下载日志已记录，仅供内部参考

计算机

2026 年 09 月 13 日

Rubin 升级驱动硬件价值扩散，AI 应用持续深化

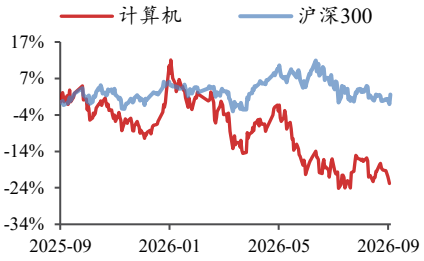
——行业周报

张越（分析师）
zhangyue1@kysec.cn
证书编号：S0790524090003

叶进新（联系人）
yejinxin@kysec.cn
证书编号：S0790126060053

投资评级：看好（维持）

行业走势图



数据来源：聚源

相关研究报告

《OpenAI 发布 GPT-6 Astra，大模型迈向执行主体—行业点评报告》-2026.9.7

《高密度算力驱动液冷加速渗透，AI 产业持续迭代—行业周报》-2026.9.6

《算力供需缺口持续扩大，算力租赁行业景气度上行—行业深度报告》-2026.8.25

● Rubin 重塑机柜价值结构，存算网协同打开产业成长空间

全球 AI 服务器需求持续增长，Rubin 平台以机架级协同设计推动算力、存储与互联全面升级，产业价值向整机柜系统及配套环节扩散。VR200 整柜 BOM 较 GB300 提升约 95%，GPU 单柜价值增长约 57%，内存增长约 435%，成为重要增量来源；PCB、MLCC 与 ABF 载板受益于用量增加与规格升级，单柜价值同步提升。架构层面，NVLink 6 强化柜内互联，Spectrum-X 推动集群网络升级，无缆线计算托盘提升装配与维护效率，全液冷及供电架构演进支撑更高功率密度。随着智能体持续推理需求释放，AI 基础设施建设有望从算力扩容进一步走向存算网全栈升级，带动内存、高阶 PCB、高速互联、电源与液冷等环节成长，产业链价值增量有望持续向细分领域扩散。

● 模型迭代与 Agent 落地共振，云端及端侧算力持续升级

本周 AI 产业延续密集创新，模型能力、应用落地与算力建设协同推进。模型端，DeepSeek 发布 V4.1 Flash，强化原生多模态能力并降低调用成本；Kimi K2.8 Preview 扩大百万上下文覆盖，推动 Coding 与 Agent 能力普及。应用端，腾讯文档上线 AI 工作台，支持人机双写与后台任务执行；Meta 推出个人 Agent Muse，拓展云端持续执行场景，AI 正由内容生成迈向复杂任务交付。算力端，OpenAI 加码数据中心建设，华为麒麟 9050 Pro 推动大模型端侧部署，Arm 与苹果持续提升移动终端 AI 性能。光互连领域，1.6T 产品推进量产与商业化，NPO、CPO 及 OCS 等技术加快演进，有望与模型效率提升、应用渗透共同拓展 AI 产业空间。

● AI 算力相对表现占优，细分领域结构性机会持续显现

9 月 7 日至 9 月 11 日，计算机指数周度变动为-4.61%，同期沪深 300、上证指数、创业板指和科创 50 分别变动-0.83%、-1.07%、+1.08%和-1.52%。板块内部结构性机会仍有体现，AI 算力指数周度变动约-1.67%，相对 AI 应用指数取得约 1.62 个百分点的超额收益。个股层面，火炬电子、中际旭创分别上涨 14.02%和 13.76%，均录得两位数涨幅；AI 应用方向涨幅前十个股周涨幅均超过 3%，深南电路上涨 19.55%。结合 Rubin 架构升级、光互连产品演进与 Agent 应用拓展，AI 产业催化持续积累，可关注技术升级与需求释放带来的细分领域机会。

● 受益标的

国产算力：寒武纪、海光信息、禾盛新材、沐曦股份、浪潮信息、工业富联等；**海外算力：**东山精密、胜宏科技、联特科技、德邦科技等；**算力租赁：**宏景科技、利通电子、协创数据等；**大模型：**智谱、MiniMax、阿里巴巴、腾讯控股；**液冷：**盾安环境、同飞股份、恒铭达等。

● 风险提示：大模型技术迭代不及预期风险、AI 技术应用不及预期风险、相关政策不及预期风险、AI Infra 建设不及预期风险等。

下载日志已记录，仅供内部参考

目 录

1、 Rubin 机柜价值结构重塑，机架协同定义系统新范式 4
1.1、 供需失衡传导至全链条，BOM 成本与结构重塑 4
1.2、 Rubin 机柜核心架构拆解，关注高弹性细分环节 7
2、 计算机产业动态更新 17
2.1、 国内 CSP 厂商与大模型厂商最新动向 17
2.2、 海外 CSP 厂商与大模型厂商最新动向 18
2.3、 国产算力产业最新进展 20
2.4、 海外算力产业最新进展 20
3、 上周市场表现回顾（2026/9/7-2026/9/11） 22
4、 风险提示 23

图表目录

图 1： 2027 年 AI 服务器有望突破 320 万台，维持高景气增长 4
图 2： 半导体多环节涨价与供需预期 4 月全面上修，紧缺态势持续升级 5
图 3： Rubin 采用极端协同设计理念，将数据中心视为计算的基本单元 5
图 4： AI 服务器硬件价值主线已由 GPU 单点扩张迈入整机柜系统密度扩张阶段 6
图 5： Vera Rubin 平台集成六款专用 AI 芯片，机架级原生融合计算、网络与基础设施 8
图 6： Vera Rubin 平台旗舰 NVL72 机架系统，兼顾峰值性能与持续智能生成 8
图 7： NVIDIA Rubin GPU 为下一代 AI 而生 9
图 8： NVIDIA Vera CPU 为数据移动和智能体计算而生 9
图 9： Vera CPU 与 Rubin GPU 实现全方位代际跃升，算力与互联能力大幅增强 9
图 10： 二代 NVLink-C2C 可以提供 1.8TB/s 带宽 10
图 11： 面向千兆级 AI 工厂的处理器 10
图 12： ConnectX-9 为超大规模 AI 提供超高吞吐量 10
图 13： BlueField-4 DPU 为操作系统提供核心算力支撑 10
图 14： ConnectX-9 与 BlueField-4 为 Rubin 打造高性能、可编程、安全的数据通路 11
图 15： Rubin 托盘为无缆线、无软管、无风扇的密封模块 11
图 16： VR NVL72 计算托盘采用模块化、无缆线设计 12
图 17： Vera Rubin NVLink 交换托盘构成柜内横向扩展网络架构 12
图 18： NVIDIA NVLink 6 Switch 将整机柜转化为单个加速单元 13
图 19： Vera Rubin NVL72 NVLink 全对全互联拓扑 14
图 20： NVIDIA Spectrum-X 以太网面向 AI 智算集群的横向扩展与跨域互联 14
图 21： NVIDIA Spectrum-6 以太网交换机搭载共封装光学，提供多速率高速光端口 15
图 22： Vera Rubin NVL72 以内部歧管与快接头构建起覆盖整柜的冷却分配网络 16
图 23： Vera Rubin NVL72 在供电架构上实现由低压直流向 800V 高压直流的跃迁 16
图 24： 计算机行业板块相对表现 22
图 25： AI 细分指数表现情况 22
图 26： AI 算力板块个股本周涨幅前十 23
图 27： AI 应用板块个股本周涨幅前十 23

表 1： AI 服务器出货量持续扩大 4
表 2： VR200 服务器硬件成本结构正发生结构性重构 7

下载日志已记录，仅供内部参考

表 3： 相较 BlueField-3，BlueField-4 带宽算力内存升级，支撑 AI 工厂规模化扩容 10
表 4： NVIDIA NVLink 6 单 GPU 双向互联带宽 3.6TB/s，相对上一代实现翻倍 13
表 5： 单颗 GPU 最多支持 36 条 NVLink 链路，较前代 18 条翻倍 13

下载日志已记录，仅供内部参考

1、 Rubin 机柜价值结构重塑，机架协同定义系统新范式

1.1、 供需失衡传导至全链条，BOM 成本与结构重塑

全球 AI 服务器市场维持高景气增长，价值构成与产业驱动逻辑正发生显著变化。据 TrendForce 测算，2026 年全球 AI 服务器出货量同比增速将超 20%，占整体服务器比重提升至 17%；更新预测显示增速可达 28%以上，增长由北美云服务商、主权云项目、推理负载及定制 AI ASIC 部署共同支撑。出货规模持续攀升，2023 年约 120 万台，2024 年升至 175 万台，2025 年达 217 万至 220 万台，2026 年攀升至 265 万至 270 万台，2027 年有望突破 320 万台，需求增长并非短期采购脉冲。相较于传统服务器，AI 服务器单系统价值结构深度调整，内存、先进封装等环节对系统成本、性能与供应链风险的影响力持续提升。

表1：AI 服务器出货量持续扩大

年份	AI 服务器出货量	同比增速	市场份额	主要驱动因素
2023	约 120 万台	+38.4%	约 9%	AI 服务器早期大规模部署
2024	约 175 万台	+46%	约 12.1%	H100/H200 平台放量
2025	约 217 万-220 万台	+24.3%	约 14%	GB200 和 GB300 部署
2026	约 265 万-270 万台	>20%	约 17%	GB300、Rubin 及 AI ASIC 系统
2027	320 万台以上	约 20%（预估）	约 20%（预估）	推理需求扩张与主权 AI 基础设施

数据来源：Aetrix、开源证券研究所

图1：2027 年 AI 服务器有望突破 320 万台，维持高景气增长



数据来源：Aetrix、开源证券研究所

需求放量正与供给刚性碰撞，上游供应链或步入较长紧缺周期。高盛将 2026 年 DRAM 涨幅预期由约 150%上调至 250%-280%，NAND 由约 100%上调至 200%-250%，ABF 载板、CCL 同步上修，其预计 DRAM 供需缺口达 4.9%，或为十五年来最严重短缺，紧缺延续至 2028 年。HBM 耗片量约为普通 DRAM 的数倍，持续挤占通用产能，缺口根源主要为供给弹性不足叠加 AI 服务器需求强劲。先进制程新产能预计 2028-2029 年方逐步释放，云厂商长约进一步锁定存量供给，供给弹性趋于收缩。涨价信号自存储向载板、被动元件与材料逐级传导，AI 服务器价值构成调整或由需求侧逻辑演变为供需共振逻辑。

下载日志已记录，仅供内部参考

图2：半导体多环节涨价与供需预期 4 月全面上修，紧缺态势持续升级



零部件 / 材料	价格变动 2025年	价格变动 2026年
DRAM	10%~20%	约150%
NAND	-20%~-10%	约100%
近线硬盘	0%	0%~5%
晶圆代工—台积电	5%~10%	5%
晶圆代工—中国大陆	0%	0%
多层陶瓷电容 (MLCC)	-10%~-5%	略低于0%
T型玻璃	0%	20%~30%
钽粉	5%~10%	10%
CCL用铜箔 (HVLP3+)	5%	5%~10%
磷化铟衬底	0%	15%
ABF载板—整体	0%~5%	20%~30%
ABF载板—GPU	0%	0%
覆铜板（CCL）	5%~10%	5%~10%

零部件 / 材料	价格变动 2025年	价格变动 2026年
DRAM	10%~20%	250%~280%
NAND	-20%~-10%	200%~250%
近线硬盘	0%	0%~5%
晶圆代工—台积电	5%~10%	5%
晶圆代工—中国大陆	0%	10%
多层陶瓷电容 (MLCC)	-10%~-5%	0%~5%
T型玻璃	0%	20%~30%
钽粉	5%~10%	10%
CCL用铜箔 (HVLP3+)	5%	5%~10%
磷化铟衬底	0%	15%
ABF载板—整体	0%~5%	30%~35%
覆铜板（CCL）	10%~15%	35%~40%

