仅供内部参考，请勿外传！

计算机

云栖大会：坚定投入 AI 模型、AI 芯片、AI 云三大基石

2026 年 9 月 22 日，云栖大会在杭州开幕，阿里巴巴集团 CEO 吴泳铭明确，阿里将围绕 AI 模型、AI 芯片与 AI 云进行长期战略投入。AI 云给出到 2032 年阿里云运营的全球数据中心规模超过 20GW 的目标，验证 AI 基础设施资本开支的中长期上行。

模型层面，公司沿基模智能水平、交互能力、商业化同步推进。基模端，大模型 RSI 软硬件两端同步发力，已初步进入模型训练、推理及芯模协同等环节中；交互端，多模态能力是提升 AI 交互的关键，以理解与生成一体化的多模态能力将补齐模型感知与交互；商业化层面，千问大模型落地多款智能终端，阿里发布 AI 手机全栈解决方案 Qwen Intelligence 与随身硬件 QwenNote，融合千问办公，加速端侧 AI 商业化落地。

算力层面，阿里实现从单芯片延伸至超节点与大规模集群的全方位布局。1）AI 芯片：发布新一代训推一体 AI 芯片真武 V900，单芯片性能是真武 M890 的 3 倍，还原生支持 FP8、FP4 低精度计算，在大幅提升算力密度的同时显著降低推理成本；2）CPU：发布倚天 720、倚天 730 和倚天 750 多款服务器 CPU 后续规划，单核性能、CPU 与 AI 芯片的协同效率将持续提升；3）超节点：发布新一代磐久超节点服务器，集成真武 V900、ICN Switch、磐脉智能网卡及镇岳 SSD 主控芯片，单一 AI 集群可扩展至 50 万卡。

云层面，围绕 Model、Harness、Context 三大核心场景，构建 AI Native Cloud、Agent Native Cloud 与 Context Engine 技术栈。1）AI Native Cloud：提升集群利用率与 Token 分发效率，并缩短模型从开发验证到规模化服务的工程路径；2）Agent Native Cloud：围绕企业级 Agent 构建，提供 Infra、Platform、Desktop 三层能力；3）Context Engine：针对企业端 Agent 落地缺乏专有语境，围绕“存储—计算处理—上下文工程”，将企业全域数据转化为实时上下文。

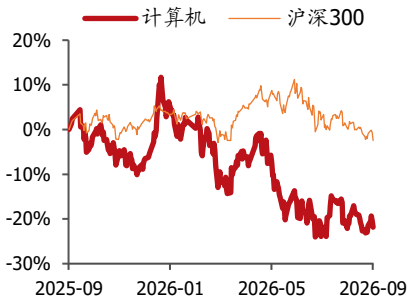
建议关注：

需求端，Coding、AI 办公等 AI 应用场景落地持续放大推理需求，吴泳铭云栖大会表示当前机器思考总量不到人类的 3%，未来将达到 1000 倍以上。供给端，阿里云给出 2032 年全球数据中心超 20GW 目标，并确立智能时代基建框架：模型端 Scaling 与多模态主线确立，芯片端从单芯走向超节点与大规模集群，云端以 Agentic Cloud 承接商业变现。智能需求中长期增长与供给框架落定，阿里中长期投入方向明确，AI 产业链投入确定性抬升，看好产业链景气度持续上行。建议关注：1）模型层：智谱，MiniMax 等；2）国产算力：寒武纪、海光信息、天数智芯、壁仞科技、沐曦股份、摩尔线程等。

风险提示：技术迭代不及预期风险；经济下行超预期风险；行业竞争加剧风险。

增持（维持）

行业走势



作者

分析师 **孙行臻**

执业证书编号：S0680526010001

邮箱：sunxingzhen1@gszq.com

研究助理 **徐瀚**

执业证书编号：S0680126090003

邮箱：xuhan@gszq.com

相关研究

1、《AI 投资进入回报兑现期：：宏大叙事延续，CAPEX/ARR 剪刀差收敛》 2026-09-23

2、《计算机：华为领航三大看点：单芯提速+系统工程创新+软硬全栈优化 》 2026-09-20

3、《计算机：AI 安全：从技术政策共振到海外业绩兑现，行业加速在即》 2026-09-12

内容目录

一、大模型：RSI 抬升智能上限，多模态拓展交互边界 3
二、算力：单芯片到超节点与大规模集群全方位布局 6
三、云：构建多层技术栈，支撑企业级 Agent 规模化落地 7
建议关注 7
风险提示 8

