

以计算加速迈进智能化未来

——IDC新一代云基础设施实践报告

目录

01

趋势：云服务能力持续跃升，加速企业数智化转型与创新

- 1.1 技术全面升级，为复杂的企业在线业务提供保障
- 1.2 软硬一体协同优化，应对AI时代激增的数据冲击
- 1.3 持续的融合创新，助力企业的国际化布局

02

挑战：企业多元业务需求与海量AI数据的冲击

- 2.1 在线业务面临性能与效率的极限挑战
- 2.2 AI数据处理与计算协同的复杂度激增
- 2.3 国际化进程中的全球布局、合规与质量一致性难题
- 2.4 安全、稳定与成本的多元保障要求

03

解决方案

- 3.1 打造极致性能体验，为传统计算业务打开新空间
- 3.2 技术和架构创新，提升AI时代的向量数据处理和协同计算效率
- 3.3 强化硬件安全设计，持续增强安全保障能力
- 3.4 全球一致的云服务能力体系，全面助力企业国际化战略

04

优秀实践分析

- 4.1 小鹏汽车
- 4.2 微帧科技
- 4.3 嘎嘎射击
- 4.4 蚂蚁集团ZOLOZ

前言

IDC分析师认为：全球AI基础设施革新的浪潮中，算力需求的爆发正在驱动云计算与边缘计算深度融合，行业定制化与智能化服务加速渗透，成本优化与绿色计算将成为竞争的关键。未来，基础设施的核心矛盾将从“资源供给”转向“效率与价值平衡”，技术迭代将围绕“弹性算力调度”“数据主权治理”“垂直场景深度适配”三大主线展开。

越来越多的企业核心数据正在向云数据中心迁移，计算密集型任务处理能力与弹性资源供给能力正成为云服务商的核心竞争力。面对企业客户的数字化转型需求，减少延迟和工作负载可移植性将是客户的关键优先事项，为垂直特定数据类型提供量身定制的云服务将创造有利的竞争优势。云提供商须为跨行业数据采集、存储和计算需求的大幅增长做好准备。

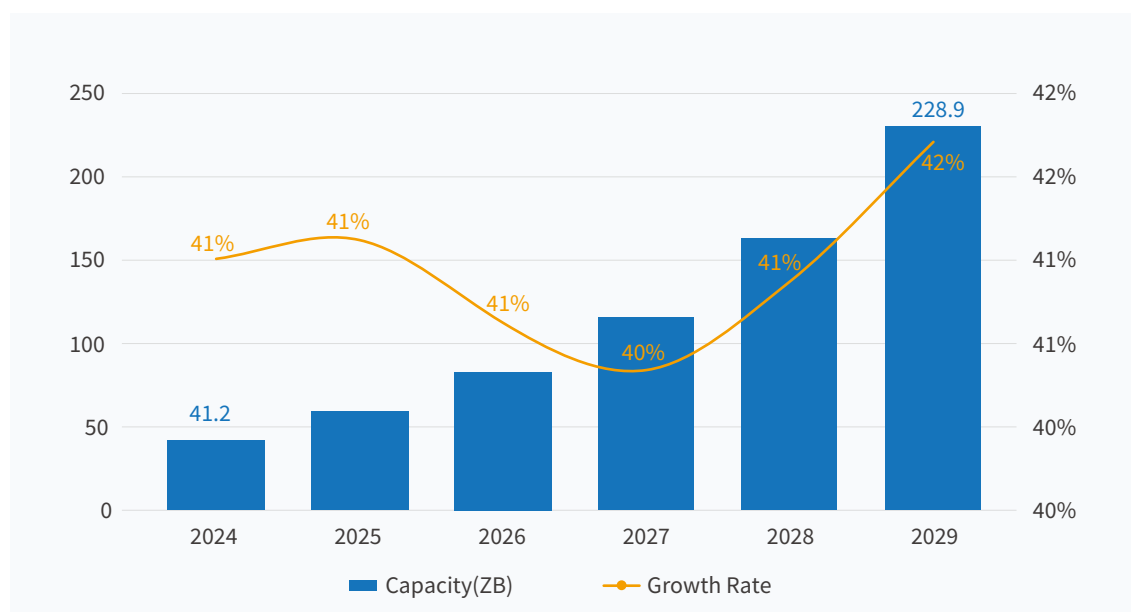
在AI高速发展和在线业务快速膨胀的时代，企业用户对云基础设施的性能、成本、稳定性、安全性等方面提出了全新的要求。为适应企业创新、降本增效以及业务出海等需要，云服务商不断通过协同创新升级全栈服务品质，同时也利用自身融合发展的经验优势，助力企业积极开展国际化布局。



01 趋势：云服务能力持续跃升 加速企业数智化转型与创新

IDC预计，云数据中心数据增长在2025年为58.1ZB，2029年将翻4倍，达到228.9ZB，2024-2029年复合年增长率为40.9%。

图1 全球云数据中心数据增长，2024-2029

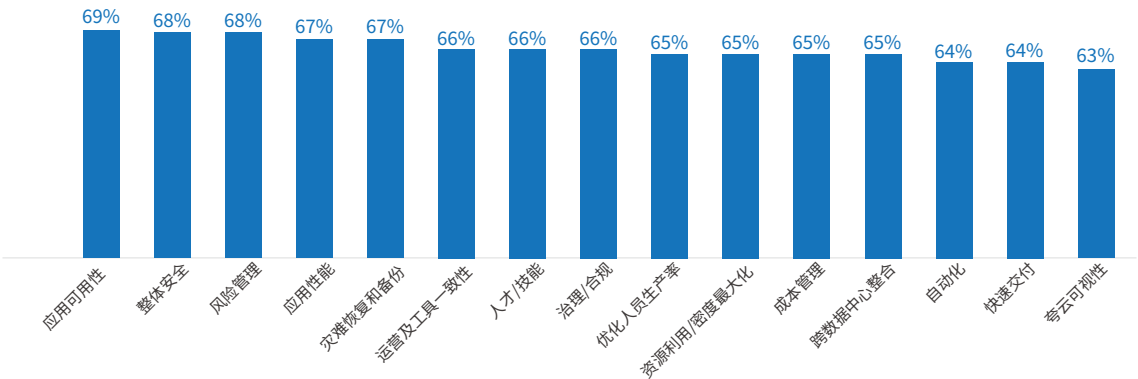


来源：IDC全球数据圈预测, 2025-2029

1.1 技术全面升级，为复杂的企业在线业务提供保障

企业在线业务的受众范围和功能复杂度在快速增加，在金融交易、电商直播、实时游戏等场景下，服务端动辄需要支持百万级并发连接和毫秒级响应要求，应对海量的网络协议处理、页面加载、安全等事务。企业云计算客户不仅对算力密度有极致追求，还期望通过连接性能和存储技术等多个方面的协同进步，实现数据库、大数据等服务平台的性能跃升。在IDC面向全球1350家企业所做的数字化进程与业务成果调研中，应用的可用性、综合安全性、应用的性能等都成为企业核心关注的目标。

