8i 实例通用性能强悍 [13]

采用多种内置加速器，g8i 实例场景化性能强悍

随云服务应用向纵深双向发展，硬件原生加速能力变得日益关键。第四代至强® 可扩展处理器内置广泛且独特的内置加速器，有助于提高性能和效率，减少另行添置专用硬件的需求。在云端和本地环境中，这些专用功能支持 AI、安全性、科学计算、数据分析、存储和网络等目前常见的严苛工作负载，可助力提供更快的处理速度、更强的数据保护和更充分的基础设施利用。除此之外，这些内置加速器还能够提高应用性能，降低成本并提升能效。

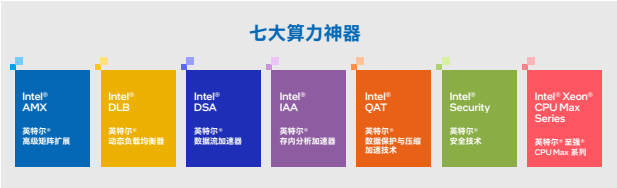


图 2-1-1 第四代英特尔® 至强® 可扩展处理器内置 “七大算力神器”

[11] 数据援引自: http://news.sohu.com/a/659794618_115128
[12] 数据来源于阿里云，如欲了解更多详情，请联系阿里云: https://www.aliyun.com/
[13] 数据援引自: https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/ali-cloud-8th-ecs-performance-improved-with-xeon.html

- **英特尔® 高级矩阵扩展（英特尔® AMX）**可大幅提升深度学习训练和推理的性能，而且集成调优功能，可支持中小型深度学习训练模型，进一步提升基于 CPU 的深度学习与训练能力。面向云实例，英特尔和阿里云深度合作，将这一能力透传到上层的虚拟机，成功在 g8i 中融入 AMX 能力；

- **英特尔® 数据保护与压缩加速技术（英特尔® QAT）**可通过卸载加密、解密和压缩释放处理器内核，降低系统资源消耗，让系统支持更多客户端运行。运用第四代至强® 可扩展处理器内置的英特尔® QAT，g8i 无需额外插卡，在极限情况下压缩 / 解压缩性能可提升高达 70 倍[14]；

- **英特尔® 存内分析加速器（英特尔® IAA）、英特尔® 数据流加速器（英特尔® DSA）**对 g8i 实例性能提升也功莫大焉，助力阿里云在逻辑推理推广、大数据、数据库等多种场景，拥有更强劲的能力和更亮眼的表现。

依托上述丰富的加速器，g8i 实例构建出多样且强劲的硬件原生加速能力。其中，深度学习训练场景性能提升 2 倍以上，推理性能提升 4 倍，加解密、压缩 / 解压缩等场景性能提升 4 倍以上，使得阿里云在统一技术架构下可获得更好的场景化性能扩展，为用户提供更高的性价比。[15]

场景化性能全面提升

加速器	场景	基准测试	性能提升
高级矩阵拓展（AMX）	深度学习（Mlperf 性能测试）	resnet50（图像识别算法） retinanet（目标识别算法） bert（自然语言处理算法）	最大207%↑ 最大124%↑ 最大173%↑
数据保护与压缩加速技术（QAT）	压缩解压缩 OpenSSL 加解密	gzip、deflate、lz4 RSA 非对称加密算法	17-69 倍 5-7 倍
存内分析加速器（IAA）	数据存储	Rocksdb 测试	最大100%

图 2-1-2 g8i 实例场景化性能全面提升 [16]

借力英特尔® 安全策略，阿里云构建全方位防护体系

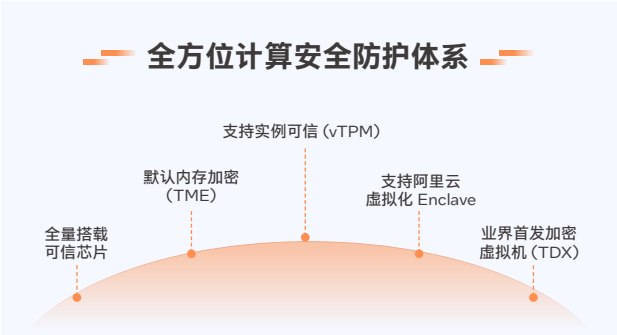


图 2-1-3 g8i 实例具备全方位计算安全防护体系

ECS g8i 全量搭载安全芯片 TPM 作为硬件可信根，实现服务器可信启动，确保零篡改；在虚拟化层面，g8i 支持虚拟可信能力 vTPM，提供实例启动过程核心组件的校验能力。在实例可信的基础上，配合英特尔® SGX 提供的基于硬件的可信执行环境（TEE）和英特尔® TME，以及阿里云自研的加密计算隔离环境 enclave，g8i 进一步强化了数据可用不可见。

同时，g8i 实例还启动了机密虚拟机能力，也即英特尔® TDX（Intel® Trust Domain Extension）的邀测，让用户无需二次开发即可将现有应用迁移至受 TDX 保护的实例，实现数据可用不可见。这也是经由阿里云和英特尔在 TDX 的架构设计、功能验证、安全分析和性能优化等方面紧密合作，实现了 TDX 技术全球首发。

持续技术创新，共享科技红利

搭载第四代英特尔® 至强® 可扩展处理器，并在其多款加速器的加持下，新一代 ECS 实例 g8i 正以强劲的性能和高效用比，支持着阿里云进一步提升服务能力，托举千行百业在更丰富的应用场景中，更直接地感受智算新引擎带来的彪悍实力和更安全的计算环境，加快关键业务云化和智能化，创造新价值。

[14、15、16] 数据援引自：http://news.sohu.com/a/659794618_115128

基于英特尔® 数据中心 GPU Flex 系列的阿里云边缘云，为云应用、转码提供超高性价比解决方案

概述

随着越来越多形式丰富的视频成为人们表达和社交的主要手段之一，视频处理和分析的价值越来越凸显，5G 网络的普及也使得移动用户的视频流量消耗大幅增加，用户持续追求更好的网络质量和更高的视频分辨率。

然而大量的视频处理需求给云服务提供商带来了巨大的挑战，提升视频处理能力、减少单宽成本等成为云服务提供商的迫切需求。同时，在单个用户平均收入需要保持持平的情况下，满足用户对高质量视频的期望也给提供商带来了巨大的效率挑战。

高处理性能可通过提高每台服务器的流密度，为包括 AV1 在内的高级流编解码器提供更佳支持，在满足成本要求方面发挥着关键作用。这种转变对于提供商来说至关重要，有助于其为客户提供具有更高清晰度和更高帧速率的高级视频内容。

挑战

全球云游戏市场的快速增长仍在继续，预计到 2026 年的复合年增长率约为 3.2%，届时其价值将达到约 32 亿美元。[17] 为了在这一领域保持持续竞争力，游戏服务提供商需要基于高效的基础设施不断创新，为用户提供一流的游戏体验。云分发模型使他们能够在自己优化的服务器基础架构上定位游戏开发，简化开发流程，提供顶级体验，同时降低成本。内容目录正在增长，为 CSP 和 CoSP 提供了更多机会。5G 的推出在未来几年内会为最终用户提供更大的带宽，预计这一使用领域的吞吐量需求将急剧增长。

解决方案

英特尔® 数据中心 GPU Flex 系列是一款通用数据中心图形处理器，其针对媒体流密度和质量进行了优化，具有服务器级的可靠性、可用性和可扩展性。与英特尔® 至强® 处理器结合使用，可提供灵活、开放的媒体解决方案，满足当今不断变化和具有挑战性的媒体交付需求。

阿里云边缘节点服务（Edge Node Service, ENS）基于运营商边缘节点和网络构建，可一站式提供“融合、开放、联动、弹性”的分布式算力资源，帮助用户业务下沉至运营商侧边缘，有效降低计算时延和成本。边缘节点服务拥有遍布全球的 2,800 多个节点，客户可以实现边缘地市资源分钟级创建，同时从终端到节点的响应时间缩短到 5-15 毫秒。[18]

阿里云和英特尔合作，在其边缘节点服务中引入第三代至强® 可扩展处理器和英特尔® 数据中心 GPU Flex 系列，推出异构计算服务 gi7s，助力提供商构建高性能基础设施，为云游戏提供标准、高吞吐量、低 TCO 的基座。平台将高游戏质量与每台服务器的高密度游戏实例结合在一起，支持英特尔® oneAPI 工具套件基于开放标准的跨架构编程，具有跨处理器和加速器的代码可移植性，可为主要视频业务包括云应用、编解码提供超高性价比方案。

[17] 数据援引自: https://www.globenewswire.com/news-release/2022/01/24/2371478/28124/en/Insights-on-the-Cloud-Gaming-Global-Market-to-2026-Featuring-Intel-Google-and-Microsoft-Among-Others.html

[18] 数据援引自: https://cn.aliyun.com/product/network/ens?from_alibabacloud=&spm=5176.28055625.J_3207526240.149.734d154aCxO4Io&scm=20140722.M_4691537._.V_1

■ 云应用

云应用测试基于边缘节点服务的异构计算实例 gi7s 展开，使用云版某播放器，测试场景为 1080p30 全屏播放影片，单张 Flex 140 的 gi7s 实例可支持 56 路。整机实例带 4 张 Flex 140，可同时支持 224 路。相对于边缘节点服务的上一代产品（基于第二代至强® 可扩展处理器和英特尔® 服务器 GPU SG1）密度提升 40%[19]。

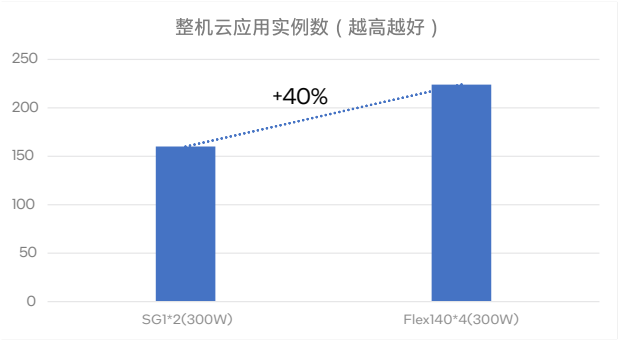


图 2-2-1 边缘节点服务两代异构计算实例性能比较[20]

■ 编解码

编解码测试基于 Flex 140。Flex 140 内置 4 个 X^e Media 硬件加速单元，自带丰富的后处理硬件单元，不需要额外硬件辅助就可提供灵活的视频后处理功能；同时作为业界第一款提供硬件 AV1 编码能力的 GPU，其可节省 30%-43% 的存储和传输需求[21]。Flex GPU 基于硬件 AVC/HEVC/AV1/VP9 的编解码方案，在提高性能的同时也可提供和软件编码相当的质量，且 Flex 140 单卡能够支持 72 路 1080p30 HEVC 和 48 路 1080p30 AVC 的视频流的同时转码[22]。

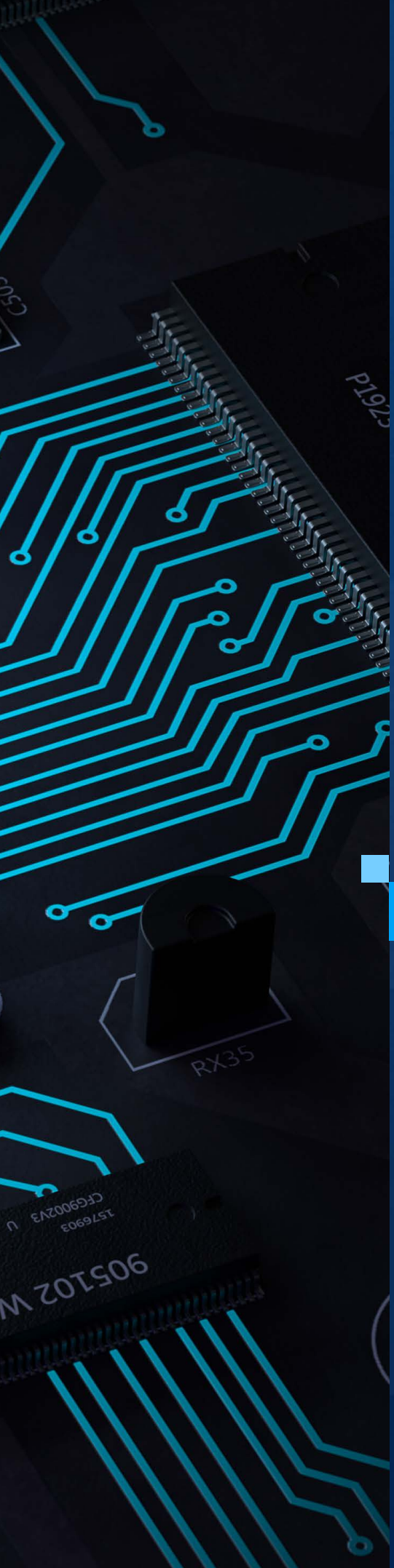
与此同时，支持 HEVC/AV1 的低时延 8K@60fps 实时转码方案，单台服务器可支持多张 Flex 140，大大节省了空间和使用成本。基于英特尔® oneAPI 工具套件的开放式软件栈，除了可支持各种通用的框架，例如 FFmpeg 外，也可提供更底层、更灵活的控制，助力业界软硬件协同创新。

展望

媒体内容提供商的技术不断演进，在致力于为客户提供卓越的质量和体验的同时，追求更低的成本。阿里云边缘节点服务引入先进的英特尔硬件，基于标准的开放式软件栈，实现视频内容相关工作负载的高可靠性、可用性和可扩展性，凭借硬件和软件技术的协同来驱动高密度、高质量的云应用和转码服务。阿里云与英特尔将继续携手探索至强® 可扩展处理器和英特尔® 数据中心 GPU 在 AI 视觉推理、虚拟桌面及云游戏等领域的应用，共同推动行业向前发展。

[19、20] 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
[21] https://bitmovin.com/av1-multi-codec-dash-dataset/
[22] https://github.com/intel/media-delivery/blob/master/doc/benchmarks/intel-data-center-gpu-flex-series/intel-data-center-gpu-flex-series.rst


RTC32
R35
C500
C509
C501
1R2
IC101
RTC27
25.0F6K
SKW23
C-100
16V
R32
C601
C602



增强数据保护

英特尔® TDX 云上落地，助阿里云第八代企业级 ECS 实例为企业云服务提供更优安全防护

“阿里云从成立的第一天起，安全可信就是第一属性。阿里云一直致力于通过各种方式使阿里云对客户更透明，且用户数据的保护是其中不可分割的重要部分。除了严密的安全规约外，我们也在通过各种硬件安全技术来实现用户‘可验证’的数据保护机制。第四代英特尔® 至强® 可扩展处理器在强劲的算力之外，其提供的英特尔® TDX 技术也有力地支持我们为客户提供了更便捷和更多样化的机密计算服务。”

刘煜堃
阿里云高级安全专家
阿里云安全团队

为满足客户希望构建具有不同可信边界级别的 TEE 环境的需求，阿里云与合作伙伴英特尔一起，在全新的第八代企业级 ECS 实例 g8i 中引入第四代至强® 可扩展处理器。新一代处理器所内置的英特尔® TDX 与 ECS g8i 实例搭载的可信平台模块（Trusted Platform Module，TPM）安全芯片相配合，可为大型互联网、新金融等业务场景提供更高安全等级的数据保护能力和云上可信运行环境，进一步帮助客户实现数据可用不可见的愿景。

解决方案 [23]

阿里云第八代企业级 ECS 实例 g8i 创新地采用了“CIPU + 飞天”的技术架构，并引入第四代至强® 可扩展处理器作为核心算力引擎，不仅性能相比上一代实例提升 60% 以上 [24]，在深度学习、人工智能推理训练、大数据等应用场景中也有显著的能力跃升，同时其还在全球范围率先支持基于英特尔® TDX 技术的机密虚拟机能力，在云服务安全性方面获得了新的突破。

■ 通用及整体化性能双提升

得益于第四代至强® 可扩展处理器拥有的澎湃算力，以及英特尔® AMX、英特尔® IAA 等提供的性能加持，新实例在各类云服务应用场景中都有着出色的性能表现，例如在深度学习训练场景中，性能较上一代实例提升 2 倍以上，推理性能则提升 4 倍；而在 RocksDB 等数据存储工作负载中，性能较上一代提升 1 倍以上。[25]

■ 全方位计算安全防护体系

随着更多企业级业务系统与云服务相融合，阿里云看到客户的大多数应用程序或工作负载都是以虚拟机或容器的方式部署到云环境中，并需要获得更大的可信边界。因此单一使用英特尔® SGX 提供的应用程序级可信边界，不仅会增加客户将全部应用程序、工作负载部署到更加安全的云环境的难度，同时逐一对应用程序、工作负载开展改造，也会带来巨大的工作量。

[23] 如欲了解更多解决方案详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/tdx-ali-cloud-ecs-instance-g8i-better-security.html
[24] 数据援引自公开媒体报道：https://developer.aliyun.com/article/1114031
[25] 数据来源于阿里云，如欲了解更多详情，请联系阿里云：https://www.aliyun.com/

而第四代至强® 可扩展处理器内置的英特尔® TDX 技术，与阿里云新实例搭载的 TPM 安全芯片相配合，并结合阿里云自研的加密计算隔离环境 enclave，为阿里云第八代企业级 ECS 实例 g8i 构建了一个基于虚拟化的硬件可信环境，即为整个虚拟化实例（包括虚拟机、容器）都构建出可信的边界，由此为客户提供了可信边界更大、更易部署的安全云环境。

▪ 基于英特尔® TDX，构建全新的 TEE 环境和机密计算方案

如图 2-3-1 所示，借助英特尔® 虚拟机扩展（Intel® Virtual Machine Extension，英特尔® VMX）技术与英特尔® 多密钥全内存加密（Intel® Multi-Key Total Memory Encryption，英特尔® MK-TME）技术，英特尔® TDX 为云实例提供了一种被称为 “信任域（Trust Domain，TD）” 的全新虚拟访客环境。TD 可与其它 TD、实例，以及底层系统软件、管理软件实现相互隔离。而这些安全策略的实施，是由运行在安全仲裁模式（Secure-Arbitration Mode，SEAM）下的 TDX 安全服务模块来完成。

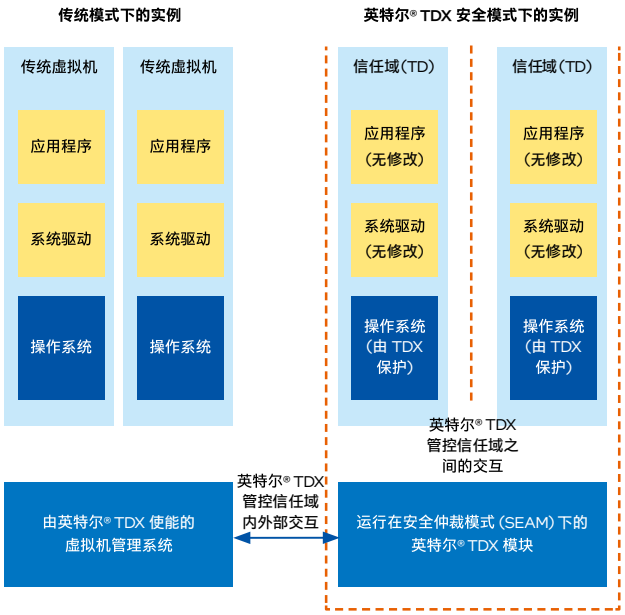


图 2-3-1 英特尔® TDX 技术架构

这一架构中，英特尔® TDX 借助英特尔® MK-TME 技术为 TD 提供了数据机密性和完整性。英特尔® MK-TME 技术支持使用多种密钥对内存进行加密：

- 一方面，其提供的私密密钥，可用于对专用内存（放置 TD 的机密数据）进行加密；
- 另一方面，其提供的共享密钥则用于对共享内存进行加密，用于与 TD 外部的代理进行通信，以执行 I/O 操作，如网络访问、存储服务、调用管理程序服务等。

作为一种全新的机密计算技术，英特尔® TDX 使 TEE 环境的可信边界获得了有效扩展，从而让不同类型下的云服务，无论是 IaaS 或是 PaaS 中的云工作负载都能通过英特尔® TDX 整体纳入机密计算的数据保护之下。一般地，客户可以选择运行两种常用的机密计算方案，机密虚拟机（TD VM）和机密容器（TD CC），阿里云第八代企业级 ECS 实例 g8i 可为客户提供机密虚拟机和机密容器两种使用模式。机密虚拟机是运行在 TD 中的虚拟机实例，而机密容器是将机密计算与云原生容器集成，以保护 Kubernetes 上运行的敏感数据和应用程序。无论哪种方式，客户都可以在云实例中轻松地搭建起自己的应用程序和数据，并受到可信赖的安全保护，使应用程序与数据都与外部环境隔离，以防止未经授权的访问。

应用与展望

为客户提供有效的数据安全和隐私保护，是阿里云等云服务提供商最重要的原则之一，而由第四代至强® 可扩展处理器、英特尔® TDX 技术以及一系列阿里云安全服务提供的数据保护矩阵，正帮助客户通过定制化的机密计算解决方案，实现更加可靠的云端数据资产保护。面向未来，阿里云还将与英特尔进一步展开深入合作，为更多行业和领域的客户构建更加安全、开放和高可靠性的云计算基础设施。

英特尔® SGX 与英特尔® TDX 落地龙蜥社区，助力探索更多样云原生机密计算场景

应对数据安全威胁，龙蜥社区给出解答

如何在加强安全的前提下更大程度发挥数据的价值，是近年来的重要课题，也让作为当前数据处理基础设施的云计算经历着一次范式转换，即从默认以 CSP 为信任基础的计算范式走向信任链与 CSP 解耦的新范式，此范式被称为隐私保护云计算，而机密计算是实现隐私保护云计算的必由之路。

机密计算是指通过在基于硬件的可信执行环境（TEE）中执行计算的方式来保护使用中的数据。为拥抱隐私保护云计算新范式，促进隐私保护云计算生态发展，龙蜥社区成立了云原生机密计算（以下简称“CNCC”）SIG（Special Interest Group），专注于机密计算底层技术，致力于通过合作共建的方式，为业界提供开源和标准化的机密计算技术及安全架构，推动云原生场景下机密计算技术的发展。

作为 CNCC SIG 的主要参与者及重要项目贡献者之一，英特尔充分利用英特尔® SGX 和英特尔® TDX，以及在其上构建的运行时虚拟机与容器等应用，加速云原生机密计算探索和实践，助力解决目前机密计算领域普遍存在的共性问题，如用户对技术认知感不足、技术门槛相对偏高、应用产品缺乏普适性以及信任和信任模型的问题。

英特尔机密计算技术在龙蜥社区的实践及规划

英特尔的机密计算技术产品组合在英特尔® SGX 和英特尔® TDX 加持下，允许企业按需选择安全级别，满足自身的业务需求和监管方面的要求。

针对云原生机密计算 SIG 活跃项目，英特尔提供英特尔® SGX Stack，并支持 Inclavare containers，以及 JavaEnclave、Occlum 等，助力业界强化“零信任”安全策略，其工作可以分为四大类：

1. 为机密计算和可信环境提供基础支撑，主要包括英特尔® SGX SDK、英特尔® SGX PSW/DCAP 安装包的适配和 Anolis 的集成。

比如，英特尔为 Anolis OS 提供了适配和优化过的英特尔® SGX SDK/PSW/DCAP 以及英特尔® TDX DCAP 软件安装包和运行时库，可以让用户更加方便高效地为上层的机密计算应用，如机密虚拟机和机密容器等提供服务。

函数	Trusted FaaS
应用	KMS / MySQL/ Nginx / PPML /eTPM /Enclave Network Gateway
服务	Attestation 体系 & 供应链安全
库 & 运行时	JavaEnclave / Occlum / Gramine / Intel SGX & PSW / DCAP / rats-tls & librats / Intel HE Stack
容器	海光 CSV 机密容器 / 袋鼠 CC/Inclavare Containers / Enclave-CC
虚拟化	QEMU / KVM / Dragonball / TDVF / OVMF
OS	Anolis 8 / Anolis 23 + ANCK 5.10 / 5.19
TEE 硬件平台	Intel TDX 1.0 & SGX 2.0 & 海光 CSV1 / 2 / 3

图 2-4-1 龙蜥社区机密计算技术图谱 [26]

2. 支持基于 LibOS 的运行时，目前使用该项技术的已有 Gramine 和蚂蚁集团的 Occlum。

在 Gramine 堆栈中使用英特尔® SGX，即以 Gramine 作为基础支撑模块，为 Enclave 机密容器提供英特尔® SGX 运行环境，并通过集成到龙蜥操作系统 Anolis OS，让用户无需修改或者重新编译代码即可部署应用。

[26] 图片来源于龙蜥社区：https://openanolis.cn/blog/detail/644006874786283522

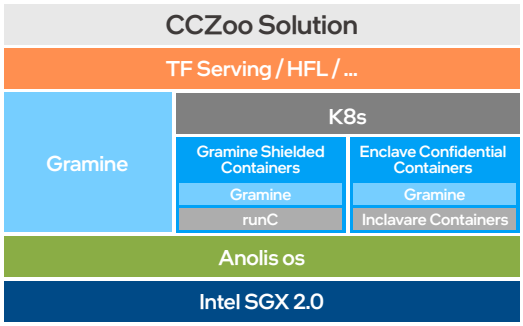


图 2-4-2 Gramine 堆栈 [27]

Occlum 也采用类似路径，让 Enclave 机密容器及其他英特尔® SGX 机密容器 / 工作负载可以选择 Occlum 作为运行英特尔® SGX 的环境。

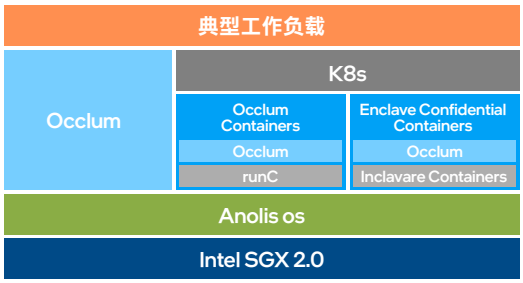


图 2-4-3 Occlum 堆栈 [28]

3. 支持虚拟化和容器，如英特尔® SGX 虚拟化、SGX 和 TDX 机密容器。

而对于英特尔® SGX 虚拟化，在 VM 或者 VM Pod 当中，英特尔® SGX SDK/PSW/DCAP 都起到了基础支撑作用，服务于具体的 Runtime。英特尔® TDX 主要提供远程证明支持，基于 OS/VM 整体加密且不必移植任何代码，提升了易用性。面对用户对远程证明的担忧，英特尔则提供了一套软件和服务，让 TDX VM 可以更加高效和安全地完成远程证明。

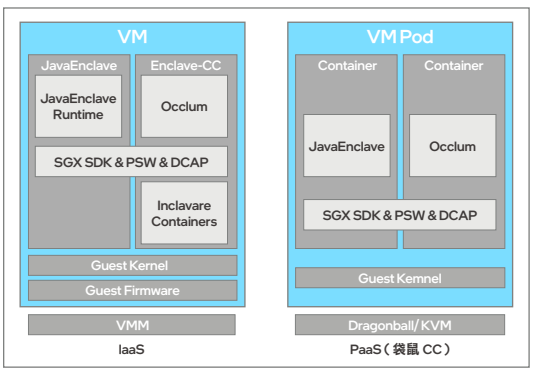


图 2-4-4 SGX 虚拟化 [29]

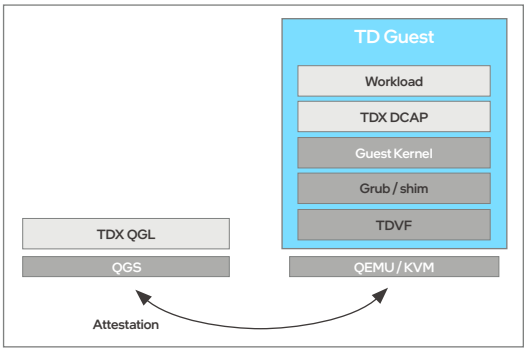


图 2-4-5 TDX VM [30]

