英特尔中国

公有云和互联网创新实践

CONTENT
目录

Large Language Model (LLM)

大语言模型

CPU 也能玩转 AI - 为 AI 提速，给安全加码

阿里云

挑战

- 算力需求激增：视频、数据库等场景算力需求激增
- 智能化应用普及：大模型推理需求爆炸式增长
- 数据安全挑战：数据隐私及安全需求增强

解决方案

阿里云引入第五代至强® 可扩展处理器，实现 ECS g8i 算力再升级，为大模型 AI 推理加速添新解，更易得、更易用、可扩展性强，满足从小模型到超大模型的各类需求。

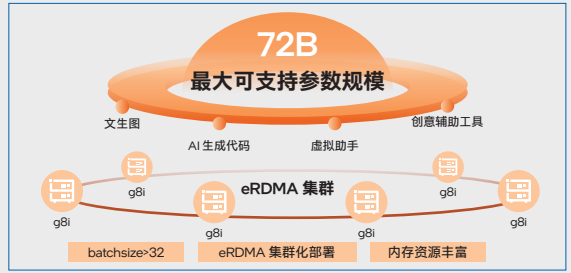
- 使用处理器内置的 AI 加速引擎 -- 英特尔® AMX 和英特尔® AVX-512，提升并行计算和浮点运算能力；
- 受益于第五代至强® 可扩展处理器显著提升的内存带宽和三级缓存共享容量，化解 AI 大模型吞吐性能挑战；
- 利用第五代至强® 可扩展处理器内置的英特尔® SGX 和英特尔® TDX 安全引擎，实现端到端的数据全流程保护。

85%
整机性能提升高达[1]

7 倍
AI 推理性能提升高达[2]

50%
中小参数模型起建成本降低[3]

第五代英特尔® 至强® 可扩展处理器具备更强通用计算和 AI 加速能力



阿里云 ECS g8i 集群可支撑 72B 参数级别的大语言模型分布式推理

[1,2,3] 数据来源于阿里云未公开的内部测试，如欲了解更多详情，请联系阿里云：https://www.aliyun.com
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。

让更加可及、经济的 AI 算力资源，在千行百业扬“千帆”

解决方案

大模型推理优化

AI 服务平台

百度智能云

挑战

- LLM 推理中大量矩阵及向量矩阵乘法对硬件的较高需求
- 满足行业离线 LLM 应用需求，并支持用户快速部署 LLM
- 解决 30B 等规模的 LLM 使用高端 GPU 成本较高等问题

解决方案

千帆大模型平台利用百度智能云平台中丰富的英特尔® 至强® 可扩展处理器资源，加速 LLM 模型推理，满足 LLM 模型实际部署需求。

- 基于至强® 可扩展处理器不断提升的算力和内存带宽，有效支持 LLM 实现端到端加速；
- 采用第四代 / 第五代至强® 可扩展处理器内置的 AI 加速引擎 – 英特尔® AMX，最大限度地利用计算资源，显著增加 AI 应用程序的每时钟指令数 (IPC)；
- 利用大模型推理软件解决方案 xFasterTransformer(xFT)，进一步加速 LLM 推理。

2.32 倍

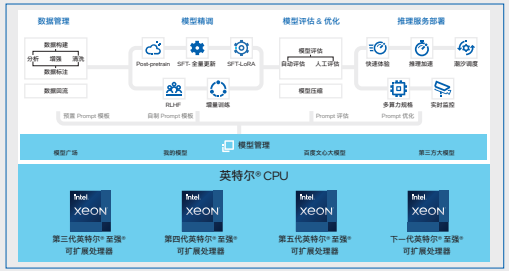
相较于第三代至强® 可扩展处理器，基于第五代至强® 可扩展处理器的 Llama-2-7b 模型输出 Token 吞吐提升达[1]

75%

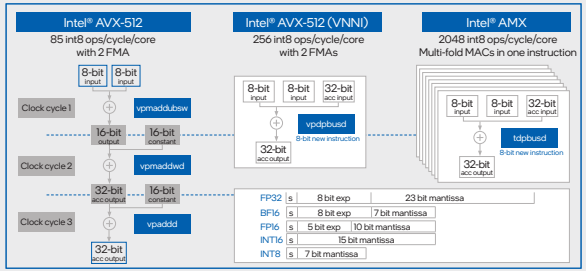
相较于第三代至强® 可扩展处理器，基于第五代至强® 可扩展处理器的 Llama-2-7b 模型首 Token 时延降低达[2]

利用充足的 CPU 资源，降低 LLM 推理服务 TCO

扫码获取全文



百度智能云千帆大模型平台可支持广泛的英特尔® CPU 选择



英特尔® AMX 可以更高效地实现 AI 加速

[1,2] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/artificial-intelligence/baidu-ai-cloud-accelerates-llm.html

用 CPU 打造智行云网大脑，网络大模型直面算力拦路虎

解决方案

大模型推理优化

网络大模型

中国电信 CHINA TELECOM

挑战

- 网络大模型在执行云网运营等应用时，需承受巨大的并发推理压力和性能要求
- GPU 算力方案会带来巨大的成本压力和能耗，且不利于 LLM 大规模应用
- 网络大模型运行过程对内存容量有较大需求，而 GPU 方案往往缺乏足够的内存容量

解决方案

中国电信网络大模型方案引入第五代至强® 可扩展处理器，借助其内置的多种 AI 加速引擎，结合英特尔开源的 xFT 分布式推理框架，有效平衡大模型推理的性能和成本。

- 采用第五代英特尔® 至强® 可扩展处理器作为方案的算力核心，利用其更多的内核数量、更强的单核性能和更大的三级缓存容量等，为大模型提供强劲算力支持；
- 利用英特尔® AMX 对 INT8 和 BF16 低精度数据类型的支持，在矩阵运算中有效提高计算速度并减少存储空间占用，更充分地利用计算资源，大幅提升网络大模型推理效能；
- 采用英特尔 AI 软件工具（如 xFasterTransformer）提升推理性能、降低部署成本并便捷地迁移模型。

< 100 毫秒

新方案已在多个运维场景推理任务中运用，辅助生成时延可有效满足业务响应时间要求[1]

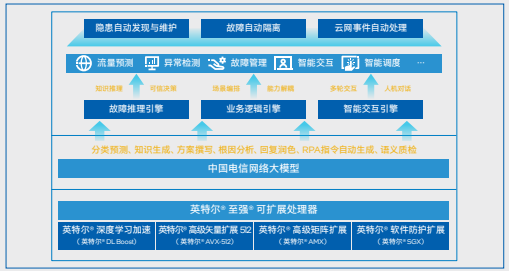
提升 10%

新方案使得运维效率有效提升，准备在中国电信现网各省公司实现规模落地[2]

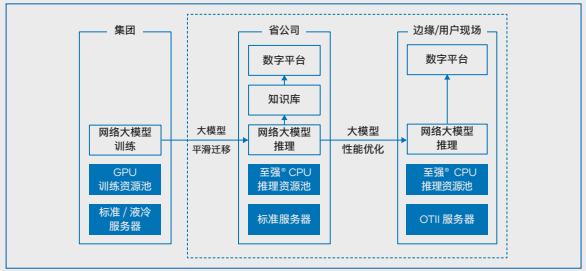
>40%

与主流 GPU 相比，CPU 平台方案可节省算力资源池建设成本[3]

扫码获取全文



基于 CPU 平台的中国电信网络大模型推理算力方案架构



面向边缘 / 用户现场的中国电信网络大模型推理部署

[1,2,3] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/telecom-actively-research-network-llms.html

* 荣获第二届“华彩杯”算力大赛 2024 年全国总决赛一等奖

看如何用 CPU 加速的 AI 大模型构建数智化供应链

- 解决方案
- 云服务器升级
- 大模型推理调优

挑战

- 巨大算力开销带来的性能挑战
- 专用模型服务器带来的成本挑战
- 专用 AI 服务器带来的灵活性挑战

解决方案

京东云推出搭载第五代至强® 可扩展处理器的新一代云服务器，以处理器内置 AI 引擎显著加速多种云上大模型推理，有效支撑 11.11 促销运行高峰。

- 利用第五代至强® 可扩展处理器及其内置的 AI 加速引擎 -- 英特尔® AMX，在提升算力的同时，高效处理大量矩阵乘法运算，提升 AI 推理性能；
- 使用英特尔® oneDNN 对 CPU、GPU 或两者使用相同的 API，抽象出指令集的其他复杂的性能优化，实现深度学习构建块的高度优化。

