intel.
XEON®

如何充分利用内置加速器的英特尔® 至强® 可扩展处理器

电子指南

目录

什么是内置加速技术？为何应该使用这种技术？

如果每次需要建立新功能时都可以利用已经内置于 CPU 的技术，而无需购买新设备，那将会是怎样一番情景？有英特尔® 至强® 可扩展处理器，您就可以做到这一点。这些 CPU 包含多种被称为“内置加速器”的功能，可以增强工作负载的性能优势。

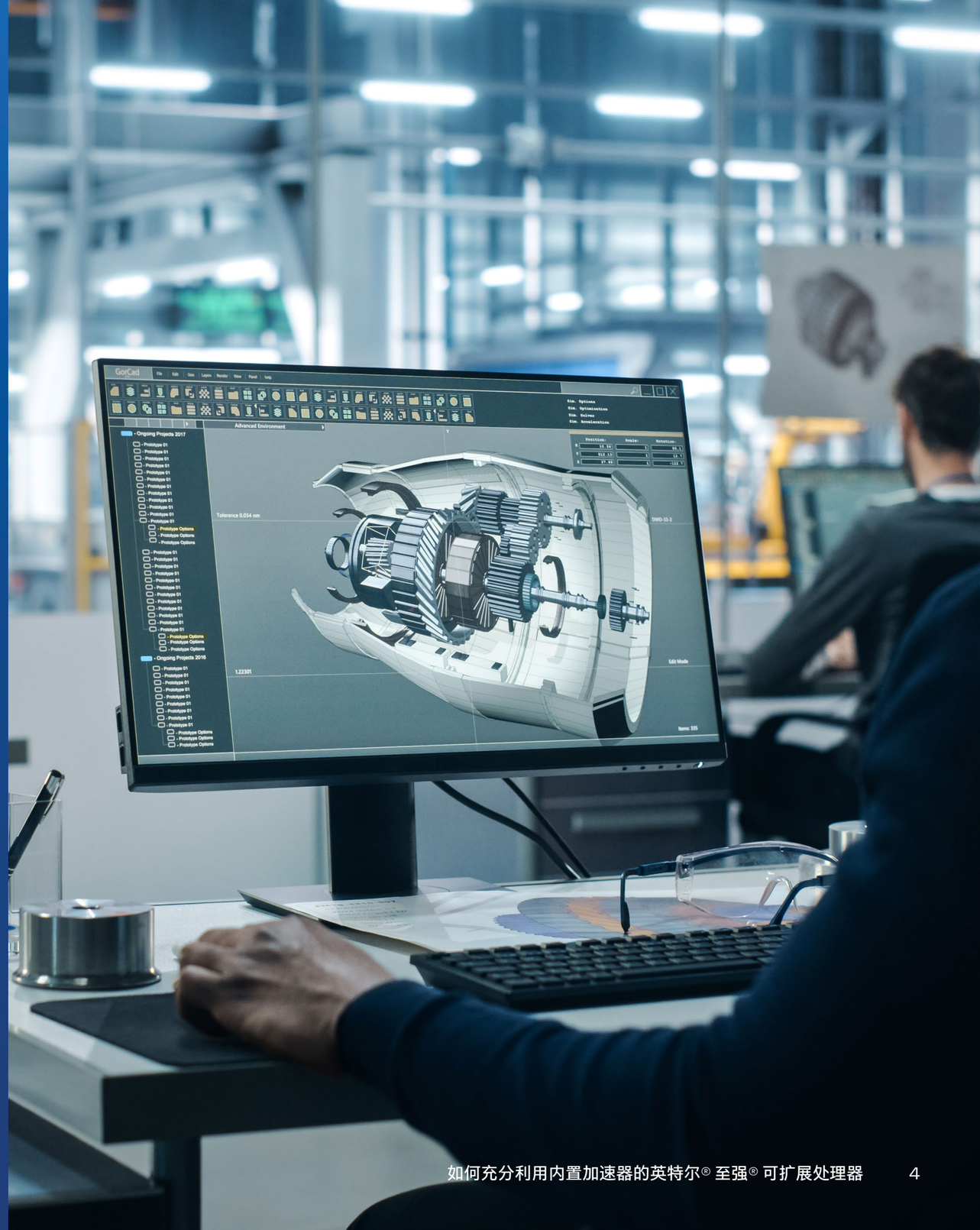
英特尔® 至强® 可扩展处理器支持广泛且独特的内置加速器，有助于提高性能和效率，减少另行添置专用硬件的需求。在云端和本地环境中，这些专用功能支持人工智能（AI）、安全性、科学计算、数据分析、存储和网络等目前最为常见的严苛工作负载。

