

昇腾AI：引领“超节点+集群”时代

——AI算力“卖水人”系列（8）

评级：推荐(维持)

刘熹(证券分析师)
S0350523040001
liux10@ghzq.com.cn

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

最近一年走势



相对沪深300表现

表现	1M	3M	12M
计算机	-5.1%	15.6%	69.3%
沪深300	2.2%	15.3%	28.3%

相关报告

《计算机行业专题研究：计算需求演进，超节点成为AI基础设施共识（推荐）*计算机*刘熹》——2025-09-24

《从Blackwell到Rubin：计算、网络、存储持续升级——AI算力“卖水人”专题系列（7）（推荐）*计算机*刘熹》——2025-09-17

《从Tokens角度跟踪AI应用落地进展——计算机行业大模型及AI应用专题（推荐）*计算机*刘熹》——2025-09-15

本篇报告解决了以下核心问题：1、华为新一代昇腾处理器、超节点、集群方案的参数及其与英伟达对比；2、华为昇腾产业链的核心构成；3、当前AI算力行业的需求端变化，华为昇腾如何受益于AI芯片国产化。

◆ **核心逻辑：华为昇腾全栈软硬件平台可对标英伟达，面向“超节点+集群”时代，在AI与国产化双轮驱动下，华为昇腾业务增长有望超预期。**

- **昇腾处理器：一年一代，引领“超节点+集群”新范式**

华为推出基于达芬奇架构的NPU，是全球首个覆盖全场景的AI芯片。华为推出基于Ascend 910C打造的CloudMatrix 384集群解决方案；截至2025年9月18日，华为Atlas 900 A3 SuperPoD超节点累计部署300多套。华为公布昇腾路线图，2026年将发布昇腾950、Atlas 950 SuperPoD、通算超节点Taishan 950 SuperPoD及企业级风冷AI超节点Atlas 850，2027年将发布昇腾960及Atlas 960 SuperPoD，基本保持每年一代、算力翻倍。华为超节点基于灵衢互联协议，实现多种组件资源池化和平等协同，目前灵衢1.0已正式在Atlas 900 A3 SuperPoD上商用验证，2025年9月，华为全面开放灵衢2.0。面向“超节点+集群”时代，华为网络、CPU、系统能力等方面优势有望进一步显现。

- **昇腾生态：CANN全面开源开放，联合产业链上下游共建生态**

1) **软件：**2025年8月，华为宣布昇腾硬件使能CANN全面开源开放，Mind系列应用使能套件及工具链全面开源。CANN计算架构可对标英伟达CUDA核心软件层。通过软硬件对标，华为昇腾致力于成为全球AI算力第二选择。

2) **超节点/服务器：**昇腾推出自有Atlas硬件（AI加速卡、AI服务器等），也向伙伴开放自有硬件用于集成与二次开发，并且昇腾部分整机伙伴背靠地方国资，具备良好资源优势。此外，昇腾服务器也获科大讯飞、美团、紫东太初等科技产业积极适配。

3) **超节点/服务器价值量：**AI芯片价值量占比最高（约70%），未来，光模块、液冷、电源、存储、PCB等零部件价值量成长性同样较强。

- **算力行业：CSP、主权国家、AI推理等需求向上，美国制裁升级推动AI芯片国产化**

需求侧：英伟达预计未来五年全球AI资本支出将达3-4万亿美元。未来，互联网、主权国家、原生AI厂商、AI推理等环节需求仍有望持续增长。甲骨文表示“AI推理市场将“远大于”AI训练市场”。主权AI正在崛起，鸿海预计未来五年主权AI领域投资有望超1万亿美元。

国产化：2024年，中国加速芯片的市场规模增长迅速，超过270万张，中国本土人工智能芯片品牌的出货量已超过82万张，市占率30%左右。2025年，中美关税摩擦经历多轮升级，8月，英伟达下令停止生产H20，我国AI算力自主可控将成为大势所趋。未来，昇腾凭借优秀的软硬件产品性能及产业生态，有望持续受益AI+国产化浪潮。

大模型训推带动AI算力需求增长，AI芯片、CPU、AI服务器及其组件、光模块、液冷、IDC等环节有望持续受益。维持对计算机行业“推荐”评级。

● 相关公司

1) **AI芯片**：海光信息、寒武纪、百度集团、英伟达、AMD。

2) **CPU**：海光信息、龙芯中科、中国长城、Intel、AMD。

3) **服务器整机**：工业富联、中科曙光、浪潮信息、华勤技术、紫光股份、神州数码、中国长城、软通动力、烽火通信、拓维信息、鸿海、纬创、广达、超微电脑、戴尔。

4) **服务器组件**：①**散热**：曙光数创、英维克、飞荣达、同方股份、申菱环境、高澜股份、奇鋕科技、双鸿、健策、VERTIV；②**铜连接**：安费诺、沃尔核材、华丰科技；③**HBM**：SK海力士、三星、美光、赛腾股份、联瑞新材；④**电源**：欧陆通、麦格米特、中国长城。

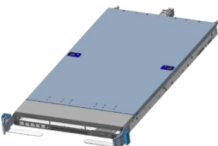
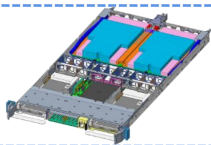
5) **算力租赁**：协创数据、宏景科技、有方科技、鸿博股份。

6) **数据中心**：云赛智联、奥飞数据、润泽科技、大位科技、光环新网、数据港、科华数据、万国数据、城地香江。

风险提示：宏观经济影响下游需求、大模型产业发展不及预期、市场竞争加剧、中美博弈加剧、相关公司业绩不及预期、各公司并不具备完全可比性，对标的相关资料和数据仅供参考。

投资建议与相关公司

- ☆ **ASIC**
 - 内地: 阿里巴巴、百度集团、芯原股份
 - 海外: 博通、Marvell
- ☆ **GPU**
 - 内地: 华为海思、海光信息、寒武纪
 - 海外: 英伟达、AMD、Intel
- ☆ **CPU**
 - 内地: 华为海思、海光信息、飞腾、龙芯中科
 - 海外: 英伟达、AMD、Intel
- ☆ **主板**
 - 内地: 沪电股份、胜宏科技、深南电路、生益科技
 - 中国台湾: 泰安电脑(神达)、技嘉、华擎、华硕、英业达、纬创、研华
 - 美国: Intel、Supermicro
- ☆ **Compute Tray * 10**
 - 内地: 工业富联
 - 中国台湾: 纬创
- ☆ **Switch Tray * 9**
 - 内地: 工业富联
 - 中国台湾: 纬创
- ☆ **Compute Tray * 8**
- ☆ **电源**
 - 台达、中国长城、欧陆通、麦格米特、泰嘉股份
- ☆ **AIDC**: 润泽科技、世纪互联、万国数据等



- ☆ **存储 NAND**
 - 公司: 长江存储、兆易创新、佰维存储、东芯股份
 - 海外: 三星、西部数据、铠侠、SK海力士、美光

- ☆ **光模块**
 - 中际旭创、新易盛、光迅科技、天孚通信、华工科技

- ☆ **算力租赁**
 - Coreweave、协创数据、有方科技、宏景科技

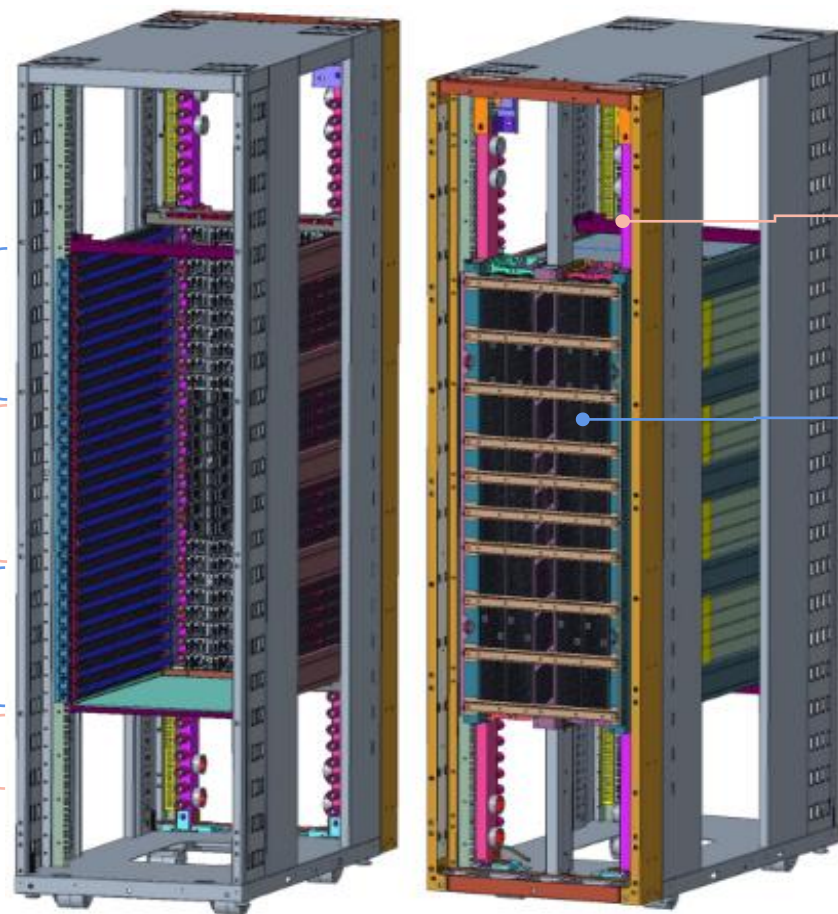
- ☆ **网络**: 思科、华为、英伟达、Arista、紫光股份、中兴通讯、星网锐捷、盛科通信

- ☆ **散热(风冷/液冷)**
 - 内地: 曙光数创、飞荣达、中航光电、英维克、同飞股份、申菱环境、高澜股份、依米康
 - 海外: 奇鋌、双鸿、健策、建准、高力、VERTIV

- ☆ **线缆(铜连接)**
 - 内地: 沃尔核材、华丰科技、兆龙互连、鼎通科技、立讯精密、神宇股份
 - 海外: 安费诺

- ☆ **整机(OEM/ODM)**
 - 1) 内地: 工业富联、浪潮信息、中科曙光、紫光股份、中国长城、华勤技术、神州数码、拓维信息、烽火通信、软通动力; 2) 中国台湾: 鸿海、广达、纬创、英业达、纬颖; 3) 美股: 戴尔、超微电脑

