英特尔中国公有云和互联网创新实践

云同行 AI 加速

Contents
目 录

趋势篇

案例篇

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

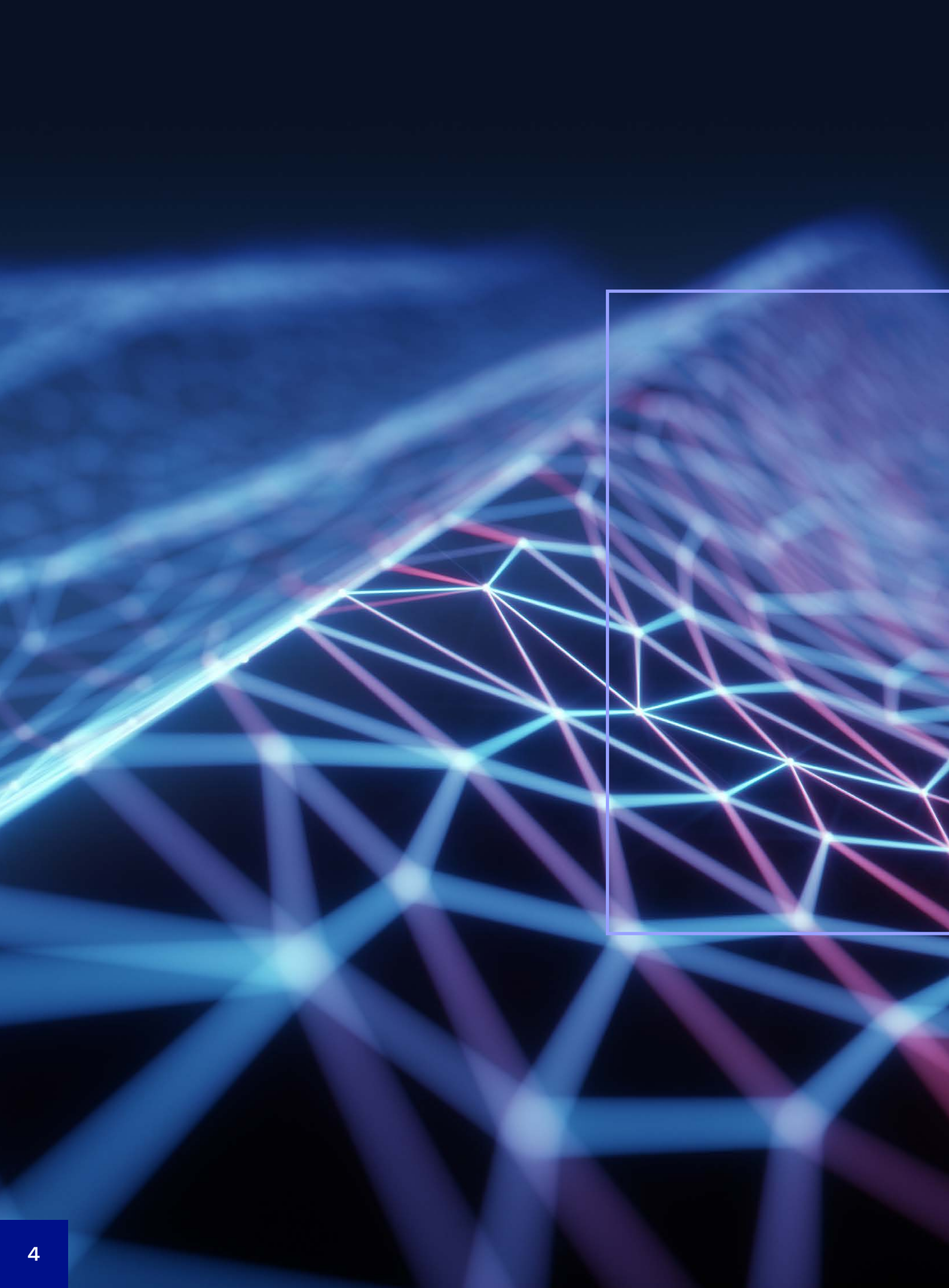
基础设施存储优化

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

云与数据中心新一轮的发现

2023 年初，ChatGPT 横空出世，狂飙科技圈。热度退却，当人们将目光聚集到其背后的支撑技术时，一个老生常谈，但从未让人产生倦意的名词——云计算，又一次浮现……

在数字经济和数字技术高速发展的今天，云计算一方面以滚雪球之势，将与基础计算相关的所有外延纳入自己的领地；另一方面，却又从未停止精细地打磨自己，让其内涵越发精致。是此，当许多技术已经坠入历史的尘埃，云计算却是“历久弥新”，不断发掘出新鲜的产品和应用。

数实融合、人工智能、产业创新，刺激云计算新一轮的增长

数字经济时代，作为数字经济底座的云计算被赋予多重涵义，它既是信息技术发展和服务模式创新的集中体现[1]，又是诸多前沿科技，如人工智能、区块链等的创新引擎。蕴藏其中的算力，更是被看作国家综合实力和国际话语权的关键要素，成为新的核心竞争力。

这种背景下，数字中国建设上升为国家重要战略，中央及各级政府出台了一系列数字产业政策，云计算作为新兴数字产业之一，为数字经济发展提供强有力的基础支撑，成为“十四五”期间重点发展产业之一[2]。

从发展阶段来看，中国云计算已进入高速增长阶段[3]，不再以上云数量来衡量云计算产业发展水平，而是转向关注云资源利用水平，以及云资源是否融入产业创新的新历程。为此，各级政府出台的政策都聚焦在对云资源的深入应用和创新上。比如，“二十大报告”（2022 年 10 月）强调利用云平台构建新的增长引擎；国务院《扩大内需战略规划纲要（2022-2035）》（2022 年 12 月）指出要推动云计算广泛、深度应用；工信部等八部门发布《关于推进 IPv6 技术演进和应用创新发展的实施意见》（2023 年 4 月）致力于鼓励 IPv6 与云计算等技术的融合创新。

在政策因势利导的作用下，中国云计算呈现持续发展态势，据研究，至 2025 年云计算市场规模将达 11,055 亿元，年复合增长率在 30% 以上[4]。

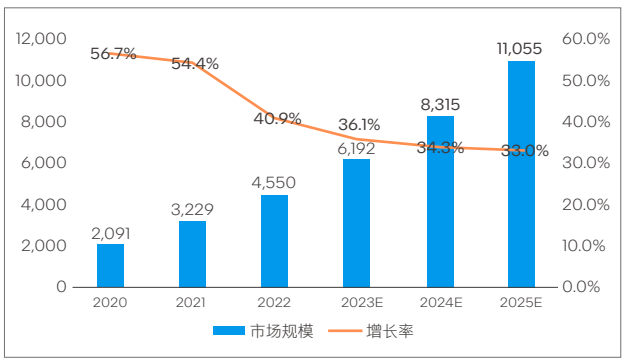


图 1-1-1 中国云计算市场规模及增速（亿元）[5]

从厂商来看，阿里、华为、腾讯和百度等头部厂商的地位相对比较稳固，几家头部厂商的市场份额占比将近 80%[6]。

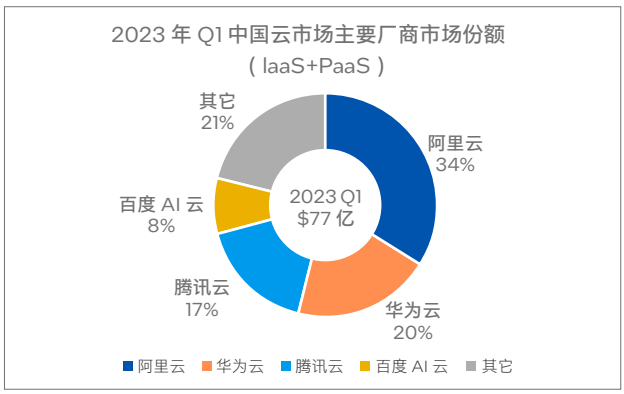


图 1-1-2 中国云计算市场份额分布[7]

[1] 引自：《云计算白皮书（2023）》，中国信通院，2023（P1 前言部分）
[2] 如欲了解更多详情请参阅：《云计算白皮书（2023）》，中国信通院，2023（P9 我国云计算发展概述）
[3] 援引自：https://zhuanlan.zhihu.com/p/386939601?utm_id=0
[4, 5] 数据援引自：中国信息通信研究院，2023 年 5 月
[6, 7] 数据援引自：Canalys estimates, Cloud Channels Analysis, June 2023

从产业来看，随着智能汽车的兴起，在中国长期处于上云、用云第二梯队[8]的汽车行业，将在换档后成为云计算市场上强劲的增长点。不仅如此，在与云计算深度融合后，汽车产业实现了智能座舱、智能驾驶、车联网、车云等多项创新。预计至 2026 年，中国汽车云行业市场规模持续增长超 800 亿元[9]。

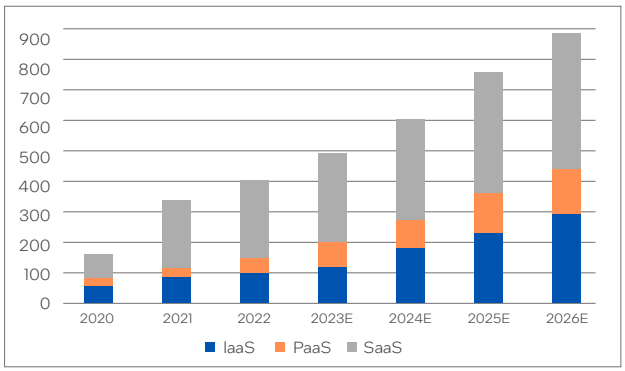


图 1-1-3 2022 年中国汽车云市场追踪报告[10]

产业数字化转型促进云资源的深度应用，触发云技术或应用模式系列变革

用起来放心（系统稳定性、云原生安全）、创起来随心（AI 驱动的应用开发）、闯起来宽心（专用高性能基础设施）、管起来省心（一云多芯）、买起来舒心（云优化治理），是产业用户的普遍期待，自然也是高速增长期的云计算该具备的特性。

面向产业创新和数智化转型驱动云资源深度应用的大趋势，包括信通院、IDC 等在内的权威机构都对云计算技术走向给予了高度关注，并进行了预测。

其中，信通院从云计算赋能千行百业转型升级的角度，发布了十个云计算新的应用方向，包括应用现代化、一云多芯、分布式云、低 / 无代码和智算服务等。而 IDC FutureScape 从云计算促进基础软硬件持续创新的角度，对中国云计算市场做了十项预测，涵盖云基础设施、云部署、云连接等多个方面。

尽管各家的描述不尽相同，但从发挥云计算的本质优势及与产业融合创新的角度来看，其焦点大同小异，无外乎用云（包括云的稳定性和安全性、基于云的创新、云在特殊场景的应用）和管云（包括云的运维和成本）。

对系统稳定、（云原生）应用安全、恢复时间趋零的至臻追求，是产业放心用云的前提

云平台的分布式属性决定了各模块之间关系错综复杂，单一节点的问题会牵一发而动全身，让平台的稳定性问题呈指数级爆发。

随着上云业务量的持续提升，企业系统面临着容量管理难、服务关系调用复杂等问题。而采用系统思维的稳保体系，即由事前规划、事中检测、事后管理形成的流程闭环，能增加平台的韧性，促进业务稳定运行。另外，可观测性应成为系统的“中枢神经”，提供实时监测和系统分析的能力。最后，机器学习和人工智能已被考虑引入稳定性保障体系，让系统稳如磐石[11]。

在安全层面，云原生革新了传统用云方式，将安全能力融入企业上云的全程。信通院认为云原生安全已成为云上安全防护的最佳路径。原因在于，云原生安全体系日趋成熟，且已从单点防护向全流程一体化防护转变，并覆盖了应用的全生命周期。

超越传统 IT 的故障恢复性能，也是产业用户对云计算的新追求。有数据显示，到 2026 年，30% 的中国 500 强企业将使用网络恢复即服务[12]。因为随着勒索软件攻击增加，企业需要更为复杂的恢复策略，而这些策略很难通过自研获得，而让网络恢复即服务成为云计算服务的一部分，可让企业用户唾手可得。

AI 驱动的开发，以低 / 无代码打破云上开发的技术壁垒，让应用创新更加随心所欲

一方面，云平台以产品化、自助式的平台，满足多场景下的应用开发。容器平台及传统 PaaS 等经过平台工程化，可演进为面向开发者的一站式平台，灵活组合 Backstage、Grafana、KubeVela 等云原生能力，对接算力、Kubernetes 等差异化基础设施，屏蔽其复杂性，使开发者可以专注于业务需求，而非代码[13]。

另一方面，人工智能技术正在加紧向基于云平台的开发渗透，其将通过自动生成代码，来满足新数字化解决方案在开发和早期部署时的功能和业务需求，从而显著提高开发人员速度。

专用高性能基础设施服务化、普惠化将推动云服务向算力服务模式加速演进，让人们在落实想法时更加宽心

人工智能与产业的深度融合，在交通、医疗、制造和城市治理等领域催化了诸多智算和科学计算应用场景。这一方面要求云计算能跨越硬件架构（CPU、GPU、FPGA），输出不同类型的服务；另一方面，需要云平台在覆盖多层级算力的基础上，促进算力服务的泛在化和普惠化。

当然，其前提是利用云计算所具备的硬件解耦、标准化封装部署等特性，促进算力能力的标准化输出，从而避免软件被固定形式的算力需求捆绑。

[8] 援引自：《云计算白皮书（2023）》，中国信通院，2023（P22 行业上云的阶梯分布）
[9, 10] 数据援引自：沙利文《2022 年中国汽车云市场追踪报告》
[11] 云计算白皮书（2023），中国信通院，2023（P20 关于“系统稳定性”的论证）
[12] IDC FutureScape:《2023 年中国云计算市场十大预测》
[13] 如欲了解更多详情请参阅：《云计算白皮书（2023）》，中国信通院，2023（P20 关于“产品化、自助式开发平台”的论证）

	云数据中心	智算中心	科学计算中心
建设目的	通用企业场景提供支撑服务	AI 相关的产业化和政府治理智能化	科学计算等场景支撑服务
主要应用领域	云计算	AI 相关场景，如知识图谱、自然语言识别、智能制造、自动驾驶、智慧农业等	基础科学研究、工业制造、模拟仿真、气象环境、天文地理等
主要投资主体	互联网、运营商等	AI 企业	政府、企业
主要芯片类型	CPU	CPU+AI 芯片	CPU+GPU 芯片
平均功率等级	4-8KW	CA9-15KW	>1MW

表 1 不同算力模式的对比[14]

一云多芯在满足多元算力需求的同时，屏蔽技术异构带来的复杂性，让云管更加省心

通俗地说，一云多芯就是用一套云操作系统来管理不同类型的芯片、架构、接口、技术栈等硬件服务器集群。

信通院《云计算白皮书（2023）》认为，一云多芯不仅可以对底层各种异构资源统一调度，还可以实现对上层应用和软件的适配操作，并保证云平台性能高效稳定。硬件芯片方面，通过屏蔽底层芯片差异实现资源池化，从而满足对各种芯片的统一调度，这不仅包含不同指令集架构的 CPU，也包括 CPU 以外的专有芯片；软件应用方面，一云多芯能够适配各种操作系统、虚拟机、容器数据库、中间件等，同时还能支撑虚拟化和云原生应用；性能调优方面，一云多芯可对不同芯片进行调优适配，提升平台整体性能，通过虚拟化产品性能调优、内核调优和部署架构优化，将性能指标差异控制在有效区间，从而高效释放算力，也因此具备了快速发展的内在驱动力。

