解决方案简介

人工智能（AI）
英特尔® 深度学习加速技术（英特尔® DL Boost）

内置加速器优化人工智能（AI）工作负载

英特尔® 深度学习加速技术（英特尔® DL Boost），无需独立附加加速器，即可帮助要求苛刻的 AI 工作负载实现性能加速。

内置 AI 加速器的 x86 数据中心级 CPU[1]。

无需增加基础设施的成本和复杂性，即可**满足服务级别协议（SLA）**。

加速计算密集型工作负载，助力高性能计算（HPC）、科学研究、医疗行业等诸多领域发展。

使用 bfloat16 与 INT8 加速对象检测、图像识别/分类和自然语言处理（NLP）[2]。

如何在不依赖专用加速器的前提下，利用 CPU 内现有加速器实现人工智能（AI）的效率和性能提升？英特尔® 至强® 可扩展处理器具有内置加速器，包括英特尔® 深度学习加速技术（英特尔® DL Boost），专为提高常见 AI 推理和训练工作负载的性能而设计。英特尔® DL Boost 内置于已为数据中心或云端传统工作负载提供了出色性能、安全性和可靠性的 CPU 封装中，从而为企业和机构提供 AI 加速能力，有助于优化医疗诊断、加速科学研究和提升电信网络性能。

优化 AI 的性能

企业和机构正积极利用大数据进行数据分析任务；而在几年前，这些任务必须在高端工作站或超级计算机上才有可能完成。如今，经济型数据中心服务器已经能够运行 AI 工作负载，这些工作负载自海量数据中提取可执行信息，然后将提取结果应用于高性能计算（HPC）、科研、金融、医疗等多个领域。

要处理 AI 工作负载，传统方法是为专用设备配备独立的图形处理单元（GPU）。虽然 GPU 可能需要用来加速训练工作负载，但当 AI 推理工作负载在搭载第三代英特尔® 至强® 可扩展处理器的服务器上运行时，GPU 通常不是必备配置。这是因为英特尔提供目前市面上少有的内置 AI 加速器的 x86 数据中心级 CPU[1]。