- ☆ **BIOS/BMC**
 - BIOS/BMC: 1) 内地: 卓易信息; 2) 中国台湾: 新唐、系微、信骅



正面

背面

注: 蓝字为零部件。上图展示以英伟达GB300 NVL72整机柜为例, 仅列示了部分参与公司

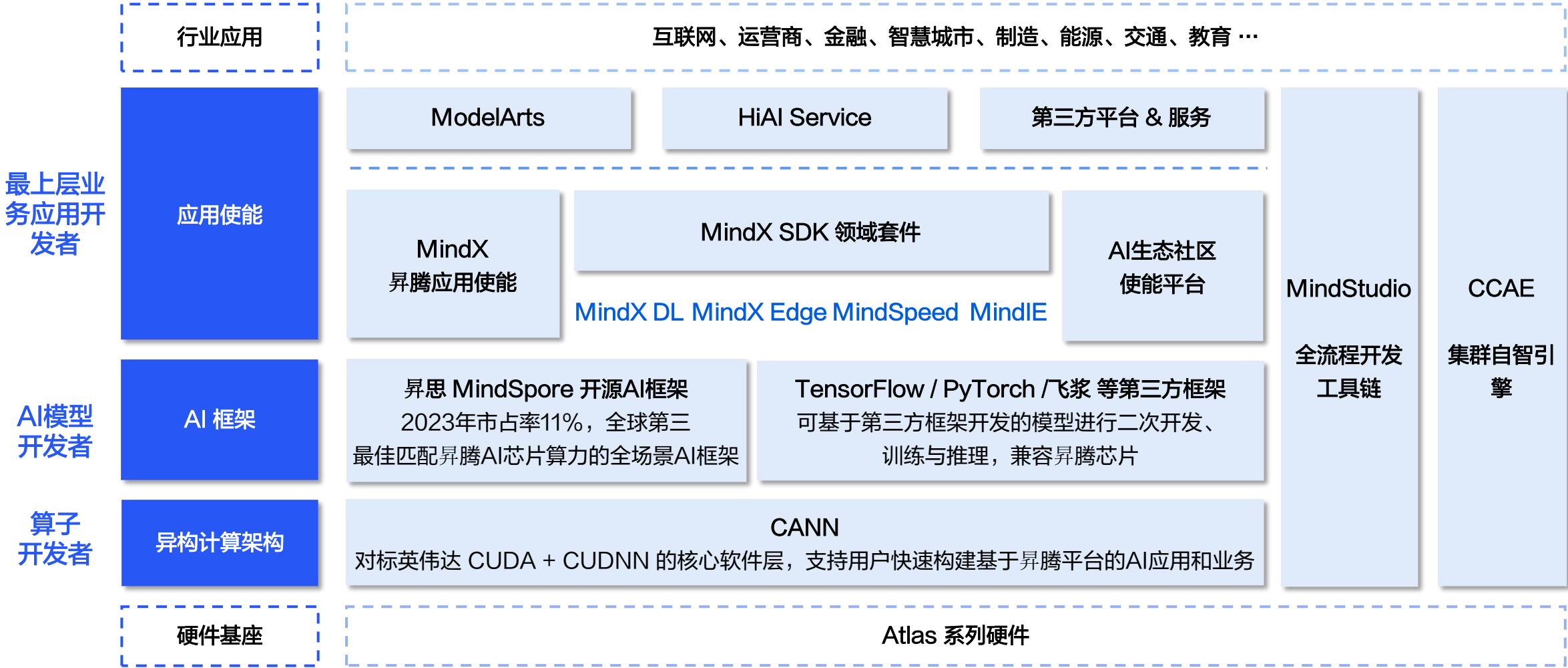
第一章 昇腾芯片

一年一代、算力翻倍，910C CM384量产

1.1 昇腾芯片：基于华为达芬奇架构，是昇腾计算的核心基座

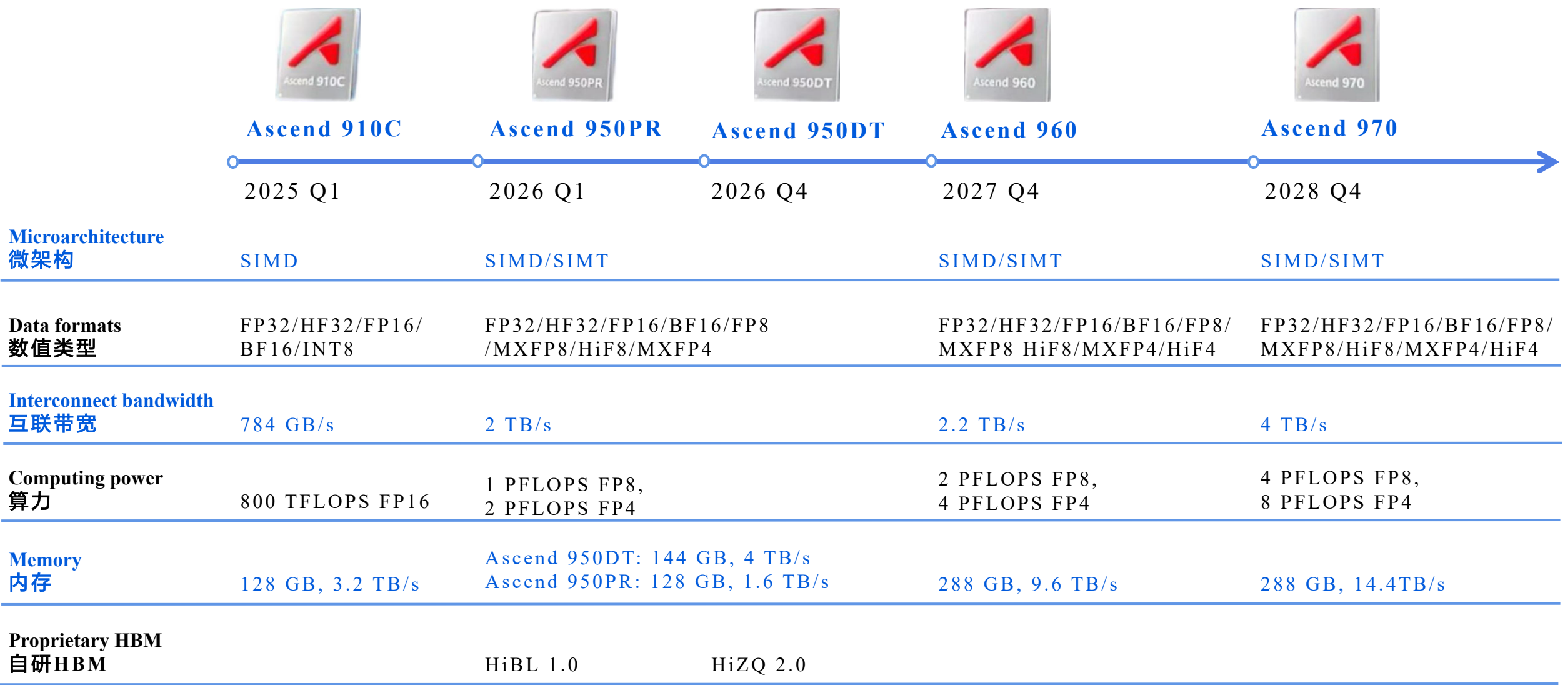
- 昇腾处理器（HUAWEI Ascend）是基于华为达芬奇架构的NPU。NPU（神经网络处理器）针对矩阵运算专门优化设计，昇腾AI芯片是构建昇腾计算产品、使能上层软件 and 应用的底座，有高算力、高能效、灵活可裁剪等特性。

图：昇腾 AI 基础软硬件平台



1.1 昇腾芯片：更易用、更多数值类型、更高带宽、更大算力

图：昇腾AI保持每年一代新品的频率持续迭代



1.1 昇腾计算：面向“端+边+云”的全场景AI基础设施

- 基于昇腾系列AI处理器，华为Atlas人工智能计算解决方案通过模块、板卡、小站、服务器、集群等丰富的产品形态，打造面向“端、边、云”的全场景AI基础设施方案，覆盖深度学习领域推理和训练全流程。

图：昇腾计算产品布局（部分）

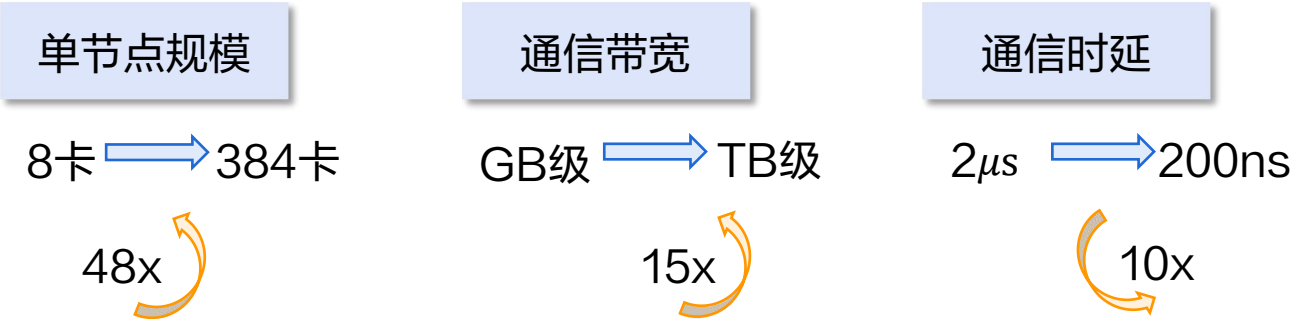
云	Atlas 300 AI 加速卡	Atlas 800 AI 服务器	Atlas 900 AI集群	端	Atlas 200I A2 加速模块
	Atlas 300I Pro 推理卡  - 算力：最高140 TOPS INT8，70 TFLOPS FP16 - CPU：8 core * 1.9 GHz - 功耗：最高72W	Atlas 800T A2训练服务器  4U AI服务器 - CPU：4*鲲鹏920 - 风冷 - 用于深度学习模型开发和训练	Atlas 900 A2 PoD  47U 机柜 - CPU：鲲鹏920 * 32 - 液冷 - 功耗：50.5kW		Atlas 200I A2 加速模块  - 算力：20 TOPS INT8，10 TFLOPS FP16 - CPU：4 core* 1.6 GHz - 功耗：25W
	Atlas 300I Duo 推理卡  - 算力：最高280 TOPS INT8、140 TFLOPS FP16 - CPU：16 core * 1.9 GHz - 功耗：150W	Atlas 800I A2推理服务器  4U AI服务器 - CPU：4*鲲鹏920 - 内部网络：可选NPU全互联机型，整机互联带宽392GB/s - 用于生成式大模型推理，风冷	Atlas 900 AI 集群  数千颗昇腾训练处理器，整合HCCS、PCIe 和 RoCE 三种高速接口	边	Atlas 500 智能小站  - 算力：20 TOPS INT8，10 TFLOPS FP16 - CPU：4 core* 1.6 GHz - 功耗：32.3-44.5W

1.2 昇腾Atlas 900 AI集群：打造首个万卡AI集群

- 昇腾384超节点（Atlas 900 A3 SuperPoD），华为 CloudMatrix 的核心愿景是构建一个「万物皆可池化、平等对待、自由组合」的 AI 原生数据中心。昇腾超节点通过总线技术实现384个NPU之间的大带宽低时延互联，解决集群内计算、存储等各资源之间的通信瓶颈，通过系统工程的优化，实现资源的高效调度，让超节点像一台计算机一样工作。
- 截至2025年9月18日，华为Atlas 900 A3 SuperPoD超节点累计部署300多套，服务于互联网、金融、运营商、电力、制造等行业的20多个客户。

Atlas 900 A3 SuperPoD

累计部署300+套，服务20+客户



1.2 昇腾Atlas 900 AI集群：打造首个万卡AI集群

- 昇腾384超节点由12个计算柜和4个总线柜构成，依托华为在ICT领域深厚的技术与工程经验，通过最佳负载均衡组网方案，该超节点可进一步扩展为包含10万卡的Atlas900 SuperCluster超节点集群，为未来更大规模的模型演进提供有力支撑。
- 昇腾384集成了384颗昇腾910C NPU和192颗鲲鹏CPU的庞大超级节点，采用全互连拓扑架构，可提供300PFLOPS的密集BF16计算能力，几乎是GB200 NVL72的两倍，也具备超过3.6倍的总内存容量和2.1倍的内存带宽。CloudMatrix384设计了三个互补的网络平面：
 - ✓ **UB平面**：核心的Scale-Up网络，以全互联（All-to-All）拓扑连接所有NPU和CPU。每颗昇腾 910C 贡献超过392 GB/s的单向带宽。它专为TP、EP等细粒度并行以及内存池的快速访问而设计。
 - ✓ **RDMA平面**：用于超级节点间的Scale-Out通信，采用RoCE协议，确保与现有生态兼容。NPU 是该平面的唯一参与者，用于KV Cache在Prefill和Decode节点间的传输、分布式训练等。
 - ✓ **VPC平面**：通过华为自研的擎天卡连接到数据中心网络，负责管理、控制、访问持久化存储等。

图：昇腾CloudMatrix 384 对比 英伟达GB200 NVL72

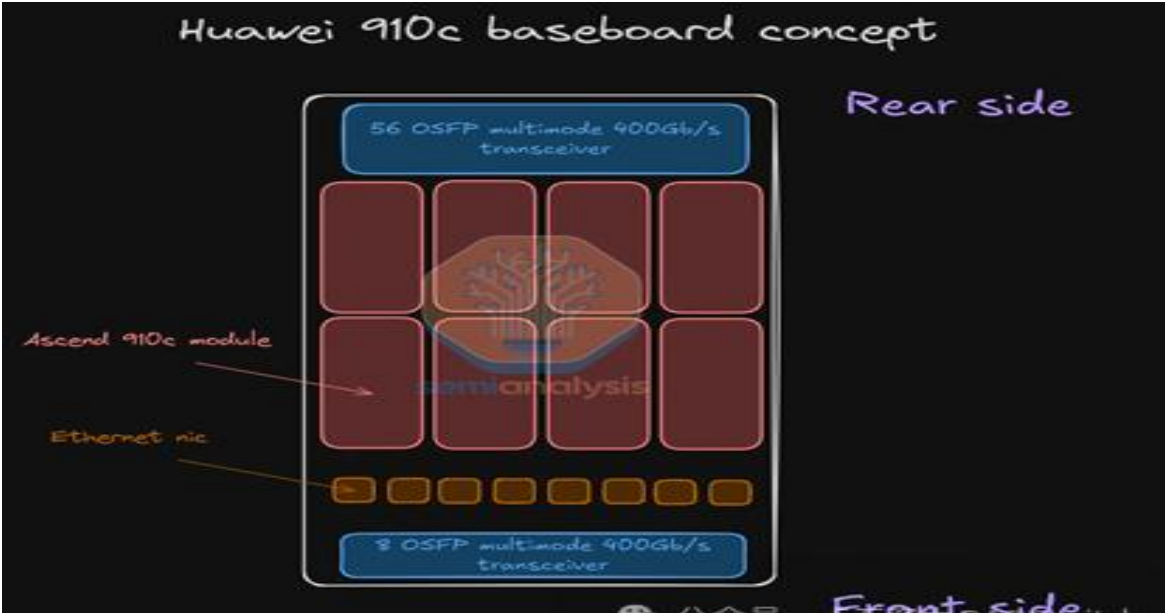
单芯片				
	Unit	GB200	Ascend 910C	HW vs NV
BF16 dense TFLOPS	TFLOPS	2500	780	0.3x
HBM capacity	GB	192	720	0.7x
HBM bandwidth	TB/s	8	3.2	0.4x
Scale Up Bandwidth	Gb/s uni-di	7200	2800	0.3x
Scale Out Bandwidth	Gb/s uni-di	400	400	1.0x

系统				
	Unit	Nvidia GB200 NVL72	Cloud Matrix CM384	HW vs NV
BF16 dense PFLOPS	PFLOPS	180	300	1.7x
HBM capacity	TB	13.8	49.2	3.6x
HBM bandwidth	TB/s	576	1229	2.1x
Scale Up Bandwidth	Gb/s uni-di	518400	1075200	2.1x
Scale Up Domain Size	GPUs	72	384	5.3x
Scale Out Bandwidth	Gb/s uni-di	28800	153600	5.3x
All-In System Power	W	145000	599821	4.1x
All-in Power per BF16 dense PFLOPS	W/TFLOP	0.81	2	2.5x
All-in Power per memory bandwidth	W per TB/s	252	488	1.9x
All-in Power per memory capacity	kW/TB	10.5	12.2	1.2x

1.2 昇腾Atlas 900 AI集群：算力超NVL72的70%，功耗约4倍+

- 为保障复杂的系统级互联得以实现，每张基板上均需要大量的互联元件。单张基板合计8张910C GPU、2张鲲鹏CPU、8张Thor-2后端网卡、8*7只Scale Up 400G LPO光模块（每张GPU对应7只）等。
- 由于纵向扩展与横向扩展网络大规模采用光模块，整个包含384张GPU的集群功耗极高。即便采用了功耗更低的线性直驱LPO光模块方案，单个CM384超节点功耗超500千瓦，是英伟达GB200 NVL72机架的4倍以上。但同时，整个超节点也提供了超出NVL72超节点70%的算力。

图：昇腾910C基板情况



表：昇腾CloudMatrix 384 功耗情况

CloudMatrix384 Power Budget at 100% Load including Scale Up&Scale Out Network(SemiAnalysis Estimates)			
item	unit power (w)	quantity	extened power (w)
AScd 910C	750	8	6000
A Accleart Bas+board	300	1	300
Motherboard	95	1	95
Kunpeng 920 CPU	180	2	360
PCIe 4.0 Retimers on Baseboard	10	8	80
PCIe 4.0 Switches	50	4	200
RAM 6408	10	32	320
Fans	60	20	1200
M.2 NVMe Boot Drive	10	2	20
NVMe	20	4	80
Storage backplane	75	1	75
OSFP 400Git/s Siph LPO Scale Up transceiver	7	56	364
Thor-2 400GbE/s Scale Out NIC	26	8	207
OSFP 400Git/s Siph LPO Scale Out transceiver	7	8	52
200Gbit/s Frontend DPU	25	2	50
QSFP112 mutimode 200Gb/s transceiver	7	2	14
Power Per Compute Chassis (pre PSU inefficiency)			9,417
PSU efficiency			94%
Power Per Compute Chasis (post PSU inefficiency)			10,018
Huawei CloudEngine 16800 Scale Up Ethernet Switch Chassis			
Switch ASICs	200	28	5600
Fans	61.2	50	3060
OSFP 400Git/s Siph LPO Scale Out trasceiver	6.5	672	4368
Power Per Scale Up Switch Chassis (per PSU inefficiency)			13,028
PSU efficiency			94%
Power Per Scale Up Switch Chassis (post PSU inefficiency)			13,860
Huawei CloudEngine 16800 Scale Out Ethernet Switch Chassis			
Switch ASICs	2000	32	64000
Fans	61.2	50	3060
OSFP 400Git/s Siph LPO Scale Out trasceiver	6.5	768	4992
Power Per Scale Out Switch Chassis (per PSU inefficiency)			14,452
PSU efficiency			94%
Power Per Scale Out Switch Chassis (post PSU inefficiency)			15,374
Cloud Matrix 384 Pod			
Compute Chassis	10018	48	480864
Scale Up Switch Chassis	13850	4	55400
Scale Out Swch Chassis (2 Ter Network)	15374	2	23062
Total Critical IT Power			559,378