图2 全球企业认为未来2年对业务成果最重要的IT事项



n = 1350

来源: Cloud Pulse Survey, IDC

- **对算力密度的极致追求：**企业希望利用有限的物理空间输出更强大的算力。这一方面体现在一些高端的云服务实例可以提供数百、数千甚至数万数量级的CPU、GPU核服务能力；另一方面，为满足大数据、数据库、3D视频处理在内的一些单核敏感型业务的需要，云服务仍将持续提升单核、单实例性能。
- **多技术融合提升连接性能：**云服务商综合利用内存/缓存、PCIe、RDMA、IP网、EIP、VPC等一系列技术升级和软硬件融合优化成果，大幅提升云、边、端不同位置服务之间的协同效率，这对于保障在线业务的体验至关重要。
- **存储方案升级应对大数据量冲击：**云服务商通过采用更高性能的存储设备和更高效的存储架构，结合对数据布局的优化，提供贴合不同在线业务需求的个性化存储服务，例如低时延块存储（数据库多副本场景，<0.1ms延迟）、高带宽弹性盘（大数据单副本场景，吞吐量达40Gbps）、高速临时存储等。

1.2 软硬一体协同优化，应对AI时代激增的数据冲击

AI预训练和推理过程需要存储和预处理海量的多模态数据，数据向量化趋势也非常显著，为保障AI应用特别是中小模型推理和传统AI搜推场景的实时响应，云基础设施也在架构层面做出了持续的优化创新。

- **分布式预处理框架和流程优化：**通过构建分布式的训练数据预处理框架，将IVF、HNSWLib、Faiss/Flat等向量化算法进行分布式改造，使其能够在多个计算节点上并行运行。通过数据分片和任务分配，充分利用集群的计算资源，提高处理大规模向量数据的能力。这需要云计算基础设施提供灵活的适配和资源编排能力，以及严格的资源隔离与共享保障。
- **利用硬件加速提升数据预处理和AI推理效率：**在处理器内部增加专用硬件加速单元和专用指令集，提升数据清洗、加密与传输效率。对于视觉模型、视频处理、数据库模糊检查查询等任务，云实例基于新一代处理器的向量指令集、矩阵加速指令集等，能够直接支持相关AI算子执行，简化系统架构，提升响应性能与可靠性，持续改善搜推广、语音/视频处理以及中小模型的推理体验。
- **云原生方案形成整体保障：**除了算力层面的降本措施，云服务商还通过持续增强弹性伸缩、Serverless以及统一运维等云原生解决方案的效能，实现硬件、算法、平台、服务的协同进化。例如，在强化弹性伸缩能力时，综合运用编排调度（例如Kubernetes）、服务监控、服务治理以及相关的配置管理等云原生能力，在提高弹性伸缩的速率的同时，持续增强自动化预测能力以及优化自适应策略。

1.3 持续的融合创新，助力企业的国际化布局

面对智能化、全球化、融合化的发展浪潮，国内云服务商凭借自身在互联网创新、跨境电商以及AI创新应用等领域的创新优势，不仅可以为自身拓展新的广阔发展空间，也为全球市场带来新的发展机遇与变革动力。

- **互联网创新推动全球市场繁荣：**国内云服务商在互联网创新方面的成果优势，一方面体现在业务的示范效应，例如在电商、社交媒体、在线教育、远程办公等互联网服务领域，国内云服务商具有丰富的融合开发和运营经验，可以通过国际合作与交流，输出到全球市场。另一方面也体现在丰富的技术融合经验，即基于自身的互联网创新业务推动云计算服务能力的发展和进步，例如综合运用云网络、分布式云架构、负载均衡、CDN等技术，保障全球用户获得流畅的访问体验。

- **跨境电商形成市场综合服务优势：**一些国内大型云服务商具备丰富的跨境电商业务运营经验，可以整合跨境物流、支付等环节的资源，为跨境电商企业提供一站式的解决方案。在信息服务方面，具备强大的数据分析工具，能够帮助跨境电商企业精准洞察不同地区消费者的需求和偏好，优化选品和营销策略，有效保障交易数据的安全合规要求。
- **AI创新应用不断挖掘新的市场潜力：**近年来，国内云服务商在AI创新应用方面不断突破，为挖掘全球市场新潜力提供了强大动力。相关的基础AI模型和AI技术栈都十分强大，特别是能够持续利用云计算的海量数据存储和强大计算能力优势，支撑智能化的云应用出海。一大批AI产品已经在市场中发挥了巨大作用，例如B端市场的国内企业生产、供应链管理、风险控制，以及C端市场的智能搜推、智能客服等。



02 挑战：企业多元业务需求与海量AI数据的冲击

2.1 在线业务面临性能与效率的极限挑战

面对企业海量在线业务以及复杂的创新场景，现有的云基础设施常常难以应对用户在弹性、低时延和高吞吐等方面的基础性能问题。

- **复杂应用的复合性能挑战：**以游戏场景为例，其涉及复杂的图形渲染，以及物理环境模拟、人工智能算法等，既需要高性能单核算力，支撑Unity和Unreal Engine等3D引擎的运行，也需要可靠的多线程并发能力，支持多玩家同步时的后台任务处理和AI推理。游戏业务的周期特性对于资源的弹性伸缩能力要求极高。此外，玩家数据的记录也涉及频繁的写操作，需要保持长连接、低时延的计算、存储服务。当前的游戏业务还广泛使用到数智驱动，利用AI加速对玩家行为进行实时分析并提供个性化推荐。
- **存算分离架构带来网络、存储性能压力：**在大数据、数据库场景中，存算分离架构使计算节点和存储节点之间的数据传输量大幅增加。大数据平台运行时，CPU负载普遍在60%以上，内存带宽利用率可高达80%以上。高负载情况下，保持算力的平稳输出非常重要，包括性能的线性度指标以及资源隔离的力度等。特别在是高密度计算环境下，单服务器下的租户更多，相关租户之间的故障隔离以及缓存、存储等资源的共享访问问题会更加突出，既要确保租户之间运行不受任何影响，又不能以牺牲用户访问性能为代价。此外，多副本存储策略在面对存算分离架构时也会带来数据一致性的挑战。一些分布式事务处理涉及多个计算节点和存储节点，进一步增加了系统设计和实现的复杂度。
- **云原生架构带来系统架构和技术栈的新问题：**例如如何处理微服务间频繁的通信、复杂的拓扑、多样化存储以及动态多变的负载等，并强化隔离保障安全。云服务商通过持续增强弹性伸缩、Serverless以及统一运维等云原生解决方案的效能，帮助用户降低成本。例如，在强化弹性伸缩能力时，综合运用编排调度（例如Kubernetes）、服务监控、服务治理以及相关的配置管理等云原生能力。

2.2 AI数据处理与计算协同的复杂度激增

传统的云存储和处理架构难以高效应对，存储和传输成本也非常高昂。同时，AI算力需求持续高速增长，为保障AI系统的实时响应，云基础设施需要在架构层面做出创新，以应对多种类型的挑战。

