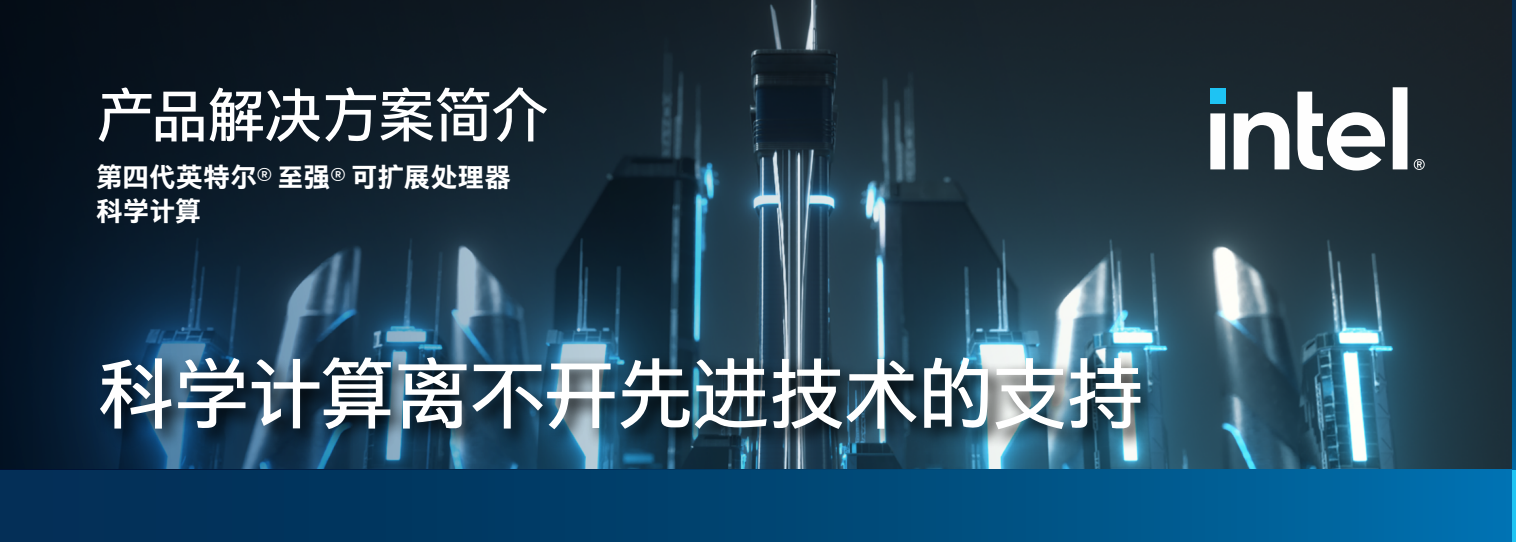
产品解决方案简介

第四代英特尔® 至强® 可扩展处理器
科学计算



科学计算离不开先进技术的支持

第四代英特尔® 至强® 可扩展处理器单核性能更高、核数更多、I/O 和内存子系统更强，并且配备了一系列内置硬件加速器，从而能为科学计算工作负载带来诸多助益。

AI 与科学计算工作负载的不断融合从新的维度带来性能挑战。除了要满足生命科学、材料科学、制造、仿真/建模以及金融等领域对科学计算工作负载的既有和不断增长的需求外，企业的基础设施还需满足企业级推理和训练对系统资源的需求，从而提供良好的最终用户体验。据此估计，全球科学计算市场将以 7.7% 的复合年增长率 (CAGR) 增长，到 2026 年将达到 592 亿美元[1]。

7.7% CAGR（到 2026 年）
全球科学计算市场增长[1]

592 亿美元（到 2026 年）
全球科学计算支出[1]

实现科学计算系统的平衡

第四代英特尔® 至强® 可扩展处理器为科学计算工作负载带来性能突破，助力缩短实现价值的时间。该平台采用全新架构，单核性能更高，每路配备多达 60 个内核，系统支持 2 路、4 路和 8 路配置。这相当于单核密度最高可达 120 个线程，比上一代产品增加了 50%。

为了与内核数增加这种情况相匹配，该平台在内存和 I/O 子系统方面也做了相应改进。DDR5 内存提供的带宽和速度最高可达 DDR4 的 1.5 倍，传输速率达到 4800 MT/s。此外，该平台每路有 80 条 PCIe Gen 5 通道，与之前的平台相比，I/O 得到显著提升。该平台还提供 CXL (Compute Express Link 1.1) 连接，支持高网络带宽并使附加加速器能够高效运行。

第四代英特尔® 至强® 可扩展处理器可为各类快速增长的工作负载提供性能加速。它内置多种针对特定应用的加速器，使 AI、数据分析、网络、存储和科学计算等领域工作负载的性能得到提升，其中包括：