4.19 倍

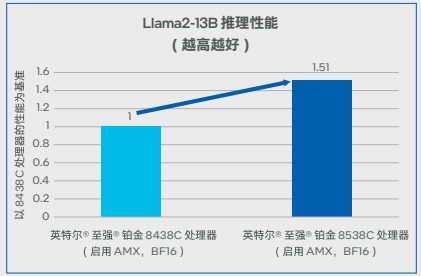
基于第五代至强® 可扩展处理器，通过英特尔® AMX 将模型转化为 BF16，JD SE-ResNext-50 推理性能提升高达[1]

51%+

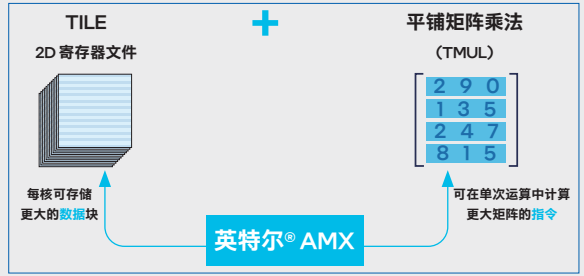
京东与英特尔联合定制优化的第五代至强® 可扩展处理器较上一代的推理性能（Token 生成速度）提升了[2]

避免采购专用硬件加速器的高昂支出

扫码获取全文



Llama2-13B 推理性能测试数据[3]



英特尔® AMX 架构

[1,2,3] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/artificial-intelligence/the-new-generation-of-jd-cloud-servers.html

中小模型推理新选择！算力性能倍增，实例全新升级

挑战

- 提供更加稳定可靠、弹性灵活、性能优越的云实例
- 不断拉升云实例算力性能，满足变化迅速的业务需求
- 支持内置 AI 加速的算力需求，兼顾速度与成本

解决方案

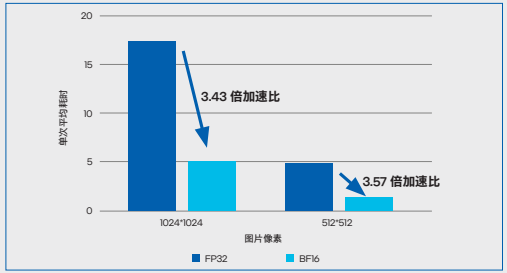
火山引擎第三代弹性计算云服务器实例 g3i 引入第五代至强® 可扩展处理器进行全新升级，通用性能与应用场景性能均大幅提升，可有力胜任高达 80 亿参数的模型推理，并兼顾速度与成本。

- 基于火山引擎最新自研 DPU2.0 架构和第五代至强® 可扩展处理器显著提升的代际性能、更高的 CPU 核心数、更快的内存以及更大的末级缓存容量，显著提升 g3i 算力性能，实现内存扩容，有效为 LLM 与更多场景提供支撑；
- 利用第五代至强® 可扩展处理器及其内置的 AI 加速引擎 -- 英特尔® AMX，在提升算力的同时，高效处理大量矩阵乘法运算，提升 AI 推理性能，胜任 80 亿参数模型推理，降低中小模型推理成本。

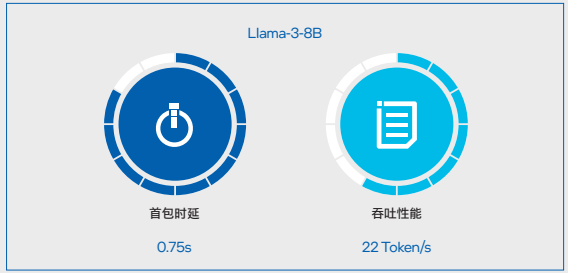
122%
火山引擎 g3i 整机算力提升[1]

75%
火山引擎 g3i 内存带宽提升[2]

3.43 倍
在 1,024*1,024 分辨率下，使用英特尔® AMX 将数据转换为 BF16，SDXL-Turbo 文生图推理可实现加速比[3]



SDXL-Turbo 文生图推理性能[4]



火山引擎 g3i 可胜任 80 亿参数的模型推理[5]

[1,2,3,4,5] 数据来源于火山引擎未公开的内部测试，如欲了解更多详情，请联系火山引擎：https://www.volcengine.com
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。

实现云端“算力 + 模型”一站式部署

挑战

- 采用高性能、高经济性的模型推理算力基础
- 利用创新的 AI 硬件加速策略，提升模型推理灵活性
- 以经过调优的模型镜像，充分调用硬件加速能力

解决方案

金山云推出搭载第四代至强® 可扩展处理器的新一代云服务器，以针对性调优的模型镜像，充分利用原生 AI 加速能力，有效提升云上大模型推理性能。

- 利用第四代至强® 可扩展处理器提供的强劲底层算力支撑，及内置 AI 加速引擎 -- 英特尔® AMX，以矩阵运算显著提升 AI 推理性能；
- 基于英特尔® MKL 及英特尔® oneDNN 搭建大模型镜像，在满足计算准确率的前提下，进一步提升模型性能。

3.97-4.96 倍

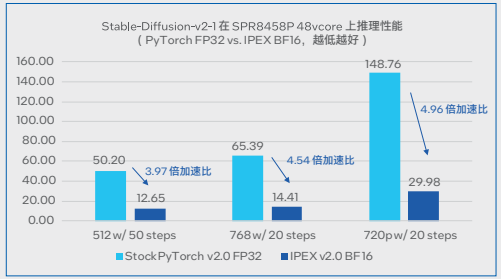
采用 IPEX 2.0 BF16 优化后，
Stable Diffusion 模型推理性能提升达[1]

2.52-2.62 倍

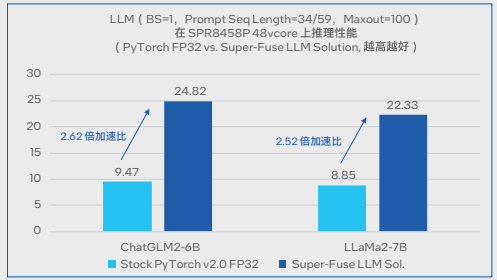
在 LLaMa2-AMX 和 ChatGLM2-AMX 性能测试中，
经英特尔 Super-Fuse 优化后，LLM 推理性能提升达[2]

灵活满足各种应用负载所需

扫码获取全文



Stable-Diffusion 模型优化前后性能对比[3]



大语言模型优化前后性能对比[4]

[1,2,3,4] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/kingsoft-cloud-accelerates-large-model-inference.html

Traditional Deep Learning

传统深度学习

激发硬件 AI 加速潜能，让每一份推荐都“算数”

案例研究

AI 模型推理加速

智能推荐系统

阿里妈妈
让每一份经营都算数

挑战

更复杂的模型结构和更丰富的组合特征，不断提升对硬件基础设施的算力需求

在有限的算力资源和严格的时延约束下，充分发挥硬件效能，持续提供强劲算力和 AI 加速

解决方案

阿里妈妈引入第五代至强® 可扩展处理器作为算力核心，并借助处理器内置的英特尔® AMX 及软件方案，为新方案提供面向 AI 推理的优化加速，为平台带来更优的推荐效果。

- 利用第五代至强® 可扩展处理器更大的末级缓存容量等为推荐系统提供强劲的算力支持；
- 英特尔® AMX 可提供矩阵类型的运算且同时支持 INT8 和 BF16 数据类型，助力阿里妈妈推荐系统在保证精度影响最小的前提下加速推理过程；
- 借助英特尔® oneDNN、算子融合等软件方案，加速矩阵运算，提升内存访问效率。

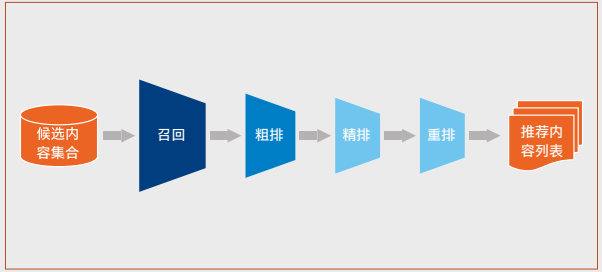
1.52 倍

基于第五代至强® 可扩展处理器的广告推荐模型，经过英特尔® AMX 和英特尔® AVX-512 优化后，相较上一代吞吐性能提升达[1]

