NDA – 内部专用

人工智能 (AI)

合作伙伴支持程序包

利用基于英特尔的解决方案，应对客户的业务挑战

演示文稿说明

演示稿名称	人工智能合作伙伴支持程序包
摘要	本程序包采用公开和非公开的英特尔内容开发而成，旨在供 PSAM 和英特尔合作伙伴使用，以指导与合作伙伴代表的深入对话。 重点围绕 AI，与 SSP VX 最高战略优先事项保持一致，并使合作伙伴销售商能够宣传英特尔的 AI 投资和开放生态系统。 该程序包经过定制，能够与合作伙伴开展对话，其内容适合所有受众角色。 这是一个持续进行的项目。我们将定期发布新版本并更新支持包，以添加最新的英特尔内容。英特尔重点将扩大，并且很快将提供更多定制包。
目标受众	合作伙伴生态系统、最终客户、销售
接下来即将推出	• 以 GEO 为中心的包 • 为合作伙伴量身定制的包

行为召唤	1. 在与合作伙伴的对话中使用本支持包 2. 与您的合作伙伴分享公开版本 3. 如果您的合作伙伴不是 IPA 的成员，请鼓励他们加入 4. 向 Amy Kircos (amy.kircos@intel.com) 提供关于您所在地区或合作伙伴所需的任何建议定制的反馈，以对即将推出的支持包产生影响

让 AI 遍布每个角落

在每个平台中实现连续的 AI 体验……
从客户端和边缘到数据中心和云。

为什么要与英特尔建立合作伙伴关系？

在英特尔，我们有机会改善地球上每个人的生活，提高每家企业的业绩

但我们并非孤军作战！

我们联手合作伙伴，**让 AI 遍布每个角落**，最大限度降低部署风险，
为客户创造真正的价值

您与英特尔合作，就是选择了完整的 AI 生态系统

我们拥有广泛的 AI 技术组合，并与硬件、软件和系统集成商建立合作关系，
紧密合作开发真实世界的解决方案，为各行各业、企业和社区提供差异化的业务成果。
帮助您发展业务。

加入我们的旅程，让 AI 遍布每个角落

英特尔 AI 行业影响

CLUB 386

"英特尔深知这是一个千载难逢的商机，可寻址市场 (TAM) 总值高达数百亿美元，因此，一直忙于在各行各业和商业细分市场构建普及 AI 所需的基础设施。

Windows Central

"做好准备，如果可以相信这个行业，那么 2024 年将是属于 AI 电脑的一年，而这一切从英特尔开始。"

EXP review

"我们开始感觉到，英特尔在同时发布性能增强型 AI 电脑和新的数据中心 CPU 之后，已经逐步形成了自己在 AI 领域的优势。"

tom's HARDWARE

"至关重要的是，英特尔的新芯片也已如期到货，这是对公司转型进程的重要确认。"

omicrono

"英特尔知道，到 2024 年，AI 将无处不在，并希望其处理器成为所有软件技术的基础，而这些技术将广泛运用在互联网和 Windows 等计算机操作系统之中。有了这些，只需点击几下，就能重新编辑自己喜欢的歌曲，或在电脑上轻松快捷地为旅行照片建模。有了英特尔酷睿 Ultra，人人都能成为艺术家、作家和音乐家。”

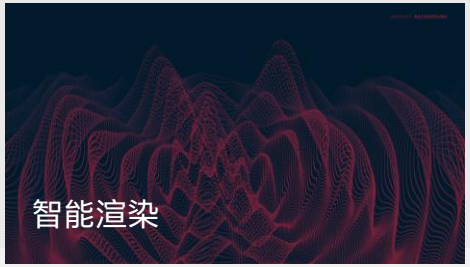
siliconANGLE

"英特尔不仅给出令人印象深刻的数字和说法，还提供了一些真实世界的例子，说明其新芯片将支持的 AI 工作负载类型。例如，餐厅将能够根据食客的个人预算和饮食需求，指导食客选择菜单，而制造商将能够建立新的系统，记录工厂车间的质量和安全问题。由英特尔芯片驱动的高级 AI 还将有助于创造出更有效的超声波系统，发现人类医生可能会忽略的问题。“

AI 正在改变全球的企业

企业如何从中受益?

您的企业可以利用 AI 来增加利润和提高效率



Gartner AI 技术成熟度曲线

2023 年人工智能 (AI) Gartner Hype Cycle™（技术成熟度曲线）确定了多项创新和技术，不仅具有显著甚至变革性的效益，还能解决易出错系统的局限性和风险。

AI 策略应该考虑哪一种创新技术提供了最可靠的投资理由。

"尽早采用这些创新技术将带来显著的竞争优势，并缓解在业务流程中利用 AI 模型的相关问题"。Gartner 总监分析师 Afraz Jaffri

2023 年 Gartner 技术成熟度曲线中人工智能的新变化

AI 发展日新月异

底层数据技术：

$3000 亿 到 2026 年，全球 GenAI 支出将超过 3000 亿美元

到 2026 年
AI 无处不在

超过 50% 的企业管理的数据将在数据中心或云端之外创建和处理

58% 的领先上市公司首席执行官正在积极投资于 AI

50% 的边缘部署将涉及 AI

AI 将像互联网一样产生颠覆性影响

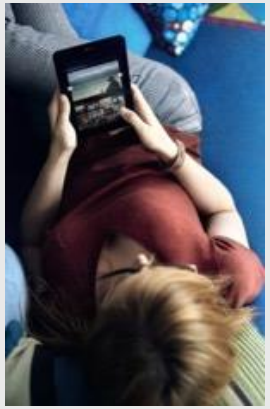
- 预计到 2040 年，**生成式 AI** 将为全球经济增加高达 $4.4 万亿的价值[2]
- **AI 推理**将推高计算成本；超过摩尔定律的速度
- **大模型**规模的增长（1T+ 参数模型）
- **更小、更灵活模型的**增长（~10B 参数）

[1] https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights
[2] 全球人工智能支出指南 (IDC)

https://chiefexecutive.net/the-rise-of-the-ai-ceo/
https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

到 2026 年，至少 50% 的边缘计算部署将涉及机器学习 (ML)，而 2022 年这一比例仅为 5%（《构建边缘计算战略》，2023 年 4 月）

业务机会
用例示例





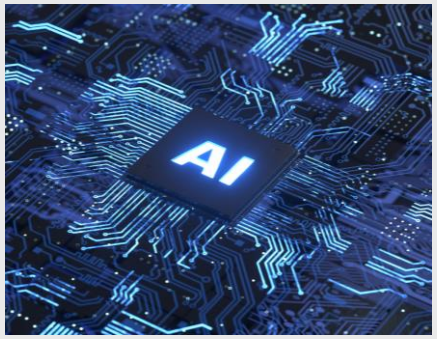




教育	健康	财务	零售	政府	能源	汽车	制造	电信
教师助手	药物研发	算法交易	产品推广	政府服务 聊天机器人	消费预测	汽车开发	工厂自动化	个性化客户 服务
学生学习支持	医生副手	客户投资组 合助理	客户对接和 情绪工具	文档搜索总结	运营绩效	多语种车内辅助	预测性维护	网络自动化
父母聊天门户	患者家庭聊 天机器人	风险/信用评估	图像购物辅助	实时语言翻译	能源交易助理	供应链优化	精准农业	运营绩效

