解决方案简介

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

金山云

挑战

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

- 如何加速数据清理、模型推理等人工智能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。
- 如何高效使用 CPU 等现有的硬件资源，并且利用客户公有云、私有云和混合云中的服务器资源，以降低硬件成本。
- 如何增强云服务器的灵活性，使其能够在复杂场景中敏捷扩展，支撑传统负载与人工智能等新型工作负载高效运行的需求。

解决方案概述

人工智能已经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，人工智能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，提升人工智能端到端流程的性能表现，从而充分释放硬件的潜能，并降低系统总体拥有成本（TCO），加速人工智能技术的创新。

为了帮助用户加速人工智能端到端流程，特别是提升人工智能推理性能，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® 高级矩阵扩展（英特尔® AMX）加速器，并融合了金山云自主创新的加速技术，能够有效提高人工智能模型的推理性能，同时发挥云服务器在敏捷性、扩展性等方面的优势，助力客户挖掘人工智能时代的价值。

金山云第七代性能保障型云服务器 X7

金山云第七代性能保障型云服务器 X7 搭载英特尔® 至强® 铂金 8458P 处理器，网络带宽升级至 100G，同时支持挂载极速云盘 ESSD，整体机型在计算、网络、存储多维度进行了深度优化，可为用户提供计算速度更快、网络吞吐更大以及存储更加高效的云服务。

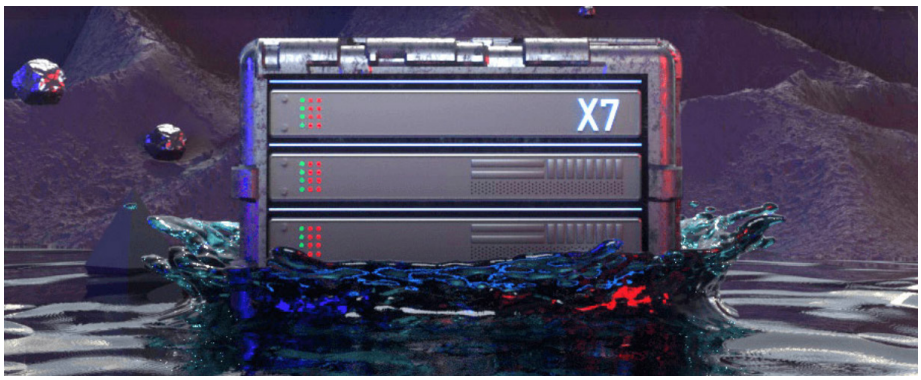

图 1. 金山云第七代性能保障型云服务器 X7

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

图 2. 第四代英特尔® 至强® 可扩展处理器

第四代英特尔® 至强® 可扩展处理器在人工智能性能上更进一步，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，进一步增强了前代技术 — 矢量神经网络指令（VNNI）和 BF16，从一维向量发展为二维矩阵，能够有效利用计算资源，提高高速缓存利用率，以及避免潜在的带宽瓶颈，从而可显着增加人工智能应用程序的每时钟指令数（IPC），为人工智能工作负载中的训练和推理带来显著的性能提升。

- 在计算方面，通过采用最新的第四代英特尔® 至强® 可扩展处理器，金山云新一代云服务器 X7 计算性能较上一代最大提升 60%[5]，同时借助内置的英特尔® AMX 原生人工智能加速能力，大幅提高了云服务器的整体性能，更加适用于计算密集型、深度学习等业务场景。
- 在内存方面，金山云新一代云服务器 X7 支持八通道 DDR5 内存，单条内存带宽高达 4800MT/s，对比上一代实例性能提升 50%[6]，更加适用于内存计算等数据密集型业务场景，服务深度学习以及人工智能相关领域。
- 在网络方面，金山云新一代云服务器 X7 的物理网络升级至 2 x 100G，单虚机内网吞吐最高提升至 100G，PPS 提升至最高 2400 万，连接数最高支持 400 万，网络性能大幅提升[7]。
- 在存储方面，金山云新一代云服务器 X7 支持挂载极速云盘 ESSD，单盘吞吐最高提升至 4GB/s，IOPS 提升至最高 100 万，访问延时降低至 0.2ms，存储能力显著优化[8]。

[1] 数据来源自第四代英特尔® 至强® 可扩展处理器的最大核数（60 核）与第三代英特尔® 至强® 可扩展处理器的最大核数（40 核）的比较。

[2] 详细配置信息请访问：intel.com/processorclaims，选择“第四代英特尔® 至强® 可扩展处理器”，查看编号“G2”。实际性能受使用情况、配置和其他因素的差异影响。

[3] 数据来源自第四代英特尔® 至强® 可扩展处理器（80 条 PCIe 5.0 通道）与第三代英特尔® 至强® 可扩展处理器（64 条 PCIe 4.0 通道）的比较。

[4] 详细配置信息请访问：intel.com/processorclaims，选择“第四代英特尔® 至强® 可扩展处理器”，查看编号“G1”。实际性能受使用情况、配置和其他因素的差异影响。