1.2 昇腾Atlas 950/960 SuperPoD计划于2026/2027年上市

- **Atlas 950 SuperPoD**: 从基础器件、协议算法到光电技术，实现系统级创新突破。通过正交架构，Atlas 950实现零线缆电互联，采用液冷接头浮动盲插设计做到零漏液，其独创的材料和工艺让光模块液冷可靠性提升一倍。其创新的UB-Mesh递归直连拓扑网络架构，支持单板内、单板间和机架间的NPU全互联，以64卡为步长按需扩展，**最大可实现8192卡无收敛全互联**。
- **Atlas 960 SuperPoD (2027Q4上市)** 规模为15,488卡 (NPU)，FP8算力为8EFLOPS，FP4算力为16 EFLOPS，互联带宽为34PB/s。

Atlas 950 SuperPoD (上市时间: 2026年Q4)

全液冷数据中心AI超节点面向超大型AI计算任务的最佳选择



规模

算力

跨柜全光互联

8,192卡 (NPU)

8EFLOPS FP8/16 EFLOPS FP4

互联带宽 16.3 PB/s

1.2 昇腾Atlas 950 SuperCluster：国内芯片支撑的超算解决方案

- 基于超节点，华为发布全球最强超节点集群，分别是Atlas 950 SuperCluster 和Atlas 960 SuperCluster，算力规模分别超过50万卡和达到百万卡。华为打造的‘超节点+集群’算力解决方案将基于中国可获得的芯片制造工艺，来满足国内高速增长的算力需求。
- Atlas 950 SuperCluster集群由64个Atlas 950超节点互联组成，把1万多机柜中的52万多片昇腾950DT组成为一个整体，FP8总算力可达524 EFLOPS。相比当前世界上最大的集群 xAI Colossus，规模是其2.5倍，算力是其1.3倍。

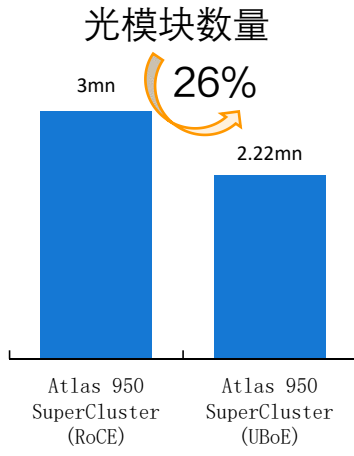
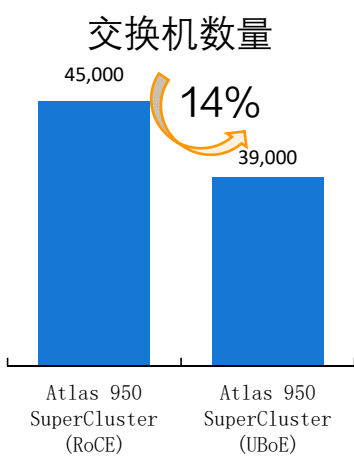
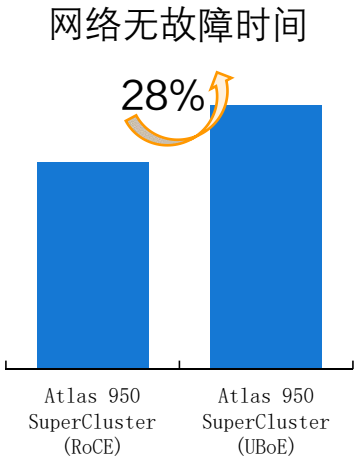
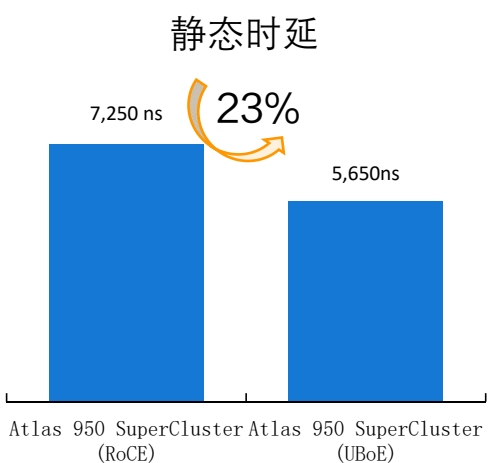
Atlas 950 SuperCluster（上市时间:2026年Q4）



524 EFLOPS FP8/1 ZFLOPS FP4

64 Atlas 950 SuperPoDs

524,288 NPUs



1.2 昇腾Atlas 960 SuperCluster：百万卡超节点性能跃升

- Atlas 960 SuperCluster预计于2027年Q4上市，集群规模进一步提升到百万卡级，提供2 ZFLOPS FP8和4 ZFLOPS FP4的性能。它同样支持UBoE与RoCE两种协议，在UBoE协议加持下，性能与可靠性同样更优，并且，静态时延和网络无故障时间优势进一步扩大，因此继续推荐UBoE组网。

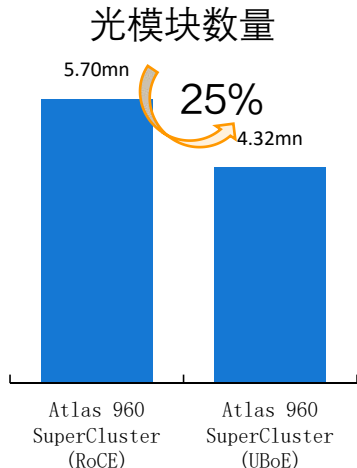
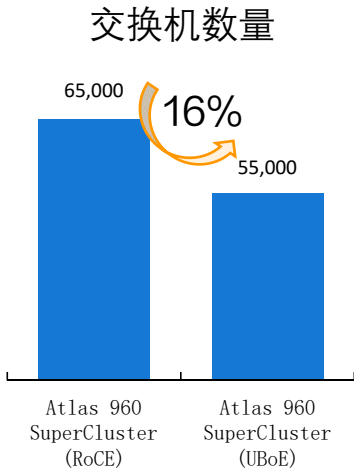
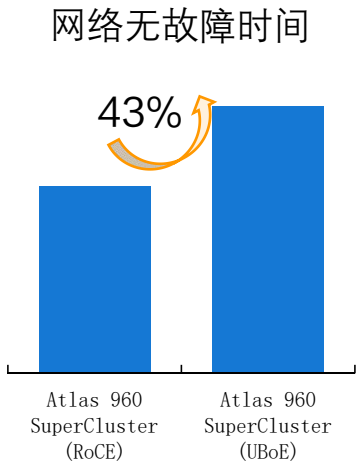
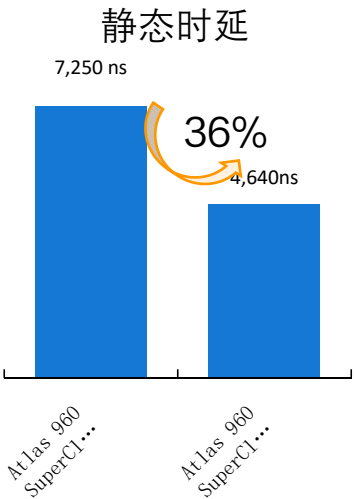
Atlas 960 SuperCluster（上市时间:2027年Q4）



2 ZFLOPS FP8 / 4 ZFLOPS FP4

64 Atlas 960 SuperPoDs

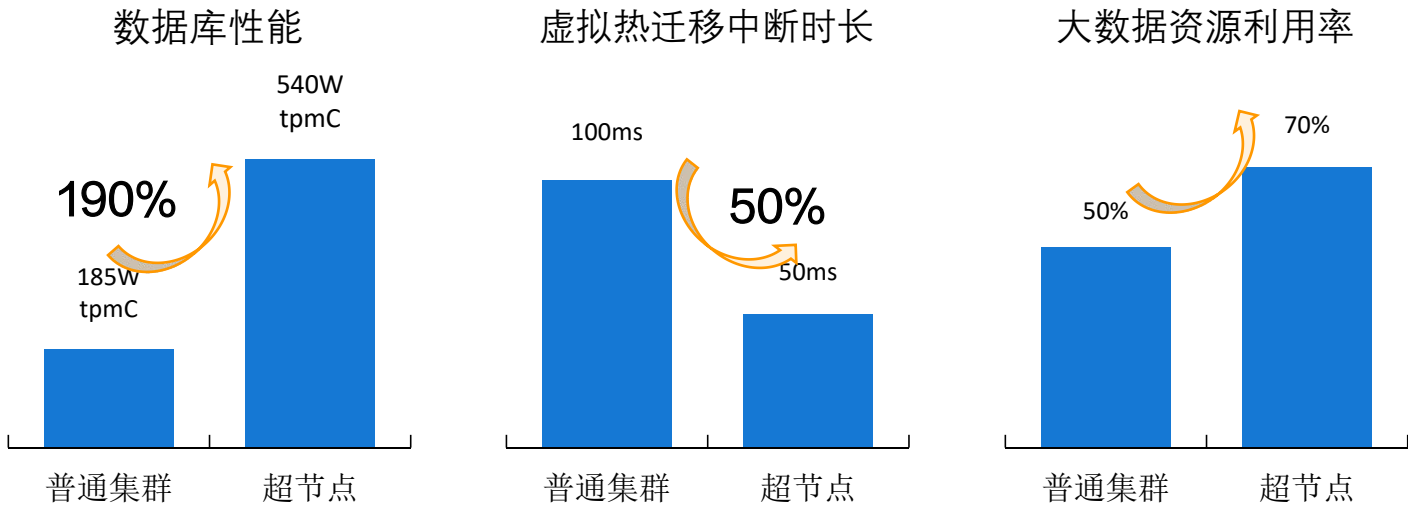
991,232 NPUs



1.2 昇腾Taishan 950 SuperPoD：业界首款通算超节点

- Taishan 950 SuperPoD——华为推出的业界首款通算超节点，为通算性能提升开辟全新路径。该产品具备百纳秒级超低时延、TB级超大带宽和内存池化能力，能大幅提升数据库、虚机热迁移和大数据场景等业务性能。该产品基于新一代鲲鹏950处理器，结合GaussDB分布式数据库，有望彻底取代各种应用场景的大型机和小型机以及Exadata数据库一体机。

Taishan 950 SuperPoD (上市时间：2026年Q1)



最大16节点 (32P)

最大内存48 TB

支持内存/SSD/DPU池化

1.2 昇腾Atlas 850：首款企业级风冷AI超节点

- Atlas 850——业界首个企业级风冷AI超节点服务器。该产品搭载8张昇腾NPU，有效满足企业模型后训练、多场景推理等需求。Atlas 850支持多柜灵活部署，最大可形成128台1024卡的超节点集群，是目前业内唯一可在风冷机房实现超节点架构的算力集群。

Atlas 850（上市时间:2026年Q4）



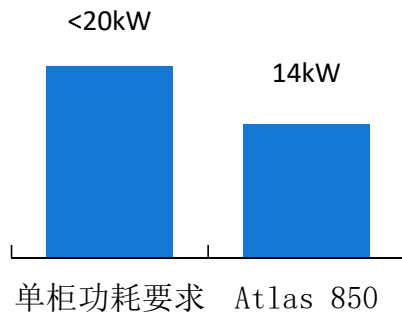
8 NPU
8 PFLOPS FP8
16 PFLOPS FP4
1,152 GB MEM

Atlas 860（上市时间:2027年Q4）



8 NPU
16 PFLOPS FP8
32 PFLOPS FP4
2,304 GB MEM

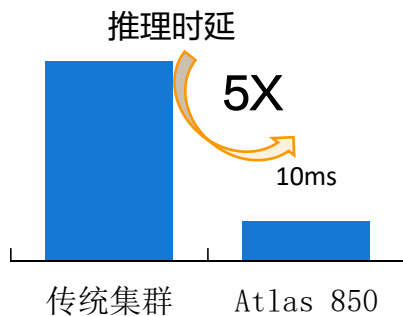
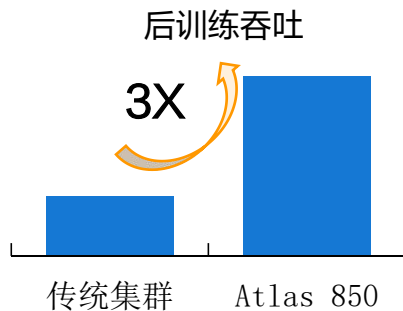
适于标准风冷机房
适用90%新建风冷机房无需改造



超节点灵活扩展
8~1024卡组合全量资源池化，灵活调配



高吞吐、低时延
倍级训练性能
推理时延满足实时交互



1.2 昇腾Atlas 350：高性能标卡驱动大模型推理

- Atlas 350标卡——昇腾950PR驱动的全新一代高性能标卡。采用最新的昇腾950PR芯片，向量算力提升2倍，支持更细粒度的Cacheline访问，在推荐推理场景可实现2.5倍性能提升，且单卡即可运行。Atlas 350支持4个灵衢端口互联，实现算力、内存等资源池化，让更大参数模型、更低时延应用可以在标卡上实现。

Atlas 350 (上市时间:2026年Q2)



算力

1.7~3.4 PFLOPS FP8
3.4~6.8 PFLOPS FP4



内存

256~512 GB
3.2~6.4 TB/S

支持2卡、4卡灵衢互联

灵活匹配业务场景

推荐推理

商品推荐|视频推荐…
2.5倍性能提升
单卡即可运行

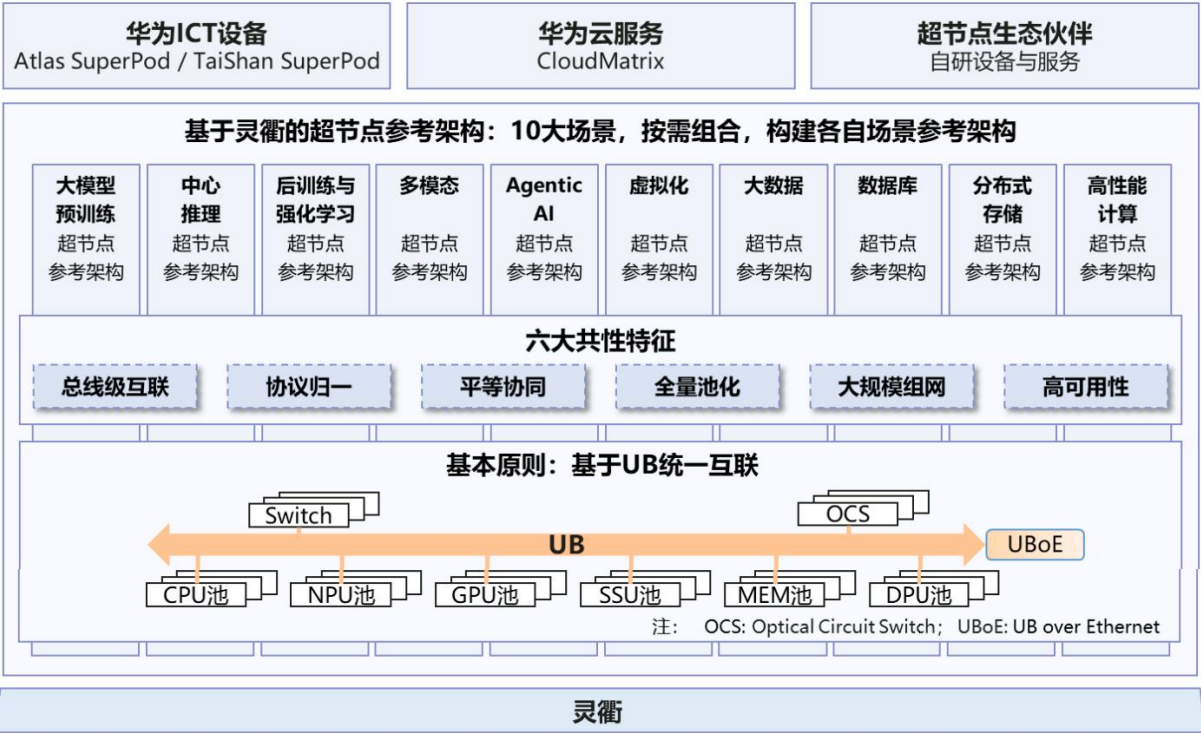
多模态生成

海报生成|短视频生成…
1.8倍性能提升
850 TFLOPS FP16算力

1.3 华为灵衢：面向超节点的互联协议

- 灵衢是一种面向超节点的互联协议，将I/O、内存访问和各类处理单元间的通信统一在 同一互联技术体系，实现数据高性能传输、算力高效协同、资源统一管理和灵活组合，是超节点参考架构的基础。基于灵衢（ UnifiedBus，简称UB ）的超节点架构，支持CPU、NPU、GPU、MEM、DPU、SSU（ Scalable Storage Unit ）和Switch中的一种或者多种组件资源池化和平等协同，构建逻辑上的一台计算机。
- 华为于2019年开始研究灵衢技术，2025年灵衢1.0正式在Atlas 900 A3 SuperPoD上商用验证，2025年9月，华为正式开放灵衢2.0协议。

