intel
xeon

至强秘笈大起底
一图读懂

第四代英特尔® 至强® 可扩展处理器

众多内置加速器 让性能更进一步

intel
xeon
4th Gen
Intel Xeon Scalable Processor

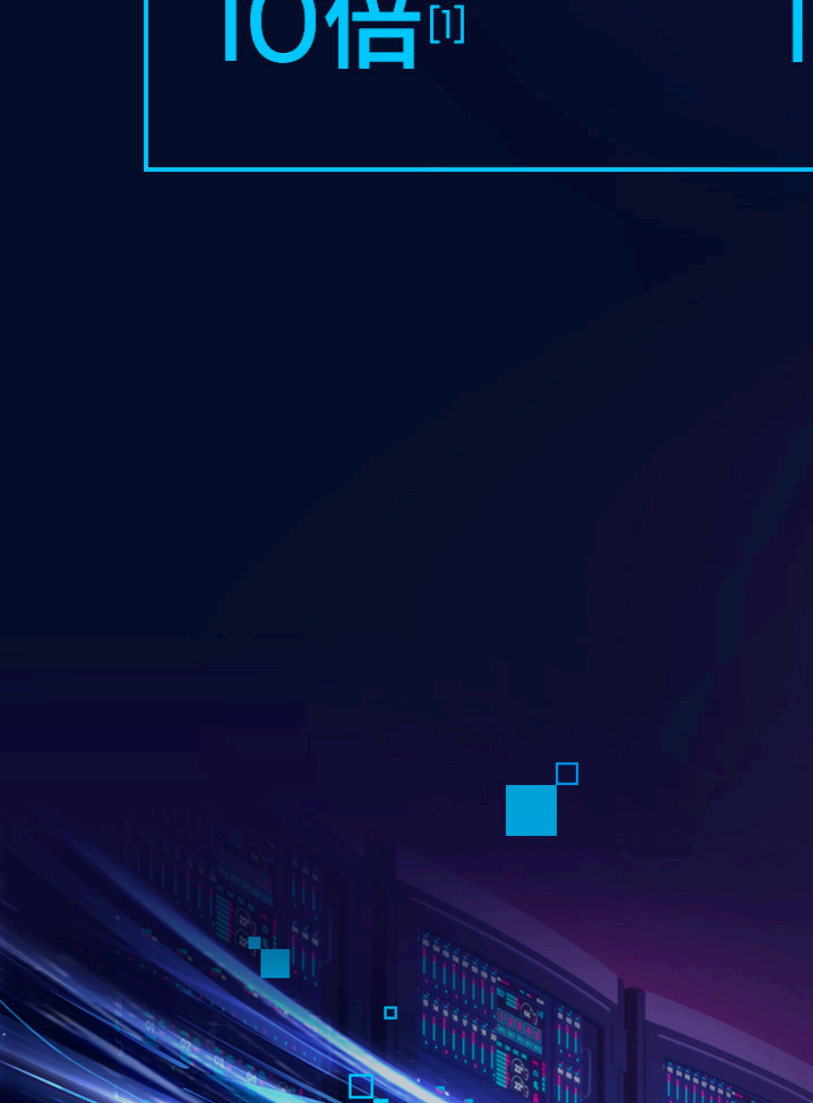
七大算力神器，加速产品技术创新

英特尔® 高级矩阵扩展
Intel® AMX

提升CPU上的AI性能
纳入调优功能
并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、推荐系统和图像识别等工作负载

PyTorch实时推理性能提升高达 10倍[1]

PyTorch训练性能提升高达 10倍[1]