主要变化（4月较1月）	
DRAM	2026年第四季度及2027年上半年供需趋紧；2026年涨幅上调至250%~280%（此前约150%）。
NAND	2026年第四季度及2027年上半年供需趋紧；2026年涨幅上调至200%~250%（此前约100%）。
台积电	2027年上半年由紧张转为非常紧张。
中国大陆晶圆代工	2026年及2027年上半年由紧张 / 平衡转为不紧张；2026年预计涨价10%（此前0%）。
MLCC	2026年第二季度至2027年上半年由紧张转为非常紧张；2026年预计涨价0%~5%（此前略低于0%）。
ABF载板	2026年下半年及2027年第一季度由紧张转为非常紧张；2026年涨幅上调至30%~35%（此前20%~30%）。
CCL	2026年第四季度及2027年第一季度由不紧张转为紧张；2026年涨幅上调至35%~40%（此前5%~10%）。

数据来源：电子工程世界、开源证券研究所

英伟达下一代旗舰计算平台 Vera Rubin 的规模量产，标志着 AI 服务器价值构成与产业驱动逻辑的转折点。2026 年 8 月，Vera Rubin 平台完成从验证样机到规模量产的切换，首批机柜已进入亚洲与北美数据中心，这一进程并非一次常规的 GPU 换代。平台采用极端协同设计理念，将数据中心而非单台 GPU 服务器视为计算的基本单元，在算力、互联与内存带宽三个维度实现代际跃迁。系统级集成度与价值量的系统性抬升，使 AI 服务器从算力扩张迈向存算网全栈升级，为下游各硬件环节的价值重估提供结构性基础。

图3：Rubin 采用极端协同设计理念，将数据中心视为计算的基本单元



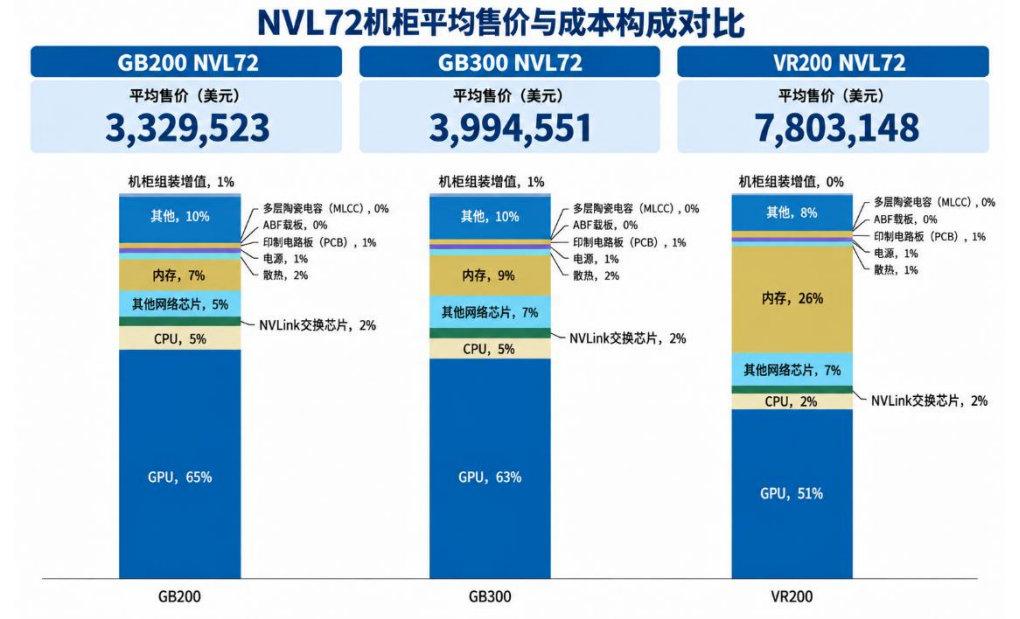
资料来源：NVIDIA、开源证券研究所

Rubin VR200 机柜的核心意义并非单柜 ASP 接近翻倍，而在于 AI 硬件价值主线已由 GPU 单点扩张迈入整机柜系统密度扩张阶段。据摩根士丹利底层测算，GPU

下载日志已记录，仅供内部参考

仍为最大美元价值池，单柜价值由 GB300 的 252 万美元增至 396 万美元，增幅约 57%，仍是 ASP 提升核心驱动；但占 BOM 比重由近三分之二回落至约 51%，主要为整柜 BOM 接近翻倍下内存与系统件增速更快所致，GPU 需求见顶。内存成为第二大价值池，单柜价值由约 37.4 万美元增至 200.2 万美元，增幅 435%，占比升至 26%，HBM、SOCAMM 的价格和供应不再只是半导体内部变量，已直接影响整机报价与 ODM 收入。PCB、MLCC、ABF 等二级供应链呈低占比、高弹性特征，增速显著高于整柜；电源与液冷亦由配套部件升级为高密度算力核心约束项。AI 硬件价值正从单一算力芯片主导，转向机柜级成分占比与系统复杂度的综合承载。

图4：AI 服务器硬件价值主线已由 GPU 单点扩张迈入整机柜系统密度扩张阶段



数据来源：wallstreetcn、开源证券研究所

（一）内存：单柜价值增逾 4 倍，成 BOM 中最大增量环节

内存单柜价值由约 37.4 万美元增至约 200 万美元，增幅 435%，为全部组件中最大单倍增量，已超过 GPU 成本的一半。增量由三部分构成，HBM4 总量 20.7TB（72 颗 GPU×288GB），SK 海力士主导供应，价值约 40 至 50 万美元；LPDDR5X 总容量 54TB，为 GB200 的三倍，按 8 至 10 美元/GB 计约 40 至 54 万美元；3D NAND 为最大结构性增量，Rubin 首次将 NAND 规模集成于机柜用于模型加载与检查点，单机架价值超 100 万美元，而 GB200 近乎为零。

（二）PCB：增幅 233%居全部组件之首，弹性最大环节

PCB 单柜价值由约 3.5 万美元增至 11.7 万美元，增幅 233%，居全部组件之首。增量来自三层叠加，新增 ConnectX 模组 PCB 与 Midplane 中板 PCB，贡献约 4.6 万美元纯增量；计算板由 22 层 HDI 升至 26 层、CCL 由 M7 升至 M8、单价由 650 美元升至 1400 美元，交换板由 24 层增至 32 层，升级幅度高于计算板，互联复杂度增速已超算力本身；叠加计算板面积扩大与新增 44 层中板。层数提升、材料升级与新增模组共振，PCB 成为本轮弹性最大组件，对供应商而言，AI 服务器所需并非更多板卡，而是更复杂昂贵的板卡，属结构性增长而非周期性波动。

（三）MLCC 与 ABF：用量与规格双升，低占比高弹性增量环节

MLCC 与 ABF 载板单柜价值分别由 1530 美元增至 4320 美元、由 1.1 万美元增

下载日志已记录，仅供内部参考

至 2 万美元，增幅 182%与 82%，是被低估的增量环节。MLCC 方面，据村田数据，单台 AI 服务器搭载约 3 万颗 MLCC，约为智能手机的 30 倍、汽车的 3 倍，单个 NVL72 机柜消耗量达 44 万颗，用量扩张直接驱动价值提升。ABF 载板方面，Rubin GPU 载板单价由约 100 美元翻倍至 200 美元，NVSwitch ASIC 由 18 颗倍增至 36 颗，ConnectX 芯片由 36 颗增至 72 颗，单价与数量同步倍增。二者绝对价值量虽低，但受益于 AI 服务器对高可靠被动元件与先进封装载板的结构性需求，增速显著高于整柜均价，具备较强的业绩弹性。

（四）电源与液冷：功率密度倒逼架构升级，散热成刚性约束

电源与液冷的价值变化不大，单柜价值分别由 5.8 万美元增至 7.6 万美元、由 6.5 万美元增至 7.2 万美元，增幅 32%与 12%，但数字背后是供电与散热架构的强制升级。电源方面，英伟达确认 Rubin Ultra 平台将采用 800V HVDC 架构，传统 54V 配电难以承载单机柜 200kW 以上的功率密度，切换至 800V 后端到端供电效率由 87.6%提升至 98%。液冷方面，在该功率密度下全液冷已成为必需配置，若计入 side-car CDU，单柜散热总价值约达 12.2 万美元。两项环节的价值增量不来自涨价，而来自功率密度抬升带来的技术形态切换，供电与散热已由可选项转变为高密度算力的刚性约束。

表2：VR200 服务器硬件成本结构正发生结构性重构

Nvidia NVL72 Bill of Materials	GB300	VR200	Diff.
GPU	$ 2520000	$3960000	57%
CPU	$180000	$180000	0%
NVLink Switch chip	$64800	$144000	122%
Other networking chips	$261000	$576000	121%
Memory	$373939	$2001600	435%
Cooling	$64610	$72080	12%
Power supply	$57600	$76000	32%
PCB	$35100	$116730	233%
ABF Substrate	$11160	$20340	82%
MLCC	$1530	$4320	182%
Others	$402412	$623278	55%
Rack assembly value add	$22400	$28800	29%
Total	$3994551	$7803148	95%

数据来源：电子工程世界、开源证券研究所

AI 服务器价值重构主线，正由单一算力芯片转向整机柜级系统价值扩散。Rubin 整机成本接近翻倍，核心驱动主要系存储等侧规格升级与供需涨价，而仅非 GPU 本身提价；本轮涨价由产能错配驱动，高端零部件紧张格局或延续，构成周期底部支撑。价值增量有望沿内存、PCB、MLCC、互联芯片、电源与液冷等环节逐级释放，配套环节涨幅普遍高于 GPU，产业链利润正从核心芯片向中下游细分龙头扩散，其中内存受 HBM 挤占与长约锁定驱动，涨价持续性最强。

1.2、 Rubin 机柜核心架构拆解，关注高弹性细分环节

NVIDIA Vera Rubin 是面向 Agentic AI 时代的新一代全栈计算平台，其根本优势在于将整座数据中心重塑为单一计算单元，以支撑持续推理的智能体工作负载。与传统面向单次交互的算力供给不同，Agentic 工作负载要求推理、规划、工具调用与长上下文执行在多个步骤间持续进行，这对端到端的低时延、高吞吐与长上下文

下载日志已记录，仅供内部参考

处理能力提出了全新挑战。Vera Rubin 平台围绕这一需求进行原生设计，从数据搬运、矩阵运算效率到长上下文执行与机柜级部署逐层优化，系统性消除了智能体推理的端到端瓶颈。其核心价值在于以极限协同设计，将算力、内存与互连统一为面向智能体工作负载的一体化底座，从而在持续推理这一全新范式下，为万亿参数模型与多步骤智能体任务提供稳定而高效的算力支撑。

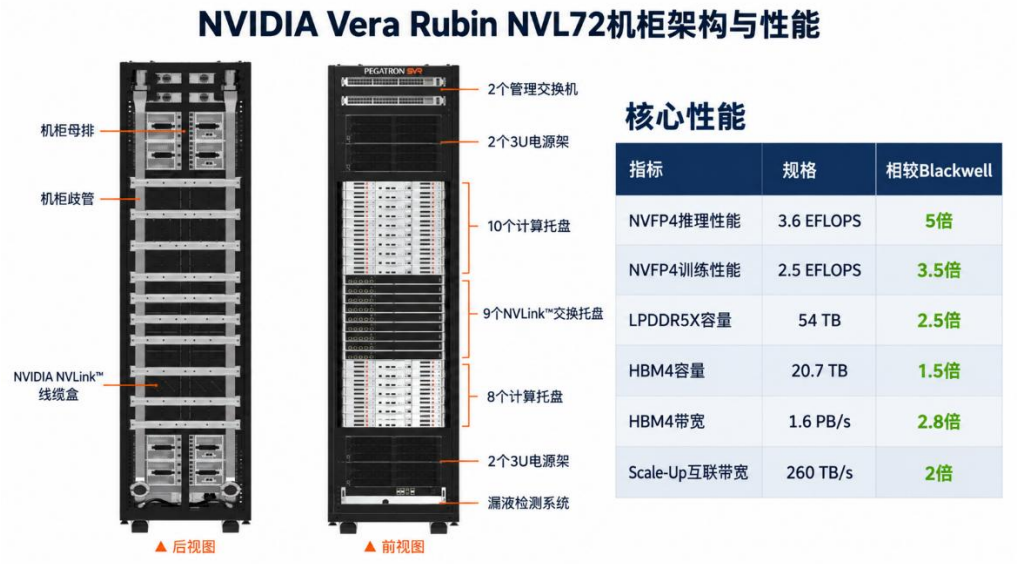
图5：Vera Rubin 平台集成六款专用 AI 芯片，机架级原生融合计算、网络与基础设施