当今 AI 面临的挑战是什么？

为什么与英特尔合作



GPU 可用性

英特尔 CPU 替代方案
解决全球 GPU 短缺问题

正当全球的信息技术公司对 Nvidia 的 GPU 涨价以及全球 GPU 短缺日益不满之际，Naver 的 AI 服务器交换机推出市面

厂商锁定

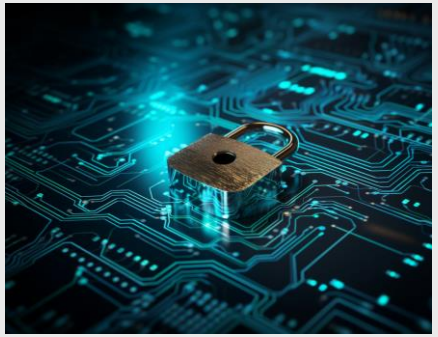
使用基于标准的开源软件，
避免受限于厂商

英特尔与各种行业标准开放框架及库合作，力图优化英特尔技术达到卓越性能，确保优质的开箱即用体验

成本

第四代英特尔® 至强®，
性价比更高

在实际工作应用中，英特尔凭借更优的性能、更低的价格和更平衡的 AI 推理平台，颠覆了业界，实现了 AI 的民主化

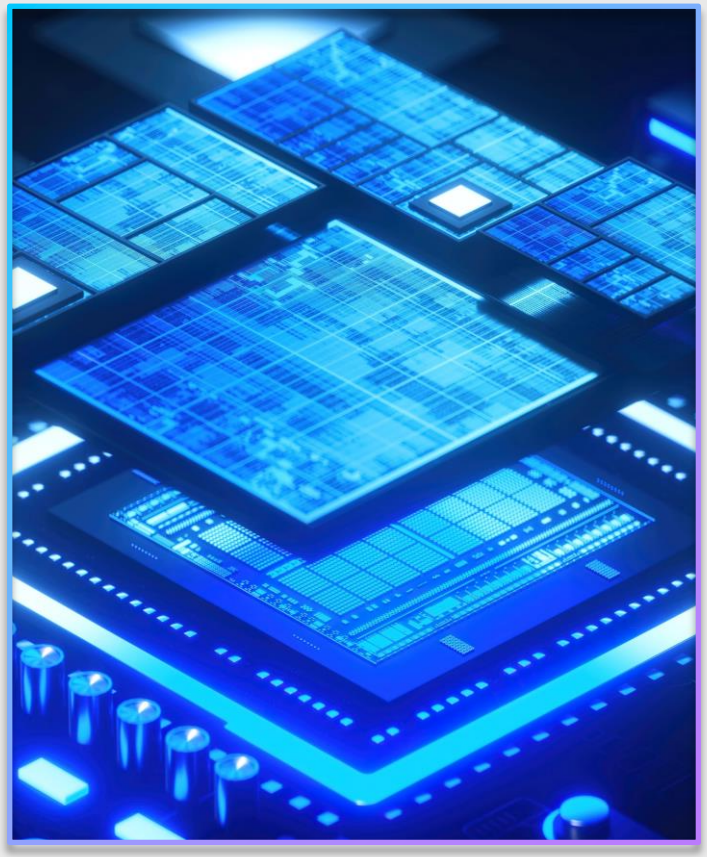


安全 AI

英特尔提供非常全面的
安全产品组合

英特尔的安全能力让您可以设适合您工作负载的信任边界，有助保护敏感数据、内容和软件 IP，以防高级攻击、篡改和盗窃

英特尔可以为 AI 提供什么：精心设计的 AI 平台



- 各行各业争相涌入 AI。英特尔是唯一一家拥有全套软硬件平台的公司，提供开放和模块化解决方案，带来有竞争力的总体拥有成本 (TCO) 和实现价值的时间，帮助合作伙伴在这个指数式增长和 AI 无处不在的时代，运筹帷幄，决胜千里。
- 我们正在将 AI 注入英特尔技术之中，以支持当今的 GenAI 工作负载，推动 AI 电脑和边缘 AI 等新兴用途，以及开拓创新，推动未来十年 AI 的发展。
- 我们凭借在知识产权、工艺、封装、安全、软件、服务、制造和代工服务方面的领先地位，能够实现 AI 的全部潜力，彻底转变不同行业，应对世界上最大的挑战。

英特尔可以为 AI 提供什么：开放且与硬件无关的软件方法

- 开放的方法以及与开发人员生态系统的深度契合，对于降低开发人员和客户的准入壁垒并开启 AI 创新而言，至关重要。我们正在加快开发开放式 AI 软件生态系统，以便打破专有壁垒。

- 我们为合作伙伴和开发人员提供抢先体验以及最快捷的路径，利用英特尔® Developer Cloud 以及集成且可扩展的硬件/软件系统和解决方案，扩展其 AI 解决方案。

英特尔可以为 AI 提供什么：值得信赖的平台和解决方案

- 只有当 AI 合乎道德并且负责任之时，才是真正广为所用。随着 AI 呈指数式增长，我们与业界通力合作，提供创新的生态系统工具和解决方案，让 AI 更加安全，帮助解决隐私问题。
- 我们正在构建平台和技术，集 AI 和安全于一体，以便合作伙伴在数据中心、云端、个人电脑和边缘中，自信地帮助确保各种 AI 工作负载的安全。

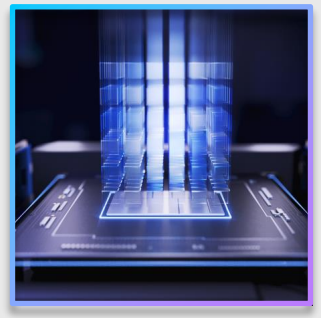
英特尔如何让 AI 遍布每个角落

改变世界的技术，助益于每一个人

英特尔的独特价值

- 开放方法
- 软硬件的专业知识
- 生态系统
- 执行

英特尔广泛的 AI 支持技术组合、对未来 AI 增强创新的独特愿景，以及对开放生态系统无与伦比的支持，正在助力**让 AI 遍布每个角落**，助益于每一个人



- 在各种工作负载中扩展 AI，让个人和组织都能使用 AI
- 异构架构、开放标准和解决方案，让客户自信地在数据中心、云、个人电脑和边缘中，确保各种 AI 工作负载的安全

AI 连续体
让 AI 无处不在

英特尔是您值得信赖的合作伙伴，**让 AI 遍布每个角落**，并在此过程中的每一步都为您的业务保驾护航

从数据中心、云和网络，到客户端和边缘

说明：英特尔® 酷睿™ Ultra 从 Meteor Lake 开始集成了 NPU 低功耗推理引擎。

与英特尔携手构建负责任的 AI

增强无障碍性

对于许多残疾人来说，独立和自主可能是一项挑战。AI 正在帮助改变这种状况，创造出替代性产品，应对日常障碍

创建环境解决方案

研究人员利用 AI 技术，可以加深对环境的认识，然后制定解决方案，建设更美好的未来

扩大教育机会

英特尔致力于解决全球 AI 技能人才短缺问题，制定了"青少年 AI"和“未来劳动力 AI"等计划，帮助学生为迎接数字革命做好准备

推进医疗保健

目前，AI 已普遍应用于医疗保健和生命科学领域，从改善患者护理到开展预防性疾病研究，不一而足

提高安全

从让自动驾驶汽车成功行驶到减少剥削儿童的现象，AI 技术正在帮助社会变得更加安全

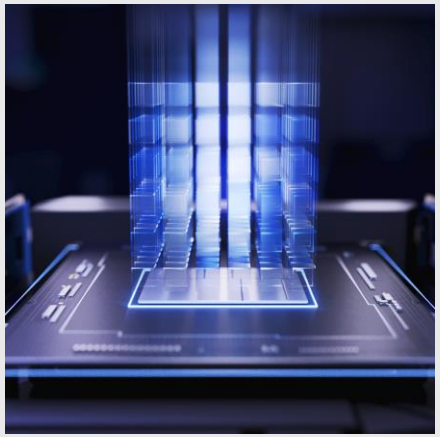
英特尔如何步步为营，助力 AI

英特尔是您值得信赖的合作伙伴，**让 AI 遍布每个角落**，
并在此过程中的每一步都为您的业务保驾护航

实现价值最大化

随处部署

安全可靠



生态系统投资

最大化
价值

实现价值最大化

为什么英特尔的开放 AI 方法适合您的 AI 业务需求

避免受限于厂商
基于标准的开源软件

MODIN
SciPy
dmlc XGBoost
PyTorch
pandas
NumPy
scikit learn
TensorFlow
Spark
Numba
SigOpt AutoML
oneAPI

利用英特尔的硬件产品组合
面向 AI 用例进行优化

intel. XEON
intel. GAUDI
intel. DATA CENTER GPU
intel. CORE

利用面向未来 AI 的软件和开放标准优化的硬件，
从客户端和边缘到数据中心和云端，创造新的机会

最大化
价值

英特尔的 AI 战略

英特尔为加速 AI 创新带来了什么

AI 应用程序和软件

开放式 | 高效 | 可访问性 | 新算法

利用内置的开放标准和软件，

英特尔® Developer Cloud **混合 AI** OpenVINO™

大规模提高性能，

数据中心
- 可扩展系统
- 加速器，至强®

网络
- 开放标准
- 网络基础设施

客户端和边缘
- AI 电脑
- NPU、GPU、CPU

并利用先进、

负责任的流程，

先进的高性能技术

开放式 AI 系统代工厂

加速每个平台，

合乎道德领导力基金会

从而确保 AI 数据的可信度和安全性。

广布 AI: 白皮书 视频

最大化
价值

英特尔 AI 产品组合

利用经过优化的硬件和软件，满足您的全部 AI 计算需求

开源软件环境

oneAPI　OpenVINO　PyTorch　XGBoost　MODIN　TensorFlow　ONNX RUNTIME　deepspeed

深度学习加速

- Intel Gaudi — 专门用于深度学习训练与推理

通用加速

- Intel Data Center GPU Flex Series — 云游戏、VDI、媒体分析、实时密集型视频
- Intel Data Center GPU Max Series — 并行计算、高性能计算、面向高性能计算的 AI

通用

- Intel Xeon / Intel Xeon Max Series — 实时、中等吞吐量、低延迟、稀疏推理
- Intel Xeon / Intel Xeon Max Series / Intel Xeon W — 中小规模训练和微调
- Intel Core i7 / Intel Xeon D / Intel Data Center GPU Flex Series — 边缘和网络 AI 推理
- Intel Core i7 / Intel Arc Graphics — 在客户端上进行推理