4. 支持同态加密，包括相关的硬件、软件和库等。

关于同态加密，英特尔也有相关的软件和库，还提供了硬件支持，比如，第三代、第四代至强® 可扩展处理器，以及英特尔® QAT 硬件加速器。

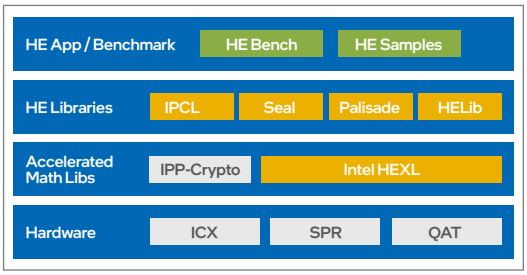


图 2-4-6 英特尔同态加密堆栈

总结与展望

英特尔携手龙蜥社区，支持机密计算 SIG，就是要通过推动开源合作，构建云原生机密计算开源技术栈，提升机密计算可用性、通用性和用户信心，加速机密计算在云上的部署和应用。

为加速这一进程，英特尔也正在持续推进机密计算技术的落地和规划，比如英特尔® SGX、英特尔® TDX DCAP 等基础框架的解包适配，已经顺利完成测试，在英特尔软件仓库发布，并已于 2022 年四季度实施开源，为开发者应用开发提供平台，同时与阿里巴巴合作，集成到 Anolis 软件库；与此同步，英特尔® SGX SDK 解包适配以及开源和 Anolis 软件库集成工作也一并完成，有力促进了云原生机密计算探索和应用实践，以及计算生态建设，更有助于各级、各类用户基于 “零信任” 安全策略，构建更安全的机密计算环境，为数据要素更好地释放动能保驾护航。

[27、28、29、30] 图片来源于龙蜥社区：https://openanolis.cn/blog/644006874786283522

英特尔® SGX 助力阿里云构建端到端隐私保护机器学习方案

面向大数据与 AI 的数据融通面临严峻安全风险

跨机构、跨行业的数据融合、联合分析和建模的需求日趋增加，数据安全风险急剧增长。这一方面是由于数据本身可复制、易传播，在传统安全模式下，数据一经分享难以追踪。另一方面，数据持续流动会导致责任划分不明确、权限控制困难、以及难以追责等问题。保证数据的安全可靠成为重中之重。然而，面向 AI 和大数据的传统安全防护方案常常会面临难以保护正在使用的数据和硬件底层等挑战。

基于英特尔® BigDL PPML 的阿里云端到端隐私保护机器学习

为帮助企业在 AI 和大数据等应用中，更好地实现端到端隐私保护，阿里云与英特尔合作，将英特尔® BigDL PPML 与阿里云 DataTrust 平台进行协同，联合验证了隐私保护机器学习的端到端工作流和相关业务场景。

■ 英特尔® BigDL PPML

BigDL 是英特尔开源的、统一的人工智能解决方案平台，其应用英特尔® SGX 可信硬件执行环境 (TEE)，并集成其他软硬件安全措施，构建了一个分布式的隐私保护机器学习平台，能够保护端到端 (包括数据输入、数据分析、机器学习、深度学习等各个阶段) 的分布式人工智能应用。

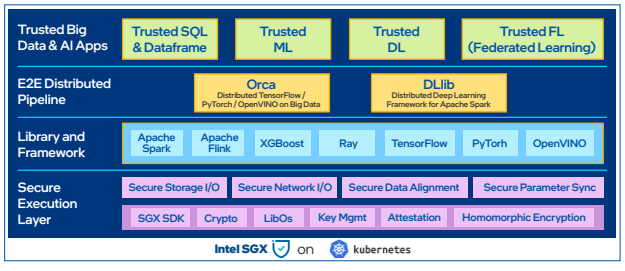


图 2-5-1 英特尔® BigDL PPML 软件栈

借助 BigDL PPML 平台，开发者可以：

- 在加密数据上开发并运行标准的分布式人工智能应用；
- 利用基于硬件的安全技术 (如英特尔® SGX) 保护计算过程以及相应的内存数据；
- 为 AI 应用提供端到端的安全和隐私保护，例如：在具备英特尔® SGX 硬件能力的 K8s 环境中创建并认证可信的集群环境；通过密钥管理系统 (Key Management System, KMS) 为分布式数据提供加密和解密能力；通过英特尔® SGX、加解密技术、TLS 和安全认证等技术实现更加安全的分布式计算和数据通信。

■ 阿里云 DataTrust 隐私保护计算平台

阿里云 DataTrust 是行业领先的基于可信执行环境、安全多方计算 (Secure Multi-Party Computation, MPC)、联邦学习 (Federated Learning, FL)、差分隐私 (Differential Privacy, DP) 等隐私增强计算 (Privacy Enhancing Technique) 技术打造的隐私增强计算平台，其以英特尔® SGX 为底座，结合 MPC、FL 等技术，基于阿里云数据中台丰富的应用场景实践，能够在保障数据安全的前提下完成多方数据联合分析、联合训练、联合预测，为企业提供立足数据业务原生的数据安全流通解决方案，助力企业业务增长。

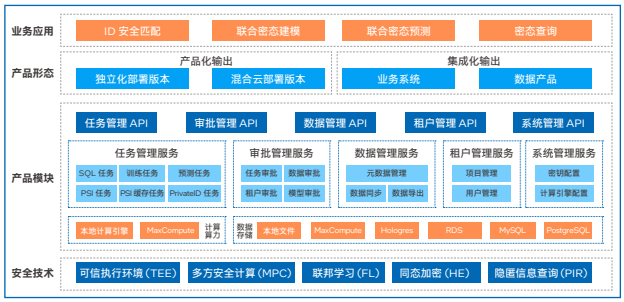


图 2-5-2 阿里云 DataTrust 架构

■ 端到端解决方案工作流程

基于隐私计算的核心功能，BigDL PPML 解决方案集成了端到端隐私保护计算工作流的更多组件，例如签鉴服务（Attestation Service）、密钥管理（Key Management）以及基于 Kubernetes 的安全容器化部署方案。

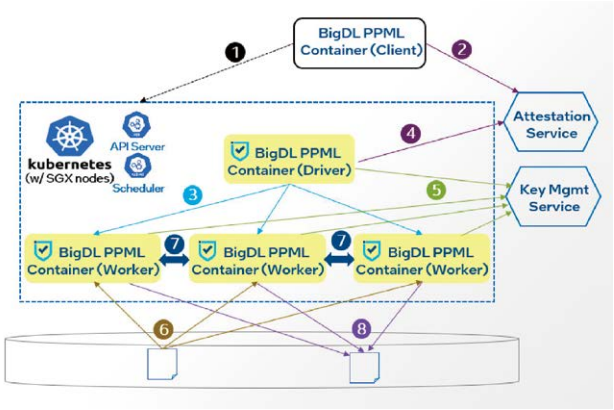


图 2-5-3 基于 BigDL PPML 的端到端安全计算工作流程图

在基于 BigDL PPML 的端到端安全计算工作流中，各个流程的功能如下：

① 用户通过 BigDL PPML 命令行向 Kubernetes 提交任务，此操作会创建一个驱动节点
② BigDL PPML 客户端认证驱动节点
③ 驱动节点创建更多的工作节点
④ 驱动节点认证工作节点
⑤ 驱动节点和工作节点从密钥管理服务获取密钥
⑥ 工作节点读取输入数据并解密
⑦ 工作节点分布式运行大数据，机器学习或者深度学习任务
⑧ 工作节点加密并写回输出数据

图 2-5-4 基于 BigDL PPML 的端到端安全计算工作流程功能

BigDL PPML 提供了实现以上工作流的集成解决方案，通过使用预制的工作流解决方案，开发者可以更加专注于业务逻辑的相关开发工作，利用 BigDL PPML 保障其应用的端到端安全性和隐私性。用户可以显著提升隐私计算应用的开发效率，大幅缩短实现隐私计算解决方案的开发时间（Time to Solution）。

收益：推动数据价值安全流动

基于英特尔® BigDL PPML 的阿里云端到端隐私保护机器学习解决方案继承了可信执行环境的优点。和传统数据安全解决方案相比，它的安全性和数据效用性十分突出，同时性能损耗较低。

通过应用该方案，企业能够构建端到端的安全保护流程，在数据输入、数据分析、机器学习、深度学习等 AI、大数据应用的多个阶段建立安全防护能力。同时，该方案实现了基于硬件底层的保护，具备更高的数据保护等级，能够防护传统安全方案难以抵抗的攻击形式，降低重要数据泄露的风险。

在该方案的支持下，企业能够提供更加安全的数据融通服务，该方案的典型应用场景包括：

- **全域精细运营**：品牌方通过联动平台、第三方等全域数据，在更好地保护个体隐私及数据安全的前提下，构建品牌数智化运营能力，优化人货场的配置，拉动业务增长；
- **联合智能风控**：行业或企事业单位在原始数据不出自身环境的前提下，通过隐私增强计算技术，实现与多方数据的联合风控，提高风控识别有效性，助力业务良性增长；
- **广告搜索推荐**：在加强消费者隐私与一方、二方数据安全防护的前提下，通过数据加持进行联合建模，提升算法准确率，提高广告投放有效性，推动业务持续高效增长。

总结与展望

BigDL PPML 隐私保护机器学习解决方案基于英特尔® SGX、BigDL 以及众多安全相关的组件共同打造，为巩固数据的安全性和大数据及 AI 工作负载性能提供了平台解决方案。

阿里云和英特尔共同验证了 BigDL PPML 解决方案的工作流程，该合作展示了应用 BigDL PPML 开发端到端隐私保护应用的最佳实践，体现了 BigDL PPML 在加速开发隐私保护应用方面的显著作用。双方将在当前合作成果的基础上，进一步强化端到端隐私保护方面的创新与实践，帮助用户实现更加安全的数据融通，在巩固安全的基础上加速数据价值挖掘。

英特尔® SGX 和英特尔® DL Boost 赋能蚂蚁集团隐私保护机器学习

隐私保护机器学习（PPML）有助于化解囤积和处理海量数据带来的隐私、安全和监管等风险，其采用加密技术差分隐私、硬件技术等，旨在处理机器学习任务的同时保护敏感用户数据和训练模型的隐私。

在英特尔® SGX 和蚂蚁集团用于英特尔® SGX 的内存安全多进程用户态操作系统 Occlum 的基础上，蚂蚁集团与英特尔合作搭建了 PPML 平台。本文将介绍这项运行在 Analytics Zoo[31] 上的解决方案，并展示该解决方案在第三代至强® 可扩展处理器上得到英特尔® DL Boost 技术助力时的性能优势。

英特尔® SGX 和 Occlum

英特尔® SGX 是英特尔的受信任执行环境（TEE），它提供基于硬件的内存加密，隔离内存中的特定应用代码和数据。英特尔® SGX 使得用户层代码可以分配内存中的受保护区域，即“飞地”，这些区域不受更高权限等级程序运行的任何影响。

与同态加密和差分隐私相比，英特尔® SGX 在操作系统、驱动、BIOS、虚拟机管理器或系统管理模型已瘫痪的情况下仍可帮助防御软件攻击。因此，英特尔® SGX 在攻击者完全控制平台的情况下仍可增强对隐私数据和密钥的保护。第三代至强® 可扩展处理器可使 CPU 受信任内存区域增加到 512GB，使得英特尔® SGX 能够为隐私保护机器学习解决方案打下坚实的基础。

蚂蚁集团一直积极探索隐私保护机器学习领域，并发起了开源项目 Occlum。Occlum 是用于英特尔® SGX 的内存安全多进程用户态操作系统（LibOS）。使用 Occlum 后，机器学习工作负载等只需修改极少量（甚至无需修改）源代码即可在英特尔® SGX 上运行，以高度透明的方式保护了用户数据的机密性和完整性。

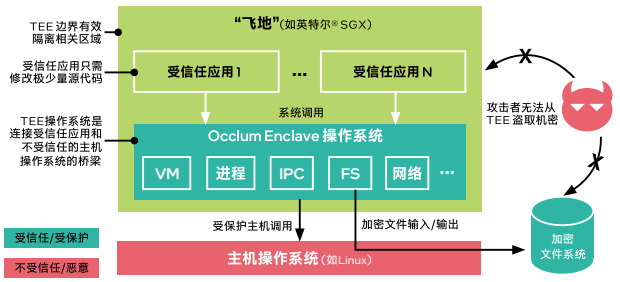


图 2-6-1 用于英特尔® SGX 的 Occlum 架构（图片来源：Occlum · GitHub）

Analytics Zoo 赋能端到端 PPML 解决方案

Analytics Zoo 是面向基于 Apache Spark、Flink 和 Ray 的分布式 TensorFlow、Keras 和 PyTorch 的统一的大数据分析和人工智能平台。使用 Analytics Zoo 后，分析框架、ML/DL 框架和 Python 库可以在 Occlum LibOS 以受保护的方式作为一个整体运行。此外 Analytics Zoo 还提供安全数据访问、安全梯度与参数管理等安全功能，可赋能联邦学习等隐私保护机器学习用例。

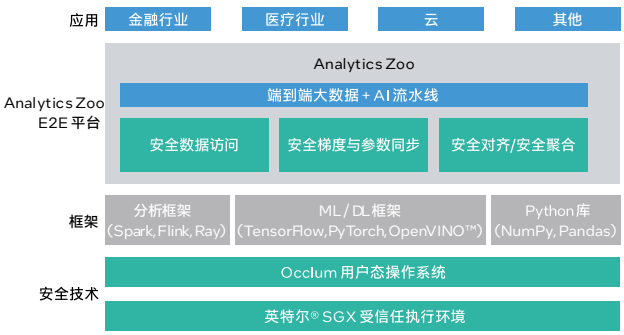


图 2-6-2 端到端 PPML 解决方案为金融服务、医疗卫生、云服务等应用领域提供安全分布式计算

[31] BigDL 2.0 已包含 BigDL 和 Analytics Zoo。

在 Analytics Zoo PPML 平台上，蚂蚁集团与英特尔共同打造了一个更加安全的分布式端到端推理服务流水线。该流水线采用 Analytics Zoo Cluster Serving 打造，后者是轻量级分布式实时服务解决方案，支持多种深度学习模型，包括 TensorFlow、PyTorch、Caffe、BigDL 和 OpenVINO™。Analytics Zoo Cluster Serving 包括 web 前端、内存数据结构存储 Redis、推理引擎（如面向英特尔® 架构优化的 TensorFlow 或 OpenVINO™ 工具套件），以及分布式流处理框架（如 Apache Flink）。

推理引擎和流处理框架在 Occlum 和英特尔® SGX “飞地”上运行。web 前端和 Redis 受到传输层安全（TLS）协议加密，因此推理流水线中的数据（包括用户数据和模型）在存储、传输、使用的过程中都受到更多地保护。

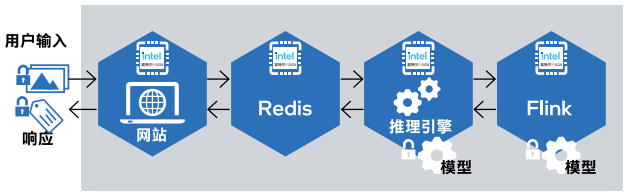


图 2-6-3 推理服务流水线

英特尔® DL Boost 加速端到端 PPML 解决方案

该解决方案执行如下端到端推理流水线：

1. RESTful http API 接收用户输入，Analytics Zoo pub/sub API 将用户输入转化成输入队列，并由 Redis 管理。用户数据受加密保护；
2. Analytics Zoo 从输入队列中抓取数据。它在分布式流处理框架（如 Apache Flink）上采用推理引擎进行推理。英特尔® SGX 使用 Occlum 来保护推理引擎和分布式流处理框架。英特尔® oneAPI 深度神经网络库（oneDNN）利用支持 INT8 指令集的英特尔® DL Boost 提高分布式推理流水线的性能；
3. Analytics Zoo 从分布式环境中收集推理输出，并送回到由 Redis 管理的输出队列。随后，解决方案使用 RESTful http API 将推理结果作为预测返回给用户。输出队列中的数据和 http 通信内容都被加密。

性能分析 [32]

与不受英特尔® SGX 保护的推理流水线相比，当推理解决方案受到英特尔® SGX 保护，ResNet50 推理流水线的吞吐量会有少许损失。而采用支持 INT8 指令集的英特尔® DL Boost 后，受英特尔® SGX 保护的推理流水线吞吐量翻了一番。

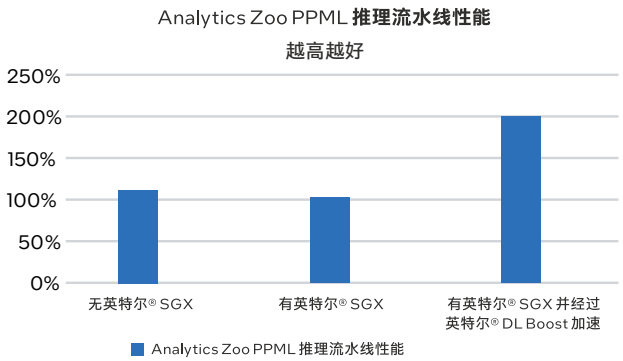


图 2-6-4 英特尔® SGX、英特尔® DL Boost 和第三代英特尔® 至强® 可扩展处理器提供高性能安全能力

基于英特尔® SGX 打造的 Analytics Zoo PPML 解决方案继承了受信任执行环境的优点。和其它数据安全解决方案相比，它的安全性和数据效用性十分突出，性能方面仅略逊于纯文本。英特尔® DL Boost 和英特尔® oneDNN 则进一步提升了 Analytics Zoo PPML 推理解决方案的性能。

总结

在日益复杂的法律和监管环境中，对于企业和组织来说，保护客户数据隐私比以往任何时候都更加重要。在隐私保护机器学习的助力下，企业和组织就能在继续探索强大的人工智能技术的同时，面对大量敏感数据处理降低安全性风险。

Analytics Zoo 隐私保护机器学习解决方案基于 Occlum、英特尔® SGX、英特尔® DL Boost 和 Analytics Zoo 打造，为助力确保数据的安全性和大数据及人工智能工作负载性能提供了平台解决方案。蚂蚁集团和英特尔共同打造并验证了这一 PPML 解决方案，并将继续合作探索人工智能和数据安全领域的最佳实践。

[32] 如欲了解性能详情，请访问：https://www.intel.cn/content/www/cn/zh/now/csp-abm-alibaba/better-together-sgx-boost-privacy.html

微服务

英特尔处理器技术特性为阿里云网关产品提供有效 HTTPS 加速，应对互联网数据新浪潮

作为互联网系统与外部网络之间重要的门户与桥梁，数据中心的网关产品一直是庞大数据流处理的重要节点之一，因此提升其密码学计算效率无疑将成为提高网络服务性能的关键。为此，阿里云与合作伙伴英特尔一起，借助面向单路和双路的第三代至强® 可扩展处理器所具备的两种全新技术特性来为旗下网关产品加速。

解决方案

随着 HTTPS 化的全面推进，承担阿里巴巴集团所有入口流量的阿里统一接入网关 Tengine（阿里云在 Nginx 的基础上，针对大访问量需求所开发的网关产品）受到了巨大的性能挑战。为此，阿里云亟待通过各类技术探索提升 Tengine 在密码学计算等方面的效率。

面向单路和双路的第三代至强® 可扩展处理器可为处理器硬件加速方案提供坚实基础。而这一处理器新增的两种全新技术能力，包括新的指令集扩展以及 Multi-buffer（多缓冲）技术，也能为密码学计算效率的提升提供巨大增益。

首先在新的指令集扩展上，新处理器在传统 AES-NI 加速指令的基础上，新增了基于英特尔® AVX-512 的英特尔® 密码操作硬件加速（英特尔 ® Crypto Acceleration）特性，其内置了 Vector AES（VAES）、Integer Fused Multiply Add（IFMA 大数计算）、Galois Field New Instructions（GFNI）等能力，同时也支持 SHA extension（SHA-NI），有效提升了对 SHA256 算法的处理性能。

其次新处理器也引入了 Multi-buffer 技术，这一技术能够集中批量请求，并进行并发处理。其与英特尔® AVX-512 这一类 SIMD 指令相配合后，能有效加速密码学的计算效率。如图 2-7-1 所示，使用 Multi-buffer 技术的新平台，在 ECDHE x25519、RSA Sign 2048、AES-CTR 等不同的密码学计算上，相比上一代平台（基于第二代至强® 可扩展处理器）均有数倍的提升 [33]。值得一提的是，英特尔® Crypto Acceleration 方案沿用并扩展了英特尔® QAT 的软件框架 QAT Engine，使英特尔® QAT 可从专用的硬件加速卡场景扩展到通用的处理器指令加速场景，有效扩展了适用范围。

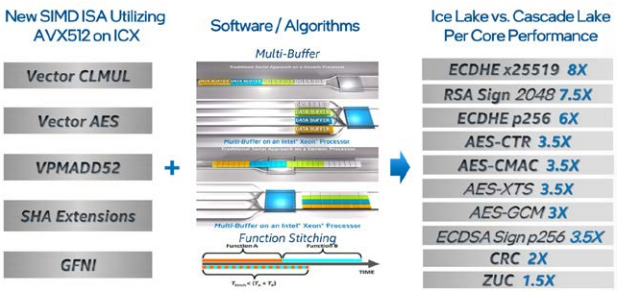


图 2-7-1 Multi-buffer 配合 SIMD 指令加速 [34]

通过龙蜥社区与英特尔的携手合作，阿里云已在其第七代云服务器引入面向单路和双路的第三代至强® 可扩展处理器的基础上，加入了上述的加速特性，并为 Tengine 网关带来了显著的性能加速。来自验证测试的数据表明[35]，面向单路和双路的第三代至强® 可扩展处理器的单核 AES 性能可加速 2.2 倍，达到上一代处理器的 3.4 倍；RSA 单核可加速 4.9 倍，达到上一代处理器的 5.1 倍；同时 Tengine 的单核 SSL/TLS 握手性能也可相应地加速 3.1 倍，达到上一代处理器的 3.2 倍。

[33, 34] 如欲了解更多详情请访问：https://openanolis.cn/blog/detail/643818207334503377
[35] 数据援引自龙蜥社区报告《Ice Lake SSL/TLS加速实践》：https://openanolis.cn/sig/crypto/doc/390714951012679780

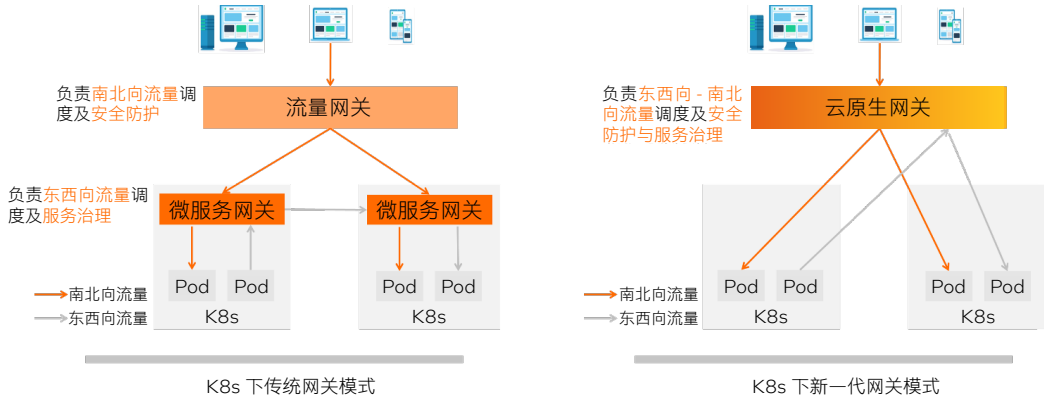


图 2-7-2 阿里云全新云原生网关产品 [36]

在 Tengine 这样的传统网关之外，阿里云也在积极探索下一代网关的研发与构建。如图 2-7-2 所示，与传统由流量网关与业务网关（或微服务网关）两层网关构建的方案不同，在容器和 K8s 主导的云原生时代，阿里云基于开源 Envoy 构建的云原生网关产品将流量网关与业务网关（或微服务网关）合二为一，并通过硬件加速、内核调优等方式，在不影响性能的前提下，帮助用户大幅降低部署网关的资源成本。

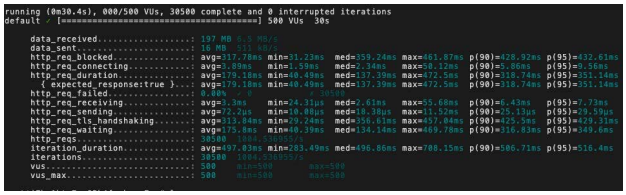
与传统网关产品相比，新的 MSE 云原生网关具有以下优势：

- 网关直连业务 Pod IP，无需经过传统的 Cluster IP 等，RT（Response Time，响应时间）更低；
- 支持 Wasm 插件市场，支持插件热加载，满足用户多语言的自定义插件诉求；
- 通过自研 Multi-Ingress Controller 组件，支持多集群 Ingress 复用同一个网关实例；
- 原生兼容原生 K8s Ingress 规范，且支持 Nginx Ingress 核心功能注解的无缝转换。

更为重要的是，上述面向单路和双路的第三代至强® 可扩展处理器所具备的加速特性，也被成功应用到 MSE 云原生网关产品上，其对于 HTTPS 硬件加速的支持，使 QPS 等性能指标得以大幅提升。

在面向某最终客户的推广中，这一加速性能获得了客户的充分认可。由于在 MSE 云原生网关中，基于新处理器的 HTTPS 硬件加速功能已被充分产品化，因此客户的部署十分便捷，只需在购买时开启相关选项即可。在验证测试中，开启后的加速效果非常明显。如图 2-7-3 所示 [37]，其中单实例（1C2G）的压测 HTTPS QPS 从 1,004 提升到 1,873，提升约 86%，TLS 握手 RT 从 313.84ms 降到 145.81ms，下降约一倍。

加速前：



加速后：

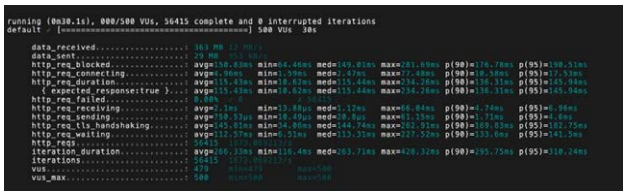


图 2-7-3 MSE 云原生网关开启 HTTPS 加速前后对比 [38]

基于这一结果，最终客户对基于面向单路和双路的第三代至强® 可扩展处理器的 HTTPS 硬件加速功能非常认可，并最终选择使用 MSE 云原生网关来替代 Nginx Ingress 网关。

总结与展望

英特尔正不断以新产品、新技术的推出来帮助阿里云等合作伙伴有效应对数据处理挑战，而基于面向单路和双路的第三代至强® 可扩展处理器打造的 HTTPS 硬件加速能力，正是这一过程中的有力成果。面向未来，英特尔还将与阿里云继续合作，持续推动新一代数据中心产品的性能提升。

[36] 图片来源：https://openanolis.cn/blog/detail/643818207334503377
[37] 数据援引自龙蜥社区博客《性能提升1倍，成本直降50%！基于龙蜥指令加速的下一代云原生网关》：https://openanolis.cn/blog/detail/643818207334503377
[38] 如欲了解更多详情请访问：https://openanolis.cn/blog/detail/643818207334503377

基于英特尔® 架构的阿里云服务网格 ASM 产品技术加速应用服务加密通信

背景

在目前非常流行的 Service Mesh 项目 Istio 中，其数据面 Envoy 无论是作为网格入口流量网关还是作为内部微服务的边车代理，都需要处理大量的 mTLS 请求。TLS 协议作为网络安全通信的基石，一次会话的处理过程总体上可分为握手阶段和数据传输阶段，握手阶段最重要的任务是使用非对称加密技术协商出一个会话密钥，然后在数据传输阶段，使用协商出的会话密钥对数据执行对称加密操作，再进行传输。