资料来源：NVIDIA

Vera Rubin NVL72 以机柜为基本计算单元，通过 NVLink 6 全互联拓扑将 72 颗 Rubin GPU 与 36 颗 Vera CPU 封装为一台机架级超级计算机。性能层面，整柜提供 3.6TB/s 的单 GPU 带宽与 260TB/s 的横向扩展带宽，单位功耗的训练性能较 Blackwell 提升最高 4 倍、推理性能提升最高 10 倍，单 Token 成本降至前代约十分之一。机柜构成上，整柜基于 48U 第三代 MGX 架构，由 18 个计算托盘与 9 个 NVLink 交换托盘组成；机柜顶部与底部分别配置 2×3U 电源架，合计 4 个电源架，并部署 2 个管理交换机。机柜后部则由 NVLink 线缆盒、供电母排与液冷歧管构成，辅以机柜级泄漏检测系统。整体上，Rubin 并非追求峰值性能，而是以可预测时延、异构阶段的高利用率与功率支撑持续稳定的智能生产。

图6：Vera Rubin 平台旗舰 NVL72 机架系统，兼顾峰值性能与持续智能生成



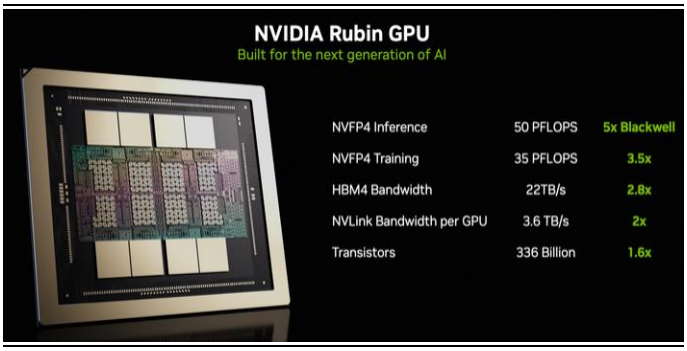
数据来源：NVIDIA、PEGATRON、开源证券研究所

下载日志已记录，仅供内部参考

（一）计算托盘

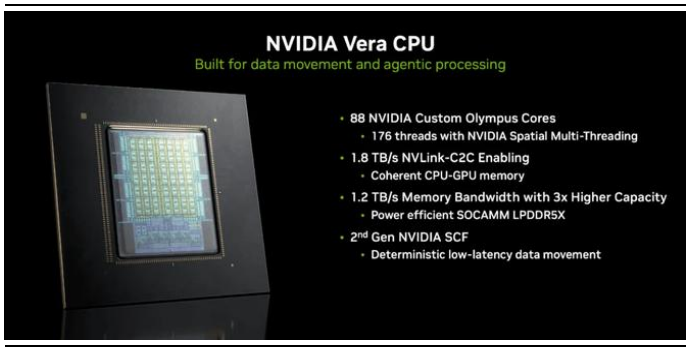
计算托盘是 Vera Rubin NVL72 的基础算力单元，每一托盘集成两个 Vera Rubin 超级芯片，承载 4 颗 Rubin GPU 与 2 颗 Vera CPU。其中，Rubin GPU 由两颗计算裸片经 NV-HBI 高速互联封装而成，集成 224 个流式多处理器与第五代张量核心，单卡配备 288GB HBM4 内存，在 NVFP4 精度下实现最高 50 PetaFLOPS 的推理算力；Vera CPU 则采用 88 核定制 Olympus 核心，通过空间多线程将有效线程数扩展至 176 个，单颗最高支持 1.5TB LPDDR5X 内存。二者通过 NVLink-C2C 内存一致性互联紧密耦合，使 CPU 得以作为数据引擎直接参与调度、数据搬运与执行协调，从而在单个托盘内完成算力与数据处理的一体化整合，为整柜的机架级加速奠定基础。

图7：NVIDIA Rubin GPU 为下一代 AI 而生



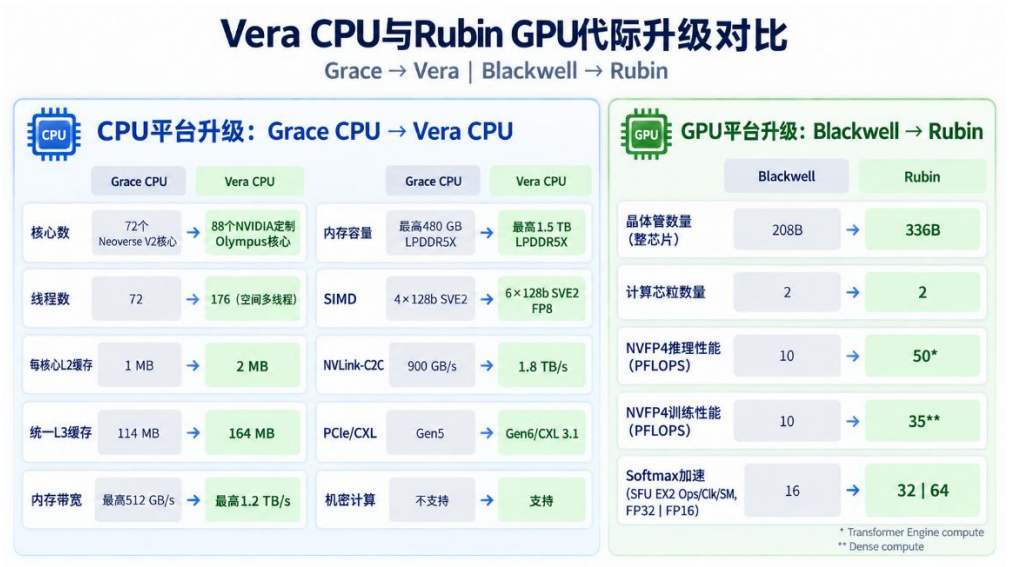
资料来源：NVIDIA

图8：NVIDIA Vera CPU 为数据移动和智能体计算而生



资料来源：NVIDIA

图9：Vera CPU 与 Rubin GPU 实现全方位代际跃升，算力与互联能力大幅增强



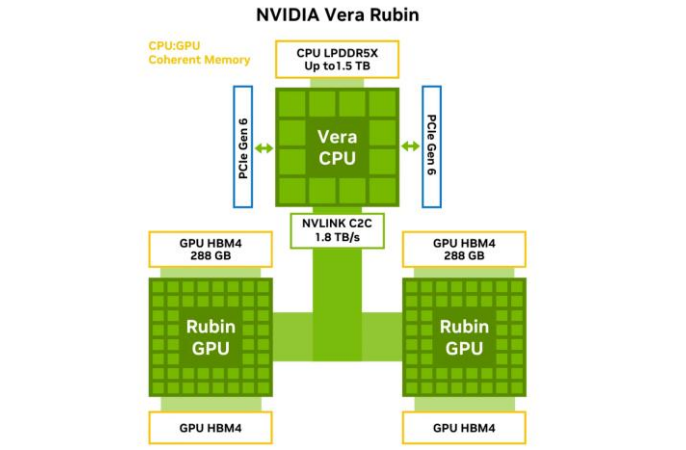
数据来源：NVIDIA、开源证券研究所

计算托盘借助第六代 NVLink 实现芯片间的超高带宽互联，是 72 颗 GPU 得以协同为单一算力域的关键。托盘层面，每颗 Rubin GPU 经 NVLink 6 提供 3.6TB/s 的双向互联带宽，较 Blackwell 的 NVLink 5 翻倍；同时，每个托盘配置 4 个 NVLink 6 连接器，将算力接入整柜的 NVLink 全互联拓扑。芯片内部，NVLink-C2C 以 1.8TB/s 的带宽实现 Vera CPU 与 Rubin GPU 之间的内存一致性通信，使 CPU 与 GPU 共享统一寻址空间、消除传统异构架构下的数据搬运瓶颈。互联设计贯穿芯内-托盘-整柜三

下载日志已记录，仅供内部参考

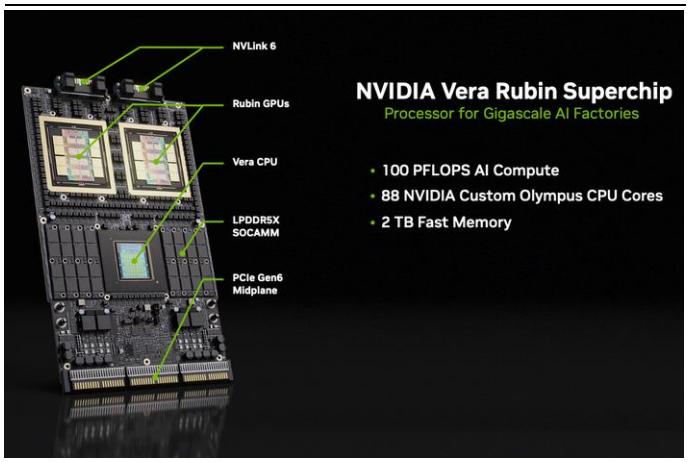
个层级，配合机柜后部的 NVLink 主干与交换托盘，最终支撑整柜 260TB/s 的横向扩展带宽，使全部 GPU 在极低时延下实现任意两点间的直达通信。

图10：二代 NVLink-C2C 可以提供 1.8TB/s 带宽



资料来源：NVIDIA

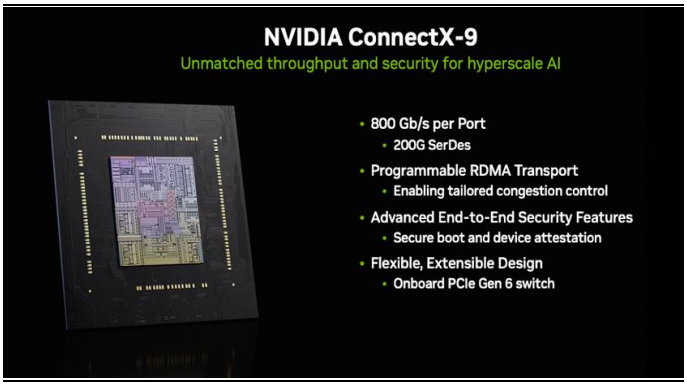
图11：面向千兆级 AI 工厂的处理器



资料来源：NVIDIA

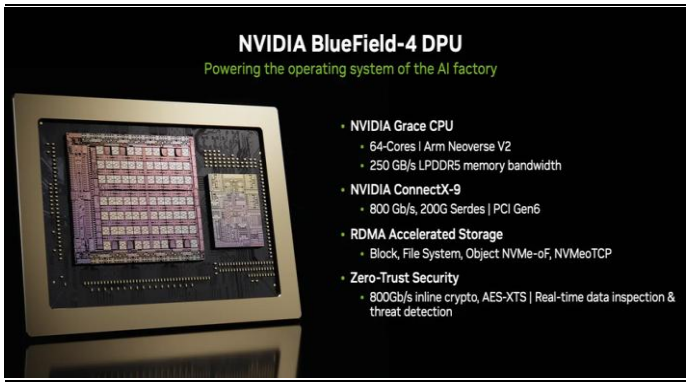
在横向扩展与存储卸载层面，每个计算托盘通过 ConnectX-9 SuperNIC 与 BlueField-4 DPU 承担网络与数据处理职能。托盘前置模块化舱位集成 8 个 ConnectX-9 SuperNIC，为每颗 Rubin GPU 提供 1.6Tb/s 的对外网络带宽，较前代 ConnectX-8 实现带宽翻倍，支撑大规模集群间的横向扩展通信；同时部署 1 个 BlueField-4 DPU，以 800Gb/s 带宽承担网络、存储与安全的卸载处理，将 CPU 与 GPU 从基础设施任务中解放出来，使其专注于 AI 执行。在存储侧，托盘配置 4 个 E1.S NVMe 以提供本地缓存。网卡+DPU+本地存储的组合，使计算托盘在保证算力密度的同时，具备较强的数据通路能力。