信通院调研显示，当前一云多芯市场已达 900 多亿元人民币，预计 2024 年能够增长到 1,300 亿元左右[15]。

FinOps 将云优化治理落到实处，助力企业优化云资源应用成本管理

产业与云计算深度融合，也带来了云上资源应用成本大幅攀升、骇人的能源消耗以及因云资源利用的不合理性导致的浪费问题。在实现“双碳”目标的大背景下，云计算领域已将降碳减排作为关键战略，来促进云计算发展的可持续性和其上各项应用的降本增效。

在云平台中，通过将财务（Finance）与 DevOps 有机融合，可在对云资源成本实现透明化、可控化和优化管理的基础上，帮用户实现更高的用云效益和业务价值。首先，FinOps 可帮用云企业获得实时的成本数据和分析报告，了解云资源的使用情况和成本分布，以此使云资源消耗和成本驱动因素透明化；其次，FinOps 内置的资源分配、审批和监控功能，能帮企业实现对云资源使用的规范和限制，避免不必要的浪费；最后，FinOps 能够帮助企业在识别成本高、效益低的云资源的基础上，优化云资源的配置和使用策略。

万物数字化时代，一切归于算力，算力取决于芯片

2023 年可被看作人工智能划时代的一年，人工智能大模型的快速发展，引发数字技术应用方式和算力资源供给的双向变革，加速了云计算朝面向大体量分布式应用体系化、工程化创新的操作系统演进[16]，意即云计算正在演变为数字世界的操作系统，其重要性不言而喻。

虽然说，云计算的核心竞争力由运算能力、匹配速度、安全性能和成本优势等四部分组成，但算力无疑是核心的核心。所以，算力就成为评估一个云平台的实力关键指标。

而作为云计算上游产业的芯片产业被视作云计算发展水平的风向标。也正因此，云计算产业的繁荣与芯片技术的兴盛息息相关。中国半导体行业协会数据显示，中国芯片行业市场在 2017 年至 2021 年间的复合年均增长率达 17.9%，规模由 5,411 亿元增长至 10,458 亿元，预计 2023 年中国芯片市场规模将增至 12,767 亿元。[17]

单位：亿元

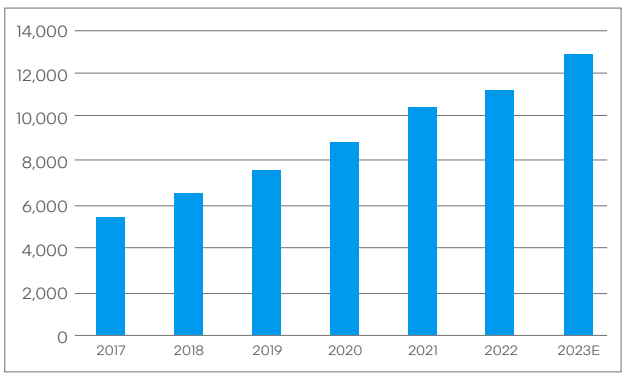


图 1-1-4 2017-2023 中国芯片市场规模预测趋势图[18]

英特尔® 至强® 可扩展平台，为云创新提供“芯”基石

英特尔一直走在超大规模云服务前沿，拥有广泛、优化的软件生态，并兼具跨多云环境的可靠性、灵活性和安全性。英特尔与全球领先云服务提供商开展的联合研发及业务合作，已经交付了数代专为云规模打造和优化的定制芯片，帮助实现从边缘到云的更全计算、更多存储、更快传输。

[14] 如欲了解更多详情请参阅：《人工智能计算中心发展白皮书》，中信所，2022
[15] 数据援引自：http://www.jwview.com/jingwei/html/07-28/551297.shtml
[16] 如欲了解更多详情请参阅：《云计算白皮书（2023）》，中国信通院，2023（P26 关于“云计算正向数字世界操作系统转变”的论证）
[17, 18] 数据来源：中国半导体行业协会、中商产业研究院整理

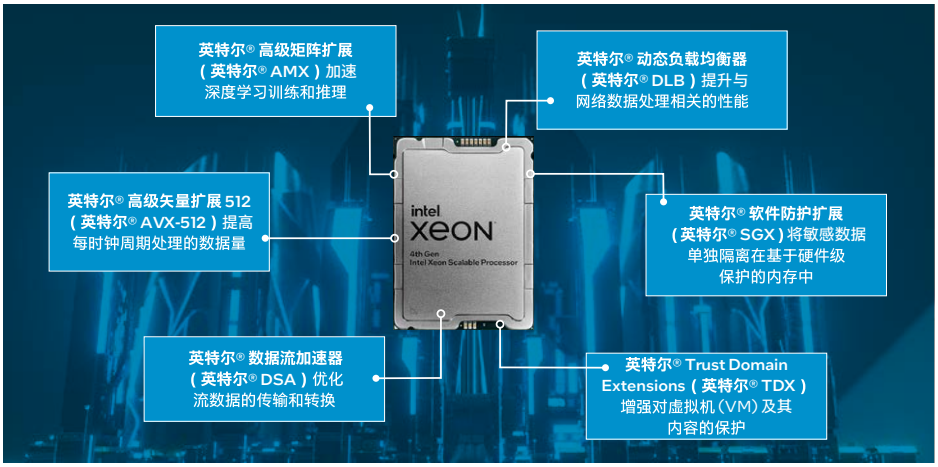


图 1-1-5　第四代英特尔® 至强® 可扩展处理器内置丰富加速引擎，重新定义云上负载性能

英特尔提供了从包括 CPU、GPU、FPGA、存储、网络等在内的全套硬件产品，到基于开源软件 (OSS) 的广泛软件堆栈，以及丰富的开发和优化等工具，可为当前和下一代工作负载提供值得信赖和可扩展的无障碍运行环境。依托于英特尔® 至强® 可扩展平台既灵活又可扩展的高性能产品组合，无论在混合云、多云还是边缘环境，企业都能够高效运行计算密集型、数据密集型等多样化工作负载。

第四代英特尔® 至强® 可扩展处理器，旨在满足紧迫的工作负载需求，同时提供高可信赖的云选择和应用可移植性。其内置众多加速器，可为 AI、数据分析、网络、存储和科学计算等快速增长的云工作负载提供性能和能效优势，从而帮助企业加速上云和优化云战略，赋能业务创新和提高投资回报率 (ROI)。

此外，第四代英特尔® 至强® 可扩展处理器还可面向企业和机构不断提升的安全性、身份与合规管理等迫切需求，加速实现机密计算，促进零信任安全策略实施，使在超大规模平台和其他分布式网络中，依托受监管工作负载的创新成为可能。

面向开发型企业内部转型的需求，特别是对超大规模平台和智能边缘环境中工作负载运行，及其所需的数百万计操作请求及高速分布式网络通信，第四代英特尔® 至强® 可扩展处理器通过内置的加速器，可以优化云平台间和云平台内的数据传输，实现负载平衡，进而通过支持关键的微服务，提升性能和保障业务跨云环境无缝运营。

对于 AI 与高级数据分析等增长尤其快速的负载，第四代英特尔® 至强® 可扩展处理器凭借其提升的性能以及内置的 AI 和密集工作负载等加速功能，能够帮助企业更高效地从数据中获取更多的洞察和价值，实现数据驱动型决策。

英特尔在为上云、用云以及优化云应用提供强劲硬件架构之外，还具备了丰富的云工具，如英特尔® Workload Optimizer、英特尔® Cloud Optimizer 和英特尔® oneAPI 工具套件等。其中许多工具都充分利用了英特尔® 至强® 可扩展处理器和其他英特尔® 处理器及平台中的硬件增强功能，不仅可助用户进一步提升云运营的性价比，而且可从提供性能优化、资源管理到定制云迁移建议，全流程支持企业实施和优化云战略，帮助大幅提高云基础设施的回报。

英特尔是云技术创新的引领者，多年来一直为云服务提供支持。如今，全球已有超过 1,500 万台服务器采用英特尔® 处理器，在主要云服务提供商的设施中运行多种多样的工作负载[19]。在此进程中，英特尔也建立起庞大的硬件和软件解决方案生态系统，提供了一致、稳定的技术堆栈，并通过数以千计来自真实场景的实战案例，提供更多选择和更好的互操作性，从而能够帮助产业用户和软硬件开发商、基础设施运营商、云服务提供商等以端到端模式，软硬一体，快速和高质量地实施云基础设施构建、应用交付模式以及从顶层到应用的优化，共同面向未来持续应对变革，全面解锁云潜力和释放数字生产力。

在下一篇 “案例实战篇” 中，我们将对英特尔与阿里云、腾讯云、百度智能云、火山引擎、京东云、美团、金山云和青云在云基础设施能力提升、绿色数据中心建设以及推荐系统性能优化、大语言模型推理优化等方面的实战经验进行分享，为广大行业伙伴提供示例参考。

[19] 如欲了解更多详情请参阅：https://www.intel.cn/content/www/cn/zh/cloud-computing/cloud-tools.html

案例篇

第四代至强® 内置英特尔® TDX，助阿里云 ECS g8i 为企业云服务提供更优安全防护

英特尔® TDX 有力支持阿里云为客户提供更便捷和更多样化的机密计算服务

扫码了解更多案例细节

挑战

如何在云环境中更有效地保护客户的数据资产是阿里云等云服务提供商的关注焦点。而伴随云服务逐渐成为各类企业核心业务系统的 IT 基座，系统在云环境中运行时的数据保护，同样受到了越来越多的关注。 虽然上一代云实例采用了英特尔® SGX 提供的应用程序级可信边界，能够保证重要代码和数据的机密性与完整性，但当客户的大多数应用程序或工作负载都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时，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、工作负载部署到安全云环境的难度，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。

解决方案

阿里云第八代企业级 ECS 实例 g8i（以下简称 "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，实现了：

■ 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。[20]

■ 全方位计算安全防护体系

由第四代至强® 可扩展处理器内置的英特尔® TDX，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 g8i 构建了一个基于虚拟化的硬件可信环境，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

构建硬件 TEE 环境的关键，是对内存中的敏感数据提供可靠的隔离和保护措施。由第四代英特尔® 至强® 可扩展处理器内置内存控制器提供的内存加密引擎，可以让客户在不修改系统和应用程序软件的情况下，使用临时密钥对运行中的内存数据进行加密，从而使敏感数据始终处于加密隔离状态。

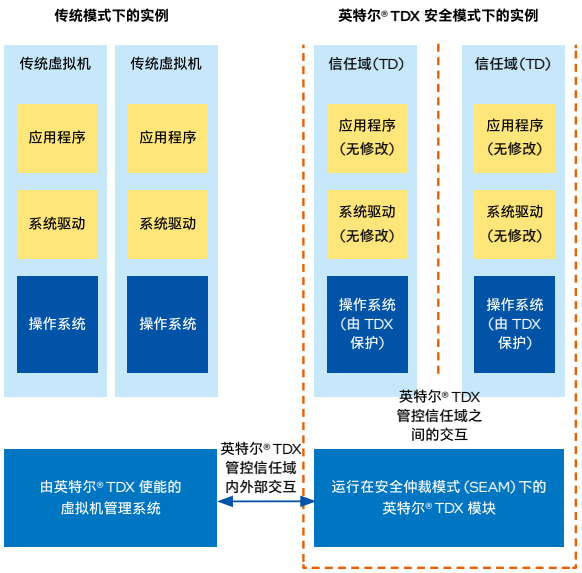


图 2-1-1 英特尔® TDX 技术架构

借助英特尔® 虚拟机扩展（英特尔® VMX）技术与英特尔® 多密钥全内存加密（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（SEAM）下的 TDX 安全服务模块来完成。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。g8i 可为客户提供机密虚拟机和机密容器两种使用模式。

方案总结

第四代至强® 可扩展处理器在为阿里云第八代企业级 ECS 实例提供强劲的算力支持之外，新处理器内置的英特尔® TDX，也为阿里云向客户提供面向虚拟化实例的机密计算新方案提供了坚实的技术保障，助力客户在不改变现有应用程序的情况下，为其 IaaS 和 PaaS 应用分别构建基于硬件设备的可信执行环境，如机密虚拟机或机密容器。同时，英特尔® TDX 技术使用便捷，客户能在阿里云环境中大规模部署并实现实时迁移，拥有更灵活和友好的保密云计算环境。

[20] 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/

第四代至强® 提供强劲加速，助阿里云 ECS 性能一路“狂飙”

阿里云 ECS g8i 通用及场景化性能双提升，构建全方位计算安全防护体系 [21]

扫码了解更多案例细节

阿里云推出了搭载第四代英特尔® 至强® 可扩展处理器的第八代企业级弹性计算实例规格族 ECS g8i，基于全面升级的第四代至强® 可扩展平台及其内置的丰富加速器，阿里云第八代 ECS 实例通用与场景化性能双双狂飙，同时构建出了更高安全等级的数据保护能力和云上可信运行环境。

第四代至强® 可扩展处理器加持，g8i 通用算力彪悍提升 [22]

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，且可依托 RDMA 协议栈的高性能、低开销特性，将 CPU 负载更多释放出来，使其更专注于业务处理。

这些独具的优势，在第四代至强® 可扩展处理器具备的 DDR5、CXL1.1、PCIe 5.0 等全新特性及内置加速器的支持下，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般表现。

多项至强® 内置加速器加持，g8i 场景化性能狂飙 [23]

在通用算力彪悍提升的基础上，g8i 实例还依托第四代英特尔® 至强® 可扩展处理器内置的丰富硬件加速器，实现了场景化性能的狂飙，其中在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。

场景化性能全面提升

加速器	场景	基准测试	性能提升
高级矩阵扩展 (AMX)	深度学习 (MLperf 性能测试)	resnet50（图像识别算法）	最大 207% ↑
		retinanet（目标识别算法）	最大 124% ↑
		bert（自然语言处理算法）	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	17-69 倍
		RSA 非对称加密算法	5-7 倍
存内分析加速器 (IAA)	数据存储	Rocksdb 测试	最大 100%

图 2-2-1 g8i 实例场景化性能全面提升

英特尔® 安全引擎，助阿里云构建全方位防护

g8i 实例在性能飙升之外，还以立体化、业界领先的计算安全防护体系，构建出又一特色优势，而英特尔® 安全引擎（英特尔® Security Engine）在其中功不可没。

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 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。

[21] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/ali-cloud-8th-ecs-performance-improved-with-xeon.html

[22、23] 数据援引自：http://news.sohu.com/a/659794618_115128

第四代至强® 内置 AI 加速，助阿里巴巴电子商务推荐系统性能显著提升

英特尔® AMX 助阿里巴巴推荐系统推理吞吐量提升达 2.89 倍 [24]

扫码了解更多案例细节

挑战

现代化推荐系统对 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 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，进而缩短时延和提高吞吐量，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。

解决方案

阿里巴巴选择第四代至强® 可扩展处理器进行性能优化，以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。

内置了创新的英特尔® AMX 加速引擎的第四代至强® 可扩展处理器代际性能大幅提升，且在 AI 性能上更进一步。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。

性能表现

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。

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 ® 可扩展处理器可将代理模型的吞吐量提升达 3 倍，同时将时延严格保持在 15 毫秒以下。[25]

