解决方案简介

英特尔® AMX 助力增强阿里云地址标准化 AI 推理性能

阿里云

intel XEON

第四代英特尔® 至强®可扩展处理器

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。
- 高效使用现有的硬件资源，并且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，以降低硬件成本。

解决方案概述

深度学习（DL）作为一项重要的人工智能（AI）技术，被广泛应用于多个领域，例如计算机视觉（CV）、自然语言处理（NLP）和推荐系统。但是，随着数据量的爆炸性增长以及 DL 模型复杂性的不断提高，推理正面临巨大的计算挑战。用户希望优化硬件、软件和算法，以提升性能，充分释放硬件的潜能，并降低总体系统成本。此外，优化 DL 算法还能帮助用户采用更复杂的 DL 模型，从而提高准确率，同时保持相同的时延。

为了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。基于第四代英特尔® 至强® 可扩展处理器的平台使用英特尔® 高级矩阵扩展（英特尔® AMX）优化端到端推理性能（比前代高出 2.5 倍）[1]，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[2]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。它使用 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。它提供 20 多种地址服务[3]，可灵活地部署在公有云、私有云或混合云上。

---

[1] 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。

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

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

采用第四代英特尔® 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了 BERT[4] 作为地址标准化搜索模块的核心模型。BERT 是一个广泛用于多任务向量召回和精排（Fine Ranking）的关键模型。为了优化 BERT 性能，阿里云 PAI 团队采用了第四代英特尔® 至强® 可扩展处理器集成的英特尔® AMX 等高级特性进行优化。

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 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，以及避免潜在的带宽瓶颈，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显著的性能提升。