在本指南中，我们将重点关注 AI、科学计算和安全性。

英特尔® 内置加速器在实际应用中的优势

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，我们的内置加速器都能够助力您的业务达到新高度。这些内置加速器具备一系列优势，包括更快的处理速度、更强的数据保护和更充分的基础设施利用。除此之外，这些内置加速器还能够**提高应用性能，降低成本并提升能效：**

性能	英特尔® 内置加速器作为专用组件，常可为目标工作负载带来更高的性能[1]。
成本节约	英特尔® 内置加速器可以优化性能，无需另行购买专用硬件。
节能	使用英特尔® 内置加速器，用户无需为服务器额外增加内核，因此有助于用户提升能效。



哪些英特尔® 内置加速器能够满足您的业务需求？

虽然英特尔® 至强® 可扩展处理器内置了全套加速器，但特定的任务/工作负载选用特定的加速器，才能更好地发挥加速作用。为了帮助您判断哪些英特尔® 技术可以更好地支持您的业务，我们将详细介绍 AI、科学计算和安全性方面的三个重点加速技术。

AI

英特尔® 深度学习加速技术（英特尔® DL Boost）能够大大提高常见的 AI 和科学计算工作负载的推理和训练性能[2]。

科学计算

英特尔® 高级矢量扩展 512（英特尔® AVX-512）是专门为加速科学、商业等领域中要求严苛的计算工作负载的性能而打造的加速器。

安全性

英特尔® 软件防护扩展（英特尔® SGX）能够通过特有的应用隔离技术保护使用中的数据。

AI:
英特尔® 深度学习加速技术

什么是英特尔® 深度学习加速技术？

英特尔® DL Boost 专为提升 AI 和深度学习相关任务和工作负载的性能和效率而设计。

英特尔® 至强® 可扩展处理器于 2019 年推出了采用英特尔® 矢量神经网络指令（VNNI）的 AI 专用加速技术，这就是现在的英特尔® DL Boost。

英特尔® DL Boost 的 VNNI 组件基于英特尔® AVX-512，它将三种指令集合而为一，从而大大缩短了完成一项任务所需的时间。

英特尔® DL Boost 最常见的用例有哪些？

英特尔® DL Boost 可以为各种 AI 推理任务加速，如图像分类、语言翻译和对象检测。

实际使用英特尔® DL Boost 的公司表现如何？

软件公司 rinf.tech 使用英特尔® DL Boost 提供更快速准确的图像分析，用以支持在零售、汽车、视频监控和商业智能用例中更好地进行实时决策。推理性能比基准快了高达 7.4 倍[3]。

慧影医疗科技公司（HYHY）使用英特尔® DL Boost 帮助其优化全周期 AI 医疗影像解决方案的性能。得益于软硬件协同加速带来的优势，该公司在医疗图像分析场景中的推理速度得到显著提升[4]。

宁波江丰生物信息技术有限公司（KFBIO）是一家专门从事数字病理系统开发和生产的企业。他们利用英特尔® DL Boost 完成了对结核杆菌标本的扫描与诊断，速度比基准快了高达 11.4 倍[5]。

