白皮书

阿里巴巴电子商务推荐系统 第四代英特尔® 至强® 可扩展处理器 助力提升性能

第四代英特尔® 至强® 可扩展处理器

推荐系统（RM）是非常重要和普遍的人工智能（AI）应用，其通常包括知识库、主题模型、用户/视频画像、实时反馈/统计、推荐引擎等基础组件，能够对于海量数据进行分析，并通过实时特征工程、在线学习、多模型融合等技术进行智能排序，从而根据用户的偏好，为用户提供个性化的内容与服务，助力提升用户价值。

为了推动推荐系统的创新，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。阿里巴巴还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英特尔® 至强® 可扩展处理器，利用该处理器搭载的英特尔® 高级矩阵扩展（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：

- **如何满足 AI 推理在吞吐量与时延方面的要求**

  阿里巴巴核心 RM 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。

- **如何确保 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，但同时也会对于推理精度带来影响。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

面临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、总体拥有成本（TCO）等方面实现平衡。为此，阿里巴巴选择了第四代英特尔® 至强® 可扩展处理器进行性能优化。

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 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。第四代英特尔® 至强® 可扩展处理器提供了现代性能和安全性，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 AI、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。

第四代英特尔® 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 RM 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。

下图列出了使用具备核心 RM 模型主要特征的代理模型时，第四代英特尔® 至强® 可扩展处理器和第三代英特尔® 至强® 可扩展处理器的代际性能对比。

图 1 显示，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 核第四代英特尔® 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 核第三代英特尔® 至强® 可扩展处理器，同时将时延严格保持在 15 毫秒以下[1]。

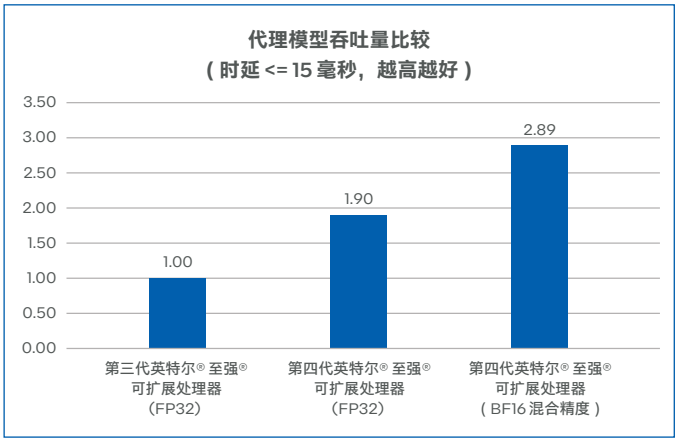

图 1. 代理模型的代际性能比较（时延 <= 15 毫秒）

收益

- 阿里巴巴能够在保证 RM 模型符合推理时延 <= 15 毫秒的同时，将推理的吞吐量提升到 2.89 倍[1]。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；
- 升级为第四代英特尔® 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，在利用第四代英特尔® 至强® 可扩展处理器集成的英特尔® AMX 等创新硬件特性，并进行软件优化之后，核心 RM 模型在性能上能够获得的巨大提升。除了 RM 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英特尔® 至强® 可扩展处理器，以释放该处理器的性能潜力。

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，这将有助于加快 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。

intel®

[1]截止到 2022 年 10 月 21 日的英特尔测试。单节点，双路第三代英特尔® 至强® 可扩展处理器，32 核，超线程启用，睿频启用，总内存 2048 GB（16x128 GB DDR4 3200 MT/秒），BIOS WLYDCRB1.SYS.0027.P80.2203310646，微代码 0xd000363，1 个 447.1G SSDSCKKB48，1 个 223.6G SSDSC2BB24，CentOS Linux 版本 8.4.2105，5.15.0-spr.bkc.pc.8.8.5.x86_64，GCC 8.5.0，TensorFlow 1.15up3，核心 RM 模型的阿里巴巴代理，1 个实例/插槽，单插槽，用户 BS=1，项目 BS=600，数据类型：FP32。**配置 2：**截止到 2022 年 10 月 21 日的英特尔测试。单节点，双路第四代英特尔® 至强® 可扩展处理器，48 核，超线程启用，睿频启用，总内存 1024 GB（16x64 GB DDR5 4800 MT/秒），BIOS EGSDCRB1.86B.0090.D03.2210040151，微代码 0xab000310，1 个 1.8T SSDSC2KG01，1 个 447.1G SSDSCKKB48，1 个 1.8T SSDPE2KX020T7，CentOS Linux 版本 8.4.2105，5.15.0-spr.bkc.pc.8.8.5.x86_64，GCC 8.5.0，TensorFlow 1.15up3，核心 RM 模型的阿里巴巴代理，1 个实例/插槽，单插槽，用户 BS=1，项目 BS=600，数据类型：FP32。**配置 3：**截止到 2022 年 10 月 21 日的英特尔测试。单节点，双路第四代英特尔® 至强® 可扩展处理器，48 核，超线程启用，睿频启用，总内存 1024 GB（16x64 GB DDR5 4800 MT/秒），BIOS EGSDCRB1.86B.0090.D03.2210040151，微代码 0xab000310，1 个 1.8T SSDSC2KG01，1 个 447.1G SSDSCKKB48，1 个 1.8T SSDPE2KX020T7，CentOS Linux 版本 8.4.2105，5.15.0-spr.bkc.pc.8.8.5.x86_64，GCC 8.5.0，TensorFlow 1.15up3，核心 RM 模型的阿里巴巴代理，1 个实例/插槽，单插槽，用户 BS=1，项目 BS=600，数据类型：FP32 和 BF16 混合精度。

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。