产品简介

AI 加速
英特尔® 至强® 可扩展处理器

英特尔® 至强® 可扩展处理器内置加速器 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]

9 成
的企业应用到 2025 年会引入 AI[2]

从数据分析和经典机器学习到语言处理和图像识别的广泛工作负载和用例，到处可见 AI 的身影。英特尔® 至强® 可扩展处理器结合了为整条 AI 流水线提供灵活卓越的算力，以及面向数据科学、模型训练和深度学习推理等特定 AI 工作负载的内置加速器。

AI 所涉范围大过深度学习，并且还在不断扩大

AI 尚处早期阶段，当前正在各个领域迅速成长壮大。从核心企业应用到自动话务台系统，经典的机器学习算法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科学到训练、验证，再到最终部署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® 可扩展处理器内置加速技术，可突破上述障碍，全面提升 AI 性能。

AI 真的就是数学，很多很多数学。

海量数学计算是每个 AI 任务和运行的核心。很多数据科学运算，如建模数据和机器学习算法，都基于统计学、代数和复矢量数学。深度学习 AI 需要进行大量矩阵乘法。所有这些 AI 应用都是“暴力运算”，涉及大型数据集和广泛的处理资源，包括 CPU、GPU、FPGA 和针对特定工作负载的定制 ASIC。

英特尔® 高级矢量扩展 512（英特尔® AVX-512）是加速 AI 的数学计算利器

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，但在深度学习训练（整个 AI 流水线的一个细分）方面，速度不如专用加速器。这是因为 CPU 是按顺序逐一处理运算的，一次只能执行一个计算。而其他类型的处理器可以并行处理运算，即同时进行多个计算。

英特尔® AVX-512 将更多运算纳入每个时钟周期，因此克服了 CPU 的架构限制，使 CPU 更像并行处理器一样工作。

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 ›

BeeKeeperAI 开发临床 AI 算法，并帮助保护数据隐私。

阅读全文 ›

复杂的 CPU 指令，精简的策略：每个时钟周期内都更聪明、更高效地工作

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 (英特尔® DL Boost) 是更聪明的神经网络数学计算

深度学习 AI 使用大量矩阵乘法来训练神经网络模型，并利用推理将这些模型应用于实际任务。在推理过程中，计算机将输入数据（如包含语音的音频信号）与模型（在本文中指语音识别模型）进行比较，然后推断数据的含义。推理在对象识别、图像分割、文本识别以及几乎所有其他深度学习 AI 任务中都会用到。

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。英特尔® DL Boost 还会加速使用 INT8 精度的深度学习工作负载。

即将到来的技术升级将进一步提升 AI 性能

第四代英特尔® 至强® 可扩展处理器将内置一种新的加速器，专门用于处理对深度学习工作负载来说至关重要的矩阵乘法。英特尔® 高级矩阵扩展（英特尔® AMX）结合了一个新指令集（将大型矩阵转换为单个运算）与二维寄存器文件（为每个内核存储更大的数据块）。

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，这意味着它们可以随时用于任何与之兼容的软件。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。

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（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 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，它们可以利用第三代英特尔® 至强® 可扩展处理器指令集架构中的各个内置加速器。

从根本上说，基于英特尔® 硬件加速 AI 就像下载一版您已经使用并投入运行的英特尔® 工具一样简单。

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 Analytics 工具套件 ›

软件优化为 AI 流水线中的应用带来的增益

约 38 至 200 倍
scikit-learn 速度提升（使用英特尔® Extension for scikit-learn）[3]

约 90 倍
pandas 速度提升（使用英特尔® 分发版 Modin）[3]

高达
3 倍
TensorFlow 速度提升（使用英特尔® oneDNN）[3]

第三代英特尔® 至强® 可扩展处理器的 AI 加速

加速深度学习 AI 工作负载

高达
1.74 倍
INT8 批推理吞吐量提升
内置英特尔® DL Boost 的第三代英特尔® 至强® 可扩展处理器与上一代处理器在 BERT-Large SQuAD 上的吞吐量对比[4]

高达
1.59 倍
INT8 实时推理吞吐量提升
内置英特尔® DL Boost 的第三代英特尔® 至强® 可扩展处理器与上一代处理器对比[5]

高达
4.5 倍
INT8 精度下每秒检测的图像的数量增幅[6] 以及高达 6 倍多 BF16 精度下进行对象检测时每秒检测的图像的数量增幅[7]
（SSD-ResNet-34，利用内置于即将上市的第四代英特尔® 至强® 可扩展处理器的英特尔® AMX）

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 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. “IDC FutureScape: Worldwide IT Industry 2020 Predictions”（IDC FutureScape：2020 年全球 IT 行业预测），2019 年 10 月。Doc #US45599219. idc.com/getdoc.jsp?containerId=US45599219.
3. “One-Line Code Changes to Boost pandas, scikit-learn, and TensorFlow Performance”（只需修改一行代码即可提升 pandas、scikit-learn 和 TensorFlow 性能），2021 年 7 月。https://www.intel.cn/content/www/cn/zh/developer/articles/technical/code-changes-boost-pandas-scikit-learn-tensorflow.html#gs.d8e6eo
4. 详情请见以下网址的 [123]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。结果可能不同。
5. 详情请见以下网址的 [122]：https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。结果可能不同。
6. 详情请见以下网址的 [41]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/vision-2022/。结果可能不同。
7. 详情请见以下网址的 [42]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/vision-2022/。结果可能不同。

0922/MP/CMD/PDF