白皮书

第四代英特尔® 至强® 可扩展处理器

第四代英特尔® 至强® 可扩展处理器助力青云 QingCloud 新一代 e4 云服务器实现性能突破

QingCloud Technologies

“第四代英特尔® 至强® 可扩展处理器内置了强大的加速器，帮助我们将 e4 云服务器的性能提升至新的水平，其不仅能够满足大数据、数据库等场景需求，在 AI 推理、云原生、高性能计算等更多场景下也有着不俗表现。通过 e4 具备的优秀计算性能、弹性伸缩和分布式存储等特点，企业能够高效开发和运行高负载的复杂应用以及 AI 服务，加速数字化转型。”

—沈鸥

青云科技副总裁

解决方案概述

为了应对数字化转型带来的重重挑战，用户希望获得更高的性能、更安全的数字化服务、更稳定的基础平台。数据中心与云服务提供商必须更加精准且前瞻性地洞察到当前行业正在发生的改变，并通过基础设施架构优化、采用新一代硬件平台、技术与服务创新等方式，提供敏捷、灵活、高性能、高可用的解决方案，为用户数字化转型之旅提供基础能力支撑，帮助用户在商业竞争中赢得先机。

为帮助企业用户更好地应对云原生趋势对 IT 架构带来的挑战，青云科技推出了搭载第四代英特尔® 至强® 可扩展处理器的新一代 e4 云服务器。该服务器利用处理器内置的英特尔® 高级矩阵扩展（英特尔® AMX）、英特尔® QuickAssist（英特尔® QAT）等特性，加速 AI、数据分析、数据加解密等场景下的处理能力。同时青云还验证了第四代英特尔® 至强® 处理器的 In-Memory Analytics Accelerator（英特尔® IAA）、英特尔® Data Streaming Accelerator（英特尔® DSA）、英特尔® Software Guard Extensions（英特尔® SGX）等高级硬件特性对多种应用场景的提升能力。例如，通过英特尔® Software Guard Extensions（英特尔® SGX）等硬件安全特性，可助力企业提升安全保护能力，帮助企业实现云化，加速数字化转型。

挑战

在用户加速拥抱数字化的背景下，越来越多的数据与应用被迁移到云端环境，云平台相应地承受着越来越大的压力，这些压力包括：

- **负载日趋复杂化、带来了多元算力需求**

  在现代化的数据中心，负载正在日趋复杂化，人工智能（AI）、数据压缩、数据加解密等负载快速增长，通用计算处理这些负载的执行效率不高，带来了较高的性能压力。这意味着数据中心需要提供多元化算力，将负载卸载到特定的加速器上，以支持上层应用使用更优架构完成每项任务。

- **基础设施规模越来越大、总体拥有成本（TCO）持续攀升**

  用户在云原生等领域的持续投资意味着基础设施规模的不断增长，这带来了大量的服务器采购、运营等成本，在强调可持续发展、精益运营的今天，只有尽可能地提升基础设施的性能密度，释放硬件潜能，才能够更好地控制 TCO 增长，实现更高的投资收益。

● 数据安全面临严峻挑战

数字化产品、应用和服务都在源源不断地产生数据，这些数据是企业的重要竞争力，也是企业创新的重要基础。但是海量的数据如果没有安全的保护机制，很可能会造成核心数据丢失或泄露，从而使企业业务遭受巨大的冲击，甚至会让企业陷进生死存亡的风险境地。数据的安全性如何有效保护，是企业亟需解决的一大难题。

青云 QingCloud 全栈云助力企业数字化转型

青云科技是一家企业级云服务商与数字化解决方案提供商。自 2012 年创立以来，坚持核心代码自研，以卓越的技术实力见长，构建起端到端的数字化解决方案，持续打造云原生最佳实践。青云科技积极布局混合云市场，无缝打通公有云和私有云，交付一致功能与体验的混合云，并于 2021 年 3 月登陆上交所科创板。

青云科技坚持自主创新、中立可靠、灵活开放的理念，立足企业现实需求，围绕“私有云、公有云、云原生、信创”四大核心业务线，帮助企业构筑坚实的数字基石，实现全场景自由计算，为数字化创新添加“云动力”。

在服务层次上，纵向跨越 IaaS、PaaS 和应用平台的全栈云架构；在服务交付形态上，以统一架构实现公有云、私有云、混合云和托管云的一致化交付与管理；在服务场景纵深上，集结云、网、边、端一体化能力，实现全域智能数据互联。

