产品简介

AI 加速
英特尔® 至强® 可扩展处理器

英特尔® AI 引擎专为英特尔® 至强® CPU 打造，全面提升 AI 流水线性能

70%
的数据中心 AI 推理都在英特尔® 至强® 可扩展处理器上运行[1]

从数据预处理、经典机器学习，到语言处理和图像识别等深度学习模型，AI 的身影到处可见，遍及广泛的工作负载和用例。配备英特尔® AI 引擎的英特尔® 至强® 可扩展处理器，结合了可服务整条 AI 流水线的强大算力，以及面向机器学习、数据分析和深度学习等特定 AI 工作负载的内置加速器。

内置强大动力，助力企业实现 AI 加速

无处不在的 AI 已遍及各种各样的关键工作负载。从核心企业应用到自动话务台系统，经典的机器学习（ML）和深度学习模型正在成为企业实现业务发展的基础构建模块。AI 能否大规模应用取决于从数据预处理到训练，再到最终部署这一系列漫长的开发流程。每个步骤又有自己的开发工具链、框架和工作负载，这些都会产生特有的瓶颈，对计算资源的要求也不同。英特尔® 至强® 可扩展处理器配备内置加速器，可在开箱后立即运行整个流水线，全面提升 AI 性能。英特尔® 加速引擎是为特定功能打造的内置加速器，用于支持要求严苛的新兴工作负载。

借助英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）加速深度学习

第四代英特尔® 至强® 可扩展处理器配备的英特尔® AMX 是面向深度学习训练的新一代技术。英特尔® AMX 不仅进一步拓展了上一代英特尔® 至强® 可扩展处理器的内置 AI 加速技术，还带来显著的性能增益，非常适合自然语言处理、推荐系统和图像识别等工作负载[2]。

英特尔® AMX 可为 AI 模型提供工作负载加速，并通过将特定 AI 工作负载整合至 CPU，而非将其卸载至独立加速器的方式，帮助客户降低总体拥有成本（TCO）[3]。

与 CPU 内核上的英特尔® 高级矢量扩展 512（Intel® Advanced Vector Extensions 512，英特尔® AVX-512）相比，英特尔® AMX 还可提高平铺乘法性能，显著提升最大吞吐量（单个周期运算量）[4]。

优化自然语言处理和推荐系统

第四代英特尔® 至强® 可扩展处理器和英特尔® AMX，无需增配其他硬件即可为自然语言处理带来显著的性能提升。多个库已集成至 TensorFlow 和 PyTorch，开发人员无需进行额外操作即可利用内置 AI 加速技术的诸多优势。开发人员还能轻松地从不同的硬件环境迁移代码，从而节省大量时间和成本。

客户成功案例：基于英特尔® 至强® 可扩展处理器实现真实场景下的加速

腾讯云借助第三代英特尔® 至强® 可扩展处理器实现实时语音合成。

了解详情 ›

欧洲核子研究组织（CERN）拥有世界上规模最大的粒子加速器。该组织利用内置英特尔® DL Boost 在不影响准确性的前提下实现了推理加速。

阅读全文 ›

通过加速深度学习推理和训练，配备英特尔® AMX 的第四代英特尔® 至强® 可扩展处理器可在平衡 TCO 的前提下提供定制化用户体验。借助能够将用户实时行为以及时间和地点等相关场景特征考虑在内的深度学习推荐系统，第四代英特尔® 至强® 可扩展处理器即可实现上述目标。

第四代英特尔® 至强® 可扩展处理器协同加速引擎驱动未来创新

无论是将英特尔® 至强® 可扩展处理器用于处理本地工作负载，还是处理云端或边缘工作负载，英特尔® 加速引擎都能够助力您的业务达到新高度。这些加速引擎具备一系列优势，包括安全性方面的处理速度更快，数据保护力更强以及基础设施利用得更充分。

英特尔® 加速引擎还有助于提高虚拟和物理 CPU 利用率，同时降低每核的解决方案许可费用。

除此之外，这些内置加速器还能够提高应用性能，降低成本并提升平台层面的效率。

英特尔® 高级矢量扩展 512（英特尔® AVX-512）是加速机器学习的强大利器

英特尔® 至强® 可扩展处理器的内核可以使用哈希算法对网站进行 SSL 加密，处理海量数据库，以及针对药物研究、芯片设计或一级方程式赛车引擎运行仿真。它们虽然全能，但需要借助 AVX-512 加速器才能更快完成深度学习训练工作负载。

英特尔® AVX-512 经过多代升级，使英特尔® 至强® 可扩展处理器能够在每个时钟周期内进行更多操作，并提供可与并行处理比肩的出色性能。英特尔® AVX-512 扩展技术属于指令集，会告诉 CPU 做什么以及如何做。它们的工作原理很复杂，但基本逻辑非常简单。首先，尽可能将多个步骤压缩为更少的运算。其次，帮助 CPU 在每个时钟周期内执行更多运算。