英特尔® AI 软件目录

工程师数据 | 创建模型 | 优化和部署

MODIN · SciPy · dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO

pandas · NumPy · scikit learn · TensorFlow · DirectML

一次编码，随处部署

Spark · Numba · SigOpt AutoML · intel Neural Compressor

大规模数据分析†

机器学习和深度学习框架、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跨架构编程模型

云与企业

客户端和工作站

边缘

intel AI ANALYTICS TOOLKIT

加速端到端数据科学和 AI

英特尔® Developer Cloud 和 英特尔® Developer Catalog

试用最新的英特尔工具和硬件，访问优化的 AI 模型

cnvrg.io

全栈机器学习操作系统

英特尔® Geti

注释/训练/优化平台

Hugging Face

英特尔优化和微调配方、优化推理模型和模型服务

注：根据预期的 AI 使用模型，针对其他层的目标组件，优化堆栈每一层的组件，最右边一列的解决方案并非使用了每个组件
† 此列表包括针对英特尔硬件优化的流行开源框架

利用参考套件加速 AI 开发

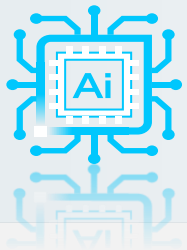
经优化的 AI 参考套件可帮助开发人员和数据科学家更快进行创新

依托 oneAPI 基于标准的开放式异构编程模型和英特尔端到端 AI 软件产品组合组件（如英特尔® AI Analytics Toolkit 和英特尔® 发行版 OpenVINO™ 工具套件）而构建，这些参考套件可帮助 AI 开发人员简化将 AI 纳入其应用的流程，增强现有智能解决方案并加速部署。

与传统模型开发工作流程相比，结果是经验证的性能提升，以及更短、更高效的工作流程

用户使用 **AI 参考套件**（旨在与企业对话式 AI 聊天机器人建立交互），便可在批量模式下体验推理，由于**采用了 oneAPI 优化，速度提升高达 45%**

AI 参考套件旨在为生命科学示范培训实现视觉质量控制检测自动化，通过 oneAPI 优化，**速度提升高达 20%，视觉缺陷检测推理速度提升 55%**



为让开发人员能够预测公用事业资产运行状况并提高服务可靠性，有一款 **AI 参考套件** 可以将**预测精度提升高达 25%**。

新闻稿

AI 参考套件电子书

随时保持
安全

安全可靠

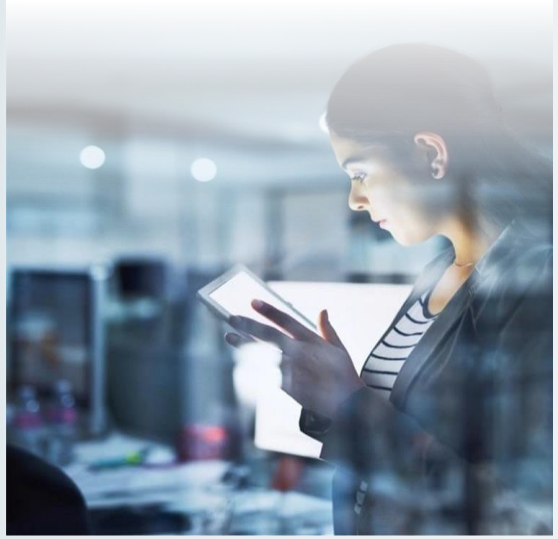
利用内置安全功能，保护您的 AI 计划并遵守法规

安全性

保护敏感数据和模型

合规性

遵守安全和隐私法规



保密性

参与多方 AI，而不会暴露私有数据

随时保持
安全

英特尔提供最全面的安全产品组合

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

应用程序隔离

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

虚拟机隔离

英特尔® Trust Authority

面向多云和混合云的独立信任验证服务

软件解决方案、云、原始设备制造商和系统集成商生态系统

英特尔安全第一的开发和生命周期支持

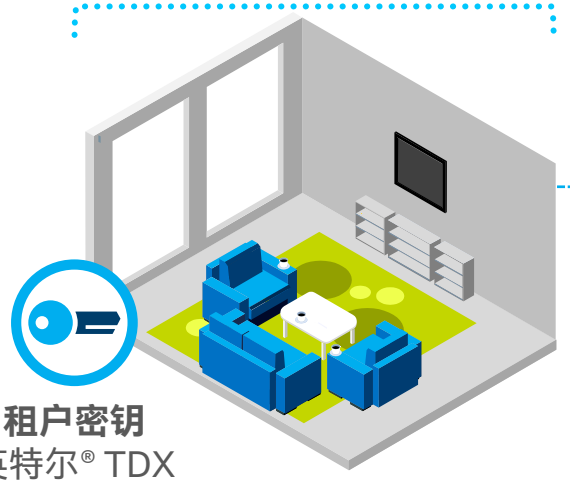
*英特尔® TDX 通过特定云提供商提供

保护内存中的数据

现今： 如果您能侦测内存，则可以看到所有经过的东西，包括用于解密数据的私钥

类比：系统内存

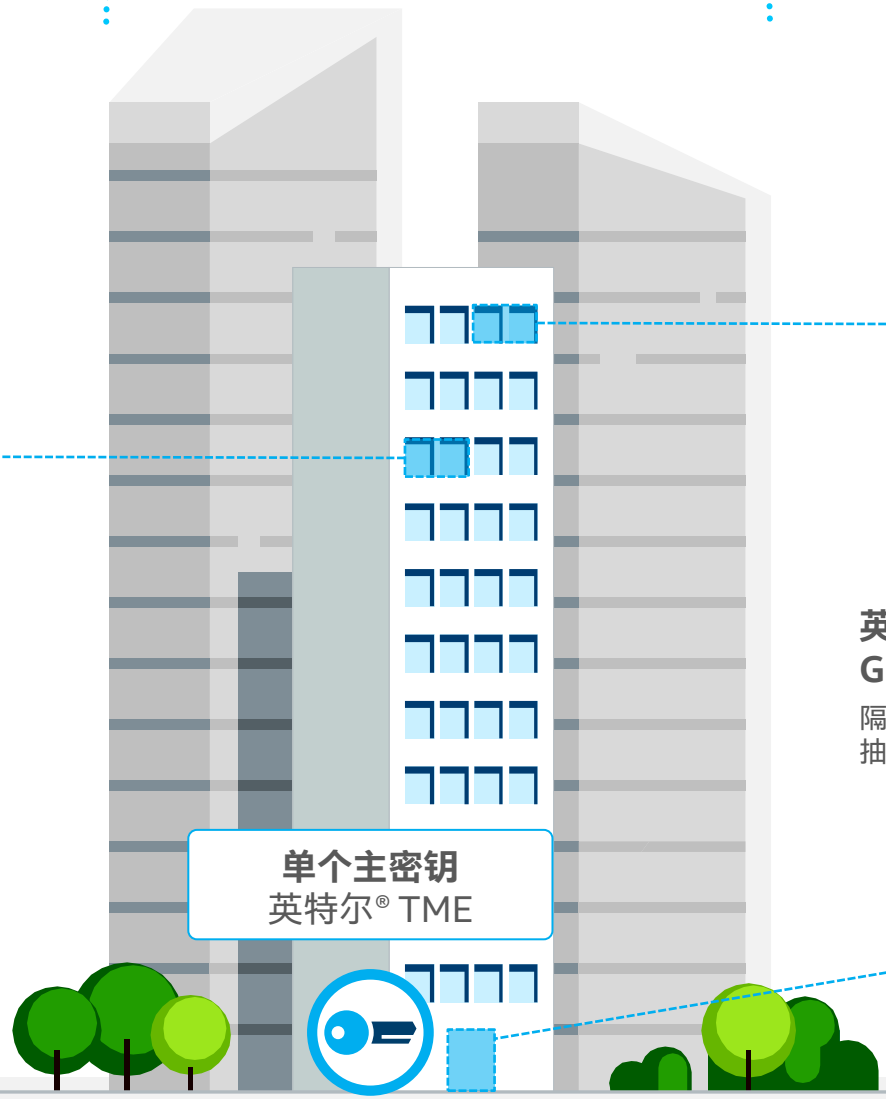
类比：虚拟机



租户密钥
英特尔® TDX

英特尔® TDX = 英特尔® Trust Domain Extensions

使用不同的密钥，单独加密每个虚拟机空间（仅要求操作系统/VMM 能够感知功能）



类比：应用

应用程序密钥
英特尔® SGX

英特尔® SGX = 英特尔® Software Guard Extensions

隔离造就单个应用程序数据空间（需要应用程序代码修改或抽象接口）

英特尔® TME = 英特尔® Total Memory Encryption

使用单个密钥，加密全部系统内存（不需要修改操作系统/应用程序）

随时保持
安全

利用英特尔® Security Engines，加速创新，加强数据保护

采用英特尔® 至强® 可扩展平台，实现机密计算

将数据付诸行动，同时帮助保持隐私

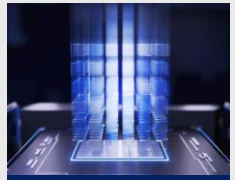
英特尔® 至强® CPU 配备多个英特尔® Security Engines，既能维持出色性能，又能保护数据机密性与代码完整性。