- **期权定价：** 解决决策时间短、应用非常复杂且要求各不相同，以及随着 AI 应用愈发普及，市场需求不断变化等问题。
- **生命科学应用：** 通过完善模型和执行大规模计算来提高仿真精确度，使科研和发现更快速高效。
- **计算机辅助工程：** 推进计算机辅助工程应用快速获得结果，助力降低成本、改善产品的安全和设计，并加速上市。

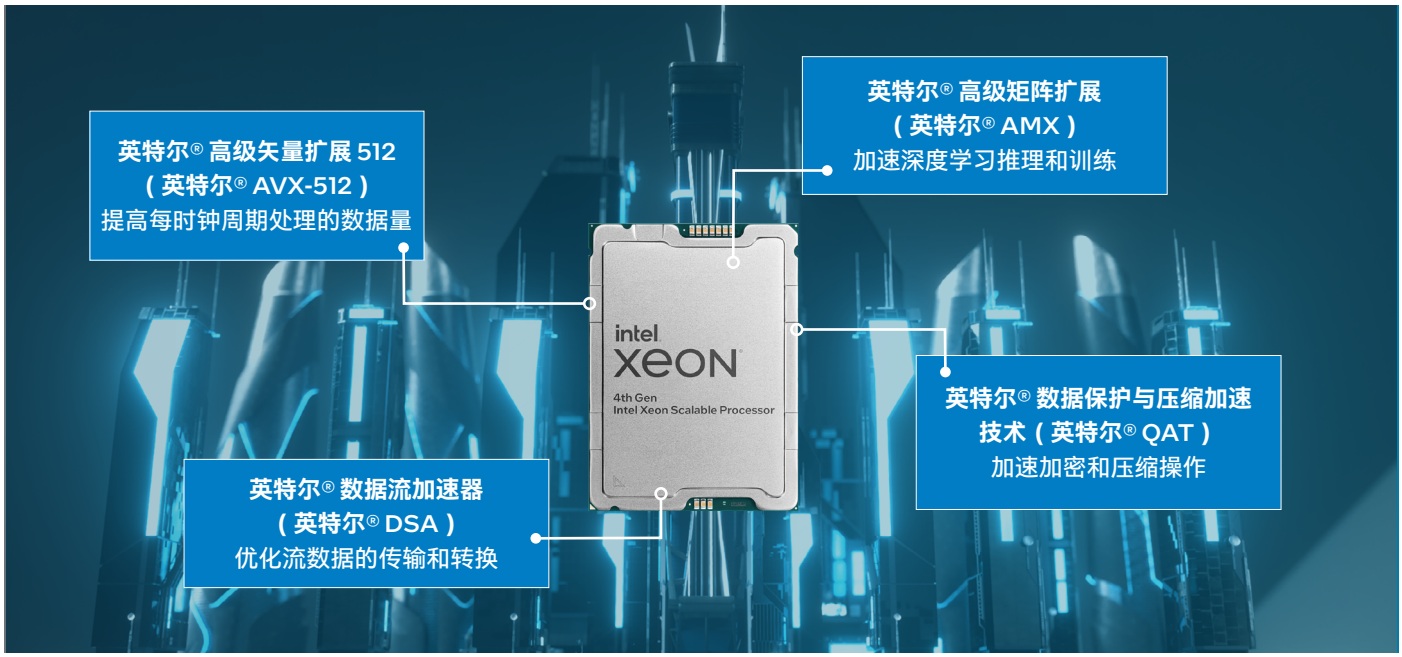
第四代英特尔® 至强® 可扩展处理器基于内置硬件加速器（包括面向科学计算的加速器，即英特尔® 科学计算引擎）引入一种实现高性能的新范式。

intel XEON

性能证明

高达 **1.56 倍** 几何平均数 为 28 个常见科学计算工作负载带来的性能提升（与上一代产品相比）[2,3]

英特尔® 科学计算引擎



性能证明

高达 **1.68 倍** 几何平均数　LAMMPS 工作负载性能提升（与上一代产品相比）[2,3]

基于内置加速器的先进功能

随着工作负载复杂性及其对计算资源的需求的提高，可以从 CPU 内核卸载某些功能，将那些执行资源留给业务关键型任务。这些功能包括 AI、安全以及常见的存储和网络功能。

直接内置于第四代英特尔® 至强® 可扩展处理器芯片的硬件加速器能够加速平台内的数据传输和处理。由于它们内置于处理器中，与独立解决方案或内核上运行的基于软件的解决方案相比，不会产生访问 PCIe 总线的时延，相应地，就节省了能耗。利用这些内置加速器的用例可以实现更好的性能并节省资本支出 (CapEx) 和运营支出 (OpEx)。