英特尔® AVX-512 和英特尔® DL Boost：CPU 内置指令集系列

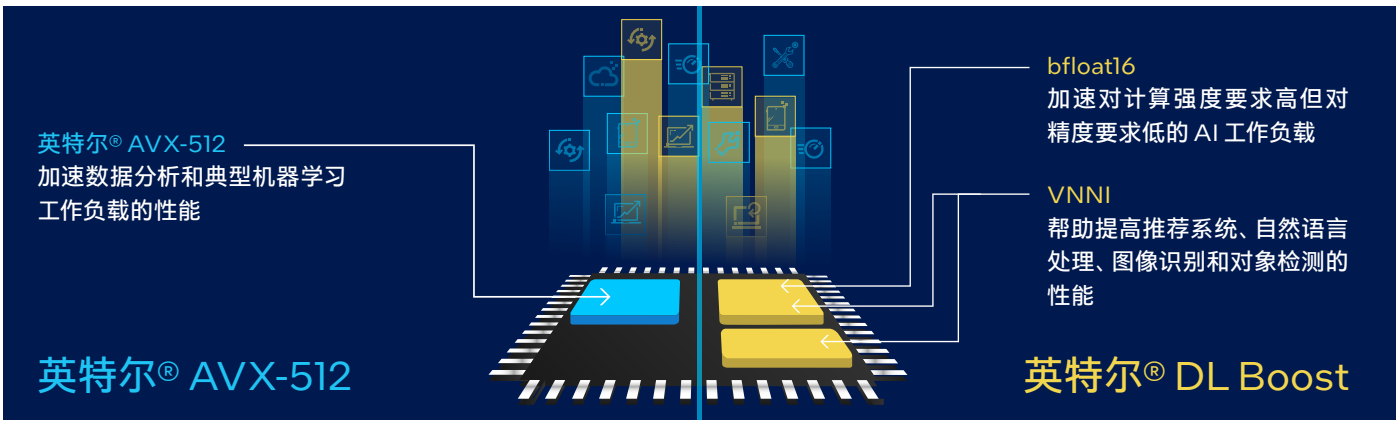

图 1. 英特尔® AVX-512 与英特尔® DL Boost 指令集专为提高 AI 性能而设计

内置于英特尔® 至强® 可扩展处理器的多项技术中，英特尔® DL Boost 可为不断增长的大量 AI 用例提供高吞吐量和低时延服务。这些用例包括疾病检测和治疗、金融交易关键数据提取、推进科研和学术研究等。由于内置于英特尔® 至强® 可扩展处理器中，英特尔® DL Boost 可为大多数用例的外部 GPU 解决方案提供可用的替代方案。因此，您可在不增加基础设施成本和复杂性的前提下，更好地满足性能服务级别协议 (SLA)。

深入了解创新历程

英特尔® 高级矢量扩展 512（英特尔® AVX-512）指令集于第一代英特尔® 至强® 可扩展处理器中率先引入，英特尔® DL Boost 正是通过扩展该指令集进行工作。英特尔® AVX-512 是基于 x86 数据中心级 CPU 构建的一套单指令多数据 (SIMD) 指令集。相较于传统的单指令单数据 (SISD) 指令，单指令多数据指令集可利用一条指令执行多个数据运算。英特尔® AVX-512 的寄存器位宽是 512 位，支持 16 个 32 位单精度浮点数或 64 个 8 位整数 (INT8)。

第二代英特尔® 至强® 可扩展处理器通过添加英特尔® AVX-512 矢量神经网络指令 (VNNI)，进一步加速英特尔® AVX-512 指令集。VNNI 是对英特尔® AVX-512 指令集的扩展，它将三条执行指令合并成一条指令，进一步提升 INT8 模型的推理性能（见图 2）。

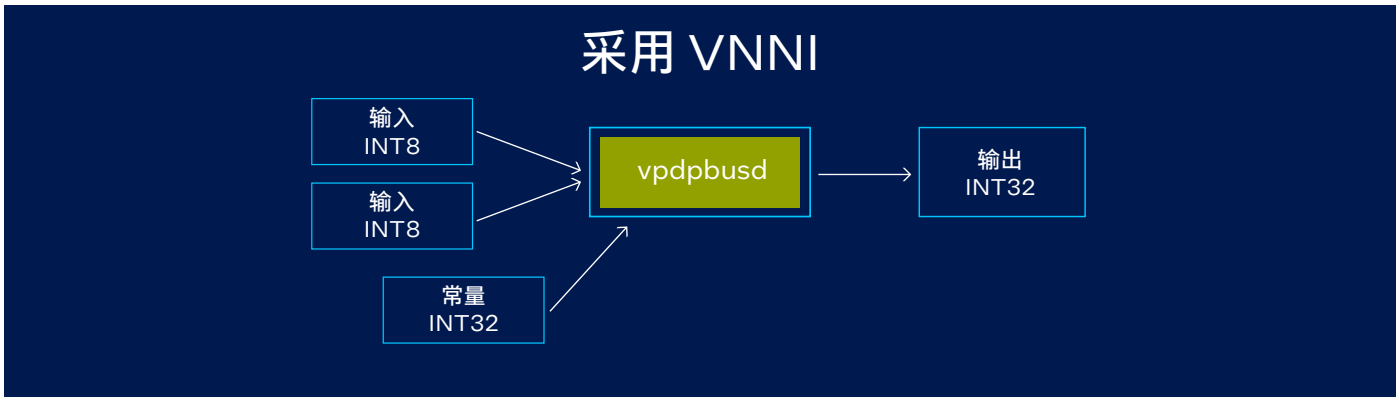

图 2. 采用 VNNI 的平台仅需 vpdpbusd 一条指令即可完成 INT8 卷积运算

相比之下，未采用 VNNI 的平台需要串行运行三条单独指令（vpmaddubsw、vpmaddwd 和 vpaddd），才能完成 INT8 卷积运算中的乘积累加运算（如图 3 所示）。

