面向Agent，阿里云全栈重构

本报记者 陈存

刚刚结束的2026阿里云峰会，被不少参会者视为阿里云成立17年以来最具颠覆性的发布会之一。

不同于以往围绕单一模型、芯片或云产品展开的技术升级，这一次，阿里云首次围绕Agent进行全栈产品发布，从底层芯片、云基础设施，到模型、推理平台，再到面向开发者的服务体系，几乎所有能力都被重新定义。

某种程度上，这场峰会更像是一次产业范式切换的公开宣告：AI正在从辅助工具走向执行主体，而阿里云，也开始从服务人转向服务Agent。

“我们已经真正迈入Agentic时代。”阿里云智能集团资深副总裁、公共云事业部总裁刘伟光在峰会上表示，“为Agent时代，阿里云全栈就绪。”

全栈AI升级

未来两到三年，云计算最大的增长机会，将来自海量Agent的运行、编排与协同。

过去几年，大模型产业的核心逻辑持续演进，从主要解决对话问题，到具备长文本理解能力和对复杂问题的思考能力。2025年之后，AI开始进入做事阶段，带动其商业化进程迈入新纪元。

就在峰会前不久，阿里最新财报显示，AI相关产品收入已连续11个季度保持三位数同比增长。AI相关产品收入在阿里云外部商业化收入中的占比首次突破30%。AI模型及应用服务ARR（年度经常性收入）已超过80亿元，预计年底突破300亿元。阿里巴巴集团首席执行官吴泳铭表示，预计未来一年，AI相关产品收入的占比将突破50%，成为推动云业务收入增长的主要引擎，阿里全栈AI技术投入已正式跨越初期培育阶段，进入正向的规模商业化回报周期。

据悉，截至2026年3月，阿里云旗下大模型平台百炼客户数量同比增长8倍，过去三个月token消耗规模较上一季度大幅提升。QuestMobile数据显示，千问成为了今年用户增长最为迅猛的APP，月活同比增长4241%。

这也是阿里云此次全面转向Agent的重要背景。在刘伟光看来，Agentic时代最核心的三个变化，一是模型能力跨越式提升，AI能够自主完成复杂的长链路任务，其本身成为能力增长的引擎；二是AI能直接交付结果而非跑完流程，甚至能直接创造“原本无法发生的产出”；三是人机交互模式变革，AI Native组织新范式诞生，超越人类劳动者数个数量级，且能24小时×7天持续工作的Agent劳动者涌现，人类员工与Agent形成混合工作网络，甚至创造出新的流程。

阿里巴巴首席财务官徐宏在财报电话会上称，阿里“看到了（投入AI的）历史机遇”“这个窗口期对我们来讲可能就是几年的时间”。阿里云智能集团首席技术官李飞飞也认为，未来两到三年，云计算最大的增长机会，将来自海量Agent的运行、编排与协同。他表示：“传统企业，以及新型的AI原生组织和企业，都在快速地从以人为中心的工作流和为人编写的SaaS化的工作流，向以Agent为中心、为Agent工作去提供支持的工作流演进。”

对AI和云的需求无穷无尽，阿里云正在建设的是“中国最大的AI工厂”。

为Agent而生的云

在本次峰会上，“为了Agent”“帮助Agent”“面向Agent”成为高频概念，产品重构、流程优化、应用落地的革新步伐也在同步迈开。

刘伟光总结了Agentic时代的四大基石：一是支撑Agent长程任务的模型能力，能为生产级Agent提供所需的跨越式迭代速度和可靠性；二是支撑Agent运行负载的Agentic Cloud；三是支撑Agent灵活调用的工具与服务，包括Skill、MCP、CLI等；四是支撑推理井喷的性能与供应能力。

四大需求正对应阿里云的五大布局，即底层芯片、Agent Cloud、全模态模型矩阵、模型服务平台和Agent应用，刘伟光宣布，阿里已率先成为完整打通五层全栈的云厂商和AI厂商。

在芯片层，会上，阿里云发布了面向Agent时代最新的并行计算芯片真武M890，以及配套的ICN Switch互联芯片，旨在为Agentic时代的工作负载提供持续的升级优化服务。

在Agent Cloud层，云的作用在于为Agent的工作负载提供关键支撑，支持Agent的开发和应用；同时，让Agent能够像人一样使用丰富的云产品。为此，阿里云针对Agent工作负载“无规律弹性、短生命周期、瞬时起量即走”的特点，重构了从训练到推理的云体系。

在模型层，阿里云正式发布千问3.7-Max，相比上一代更强调Agent能力，据介绍，3.7-Max可全自主完成35小时的超长程智能体复杂任务。

在模型服务和应用层，阿里云上线“千问云”官网，主打“为Agent而生”，将模型选型、模型调用、认证配置、用量查询等完整链路能力封装为Skills，便于Agent辅助直接调用。此外，阿里云还同步推出了Skills门户，将购物、出行、支付等常用功能统一封装为可被Agent调用的标准化模块。