作为青云全栈云方案的重要基石，青云 QingCloud 新一代 e4 云服务器实现了性能的大幅提升。该服务器基于第四代英特尔® 至强® 可扩展处理器，实现 CPU 性能提升 50%，存储 IOPS 性能提升 35%，网络延迟降低 30%[1]，支持 AMX、QAT 等指令集扩展，可广泛应用于 AI 推理、高性能数据库、高性能计算、大数据、计算密集型开发测试等业务场景。

图 1. 青云 QingCloud 的全栈云产品架构

[1] 数据援引自青云科技内部测试结果。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。

第四代英特尔® 至强® 可扩展处理器加速多种工作负载性能

为了给用户提供高性能的基础算力支撑，青云科技利用第四代英特尔® 至强® 可扩展处理器内置的多种高级硬件特性，优化应用负载性能，释放了处理器在性能、稳定性、扩展性、安全性等方面的潜力，铸就卓越基础设施平台。

第四代英特尔® 至强® 可扩展处理器通过创新架构增加了每个时钟周期的指令，每个插槽多达 56 个核心，支持 8 通道 DDR5 内存，有效提升了内存带宽与速度，并通过 PCIe 5.0（80 个通道）实现了更高的 PCIe 带宽提升。第四代英特尔® 至强® 可扩展处理器提供了现代性能和安全性，可根据用户的业务需求进行扩展。借助内置的加速器，用户可以在 AI、分析、云和微服务、网络、数据库、存储等类型的工作负载中获得优化的性能。通过与强大的生态系统相结合，第四代英特尔® 至强® 可扩展处理器能够帮助用户构建更加高效、安全的基础设施。