图：基于灵衢的超节点参考架构



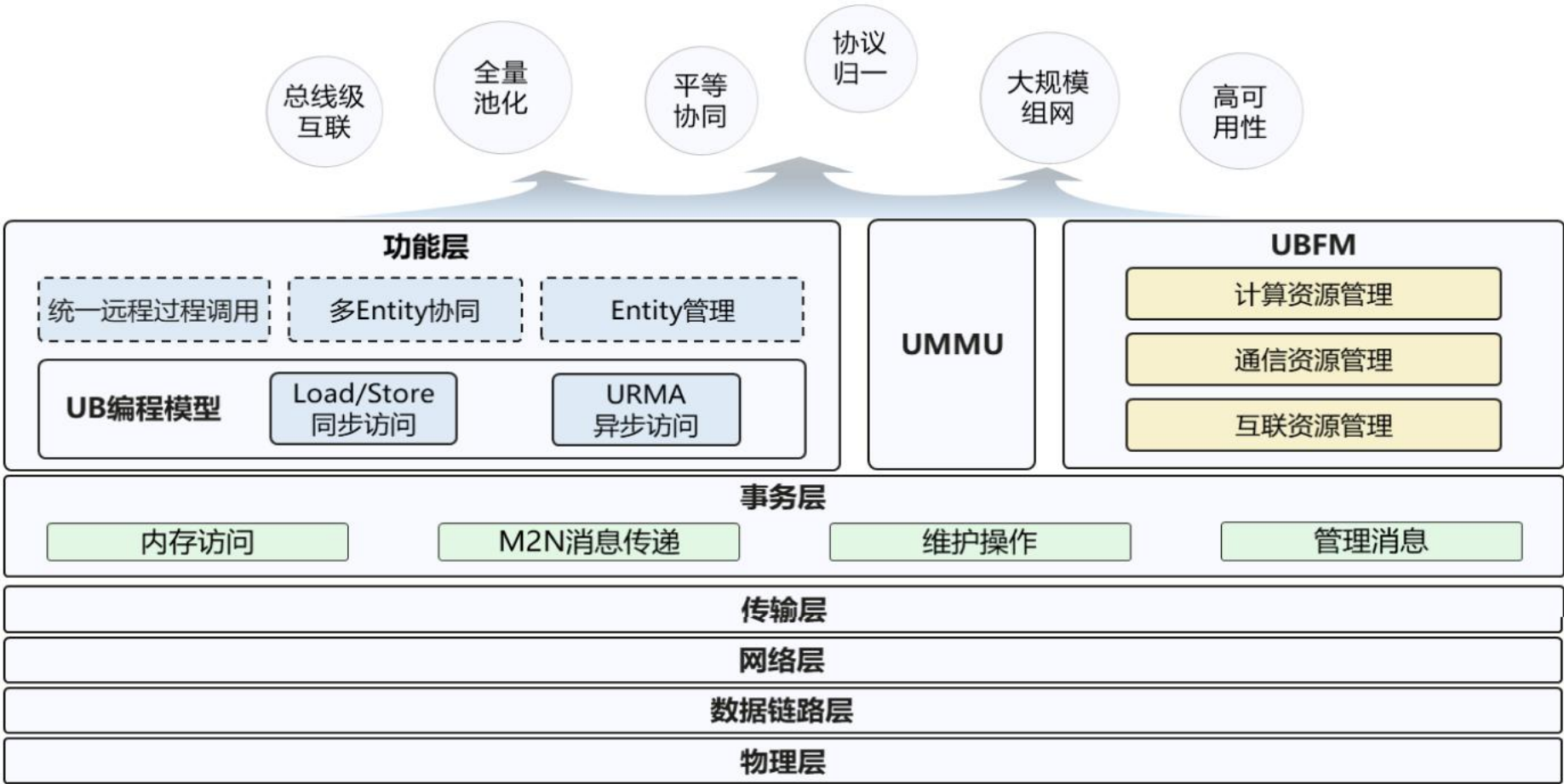
表：基于灵衢的超节点参考架构具备六大特征

超算中心	算力规划
总线级互联	基于灵衢的总线级互联，提供百ns同步内存语义访问时延和2~5us异步内存语义访问时延，满足算力单元高并发的访问需求；提供组件间TB/s级带宽，相比传统数据中心网络带宽至少提升10倍。
协议归一	基于灵衢的协议归一，支持超节点内不同类型、不同距离的组件统一互联，访问无协议转换开销，组件包括CPU、NPU、GPU、MEM、DPU、SSU和Switch等；提供统一的编程模型。
平等协同	基于灵衢的平等协同机制，支持超节点内所有组件去中心化的互相访问、调用和协同工作，提升组件间访存和通信性能。
全量池化	基于灵衢和Linux操作系统的灵衢扩展组件，提供超节点的设备管理、内存管理、通信和虚拟化等功能，支持超节点资源的高效池化管理和调用，提升资源弹性和利用率。
大规模组网	支持超节点以大于90%的线性度从单节点扩展到8192卡，未来还将持续提升至15488卡，甚至更大规模；支持超节点通过UBoE构建百万卡规模的集群，兼容以太网网。
高可用性	基于灵衢的可靠机制，支持超节点内应用无感知的us级检错和容错，在8192卡超节点范围内实现光互连MTBF（Mean Time Between Failures）大于6000小时。

1.3 华为灵衢：面向超节点的互联网协议

- 灵衢提供分层的协议栈，从下到上由物理层、数据链路层、网络层、传输层、事务层、功能层以及UMMU、UBFM（UB Fabric Manager）组成。其中，Entity为功能实体，是全局通信的基本单元；URMA（Unified Remote Memory Access）为统一远程内存访问。

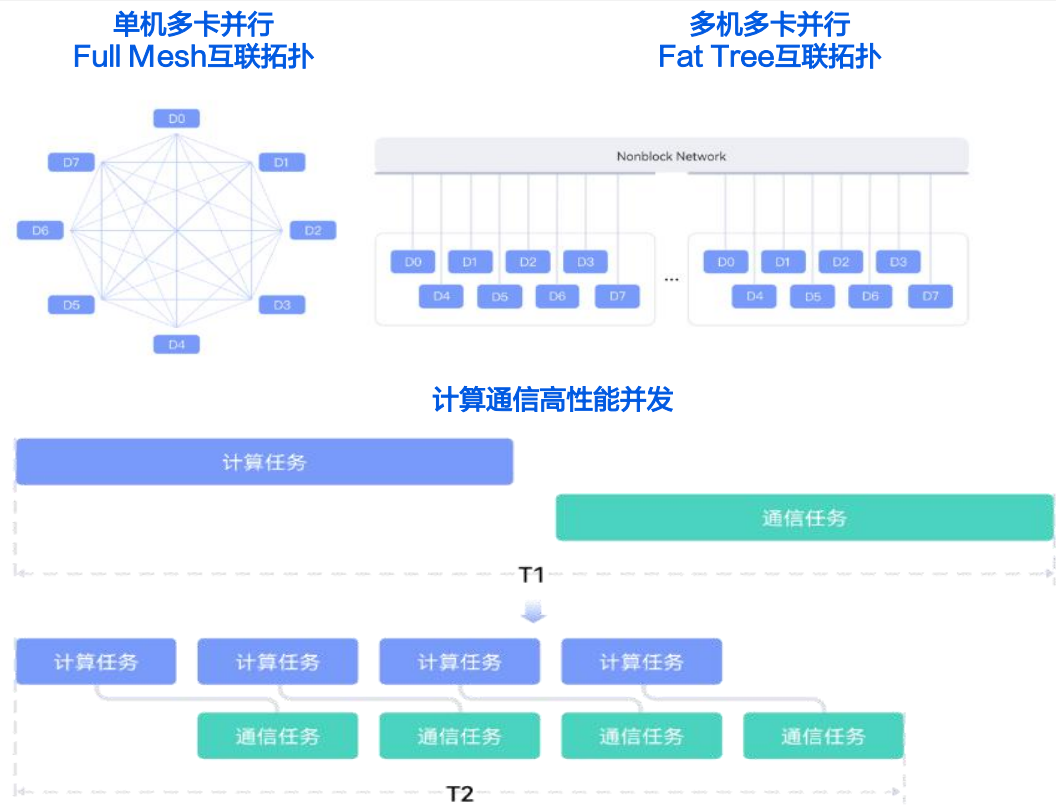
图：灵衢协议栈



1.3 HCCL、华为交换机为昇腾AI提供强大的通信基础

- **卡间互联：**集合通信库（ Huawei Collective Communication Library ，简称HCCL ）是基于昇腾硬件的高性能集合通信库，提供单机多卡以及多机多卡间的数据并行、模型并行集合通信方案。HCCL支持AllReduce、Broadcast、Allgather、ReduceScatter、AlltoAll等通信原语，Ring、Mesh、HD等通信算法，在HCCS、RoCE和PCIe高速链路实现集合通信。
- **交换机：**华为推出面向智能时代的CloudEngine数据中心交换机，定义了智能时代数据中心交换机的三大特征：400GE超宽、0丢包以太网、全生命周期自动管理，助力客户加速智能化转型。

图：昇腾高性能集合通信



图：华为推出多款数据中心交换机

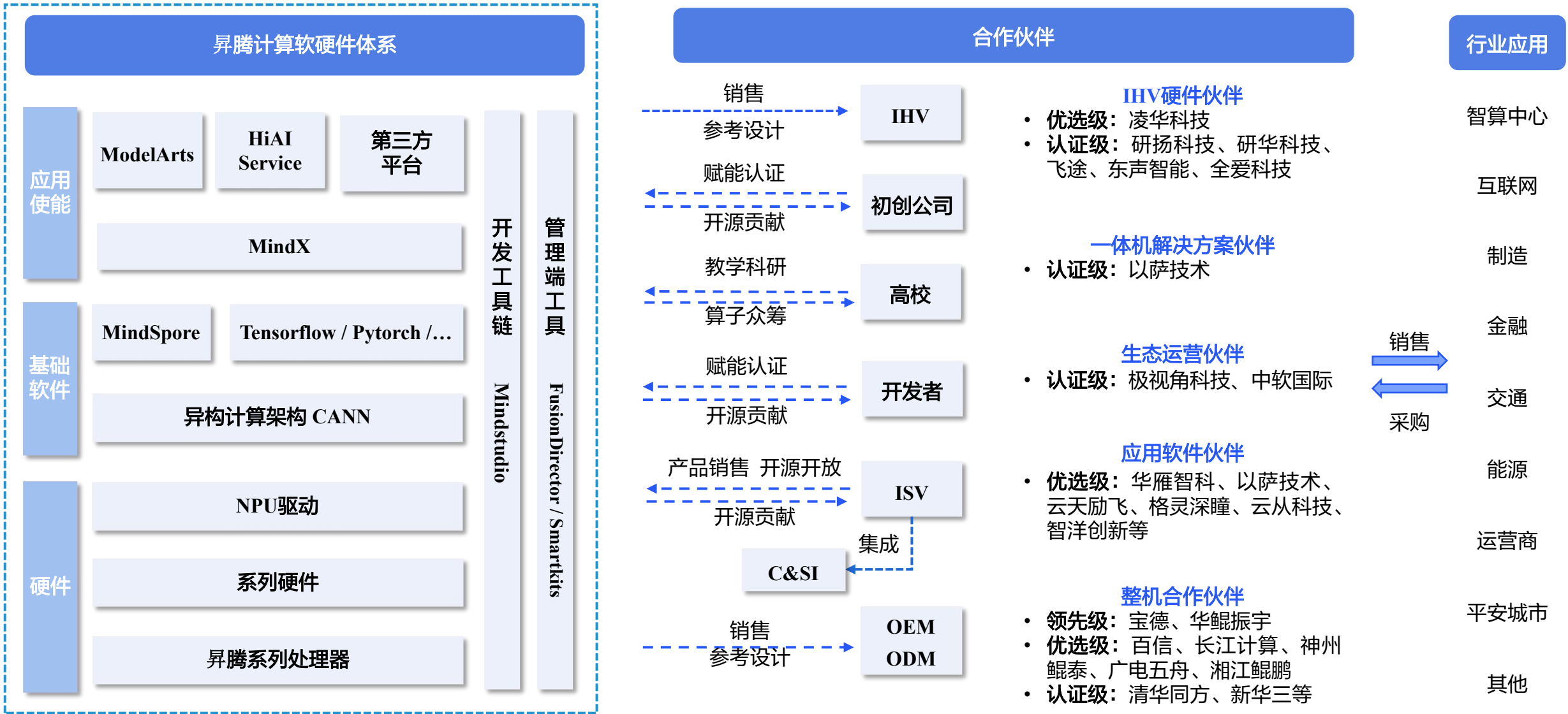
CloudEngine XH16800系列数据中心交换机

- 面向AI场景的数据中心核心交换机
- 在算力效率方面，CloudEngine XH16800 系列交换机支持NSLB(网络级负载均衡)，实现训练效率提升20%。
- 在算力可用率方面，支持算网CCAIE一体化运维等功能，排障效率提升90%。