步骤越少意味着处理速度越快

数学计算可以很聪明，也可以很优雅。英特尔® AVX-512 使用大量聪明、简便的数学计算将常见的计算运算压缩、组合、融合到更少的步骤中。举个简单的例子：您可以指示 CPU 执行 3x3x3x3x3 这样的计算，这个计算过程需要五个时钟周期。或者您可以创建一条 3^5 指令，使 CPU 能在一个周期内完成计算。AVX-512 采用的就是这种逻辑，并将其应用于数百个针对具体工作负载的运算，包括 AI 中一些极其复杂的运算。

位数越多，处理速度越快

AVX-512 中的“512”指的是第二种方式，这些指令增加了 CPU 在每个时钟周期能够处理的位数。四十年前，16 位 PC 是主流，但很快就被 32 位设备取代。如今，智能手机的运行位数达到 64 位。位数指的是寄存器的数量。寄存器是 CPU 在每个时钟周期内可以寻址的 CPU 存放数据的内存插槽。AVX-512 将寄存器的数量扩展到 512 位。当应用利用英特尔® AVX-512 时，只需扩展寄存器数量，就可以使运行速度比 CPU 的基础 64 位快高达 8 倍，这就好像是从 1 一直数到 96 与 8、16、24 这样按 8 的倍数数到 96 的对比。

英特尔® 深度学习加速技术（Intel® Deep Learning Boost，英特尔® DL Boost）是更聪明的神经网络数学计算

训练深度学习模型可能需要数小时或数天的算力。而深度学习推理可能需要几分之一秒到几分钟，具体取决于模型的复杂程度和对结果的准确度的要求。当训练或推理扩展到数据中心级计算时，时间、能耗和性能预算会显著上浮。

英特尔® DL Boost 使用多条英特尔® AVX-512 指令，支持 INT8 和 BF16 数据类型，可加速深度学习工作负载。它将三个运算合并成一个矢量神经网络指令（VNNI）集，从而减少了每个时钟周期的运算量，同时充分发挥英特尔® 至强® 可扩展处理器的计算潜能。VNNI 可通过使用 INT8 精度来加速深度学习（DL）推理。

第四代英特尔® 至强® 可扩展处理器的推出也势必为性能带来更大提升。在英特尔® AMX 和 AVX-512 的协同助力下，第四代英特尔® 至强® 可扩展处理器与第三代英特尔® 至强® 可扩展处理器相比，前者执行平铺乘法运算时的最大吞吐量（单个周期运算量）更高[5]。

更低功耗的引擎运行更强大的 AI 工作负载

由于英特尔® 至强® 可扩展处理器配备英特尔® AI 引擎，所需的硬件资源更少，可为运行 AI 工作负载提供更强大、更节能的解决方案。

英特尔® 至强® 可扩展处理器配备内置加速引擎，可实现更出色的工作负载成果，例如降低当下要求严苛的 AI 工作负载的总体拥有成本（TCO）并提高其投资回报（ROI）[6]。

平均算下来，采用英特尔® 至强® 可扩展处理器的系统比需要集成 GPU 的同类系统成本低 17%[7]。

英特尔® 至强® 可扩展处理器几乎是自动为 AI 加速

英特尔® 至强® 可扩展处理器的 AI 加速技术内置于 CPU 的指令集架构（ISA）中，这意味着它可以随时用于任何与之兼容的软件。英特尔软件工程师正在不断优化开源 AI 工具链，并将这些优化传递回社区。例如，TensorFlow 2.9 出货时默认附带英特尔® oneAPI 深度神经网络库（Intel® oneAPI Deep Neural Network Library，英特尔® oneDNN）优化。下载最新版本 TensorFlow，它会自动应用英特尔的优化方案[8]。

对于 AI 流水线中的其他应用，数据科学家和开发人员可以下载免费的开源英特尔® 分发版工具、库和开发环境，它们可以利用英特尔® 至强® 可扩展处理器指令集架构中的各个内置加速器。

这样一来，数据科学家和 AI 开发人员就无需专门就英特尔® AVX-512 对自己的工具重新编码和编译，因为我们已经为他们做了这个工作。

当前，企业和机构需要从自身的基础设施中获得更多的工作负载性能，并以更加节能和经济的方式实现这一目标。英特尔® 至强® 可扩展处理器的专用英特尔® AI 加速引擎能助您让自身业务中关键 AI 工作负载尽可能多地发挥价值。