英特尔® 数据保护与压缩加速技术
Intel® QAT

通过卸载加密、解密和压缩释放处理器内核
让系统能够支持更多客户端运行
或实现降低能耗的目的

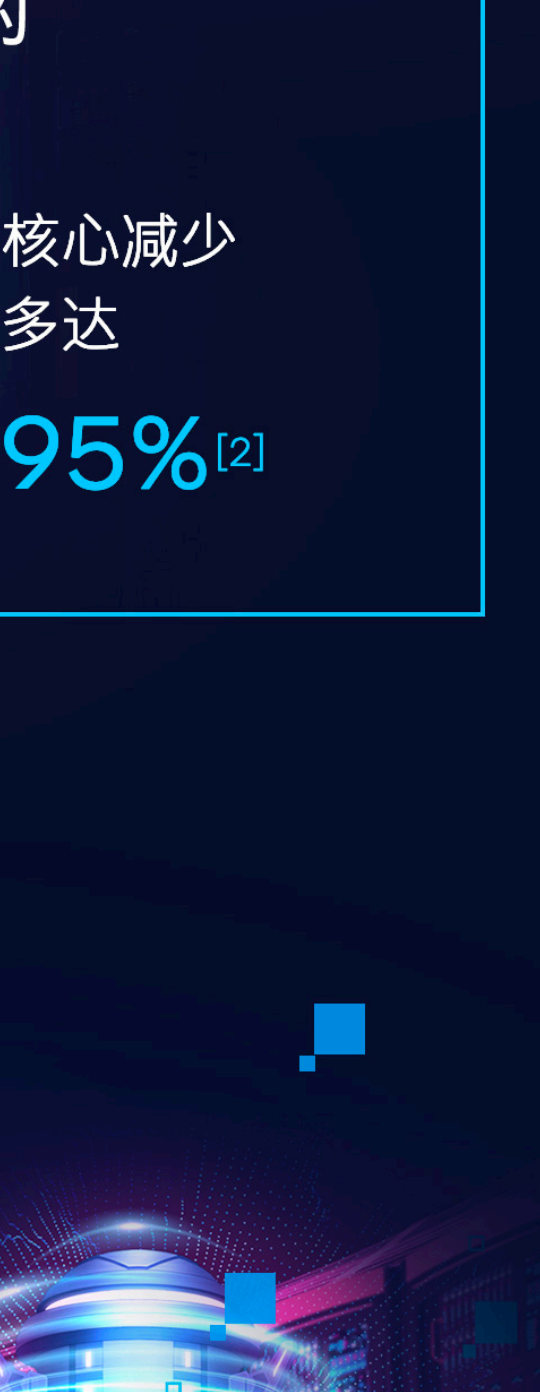
可将l级压缩吞吐量提高 2倍[2]

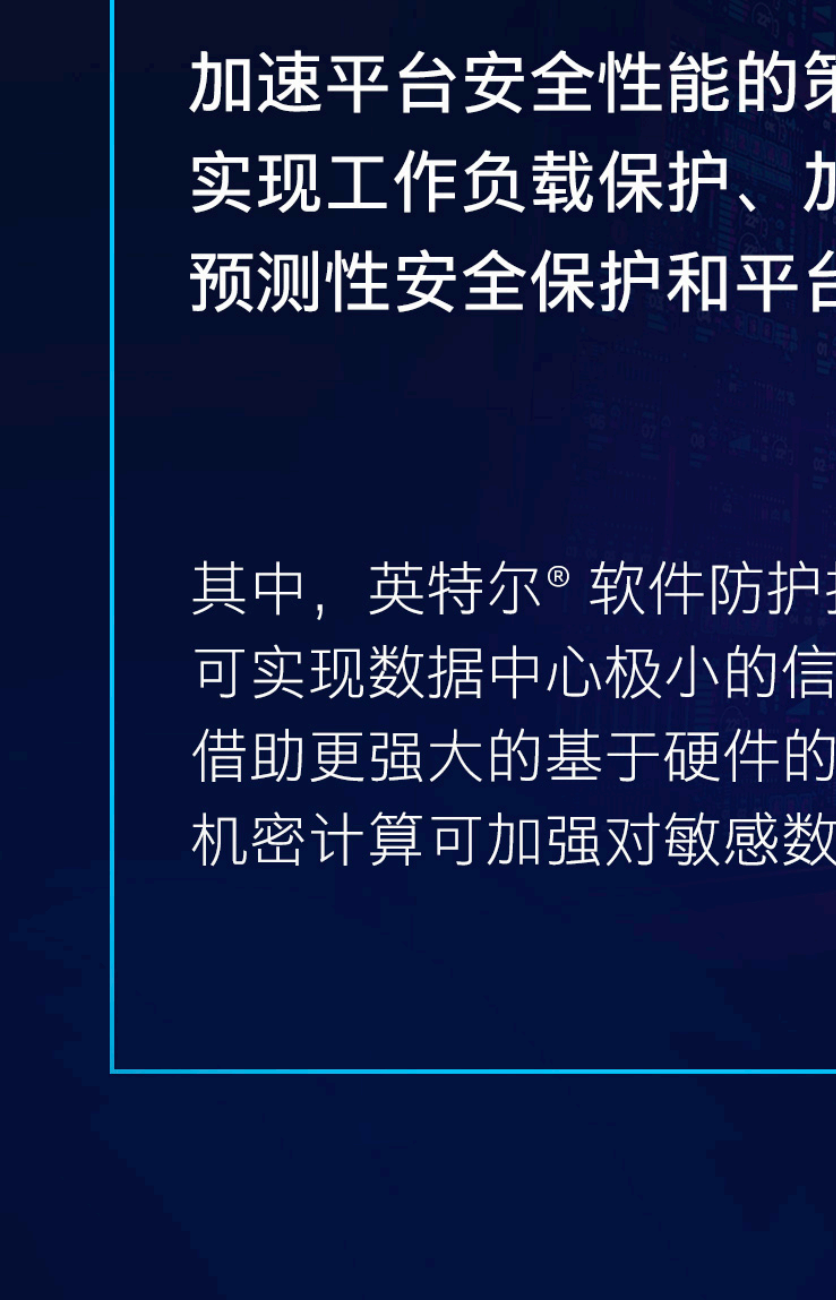
核心减少多达 95%[2]