但同时，mTLS 加密算法的应用会带来较高的资源消耗。尤其在握手阶段的非对称加解密的操作，需要消耗大量的 CPU 资源，也会增加微服务调用之间的时延时间和入口网关的服务响应时间，这在大规模微服务场景、计算资源有限的边缘计算场景等场景下，会带来棘手的性能挑战。因此，大量用户希望通过服务网格技术实现更高的安全防护能力，但也对该技术所带来的性能压力心存顾虑。

采用英特尔® Crypto Acceleration 优化的阿里云 ASM

阿里云服务网格产品ASM（Alibaba Cloud Service Mesh）提供了一个全托管式的服务网格平台，兼容社区开源 Istio，用于简化服务的治理，包括服务调用之间的流量路由与拆分管理、服务间通信的认证安全以及网格可观测性能力，能够有效减轻开发与运维的工作负担。ASM 增强了多协议支持以及动态扩展能力，提供精细化服务治理，完善零信任安全体系，增加大规模集群支持能力，并融合了 Multi-Buffer 等技术来持续提升性能，降低服务网格使用门槛，助力客户在生产环境中进行大规模落地。

作为业内首个全托管 Istio 兼容的服务网格产品，阿里云 ASM 在架构上保持了与社区、业界趋势的一致性，控制面的组件托管在阿里云侧，与数据面侧的用户集群独立。

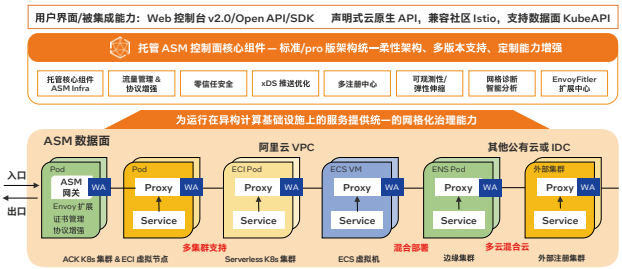


图 2-8-1 阿里云服务网格产品 ASM 架构图

阿里云服务网格 ASM 基础设施层为基于第三代英特尔® 至强® 可扩展处理器的阿里云第七代 ECS 服务器。基于该处理器的阿里云第七代 ECS 服务器相较于上一代产品[39]，单核性能提升 30%，整机算力提升 50% 以上，这为阿里云服务网格 ASM 奠定了坚实的性能基础。

该服务器提供了若干不同的实例类型，并集成了搭载 Multi-Buffer 技术的最新 Envoy 上游版本，提供了基于英特尔® Multi- Buffer 技术的 TLS 加解密性能优化能力，为最终客户的落地实践提供了卓越的平台支撑。

[39] 更多信息详见 https://developer.aliyun.com/article/783678。
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。

■ 英特尔® Crypto Acceleration 及 Multi-Buffer 解决方案

第三代至强®可扩展处理器引入了英特尔®Crypto Acceleration，其可提供公钥加密（Public-Key Cryptography）功能，通过新的指令集 AVX-512_IFMA，提供对公钥加密中常见的“大数”乘法的支持。英特尔® Crypto Acceleration 还搭载了 Multi-Buffer 多缓冲区处理技术，目前英特尔® Multi-Buffer 技术通过英特尔® 集成性能基元（英特尔® IPP）集成的 Cryptography Multi-buffer Library 加密库向上对 TLS 应用提供接口调用，该库基于英特尔® AVX-512 操作提供了 RSA、ECDSA 等算法的多缓冲区优化版本。

Envoy 使用的 TLS 实现库是 BoringSSL Library，其提供了一个名为 Private Key Provider 的框架，用户只需按照 Private Key Provider 框架要求实现相关的功能接口，可以集成外部自定义的加解密操作实现。本方案针对英特尔® Multi-Buffer 技术实现了一个名为 CryptoMB 的 Private Key Provider Extension。

同时为了利用英特尔® Multi-Buffer 技术的优势，可以并行处理 8 个加密操作，BoringSSL 中的 TLS 握手的过程被重构实现为异步模式。在这些异步操作中，还可以结合使用 AVX-512 指令处理，大大提高了整体性能。为了平衡 TLS 握手请求处理吞吐量和时延的关系，英特尔还引入了一个计时器的变量进行控制。在 TLS 操作填满 8 个缓冲区或者 Timer 计时器超时两个条件满足其一时，当前缓冲的所有 TLS 操作都将会被一次性处理。

Envoy 1.20 及后续的版本已经集成了英特尔® Multi-Buffer 技术，Envoy 的配置文件可以根据运行的平台是否支持 Crypto Acceleration 架构功能进行配置。一旦启用该功能，Envoy 的 TLS 配置除了 Private Key 的路径信息，还需要指定 CryptoMB Private Key Provider 以及计时器信息。这些配置也可以集成到外部控制面实现中，如最新的 Istio 项目 1.14 版本已经支持该功能，并且实现了对网格级别、网关级别或者特定工作负载级别的配置，实现了灵活、精细化的功能管理。

■ 英特尔® Multi-Buffer 在阿里云 ASM 的落地实践

为了更好地通过英特尔® Multi-Buffer 技术来加速云服务网格的加解密性能，阿里云 ASM 通过如图 2-8-2 所示步骤启用英特尔® Multi-Buffer 技术：

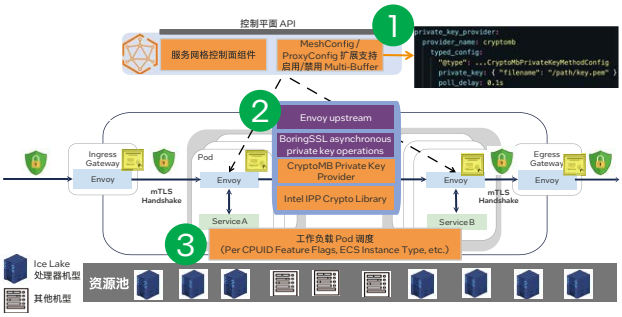


图 2-8-2 阿里云 ASM 启用英特尔® Multi-Buffer 流程图

目前，英特尔® Multi-Buffer 已经在阿里云 ASM 产品的最新版本中对外开放，在阿里云 ASM 控制台中，用户可以通过性能优化开关一键启用此功能，从而实现加解密性能的提升。

为了测试英特尔® Multi-Buffer 对于性能的影响，阿里云将通用型实例规格族 g7 作为 Kubernetes 节点，并验证英特尔® Multi-Buffer 启用前后，阿里云 ASM QPS 的变化。在启用 Multi-Buffer 功能后，阿里云 ASM 的 QPS 有 75% 的性能提升[40]。如果使用的是弹性裸金属节点，提升的性能幅度将更高。

总结和展望

除了 Multi-Buffer 和 AVX-512 之外，阿里云和英特尔在云原生服务网格技术方面还开展了更为广泛的合作。双方正在探索基于第四代至强® 可扩展处理器的深度优化，该处理器内置了针对网络安全专用的加密操作加速器英特尔® QAT，能够将高性能安全性、私钥保护和压缩 / 解压缩等场景的负载从 CPU 卸载到 QAT 中，有效提升应用程序和平台的性能。

[40] 数据援引自阿里云内部测试结果。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。


A 3
B 3
A 4
B 4
EXIT

绿色高效数据中心

阿里云携手英特尔构建绿色高效数据中心，推动液冷技术普惠发展

“浸没式液冷方案能有效降低数据中心能耗水平，并在提升机柜功耗密度的同时，压降成本、故障率以及噪音等，是我们阿里云打造绿色数据中心的重要技术方向之一。来自英特尔产品与技术的支持及与其开展的合作，帮助我们的方案在材料兼容性、芯片电气特性等方面取得巨大突破，使方案在落地实践后取得了巨大成功。”

钟杨帆
资深技术专家
阿里云基础设施服务器研发事业部

为借助更为先进的散热技术来降低 PUE 和 TUE 水平，同时应对机柜功耗密度、散热设备运维成本、设备故障率以及噪音等更多挑战，突破目前多数的数据中心 PUE 值都在 1.5 以上的瓶颈，在液冷技术领域始终位于业界前列的阿里云，正与合作伙伴英特尔协同创新，基于至强® 可扩展平台进行紧密技术协作，推动浸没式液冷技术在数据中心的实践与运用。目前，阿里云单相浸没式液冷技术方案已在阿里云数据中心实现大规模部署并取得了良好的效果，包括 PUE 值达到了极低的 1.09 等[41]，有力证明了该方案可成为阿里云实现双碳目标的有力抓手。

解决方案

多年来，英特尔都通过基于至强® 可扩展平台的软硬件产品组合为阿里云数据中心提供强劲算力引擎，并面向各类云上应用共同实施优化。如大家所熟知的，传统的服务器及其芯片等器件在过去几十年中都是基于风冷设计，浸没式液冷彻底改变了服务器中各种器件的工作环境和使用条件，是一个从未被探索，有大量的问题亟需研究和解决的领域，需要数据中心的各个参与方一起深度协同，来对方案进行设计、验证和优化。

从 2015 年前后起，英特尔与阿里云就开始在浸没式液冷技术领域开展广泛合作，协同开发浸没式液冷服务器，并在材料兼容性、芯片电气特性、服务器系统结构设计以及产业链拓展等方面获得了巨大成功。

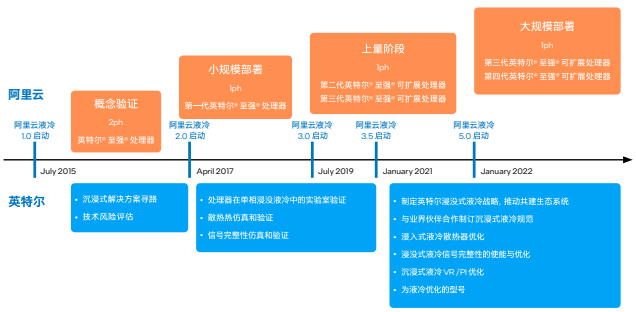


图 2-9-1 英特尔与阿里云在浸没式液冷领域的合作历程

- **材料兼容性**：浸没式液冷中的芯片需要完全浸没在冷却液中工作，芯片浸泡在液体中可能会发生物理特性变化，甚至与液体发生化学反应。即便是非常缓慢的化学反应和物理特性变化都会影响芯片长期运行的可靠性，因此方案必须对浸没在液体环境中的器件开展充分的材料兼容性分析和验证，避免处理器等器件的材料在冷却液中发生特性变化和性能下降。为此，英特尔针对芯片材料兼容性设计了一整套测试方案和数据分析方法，并通过大量的实验来验证至强® 可扩展处理器等硬件产品在浸没式液冷环境下运行的可靠性。

[41] 数据来源于阿里云，如欲了解更多详情，请联系阿里云：https://www.aliyun.com

- **芯片电气特性**：传统风冷服务器中有大量电信号是以空气为介质传输的，而在浸没式液冷方案中这些信号的传输介质就从空气变成了液体。由于空气与液体的电气特性不同，这些电信号尤其是芯片间互连的高速接口信号可能在液体环境中出现波形严重失真、时序错误等信号完整性问题。为应对这些问题，英特尔和阿里云的工程师们对至强® 可扩展处理器、英特尔® Agilex™ FPGA 芯片等的高速接口电路在浸没式液冷的工作环境中重新做了信号完整性仿真分析和测试验证，并与其它零部件和服务器整机厂商合作改进了高速信号连接方案，确保几十乃至上百 G 赫兹、皮秒级别的高速信号在浸没式液冷方案中也能具备与风冷方案相同的信号完整性和系统可靠性指标。

- **服务器系统结构设计**：为了让浸没式液冷方案实现更高的机柜功耗密度，支持更高 TDP 的处理器，也需要对浸没式液冷服务器系统结构做优化设计。英特尔和阿里云的工程师们为此搭建了浸没式液冷服务器系统散热仿真模型和测试验证平台，对服务器内部的液体流场和温度分布、液体的自然对流和强制对流等效应做了深入研究。同时，双方也与各个合作伙伴携手，共同开发高性能的浸没式液冷散热器方案，优化了服务器系统结构设计。

- **产业链拓展**：英特尔与阿里云在业界积极推进液冷技术的标准化，以及跨区域的行业合作。目前，英特尔在 OCP（Open Compute Project，开放计算项目）组织中已发布了多个关于液冷技术的白皮书和设计规范，同时也与阿里云一起，通过与 ODCC（Open Data Center Committee，开放数据中心委员会）组织的紧密合作，全力支持本土标准的制订和技术白皮书的推广。

应用成果

现在，单相浸没式液冷方案已在阿里云多个数据中心中获得成功部署与验证，来自一线应用的数据表明，方案在散热效率等多方面有着显著优势[42]：

- 采用浸没式液冷方案的数据中心 PUE 值可达到极低的 1.09，对比风冷方案的年均 1.5，下降幅度达 34.6%；
- 采用浸没式液冷方案，机柜功耗密度得到显著提升，**单机柜功率可达 100 千瓦以上**；
- 与风冷方案相比，浸没式液冷方案的**资源利用率提升 50% 以上**；
- 与风冷方案相比，浸没式液冷方案的**设备故障率下降了 50%**。

同时，这一方案也为阿里云带来了巨大的成本收益和环境噪音大幅下降等优势。一方面，与风冷方案相比，浸没式液冷方案所需的基础设备更为简洁，无需建设大型冷却塔，而其管道、TANK 也可重复利用且工作寿命非常可观，当服务器等器件需要更新换代时，液冷系统可以跨代使用，无疑可节省大量时间和物料成本；另一方面，浸没式液冷方案不需要风扇等设备，因此工作起来非常安静，对降低环境噪音有着很大助益。

未来展望

阿里云与英特尔携手共建的浸没式液冷方案的成功落地应用，无疑为建设绿色数据中心提供了可参考的实践案例和新的行业标杆。面向未来，双方还将在绿色数据中心技术领域开展进一步合作，使得数据中心更好地实现绿色发展可持续，引领更多行业伙伴共推液冷技术普惠发展，助力“双碳”战略实施，共筑美好未来。

[42] 数据来源于阿里云，如欲了解更多详情，请联系阿里云：https://www.aliyun.com

人工智能优化

英特尔® AMX 助力增强阿里云地址标准化 AI 推理性能

挑战

为了给用户提供高效、精准的地址标准化服务，阿里云机器学习平台团队希望能够在算力平台的构建中，重点化解如下挑战：

- 加速数据清理、模型推理等多个工作负载，加快平台的一站式性能；
- 高效使用现有硬件资源，并充分利用阿里巴巴公有云、私有云和混合云中的服务器资源，降低硬件成本。

解决方案

为提高地址标准化服务的性能，阿里云机器学习平台 (PAI) 团队与英特尔和阿里巴巴达摩院的 NLP 团队开展创新协作。基于第四代至强® 可扩展处理器内置的 AI 加速引擎 -- 英特尔® 高级矩阵扩展 (英特尔® AMX)，优化端到端推理性能 (比前代提升达 2.5 倍)[43]，并且准确率也保持在可接受的范围内。

■ 阿里云地址标准化服务

阿里云地址标准化[44] 是一种高效的标准地址算法服务 (AaaS)，由阿里巴巴达摩院的 NLP 团队依托阿里云海量的地址语料库而开发。该 AaaS 是一个一站式闭环地址数据处理服务平台，其使用 NLP 算法，针对各行业业务系统所登记的地址数据进行纠错、补全、归一、结构化和标签化，实现地址库的清洗和标准化，可提供 20 多种地址服务[45]，并可灵活地部署在公有云、私有云或混合云上。

■ 采用第四代至强® 可扩展处理器优化地址标准化服务

阿里云地址标准化采用 BERT[46] 作为地址标准化搜索模块的核心模型。为了优化 BERT 性能，阿里云 PAI 团队采用了第四代至强® 可扩展处理器内置的英特尔® AMX 等高级特性。

第四代至强® 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 56 个核心，支持 8 通道 DDR5 内存，有效提升了内存带宽与速度，并通过每 PCIe 5.0 (80 个通道) 实现了更高的 PCIe 带宽提升。第四代至强® 可扩展处理器在 AI 性能上更进一步。该处理器内置了创新的英特尔® AMX 加速引擎。英特尔® AMX 针对广泛的硬件和软件优化，它进一步增强了前代技术——向量神经网络指令 (VNNI) 和 BF16，从一维向量发展为二维矩阵，以便最大限度地利用计算资源，提高高速缓存利用率，避免潜在的带宽瓶颈，显著增加人工智能应用程序的每时钟指令数 (IPC)，为 AI 工作负载中的训练和推理提供显著的性能提升。

图 2-10-1 英特尔® AMX 简介

[43] 测试数据配置：单节点，双路英特尔® 至强® 铂金 8369B 处理器 (32 内核) 以及双路英特尔® 至强® 铂金 8475B 处理器 (48 内核)，超线程启用，睿频启用，1 实例 / 内核，BS=32，seq_len=24，数据类型 :INT8 1 实例 / 内核，BS=32，seq_len=24，数据类型：INT8。

[44] https://www.aliyun.com/product/addresspurification/addrp

[45] https://help.aliyun.com/document_detail/169746.html

[46] BERT: 用于语言理解的预训练的深度双向 Transformer (Devlin J、Chang M W、Lee K 等，ACL 2019)。

阿里云地址标准化服务还利用 Blade 优化地址标准化的推理性能，Blade 是阿里云机器学习 PAI 团队引入的一款通用推理优化工具。Blade 集成了多种优化方法，包括计算图优化，优化库，例如英特尔® oneAPI 工具套件，英特尔® 深度神经网络库（英特尔® oneDNN）、BladeDISC 编译器、Blade 高性能运算符库、自定义后端和 Blade 混合精度。

通过将英特尔® Custom Backend [47] 作为 Blade 的软件后端，阿里云地址标准化服务可提升量化和稀疏化推理方面的模型性能。它主要包括 3 级优化：首先使用原始高速缓存策略优化内存，然后优化图融合，最后在运算符级别，构建一个包括自定义和稀疏内核的高效运算符库。

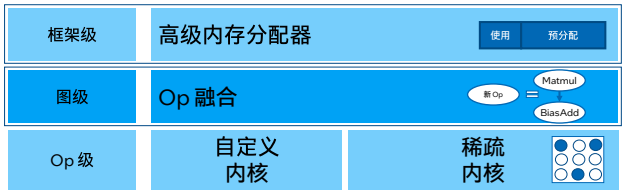


图 2-10-2 英特尔 ® Custom Backend 结构

英特尔® AMX 极大地改进了 INT8 的功能，英特尔® Custom Backend 也可利用英特尔® oneDNN 支持 INT8。因此，相比 VNNI，基于英特尔® AMX 的 INT8 量化可显著提升模型性能。

阿里云和英特尔还对地址标准化模型进行了调整，以提升 PAI Blade 的推理性能。测试数据显示，相比采用 INT8 量化的前代平台，阿里云地址标准化服务 BERT 在第四代至强® 可扩展处理器上性能提升达近 2.5 倍 [48]。

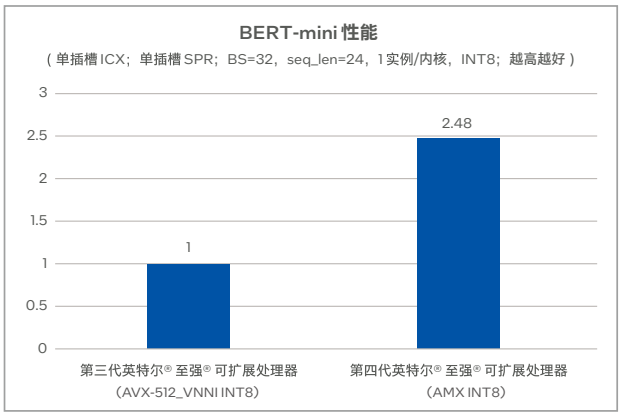


图 2-10-3 BERT 模型的推理性能 [49]

在精度方面，基于 CCKS2021 中文 NLP 地址相关性任务 [50] 的验证显示，基于 FP32 优化的精度仍保持在 78.72，而基于 INT8 优化的模型精度为 78.85。

收益

- 在保证模型精度的前提下，显著提升了 BERT 模型性能，有助于提供更加高效的地址标准化服务；
- 通过软件优化充分释放了硬件潜力，有效利用服务器资源，从而降低了地址标准化服务的 TCO。

展望

为提升更多 DL 模型的端到端性能，英特尔和阿里云正在扩大与客户之间的协作，探索以创新方式优化软硬件集成，以加速 DL 模型的性能，更大限度地发挥英特尔技术的价值。英特尔还希望与行业合作伙伴开展更深入的协作，不断为 AI 技术的部署和实施做出自己的贡献。

[47] https://github.com/intel/neural-compressor/commits/inc_with_engine

[48、49] 测试数据配置：单节点，双路英特尔® 至强® 铂金 8369B 处理器（32 内核）以及双路英特尔® 至强® 铂金 8475B 处理器（48 内核），超线程启用，睿频启用，1 实例 / 内核，BS=32，seq_len=24，数据类型 :INT8 1 实例 / 内核，BS=32，seq_len=24，数据类型 :INT8。

[50] https://tianchi.aliyun.com/competition/entrance/531901/introduction

英特尔助力构建开源大规模稀疏模型训练 / 预测引擎 DeepRec

DeepRec (PAI-TF) 是阿里巴巴集团统一的开源推荐引擎，主要用于稀疏模型训练和预测，可支撑千亿特征、万亿样本的超大规模稀疏训练，在训练性能和效果方面均有明显优势。为解决当前主流的开源引擎对超大规模稀疏训练场景的支持尚有一定局限的问题，DeepRec 基于 TensorFlow1.15 针对稀疏模型场景进行了深度定制优化，主要措施包含以下三类：

- **模型效果：**主要通过增加 EmbeddingVariable (EV) 动态弹性特征功能以及改进 Adagrad Optimizer 来实现优化。EV 功能解决了原生 Variable size 大小难以预估、特征冲突等问题，并提供了丰富的特征准入和淘汰策略等进阶功能；同时，针对特征出现频次进行冷热自动配置特征维度问题，增加了高频特征表达力，缓解了过拟合，能够明显提高稀疏模型效果；

- **训练和推理性能：**针对稀疏场景，DeepRec 在分布式、子图、算子、Runtime 等方面进行了深度性能优化，包括分布式策略优化、自动流水线 SmartStage、自动图融合、Embedding 和 Attention 等图优化、常见稀疏算子优化、内存管理优化，大幅降低了内存使用量，显著加速了端到端训练和推理性能；

- **部署及 Serving：**DeepRec 支持增量模型导出和加载，实现了 10TB 级别的超大模型分钟级别的在线训练和更新上线，满足了业务对时效性的高要求；针对稀疏模型中特征存在冷热倾斜的特性，DeepRec 提供了多级混合存储 (可达四级混合存储，即 HBM + DRAM + PMem + SSD) 的能力，可在提升大模型性能的同时降低成本。

英特尔技术助力 DeepRec 实现高性能

英特尔与阿里巴巴 PAI 团队的紧密合作在实现以上三个独特优势中都发挥了重要作用，DeepRec 三大优势也充分体现了英特尔技术的巨大价值。

在性能优化方面，英特尔超大规模云软件团队与阿里巴巴紧密合作，针对 CPU 平台，从算子、子图、框架、runtime 等多个级别进行优化，充分利用英特尔® 至强® 可扩展处理器的各种新特征，更大程度发挥硬件优势。

为了提升 DeepRec 在 CPU 平台的易用性，还搭建了 modelzoo 来支持绝大部分主流推荐模型，并将 DeepRec 的独特 EV 功能应用到这些模型中，实现了开箱即用的用户体验。

■ 框架优化

DeepRec 集成了英特尔开源的跨平台深度学习性能加速 oneDNN(oneAPI Deep Neural Network Library)， 并且将 oneDNN 原有的线程池修改，统一成 DeepRec 的 Eigen 线程池，减少了线程池切换开销，避免了不同线程池之间竞争而导致的性能下降问题。oneDNN 已经针对大量主流算子实现了性能优化，包括 MatMul、BiasAdd、LeakyReLU 等在稀疏场景中的常见算子，能够为搜广推模型提供强有力的性能支撑，并且 oneDNN 中的算子也支持 BF16 数据类型，与搭载 BF16 指令集的第三代至强® 可扩展处理器同时使用，可显著提升模型训练和推理性能。

在 DeepRec 编译选项中，只需加入“--config=mkl_threadpool”，便可轻松开启 oneDNN 优化。

■ 算子优化

oneDNN 虽可用来大幅提升计算密集型算子的性能，但搜索广告推荐模型中存在着大量稀疏算子，如 Select、DynamicStitch、Transpose、Tile、SparseSegmentMean 等，这些算子的原生实现大部分存在一定的访存优化空间，对此可采用针对性方案实现额外优化。该优化调用 AVX-512 指令，只需在编译命令中加入“--copt=-march=skylake-avx512”即可开启。

■ 子图优化

图优化是当前 AI 性能优化的主要有效手段之一。同样的，当 DeepRec 应用在大规模稀疏场景下时，通常存在着以 embedding 特征为主的大量特征信息处理，并且 embedding 中包含了大量小型算子；为了实现通用的性能提升，优化措施在 DeepRec 中加入了 fused_embedding_lookup 功能，对 embedding 子图进行融合，减少了大量冗余操作，同时配合以英特尔® AVX-512 指令加速计算，最终 embedding 子图性能提升显著。

通过在 tf.feature_column.embedding_column (..., do_fusion=True) API 将 do_fusion 设置为 True，即可开启 embedding 子图优化功能。

■ 模型优化

基于 CPU 平台，英特尔在 DeepRec 构建了涵盖 WDL、DeepFM、DLRM、DIEN、DIN、DSSM、BST、MMoE、DBMTL、ESMM 等多个主流模型的独有推荐模型集合，涉及召回、排序、多目标等多种常见场景；并针对硬件平台进行性能优化，相较于其他框架，为这些模型基于 Criteo 等开源数据集在 CPU 平台上带来极大的性能提升。

其中表现最突出的当属混合精度的 BF16 和 Float32 的优化实现。通过在 DeepRec 中增加自定义控制 DNN 层数据类型的功能，来满足稀疏场景高性能和高精度的需求；开启优化的方式如图 2-11-1 所示，通过 keep_weights 保留当前 variable 的数据类型为 Float32，用于防止梯度累加导致的精度下降，而后再采用两个 cast 操作将 DNN 操作转换成 BF16 进行运算，依托第三代至强® 可扩展处理器所具备的 BF16 硬件运算单元，提升 DNN 运算性能，同时通过图融合 cast 操作进一步提升性能。

```
# DNN Layers
dnn_scope = tf.variable_scope('dnn_layers')
with dnn_scope.keep_weights(dtype=tf.float32) if ENABLE_BF16 else dnn_scope:
  if ENABLE_BF16:
    net = tf.cast(net, dtype=tf.bfloat16)

  net = self.dnn(net, self.dnn_hidden_units, "hiddenlayer") # BF16 datatype

  if ENABLE_BF16:
    net = tf.cast(net, dtype=tf.float32)
```

图 2-11-1 混合精度优化开启方式

为了能够展示 BF16 对模型精度 AUC(Area Under Curve) 和性能 Gsteps/s 的影响，针对现有 modelzoo 的模型都应用以上混合精度优化方式。阿里巴巴 PAI 团队使用 DeepRec 在阿里云平台的评测表明[51]，基于 Criteo 数据集，使用 BF16 优化后，模型 WDL 精度或 AUC 可以逼近 FP32，并且 BF16 模型的训练性能提升达 1.4 倍，效果显著。

同时，针对超大规模稀疏训练模型 EV 对存储和 KV 查找操作的特殊需求，基于英特尔® 傲腾™ 持久内存（简称“PMem”）的内存管理和存储方案，可支持和配合 DeepRec 多级混合存储方案，满足了大内存和低成本需求；可编程解决方案事业部团队使用 FPGA 实现对 Embedding 的 KV 查找功能，大幅提升了 Embedding 查询能力，同时可释放更多的 CPU 资源。

结合 CPU、PMem 和 FPGA 的不同硬件特点，从系统角度出发，针对不同需求更加充分地发挥英特尔软硬件优势，可加速 DeepRec 在阿里巴巴 AI 业务中的落地，并为整个稀疏场景的业务生态提供更优的解决方案。