- **数据处理挑战：**除AI海量、多模态特征外，工业制造、医疗等领域的AI应用所涉及的异构数据多，格式和标准不统一，进一步导致管理和存储成本的上升。在海量AI数据预处理过程中，非结构化数据清洗、数据标注等工作的效率往往不高，也严重影响了AI应用目标的达成。
- **协同计算挑战：**规模化的AI训练和推理任务，往往采用分布式架构执行，参数同步时的通信需求巨大，传统云网络在高并发实时通信场景下难以满足 PB 级数据传输的低延迟需求。同时，很多AI任务混合使用CPU、GPU、TPU等算力，但一些云平台对大量的异构算力缺乏统一管理 and 调度框架，算力有效利用率长期难以提升。
- **体验与效率挑战：**大量的AI模型服务以API形式向外输出，当云架构设计不合理时，用户端的高并发请求极易导致服务崩溃。值得注意的是，AI任务的多样化意味着并非所有AI任务都适合采用GPU集群方式进行处理。例如一些实时推荐、召回分析、游戏状态管理等场景，需要在数据库的域内完成模糊查找、匹配等操作，以满足实时性和数据不出域的要求。

2.3 国际化进程中的全球布局、合规与质量一致性难题

全球化业务对企业云计算的规模、弹性以及各项性能指标的要求更高，技术实现相对复杂且成本高昂。各国家、地区的政策法规、基础设施差异大，同时还要尽量保障全球服务质量和体验的一致性，这无疑增加了企业全球化发展的难度。

- **云服务与算力网点布局问题：**在本土之外构建云基础设施，可能涉及土地、供电、人力等多项工作，组织协调难度极大。不同国家基础设施发展水平不一，无论是中企出海还是外企入华，都会因不熟悉当地环境而困难重重，难以获得理想的云服务与算力支持。此外，合规风险问题不容忽视，各国对数据存储、跨境传输等有严格规定，稍有不慎将面临巨额罚款和声誉损失。
- **规模、性能与弹性需求：**不同市场的业务规模差异大，在海外拓展初期，业务波动往往较大，各地区的增长难以预测，这给云计算的技术需求规划也带来了难题。企业既要避免初期大规模投入造成资源浪费，又要确保业务高峰时有足够云计算资源支撑。部分地区的云服务性能难以保证，数据的传输延迟和丢包率高，严重影响用户端的响应速度和综合体验。

- **服务质量与体验一致性：**很多大型企业，对全球化业务的服务质量和体验一致性要求极高，确保用户身处任何国家都期望获得相同体验、相同水准的服务。但不同国家文化背景、使用习惯和技术水平差异性，都会严重对上述目标产生严重影响。这需要企业投入资源，兼顾全球标准的建立以及本地化适配问题。

2.4 安全、稳定与成本的多元保障要求

云计算的安全性与稳定性直接影响用户信任度，任何数据泄露或业务中断均可能导致灾难性后果。

- **数据安全性——信任与合规的双重考验：**云计算的多租户架构和分布式存储特性增加了数据被跨域非法访问的风险，在金融、医疗、零售等场景中，数据泄露可能引发灾难性后果。此外，跨境数据传输需满足各国严格的法规要求，稍有不慎便可能面临巨大的风险。
- **应用稳定性——AI与线上业务的双重冲击：**AI应用的普及加剧了云计算的不确定性。模型训练需要应对海量数据冲击，推理服务对延迟非常敏感。同时，大量的线上业务依赖7×24小时高可用性，业务中断或数据丢失都可能引发巨额经济损失。
- **运维复杂度——人力与资源的双重负担：**云计算的分布式架构和动态资源调度需求增加了运维难度。大型企业可能使用跨区域、跨云环境下数以千计的实例，处理自动化扩容、故障转移等复杂任务。企业需投入大量的云原生开发和运维人才，采购昂贵的监控与自动化工具等。
- **成本控制——算力性价比难题：**企业多云集群和异构计算资源的效率和适配不足，使算力成本长期居高不下，弹性能力的缺失造成的大量云实例闲置，也加剧了浪费现象。在企业加速业务创新的背景下，大型企业的多业务线体系需要频繁地应付新业务上线部署时千奇百怪的需求，进一步产生了额外的高额开支和大量的重复建设。

03 解决方案

阿里云第九代企业级ECS实例（ECS g9i）基于“CIPU+飞天”技术架构，搭载英特尔®至强® 6性能核处理器（P-core），为包括数据库、大数据、游戏在内的多类场景带来显著的性能和体验提升。

CIPU作为新一代云计算管控和加速中心，向下对数据中心的计算、存储、网络资源快速云化并进行硬件加速，向上接入飞天云操作系统实现弹性多租户的安全隔离、性能优化和成本控制，全方位助力企业应对在线业务的高并发、低延迟、高可用挑战。

作为最新一代的至强® 产品，英特尔®至强® 6 性能核处理器经过精心优化，具备更出色的单核性能，更适合公有云工作负载，能够为浮点运算、事务型数据库和科学计算等工作负载提供更高的单个 vCPU 性能。在核心数量、算力密度、内存与 I/O 升级和AI 深度优化，以及多场景性能上，至强® 6性能核处理器均有全面突破。

图3. 英特尔®至强®代际演进

	代际规格		
	第二代英特尔®至强®可扩展处理器	第四代英特尔®至强®可扩展处理器	英特尔®至强®6900 性能核处理器
最高单路内核数	28 核	60 核	128 核
内存	6通道 高达2933MT/s(DDR4)	8通道 高达4800MT/s(DDR5)	12通道 高达6400MT/s(DDR5)
PCIe	PCIe 3.0, 48条通道	PCIe 5.0, 80条通道	支持PCIe 5.0, 多达96条通道
AI加速器	AVX-512(VNNI)	AMX/TMUL(INT8& BFloat 16)&AVX-512 (VNNI/INT8)	AMX/TMUL(INT8& BFloat 16,FP16)&AVX-512 (VNNI/INT8)

来源: Intel

- 至强®6 性能核处理器专为计算密集型工作负载而设计，计算单元基于英特尔 3 制程工艺，最高配备 128 个性能核，当前阿里云 ECS g9i 采用高性能 120 核定制款。
- 支持多达 12 个通道 6400MT/s 的 DDR5，拥有高达 504MB 的三级共享缓存，96 条 PCIe 5.0 及 64 条 CXL 2.0 通道。
- 新增对 FP16 数据格式的支持，可为 AI 推理和机器学习 (ML) 等计算密集型工作负载提供优于其他通用 CPU 的出色性能。同时，凭借其出色的 AI 推理性能，英特尔®至强®6 性能核处理器也是全球众多出色 AI 加速平台用于支持数据处理的机头 CPU 优选之一。