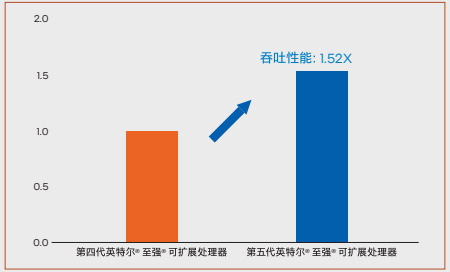
提升智能推荐系统准确性和效率

更精准的商品匹配和信息推荐

扫码获取全文



典型的阿里妈妈推荐系统架构



第五代至强® 可扩展处理器带来的吞吐性能提升[2]

[1,2] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/alimama-recommendation-system-upgrade.html

AI 辅助提升 DDR5 内存可靠性，让数据中心一直“在线”

解决方案

AI 辅助故障分析

数据中心内存故障

阿里云

解决方案

阿里云携手英特尔合作改进 DDR5 内存可靠性，联合开发了面向 DDR5 的内存故障预测和预防解决方案，帮助提升服务器的可靠性和业务的正常运行。

- 方案在 BMC 中集成英特尔® MRT 技术提供 AI 辅助的实时预测和内存故障分析，其利用多维模型和人工智能算法，在微观层面检测内存故障，使得数据中心提前预警和主动预测潜在的内存故障风险；
- 在平台中引入第五代至强® 可扩展处理器，助力阿里云数据中心为不同工作负载提供更加强劲的算力支持。

57%

基于第五代至强® 可扩展处理器，方案经过迭代优化后，预期能够预测的不可纠正错误 (UE) 达[1]

74%

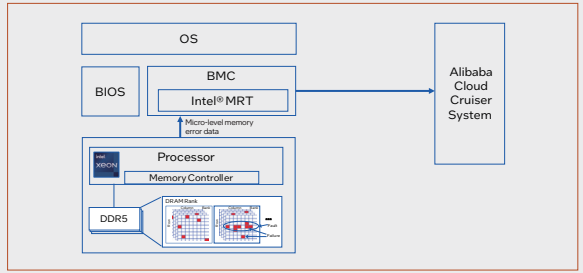
基于第五代至强® 可扩展处理器，方案经过迭代优化后，预期能够预测的可纠正错误 (CE) 达[2]

快速且全面的硬件监控服务

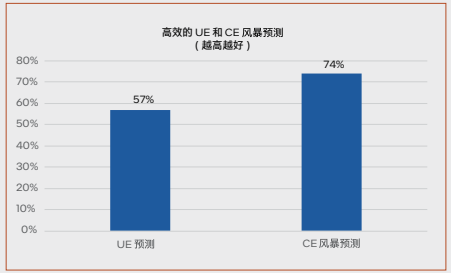
挑战

- DDR5 引入了新的架构和信号传输方式，需要更复杂的电路设计和优化
- DDR5 内存模块容量更大，增加了故障的风险
- In-DRAM 纠错码（ECC）导致主机错误观察不够明确

扫码获取全文



解决方案架构图



高效的 UE 和 CE 风暴预测[3]

[1,2,3] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/now/data-centric/ddr5-memory-reliability.html

AI+游戏，让消消乐玩法更多样，体验更顺畅

案例研究

AI 模型推理优化

游戏开发与运营

Happy Elements 腾讯云

挑战

- 游戏服务器需要足够算力处理大量的游戏数据和用户请求
- 满足模型推理性能需求的同时降低模型推理的单位成本
- 服务器需具备足够的灵活性，以适应不断变化的游戏内容和用户需求

解决方案

乐元素引入基于第五代至强® 可扩展处理器的新一代腾讯云实例 S8，并采用处理器内置的 AI 加速引擎，软硬结合加速 AI 推理，提升开发效率和游戏体验。

- 利用基于第五代至强® 可扩展处理器的腾讯云实例 S8 获得平衡、稳定的计算、内存和网络资源；
- 采用处理器内置 AI 加速引擎 -- 英特尔® AMX， 高效处理矩阵乘法运算，加速基于 CPU 的 AI 推理，避免使用独立加速器带来的成本和复杂性；
- 借助英特尔® oneDNN 这一开源、跨平台的库，开发人员可对 CPU、GPU 使用相同的 API，从而抽象出指令集和其他复杂的性能优化，显著降低编程难度。

3.44 倍

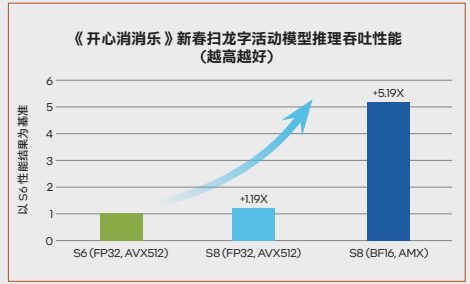
相较于第三代至强®可扩展平台，基于第五代至强®可扩展平台+英特尔®AMX将模型转化为 BF16，推理性能提升达[1]

5.19 倍

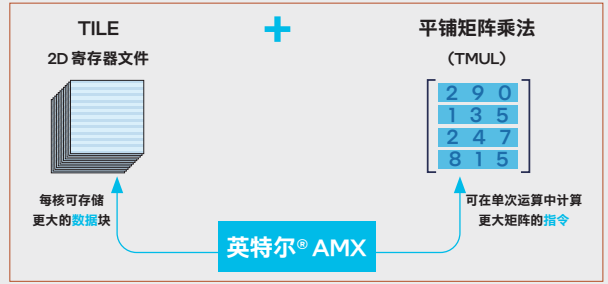
基于第五代至强® 可扩展处理器 + 英特尔® AMX，新春扫龙字活动模型推理性能提升达[2]

灵活应对更多AI 扩展应用

扫码获取全文



《开心消消乐》新春扫龙字活动模型测试数据[3]



英特尔® AMX 架构

[1,2,3] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/artificial-intelligence/s8-accelerates-happy-elements-game-ai-inference.html

ERNIE –Tiny 用“芯”瘦身 加速 NLP 应用商业落地

解决方案

模型量化

自然语言处理

飞桨

挑战

在保持精度的前提下，实现更短的 AI 推理运算时间和更少的算力需求

在既有 CPU 平台上高效率完成推理任务，减少对昂贵的专用 AI 算力设备的需求，降低 TCO

解决方案

百度 ERNIE-Tiny 使用内置英特尔® AMX 的第四代英特尔® 至强® 可扩展处理器，配合多项优化措施，充分利用处理器带来的性能加速“红利”，大幅提升推理效率。

- 采用第四代英特尔® 至强® 可扩展处理器作为 ERNIE –Tiny 推理工作的算力输出引擎，为高强度工作负载提供更可靠的全局加速；
- 以第四代至强® 可扩展处理器内置的 AI 加速技术-- 英特尔® AMX，大幅提升 ERNIE-Tiny 推理性能；
- 利用英特尔® oneDNN 实现对英特尔® AMX 的调用，有效助力用户提升 AI 应用及框架性能。

2.66 倍

采用第四代英特尔® 至强® 可扩展处理器的 ERNIE-Tiny 吞吐量提升达[1]

减半

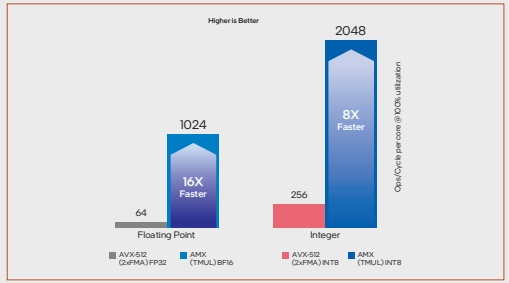
ERNIE –Tiny Medium 版与基础版 ERNIE 3.0 相比，其网络层数[2]

2,048 次 INT8 运算

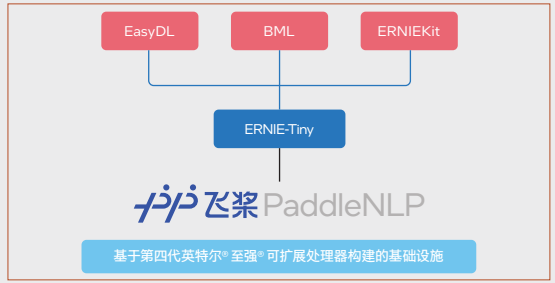
1,024 次 BF16 运算

英特尔® AMX 每个物理核在每个时钟周期可实现[3]