了解内置英特尔® 加速引擎的英特尔® 至强® 可扩展处理器还可为您业务中关键 AI 工作负载带来怎样的强大动力。

了解更多信息

基于英特尔® 至强® 可扩展处理器的 AI 和深度学习 ›

英特尔® AVX-512 ›

英特尔® 深度学习加速技术 ›

英特尔® AI 分析工具套件 ›

第四代英特尔® 至强® 可扩展处理器的 AI 加速

加速深度学习 AI 工作负载

与上一代产品相比，第四代英特尔® 至强® 可扩展处理器凭借英特尔® AMX，在使用 SSD-ResNet34 进行深度学习推理时，AI 工作负载速度提升高达 3 至 5 倍；在使用 ResNet50 v1.5 进行训练时，速度提升高达 2 倍[9]。

想要立即在云端或在自有基础设施上加速 AI 工作负载？英特尔面向 AI 和机器学习的优化方案可以帮到您。

了解更多信息 ›

[1] 基于英特尔对截至 2021 年 12 月运行 AI 推理工作负载的全球数据中心服务器装机容量的市场建模。

[2] 采用英特尔® AMX（BF16）的推理性能：性能预测基于非量产的双路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids），56C，350W TDP，共配置 1 TB（8 通道/64 GB/4800）的 DDR5 内存，使用 BKC 46，采用英特尔® AMX/int8 和 BF16，CentOS Stream 8，经 oneDNN 优化的英特尔 AMX 内核，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake），28C，250W（8380H）。由于推理性能在不同路数处理器上的结果呈线性扩展，双路处理器的测试数据为八路处理器的测试结果乘以 0.25；配置：单节点，8 个第三代英特尔® 至强® 铂金 8380H 处理器（28C, 250W），基于英特尔参考平台（之前代号为 Cooper City），总内存 384 GB（48 个插槽/64GB/2933），ucode 0x7002302，启用超线程，启用睿频，Ubuntu 20.04 LTS，Linux 5.4.0-29-generic，英特尔® 固态盘 800 GB 操作系统驱动程序；测试结果可能不同。基于英特尔于 2022 年 1 月 27 日进行的测试。对象检测（RT）：采用 SSD-RN34，BS=1，56，BF16，内部版面向英特尔® 架构优化的 TensorFlow 2.8，Squad 1.1 数据集。采用英特尔® AMX（BF16）的训练性能：性能预测基于非量产的单路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids）在 ResNet-50 v1.5 上进行深度学习训练，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake）。

[3] 采用英特尔® AMX（BF16）的推理性能：性能预测基于非量产的双路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids），56C，350W TDP，共配置 1 TB（8 通道/64 GB/4800）的 DDR5 内存，使用 BKC 46，采用英特尔® AMX/int8 和 BF16，CentOS Stream 8，经 oneDNN 优化的英特尔 AMX 内核，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake），28C，250W（8380H）。由于推理性能在不同路数处理器上的结果呈线性扩展，双路处理器的测试数据为八路处理器的测试结果乘以 0.25；配置：单节点，8 个第三代英特尔® 至强® 铂金 8380H 处理器（28C, 250W），基于英特尔参考平台（之前代号为 Cooper City），总内存 384 GB（48 个插槽/64GB/2933），ucode 0x7002302，启用超线程，启用睿频，Ubuntu 20.04 LTS，Linux 5.4.0-29-generic，英特尔® 固态盘 800 GB 操作系统驱动程序；测试结果可能不同。基于英特尔于 2022 年 1 月 27 日进行的测试。对象检测（RT）：采用 SSD-RN34，BS=1，56，BF16，内部版面向英特尔® 架构优化的 TensorFlow 2.8，Squad 1.1 数据集。采用英特尔® AMX（BF16）的训练性能：性能预测基于非量产的单路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids）在 ResNet-50 v1.5 上进行深度学习训练，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake）。

[4] https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/vision-2022/，第 [41] 和 [42] 项基准测试。结果可能不同。

[5] https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/vision-2022/，第 [41] 和 [42] 项基准测试。结果可能不同。

[6] 与上一代产品 (fp32) 相比，配备英特尔® AMX (bf16) 的第四代英特尔® 至强® 可扩展处理器在使用 Hugging Face 进行文档级情感分析 (DLSA) 时，端到端实时推理性能加速可高达 6 倍。

新配置：单节点，2 个英特尔® 至强® 铂金 8480+ 处理器，平台内存配置 1024 GB DDR5，微代码：0x2b000041，禁用超线程，启用睿频，Ubuntu 22.04.1 LTS，5.15.0-47-generic，1 个 1.92 TB 英特尔® NVMe 固态盘，基于英特尔于 2022 年 9 月 8 日进行的测试。

基准配置：单节点，2 个英特尔® 至强® 铂金 8380 处理器，Ubuntu 22.04.1 LTS, BIOS 版本：WLYDCRB1.SYS.0021.P25.2107280557，禁用超线程，启用睿频，5.15.0-47-generic，微代码：0xd000363，512 GB RAM (16 x 64 GB 3200 Mt/s)，1 个 1.92 TB 英特尔® NVMe 固态盘，基于英特尔于 2022 年 8 月 31 日进行的测试。