英特尔® 安全技术
Intel® Security

加速平台安全性能的策略组合
实现工作负载保护、加密运算加速、预测性安全保护和平台安全启动

其中，英特尔® 软件防护扩展（Intel® SGX）可实现数据中心极小的信任边界
借助更强大的基于硬件的内存保护
机密计算可加强对敏感数据的隔离





英特尔® 动态负载均衡器
Intel® DLB

可根据系统负载变化
动态地在多个CPU内核/线程上高效分配网络数据以进行处理
且能够还原在多个CPU内核上同时处理网络数据包的顺序

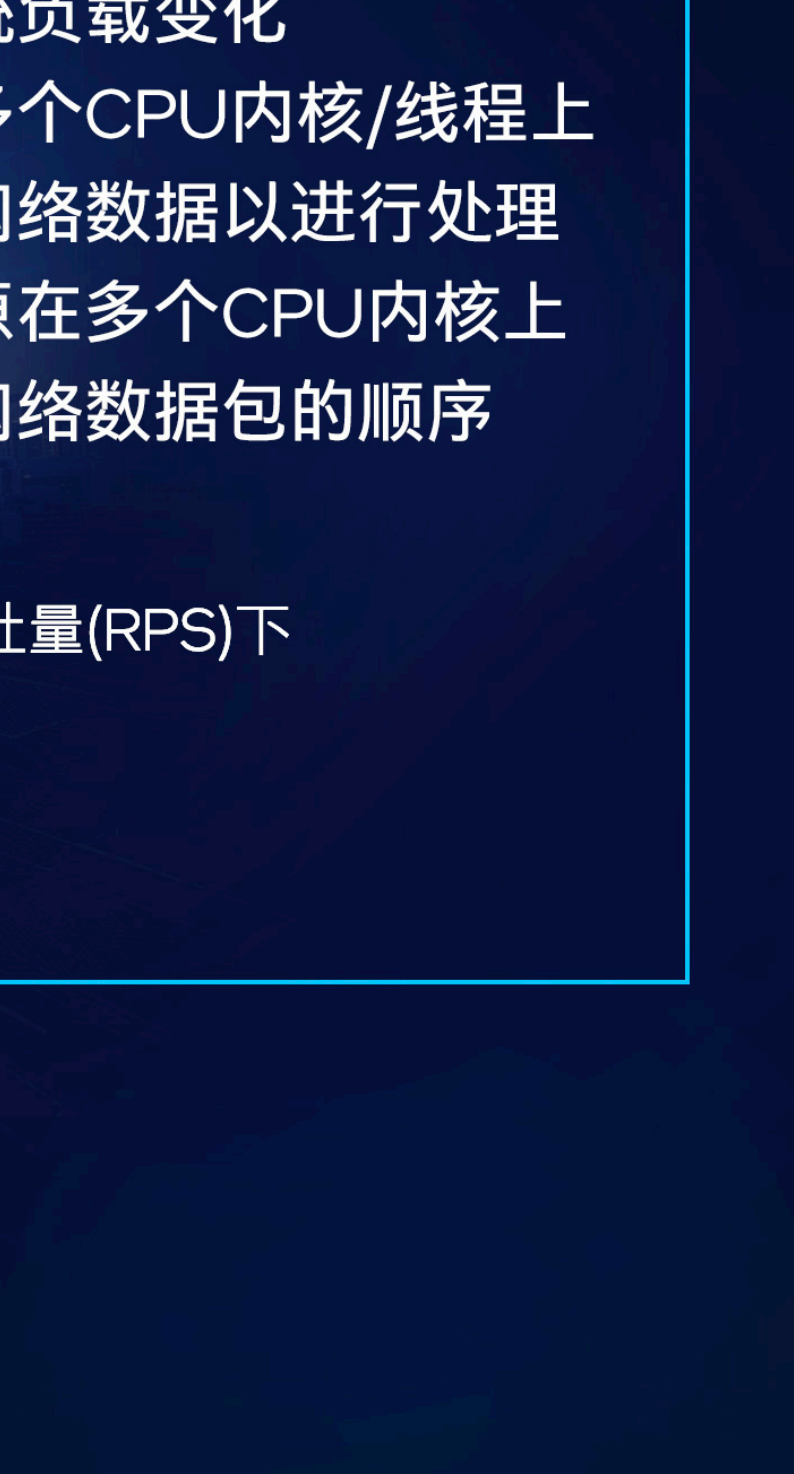
在相同的吞吐量(RPS)下时延可降低 96%[3]