产品简介

借助英特尔®技术，加深对关键业务成果的洞察：

业务简介

利用
英特尔® SGX 和 英特尔® TDX，
拥抱机密计算



实现 AI 生态系统

利用 AI 开发人员偏爱的现代软件工具，以优化的性能，推动新机遇和关键业务成果

开放式
可编程性

PyTorch
TensorFlow
python
Hugging Face
UXL
SYCL

选择余地
兼容性

oneAPI
OpenVINO
SYCLomatic
OpenXLA
OpenAI Triton
MLIR

信任
推断推动

安全 AI

安全工作负载

安全模型

保护使用中的数据

规模化
开发和测试

英特尔® Developer Cloud

cloud.intel.com

大中小型模型

全系统、全集群

最新英特尔 CPU，加速器和软件

面向 AI 的开放、加速、互联计算

多厂商

多架构

硬件/架构

intel Xeon
intel Gaudi
intel Data Center GPU Flex Series
intel Data Center GPU Max Series
intel Core Ultra
intel Arc Graphics
intel Ethernet

[“AI 无处不在”的客户故事]

让 AI 遍布
每个角落

在英特尔® 至强® 上运行 AI

云与企业

边缘

云与企业

第五代英特尔® 至强®：专为 AI 设计的处理器

第五代英特尔® 至强® 处理器的内核皆有 AI 加速功能，无需添加独立加速器，就能满足苛刻的端到端 AI 工作负载要求

更出色的 AI 推理性能

与上一代相比[1]

提升高达 42%

通用计算性能提升

与上一代相比[1]

平均 21%

更快的自然语言处理

与上一代相比[1]

提升高达 23%

英特尔执行副总裁兼数据中心与人工智能事业部总经理
Sandra Rivera

"我们的第五代英特尔® 至强® 处理器专为 AI 而设计，可为在云端、网络和边缘用例中部署 AI 功能的客户提供更高的性能。我们在经过验证的基础上，推出第五代英特尔® 至强®，有助于以更低的总体拥有成本，实现快速采用和扩展，这正是我们与客户、合作伙伴和开发人员生态系统长期合作的结果。

更多信息

网站

产品简介

有关工作负载和配置的信息，请访问 intel.com/processorclaims：第五代英特尔® 至强® 可扩展处理器。结果可能会有所不同。

英特尔® 至强®：在真实世界的 AI 应用中，CPU 性能遥遥领先

在实际工作应用中，英特尔为 AI 推理提供了性能更强、价格更低、更均衡的平台，从而颠覆了行业，实现了 AI 的民主化：

- 更大的高速缓存有助于数据局部性，更大的内存容量可解决更大的问题
- 更高的内核频率、多个标量端口和无序执行，有助于加速单线程或多线程标量计算
- 英特尔® 高级矢量扩展 512（英特尔® AVX-512），有助于非 DL 矢量计算
- 英特尔® 高级矩阵扩展（英特尔® AMX），内置 AI 加速硬件支持

技术全文　　信息图表

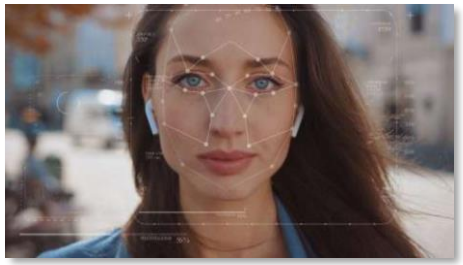
揭开 GPU 的神秘面纱：内置加速器的 CPU 如何彻底改变 AI

第四代英特尔® 至强® 可扩展处理器，带有用于 AI 推理的加速器

诸如英特尔® AVX-512 和英特尔® AMX 之类的加速器旨在提高性能、减少延迟并增加内存带宽，使其非常适合运行要求苛刻的推理 AI 工作负载

英特尔® Advanced Matrix Extensions（英特尔® AMX）

大幅加快深度学习训练和推理，非常适合自然语言处理、推荐系统和图像识别等工作负载



网站 | 解决方案简介
视频 | 用户指南

英特尔® Advanced Vector Extensions 512（英特尔® AVX-512）

可在端到端 AI 工作流程（如数据准备）中，加速经典机器学习和其他工作负载

网站 | 解决方案简介
视频 | 用户指南和下载

内置加速器以及为什么应该使用

云与企业

配备英特尔® AMX 的第四代英特尔® 至强® 可扩展处理器性能超越 AMD EPYC

利用更快、更个性化的 AI，推动收入增长，改善客户体验

- 更好地为业务决策提供信息，推动收入增长
- 改善客户保留和获取
- 提高互动，改善转化率
- 减少企业的重复性任务、 成本及时间

intel XEON 对比 EPYC AMD

[了解配备英特尔® Advanced Matrix Extensions（英特尔® AMX）的第四代英特尔® 至强® 可扩展处理器如何超越 AMD EPYC]

AI 工作负载：英特尔® 至强® 可扩展处理器上的 VMware

英特尔® Advanced Matrix Extensions（英特尔® AMX）

“您可以使用内置 AI 加速的 CPU，运行整个端到端 AI 流水线——数据准备、训练、优化、推理。”

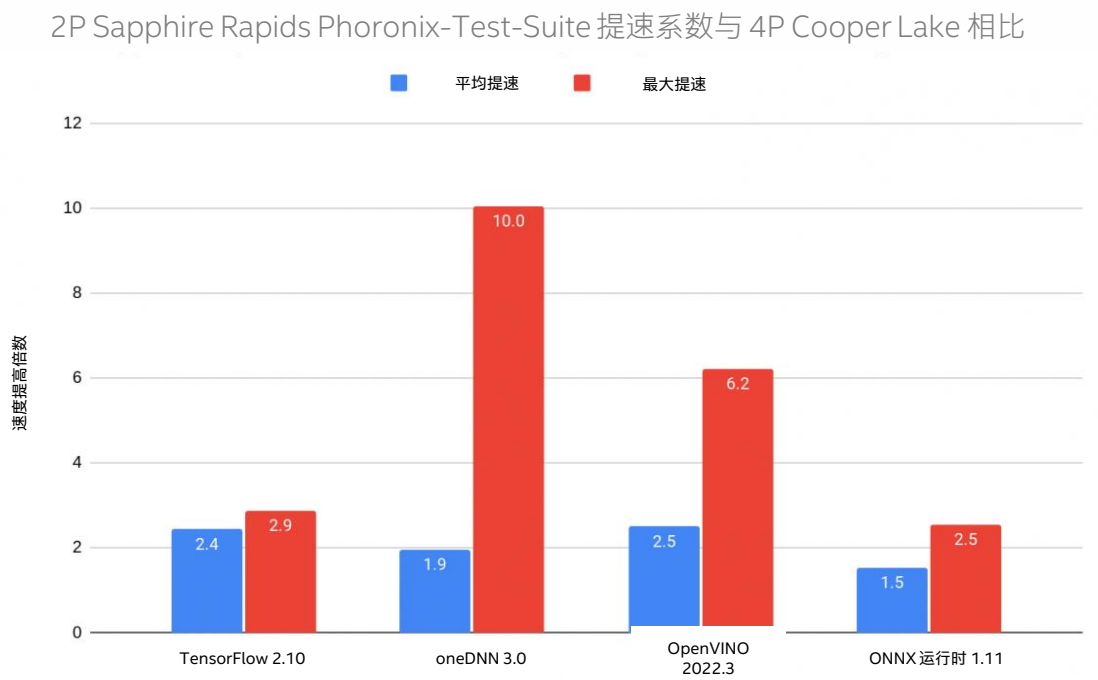
“要想提高 AI/机器学习工作负载的性能，您可以做的一件事是让 CPU 的 AMX 指令做部分 AI/机器学习工作，减少对昂贵且难以采购的 GPU 的需求。”

全文来自 VMware 主任工程师 Earl Ruby

AI 工作负载：红帽 英特尔® 至强® 可扩展处理器

红帽企业 Linux 利用以下装备，实现重大性能提升
第四代英特尔® 至强® 可扩展处理器

AMX



我们的结果显示，
第四代提速系数平均在
1.5 倍，最高可达 10 倍[1]

Red Hat

[1]https://www.redhat.com/en/blog/red-hat-enterprise-linux-achieves-significant-performance-gains-intels-4th-generation-xeon-scalable-processors

云与企业

案例研究

	挑战	解决方案/结果	英特尔产品	更多信息
Tencent Cloud 云计算服务搜索引擎	如何处理大规模查询并用搜索结果及时回复	腾讯可以使用优化的 BERT 模型，提供更好的服务体验，帮助降低 TCO	第四代英特尔® 至强® + 英特尔® AMX	案例研究
美团 Meituan 领先的零售技术公司	经济高效的视觉 AI 服务	美团将其在线资源的整体效率提高了 3 倍以上并节省了 70% 的服务成本	第四代英特尔® 至强® + 英特尔® AMX + 英特尔® IPP + 英特尔® Extension for PyTorch（英特尔® IPEX）	案例研究
SIEMENS 医疗图像处理	提高放射治疗专业人员的效率	利用基于 AI 的自动轮廓技术，为放射治疗专业人员提供支持，不仅提高工作负载效率，改善一致性，而且有助腾出员工专注于增值工作	第四代英特尔® 至强® + 英特尔® AMX® + OpenVINO™	案例研究 视频
Alibaba Cloud 领先的云计算提供商	提高地址净化服务的性能	更快速的端到端性能，为阿里巴巴在物流、电子商务、能源、零售和金融领域的客户带来更好的业务成果。使用内置加速器帮助阿里巴巴控制 TCO	第四代英特尔® 至强® + 英特尔® AMX® + 英特尔® oneDNN	案例研究