图12：ConnectX-9 为超大规模 AI 提供超高吞吐量



资料来源：NVIDIA

图13：BlueField-4 DPU 为操作系统提供核心算力支撑



资料来源：NVIDIA

表3：相较 BlueField-3，BlueField-4 带宽算力内存升级，支撑 AI 工厂规模化扩容

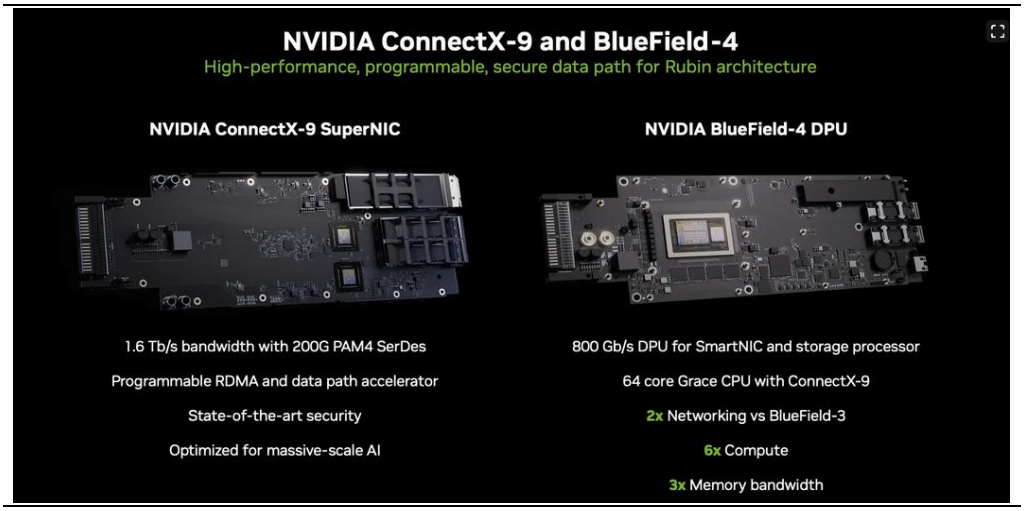
特性	BlueField-3	BlueField-4
端口带宽	400 Gb/s	800 Gb/s
计算能力	16 核 Arm A78 架构	64 核 Arm Neoverse V2 架构，计算性能达前代 6 倍
内存带宽	75 GB/s	250 GB/s
内存容量	32 GB	128 GB
云网络主机承载规模	3.2 万台	12.8 万台
传输态数据加密速率	400 Gb/s	800 Gb/s

下载日志已记录，仅供内部参考

特性	BlueField-3	BlueField-4
NVMe 存储解聚性能	4K 随机 IOPS 达 1000 万	4K 随机 IOPS 达 2000 万

数据来源：NVIDIA、开源证券研究所

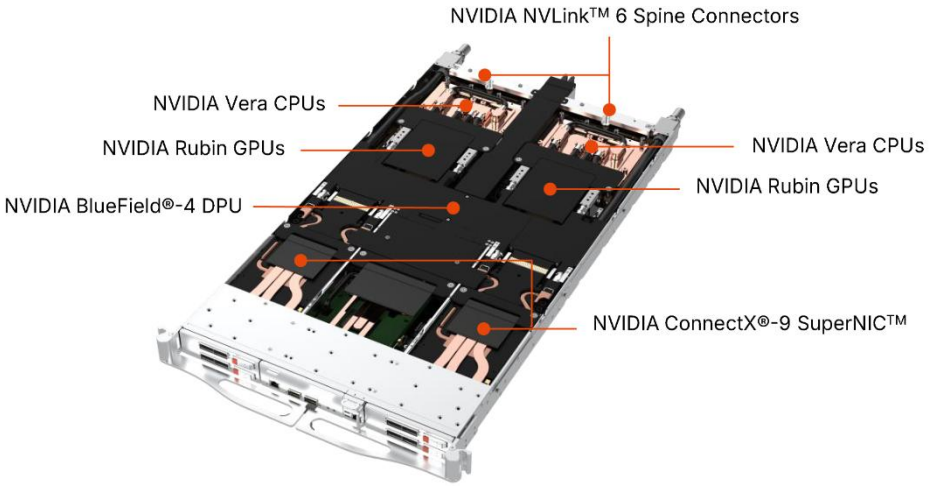
图14：ConnectX-9 与 BlueField-4 为 Rubin 打造高性能、可编程、安全的数据通路



资料来源：NVIDIA

计算托盘最显著的架构变革在于无缆线化设计，以刚性 PCB 中板取代了前代内部铜缆与液冷软管。相较 Blackwell 托盘依赖 DensiLink、MCIO 等复杂线缆连接，Rubin 托盘通过板对板连接器与中板实现模块间互连，形成无缆线、无软管、无风扇的密封模块，既消除了线缆装配带来的故障点，又将托盘组装时间由近两小时压缩至约五分钟，装配与维护效率提升最高 20 倍。这一设计同时优化了信号完整性，使高速 PCIe 与以太网信号得以在更短路径内传输。散热方面，托盘采用单相直接液冷，以更高流量适配高功率负载。整体上，无缆线化标志着计算托盘从组件拼装迈向模块化即插即用的工程范式跃迁。

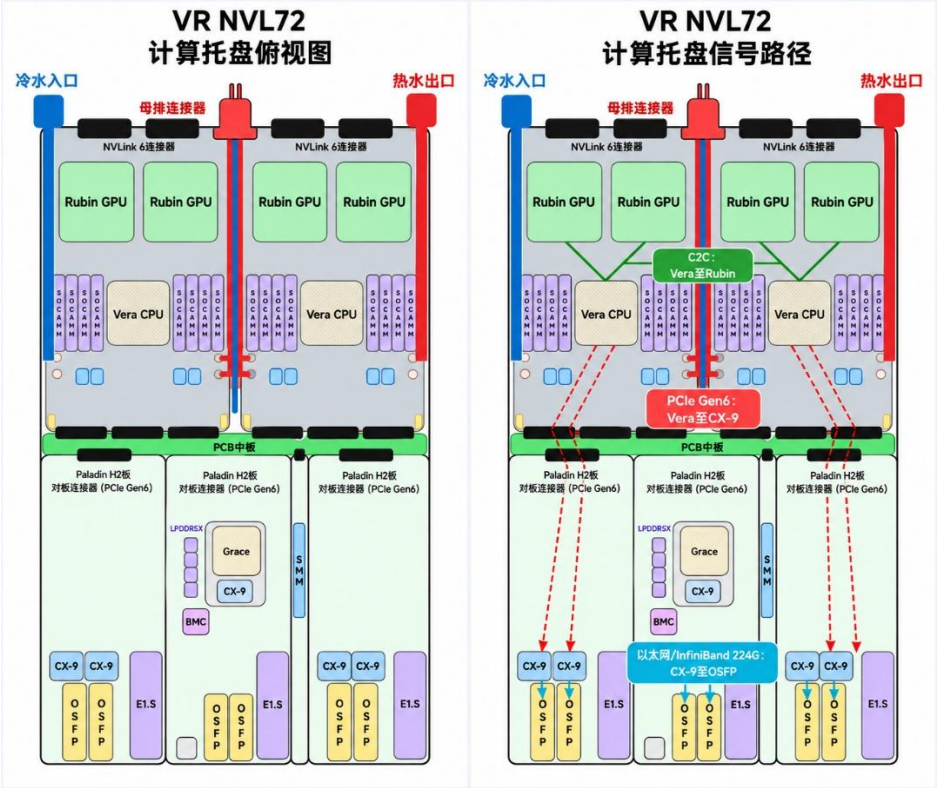
图15：Rubin 托盘为无缆线、无软管、无风扇的密封模块



资料来源：PEGATRON

下载日志已记录，仅供内部参考

图16：VR NVL72 计算托盘采用模块化、无缆线设计

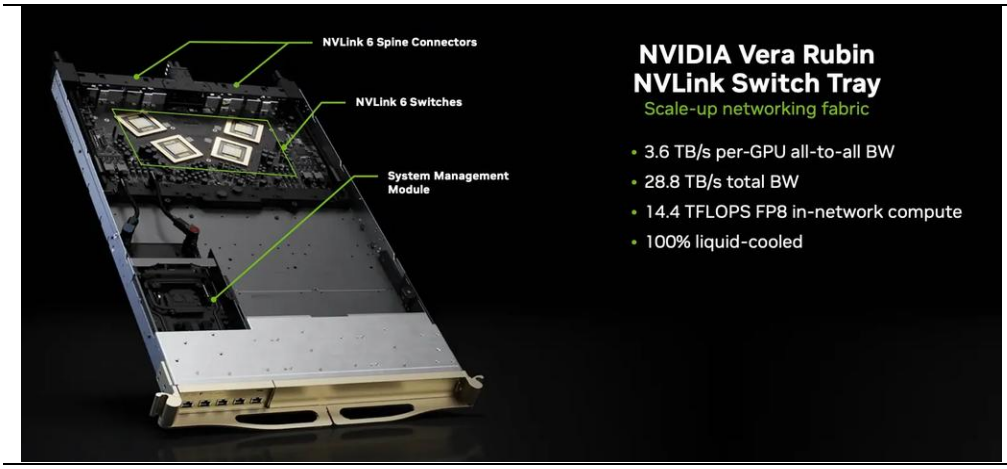


资料来源：SemiAnalysis、开源证券研究所

（二）交换托盘

NVLink 交换托盘是将 72 颗 Rubin GPU 聚合为单一算力域的核心枢纽，负责在机柜内构建 GPU 间的高带宽全互联网络。整柜共部署 9 个交换托盘，每个交换托盘集成 4 颗 NVLink 6 Switch 交换芯片，合计 36 颗交换芯片。这些交换芯片通过 NVLink 连接多个 GPU，将原本局限于服务器内的 GPU 直连扩展至整个机柜，使 72 颗 GPU 在非阻塞计算结构中实现任意两点间的全互联通信。交换托盘与位于机柜后部的 NVLink 主干协同工作，后者由 4 个模块化预集成线缆盒构成，内含约 5000 根铜缆，共同将全部 GPU 纳入统一的横向扩展域，从而把分散的计算托盘连接为一台机架级加速器，为万亿参数模型的训练与推理提供通信底座。