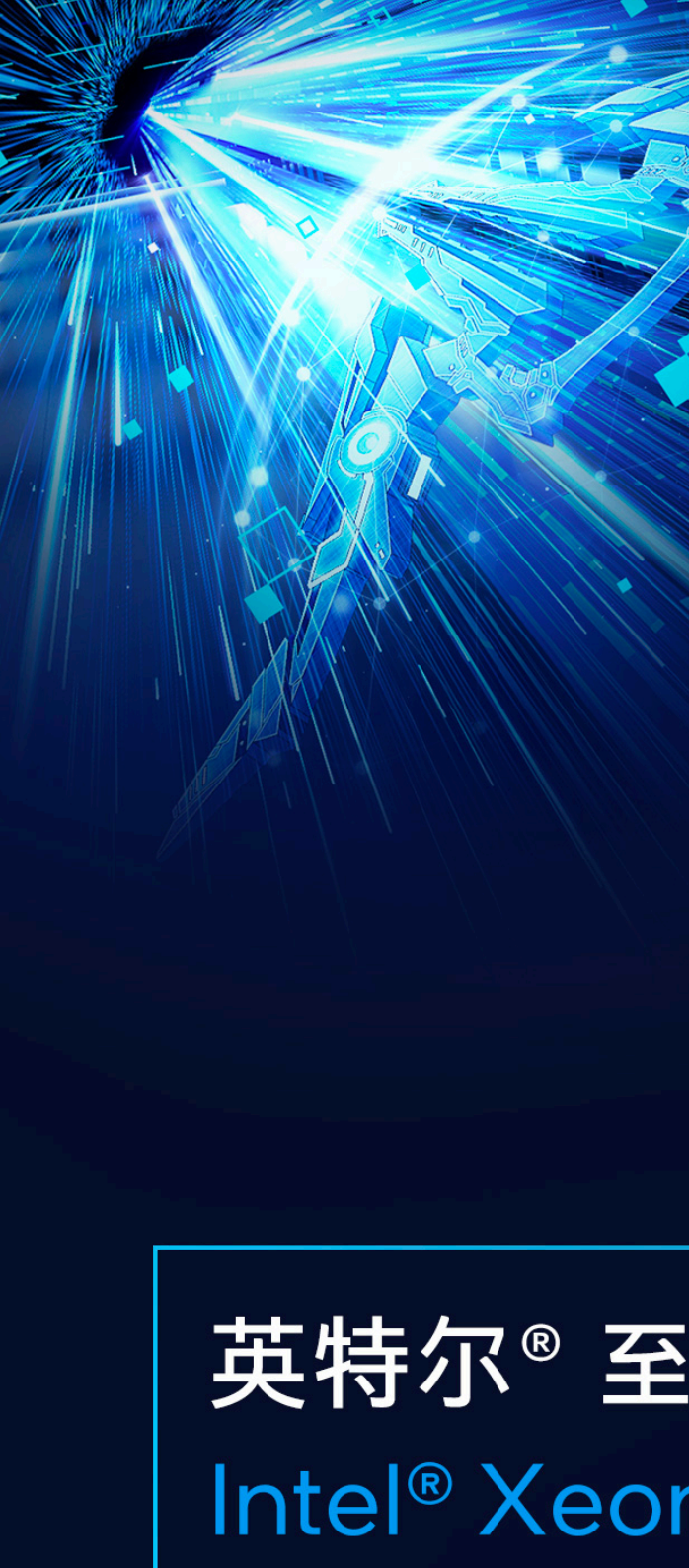
英特尔® 数据流加速器
Intel® DSA

专为数据移动任务所设计
通过优化数据移动和转移操作
加速CPU、内存和缓存
以及所有附加的内存、存储、网络设备之间的数据移动

大型数据包顺序读取的延迟降低 37%[4]

IOPS提升高达 1.7倍[4]





英特尔® 存内分析加速器
Intel® IAA

可提高数据查询吞吐量
同时减少内存占用
高效处理内存数据库和大数据分析工作负载

加速运行实时分析
可将RocksDB性能提升 3倍[5]

英特尔® 至强® CPU Max系列
Intel® Xeon® CPU Max Series

集成高带宽内存（HBM）的x86处理器
能够在无需变更代码的情况下
为科学计算和AI工作负载加速

英特尔® 至强® CPU Max系列