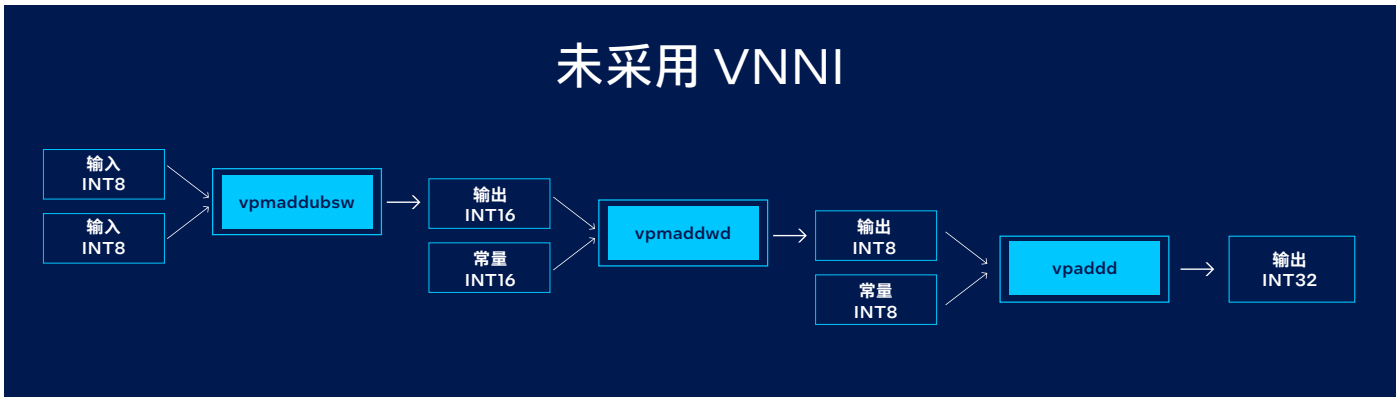

图 3. 未采用 VNNI 的平台需要分别运行三条指令，导致 AI 处理任务的时延增加

而采用 VNNI 的第二代英特尔® 至强® 可扩展处理器通过合并上述运算，提升了计算资源的利用率，同时更好地利用高速缓存，避免了潜在的带宽瓶颈。

浮点 32 (FP32) 可为推理工作负载提供高数值精度（见图 4 中的小数/尾数），它对准确度的要求更高，各代英特尔® 至强® 可扩展处理器均可为其提供支持。第三代英特尔® 至强® 可扩展处理器支持脑浮点格式（bfloat16 或 BF16）[3]，性能进一步升级。但大多数 AI 训练工作负载无需达到计算密集型 FP32 所具备的准确度水平。bfloat16 指令将 FP32 数据转化为 bfloat16，对于要求高计算强度但对精度要求较低的工作负载而言，bfloat16 是更高效的编码格式。

