申万宏源研究 SWS RESEARCH

2026 年 08 月 15 日

关注华为超节点进展！国产大模型多更新！

——计算机行业周报 20260810-20260815

看好

相关研究

《从美股软件行情看 FDE！——计算机行业周报 20260803-20260807》 2026/08/08

《DeepSeek-V4-Flash 继续展现国产模型超高性价比！——计算机行业周报 20260727-20260731》 2026/08/01

证券分析师
黄忠煌 A0230519110001
huangzh@swsresearch.com
洪依真 A0230519060003
hongyz@swsresearch.com
刘洋 A0230513050006
liuyang2@swsresearch.com
研究支持
崔航 A0230524080005
cuihang@swsresearch.com
曹峥 A0230525040002
caozheng@swsresearch.com
陈晴华 A0230525100001
chenqh@swsresearch.com
联系人
王开元 A0230125030001
wangky@swsresearch.com

申万宏源研究微信服务号

本期投资提示：

- **本周周报重点：1）华为超节点；2）智谱、DeepSeek 等大模型发布。**
- **当前国产算力确定性最高供应商之一——华为，超节点能力成熟。华为依托全栈自研打造 Atlas 950 SuperPoD 超节点，构建国产 AI 算力全新范式。**硬件端推出差异化 Ascend 950PR、950DT 芯片，分别适配推理与训练场景，搭配自研灵衢 UB 统一总线搭建三级全 Mesh 互联体系，单柜算力、互联带宽对标英伟达 GB200 NVL72，1024 卡集群采用 800G 光模块实现低时延跨柜互联。产品于 2026 WAIC 真机亮相，计划四季度上市，上一代 Atlas 900 超节点已商用 750 余套，目前已获科大讯飞、美团等头部企业深度适配验证，覆盖多行业大模型训练与高并发推理场景。
- **智谱发布 GLM 5.3：后训练 Scaling 驱动代码及 Agent 能力再跃升。**本次提升官方明确表示来自后训练 Scaling，在复杂编程能力、长程任务、网络安全方面表现强劲。**在 Coding 等任务可执行、结果可验证的场景中，经过充分训后的较小参数模型能够以更低延迟和成本接近甚至超过更大参数的通用模型，**但在跨代码库理解、复杂系统重构和长程规划等高难度任务中，预训练规模仍影响模型能力上限。
- **DeepSeek 发布 V4 Pro 正式版。**V4 Pro 的 Agent 成绩至少包含四层共同作用：模型后训练决定规划与工具调用倾向；Harness 决定上下文组织和循环；推理预算决定思考长度；工具与沙箱决定可执行边界。**上述解释了官方跑分与部分外部榜单可能出现落差**：外部评测若使用不同的 Agent Loop、上下文管理、失败重试、工具 schema 或最大交互步数，测到的是另一套系统。
- **重点推荐主线：**1）数字经济领军；2）AIGC 应用；3）AIGC 算力；4）数据要素；5）信创弹性；6）港股核心；7）智联汽车；8）新型工业化；9）医疗信息化。详细标的请见正文。
- **风险提示：行业若激进扩人策略，可能影响盈利增长。外部因素影响供应链稳定性。**

1. 华为：全栈自研，国内超节点领跑者

当前国产算力确定性最高供应商之一——华为。超节点是智算核心载体，华为具备全栈自研完整优势值得重点关注。公司新一代 Atlas 950 超节点搭载差异化昇腾 950 系列芯片，自研灵衢统一总线搭建三级高速互联，单机柜算力、互联能力对标海外头部产品。

昇腾芯片是华为 AI 算力战略的基础。华为坚持全栈自研，昇腾芯片持续迭代，并与超节点架构协同演进，构成华为 AI 算力战略的核心技术主线。华为自 2018 年发布 Ascend 310 芯片、2019 年发布 Ascend 910 芯片以来，持续推进昇腾 AI 芯片迭代；2025 年，Ascend 910C 芯片伴随 Atlas 900 超节点规模部署进入广泛应用；2026 年，进一步推出 Ascend 950 系列，包括 Ascend 950PR 和 Ascend 950DT，基于新一代芯片打造的 Atlas 950 超节点也已正式亮相。展望未来，至 2028 年，华为还规划推出 Ascend 960 和 Ascend 970 系列，持续提升 AI 芯片的算力、内存带宽与互联能力。

最新一代 Ascend 950 系列进一步体现了华为从 AI 芯片、存储、互联到超节点及基础软件的全栈技术能力。Ascend 950 系列包括 950PR 和 950DT 两款芯片，两者采用相同的 Ascend 950 Die，但针对大模型不同计算阶段进行差异化设计。

Ascend 950PR 主要面向推理 Prefill 阶段和推荐业务场景，采用了华为自研的低成本 HBM，HiBL 1.0，相比高性能、高价格的 HBM3e/4e，能够大大降低推理 Prefill 阶段和推荐业务的投资，最大支持 128GB 片上内存和 1.6TB/s 内存带宽。**Ascend 950DT 则重点面向推理 Decode 和模型训练场景，进一步强化内存及互联能力。**芯片内存容量达到 144GB，内存带宽达到 4TB/s。