[5,6,7,8] https://www.ksyun.com/nv/activity/X7launch，截止 2023 年 6 月。

采用英特尔® AMX 优化人工智能推理性能

得益于第四代英特尔® 至强® 可扩展处理器内置的英特尔® AMX 技术，金山云新一代云服务器 X7 加速了人工智能推理性能，并在 AIGC 等负载中有着卓越的表现。

金山云测试了金山云新一代云服务器 X7 在 Stable-Diffusion 模型推理中的性能表现。Stable-Diffusion 是一种基于机器学习的生成式人工智能模型，能够根据文本生成高分辨率图像。Stable-Diffusion 一般需要数秒完成图片生成，计算量极大，其主要性能瓶颈在多头注意力计算部分（MHA）。

第四代英特尔® 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® 基于 PyTorch 优化的 Intel-Extension-for-PyTorch（IPEX）插件在 2.0 版本发布了基于至强® 可扩展处理器平台的 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。

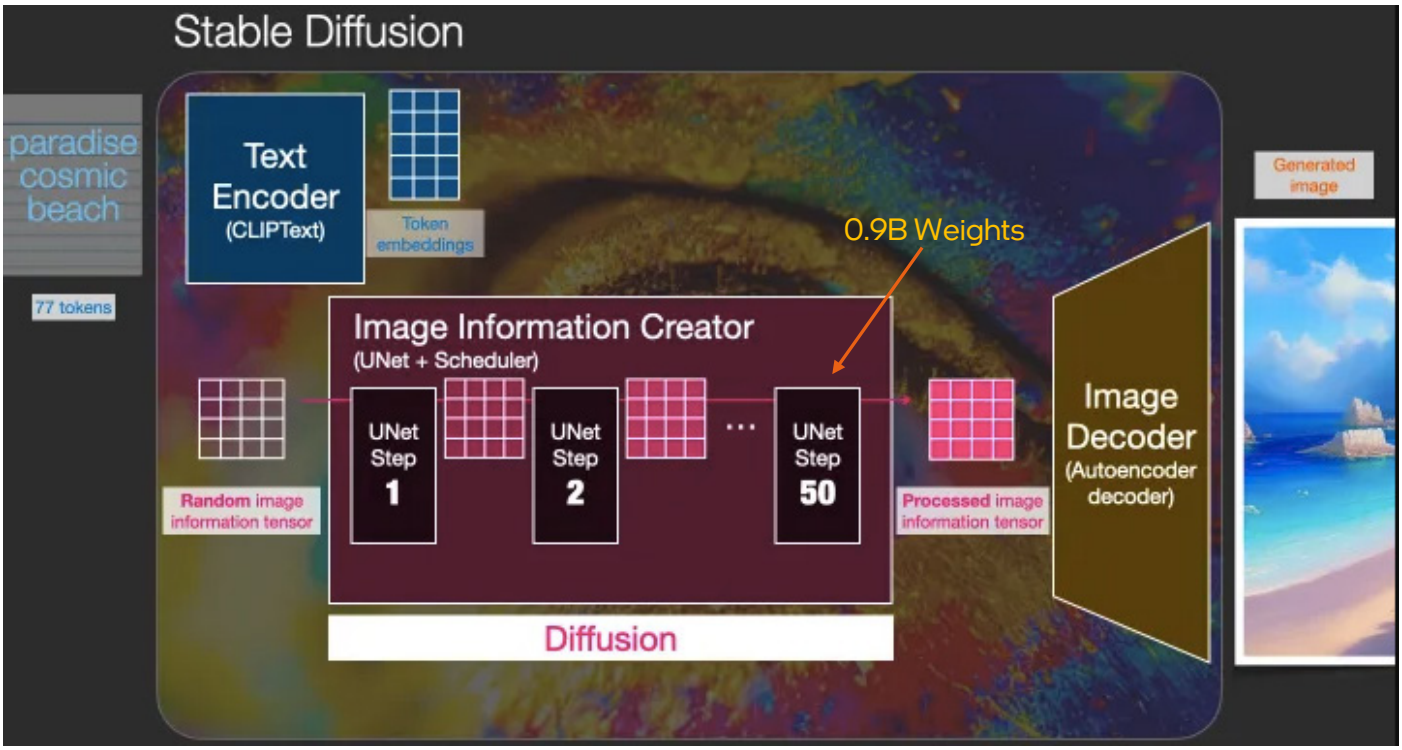

图 3. Stable-Diffusion 模型结构

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 4 所示，相较优化之前的模型，在使用 IPEX 2.0 BF16 优化之后，Stable-Diffusion 模型推理性能提升了 3.97 倍 - 4.96 倍[9]。