CloudEngine 16800系列数据中心交换机

- 面向智能时代的数据中心交换机，适用于一体化大数据中心，以及“东数西算”等跨区域算力调度场景。
- 拥有高密400GE与智能调度算法实现算力全面加速，超融合以太技术统一通用计算存储、超算，全方位降低TCO。

1.4 广泛构建合作伙伴体系，助力昇腾AI产业链繁荣



1.4 华为昇腾智算中心建设经验丰富

- **昇腾人工智能计算中心建设方案：**昇腾AI提供从底层硬件到顶层应用使能的人工智能全栈能力，助力人工智能计算中心的建设。依托人工智能计算中心，打造公共算力服务平台、应用创新孵化平台、产业聚合发展平台、科研创新和人才培养平台，以“1中心+4平台”赋能产业发展。
- **智算中心建设提速，昇腾处理器份额领先。**昇腾处理器性能国内领先，成为国内智算中心的主流选择，18个国家新一代人工智能创新发展试验区中，已有12个城市披露采用昇腾AI芯片，占比高达2/3。深圳鹏城云脑II、武汉人工智能计算中心快速交付，核心建设周期 3.5 ~ 4个月。

图：昇腾人工智能计算中心建设方案



图：部分使用昇腾芯片的智算中心

超算中心	项目状态	算力规划
北京昇腾人工智能计算中心	已运营	一期100P
青岛人工智能计算中心	已运营	一期100P
天津人工智能计算中心	已运营	300P
重庆人工智能计算中心	已运营	400P
广州人工智能计算中心	已运营	99P
沈阳人工智能计算中心	已运营	一期100P，总计400P
南京人工智能计算中心	已运营	40P
成都智算中心	已运营	300P
杭州人工智能计算中心	已运营	一期40P，二期100P
西安人工智能计算中心	已运营	300P
武汉人工智能计算中心	已运营	200P
中原人工智能计算中心	已运营	一期100P，未来300P
鹏城云脑II	已运营	1000P

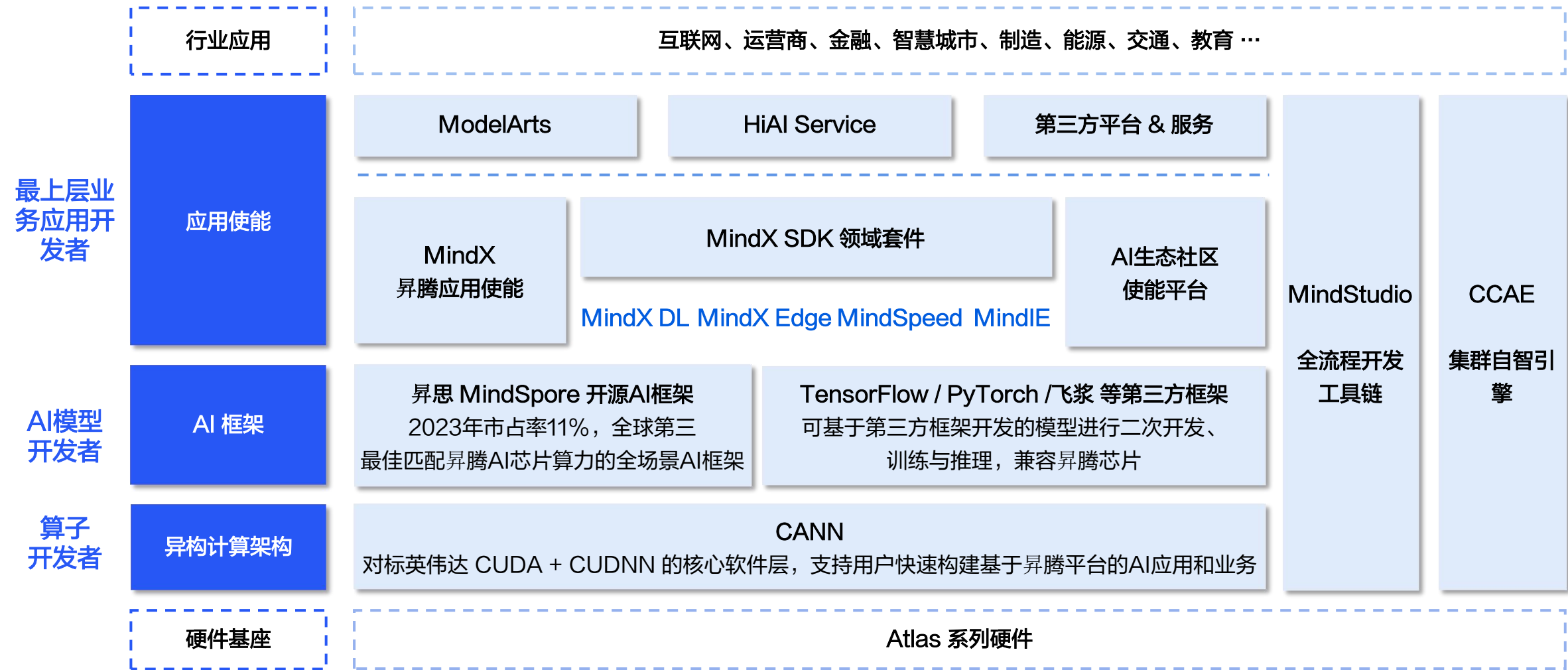
第二章 昇腾软件架构

构建全栈AI软硬件平台，赋能昇腾生态蓬勃发展

2.1 昇腾软件：构建全栈AI软硬件平台，CANN全面开源开放

- 昇腾处理器的卓越性能决定其算力水平，但其相关的软件生态是推升AI芯片计算生态发展的关键。2025年8月，华为宣布：华为昇腾硬件使能CANN全面开源开放，Mind系列应用使能套件及工具链全面开源，支持用户自主的深度挖潜和自定义开发。

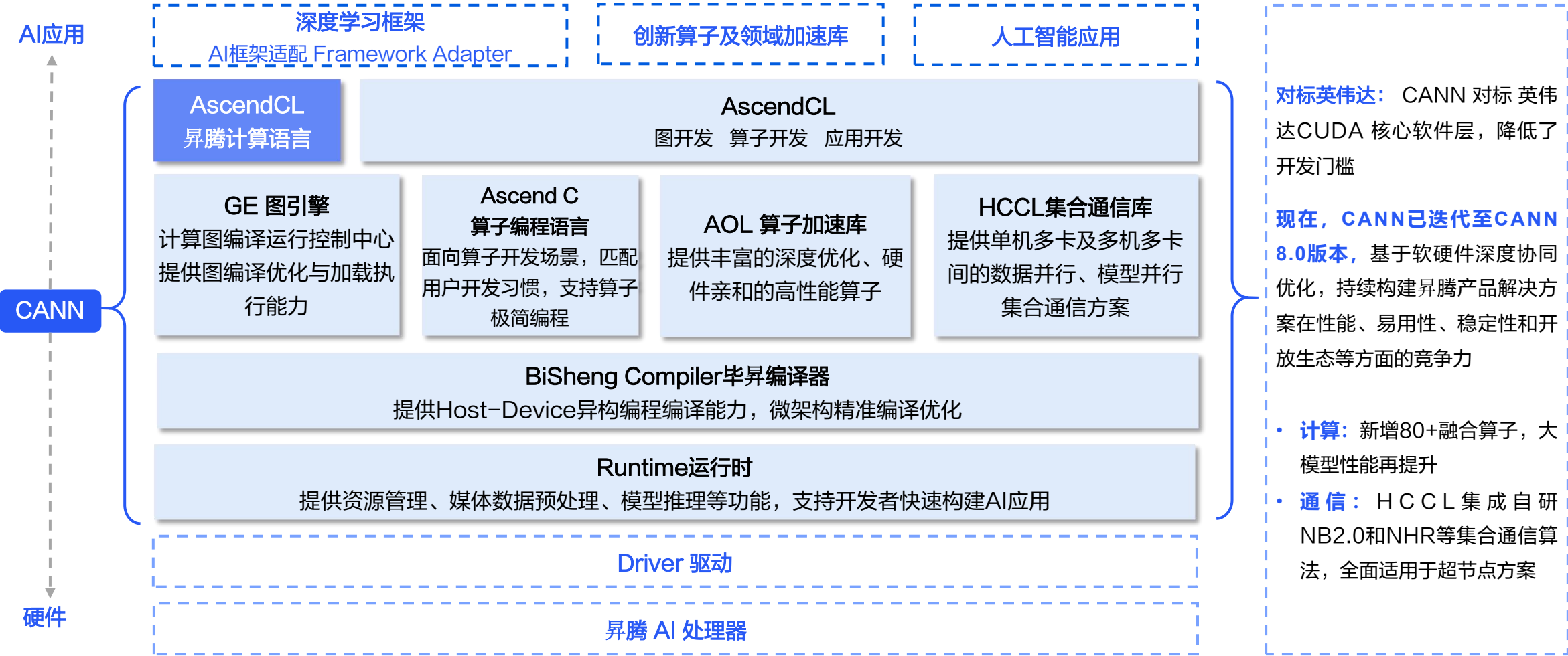
图：昇腾全栈 AI 软硬件平台，赋能昇腾生态不断发展



2.2 昇腾CANN：对标英伟达CUDA核心软件层

- 昇腾CANN对上支持AI框架，对下服务AI处理器，对标英伟达CUDA核心软件层。英伟达CUDA是一种通用并行计算架构，提供硬件直接访问接口，适配多种AI框架等。

图：昇腾 CANN 为昇腾AI软硬件平台的核心，可提升昇腾处理器计算效率



2.2 CANN8.0版独创通信算法/算子与丰富API，释放硬件价值

- CANN 8.0，新增NB 2.0等十几类通算融合算子，通过自适应通信域优化，充分利用超节点通信全链路，实现更细粒度的通信计算并行，将模型训练性能提升20%以上。新增的通信、矩阵运算、数据搬运等API，提供20类200+ API，算子开发效率提升30%，大幅降低算子开发工作量，Matmul等典型融合算子开发代码量缩减30%以上。

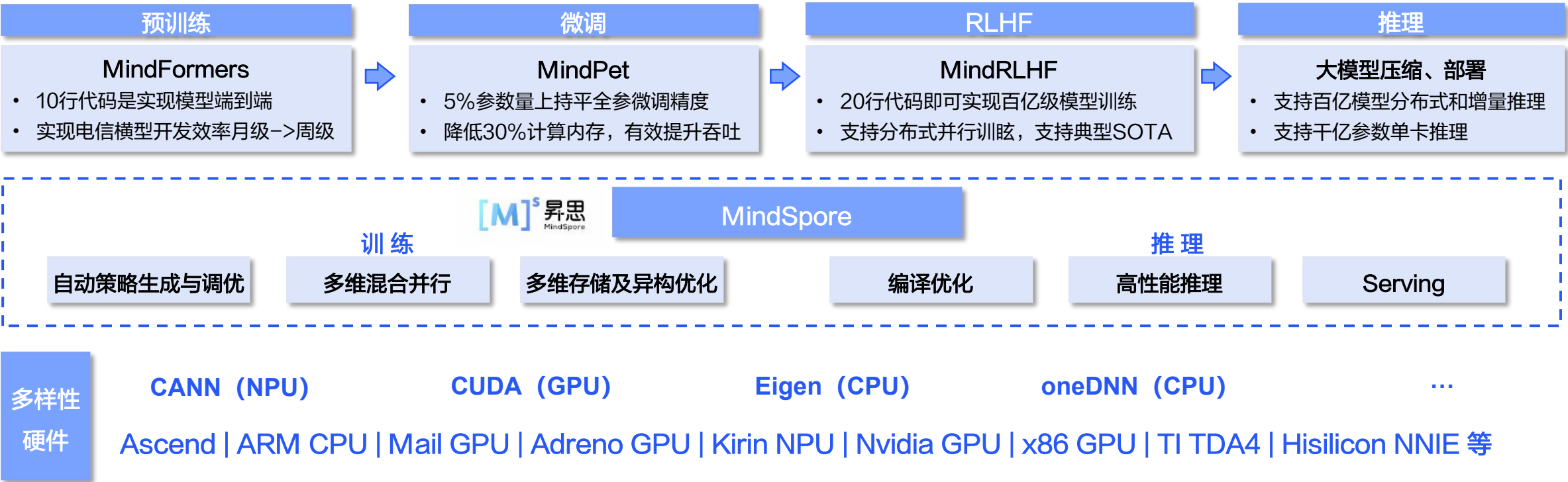
图：CANN 8.0独创的通信算法、算子与丰富的API



2.3 昇思MindSpore：适配DeepSeek等大模型，打造大模型训推平台

- 昇思MindSpore是华为推出的面向全场景AI计算框架，打造大模型最佳训推平台之一。MindSpore在机器学习开发中起上承应用（北向）、下接芯片（南向）的桥梁意义。MindSpore 1.0实现业界首个全场景AI框架，1.5版本开始原生支持大模型，2.0版本提供大模型套件支持一站式训练，2025年4月，昇思MindSpore 2.6版本正式发布。该版本全面支持类DeepSeek V3/R1 MoE模型训练推理全流程，推出训推一体的强化学习套件使能后训练范式创新，易用性上实现主流SOTA模型的Day0迁移。
- MindSpore 适配大模型，市场份额位于第一梯队。昇思MindSpore面向政企、金融、运营商、电力、交通等行业提供了端到端的训推一体解决方案，赋能千行百业。