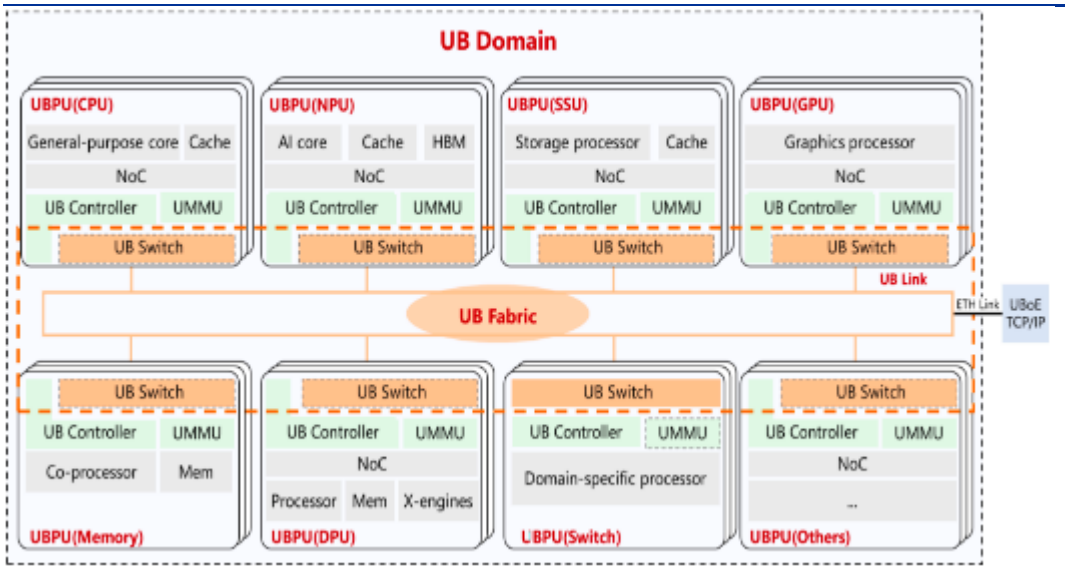
表 1：华为芯片参数情况

指标	Ascend 910C	Ascend 950PR	Ascend 950DT	Ascend 960	Ascend 970
发布时间	2025 Q1	2026 Q1	2026 Q4	2027 Q4	2028 Q4
微架构	SIMD	SIMD/SIMT	SIMD/SIMT	SIMD/SIMT	SIMD/SIMT
数值类型	FP32/HF32/FP16/BF16/INT8	FP32/HF32/FP16/BF16/FP8/MXFP8/HiF8/MXFP4	FP32/HF32/FP16/BF16/FP8/MXFP8/HiF8/MXFP4	FP32/HF32/FP16/BF16/FP8/MXFP8 HiF8/MXFP4/HiF4	FP32/HF32/FP16/BF16/FP8/MXFP8/ HiF8/MXFP4/HiF4
互联带宽	784 GB/s	2 TB/s	2 TB/s	2.2 TB/s	4 TB/s
算力	800 TFLOPS FP16	1 PFLOPS FP8, 2 PFLOPS FP4	1 PFLOPS FP8, 2 PFLOPS FP4	2 PFLOPS FP8, 4 PFLOPS FP4	4 PFLOPS FP8, 8 PFLOPS FP4
相当于英伟达	H100	H100	H200	B100	B 系列
内存	128 GB, 3.2 TB/s	128 GB, 1.6 TB/s	144 GB, 4 TB/s	288 GB, 9.6 TB/s	288 GB, 14.4 TB/s

资料来源：快科技，申万宏源研究

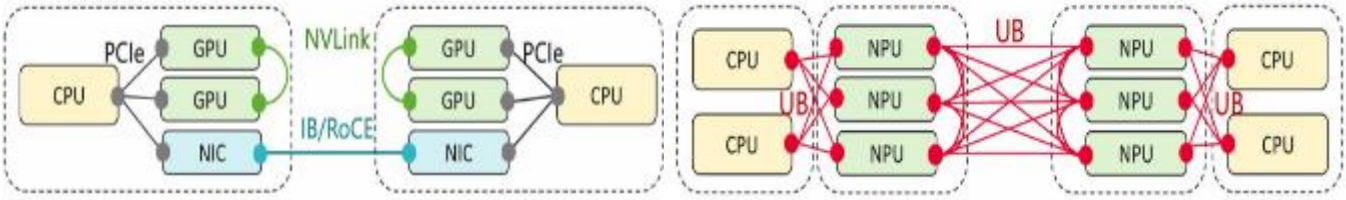
在昇腾芯片持续迭代的基础上，华为进一步通过自研灵衢（UnifiedBus，UB）总线构建超节点高速互联能力，成为构建超节点高速互联的关键技术。灵衢是华为自研的统一互联协议，通过统一的总线架构连接超节点内部不同类型、不同距离的计算、存储与网络组件，包括 CPU、NPU、GPU、MEM、DPU、SSU 及 Switch 等，实现异构资源之间的统一互联。AI 服务器及集群通常需要 PCIe、NVLink 以及 IB/RoCE 等多套互联协议分别承担芯片内、服务器内及跨节点通信，**而华为通过一套 UB 体系覆盖上述不同层级的互联需求，从而简化超节点内部互联架构，并为算力、内存及其他资源的统一协同提供底层支撑。**

图 1 华为 UB 构成



资料来源：灵衢官网，申万宏源研究

图 2：灵衢统一总线 UB 实现统一互联



资料来源：论文《UB-Mesh: a Hierarchically Localized nD-FullMesh Datacenter Network Architecture》，申万宏源研究

凭借昇腾 950DT 芯片与灵衢 UB 协同，Atlas 950 单柜的 Scale-up 互联能力已与 NVIDIA GB200 NVL72 达到同一量级。Atlas 950 单柜搭载 64 颗 Ascend 950DT NPU 和 16 颗鲲鹏 950 CPU，mxFP4、mxFP8/HiF8 及 FP16/BF16 算力分别达到 114.1、58.8 和 31.1 PFLOPS，并配置约 6.1TB NPU 片上内存，单颗 950DT 峰值内存带宽达 4.0TB/s。互联方面， Atlas 950 单柜最高支持 64×1.68TB/s 双向 UB Link，GB200 NVL72 的 NVLink 整柜带宽为 130TB/s，均达到百 TB/s 量级，体现出华为通过“昇腾芯片+高速互联”强化系统级算力的技术能力。

表 2：Atlas 950 SuperPoD 单柜与 NVIDIA GB200 NVL72 单机柜对比

指标	华为 Atlas 950 SuperPoD 单柜	NVIDIA GB200 NVL72 单机柜

机柜规格	44OU，约 47U	Rack-scale，单机柜
AI 芯片	64 × Ascend 950DT NPU	72 × Blackwell GPU
CPU	16 × 鲲鹏 950 CPU，单颗最高 96 核、2.3GHz	36 × Grace CPU
AI 算力	114.1 PFLOPS（mxFP4）；58.8 PFLOPS（mxFP8 / HiF8）；31.1 PFLOPS（FP16 / BF16）	1,440 / 720 PFLOPS（NVFP4）；720 PFLOPS（FP8/FP6）；360 PFLOPS（FP16/BF16）
高带宽内存（HBM）	总容量：64 × 96 GB；单颗峰值内存带宽：4.0 TB/s	13.4 TB HBM3E 带宽 576 TB/s
CPU 内存	最多支持 192 个内存插槽，内存速率最高 6400 MT/s；单条 DDR5 内存最高支持 64 GB	17 TB LPDDR5X，带宽 14 TB/s
本地存储	最多支持 16 块 2.5 英寸 NVMe SSD + 16 块 2.5 英寸 SATA SSD；或 32 块 2.5 英寸 NVMe SSD	单计算托盘：4 × 3.84 TB E1.S NVMe（RAID 0）+ 1 × 1.92 TB M.2 NVMe 系统盘；共 18 个计算托盘
Scale-up 互联	最大 64 × 1.68TB/s 双向 UB Link	NVLink：130 TB/s
机柜功耗	约 100kW	约 120kW