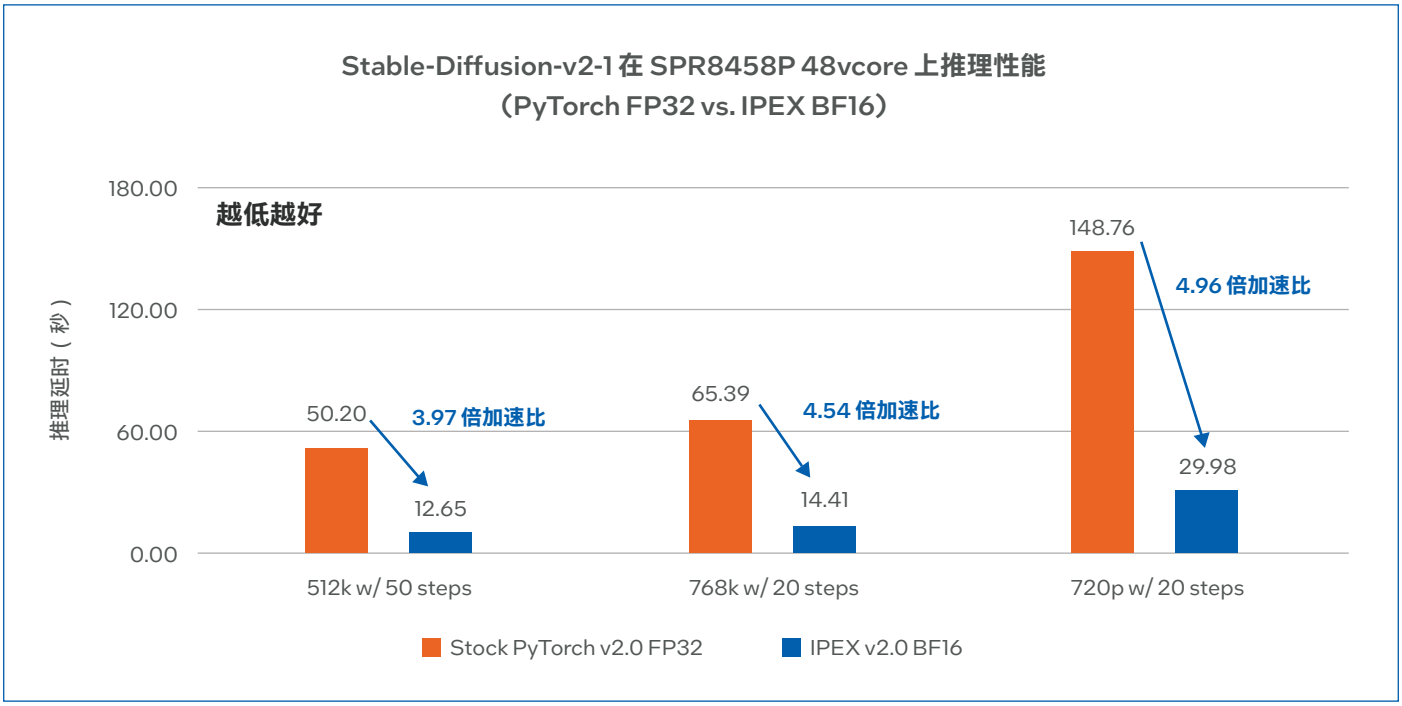

图 4. Stable-Diffusion 模型优化前后性能对比[10]

收益

通过自研技术的创新以及第四代英特尔® 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：

- 更高的性能，能够满足广泛实际应用场景的对于性能的需求。特别是在人工智能性能方面，金山云新一代云服务器 X7 能够有效加速 AIGC 等应用的运行。
- 通过英特尔® AMX 的应用以及算法优化，充分释放了硬件潜力，有效利用服务器资源，从而降低了端到端人工智能应用流程的 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。

[9,10] 数据援引自截止 2023 年 6 月金山云内部测试结果。测试配置：英特尔® 至强® 铂金 8458P 处理器，48vcore，HuggingFace stabilityai/stable-diffusion-2-1。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。

展望

云服务器已经成为用户扩展人工智能创新，承载模型训练、模型推理等应用需求的重要选择，通过采用内置英特尔® AMX 加速器的四代英特尔® 至强® 可扩展处理器，金山云第七代性能保障型云服务器 X7 能够显著加速 AIGC 等模型的性能表现，在端到端人工智能流程中的优势突出。而且，该方案不需要部署独立的加速器，因此在经济性方面有着更佳的表现。

在当前合作成果的基础上，英特尔与金山云还将对第七代性能保障型云服务器 X7 进行进一步合作优化，包括验证服务器在更多场景中的性能表现、通过软件与算法优化进一步释放硬件潜力等，进而为用户提供更加卓越的云服务。

关于金山云

金山云创立于 2012 年，目前稳居中国公有云互联网云服务商前三（市场排名数据来源：IDC《中国公有云服务市场（2020Q3）跟踪》报告），业务范围遍及全球多个国家和地区。依托金山集团 35 年的企业级服务经验，金山云坚持技术立业，逐步构建了完备的云计算基础架构和运营体系。通过与大数据、数据库、边缘计算、AR/VR 等前沿技术有机结合，金山云在所深耕的行业提供超过 150 种解决方案，已广泛用于互联网、金融、医疗、公共服务等领域，累计为 597 家优质客户提供高品质的云服务。

关于英特尔

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

intel®

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

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

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。

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

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。