白皮书

英特尔® 至强® 可扩展处理器
英特尔® 高级矩阵扩展（英特尔® AMX）

基于英特尔® 至强® 可扩展处理器的新一代京东云服务器加速大模型推理 助力构建数智化供应链基础设施

概述

在 2023 年 11.11 促销活动中，京东成交额、订单量、用户数均创新高，超 60 个品牌销售破 10 亿元，近 2 万个品牌成交额同比增长超 3 倍，新商家成交单量环比增长超 5 倍[1]。在促销期间，京东云每秒用户访问峰值同比提升了 170%，京东云智能客服累计咨询服务量超 14 亿次……亮眼的成绩背后，离不开京东在大模型等创新技术方面的投入，这催生了卓越的智能营销、智能服务体系，为消费者带来了领先的服务体验。

为了给上层智能应用提供坚实的支撑，京东云研发并上线了基于第五代英特尔® 至强® 可扩展处理器的新一代京东云服务器，比上一代服务器的整体性能提升了 23%。除了整体性能的提升之外，京东云还与英特尔密切合作，采用处理器内置的英特尔® 高级矩阵扩展（英特尔® AMX）等高级硬件能力，以及英特尔® oneAPI，英特尔® 深度神经网络库（英特尔® oneDNN）等软件技术，提升人工智能（AI）推理等方面的性能表现，并在 AI 视觉模型 ResNeXt-50、大语言模型 Llama v2-13B 等模型推理中得到成功验证。

挑战

作为一家技术驱动的公司，京东致力于利用 AI 赋能智能商业、智能金融等业务，以创造全球领先的智能商业体。目前，AI 技术已经在数百个京东商业场景中得到成功应用，为商家与消费者带来了卓越的技术服务。京东高度重视大模型的创新以及在实际业务中的应用，例如，在智能营销方面，京东利用大模型赋能产业洞察与市场营销，利用人工智能内容生成（AIGC）技术实现商品内容生成，从而显著降低营销成本，提升营销效率；在智能服务方面，由大模型赋能的智能客服融汇了京东服务过程中产生的超亿级高质量对话数据，提升了多模态理解能力，能够满足大量场景的客服咨询需求。

大模型训练、模型微调、模型推理属于典型的算力密集型应用，基础设施的算力水平，对于模型质量与效率，起到了举足轻重的作用。为了给快速扩展的大模型应用提供强大的基础能力支撑，京东云搭建了面向大模型应用的数智平台，通过极致的技术创新，在各类基础云产品服务层面，提供了稳定的数智能力的支持，也实现了极致的成本优化，助力了以上数百个场景的 AI 应用升级。

模型推理是大模型应用的重要一环，京东云希望进一步优化推理服务器，以更好地应对大模型推理在性能、成本、灵活性等方面带来的挑战。

性能挑战

京东云拥有大量的大模型应用场景，以及海量的应用需求，意味着模型推理会带来巨大的算力开销。要提升模型推理性能，一方面可以采用具备更高算力的硬件基础设施，另一方面也可以利用将模型转化为低精度格式、采用加速指令集等方式，以更好地释放算力。

成本挑战

大模型推理会带来较高的成本支出，特别是当部署专用的模型服务器时。京东云希望在满足模型推理的性能需求时，能够尽可能地降低模型推理的单位成本，从而寻找更适合推理的算力选项。

灵活性挑战

专用的 AI 服务器在应用场景方面有较大的局限性，京东云希望能够提升基础设施的敏捷性，从而灵活地满足各种长尾化的模型推理需求，且支持在数据中心内部不同的工作负载间进行灵活切换。

