产品简介

英特尔® AI 引擎
第五代英特尔® 至强® 可扩展处理器

第五代英特尔® 至强® 可扩展处理器和英特尔® AI 引擎全面提升 AI 管线性能

65%
的数据中心 AI 推理都在英特尔® 至强® 处理器上运行[1]

高达
14 倍
的 SSD-ResNet34 实时对象检测推理性能提升
这是内置 AMX BF16 的第五代英特尔® 至强® 处理器与第三代英特尔® 至强® 处理器的比较结果[2]

高达
9.9 倍
的实时自然语言处理推理（BERT-large）性能提升和 7.7 倍的每瓦性能提升
这是内置 AMX BF16 的第五代英特尔® 至强® 处理器与第三代英特尔® 至强® 处理器的比较结果[3]

高达
8.7 倍
的推荐系统批量推理（DLRM）性能提升和 6.2 倍的每瓦性能提升
这是第五代英特尔® 至强® 处理器与第三代英特尔® 至强® 处理器的比较结果[4]

从数据预处理、经典机器学习（ML），到自然语言处理和图像识别等深度学习使用，AI 的身影到处可见，遍及广泛的工作负载和用例。英特尔® 至强® 可扩展处理器可为整个 AI 管线提供强大算力。其内置的加速器面向机器学习、数据分析和深度学习等特定 AI 工作负载进行了优化。

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，经典的机器学习和深度学习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管线，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

第五代英特尔® 至强® 可扩展处理器在通用计算方面表现出色，将持续为支持当下各种关键工作负载奠定有力基础。这些处理器采用了英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）这一内置 AI 加速器，可加速基于 CPU 的深度学习推理和训练。在很多情况下，这能够消除独立加速器导致的额外成本和复杂性。新一代英特尔® 至强® 处理器非常适合参数量在 200 亿（20B）以下的大语言模型（LLM），通常可满足客户服务级别协议（SLA）要求[5]。英特尔® AMX 在迁移学习和调优方面也非常出色，只需短短 4 分钟（而非数小时或数天）即可完成模型训练，而无需借助其他硬件。65% 的数据中心推理都在英特尔® 至强® 处理器上运行，因此客户可受益于其现有的通用 AI 架构，而无需为迁移到 GPU 基础设施的复杂性而苦恼。

第五代英特尔® 至强® 可扩展处理器协同英特尔® 加速引擎驱动未来创新

无论是将英特尔® 至强® 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，内置英特尔® 加速引擎的英特尔® 至强® 处理器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，包括数据保护力更强、基础设施利用得更充分。

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

腾讯云借助英特尔 至强® 可扩展处理器实现实时语音合成。

了解详情 ›

Gunpowder 使用第四代英特尔® 至强® 处理器运行 Google Cloud C3 实例，以加快渲染速度。

阅读全文 ›

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® 高级矩阵扩展助力加速深度学习

英特尔® AMX 是英特尔基于第五代英特尔® 至强® 可扩展处理器实现深度学习训练和推理升级的新利器。英特尔® AMX 非常适合自然语言处理、推荐系统和图像识别等工作负载。与第三代英特尔® 至强® 处理器相比，内置 AMX BF16 的第五代英特尔® 至强® 处理器可使实时对象分类推理性能提升高达 7.2 倍，每瓦性能提升高达 5.3 倍[6]。

英特尔® AMX 可为 AI 模型提供工作负载加速，并帮助更多客户在已有的平台上满足多种 SLA 要求。第五代英特尔® 至强® 可扩展处理器可为科学计算和 AI 等矢量和矩阵运算工作负载提供更高的睿频，并增加了五个级别的睿频深度。

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 可提高矩阵乘法运算性能，显著提升吞吐量（单个周期运算量）[7]。这有助于加速完成深度学习训练工作负载，让更多客户在其已开展业务的平台上满足其 SLA 要求。

有效支持自然语言处理和生成式 AI

内置英特尔® AMX 的第五代英特尔® 至强® 可扩展处理器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。英特尔的多个库已面向 TensorFlow 和 PyTorch 进行了优化并与之集成，使开发人员能够享受到开箱即用的内置 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。

通过加速深度学习推理和训练，配备英特尔® AMX 的第五代英特尔® 至强® 可扩展处理器可满足您的 SLA 要求，同时平衡总体拥有成本（TCO）。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，第五代英特尔® 至强® 可扩展处理器即可实现上述目标。

第五代英特尔® 至强® 处理器运行模仿以人为中心的内容的生成式 AI 模型，例如大语言模型和文本生成图像模型。

英特尔® AVX-512 是加速机器学习的强大利器

英特尔® 至强® 处理器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。

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，并提升并行处理应用的性能。英特尔® AVX-512 指令集架构（ISA）提供多种扩展技术，可提高 AI、科学计算、网络和存储等各种工作负载的性能。

新一代英特尔® 至强® 可扩展处理器的睿频深度从四级提高至五级。借助英特尔® AMX 和英特尔® AVX-512，能够提升某些科学计算和 AI 工作负载的睿频。

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。第五代英特尔® 至强® 可扩展处理器内置的英特尔® 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 3 x 3 x 3 x 3 x 3 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。英特尔® 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 中一些非常复杂的运算。

位数越多，处理速度越快

英特尔® 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。顾名思义，英特尔® AVX-512 将寄存器数量扩展到了 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。

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，也可以帮助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）。

英特尔® 至强® 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。

对于 AI 管线中的其他应用，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。数据科学家和 AI 开发人员何必专门就英特尔® AVX-512 对自己的工具重新编码和编译？这项工作可以由我们完成。

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® 可扩展处理器内置专用英特尔® 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

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® AI 分析工具套件 ›

基于英特尔® 硬件和软件实施开发 ›

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 ›

intel® XEON®

1. 基于英特尔对截至 2022 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. 详情请见以下网址的 [A21]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。
3. 详情请见以下网址的 [A19]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。
4. 详情请见以下网址的 [A20]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。
5. 基于英特尔 2023 年 12 月进行的内部建模。
6. 详情请见以下网址的 [A22]：intel.com/processorclaims（第五代英特尔® 至强® 可扩展处理器）。结果可能不同。
7. 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。

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 www.Intel.cn/PerformanceIndex。

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。

配合工作负载/配置信息请见 www.intel.com/processorclaims（第五代英特尔至强可扩展处理器）。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（英特尔 AVX）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。