除了以更多内核为 ECS g9i 提供澎湃算力外，英特尔至强®6 性能核处理器还能助力主流工作负载在 ECS g9i 上实现显著增益。与基于第五代英特尔®至强®处理器的 ECS g8i 相比，ECS g9i 的在线游戏性能提升 15%，数据库性能提升 17%，Web 应用性能提升 20%。

3.1 打造极致性能体验，为传统计算业务打开新空间

1、在线业务

至强®6 性能核处理器采用 Chiplet 架构，处理器拥有 3 个计算芯粒，2 个 IO 单元芯粒，所有芯粒之间通过 EMIB 多芯片互连桥接封装技术 (Embedded Multi-die Interconnect Bridge) 技术进行高速连接。单个计算芯粒拥有最多 43 个内核，可以在同一计算芯粒内高效地处理在线业务，对于大规模在线业务 (如 web 服务、缓存层、云原生微服务) 可以在一个 Die 内分布完成，减少跨 Die 通信延迟，更适合低延迟、大吞吐云业务场景。EMIB 高速封装桥接技术使计算单元之间通过高速桥互联，同一个计算芯粒内无“非一致性内存访问” (NUMA) 问题。在 SNC (Sub NUMA Cluster) 模式下，关键数据驻留在同一个芯粒的本地内存 / 缓存中，避免跨芯粒访问的内存延迟开销，对高并发在线负载和数据库事务处理更有优势。

(1) 典型 Web 业务

为应对 Web 业务高并发、低延迟、动态流量波动大的特性，ECS g9i 提供了强大的单核计算性能和弹性伸缩能力，从容应对突发流量。完备的 QoS 保障措施，能够更好地实现多租户环境下的性能稳定与资源公平分配。通过在硬件架构上的一系列创新，至强®6 性能核处理器拥有强大的硬件加速能力，可为计算、网络资源带来显著提升。

- **弹性伸缩：**ECS g9i 对弹性伸缩能力进行了大幅优化。单实例创建时间从 10s 缩减至 4s，单规格单可用区的 1 万实例交付时间从 5 min 降低到 1min，从触发弹性开始，30s 即可实现节点就绪。基于至强® 6 性能核处理器所提供的英特尔® Speed Select 技术 (英特尔® SST)，ECS g9i 允许用户调整 CPU 核心数量与时钟频率，以满足特定计算需求，例如可以从 120 核改配到 96 核，以提高单核性能。在线主频变配 + 芯粒化架构缓解资源紧张问题，一台服务器可实现灵活多档配置，无需重新调度到其他机器或可用区，用户无需担心可用区内新实例资源不足的问题。对于运行 7x24 的云原生业务和关键服务系统，可持续提供 SLA，业务连续性更强，无需“停机换车”。
- **QoS 保障：**ECS g9i 通过硬件虚拟化、CPU 的资源调度管理等多种优化手段，实现了对存储、网络、内存及计算资源的精细化隔离与保障，确保了项目租户间的资源隔离，互不干扰。首先通过英特尔® 资源调配技术 (英特尔® RDT, Intel Resource Direct Technology) 的能力，提供了包含丰富组件的框架，可以跟踪和控制平台上同时运行的多个应用程序、容器或 VM 使用的缓存和内存带宽，确保复杂环境中关键业务负载的性能。英特尔还提供了基于硬件的动态内存带宽控制技术 (HWDRC)，相对于静态的内存带宽分配，HWDRC 进一步提高了内存带宽利用率。英特尔® 数据直接 I/O 技术 (英特尔® Data Direct I/O) 是英特尔® 集成 I/O 的关键功能，能够使处理器缓存成为 I/O 数据的主要目的地和来源，而无需像传统 DMA 一样通过主内存中转，进一步实现更高的带宽、更低的延迟和更低的功耗。另外，在系统设计层面，基于 g9i 的底层硬件，也实现支持处理器双单路硬件架构通过独立供电与运行单元设计，在单路故障时仍可维持另一单路稳定运行，实现功耗隔离与故障隔离，保障系统持续稳定运行。
- **QAT 硬件加速：**英特尔® 数据保护与压缩加速技术 (英特尔® QAT) 专用加速引擎提供了带外的独立于 CPU 核心的额外压缩 / 解压算力和加解密算力，有效卸载 CPU 负载。QAT 支持丰富的压缩算法，既支持传统的 deflate 系列压缩算法，包括 zlib, gzip 等格式，也支持压缩速度极快的 LZ4，以及速度和压缩率比较均衡的 ZSTD 等新型的压缩算法，能满足客户大多数的算法需求和多样化的业务场景；QAT 支持主流的对称加解密和非对称加解密算法，可广泛用于加速 TLS 协议的握手阶段以及后续数据块的加解密，提升 Nginx 服务以及其他使用 HTTPS/TLS 协议的服务性能，适用于公司网关、加解密专用节点等场景；QAT 还采用了 plugin 的方式，灵活适配 OpenSSL，让众多使用 OpenSSL 的客户能无缝切换到使用 QAT，轻松启用硬件加速能力。
- **eRDMA 网络技术：**通过弹性远程直接内存访问，端到端网络延迟低至 8 微秒，较传统 VPC 降低三分之二，有效支持大规模组网和高弹性需求。依托于 CIPU 架构，ECS g9i 实现了网络资源的灵活调配优化，满足各种高并发、低延迟的网络需求。

(2) 数据库业务

当前的在线数据库业务具有数据量大、计算量大、数据随机访问多、多核并发等典型特征，从不同类型数据库的性能需求看，MySQL 高并发场景依赖多核与高主频平衡；Redis 对单核性能（CPU 主频）、内存带宽、时延都要求极高；PostgreSQL 的复杂查询依赖多核并行执行与高缓存命中率，ClickHouse 分析型数据库则更加需要向量计算加速。

- **优化与性能提升：**ECS g9i 针对关系型数据库、Redis、向量数据库以及压缩 / 解压缩等数据处理与分析场景均具备针对性的优化措施，数据库场景的综合性能较上一代提升 17%。
- **单核可用更大 L3 缓存：**至强®6 性能核处理器的 L3 共享缓存高达 504MB，相比前代提升显著。在现代数据库(如 MySQL、PostgreSQL、Redis)中，大量查询操作涉及对小数据块的随机访问。更大的 L3 缓存意味着这些热数据可以更大概率“命中”缓存而不用落到主内存，减少访问延迟。在单核查询时，可以直接从超大共享 L3 中抓取数据，避免等待 DRAM，显著加速 QPS(每秒查询数)和响应时间，降低访问延迟。比起“小 L3 + 多核”的设计，至强®6 的“大 L3 + 强核”架构，能显著降低 P99 查询延迟，稳定服务体验。

2、离线数据处理业务

针对离线数据处理业务的高吞吐计算能力、海量存储弹性扩展、高效数据传输、低延迟网络以及高容错与数据可靠性保障等要求，ECS g9i 支持弹性临时盘和高性能存储解决方案，在本地 Shuffle 测试场景中，单实例处理性能较上一代提升 12%，整机处理性能提升 10%。在数据挖掘、机器学习、数据分析等领域中，ECS g9i 可以为用户提供更快速、更准确的数据处理能力。