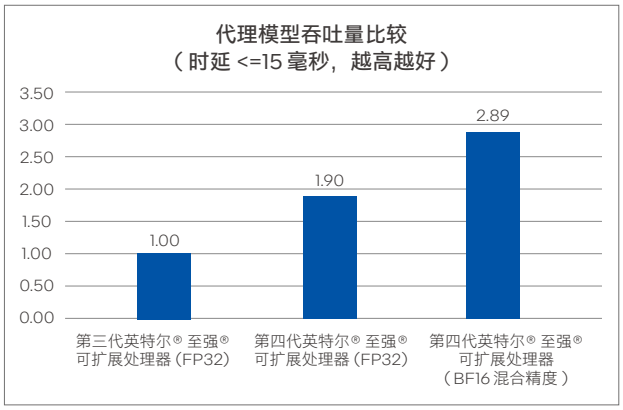


图 2-3-1 代理模型的代际性能比较 (时延 <=15 毫秒)[26]

方案总结

阿里巴巴与英特尔联合验证了，在利用第四代至强® 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。除了推荐模型之外，阿里巴巴还将探索在更多 AI 推理工作负载中使用第四代至强® 可扩展处理器，以释放该处理器的性能潜力。

[24、25、26] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/alibaba-e-comm-recommendation-system-enhancement.html

第四代至强® 内置英特尔® DLB，助力腾讯云实现高精度的网络限速

使用英特尔® DLB 的 Atomic Queue 实现无锁限速方案，助力腾讯云优化网络资源调度和分配

扫码了解更多案例细节

挑战

网络资源分配的常用方法是在网关对每个用户的带宽及并发控制和请求进行限速，以保护系统不会因为单位时间内的请求数量超载而造成拥塞，令牌桶算法是常见的限速机制之一。而开发者通过更改令牌桶的使用方式，配合一定的算法，降低 “锁” 竞争的概率，减少 “锁” 对性能的影响，这种方法称为轻量化锁。

轻量化锁限速方案包含两个关键参数：一是全局令牌桶产生令牌的速率，即限速后的目标速率；二是批量大小，当本地桶中令牌数量不足时，从全局桶预取令牌的数量。

全局令牌桶产生令牌的速率较低时，存在一种情况，即在单位时间内产生的令牌数无法满足所有本地令牌桶的批量预取请求。无法得到补充的本地令牌桶将因没有足够的令牌而导致报文被丢弃。然而，其他的本地令牌桶中却可能仍有未消耗的令牌，这些被丢弃的报文并没有超出限定的速率，导致限速后的速率低于目标速率。

以上原因会带来限速后的速率波动，让精度成为限速方案优化时必须关注的问题。

解决方案

第四代至强® 可扩展处理器引入了英特尔® DLB 技术，可有效解决高并发软件架构遇到的性能挑战。利用 DLB 的 Atomic Queue 特性，可在多核心的场景下实现无锁限速方案。

将待处理的网络报文按照其所属的限速网络数据流进行分组，英特尔® DLB 的 Atomic Queue 能够把属于同一分组的报文调度到同一个处理器核心进行处理；另外，Atomic Queue 还会为每一条流动态地选择处理器核心，当有多条网络数据流时，流量能够较为均匀地分散到各个处理器核心，确保处理器中多个核心的负载均衡。

在无锁限速方案中，处理器核心被分成了两组，从队列操作的角度，分别被称为生产者和消费者。生产者为每个报文生成 Atomic Queue 所需的 Flow ID，随后将报文入队到 DLB 的 Atomic Queue 中。DLB 在消费者线程间分发消息，同时保证原子性。消费者从 Atomic Queue 获取报文之后，以无锁的方式安全地访问 Flow ID 对应的全局令牌桶，完成限速相关操作。

在无锁限速方案中，由于只使用了全局令牌桶，因此不存在低速率时本地令牌桶预留令牌导致的限速后速率偏低，以及预取令牌导致的限速后速率偏高的精度问题。

性能表现

测试中以目的 IP 地址区分不同的需要限速的网络数据流，通过网络测试仪向被测设备发送不同目的 IP 地址的网络数据流，数据包在被测设备处理后返回给网络测试仪；网络测试仪每 2 秒统计一次数据包的接收速率，连续统计 20 次（共 40 秒），然后记录结果。

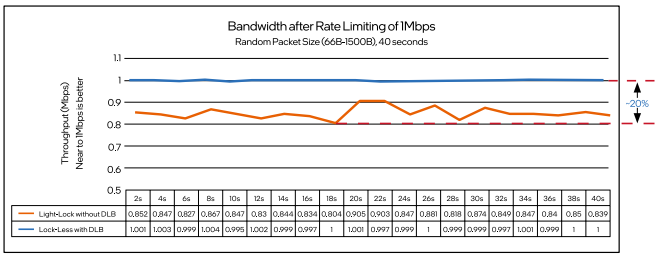


图 2-4-1 测试对比结果[27]

从图中可以看到[28]，无锁方案整体限速非常稳定且准确，整体误差小于 1%；而轻量化锁方案，限速后的流量速率偏小，且有大幅度波动，甚至出现了大于 20% 的误差，这充分说明使用基于英特尔® DLB 的无锁限速方案相比轻量化锁限速方案，能够获得更高的限速精度。

方案总结

基于上述阐述和测试可看到，基于英特尔® DLB 的无锁限速方案，借助英特尔® DLB 软件开发包提供的丰富开发库，可在 Linux 系统的内核态、用户态以及 DPDK 框架中实现精准限速，且可灵活地应用于不同场景。

[27, 28] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/high-precision-network-rate-limiting-with-dlb.html

集成英特尔® Neural Compressor，腾讯云 TACO Kit 为 AI 应用提供高效异构加速服务

优化后，在 bert-base-uncased-mrpc 场景中，推理性能提升达 2.39 倍 [29]

扫码了解更多案例细节

挑战

从云端、边缘到终端设备，更广泛的应用场景意味着 AI 的部署环境正变得更为复杂且多元化。而要在异构硬件平台上运行全栈软件，用户不仅需要基于不同的硬件基础设施来设计高效稳定的开发和部署方案，还需要根据业务场景、软件框架的不同来实施复杂的调优过程，任何环节的缺失和短板，都可能无法最大化发挥软、硬件的潜力，这不仅将抬高用户的技术准入门槛，也会大幅提升 AI 应用的构建成本。

解决方案

英特尔和腾讯云通过深入合作，以硬件异构、软件同构的构建模式，携手为用户提供了高性能的异构加速解决方案。

■ 腾讯云打造全新的异构计算加速套件 TACO Kit

腾讯云面向不同角色用户，包括 AI 方案设计者、AI 开发人员以及 AI 使用者推出的全新异构计算加速软件服务，计算加速套件 TACO Kit，以一系列软硬件协同优化组件和特有的硬件优化方案，为用户提供支持异构硬件的跨平台统一软件视角，并借助多元化异构、高性能加速框架、离线的虚拟化技术以及灵活的商业模式等优势，实现了对多元算力的轻松驾驭，从而助力用户的 AI 应用实现全方位、全场景的降本增效。

而作为异构加速服务的入口，TACO Kit 内置的 AI 推理加速引擎 TACO Infer 则能针对用户 AI 应用中不同的训练和服务框架、不同的优化实践和使用习惯、不同的软件版本和硬件偏好，以计算加速、无感接入和鲁棒易用的特性和优势，帮助用户一站式解决 AI 模型在生产环境中部署与应用的痛点。

■ 英特尔® Neural Compressor 助力 TACO 加速推理性能

英特尔® Neural Compressor 可通过插件的方式集成到 TACO Kit 中。得益于英特尔® Neural Compressor 提供的优势特性，TACO Kit 在与之实现集成后，能够利用量化压缩技术来为不同的深度学习框架提供统一的模型优化 API，实现便捷的模型推理优化过程（由 FP32 数据类型量化为 INT8 数据类型）。同时，其内置的精度调优策略可根据不同的模型内部结构生成精度更佳的量化模型。该过程不仅大幅降低了用户进行模型量化的技术壁垒，也有效提升了 AI 模型的推理效率。

在云端部署时，量化后的模型可通过至强® 可扩展平台内置的英特尔® DL Boost 来获得行之有效的硬件加速。借助英特尔® DL Boost 所提供的 AVX-512_VNNI（矢量神经网络指令）指令集，量化为 INT8 数据类型的模型能获得更高的推理效率。

性能表现

推理性能加速结果如图 2-5-1 所示 [30]，在保持精度水平基本不变的情况下，各个深度学习模型的推理性能均获得了显著的提升，提升幅度从 55% 到 139% 不等。在其中的 bert-base-uncased-mrpc 场景中，推理性能达到了基准值的 2.39 倍，获得了令人满意的成果。

方案总结

腾讯云与英特尔联合验证的结果表明，将英特尔® Neural Compressor 以插件形式集成到 TACO Kit 中，得益于其提供的量化技术及其它性能调优特性，以及英特尔® DL Boost 对量化后模型提供的硬件加速，AI 模型的推理性能可获得显著提升。

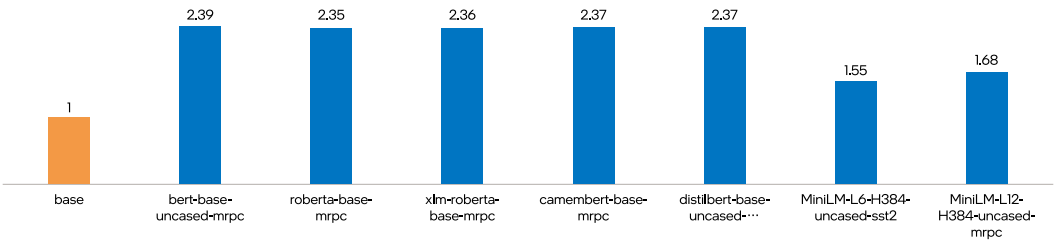


图 2-5-1 集成英特尔 ® Neural Compressor 的 TACO Kit 所带来的推理性能加速 [31]

[29、30、31] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/neural-compressor-tencent-cloud-taco-kit-ai.html

第四代至强® 可扩展平台全面升级，助百度智能云打造新一代云智一体架构产品

基于英特尔® 架构的处理器及内置加速器为百度智能云提供算力与 AI 能力支持

扫码了解更多案例细节

挑战

为了与用户需求形成更深层次的融合，百度智能云计划打造具有性价比的异构算力和高效的 AI 开发运行能力。而在这一过程中，百度智能云也面临着新的需求与挑战：

- **更强劲且支持 AI 加速的算力需求：**AI 应用对算力需求的持续提升，以及对性能功耗比的更多关注，正推动百度智能云寻求能支撑多种算力，覆盖更广维度计算场景，且能够有效实现 AI 加速的多元化算力设备；
- **更多面向 AI 计算的内存需求：**AI 模型参数量的持续增大，对内存容量提出更高要求，内存墙无形中成为一部分 AI 应用的瓶颈。同时，传统 AI 算力设备内存扩容能力有限，且传输速率升级无法兼顾低成本与算力的高速增长；
- **更高的安全和数据合规性要求：**元宇宙等融合虚拟资产的出现以及企业对数据收集、脱敏、标注等流程安全性的关注，推动百度智能云在数据安全性上投入更多关注，这不仅涉及到算力部署形态的调整，也对算力设备提出了物理级别的安全防护要求。

解决方案

■ 全新处理器架构及性能升级，为 AI 应用提供充沛算力

在用户所聚焦的算力需求上，第六代 BCC/BBC 产品引入第四代至强® 可扩展处理器来满足用户在不同场景下的需求。搭载第四代至强® 可扩展处理器的第六代 BCC/BBC 产品可支持最大 3.1GHz 基频，3.4GHz 全核睿频的高主频实例规格族，让用户能够游刃有余地选择更贴近业务能力的算力底座。[32]

■ 全新处理器内置 AI 能力，为 AI 应用注入有效加速

第四代至强® 可扩展处理器内置英特尔® AMX，可帮助百度智能云大幅升级 AI 性能。值得一提的是，为更有效地实现对英特尔® AMX 指令的调用，百度智能云也引入了英特尔® oneAPI 工具套件。

■ 借力傲腾™ 持久内存，满足 AI 计算所需大容量内存

在算力提升之外，第六代 BCC/BBC 产品也针对 AI 计算中关键的大容量内存需求，引入傲腾™ 持久内存 300 系列。在大容量与成本优势之外，傲腾™ 持久内存 300 系列也能对 CXL 协议起到支持和推动作用。

■ 以英特尔® SGX 为数据安全性提供硬件级保障

内置于第四代至强® 可扩展处理器中的英特尔® SGX，能在内存等硬件环境中构造出一个可信的安全“飞地”(Enclave)，为敏感数据和代码提供独立于操作系统和硬件配置的、增强的安全防护。通过英特尔® SGX 的加持，第六代 BCC/BBC 产品的实例可在很大概率上保证云上运行业务的代码、数据不被 OS、虚机监控器等监视、修改，从而能提供对业务过程数据安全性的重要保障。

方案总结

随着第四代至强® 可扩展处理器等产品与技术在百度智能云第六代 BCC/BBC 产品中获得充分融合，由第六代 BCC/BBC 产品提供的云实例也在生命科学、自动驾驶以及工业制造等多个领域的 AI 场景中获得落地部署，并在实践应用中获得了用户的良好反馈。

> “未来的云服务将在实现各行各业数字化转型的基础上，聚焦并持续深化推动其智能化升级进程。我们基于云智一体 3.0 架构的第六代云服务器、裸金属等产品，就是从 AI IaaS 层和 AI PaaS 层能力入手，提供用户所需的极致算力和高效 AI 开发能力。第四代至强® 可扩展处理器、英特尔® AMX 等产品与技术，为我们的新产品提供了从算力、AI 加速、内存扩容到数据安全等的全面加持。”
>
> **谢广军**
> 百度副总裁
> 百度智能云

[32] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/baidu-build-cloud-intelligence-integration-product.html

第四代至强® 内置 AI 加速，助轻量化 ERNIE 3.0 基于 CPU 平台显著提升推理性能

英特尔® AMX 助百度 ERNIE -Tiny 性能提升达 2.66 倍 [33]

扫码了解更多案例细节

挑战

NLP 是 AI 领域的重要分支。作为拥有强大互联网基础的领先 AI 公司，百度凭借其旗下飞桨文心·NLP 大模型所具备的创新技术优势，在语言理解、语言生成等 NLP 场景中已获取了明显的市场优势。ERNIE 3.0 作为百度飞桨文心·NLP 大模型的重要组成部分，在各种 NLP 应用场景，尤其是中文自然语言理解和生成任务中展现出卓越的性能。在其实际落地应用过程中，许多细分领域根据自身业务特点提出了特定化需求。为此，百度推出多个 ERNIE 3.0 轻量化版本 ERNIE-Tiny，在通用平台上即可高效率完成推理作业。与此同时，引入更强的通用计算平台和优化方案，也是助力 ERNIE-Tiny 获得更优效率的另一项重要手段。

解决方案

■ 引入第四代英特尔® 至强® 可扩展处理器为 ERNIE 3.0 带来更强 AI 加速引擎

英特尔® AMX 采用全新的指令集与电路设计，在实际工作负载中，其能同时支持 BF16 和 INT8 数据类型，与上一代 AI 加速引擎相比，大幅提升 AI 工作负载的效率，有助于提升 ERNIE-Tiny 在推理环节的性能表现。

■ 利用英特尔® oneDNN 实现对英特尔® AMX 指令的调用

为了让英特尔® AMX 的加速能力能直接作用于 ERNIE-Tiny，百度与英特尔一同借助英特尔® oneDNN 来实现英特尔® AMX 指令的调用。作为开源的、跨平台的性能库，英特尔® oneDNN 可有效助力用户提升其 AI 应用与框架在英特尔® 架构平台上的性能，且其也已加入了对英特尔® AMX 的支持。