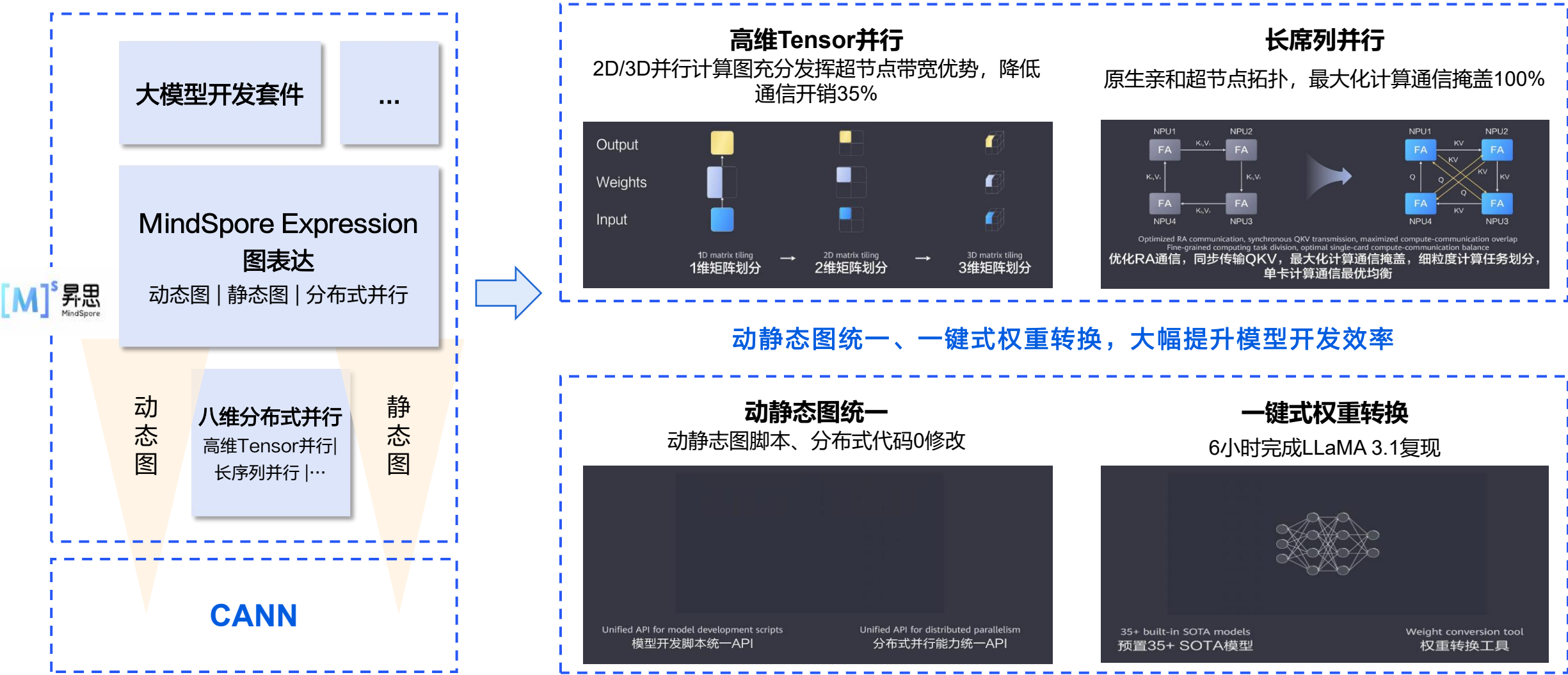
图：昇思MindSpore是华为推出的面向全场景AI计算框架



2.3 MindSpore原生亲和超节点架构，大幅缩短模型开发训练时间

昇思MindSpore 2.6版本：基于超节点原生并行创新、打造动静统一API

面向超节点架构，创新高维Tensor并行与长序列并行
提升万亿参数模型训练性能20%



2.4 MindStudio：训练模型性能提升340%，提升客户迁移昇腾信心

- MindStudio是华为面向昇腾AI开发者提供的一站式开发环境和工具集。MindStudio致力于提供端到端的昇腾AI应用开发解决方案，使开发者能够在在一个工具上高效完成算子开发、训练开发和推理开发。
 - ✓ 算子开发：支持通过C/C++语言实现算子自动生成，并提供算子异常检测、上板调试及性能数据展示定位算子瓶颈。
 - ✓ 训练开发：针对不同的训练框架提供灵活多样的调优手段，算子性能提升20-80%，调优效率从周级向天级。
 - ✓ 推理开发：可实现一站式调试调优，大模型稀疏量化压缩，分钟级完成主流模型迁移，提高客户向昇腾迁移的信心。

图：MindStudio全流程开发工具链



2.5 昇腾推理解决方案：提升吞吐量，降低推理成本

- 面向AI模型推理场景，昇腾推理解决方案包含Atlas及伙伴推理硬件、异构计算架构CANN、昇腾推理引擎MindIE、行业应用开发套件MindX SDK，边缘部署使能MindX Edge等。昇腾推理基础软件提供系统性联合优化来降低时延，提升吞吐效率，在满足时延条件下，最大实现6x吞吐提升，大大降低了每百万tokens的推理成本。

图：昇腾推理解决方案



2.5 MindIE：推理引擎MindIE关键特性升级，提升AI落地经济性

- 为了更好的满足企业AI落地经济性，昇腾推理引擎MindIE升级了关键特性。率先支持自适应PD分离部署，根据负载自动调整计算、访存比例，实现推理资源按需分配，吞吐提升50%以上。提供单轮多Token的启发式并行解码能力，输出时延降低50%，支持外挂知识库、小参数模型和自回归模型，提高推理准确率。最后，MindIE还支持将长序列自动拆分为多个短序列进行推理，以有限的资源实现更高效的推理。

图：推理引擎MindIE关键特性升级



第三章 晶圆制造

AI自主可控大势所趋，国产Fab厂持续发展

3.1 晶圆生产过程：半导体设计、晶圆制造、测试封装

- 半导体设计、晶圆制造、测试封装是芯片制造最重要的三个环节。晶圆代工模式由Fabless（无晶圆厂，半导体设计公司），以及Foundry（晶圆代工厂）组成，晶圆代工厂会将制造好的芯片交由封测厂商进行芯片的封装和测试。头部的晶圆代工厂主要是台积电、三星、联电、中芯国际等。

图：半导体产业链分析图

产业链环节	核心研发	半导体制造上游			半导体设计和制造		
		设备	原材料	EDA/IP	IC设计和IDM	逻辑芯片制造	封装和测试
内容	半导体行业是研发最为密集的行业之一，研发推动行业技术的发展	大多为单一供应源，备货周期长	前端包括硅片、金属以及各类化学用品和气体	EDA是实现IC设计流程自动化的必备软件工具，该流程极为模块化，包含数十亿个组件	通过架构和设计，以实现芯片产品需求	摩尔定律驱动	通过引线键合、倒装等技术使芯片与外部电路连接
	研发人员始终致力于提高半导体产品的算力和速度且降低成本						
相关公司	玩家大多为学术机构或政府机构	需投入大量时间和金钱，打造洁净厂房，放置设备	后端包括封装材料或基材	IP是指可以授权给第三方的半导体IP块/核	芯片设计主要包括逻辑、存储和模拟芯片	(如28纳米 vs.7纳米)划分	劳动密集型，通常外包给OSAT行业
相关公司	ITRI SEMATECH CEA-Leti Semiconductor Research Corporation IMEC	KLA ASML 致同一行 上海微电子 北方华创 中微半导体	Siltronic GlobalWafers SUMCO Shin-Etsu 江丰电子 中环股份 金宏气体 沪硅产业	新思科技 Cadence Mentor Graphics 华大九天 国微思尔芯 概伦电子 芯原微电子	英特尔 恩智浦 瑞萨电子 英伟达 三星 海思 寒武纪 紫光展锐 合肥长鑫 瑞芯微	台积电 三星 联华电子 GlobalFoundries 中芯国际 华虹半导体 武汉新芯	日月光集团 安靠科技 长电科技 通富微电 华天科技

注：“相关公司”仅包含部分参与企业，其中蓝色字体为中国大陆公司。

3.2 多家代工厂发展先进制程，台积电为领头羊

- 行业追求缩小支制程来提升晶体管密度，从而增加计算能力。先进制程与成熟制程的分界线为28nm，28nm以下的制程被称为先进制程，发展先进制程有技术与资本双重障碍。索尼、Sharp、Infineon、AMD、德州仪器、意法半导体、格罗方德和联电陆续退出先进制程研发，目前，7nm及以下的制程工艺，仅台积电、英特尔、三星三家厂商可以掌握。头部的3大玩家中，也只有台积电的先进制程是最成熟和稳定的。

表：全球主要晶圆厂制程节点技术路线

公司	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024E	2025E	2026E	>>
台积电	28nm			20nm	16nm		10nm	7nm	7nm+	5nm 6nm		3nm	N4P N3E		N2 (GAA)	N2P (GAA)	A14 (GAA)
三星		28nm		22nm	14nm		10nm	8nm	7nm EUV 6nm	5nm 6nm	4nm	3GAE		3GAP	SF2	SF2P	SF1.4
Intel	22nm			14nm		14nm+	14nm++		10nm	10nm+	10nm++ 7nm+		7nm++	20A 18A			
中芯国际	40nm				28nm				14nm		N+1	N+2			---		

3.3.1 台积电：先进制程技术的引领者，推进研发2nm工艺

- 台积电一直是全球先进制程技术的引领者。在2025年至2026年间，台积电即将推出的几项关键工艺技术，包括N3X、N2、N2P，以及革命性的A16工艺。
- 2025H2：①N3X：面向极致性能的3纳米级工艺，通过降低电压至0.9V在相同频率下减少功耗，同时在相同面积下提升性能或提高晶体管密度；②N2：台积电首个采用全栅（GAA）纳米片晶体管的节点，显著提升PPA（性能、功耗、面积）特性，相较于N3E有明显进步。

图：台积电先进制程规划图



图：台积电先进制程工艺性能分析

工艺节点	相比于	电源	性能	密度	量产时间
N3	N5	-25	10	?	Q4 2022
N3E	N5	-34	18	1.3x	Q4 2023
N3P	N3E	-5	5	1.04x	H2 2024
N3X	N3P	-7	5	1.10x	H2 2025
N2	N3E	-25	10	1.15x	H2 2026
N2P	N2	-30	15	1.15x	H2 2026
N2P	N3E	-5	5	?	H2 2026
A16	N2P	-15	8	1.07x	H2 2026

3.3.2 中芯国际：中国大陆集成电路制造业领导者

- 中芯国际是世界领先的集成电路晶圆代工企业之一，也是中国大陆集成电路制造业领导者，拥有领先的工艺制造能力、产能优势、服务配套，向全球客户提供8英寸和12英寸晶圆代工与技术服务。根据全球各纯晶圆代工企业最新公布的 2023年销售额情况排名，中芯国际位居全球第四位，在中国大陆企业中排名第一。

图：全球主要晶圆代工厂工艺路线图（2024年10月）



第四章 需求端

CSP厂商资本开支同比提升，云端营收持续增长

4.1 中美关税摩擦经历多轮升级，自主可控大势所趋

- 2025年中美关税摩擦在短短数月内经历了多轮升级，自主可控为大势所趋。
- 2025年2月1日，美国总统特朗普签署行政命令，宣布对中国所有进口商品额外征收10%的关税；随后4月2日特朗普在白宫签署两项关于所谓“对等关税”的行政令，宣布美国对所有贸易伙伴设立10%的“最低基准关税”，并对某些贸易伙伴征收更高关税。中国的迅速反制在美方行动后立即展开。5月12日，《中美日内瓦经贸会谈联合声明》发布，关税摩擦有所缓和。

表：中美博弈发展情况

时间节点	美国对中国制裁行动	中国对美国反制行动
2025 年 2 月 1 日	对所有中国商品加征 10% 关税	2 月 4 日宣布对美煤炭、LNG 加征 15%，原油等加征 10%
2025 年 3 月 3 日	再次加征 10%，累计 20%	3 月 4 日宣布对美部分农产品加征 10 – 15% 关税
2025 年 4 月 2 日	实施 “对等关税” 34%，累计 54%	4 月 4 日宣布对美所有商品加征 34% 关税
2025 年 4 月 8 日	提高至 84%，累计 104%	4 月 9 日提高至 84%
2025 年 4 月 10 日	提高至 125%，累计 145%	4 月 11 日提高至 125%
2025 年 4 月 15 日	宣布最高达 245% 的关税	-
2025 年 4 月 16 日	英伟达当地时间15日发布8-K文件称，公司日前收到美国政府通知，H20芯片和达到H20内存带宽、互连带宽等的芯片向中国等国家和地区出口需要获得许可证	
2025 年 5 月 12 日	《中美日内瓦经贸会谈联合声明》发布，协议核心是美对华关税从145%降至10%，中方同步从125%降至10%，但各自保留部分筹码。双方承诺在90天内暂停24%的关税加征，仅保留10%的剩余关税。美方取消对香港、澳门商品的加征关税，中方则暂停非关税反制措施。	
2025 年 5 月 13 日	美国商务部正式发布文件，启动撤销拜登签署的《AI扩散规则》，同时宣布采取额外措施加强对全球芯片出口管制。发布指导意见，指出在世界任何地方使用华为昇腾（Ascend）芯片违反了美国的出口管制。 5月19日，中方注意到，近日美方对指南新闻稿相关表述进行了调整，但指南本身的歧视性措施和扭曲市场本质并没有改变。美方滥用出口管制措施，以莫须有的罪名对中国芯片产品加严管制，甚至干涉中国公司在中国境内使用中国自己生产的芯片，美方的手伸得太长，是典型的单边霸凌行径，中方坚决反对。	

4.1 中美关税摩擦经历多轮升级，自主可控大势所趋

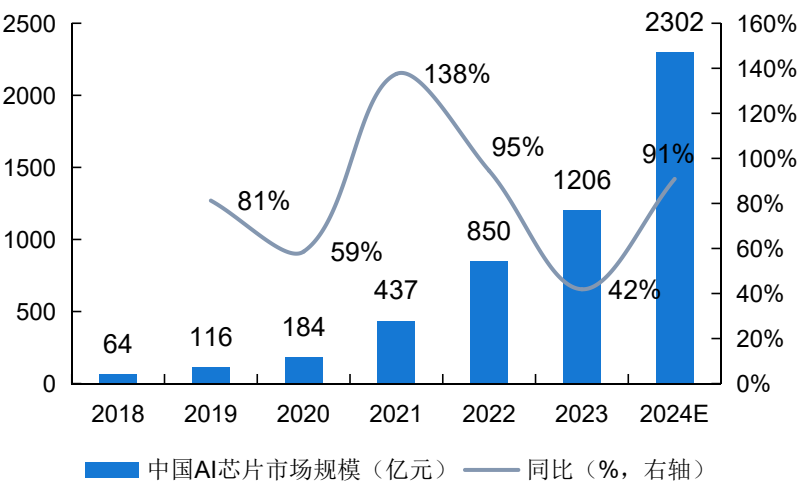
- 中国对美国实行反制关税，或将加速我国自主可控的进程。2025年4月4日，我国宣布对所有美国进口商品征收 34% 关税，并加强对稀土出口限制，我们认为这一举措或将增加进口软硬件产品的销售成本，从而加速国内市场向自主可控产业链转移。
- 近年，我国持续推动“自主可控”、“科技自立”。2022年国资委发布79号文，要求2027年底前，实现所有中央企业信息化系统安可信创替代。2024年，我国发布多项政策，提出抓紧打造自主可控的产业链供应链，四大协会建议国内企业审慎选择采购美国芯片。

表：国家政策从围绕信创安全逐步转向供应链自主可控

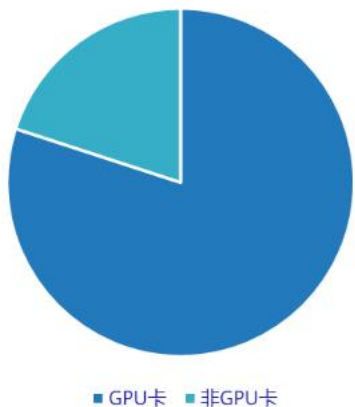
政策	发布单位	时间	内容
四大协会建议国内企业审慎选择采购美国芯片	四大部门	2024.12	中国汽车工业协会、中国半导体行业协会、中国互联网协会、中国通信企业协会四大协会联合发声表示抗议，认为美国芯片产品不再安全、不再可靠，并建议国内企业审慎选择采购美国芯片。
《关于推动未来产业创新发展的实施意见》	工信部等七部委	2024.1	深入推进5G、算力基础设施、工业互联网、物联网、车联网、千兆光网等建设，前瞻布局6G、卫星互联网、手机直连卫星等关键技术研究，构建高速泛在、集成互联、智能绿色、安全高效的新型数字基础设施。
《中共中央关于进一步全面深化改革、推进中国式现代化的决定》	中共中央	2024.7	抓紧打造自主可控的产业链供应链，健全强化集成电路、工业母机、医疗装备、仪器仪表、基础软件、工业软件、先进材料等重点产业链发展体制机制，全链条推进技术攻关、成果应用。
《新产业标准化领航工程实施方案（2023-2035年）》	工信部等四部委	2023.8	优化产业科技创新和标准化布局联动机制，协同推进技术研发、标准研制和产业发展。加强关键技术领域标准研究，推动先进适用的科技创新成果形成标准，促进科技创新成果高效转化。
《国资委79号文》	国资委	2022.9	要求所有中央企业2022年11月底前将安可替代总体方案报送国资委；自2023年1月起，每季度末向国资委报送信创系统替换进度。要求2027年底前，实现所有中央企业信息化系统安可信创替代。