扫码获取全文



与英特尔® AVX-512 相比，英特尔® AMX 可带来 8 倍以上的效率提升[4]



ERNIE-Tiny 对外能力输出

[1,2,3,4] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/artificial-intelligence/spr-built-in-amx-baidu-ernie-performance-increase.html

技术篇：英特尔 AI 产品组合

开放式软件环境

oneAPI　OpenVINO　TensorFlow　PyTorch　ONNX RUNTIME　XGBoost　MODIN　DeepSpeed

深度学习加速

intel Gaudi　专用于深度学习训练和推理加速

通用加速

intel Data Center GPU Flex Series　AI 视觉推理、 VDI、媒体分析

intel Data Center GPU Max Series　并行计算、科学计算、面向科学计算的 AI、数据中心

通用计算

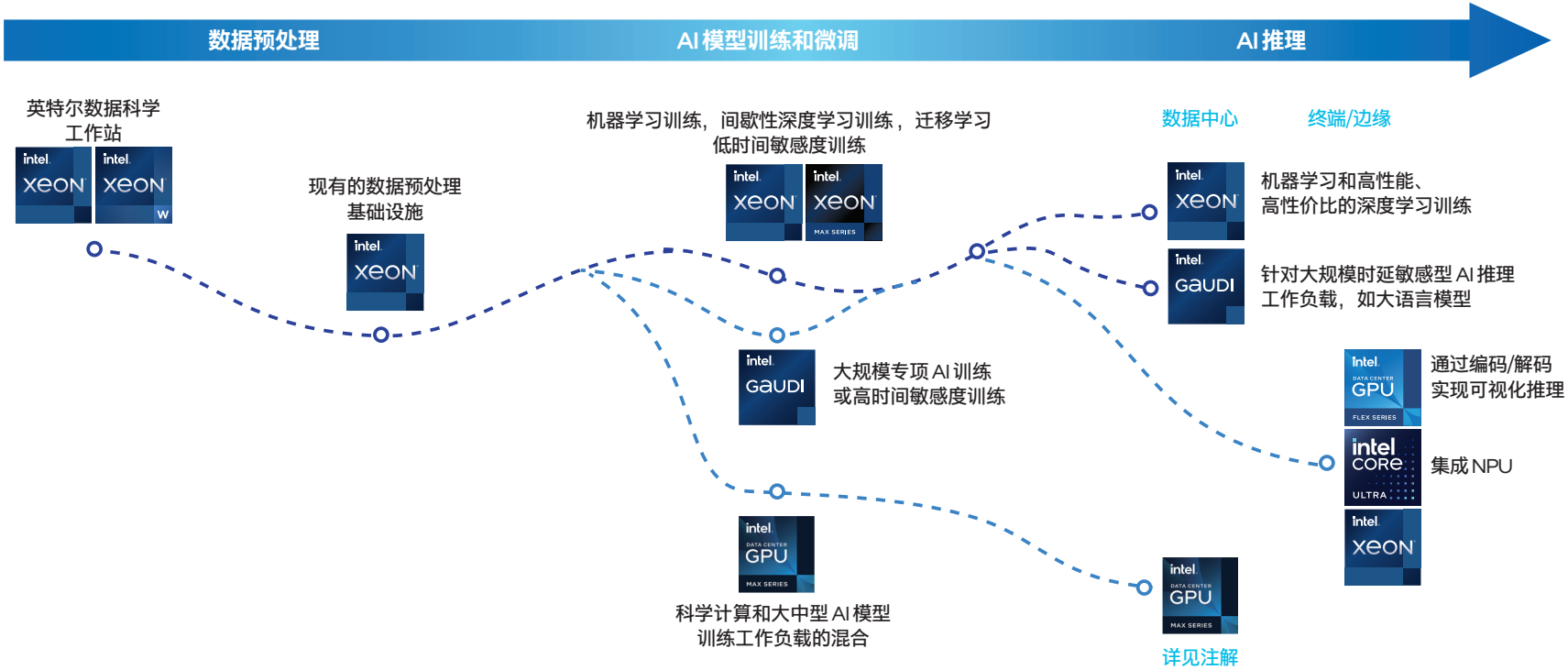
intel Xeon　intel Xeon Max Series　实时，中等吞吐量，低时延和稀疏推理

intel Xeon　intel Xeon Max Series　intel Xeon W　intel Ethernet　中小型训练和微调

intel Core Ultra　intel Xeon D　intel Data Center GPU Flex Series　intel Arc Graphics　intel Core　intel Agilex　intel Ethernet　边缘和网络 AI 推理

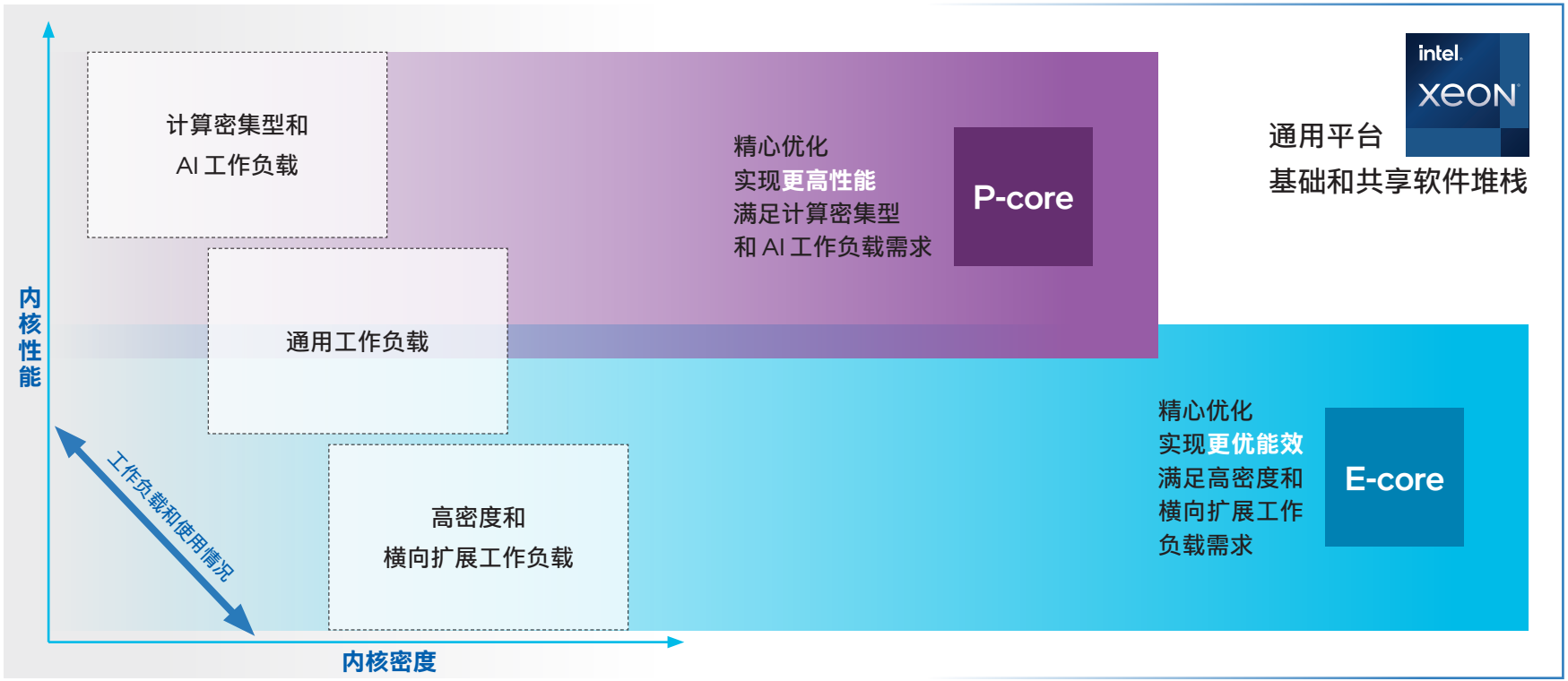
intel Core Ultra　intel evo Edition　intel Arc Graphics　终端推理