总结

英特尔为 AI 应用提供了多样化的硬件选择，为客户选择更优性价比的 AI 方案提供了可能。与此同时，英特尔与阿里巴巴及广大客户正一同基于多样化硬件实施软硬一体的创新协作和优化，从而更充分地发挥英特尔技术和平台的价值。英特尔也期望继续和业界伙伴展开更深入地合作，持续为 AI 技术的部署落地贡献力量。

[51] 如欲了解更多性能测试详情，请访问 https://github.com/alibaba/DeepRec/tree/main/modelzoo/WDL

分布式 AI 推理助力阿里云实时计算

Apache Flink 是实时流计算的代表性框架，其可实现毫秒级低时延的实时流数据计算，且拥有丰富的使用场景和活跃的用户社区。基于这一框架，阿里云构建了实时计算 Flink 版 (Alibaba Cloud Realtime Compute for Apache Flink，Powered by Ververica)，来为业界越来越多的实时 AI 场景提供多途径支持，包括 Alink 机器学习算法库和 AI Flow 大数据 AI 平台等。同时，依托与英特尔在大数据及 AI 领域的紧密合作，阿里云在实时计算 Flink 中集成了 Analytics Zoo Cluster Serving 来构建 AI 推理解决方案。

阿里云实时计算 Flink 版

阿里云实时计算 Flink 版是阿里云基于 Apache Flink 构建的企业级、高性能实时大数据处理系统，其运行于 Yarn 或者 Kubernetes 提供的调度系统上。Kubernetes 运行在阿里云的基础架构上，包括阿里云神龙服务器、阿里云 ECS 实例，以及阿里云存储解决方案等。阿里云实时计算 Flink 版总体架构如图 2-12-1 所示：

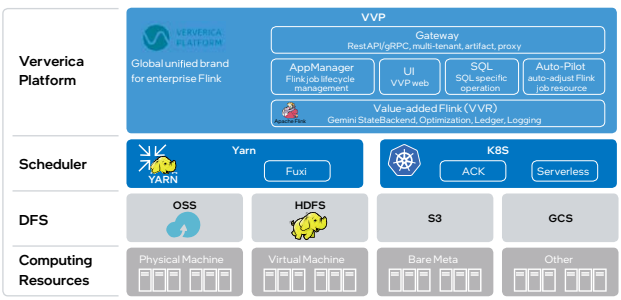


图 2-12-1 阿里云实时计算 Flink 版总体架构

阿里云实时计算 Flink 版性能优越，可 100% 兼容开源 Flink 接口，提供百万级作业吞吐，计算时延低至毫秒级，并支持数十种作业指标监控；同时，其还具有作业智能调优功能，提供一站式开发界面和智能诊断系统。

Analytics Zoo Cluster Serving

Analytics Zoo [52] 是英特尔开源的、统一的大数据分析和人工智能平台，其可将 TensorFlow、Keras、PyTorch 等 AI 训练和推理框架无缝扩展到分布式 Apache Spark、Flink、Ray 等大数据平台上运行。用户可以利用 Analytics Zoo 提供的分布式 AI 推理解决方案 Cluster Serving，快速构建运行于 Apache Flink 之上的实时 AI 推理服务，仅需极少量代码和指定极少量信息，即可构建一个可扩展的推理流水线。

使用阿里云实时计算 Flink 版和 Analytics Zoo Cluster Serving 构建 AI 推理解决方案

在阿里云实时计算 Flink 版 2.4.2 及以上版本的内置函数中，集成了 Analytics Zoo Cluster Serving 的 AI 推理功能。

用户在 Flink 应用程序调用 `CLUSTER_SERVING ` 函数，只需要提供模型文件和数据文件，即可使用 TensorFlow、PyTorch 和 OpenVINO™ 工具套件的推理模型，完成机器学习端到端的应用。

开发者可以参考以下语法要求来定义 Cluster Serving 的数据：

`CLUSTER_SERVING ` 函数语法：

LOAD MODULE `cluster-serving `;

CLUSTER_SERVING (uri, `data `)

其中，`CLUSTER_SERVING ` 函数参数定义如下：

参数	数据类型	说明
uri	String	数据 ID
data	String	数据

表 2 `CLUSTER_SERVING ` 函数参数定义

[52] BigDL 2.0 已包含 BigDL 和 Analytics Zoo。

使用 Analytics Zoo Cluster Serving 搭建推理解决方案示例

如下将介绍使用 Analytics Zoo Cluster Serving 在阿里云实时计算 Flink 版搭建分布式推理的方法。本样例将采用一个简单深度学习模型 AutoEncoder，模型输入为一个向量（一维数组），通过自动编码模型输出一个向量。

以输入格式为 input_shape=(?,128) 的 TensorFlow Saved Mode 模型为例，在阿里云实时计算 Flink 版使用 Cluster Serving 实现在线推理的步骤如下：

1. 登录阿里云 OSS 控制台，上传测试数据 input.csv 文件至 oss: //***/input.csv 目录；
2. 在阿里云实时计算 Flink 版控制台，创建 Flink 源表和结果表，加载 `cluster-serving ` Function Module，并且使用 `CLUSTER_SERVING ` 推理函数预测输入数据插入结果表；
3. 在作业高级配置的更多 Flink 配置和 Preview Session 集群 ** 其他配置 ** 均增加以下配置：pipeline.global-job-parameters:'"modelPath:""[模型目录地址]"""' 其中，模型目录地址为模型文件的 OSS 存放目录，例如 oss://***/tf_auto/；
4. 在作业运维界面，单击目标作业操作列下的运行。

阿里云实时计算 Flink 版和 Analytics Zoo Cluster Serving 的应用场景

通过在阿里云实时计算 Flink 版中集成 Analytics Zoo Cluster Serving，用户可以在诸多场景中方便地进行分布式 AI 推理服务。例如，通过实时行动轨迹的搜索来协助病毒传播防控，或者通过实时过滤恶意点击来提高推荐系统的有效性。

阿里云实时计算 Flink 版集成的 Analytics Zoo Cluster Serving 还可用于提高推荐系统的有效性。为了满足不同用户的个性化需求，电商平台会根据用户的兴趣爱好推荐合适的商品，从而满足客户千人千面的商品需求。商家为了获取更多的平台曝光量，可能会别有用心地利用平台的推荐机制，增加自家商品的曝光机会，典型手法就是“抱大腿”攻击，其本质是通过大量协同点击目标商品和爆款商品，建立目标商品与爆款商品之间的关联关系。而使用 Analytics Zoo Cluster Serving 和 Apache Flink 可实现对恶意流量的实时识别，并保护识别程序的数据安全。

图 2-12-2 为一个可以实现上述场景的实时 AI 训练和推理方案的技术架构：

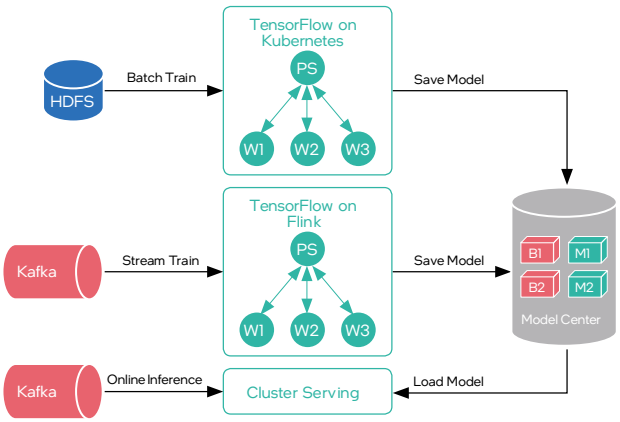
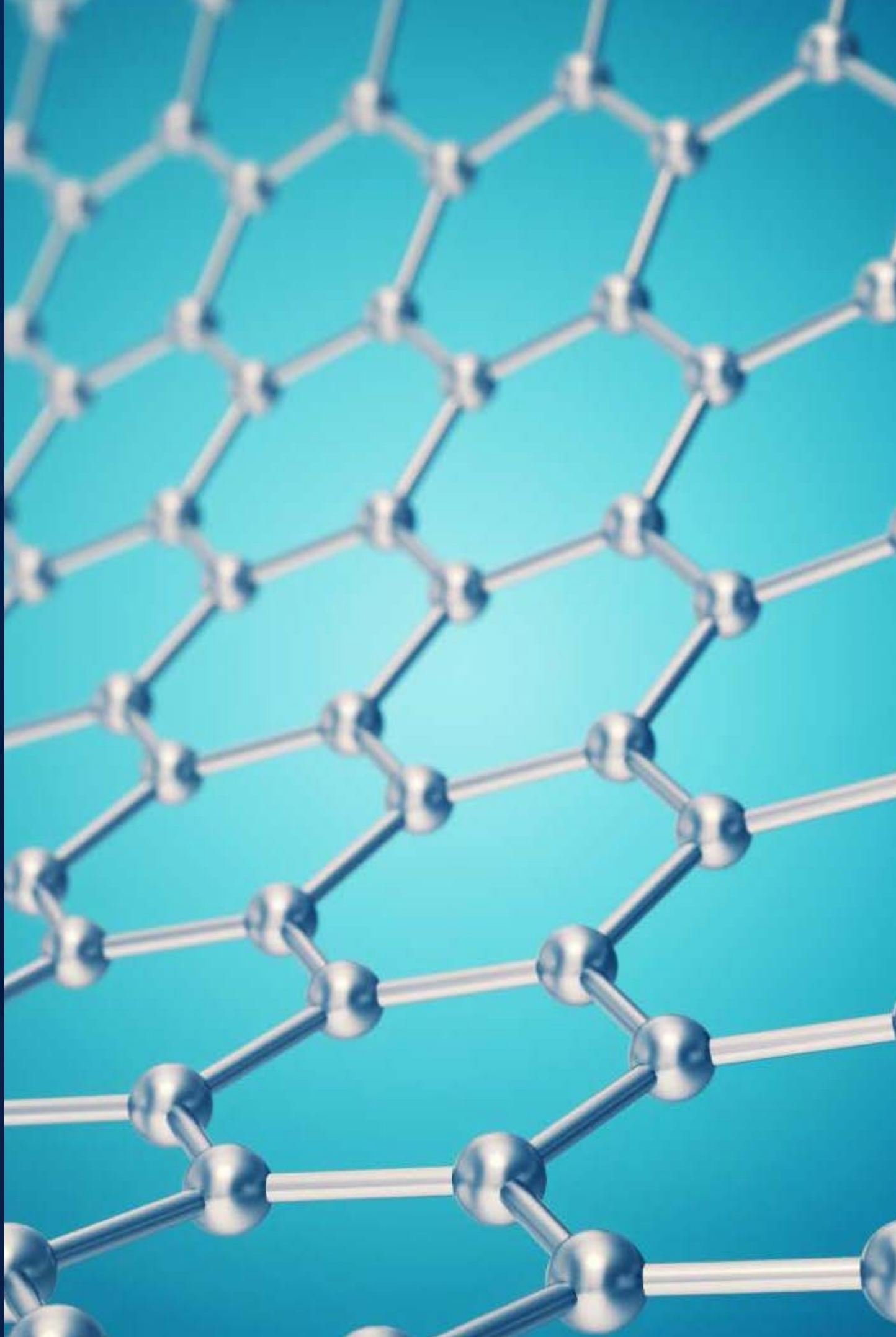


图 2-12-2 实时 AI 训练和推理架构概览

总结

实时计算 Flink 版已经广泛应用于大数据分析实时化的场景，其与深度学习相结合，可支持构建更多 AI 应用场景，例如实时风控、实时异常检测和实时推荐等。阿里云和英特尔紧密合作，在阿里云实时计算 Flink 版中集成了 Analytics Zoo Cluster Serving，使阿里云平台可提供产品化的实时 AI 推理能力，为高效进行实时大数据开发和方案部署提供了一体化平台。



科学计算加速

基于阿里云 E-HPC 的英特尔® 模拟和仿真精选解决方案

“科学计算上云，首先要考虑的是满足性能，要具备物理机的无损性能，同时提供低时延、高带宽网络和并行文件存储。阿里云通过神龙架构弹性裸金属超级计算集群 SCC 满足了上述三者需求，充分释放基于英特尔® 至强® 可扩展处理器的服务器平台算力，并针对云平台优化。除此之外，现代仿真企业客户需要长时间稳定地使用科学计算仿真资源，还需要满足超越性能之上的需求。这时候，阿里云弹性高性能计算 E-HPC 弹性伸缩、主动运维、热迁移等云特性，满足了客户希望算力随需伸缩和自动化运维的需求，这才是云上科学计算服务的关键所在。”

何万青博士
阿里云高性能计算负责人

模拟仿真已成为大量行业用户在进行目标系统设计时，为满足功能、性能、功耗和其他指标要求所需要的一项重要业务流程。由于实际系统的复杂度、精细度的快速提升，需要进行模拟仿真的需求不断增长，且效率要求也不断提升，使得模拟仿真对算力基础设施提出了苛刻要求。在众多行业场景中，用户已经广泛部署涵盖数十乃至上百节点的科学计算集群，对上千场景节点进行并行仿真。然而，传统的用于模拟仿真的科学计算集群的组建和应用存在性能短板、部署流程复杂、无法敏捷扩展算力和运营成本高昂等挑战。

基于阿里云 E-HPC 的英特尔® 模拟和仿真精选解决方案

阿里云 E-HPC 弹性高性能计算是典型的云超算系统，由 PaaS 层的 E-HPC 及其算力底座 SCC（Super Computing Cluster，超级计算集群）组成。通过 E-HPC 平台，用户可灵活定制基于任何 ECS 和异构实例构成的科学计算集群，满足不同应用特征的性价比要求。在弹性裸金属服务器基础上，阿里云 E-HPC 加入高速 RDMA 互联支持，提供了高带宽、低时延优质网络，降低了大规模集群的网络瓶颈。

阿里云 E-HPC 可以满足模拟仿真负载对于算力的苛刻需求。在集群内，各节点间通过 RDMA 网络互联，提供高带宽、低时延网络，保证了科学计算和人工智能、机器学习等应用的高度并行需求。同时，RoCE（RDMA over Convergent Ethernet，基于融合以太网的 RDMA）在实现了更高网络性能的同时，能够支持更广泛的基于 Ethernet 的应用，实现真正的云上科学计算。

阿里云 E-HPC 拥有多种实例类型，部分实例采用了第三代至强® 可扩展处理器。该处理器能够在计算、存储和网络应用中，为计算密集型工作负载提供高性能和可扩展性。得益于英特尔® 超级通道互联（英特尔® UPI）、英特尔® Infrastructure Management 技术（英特尔® IMT）、英特尔® AVX-512、英特尔® 高级加密标准新指令 AES-NI 等领先功能，该处理器可满足严苛的 I/O 密集型工作负载的需求，帮助阿里云打造高性能、高安全的敏捷服务和突破性功能。

通过使用内存计算和针对至强® 可扩展处理器优化的开源库，阿里云 E-HPC 能够支持用户处理大型数据集，同时运行广泛的模拟与仿真应用。此外，该解决方案也支持用户创建逼真的交互可视化，以便更快地获取洞察，更高效地展示新产品设计和研究突破。

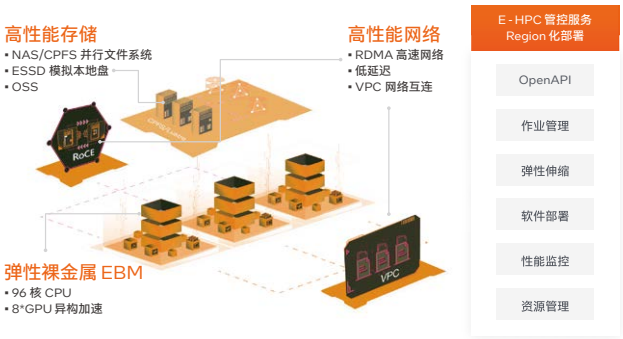


图 2-13-1 阿里云弹性高性能计算平台 E-HPC 架构

验证：针对模拟和仿真工作负载进行验证的强大性能表现[53]

阿里云和英特尔选取了高性能 Linpack（HPL）、高性能 Conjugate Gradient（HPCG）、DGEMM 和 STREAM 四个知名行业性能指标进行基准性能验证，满足或超过了其设计与测试标准，并针对模拟和建模应用工作负载显示出了纵向和横向的性能扩展潜力。

- HPL：HPL 是针对现代并行计算提出的测试方式，测试通过在分布式内存上求解双精度数学运算，来衡量受测系统的并行计算能力；
- HPCG：评测模拟真实应用的数据访问模式，如稀疏矩阵运算、测试内存子系统和内部互连，它还提供了查看单个节点性能和整个系统的综合性能的能力；
- DGEMM：一个双精度通用矩阵乘法工作负载，用于测量处理器和内存的计算能力；
- STREAM：用于测量简单矢量内核的可持续内存带宽和相应的计算速率。

双方在阿里云高主频计算型超级计算集群实例规格族 scchfc6 上进行了测试，该实例采用了英特尔® 至强® 铂金 8269 处理器，处理器与内存配比为 1:2.4，存储为 I/O 优化实例，同时支持 RoCE 网络和 VPC 网络。

测试数据显示，基于英特尔® 至强® 可扩展处理器的阿里云 E-HPC 可满足基于模拟和建模的英特尔® 精选解决方案指定的最低工作负载优化性能和功能水平，能够支持用户更高效地运行模拟与仿真应用。

Alibaba SCC		ecs.scchfg6.20xlarge		英特尔性能要求
评测指标	执行方法	阿里云结果：3.1 GHz 英特尔® 至强® 铂金 8269 处理器 @ 3.1 GHz（Cascade Lake）@ 30 Gbit/s 以太网带宽		基于英特尔® 至强® 金牌 6248 处理器（2.50 GHz，20 核 /40 线程）或英特尔® 至强® 金牌 6148 处理器（2.40 GHz，20 核 /40 线程）
HPL	跨所有 4 个节点	10586.9	GFLOP/s	超过 7700 GFLOP/s
HPCG	跨所有 4 个节点	137.31	GFLOP/s	超过 127 GFLOP/s
	在每个节点上	39.12	GFLOP/s	超过 32 GFLOP/s
DGEMM	在每个节点上	3032.78	GFLOP/s	超过 2048 GFLOP/s
STREAM	在每个节点上	192.237	GB/s	超过 164000.00 MB/s

表 3 基于阿里云 E-HPC 的英特尔® 模拟和仿真精选解决方案测试数据

总结：为模拟仿真提供弹性的高性能算力

阿里云与英特尔合作验证了基于阿里云 E-HPC 的英特尔® 模拟和仿真精选解决方案的性能指标，为希望在云上科学计算集群中快速部署模拟仿真应用的最终用户提供了参照。用户可以参考该精选解决方案，在阿里云 E-HPC 中快速搭建模拟和仿真应用，实现高性能、高扩展性、高安全性等优势的统一。

[53] 截至 2021 年 8 月的阿里云与英特尔内部测试结果。阿里云 E-HPC 配置：英特尔® 至强® 铂金 8269 处理器 @3.1 GHz(Cascade Lake) @30 Gbit/s 以太网带宽。英特尔® 精选解决方案 Plus 配置：基于英特尔® 至强® 金牌 6248 处理器 (2.50GHz，20 核 /40 线程) 或英特尔® 至强® 金牌 6148 处理器 (2.40GHz，20 核 /40 线程)。

开发软件优化

基于至强® 可扩展平台，多重优化方案助阿里巴巴 Noslate 性能加速

“随着更多基于 Node.js 的业务应用被部署到云原生、Serverless 等新一代云服务架构中，阿里巴巴也在积极探索对 Node.js 的优化改进并推出了 Noslate 产品。为了让 Noslate 获得更优性能，我们与英特尔一起基于第四代英特尔® 至强® 可扩展处理器等产品开展了一系列针对性的优化，并取得了显著的性能提升。”

李三红
程序语言与编译器技术总监
阿里云

方案背景

随着云原生、Serverless 等新一代云服务架构获得越来越多的部署，为使各类应用在新架构中具有更好的性能与弹性，进而满足业务在泛终端、全栈交付等领域的需求，阿里巴巴借助长期的技术积累与创新探索，在 Node.js 的基础上打造了 Noslate 这一面向 Serverless 架构和云原生场景的 JavaScript 容器方案。

作为一款高效、弹性和完全可定制的 Serverless 运行框架，Noslate 可为云上各类工作负载提供大量的高级功能和优化组件，例如 Node.js 发行版、Noslate Workers 和 Noslate Debugge。

性能的持续优化与应用模式的不断成熟，让 Noslate 在阿里巴巴众多业务场景中承担越来越重要的角色，并逐渐成为主流的服务端负载之一。而包括一系列至强® 可扩展处理器在内的英特尔先进产品与技术，一直以来都是阿里巴巴旗下阿里云的重要性能引擎。借助英特尔在 Node.js 上的优化经验，阿里巴巴和英特尔携手制订的优化方案围绕基于英特尔® 架构的平台及相关软硬件技术展开。后续的验证测试数据表明，优化方案能在真实工作负载下，令 Noslate 执行性能获得显著提升，启动时间得以大幅降低。

优化方案

基于英特尔® 架构平台和 Noslate 的优化方案可分为 11 个优化项，其中优化项 1、2 是针对英特尔® 架构平台所具备的向量化硬件特性而专门设计，其它优化项则可用于英特尔及其它平台（*如欲了解更多其它 9 个优化项详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/alibaba-noslate-performance-accelerate-optimize.html*）。但一些测试同样表明，这些优化项在基于英特尔® 架构的平台上有着更优的表现。

优化一：利用异步 Nginx 优化加速加解密性能

在信息隐私愈发受到重视的今天，互联网应用普遍都采用 HTTPS（Hypertext Transfer Protocol Secure，超文本传输安全协议）等更安全的网络协议来巩固数据安全。但这也带来了巨大的加解密计算工作量，进而对应用整体性能带来影响。英特尔® QAT 面向 Nginx 加入了专门的优化，令其可工作在异步模式下，即应用无需等待加解密任务完成就能够继续下一步工作，这显然可以大幅提升应用中 HTTPS 通信的性能。值得一提的是，由于英特尔® QAT 能良好地工作在软件模式下，因此 Noslate 不需要加入专门的英特尔® QAT 驱动程序来实现此优化。在面向阿里巴巴

真实工作负载的测试中，这一优化能帮助 ghost.js 提升其 HTTPS 工作负载的 TPS（Transactions Per Second，每秒事物处理量）性能 15% 以上 [54]。

优化二：利用 SIMD 指令提升缓冲区换位操作的性能

与 Node.js 一样，在 Noslate 等构建的服务端应用中，未被应用处理的二进制数据会被暂时存放在缓冲区（Buffer）中，因此 Noslate 也定义了许多缓冲区交换 API 函数来对这些数据进行操作。例如需对缓冲区中的数据进行字节顺序调整时，就会用到 Buffer.swap16()、Buffer.swap32() 及 Buffer.swap64() 等函数。

传统上，这种交换操作只能采用顺序方式，在数据量较大的情况下显然缺乏效率。而优化项是利用基于英特尔® 架构的处理器所具备的 SIMD 硬件指令，包括英特尔® AVX-512 等，让上述函数的工作模式从顺序方式变为并行方式，从而有效提升交换效率。在实战中，长度超过 128 字节的缓冲区能通过这一优化项获得更佳优化效果。

方案验证 [55]

为验证上述 11 个优化项的效果，阿里巴巴与英特尔一起基于第四代至强® 可扩展处理器，在云实例上开展了交叉测试，并覆盖了 Ubuntu、Anolis 等不同的操作系统。测试方案基于英特尔开发的 Node.js 性能测试场景展开，这些测试项都来自现实世界开源的 Node.js 应用代码或 Node.js 官方的性能测试案例。测试场景包括 Ghost.js、Web Tooling 和 Function Computing Cold Start up（**函数式计算冷启动时间**）。

结合以上不同测试场景，在每个工作负载都加入所有可用优化项后，阿里巴巴与英特尔基于阿里云最新的第八代企业级 ECS 实例 g8i（配置第四代至强® 可扩展处理器）开展了测试。测试结果中，Full Server Load 意味着对运行相同工作负载的多个节点实例（基于超线程数的 64 个实例）进行测试。测试结果如下：

■ 吞吐量获益（Ubuntu 操作系统）

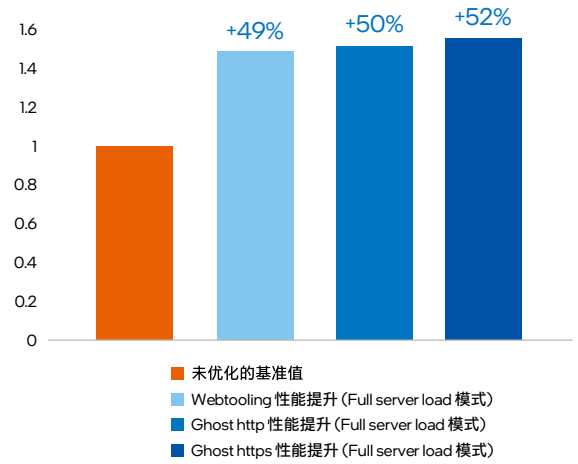


图 2-14-1 Ubuntu 操作系统中优化方案带来的吞吐量获益（归一化）

■ FC 启动时间降低（Anolis 操作系统）

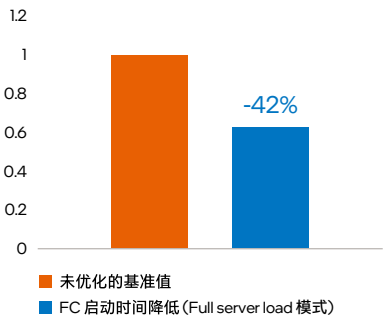


图 2-14-2 Anolis 操作系统中优化方案带来的 FC 启动时间降低（归一化）

综合来看，上述 11 项优化项在基于第四代至强® 可扩展处理器的平台上都能为 Noslate 工作负载带来显著性能提升，在不同的操作系统中，不同测试场景下的吞吐量获益可达 49% 至 61%，而 FC 启动时间也能分别降低 38% 和 42%。

展望

面向未来，双方计划以龙蜥社区为合作载体，进一步在 Node.js、WebAssembly 等技术领域上，围绕第四代至强® 可扩展平台的特性开展优化，并构建标准化、有代表性的性能测试集来提升真实工业级框架的性能。

[54] 如欲了解更多性能配置详情，请参阅白皮书《Optimize Node.js for Noslate on Intel platforms》，https://github.com/noslate-project/node-benchmark/blob/intel-whitepaper/paper/Intel_Optimization_Noslate_SPR_Update.pdf

[55] 如欲了解测试配置及更多详细测试结果，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/alibaba-noslate-performance-accelerate-optimize.html

优势互补，英特尔助阿里巴巴 Dragonwell11 与 VectorAPI 实现融合，探索 Java 性能提升新途径

方案背景

阿里巴巴多项业务应用大多通过 Java 语言开发，而 Java 应用的开发及运行效率与 Java 软件开发工具包性能息息相关，因此阿里巴巴对相关工具包的优化和革新也尤为关注。

为此，阿里巴巴在开源 OpenJDK（Java 软件开发工具包 JDK 的开源版本）的基础上推出了 OpenJDK 的下游版本 Alibaba Dragonwell，并与 OpenJDK 社区紧密合作，将更多的 Dragonwell 定制功能带到上游。随着 AI、大数据处理以及音视频处理等高密度计算负载在阿里巴巴的业务场景中占据更重要的位置，如何让 Dragonwell 与数据中心基础算力设施产生共鸣，从而以更强的数据处理性能来应对新场景、新业务的需求，形成优势互补的效果，也成为阿里巴巴 Java 开发团队所关注的焦点。

Java VectorAPI 是一种利用向量计算能力，让 Java 应用在特定领域实现优异计算性能的利器，为了让 VectorAPI 在 Dragonwell 中发挥所长，长期致力于 OpenJDK 发展并贡献了大量优化特性的英特尔与阿里巴巴一起，通过深度技术合作，将 VectorAPI 成功移植到 Dragonwell 中。