请参阅案例研究链接了解工作负载和配置。结果可能会有所不同。

关于英特尔 AI 技术的案例

FOX SPORTS

“我们减少了数周的设置时间”

“对于我们来说，英特尔® 至强® 处理器是我们部署技术的基石。我们只在英特尔® 至强® CPU 上运行，这使得我们能够随处运行：虚拟机中、专用的本地裸机中、云中。”

案例研究

SIEMENS

与上一代相比，自动轮廓算法的 AI 推理 时间提速[1] 35 倍

在能源消耗方面与上一代相比[2]，节省 20%

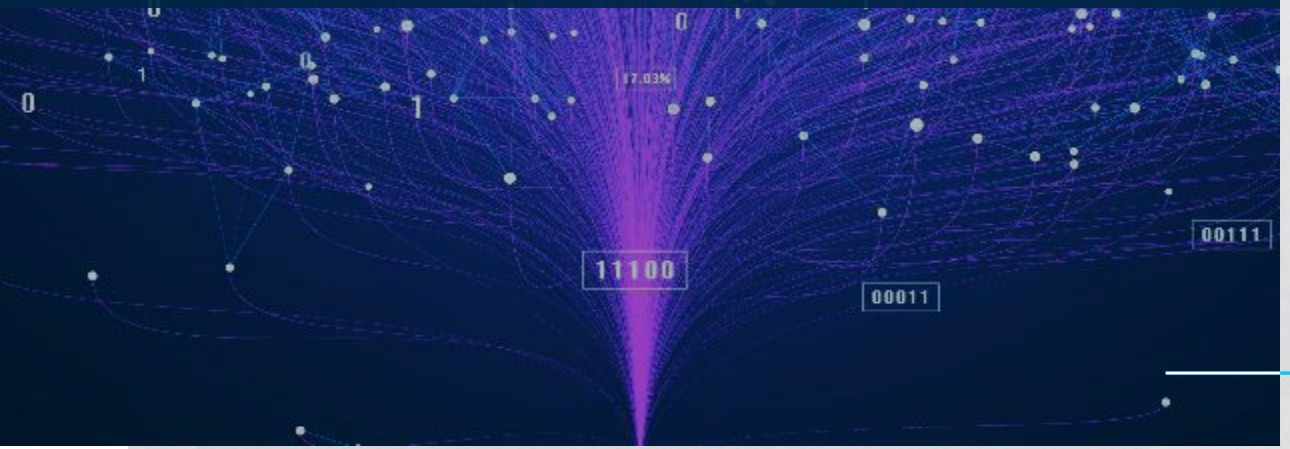
案例研究
视频

[1,2]请参阅 案例研究链接，了解工作负载和配置。结果可能会有所不同。

让 AI 遍布
每个角落

边缘 AI

将智能流程引入智能边缘，为您的业务创造真正的价值。



利用边缘计算与 AI 的合力，创造更好的业务成果，改善客户体验。

在边缘处理数据

到 2025 年，75% 的企业生成数据将在边缘创建和处理[1]

降低延迟

边缘计算在存储和处理前，
无需往返于云端，
从而缩短了洞察时间，
提高了效率

风险降低

边缘的数据在物联网设备本身存储和处理，能够对事件做出实时快速反应，从而更好地降低业务风险，增强安全性

降低成本

边缘计算将数据保存在边缘，
不仅存储和处理更具成本效益，
而且更快获取洞察力，
从而简化业务流程

边缘计算和处理为就地利用数据创造了机会

[1]来源：https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders

边缘

边缘 AI 跨行业实现变革性用例

教育

- 利用对课堂环境的关键洞察来改善学习环境
- 借助行为洞察制定更具吸引力的教学计划
- 通过基于 AI 的视频监控来改善校园安全

能源

- 通过基于 AI 的设备监控来降低环境影响
- 通过自动监控，降低能源成本

政府

- 优化人事管理
- 利用基于 AI 的视频分析增强建筑安全保障
- 减少能源浪费

医疗和生命科学

- 缩短获取洞察的时间，以方便诊断和医学检验
- 提高诊断准确性
- 提供更优质的患者护理

制造

- 改进质量控制流程
- 确保员工健康和安全
- 降低维护成本
- 启用预测性维护

零售

- 利用富有吸引力的个性化广告增加商店客流量
- 通过智能促销增加销量

交通运输

- 提高物流准确性和效率
- 通过智能包裹处理，降低运输成本，减少退货

酒店服务

- 改进餐厅客户体验并实现个性化
- 简化 QSR 订购和客户队列
- 优化食物制备并避免浪费

案例研究

- VSBLTY
- Convergint
- Taco Bell

英特尔正利用边缘 AI 跨不同环境来改进体验

边缘

借助内置的 AI 和安全功能加速关键边缘工作负载

第 4 代英特尔® 至强® 可扩展处理器性能与第 3 代相比

性能提升	利用英特尔® AMX，图像分类的 AI 推理性能提升	英特尔® AMX，物体检测的 AI 推理性能提升
1.33 倍	**3.01 倍**	**4.25 倍**

面向物联网边缘计算的**第四代英特尔® 至强® 可扩展**处理器具有出色的性能、内存、I/O 和资源可管理性，支持工作负载整合及**全新 AI 指令，实现边缘深度学习训练和推理**

[了解详情]

[AI 应用于生产的成功案例]

有关工作负载和配置的信息，请访问 intel.com/processorclaims：第四代英特尔® 至强® 可扩展处理器。结果可能会有所不同。

让 AI 遍布
每个角落

电脑上的 AI

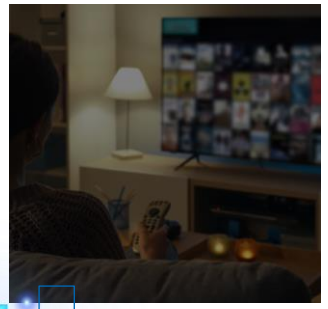
客户端和
工作站

客户端和
工作站

用例：电脑上的 AI

创作者：照片视频搜索和编辑

更快、更自然的过滤器、
更高质量的预览和更快的导出时间，
以及更快的自动搜索。



协作/直播

全新的 AI 功能，可用于下一代视频会议、直播和协作，延长电池续航时间。

主流游戏

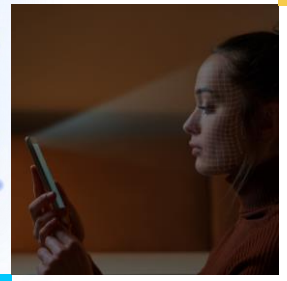
用于游戏内 3D 动画的
全新 AI 功能，增添了现实感、
转录和聊天翻译。

电脑上的 AI

”让平凡之事造就不平凡“

工作效率

用于写作、创作、编码和离线功能（例如文本和语法预测）的 AI 助理。



可用性

AI 辅助视听功能满足用户的不同需求，
让用户更轻松在电脑上创作，
提高工作效率。

创作者：文本转图像

全新的 AI 效果和功能，仅需几个
描述词（营销、广告、设计），
即可创建图像。

*英特尔 2023 年 12 月 14 日“广布 AI”活动新闻发布会

客户端和
工作站

英特尔® 酷睿™ Ultra 引领 AI 电脑时代

全新处理器是 AI 电脑的核心，在各种操作系统和应用程序中利用 AI 功能

首款基于英特尔 4
制程技术打造的处理器
40 年来最大的架构转变

内置英特尔锐炫™ GPU3，
最多提供 8 个 Xᵉ 内核
图形性能比上一代提升
高达 2 倍[1]

英特尔最新的 NPU，
英特尔® AI Boost，
专为低功耗长时间运行的
AI 工作负载而打造
与上一代相比，
功耗提升高达 2.5 倍[1]

英特尔执行副总裁兼客户端
计算事业部总经理 Michelle
Johnston Holthaus

“英特尔® 酷睿™ Ultra 的推出标志着，英特尔在电脑上启用 AI 的规模和速度无与伦比。到 2028 年，AI 电脑将占电脑市场的八成[2]，再加上我们庞大的软硬件合作伙伴生态系统，英特尔已经做好充分的准备，提供下一代计算。”

更多信息

网站

产品简介

[1]有关工作负载和配置的信息，请访问 intel.com/processorclaims：英特尔酷睿 Ultra 7 165H 性能。结果可能会有所不同。

[2] 来源：Boston Consulting Group

客户端和
工作站

案例研究：利用英特尔® 酷睿™ Ultra 处理器，推动患者护理

CPU 驱动的超声成像应用程序提供更易访问、更具成本效益的成像技术

情况

三星 Medison 是医疗保健创新领域的先锋。他们的超声成像应用使用 AI，实现最有效的患者护理。

挑战

此前，他们的应用程序运行于由竞争对手独立 GPU 加速的前代英特尔酷睿处理器。

解决方案

三星测试了内置 GPU 引擎的全新英特尔® 酷睿™ Ultra 处理器。他们发现，与上一代 CPU + dGPU 组合相比，AI 性能显著提升。三星 Medison 利用英特尔® 酷睿™ Ultra，可以在仅基于 CPU 的下一代超声波设备中提供高级 AI 功能。