英特尔® XPU 平台：满足 AI 之旅各阶段需求



注：不限于以上所示典型的部署路径，所有计算平台均适用

英特尔® 至强® 6 处理器家族



英特尔® 至强® 6 性能核处理器

每颗 CPU 集成多达 **128 个**性能核

更高内存带宽

多达 12 个通道 DDR5（高达 6400MT/s）
采用 MRDIMM 内存（高达 8800MT/s）

更大三级共享缓存

高达 504 MB

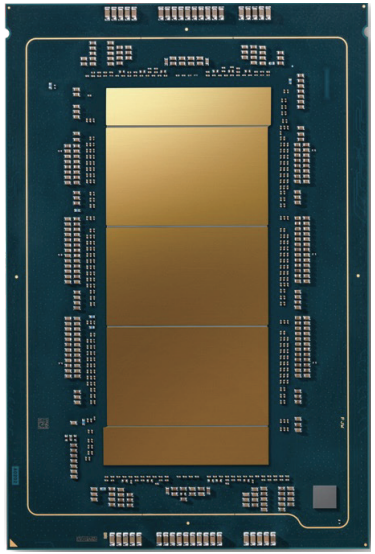
多达 96 条通道 PCIe 5.0
R1S：支持 136 条通道 PCIe 5.0

Compute Express Link 2.0（CXL 2.0）

多达 64 条通道
支持 Type 3 “Flat” 内存模式

英特尔® UPI 2.0

高达 6 UPI 2.0
高达 24 GT/s



计算单元（Compute Die）
基于英特尔 3 制程工艺

支持单路到八路设计

（高端 6900P 系列最多支持双路）

内置 AI 与科学计算加速引擎

英特尔® AMX——
（增加对 FP16 数据格式的支持）
英特尔® AVX-512

其它内置加速引擎

英特尔® QAT / 英特尔® IAA
英特尔® DSA / 英特尔® DLB

硬件增强型安全特性

英特尔® SGX / 英特尔® TDX

英特尔® Scalable Vector Search（SVS）

可调用英特尔® AMX 加速能力优化大模型
应用的好搭档——向量数据库的性能表现

英特尔® 至强® 6 能效核处理器

多达 **288** 个内核[1]（每个处理器）

更大三级缓存（L3）：高达 **216 MB**

更多内存通道：多达 **12 条**

英特尔® UPI 2.0 速度：高达 **24 GT/s**

多达 **188 条 PCIe 5.0 通道**
（双路服务器）

Compute Express Link（CXL）2.0 通道
Type 3 内存支持

AI 和科学计算
英特尔® 高级矢量扩展 2（VNNI/INT8）
英特尔® 数据流加速器（英特尔® DSA）

安全性
英特尔® 软件防护扩展（英特尔® SGX）
英特尔® 信任域扩展（英特尔® TDX）
英特尔® 密码操作硬件加速

存储和分析
英特尔® 数据流加速器（英特尔® DSA）
英特尔® 存内分析加速器（英特尔® IAA）
英特尔® 数据保护与压缩加速技术（英特尔® QAT）

网络
英特尔® 数据保护与压缩加速技术（英特尔® QAT）
英特尔® 动态负载均衡器（英特尔® DLB）

Web 和微服务 | 数据分析 | 媒体 | 网络

数据库 | 存储 | AI 推理 | 边缘

[1]已推出的 6700E 系列每个 CPU 拥有多达 144 个内核，25 年 1Q 发布的 6900E 系列最高达 288 个内核

第五代英特尔® 至强® 可扩展处理器

多达 **64** 个内核（每个处理器）

更高内存带宽：高达 **5,600 MT/s**

更大三级缓存（LLC）：高达 **3 倍**[1]（PCIe 5）

UPI 2.0 速度：高达 **20 GT/s**

Compute Express Link（CXL）1.1*
Type 3 内存支持

无需更改代码即可直接兼容
第四代英特尔® 至强® 可扩展处理器

英特尔® AMX
更高的 AMX 频率，全新许可水平
每个内核均内置 AI 加速器

集成 IP 加速器
英特尔® 数据保护与压缩加速技术（英特尔® QAT）
英特尔® 存内分析加速器（英特尔® IAA）
英特尔® 数据流加速器（英特尔® DSA）
英特尔® 动态负载均衡器（英特尔® DLB）

更高的能效
内置加速器提供高效计算/ 更高的每瓦性能
经优化的电源模式 2.0 / 针对工作负载优化的 SKU
无缝固件升级缩短停机时间

更全面的机密计算产品组合
英特尔® SGX
英特尔® TDX
英特尔® Trust Authority

与第四代至强® 相比 平均性能提升[2]	21%
与第四代至强® 相比， AI 推理性能提升[3]	42%
运行参数量在 200 亿以下的 LLM 时，词元处理时延低于[4]	100ms

[1]详情请见以下网址的 [G11]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。
[2,3]详情请见以下网址的 [G1, A16]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。
[4]基于英特尔 2023 年12 月进行的内部建模。详情请见以下网址的[A1, A2, A16]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。

第四代英特尔® 至强® 可扩展处理器

英特尔® 高级矩阵扩展（英特尔® AMX）

英特尔® 存内分析加速器（英特尔® IAA）

英特尔® 数据流加速器（英特尔® DSA）

英特尔® 动态负载均衡器（英特尔® DLB）

高达 1.53 倍
平均性能增益[1]
（与上一代产品比较）

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

英特尔® 数据保护与压缩加速技术（英特尔® QAT）

英特尔® 安全引擎
英特尔® SGX
英特尔® TDX
…

面向 vRAN 的英特尔® 高级矢量扩展

英特尔® 至强® CPU Max 系列（配备 HBM）

高达 10 倍
PyTorch 实时推理和训练性能提升[2]
启动英特尔® AMX (BF16) 时与上一代产品 (FP32) 的比较结果

创新、设计和交付均坚持工作负载优先

CPU 内置多种加速器

更出色的性能与能效，更好满足客户所需

更全面的机密计算产品组合

[1,2] 如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon-accelerated/4th-gen-xeon-scalable-processors-product-brief.html

第四代英特尔® 至强® 可扩展处理器内置七大加速器

英特尔® 存内分析加速器（英特尔® IAA）
优化内存占用和
查询吞吐量

英特尔® 高级矩阵扩展（英特尔® AMX）
加速深度学习
推理与训练

英特尔® 数据保护与压缩加速技术
（英特尔® QAT）
加速加密与压缩操作

英特尔® 动态负载均衡器（英特尔® DLB）
提升与网络处理
相关的性能

英特尔® 至强® CPU Max 系列
集成高内存带宽，为科学计算与 AI 工作负载
大幅提升数据吞吐量

英特尔® 数据流加速器（英特尔® DSA）
优化数据流的
传输和转换

英特尔® 安全技术（英特尔® Security）
帮助保护数据机密性与代码完整性

英特尔® 至强® 可扩展处理器内置 AI 加速能力的演进

内置 AI 加速能力的数据中心级 CPU

第二代至强® 可扩展处理器（Cascade Lake）
英特尔® DL Boost（AVX-512_VNNI）
全新内存存储层次结构

第三代至强® 可扩展处理器（Cooper Lake）
英特尔® DL Boost（AVX-512_BF16）

第三代至强® 可扩展处理器（Ice Lake）
英特尔® DL Boost（AVX-512_VNNI）和英特尔® Software Guard Extensions（英特尔® SGX），
支持领先 AI 应用，如联邦学习

第四代至强® 可扩展处理器（Sapphire Rapids）
第五代至强® 可扩展处理器（Emerald Rapids）
英特尔® Advanced Matrix Extensions（AMX）进一步扩展了至强® 可扩展处理器上的内置 AI 加速功能

英特尔® AVX - 512

VPMADDUBSW
VPMADDWD
VPADDD

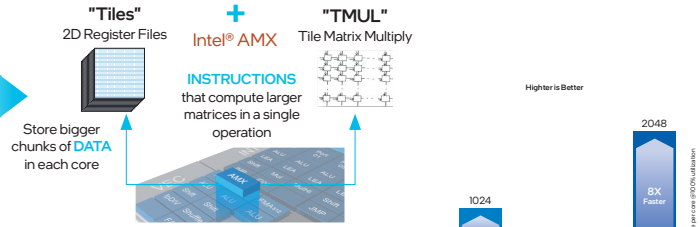
第一代至强® 可扩展处理器

英特尔® DL Boost（VNNI）

VPDPBUSD
（8-bit new instruction）

更高效的推理加速

第二代和第三代至强® 可扩展处理器



将三条指令合而为一，可最大限度地利用计算资源，提高缓存利用率

1.74x
推理表现速度提升[1]
（BERT, 第三代 vs 第二代）

相比英特尔® AVX-512，
英特尔® AMX 可提供超过
8x operations/clock/core

领先性能

[1] 如欲了解更多详情，请访问：https://www.intel.cn/content/www/cn/zh/now/data-centric/3rd-gen-intel-xeon-scalable-processors.html