- **性能：** 专用的加速器大幅提升目标工作负载的吞吐量。
- **设备成本：** 由于加速器内置于第四代英特尔® 至强® 可扩展处理器中，因此无需另外的设备投资。
- **运营成本：** 由于内置加速器减少了在机架中增加内核的需求，因此可以节省大量能源。

英特尔® 高级矩阵扩展（英特尔® AMX）：加速深度学习

事实证明，机器学习可以卓有成效地进行科学计算工作负载调优，实现效率与效能的提升。英特尔® 高级矩阵扩展（Intel® Advanced Matrix Extensions，英特尔® AMX）是一种内置的硬件加速器，可以通过加速深度学习算法的核心——张量处理，显著提高推理和训练性能。该技术包括 TILE 和 TMUL（平铺矩阵乘法）两部分，前者由一组可扩展的 2D 寄存器组成，每核最多 8 个 TILE，可存储比上一代产品更大的数据块；后者是一组矩阵乘法指令，是 TILE 上的首批算子。英特尔® AMX 使深度学习软件能够在给定时间段内完成更多推理，或者更快地部署解决方案，从而加速实现价值。

英特尔® 高级矢量扩展 512（英特尔® AVX-512）：最新 x86 矢量指令集

经过多代技术发展，精度逐渐提升的矢量化技术有助于在更大的数据集上更快完成计算。英特尔® 高级矢量扩展 512（Intel® Advanced Vector Extensions 512，英特尔® AVX-512）作为最新 x86 矢量指令集，构建于前几代技术的矢量处理能力基础上，可加速完成数据密集型工作负载。借助两个 512 位融合乘加 (FMA) 单元，科学计算应用在 512 位矢量内的每个时钟周期可打包 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，以满足苛刻的计算工作负载需求，推动商业智能。与英特尔® 高级矢量扩展 2（Intel® Advanced Vector Extensions 2，英特尔® AVX2）相比，英特尔® AVX-512 使数据寄存器的宽度和数量以及融合乘加单元的宽度都增加了一倍。

英特尔® 数据流加速器（英特尔® DSA）：优化流数据传输

数据传输和转换操作对存储、网络和数据密集型工作负载（例如科学计算中的数据分析）的性能来说至关重要。英特尔® 数据流加速器（Intel® Data Streaming Accelerator，英特尔® DSA）能够卸载大规模部署中会产生开销的常见数据传输任务，藉此提升这些功能的性能。通过承担包括校验、内存比较和检查点在内几乎所有的数据传输操作，英特尔® DSA 可以使 CPU 内核资源避免在数据移入移出内存、存储和网络子系统方面产生开销。英特尔® DSA 优化了跨 CPU、内存和缓存以及各种附加内存、存储和网络设备的流数据传输。

英特尔® 数据保护与压缩加速技术（英特尔® QAT）：提升加密和压缩速度

减少加密和数据压缩相关开销对于提升集群整体性能发挥着重要作用。英特尔® 数据保护与压缩加速技术（Intel® QuickAssist Technology，英特尔® QAT）作为硬件加速器内置于第四代英特尔® 至强® 可扩展处理器中，可实现更快的动态数据加解密、更高效的数据压缩。与前几代技术相比，这一技术的最新版本在加解密算法、安全哈希、公钥加密和压缩/解压缩方面的表现更出色。它通过从处理器内核卸载这些任务，释放出内核资源用于处理其他工作，因此提升了总吞吐量。英特尔® QAT 有助于实现零信任安全策略，在各种基础设施中对各个阶段（静态下、传输中、使用中）的数据实施保护，而无损关键工作负载的性能。

主要技术

多达 60 个内核	多达 8 条内存通道	多达 80 条通道
（每路）	（DDR5，传输速率高达 4800 MT/s）	（每路的 PCIe 5.0）
高达 50% 的增幅[2]；系统支持 2 路、4 路或 8 路配置	高达 50% 的内存带宽和速度提升[2]	I/O 容量增幅[2]

第四代英特尔® 至强® 可扩展处理器具有更出色的单核浮点运算性能和一系列内置硬件加速器

开发人员赋能和支持

英特尔® oneAPI 工具套件是英特尔长期坚守对科学计算软件生态系统的承诺并不断演进的产物，它提供编译器、库和性能工具，能够简化面向英特尔® 架构优化的高质量软件的开发路径。这些工具套件为那些想要利用第四代英特尔® 至强® 可扩展处理器内置加速器的开发人员提供了捷径，以及基于标准的开源软件开发堆栈。开发人员可以利用英特尔® oneAPI 工具套件生成代码，全面提高各英特尔® 架构（包括内置加速器的 CPU、GPU 和 FPGA）的性能。