[1] https://www.iimedia.cn/c400/96750.html

解决方案：基于第五代英特尔® 至强® 可扩展处理器的京东云服务器

新一代京东云服务器的设计经验，源于京东云大规模数据中心的建设与运维经验，并通过了京东集团内部多业务场景的应用打磨和对外服务的千锤百炼。新一代京东云服务器搭载了第五代英特尔® 至强® 可扩展处理器，该处理器凭借内置加速器实现单核性能提升，能够轻松应对要求严苛的工作负载。第五代英特尔® 至强® 可扩展处理器拥有更可靠的性能，更出色的能效。它在运行各种工作负载时均可实现显著的每瓦性能增益，在 AI、数据中心、网络和科学计算的性能和总体拥有成本（TCO）方面亦有更出色的表现。相较上一代产品，第五代英特尔® 至强® 可扩展处理器可在相同功耗范围内提供更高的算力和更快的内存。此外，它与上一代产品的软件和平台兼容，因此部署新系统时可大大减少测试和验证工作。

第五代英特尔® 至强® 可扩展处理器每个内核都具备 AI 加速功能，无需添加独立加速器，就可处理要求严苛的 AI 工作负载，包括对参数量多达 200 亿的模型进行推理和调优[2]。

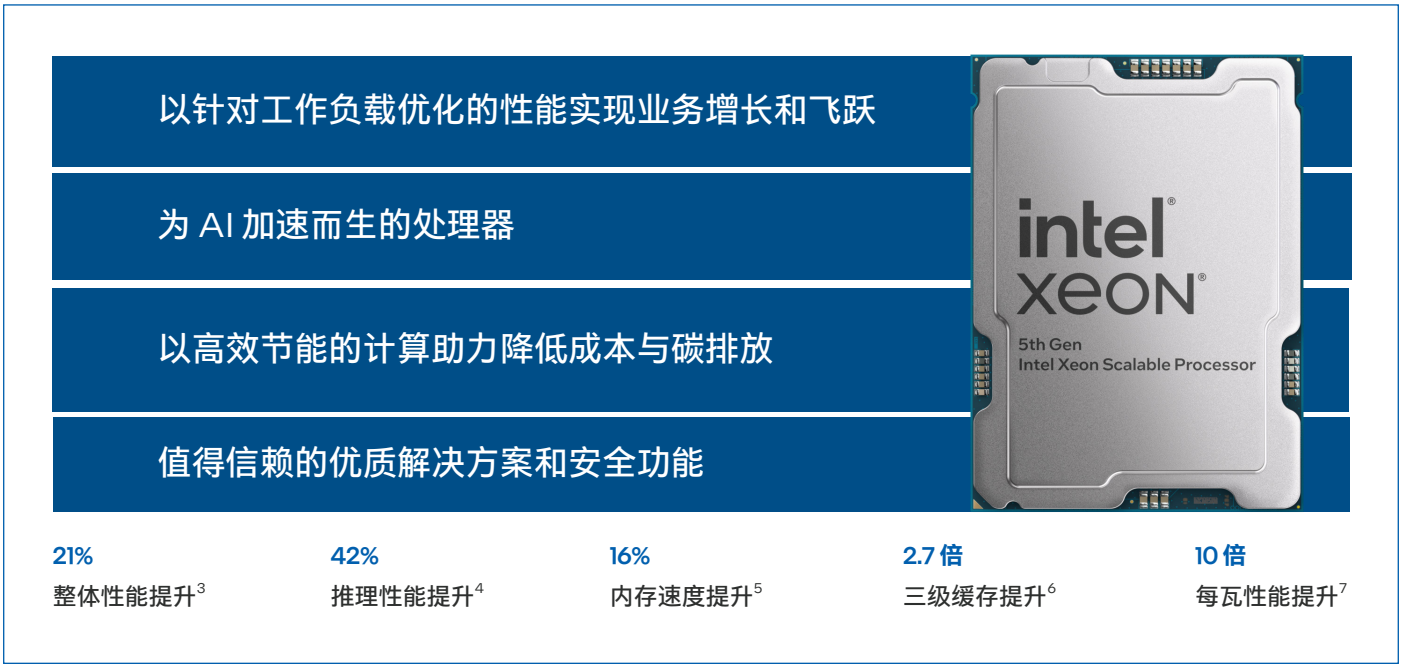

图 1. 第五代英特尔® 至强® 可扩展处理器具备更强大性能

为了进一步提升新一代京东云服务器在模型推理等负载中的性能表现，京东云与英特尔深度利用英特尔® AMX 以及英特尔® oneAPI、英特尔® oneDNN 来进行性能优化。英特尔® AMX 作为英特尔® 至强® 可扩展处理器内置的加速器，可加速基于 CPU 的深度学习推理，避免了使用独立加速器带来的成本和复杂性。英特尔® AMX 在迁移学习和再训练方面同样出色，用户无需额外添置硬件即可使模型保持最新状态。