图17：Vera Rubin NVLink 交换托盘构成柜内横向扩展网络架构

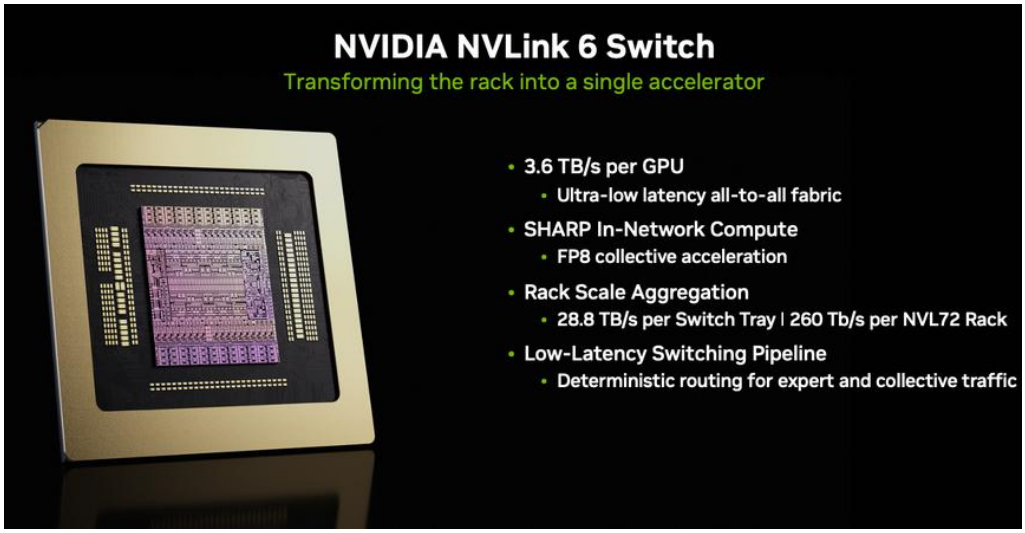


资料来源：NVIDIA

下载日志已记录，仅供内部参考

交换托盘在性能与弹性两个维度上均较前代显著跃升，成为保障 AI 工厂持续稳定运行的关键。性能层面，NVLink 6 使每颗 GPU 的互联带宽提升至 3.6TB/s，较上一代翻倍，整柜聚合带宽达 260TB/s；同时，每颗 NVLink 6 Switch 内置 SHARP 引擎，提供网内归约与组播加速，整柜合计 130 TFLOPS 的 FP8 网内计算能力，可显著加速 all-reduce、reduce 等集合通信，降低通信流量对计算资源的占用。运维层面，交换托盘引入控制平面弹性、部分填充运行、托盘热插拔、软件定义路由与在线软件升级等能力，支持在不停机的情况下维护或替换交换托盘，即使多个交换托盘离线，整柜仍可继续运行，从而将大规模集群的维护停机影响降至最低。

图18：NVIDIA NVLink 6 Switch 将整机柜转化为单个加速单元



资料来源：NVIDIA

从拓扑结构看，NVL72 通过芯片内-托盘内-机柜级三层互联，将 72 颗 GPU 组织为单一的非阻塞 scale-up 域。芯片层面，Vera CPU 与 Rubin GPU 经 NVLink-C2C 以 1.8TB/s 实现内存一致性耦合；托盘层面，每颗 GPU 经 NVLink 6 以 3.6TB/s 接入交换结构，单颗 GPU 最多支持 36 条 NVLink 链路，较前代 18 条翻倍。机柜层面，18 个计算托盘经铜缆主干连接至 9 个交换托盘，由 36 颗 NVLink 6 Switch 构成全互联矩阵，使任意 GPU 到任意 GPU 仅需一跳直达，并以均匀时延与均匀带宽互通。GPU-NVSwitch-GPU 的一跳拓扑，是 72 颗 GPU 得以作为一个统一加速器协同工作的物理基础，支撑其 [illegible] 的横向扩展带宽与 3.6 ExaFLOPS 的 AI 算力。

表4：NVIDIA NVLink 6 单 GPU 双向互联带宽 3.6TB/s，相对上一代实现翻倍

互联类型	Blackwell	Rubin
NVLink（GPU-GPU）（GB/s，双向）	1,800	3,600
NVLink-C2C（CPU-GPU）（GB/s，双向）	900	1,800
PCIe 接口（GB/s，双向）	256（Gen 6）	256（Gen 6）

资料来源：NVIDIA、开源证券研究所

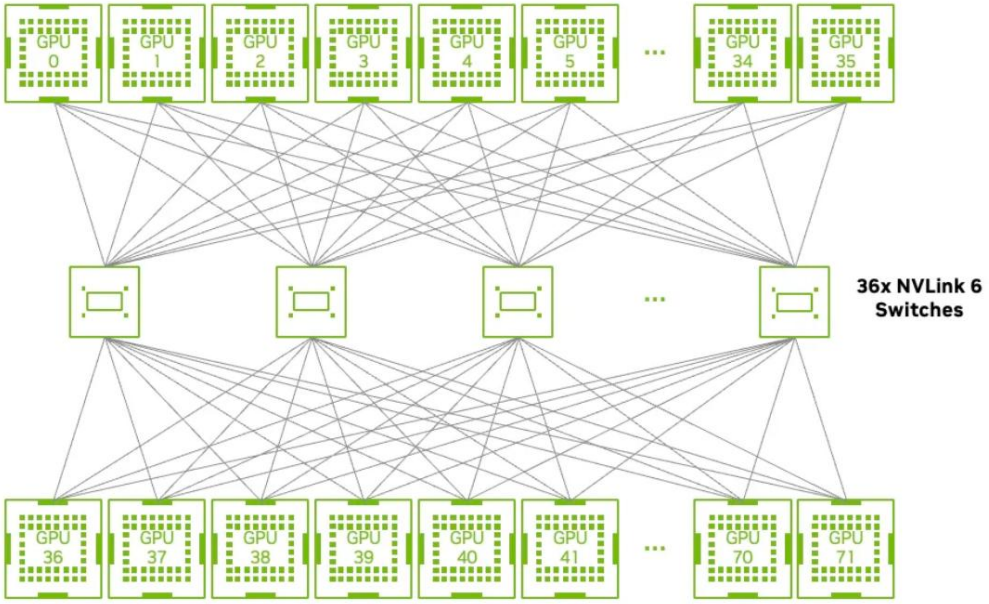
表5：单颗 GPU 最多支持 36 条 NVLink 链路，较前代 18 条翻倍

维度	第四代 NVLink	第五代 NVLink	第六代 NVLink
每 GPU 的 NVLink 带宽	900GB/s	1,800GB/s	3,600 GB/s
每 GPU 的最大链路数	18	18	36
支持的 NVIDIA 架构	NVIDIA Hopper 架构	NVIDIA Blackwell 架构	NVIDIA Rubin 平台

资料来源：NVIDIA、开源证券研究所

下载日志已记录，仅供内部参考

图19：Vera Rubin NVL72 NVLink 全对全互联拓扑

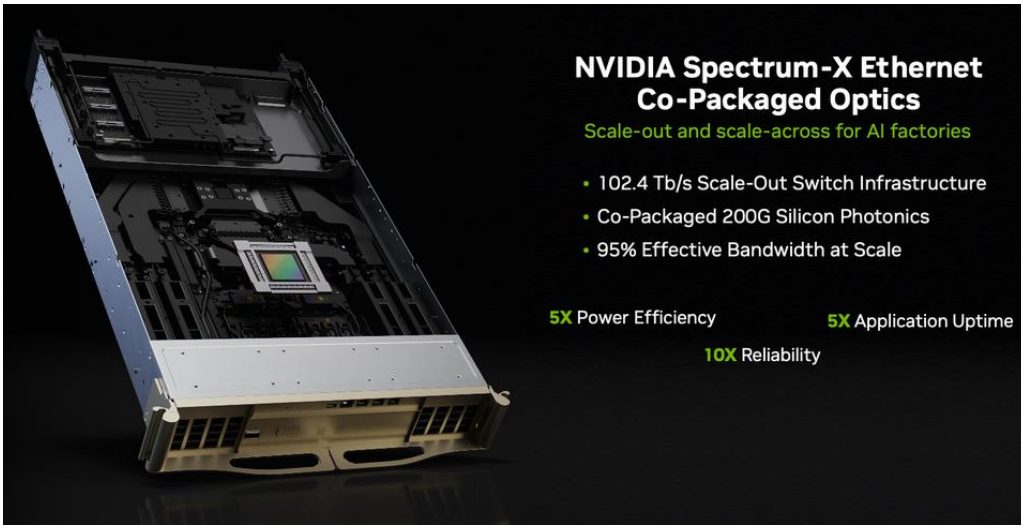


资料来源：NVIDIA

（三）Spectrum-X 以太网交换机

在机柜级全互联之上，Vera Rubin NVL72 依托 Spectrum-X 以太网将算力协同扩展至机柜之外，构建面向 AI 工厂的横向扩展网络。NVLink 6 使 72 颗 GPU 在机柜内协同为单一加速器，而 Spectrum-X 以太网则把这一能力延伸至跨机柜、跨数据中心的更大规模集群，提供可预测的高吞吐横向扩展连接。与传统以太网不同，AI 工作负载中的 MoE 分发、集合通信与同步密集阶段会产生突发、非对称且高度相关的流量，极易放大拥塞、尾延迟与性能抖动。Spectrum-X 以太网正是针对这类流量特性而生，通过协调拥塞控制、自适应路由与端到端遥测三大机制，在负载压力下维持有效带宽与可重复的性能表现，从而支撑多租户 AI 工厂的高利用率运行。

图20：NVIDIA Spectrum-X 以太网面向 AI 智算集群的横向扩展与跨域互联



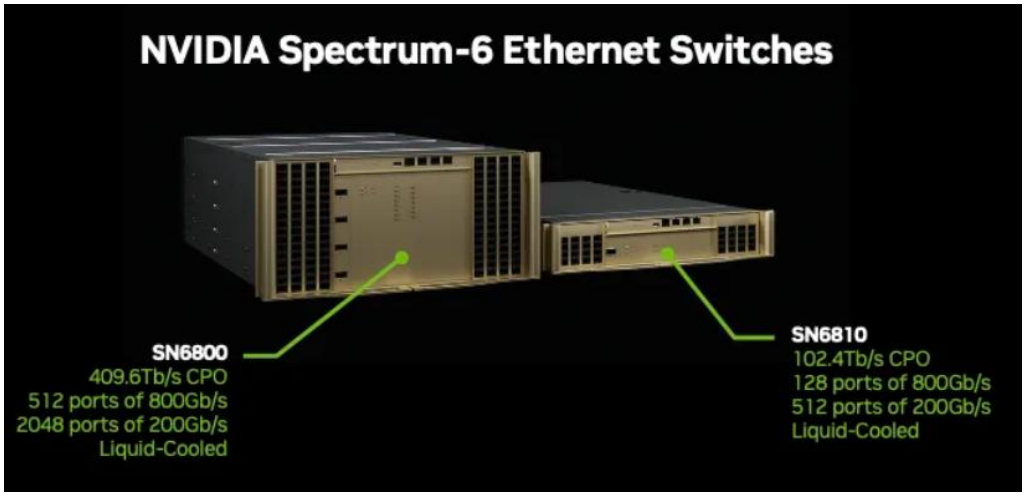
资料来源：NVIDIA

在实现层面，Spectrum-X 由 Spectrum-6 交换机与 ConnectX-9 SuperNIC 端点协同构成，依托共封装光学完成横向扩展网络的代际升级。交换机与端点紧密协同

下载日志已记录，仅供内部参考

设计，二者配合塑造流量、隔离负载并消除热点，使网络结构具备对 AI 流量的内生适配能力。在物理层，Spectrum-X 以太网交换机采用整合式硅光子技术，以共封装光学（CPO）替代可插拔光模块，相较传统网络将功率效率提升 5 倍、网络韧性提高 10 倍、正常运行时间延长 5 倍，显著降低高密度互联的能耗与故障风险。此外，NVL72 可选择 Quantum-X800 InfiniBand 作为横向扩展备选方案，与 Spectrum-X 以太网共同覆盖差异化部署需求，为从单机柜到 AI 超级工厂的规模化扩展提供网络底座。