可为实际应用性能提升 3.7倍[6]

提高工作负载性能，实现可持续发展

通用计算*

intel xeon

53%
平均性能提升[7]

人工智能*

多达10倍
实时推理和训练性能提升[1]

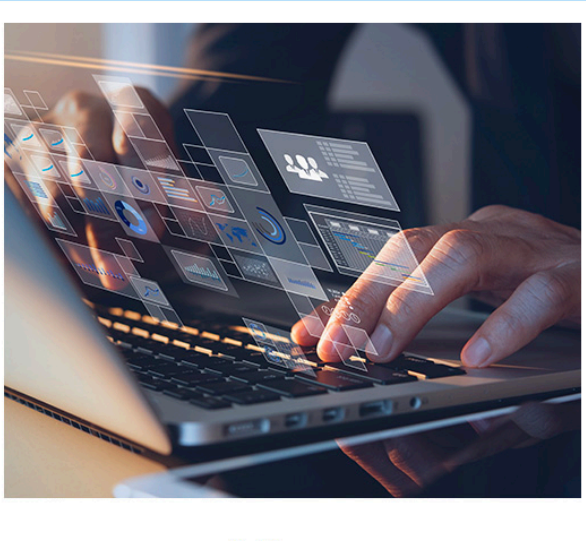
网络5G vRAN*

多达2倍
vRAN工作负载容量提升（在相同功率包络下）[11]

网络&存储*

多达2倍
数据压缩吞吐量（核心减少95%）[2]

数据分析*



多达3倍
性能提升[5]

科学计算**

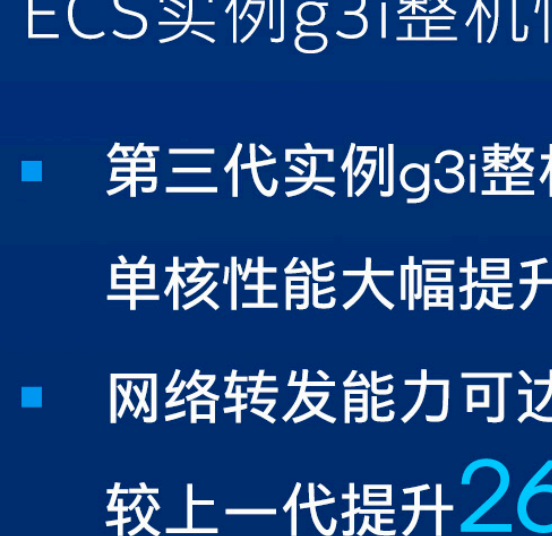
多达3.7倍
工作负载优化[6]

平均每瓦性能提升了2.9倍[8]
在优化电源模式下每个CPU节能高达70瓦[9]
总体拥有成本降低52%到66%[10]