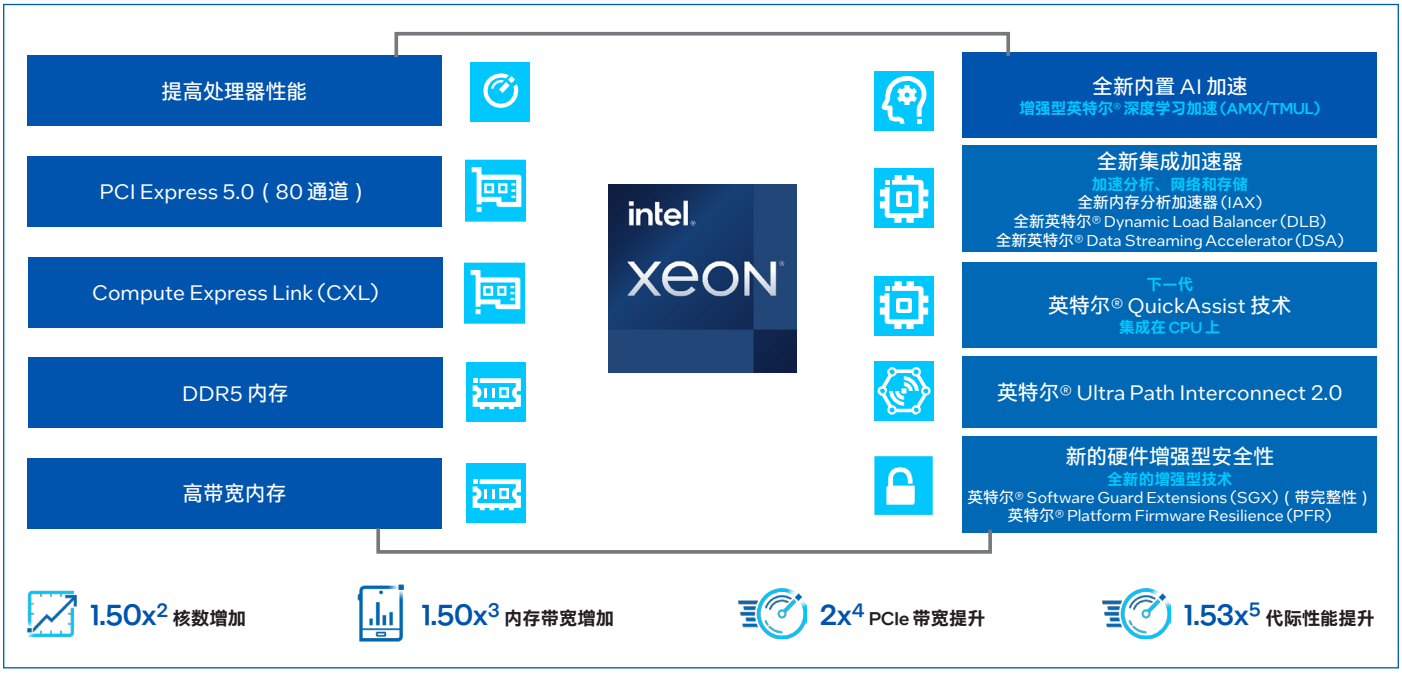

图 2. 第四代英特尔® 至强® 可扩展处理器为数据中心提供多种优势

第四代英特尔® 至强® 可扩展处理器内置了多种高级硬件特性，能够满足用户的多样化算力需求。其中，英特尔® AMX 针对广泛的硬件和软件优化，通过提供矩阵类型的运算，显著增加了人工智能应用程序的每时钟指令数（IPC），可为深度学习推理和训练提供显著的性能提升；英特尔® QAT 面向高性能安全性、私钥保护和压缩/解压缩等场景，能够将相关负载从 CPU 卸载到 QAT 中，有效提升应用程序和平台的性能；英特尔® DSA 优化存储、网络和分析中常见的流数据移动和转换操作；英特尔® IAA 可以加速数据库查询吞吐量和其他类型的工作负载，减少内存占用；英特尔® SGX 能够更有效地抵御多种类型的攻击，显著加强数据安全，满足对于机密计算的广泛需求；英特尔同时首次提供了板载 HBM 内存的英特尔® 至强® Max 系列处理器，为更广阔的市场带来了高带宽内存。

[2] 数据来源自第四代英特尔® 至强® 可扩展处理器的最大核数（60 核）与第三代英特尔® 至强® 可扩展处理器的最大核数（40 核）的比较。

[3] 详细配置信息请访问：intel.com/processorclaims，选择“第四代英特尔® 至强® 可扩展处理器”，查看编号“G2”。实际性能受使用情况、配置和其他因素的差异影响。

[4] 数据来源自第四代英特尔® 至强® 可扩展处理器（80 条 PCIe 5.0 通道）与第三代英特尔® 至强® 可扩展处理器（64 条 PCIe 4.0 通道）的比较。

[5] 详细配置信息请访问：intel.com/processorclaims，选择“第四代英特尔® 至强® 可扩展处理器”，查看编号“G1”。实际性能受使用情况、配置和其他因素的差异影响。

青云科技与英特尔重点从以下几个方面入手，验证第四代英特尔® 至强® 可扩展处理器对于常见负载的加速能力：

采用英特尔® AMX 提升 AI 性能

第四代英特尔® 至强® 可扩展处理器内置英特尔® AMX，无需配置额外的硬件即可加速深度学习推理和训练。英特尔® AMX 针对广泛的硬件和软件优化，它进一步增强了前代技术 — 矢量神经网络指令（VNNI）和 BF16，从一维向量发展为二维矩阵，以便最大限度地利用计算资源，提高高速缓存利用率，避免潜在的带宽瓶颈。

青云科技利用英特尔® AMX 提升基于 CPU 的 AI 性能，其支持中小型深度学习训练模型，大幅提高深度学习训练和推理性能，适合自然语言处理、推荐系统和图形识别等工作负载。数据如图 3 所示，在采用英特尔® AMX 优化之后，在满足精度的前提下，AI 推理性能模型，包括 Bert、Resnet 等，可以提升 4-5 倍[6]。

青云科技同时测试了在 e4 云主机上，通过使用英特尔® AMX，进行大模型 ChatGLM-6B 推理的性能表现。数据如图 4 所示，英特尔® AMX 能够为 e4 云主机带来显著的性能提升，e4 云主机 FP32（启用 AMX）相较于 e4 云主机 FP32（未启用 AMX）推理性能提升了约 6.26 倍，推理时延减少了 84.6%[7]。此外，e4 云主机使用 BF16+FP16 模式配合英特尔® AMX 特性有着更佳表现。e4 云主机 BF16+FP16（启用 AMX）相较于 e4 云主机 FP32（启用 AMX）的推理性能提升了约 1.32 倍，相应的推理时延减少了约 25.6%；e4 云主机 BF16+FP16（启用 AMX）相较于 e4 云主机 FP32（未启用 AMX）的推理性能提升了 6.85 倍，推理时延减少了 88.6%[8]。与使用 GPU 运行 ChatGLM-6B 推理相比，使用 e4 云主机搭配英特尔® AMX 加速器进行 ChatGLM-6B 推理有更好的总体拥有成本（TCO）。