■ 内存性能优化

ERNIE-Tiny 推理过程中有许多串行操作，每次运算都会先读数据再写数据，循环往复会消耗大量操作时间。优化方案则是将矩阵乘法与元素的运算及激活融合在一起，即把连续的操作合并为一个大操作，使内存的运行效率显著提升。

同时，方案中针对多线程的优化也被证明可助力 ERNIE 3.0 提升推理计算性能，与上一版本相比，方案进一步优化了多线程的效率，并提升了多核的扩展性。

性能表现

测试在第四代至强® 可扩展平台与第三代至强® 可扩展平台之间展开。后者使用英特尔® AVX-512_VNNI 对模型进行了 INT8 量化提速，而前者则启用英特尔® AMX 技术进行加速。测试结果显示，ERNIE-Tiny 性能获得显著提升，对比上一代至强® 可扩展平台，吞吐量提升达 2.66 倍 [34]。

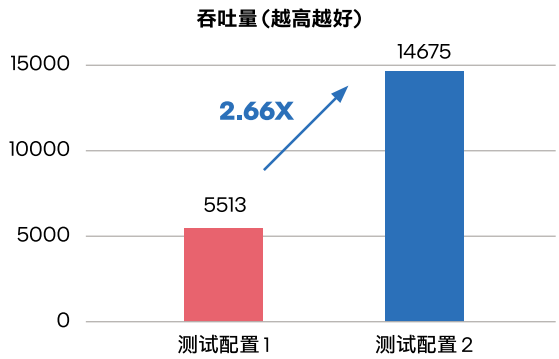


图 2-7-1 ERNIE-Tiny 在不同平台上的性能对比 [35]

方案总结

百度与英特尔本次协作优化的成功，再一次证明各个行业用户在通用的 CPU 平台上也能同样方便地部署高效能的 ERNIE-Tiny，用以应对越来越多的 NLP 应用需求。使用这一方案，用户不必额外采购昂贵的专用 AI 算力设备，这将大幅降低企业借助 NLP 能力提升业务效率的门槛，并加速更多 NLP 技术与应用的商业落地过程。

> “第四代至强® 可扩展处理器及英特尔® AMX 的引入，使得轻量版 ERNIE 3.0 在通用 CPU 平台上也能获得令人满意的推理效能，能帮助更多用户在其既有 IT 设施中更为方便地部署 ERNIE 3.0，从而进一步普及其应用范围。”
>
> **孙宇**
> 百度杰出架构师
> 百度自然语言处理部

[33、34、35] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/artificial-intelligence/spr-built-in-amx-baidu-ernie-performance-increase.html

英特尔® 数据中心 GPU Flex 系列助火山引擎打造高画质、低时延云游戏体验

火山引擎引入英特尔® 数据中心 GPU Flex 系列，作为云端渲染和编码工作核心引擎

扫码了解更多案例细节

挑战

面对海量玩家在线时，对实时渲染和编码工作负载的高需求，火山引擎需开发新一代云游戏解决方案，在保证其支撑的云游戏大作具有更华丽炫酷的画面品质和更加流畅顺滑的操作体验的同时，具备以下效能表现：

- **更快的端到端响应速度**：为让玩家体验云游戏时，不因响应时延过长而影响游戏体验，端到端响应时延压降至 70 毫秒以内[36]；
- **更优的网络状态调节能力**：为使云游戏在不同移动网络环境下依然能保持良好的流畅感，新方案计划加入网络状态动态调节能力来确保网络延迟一致性；
- **更低的单路云游戏实例成本**：计划通过提升 GPU 处理效能，降低硬件和网络带宽成本。

解决方案

火山引擎首先借助其在边缘计算领域的经验积累和服务能力，在新方案中将云游戏实例广泛地部署到全国乃至全球范围内的边缘节点，并通过自研的智能调度技术来为玩家分配适宜的云游戏实例，就近接入、渲染和编码，为玩家构建低时延的交互体验。

其次，新方案也引入了经过亿级 DAU 产品验证打磨过的自研 RTC 产品 ByteRTC，从而大幅优化全网络链路的延迟状况，并支持最高 4K/60FPS 的视频流输出，打造更佳画质。

在降低单路成本方面，火山引擎引入英特尔® 数据中心 GPU Flex 系列，与至强® 可扩展平台一起作为新方案的核心处理引擎，在提供强大渲染能力和硬件编码能力的同时，提升云游戏实例中渲染和编码工作负载的效能。同时，英特尔也为 Flex 系列 GPU 产品提供了丰富的软件栈，来提升其渲染与编码性能以及可运行的游戏种类。

性能表现

现在，这一全新的云游戏解决方案已在抖音云游戏平台等互联网产品中获得了成熟化运营，并面向玩家推出了《航海王热血航线》等一系列精品云游戏大作[37]。

在英特尔® Flex 140 GPU 卡的加持下，游戏在 720p60 帧场景中可获得 60 路的编码能力和 20 路的渲染能力，远高于同等能耗的其它 GPU 产品。而在 1080p60 帧场景中，编码能力和渲染能力则分别达到 28 路和 10 路。[38]

同时在英特尔® Flex 140 GPU 卡提供的硬件编码加持下，新方案使用 H.265 编码格式相比于传统的 H.264 编码格式，能使云游戏在同等画质下减少约 40% 的网络带宽[39]，无疑可有效帮助客户压降成本。

《航海王热血航线》	分辨率	GPU 编码	GPU 渲染	单 GPU 卡并发数
FLEX 140	720P 60FPS	15 * 4	10 * 2	20（路）
	1080P 60FPS	7 * 4	5 * 2	10（路）
FLEX 170	720P 60FPS	15 * 2	30	30（路）
	1080P 60FPS	7 * 2	15	14（路）

表 2 新方案为《航海王热血航线》带来的性能提升[40]

方案总结

双方计划通过更深入的技术协作，完善和优化英特尔® 数据中心 GPU Flex 系列、至强® 可扩展处理器等在云游戏解决方案中的运用，进一步满足云游戏服务在处理性能、端到端时延、部署成本以及服务稳定性等方面的高品质要求，进而在推动云游戏体验持续提升的同时，为云游戏产业乃至数字经济的发展提供更强动力。

> “借助云服务、边缘计算等技术的加持，我们的云游戏解决方案正帮助更多玩家不受终端配置束缚，体验到更多的高品质游戏大作。而英特尔® 数据中心 GPU Flex 系列产品在音视频渲染与编码等方面的卓越表现，使新方案能在更低的成本下为玩家提供更高画质的视觉效果和低时延的操作互动，获得了玩家的高度认可。”
>
> **梁宇**
> 云游戏架构师
> 火山引擎

[36、37、38、39、40] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/volcano-engine-create-cloud-gaming-experiences.html

基于至强® 可扩展处理器的火山引擎 g2i 实例助力加速 VASP 医药分子模型运算

某生物医药科技公司利用英特尔® MPI 库，对 VASP 分子训练模型进行 NUMA 亲和性优化

扫码了解更多案例细节

挑战

VASP（Vienna Ab-initio Simulation Package）是维也纳大学 Hafner 小组开发的进行电子结构计算和量子力学 - 分子动力学模拟的软件包。它是材料模拟和计算物质科学研究中最流行的商用软件之一。某医药公司在使用 VASP 生物医药分子模型优化自己的算法时遇到了性能低下的问题。

解决方案

火山引擎和某医药科技公司一起针对 VASP 分子模型的特点，进行技术匹配和测试，并发现英特尔® oneAPI 工具套件中的 MPI 库能够助其获得理想结果。

■ 测试环境

- 采用基于面向单路和双路的第三代英特尔® 至强® 可扩展处理器的火山引擎 g2i 实例。
- 英特尔® oneAPI 工具套件的 MPI 库。MPI（Message Passing Interface），是开发者们在科学计算程序中，用于在参与计算的不同 CPU 或服务器节点之间进行消息传递的一组规范或接口。通过这组接口，能帮助开发工程师们在不同的计算平台上快速编写可跨平台移植的并行计算程序，提升开发效率。

■ 测试流程

- 步骤 1，安装英特尔® oneAPI 工具套件，并使能环境变量；
- 步骤 2，基于英特尔® oneAPI 工具套件中的英特尔® Compiler 和英特尔® MKL（英特尔® 数学核心函数库），编辑 VASP 软件包的 makefile，编辑相关库的地址，打开编译器优化配置，编译构建 VASP 程序；
- 步骤 3，获取测试用例，运行 VASP 程序。

■ 优化方案

- 基于业务模型提供定制化解决方案，其中包括使用英特尔® oneAPI 工具套件提升性能 [41]；
- 客户测试中遇到了一个异常 case，导致容易超时甚至运算失败等问题，而且在不同的配置下最终运算结果会有小幅度的差异；
- 通过英特尔与火山引擎联合定位分析，最终选择了新版本的英特尔® MPI，结合测例中的 NPAR 参数调优：

  - VASP 官方推荐实践：$NPAR \approx \sqrt{\#\ of\ cores}$
  - 调优后的最佳实践：去掉 NPAR 这个参数，可以获得平衡的性能以及稳定的结果输出。

性能表现 [42]

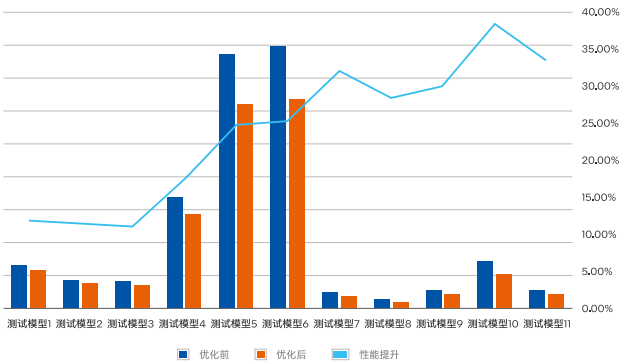


图 2-9-1 优化性能提升对比（单位：秒，越低越好）

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.cn/PerformanceIndex

方案总结

火山引擎向该生物医药科技公司提供了英特尔® oneAPI 工具套件中的 MPI 库，广泛赋能更加数字化和智能化的药物研发效率升级，并对 VASP 分子训练模型进行了 NUMA 亲和性优化，极大地提升了运算性能，从而可以进一步提高研发成功率，并降低研发成本。

[41] 测试日期为 2022 年 9 月，该数据由字节跳动提供，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。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.cn/PerformanceIndex

[42] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/volcano-engine-vasp-pharmaceutical-molecular-model.html

采用基于英特尔® 架构处理器的京东云绿色数据中心高密度算力方案

第四代至强® 可扩展处理器助京东云数据中心液冷机架解决方案实现 1.1 的 PUE[43]

扫码了解更多案例细节

挑战

为了满足日益增长的算力需求，同时降低能耗，京东云希望能够解决以下挑战，从而持续改善机架式服务器的空间利用率与功率密度：

■ 12V 总线电压设计的损耗较高

在新一代服务器的功率已经远超前代产品的背景下，12V 的总线电压对于供电效率的制约已经愈发明显。12V 总线电压设计会降低传输至负载的功率，严重影响系统功效。

■ 功率瓶颈导致机柜空间浪费

多种部件功耗提升导致机柜总功率显著增加，既有供电系统难以满足如此高功率的供电需求，导致数据中心必须减少机柜的服务器部署数量，以应对功率方面的限制，这导致了约三分之二的空间闲置。

■ 传统冷却方式的效率不足

传统的空气冷却系统无法在高密度集群垂直机架阵列的 IT 设备入口处提供均匀温度的空气，且冷却效率相对较低，在经济性、供电、噪音等方面都会带来巨大的困扰。

■ 提升服务器的算力能效比

只有提升服务器的算力能效比，并更加灵活地调用服务器的算力资源，才能有助于降低特定任务的能耗。

解决方案

■ 采用第四代英特尔® 至强® 可扩展处理器提升性能密度与能耗

京东云绿色数据中心高密度算力方案采用第四代至强® 可扩展处理器。除了释放第四代至强® 可扩展处理器的基础算力优势之外，京东云还灵活应用处理器内置的英特尔® IAA、英特尔® AMX、英特尔® DSA、英特尔® QAT 等高级硬件能力，提升数据中心面向多种负载的能效。

■ 采用模块化服务器设计实现更加高效的散热设计，并发布从数据中心级到微处理器级的冷板液冷整体解决方案

该方案组建冷却液回路，利用 CDU 分配冷却液。在通过冷板收集计算节点的热量后，冷却液不断流向另一个冷的 CPU，并通过另一个连接器离开服务器冷板管道，实现液冷计算节点的液冷循环。

■ 改善供电方案

京东云还推出搭载了第四代至强® 可扩展处理器的大功率、高利用率整机柜液冷服务器，通过将整机柜供电的总线电压从 12V 提高到 54V，降低整机柜服务器内部的连接阻抗和导电阻抗，从而大幅降低电源输出端到终端设备的全链路传输损耗。

方案收益[44]

- 铜排效率改善 3.8%，铜排损耗减少 770W；
- 54V 转 12V 模组效率达到 97.7%；
- 节省数据中心 TCO；
- 降低碳排放，京东云数据中心的液冷机架解决方案已实现 1.1 的 PUE。

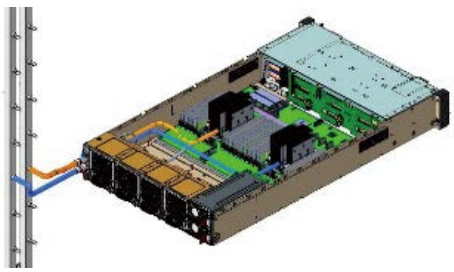


图 2-10-1　京东云服务器冷板液冷方案

方案总结

通过在整机柜中采用 54V 总线电压设计，以及全机柜液冷方案，京东云能够显著降低服务器供电损耗，提升电源链路效率，降低碳排放，助力打造高性能、高密度、可持续的绿色数据中心。除了供电、冷却方面的节能减排措施之外，第四代至强® 可扩展处理器等新一代硬件将有助于提升性能密度与能效，助力降低服务器在多种负载下的能耗。

[43、44] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/high-density-computing-power-solution-jd-cloud.html
截止至 2023 年 4 月京东云和英特尔联合测试得出的数据，通过比较 54V 天枢机架和 12V 传统机架的数据得出。测试配置：第四代英特尔® 至强® 可扩展处理器 ES2XCC，48 核，2x350W；2,048GB 总内存 (32x64GB)；2x240GB M.2+16x2TB NVMe。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

第四代至强® 内置 AI 引擎，助美团加速视觉 AI 推理服务，优化成本

英特尔® AMX 助美团将主流视觉模型推理性能提升最高达 4.13 倍 [45]

扫码了解更多案例细节

挑战

在美团，视觉 AI 已成为推动商业模式创新，为用户带来更加精准、个性化的互联网服务，并提升竞争优势的优先选择。但美团的视觉 AI 推理也在算力和成本等层面面临如下挑战：

■ 性能

美团高速增长的业务与用户量正使得越来越多的应用需要通过视觉 AI 构建智能化流程。美团需要在保证视觉 AI 推理精度的同时，提升视觉 AI 推理的吞吐率。

■ 成本

面向海量数据的视觉 AI 推理意味着大规模的基础设施投入。对于低流量长尾模型推理服务，采用 CPU 进行推理通常是更具成本效益的选择。

■ 灵活性

美团希望提升视觉 AI 服务的敏捷性，通过灵活地跨多种架构进行资源调度，满足长尾化场景的 AI 推理需求。

解决方案

为进一步提升视觉 AI 推理服务的性能表现，美团采用第四代英特尔® 至强® 可扩展处理器，并利用处理器内置的英特尔® AMX 加速引擎，以及英特尔® Integrated Performance Primitives（英特尔® IPP）等软件套件进行优化。