携手行业标杆，加速产业发展与创新

阿里云第八代ECS实例：整机性能大幅提升

基于第四代英特尔® 至强® 可扩展处理器，阿里云第八代实例的算力得到大幅提升。

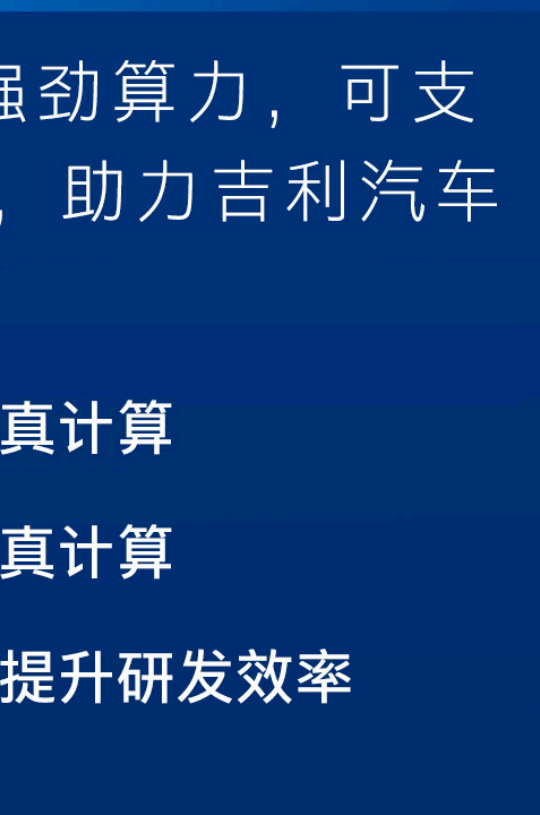


- 数据库的读写性能提升15%以上
- 加解密、压缩解压缩性能提升4倍以上
- AI推理的能力提升4倍以上
- 训练场景性实现2倍性能提升
- 整体算力最大提升60%

火山引擎第三代ECS实例g3i搭载第四代英特尔® 至强® 可扩展处理器，整机性能最高提升70%

基于第四代英特尔® 至强® 可扩展处理器，火山引擎第三代ECS实例g3i整机性能提升。

- 第三代实例g3i整机算力提升70%以上，单核性能大幅提升
- 网络转发能力可达3600万pps，较上一代提升260%
- 更加适合游戏、数据分析、中小型数据库、微服务、网络密集型业务