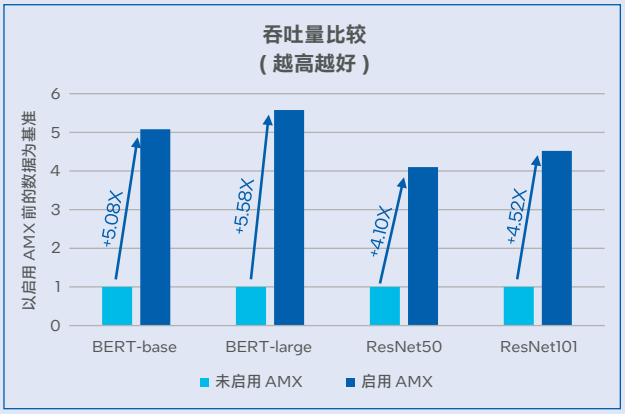

图 3. 启用英特尔® AMX 前后的吞吐量比较

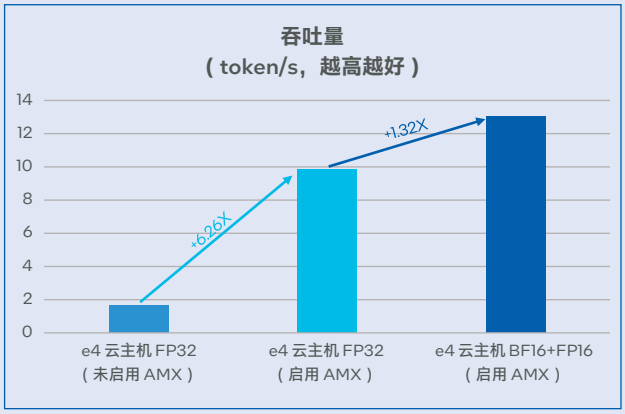

图 4. ChatGLM-6B 启用英特尔® AMX 前后的吞吐量比较

[6] 截至青云科技 2023 年 2 月的测试数据。测试配置：双路英特尔® 至强® 铂金 8480+ 处理器 @ 2.0 GHz，启用睿频加速技术，1024 GB 总内存（32x32 GB 4800MT/s），Ubuntu 22.04。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。

[7,8] 截至青云科技 2023 年 7 月的测试数据。测试配置：e4 云主机，32 核 64g。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。

采用英特尔® QAT 优化数据压缩与加解密性能

在虚拟机实时迁移、分布式存储、负载均衡等应用负载中，数据压缩与加解密是非常重要的一个处理流程。例如，为了节省存储空间，存储系统开启压缩功能可有效地提高存储资源使用率，同时大幅降低采购成本；在负载均衡业务中，HTTPs 在身份验证、加密通讯等方面的特性带来了巨大的加解密计算需求；在虚拟机迁移中，开启压缩功能进行实时迁移，会在迁移前对内存中的数据进行压缩，有助于提升虚拟机迁移效率。但同时，数据压缩与加解密带来了巨大的性能消耗，不仅影响应用的效率，而且占用了大量的计算资源。

英特尔® QAT 是英特尔面向高性能安全性、私钥保护和压缩/解压缩等场景推出的一个硬件加速技术，能够将相关负载从 CPU 卸载到 QAT 中，有效提升应用程序和平台的性能。英特尔® QAT 能够以硬件方式支持多种对称数据加密（如 AES）、非对称公钥加密（如 RSA、椭圆曲线加密）和数据压缩服务，在不额外增加 CPU 负载的前提下，提高数据压缩与加解密效率。

青云科技首先验证了英特尔® QAT 对于 ZFS 存储系统压缩的加速效果。在英特尔的协助下，青云科技自主开发了支持 QAT + ACOMP 的 ZFS 压缩补丁，其利用内核态提供的 acomp 接口，并调用 deflate 算法来实现压缩与解压缩。测试数据如图 5 所示，在常规压力测试下，英特尔® QAT 能够大幅降低写吞吐与读吞吐的 CPU 使用率。