解决方案

为应对海量数据时代更多、更大的密集型计算负载，今天的处理器正通过 SIMD 硬件指令来提升效率，其能让一个指令单次操作多条数据，显著提升处理效率。

英特尔在 1996 年就率先加入了对 SIMD 指令的支持。在今天的至强® 可扩展平台上，英特尔® AVX-512 已能为应用提供 512 位的寄存器和 512 位的 FMA（Fused Multiply and Add，融合乘加）单元，使之可同时执行 32 次双精度、64 次单精度浮点运算，或操作八个 64 位和十六个 32 位整数，从而在 AI、大数据处理、多媒体处理以及游戏娱乐等高密度计算场景中获得数倍以上的性能提升。

为了让 SIMD 硬件指令集发挥效能，传统上（例如在较老版本的 Java 中），开发者在直接针对 SIMD 编程时，需要通过 JNI（Java Native Interface，Java 原生接口）调用由 C/C++ 或汇编语言构建的库来实现。使用 JNI 不仅会带来不容忽视的额外性能开销，也会因混合编程模式而增加系统管理和维护的复杂度。

随着 Java 编译器 / 解释器（如 JVM(Java Virtual Machine，Java 虚拟机））的迭代升级，此类代码也可由编译器 / 解释器提供的自动向量化功能完成。这虽然降低了开发工作量，但由于自动生成的代码往往未经充分优化，因此在性能上往往无法达到预期效果。

VectorAPI 的出现，为上述问题带来了解决之道。其不仅能让 Java 开发者直接面向 SIMD 进行编程，让代码获得充分优化，又避免了 JNI 这一中间件的介入，降低了性能开销和系统复杂度。

这使得 Java 应用能够充分获得 SIMD 带来的计算处理优势，如图 2-15-1 示例代码所示，使用 VectorAPI 后，在内置英特尔® AVX-512 的处理器上，一次加法可处理（512/32=16）16 个浮点数，而传统加法一次只能处理一个浮点数。

```
void add (float[] A, float[] B, float[] C) {
    for (int i = 0; i < C.length; i++) {
        C[i] = A[i] + B[i];
    }
}

public class AddClass<S extends Vector.Shape<Vector<?>, ?>>> {
    private final FloatVector.FloatSpecies<S> spec;
    AddClass (FloatVector.FloatSpecies<S> v) {spec = v; }

    // vector routine for add
    void add (float[] A, float[] B, float[] C) {
        int i=0;
        for (; i + spec.length() < C.length; i += spec.length()) {
            FloatVector<S> av = spec.fromArray(A, i);
            FloatVector<S> bv = spec.fromArray(B, i);
            av.add (bv).intoArray(C, i);
        }

        // clean up loop
        for (; i < a.length; i++) C[i] = A[i] + B[i];
    }
}
```

图 2-15-1 使用 VectorAPI 的示例代码

随着内置英特尔® AVX-512 的英特尔® 处理器，包括第三代及第四代至强® 可扩展处理器在各类数据中心中承担更重要的算力任务，在 AI、数据库以及图像 / 多媒体处理等一系列实战中，VectorAPI 也为不同的 Java 应用带来了数倍的性能提升。例如在某种 Parquet 文件（一种广泛用于大数据处理和分析领域的列式存储格式）解码器中，其被验证可带来 4-8 倍的性能提升 [56]。

得益于以上的优异性能表现，业界对 VectorAPI 的发展也表现出巨大的热情和关注。从其作为孵化器项目被 OpenJDK16 引入开始，每一次 OpenJDK 版本的升级也会对其进行同步升级，例如目前 OpenJDK 20 中部署的 VectorAPI 版本为 JEP（JDK Enhancement Proposals，JDK 增强提案）438，Fifth Incubator。

但在实践中，VectorAPI 的推广还存在一个巨大障碍。业界目前使用最广泛的 Java LTS（Long-Term Support，长期支持）版本 JDK11 并不支持 VectorAPI，需要到下一个 Java LTS 版本 JDK17 中才予以支持，而在成熟的生产环境中升级 Java 版本显然是一件代价高昂的事情。

虽然各头部企业已着手尝试解决这一问题，例如阿里巴巴就在其内部使用的 AJDK 中加入了对 VectorAPI 的支持，但这与 OpenJDK 社区级别的实现还是有着巨大差别，且后期升级维护也会出现困难。为此，阿里巴巴与英特尔一起，结合双方在 OpenJDK 发展、Java 开发以及面向 SIMD 硬件指令集调优等方面的技术优势，将 VectorAPI（JEP 338）移植到 Dragonwell11（基于 JDK11）上，既让基于 Dragonwell11 的大量 Java 应用能利用到 VectorAPI 的强大功能，又保护了现有的投资，避免了升级 JDK 带来的风险。

但这并非易事，双方的协作面临着一系列的挑战。包括：

- 移植代码量非常巨大，根据统计涉及代码行数高达 29 万行；
- 移植工作需要让 Dragonwell11 和 VectorAPI 分别能方便地跟踪到上游 OpenJDK11 和后续 VectorAPI 版本的修正、增强和演进，具有相当强的技术挑战；
- 移植后的应用性能要和上游 OpenJDK 接近，并使得 OpenJDK11 及后续版本的性能改动可以方便引入到 Dragonwell11 中；
- 移植工作要保证现有基于 Dragonwell11 开发部署的大量业务具有稳定可靠性。

为了让移植工作达到预期目标，阿里巴巴与英特尔双方都投入了大量资源，包括上游 OpenJDK 社区 VectorAPI 的关键贡献者也积极参与其中。双方携手对海量代码进行了细致地梳理、排查和分析，并就移植过程开展了不同层面、不同维度的讨论，并确定了最终的实施方案。

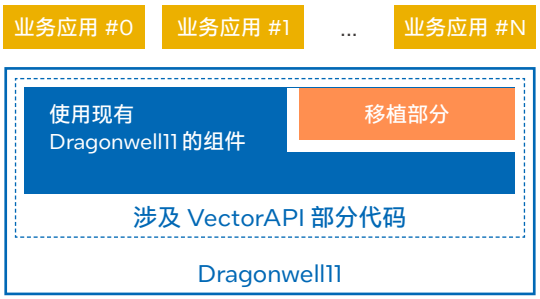


图 2-15-2 尽量保持 Dragonwell11 原有结构的移植方案

如图 2-15-2 所示，为了确保上层业务的稳定可靠性，整个移植方案尽量保持了 Dragonwell11 的原有结构，只有涉及 VectorAPI 的部分才采用上游的实现。而对一些上游的改动，方案也对其实现进行了有针对性地分析，如能使用现有 Dragonwell11 的组件就不进行改动，从而把对 Dragonwell11 的影响降到最小。

在第一阶段移植工作结束后，双方还一起利用 OpenJDK 内建的测试集，对移植后的 Dragonwell11 开展了完整的覆盖性测试，以确保移植质量。测试结果表明，移植后的 Dragonwell11 在性能增幅、兼容性等多个维度都符合预期目标。

总结与展望

目前，经过阿里巴巴与英特尔双方的努力，VectorAPI 已经成功合并到 Dragonwell11 的主分支（master）上，且完全兼容 VectorAPI 1st Incubator(JEP 338)。双方将继续把 VectorAPI 近年来的升级版本，包括 JEP 414、JEP 417、JEP 426、JEP 438 的功能都逐一移植到 Dragonwell11 上。

[56] 相关数据援引自：https://github.com/apache/parquet-mr/pull/1011

案例篇

千行百业落地

第四代英特尔® 至强® 可扩展处理器助力阿里巴巴电子商务推荐系统实现性能突破

为了推动推荐系统创新，充分释放推荐系统在提升用户体验、助力商业价值挖掘等方面的价值，阿里巴巴构建了核心推荐模型，以负责处理天猫和淘宝全球庞大客户群发出的数亿实时请求，并与英特尔合作实施了一项优化下一代推荐系统的联合方案。该方案采用了第四代至强® 可扩展处理器，利用该处理器内置的英特尔® AMX 高级硬件特性，为阿里巴巴的核心推荐模型带来 AI 推理性能突破，并保证足够的精度。

挑战

现代化推荐系统对于 AI 算力有着较高的要求，为了实现性能与成本的平衡，阿里巴巴在推荐系统中采用了 CPU 处理 AI 推理等工作负载。但同时，这一推荐系统面临着如下 AI 推理挑战：

- **AI 推理在吞吐量与时延方面的要求**
  阿里巴巴核心推荐模型不仅需要在单位时间内处理海量的请求，还必须确保处理时间在严格的时延阈值范围内，以实现出色的用户体验。

- **确保 AI 推理精确性，保证推荐质量**
  较低精度的数据类型有助于缩减数据大小，优化内存访问，进而缩短时延并提高吞吐量，但同时也会对推理精度带来影响。阿里巴巴希望能够在优化推理性能的同时，确保推荐质量达到理想的水平。

采用第四代英特尔® 至强® 可扩展处理器提升推荐性能

面对爆炸式增长的用户数据，以及不断扩展的业务处理压力，阿里巴巴希望能够持续提升核心推荐系统的性能，同时在基础设施的灵活性、敏捷性、TCO 等方面实现平衡。为此，阿里巴巴选择了第四代至强® 可扩展处理器进行性能优化。

第四代至强® 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 56 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。第四代至强® 可扩展处理器在 AI 性能上更进一步。该处理器内置了创新的英特尔® AMX 加速引擎。英特尔® AMX 架构和指令的功能类似于脉动阵列，提供矩阵类型的运算，可以高效处理两个矩阵之间的乘法，同时支持 INT8 和 BF16 数据类型，能够确保该 CPU 像高端通用图形处理器（GPGPU）一样处理 DNN 工作负载，显著增加了人工智能应用程序的每时钟指令数（IPC），可为 AI 工作负载中的训练和推理提供强劲动力。

阿里巴巴还使用英特尔® oneAPI 深度神经网络库（英特尔® oneDNN），将 CPU 微调到峰值效率。oneDNN 是英特尔® oneAPI 工具套件的一部分，并集成到 TensorFlow 和 PyTorch 框架等许多工业软件中，它抽象出指令集和其他复杂的性能优化，提供了高度优化的深度学习构建块实现。通过这一开源、跨平台的库，深度学习应用程序和框架开发人员可以在 CPU、GPU 或两者之间使用相同的 API。

阿里巴巴与英特尔合作，集成上述所有硬件和软件特性，并将其应用于阿里巴巴核心推荐模型的整个堆栈。

优化后的软件和硬件已经部署在阿里巴巴的真实业务环境中，它们成功通过了一系列验证，符合阿里巴巴的生产标准，包括应对阿里巴巴双十一购物节期间的峰值负载压力。阿里巴巴发现，与既有 CPU 平台相比，这代平台的端到端性能提高了一个数量级。

图 3-1-1 列出了使用具备核心推荐模型主要特征的代理模型时，第四代至强® 可扩展处理器和第三代至强® 可扩展处理器的代际性能对比。在 AMX、BF16 混合精度、8 通道 DDR5、更大高速缓存、更多内核、高效的内核到内核通信和软件优化的配合下，主流的 48 核第四代至强® 可扩展处理器可以将代理模型的吞吐量提高近 3 倍，超过主流的 32 核第三代至强® 可扩展处理器，同时将时延严格保持在 15 毫秒以下。[57]

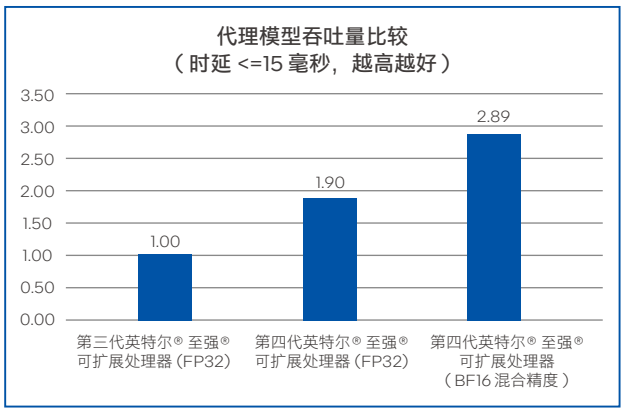


图 3-1-1 代理模型的代际性能比较（时延 <=15 毫秒）[58]

收益

- 阿里巴巴能够在保证推荐模型符合推理时延 <=15 毫秒的同时，将推理的吞吐量提升达 2.89 倍。[59] 同时在将模型量化到 BF16 之后，AI 推理精度依然能够满足需求；
- 升级为第四代至强® 可扩展处理器带来的性能收益远高于硬件成本，有助于阿里巴巴降低 TCO，获得更高的投资收益；
- 基于 CPU 的推理方案具备媲美高端 GPGPU 的性能表现，同时在成本、灵活性等方面具备更强的优势。

展望

阿里巴巴与英特尔联合验证了，利用第四代至强® 可扩展处理器内置的英特尔® AMX 等创新硬件特性，并进行软件优化之后，核心推荐模型在性能上能够获得巨大提升。未来，阿里巴巴还将与英特尔围绕数据中心基础设施架构优化、技术创新等领域进行深度合作，加速第四代英特尔® 至强® 可扩展处理器等新一代硬件的优化实践，加快 AI 等应用的运行，向阿里巴巴海量客户提供更高效的服务，助力以数据为中心的商业变革。

[57、58、59] 如欲了解更多性能测试详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/alibaba-e-comm-recommendation-system-enhancement.html

基于第三代至强® 可扩展处理器的阿里云弹性计算服务助用友优化智能 OCR 性能，建立更敏捷、更经济的 AI 中台

“数智化创新应用的一大重心是实现智能化驱动。通过提供 AI 中台，我们能够支撑企业智慧大脑、赋能管理者商业创新和智慧管理，提升员工工作效率与用户体验。高效的 AI 中台有赖于强大的基础设施，得益于第三代英特尔® 至强® 可扩展处理器的应用与软件优化，我们能够满足智能 OCR 等应用对于性能的需求，为用户提供更高性能、更低成本、更高灵活性的智能化服务。”

方高林
用友智能中台总经理

传统 OCR 应用依赖模板匹配、特征提取等技术，对于图片质量、应用环境的要求较为苛刻，在图像清晰度不高、字体变化、模糊或背景干扰等复杂环境下，难以达到理想的识别准确率。在此背景下，AI 结合传统光学字符识别的智能 OCR 技术，凭借高速、准确、低成本等优势，成为 OCR 应用的新选择。与传统 OCR 技术相比，智能 OCR 技术不仅能够通过深度学习网络，可靠、快速地完成海量样本的训练，应对复杂的识别任务，而且还能够与自然语言处理、知识图谱等技术进行融合，实现更佳的识别效果。但同时，智能 OCR 在模型推理等方面也带来了较高的算力需求，虽然基于独立 GPU 的推理方案性能较强，但也给企业带来了部署成本高昂、无法充分利用现有 CPU 平台等挑战。

基于采用第三代至强® 可扩展处理器的阿里云弹性计算服务的用友智能 OCR 应用

用友智能 OCR 应用集成于用友商业创新平台 YonBIP 的 PaaS 云平台 (iuap) 中，用户能够在 YonBIP 的智能报销、智慧协同等应用便捷调用智能 OCR 能力。为了充分利用基于 CPU 的 AI 推理方案在成本、灵活性等方面的优势，用友将基于第三代至强® 可扩展处理器的云服务器作为智能 OCR 模型推理的基础平台，并在阿里云弹性计算服务中进行了性能验证，证明能够满足实际场景对于 OCR 推理的性能需求。

■ 采用第三代至强® 可扩展处理器的阿里云弹性计算服务加速智能 OCR 推理

为了充分释放第三代至强® 可扩展处理器的 AI 推理潜能，加速智能 OCR 应用，用友在阿里云第七代 ECS 云服务器上进行了优化与验证。阿里云第七代 ECS 云服务器依托第三代神龙架构，提供稳定可预期的超高性能，同时通过芯片快速路径加速手段，完成存储、网络性能以及计算稳定性的数量级提升。本次验证使用的阿里云实例采用了英特尔® 至强® 铂金 8369B 处理器，规格为 16C32G 单实例。

用友采用如下步骤进行性能优化：

- 首先，将模型从原框架迁移到 OpenVINO™ 框架，从而将 OCR 模型推理时延降低 67%[60]。OpenVINO™ 工具套件是用于快速开发应用程序和解决方案，以解决各种任务（包括人类视觉模拟、自动语音识别、自然语言处理和推荐系统等）的综合工具套件。该工具套件基于最新一代的人工神经网络，包括卷积神经网络（CNN）、递归网络和基于注意力的网络，可跨英特尔® 硬件扩展计算机视觉和非视觉工作负载，从而大幅提高性能；

- 其次，用友针对 OCR 模型的实际特征与应用环境，对于 OpenVINO™ 的参数、整体调用流程进行优化，从而在上一步的基础上，将 OCR 模型推理时延降低了 69%[61]；

- 最后，用友采用第三代至强® 可扩展处理器集成的 VNNI 指令集，对整个推理过程进行加速，进一步将推理时延降低了 34%[62]。

通过如上三个步骤的优化，OCR 模型最终的推理时延降低了 93.1%[63]，可满足实际应用对于 OCR 的性能要求。

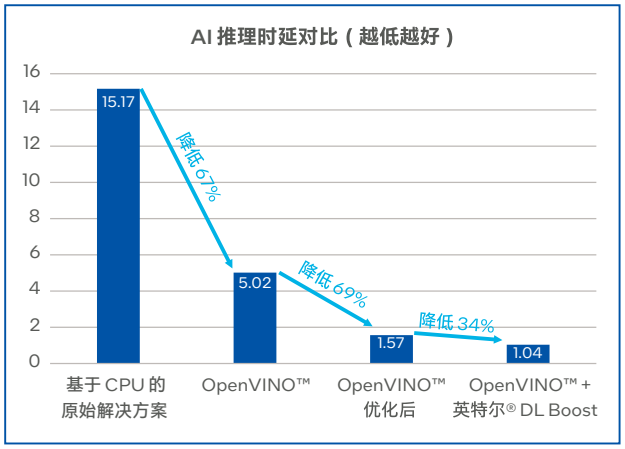


图 3-2-1 基于第三代至强® 可扩展处理器的用友 OCR 方案优化前后的性能比较[64]

收益：基于 CPU 的方案可满足特定 OCR 场景对于 SLA 的需求

在基于第三代至强® 可扩展处理器的云服务器上，用友验证了在综合利用 OpenVINO™ 工具套件以及处理器内置硬件特性进行优化后，OCR 推理模型性能实现了有效提升，能够保证端到端的 OCR 应用达到特定的 SLA 目标。例如，基于该 OCR 方案的发票识别端到端应用响应速度低于 2 秒，满足 3 秒之内的 SLA 指标要求。[65]

对于客户而言，这一组合方案能够带来如下价值：

- 通过性能优化，显著提升了基于 CPU 的 AI 推理性能，无需使用专门的基于 GPU 的硬件来进行推理，不仅能够降低硬件的采购成本，相应的空间、功耗、软硬件调优等成本也得到显著降低，有助于提升 OCR 应用的投资回报率（ROI）；

- 该方案能够有效利用现有的 CPU 服务器资源，用户无需额外采购 / 租用 GPU 服务器，提升了基础设施的灵活性，也避免了 GPU 服务器部署、维护等带来的资源损耗；

- 用友智能 OCR 服务能够在 BIP 平台上，以软件即服务（SaaS）的方式快速提供给客户，实现快速部署以及定制化支持；

- 该应用实践为用友 YonBIP 用户的云实例 / 硬件选型提供参考，用户可以根据实际的性能需求，选择更适用的云实例。

[60、61、62、63、64] 测试结果援引自用友于 2022 年 8 月开展的测试。测试配置：阿里云实例，英特尔® 至强® 铂金 8369B 处理器，16 核，32 线程。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

[65] 如欲了解更多性能测试详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/yonyou-smart-ocr-performance-optimization.html

金蝶云基于阿里云平台和英特尔软硬件，构建更高效 PaaS 服务

“Tair 持久内存型实现了数据的持久化存储，不仅满足了金蝶云苍穹平台各项业务对高速缓存的需求，还降低了内存数据库的构建与交付成本，实现了性能、成本的均衡，这让我们更有信心应对各种业务挑战。我们也希望通过与英特尔、阿里云等伙伴的合作，打造更加卓越的 PaaS 平台，助力用户实现数字化价值创造与产业共生。”

金蝶云

目前，部署 Redis 等高性能的内存数据库已成为 PaaS 平台数据缓存组件优化的重要方向。通过使用高性能的内存读写数据，内存数据库能够大幅提升读写速度，使得 PaaS 平台的各项服务都倾向于使用内存数据库进行高速缓存。然而，这也导致内存数据库需要即时处理的数据量快速增长，系统压力不断攀升，内存数据库在高性能、经济性与数据可靠性等方面都需要进一步进行优化，解决 TCO 快速上升、难以支持在非易失性场景中的应用、性能出现瓶颈等问题。

解决方案：基于英特尔® 架构的金蝶云苍穹平台

为了化解 PaaS 平台数据缓存组件在经济性、持久性等方面的挑战，金蝶与阿里云、英特尔合作，使用了基于傲腾™ 持久内存与第三代英特尔® 至强® 可扩展处理器的阿里云云原生内存数据库 Tair，作为 PaaS 平台的缓存组件，并在阿里云平台上进行了验证。

■ 金蝶云苍穹平台

金蝶云苍穹采用领先的云原生技术和中台架构，是企业级 IT 应用服务平台。它不同于互联网企业的 PaaS 平台，并不负责技术层面的资源管理，而是提供了诸多应用基础服务与组件，实现对含有业务模型的业务应用的开发、部署、升级等全生命周期服务与管理。

■ 阿里云云原生内存数据库 Tair 持久内存型

云原生内存数据库 Tair 是基于 Tair 产品研发的云上托管键值对缓存服务，是 Tair 持续创新过程中的一个里程碑。Tair 持久内存型基于傲腾™ 持久内存，提供大容量、兼容 Redis 的内存数据库产品，单实例成本对比云数据库 Redis 社区版最高可降低 30%[66]，且数据持久化不依赖传统磁盘，保证每个操作持久化的同时提供近乎 Redis 社区版的吞吐和时延，可显著提升业务数据可靠性。

■ 经过测试验证的性能与性价比表现

为了验证该方案在性能、性价比等方面的表现，金蝶启动了适配与压力测试。该测试基于阿里云通用型实例规格族 g7 进行。根据 3、5、8 时间分布原则，Tair 持久内存型 95% 以上的请求处理时间在 5 秒以下，性能达到设计目标。Tair 持久内存型与 Redis 社区版的对比性能测试结果显示，Tair 持久内存型的性能要稍高于 Redis 社区版[67]。

在成本方面，由于傲腾™ 持久内存的同容量价格低于 DRAM 内存，因此在同等规格的缓存服务按照包年或包月的付费方式计算时，Tair 持久内存型最高可节省 28% 的缓存服务成本[68]。

[66] 如欲了解更多详情请访问：https://help.aliyun.com/document_detail/183956.html
[67] 如欲了解更多详情请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/kingdee-cloud-cangqiong-platform-on-alibaba-cloud.html
[68] 数据由阿里云基于英特尔® 傲腾™ 持久内存与常规内存的购买价格计算得出。

在 K8S 节点负载、缓存负载、数据库负载监控等测试中，Tair 持久内存型的 CPU 占用率均未达到瓶颈，且显著低于 Redis 社区版，这一方面验证了第三代至强® 可扩展处理器的强劲性能，另一方面也证明 Tair 持久内存型具备更好的资源利用能力。

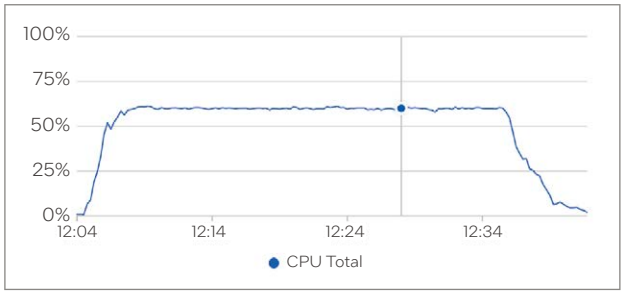


图 3-3-1 Tair 持久内存型 K8S 节点负载监控——CPU 占用率 [69]

在缓存负载监控中，Redis 社区版与 Tair 持久内存型两者的网络带宽使用率最大值分别为 98.92%、99.3%，已经达到网络带宽的性能瓶颈。[70]

总体来看 [71]，K8s 节点、Pod、数据库服务 PostgreSQL 等负载均正常。ESSD PL1 磁盘吞吐量突发峰值可以达到 950MB/s，官方承诺最大吞吐量为 350MB/s，可以满足 PostgreSQL 的大量 WAL 写需求。

■ 提供强大的 AI 服务扩展能力

在该方案的基础上，金蝶云苍穹还通过 AI 服务云提供了多种 AI 服务。以视觉识别服务为例，该服务基于 OCR、NLP 等技术，提供了开箱即用的预置文字识别、自定义模板文字识别、文档差异分析等服务能力，帮助企业以低代码方式在业务单据中嵌入文字识别、文档智能分析等能力，实现智能化、便捷化的业务场景。

为了提升视觉识别模型的推理性能，金蝶云苍穹在基于第三代至强® 可扩展处理器的阿里云实例上，采用了 OpenVINO™ 工具套件进行性能优化。

为了验证此方案的性能表现，金蝶在阿里云 ecs.g7.8xlarge 上进行了测试。该实例搭载英特尔® 至强® 铂金 8369B 处理器，提供了 32 个 vCPU、128GB 内存、16 Gbit/s 的网络基础带宽。在该实例上，金蝶首先验证了通过 OpenVINO™ 工具套件将模型从 FP32 量化为 IN8 之后，两者的性能对比。测试结果如图 3-3-2 所示，量化为 INT8 之后，性能提升约 58% [72]。

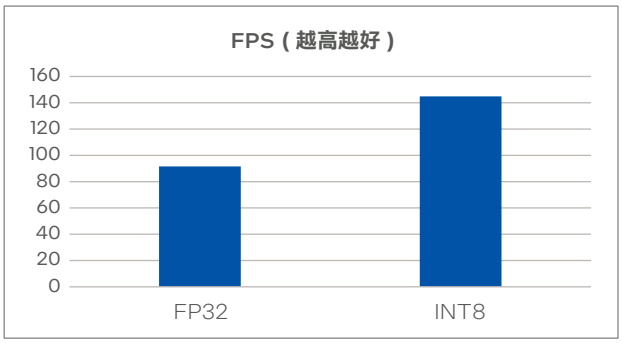


图 3-3-2 使用 CPU 推理时，FP32/INT8 性能差异 (input_size=32x544) [73]

金蝶还对比了直接使用 CPU 进行推理 (ovms,INT8) 与使用某主流 GPU 进行推理 (tensorflow/serving, FP32) 的性能。测试数据如图 3-3-3 所示。数据显示，在通过 OpenVINO™ 工具套件进行 INT8 量化之后，性能相比 GPU 有着较大优势 [74]。

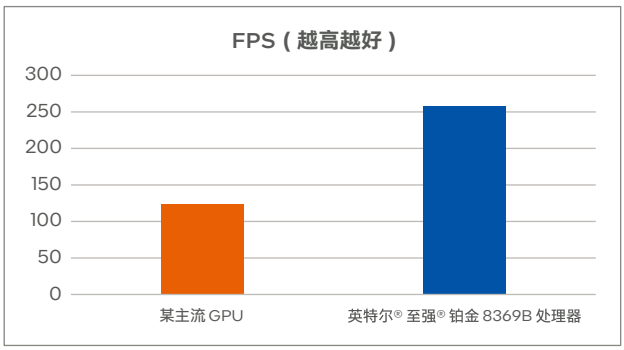


图 3-3-3 金蝶视觉识别服务 AI 推理性能对比 [75]

总结与展望

基于 Tair 持久内存型的金蝶云苍穹平台证明了英特尔® 傲腾™ 持久内存和第三代英特尔® 至强® 可扩展处理器等软硬件的应用潜力。金蝶、阿里云和英特尔三方还将围绕方案调优、场景化落地等方向，进一步推动在 PaaS 平台方面的合作，未来还将下沉到金蝶 SaaS 服务的所有场景。