英特尔® AMX 引入了一种用于矩阵处理的新框架（包括了两个新的组件，一个二维寄存器文件，其中包含称为 “tile” 的寄存器，以及一组能在这些 tile 上操作的加速器），从而能高效地处理各类 AI 任务所需的大量矩阵乘法运算，提升其在训练和推理时的工作效能。例如在向量检索的过程中，如存在 n 个 batch 任务，进行相似度计算时就需要对 n 个输入向量 x 和 n 个数据库中向量 y 进行比对，这其中的距离计算会产生大量的矩阵乘法，而英特尔® AMX 能够针对这一场景实现有效加速。

[2] 基于截至 2023 年 12 月英特尔的内部建模。

[3] 与第四代英特尔® 至强® 处理器相比的平均性能提升，以 SPEC CPU rate、STREAM Triad 和 LINPACK 的几何平均值为衡量标准。请参阅 intel.com/processorclaims 上的 [G1]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[4] 与第四代英特尔® 至强® 处理器相比，取得 1.19 倍到 1.42 倍的性能提升（ResNet50v1.5、BERT-Large、SSD-ResNet34、RNN-T（仅 BF16）、Resnext101 32x16d、MaskRCNN（仅 BF16）、DistilBERT）。请参阅 intel.com/processorclaims 上的 [A15-A16]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[5] 请参阅 intel.com/processorclaims 上的 [G12]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[6] 请参阅 intel.com/processorclaims 上的 [G11]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

[7] 使用内置加速器在 AI、数据和网络工作负载上进行测量，取得 1.46 到 10.6 倍的每瓦性能提升。请参阅 intel.com/processorclaims 上的 [A19-A25]、[D1]、[D2]、[D5] 和 [N16]：第五代英特尔® 至强® 可扩展处理器。结果可能有所差异。

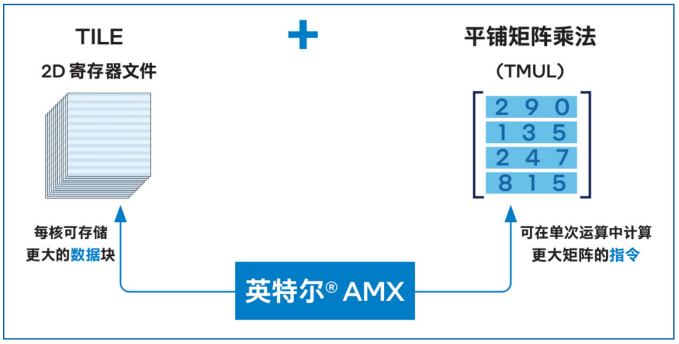

图 2. 英特尔® AMX 架构

英特尔® oneDNN 则提供了深度学习构建块的高度优化实现。借助这一开源、跨平台的库，深度学习应用程序和框架开发人员可以对 CPU、GPU 或两者使用相同的 API，从而抽象出指令集和其他复杂的性能优化。在模型性能优化中，通过使用英特尔® oneDNN，操作者只需要调用包含一些后期操作的 MatMul 基元并传递几个参数，oneDNN 即可完成其余的工作，例如配置块寄存器文件、从内存加载数据、使用后期操作执行矩阵乘法计算、将结果存储回内存中，最后释放块寄存器文件。通过使用英特尔® oneDNN，降低了工程师开发的难度和工作量，提高编程效率。

通过上述优化措施，京东云能够在大模型推理（Llama2-13B 及以下的模型）、Stable Diffusion、Vector Search、小型深度学习模型推理、数据分析/机器学习等场景里提供满足客户性能和服务质量（QoS）需求的解决方案。

