英特尔® 至强® 可扩展处理器凭借内置加速器解决重大工作负载挑战的五大方式

工作负载在不断演变，计算机架构也在持续发展。曾经，增加 CPU 内核或选择更高频率的 CPU 就可以提高工作负载效率。而现在，这些方法再也无法保证过去能够实现的性能效率优势。如今，专门为人工智能 (AI)、安全性、科学计算、网络、存储和数据分析等特定功能而打造的集成加速器可以实现更多价值。

现有的和未来的英特尔® 至强® 可扩展处理器支持广泛而独特的内置硬件加速器，满足云端和企业部署中有关现代工作负载的需求。无论您是希望提升性能、降低成本，还是希望提高能效，**本文介绍的五种方式都能够通过内置加速器的英特尔® 至强® 可扩展处理器帮助您所在的企业解决重大工作负载挑战。**

01

减少另行添置硬件的需求

硬件越多意味着系统成本越高；同时，也可能意味着企业在增加设备时，更容易因为遇到潜在瓶颈而导致效率低下。英特尔® 至强® 可扩展处理器内置多种加速器，开箱即可为用户提供所需的性能。这样一来，企业就不必另行购买和集成硬件，因此避免了一大笔开支。

虽然为了满足某些工作负载要求，需另行添加专用硬件，但在多种情况下，英特尔的内置加速技术本身就足以支持您高效运行多种工作负载。

例如，内置 AI 加速器的英特尔® 至强® 可扩展处理器就可以在执行其他关键任务的同一硬件上运行 AI 训练和推理及许多经典机器学习应用等复杂工作负载。英特尔® 至强® 可扩展处理器已经针对数据科学家使用的 TensorFlow 和 PyTorch 等热门 AI 框架进行了优化。

02

更快完成工作负载

英特尔直接作用于那些热门的 AI 工具、框架和解决方案，优化它们在英特尔® 产品上的性能。这有助于企业通过 CPU 实现出色的训练和推理。英特尔® 至强® 可扩展处理器内置英特尔® 深度学习加速技术（英特尔® DL Boost），这种内置加速器能够提高常见 AI 工作负载的性能。

客户在第三代英特尔® 至强® 可扩展处理器上使用面向英特尔® 架构优化的 TensorFlow 和 英特尔® DL Boost 所获得的 AI 推理性能要比在第二代英特尔® 至强® 可扩展处理器上高出 11 倍以上[1]。即将推出的第四代英特尔® 至强® 可扩展处理器内置英特尔® 高级矩阵扩展（英特尔® AMX），每秒 INT8 图像推理次数是上一代产品的 4.5 倍[2]。

11 倍

以上的 AI 推理性能提升
（与第二代英特尔® 至强® 可扩展处理器相比）

4.5 倍

每秒 INT8 图像推理次数提升
（与上一代产品相比）

> “我们发现功耗相差无几，但性能提升巨大。主要差异在于能效上的显著提升。”
>
> **Patrick Kennedy**
> *ServeTheHome*

03

提高能效

由于英特尔的内置加速器有助于大幅提高能效，因此您可以提高各种工作负载的性能，而无需在服务器机架上增加独立加速器。

近期，来自 ServeTheHome（一个专为 IT 专业人士提供指南的网站）的 Patrick Kennedy 强调了这一优势。针对基于英特尔® DL Boost 运行 AI 工作负载的表现，他指出：“我们发现功耗相差无几，但性能提升巨大。主要差异在于能效上的显著提升[3]。”

04

在保持高性能的同时保护敏感数据

英特尔® 至强® 可扩展处理器支持机密计算解决方案，因此可以更好地保护本地、边缘和云端的数据。英特尔® 软件防护扩展（英特尔® SGX）是针对安全性的内置加速器，它是目前市场上一种大量部署、经广泛研究且久经实战考验的数据中心用机密计算技术。

英特尔® SGX 有助于保护使用中的敏感数据和应用代码。这有助于抵御可能导致业务运营中断、关键数据受损或破坏合规性的违规、泄露或攻击事件。机密计算技术的受攻击面越大，敏感数据面临的风险就越大。英特尔® SGX 在系统中的受攻击面非常小[4]。

此外，英特尔® 密码操作硬件加速作为至强® 内核架构中的指令集，采用单指令多数据流 (SIMD) 技术，能够在每个时钟周期内处理更多加密运算，因此使一切运转得更快。这能够为需要进行可靠数据加密的应用提高总吞吐量，并尽可能降低对性能和用户体验的影响。

05

为科学计算工作负载带来更大内存容量和带宽

商界、科学界及学术界对提高计算性能的需求从未像现在这么强烈。英特尔帮助众多企业和机构设计了能够执行严苛工作负载的系统架构。无论这些企业和机构是在努力攻克医学、经济方面的重大挑战，还是在解决工程领域的难题，英特尔® 至强® 可扩展处理器内置的科学计算加速器都可以提高工作负载性能，完成非常先进的计算任务，并且速度比之前更快。

对于建模和仿真等数据密集型工作负载，英特尔® 至强® 可扩展处理器不仅支持代码利用英特尔® 高级矢量扩展 512（英特尔® AVX-512），而且还能提供更大的系统内存容量和带宽。这有助于提高在现有硬件上执行复杂工作负载的速度。

英特尔® 至强® 可扩展处理器内置的加速器专为特定细分市场设计和优化，不仅可以实现基于硬件的高性能工作负载加速，而且具有出色的成本效益和能效。

第四代英特尔® 至强® 可扩展处理器将内置多种加速器，包括英特尔® 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）。这些加速技术将不断提升性能、优化结果，使客户能够比之前更快、更可持续、更成功。

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 [118]：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 [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。

[3] Kennedy, Patrick。“Deep Dive into Lowering Server Power Consumption”（深入探讨降低服务器功耗问题），Serve The Home，2022 年 2 月 21 日。https://www.servethehome.com/deep-dive-into-lowering-server-power-consumption-intel-inspur-hpe-dell-emc/2/。

[4] 详情请见 https://www.intel.cn/content/www/cn/zh/architecture-and-technology/software-guard-extensions-enhanced-data-protection.html。