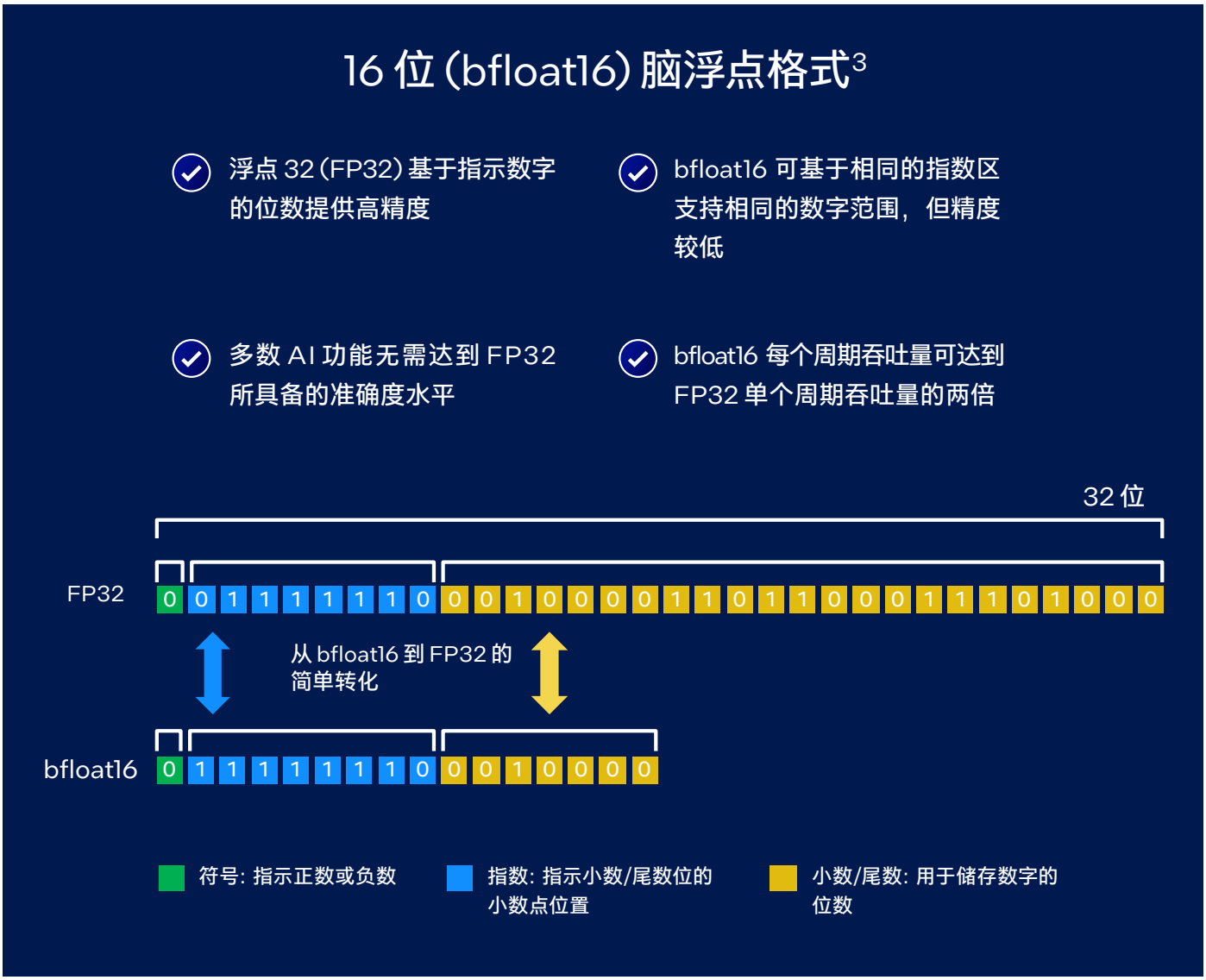

图 4. bfloat16 可在提升吞吐量的同时，始终保持 AI 工作负载所需的准确度水平[3]

相较于使用 FP32 对要求较低准确度的工作负载进行编码，使用英特尔® DL Boost 和 INT8 可获得高达 4.3 倍的性能优化，使第三代英特尔® 至强® 可扩展处理器上运行的工作负载的性能得到显著提升[2]。英特尔® DL Boost 可实现常见对象检测、图像识别、自然语言处理 (NLP) 和图像分类工作负载的推理加速（见图 5）[4]。英特尔® DL Boost 采用 bfloat16，可将 AI 训练性能提高达 1.93 倍[3,5,6]。