英特尔® DL Boost 具备哪些性能优势？

使用面向英特尔® 架构优化的 TensorFlow **和英特尔® DL Boost 的客户将获得超过**

11 倍 的 AI 推理性能提升

基于第三代英特尔® 至强® 可扩展处理器与第二代英特尔® 至强® 可扩展处理器的比较[6]。

英特尔® DL Boost 还可以为 AI 工作负载带来更出色的性能功耗比，使企业和机构能够降低成本和能耗。

科学计算：
英特尔® 高级矢量扩展 512

什么是英特尔® 高级矢量扩展 512？

英特尔® AVX-512 是能够增强性能的通用加速器，具有广泛用途。它有超宽 512 位矢量运算能力，特别适合处理科学计算领域常见的严苛计算任务。

英特尔® AVX-512 最常见的用例有哪些？

英特尔® AVX-512 广泛用于教育、市政、金融、企业、工程和医疗等行业，可处理各种复杂的任务。这其中包括 AI、科学仿真、3D 建模和分析、金融分析、音视频处理、加密和数据压缩。

英特尔® AVX-512 支持对金融服务工作负载进行实时分析，从而提升客户体验、合规性和数据安全。

英特尔® AVX-512 支持在现有硬件上运行复杂的工作负载，从而为 3D 建模和仿真等任务加速。

实际使用英特尔® AVX-512 的公司表现如何？

纽约州立大学布法罗分校的计算研究中心使用英特尔® AVX-512 为纽约州西部的企业提供丰富的计算资源，例如，帮助 Marion Surgical 公司利用虚拟现实 (VR) 和增强现实 (AR) 教外科医生进行复杂手术[7]。

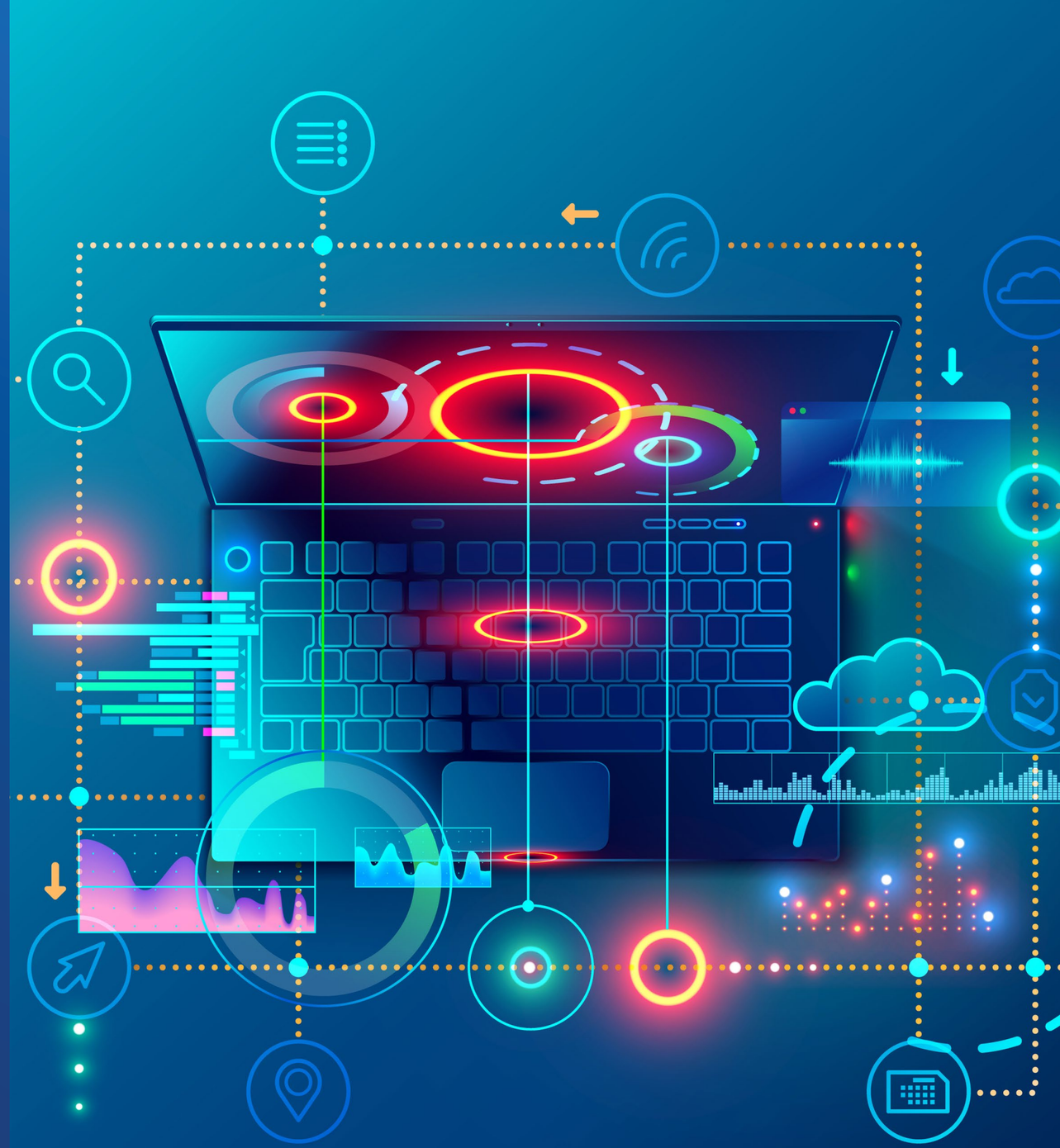
麻省理工学院和哈佛大学的博德研究所使用英特尔® AVX-512 帮助其在 Google Cloud N1 和 N2 实例上运行基因组学工作负载时提高处理速度、降低成本[8]。