第四代至强® 可扩展处理器代际性能大幅提升，同时借助其内置的加速器，用户可以在 AI、分析、云和微服务等类型的工作负载中获得优化的性能。该处理器内置的英特尔® AMX 加速引擎，采用了全新的指令集与电路设计，通过提供矩阵类型的运算，显著增加了 AI 应用程序的每时钟指令数（IPC），可大幅提升 AI 工作负载中的训练和推理性能。在实际的工作负载中，英特尔® AMX 可同时支持 BF16 和 INT8 数据类型。

此外，美团还结合了英特尔® PyTorch 扩展（IPEX）加速 PyTorch。IPEX 是英特尔发起的一个开源扩展项目，它基于 PyTorch 的扩展机制实现，通过提供额外的软件优化更充分地发挥硬件特性，帮助用户在原生 PyTorch 的基础上更有效地提升英特尔® 处理器上的深度学习推理和训练的计算性能。

性能表现

在多个视觉 AI 模型中，美团通过采用英特尔® AMX，动态将模型数据类型从 FP32 转换为 BF16，从而在可接受的精度损失下，增加吞吐量并加速推理。为了验证优化后的性能提升，美团将使用英特尔® AMX 加速技术转换后的 BF16 模型，与基准 FP32 模型的推理性能进行了比较。测试数据如图 2-11-1 所示，在将模型转化为 BF16 之后，模型推理性能可实现提升达 3.38-4.13 倍，同时 Top1 和 Top5 精度损失大部分可以控制在 0.01%-0.03%。[46]

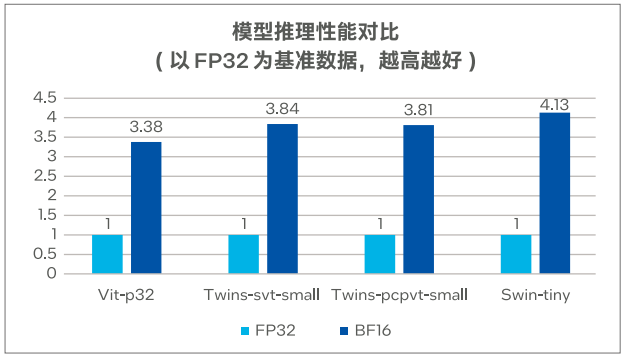


图 2-11-1 FP32/BF16 模型推理性能对比 [47]

得益于性能的提升，美团能够更加充分地释放现有基础设施的潜能，降低视觉 AI 服务的投资规模，线上资源效率整体提升 3 倍以上，并节省 70% 的服务成本，实现了资源的敏捷调度，能够支撑视觉 AI 服务的高效创新。[48]

方案总结

美团的视觉 AI 推理优化实践证明，内置英特尔® AMX 加速引擎的第四代至强® 可扩展处理器可有效提升 AI 推理性能，并降低视觉 AI 推理服务的 TCO。在现有工作成果的基础上，美团与英特尔还致力于进一步利用硬件创新、软件优化持续提升推理性能，充分释放视觉 AI 服务的价值。

[45、46、47、48] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/meituan-visual-ai-reasoning-service-optimize-cost.html

第四代至强® 助金山云第七代性能保障型云服务器 X7 优化，显著加速 AIGC 模型推理

在搭载英特尔® 至强® 铂金 8458P 处理器的新一代云服务器 X7 上，Stable-Diffusion 推理性能提升达 3.97 - 4.96 倍[49]

扫码了解更多案例细节

挑战

生成式人工智能（AIGC）等创新浪潮驱动了 AI 的新一轮增长，模型训练和模型推理成为云服务器的重要负载。要满足 AI 领域的市场需求，云服务提供商需要解决以下挑战：

- 加速数据清理、模型推理等 AI 端到端工作流程中的多种工作负载，加快平台的一站式性能；
- 高效使用 CPU 等现有的硬件资源，并利用客户公有云、私有云和混合云中的服务器资源，以降低硬件成本；
- 增强云服务器的灵活性，使其能够在复杂场景中敏捷扩展，支撑传统负载与 AI 等新型工作负载高效运行的需求。

解决方案

AI 已成为推动数字化创新的重要动力，伴随着 AIGC 等应用的快速落地，深度学习模型规模与复杂度不断提升，数据量也持续增长，AI 算力供给与需求之间的矛盾正在日趋凸显。用户希望优化硬件、软件和算法，在保证模型精度和时延等指标的前提下，提升 AI 端到端流程的性能表现，从而充分释放硬件的潜能，并降低系统 TCO，加速 AI 技术的创新。

为帮助用户加速 AI 端到端流程，特别是提升推理性能，基于第四代英特尔® 至强® 可扩展处理器的金山云第七代性能保障型云服务器 X7 进行了针对性优化。服务器采用了处理器内置的英特尔® AMX 加速器，并融合金山云自主创新的加速技术，能够有效提高 AI 模型的推理性能，同时发挥云服务器在敏捷性、扩展性等方面的优势，助力客户挖掘 AI 时代的价值。

性能表现

金山云测试了新一代云服务器 X7 在 Stable-Diffusion 模型推理中的性能表现。第四代至强® 可扩展处理器在 Stable-Diffusion 模型推理中有着卓越的性能表现，这源于其在算法上面的优化。针对该模型的 MHA 计算瓶颈，英特尔® PyTorch 扩展包（IPEX）插件在 2.0 版本发布了基于至强® 可扩展平台的 Flash Attention 算法，主要内容包括以合适的尺寸拆分矩阵计算，实现更高效的缓存利用；使用张量 AMX BF16 加速 MHA 矩阵计算，达到更快的速度；将计算缓存区与线程绑定，实现更少的内存开销。

在搭载英特尔® 至强® 铂金 8458P 处理器的金山云新一代云服务器 X7 上，双方对 Stable-Diffusion 模型推理性能进行了测试。测试数据如图 2-12-1 所示，相较优化之前的模型，在使用 IPEX 2.0 BF16 优化之后，Stable-Diffusion 模型推理性能提升达 3.97 - 4.96 倍[50]。

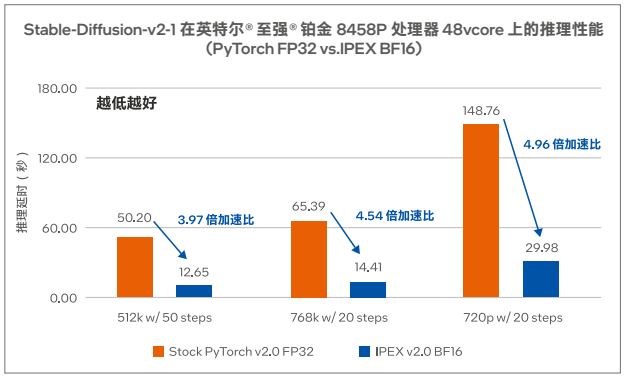


图 2-12-1 Stable-Diffusion 模型优化前后性能对比[51]

方案总结

通过自研技术的创新以及第四代至强® 可扩展处理器的采用，金山云第七代性能保障型云服务器 X7 在各大业务场景上性能较上一代均有大幅提升，这能够为用户的云上业务带来更高的收益：

- 以更高的性能满足广泛实际应用场景对性能的需求，特别是在 AI 性能方面，金山云新一代云服务器 X7 能够有效加速 AIGC 等应用的运行；
- 通过应用英特尔® AMX，加持算法优化，充分释放硬件潜力，有效利用服务器资源，从而降低端到端 AI 应用流程的 TCO；
- 不受限于特定应用类型，能够灵活应对深度学习、数据库、高网络收发包等负载的支撑需求，实现更高的敏捷性与扩展性。

[49、50、51] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/kingsoft-cloud-7th-gen-performance-cloud-server.html

第四代至强® 及其多种内置加速器，助青云 QingCloud 新一代 e4 云服务器实现性能突破

青云科技采用英特尔® AMX 优化之后，在满足精度的前提下，AI 模型推理性能提升达 4-5 倍[52]

扫码了解更多案例细节

挑战

在用户加速拥抱数字化的背景下，越来越多的数据与应用被迁移到云端环境，云平台承受着越来越大的压力，这些压力包括：

- **负载日趋复杂化、带来了多元算力需求**：现代化的数据中心需要提供多元化算力，将负载卸载到特定的加速器上，以支持上层应用使用更优架构完成每项任务。
- **基础设施规模越来越大、TCO 持续攀升**：只有尽可能地提升基础设施的性能密度，释放硬件潜能，才能够更好地控制 TCO 增长，实现更高的投资收益。
- **数据安全面临严峻挑战**：数据是企业的重要竞争力，也是企业创新的重要基础。数据的安全性如何有效保护，是企业亟需解决的一大难题。

解决方案

■ 基于第四代至强® 可扩展处理器加速多种工作负载性能

为了给用户提供高性能的基础算力支撑，青云科技利用第四代至强® 可扩展处理器内置的多种高级硬件特性，优化应用负载性能，释放了处理器在性能、稳定性、扩展性、安全性等方面的潜力，铸就卓越基础设施平台。

- 采用英特尔® AMX 优化基于 CPU 的 AI 性能，在满足精度需求的前提下，AI 模型推理性能得到了大幅提升；
- 采用英特尔® QAT 优化数据压缩与加解密性能，青云科技有效提升了 ZFS 存储系统压缩性能、OpenResty HTTPS 服务性能，并降低了虚拟机实时迁移耗时；
- 采用英特尔® SGX 构建可信的密钥管理服务，提供密钥计算、交换等复杂的安全计算环境，更有效地抵御多种类型的攻击，保护应用与数据的安全；
- 采用英特尔® IAA 加速数据库，青云科技有效提升了 MongoDB、ClickHouse 性能。

■ 利用 HBM 内存加速应用的内存访问

英特尔® 至强® CPU Max 系列是唯一一款基于 x86 的高带宽内存处理器，为解锁和加速受内存限制的科学计算和 AI 工作负载而设计。青云科技选择 HPL、VASP、Lammps 三个软件，测试在科学计算集群中， HBM 内存相较于 DDR 内存的性能提升。以 VASP 为例，随着核心数的增加，使用 HBM 内存的提升效果很明显，使用 HBM 内存 22 核心的性能基本上和 44 核心的 DDR 内存计算效率持平[53]。

性能表现

青云科技测试数据显示，在采用英特尔® AMX 优化之后，在满足精度需求的前提下，AI 模型推理性能，包括 Bert、ResNet 等，可以提升 4-5 倍[54]。

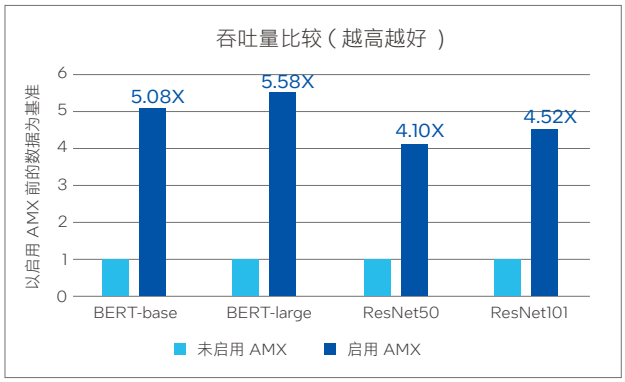


图 2-13-1 启用英特尔® AMX 前后的吞吐量比较[55]

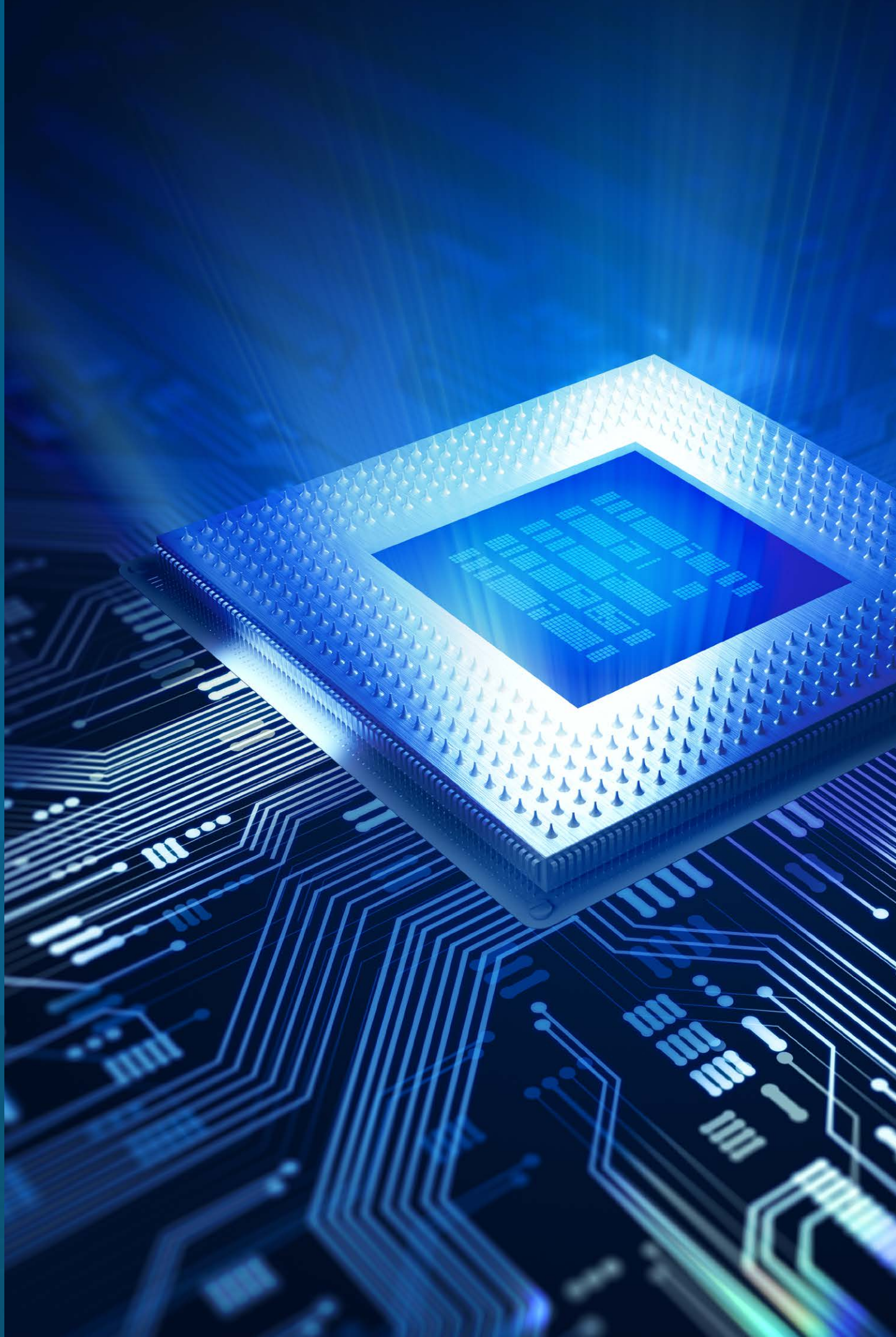
方案总结

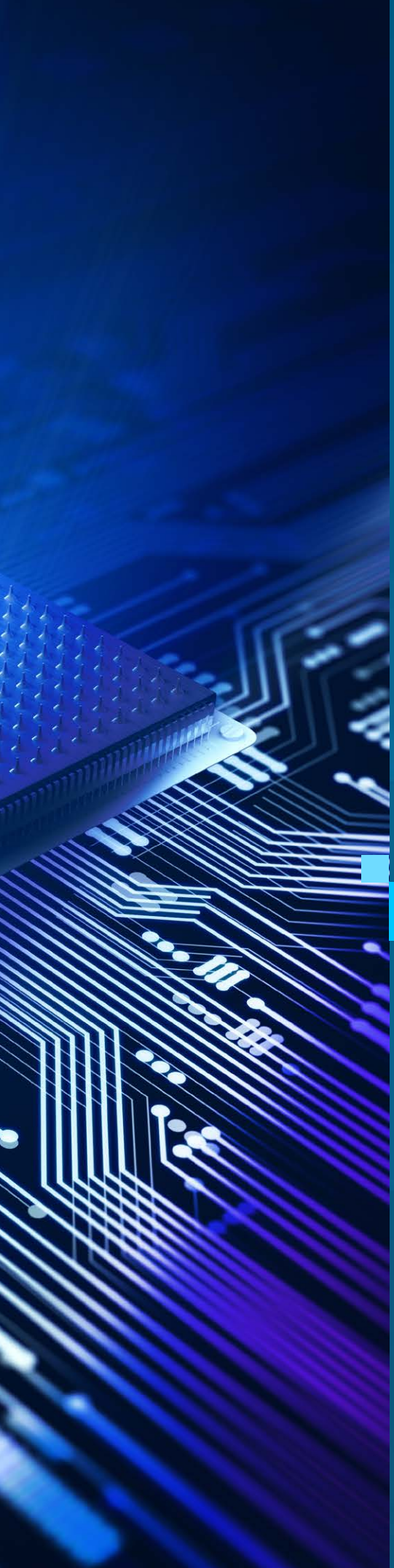
通过搭载第四代至强® 可扩展处理器，并利用处理器内置的高级硬件特性，青云 QingCloud 新一代 e4 云服务器实现了巨大的性能飞跃，满足了企业对即时数据高并发、高吞吐量处理、低时延等需求，通过提供更高性能、更稳定、更高性价比的基础支撑，帮助企业实现云化，加速数字化转型。