随后，青云科技测试了负载均衡 HAproxy 的性能表现，使用 ab 工具，不断加大线程数量，对相同配置的 HAproxy 进行压测，对比 QAT 启用前后的性能表现，测试数据显示，在同样的压力测试下，开启 QAT 之后，CPU 消耗更少，另外在大压力的情况下，开启 QAT 之后，系统吞吐量增加了约 15%，时延降低了约 10%[10]。

在 API 网关 OpenResty 的 HTTPs 服务性能测试中，青云科技测试了 QAT 启用前后的 OpenResty 的数据吞吐性能，数据如图 6 所示，OpenResty 使用 QAT 后加解密效率提升明显，其中 4 核 8 线程时提升率最高，达到未启用前的 6.8 倍[11]。

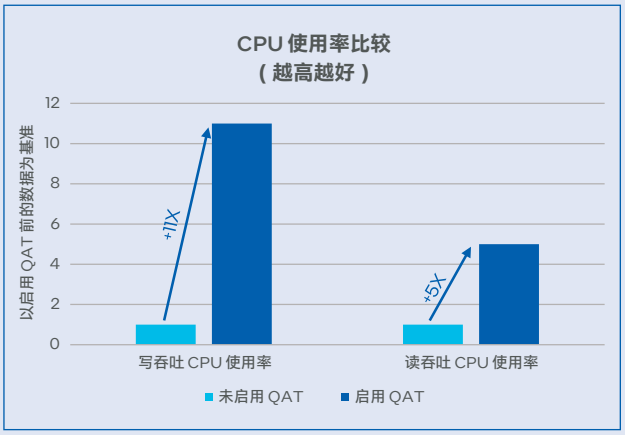

图 5. ZFS 存储系统压缩性能对比[9]

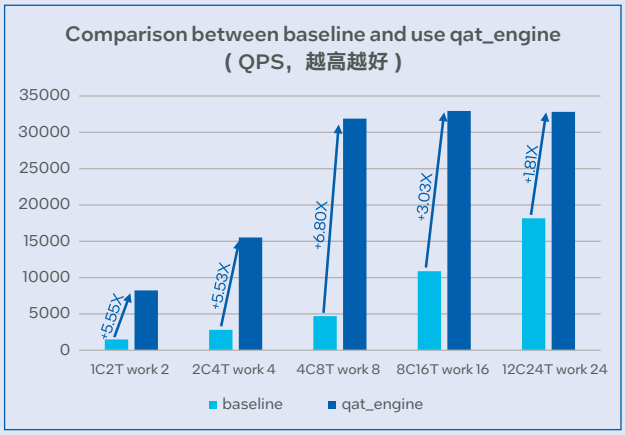

图 6. OpenResty HTTPs 服务性能测试

[9] 截至青云科技 2023 年 5 月的测试数据。测试配置：双路英特尔® 至强® 铂金 8458P 处理器 @ 3.80 GHz，512 GB 总内存 (16x32 GB 4800 MT/s)，480 GB SATA，3 TB 硬盘，Ubuntu 22.04。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。

[10] 截至青云科技 2023 年 4 月的测试数据。测试配置：第四代英特尔® 至强® 可扩展处理器，4 核 8G 计算型 e4，Ubuntu 22.04.1。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。

[11] 截至青云科技 2023 年 5 月的测试数据。测试配置：英特尔® 至强® 铂金 8458P 处理器，Ubuntu 22.04.2 LTS。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。

在虚拟机实时迁移测试中，青云科技利用 QAT 技术将压缩解压缩卸载至 QAT，来释放更多的计算资源，以及提升压缩速度，从而达到提升实时迁移效率的目的。测试数据如图 7 所示，在无负载情况下，使用 QAT 压缩相比原压缩方式，迁移时间减少约 66%，压缩率增加约 13 倍，数据压缩时的 CPU 使用率降低约 81%[12]。