欧洲核子研究组织 **CERN** 从事基本粒子研究的研究人员曾使用英特尔® AVX-512 通过量化来加速仿真工作负载，使性能提高了 1.8 倍，准确性也略有提升[9]。

英特尔® AVX-512 具备哪些性能优势？

由于科学计算工作负载涉及大量数据，因此 CPU 的性能对于确保及时获得准确结果至关重要。

与上一代解决方案相比，英特尔® AVX-512 有助于提高至强® 可扩展处理器的矢量处理能力，使企业和机构能够更快地处理密集的工作负载。借助多达两个 512 位融合乘加（FMA）单元，应用程序在 512 位矢量内的每个时钟周期每秒可打包 32 次双精度和 64 次单精度浮点运算，以及八个 64 位和十六个 32 位整数。因此，与英特尔® 高级矢量扩展 2.0（英特尔® AVX2）相比，数据寄存器的宽度、数量以及融合乘加单元的宽度都增加了一倍。



安全性：
英特尔® 软件防护扩展（SGX）

什么是英特尔® 软件防护扩展？

英特尔® SGX 提供基于硬件的安全解决方案，可通过应用隔离技术帮助保护使用中的数据。开发人员可以通过保护选定的代码和数据来让这些代码和数据免受检查或修改，在飞地内执行涉及敏感数据的操作，帮助提高应用的安全性或保护数据的机密性。这有助于减小系统的受攻击面，再增加一层防护。

英特尔® SGX 最常见的用例有哪些？

英特尔® SGX 支持机密计算解决方案，可以更好地保护本地、边缘和云端的数据。该加速器为希望保护敏感数据和代码的企业和机构提供支持，有助于确保符合与数据隐私、主权和保密性相关的法律法规。

实际使用英特尔® SGX 的公司表现如何？

英国金融机构，**全英房屋抵押贷款协会（Nationwide Building Society）**，使用英特尔® SGX 建立“了解客户”系统，该系统**支持在飞地内更安全地处理多个机密数据集，**使其符合数据监管规定[10]。