[69, 70, 71, 72, 73, 74, 75] 如欲了解更多详情请访问: https://www.intel.cn/content/www/cn/zh/cloud-computing/kingdee-cloud-cangqiong-platform-on-alibaba-cloud.html

融合英特尔® oneAPI 工具套件，阿里云弹性高性能计算助深势科技加速 LAMMPS 工作负载

> “得益于英特尔® oneAPI 工具套件的支持，深势科技成功地在阿里云弹性高性能计算上对数以百万计的原子开展了分子动力学模拟。”
>
> **张林峰**
> 联合创始人兼首席科学家
> 深势科技

> “弹性高性能计算为个人用户、教育和科研机构以及公共机关提供了一个快捷、灵活且更加安全的云计算平台。结合英特尔® oneAPI 工具套件，弹性高性能计算可以帮助客户在英特尔® 至强® 可扩展处理器上构建一个高效的分析计算平台。”
>
> **陶锦中**
> 高性能计算解决方案产品经理
> 阿里云

LAMMPS 是广泛运用的开源分子动力学模拟软件，其在处理效率、并行计算以及灵活性等方面的优势，正在深势科技的工作中起到关键作用。但分子模拟建模时，其内在固有的复杂性和瞬息万变的动态变化，使 LAMMPS 工作负载对算力有着十分苛刻的需求，进而使深势科技在选择计算平台和优化方案时面临巨大挑战。

解决方案

为了向各行业的科研、生产提供快捷、弹性、安全的科学计算服务，具有公有云、专有云等多种产品形态的阿里云弹性高性能计算正在为用户提供高性能、跨架构的一站式 HPC/HPDA 云原生平台服务。

得益于阿里云弹性高性能计算提供的丰富实例阵列，深势科技不仅能为 LAMMPS 工作负载选择更适合的实例，同时其配置的第三代英特尔® 至强® 可扩展处理器也可为用户提供所需的强劲算力。与此同时，深势科技在方案中也引入了英特尔® oneAPI 工具套件来推动 LAMMPS 工作负载在阿里云弹性高性能计算上的进一步性能优化。

利用工具套件中的英特尔® oneAPI DPC++/C++ 编译器（Intel® oneAPI DPC++/C++ Compiler）、英特尔® MPI 库（Intel® MPI Library）以及英特尔® VTune™ Profiler 等，深势科技对代码进行重新编译，并通过热点分析等方法进行深层次优化，使 LAMMPS 工作负载性能获得显著提升。

■ 采用英特尔® oneAPI DPC++/C++ 编译器、英特尔® MPI 库，提高运算效率

英特尔® oneAPI DPC++/C++ 编译器是英特尔® oneAPI 工具套件中的核心组件之一，其提供了对 DPC++ 编程语言（一种基于 C++，专门用于并行计算和加速器程序的编程语言）的强大支持。而英特尔® MPI 库是一个面向开源 MPICH 规范的多功能消息传递库，其可帮助用户在基于英特尔® 架构的科学计算硬件平台（或兼容平台）上创建、维护和测试各类复杂的应用程序，带来性能上的增益。

深势科技利用英特尔® oneAPI DPC++/C++ 编译器和英特尔® MPI 库来加快 LAMMPS 的运算速度，优化工作主要通过如下 2 个步骤实现：

1. 将默认 GCC 编译器替换为英特尔® oneAPI DPC++/C++ 编译器，并使用开源的 MPICH 库，这使 LAMMPS 的执行时间从 16 分 22 秒缩短至 14 分 48 秒，性能提高约 9.6%[76]；

[76] 如欲了解更多性能详情请访问：https://www.intel.com/content/www/us/en/developer/articles/technical/alibaba-cloud-ehpc-dp-intel-winning-cloud-solution.html

2. 将开源 MPICH 库替换为英特尔® MPI 库，同时保留英特尔® oneAPI DPC++/C++ 编译器，这令 LAMMPS 的执行时间从 14 分 48 秒进一步缩短到 13 分 43 秒，性能提高约 7.3%[77]。

如表 4 所示，上述两个步骤可将 LAMMPS 工作负载的执行时间从 16 分 22 秒缩短到优化后的 13 分 43 秒，性能整体提升约 16.2%[78]。

编译器 + MPI 库	执行时间（hh：mm：ss）
使用 GCC 编译器和 MPICH 库编译 LAMMPS	0:16:22
使用英特尔® oneAPI DPC++/C++ 编译器和 MPICH 库编译 LAMMPS	0:14:48
使用英特尔® oneAPI DPC++/C++ 编译器和英特尔® MPI 库编译 LAMMPS	0:13:43

表 4 两个步骤缩短 LAMMPS 运行时间[79]

■ 利用英特尔® VTune™ Profiler 工具，分析和消除性能瓶颈

作为英特尔® oneAPI 基础套件的一部分，高级性能分析工具英特尔® VTune™ Profiler 可用于优化基于英特尔® 架构的处理器、GPU 和 FPGA 等之上的应用性能、系统性能和系统配置。

如图 3-4-1 所示，在优化过程中，深势科技借助英特尔® VTune™ Profiler 对 LAMMPS 的计算进程进行了检测，并发现最耗时的热点是 PMPI_Wait，占用了约 27 秒的处理器执行时间。根据这一结果，深势科技对 PMPI_Wait 实施微调后，处理器执行时间可减少至约 11 秒。[80]

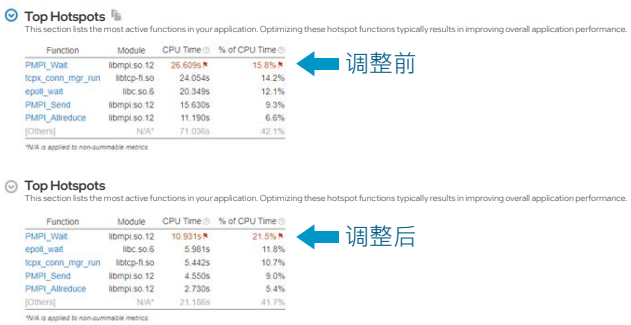


图 3-4-1 LAMMPS 计算进程中的热点瓶颈调整前后[81]

在上述优化基础上，深势科技继续通过微调进程和线程的组合，例如将当前工作负载的线程数增加到 2，并将进程数减少到 16，可进一步减少 LAMMPS 工作负载的执行时间。经过几轮组合微调，最高可将工作负载执行时间从 13 分 43 秒减少到 8 分 58 秒，性能提升约 34.6%[82]。

进程数	线程数	执行时间（hh：mm：ss）	平均处理器利用率
32	1	0:13:43	85.5%
16	2	0:08:58	87.8%
8	4	0:10:00	90.6%

表 5 通过微调进程和线程的组合优化 LAMMPS 运行[83]

融合上述英特尔® oneAPI DPC++/C++ 编译器、英特尔® MPI 库和英特尔® VTune™ Profiler 的优化后，如图 3-4-2 所示，LAMMPS 工作负载的整体执行时间从初始的 16 分 22 秒减少至 8 分 58 秒，性能提升达 45.2%，优化效果非常显著[84]。

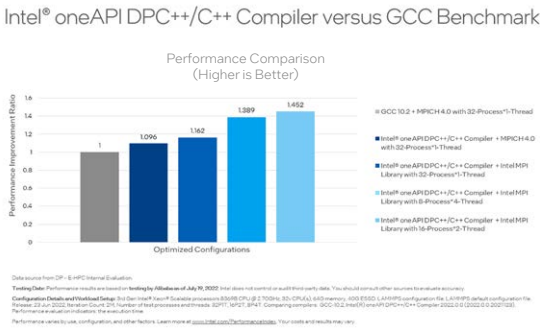
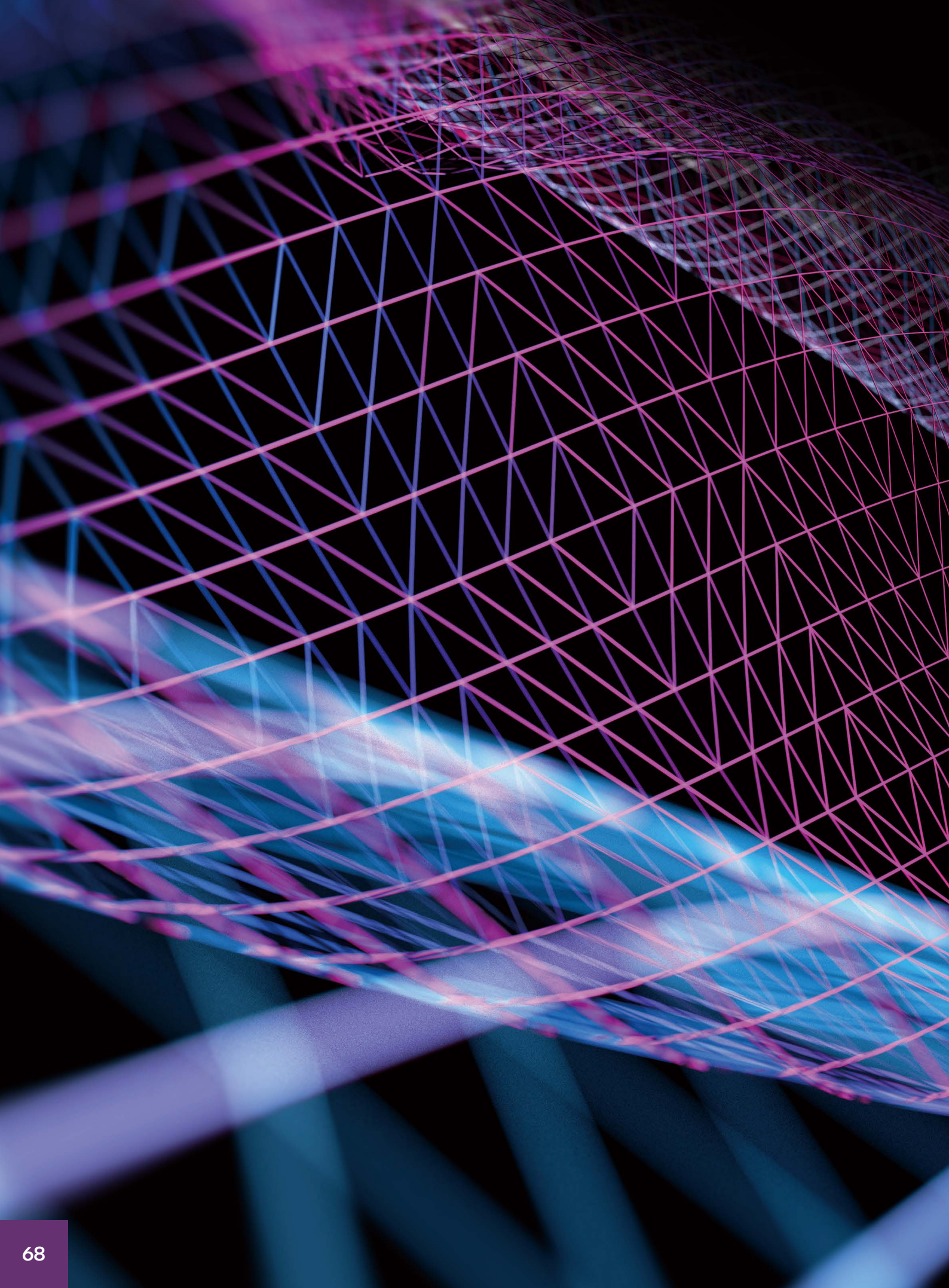


图 3-4-2 组合优化提升 LAMMPS 性能[85]

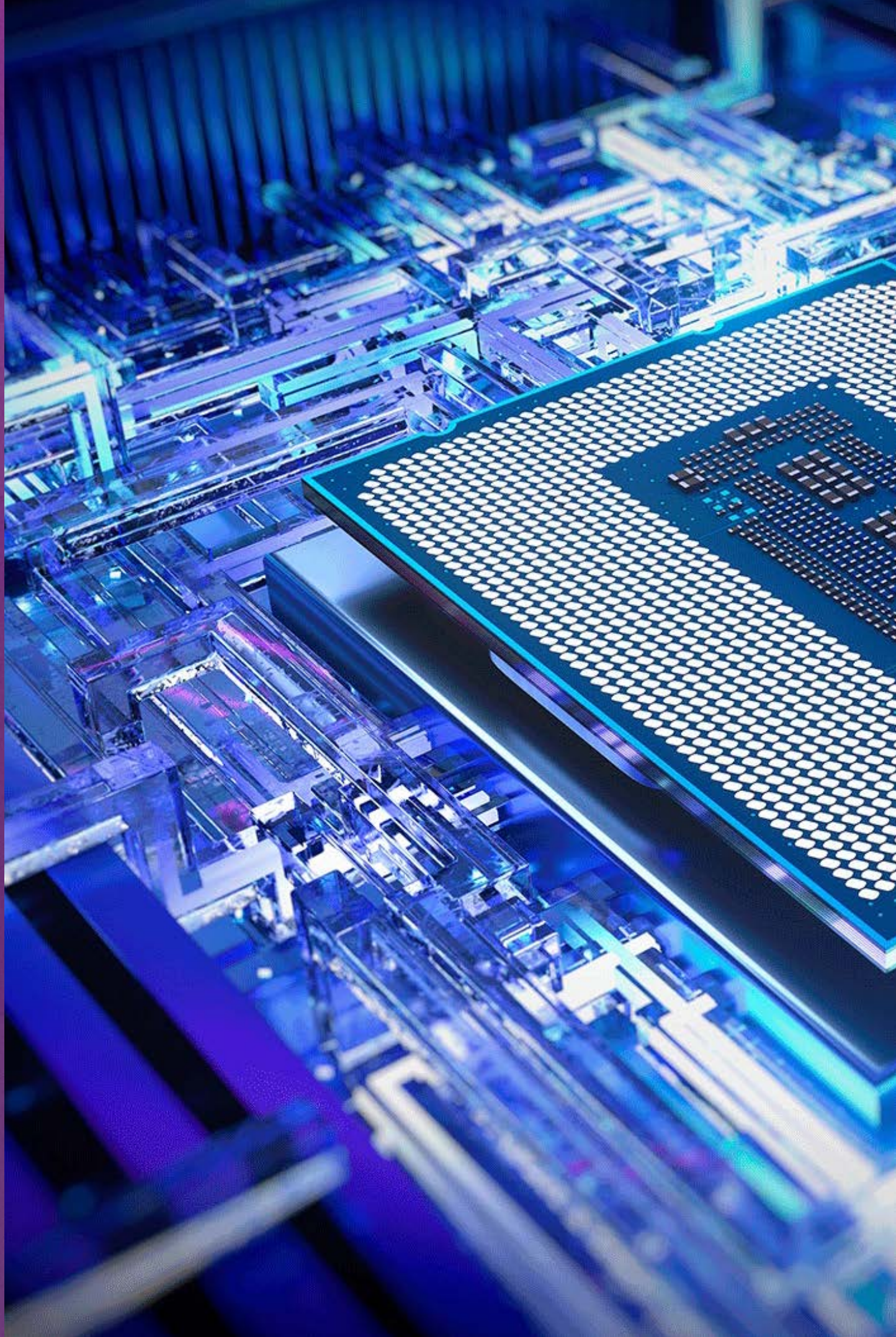
总结与展望

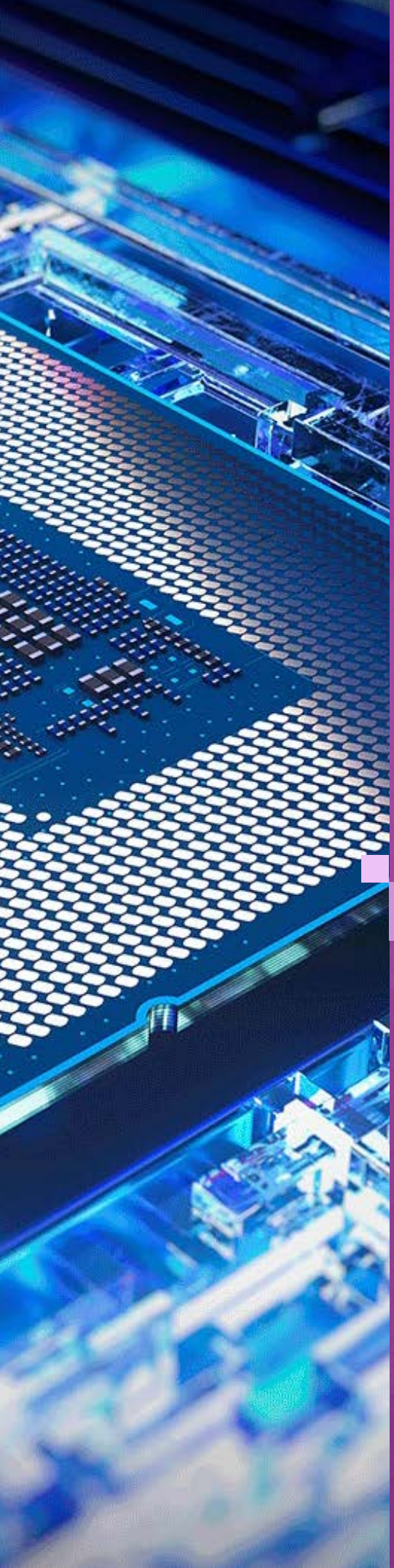
基于阿里云弹性高性能计算平台，在英特尔® oneAPI 工具套件等的助力下，深势科技成功地实现了 LAMMPS 应用性能的大幅提升。在实践中，阿里云弹性高性能计算用户获得了云原生、全栈、高性能的计算 PaaS 服务，该服务快速、灵活且更加安全，并支持互操作性。与此同时，英特尔® oneAPI 工具套件也为包括科学计算在内的、日益复杂的应用负载在基于英特尔® 架构的平台上实现更好地性能提升提供了更佳途径。这一成功实践，为未来科学计算技术进一步推动社会经济发展提供了良好标杆。

[77、78、79、80、81、82、83、84、85] 如欲了解更多性能详情请访问：https://www.intel.com/content/www/us/en/developer/articles/technical/alibaba-cloud-ehpc-dp-intel-winning-cloud-solution.html



产品篇





以数据为中心的硬件产品组合

第四代英特尔® 至强® 可扩展处理器

第四代英特尔® 至强® 可扩展处理器旨在为人工智能、数据分析、存储和科学计算等方面快速增长的工作负载提供性能加速。该处理器基础性能进一步大幅提升，具有很强的灵活性，且具备多种内置加速器。同时利用先进的安全技术，即使面对敏感或受监管的数据，也能解锁新的商业合作机会和洞察。使用这款处理器可跨多个云和边缘环境进行扩展，满足自身的部署需求。

- 全新内置加速器
- 80 条 PCIe 5.0 通道
- 支持 1 至 8 路配置
- 更高的单核性能
  每路多达 60 个内核
- 英特尔® UPI 2.0
  （高达 16 GT/s）

- 增加三级缓存
  （LLC）共享容量
- Compute Express
  Link (CXL) 1.1
- 8 通道 DDR5
  传输速率高达 4,800 MT/s（1DPC）
  传输速率高达 4,400 MT/s（2DPC）
  每路 16 个 DIMM
  全新 RAS 功能（增强型 ECC、ECS）
- 高带宽内存（HBM）
  （64GB / 每路）
- 经优化的电源模式

第四代英特尔® 至强® 可扩展处理器的新特性或新功能

PCI Express Gen5 (PCIe 5.0)

带来全新的 I/O 速度，可在 CPU 和互联设备之间实现更高的吞吐量。第四代至强® 可扩展处理器具有多达 80 条 PCIe 5.0 通道，非常适合高速网络、高带宽加速器和高性能存储设备。PCIe 5.0 的 I/O 带宽是 PCIe 4.0 的两倍[86]，仍具备向后兼容性并提供用于 CXL 连接的基础插槽。

DDR5

以更高内存带宽克服数据瓶颈，提高计算性能。与 DDR4 相比，DDR5 的带宽提高多达 1.5 倍[87]，因此有机会提升性能、容量和能效并降低成本。借助 DDR5，第四代至强® 可扩展处理器提供的速率可高达 4,800 MT/s (1 DPC) 或 4,400 MT/s (2 DPC)。

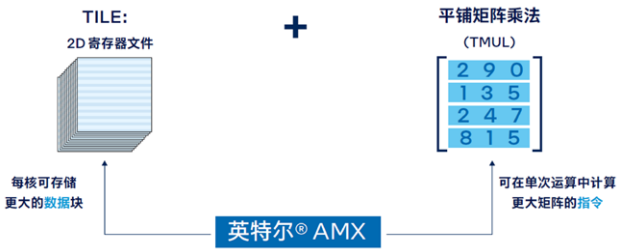
CXL

借助面向下一代工作负载的 CXL 1.1，降低数据中心的计算时延并帮助减少 TCO。CXL 是另一种跨标准 PCIe 物理层运行的协议，可以在同一链路上同时支持标准 PCIe 设备和 CXL 设备。CXL 可带来的一大关键能力是在 CPU 和加速器之间创建统一且一致的内存空间，它将革新未来数年数据中心服务器架构的构建方式。

[86、87] https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors-product-brief.html

英特尔® 高级矩阵扩展（英特尔® AMX）

英特尔® AMX 是内置于第四代英特尔® 至强® 可扩展处理器的加速器，可优化深度学习 (DL) 训练和推理工作负载。借助英特尔® AMX，第四代英特尔® 至强® 可扩展处理器可在优化通用计算和 AI 工作负载间快速转换。开发人员可编写非 AI 功能代码来利用处理器的指令集架构 (ISA)，也可编写 AI 功能代码，以充分发挥英特尔® AMX 指令集的优势。



英特尔® AMX 架构由两部分组件构成：

- 第一部分为 TILE，由 8 个 1KB 大小的 2D 寄存器组成，可存储大数据块；
- 第二部分为平铺矩阵乘法 (TMUL)，它是与 TILE 连接的加速引擎，可执行用于 AI 的矩阵乘法计算。

功能

- 提供广泛的软硬件优化，提升 AI 加速能力
- 同时支持 INT8 和 BF16 数据类型

商业价值

- 为 AI/ 深度学习推理和训练工作负载带来显著性能提升
- 通过硬件加速使常见应用更快交付

软件支持

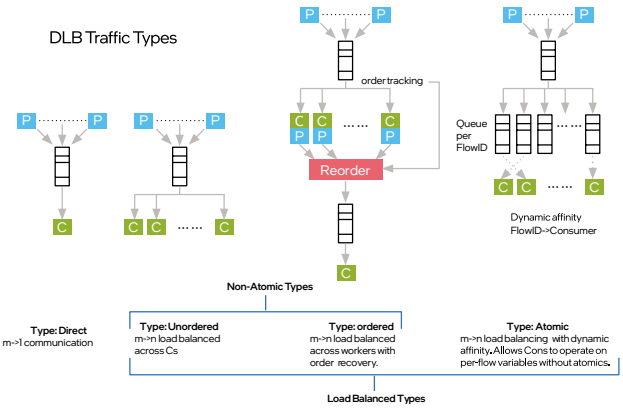
- 市场上的主流框架、工具套件和库 (PyTorch、TensorFlow)、英特尔® oneAPI 深度神经网络库 (英特尔® oneDNN)

用例

- 图像识别、推荐系统、机器 / 语言翻译、自然语言处理 (NLP)、媒体处理和分发

英特尔® 动态负载均衡器（英特尔® DLB）

英特尔® DLB 是一个硬件队列管理器和负载平衡器，开发人员能通过它获得硬件辅助队列，帮助实现每秒数百万个传入请求的负载均衡。在多核英特尔® 至强® 可扩展处理器上处理网络数据时，英特尔® DLB 有助于提高系统性能，它实现了在多个 CPU 内核 / 线程上高效地分配网络处理，并根据系统负载的变化而动态地在多个 CPU 内核上分配网络数据以进行处理。同时，英特尔® DLB 能够还原在多个 CPU 内核上同时处理网络数据包的顺序。



英特尔® DLB 的四种队列模型

功能

- 当网卡 (NIC) 静态负载分配机制引发负载不均衡时，在内核间实现数据负载的动态再分配

商业价值

- 提升系统在多核英特尔® 至强® 可扩展处理器上处理网络数据的性能
- 提升分布式处理、动态负载均衡和动态调整网络处理顺序的性能

软件支持

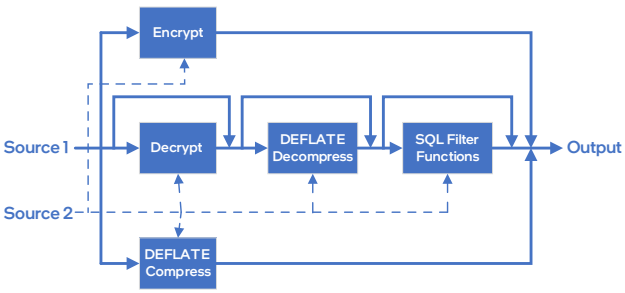
- 英特尔® Data Mover Library

用例

- IPSec 安全网关、VPP 路由器、用户平面功能 (UPF)、vSwitch、流数据处理、大象流处理

英特尔® 存内分析加速器（英特尔® IAA）

英特尔® IAA 专为提升数据库和数据分析性能而设计。它可以提高内存数据库和高级分析工作负载的查询吞吐量并降低其占用的内存空间，进而加速数据传输；可以减少对 CPU 内核的依赖，从而提高 CPU 内核利用率。其适用于内存数据库、开源数据库和数据存储（如 RocksDB、Redis、Cassandra 和 MySQL）。与在没有加速功能的 CPU 内核上使用软件进行压缩相比，借助英特尔® IAA，客户在运行开源的 RocksDB 数据库引擎时可以获得更高的数据解压缩吞吐量。



https://www.intel.com/content/www/us/en/content-details/721858/intel-in-memory-analytics-accelerator-architecture-specification.html

功能

- 加速数据分析原语的内置加速器 IP、循环冗余校验（CRC）计算、压缩和解压缩

商业价值

- 提高内存数据库和数据分析工作负载的查询吞吐量
- 减少数据分析工作负载所占用的内存和带宽，释放更多 CPU 空间

软件支持

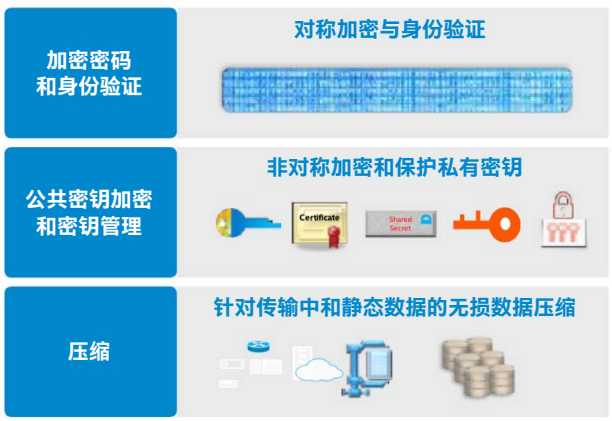
- 英特尔® Query Processing Library 和英特尔® Data Mover Library

用例

- 商业内存数据库、开源内存数据库（RocksDB、Redis、Cassandra、MySQL、MongoDB）和用于大数据分析的列式格式

英特尔® 数据保护与压缩加速技术（英特尔® QAT）

英特尔® QAT 可提升性能，从而满足当今网络工作负载的需求，使系统能够服务更多客户端。它可以大大提高密码操作（包括对称和非对称加解密）工作负载的速度。与在没有加速功能的 CPU 内核上运行软件相比，使用 RSA4K 的英特尔® QAT 可以提高开源的 NGINX Web 服务器上的客户端密度。英特尔® QAT 可加速 SQL Server 数据库备份。借助英特尔® QAT，SQL Server 客户可以提高备份操作速度，减少备份存储容量。同时，英特尔® QAT 可加速密码操作和数据压缩 / 解压缩，使存储工作负载和应用的性能得到提升。例如，与在没有加速功能的 CPU 内核上运行相同的压缩算法相比，将英特尔® QAT 作为卸载引擎可以大幅提高压缩吞吐量。