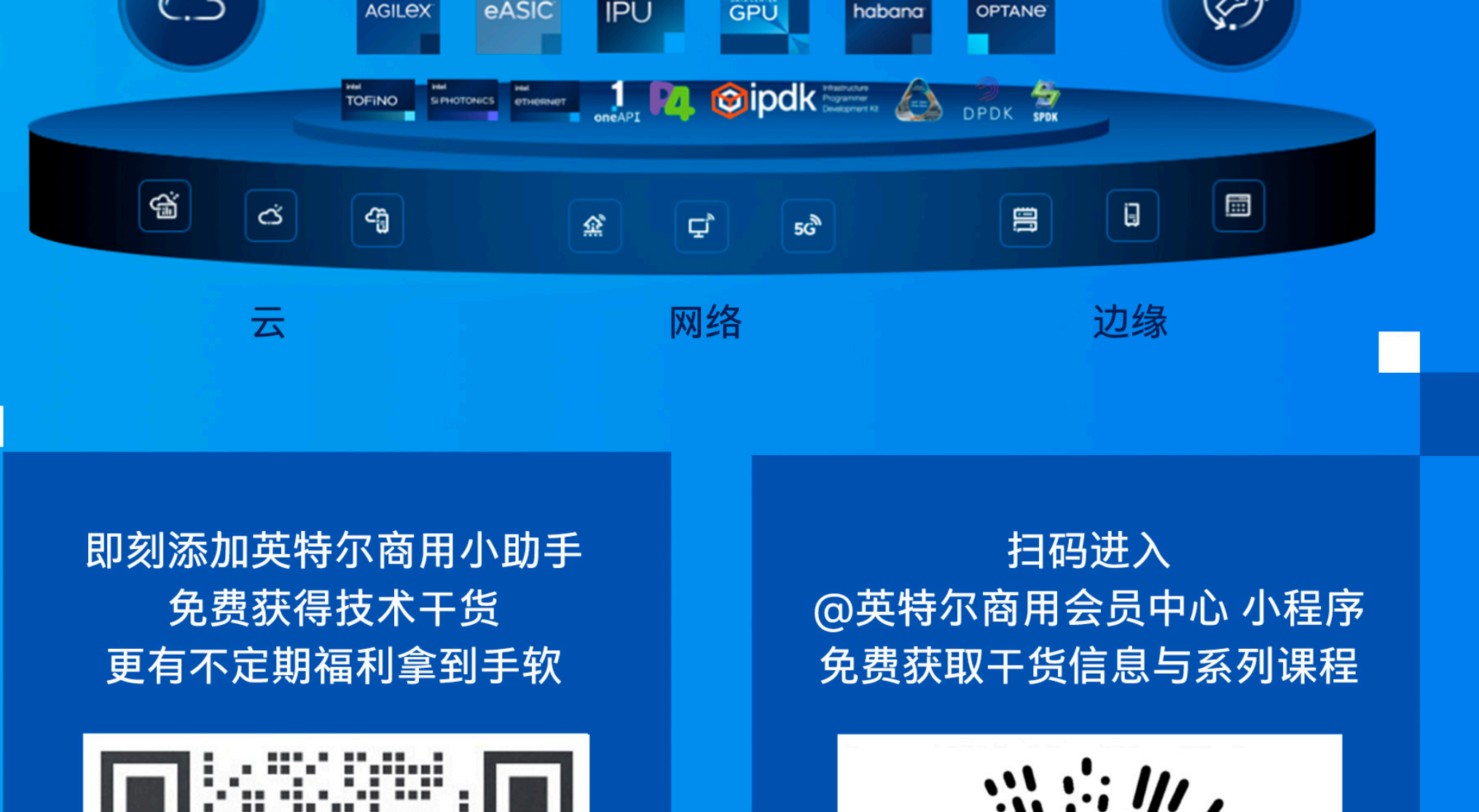
吉利汽车仿真业务：赋能研发数字化转型

第四代英特尔® 至强® 可扩展处理器的强劲算力，可支持仿真业务的高效运行，推进研发测试，助力吉利汽车数字研发。

- 超低风阻开发仿真计算
- 安全性能开发仿真计算
- 助力吉利仿真，提升研发效率

布局多元算力，加速千行百业云数智转型

面对日益多元的数字化创新需求
英特尔提出XPU产品战略
打造从云到端的全面产品组合
提供从CPU到GPU、FPGA、IPU等多种异构算力资源
为不同业务和应用场景
提供定制化算力服务



即刻添加英特尔商用小助手
免费获得技术干货
更有不定期福利拿到手软

扫码进入
@英特尔商用会员中心 小程序
免费获取干货信息与系列课程

---

*第四代英特尔® 至强® 可扩展处理器vs第三代英特尔® 至强® 可扩展处理器
**英特尔® 至强® CPU Max系列vs英特尔® 至强® 8380

[1]请访问 intel.com/processorclaims，查看4th Gen Intel® Xeon® Scalable processors中的[A17]，结果可能会有调整。

[2]请访问 intel.com/processorclaims，查看4th Gen Intel® Xeon® Scalable processors中的[D1]，结果可能会有调整。

[3]请访问 intel.com/processorclaims，查看4th Gen Intel® Xeon® Scalable processors中的[W6]，结果可能会有调整。

[4]请访问 intel.com/processorclaims，查看4th Gen Intel® Xeon® Scalable processors中的[N18]，结果可能会有调整。

[5]请访问 intel.com/processorclaims，查看4th Gen Intel® Xeon® Scalable processors中的[N16]，结果可能会有调整。

[6]英特尔® 至强® 8380：截至英特尔2022年10月7日的测试。1-节点，2个英特尔® 至强® 可扩展处理器8380，启用超线程，启用睿频，总内存 256 GB (16x16GB 3200MT/s DDR4)，BIOS版本 SE5C620.86B.01.01.0006.2207150335，ucode版本=0xd000375，Rocky Linux 8.6，Linux版本4.18.0-372.26.1.el8_6.crt1.x86_64，YASK v3.05.07

英特尔®至强®CPU Max系列：截至英特尔2022年第36周的测试。1-节点，2x 英特尔®至强®CPU Max系列，启用超线程，启用睿频，SNC4，总内存128 GB (8x16GB HBM2 3200MT/s)，BIOS版本 SE5C7411.86B.8424.D03.2208100444，ucode版本=0x2c000020，CentOS Stream 8，Linux 版本 5.19.0-rc6.0712.intel_next.1.x86_64+server，YASK v3.05.07。