资料来源：华为官网，英伟达官网，申万宏源研究

2. Atlas 950 SuperPoD 构建了从节点内、单柜内到跨机柜的三级互联体系

Atlas 950 SuperPoD 构建了从节点内、单柜内到跨机柜的三级互联体系，通过分层扩展将 8 卡、64 卡进一步扩展至 1024 卡乃至 8192 卡 Scale-up 域。华为官方将其互联拓扑划分为 X、Y、Z 三个层级，分别对应单个 Compute Tray 8 卡间 1D-FullMesh、单柜 64 卡 2D-FullMesh 以及跨柜 Clos。

表 3：Atlas 950 8 卡、64 卡、1024 卡、8192 卡的互联方式

卡数	物理层级	拓扑	互联方式	互联协议
8	一个计算板 Tray	1D Full-Mesh 全电互联	板上直接相连	UB-Mesh
64	一台 64 卡 Rack	2D Full-Mesh 全电互联	柜内 8 个 Tray 通过 4 个 25.6T 交换板全互联	UB-Mesh
1024	一个 1024 卡超节点	2D Full-Mesh（电）+一层 Clos（光）	16 台 64 卡 Rack 通过 4 台 UB 交换柜互联	UB
8192	一个 8192 卡超节点	2D Full-Mesh（电）+一层 Clos（光）	128 台 64 卡 Rack 通过 128 台 100T UB 交换柜互联	UB/UBoE

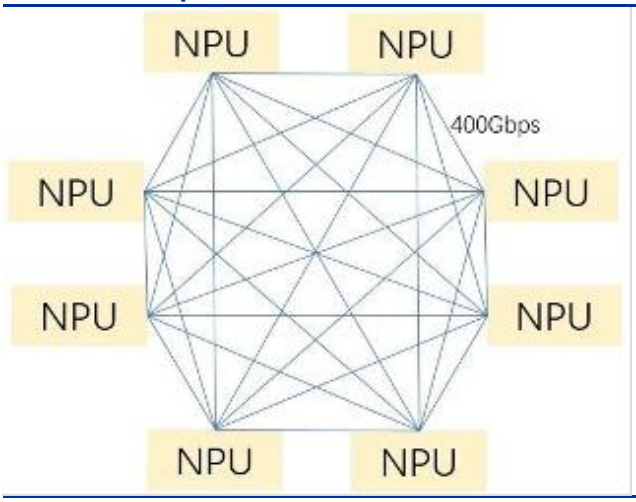
资料来源：《华为：基于 UB 的超节点参考架构白皮书》，申万宏源研究

节点内：8 颗 NPU 通过 HCCS 组成 Mesh 全互联，实现节点内部的高速直连。每个计算单元包含 8 颗 Ascend 950DT，8 颗 NPU 之间通过 HCCS 形成 Full-Mesh，构成整个互联体系的 X 轴；每颗 NPU 均可与其余 7 颗 NPU 直接通信，从而降低节点内数据交换时延。**若按照每颗 NPU 与其他 7 颗 NPU 之间按照 400Gbps 粒度互联，则单 NPU 节点内 Mesh 互联总带宽为：7 × 4 × 112Gbps = 3.136Tbps。**

转发机制：NPU 通过 IO DIE 内置的 UB On-Chip Switch 完成跨芯片数据转发。数据进入NPU端口后，可根据路由表判断目标位置；若目标并非本芯片，流量可直接在IO DIE

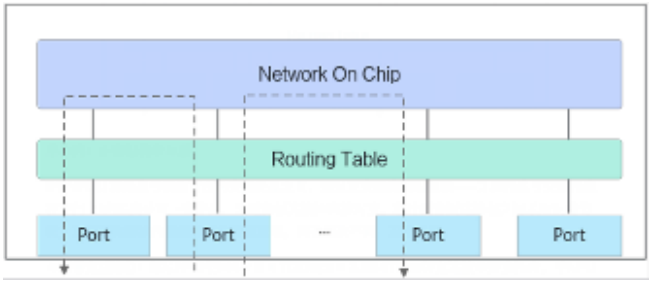
内部转发至对应输出端口，无需进入计算 DIE，也不会占用 DRAM 访问带宽。通过将部分交换功能集成至芯片 IO 侧，可减少数据转发对计算和存储资源的占用。

图 3：8 卡互联方式：每颗 NPU 与其他 7 颗 NPU 之间实现 400Gbps 互联



资料来源：智联网络架构师（公众号），申万宏源研究

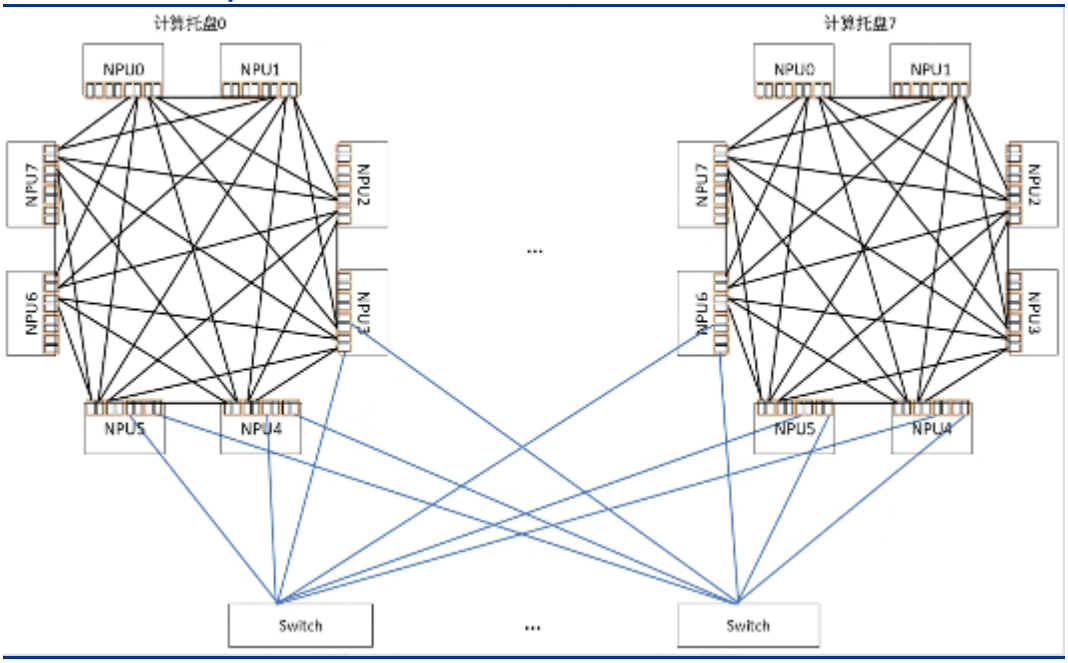
图 4：板内任意 2 颗 NPU 之间均可提供 3.136Tbps 的端到端互联带宽