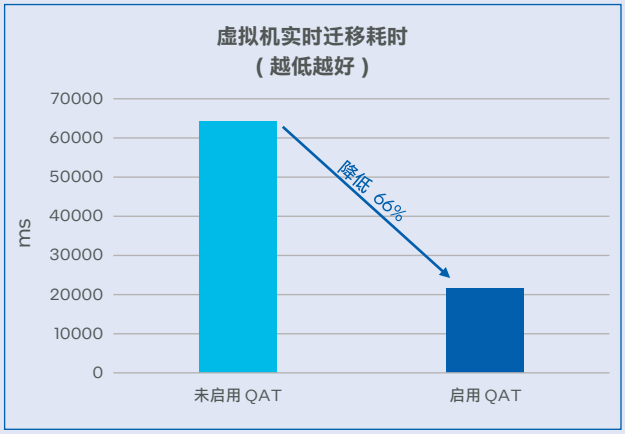

图 7. QAT 启用前后虚拟机实时迁移耗时比较

采用英特尔® IAA 加速数据库

MongoDB、ClickHouse 是当前常见的数据库应用，其中，MongoDB 是免费开源的跨平台 NoSQL 数据库，ClickHouse 则是一个用于联机分析处理（OLAP）的开源列式数据库。为了在优化数据库性能的同时持续提升数据库压缩率，青云科技采用了英特尔® IAA 进行优化。

英特尔® IAA 是一款硬件加速器，结合分析原始函数，能够提供出色的吞吐量压缩和解压缩性能。英特尔® IAA 主要针对大数据和内存分析数据库等应用程序，以及内存页压缩等应用程序透明用途，能够在分析查询处理期间过滤数据。英特尔® IAA 支持零压缩等轻量级压缩方案以及霍夫曼编码和 Deflate 等较重的压缩算法。对于 Deflate 格式，它支持对压缩流进行索引，以实现高效的随机访问。

MongoDB 的吞吐量测试数据如图 8 所示，对比 Zlib 压缩算法，IAA 将性能提升了 85.63% ~ 548.57%，对比 Snappy 算法，其性能最高可提升 49.91%[13]。此外，IAA 在 MongoDB 中相较于上述两种算法，拥有更大的压缩比，更低的时延，能够加速大数据查询过程。

ClickHouse 的测试数据显示，与 LZ4 相比，IAA 方案提供了 62% 的压缩效果和 35% 的 QPS 增强效果。与 ZSTD 相比，IAA 方案可提供 50% 的 QPS 提高效果，压缩率下降 16%[14]。

[12] 截至青云科技 2023 年 5 月的测试数据。测试配置：双路英特尔® 至强® 铂金 8458P 处理器 @ 3.80 GHz，512 GB 总内存（16x32 GB 4800 MT/s），480 GB SATA，3 TB 硬盘，Ubuntu 22.04。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。

[13] 截至青云科技 2023 年 5 月的测试数据。测试配置：英特尔® 至强® 铂金 8458P 处理器，Ubuntu 22.04.1 LTS。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。

[14] 截至青云科技 2023 年 5 月的测试数据。测试配置：英特尔® 至强® 铂金 8458P 处理器，512 GB 总内存（16x 32 GB 4800 MT/s），Ubuntu 22.04.1 LTS。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。