图表目录

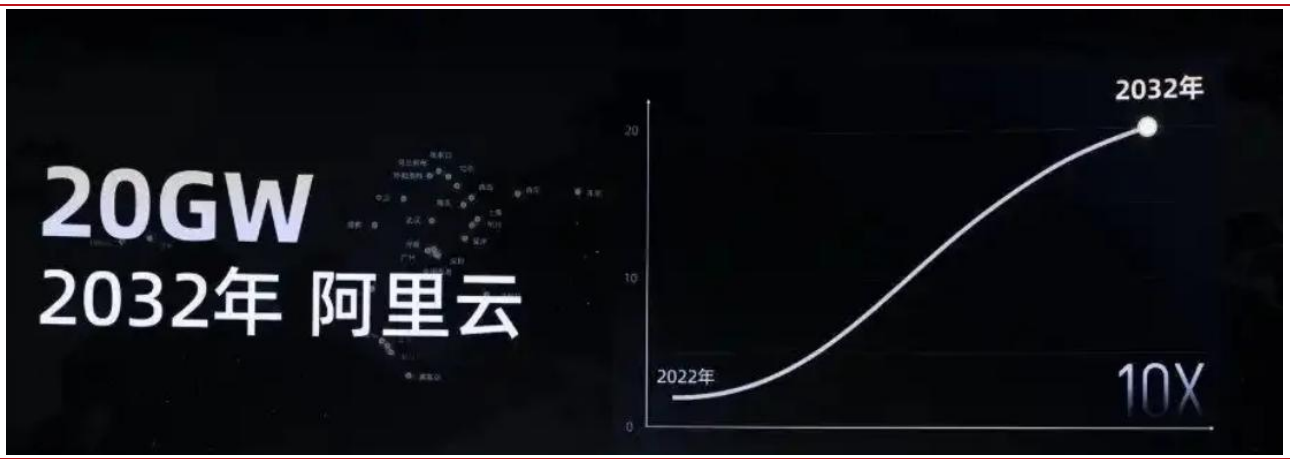
图表 1：2032 年阿里云运营的全球数据中心目标规模将超过 20GW 3
图表 2：阿里大模型预训练持续 Scaling，未来将推出 5-10T 大模型 4
图表 3：阿里巴巴 AI 大模型全景图 5
图表 4：平头哥新一代训推一体 AI 芯片真武 V900 6
图表 5：新一代磐久超节点服务器 6
图表 6：阿里 Agent Native Cloud 架构 7



一、大模型：RSI 抬升智能上限，多模态拓展交互边界

2026 年 9 月 22 日，云栖大会在杭州开幕，阿里巴巴集团 CEO 吴泳铭明确，阿里将围绕 AI 模型、AI 芯片与 AI 云进行长期战略投入。三者被定义为“机器思考能够规模化供给的前提”：芯片生产 Token，模型决定 Token 的智能水平，云负责把 Token 送到需求端。其中 AI 云给出到 2032 年阿里云运营的全球数据中心规模超过 20GW 的目标，验证 AI 基础设施资本开支的中长期上行。

图表1：2032 年阿里云运营的全球数据中心目标规模将超过 20GW



资料来源：电商派 Pro 公众号，国盛证券研究所

模型层面，公司沿基模智能水平、交互能力、商业化同步推进：基模层面，以 RSI 提升能力上限；交互层面，以理解与生成一体化的多模态能力补齐模型感知与交互；商业化层面，以端侧 AI 方案打通物理场景，商业化兑现提速：

1）大模型 RSI 软硬件两端同步发力，已初步进入模型训练、推理及芯模协同等环节中。基于新一代架构的 Qwen4 已在训练中，未来 Qwen4.5、Qwen5 等版本将扩展至 5-10T 参数。