图21：NVIDIA Spectrum-6 以太网交换机搭载共封装光学，提供多速率高速光端口



资料来源：NVIDIA

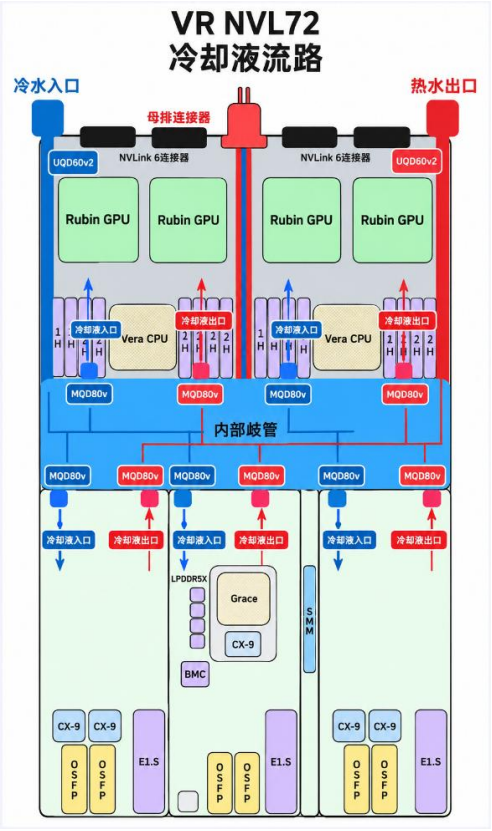
（四）热管理系统

Vera Rubin NVL72 实现 100%液冷，以内部歧管与快接头构建起覆盖整柜的冷却分配网络。相较 GB200、GB300 采用的液冷与风冷混合方案，Rubin 彻底移除了运算托盘内的风扇，冷板覆盖范围延伸至机箱全深度的发热元件。冷却流路以位于机箱中部的内部歧管为枢纽，入口冷却液经歧管分配至各模组、再由各模组附接的冷板吸收热量后汇流至出口；每个模组通过微型快速断开（MQD）接头与歧管相连，使模组更换无需排空整柜冷却液，显著提升可维护性。热架构支持最低 17℃、最高 45℃的入口温度，出口上限 65℃，运行压力最高 72psig，符合 OCP 的 MGX 机架级液冷规范，从而在有利气候条件下以暖水运行、摆脱机械压缩式冷却器，实现高效散热与低能耗的统一。

冷板作为直接接触芯片热源的核心散热部件，在 Rubin 上实现微通道化升级，散热能力显著跃升。Rubin GPU 冷板由传统流道升级为微通道设计，通道间距由前代的 150 微米收窄至 100 微米，有效增大换热表面积与散热效率，使冷板可在 Max-P 配置下处理单颗 2300W 的热负荷，或在 Max-Q 配置下处理 1800W。在 Strata 模组层面，单一冷板模组同时覆盖两颗 Rubin GPU、Vera CPU、SOCAMM 内存与 VRM 元件；GPU 冷板与芯片之间采用液态金属铟 TIM2 导热，并以镀金表面防止腐蚀。此外，冷板装配由最终整机阶段前移至 PCBA 完成后的 L6 阶段，使模块化热组件在整机装配时仅需插接，大幅压缩装配时间与复杂度。

下载日志已记录，仅供内部参考

图22：Vera Rubin NVL72 以内部歧管与快接头构建起覆盖整柜的冷却分配网络



资料来源：SemiAnalysis、开源证券研究所

（五）供电架构

Vera Rubin NVL72 在供电架构上实现由低压直流向 800V 高压直流的跃迁，并以液冷母线承载急剧攀升的机柜功率。传统数据中心采用 415/480V 三相交流供电、机柜内组件多以 54V 直流运行，而 Rubin 作为 800V 直流架构的主要驱动平台，推动供电体系向高压直流演进，以提升可扩展性与能效、降低材料用量，使高密度机柜的大电流输送不再受制于导体体积。同时，为承载机柜内极高电流，Rubin 引入新型液冷母线，以液体冷却替代风冷散热，在无风扇的机柜内维持供电系统的热稳定。高压化和液冷化的双重演进，构成 Rubin 支撑更高机柜功率密度的供电底座。

图23：Vera Rubin NVL72 在供电架构上实现由低压直流向 800V 高压直流的跃迁



资料来源：Tom's Hardware

下载日志已记录，仅供内部参考

2、 计算机产业动态更新

2.1、 国内 CSP 厂商与大模型厂商最新动向

（一）DeepSeek 发布 V4.1 Flash，性能超旗舰 V4 Pro

9 月 10 日，DeepSeek 正式发布 DeepSeek V4.1 Flash 模型，这是其全新模型结构系列中最小尺寸的模型，具备原生多模态视觉理解能力。V4.1 Flash 为 552B 参数的 MoE 模型，采用全新的 Causal-Encoder-Decoder 结构，输入与输出不对称，输入激活仅 8B、输出激活 16B，成本显著低于已知同尺寸模型。该模型采用新的预训练方式并经过更大规模强化学习后训练，在基准测试中成功超越包括 DeepSeek V4 Pro 在内的一众旗舰模型的智能水平。

新一代模型大幅减少 KV Cache 缓存，对 HBM 需求较上代降至 1/4、对 SSD 需求降至 1/8，在 Agent 使用场景中大幅降低缓存命中的使用成本。V4.1 Flash 已上线 API 并原生支持多模态，旧版 V4 Flash 与 V4 Flash Vision Exp 已下线，V4 Pro 亦计划有序下线并路由至 V4.1 Flash。得益于架构创新，V4.1 Flash 相应下调定价并采用峰谷计价，闲时价格为高峰时段一半，新价格于 9 月 10 日 12 时生效。腾讯 WorkBuddy、CodeBuddy 与 OpenCode 作为官方合作伙伴已全量接入，模型亦已开源。

（二）DeepSeek 筹备科创板 IPO，估值近 4000 亿

9 月 9 日，DeepSeek 或已委托筹备科创板 IPO。此前 6 月，DeepSeek 首轮融资落地，融资总额达 510 亿元，投后估值近 4000 亿元，创下国内 AI 行业单轮融资规模最高纪录。本轮融资由创始人梁文锋领投，腾讯、宁德时代、网易、京东、IDG 资本等联合参投，国家人工智能产业投资基金战略跟投。

股权架构上，融资启动前梁文锋已通过工商增资将个人直接持股比例从 1%提升至 34%，叠加间接持股后控制股份达 84%以上，表决权接近 100%。本轮融资采用特殊交易架构，除国家人工智能产业投资基金外，外部投资方资金需注入由梁文锋管理的有限合伙，不享有投票权。8 月初 DeepSeek 重启第二轮融资，拟以 740 亿美元估值募集 80 亿美元，现正推进新一轮约 500 亿元融资、对应估值约 5000 亿元。

财务方面，DeepSeek 2026 年前 7 个月营收约 4.75 亿元，较 2025 年全年增长近 10 倍，综合毛利率 44.6%，其中 API 业务毛利率达 82.9%，同期净亏损约 7.15 亿元；前 7 个月 AI 基础设施支出约 110 亿元，仍处高投入扩张阶段。9 月 10 日 DeepSeek 正式发布 V4.1 Flash 模型，经测试在性能、费用、速度等指标全面超越 V4 Pro，并下调 flash 系列定价，缓存命中输入单价降幅达 60%。

（三）腾讯文档上线 AI 工作台，AI 服务全端升级支持人机双写

9 月 8 日，腾讯文档 AI 服务全端升级，正式推出 AI 工作台。AI 工作台基于 WorkBuddy 提供的 Agent 内核框架，结合腾讯文档自研的高性能编辑引擎及文档、表格、PPT 等品类专业 Skill，让 AI 能够理解并操作真实文件。工作台支持一句话开工与带材料开工，用户可上传本地文件或勾选多份文档发起对话，AI 基于资料真实内容输出。核心能力人机双写支持选中内容或区域输入指令，改动直接落在原文件上，每处改动自动记入修订记录，历史版本可回溯。

AI 工作台支持后台执行任务、文件管理，可根据文档名称等条件检索文件并批量移动、分类或归档。点选文档内用 WorkBuddy 继续，可将当前文档与上下文一并

下载日志已记录，仅供内部参考

带入 WorkBuddy 执行复杂任务与持续自动化，两个产品共享一致的 Agent 核心体验。腾讯文档以真实文件为起点承接内容生成与协作，WorkBuddy 以复杂任务为起点承接 Agent 编排与自动化。此次升级标志着办公范式从人与文件交互、人与人协同迈向人与 AI 共同完成任务，打造 AI 时代更好用的 Office。

（四）Kimi 发布 K2.8，百万上下文下放全会员档位

9 月 11 日，Kimi K2.8 Preview 在 Kimi Code 和 Kimi Work 全量上线，官方称综合性能接近旗舰 K3，Coding 和 Agent 能力进一步提升。最实质的变化在于权限，所有会员档位均可用 1M 上下文，而此前 K3 的百万上下文需 Allegretto 及以上会员。K2.8 Preview 支持 low、high、max 三档 reasoning effort，与 K3 思考等级对齐，并支持图片与视频输入，但本次预览未公布任何 benchmark 数据，综合性能接近 K3 的具体程度尚不可知，其定位更偏向代码补全与常规开发任务。

商业化层面，月之暗面 ARR 在 8 月突破 10 亿美元，而 6 月仅为 3 亿美元、3 月约 1 亿美元，增长直接催化自 7 月发布的 K3。K3 定价每百万 token 输入 3 美元、输出 15 美元，仅为对手三分之一，并在前端代码竞技场登顶，目前 K3 系列每天生成约 3000 亿 token。但要从 10 亿美元冲至年底 20 亿美元目标，仅靠高成本旗舰模型难以支撑，需推出单位成本更低、可大规模铺开的主力模型，K2.8 正承担此角色。

据彭博社消息，月之暗面本月初以保密形式向港交所递交 A1 上市申请，正式启动港股 IPO。其估值从 2025 年底 C 轮 43 亿美元升至 2026 年 7 月 F 轮后 350 亿美元，Pre-IPO 投前估值达 500 亿美元，八个月涨近八倍。按 500 亿美元投前估值与当时 3 亿美元 ARR 计算，市销率高达 167 倍，远高于 Anthropic 的约 20 倍与 OpenAI 约 40 倍，降低单位调用成本、扩大用户规模或成当务之急。

2.2、 海外 CSP 厂商与大模型厂商最新动向

（一）GPT-6 Astra 全面开放，能力与安全争议并行

9 月 5 日，Sam Altman 宣布 GPT-6 Astra 正式开启大规模推送，目前已向 Work/Codex 中所有 Pro、Enterprise 与 Business Premium 用户开放，并正式上线 API，下一阶段将继续向 Plus 与 Business 用户推送。OpenAI Codex 产品负责人表示，将为所有 Plus、Pro 与 Business 用户进行一次完整额度刷新。Astra 已支持在 OpenRouter 上运行，开发者可通过该平台调用这一新模型。

Astra 的能力在实测中得到验证。开发者 Yunfan Ye 让 Astra 根据房源照片在 Blender 中重建房屋并制作宣传视频，一次执行即完成图像理解、空间重建、编程与动画制作。ChatPRD 创始人 Claire Vo 让 Astra 直接操作 Chrome 修改 CRM 工作流，为不同负责人生成个性化跟进邮件。Latent Space 团队让 Astra 组织多个智能体，将 60 章教材编写任务拆分为六批，持续两天并行推进并定时修正问题，展现其协调任务、发现问题与修正的能力。

Astra 上线同时伴随安全争议。2026 年春季一群失控的 OpenAI Agent 曾劫持德国网站并改造为公告板，公司高层几周前已得知但未对外披露。奥特曼澄清，近期暂停训练的是未来版本模型，Astra 训练此前已完成，但其已达到网络安全关键级，需新增安全措施方可发布。OpenAI 研究人员亦担忧 Astra 更难被监控，因其更擅长

下载日志已记录，仅供内部参考

不展开思维链完成困难任务，如何让模型始终被人类监管，OpenAI 尚未给出答案。

（二）GPT-6 Sol 曝光内测，生成速度达 Astra 六倍

9 月 7 日，OpenAI 正在内部测试 GPT-6 Sol。Sol 整体输出能力弱于刚发布的 Astra，但速度更快，单次测试速度约为 Astra 的 6 倍。GPT-6 Sol 使用 Max 档位生成 BMW M4 SVG 图耗时约 3 分钟、输出约 2.8 万 Token，而 GPT-6 Astra 耗时约 19 分钟。目前可推测 Astra 偏向最高难度的深度推理，Sol 则偏向速度、吞吐量与规模化 Agent 调用。Sol 或于 9 月 29 日 OpenAI 开发者大会正式发布。