英特尔® 高级矩阵扩展（英特尔® AMX）

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

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

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

高达

10 倍

与第三代至强® 可扩展处理器相比，第五代至强® 可扩展处理器可使推理工作负载性能提升[1]

高达

1.23-1.35 倍

实时推理性能提升[2]

高达

1.2-1.38 倍

实时推理每瓦性能提升[3]

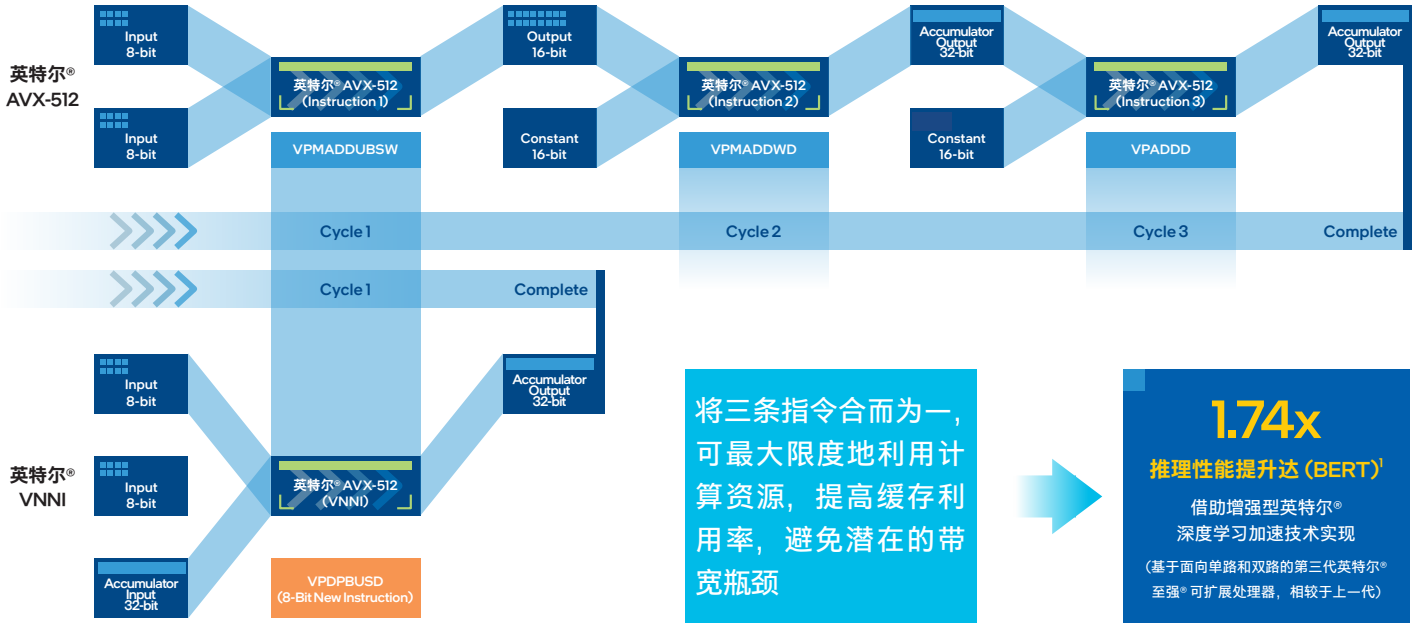
与上一代产品相比，内置英特尔® AMX 的第五代至强® 可扩展处理器

[1,2,3] 有关性能和基准测试结果的更完整信息，请访问：https://www.intel.cn/content/www/cn/zh/customer-spotlight/cases/accelerate-ai-workloads.html
结果可能不同。

英特尔® 深度学习加速

矢量神经网络指令（VNNI）

扩展英特尔® AVX-512 以加速 CPU 平台上的 AI / 深度学习推理



将三条指令合而为一，可最大限度地利用计算资源，提高缓存利用率，避免潜在的带宽瓶颈

1.74x

推理性能提升达（BERT）[1]

借助增强型英特尔®
深度学习加速技术实现

（基于面向单路和双路的第三代英特尔®
至强® 可扩展处理器，相较于上一代）

[1] 如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/now/data-centric/3rd-gen-intel-xeon-scalable-processors.html

英特尔® 深度学习加速

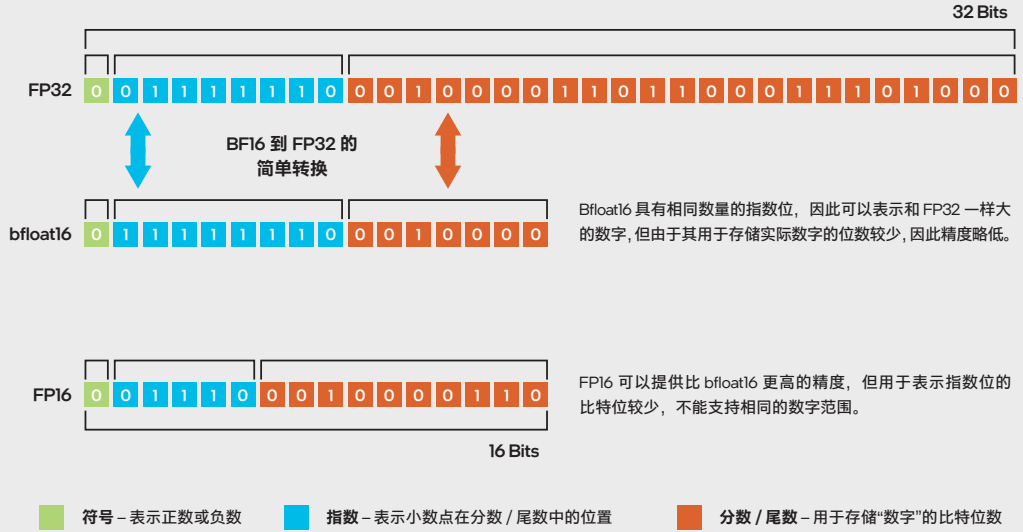
- 依据表示数字的比特位数，FP32 可提供更高的精度
- 许多 AI 功能并不需要 FP32 提供的精度水平
- bfloat16 支持基于相同指数域的相同范围的数字，但精度略低
- 从 FP32 转换到 bfloat16 比转换到 FP16 更简单
- 与 FP32 相比，使用 bfloat16 可实现每周期两倍的吞吐量

脑浮点数 (bfloat16)

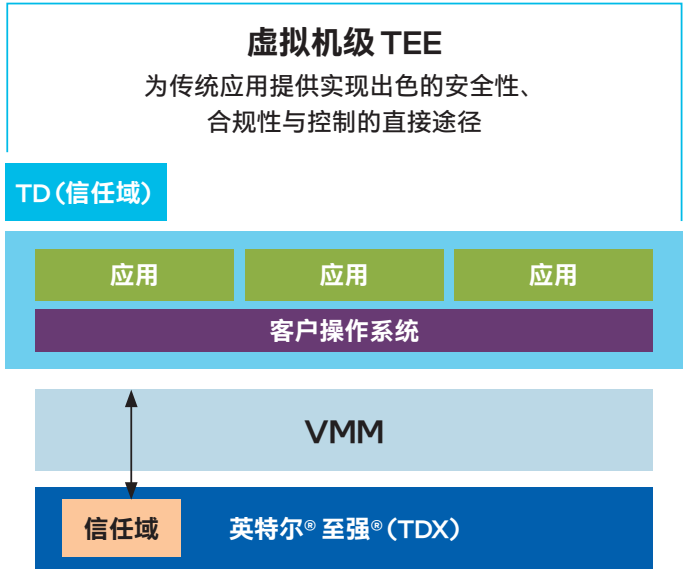
示例：

Number: 0.5658097267150878906 2596

As FP32: 0.565809726715087890625



英特尔® Trust Domain Extensions（英特尔® TDX）



建立数据主权和控制	将数据和知识产权隔离在保密的虚拟机中，并将受保护数据的访问权限限制在获得明确许可的软件或管理员范围内；
保护数据和知识产权	通过在 VM 内的 TEE 对机密数据进行硬件增强隔离，帮助减少攻击面并降低外部实体破坏、篡改或窃取数据的风险；
简化监管合规	英特尔® TDX、英特尔® SGX 等可以帮助受严格的数据隐私法规约束的组织通过加密和安全区来满足合规标准；
在可信环境中部署 AI	英特尔® TDX 凭借强大的隔离、完整性和保密功能，帮助保护应用程序、数据和 AI 模型免受未经授权的访问。