性能验证：实现显著的代际性能提升

为了验证在典型的模型推理负载中，基于第五代英特尔® 至强® 可扩展处理器的新一代京东云服务器带来的代际性能提升，京东云进行了性能测试。

Llama2-13B

LLaMa2 是 Meta 发布的免费可商用版本的大模型，Llama 2 模型系列包含 70 亿、130 亿和 700 亿三种参数变体。LLaMa2 相比第一代在预训练语料库大小上增加了 40%，Llama 2 接受了 2 万亿个 Token 的训练，精调 Chat 模型在 100 万人类标记数据上训练，上下文长度是第一代的两倍，并采用了分组查询注意力机制等优化结构。

Llama2-13B 推理性能测试数据如图 3 所示，对比京东与英特尔联合定制优化的第四代英特尔® 至强® 可扩展处理器，京东与英特尔联合定制优化的第五代英特尔® 至强® 可扩展处理器的推理性能（Token 生成速度）提升了 51%[8]，可用于问答、客服和文档总结等多种 AI 场景。对于更高参数模型甚至是 70B Llama2，第五代英特尔® 至强® 可扩展处理器仍可胜任。京东深入优化的通用算力更易获得并具有弹性伸缩等优势，可为不同 AI 场景应用提供更多选择。

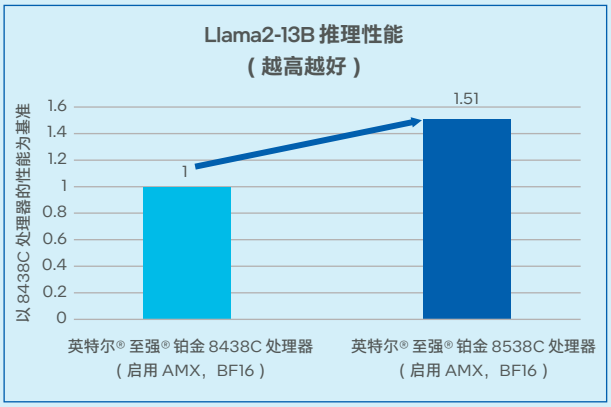

图 3. Llama2-13B 推理性能测试数据

[8] 京东截止至 2024 年 2 月的内部测试结果。测试配置：基准配置 — 双路英特尔® 至强® 铂金 8438C 处理器 @ 2.6 GHz，768 GB 总内存 (24x32 GB DDR5 4800 MT/s)，CentOS Stream 9；新配置 — 双路英特尔® 至强® 铂金 8538C 处理器 @ 2.6 GHz，512 GB 总内存 (16x32 GB DDR5 4800 MT/s)，CentOS Stream 9。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。

SE-ResNeXt-50

ResNeXt 是一种卷积神经网络（CNN）架构，被广泛用于处理视觉和文本数据的深度学习问题中，该模型在大数据集上表现出色，并且也是迁移学习应用的合适选择。SE_ResNeXt 是将 Squeeze-and-Excitation（SE）模块应用在 resnext 中的 residual block 上得到的模型，京东将其用于实现 AI 图像分类，支撑零售、安防、5G 多媒体、娱乐等多项应用。

SE-ResNext-50 推理性能的测试数据如图 4 和图 5 所示，通过英特尔® AMX 将模型从 FP32 转化为 BF16，第五代英特尔® 至强® 可扩展处理器可带来超过 4 倍的性能提升；而在相同的数据精度下，第五代英特尔® 至强® 可扩展处理器的代际性能提升达到 1.38 倍[9]。