瑞士再保险集团（Swiss Re Group）作为全球最大的再保险提供商之一，使用英特尔® SGX 进行机密计算概念验证，成功地探索了一条对多方**机密数据提供额外保护**的道路[11]。

加利福尼亚大学旧金山分校使用英特尔® SGX 开发了能够保护隐私的分析方法，加速了临床算法的开发和验证。该平台将提供“零信任”环境，有助于**保护算法的知识产权和医疗数据隐私**[12]。

英特尔® SGX 具备哪些安全性优势？

英特尔® SGX 为机密计算提供了一个关键的构建模块，它能够限制访问使用中的敏感数据和代码，使其免受其他软件的检查和破坏。只有英特尔® SGX 能够灵活支持虚拟、裸机和云原生容器部署。

下一代英特尔® 内置加速器

英特尔® 至强® 可扩展处理器目前已配备了英特尔® 内置加速器，并且即将推出下一代内置加速器。

第四代英特尔® 至强® 可扩展处理器将内置英特尔® 高级矩阵扩展（英特尔® AMX）、英特尔® QuickAssist 技术（英特尔® QAT）和英特尔® 数据流加速器（英特尔® DSA）等内置功能。下面是上述功能的简要介绍。

英特尔® AMX

作为可提升深度学习性能的下一代英特尔® DL Boost，英特尔® AMX 配备了矩阵乘法指令集，将大幅提升 AI 推理和训练性能，每秒 INT8 图像推理次数是上一代产品的 4.5 倍[13]。

英特尔® QAT

英特尔® QAT 现在作为英特尔® 至强® 可扩展处理器的一部分，将为用户提供更快的数据加密和更高效的数据压缩，适用于从网络到企业、从云端到存储、从内容分发到数据库各种应用场景。

英特尔® DSA

英特尔® DSA 是一款高性能加速器，旨在优化网络、数据处理密集型应用和高性能存储中常见的流数据传输和转换操作。

结论

英特尔长期致力于创新和集成业务，在针对英特尔® 至强® 可扩展处理器开发这种新型内置加速器方面具有得天独厚的优势。针对各种工作负载打造的集成电路将为客户创造更大的商业价值。

无论您是希望提高性能，支持可持续发展计划，还是希望能够保护敏感数据，英特尔的内置加速器家族都能提供相应的解决方案，让您无需另行购买硬件。英特尔的内部测试和实际用例均表明，与市面上其他 CPU 选择方案相比，这些加速器带来的价值不可小觑。

若要详细了解英特尔® 至强® 可扩展处理器，请访问
https://www.intel.cn/xeonscalable。

[1] 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/。
[2] 详情请见以下网址的 [121]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。
[3] 详情请见以下网址中文件的第 12 页：https://www.intel.com/content/dam/www/public/us/en/documents/product-overviews/dl-boost-product-overview.pdf。
[4] 详情请见 https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/hyhy-customer-story.html。
[5] 详情请见 https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/kfbio-ai-customer-story.html。
[6] 详情请见以下网址的 [118]：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/。
[7] 详情请见 https://www.intel.com/content/www/us/en/customer-spotlight/stories/university-at-buffalo-customer-story.html。
[8] 详情请见 https://www.intel.cn/content/www/cn/zh/newsroom/news/broad-institute-intel-google-advance-biomedical-research.html。
[9] 详情请见 https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/cern-inference-customer-story.html。
[10] 详情请见 https://www.intel.cn/content/www/cn/zh/customer-spotlight/stories/nationwide-building-society-customer-story.html。
[11] 详情请见 https://www.intel.com/content/www/us/en/customer-spotlight/stories/swiss-re-customer-story.html。
[12] 详情请见 https://www.intel.com/content/www/us/en/newsroom/news/ucsf-propel-medical-device-innovations.html。
[13] 详情请见以下网址的 [41] 和 [42] 基准测试：https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/vision-2022/。结果可能不同。