[52, 53, 54, 55] 如欲了解更多性能详情，请访问：https://www.intel.cn/content/www/cn/zh/cloud-computing/qing-cloud-cloud-server-performance-breakthroughs.html

扫码查看英特尔官网，
了解更多英特尔公有云和互联网创新实践

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[56]，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[57]，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)。

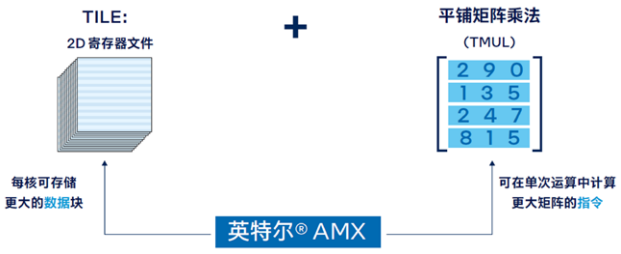
■ CXL

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。

[56、57] 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

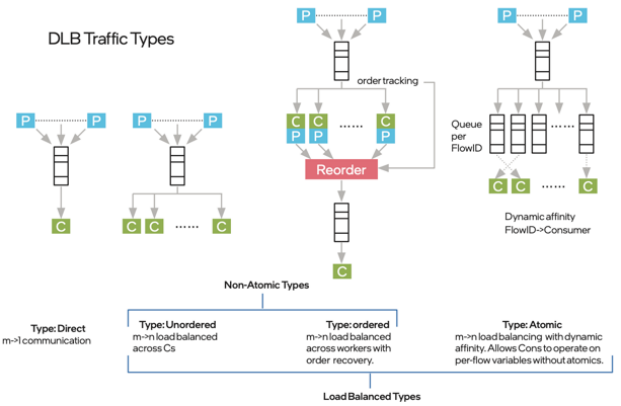
- 市场上的主流框架、工具套件和库 (PyTorch、TensorFlow)、英特尔® oneAPI 深度神经网络库 (英特尔® oneDNN)

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

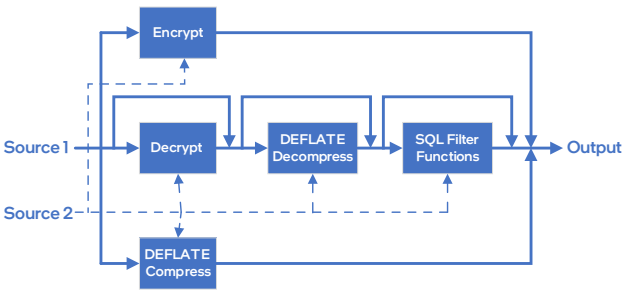
- 英特尔® Data Mover Library

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

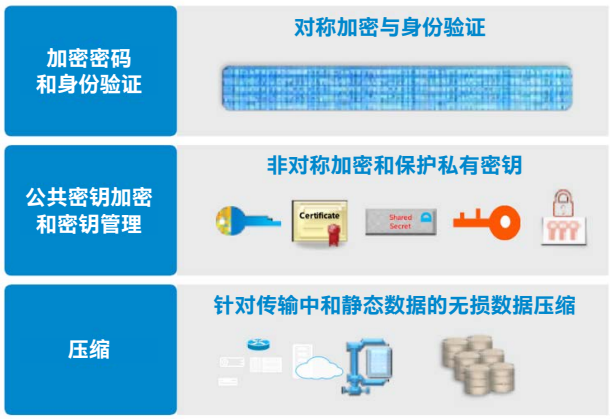
- 英特尔® Query Processing Library 和英特尔® Data Mover Library

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

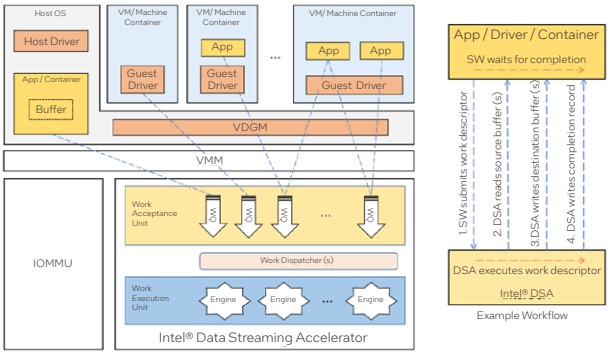
- 加速密码操作的英特尔® QAT 引擎

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

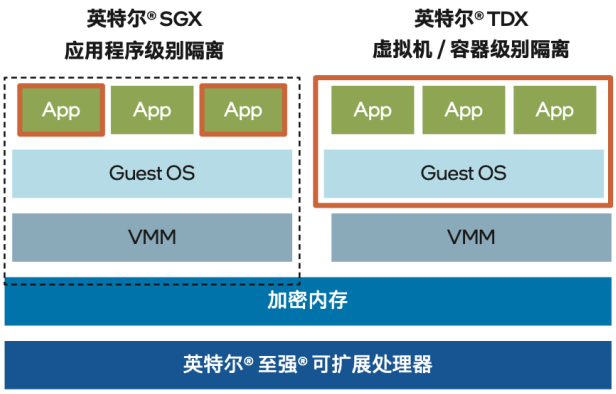
- 英特尔® Data Mover Library

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[58]。

“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。

[58] 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。[59]

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）**。

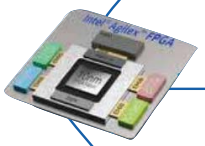
	英特尔® 数据中心 GPU Flex 140	英特尔® 数据中心 GPU Flex 170
目标工作负载	媒体处理和交付、基于 Windows 和 Android 的云游戏、虚拟桌面基础设施、AI 视觉推理[2]	
显卡外形规格	半高、半长、单宽、被动散热	全高、四分之三长、单宽、被动散热
显卡 TDP	75 瓦	150 瓦
每卡 GPU 数量	2	1
GPU 微架构	Xᵉ HPG	
Xᵉ 内核数量	16 个（8 个/GPU）	32
Fixed Function Media	4（2 个/GPU）	2
光线追踪	是	
峰值算力（脉动阵列浮点运算）	8 TFLOPS (FP32)/105 TOPS (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	