简化将现有应用程序移植和迁移到机密计算环境的过程，在大多数情况下，无需更改应用程序代码，即可激活虚拟机内由英特尔® TDX 支持的可信域。

如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/products/docs/accelerator-engines/trust-domain-extensions.html

英特尔® Software Guard Extensions（英特尔® SGX）

采用英特尔® SGX 的机密人工智能应用，保护使用中的数据和代码

基于硬件的
可信执行环境（Trusted Execution Environment）

加密敏感数据

隔离

保护敏感数据和代码，不受所有其他软件、云租户或管理员的影响，即使是恶意的访问

未加密的
数据和结果

已验证的应用
或模型代码

加密或匿名的结果

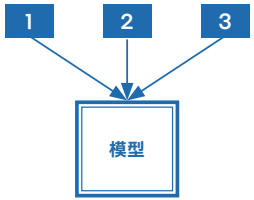
鉴证

加密验证 TEE 是否真实、配置正确且仅运行预期的软件负载

人工智能应用场景

集中式多方

示例：多家医院汇集受监管的患者数据，以进行诊断模型训练

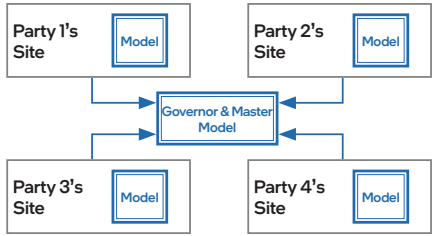


受监管的数据

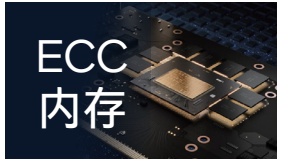
示例：智慧城市摄像头捕获的受严格数据处理法规约束的个人身份信息（PII）

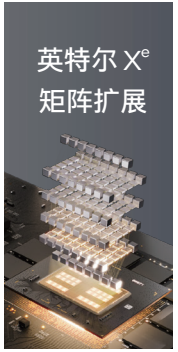
联邦学习

示例：银行合作进行反洗钱，但数据太大且敏感，无法移动

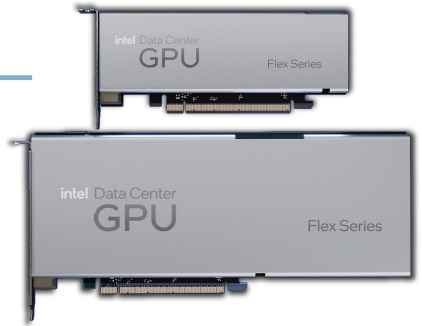


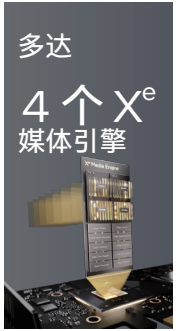
英特尔® 数据中心 GPU Flex 系列





Flex 140	16 Xe 内核 16 光线追踪单元	75W 半高 PCIe	
Flex 170	32 Xe 内核 32 光线追踪单元	150W 全高 PCIe	





面向智能视觉云的 GPU 解决方案，支持基于标准的开放式软件堆栈，针对密度和质量进行了优化，具有关键的服务器功能，可实现高可靠性、可用性和可扩展性，有助于减少数据中心使用不同解决方案并管理异构或专有环境的需求，支持的工作负载包括：

AI 视觉推理 **媒体处理和交付** **云游戏** **虚拟桌面基础设施**

如欲了解更多详情，请访问： https://www.intel.cn/content/www/cn/zh/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html

基于英特尔® 架构的 AI 软件工具组合

工程数据	模型创建	优化 & 部署

MODIN · SciPy · pandas · NumPy · Spark · Numba

大规模数据分析†

dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO · scikit-learn · TensorFlow · DirectML · SigOpt AutoML · intel Neural Compressor · W3C WebAssembly WebGPU WebNN

机器学习 & 深度学习框架，优化和部署工具†

oneAPI

英特尔® oneAPI Deep Neural Network Library

英特尔® oneAPI Collective Communications Library

英特尔® oneAPI Math Kernel Library

英特尔® oneAPI Data Analytics Library

适用于 CPU、GPU 和其他加速器的开放式、多架构编程模型

云 & 企业 · 终端 & 工作站 · 边缘

intel AI TOOLS	端到端加速数据科学和 AI 应用
英特尔® Developer Cloud and 英特尔® Developer Catalog	体验最新的英特尔工具和硬件，使用经优化的 AI 模型
cnvrg.io	全栈式机器学习操作系统
英特尔® Geti™	数据标注、模型训练及优化的平台
Hugging Face	英特尔的优化和微调指南，经优化的推理模型和模型服务

注：堆栈中每一层的组件均基于预期的 AI 使用模型，有针对性地对其他层的目标组件进行优化，但并非每个组件都被最右列的解决方案所使用。

†本列表包括面向英特尔硬件进行优化的主流开源框架。

英特尔® oneAPI AI Analytics 工具套件

利用面向英特尔® 架构优化的库加速端到端人工智能和数据分析管道

显著优势

- 利用面向英特尔® 架构优化的深度学习框架和工具提升训练和推理性能
- 使用计算密集型 Python 包为数据分析和机器学习工作流提供落地加速

性能加速

简化端到端工作流程

提高生产力

加快开发

深度学习

- 面向英特尔® 架构优化的 TensorFlow
- 面向英特尔® 架构优化的 PyTorch
- 英特尔® 低精度优化工具（英特尔® LPOT）
- 面向英特尔® 架构优化的 Model Zoo

数据分析 & 机器学习

加速数据库

- 面向英特尔® 架构优化的 Modin
- HEAVY.AI Backend(formerly OmniSci)

面向英特尔® 架构优化的 Python

- XGBoost
- Scikit-learn
- Daal-4Py
- NumPy
- SciPy
- Pandas

示例及端到端工作负载

CPU　GPU　**支持的硬件架构**

硬件支持因个别工具而异。架构支持将随着时间的推移而扩大。

点击或通过如下链接获取工具包

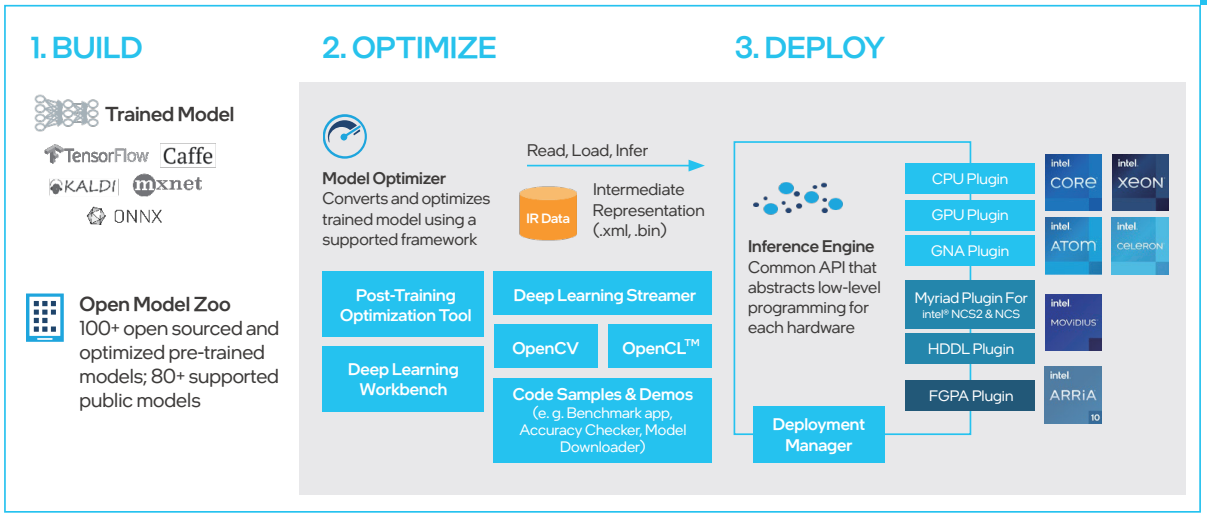
- Intel Installer
- Docker
- Apt, Yum
- Conda
- Intel® DevCloud