- **存储性能升级：**ECS g9i 弹性临时盘单盘最大支持 100 万 IOPS 和 4GB/s 吞吐。在大数据缓存加速场景中，这种高性能的存储解决方案可以显著提升系统的处理效率。支持 NVMe ESSD 云盘，访问时延降低 10%。
- **I/O 带宽：**ECS g9i 基于“飞天 + CIPU”技术架构，全面优化升级云上普惠弹性 eRDMA 网络，依托 CIPU 高速双上联网络能力，其时延降为传统 VPC 的三分之一，低至 8 微秒，可实现大规模组网及高弹性等优势，这有助于提升离线数据处理的 I/O 带宽和效率。
- **高性能存储：**ECS g9i 支持存算分离架构，用户可以根据业务需求灵活配置计算和存储资源。例如在构建数据存储方案时，可以根据冷热数据的不同需求，将数据存储于 OSS(对象存储服务)

务)中,并通过 EED(临时缓存盘)提供高性能的临时存储服务。这种架构有助于提升离线数据处理的灵活性和效率。

- **内存通道升级:** 至强®6 性能核处理器内存通道从 8 通道升级到 12 通道,能够输出更大的内存带宽和更高的数据传输速率,有助于进一步提升数据预处理任务的效率和体验,特别是对于需要处理大量数据的离线任务来说至关重要。

3、游戏 / 量化交易业务

游戏和量化交易业务均对实时性、计算性能与稳定性要求严苛。例如游戏后台必须持续低时延响应玩家操作,以保障足够流畅的体验,同时,高帧率渲染和复杂环境模拟对后台的计算资源需求极大。量化交易则依赖毫秒级行情捕捉、高频策略计算与精准订单执行,任何延迟或计算错误均可能导致收益损失或风险失控。

ECS g9i 除提供高性能、弹性、QoS 以及硬件加速等强大的在线业务支撑能力外,还特别具备以下保障优势:

- **高主频时的稳定性能:** 至强®6 性能核处理器在高主频(包括睿频)工作模式下,频率抖动极其微小,能够更好地在高主频下保持持续稳定的性能输出。
- **可靠性:** 高品质的硬件组件和先进的散热技术,确保了服务器的低故障率和高可靠性,最大程度避免故障导致的业务中断和数据丢失。
- **海量规格智能筛选:** 在游戏业务高弹性场景下,针对海量规格实现了智能筛选,动态规格供给模式有效减少了资源碎片,降本幅度可达 20%。

3.2 技术和架构创新,提升 AI 时代的向量数据处理和协同计算效率

1、数据预处理场景

阿里云 ECS g9i 在硬件层面与 SIMD 指令集深度集成,支持英特尔®高级矢量扩展 512(英特尔®AVX-512)指令集,每个内核拥有两个 512 位融合乘加(FMA)单元,显著加速 AI、科学计算和数据库等工作负载中常见的向量运算。ECS g9i 标配支持处理器内置的英特尔®高级矩阵扩展(Intel®

Advanced Matrix Extensions，英特尔® AMX 计算加速引擎，并新增支持 FP16 指令集，显著提升 AI 数据预处理等场景的执行效率。同时，基于 SIMD 指令集开发的软件能够在 ECS g9i 上无缝运行，无需重写代码，方便开发者利用 SIMD 指令集的优势进行应用优化。

英特尔在 SIMD 指令集方面持续创新，从 AVX 到 AMX、VNNI 等，在多媒体处理、量化交易、科学计算和人工智能等广泛领域被开发者所使用，推动计算架构从“标量时代”向“向量时代”跃迁。

2、训练场景

至强®6 性能核处理器的 Chiplet 架构集成 3 个计算单元与 2 个 IO 单元，每个计算单元支持多达 43 个核，并利用 EMIB 高带宽互联与统一内存一致性机制，带来类似单芯片大 Die 的计算性能和响应效率。在很多 AI 任务中，处理器可以直接支持对模型参数的操作，无需对数据进行跨 Die 传输，从而提升了响应效率，增强了数据一致性保障。192 条 PCIe 5.0 通道（单路 96 条）使 GPU 与 CPU 通信延迟大大降低，消除 GPU 数据吞吐瓶颈。MoE 大模型的专家层参数可在 CPU 内完整驻留，硬件级内存一致性协议（如 UPI 2.0）确保多 GPU 共享参数时无需冗余同步，减少训练迭代延迟。目前主流的 AI 计算架构在采用 GPU 集群的同时，均大量使用了至强®系列处理器作为机头 CPU 提供协同计算、管理和资源分配等能力来提升系统性能和能效，使工作负载的性能和总体拥有成本（TCO）之间达到更为理想的平衡点。

凭借持续迭代的 x86 架构技术优势，覆盖芯片设计、软硬件协同优化的系统级解决方案及全链条生态整合能力，英特尔目前在人工智能服务器 CPU 的市场占据 80% 以上的份额*。

3、推理场景

推荐系统

推荐系统作为 AI 在商业领域的成熟应用，是电商、广告等多行业的核心技术支持。随着数据量的激增，精准洞察用户需求、实现高效匹配成为竞争关键。现代推荐系统融合深度学习与大数据分析技术，构建起用户需求与内容展示间的精准连接，既为用户提供个性化服务，又增强企业产品曝光、驱动业务增长。算法的创新带来模型与特征处理复杂度提升，对硬件算力提出更高要求。至强®6 性能核处理器内置的英特尔®AMX，搭配深度神经网络库（英特尔®oneAPI Deep Neural Network Library，英特尔®oneDNN）、算子融合等软件方案，实现对智能推荐系统 AI 推理的优化加速，助力提升系统准确性与效率。