功能

- 加速密码操作和数据压缩 / 解压缩

商业价值

- 卸载并加速压缩 / 解压缩，使 CPU 使用效率得到提升
- 以更少的开销在设备之间实现更多加密连接和 Web 安全连接

软件支持

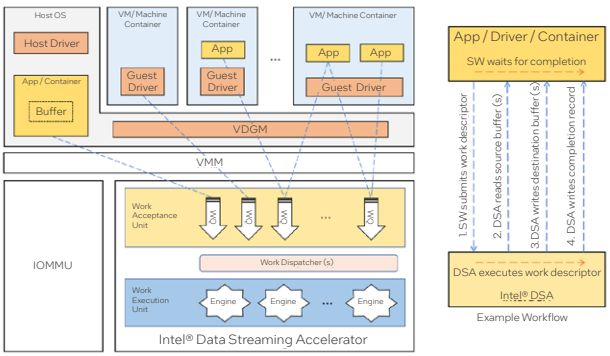
- 加速密码操作的英特尔® QAT 引擎

用例

- 分布式存储系统、文件系统、RocksDB、数据湖、Apache Spark、Hadoop、NGINX、IPSec

英特尔® 数据流加速器（英特尔® DSA）

英特尔® DSA 是新一代直接内存访问（DMA）引擎。它通过加速数据传输和转换操作（例如数据完整性校验和去重）大幅提升吞吐量。英特尔® DSA 在 CPU 上（内存、缓存和处理器内核之间）以及 CPU 之外（附加内存、存储和网络资源）都能发挥作用。这种对性能的提升使 I/O、数据传输和数据包处理更高效。



https://www.intel.com/content/www/us/en/developer/articles/technical/scalable-io-between-accelerators-host-processors.html

功能

- 优化流数据传输和转换操作

商业价值

- 提高面向 NVMe/TCP 的数据保护力度，通过卸载基于 CPU 的工作负载提升数据存储应用的效率

软件支持

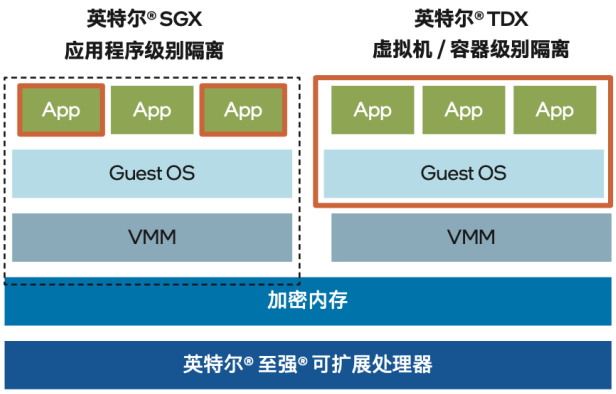
- 英特尔® Data Mover Library

用例

- 虚拟化、非透明桥之间的快速复制、ERP、内存数据库

英特尔® 安全引擎

英特尔® 至强® 可扩展处理器配备多个英特尔® 安全引擎，为各种数据（包括敏感、保密和处于监管之下的数据）保驾护航，使其可用于分析，进而帮助企业加速创新步伐，在维持出色性能的同时，帮助保护数据机密性与代码完整性。



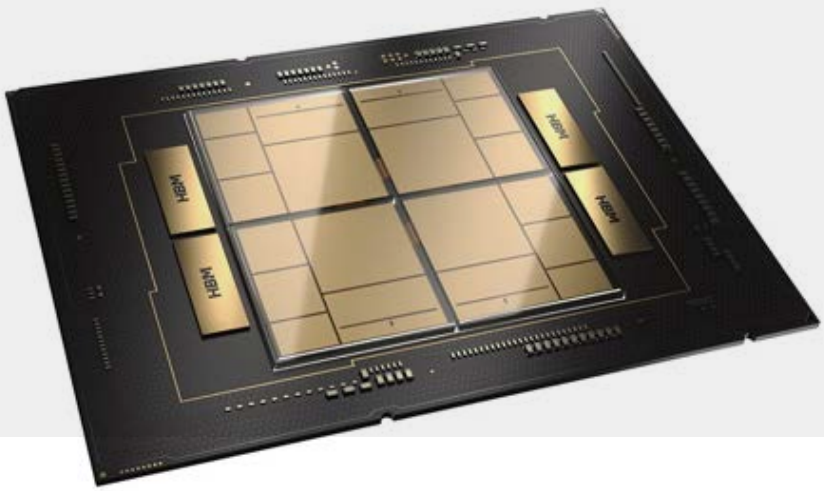
英特尔® SGX 经过广泛部署和研究，是数据中心可信执行环境（TEE）的重要技术实现，能够大幅减少系统内的攻击面。英特尔® SGX 提供基于硬件的安全解决方案，可通过专用应用隔离技术帮助保护使用中的数据。开发人员可以通过保护选定的代码和数据不被查看或修改，在“飞地”内执行涉及敏感数据的操作，帮助提高应用的安全性和保护数据的机密性。

英特尔® TDX 将进一步提升保护级别。这一全新工具于 2023 年开始通过特选云服务提供商为企业在虚拟机（VM）层面提供隔离边界和机密保障。英特尔® TDX 可将客户机操作系统和虚拟机应用都与云端主机、系统管理程序和平台的其他虚拟机隔离开来。虽然英特尔® TDX 的信任边界比英特尔® SGX 应用层面的隔离边界大，但英特尔® TDX 能使机密虚拟机比应用安全“飞地”更易于进行大规模部署和管理。

英特尔® 至强® CPU Max 系列

英特尔® 至强® CPU Max 系列采用全新微架构，支持一系列可提升平台能力的特性，包括更多内核、先进的 I/O 与内存子系统，以及可加速重大发现的内置加速器。英特尔® 至强® CPU Max 系列具有以下特性：

- **多达 56 个 P-core（性能核）：**内核由 4 个小芯片构成，采用英特尔的嵌入式多芯片互连桥接（EMIB）技术连接，功耗为 350W；
- **64GB 高带宽封装内存及 PCIe 5.0 和 CXL 1.1 I/O。**英特尔® 至强® CPU Max 系列每核均具备 HBM 容量，可满足大多数常见科学计算工作负载的要求；
- 与其他 CPU 相比，在使用 Numenta 的 AI 技术进行自然语言处理时，其 HBM 优势可带来高达 20 倍的性能提升[88]。



“仅 HBM”模式

该模式支持内存容量需求不超过 64GB 的工作负载以及每核 1 至 2GB 的内存扩展能力，同时无需更改代码和另购 DDR，即可启动系统。

“HBM Flat”模式

该模式可为需要大内存容量的应用提供灵活性，它通过 HBM 和 DRAM 提供一个平面内存区域（flat memory region），适用于每核内存需求大于 2GB 的工作负载。使用该模式时可能需要更改代码。

“HBM 缓存”模式

旨在提升内存容量需求大于 64GB 或每核内存需求大于 2GB 的工作负载的性能。使用该模式时，无需更改代码，且 HBM 可缓存来自 DDR 的事务。

[88] https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon/xeon-max-series-product-brief.html

英特尔® 数据中心 GPU Flex 系列

英特尔® 数据中心 GPU Flex 系列是面向智能视觉云的灵活、强大且开放的 GPU 解决方案，可为视觉云工作负载提供出色的计算密度和能效。该系列产品基于英特尔® Xᵉ HPG（高性能显卡）微架构打造，内置视觉处理和 AI 加速技术。其提供的功能和优势包括：

- **支持开放、灵活、基于标准的软件堆栈以及 oneAPI 统一编程**，其中包括用于构建高性能、跨架构媒体应用和解决方案的开源组件与库、工具及框架。这种开放的方法有助于生态系统摆脱使用专有编程模型带来的技术和经济负担；
- **开创性地在 GPU 内配置了基于硬件的开源 AV1 编码器**，在相同质量下将带宽提高 30%，从而每年每十万名观众节省 2,300 万美元，或者在相同带宽下提高流媒体质量。[89]

该系列将以两种 SKU 形式提供：**英特尔® 数据中心 GPU Flex 系列 170（峰值性能更高）和英特尔® 数据中心 GPU Flex 系列 140（密度更高）**。

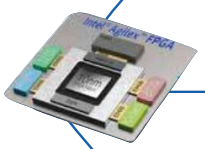
	英特尔® 数据中心 GPU Flex 140	英特尔® 数据中心 GPU Flex 170
目标工作负载	媒体处理和交付、基于 Windows 和 Android 的云游戏、虚拟桌面基础设施、AI 视觉推理[2]	
显卡外形规格	半高、半长、单宽、被动散热	全高、四分之三长、单宽、被动散热
显卡 TDP	75 瓦	150 瓦
每卡 GPU 数量	2	1
GPU 微架构	Xᵉ HPG	
Xᵉ 内核数量	16 个（8 个/GPU）	32
Fixed Function Media	4（2 个/GPU）	2
光线追踪	是	
峰值算力（脉动阵列浮点运算）	8 TFLOPS (FP32)/105 TOPS (INT8)	16 TFLOPS (FP32)/250 TOPS (INT8)
内存类型	GDDR6	
内存容量	12 GB（6 GB/GPU）	16 GB
虚拟化（实例）	SR-IOV（62 个）	SR-IOV（31 个）
操作系统	Linux（Ubuntu、CentOS、Debian）、Windows Server 2019/2022、Windows Client 10、Red Hat® Enterprise Linux	
主机总线	PCIe Gen 4	
主机 CPU 支持	第三代 / 第四代英特尔® 至强® 可扩展处理器	

[89] https://www.intel.cn/content/www/cn/zh/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html

英特尔® FPGA 和 SoC FPGA

英特尔® FPGA 提供各类可配置的嵌入式 SRAM、高速收发器、高速 I/O、逻辑模块和路由。嵌入式知识产权（IP）与出色的软件工具相结合，减少了 FPGA 开发时间、功耗和成本。在广泛的边缘和数据中心应用中实现实时人工智能。

英特尔® FPGA 产品家族



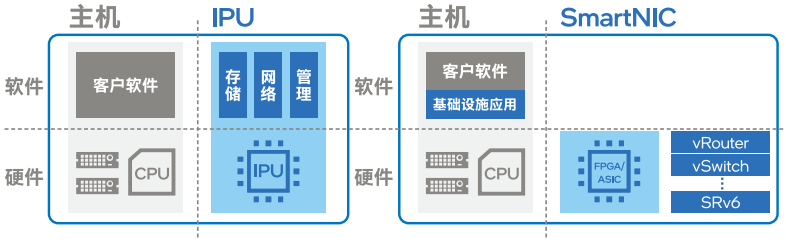
英特尔® Agilex™ FPGA

实现更出色的敏捷性和灵活性

- 高性能计算
  - 10 纳米制程技术
  - 超高速协议（内存一致性加速）
  - 超高带宽
- 先进三维封装
  - 任意点对点集成（任意工艺节点、任意供应商、任意 IP）
- 支持开发人员
  - 英特尔® Quartus® Prime（面向硬件开发人员的设计工具）
  - 面向软件开发人员的 OneAPI

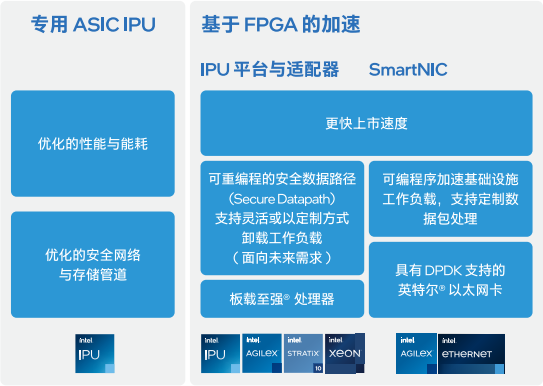
英特尔® 基础设施处理器（IPU）和 SmartNIC

英特尔® IPU 是具有强化的加速器和以太网连接的高级网络设备，它使用紧密耦合、专用的可编程内核加速和管理基础架构功能。IPU 提供全面的基础架构分载，并可作为运行基础架构应用的主机的控制点，从而提供一层额外防护。



英特尔 ® SmartNIC 是具有可编程加速器和以太网连接的可编程网络适配器卡，可以加速主机上运行的基础架构应用。

广泛的基础设施加速组合



英特尔® 以太网网络适配器

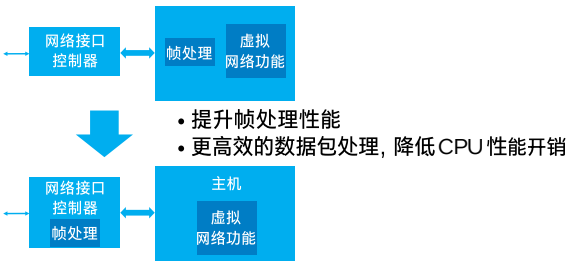
英特尔 ® 以太网产品发展路线图

应用设备队列

可为应用提供专用的网络队列，为关键流量创建专有 “快速通道”，提高应用响应时间的可预测性，降低时延并增加吞吐量。

动态设备个性化

动态设备个性化 (DDP) 技术旨在提高包处理效率，与数据平面开发工具套件 (DPDK) 结合使用时，可以减少时延，并提高云、通信和网络边缘工作负载的性能。带有 DDP 的英特尔® 以太网网络适配器 800 系列提供了重新配置数据包处理管道的功能，具备支持更广泛流量类型的能力。

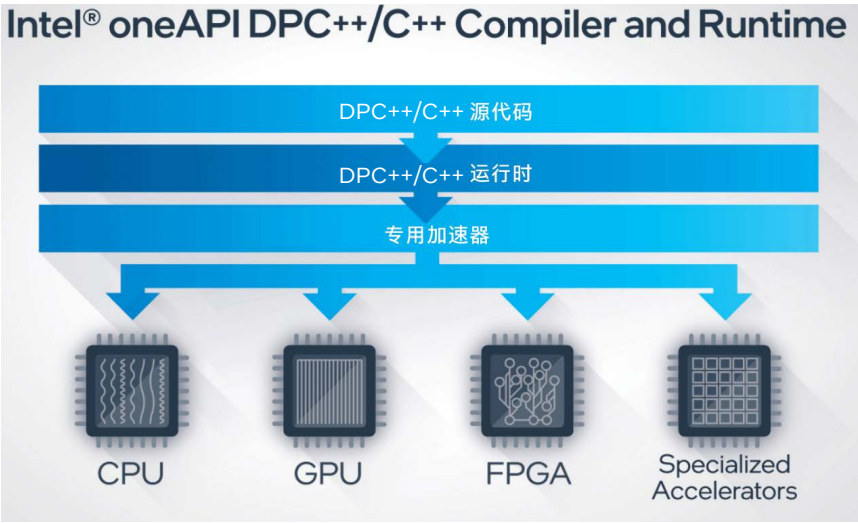


软件及系统级优化

基础设施算力优化

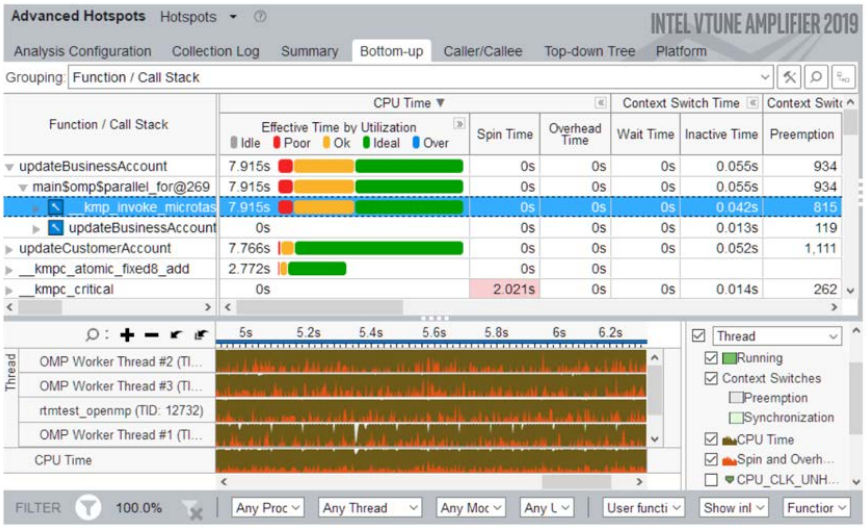
英特尔® oneAPI DPC++/C++ 编译器

英特尔® oneAPI DPC++/C++ 编译器提供了一个面向未来的编程模型，能够编译 ISO C++、Khronos SYCL 和 DPC++ 源代码，并可在包括 CPU、GPU 和 FPGA 的各种硬件上重用代码。英特尔® oneAPI DPC++/C++ 编译器可消除硬件锁定问题，提供了一个基于标准的开放、跨行业的统一编程模型。



英特尔® VTune™ Amplifier

英特尔® VTune™ 可视化性能分析（英特尔® VTune™ Amplifier）是一个通过图形用户界面，分析和优化程序性能的工具，且无需重新编译。其能够准确剖析 C、C++、Fortran、Python、Go、Java 或各种编码语言组合；提供各种数据来优化处理器、内存和存储；通过提供快速解答，采用多元化的分析将数据转化为洞察力，且缩短优化代码所需的时间。

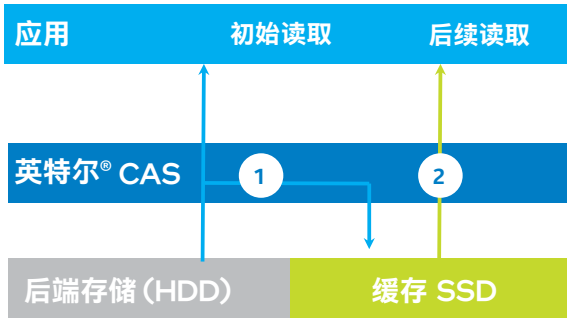


基础设施存储优化

英特尔® 高速缓存加速软件（英特尔® CAS）

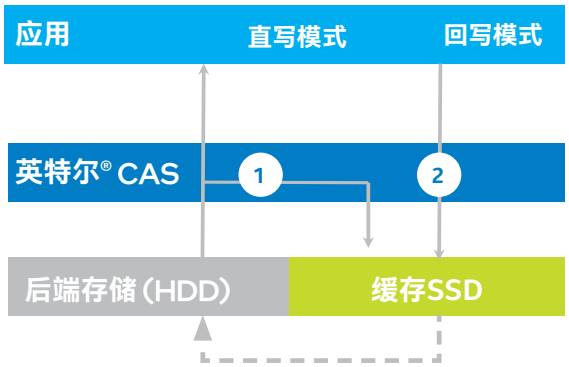
英特尔® 高速缓存加速软件作为一款服务器端缓存软件，可通过与内存进行互操作，以及与高性能固态盘（SSD）相结合，通过智能缓存管理，将最活跃的数据放入高性能固态盘介质，来提高应用程序性能，解决数据中心 I/O 性能瓶颈问题。

读取工作流程



- 数据从后端存储读取并复制到固态盘上的缓存内
- 后续读取以高性能固态盘速度返回

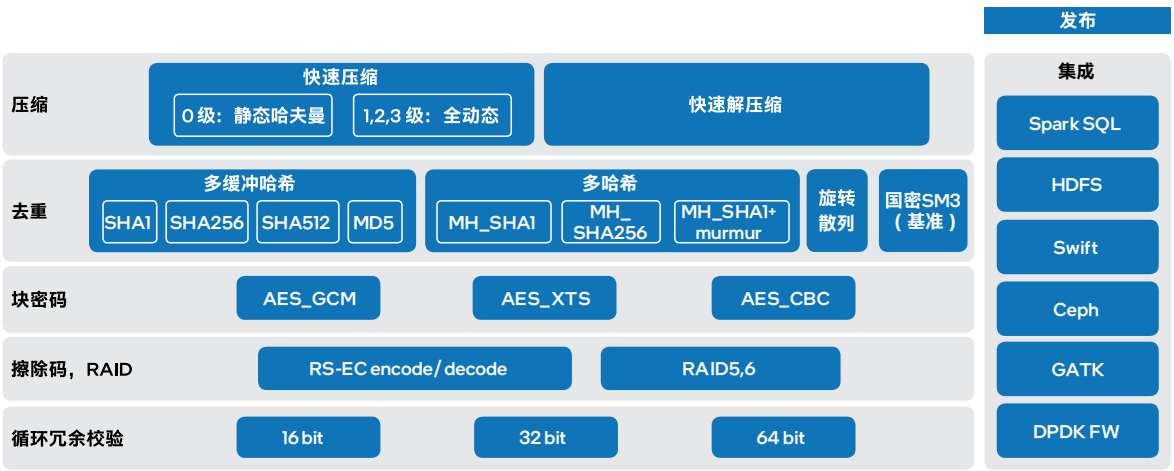
写入工作流程



- 所有数据同步写入后端存储和缓存
- 所有数据首先写入缓存，后续适时写入后端存储

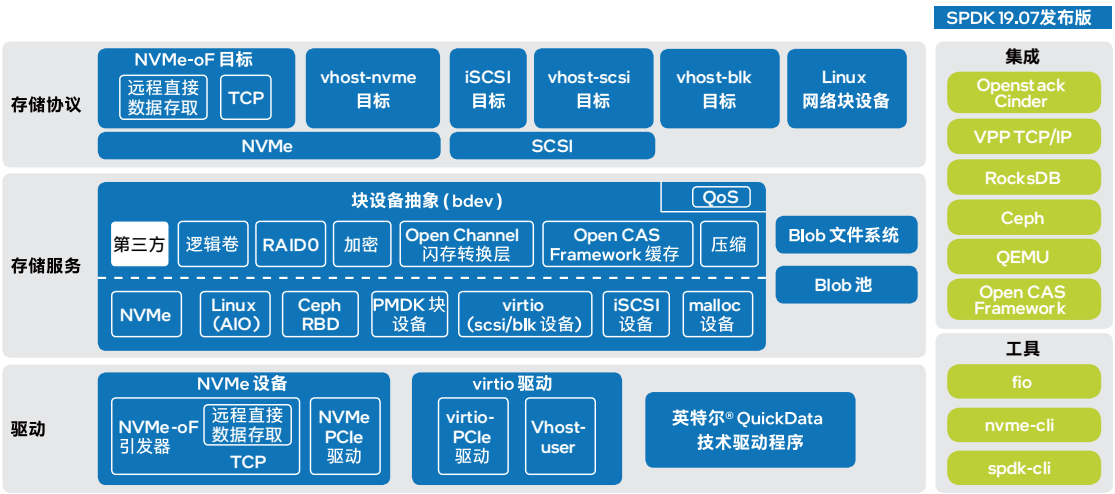
英特尔® 智能存储加速库（英特尔® ISA-L）

英特尔® 智能存储加速库基于英特尔® 架构，可为存储可恢复性、数据完整性、数据安全性提供优化，并加速数据的压缩。具体可以实现：RAID、Erasure Code 纠删码、CRC（cyclic redundancy check）、Multi-buffer Hashing（MbH）（包括 MD5、SHA1、SHA256 和 SHA512）、加密功能及压缩功能。



存储性能开发套件（SPDK）

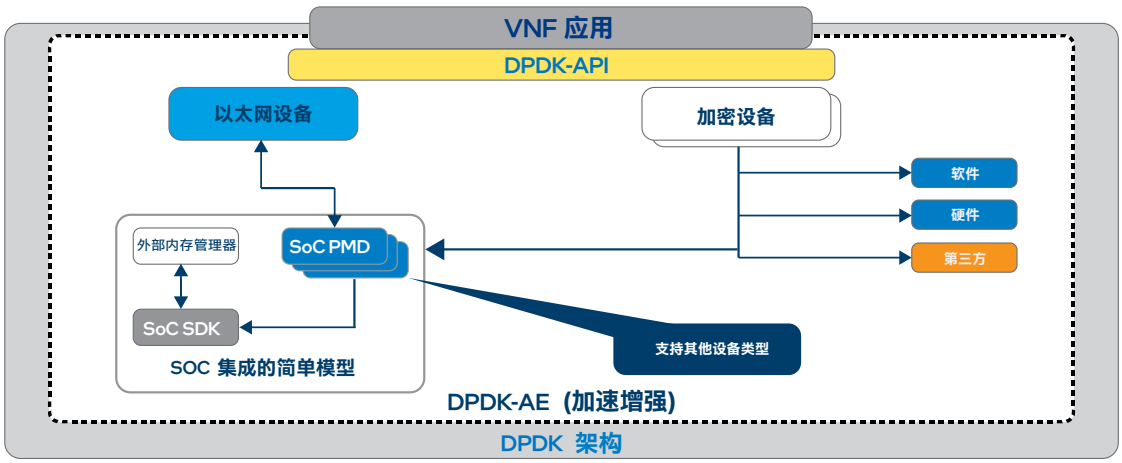
SPDK 是一套用于编写高性能、可扩展、用户模式存储应用程序的工具和库。它通过将所有必需的驱动程序移入用户空间，避免系统调用；提供完整的块堆栈作为驱动用户空间库，它执行许多与操作系统中的块堆栈相同的操作；提供基于这些组件的 NVMe、ISCSI 和 vHOST 服务器，这些组件能够通过网络或其他进程提供磁盘服务，来实现高性能。



基础设施网络优化

数据平面开发套件（DPDK）

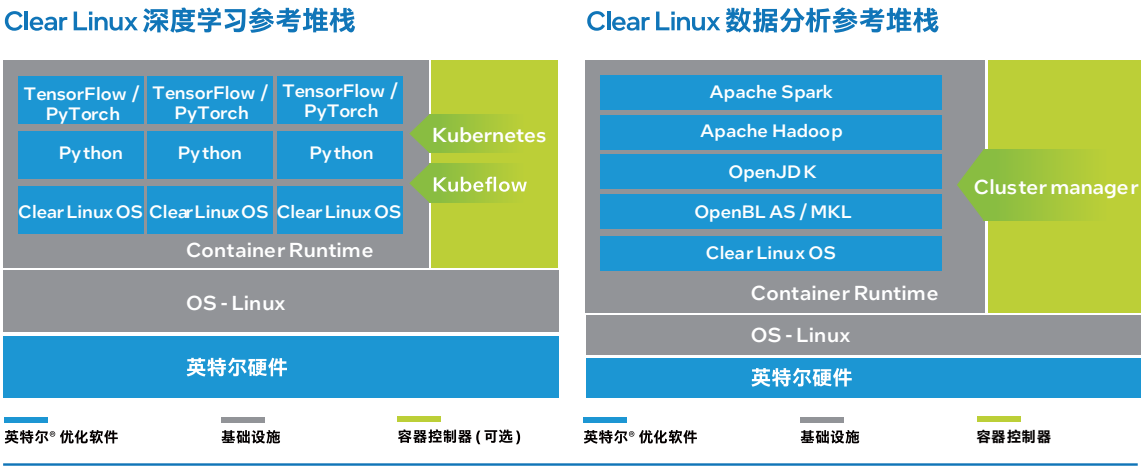
DPDK 是英特尔推出的一种高速网络数据包软件开发套件，现已开源。初期主要支持英特尔® 处理器及网卡系统，现已支持部分非英特尔® 架构处理器，以及部分非英特尔的网卡，能够通过旁路 Linux 系统网络协议栈，直接对网卡进行读写，结合多核处理器中不同核心的绑定，能够实现网络小包流量下的线速收发。DPDK 可以极大提高数据处理性能和吞吐量，为数据平面应用程序提供更多时间。