获取
详情：
了解详情

客户端和
工作站

AI 电脑加速计划

AI 电脑加速计划旨在为独立硬件开发商 (IHV) 和独立软件开发商 (ISV) 提供英特尔资源，包括 AI 工具链、培训、联合工程、软件优化、硬件、设计资源、技术专业知识、联合推广和销售机会。

英特尔引领 AI 的发展方向

参与

100+

AI 独立软件供应商合作伙伴助力优化 AI 电脑

超过

300

2024 全年由独立软件供应商提供的 AI 加速功能

超过

100m

到 2025 年配备 AI 加速器的处理器数量

立即联系，了解详情！

ai.pc.acceleration.program@intel.com

在 Gaudi2 上运行 AI

Gaudi2：基础模型高效训练和推理的理想选择

Gaudi2 是专为深度学习的性能、效率和可扩展性而设计，满足 LLM (GPT) 和 GAI (Stable Diffusion) 等大规模基础模型的需求

要求	Gaudi2
速度	在训练和推理方面， 比 A100 **提速 1.5 - 2 倍**
内存	凡 Gaudi2 设备均具有 **96 GB 片上高带宽内存**，因而在内存上更容易适合大型基础模型，并大规模训练和部署模型
可扩展性	利用 **片上集成的 24x100GbE 端口**、服务器内 8 张卡片之间的直接全面连接，以及服务器内和服务器之间基于 ROCEv2 的开放通讯，提升效率
易用性	使用 SynapseAI、PyTorch 和 DeepSpeed，迁移或构建模型，只需**更改极少的代码**即可
功耗	**与 A100 相比，吞吐量/功率提高约 1.8 倍**
成本效益	基于特制的第一代 Gaudi 架构，相较于 Amazon 云上的 A100，**其性价比提升 40%**

来源：https://habana.ai/habana-claims-validation/

Gaudi2：加速生成式 AI 和大型语言模型

Gaudi2 深度学习加速器在深度学习训练和推理方面，表现出竞争力，
与 Nvidia A100 相比，
性能加快多达 2.4 倍[1]

GAUDI®2 对比 NVIDIA

[1] 新闻稿

Habana Gaudi2 和**第四代英特尔® 至强® 可扩展处理器**为 AI 训练提供领先的性能和**优化的成本节约**[2]

[2] 新闻室　　技术文章

英特尔线上研讨会录音，讨论了 **Habana® Gaudi®2** AI 处理器**在发挥生成式 AI 和大型语言模型 (LLM) 的潜力方面的尖端能力**

线上研讨会

英特尔 Gaudi AI 加速器利用 FP8 软件，在 GPT-3 上实现 2 倍性能飞跃

深度学习创新：英特尔、Habana Labs 和 Hugging Face

英特尔与 Hugging Face 持续合作的重点是扩大训练推理解决方案的采用，
而这些解决方案已经在最新的英特尔® 至强® 可扩展和 Habana Gaudi® 及 Gaudi®2 处理器上经过优化

habana

Hugging Face

该合作将英特尔® AI 工具套件中最先进的深度学习创新技术带入了 Hugging Face 开源生态系统，
并为未来英特尔® 架构中的创新驱动因素提供了参考

英特尔、Habana Labs 和 Hugging Face 推进深度学习软件

开始使用 Habana Gaudi 上的 Transformers

更快速训练和推理：Habana Gaudi®-2 对比 Nvidia A100 80GB 基准测试

云与企业

民主化 AI：英特尔、Habana Labs 和 Hugging Face

habana

Hugging Face

与 Nvidia A100 相比，
Habana® Gaudi®2 在 1760 亿参数模型
上运行推理，提速 20%[1]

与同等 A100 服务器相比， 在 Gaudi2 服务器上运行流行计算机视觉工作负载时，每瓦特吞吐量提升
1.8 倍[1]

阅读公告，并在此处观看座谈会：
迎接生成式 AI 的计算和可持续性挑战

播客

Hugging Face 和英特尔：推动实用、更快、民主化和合乎道德的 AI 解决方案

Twitter/X 对话

民主化大型语言模型如何推动 AI 开发

[1] 性能因使用、配置和其他因素而异；工作负载和配置详细信息，可访问：Gaudi2 HL-225H SYS-820GH-THR2 的 Supermicro L12 验证报告，2022 年 10 月 20 日

在英特尔® Data Center GPU Max Series 上运行 AI

云与企业

英特尔® Data Center GPU Max Series：突破性的性能

英特尔性能和密度最高的独立 GPU

英特尔的基础 GPU 计算构建模块特性：

- 高达 408 MB 的 L2 缓存基于独立 SRAM 技术，64MB 的 L1 缓存及高达 128GB 的高带宽内存
- 凡 Max 系列 GPU 内置高达 128 个光线追踪单元，用以加速科学可视化和动画
- 增强 AI 的英特尔® Xe Matrix Extensions (XMX) 具备深度收缩阵列，能够在单个设备中实现矢量和矩阵功能
- oneAPI 基于标准、多架构编程和工具，提升性能和生产力，克服专有编辑模型锁定
- 以下配置带来突出性能表现：
  - 在英特尔® 至强® Max CPU 上运行 LAMMPS（大规模原子/分子大规模并行模拟器）工作负载上，内核卸载到六个 Max 系列 GPU 并由英特尔 oneAPI 工具优化，与第三代英特尔® 至强® 处理器相比，性能提升高达 12.8 倍[1]

英特尔® Data Center GPU Max Series 专为 **AI 和高性能计算**所用的数据密集型计算模型实现**突破性能**而设计。英特尔® Max Series GPU 在 SoC 的构造方面能够**做到更大的灵活性和模块化。**

产品简介　网站　技术文章

oneAPI

整个英特尔 Max 系列产品家族经由 oneAPI 的统一，以提供通用、开放、基于标准的编程模型，进而发挥生产力和性能。

开发人员使用 oneAPI 优化的深度学习框架和机器学习库，可以实现数据分析和机器学习工作流程的插入式加速。

1 性能因使用情况、配置和其他因素而异。 工作负载和配置详情，请访问：产品简介

案例研究：英特尔® Data Center GPU Max Series 上的 Aurora 超级计算机

更快解决世界上最具挑战性的问题……

美国能源部阿贡国家实验室 (ANL) 的 Aurora 超级计算机预计将成为业界首批具有超过十亿亿次持续双精度性能和超过二十亿亿次峰值双精度性能的超级计算机之一。**Aurora 还将率先展示在单一系统中集 Max 系列 GPU 与 CPU 于一体所具有的强大性能**，共有超 10,000 个刀片，各含 6 个 Max 系列 GPU 和两个至强® Max CPU

用于机器学习的 Aurora 刀片演示视频

行为召唤

行为召唤

教育

了解英特尔技术如何满足您的 AI 需求，以及英特尔® 至强® 产品线在哪些领域帮助您拓展业务

利用 AI 训练资料，
[了解详情]

互动

从技术领域会议开始

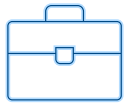
要安排技术披露会，
请发送电子邮件到：
cloud.insider.program@intel.com

英特尔® 合作伙伴联盟如何提供帮助

开始使用英特尔® 合作伙伴联盟

成为英特尔® 合作伙伴联盟会员，即可获得专属业务发展机会，
例如进入我们的全球市场、高级培训和促销支持，
所有这些均根据您的需求定制

培训和能力

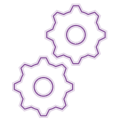


加入英特尔® 合作伙伴培训计划可获得有关高级技术、能力课程计划和学习奖励方面的专业培训

营销资源

进入英特尔® 解决方案市场和英特尔® 合作伙伴营销工作室，有助为您的产品和服务创造更多需求

有价值的奖励



通过参加符合条件的活动赚取积分，提升您的会员等级，并访问其他资源以拓展业务

如果还不是会员
立即加入

会员享有的权益

赢取积分

英特尔® 合作伙伴联盟中最受欢迎的独有权益之一，我们将向合作伙伴奖励积分，以对他们与英特尔一起取得的业绩以及积极参与高优先级活动做出表彰。
在英特尔合作伙伴联盟内，会员有 1,000 多种方式赚取积分，并有数百种机会兑换奖品。