* 注：数据来自《IDC 中国加速服务器市场追踪》。

传统 AI

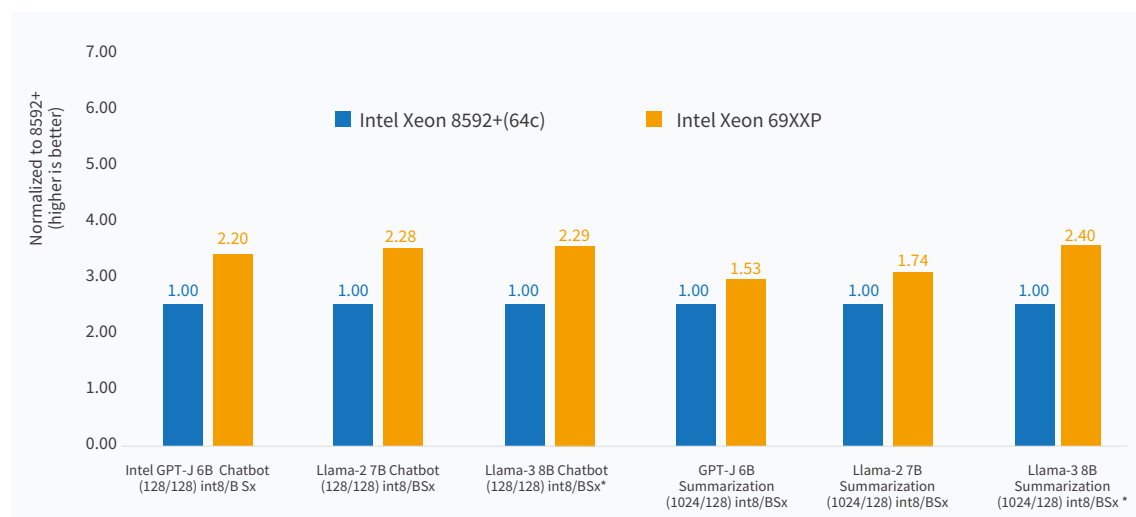
受限于算力限制和成本要求等各种原因，NLP、CV 等传统 AI 模型仍占据重要位置，满足多种业务和场景需求的精度和效果。英特尔® AMX 以及英特尔® oneDNN 能够显著优化模型中的矩阵乘法和卷积等计算，并借助算子融合、缓存分块等软件优化方案，结合英特尔® AVX-512 等向量加速指令集，进一步为传统 AI 场景提供优化加速。

大语言模型 (Large Language Models, LLMs)

随着 LLM 模型参数量的逐渐增加，KV cache 数据量也逐渐增加，意味着对 GPU 显存容量的要求也越来越高。实际推理场景中，对于 KV cache 的存储优化能大大提升推理系统的性能。为此，英特尔® QAT 重构了 KV Cache 存储范式。结合至强® 6 性能核处理器的多层统一内存架构，将高频访问的 KV 索引驻留 GPU 显存，温数据存储在 CPU 内存中，而历史上下文冷数据压缩后存入本地硬盘或者远端对象存储中，QAT 的超高压压缩解压性能，使磁盘中的压缩数据解压缩到内存并传输到显存中的延迟大大降低，提高用户体验的同时，节约了大量的 GPU 算力，实现了以存代算的总体设计模式。

另外，在很多领域，综合成本和效果两方面因素，LLM 常通过参数压缩量化、蒸馏等技术推动模型轻量化部署，尤其是 14B 参数以下的模型推理速度快，微调也更加方便，在很多中小企业或特定垂直行业部署广泛。针对此类需求，英特尔® AMX 对矩阵计算、卷积、注意力机制等 LLM 核心算子进行性能优化，结合英特尔® oneDNN 和高效的算子融合、缓存分块等软件优化技术，大幅提升 LLM 推理性能。内部测试数据显示，至强® 6 性能核处理器针对不同 LLM 的推理性能相比上一代处理器均有革命性提升。

图4 英特尔处理器在中小规模LLM推理性能的表现



来源: Intel

3.3 强化硬件安全设计，持续增强安全保障能力

ECS g9i 搭载的至强®6 性能核处理器从架构层面强化了内部的一系列硬件安全和可靠性设计，通过与云软件和虚拟化软件的深度优化，可以为用户提供丰富的安全增强能力和可靠性保障。

端到端安全防护

随着云部署成为主流，企业和用户对数据安全性的要求不断提升。英特尔早在第四代至强®处理器中就引入了英特尔®信任域扩展技术（英特尔®TDX），它基于硬件的可信执行环境，通过创建硬件隔离的虚拟机（信任域）来保护敏感数据和应用程序，防止未经授权的访问。异构系统的广泛应用，数据安全需求也延伸至 GPU 等加速器。为应对这一挑战，英特尔在至强®6 处理器中推出了 TDX Connect 技术，实现 CPU 与加速器协同构建统一的数据安全防护网。无论数据位于服务器 CPU 中处理，还是在加速器中运行模型与计算任务，TDX Connect 都能提供端到端的保护。它确保用户数据在云平台处理的全流程始终受到硬件级防护，兼顾安全性与高效性，有效隔离主机操作系统、虚拟机管理器和底层固件等系统软件对用户数据的越权访问，同时确保数据在高速 I/O 链路中传输的安全。ECS g9i 中深度集成了 TDX 技术，推出了面向机密计算场景的虚拟机产品。这些实例利用 TDX 提供的硬件级隔离能力，构建了真正意义上的“信任域”，确保客户在云端运行的应用和数据，即使在共享的多租户环境中，也能获得媲美物理隔离级别的安全保障。

稳定增强

至强®6 性能核处理器支持 ECS g9i 硬件实现双单路设计，能够有效降低故障时的爆炸半径，并实现故障时自动切换，降低业务中断风险。相当于为云计算系统安装了两个相互独立的“心脏”，给稳定性上了“双保险”，推动安全从“事后补救”转向“事前预防”。

此外，至强®6 性能核处理器还进一步强化了 RAS (Reliability, Availability, Serviceability) 特性。通过一系列创新功能显著提升系统的稳定性、可靠性和可维护性，旨在最大化平均故障间隔时间 (MTTF) 并最小化平均修复时间 (MTTR)，使系统在故障发生时能够快速恢复并继续运行：

- **可靠性：**至强®6 性能核处理器具备出色的错误检测与纠正能力，能实时监测数据传输和计算过程，及时纠错，保障高并发和大数据量时的数据准确性。硬件冗余设计也确保了关键部件有备份，故障时备份可迅速接管，提升系统可靠性。
- **可用性：**具备良好温度适应性、电压稳定性与抗干扰能力，能长时间、高负载完成计算任务。还支持热插拔，方便在系统运行时更换故障部件，减少维护时间，提高系统可用性。

可维护性：至强®6 性能核处理器提供远程管理功能，管理员可远程监控、管理处理器状态，进行故障诊断与修复，提升维护效率。此外，配备专业诊断工具，能提供详细系统信息和故障日志，助力技术人员快速定位并解决硬件问题。

3.4 全球一致的云服务能力体系，全面助力企业国际化战略

阿里云 ECS 已通过全球 29 个数据中心、88 个可用区为用户提供安全、完整的算力规格，支撑广泛的应用负载运行，ECS 具备一致的、跨区可用的高可用架构自动恢复能力，并提供统一的开发与运维体验，相关产品已通过多方国际安全认证。未来，全球各节点都会全面部署搭载至强®6 性能核处理器的 ECS g9i 实例产品，结合各节点前期持续使用的英特尔®至强®系列处理器，用户完全无需担心系统和数据的平滑迁移问题。阿里云 ECS g9i 产品及其配套的全球化、体系化服务能力，将为中国企业的国际化战略提供以下强大保障：