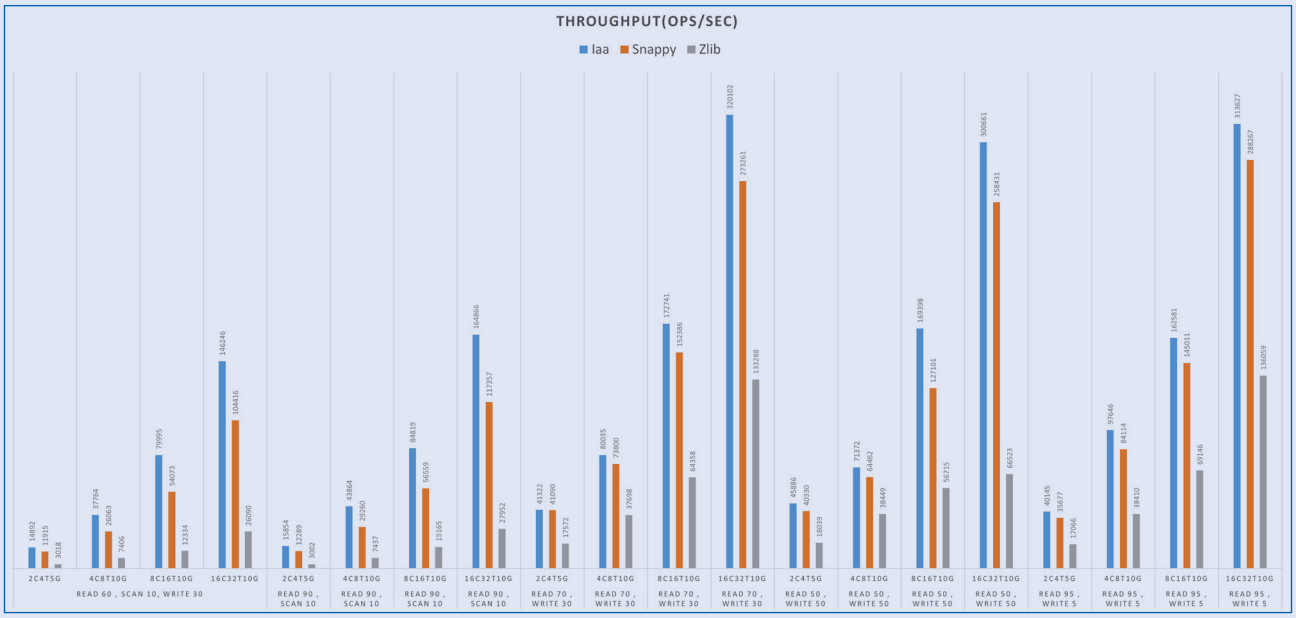

图 8. MongoDB 吞吐量测试（越高越好）

通过英特尔® SGX 构建可信环境

在云生态中，用户广泛面临着病毒、木马、网络攻击、数据窃取等安全威胁，大部分传统的安全方案主要依赖于特权代码来实现工作负载的隔离和数据的保护，难以防范利用特权代码漏洞的攻击，在安全防护能力方面亟待进一步提升。为了解决此问题，青云验证了第四代英特尔® 至强® 可扩展处理器内置的英特尔® SGX 的安全功能，打造了可信计算环境的能力。

英特尔® SGX 能够帮助用户构建基于硬件的数据中心可信执行环境（TEE）。通过将特权代码排除在受信任的范围之外，英特尔® SGX 能够更有效地抵御多种类型的攻击。它可显著加强数据安全，满足对于机密计算的广泛需求。英特尔® SGX 提供了一种基于硬件的内存加密机制，将内存中的特定应用代码和数据隔离开来。英特尔® SGX 允许为用户级代码分配专用内存区域 — 飞地（Encalve），以免受到拥有更高权限的进程的影响。

青云科技的验证显示，在云服务器上，可以启用英特尔® SGX 技术来构建可信的密钥管理服务，提供密钥计算、交换等复杂的安全计算环境，保护应用与数据的安全。

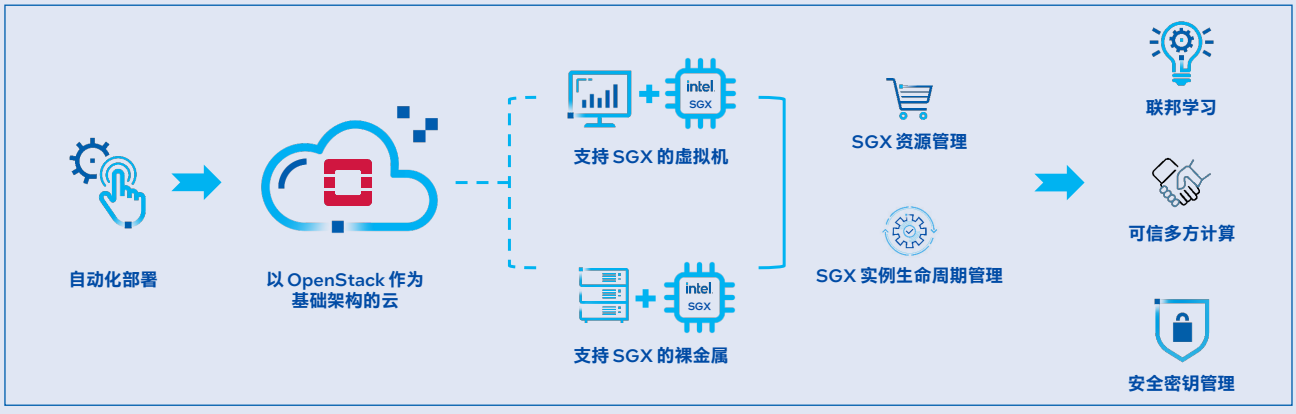

图 9. 英特尔® SGX 可支持构建可信环境

利用 HBM 内存加速应用的内存访问

英特尔® 至强® Max 系列处理器是唯一一款基于 x86 的高带宽内存处理器，为解锁和加速受内存限制的 HPC 和人工智能工作负载而设计。英特尔®至强® Max 系列处理器通过高带宽内存（HBM）为英特尔® 至强® 可扩展处理器提供增强功能，旨在释放建模、人工智能、深度学习、高性能计算（HPC）和数据分析等数据密集型工作负载的性能并提升发现速度。