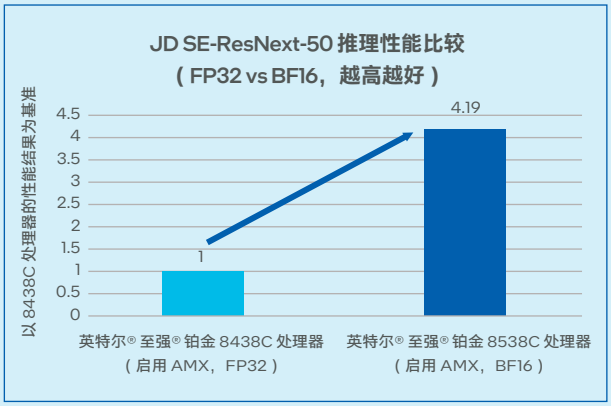

图 4. JD SE-ResNext-50 推理性能（FP32 VS BF16）

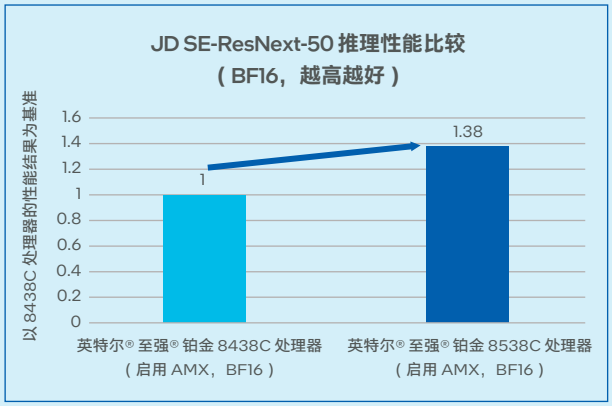

图 5. JD SE-ResNext-50 推理性能代际比较（BF16）

收益

通过硬件升级以及软件优化，基于第五代英特尔® 至强® 可扩展处理器的京东云服务器能够显著提升在模型推理等负载中的性能表现，从而带来以下收益：

- **加速模型推理等应用运行，提升 QoS：** 第五代英特尔® 至强® 可扩展处理器不仅能够在大模型推理、小型深度学习模型推理等场景中，提供强大的性能，而且可有效降低推理延迟，保证良好的用户体验。
- **降低模型推理的成本与门槛：** 基于第五代英特尔® 至强® 可扩展处理器的模型推理方案避免了在采购专用硬件加速器方面的高昂支出，而且在部署、运维等方面有更高的成本优势，有助于提升投资收益。
- **提升应对多种负载的灵活性：** 该方案建立在通用服务器的基础上，除了可用于模型推理之外，还能灵活满足数据分析、机器学习等应用的需求。

[9] 京东截止至 2024 年 2 月的内部测试结果。测试配置：基准配置 — 双路英特尔® 至强® 铂金 8438C 处理器 @ 2.6 GHz，768 GB 总内存（24x32 GB DDR5 4800 MT/s），CentOS Stream 9；新配置 — 双路英特尔® 至强® 铂金 8538C 处理器 @ 2.6 GHz，512 GB 总内存（16x32 GB DDR5 4800 MT/s），CentOS Stream 9。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

算力是大模型等 AI 应用实现价值跃升的重要支点，但在大模型落地过程中，面临算力需求指数级增长、算力成本高、异构复杂度高等挑战。在此背景下，采用涵盖通用服务器在内的多元异构 AI 算力系统，灵活地应对不同 AI 任务成为关键。基于第五代英特尔® 至强® 可扩展处理器的新一代京东云服务器则证明，通用服务器不仅在模型推理性能上有着卓越的表现，而且提供了更高的敏捷性，能够成为用户搭建 AI 算力系统的理想之选。

面向数智时代快速发展的 AI 应用需求，京东云与英特尔将在云基础设施的构建与优化方面携手进行更深入的合作，这不仅包括采用新一代硬件提升基础设施算力，也包括在电源设计、冷却技术等方面进行协同创新，从而全面优化云数据中心在算力、能效、成本等方面的表现，助力数智化变革。

关于京东

京东集团定位于“以供应链为基础的技术与服务企业”，目前业务已涉及零售、数字科技、物流、技术服务、健康、保险、产发、智联云和海外等领域，其中核心业务为零售、数字科技、物流、技术服务四大板块。京东集团奉行客户为先、诚信、协作、感恩、拼搏、担当的价值观，以“技术为本，致力于更高效和可持续的世界”为使命，目标是成为全球最值得信赖的企业。

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 京东云