- **全球部署：**除上述全球数据中心和可用区资源外，阿里云还拥有超 3200 个边缘节点提供本地化就近部署能力，资源可灵活弹性扩展，满足对网络时延要求高的场景，可为中国企业遍布全球的分支机构或生产基地等提供边缘计算资源和网络接入。2025 年，阿里云已在海外投资新建多个中心区域，涉及欧洲、南美、北美、东南亚等地域，这些数据中心即将陆续开服，为中国企业当地业务开展提供更多更优质的选择
- **弹性大规模：**弹性架构能够有效应对出海企业的业务流量高峰，确保业务的稳定性和连续性。在应用层，阿里云推出了 ACK 容器集群，基于用户负载特征的预调度技术，ACS Pod 每分钟扩容速度提升至 10000 个，单集群最大可支持 15000 节点。
- **跨区可用性和一致性：**企业在不同国际化阶段的建设需求和策略有所不同：在起步阶段，企业通常会选择一个局部地域进行集中化部署，以便快速在全球开展业务，服务全球客户；在扩张阶段，企业一般采用单元化架构，按国家或地域维度来建设，从而更好地支持本地业务发展；当企业进入成熟期后，为了给客户提供更优质的服务，会更加注重服务的稳定性和运营的精细化，进而更加关注业务数在全球多活容灾、成本控制和性能优化等能力。阿里云在全球不同地域提供的云产品具有高度一致的服务体验和服务质量，有效满足企业客户在全球化不同阶段的高标准服务需求。例如，阿里云 ACK One 作为面向混合云、多集群、容灾等场景推出的企业级云原生平台，可以连接和管理任何地域、任何基础设施上的 Kubernetes 集群，并提供一致的管理，支持对应用、流量、安全、存储、可观测等进行统一管控。

04 优秀实践分析

4.1 小鹏汽车

在智能驾驶的竞速赛道上，数据是驱动进化的核心燃料。小鹏汽车，作为中国智能电动汽车的先锋，其海量车机视频数据的实时处理与分析能力，直接决定了算法迭代与用户体验优化的速度。面对日均 PB 级数据的切分、转码等预处理挑战，小鹏汽车选择与阿里云弹性计算深度合作，基于第九代企业级实例 ECS g9i，构建了高效、稳定的数据处理基座。

业务痛点：海量数据的预处理之难

原始数据体量巨大、格式各异，必须经过高效的切分、转码、标注等预处理，才能“喂食”给后续的 AI 模型训练与仿真验证流程。随着数据的指数级增加，以往的数据处理逐渐面临性能上的挑战。同时为了不影​​响模型迭代周期，数据处理的时效也需要更加快速。

解决方案与客户价值：

第九代企业级实例 g9i，基于自研“CIPU+ 飞天”技术架构，搭载英特尔®至强®6 处理器，软硬协同优化，助力小鹏提升数据预处理效率，显著降低算力成本：

- **性能与性价比提升：**相比第八代实例，ECS g9i 单核算力最大提升 20%，通过 3.6 GHz 全核睿频，504MB 超大 L3 缓存，可保障高并发场景流畅响应。采用原生硬件加密 QAT 技术，集成 SSL/TLS 加密、压缩 / 解压缩加速引擎，可从 CPU 卸载计算密集型对称和非对称加密以及数据压缩 / 解压缩操作，释放 CPU 资源，使海量自动驾驶数据出入数据库时的透明压缩和解压等任务更高效，显著降低算力成本。
- **稳定可靠，持续运行：**ECS g9i 的双单路设计，能够有效降低故障时的爆炸半径，并实现故障时自动切换，降低业务中断风险，为小鹏 7x24 小时不间断的数据处理业务提供了坚实保障，确保了研发进程不受中断。

品 4.2 微帧科技

微帧科技 (Visionular) 作为领先的视频编码与超高清服务提供商, 深耕视频编码、深度学习、图像处理等领域 20 余年。公司拥有自研的核心编码引擎, 为长视频、短视频、直播、RTC 等全视频场景提供智能化解决方案, 目前已与国内外 100 多家头部企业达成合作, 每月处理视频超 15 亿分钟, 覆盖从云端到终端的全链路视频服务。

业务痛点:

● 高性能算力需求: 释放自研编码内核技术潜力

微帧科技自研的编码内核需要高性能算力平台, 满足超高清视频处理的计算强度需求。在视频编码前, 需要运用降噪、画质增强等 AI 预处理技术对原始视频进行优化, 这需要强大的算力支持, 用以在压缩过程中保障画质。

● 高弹性资源适配: 应对算力密集型业务潮汐波动

作为算力密集型服务, 编解码业务面临显著的流量潮汐特征, 需构建可快速扩缩容的海量算力池, 实现资源供给与业务需求的动态匹配。

● 成本优化诉求: 平衡算法迭代与资源效率的双重目标

在支撑算法引擎持续迭代与应对业务流量波动的过程中, 需通过弹性资源调度机制优化算力成本, 避免资源闲置损耗, 实现技术投入与运营效率的平衡。

解决方案:

● 高性能算力底座: ECS g9i 为视频编码前的 AI 预处理提供了有力保障, 确保降噪、画质增强等操作能够快速、精准执行, 充分释放解码引擎的技术潜能。

● 全球化算力网络, 多地域部署灵活覆盖: 依托阿里云在国内和国际核心地域的完善基础设施, 微帧科技得以构建大规模算力服务网络, 支持业务就近部署, 确保为用户提供低延迟、高可用的优质算力供给。

● 弹性资源调度, 抢占式实例即开即用: ECS g9i 结合抢占式实例模式, 通过智能调度实现算力资源的“即开即用、动态弹性”, 使业务能够灵活、高性价比地应对业务潮汐需求。

客户价值：

- **编解码效率提升：** ECS g9i 为微帧科技的 WZ264/265 及 AV1 编码效率带来了突破性提升。其中，AV1 推理耗时优化超 50%，4K 处理效率提升 35%，显著提高了视频编解码的处理速度和质量。
- **业务全面加速：** 依托阿里云国内核心地域及基础设施，构建就近接入网络，为业务提供低延迟算力支撑，加速了 AI 预处理和编码全流程。
- **智能调度降本增效：** 通过采用任务触发弹性 + Spot 实例策略，微帧科技在非实时任务处理上成本最多能实现降低 60%，资源利用率可提升至 85%。



品 4.3 嘎嘎射击

厦门真有趣信息科技有限公司，专注于移动游戏研发、全球化发行及原创 IP 孵化，累计服务超 2 亿全球玩家。其多人竞技射击手游作品嘎嘎射击融合了快节奏动作与视觉，推出即登上热门游戏榜单。

业务痛点：

- **并发性能瓶颈：**游戏内物理引擎计算、实时弹道模拟、百人同屏战斗等场景对 CPU 算力需求极高。原有基于老旧实例的基础设施在高峰时段 CPU 利用率长期饱和，导致服务器响应延迟显著增加，玩家竞技体验受到影响。
- **稳定性危机：**DAU 快速攀升，尤其在版本更新或大型活动期间，流量洪峰使老旧实例不堪重负，频繁出现服务抖动，运维团队疲于应对突发故障，游戏口碑面临风险。
- **成本与弹性失衡：**传统服务器架构难以在业务高峰（如新赛季开启）与低谷期之间实现资源的敏捷伸缩，既无法保障高峰体验，又在低谷期造成资源闲置与成本浪费。