资料来源：昇腾 950 NPU 架构白皮书，申万宏源研究

64 卡机柜：8 个 NPU 计算托盘通过正交架构实现跨板高速互联。计算托盘与机柜后侧交换托盘呈垂直正交排列，并通过背板连接器直接对接，芯片至背板之间采用铜缆传输，在节点内 8 卡 Mesh 的基础上进一步完成不同计算托盘之间的互联。按每颗 NPU 配置 4 组、每组 8×112Gbps 链路测算，连接交换托盘的聚合带宽约为 **4×8×112Gbps=3.584Tbps**；结合节点内 3.136Tbps 的聚合链路，单颗 NPU 双向互联总带宽测算约为 **2×（3.136+3.584）/8=1.68TB/s。**

图 5 Atlas 950 SuperPoD 单柜 64 卡互联架构

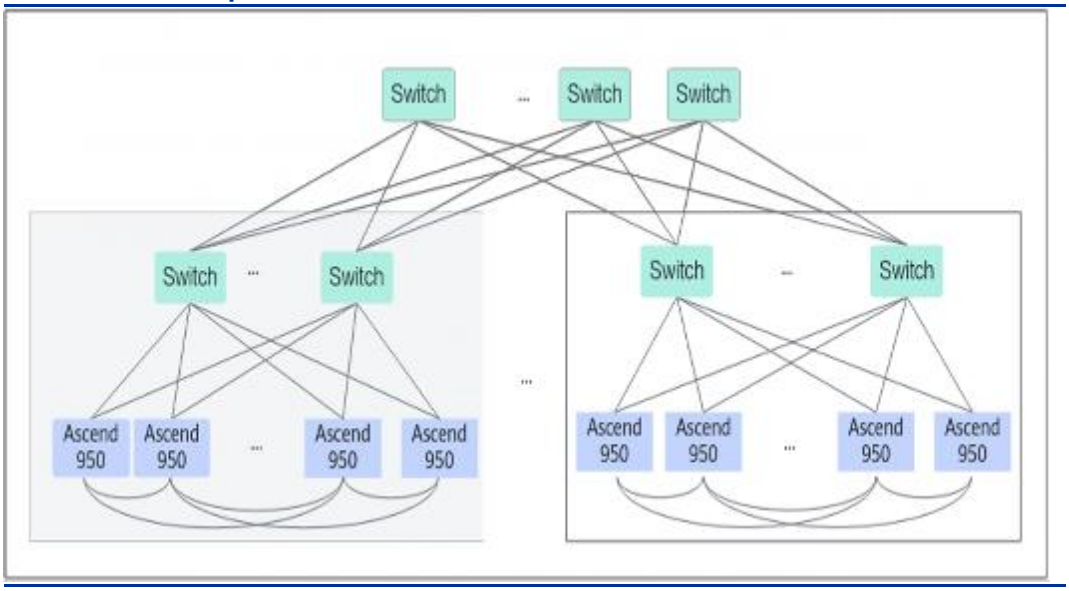


资料来源：智联网络架构师（公众号），申万宏源研究

柜间：系统由电互联切换至光互联，并通过中央 UB 交换柜构建全局 Clos 网络。1024 卡系统由 16 个计算柜和 4 个灵衢交换柜组成，多个计算柜通过中央 UB 交换柜实现跨柜连接，从而将单柜内部的高速互联继续扩展至整个千卡集群。

柜内与柜间采用约 2:1 的带宽配比。按每个灵衢交换托盘配置 32 个 800G 级端口、每端口采用 8×112Gbps 链路测算，4 个交换托盘对应的柜间聚合链路带宽约为 **4×32×8×112Gbps**；单柜 64 颗 NPU 连接跨计算托盘网络的聚合链路带宽约为 **64×4×8×112Gbps**，由此测算柜内跨托盘互联带宽约为柜间互联带宽的 2 倍。

图 6 Atlas 950 SuperPoD 1024 卡互联架构

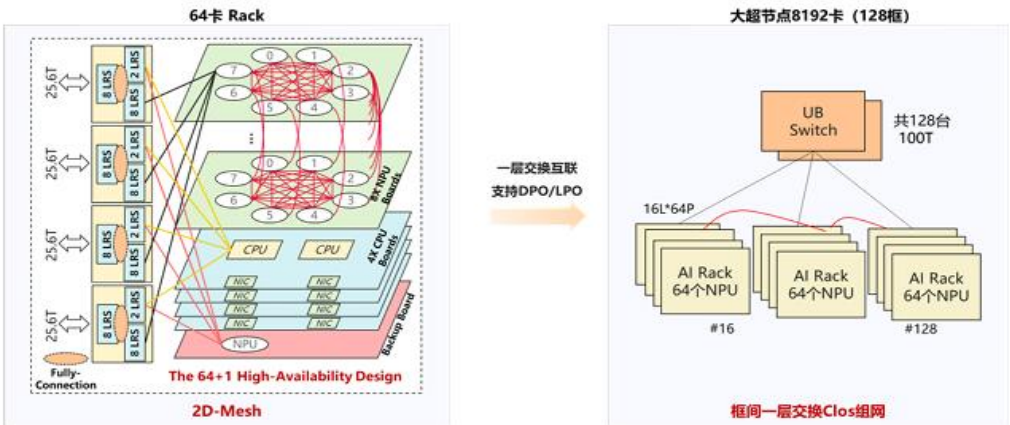


资料来源：昇腾 950 NPU 架构白皮书，申万宏源研究

光模块：1024 卡系统通过大规模 800G 光模块完成计算柜与 UB 交换柜之间的光互联。1024 卡集群共使用约 4096 个 800G 光模块，包含计算柜侧与 UB 交换柜侧；柜间连接采用 LPO（线性驱动可插拔光模块）方案，基于华为光电协同设计进一步优化系统成本、时延与功耗。

与 1024 卡相同，Rack 内采用 2D-FullMesh 组网， Rack 间采用一层 UB Switch 互连，支持从 64 卡线性扩展到 8192 卡。