- 模型自我训练：Qwen3.8-Max 通过自主搭建训练流程、构造训练数据，并自主设计实验、定位缺陷，在人类“零参与”的情况下持续迭代超 1 个月，完成 33 轮有效迭代。
- 推理优化：Qwen3.8-Max 在平头哥新款 GPU 上，自主适配优化了下代架构的 Qwen3.8-Flash 新模型的推理框架，实现单实例推理吞吐量提升 96%。
- 芯片模型协同设计：Qwen3.8-Max 基于真实的总线模块规范，自主展开前端、验证、后端的全链路自迭代，完成了减少 42%面积的物理实现优化。

图表2：阿里大模型预训练持续 Scaling，未来将推出 5-10T 大模型



资料来源：云栖大会官网，国盛证券研究所

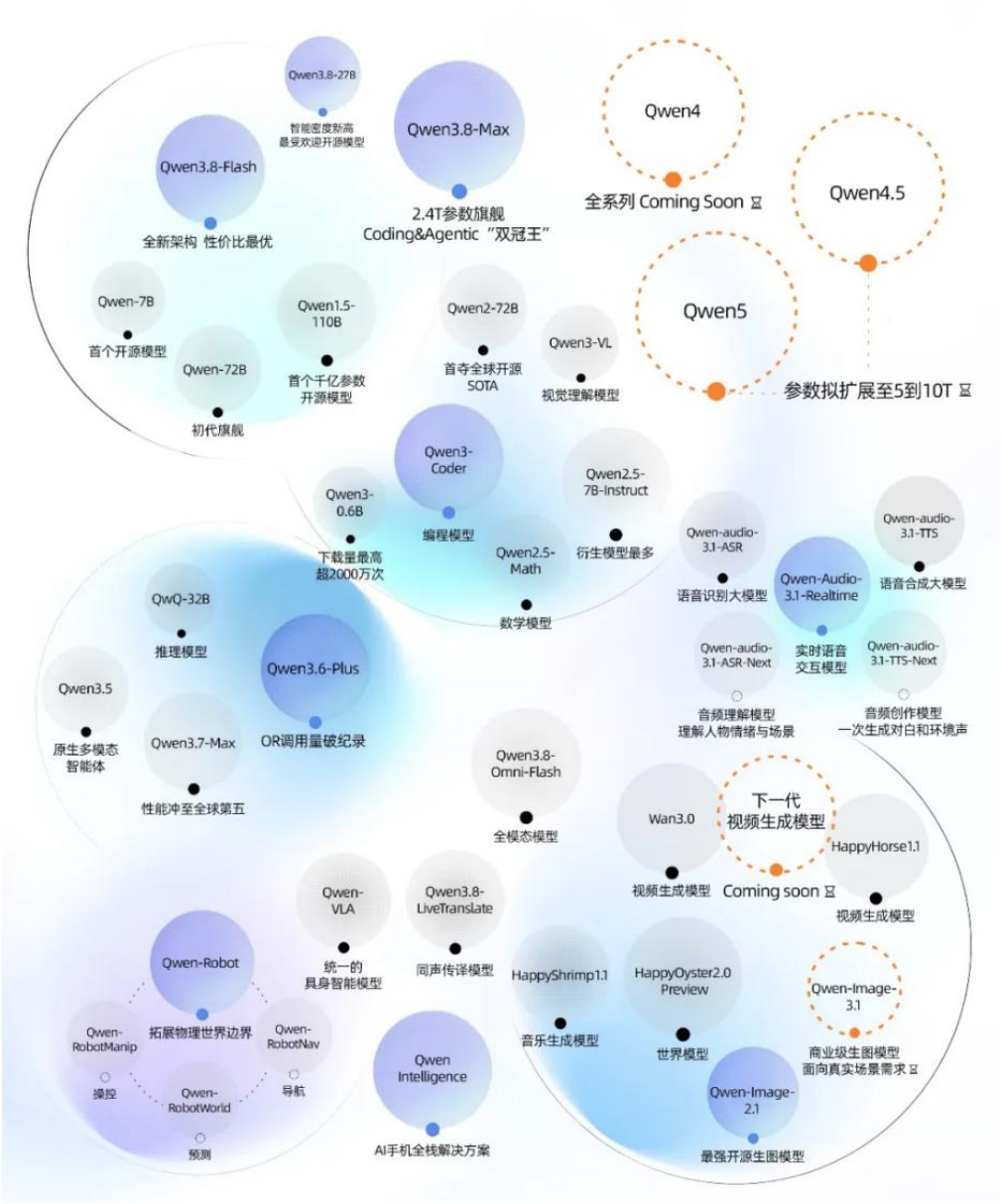
2）多模态能力是提升 AI 交互的关键，Qwen 团队将其确立为基模智能水平之外的重要研发方向，大会期间图像、语音等类型生成模型相继发布：