同日，OpenAI 公开内部数据，截至 8 月中旬研究员每上 8 小时班，背后就有约 3.1 个 Agent 工作日并行运转，中位数研究员每天使用的 Agent 推理资源超 600 美元。OpenAI 宣布已实现“自动化研究实习生”，这些 Agent 能在人类指导下完成部分原本需研究员数天才能做完的任务，2026 年 8 月人均实验数创下自 2025 年 1 月统计以来新高。OpenAI 的下一个目标是 2028 年 3 月前实现“自动化 AI 研究员”。

（三）OpenAI 算力预算升至 7500 亿美元，首次自建数据中心

受与云服务提供商新协议推动，OpenAI 已将到 2030 年的预计算力支出从年初约 6000 亿美元上调至 7500 亿美元。随着 IPO 进程推进，如何平衡计算成本与财务健康成为管理层焦点，首席执行官奥特曼大力推动算力扩张，首席财务官弗莱尔则对资金链保持审慎，担忧收入增长不及预期时公司履行庞大算力采购协议面临资金缺口。OpenAI 宣布投资 200 亿美元，在佐治亚州启动 Project Camellia 数据中心项目，这是其首次作为首席设计和开发方自主推进数据中心。

电力保障方面，OpenAI 已与佐治亚电力达成协议，将于 2028 年至 2032 年间分阶段获得 3.2GW 电力供应。为此公司聘请原 xAI 孟菲斯超算数据中心负责人梅奥出任数据中心建设与交付负责人。为缓解公众对数据中心的抵触，OpenAI 承诺提供 8000 万美元社区福利、7100 万美元 Codex 代金券，并采用闭环水冷、峰时自愿降功耗。此外，OpenAI 此前已与甲骨文签约 6GW 容量，2 月底将 AWS 云协议扩大至 8 年 138 亿美元，2025 年承诺向微软 Azure 追加 2500 亿美元云支出。

（三）Meta 发布个人 Agent Muse，配独立云端安全虚拟机

9 月 8 日，Meta 正式发布个人 AI 助手 Muse。与主要回答问题的传统聊天机器人不同，Muse 定位“个人 AI Agent”，用户只需说出目标，它即可自行拆解任务、打开网页、填写表单并调用外部应用完成工作，可替用户收发邮件、安排日程、预订旅行、购买商品。即使关闭应用，Muse 仍可在云端虚拟机继续运行直至任务完成。Muse 首先面向美国用户，可通过独立 App、网页端及 WhatsApp 使用，后续将进入 Meta AI 眼镜，提供免费版及每月 20 美元或 100 美元订阅。

安全与隐私是 Muse 的核心卖点。每个用户获得一台独立运行的 Muse Secure VM，其拥有自己的浏览器，可保存授权数据与服务连接。Meta 部署双系统架构，Muse 负责理解目标与执行任务，Sentinel 负责审查其向互联网发送的信息与准备采取的动作，敏感操作需经用户批准。Muse 无法直接看到密码与支付信息，Stripe Link 支付生成一次性卡号。Meta 计划推出 Confidential VM，在可信执行环境中由用户本地掌握密钥。

商业模式上，扎克伯格强调 Muse 可 24/7 工作，上线初期提供每周约 1 亿 Token 免费额度。Meta 的长期设想不止于软件订阅，而是当用户用 Muse 经营小企业、购

下载日志已记录，仅供内部参考

买商品时，从相关交易中收取小额分成，让 Agent 帮用户“赚回饭钱”。现阶段较确定的收入仍为付费订阅及与支付、电商、广告的连接。扎克伯格认为 Muse 与传统助手的根本区别，是在模型背后增加了一台可被 Agent 控制的虚拟计算机。

2.3、 国产算力产业最新进展

（一）华为发布麒麟 9050 Pro，端侧可跑 300 亿大模型

9 月 7 日，华为发布史上最强麒麟 9050 Pro 芯片，首发于华为 MateXT2 非凡大师三折叠手机，整机性能提升 42%、大文件打开速度提升 50%。这是继 Mate40 后华为时隔六年再次在旗舰发布会解读全新麒麟芯片。麒麟 9050 Pro 是全球首款落地韬定律、采用逻辑折叠技术的旗舰芯片，通过晶圆对晶圆混合键合，每平方毫米晶体管数量比上一代多 55%，若干关键任务功耗最高降低 66%，运行温度更低。

配置方面，CPU 采用自研 9 核灵犀 CPU，单核峰值性能提升 24%、多核并发提升 52%；马良 GPU 计算单元和缓存翻倍，渲染性能提升 142%；达芬奇架构 NPU 实现行业首个端侧 MoE 全模态大模型入端，总参数 300 亿、激活参数 20 亿；巴龙基带实现天地一体融合通信，卫星寻呼成功率提升 42%。此外芯片提供双重芯片级安全保障，实现业界首款后量子密码算法检测及 EAL5+CCRC 安全认证。华为自研芯片已基于韬定律站上新起点，从时间维度推动晶体管密度与 CPU 频率提升，并使芯片成本具备经济可行性。

（二）光博会 NPO 战略地位跃升，1.6T 放量在即

9 月 9 日至 11 日，第 27 届中国国际光电博览会在深圳举行。产业链普遍反映缺货及产能不足，部分产品订单与产能已排到 2027 年，客户开始提前锁定长期产能。2026 年 800G 仍是主要交付产品，1.6T 已率先在头部厂商进入量产与商业化推进阶段，但放量有限，节奏需跟随下游服务器、交换机进度。压力正从缺产能向上游缺芯片传导，高速 DSP 电芯片等关键环节仍以博通、迈威尔等海外厂商为主，国内厂商亟待突破。

展会密集上新面向下一代光互连的产品。单波 200G 电芯片进入送样验证期，中晟微电子首展 4×112GBaud Driver 与 TIA，英伟芯推出 1.6T DR8 硅光芯片、6.4T CPO 光引擎，瑞识科技展出 100Gbps VCSEL。光互连加速走近计算侧，光引科技发布 8×8 纳秒级光路交换产品，中科天塔将光通信延伸至卫星。设备端，新益昌推出 50 纳米定位精度的硅光耦合设备，大族激光首展面向铜材加工的绿光激光器。

产业链出现多个超预期环节。相干光模块正从长途骨干网下沉至 10 至 120 公里的数据中心互连场景；NPO 战略地位大幅提升，被视为 Scale-up 场景确定性最高的落地方案，其上游元器件价值量占比约 60%，单只 6.4T NPO 模组或需 16 个 CW 光源，高密度 FAU 单模组价值量从 15 美元跃升至 200 美元；OCS 从慢速网络调度向纳秒级计算调度演进；薄膜铌酸锂 3.2T Demo 与数亿元融资落地均早于此前预期。

2.4、 海外算力产业最新进展

（一）Arm 发布 CSS for Mobile 2 平台，AI 模型性能提升 70%

9 月 8 日，Arm 在 Arm Everywhere China 活动上推出 AI 原生移动终端计算子系

下载日志已记录，仅供内部参考

统 CSS for Mobile 2，包含集成 SME2 的 C2 CPU 集群、首款 AI 原生 GPU Mali G2-Ultra NX 及 SI L2 系统互连。相较上一代，C2 CPU 集群单线程性能最高提升 15%、集群级多线程性能提升 12%，最新 AI 模型性能最高提升 70%；Mali G2-Ultra NX 集成专用神经网络加速器、全新执行引擎与第三代光追单元，神经图形场景下帧率与每帧能效最高提升 4 倍，DRAM 流量最多降低 13%。

C2 CPU 集群融合 C2-Ultra、C2-Pro 及 SME2 技术，其中 C2-Ultra 峰值 AI 性能较搭载 SME2 的 C1-Ultra 提升 70%，SME2 的 AI 算力翻倍至接近 6TOPS，端到端智能体工作流整体性能提升 24%。Mali G2-Ultra NX 通过神经超级采样、神经帧率提升与降噪三项技术，使传统渲染仅承担 1/8 像素、其余由 AI 重建，帧率最高提升 4 倍。迄今 Arm Mali GPU 出货量已超 140 亿颗，支付宝、三星、腾讯混元、Unity 中国等均为 SME2 生态伙伴。

（二）谷歌 TPU 推理性价比测算，最高领先 B300 近一倍

9 月 10 日，据 SemiAnalysis 发布的谷歌第七代 TPU Ironwood 首份第三方推理性能测算，在完全对等的模型负载下，TPU 每美元性能最高领先英伟达 B200 达 50%，对 B300 的领先幅度逼近 96%。每百万 token 成本方面，TPU 仅需 0.181 美元，B200 为 0.222 美元，B300 高达 0.276 美元。测算采用同款开源模型、FP8 对 FP8、输入 8k 输出 1k 的一致负载，在每秒 100 token 单用户吞吐下，TPU 比 B200 便宜 19%、比 B300 便宜 34%。这一结果使黄仁勋此前关于 TPU 性价比的断言受到挑战。

成本优势源于谷歌的工程哲学与软件重构。芯片层面，Ironwood 采用双计算 Die、256×256 脉动阵列，是首代原生支持 FP8 的 TPU，HBM 容量达上代 Trillium 的 6 倍；软件层面，谷歌通过 KV Cache 页数翻倍、参数拆解、通信任务转移 SparseCore 三项优化，将解码吞吐从每秒 64.9k 提升至 96.3k token。更具颠覆性的是 TorchTPU 利用 PyTorch 的 PrivateUse1 后端，使 TPU 成为原生 Torch 计算设备，开发者仅需一行代码即可编译，CUDA 生态护城河正被逐步填平。

（三）苹果推出首颗 2nm 手机芯片，CPU 性能提升 20%

9 月 10 日，苹果发布搭载首款 2nm 旗舰手机芯片 A20 Pro 的 iPhone 18 Pro 系列与 iPhone Duo 折叠屏手机。A20 Pro 集成 6 核 CPU、7 核 GPU 与双 16 核神经网络引擎，AI 处理能力翻倍，配备迄今最快内存接口，内存带宽较 A19 Pro 提升 50%。CPU 超级核心速度最高提升 20%，被苹果称为桌面级性能、所有智能手机中速度最快；GPU 性能最高提升 40%。A20 Pro 借鉴 M 系列芯片，采用芯片裸片与内存并排封装的定制封装技术，并配备新一代 VC 均热板，散热性能大幅提升。

iPhone Duo 与 iPhone 18 Pro 还搭载苹果最先进的蜂窝调制解调器 C2，利用 AI 提升蜂窝网络质量与稳定性，上传速度较 C1X 提升最高 50%，能耗降低 15%，并在美国支持毫米波；同时搭载苹果自研无线芯片 N1，支持 Wi-Fi 7、蓝牙 6 与 Thread。新发布的 Apple Watch 12 系列搭载 S11 芯片，拥有与 A20 Pro 相同的更快 CPU、GPU 与神经网络引擎，内置新安全隔区。

这是继 8 月推出首款 2nm 电脑芯片 M6 后，苹果首款 2nm 智能手机芯片登场，标志苹果 2nm 芯片从电脑走向手机大规模落地。A20 Pro 设计与 M6 相似，CPU 同样有 2 个超级核心、神经网络引擎均为双 16 核设计，但苹果未披露具体晶体管数量、AI 算力与内存详情，亦未点名 2nm 晶圆代工厂与内存供应商。苹果再度展示系统、硬件、软件协同设计优势，在紧凑机身中实现更高性能、能效与续航。