前端行业场景各具特色，后端模型生态百花齐放。各行业都有自己的业务流程，每个模型也有不同的能力边界、调用方式和使用成本。刘伟光表示，Agent突破临界点之后可以24小时不间断工作，对AI和云的需求无穷无尽，通过五层全栈的服务升级，阿里云正在建设的是“中国最大的AI工厂”。

阿里云的增长引擎，正在全面切换为以Token为计量单位的AI收入。

跑通Token生意

更值得注意的信号来自资本市场：Agent改变的不只是技术架构，也是云的商业模式。在Agent的市场体系里，能不能产生商业价值与能不能完成任务几乎是同等重要的两件事。

过去，云厂商主要售卖CPU、存储、带宽与虚拟机；如今，越来越多企业开始直接购买模型调用能力、推理能力，以及Agent执行任务的能力。会上，阿里云明确提出，Agent驱动的MaaS收入将取代ECS成为最大的产品线——这意味着阿里云的增长引擎，正在全面切换为以Token为计量单位的AI收入。

阿里云百炼技术负责人于文渊指出，传统模型调用通常只是一问一答，但Agent不同，它需要理解目标、拆解任务、规划路径、调用工具、读取数据、执行任务，并基于任务反馈开展持续推理。这意味着，Agent带来的并不是简单的模型调用增长，而是Token消耗规模的指数级提升。

而厂商需要思考的，不再只是提供API调用服务，更是围绕Agent的真实调用方式，把上下文复用、资源调度、批量推理、预留资源和订阅服务组合起来，让Agent从原始的Token消耗逐步走向成本可控、容量确定、峰值可靠。

李飞飞强调：“高效地产生Token，并将其转化为生产应用的智能和可执行的Action，是Agent Cloud要解决的核心问题。”于文渊也表示：“百炼的定位已经不是模型API服务平台，而是Agent推理服务平台。”

除此之外，阿里云也在进一步强化开放生态，它试图扮演的角色，是AI时代不可替代的基础设施平台。本次峰会上，阿里云宣布MiniMax、智谱、Kimi、阶跃星辰、爱诗科技等多家头部模型厂商即将接入百炼平台。

从财报中对AI重要性的多次强调，到云峰会上全栈适配Agent的大胆迈步，阿里云的行动正在向市场传递一个信号：时代在改变，而企业也需“再造”基因。站在这个关键节点，谁能抢先占据有利身位，谁就能把握住未来。

全国首个人形机器人全生命周期管理服务平台发布

本报讯 5月22日，人形机器人全生命周期管理服务平台工作推进会在北京召开。会上发布了全国首个人形机器人全生命周期管理服务平台，覆盖全国百余家人形机器人企业。

据了解，人形机器人全生命周期管理服务平台是在工业和信息化部科技司指导下，由人形机器人与具身智能标准化技术委员会（以下简称“标委会”）牵头搭建的全国首个人形机器人全生命周期管理服务平台。该平台以构建全链条、全周期、全要素的一体化治理体系为核心使命，以“1个平台、1套生态、3种能力”为架构，打造人形机器人全生命周期管理服务底座，建立覆盖“研发—生产—准入—销售—使用—维护—报废—回收”的全链条管理服务体系，形成“源头可溯、全程可控、风险可防、责任可究”的闭环治理机制，为政府精准监管提供技术抓手，为企业降本增效提供服务支撑，为产业规范有序发展筑牢安全屏障。截至目前，平台已覆盖全国100余家人形机器人企业，完成200余个产品型号、2.8万余台机器人的全生命周期赋码。

开展人形机器人全生命周期管理工作，是统筹人形机器人发展和安全的重要一步。据悉，工业和信息化部科技司将推动场景应用与安全管理协同，启动人形机器人与具身智能实景实训专项行动，强化整机“身份证”在实际场景中的管理应用；推动标准体系与平台建设协同，发挥标委会作用，加快制定全生命周期管理所需的细分标准，为平台运行提供服务保障；推动部门联动与行业治理协同，用好国家人工智能产业创新应用先导区等机制，加强央地联动、部门协同，构建行业治理体系，共同提升系统化、精细化治理水平。

会上，专家解读了《人形机器人全生命周期管理规范》标准，该标准规定了人形机器人整机的身份编码规则，规定了生产、流通、维护、回收的全生命周期管理要求，适用于人形机器人制造商、服务商、销售商、使用者、回收机构等相关方。通过标准牵引，建立符合产业发展需求的全生命周期管理机制，解决安全、管理、治理等核心问题，加速人形机器人应用落地。