Cloud Insider 社区

英特尔® Cloud Insider 社区可提供不断更新的世界一流的云内容和工具。会员有机会与同行和生态系统建立联系，将创新型联合云解决方案推向市场

了解详情

行业洞察

黄金会员和钛金会员可访问精心策划的季度行业洞察，以帮助推动自身增长

了解详情

经济激励措施

会员可获得市场开发基金和激励计划的强大支持，以帮助您快速成功地进行产品营销

与您的英特尔代表交谈，了解英特尔® 合作伙伴联盟加速器计划以及更多财务奖励

资源

如何访问英特尔® 合作伙伴联盟客户支持

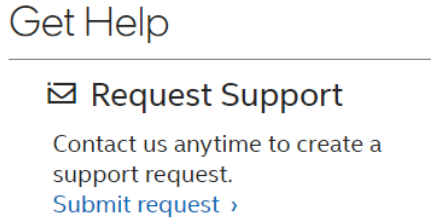
英特尔 Virtual Assistant

这个聊天机器人位于每个合作伙伴联盟网页的右下角，提供可解答大多数问题的自助服务或实时支持代理的快速链接。

“获取帮助”大型条幅广告

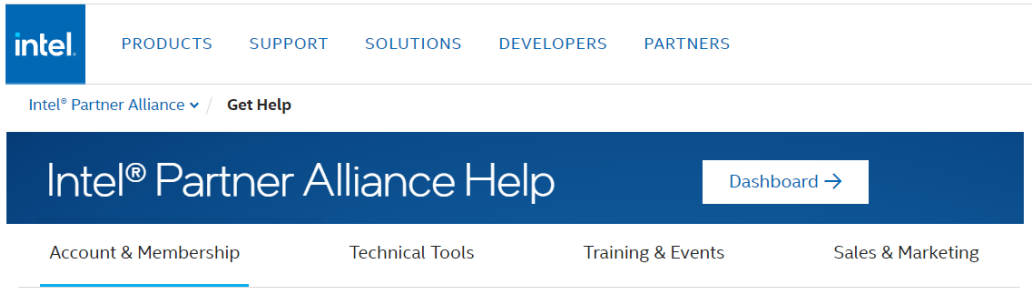
提交在线支持请求。

此链接位于合作伙伴联盟网站内大多数页面的页脚处。



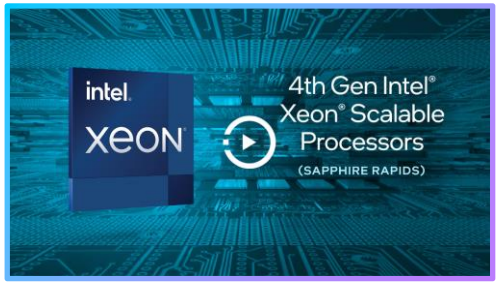
合作伙伴联盟“获取帮助”页面

获取帮助页面提供了有关合作伙伴联盟成员可用的大多数工具和权益的详细自助指南。



Cloud TV

英特尔® Cloud TV 将提供云计算新闻、趋势和战略，助力您取得成功



Sapphire Rapids
在云端

利用云
强化 AI

迈上快速路径，
随时随地扩展 AI

使用云技术的
AI 推理

云端 AI

采用 AI 的第四代英特尔® 至强® 可扩展处理器

信息和资源

产品简介

第四代英特尔® 至强® 可扩展处理器

面向英特尔® 至强® CPU 的英特尔 AI Engines 提升整个 AI 流水线的性能

技术论文

利用机密计算加速 AI 推理

第四代英特尔® 至强® 上的可扩展端到端企业 AI

利用技术创新者和英特尔® 技术，简化您的 AI 计划

信息图表

以经济高效的方式，快速部署高性能 AI

更快的 AI 投资回报率

案例研究

富士通 | 西门子 | BCM | ai.io

视频

英特尔 AI 流水线视频

英特尔® AMX：AI 的下一个重大步骤

英特尔 AI 加速器视频

第四代英特尔® 至强® 云 AI 视频

公文包

利用技术创新者和英特尔® 技术，简化您的 AI 计划

更多资源

性能指标

第四代英特尔® 至强® 可扩展处理器

目录

AI 推理软件和解决方案目录

其他培训

幻灯片内的在线培训链接

业务报告

2022 年 AI 技术成熟度曲线

在数字优先经济中实现数字化转型：成为人工智能的颠覆者

第四代英特尔® 至强® 可扩展处理器，准备好加速数据中心的性能和功能

内部

合作伙伴资源指南

将英特尔的外部可用资源页面整合到单个参考文档中，用于上市和进入市场活动

目的： 与 IPA 合作伙伴和 CNDA 下的合作伙伴一起使用此演示文稿，共同推动英特尔的整体参与

目标受众：PSAM 和 SDM

访问 PRG：goto/resourceguide（内部访问）

内部资源

非公开内容

资料类型	标题和链接
Gold Deck	AI Gold Deck
Gold Deck	SPR AI Gold Deck
销售演示文稿	英特尔 Gaudi 非 NDA 销售演示文稿
开发人员资源	英特尔 Gaudi 开发人员资源
快照	Max 系列 GPU
Gold Deck	AMX 深度探讨演示文稿
销售维基	英特尔® 至强® 按需销售维基
销售公文包	第四代英特尔® 至强® 优先用例及价值主张
推介卡	英特尔® AMX 推介卡
销售公文包	利用第四代英特尔® 至强® 处理器加速，发挥工作负载性能
销售加速仪表板	SPR 销售加速

培训资产

AI 培训资产

人工智能
人工智能：利用第四代英特尔® 至强® 处理器加速工作负载 全部
深入探讨使用 Fortanix Confidential AI 确保按需 AI 工作负载 开发运维人员、云架构师
为什么选择云端的英特尔 AI? 开发运维人员、云架构师
AI 云部署选项 云架构师、高管
云服务提供商 (CSP) AI 产品组合 云架构师、高管
实现从数据中心到边缘的 AI 性能 开发运维人员、云架构师
第四代英特尔® 至强® 平台简介 全部

法律通告与免责声明

通告和免责声明.


intel

备份

软硬件的合力

查看合力幻灯片

系统主板	英特尔® 服务器主板 S2600STB	M50CYP2SB1U Coyote Pass	英特尔公司/阿彻城
CPU	英特尔® 至强® Platinum 8270 CPU @ 2.7 GHz	英特尔® 至强® Platinum 8380 CPU @ 2.3 GHz	英特尔® 至强® Platinum 8490H @ 1.9 GHz
插槽，物理内核/插槽	2, 26	2, 40	2, 60
超线程/睿频设置	已启用/打开	已启用/打开	已启用/打开
内存	12x16GB DDR4 2933MHz	16x16GB DDR4 3200MHz	16x16GB DDR5 4800MHz
操作系统	UB-18.04 LTS	UB-22.04 LTS	UB-22.04 LTS
内核	5.3.0-24-generic	5.19.0-38-generic	5.19.0-41-generic
软件	英特尔® 发行版 OpenVINO™ 工具套件 2021.4	英特尔® 发行版 OpenVINO™ 工具套件 2022.3	英特尔® 发行版 OpenVINO™ 工具套件 2023.0
BIOS	SE5C620.86B.02.01. 0013.121520200651	SE5C620.86B.01.01.0006.2207150335	EGSDREL1.SYS.9409.P31.2302280828
BIOS 发布日期	2020 年 12 月 15 日	2022 年 7 月 15 日	2023 年 2 月 28 日
BIOS 设置	选择优化的默认设置，保存并退出	选择优化的默认设置，保存并退出	选择优化的默认设置，保存并退出
测试日期	2021 年 6 月 18 日	2023 年 6 月 20 日	2023 年 5 月 25 日
精度和批量大小	int8/批次 1	int8/批次 1	Int8/批次 1
推理请求数	52	80	120
执行流数	52	80	120
功率 (TDP)/插槽	205W	270W	350W

工作负载（模型：输入 HxW）：
Inception-v4: (299x299); Resnet-50: (224x224); Unet-camvid-onnx-0001: (368x480); Yolo-v3-tiny: (416x416)

最大化
价值

加速创新

了解如何优化开放标准工具的性能

更快构建

优化的预训练模型

- LLM - Bloom
- LLM – Llama 2
- 对象检测
- 对象识别
- 分割
- 图像处理
- 文本检测与处理
- 文本转语音

下载 LLM >　　下载 OpenVINO™ >

利用熟悉的模型来优化

行业标准框架

PyTorch
TensorFlow
scikit learn
XGBoost
MODIN
PaddlePaddle

下载英特尔® AI 工具套件 >

变得更轻松

生态系统解决方案

snowflake
Hugging Face
HAZELCAST
ahana
ANACONDA.
EROSPIKE
CLOUDERA
C3.ai
intel GeTi
AIBLE

随处部署

整个企业从概念到生产的速度更快

利用

端到端 AI 流水线软件构建

工程师数据 → 创建模型 → 优化和部署

利用 OpenVINO™

一次编写，随处部署

OpenVINO™

优化以部署

在任意架构

数据中心　云　客户　边缘应用设备

简化 AI 工作流程

MODIN · SciPy · dmlc XGBoost · PyTorch · ONNX RUNTIME · OpenVINO · Hugging Face

pandas · NumPy · scikit learn · TensorFlow · DirectML · cnvrg.io

Spark · Numba · SigOpt AutoML · intel Neural Compressor

开放式

julia · SYCL · Java · OpenCL · OpenXLA · OpenMP · C · C++ · oneAPI

高效

解决方案	预配置容器	AI 工具选择器
工具	优化的扩展	OpenVINO™
	Modin	CNVRG
参考资料	AI 参考套件	Hugging Face 协作

可访问性

生态系统参与

业界与学界	解决方案市场	高触摸支持

开发人员培训

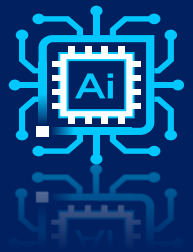
MLOPS 训练	卓越中心	文档和教程
培训视频	峰会和黑客马拉松	起飞计划

利用参考套件加速 AI 开发

好处

使用开源机器学习套件加速创新

发布到开源社区的 AI 模型已经过设计、训练并在数千个模型中测试，从而发布最适合用例的一个模型。数据科学家可以使用其行业的数据进一步定制和微调模型。



专为机器学习流水线设计和优化

每个参考套件都包括一个用户指南，用于加速企业中的 AI 部署，包括：

- 数据摄取
- 数据预处理
- 机器学习建模
- 超参数调优
- 模型投入使用和部署
- 基准测试

使用更少的计算资源构建更多模型

所有 AI 模型均利用英特尔库、框架和工具进行了优化，以满足 AI 开发需求，并在 oneAPI 的支持下利用更少计算资源提高训练速度和推理性能。AI 参考套件使用英特尔 AI 软件组合中的组成部分，包括英特尔® AI Analytics Toolkit 和英特尔® 发行版 OpenVINO™ 工具套件。