[59] 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



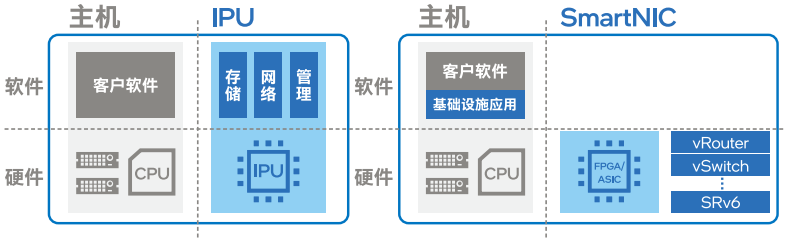
英特尔® Agilex™ FPGA

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

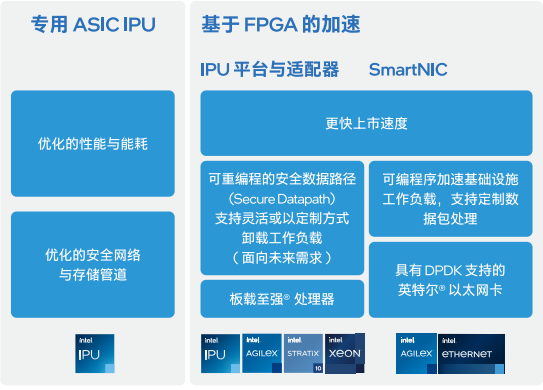
英特尔® 基础设施处理器（IPU）和 SmartNIC

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

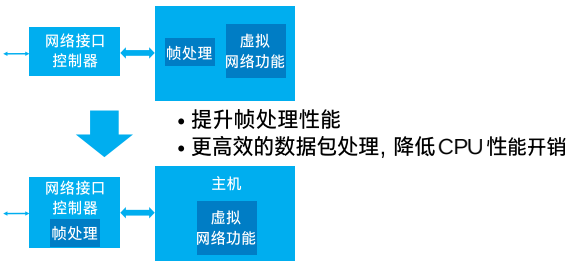
英特尔 ® 以太网产品发展路线图

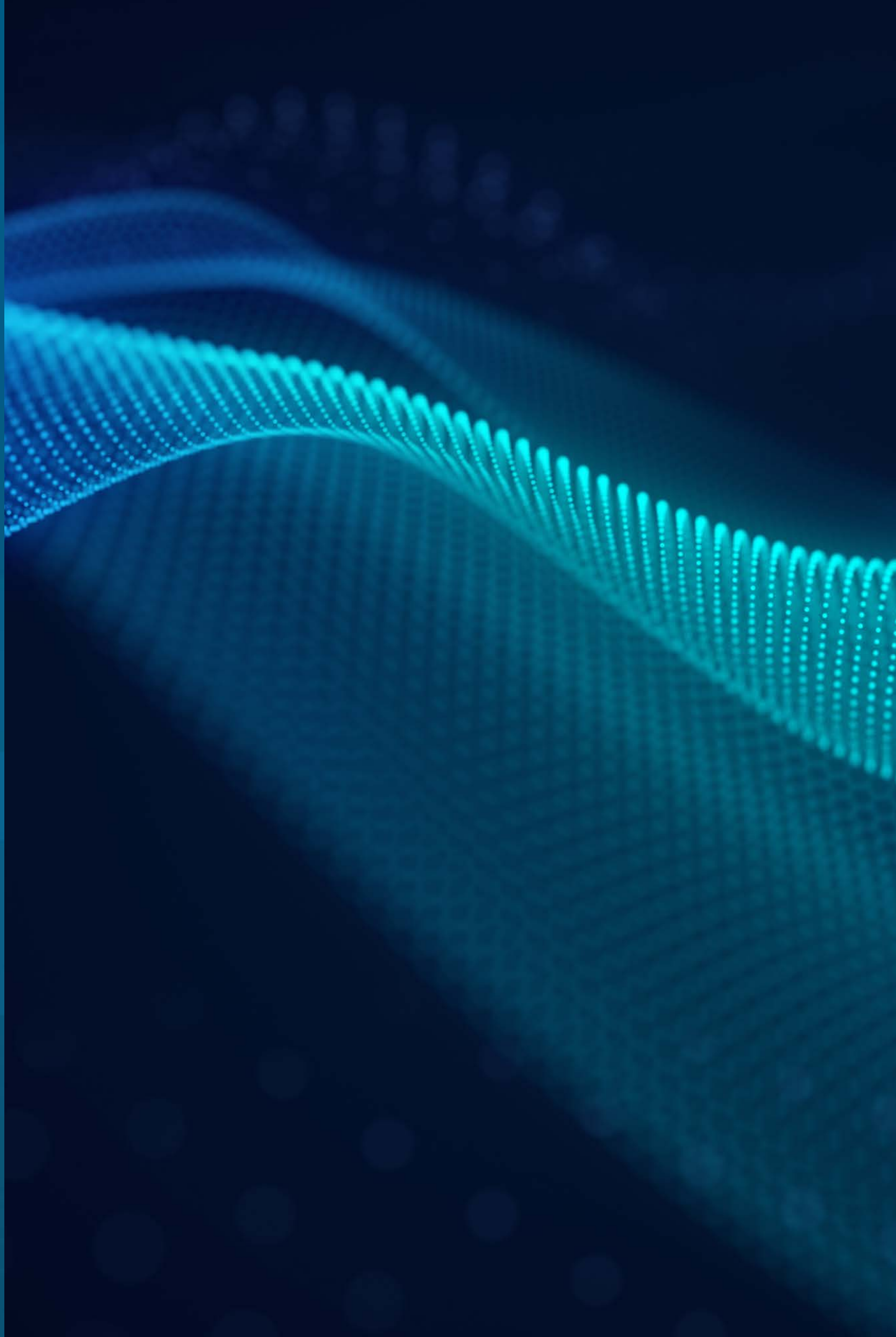
应用设备队列

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。



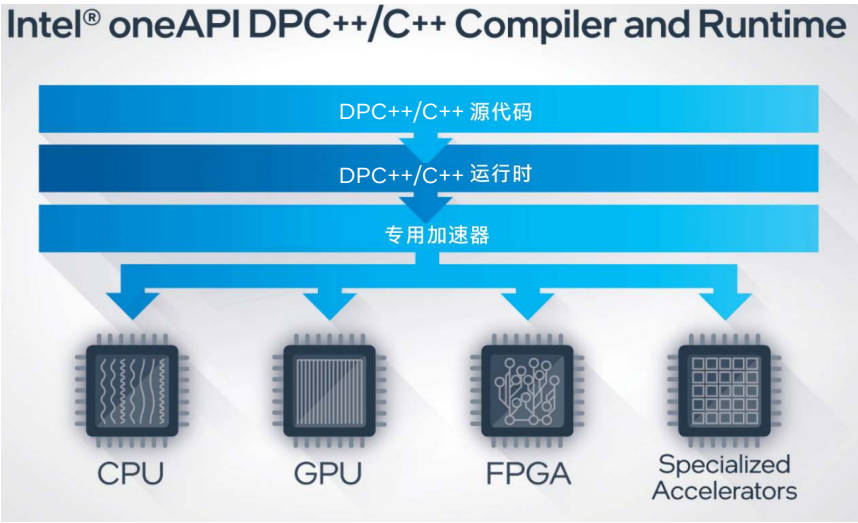


软件及系统级优化

基础设施算力优化

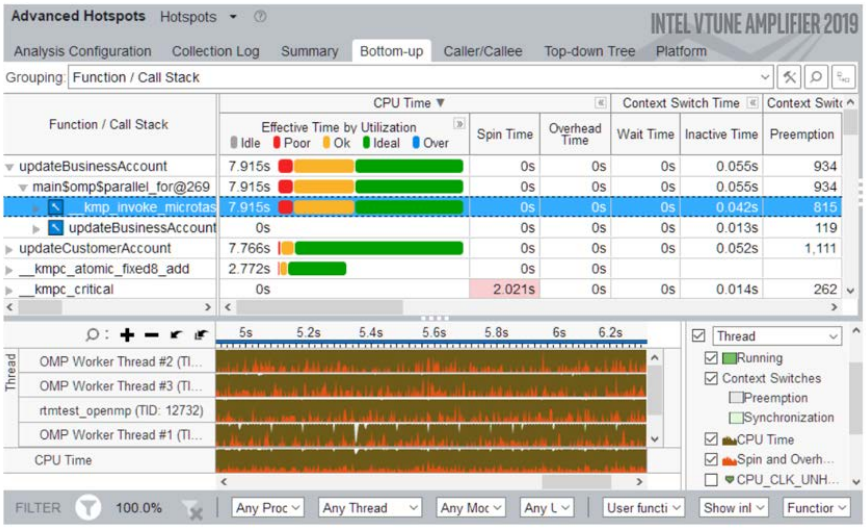
英特尔® oneAPI DPC++/C++ 编译器

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。

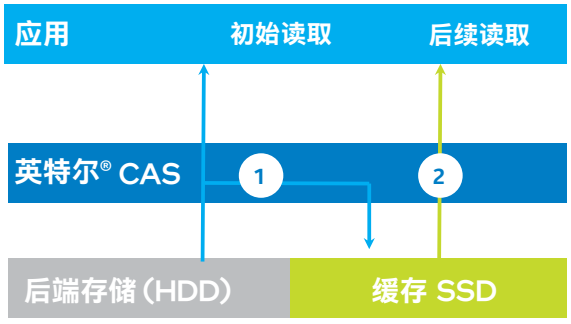


基础设施存储优化

英特尔® 高速缓存加速软件（英特尔® CAS）

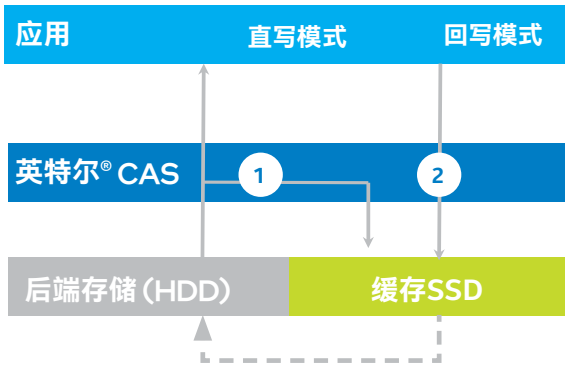
英特尔® 高速缓存加速软件作为一款服务器端缓存软件，可通过与内存进行互操作，以及与高性能固态盘（SSD）相结合，通过智能缓存管理，将最活跃的数据放入高性能固态盘介质，来提高应用程序性能，解决数据中心 I/O 性能瓶颈问题。

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

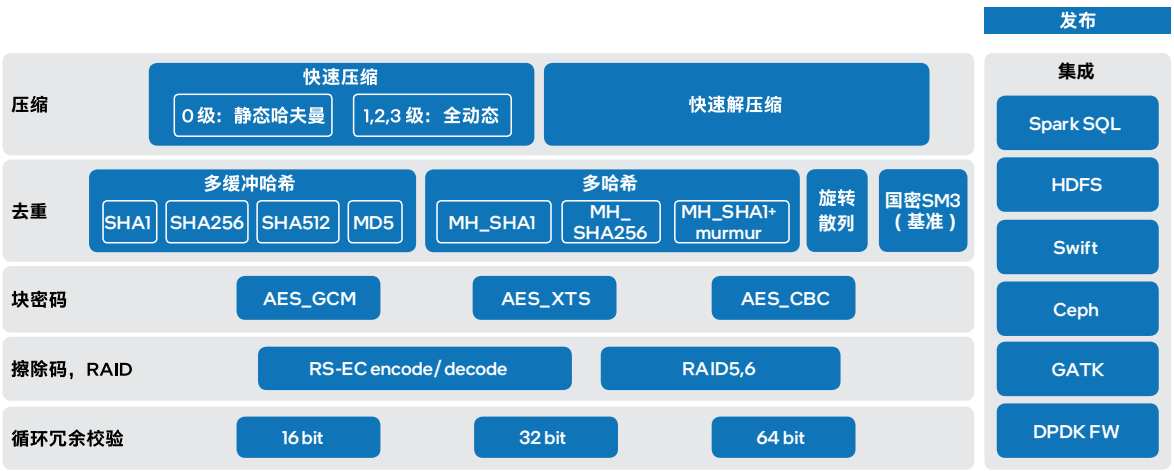
写入工作流程



- 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

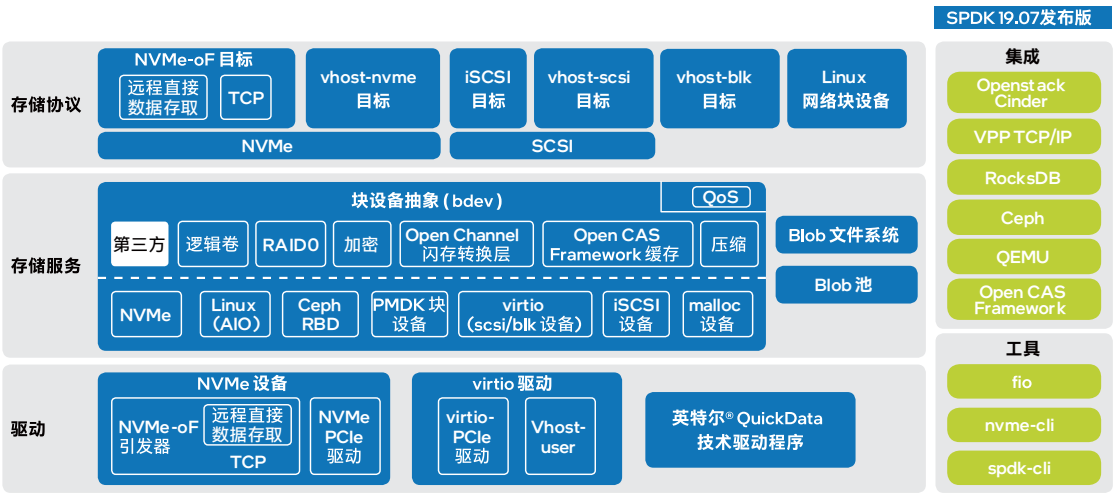
英特尔® 智能存储加速库（英特尔® ISA-L）

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）

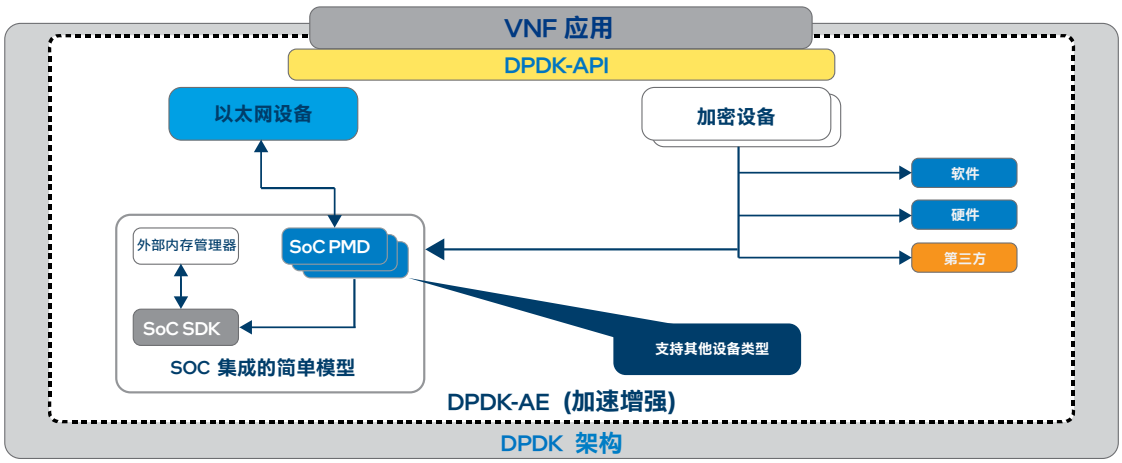
SPDK 是一套用于编写高性能、可扩展、用户模式存储应用程序的工具和库。它通过将所有必需的驱动程序移入用户空间，避免系统调用；提供完整的块堆栈作为驱动用户空间库，它执行许多与操作系统中的块堆栈相同的操作；提供基于这些组件的 NVMe、ISCSI 和 vHOST 服务器，这些组件能够通过网络或其他进程提供磁盘服务，来实现高性能。



基础设施网络优化

数据平面开发套件（DPDK）

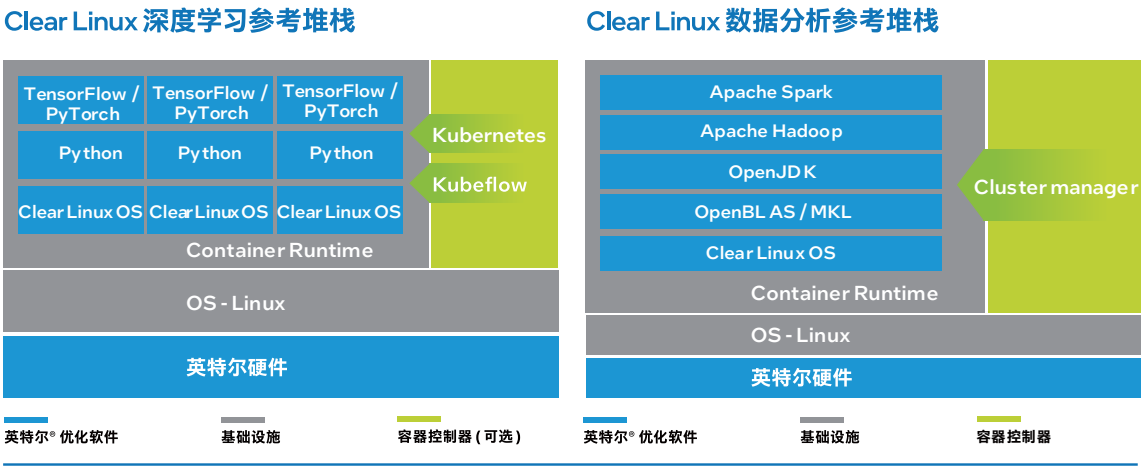
DPDK 是英特尔推出的一种高速网络数据包软件开发套件，现已开源。初期主要支持英特尔® 处理器及网卡系统，现已支持部分非英特尔® 架构处理器，以及部分非英特尔的网卡，能够通过旁路 Linux 系统网络协议栈，直接对网卡进行读写，结合多核处理器中不同核心的绑定，能够实现网络小包流量下的线速收发。DPDK 可以极大提高数据处理性能和吞吐量，为数据平面应用程序提供更多时间。