大会现场进行了人形机器人全生命周期管理服务平台签约仪式，签约单位包括北京、武汉、成都、宁波等人工智能20城（A20）工作机制成员及30余家人形机器人头部企业，共同推动技术创新与安全规范并行。依托A20工作机制，联合标委会委员单位及中国人形机器人百人会，强化人形机器人全生命周期管理机制推广应用，支持地方提升属地管理能力，夯实协同治理基础，加强行业自律引导，凝聚政企共治合力。（杨鹏岳）

京东工业发布自研垂类大模型“工小智”

本报讯 5月20日，京东工业在北京正式发布面向中小企业的“AI智采管家”——工小智。

现场演示中，用户只需拍摄一张故障电源照片，工小智就能识别规格、工况、价格和货期，直接完成购买；上传一张含15行商品的采购清单，系统能在数秒内匹配出13条精准结果，对剩余两条给出合理推测……

工小智背后是京东工业自研的垂类大模型，在商品审核、同品识别、图纸解析等21个智能体场景中，以8B参数量的模型击败了千亿级通用模型。

中小企业采购之痛，本质是标准化之痛。京东工业副总裁、战略和业务发展负责人丁德明指出，中国工业品有3000多个末级类目，其中80%缺乏统一标准。例如，同一种水泵扬程可能虚标，不同品牌同一型号未必通用，过去靠上百人团队人工修正，现在通过AI Agent交叉验证，效率呈指数级提升。

这种能力直接转化为看得见的收益。据京东工业联合国务院发展研究中心大数据研究院的测算，一家年营收5000万元的设备制造企业，采购成本约占营收的50%~60%，净利润只有5%~10%。若通过数智化供应链降低5%的采购成本，就等于直接创造100万元以上利润，相当于多做2000万元交易。

丁德明表示，京东工业不是一家SaaS软件公司，而是一家AI驱动的供应链技术与服务公司，企业的商业模式可以通过商品交易变现。

“工具免费铺开，沉淀行业数据，迭代模型能力，最终通过闭环履约和商品销售盈利。”丁德明指出，这种模式的优势在于闭环能力。京东工业拥有9000万SKU、1600多个自营仓和完整的物流网络，客户在工小智上完成选型后，可以直接下单、履约、对账，无须跳转。据测算，端到端数智化采购可节省20%~30%成本，交易时间减少30%。

未来，京东工业的AI产品矩阵还将延伸至设计端。京东工业AI与数据智能部总经理王宇辰透露，未来两个月，京东工业将推出“AI选型”和“AI数据透视”等创新产品，真正打通“设计—选型—采购”全链路。（谷 月）

矩阵超智推出MATRIX-3人形机器人

本报讯 5月18日，在上海张江集成创新园举办的“2026矩阵超智科技日（AI DAY）”上，矩阵超智推出全能旗舰级人形机器人MATRIX-3。

据介绍，MATRIX-3实现了多项技术突破。

一是以数据上量实现通用具身人工智能的WAVE物理基座大模型。该基座模型的硬核逻辑在于，机器人真正需要学习的不只是画面内容，更是动作带来的后果。WAVE通过融合大量的自然世界数据，使机器人率先实现“零样本泛化”与“失误学习闭环”的跨越，从“看懂”到“会做”，再到“做得稳”的质变。

二是仿生超能直线关节和运动控制算法。采用串并联直线关节构型设计，以及旋转+直线混合驱动，机体全身自由度33DOF，其中头部4DOF、腰部3DOF、手臂7x2DOF、腿部6x2DOF，执行器推力密度3600N/kg，双臂负载15kg，经历了十万级测试验证，提升负载、可靠性与末端操作精度，确保在长时间、高负荷工业作业下的稳定性。

三是27维自由度灵巧手。经过8000多小时循环测试，可实现拟人化的捏、夹、旋转等微米级精密灵巧操作，适用于从工业精密制造到家庭服务多场景服务。

四是安全“软护甲”3D针织仿生皮肤。首创柔性织物融合躯体的“类人肌体”，具备触觉力、压力、温度感知能力，消除了金属机器人的冷硬感，提升人机交互安全性与亲和力，在商业零售服务与高端制造场景中具备天然的适配感。

活动现场，矩阵超智位于上海张江人形机器人谷的MFH超智工厂正式启用亮相。矩阵超智创始人、CEO张海星表示，目前MFH超智工厂已具备在年内交付5000台产品的能力，并已启动针对真实工业场景的深度压力测试。随着产线的持续升级，预计在2027年实现10万台级别的量产能力，以规模效应加速通用劳动力的普及。

据悉，MATRIX-3全能旗舰级人形机器人售价58万元起，MATRIX-3 PRO售价68万元起，均包含1年基础服务包。（赵 晨）