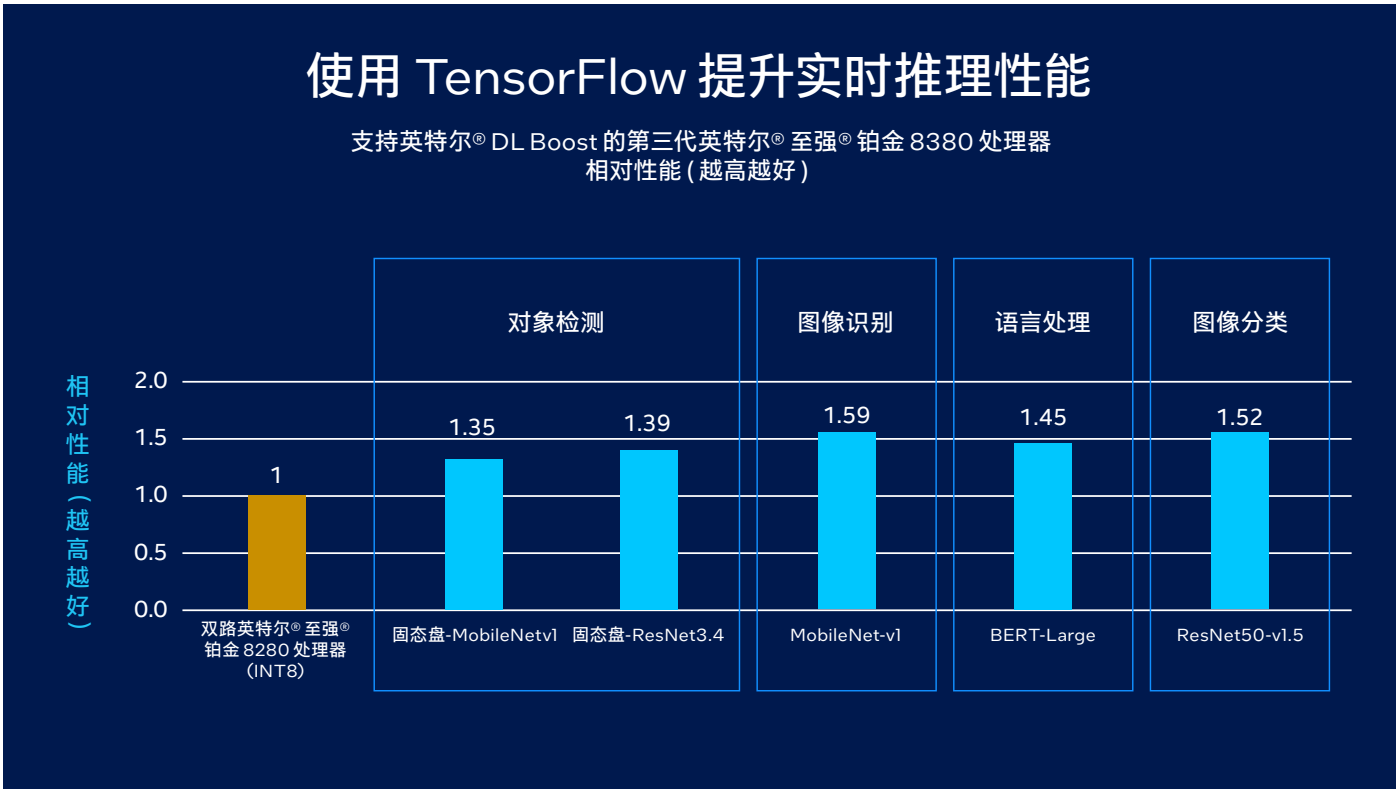

图 5. 与上一代处理器相比，支持英特尔® DL Boost 的第三代英特尔® 至强® 可扩展处理器的 INT8 实时推理吞吐量提升高达 1.59 倍[4]

使用英特尔® DL Boost 与 BF16 指令加速第三代英特尔® 至强® 可扩展处理器，并实现[3]：

1.7 倍
自然语言处理训练性能提升[6]

1.93 倍
图像分类训练性能提升[5]

……相较于第二代英特尔® 至强® 可扩展处理器。

面向英特尔® 加速器优化的主流框架和库

TensorFlow、Apache MXNet、PyTorch 和 PaddlePaddle（百度飞桨）等 AI 框架和包括 scikit-learn 和 XGBoost 在内的库，是迅速实施 AI 部署并高效运行 AI 工作负载的主流解决方案。这些框架和库为支持英特尔® DL Boost 等英特尔® 至强® 可扩展处理器内置的加速器专门设计，也可支持英特尔® 分发版 OpenVINO™ 工具包等优化工具。