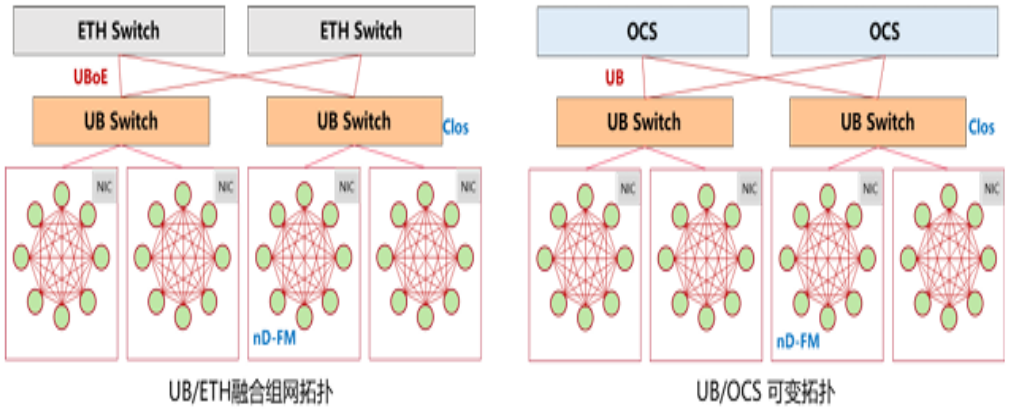
图 7 Atlas950 SuperPoD 8192 卡互联结构



资料来源：《华为：基于 UB 的超节点参考架构白皮书》，申万宏源研究

为了进一步扩大组网规模，UB 除了支持采用多级 UB Switch 扩展组网之外，还支持通过 UBoE 与以太 Switch 对接，实现融合组网，以及通过 OCS 组网，实现可变拓扑，匹配业务动态流量。

图 8 UB 融合组网和光交换组网示意



资料来源：《华为：基于 UB 的超节点参考架构白皮书》，申万宏源研究

3. 超节点加速场景落地，950 生态获头部客户验证

Atlas 950 SuperPoD 主要面向大型 AI 数据中心，重点满足超大规模模型训练和中心侧高并发推理需求，并可进一步应用于互联网、运营商、金融、科研及智算中心等高算力场景。截至 WAIC 2026，昇腾已展示 20 多个行业标杆落地案例、覆盖 60 多种真实业务场景；同期，上一代 Atlas 900A3（昇腾 384 超节点）已商用部署 750 余套。Atlas 950 SuperPoD 于本届 WAIC 首次公开真机亮相，并计划于 2026 年第四季度上市。

图 9 2026WAIC 昇腾 950 超节点真机首次亮相

资料来源：华为官网，申万宏源研究

新一代 950 生态亦开始获得头部客户验证：

科大讯飞已经与华为团队针对 950 芯片进行深度对接，在昇腾 950 平台上联合攻坚更高效模型结构、混合 Attention 机制、智能体强化学习等关键技术。预计在 2026 年 1024 开发者节期间，在昇腾 950 平台上发布中国首个对标业界最先进主流模型的旗舰大模型。

美团 LongCat-2.0 已与昇腾开展深度联合优化。围绕 Attention/MoE 模型结构、芯片算子与缓存机制以及 PD 分离部署进行协同适配，面向万亿参数 MoE 模型实现百万级长文本推理，显示昇腾已开始进入头部互联网厂商大模型推理的核心业务场景。

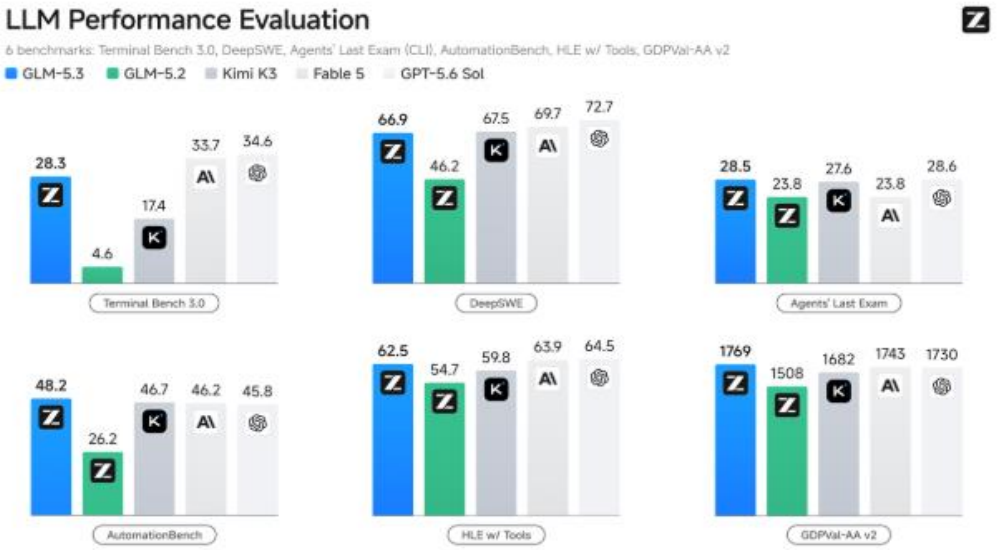
4. 智谱发布 GLM 5.3：后训练 Scaling 驱动代码及 Agent 能力再跃升

2026 年 8 月 14 日智谱正式发布 GLM 5.3。本次提升官方明确表示来自后训练 Scaling：数十倍的长程任务环境、更丰富多样的环境类型、超长的后训练时间。在复杂编程能力、长程任务、网络安全方面表现强劲。

模型保持 1M 上下文和 128K 最大输出，当前仅支持处理文本模态信息，支持多档思考、流式输出、函数调用、上下文缓存、结构化输出和 MCP。发布当日 GLM 5.3 已面向全部 GLM Coding Plan 用户开放，API 显示即将上线。

能力层面编码和 Agent 显著提升，编程体感相比 GLM 5.2 提升约 50%，部分场景超过 Kimi K3 和 Claude Fable 5：在衡量模型真实终端复杂任务执行能力的 Terminal-Bench 3.0 中，得分从 4.6 提升至 28.3（K3 为 17.4）；在聚焦长程软件工程任务的 DeepSWE 上，从 46.2 提升至 66.9（K3 为 67.5）；在 Agents' Last Exam 上，从 23.8 提升至 28.5（K3 为 27.6）。在覆盖 44 种职业、考察真实高价值知识工作的 GDPval-AA v2 中，GLM-5.3 得分达到 1769（K3 为 1682），展现出伴随编程能力提升而出现的更强专业任务执行能力。在衡量 AI Agent 真实软件漏洞分析与复现能力的 CyberGym 上，GLM-5.3 得分 84.5，高居全球第一（Mythos 5 为 83.8），体现出 Coding 能力向网络安全等复杂专业任务的进一步延伸。