- 发布多款多模态生成大模型：发布图像生成模型 Qwen-Image-3.1、音乐生成模型 Happy Shrimp1.1 版本发布，世界模型 HappyOyster 2.0 Preview 亮相，下一代视频生成模型将于 11 月发布。
- Qwen-Audio-3.1 语音大模型升级：发布音频理解模型 Qwen-Audio-3.1-ASR-Next、音频创作模型 Qwen-Audio-3.1-TTS-Nex，实时交互模型 Qwen-Audio-3.1-Realtime 交互能力提升。

3）千问大模型落地多款智能终端，阿里发布 AI 手机全栈解决方案 Qwen Intelligence 与随身硬件 QwenNote，融合千问办公，加速端侧 AI 商业化落地：

- AI 手机全栈解决方案 Qwen Intelligence：基于千问大模型，覆盖任务规划、跨应用执行、影像创作等 AI 手机需求。Qwen Intelligence 提供了三套专为手机场景深度优化的模型与 Agent 解决方案，均达到行业 SOTA 水平，分别承接规划、操作和创作。方案将在 9 月 28 日发布的荣耀新机 Magic9 落地。
- 硬件产品 QwenNote：发布由阿里千问语音大模型驱动的首款 QwenNote 产品 A2，可获取用户在各类线下场景中的信息，并语音调用千问办公。

图表3：阿里巴巴 AI 大模型全景图



资料来源：阿里巴巴公众号，国盛证券研究所

二、算力：单芯片到超节点与大规模集群全方位布局

算力层面，阿里完成从 AI 芯片、CPU、网卡、ICN 互联芯片到超节点方案的全方位布局，算力供给从单芯片延伸至超节点与大规模集群。

- 单芯层面：1）AI 芯片：发布新一代训推一体 AI 芯片真武 V900，芯片基于自研并行计算架构设计，搭载 216GB 大容量显存，片间互联带宽达 1200GB/s，单芯片性能是真武 M890 的 3 倍。芯片还原生支持 FP8、FP4 低精度计算，在大幅提升算力密度的同时显著降低推理成本，可应用于高精度训练、低精度和超低精度推理的全场景。2）CPU：将于 2027 年推出倚天 720 和倚天 730 两代服务器 CPU，单核性能持续提升；还将继续推出基于第二代自研核的倚天 750，该芯片支持平头哥自研 ICN 片间互联总线协议，可与真武 AI 芯片通过 ICN 总线直接互联，进一步提升 CPU 与 AI 芯片的协同效率。

图表4：平头哥新一代训推一体 AI 芯片真武 V900



资料来源：云栖大会官网，国盛证券研究所

- 超节点：发布新一代磐久超节点服务器，集成真武 V900、ICN Switch、磐脉智能网卡及镇岳 SSD 主控芯片，结合阿里云新一代智算中心网络架构，实现算、存、网全栈系统级协同，单一 AI 集群可扩展至 50 万卡。

图表5：新一代磐久超节点服务器



资料来源：云栖大会官网，国盛证券研究所

三、云：构建多层技术栈，支撑企业级 Agent 规模化落地

阿里云围绕 Model、Harness、Context 三大核心场景，构建 AI Native Cloud、Agent Native Cloud 与 Context Engine 技术栈，并推出 CPFS、AgentCore、Agent Context 等系列产品，支撑企业级 Agent 规模化落地。

- AI Native Cloud：聚焦存储、缓存、网络、边缘智能与 AI 平台等面向 AI 负载的云基础设施。该层一方面提升集群利用率与 Token 分发效率，另一方面压缩模型从开发验证到规模化服务的工程路径。