操作系统和编排层优化

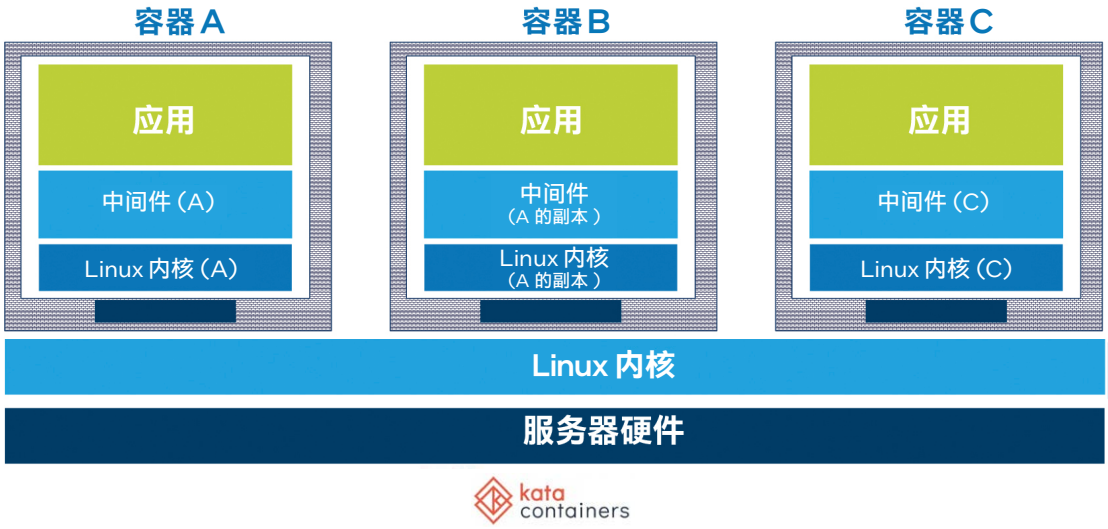
Clear Linux

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。



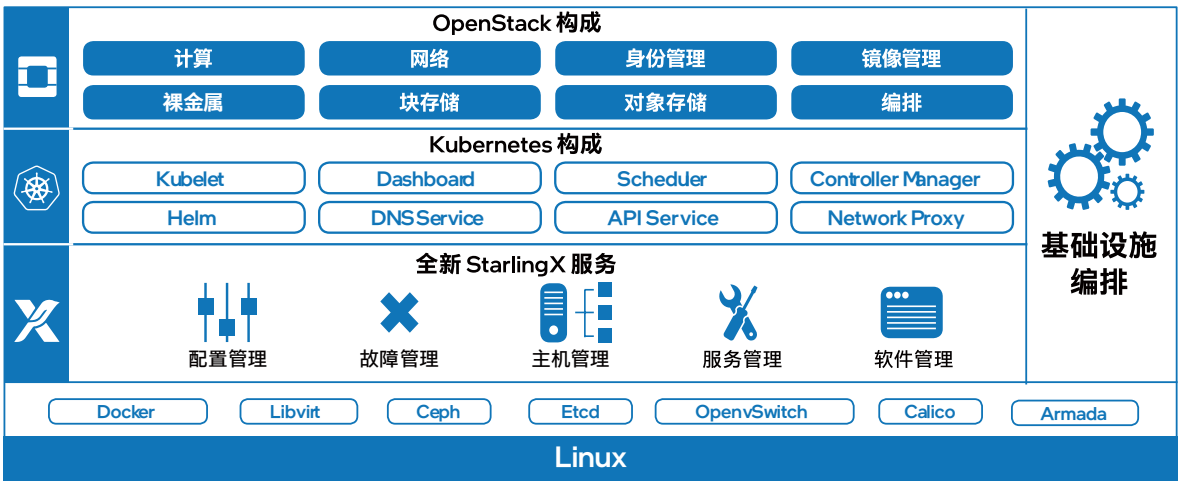
Kata Container

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。



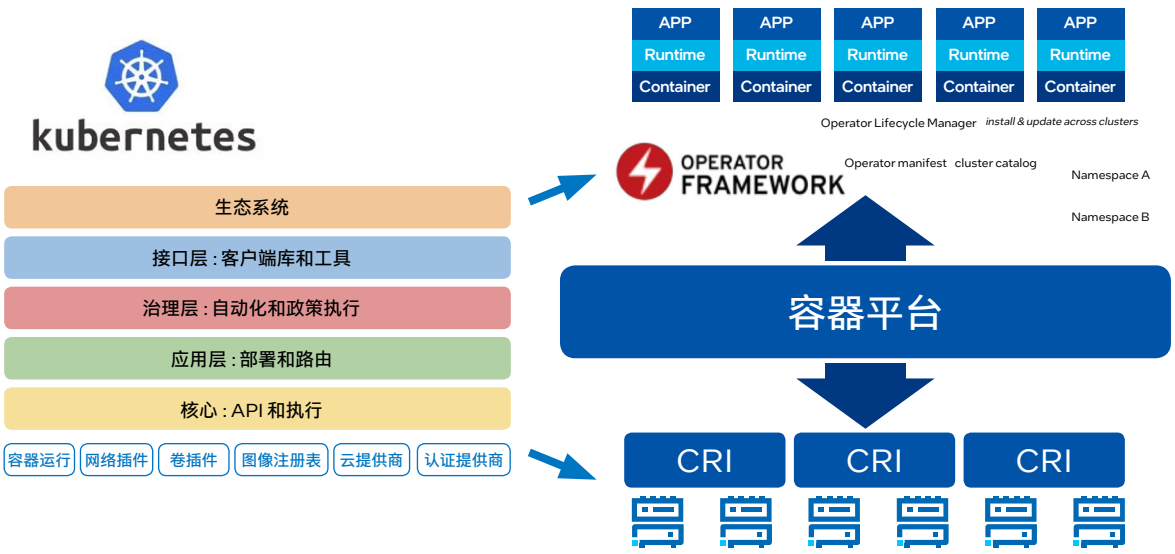
StarlingX

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

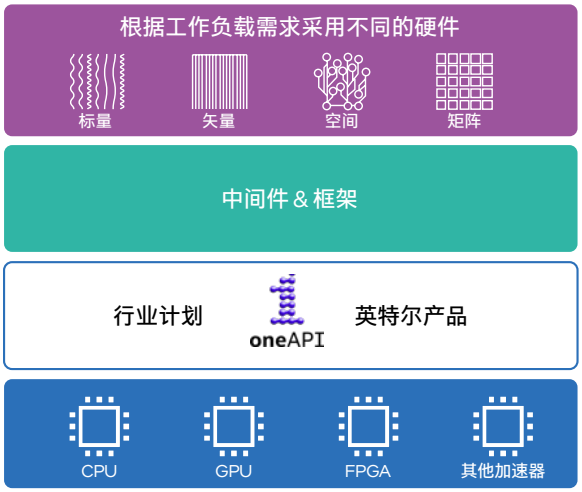
Kubernetes 是领先的容器编排解决方案。作为该项目的积极贡献者，英特尔使用多项数据中心关键技术，帮助其构建功能模块和全栈解决方案，比如：提供硬件设备插件、高级容器网络功能等技术，来解锁新的使用模式；逐层优化软件堆栈，确保最终用户获得底层硬件的全部优势；携手生态系统供应商合作，确保 Kubernetes 解决方案得以优化。



分析及 AI 性能优化

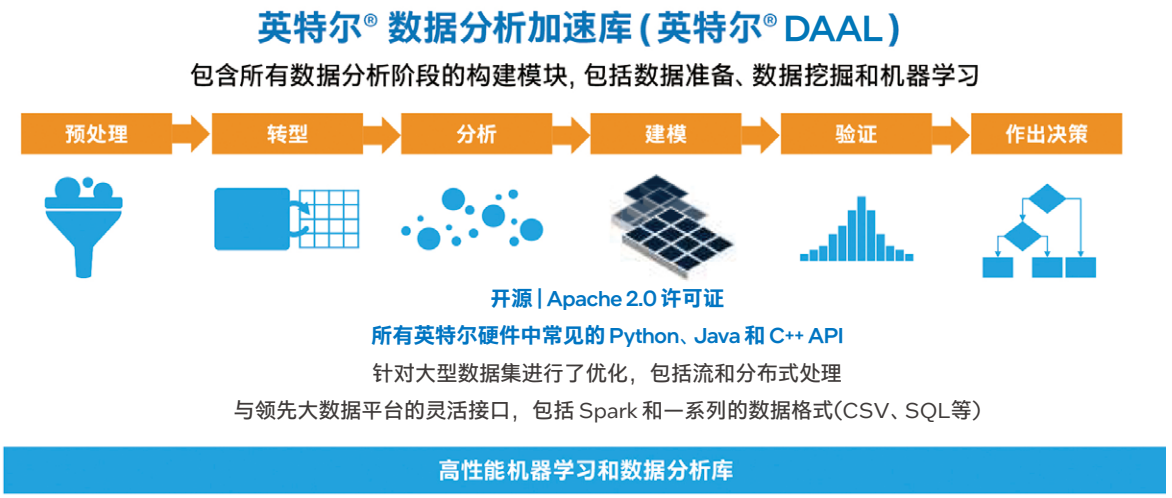
英特尔® oneAPI 工具套件

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。



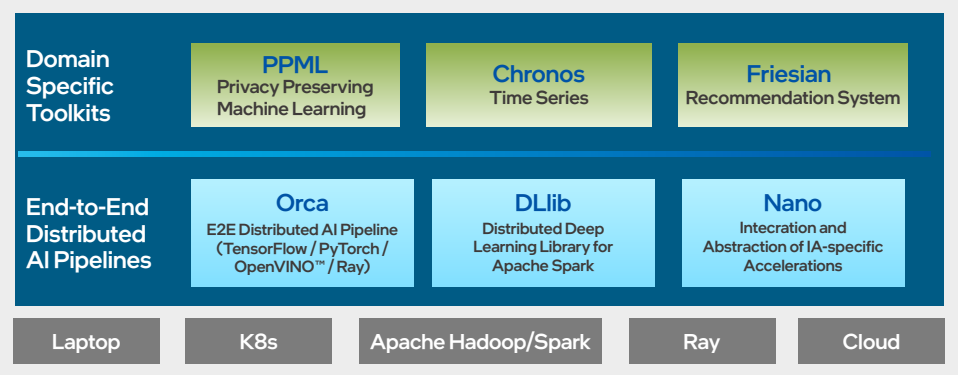
BigDL

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。

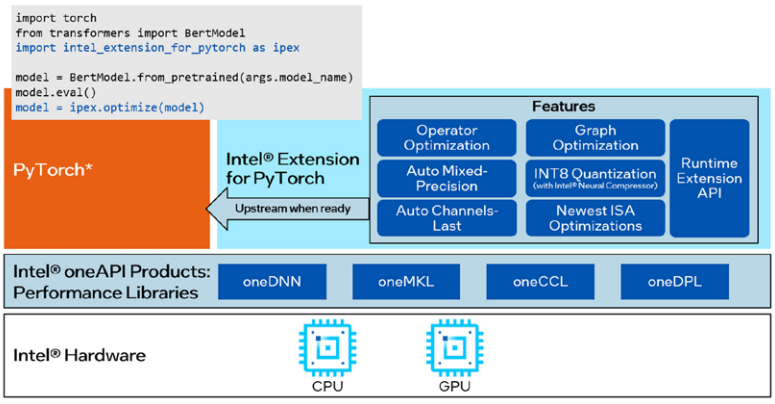
面向英特尔® 架构优化的 Caffe，集成了英特尔® 数学核心函数库，专门面向高级矢量扩展指令集英特尔® AVX2 和英特尔® AVX-512 做了优化，且具备更多处理器优化功能，展现了更优的性能，并支持多节点分布程序训练。

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。



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

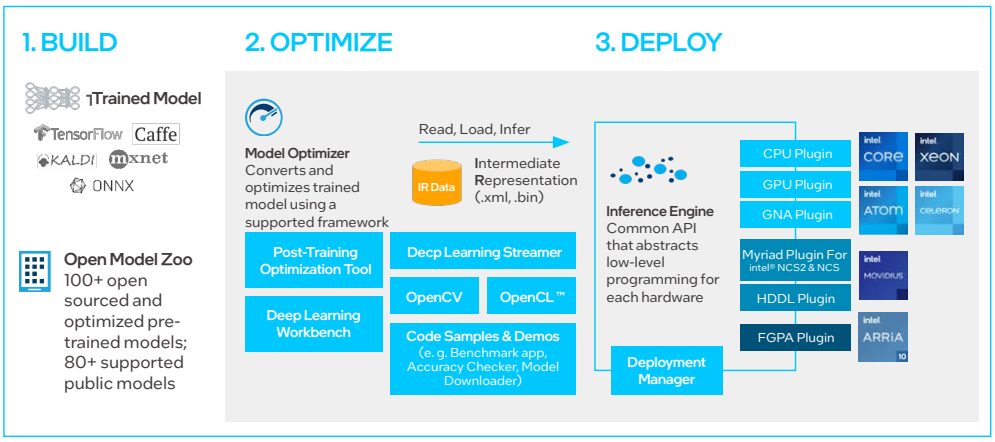
OpenVINO™ 工具套件

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

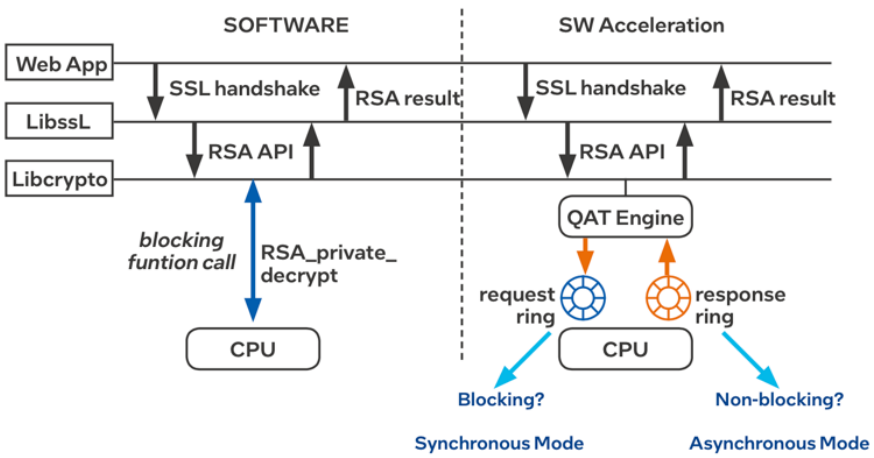
简化开发、易于使用

一次编写、随处部署



英特尔® Crypto-NI

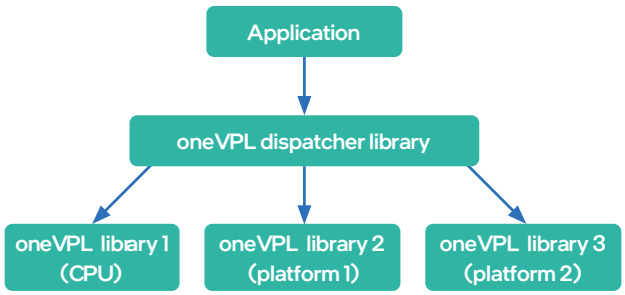
英特尔® Crypto-NI 是英特尔® 至强® 可扩展处理器中关于加解密领域的指令集，在之前英特尔® 至强® 可扩展处理器已具备的英特尔® AES-NI 指令集群上，又加入了 Vectorized AES、Integer Fused Multiply Add 等新指令。该方案使用的主要软件为 IPP Cryptography Library、Intel® Multi-Buffer Crypto for IPsec Library 和 QAT Engine。这些库基于新指令集提供了批量提交多个 SSL 请求的功能和并行异步处理机制，从而大幅提升性能。



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。

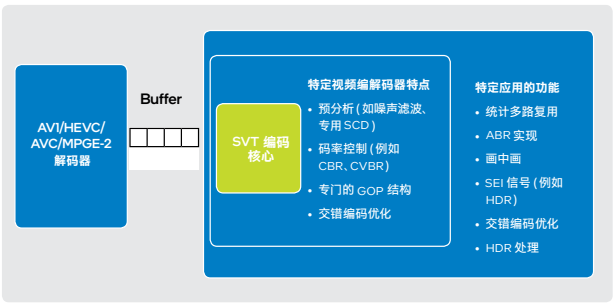


英特尔® oneVPL 的调度机制

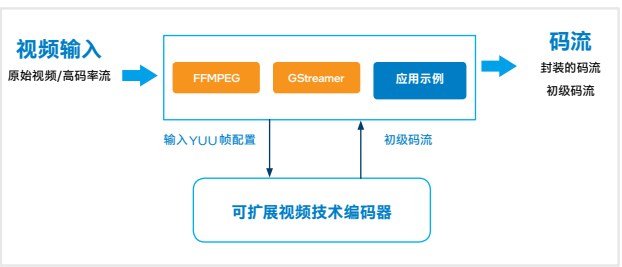
- 提供了与英特尔® Media SDK 核心 API 的兼容性
- 具备与英特尔® Media SDK相同的视频编解码器和滤波器
- 支持在通用处理器、集成显卡 GPU、独立显卡 GPU 以及其他硬件加速器中的部署
- 改进了视频处理初始化模式，可用于支持更广泛的视频处理实现方式
- 提供了新的内存抽象和优化方式，以及对解码性能的优化

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。

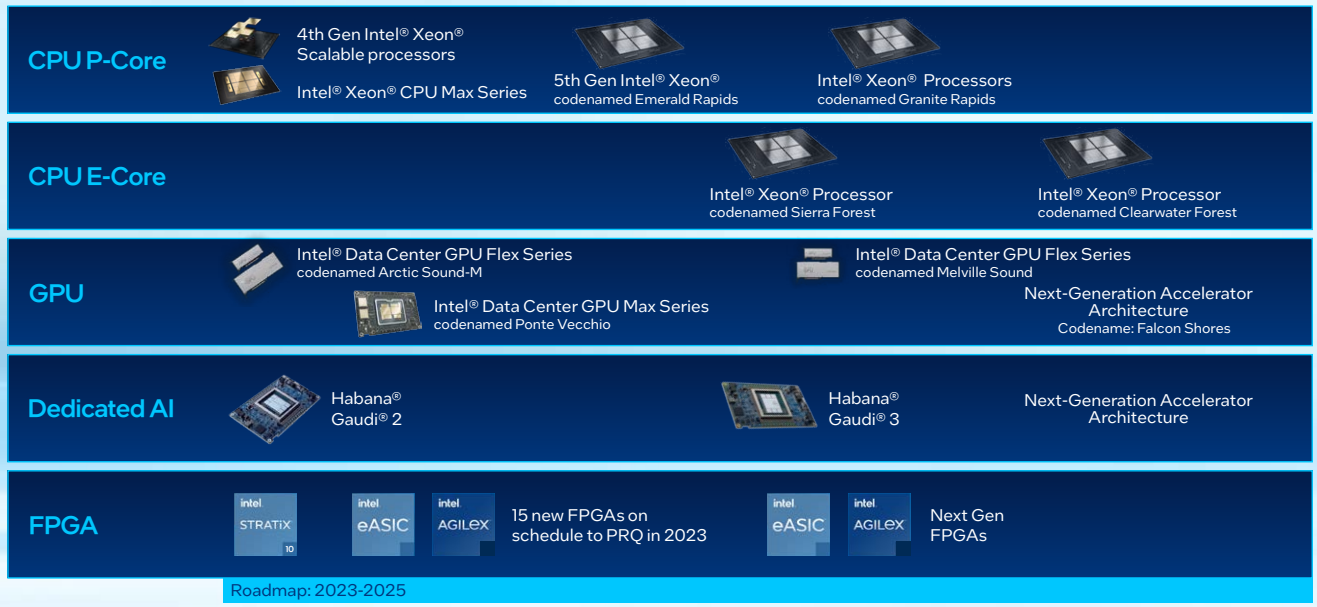


SVT 编码器核心和可视化编码器组件



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

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。