图 10 GLM 5.3 各项 benchmark 得分情况



资料来源：智谱公众号，申万宏源研究

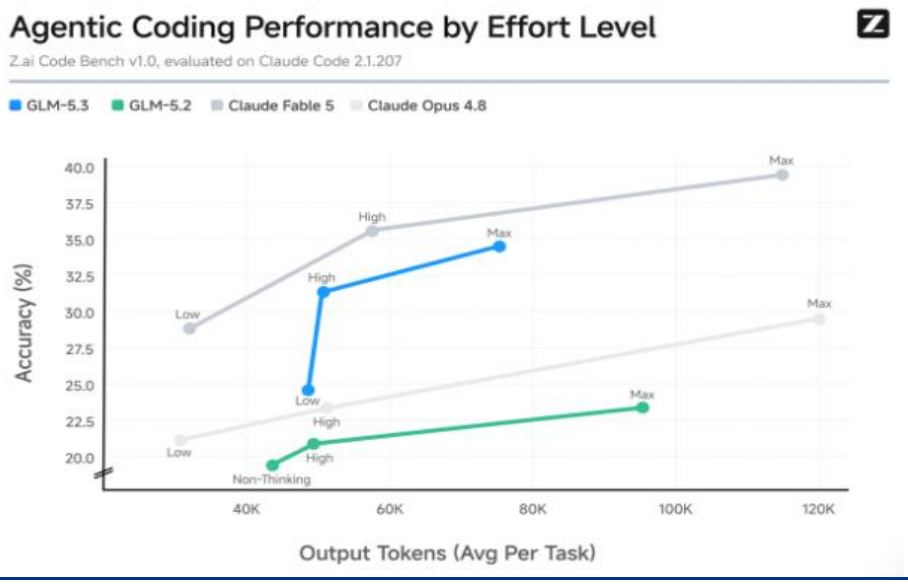
同一基座实现能力跃升，后训练 Scaling 成为重要增量路径。GLM 5.3 的训练不再局限于独立编程题，而是覆盖问题识别、方案分析、代码实现、运行验证和最终交付的完整流程。部分训练任务相当于资深工程师数日的工作量，模型需要在真实计算机群、存储系统、内部文档和代码库中完成任务。与传统 SFT 或单轮偏好优化相比，此类训练更接近企业真实生产环境，也更容易暴露长任务中的目标漂移、工具调用错误和局部最优问题。

我们认为 GLM 5.3 与 GLM 5.2 共用同一基础模型，能力提升全部来自后训练，**说明在基座不变的情况下，扩大真实任务环境、可验证奖励和长程强化学习规模，仍能显著提高模型能力的兑现效率。**预计未来随着基础模型预训练参数扩大的情况下，叠加公司自身强后训练能力，有望实现更强大的模型智能水平。

并且，GLM 5.3 在能力和效率之间取得较好平衡。在智谱推出的 Z.ai Code Bench（用于在贴近真实用户场景的条件下评估编程智能体）中，GLM 5.3 在能力表现和 token 效率上取得提升，在 Max 档位下 GLM 5.3 准确率达 34.5%，单任务平均输出约 7.5 万 token，GLM 5.2 的准确率为 23.4%，单任务平均输出约 9.6 万 token；而在 High 档位下，GLM 5.3 准确率达 31.4%，单任务平均输出约 5 万 token，Claude Opus 4.8 准确率 29.5%，单任务平均输出约 12 万 token。

上述情况同样说明，高质量后训练能够减少无效推理、重复尝试和错误工具调用。**在 Coding 等任务可执行、结果可验证的场景中，经过充分训后的较小参数模型能够以更低延迟和成本接近甚至超过更大参数的通用模型，**但在跨代码库理解、复杂系统重构和长程规划等高难度任务中，预训练规模仍影响模型能力上限。

图 11 Z.ai Code Bench 测评



资料来源：智谱公众号，申万宏源研究

5. DeepSeek 发布 V4 Pro 正式版

表 4：复盘模型、产品、价格同步升级

时间	事件	官方事实	研究含义
4 月 24 日	V4 预览版	Pro 1.6T/49B 激活、Flash 284B/13B 激活；1M 上下文；开放权重。	建立长上下文和 Agent 能力底座。
7 月 31 日	V4 Flash 0731	结构与规模基本不变，重点重做后训练；原生 Responses API、适配 Codex。	卖方关注后训练效率、低成本长程 Agent 与应用渗透。
8 月 13 日	V4 Pro 0813 正式版	APP、网页与 API 更新；模型名不变；思考强度扩展为 low/high/max；Agent 跑分显著提高。	旗舰模型完成从 Preview 到 GA，能力增量明显偏执行。
8 月 13 日	Harness 开源	Developer Preview；MIT；基于 Cordis；Everything is a Plugin。	DeepSeek 开始争夺模型之上的开发者入口和运行时标准。
8 月 17 日	API 峰谷价生效	高峰 9:00–12:00、14:00–18:00；其余为空闲时段，空闲价为高峰一半。	以价格做资源调度，同时提高高价值 Agent 调用的商业化单价。

资料来源：公司官网，申万宏源研究

一、V4 Pro 0813：提升最明显的是动手能力。

官方基准中，V4 Pro 0813 相比 Pro Preview 的提升集中于终端、代码仓库、网络安全、长周期软件工程、工具使用和 GUI 自动化。DeepSWE 从 12.8 提升至 62.7，CyberGym 从 52.7 提升至 83.3，Terminal-Bench 2.1 从 72.1 提升至 87.9，AutomationBench 从 12.8 提升至 31.8。方向上，**结论：结构未必需要大改，后训练、工具协议和执行框架可以显著改变 Agent 任务表现。**