英特尔还提供大量的开放标准工具包，以便您对这些内置加速器进行调整和优化。英特尔® oneAPI 深度神经网络库（英特尔® oneDNN）、英特尔® oneAPI 工具包和英特尔® 分发版 OpenVINO™ 工具包都是基于常见软件编程模型的 oneAPI 工具包。借助这些开放的工具，您可以使用熟悉的标准接口开展工作，畅享“一次编写，随处运行”代码效率的诸多益处。

提供强大的 AI 性能，惠及医疗、科研和工业等领域

内置英特尔® DL Boost 的英特尔® 至强® 可扩展处理器为 AI 工作负载提供的加速功能具有灵活和高性价比的特点，可惠及工业和科学研究等众多垂直领域。以下为近期的三个客户案例。有关企业和机构如何通过英特尔® DL Boost 和英特尔® 至强® 可扩展处理器加速 AI 工作负载的更多案例，请访问英特尔客户聚焦网站，https://www.intel.cn/content/www/cn/zh/customer-spotlight/overview.html 。

助力结核病检测

宁波江丰生物信息技术有限公司（江丰生物）是一家专业从事数字病理系统开发和生产的高科技生物信息技术企业。其一体化数字病理诊断系统高度融合医疗、教学、科研以及信息服务。

江丰生物目前已拥有高效的深度学习（DL）解决方案，可借助 GPU 对结核杆菌（M. Tb）进行扫描，但其工程师需要迅速完成扫描和诊断。该公司曾使用内置英特尔® DL Boost 的第二代英特尔® 至强® 可扩展处理器，对其代码库进行优化，帮助上海市公共卫生临床中心（公卫中心）加速处理结核病诊断病例。

据公卫中心称，江丰生物诊断系统的平均检测精度可达 86.8%，分类准确度可达 88.9%[7]。从输入样本到生成报告，整个工作流程可在 80 秒内完成对单个病例的诊断。

随后，江丰生物又借助面向英特尔® 架构优化的 PyTorch 1.6，而非上一版 PyTorch 1.4，完成对其结核杆菌扫描算法的多项优化。在英特尔® 至强® 金牌 6252 处理器上进行的基准测试表明，相较于 PyTorch 1.4，集成英特尔® oneDNN 的 PyTorch 1.6 可将推理速度**提升 11.4 倍**[8]。如欲了解更多信息，请参见：“KFBIO Accelerates Tuberculosis Detection with AI（江丰生物利用 AI 加速结核病检测）”
https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/kfbio-ai-customer-story.html。

提升电信网络质量

SK 电讯（SK Telecom, SKT）是韩国最大的移动通讯运营商。为了对 SK 电讯网络产生的海量数据进行高效分析，SK 电讯和英特尔的工程师搭建了一条端到端网络 AI 管道用于网络质量预测。整条管道在基于内置英特尔® AVX-512 和英特尔® DL Boost 的英特尔® 至强® 可扩展处理器的统一服务器集群上运行。BigDL 2.0 中的 Analytics Zoo 软件负责内存数据管道的搭建和分布式模型的训练及推理。与 SK 电讯过去基于传统 GPU 的实施方案相比，这一 AI 管道分别为深度学习训练和深度学习推理带来了**高达 4 倍**和**高达 6 倍**的性能提升[9]。这些性能增益使 SK 电讯能够更快地预测及监测网络质量有无降级或异常变化，并在必要时采取主动措施，以提供高质量的 5G 服务。如欲进一步了解该客户，请参见“SK Telecom: AI Pipeline Improves Network Quality（SK 电讯：AI 管道提升网络质量）”，https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/sk-telecom-ai-customer-story.html。