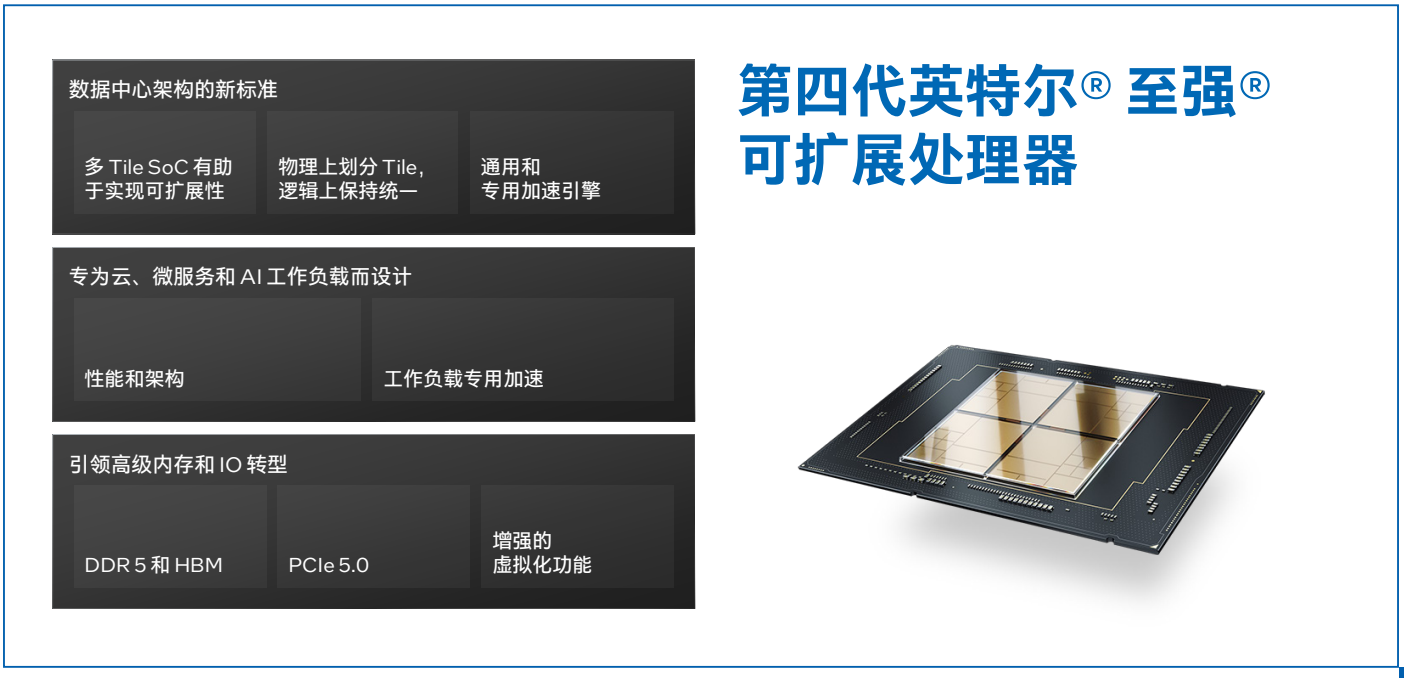

图 1. 第四代英特尔® 至强® 可扩展处理器

[4] 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）。

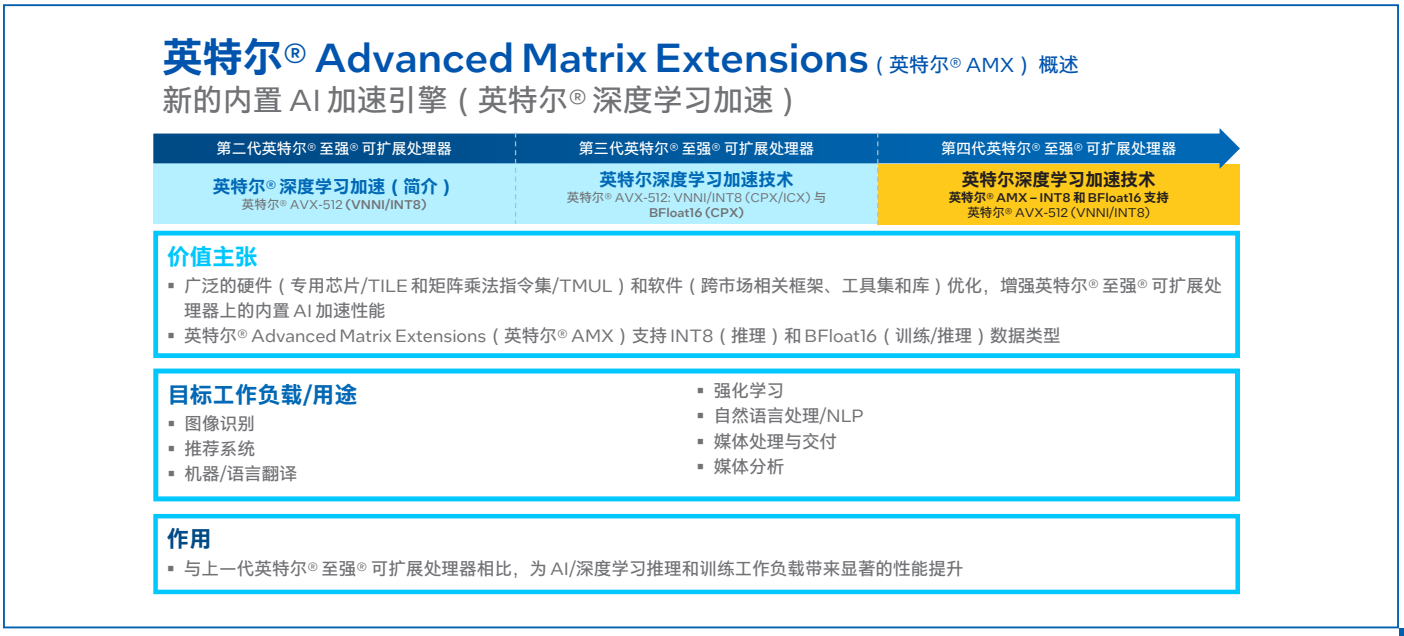

图 2. 英特尔® AMX 简介

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，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[5] 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 3. 英特尔® Custom Backend 结构

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

英特尔® AMX 极大地改进了 INT8 的功能，英特尔® Custom Backend 也可以利用英特尔® 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 在第四代英特尔® 至强® 可扩展处理器上实现了 2.5 倍的性能提升[6]。

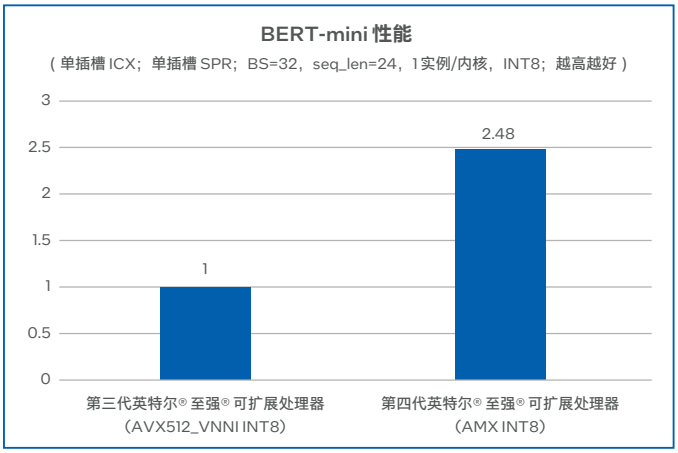

图 4. BERT 模型的推理性能[7]

在精度方面，基于 CCKS2021 中文 NLP 地址相关性任务[8] 的验证显示，基于 FP32 的优化的精度仍保持在 78.72，而基于 INT8 的优化的模型精度为 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

在阿里云地址标准化的 AI 推理优化过程中，使用英特尔® 深度学习加速充分释放第四代英特尔® 至强® 可扩展处理器在处理 AI 推理工作负载方面的巨大潜力，从而帮助阿里云显著提升端到端推理性能，并解决实际的业务问题。对于用户而言，该解决方案有助于降低部署专用加速器（例如独立显卡）时的开销，以及更有效地控制地址标准化的总体拥有成本（TCO）。

为了提升更多 DL 模型的端到端性能，英特尔和阿里云正在扩大他们与客户之间的协作，探索以创新方式优化软硬件集成，以加速 DL 模型的性能，最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。

[6,7] 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。

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

性能因使用情况、配置和其他因素而异。更多信息请访问 www.Intel.com/PerformanceIndex

性能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。

您的成本和结果可能会有所不同。

英特尔技术可能需要启用硬件、软件或服务激活。

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。