英特尔® oneAPI 基础工具套件

内核编译器、库（包括英特尔® oneAPI 数学核心函数库）和其他工具，用于开发以数据为中心的高性能应用

英特尔® oneAPI HPC 工具套件

英特尔® Fortran 编译器；支持 OpenMP 将指令卸载到 GPU；可通过消息传递接口（MPI）进行扩展

英特尔® AI 分析工具套件

优化的框架和 Python 库，可加速数据科学和分析管道

英特尔® oneAPI 渲染工具套件

渲染和光线追踪库，用于打造高性能、高保真视觉体验

由开源工具、API 和驱动程序等组成的大型开放式生态系统为基于 oneAPI 的开放标准代码开发提供了便利。这种灵活性有助于企业和机构降低将新服务和解决方案推向市场的复杂性、成本和时间要求，简化了新架构的落地，并使工程师和程序员能够将精力放在创新而不是维护代码上。

利用既有实现方案轻松集成

与英特尔合作，企业可以利用他们已经熟悉和正在使用的大规模合作伙伴生态系统缩短部署时间。全球各地的硬件和软件供应商以及解决方案集成商都在使用英特尔® 至强® 可扩展处理器构建其产品，并通过数以千计来自真实场景的实现案例提供更多选择和更好的互操作性。

性能证明

更高的 VASP 性能[2,3]

高达 **1.61 倍** 几何平均数 第四代英特尔® 至强® 可扩展处理器与上一代产品相比

高达 **2.01 倍** 几何平均数 英特尔® 至强® CPU MAX 系列与双路第三代英特尔® 至强® 可扩展处理器相比

为满足各种科学计算用例而设计

凭借高性能、DDR5 带来的更高内存带宽，以及 PCIe Gen 5 和 CXL 1.1 实现的先进 I/O，第四代英特尔® 至强® 可扩展处理器可为一系列实际用例加速。借助英特尔先进的软件库和编译器，开发人员能够更快速地构建代码，开发性能更佳且开箱即用的科学计算应用。借助强大的英特尔® AVX-512 技术和每内核 2 个 FMA 单元，代码和模型可满足严苛的计算工作负载要求。利用英特尔® MPI 库，工作负载能够在多个科学计算集群中进行扩展。此外，您还可配置英特尔® 傲腾™ 持久内存，在更大的内存中支持大型计算任务。

利用支持科学计算工作负载的英特尔® 技术实现更多可能

提升带宽：与仅采用 DDR5 的平台相比，全新英特尔® 至强® CPU Max 系列通过消除建模、AI、科学计算和数据分析等内存敏感型工作负载的瓶颈，将性能提升高达 4 倍。这是英特尔首款将高带宽内存和加速器整合到处理器封装中的 x86 CPU，其中 HBM2e 内存容量最高可达 64 GB。它减少了对 DDR 的依赖，可支持最新软件工具并且具有出色的代码复用性，因此降低了 TCO。

扩大影响：旗舰产品英特尔® 数据中心 GPU Max 系列采用英特尔先进的 IP 和封装技术，旨在加速 AI、科学计算和高级分析工作负载，满足 E 级时代的要求。该系列基于英特尔® Xᵉ HPC 架构，GPU 中配备有高带宽缓存。在 oneAPI 开放生态系统的支持下，GPU 展现了出色的灵活，既可处理 SIMT（Single Instruction Multiple Threads，单指令多线程），也可处理 SIMD（Single Instruction Multiple Data，单指令多数据），它的封装内集成了多项 IP 创新技术，包括高带宽内存。

微秒级数据访问：DAOS（分布式异步对象存储）是一种开源的软件定义横向扩展对象存储系统，可以在单一存储层中经济高效地为科学计算和 AI 应用提供高带宽、低时延和高 IOPS 的存储容器。DAOS 原生支持结构化、半结构化和非结构化数据集，同时还摆脱了传统分布式存储的局限性。

了解更多信息

www.intel.cn/xeon/scalable

www.intel.cn/hpc

[1] Intersect360 Research，2022 年 5 月 20 日，“Total HPC Market Revenue Grew 5.2% to $41.0 Billion in 2021, Says Intersect360 Research”（Intersect360 Research 指出，2021 年科学计算市场总收入增长 5.2%，达到 410 亿美元）。https://www.hpcwire.com/off-the-wire/total-hpc-market-revenue-grew-5-2-to-41-0-billion-in-2021-says-intersect360-research/。

[2] 与双路第三代英特尔® 至强® 可扩展处理器的对比。

[3] 详情请见 https://edc.intel.com/content/www/cn/zh/products/performance/benchmarks/processors/（第四代英特尔® 至强® 可扩展处理器）。结果可能不同。

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。

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 https://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。

1122/JAW/MESH/PDF