4.2 GPU：AI带动需求提升，2024年国产占比升至30%

图：2018-2024年中国AI芯片市场规模预测

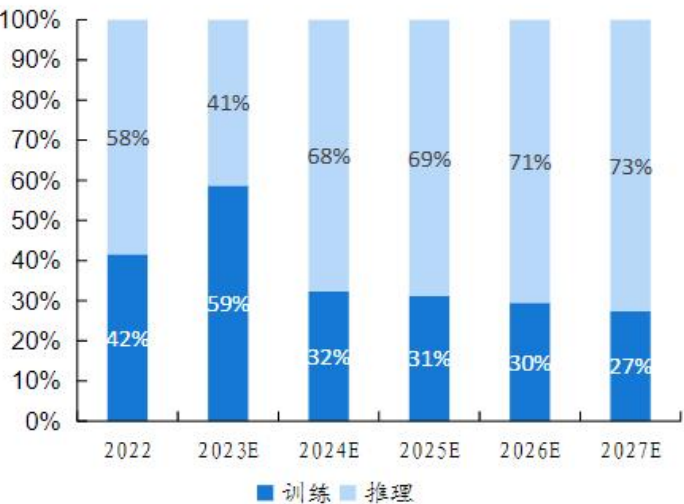


图：2024H1，中国人工智能芯片市场份额

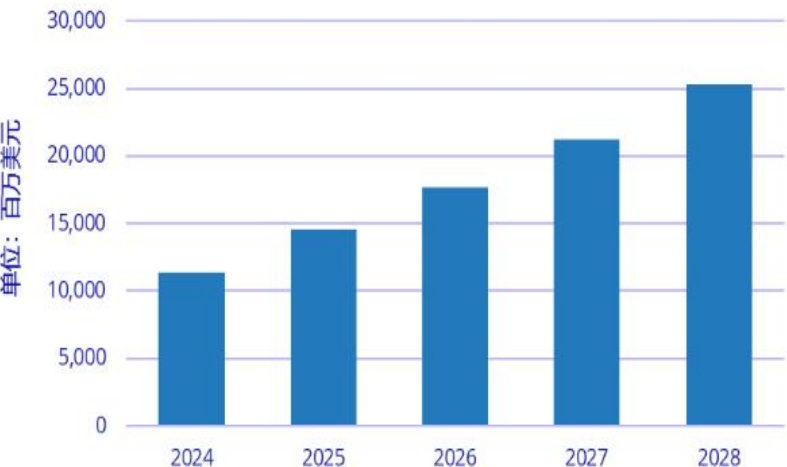


- 中国加速芯片市场持续增长。据IDC，2024年，中国加速芯片的市场规模增长迅速，超过270万张。从技术角度来看，GPU卡占据70%的市场份额；从品牌角度来看，中国本土人工智能芯片品牌的出货量已超过82万张，市占率30%左右。
- 从行业的角度看，互联网依然是最大的采购行业，据IDC数据，2024年占整体加速服务器市场超过65%的份额，其余行业均有不同幅度的增长。

图：2022-2027E 中国AI服务器工作负载预测



图：2024-2028，中国加速服务器市场规模

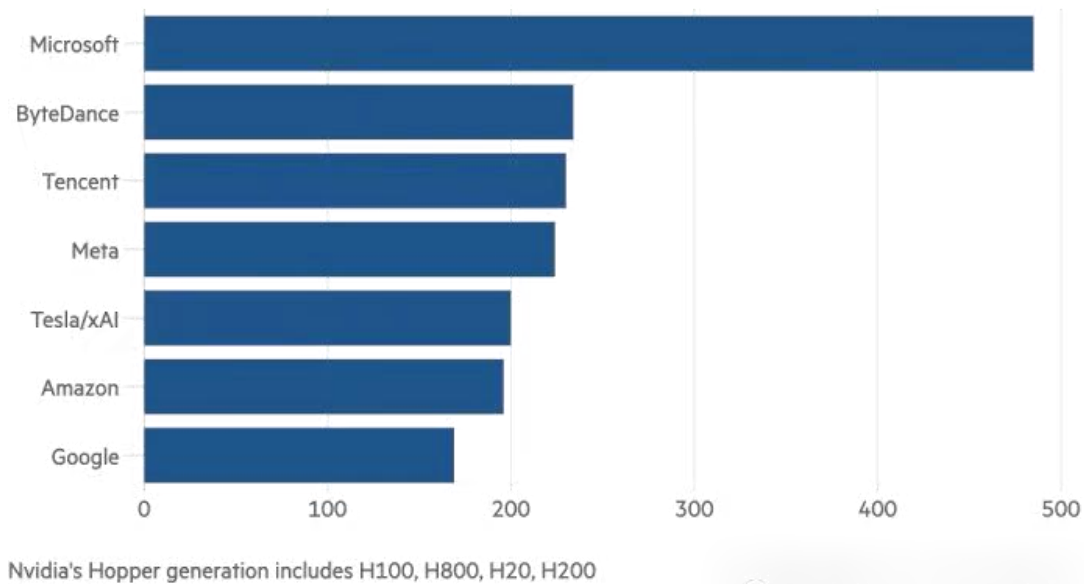


- 2024年，中国AI服务器增长迅速。IDC数据显示，2024年中国加速服务器市场规模达到221亿美元，同比2023年增长134%。其中GPU服务器依然是主导地位，占比达到69%。同时ASIC 和 FPGA等非GPU加速服务器高速增长，占比超过30%。

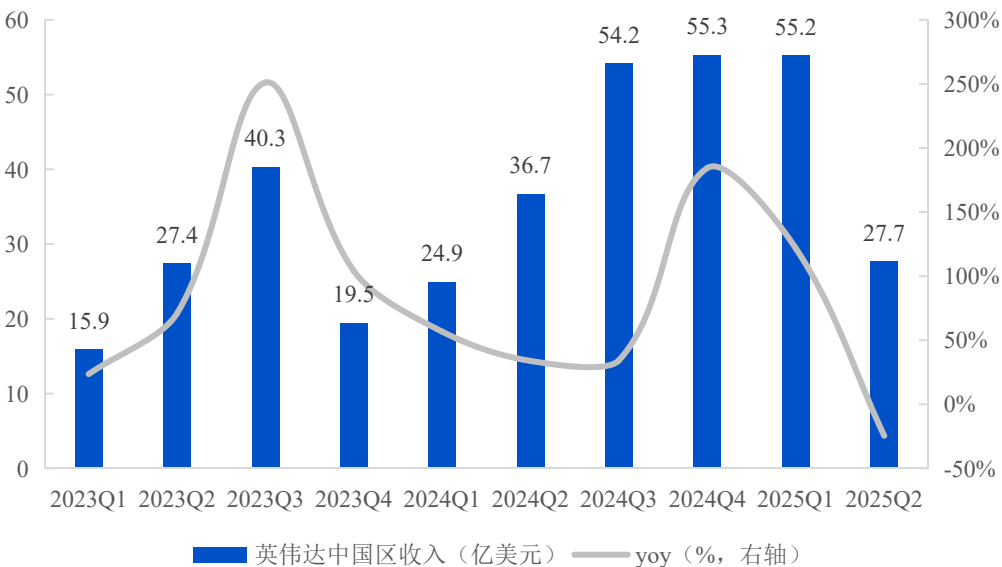
4.2 GPU：英伟达H20芯片受阻，国产替代空间打开

- 中国互联网对AI芯片需求量较高。科技咨询公司Omdia估计，2024年微软大约购买了48.5万张英伟达Hopper系列GPU，字节跳动23万张、为第二大英伟达GPU买家，腾讯23万张。
- 英伟达H20芯片在华销售多番遇阻。2025年4月15日，据彭博社报道，英伟达表示美国政府限制其H20芯片对中国的出口。根据PanSemicon公众号信息，2025年7月14日，英伟达宣布美国政府承诺将向其授予许可证，公司将恢复H20在华销售。然而，英伟达在中国市场遭遇新的障碍，2025年7月31日，国家互联网信息办公室约谈英伟达，要求其就对华销售的H20算力芯片漏洞后门安全风险问题进行说明并提交相关证明材料。2025年8月22日，根据PanSemicon公众号援引The information信息，英伟达下令停止生产H20。

图：2024年英伟达Hopper系列GPU出货情况



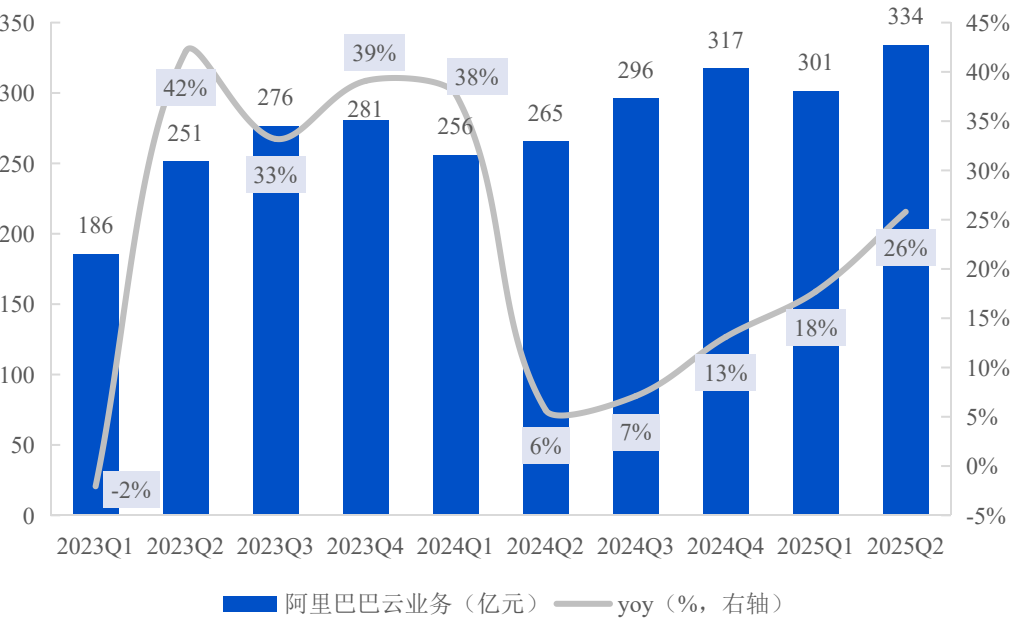
图：英伟达中国区收入情况



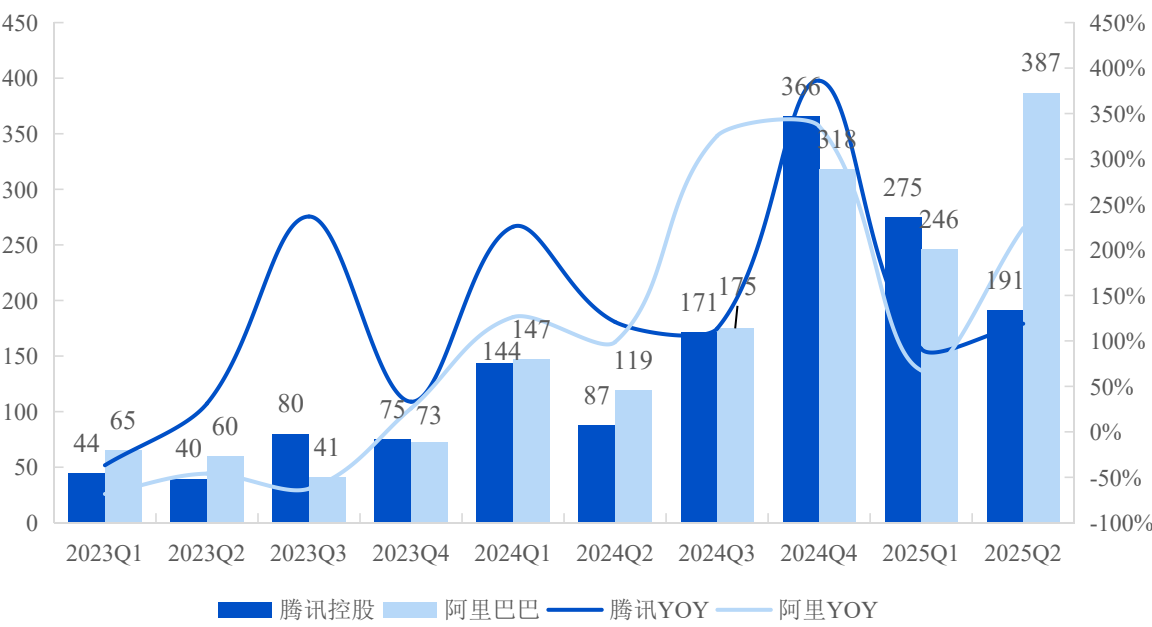
4.3.1 互联网：国内厂商云业务营收持续增长，驱动资本开支高增

- 2025Q2，阿里巴巴云业务收入334亿元，同比+26%，得益于AI需求的激增以及客户为支持AI工作负载而加大对公共云服务的采用。AI相关产品收入已连续第八个季度保持三位数增长，AIDC（阿里国际）团队的收入同比增长19%。
- 2025Q2，阿里巴巴资本开支387亿元，并在财报电话会上重申三年3800亿元人民币AI资本开支计划；腾讯控股资本开支同比增长119%，至191亿元，自2024Q4战略加速以来，腾讯累计资本开支达832亿元，主要用于支持AI相关业务的发展。