深入了解

AI 参考套件库

AI 工作负载：
Hugging Face 训练/推理基准测试

Hugging Face

[了解如何利用在 AWS 上运行的第四代英特尔® 至强® 服务器集群，加速 PyTorch 训练作业]

[查看 AWS 上逐代的 AI 推理改进]

与上一代英特尔® 至强® 相比，速度提高了 8 倍，并具有近线性扩展[1]

得益于[第 4 代英特尔® 至强®] 和 Hugging Face Optimum 的组合，您只需对自己的代码稍作修改，即可令预测提速 3 倍[2]

Hugging Face

第四代英特尔® 至强® CPU 具有卓越的推理性能，特别是在与 "Hugging Face Optimum"技术相结合时。这是在使深度学习更容易访问、更具成本效益的道路上迈出的又一步。[2]

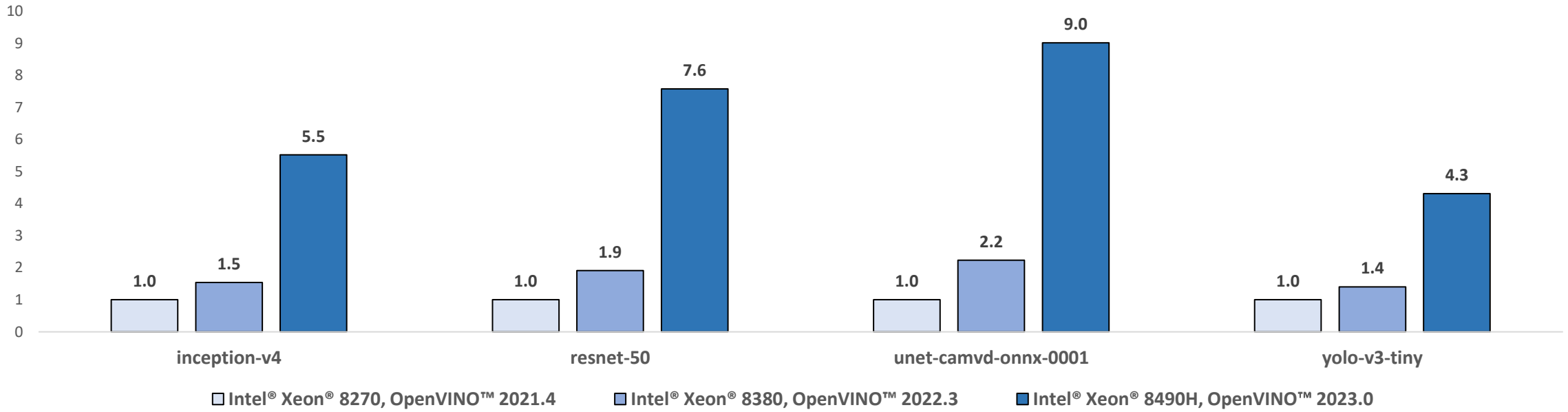
[1] https://huggingface.co/blog/intel-sapphire-rapids [2] https://huggingface.co/blog/intel-sapphire-rapids-inference

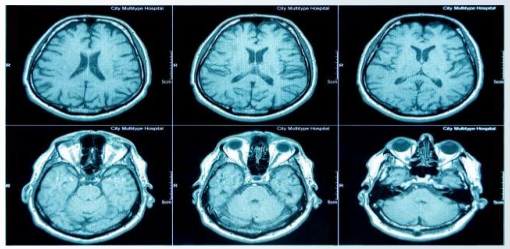
生态系统投资

英特尔投资于持续改进 AI 推理能力

采用英特尔® Advanced Matrix Extensions 的 OpenVINO™ 以指数方式改进 AI 推理性能

三代英特尔® 至强® CPU 对比（精度：INT8，批量大小：1，越大越好）





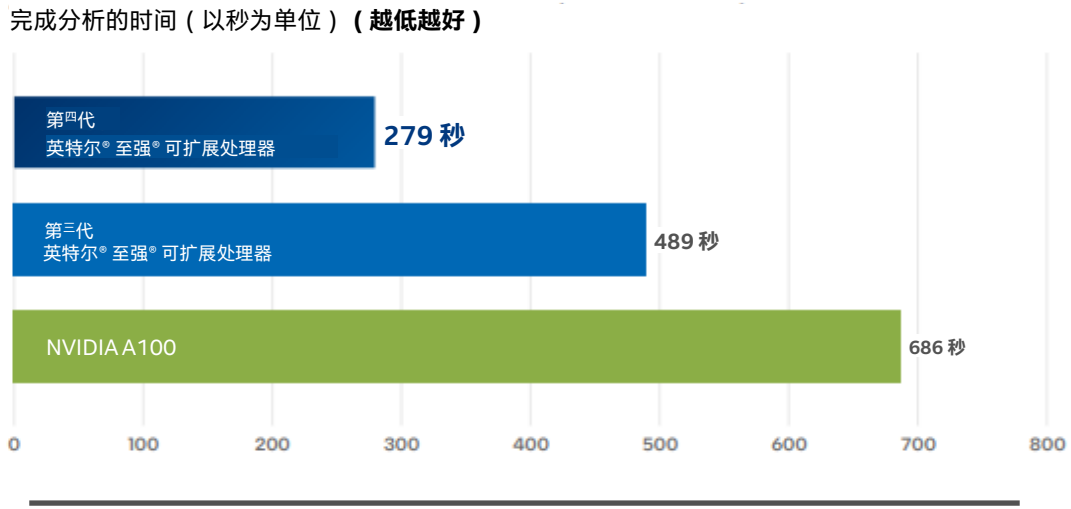
请参阅备份以了解系统配置详细信息、工作负载和定价。结果可能会有所不同。
AI 工作负载包括图像分类、高分辨率语义分割和物体检测。

云与企业

单细胞基因组学：提高英特尔® 至强® 可扩展处理器的性能

1.3M 单细胞基因组分析

完成分析的时间（以秒为单位）**（越低越好）**



在面向 CPU 优化的软件上运行第三代英特尔® 至强® 可扩展处理器，性能**比 NVIDIA A100 GPU 提速多达 1.4 倍**[1]

在面向 CPU 优化的软件上运行第四代英特尔® 至强® 可扩展处理器，性能**比 NVIDIA A100 GPU 提速多达 2.5 倍**[2]

CPU 性能优势

更快的端到端机器学习性能

NVIDIA 最近发布了一份对 130 万个单元的分析报告，对比了 NVIDIA A100 GPU 与英特尔® 至强® 处理器的性能。测试表明，GPU 的性能是 CPU 性能的 30 倍。
我们使用更好的并行算法并根据底层架构调整性能，再配合英特尔® oneAPI Data Analytics Library (oneDAL) 和 Katana Graph 来加速流水线之后，可以证明，经优化的**英特尔® 至强® 可扩展处理器具有 1.4 倍-2.5 倍的性能优势。**

1,2 性能因使用情况、配置和其他因素而异。工作负载和配置详情，请访问：基准测试简介

提高 AI 推理性能

Tencent Cloud

2-3 倍 在使用英特尔® AMX 的情况下，腾讯搜索应用使用的 BERT 模型的 AI 吞吐量对比上一代

腾讯可以使用优化的 BERT 模型提供更好的服务体验并帮助降低总拥有成本

[案例研究]

美团 Meituan

3.4 倍 美团的计算视觉平台，使用和不使用英特尔® AMX 进行 Bfloat16 优化时的 AI 吞吐量对比

美团将其在线资源的整体效率提高了 3 倍以上，并节省了 70% 的服务成本

[案例研究]

vmware®

5.7 倍 自然语言处理 (NLP) vSphere/vSAN 8.0，使用第四代英特尔® 至强® 可扩展处理器配合英特尔® AMX

英特尔提供广泛的开放式和免费使用的工具、优化库和行业框架，以提供最佳的开箱即用性能和端到端生产效率

[案例研究] [文章和演示]

请参阅案例研究链接了解工作负载和配置。结果可能会有所不同。

英特尔® 至强® 可扩展处理器支持增长最快的工作负载的 5 种途径

1. 在现有硬件上更快运行工作负载
2. 提高各种工作负载的性能
3. 降低功耗
4. 增强对最敏感数据的保护
5. 为您的数据最密集型工作负载，增加更多处理能力

查看完整列表文章：此处