[7] HP Linpack、Stream Triad、SPECrate2017_fp_base est、SPECrate2017_int_base est的几何平均值。[G2, G4, G6] 参见intel.com/processorclaims：第四代英特尔至强可扩展处理器。

[8] 如下工作负载的几何平均值：RocksDB（IAA vs ZTD）、ClickHouse（IAA vs ZTD）、SPDK大型媒介与数据请求代理（DSA vs 开箱即用）、图像分类ResNet-50（AMX vs VNNI）、物体识别SSD-ResNet-34（AMX vs VNNI）、QATzip（QAT vs zlib）

[9] 1-节点，英特尔参考验证平台，2个英特尔® 至强8480+ (56C, 2GHz, 350W TDP)，启用超线程，启用睿频，总内存：1 TB（16个插槽/ 64GB/ 4800 MHz），1个P4510 3.84TB NVMe PCIe Gen4驱动器，BIOS：0091.D05，(ucode:0x2b0000c0)，CentOS Stream 8, 5.15.0-spr.bkc.pc.10.4.11.x86_64，Java Perf/Watt w/ openjdk-11+28_linux-x64_bin，112个实例，1550MB 初始堆/最大堆大小，英特尔截至2022年1月的测试。

[10] ResNet50图像分类

新配置：1-节点，2个预生产的第四代英特尔®至强®可扩展处理器8490H（60核），含英特尔® 高级矩阵扩展（英特尔AMX），预生产超微SYS-221H-TNR，1024GB DDR5内存（16x64 GB），微代码0x2b0000c0，启用超线程，启用睿频，关闭SNC，CentOS Stream 8, 5.19.16-301.fc37.x86_64，1x3.84TB P5510 NVMe，10GbE x540-AT2，英特尔TF 2.10，AI Model=Resnet 50 v1_5，最高得分：BS1 AMX 1 核/实例（最高15ms SLA），使用物理内核，英特尔在2022年11月进行的测试。基准：1-节点，2个第三代英特尔®至强®可扩展处理器8380（40核），超微SYS-220U-TNR，1024GB DDR4内存（16x64 GB），微代码0xd000375，启用超线程，启用睿频，关闭SNC，CentOS Stream 8, 5.19.16-301.fc37.x86_64，1x3.84TB P5510 NVMe，10GbE x540-AT2，英特尔 TF 2.10，AI Model=Resnet 50 v1_5，最高得分：BS1 INT8 2 核/实例（最高15ms SLA），使用物理内核，英特尔在2022年11月进行的测试。

包含50台配备第三代英特尔至强处理器8380的服务器编队（RN50 w/DLBoost），截至2022年11月的估计：
资本成本：164万美元
运营成本（4年，包括电力和散热成本，基础设施和硬件维护成本）：73.99万美元
能耗kWh（4年，每台服务器）：44627，PUE 1.6
其它假设：电价$0.1/kWh，kWh对应千克CO2 系数为0.42394

包含17台配备第四代英特尔至强处理器8490H 的服务器编队（RN50 w/AMX），截至2022年11月的估计：
资本成本：79.94万美元
运营成本（4年，包括电力和散热成本，基础设施和硬件维护成本）：27.53万美元
能耗kWh（4年，每台服务器）：58581，PUE 1.6

人工智能——通过部署更少的基于第四代英特尔至强处理器的服务器，在实现相同性能要求的情况下，TCO降低达55%。请参阅 intel.com/processorclaims 上的 [E7]：第四代英特尔至强可扩展处理器。结果可能会有所不同。

数据库——通过部署更少的基于第四代英特尔至强处理器的服务器，在实现相同性能要求的情况下，TCO降低达52%。请参阅 intel.com/processorclaims 上的 [E8]：第四代英特尔至强可扩展处理器。结果可能会有所不同。

HPC——通过部署更少的基于英特尔至强CPU Max系列处理器的服务器，在实现相同的性能要求的情况下，TCO降低达66%。请参阅 intel.com/processorclaims 上的 [E9]：第四代英特尔至强可扩展处理器。结果可能会有所不同。

[11] 截至2022年8月30日的估计，使用FlexRAN™软件测试，对比了第四代英特尔® 至强® 可扩展处理器和第三代英特尔® 至强® 可扩展处理器的架构提升，拥有类似的内核数量、插槽功率和频率。结果可能不同。