操作系统和编排层优化

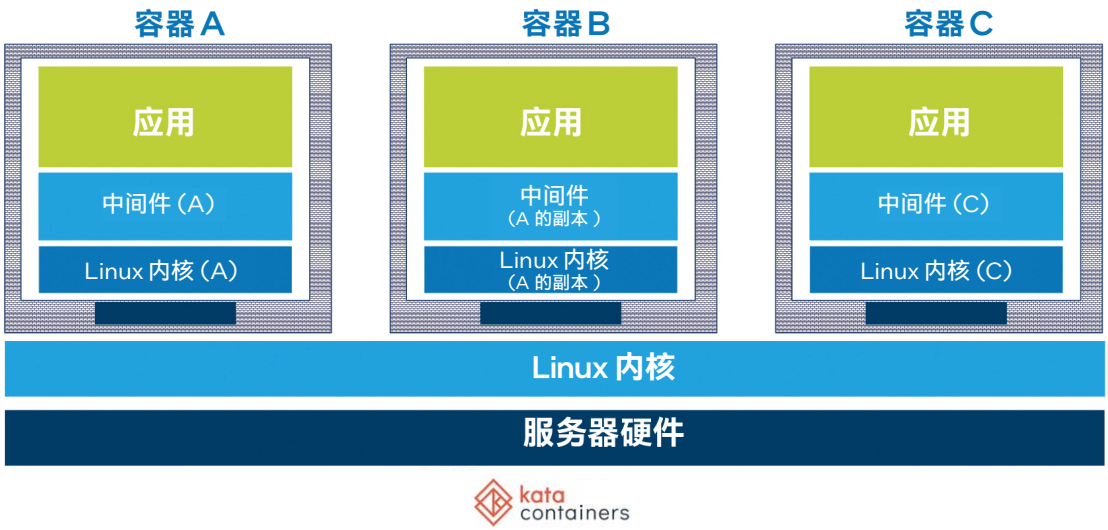
Clear Linux

Clear Linux 系统是英特尔开源的创新 Linux 发行版，它兼顾了从云到边缘计算的应用需求，既追求更优性能，又强化了安全性，还便于用户定制，且更易于管理。它采用滚动更新方式，在使其核心保持与上游 Linux 接近的同时，将所有针对英特尔® 架构平台的功能与优化整合进一整套 Linux 发行版之中。这些优化涉及到了 Linux 操作系统本身，以及与云计算和深度学习相关的功能和框架，其目的就是让它们能够更充分地利用英特尔® 架构平台带来的性能和功能优势。



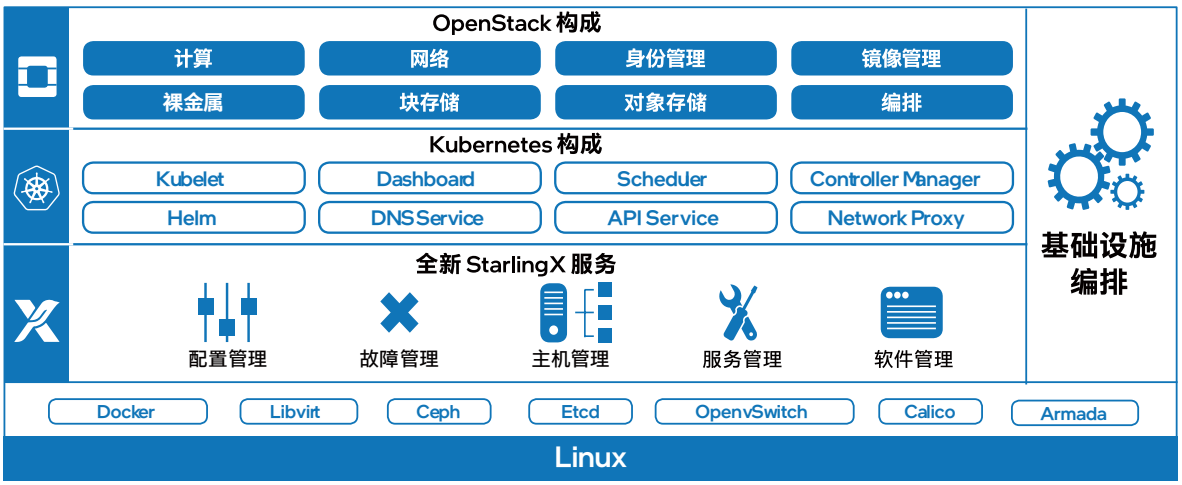
Kata Container

Kata Container 是一个创新的安全容器技术，它整合了英特尔的 Clear Containers 和 Hyper.sh 的 runV，在能够充分利用英特尔® 架构平台性能优势的同时，还支持其他架构的硬件。它还符合 OCI（Open Container Initiative）规范，可无缝地与 Docker 及 Kubernetes 对接。Kata Container 更核心的亮点就是采用轻量级虚拟化作为容器的隔离，使得它兼具容器的速度和虚拟机的安全隔离，这一点解决了长期以来困扰容器发展的安全隔离性不足问题，大大促进了云原生的发展。



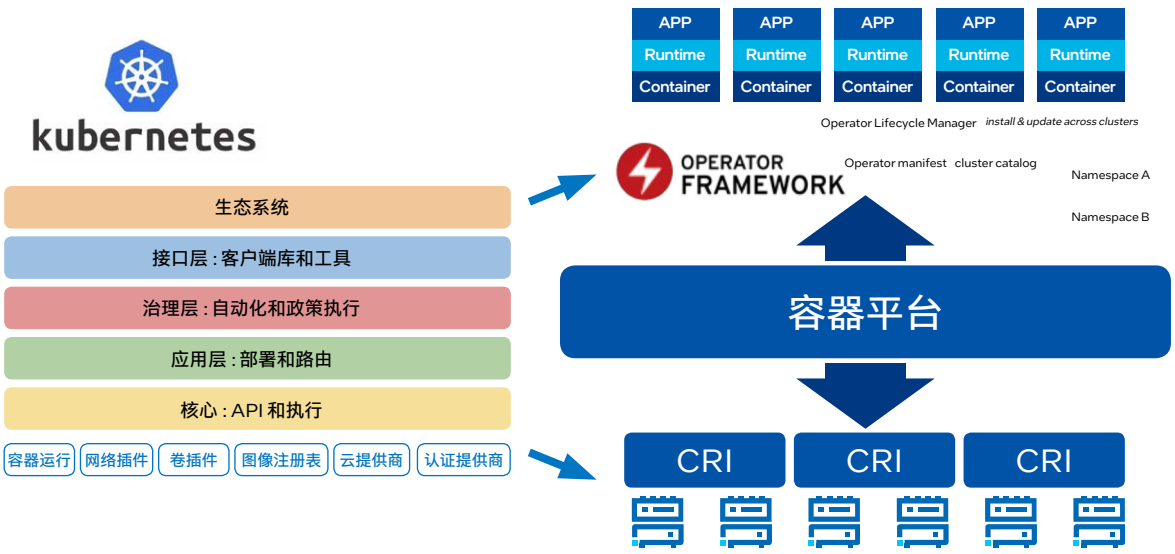
StarlingX

作为一个完整的边缘计算基础架构软件堆栈，StarlingX 不仅继承了 OpenStack 成熟完备的云服务管理能力，还与例如 Ceph、OVS、Kubernetes、DPDK 等众多优秀开源项目所提供的核心能力相结合，具备了从控制、计算到存储的全方位边缘云部署和管理能力。同时，其灵巧便捷的特性，也更适于在网络边缘进行部署。



Kubernetes

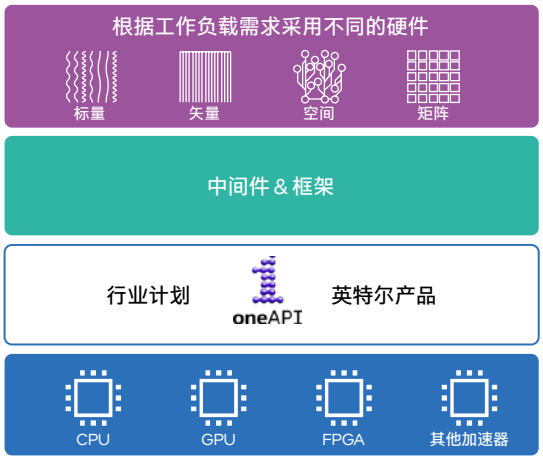
Kubernetes 是领先的容器编排解决方案。作为该项目的积极贡献者，英特尔使用多项数据中心关键技术，帮助其构建功能模块和全栈解决方案，比如：提供硬件设备插件、高级容器网络功能等技术，来解锁新的使用模式；逐层优化软件堆栈，确保最终用户获得底层硬件的全部优势；携手生态系统供应商合作，确保 Kubernetes 解决方案得以优化。



分析及 AI 性能优化

英特尔® oneAPI 工具套件

英特尔® oneAPI 工具套件是基于新一代标准的英特尔® 软件开发工具，用于跨各种架构构建和部署以数据为中心的高性能应用程序。它能够通过充分利用一流的硬件特性加速计算进程，并全面兼容现有的编程模型和代码库，可确保开发者已经编写的应用能够在 oneAPI 上无缝运行。此外，开发者只需一个代码库，便可以将应用轻松迁移到新系统和加速器上，大幅缩短了迁移时间，减轻了迁移工作量。



英特尔® 数据分析加速库（英特尔® DAAL）

英特尔® DAAL 提供 Linux、OS X 和 Windows 三种版本，可面向数据分析涉及的所有阶段（预处理、转换、分析、建模、验证和决策制定）提供高度优化的算法构建模块，以提升线下、流和分布式分析的效率。英特尔® DAAL 可为常见的数据平台，包括 Hadoop、Spark、R、Matlab 等提供良好支持，能从这些平台高效获取数据。此外，它还内置有数据管理功能，让应用可以直接访问各种来源（包括文件、内存缓冲、SQL、数据库、HDFS 等）的数据。

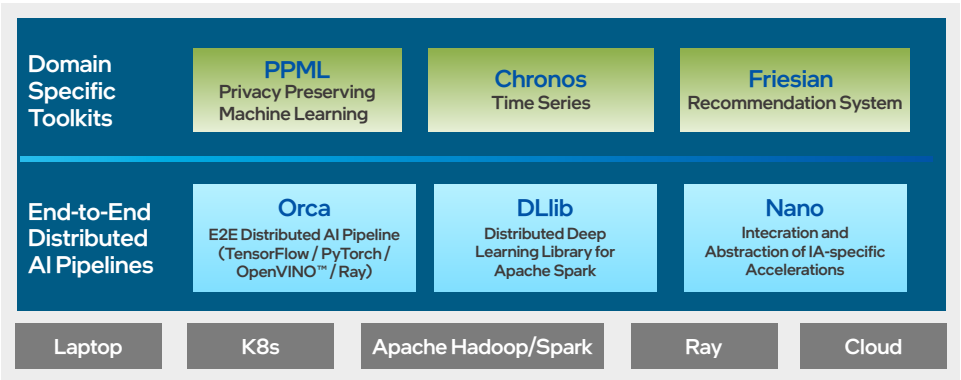
BigDL

BigDL 是英特尔开源的统一的大数据和人工智能平台，BigDL 可以将用户的数据分析或者 AI 应用无缝地从笔记本扩展到集群和云端。

BigDL 的特性包括：

- Orca：在 Spark 和 Ray 上构建分布式的大数据和 AI（PyTorch / TensorFlow）流水线
- Nano：在 XPU 上对 PyTorch / TensorFlow 应用进行透明加速
- Chronos：可扩展的自动时间序列数据分析应用
- Friesian：构建端到端推荐系统
- PPML：在 SGX/TDX 上构建更加安全的大数据和 AI 应用

此外，BigDL 还发布了大语言模型（LLM）的库，可以在英特尔平台上进行大语言模型的高效推理。



英特尔® MKL-DNN

英特尔® MKL-DNN 是专为在英特尔® 架构上加快深度学习框架运行速度而设计的一个性能增强库，它包含了高度矢量化和线程化的构建模块，支持利用 C 和 C++ 接口实施深度神经网络，拥有广泛的深度学习研究、开发和应用生态系统，适用于：Caffe、TensorFlow、PyTorch、Apache MXNet、BigDL、CNTK、OpenVINO™ 工具套件等丰富的深度学习软件产品。

分发详情

- 开源
- Apache 2.0 许可证
- 所有英特尔硬件中的常见 DNN API。
- 快速的发布周期，与 DL 社区迭代，为行业框架集成提供最佳支持。
- 基于流行的英特尔® MKL 函数库，经过高度矢量化和线程化，可实现最高性能。

github.com/01org/mkl-dnn

范例：

直接 2D 卷积 | 本地响应标准化（LRN） | 整流线性单元神经元激活（ReLU） | 最大池化 | 内积

加速深度学习模型的性能

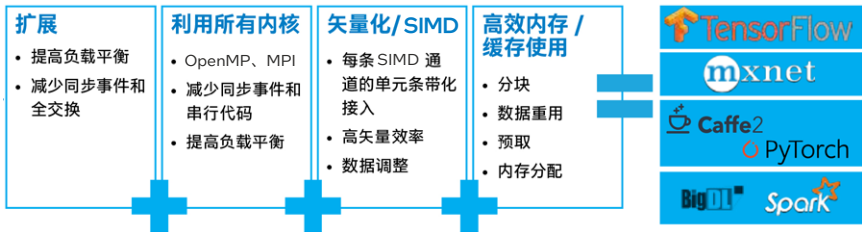
面向英特尔® 架构优化的深度学习框架

面向英特尔® 架构优化的 TensorFlow，通过计算图、内存池分配器与多个线程库等组件的优化， 能够确保深度学习工作负载在各种情况下都可利用英特尔® MKL-DNN 基本运算单元高效运行。

英特尔® Python 分发版提供了编写 Python 原生扩展所需的一切，如 C++ 和 Fortran 编译器、数学库和分析器，并且集成 NumPy、SciPy scikit-learn、pandas、Jupyter、matplotlib、mpi4py 等多个高性能数据分析和数学库，能够满足计算密集型应用需求。

面向英特尔® 架构优化的 Caffe，集成了英特尔® 数学核心函数库，专门面向高级矢量扩展指令集英特尔® AVX2 和英特尔® AVX-512 做了优化，且具备更多处理器优化功能，展现了更优的性能，并支持多节点分布程序训练。

英特尔开源的统一大数据和人工智能平台 BigDL 可以无缝、直接运行在现有的 Apache Spark 和 Hadoop 集群之上，是在处理器平台上实现大数据分析 +AI 应用的关键。BigDL 支持 PyTorch、TensorFlow、OpenVINO™ 等主流 AI 应用框架，可以将用户程序从笔记本无缝扩展到大数据集群上。



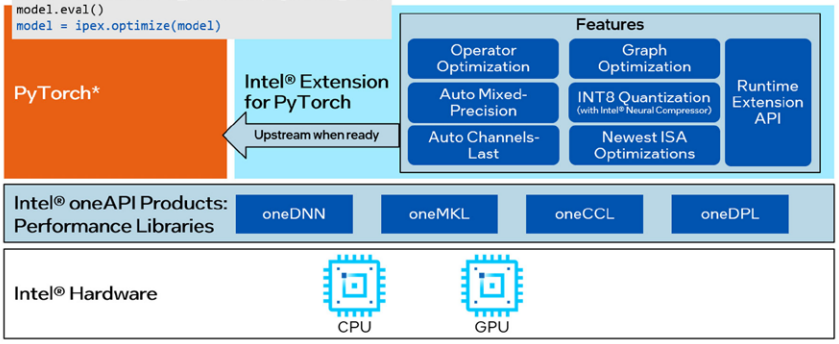
英特尔® Extension for PyTorch（IPEX）

为了提升 PyTorch 在英特尔硬件上的性能，英特尔推出了**英特尔® Extension for PyTorch**。该优化版利用了英特尔 CPU 上的英特尔® AVX-512 矢量神经网络指令（AVX-512_VNNI）、英特尔® 高级矩阵扩展（英特尔® AMX）以及英特尔独立 GPU 上的英特尔® X^e 矩阵扩展（英特尔® XMX）AI 引擎。此外，通过 PyTorch xpu 设备，英特尔® Extension for PyTorch 可以在英特尔 GPU 上为 PyTorch 提供轻松的 GPU 加速。

英特尔® Extension for PyTorch 提供了针对 eager 模式和 graph 模式的优化。在 eager 模式下，PyTorch 前端通过自定义 Python 模块（例如融合模块）、最优优化器和 INT8 量化 API 进行扩展。通过扩展图融合通道将 eager 模式模型转换为 graph 模式，可以进一步提升性能。在 graph 模式下，融合减少了运算符 / 内核调用开销，从而提高了性能。在 CPU 上，英特尔® Extension for PyTorch 根据 ISA（指令集架构）自动将运算符分派到其底层内核中，ISA 检测并利用英特尔硬件上可用的矢量化和矩阵加速单元，自动混合 float32 和 bfloat16 之间的运算符数据类型精度，以减少计算工作量和模型大小。英特尔® Extension for PyTorch 运行时扩展通过更细粒度的线程运行时控制和权重共享带来更高的效率。在 GPU 上，优化的算子和内核通过 PyTorch 调度机制实现和注册，这些运算符和内核通过英特尔 GPU 硬件的矢量化和矩阵计算功能进行加速。

```
import torch
from transformers import BertModel
import intel_extension_for_pytorch as ipex

model = BertModel.from_pretrained(args.model_name)
model.eval()
model = ipex.optimize(model)
```



https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-for-pytorch.html

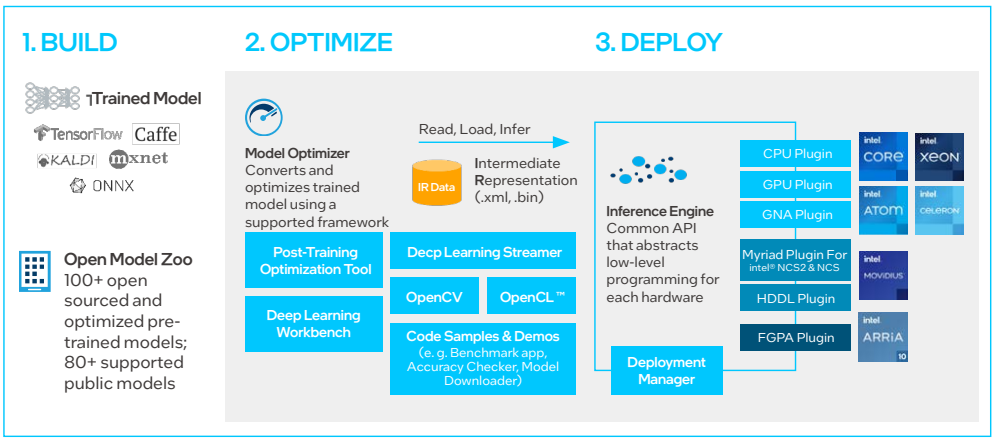
OpenVINO™ 工具套件

OpenVINO™ **工具套件**是一款加速深度学习推理及部署的软件工具套件，用以加快高性能计算机视觉处理和应用。该工具允许异构执行，支持 Windows 与 Linux 系统、Python 和 C++ 语言，提供预先转换的 Caffe、TensorFlow、MxNet 模型的 MO 文件与超过 20 个预先训练的模型，可帮助快速实现个性化的深度学习应用。通过使用 OpenCV、OpenVX 的基础库，它还便于创建特定的算法，实现定制化和创新型应用的开发。

高性能、深度学习推理部署

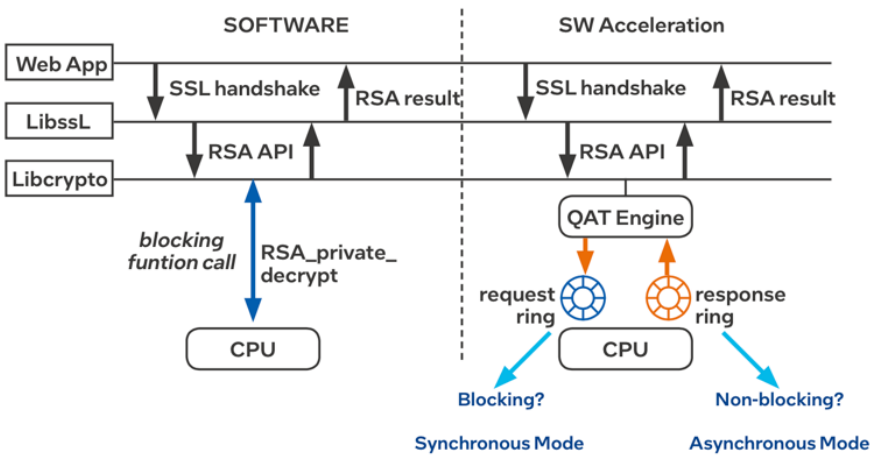
简化开发、易于使用

一次编写、随处部署



英特尔® Crypto-NI

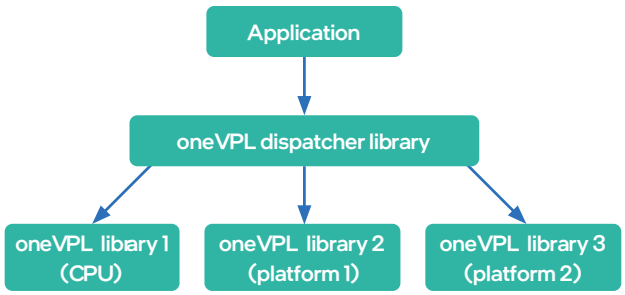
英特尔® Crypto-NI 是英特尔® 至强® 可扩展处理器中关于加解密领域的指令集，在之前英特尔® 至强® 可扩展处理器已具备的英特尔® AES-NI 指令集群上，又加入了 Vectorized AES、Integer Fused Multiply Add 等新指令。该方案使用的主要软件为 IPP Cryptography Library、Intel® Multi-Buffer Crypto for IPsec Library 和 QAT Engine。这些库基于新指令集提供了批量提交多个 SSL 请求的功能和并行异步处理机制，从而大幅提升性能。



媒体服务应用优化

英特尔® oneVPL

英特尔® oneVPL（英特尔® oneAPI Video Processing Library）是继英特尔® Media SDK 推出的下一代视频处理软件，其为视频编解码及其它通用视频处理提供了统一的、以视频为中心的 API 接口，并支持跨各种硬件加速器工作，可帮助用户在更多硬件加速器和更广泛的应用场景中获得性能提升和编程灵活性，非常适用于视频广播、直播流媒体、视频点播、云游戏和远程桌面解决方案等场景。

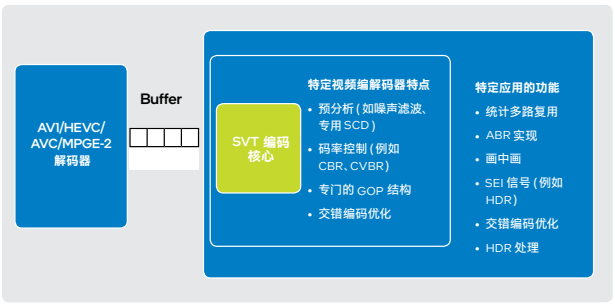


英特尔® oneVPL 的调度机制

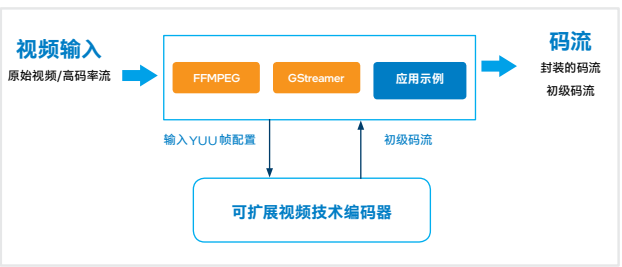
- 提供了与英特尔® Media SDK 核心 API 的兼容性
- 具备与英特尔® Media SDK相同的视频编解码器和滤波器
- 支持在通用处理器、集成显卡 GPU、独立显卡 GPU 以及其他硬件加速器中的部署
- 改进了视频处理初始化模式，可用于支持更广泛的视频处理实现方式
- 提供了新的内存抽象和优化方式，以及对解码性能的优化

英特尔® SVT

可扩展视频技术（SVT）是英特尔基于软件的视频编码架构，可使编码器在英特尔® 至强® 可扩展处理器上实现性能、时延和视觉质量之间的更佳平衡，且允许编码器根据质量和时延来调整应用程序的性能目标。英特尔® SVT 编码器具有多档性能和质量的预设值，能够满足各种质量需求下的视频云应用程序，包括视频点播（VOD）、广播、流媒体、监视、云图形和视频会议等。

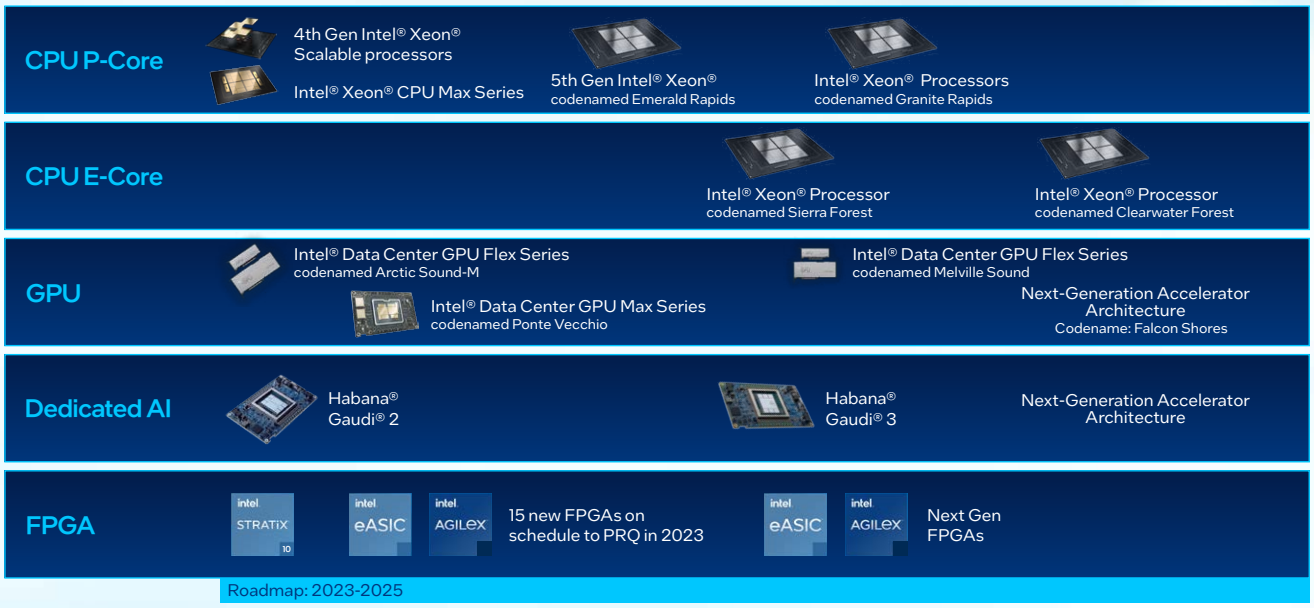


SVT 编码器核心和可视化编码器组件



SVT 编码器和示例应用程序之间的接口

英特尔数据中心与 AI 产品架构演进



英特尔® 至强® 演进路线图

关注英特尔数据中心微信公众号、商用小助手，
随时了解最新活动与资讯

扫码查看英特尔官网，
了解更多英特尔在云中的技术实践

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。

没有任何产品或组件是绝对安全的。

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。

预测或模拟结果使用英特尔内部分析或架构模拟或建模，该等结果仅供您参考。系统硬件、软件或配置中的任何差异将可能影响您的实际性能。

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。

优化声明：英特尔编译器针对英特尔微处理器的优化程度可能与针对非英特尔微处理器的优化程度不同。这些优化包括 SSE2、SSE3 和 SSSE3 指令集和其他优化。对于非英特尔微处理器上的任何优化是否存在、其功能或效力，英特尔不做任何保证。

本产品中取决于微处理器的优化是针对英特尔微处理器。不具体针对英特尔微架构的特定优化为英特尔微处理器保留。请参考适用的产品用户与参考指南，获取有关本声明中具体指令集的更多信息。

声明版本：#20110804

本文中提供的所有信息可在不通知的情况下随时发生变更。关于英特尔最新的产品规格和路线图，请联系您的英特尔代表。

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。

描述的产品可能包含可能导致产品与公布的技术规格有所偏差的、被称为非重要错误的设计瑕疵或错误。一经要求，我们将提供当前描述的非重要错误。

英特尔运营所需的任何商品和服务预测仅供讨论。就与本文中公布的预测，英特尔不负有任何购买责任。