图 12 DeepSeek Benchmark 情况

Benchmarks	DeepSeek-V4-Pro-0813	DeepSeek-V4-Flash-0731	DeepSeek-V4-Pro-Preview	DeepSeek-V4-Flash-Preview	GLM-5.2	Kimi-K3	Opus-4.8	Fable 5 (w/ fallback)
HLE (wo/w tools)	42.7/60.0	37.8/51.5	37.7/48.2	34.8/45.1	40.5/54.7	43.5/56.0	49.8/57.9	53.3/63.0
Terminal Bench 2.1	87.9	82.7	72.1	61.8	81.0	88.3	85.0	88.0
NL2Repo	61.5	54.2	38.5	39.4	48.9	-	69.7	-
Cybergym	83.3	76.7	52.7	38.7	-	80.0	78.3	83.1
DeepSWE	62.7	54.4	12.8	7.3	46.2	67.5	58.0	70.0
Toolathlon-Verified	74.1	70.3	55.9	49.7	59.9	76.5	76.2	77.9
Agents' Last Exam	25.7	25.2	16.5	15.8	23.8	27.6	25.7	-
AutomationBench (Public)	31.8	25.1	12.8	10.8	12.9	30.8	27.2	29.1
DSBench-FullStack	71.1	68.7	41.8	37.0	61.8	73.7	71.6	77.2
DSBench-Hard	67.2	59.6	31.1	25.8	54.5	63.0	71.7	68.3

资料来源：公司官网，申万宏源研究

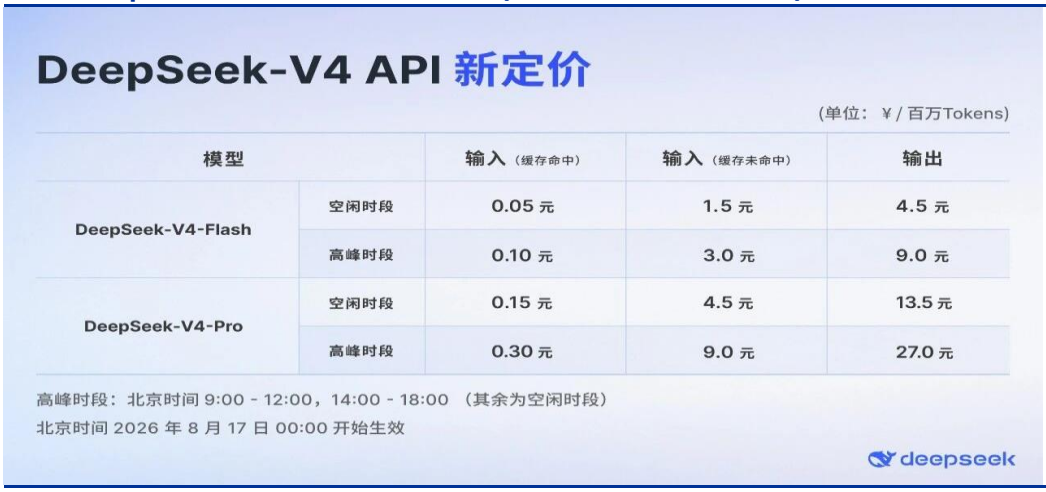
跑分拆成“模型 × Harness × 预算 × 工具环境”：

DeepSeek 明确披露，公开 Code Agent 任务使用 DeepSeek Harness 极简模式，并采用 max 思考强度、top_p=0.95、temperature=1.0。极简模式仅保留 Shell 与文件编辑等核心工具，目的是减少产品层噪声、方便基准测试。由此，**V4 Pro 的 Agent 成绩至少包含四层共同作用：模型后训练决定规划与工具调用倾向；Harness 决定上下文组织和循环；推理预算决定思考长度；工具与沙箱决定可执行边界。**

这也解释了官方跑分与部分外部榜单可能出现落差：外部评测若使用不同的 Agent Loop、上下文管理、失败重试、工具 schema 或最大交互步数，测到的是另一套系统。

二、 API 提价从普惠价格转向能力分层+峰谷调度

图 13 DeepSeek V4 API 人民币新定价（2026 年 8 月 17 日生效）



资料来源：公司官网，申万宏源研究

名义涨幅很大，但实际账单取决于工作流结构。

假设一个 Agent 任务消耗 100 万输入 tokens，其中 75% 命中缓存，并产生 20 万输出 tokens，则旧价成本约为 1.97 元，空闲时段约为 3.94 元，高峰约为 7.88 元，对应约 2

倍和 4 倍。若只看输入且缓存命中率为 75%，混合输入成本由 0.77 元升至空闲 1.24 元、高峰 2.48 元；输出越长、重试越多，账单越受输出价格上调影响。