青云科技选择 HPL、VASP、lammps 三个软件，测试在高性能计算（HPC）集群中，HBM 内存相较于 DDR 内存的性能提升。以 VASP 为例，该软件是电子结构计算和量子力学-分子动力学模拟软件包，是材料模拟和计算物质科学研究中较为流行的商用软件之一。测试数据如图 10 所示，随着核心数的增加，使用 HBM 内存的效果提升很明显，使用 HBM 内存 22 核心的性能基本上和 44 核心的 DDR 内存计算效率持平[15]。

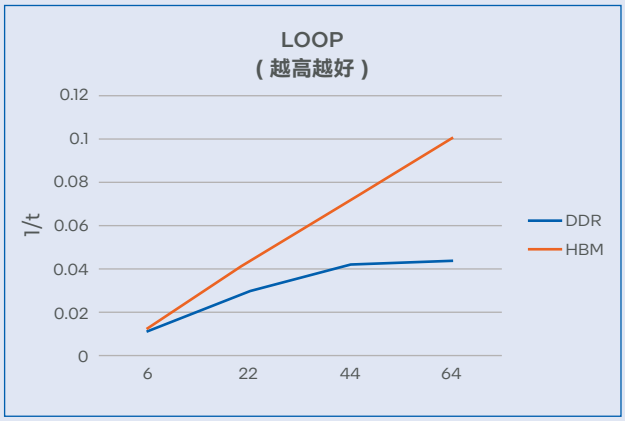

图 10. HBM 内存与 DDR 内存性能对比
（横坐标为核心数，纵坐标为时间的倒数）

展望

通过搭载第四代英特尔® 至强® 可扩展处理器，并利用处理器集成的高级硬件特性，青云 QingCloud 新一代 e4 云服务器实现了巨大的性能飞跃，满足了企业对即时数据高并发、高吞吐量处理、低延迟等需求，通过提供更高性能、更稳定、更高性价比的基础支撑，帮助企业实现云化，加速数字化转型。

未来，青云科技还将与英特尔深化合作，进一步针对云计算、隐私计算、数据库、大数据、AI 等具体场景推动软硬件协同优化，释放第四代英特尔® 至强® 可扩展处理器的潜能，为各行业的不同应用提供专业稳定的系统支撑。双方还将在云平台、云存储、人工智能、软硬件等多个领域展开了全面的深度合作，共同发挥所长、为中国云计算产业的创新发展高效赋能。

[15] 截至青云科技 2023 年 5 月的测试数据。测试配置：英特尔® 至强® MAX 9432 处理器，500 GB 总内存（128 GB HBM + 372 GB DDR5），Ubuntu 22.04，英特尔® OneAPI2023，vasp5.4.4。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。

关于青云科技

北京青云科技股份有限公司（简称：青云科技），是一家企业级云服务商与数字化解决方案提供商。自 2012 年创立以来，坚持核心代码自研，以卓越的技术实力见长，构建起端到端的数字化解决方案，持续打造云原生最佳实践，以中国科技服务数字中国。青云科技积极布局混合云市场，无缝打通公有云和私有云，交付一致功能与体验的混合云，并于 2021 年 3 月登陆上交所科创板。

关于英特尔

英特尔（NASDAQ: INTC）作为行业引领者，创造改变世界的技术，推动全球进步并让生活丰富多彩。在摩尔定律的启迪下，我们不断致力于推进半导体设计与制造，帮助我们的客户应对最重大的挑战。通过将智能融入云、网络、边缘和各种计算设备，我们释放数据潜能，助力商业和社会变得更美好。如需了解英特尔创新的更多信息，请访问英特尔中国新闻中心 newsroom.intel.cn 以及官方网站 intel.cn。

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.com/PerformanceIndex

性能测试结果基于配置信息中显示的日期进行测试，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。

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。

英特尔未做出任何明示和默示的保证，包括但不限于，关于适销性、适合特定目的及不侵权的默示保证，以及在履约过程、交易过程或贸易惯例中引起的任何保证。

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。

© 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在美国和/或其他国家的商标。其他的名称和品牌可能是其他所有者的资产。