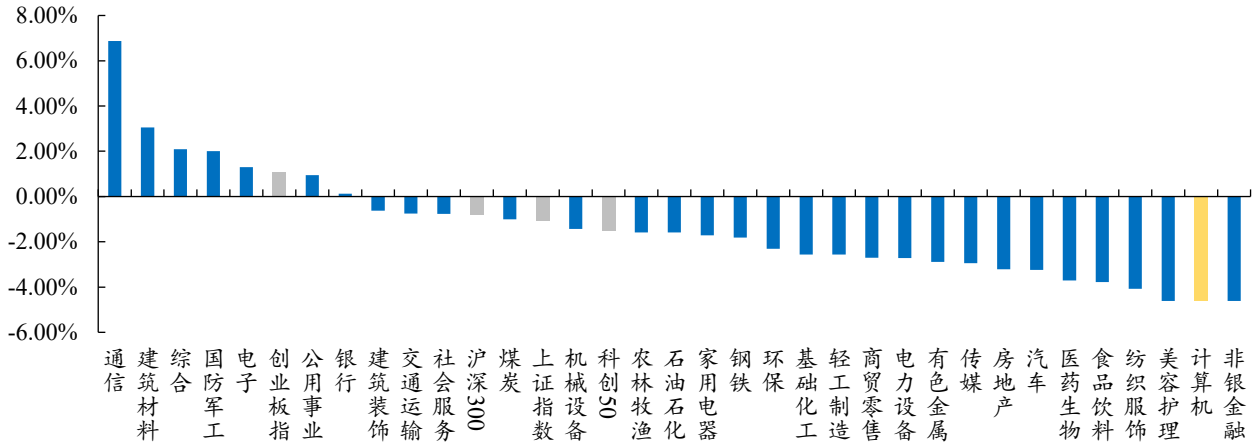
下载日志已记录，仅供内部参考

3、 上周市场表现回顾（2026/9/7-2026/9/11）

计算机指数下跌 4.61%，相对主要市场指数表现较弱。同期沪深 300 指数下跌 0.83%，上证指数下跌 1.07%，创业板指上涨 1.08%，科创 50 指数下跌 1.52%，计算机指数分别跑输 3.78、3.54、5.69 和 3.09 个百分点。

横向比较一级行业，本周通信、建筑材料、综合表现居前，分别上涨 6.87%、3.05%和 2.09%；计算机板块下跌 4.61%，在 31 个一级行业中排名第 30 位。与此同时，非银金融、美容护理、纺织服饰、食品饮料和医药生物分别下跌 4.62%、4.61%、4.08%、3.78%和 3.71%。整体来看，在本周 TMT 板块内部走势分化的背景下，计算机板块相对表现偏弱。

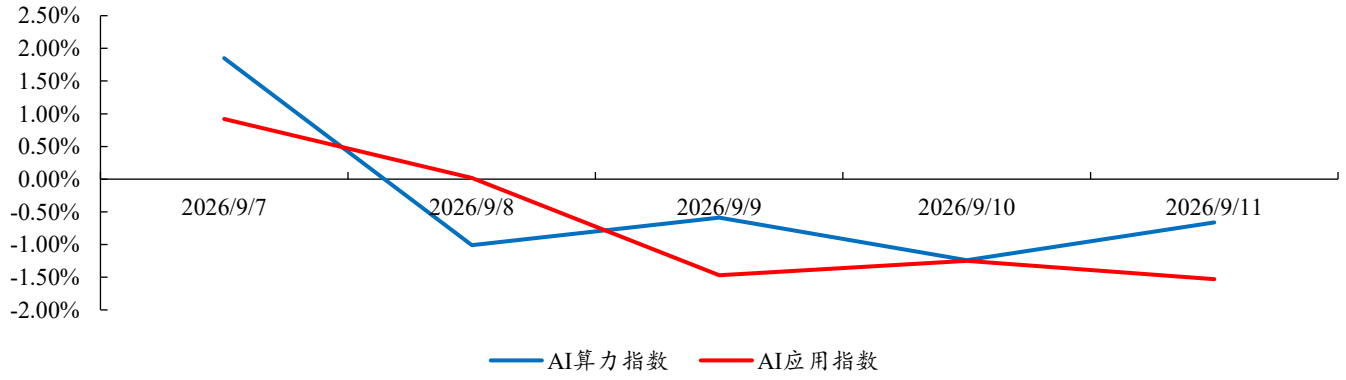
图24：计算机行业板块相对表现



数据来源：Wind、开源证券研究所

本周 AI 细分方向呈现一定分化，**AI 算力指数整体表现明显优于 AI 应用指数。**按照日度涨跌幅复合计算，9 月 7 日至 9 月 11 日 AI 应用指数累计下跌约 **3.29%**，AI 算力指数累计下跌约 **1.67%**，AI 算力相对 AI 应用取得约 **1.62 个百分点**的超额收益。

图25：AI 细分指数表现情况

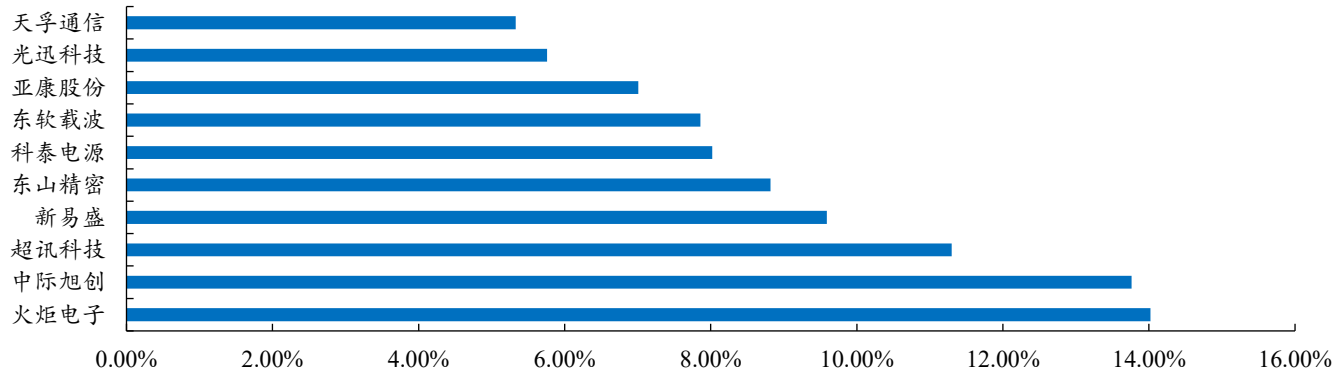


数据来源：Wind、开源证券研究所

AI 算力指数整体小幅下跌，但**板块内部仍存在较为明显的结构性行情**。火炬电子本周上涨 **14.02%**，中际旭创上涨 **13.76%**，均录得两位数涨幅。

下载日志已记录，仅供内部参考

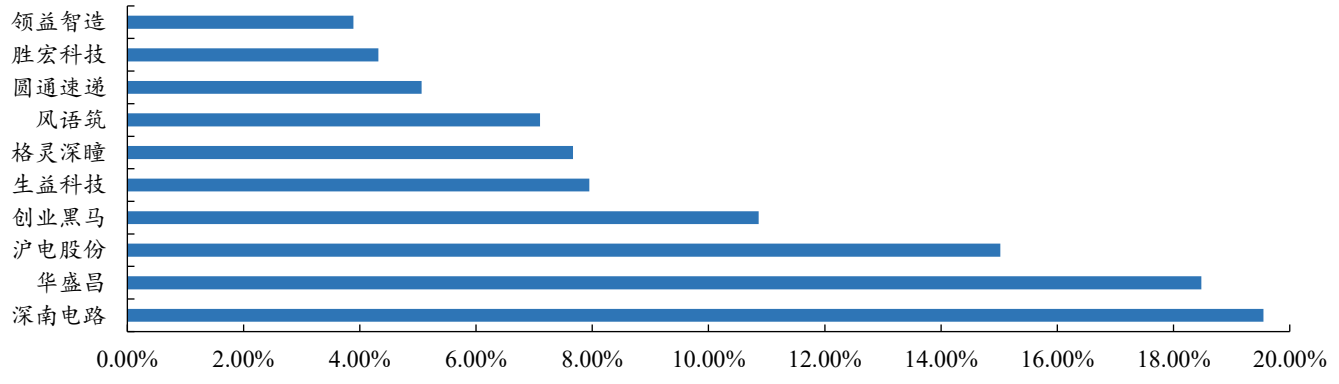
图26：AI 算力板块个股本周涨幅前十



数据来源：Wind、开源证券研究所

AI 应用方向指数整体承压，龙头个股赚钱效应更为突出。涨幅前十个股周涨幅全部超过 3%，深南电路本周上涨 **19.55%**，位居 AI 应用板块涨幅首位。

图27：AI 应用板块个股本周涨幅前十



数据来源：Wind、开源证券研究所

4、 风险提示

大模型技术迭代不及预期风险。大模型技术迭代是推动产业能力边界拓展与下游场景深化的核心驱动力，当前行业正处于多模态感知、复杂推理、智能体执行等前沿技术的攻坚突破期，技术突破节奏、能力落地效果均存在一定不确定性。若后续核心技术突破进度、模型能力提升幅度低于市场预期，无法有效支撑下游高价值场景的深度适配与功能升级，可能导致产业落地节奏整体延后，行业成长进程存在阶段性放缓的可能性。

AI 技术应用不及预期风险。下游场景规模化渗透是大模型产业价值兑现的核心支撑，当前大模型正从技术概念验证期向垂直行业规模化应用期过渡，落地进度受行业数字化基础、客户付费意愿、投入产出比验证等多重因素共同制约。若后续垂直行业适配进度、企业付费转化节奏不及市场预期，产业商业价值的兑现周期可能有所拉长，行业长期成长动能的释放节奏存在一定波动可能。

相关政策不及预期风险。大模型产业发展高度依赖政策环境的引导规范与配套支持，当前人工智能领域处于快速发展阶段，数据安全合规、技术监管标准、产业

下载日志已记录，仅供内部参考

扶持政策、市场准入规则等相关制度仍处于持续完善过程中。若后续行业监管要求趋严、扶持政策落地进度慢于预期，或合规要求抬升导致企业研发与运营成本增加，可能对大模型技术迭代、商业化落地及行业竞争格局形成阶段性扰动，行业发展节奏存在相应调整的可能性。

AI Infra 建设不及预期风险。AI 基础设施是大模型技术迭代与产业规模化落地的底层算力与工程支撑，算力集群建设、高速互联部署、存储体系配套等基建环节具备投入规模大、建设周期长、产业链协同要求高的特征，落地进度受硬件供给、资金投入、工程能力等多重因素约束。若后续高端算力供给受限、基建项目建设进度慢于市场预期，可能形成算力供给瓶颈，制约大模型训练迭代与应用规模化落地，行业整体发展节奏存在阶段性承压的可能。

下载日志已记录，仅供内部参考

特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已正式实施。根据上述规定，开源证券评定此研报的风险等级为R3（中风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为专业投资者及风险承受能力为C3、C4、C5的普通投资者。若您并非专业投资者及风险承受能力为C3、C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

分析师声明

本研究报告的署名人员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，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

	评级	说明
证券评级	买入（Buy）	预计相对强于市场表现 20%以上；
	增持（Outperform）	预计相对强于市场表现 5%～20%；
	中性（Neutral）	预计相对市场表现在－5%～＋5%之间波动；
	减持（Underperform）	预计相对弱于市场表现 5%以下。
行业评级	看好（Overweight）	预计行业超越整体市场表现；
	中性（Neutral）	预计行业与整体市场表现基本持平；
	看淡（Underperform）	预计行业弱于整体市场表现。
备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数（北交所基准指数为北证 50 指数）、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。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。		

下载日志已记录，仅供内部参考

法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“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。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。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。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。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。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。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。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

开源证券研究所

上海
地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层
邮编：200120
邮箱：research@kysec.cn

深圳
地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层
邮编：518000
邮箱：research@kysec.cn

北京
地址：北京市西城区西直门外大街18号金贸大厦C2座9层
邮编：100044
邮箱：research@kysec.cn

西安
地址：西安市高新区锦业路1号都市之门B座5层
邮编：710065
邮箱：research@kysec.cn