三个判断：商业化、资源配置、需求信号要分开。

商业化：在能力接近顶级模型后提高 API 单价，说明 DeepSeek 开始从低价获客转向分层变现。性价比仍可能存在，但需要用成功任务成本重新比较。

资源配置：峰谷价把批处理、数据清洗、离线评测等可延迟负载引导到空闲时段，也有利于释放高峰并发。缓存命中仍显著低于未命中输入，说明高复用工作流仍得到相对激励。

需求信号：涨价与峰谷机制可能说明供需约束和高峰资源紧张。

6. 风险偏好判断以及重点标的

6.1 数字经济领军

海康威视+金山办公+恒生电子+中控技术+德赛西威+启明星辰+科大讯飞+华大九天（计算机+电子）+同花顺+广联达+新大陆

6.2 AIGC 应用

金山办公+鼎捷数智+税友股份+同花顺+东方财富+虹软科技+万兴科技+润达医疗（申万医药+计算机）+科大讯飞+云从科技+福昕软件+萤石网络+道通科技

6.3 AIGC 算力

浪潮信息+海光信息+神州数码+中科曙光+寒武纪

6.4 数据要素

税友股份+博思软件+广电运通

6.5 信创弹性

海光信息+太极股份+纳思达+软通动力+索辰科技+达梦数据

6.6 港股核心

中国软件国际+金蝶国际+中国民航信息网络+联想集团

6.7 智联汽车

德赛西威+虹软科技+中科创达+豪恩汽电+经纬恒润

6.8 新型工业化

鼎捷数智+思看科技+中控技术+赛意信息+能科科技+用友网络+柏楚电子（申万机械）

6.9 医疗信息化

润达医疗（医药+计算机）+嘉和美康+创业慧康

7. 风险提示

行业若采取激进扩人策略，存在影响盈利增长的风险。如果 2026 年企业主依然选择长期激进扩张，会影响行业年度盈利和增长。

全球外部因素影响供应链稳定，存在业绩不及预期的风险。如果外部物流环境长期不佳，或影响短期季度的同比增速，递延翻转结论。

8. 计算机重点公司估值表

表 5：计算机重点公司估值表

股票代码	股票简称	2026/8/14	归母净利润（亿元）				PE			
		总市值（亿元）	2025A	2026E	2027E	2028E	2025A	2026E	2027E	2028E
688041.SH	海光信息	6,531	25.4	44.5	67.3	96.2	257	147	97	68
688256.SH	寒武纪-U	6,867	20.6	56.1	116.7	192.7	333	122	59	36
300059.SZ	东方财富	3,081	120.8	146.7	163.8	178.8	25	21	19	17
002415.SZ	海康威视	3,197	142.0	172.6	199.9	227.8	23	19	16	14
300033.SZ	同花顺	1,733	32.1	43.0	51.2	59.0	54	40	34	29
603019.SH	中科曙光	1,307	21.8	28.8	36.5	46.4	60	45	36	28
688111.SH	金山办公	1,190	18.4	25.7	26.3	31.4	65	46	45	38
002230.SZ	科大讯飞	1,004	8.4	12.6	16.9	20.5	120	80	60	49
0992.HK	联想集团	3,605	132	149	180	251	27	24	20	14
603296.SH	华勤技术	1,284	40.5	52.5	65.3	80.2	32	24	20	16
000977.SZ	浪潮信息	1,157	24.1	38.9	54.2	72.1	48	30	21	16
002920.SZ	德赛西威	541	24.5	28.6	34.0	40.3	22	19	16	13
688777.SH	中控技术	753	4.4	7.5	10.3	13.5	171	100	73	56
600570.SH	恒生电子	419	12.3	14.1	16.7	20.0	34	30	25	21
301269.SZ	华大九天	525	0.6	1.3	1.7	2.0	862	399	303	258
600588.SH	用友网络	398	-13.9	-0.9	2.6	5.1	-	-	156	78

301236.SZ	软通动力	404	2.1	3.5	4.8	6.3	196	116	84	64
688188.SH	柏楚电子	417	11.1	13.9	17.5	22.2	37	30	24	19
002152.SZ	广电运通	248	8.6	9.7	11.7	13.8	29	26	21	18
0268.HK	金蝶国际	239	1	4	7	11	257	57	32	22
300496.SZ	中科创达	276	4.5	6.3	7.9	9.6	61	44	35	29
0696.HK	中国民航信息网络	214	23	25	27	30	9	8	8	7
000034.SZ	神州数码	254	5.2	8.4	10.9	14.4	49	30	23	18
002180.SZ	奔图科技	256	-7.2	6.7	9.7	13.3	-	38	26	19
688692.SH	达梦数据	276	5.2	6.6	8.5	10.7	53	42	32	26
688208.SH	道通科技	178	9.4	11.6	14.8	17.5	19	15	12	10
688475.SH	萤石网络	235	5.7	7.3	8.7	10.1	41	32	27	23
603171.SH	税友股份	186	1.4	2.5	3.5	4.6	137	75	53	40
000997.SZ	新大陆	180	10.1	11.7	14.9	18.2	18	15	12	10
002410.SZ	广联达	151	4.1	5.8	7.0	8.0	37	26	22	19
002439.SZ	启明星辰	169	-5.7	0.2	1.2	2.3	-	926	146	75
688088.SH	虹软科技	134	2.6	3.2	4.1	5.3	52	42	33	25
688326.SH	经纬恒润-W	80	1.0	2.6	4.5	6.5	80	31	18	12
688327.SH	云从科技-UW	111	-5.6	-2.8	-0.5	0.7	-	-	-	154
300624.SZ	万兴科技	106	-0.9	0.2	0.8	1.1	-	450	138	96
002368.SZ	太极股份	89	-7.6	-0.3	1.5	3.4	-	-	58	26
301488.SZ	豪恩汽电	105	0.7	1.1	1.4	1.7	140	96	77	62
688583.SH	思看科技	115	1.0	1.5	2.3	2.9	120	74	50	39
300378.SZ	鼎捷数智	98	1.6	2.1	2.6	3.1	60	46	38	32
603859.SH	能科科技	111	2.3	2.8	3.5	4.3	49	40	32	26
0354.HK	中国软件国际	82	3	6	7	9	26	14	11	9
300525.SZ	博思软件	66	1.7	2.2	2.5	-	38	30	27	-
603108.SH	润达医疗	63	-5.5	-	-	-	-	-	-	-
688507.SH	索辰科技	154	0.3	0.8	1.0	1.1	490	202	153	138
300687.SZ	赛意信息	121	-1.1	1.2	1.8	2.2	-	102	67	56
300451.SZ	创业慧康	56	-4.0	-1.4	-0.1	1.0	-	-	-	59
688095.SH	福昕软件	59	0.3	1.2	1.8	2.6	196	50	34	23
688246.SH	嘉和美康	26	-2.5	-0.7	0.0	0.5	-	-	630	50

资料来源：Wind，申万宏源研究；注：盈利预测取 Wind 一致预期，港股数据采用人民币

信息披露

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

与公司有关的信息披露

本公司隶属于申万宏源证券有限公司。本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司关联机构在法律许可情况下可能持有或交易本报告提到的投资标的，还可能为或争取为这些标的提供投资银行服务。本公司在知晓范围内依法合规地履行披露义务。客户可通过 compliance@swsresearch.com 索取有关披露资料或登录 www.swsresearch.com 信息披露栏目查询从业人员资质情况、静默期安排及其他有关的信息披露。

机构销售团队联系人

华东团队	茅炯	021-33388488	maojiong@swhysc.com
华北团队	肖霞	15724767486	xiaoxia@swhysc.com
华南团队	王维宇	0755-82990590	wangweiyu@swhysc.com
华北创新团队	潘烨明	15201910123	panyeming@swhysc.com
华东创新团队	朱晓艺	18702179817	zhuxiaoyi@swhysc.com

股票投资评级说明

证券的投资评级：

以报告日后的 6 个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：

买入（Buy）　　：相对强于市场表现 20%以上；

增持（Outperform）　　：相对强于市场表现 5%～20%；

中性（Neutral）　　：相对市场表现在 - 5%～ + 5%之间波动；

减持（Underperform）　　：相对弱于市场表现 5%以下。

行业的投资评级：

以报告日后的 6 个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：

看好（Overweight）　　：行业超越整体市场表现；

中性（Neutral）　　：行业与整体市场表现基本持平；

看淡 (Underweight)　　：行业弱于整体市场表现。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。申银万国使用自己的行业分类体系，如果您对我们的行业分类有兴趣，可以向我们的销售员索取。

本报告采用的基准指数　　：沪深 300 指数（A 股）、恒生中国企业指数（H 股）、纳斯达克指数（美股）

法律声明