软件配置：英特尔面向 PyTorch 的扩展程序 (IPEX)：v1.13.0+cpu，Transformers v4.21.0，深度学习模型：Bert-large-uncased https://huggingface.co/bert-large-uncased，4 项调优实例：20 (Ice Lake) 项和 28 (Sapphire Rapids) 项推理实例，数据集：IMDB (25K 用于调优，25K 用于推理)，批量大小：256（IMDB 数据集）/1024（SST-2 数据集），序列长度：512（IMDB 数据集）

[7] 详情请见以下网址的 [100]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/

[8] 采用英特尔® AMX（BF16）的推理性能：性能预测基于非量产的双路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids），56C，350W TDP，共配置 1TB（8 通道/64 GB/4800）的 DDR5 内存，使用 BKC 46，采用英特尔® AMX/int8 和 BF16，CentOS Stream 8，经 oneDNN 优化的英特尔 AMX 内核，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake），28C，250W（8380H）。由于推理性能在不同路数处理器上的结果呈线性扩展，双路处理器的测试数据为八路处理器的测试结果乘以 0.25；配置：单节点，8 个第三代英特尔® 至强® 铂金 8380H 处理器（28C, 250W），基于英特尔参考平台（之前代号为 Cooper City），总内存 384 GB（48 个插槽/64GB/2933），ucode 0x7002302，启用超线程，启用睿频，Ubuntu 20.04 LTS，Linux 5.4.0-29-generic，英特尔® 固态盘 800 GB 操作系统驱动程序；测试结果可能不同。基于英特尔于 2022 年 1 月 27 日进行的测试。对象检测（RT）：采用 SSD-RN34，BS=1，56，BF16，内部版面向英特尔® 架构优化的 TensorFlow 2.8，Squad 1.1 数据集。采用英特尔® AMX（BF16）的训练性能：性能预测基于非量产的单路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids）在 ResNet-50 v1.5 上进行深度学习训练，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake）。

[9] 采用英特尔® AMX（BF16）的推理性能：性能预测基于非量产的双路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids），56C，350W TDP，共配置 1TB（8 通道/64 GB/4800）的 DDR5 内存，使用 BKC 46，采用英特尔® AMX/int8 和 BF16，CentOS Stream 8，经 oneDNN 优化的英特尔 AMX 内核，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake），28C，250W（8380H）。由于推理性能在不同路数处理器上的结果呈线性扩展，双路处理器的测试数据为八路处理器的测试结果乘以 0.25；配置：单节点，8 个第三代英特尔® 至强® 铂金 8380H 处理器（28C, 250W），基于英特尔参考平台（之前代号为 Cooper City），总内存 384 GB（48 个插槽/64GB/2933），ucode 0x7002302，启用超线程，启用睿频，Ubuntu 20.04 LTS，Linux 5.4.0-29-generic，英特尔® 固态盘 800 GB 操作系统驱动程序；测试结果可能不同。基于英特尔于 2022 年 1 月 27 日进行的测试。对象检测（RT）：采用 SSD-RN34，BS=1，56，BF16，内部版面向英特尔® 架构优化的 TensorFlow 2.8，Squad 1.1 数据集。采用英特尔® AMX（BF16）的训练性能：性能预测基于非量产的单路第四代英特尔® 至强® 可扩展处理器（之前代号 Sapphire Rapids）在 ResNet-50 v1.5 上进行深度学习训练，对比第三代英特尔® 至强® 可扩展处理器（之前代号 Cooper Lake）。

一般提示和法律声明

实际性能受使用情况、配置和其他因素的差异影响。更多信息请见英特尔的性能指标网页。

性能测试结果基于配置信息中显示的日期进行的测试，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。

加速器是否可用视 SKU 而定。更多产品详情，请见英特尔产品规格页面。

英特尔高级矢量扩展技术（英特尔 AVX 技术）为某些处理器操作提供较高的吞吐量。由于处理器功率特性不尽相同，因此利用 AVX 指令可能会导致 a) 某些部件以低于额定频率的频率运行，b) 采用英特尔睿频加速技术 2.0 的某些部件无法实现任何或最高的睿频。产品性能会基于硬件、软件和系统配置的变化有所变化，您可以访问 https://www.intel.cn/content/www/cn/zh/architecture-and-technology/turbo-boost/intel-turbo-boost-technology.html 了解更多信息。

英特尔致力于尊重人权，坚决不参与谋划践踏人权的行为。参见英特尔的《全球人权原则》。英特尔的产品和软件仅限用于不会导致或有助于违反国际公认人权的应用。

英特尔技术可能需要启用硬件、软件或激活服务。