降低科学和学术研究成果的时间成本

欧洲核子研究组织（European Organization for Nuclear Research, CERN）的物理学家和研究人员借助一系列粒子加速器设施，进行有关物质最基本成分—基本粒子的相关研究。该组织的大型强子对撞机（Large Hadron Collider, LHC）是世界上规模最大的粒子加速器，为满足该加速器的未来需求，其研究人员与英特尔合作，开展深度学习推理工作负载的加速工作。研究人员使用英特尔® oneAPI AI Analytics Toolkit（AI Kit）提升了第二代英特尔® 至强® 可扩展处理器的推理性能。英特尔® DL Boost 所提供的内置 AI 加速功能是实现项目性能增益的关键所在，且可在无损准确度的情况下加速推理进程。

研究人员采用条件生成对抗网络（GAN），对面向未来潜在粒子加速器的热量仪进行了仿真，证明了英特尔® DL Boost 能够带来的性能增益：较之以前，所需的计算资源大幅降低。

相较于 FP32 推理模型，研究团队使用了速度更快的由英特尔® DL Boost支持的 INT8 推理模型，为欧洲核子研究组织的复杂 GAN 模型推理实现 **1.8 倍**增益，推理准确度也略有提升[10]。

这项工作影响深远。据欧洲核子研究组织的研究人员称，全球大型强子对撞机计算网格（WLCG）中半数以上的计算都用于仿真。性能、成本和准确度对于部署 WLCG 训练的模型而言至关重要[11]。此外，欧洲核子研究组织训练条件生成式对抗网络的方法，以及使用英特尔® DL Boost 在无损准确度的情况下实现加速，都为需进行相似蒙特卡罗仿真的其他领域应用开辟了新的可能。请访问“CERN Accelerates Simulation Workloads with AI（CERN 借助 AI 加速仿真工作负载）” https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/cern-inference-customer-story.html，了解更多详情。

CPU 内置 AI 加速

各代英特尔® 至强® 可扩展处理器均支持在基于 CPU 的平台运行更高 AI 工作负载和更多用例类型。无需借助专用加速平台，您可使用内置英特尔® DL Boost 的英特尔® 至强® 可扩展处理器，为 Amazon Web Services（AWS）、Microsoft Azure 和 Google Cloud Platform（GCP）等多个主流云服务平台上运行的 AI 工作负载加速。诸如英特尔® DL Boost 的内置加速器可以提高常见 AI 工作负载的推理和训练速度，且较基于 GPU 的专用平台而言，可提供更高的灵活性和资源利用率。因为这些加速器内置于您所熟知且信任的英特尔® 至强® 可扩展处理器中，而这类处理器已在数据中心或云端广泛用于支持传统工作负载。

了解有关英特尔® DL Boost 及其他英特尔® AI 加速器的更多信息：
https://www.intel.cn/content/www/cn/zh/artificial-intelligence/overview.html

intel XEON

[1] 英特尔，“第三代英特尔® 至强® 可扩展处理器”，https://www.intel.cn/content/www/cn/zh/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html。

[2] 来源：Lisa Spelman，Ice Lake Press Workshop（Ice Lake 新闻研讨会），第 39 页，英特尔附录第 15 项声明，“3rd Gen Intel® Xeon® Scalable Platform Technology Preview（第三代英特尔® 至强® 可扩展平台技术预览）”，2021 年 11 月。https://newsroom.intel.com/wp-content/uploads/sites/11/2021/04/3rd-Gen-Intel-Xeon-Scalable-Platform-Press-Presentation-281884.pdf。

[3] 仅第三代英特尔® 至强® 可扩展处理器 Cooper Lake（代号）版本支持 bfloat16。

[4] 来源：英特尔，“第三代英特尔® 至强® 可扩展处理器性能指标第 119 至 123 项声明”，https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。

[5] 来源：英特尔，“第三代英特尔® 至强® 可扩展处理器性能指标第 9 项声明”，https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。

[6] 来源：英特尔，“第三代英特尔® 至强® 可扩展处理器性能指标第 1 项声明”，https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。

[7] 测试由上海市公共卫生临床中心和江丰生物共同完成。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。