- Agent Native Cloud：围绕企业级 Agent 构建，提供 Infra、Platform、Desktop 三层能力。1）Agent Infra：以算力、运行环境、数据、网络为底座，为 Agent 提供安全、弹性、低成本的运行环境。2）Agent Platform：提供一整套企业级 Agent PaaS，本质是统一的控制平面，将分散的 Agents 纳入企业级治理与持续进化体系。3）Agent Desktop：让 Agent 进入真实业务界面，与人在同一工作环境协同。

图表6：阿里 Agent Native Cloud 架构



资料来源：云栖大会官网，国盛证券研究所

- Context Engine：针对企业端 Agent 落地缺乏专有语境，围绕“存储—计算处理—上下文工程”，将企业全域数据转化为实时上下文。1）底层：以多模态数据存储承接对话、视频帧、点云、轨迹与 Agent 反馈；2）中间层：覆盖数据集成、转换、治理、质量与建模；3）上层：通过全域知识库、多模态语义整合、持续记忆管理和实时上下文装配，把分散数据组织成 Agent 可按需调用、动态更新的上下文。

建议关注

需求端，Coding、AI 办公等 AI 应用场景落地持续放大推理需求，吴泳铭云栖大会表示当前机器思考总量不到人类的 3%，未来将达到 1000 倍以上。供给端，阿里云给出 2032 年全球数据中心超 20GW 目标，并确立智能时代基建框架：模型端 Scaling 与多模态主线确立，芯片端从单芯走向超节点与大规模集群，云端以 Agentic Cloud 承接商业变现。智能需求中长期增长与供给框架落定，阿里中长期投入方向明确，AI 产业链投入确定性抬升，看好产业链景气度持续上行。建议关注：1）模型层：智谱，MiniMax 等；2）国产算力：寒武纪、海光信息、天数智芯、壁仞科技、沐曦股份、摩尔线程等。

风险提示

技术迭代不及预期风险：若技术迭代不及预期，则对产业链相关公司会造成一定不利影响。

经济下行超预期风险：若宏观经济景气度下行，固定资产投资额放缓，影响企业再投资意愿，从而影响消费者消费意愿和产业链生产意愿，对整个行业将会造成不利影响。

行业竞争加剧风险：若相关企业加快技术迭代和应用布局，整体行业竞争程度加剧，将会对目前行业内企业的增长产生威胁。

免责声明

国盛证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，但本公司及其研究人员对该等信息的准确性及完整性不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，可能会随时调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本公司力求报告内容客观、公正，但本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。投资者应当充分考虑自身特定状况，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。

本报告版权归“国盛证券股份有限公司”所有。未经事先本公司书面授权，任何机构或个人不得对本报告进行任何形式的发布、复制。任何机构或个人如引用、刊发本报告，需注明出处为“国盛证券研究所”，且不得对本报告进行有悖原意的删节或修改。

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的任何观点均精准地反映了我们对标的证券和发行人的个人看法，结论不受任何第三方的授意或影响。我们所得报酬的任何部分无论是在过去、现在及将来均不会与本报告中的具体投资建议或观点有直接或间接联系。

投资评级说明

投资建议的评级标准		评级	说明
评级标准为报告发布日后的 6 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以摩根士丹利中国指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准。	股票评级	买入	相对同期基准指数涨幅在 15%以上
		增持	相对同期基准指数涨幅在 5%~15%之间
		持有	相对同期基准指数涨幅在-5%~+5%之间
		减持	相对同期基准指数跌幅在 5%以上
	行业评级	增持	相对同期基准指数涨幅在 10%以上
		中性	相对同期基准指数涨幅在-10%~+10%之间
		减持	相对同期基准指数跌幅在 10%以上

国盛证券研究所

北京
地址：北京市丰台区丽泽路 22 号院新华社国家金融信息大厦西塔 B 座 38 层
邮编：100077
邮箱：gsresearch@gszq.com

上海
地址：上海市浦东新区南洋泾路 555 号陆家嘴金融街区 22 栋
邮编：200120
电话：021-38124100
邮箱：gsresearch@gszq.com

南昌
地址：南昌市红谷滩新区凤凰中大道 1115 号北京银行大厦
邮编：330038
传真：0791-86281485
邮箱：gsresearch@gszq.com

深圳
地址：深圳市福田区福华三路 100 号鼎和大厦 24 楼
邮编：518033
邮箱：gsresearch@gszq.com