了解更多: software.intel.com/oneapi/ai-kit

OpenVINO™ 工具套件 - 由 oneAPI 提供支持

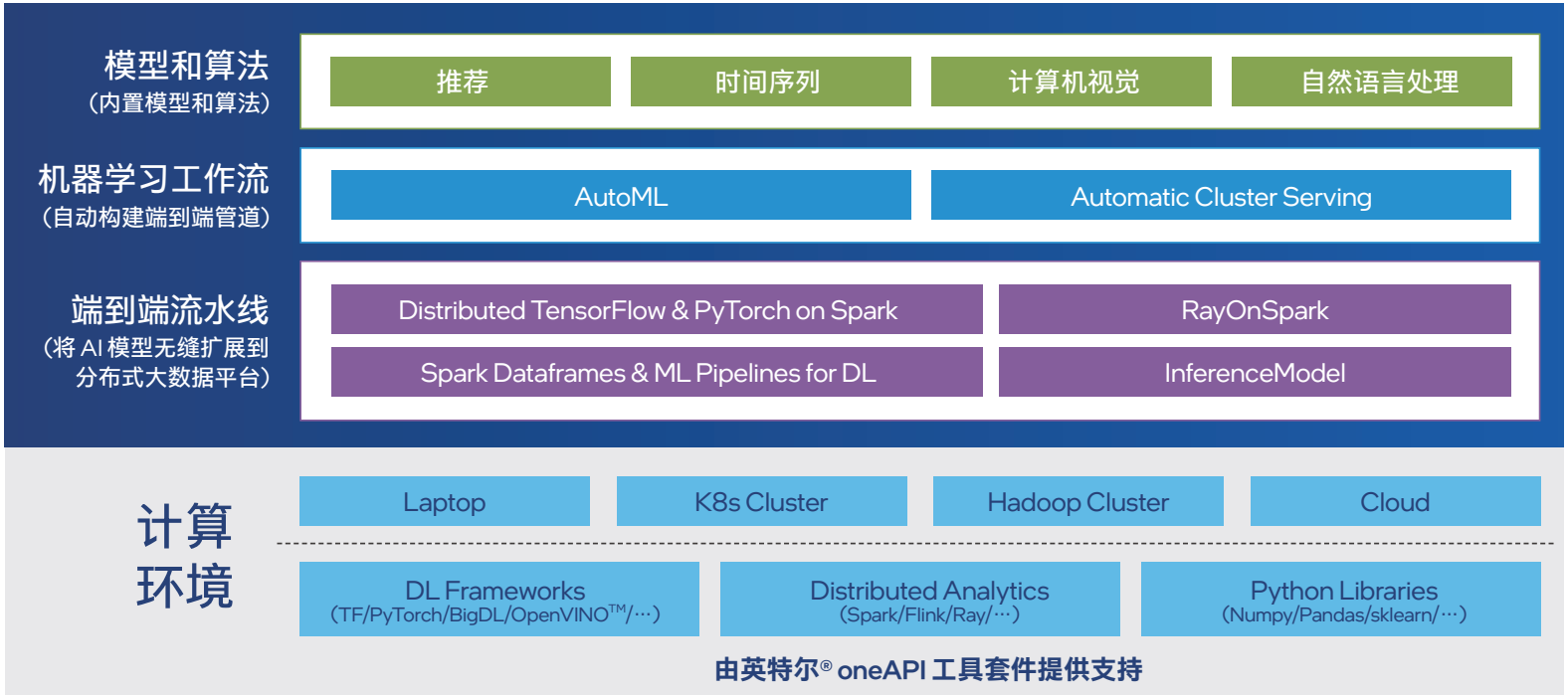
旨在使用高性能人工智能和计算机视觉推理实现更加快速和准确的实际结果，部署在从边缘到云的、基于英特尔® XPU 架构（CPU、GPU、FPGA、VPU）的生产环境中

- 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



了解更多：software.intel.com/openvino-toolkit

BigDL*：统一的大数据分析和 AI 平台

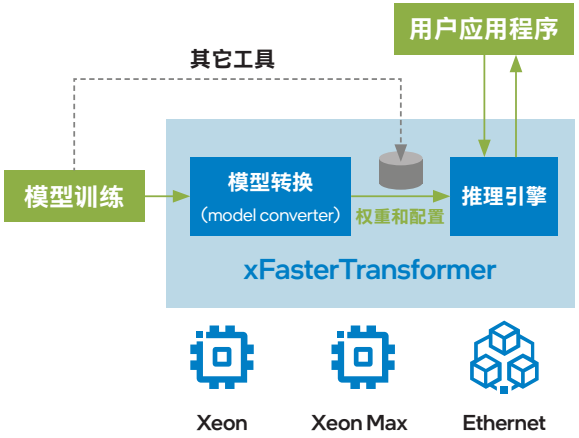


了解更多: https://github.com/intel-analytics/analytics-zoo

* 指 BigDL 2.0 已包含 BigDL 和 Analytics Zoo。

xFaster Transformer (xFT)

为大语言模型推理加速



代码以 Apache 许可证开源在
https://github.com/intel/xfastertransformer

更高性能

- 释放至强® 和至强® Max 系列处理器的 DRAM 和 HBM 带宽潜能

更好扩展性

- 支持跨 Socket、跨节点分布式推理
- 支持高达 70B LLM 模型 (Qwen-72B)

更强兼容性

- 支持多种 LLM 模型，如 LlaMA 1/2, ChatGLM 1/2/3, Baichuan, OPT, Qwen
- 支持不同规模 LLM 模型，如6B, 7B, 13B, 30B 等
- 支持 BF16、FP16、INT8、W8A8、INT4 等
- 兼容 Faster Transformer 模型格式
- 兼容 Hugging Face 与 PyTorch

更好 TCO

- 内存容量需求低
- 通过精细的内存规划进行优化，以支持更大模型

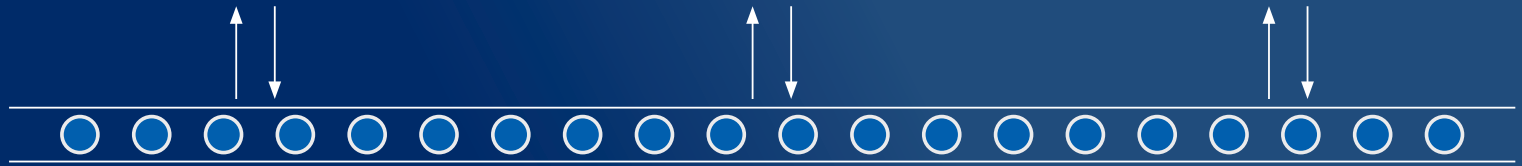
大数据分析+人工智能端到端流水线

从笔记本电脑无缝扩展到分布式大数据平台

使用样本数据在
笔记本电脑上制作原型

在承载历史数据的
集群上进行试验

使用分布式数据流水线
进行**生产**部署



- 轻松构建将 AI 模型与大数据融合对接的端到端流水线原型
- 从笔记本电脑到分布式集群的**“零”**代码更改
- 可在**生产环境中**的 Hadoop/K8s 集群上无缝部署
- 实现从机器学习到大数据应用的**流程自动化**

英特尔 AI 实战视频课程

- 至强® AI 实战课 CCF 联合专场
- 英特尔® 至强® RAS 为 AI 服务器护航
- 英特尔® 至强® CPU 让 AI 部署无处不在
- 大模型时代的云服务安全利器
- 从 OCR 起步推进企业 AI 应用落地

法律声明

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。
在特定系统的特殊测试中测试组件性能。硬件、软件或配置的差异将影响实际性能。当您考虑采购时，请查阅其他信息来源评估性能。关于性能和基准测试程序结果的更多信息，请访问 www.intel.com/benchmarks。
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。
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
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。
此处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。