[8] 江丰生物结核杆菌扫描 detectron2 模型在第二代英特尔® 至强® 金牌 6252 处理器上的吞吐量性能：**配置：新平台：**测试 1：使用 PyTorch 1.6 的单个实例：截至 2020 年 5 月 22 日英特尔所做测试。双路第二代英特尔® 至强® 金牌 6252 处理器，24 核，英特尔® 超线程技术启用，英特尔® 睿频加速技术启用，192 GB 总内存（12 个插槽/16 GB/2,666 MHz），BIOS：SSE5C620.86B.02.01.0008.031920191559（ucode：0x500002c），Ubuntu 18.04.4 LTS，内核 5.3.0-51-generic，已缓解。测试 2：使用 PyTorch 1.6 的 24 个实例：截至 2020 年 5 月 22 日英特尔所做测试。双路第二代英特尔® 至强® 金牌 6252 处理器，24 核，英特尔® 超线程技术启用，英特尔® 睿频加速技术启用，192 GB 总内存（12 个插槽/16 GB/2,666 MHz），BIOS：SSE5C620.86B.02.01.0008.031920191559（ucode：0x500002c），Ubuntu 18.04.4 LTS，内核 5.3.0-51-generic，已缓解。**基准配置：**使用 PyTorch 1.4 的单个实例：截至 2020 年 5 月 22 日英特尔所做测试。双路第二代英特尔® 至强® 金牌 6252 处理器，24 核，英特尔® 超线程技术启用，英特尔® 睿频加速技术启用，192 GB 总内存（12 个插槽/16 GB/2,666 MHz），BIOS：SSE5C620.86B.02.01.0008.031920191559（ucode：0x500002c），Ubuntu 18.04.4 LTS，内核 5.3.0-51-generic，已缓解。来源：https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/kfbio-ai-customer-story.html。

[9] SK 电讯于 2020 年 2 月进行的测试：Analytics Zoo 服务器为英特尔® 服务器系统 R2208WFTZSR，由 2.6 GHz 英特尔® 至强® 金牌 6240 处理器（微代码：0x400002c）提供支持。该服务器拥有 3 个节点和 6 个插槽。英特尔® 超线程技术和英特尔® 睿频加速技术均已启用。总内存为 256 GB。操作系统为 CentOS 7.8（内核 3.10.0），服务器运行 SK 电讯 Lightning DB 应用。其他软件包括 Analytics Zoo v0.7、TensorFlow v1.15、Pandas v0.25.3、NumPy v1.18.0 和 Dask v2.7.0。GPU 服务器为 HPE DL380 Gen 9，由 2.4 GHz 英特尔® 至强® 处理器 E5-2680 v4（微代码：0xb00001e）和 NVIDIA P100 GPU（AI 训练）/K80（AI 推理）提供支持。该服务器拥有 1 个节点和 2 个插槽。英特尔® HT 技术和英特尔® 睿频加速技术均已启用。总内存为 256 GB。操作系统为 CentOS 7.3（内核 3.10.0），服务器运行 SK 电讯 Lightning DB 应用。其他软件包括 TensorFlow GPU v1.12、Pandas v0.25.1、NumPy v1.14.5 和 Dask v2.7.0。来源：https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/sk-telecom-ai-customer-story.html。

[10] Florian Rehm 等，“Reduced Precision Strategies for Deep Learning: A High Energy Physics Generative Adversarial Network Use Case（深度学习精度降低策略：高能物理生成式对抗网络用例）”。来自与英特尔合作的 CERN openlab， 2021 年 3 月，https://cds.cern.ch/record/2758899/?ln=zh_CN。

[11] 英特尔，“CERN Accelerates Simulation Workloads with AI（CERN 借助 AI 加速仿真工作负载）”，2021 年 1 月，https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/kfbio-ai-customer-story.html。

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。

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。

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。

© 英特尔公司版权所有。英特尔、英特尔标识以及其他英特尔商标是英特尔公司或其子公司的商标。其他的名称和品牌可能是其他所有者的资产。

0622/SO/PRW/PDF　　请回收利用