解决方案：

为应对上述挑战，真有趣采用阿里云基于英特尔®至强®6 处理器的 ECS g9i 实例构建高性能游戏架构，结合阿里云全栈云服务实现端到端优化：

- **高性能计算底座：**ECS g9i 单核算力最大提升 20%，搭配 3.6G Hz 全核睿频与 504MB 超大 L3 缓存通过“飞天”云操作系统的智能调度，实现跨可用区算力负载均衡，确保团战场景下逻辑处理延迟缩小，消除操作卡顿。
- **弹性扩展体系：**基于阿里云弹性伸缩，实时监测 CPU 利用率、开服峰值期快速扩展，资源部署效率提升 30%，低谷期自动缩减至基础集群，显著降低 IT 成本。

客户价值：

- **极致性能支撑爆款落地：**ECS g9i 助力《嘎嘎射击》开服首周 DAU 突破 20 万，多人团战场景性能提升 30%。
- **资源效率与成本双重优化：**算力成本降低 26%，CPU 资源利用率显著提升，实现“峰值不卡顿、低谷不浪费”的精细化运营。

品 4.4 蚂蚁集团ZOLOZ

蚂蚁数字科技是蚂蚁集团的技术商业化业务部门，ZOLOZ 是蚂蚁数科的身份安全品牌，产品囊括可信身份认证、反洗钱、终端安全等，为中国、印尼等 14 个国家和地区的 70 余家合作伙伴提供技术服务。ZOLOZ RealId 产品采用深度学习模型提供实人认证、人脸识别、证件核验等组合能力，帮助客户解决数字化渠道下实人认证相关的体验和安全问题，应用于金融业务核身、人证核验、人脸登录等多种场景。

业务痛点：

让金融机构能够快速构建专属的智能体，以“数字员工”的身份处理客户咨询、执行交易、分析报告、管理风险：

- **性能瓶颈：**复杂的金融问答、海量报告的即时分析，要求大语言模型具备极低的响应延迟，保障客户体验和交易机会。
- **精准度挑战：**金融场景对信息的解析和意图的理解要求“零容错”。智能体必须在纷繁复杂的金融术语和上下文中，实现超高精度的信息提取与判断。
- **安全隐忧：**智能体掌握着敏感的客户数据和交易权限。如何防范模型本身的安全漏洞、抵御针对 AI 的新型攻击、确保交互内容的安全合规，成为决定其能否被信任的核心。

解决方案：

阿里云 +Zoloz 方案面向智能体开发，提供 MaaS、Paas、Saas 服务以及 AI safety 领域。在选择计算实例时关注算力性能、响应时长、成本优化。

● MaaS 层：

- 通过 ZOLOZ 大模型平台提供全生命周期工具链，快速部署 LLM 模型推理服务全面支持 20B 参数以下的模型；利用 CPU 资源减少对 AI 加速卡的需求，降低推理服务的总拥有成本 (TCO)。在离线推理场景中，ZOLOZ 选择了小模型推理服务 CPU 化的策略：从阿里云 ECS 7 第七代英特尔代到目前的 9 代，视觉识别模型实际测试整体效率提升了四倍，综合计算成本降低了 20%。

- ZOLOZ FinLLM 结合提示工程与海量金融数据的指令微调,输出精准、上下文感知的专业内容,处理市场情绪分析、行情预测、风险评估、自动化客服等关键金融任务,保证任务处理时间在时延阈值范围内。在这个过程中,ECS g9i 在模型推理加速方面提升了 2.3 倍,推理消耗时长降低了 35%,算力成本降低了 72%。
- **PaaS 层:** Zoloz 的智能体平台支持金融客户搭建智能体应用,构建企业级智能知识库平台——RAG。在文档提取洞察、智能合规审计等场景中,实测智能体表现优异,任务平均处理时间大幅降低。搭载至强® 6 处理器,在承载同等规模智能体服务时,所需服务器节点数量减少了 30%,金融机构由此可降低 25% 的硬件采购与运维成本。
- **安全应用:** 在线支付的“刷脸”瞬间,需要同时完成活体检测、人脸比对和交易风险评估。英特尔处理器的强大算力确保了 ZOLOZ 服务在极端并发压力下,保持低于 100 毫秒的响应速度。蚂蚁数科基于“以 AI 对抗 AI”的创新理念,构建的高效智能检测防御体系——ZOLOZ Anti-Deepfake 攻防互动系统,已部署在 ECS g9i 平台上,可为 AI 推理带来最高达 3.3 倍的性能提升,并将每瓦性能提升高达 1.7 倍。



05 结语

自 2009 年起，阿里云与英特尔就已开启深度技术合作，强化软硬协同和性能挖掘。作为阿里云核心计算产品，ECS 实例已历经十五年商业化历程。至今，双方合作深度贯穿至强® 处理器家族历代产品迭代和演进，持续为云计算算力升级提供底层支撑。在数字经济蓬勃发展、企业数字化转型加速推进的时代，阿里云与英特尔的合作成果显著，意义深远，成为推动产业升级、赋能行业发展的关键力量。

面对企业加速构建 AI 能力体系的新需求，双方的合作不仅是技术层面的优势互补，更体现在产业生态的协同共建，通过深度融合与协同创新，共同打造更高效、更智能、更安全的全栈云服务，帮助企业客户更从容地应对 AI 时代内部运营效率和外部市场环境的双重挑战，更好地支撑业务创新发展。

面向未来，阿里云依托英特尔高性能处理器的优势技术，正在以前所未有的强度投入 AI 技术研发和云基础设施建设。双方将继续秉持合作共赢的理念，进一步强化技术研发与产业协同，持续为客户创造价值，携手谱写 AI 时代的云智能新篇章。



关于 IDC

国际数据公司（IDC）是在信息技术、电信行业和消费科技领域，全球领先的专业的市场调查、咨询服务及会展活动提供商。IDC帮助IT专业人士、业务主管和投资机构制定以事实为基础的技术采购决策和业务发展战略。IDC在全球拥有超过1100名分析师，他们针对110多个国家的技术和行业发展机遇和趋势，提供全球化、区域性和本地化的专业意见。在IDC超过50年的发展历史中，众多企业客户借助IDC的战略分析实现了其关键业务目标。IDC是IDG旗下子公司，IDG是全球领先的媒体出版、会展服务及研究咨询公司。

IDC China

IDC中国（北京）：中国北京市东城区北三环东路36号环球贸易中心E座901室

邮编：100013

+86.10.5889.1666

Twitter: @IDC

blogs.idc.com

www.idc.com

版权声明

凡是在广告、新闻发布稿或促销材料中使用IDC信息或提及IDC都需要预先获得IDC的书面许可。如需获取许可，请致信 gms@idc.com。翻译或本地化本文档需要IDC额外的许可。

获取更多信息请访问www.idc.com，更多有关IDC GMS信息，请访问<https://www.idc.com/prodserv/custom-solutions>。

版权所有2025 IDC。未经许可，不得复制。保留所有权利。