图：2023Q1-2025Q2国内CSP厂商云业务收入（亿元）



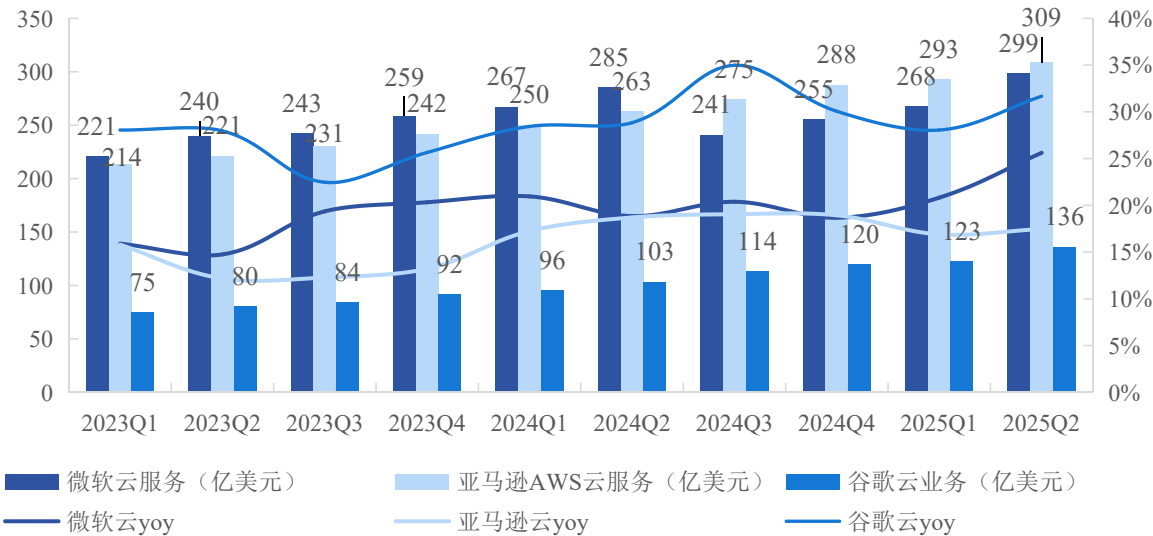
图：2023Q1-2025Q2阿里巴巴与腾讯控股资本开支（亿元）



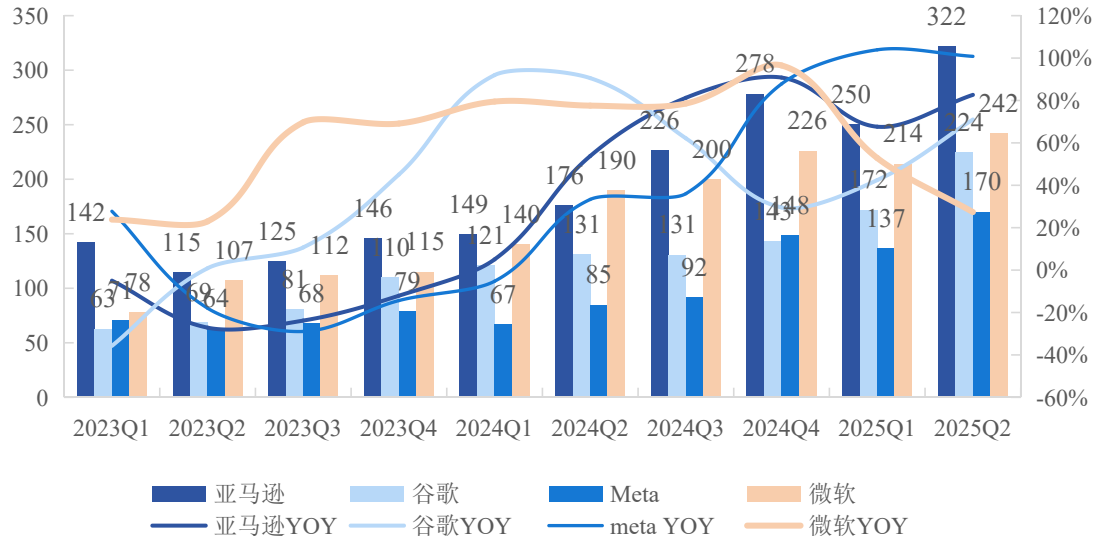
4.3.1 互联网：北美CSP纷纷上修CapEx，主要用于AI基础设施建设

公司	云业务情况、CapEx指引调整及其主要投向
谷歌	- 二季度CapEx同比+71%至224.46亿美元，约三分之二用于服务器，三分之一用于数据中心及网络设备。7月23日，公司将2025年CapEx预期从750亿美元上调至850亿美元，主要用于AI基础设施建设。 - 谷歌云业务增速领先，Q2营收达136.24亿美元，同比+32%，增速较Q1的+28%进一步加速。AI业务全面发展，目前每月处理超980万亿tokens（5月为480万亿tokens/月），Gemini应用目前MAU已超4.5亿，用户增长和参与度仍在提升，日请求量较Q1增长超50%。
微软	- 二季度CapEx（含融资租赁）同比+27%至242亿美元，接近一半的支出用于服务器（包括CPU和GPU）。微软7月30日表示，预计Q3的CapEx将超300亿美元。 - 智能云业务整体营收为299亿美元，同比+26%，其中Azure相关云服务收入增长39%，全年营收首次突破750亿美元大关；微软云业务收入首次突破1680亿美元，同比增长23%。
Meta	Meta二季度CapEx（含融资租赁）为170亿美元。7月30日，公司调整2025年CapEx区间为660–720亿美元（原区间为640–720亿美元），并表示2026年CapEx仍将显著增长，主要用于扩展生成式AI所需的服务器、网络和数据中心建设。
亚马逊	- 亚马逊二季度CapEx，同比+83%至322亿美元，公司表示2025年的资本支出预计将达1000亿美元，绝大部分用于AI和云服务AWS。 - 二季度AWS云服务营收同比+17.5%，增速虽不及谷歌与微软云业务，但CEO表示当前AWS在AI领域面临“需求超过供给”局面，电力才是最大瓶颈。

图：2023Q1–2025Q2海外CSP云业务收入（亿元）



图：2023Q1–2025Q2海外CSP资本开支（亿元）



4.3.2 大型AI算力集群陆续落地，主权AI成为新增需求方

- 主权AI正在崛起。英伟达在8月27日财报电话会上表示，今年有望实现200亿美元的主权AI收入。英伟达认为，主权AI正在崛起，比如，欧盟计划投资200亿欧元，在法国、德国、意大利和西班牙建立20个AI工厂，包括5个超级工厂，将AI计算基础设施提升十倍。鸿海预计未来五年主权AI领域投资有望超1万亿美元，主要包括美国的Stargate（5000亿美元）、欧盟InvestAI（2000亿欧元）以及沙特Humain AI（1000亿美元）。

表：大型AI算力集群与主权AI进展

分类	公司	集群名称	进展
大型AI算力集群	OpenAI	星际之门等	7月21日，Sam Altman表示今年年底前将上线的GPU数量会远超过100万块，且未来上线数量可能提高100倍。
	xAI	Colossus	7月23日，马斯克宣布Colossus 2将在几周后上线，首批有55万块GB200/GB300，xAI计划在未来五年内将部署相当于5000万个H100等级的AI GPU。
	Meta	Prometheus	是Meta即将投入使用的第一个GW级别（Gigawatt-scale）的AI集群，预计将于2026年上线。
		Hyperion	将在未来几年扩展到5GW。
主权AI	美国星际之门（5000亿美元）		7月22日，OpenAI宣布将与甲骨文共同开发4.5GW额外的“星际之门”数据中心容量。目前，位于德克萨斯阿布林的星际之门1号站已经上线运行。 8月19日，软银收购鸿海俄亥俄州AI服务器生产基地，计划将其改造为专门生产用于“星际之门”（Stargate）计划的AI服务器及相关设备的核心据点。
	欧洲（2000亿欧元）		7月31日，欧洲首个“星际之门”数据中心项目将在挪威纳尔维克设立，规划容量为230MW，并有意额外拓展290MW，OpenAI将作为算力的初始采购方。
	中东（沙特1000亿美元）		- 阿联酋：当地时间5月22日，OpenAI宣布推出Stargate UAE（星际之门阿联酋）项目，首批200MW产能将于2026年投入使用，该场地最终将容纳5GW的数据中心。 - 沙特阿拉伯：Humain宣布与英伟达建立战略合作伙伴关系，在沙特建设AI工厂。预计未来五年，新工厂将由数十万块英伟达最先进的GPU驱动，发电能力高达500MW，首个部署阶段将包括一台配备18000块英伟达GB300 Grace Blackwell AI超级计算机，并搭载英伟达InfiniBand网络。AMD也与humain达成合作，5年内共同投资达100亿美元，部署500MW的AI计算能力。

- 目前，中国移动、中国电信均已规划针对国产算力的万卡算力集群，且在不断加大在国产AI算力的投入。我们认为，随着国产AI芯片的出货量增长，相关服务器厂商有望持续受益。

表：2023-2024年运营商AI服务器采购情况

公司	项目名称		总数量	总金额	AI服务器数量	国产化比例
中国移动	《2023年至2024年新型智算中心(试验网)采购》	标包1-11	人工智能服务器1204台、高性能无损交换机324台（含集群管理软件2050端口）、集群管理服务器2台、RoCE交换机204台、分布式文件存储119PB、智算资源池化软件（虚拟化软件）608套	-	人工智能服务器1204台	-
		标包12	扣卡风冷型特定场景AI训练服务器106台，扣卡液冷型特定场景AI训练服务器1144台	24.74亿元	人工智能服务器及配套产品1250台	-
	《2024年至2025年新型智算中心采购招标》		7994台人工智能服务器及配套产品、60台白盒交换机，总计8054台设备	191亿元	训练服务器7994台	中标候选人： ①昆仑技术21.05%； ②华鲲振宇17.54%； ③宝德计算机份额15.79%； ④百信份额14.04%； ⑤长江计算份额12.28%； ⑥神州鲲泰份额10.53%； ⑦湘江鲲鹏份额8.77%
中国联通	《2024年中国联通人工智能服务器集中采购项目》		2503台人工智能服务器，688台关键组网设备RoCE交换机，总计3191台	-	人工智能服务器2503台	昆仑、宝德、虹信和长江4家入围
中国电信	《AI算力服务器（2023-2024年）集中采购项目》		AI服务器 4175台	84.63亿元	AI服务器4175台	I系列规模为2198台，G系列规模为1977台。I系列CPU采用Intel至强可扩展处理器，G系列CPU采用鲲鹏处理器
	《服务器（2024-2025年）集中采购项目》		服务器15.6万台	-	GPU服务器13135台	G系列（国产化系列）数量达10.53万台，占比达67.5%；GPU服务器中，A系列6295台，G系列6840台

4.3.4 智算：政府引导智算中心发展，促进国产AI芯片需求增长

- 政策引导智算中心发展，加速算力设施国产化进程。如青岛海之心智算中心要求国产人工智能加速卡占比超70%，宁夏推动华为等自主可控产业园区落地等。2023年10月8日，工信部、网信办等六部门联合印发《算力基础设施高质量发展行动计划》，提出计算力、运载力、存储力、应用赋能四方面到2025年量化指标。我们预计国产GPU芯片需求有望持续增长。

表：《算力基础设施高质量发展行动计划》量化指标

	序号	指标	2023年	2024年	2025年
计算力	1	算力规模(EFLOPS)	220	260	300
	2	智能计算中心(个)	30	40	50
	3	智能算力占比(%)	25	30	35
运载力	4	重点应用场所光传送网(OTN)覆盖率(%)	50	65	80
	5	SRv6 等创新技术使用占比(%)	20	30	40
	6	国家枢纽节点数据中心集群间网络时延达标率(%)	65	75	80
存储力	7	存储总量(EB)	1200	1500	1800
	8	先进存储容量占比(%)	25	28	30

表：我国引导智算中心发展相关政策

部门	政策/公告	内容
宁夏回族自治区人民政府办公厅	《自治区人民政府办公厅关于促进全国一体化算力网络国家枢纽节点宁夏枢纽建设若干政策的意见》	加大信创企业扶持力度，对基础软硬件实现国产化率90%以上的数据中心，给予企业最高不超过1000万元奖励。重点加强基础芯片、自主指令集的产学研及配套产业建设。
宁夏回族自治区发展改革委	《全国一体化算力网络国家枢纽节点宁夏枢纽建设2023年工作要点》	推动龙芯中科、中电子、中兴、华为等自主可控产业园区落地，鼓励以国产化CPU、GPU、操作系统等自主可控产品为底座的信创云平台自主研发，打造安全可信计算、网络和存储能力。
青岛人工智能产业园	《青岛“海之心”人工智能计算中心项目》	要求国产人工智能加速卡占比≥70%
成都市经信局	《成都市围绕超算智算加快算力产业发展的政策措施》	鼓励智算中心建设国产自主可控、安全可靠的人工智能算力基础设施和技术路线生态。
南京麒麟科创园	《南京智能计算中心算力推广办法（试行）》	以“算力券”和“算力折扣”两种形式为广大企业提供更优质的算力资源、更有力的服务保障。
北京市人民政府	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》	在人工智能产业聚集区新建或改建升级人工智能商业化算力中心，加强国产芯片部署应用，推动自主可控软硬件算力生态建设
成都市经信局	《成都市加快大模型创新应用推进人工智能产业高质量发展的若干措施》	引导国家超算成都中心、成都智算中心合理扩容，支持鲲鹏、昇腾、海光等自主可控芯片部署，提高自主研发算力设备比例。

第五章

投资建议与风险提示

大模型训推带动AI算力需求增长，AI芯片、CPU、AI服务器及其组件、光模块、液冷、IDC等环节有望持续受益。维持对计算机行业“推荐”评级。

● 相关公司

1) **AI芯片**：海光信息、寒武纪、百度集团、英伟达、AMD。

2) **CPU**：海光信息、龙芯中科、中国长城、Intel、AMD。

3) **服务器整机**：工业富联、中科曙光、浪潮信息、华勤技术、紫光股份、神州数码、中国长城、软通动力、烽火通信、拓维信息、鸿海、纬创、广达、超微电脑、戴尔。

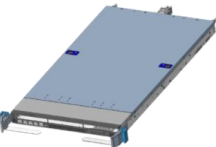
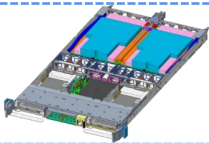
4) **服务器组件**：①**散热**：曙光数创、英维克、飞荣达、同方股份、申菱环境、高澜股份、奇鋆科技、双鸿、健策、VERTIV；②**铜连接**：安费诺、沃尔核材、华丰科技；③**HBM**：SK海力士、三星、美光、赛腾股份、联瑞新材；④**电源**：欧陆通、麦格米特、中国长城。

5) **算力租赁**：协创数据、宏景科技、有方科技、鸿博股份。

6) **数据中心**：云赛智联、奥飞数据、润泽科技、大位科技、光环新网、数据港、科华数据、万国数据、城地香江。

5.1 投资建议与相关公司

- ☆ **ASIC**
 - 内地：阿里巴巴、百度集团、芯原股份
 - 海外：博通、Marvell
- ☆ **GPU**
 - 内地：华为海思、海光信息、寒武纪
 - 海外：英伟达、AMD、Intel
- ☆ **CPU**
 - 内地：华为海思、海光信息、飞腾、龙芯中科
 - 海外：英伟达、AMD、Intel
- ☆ **主板**
 - 内地：沪电股份、胜宏科技、深南电路、生益科技
 - 中国台湾：泰安电脑（神达）、技嘉、华擎、华硕、英业达、纬创、研华
 - 美国：Intel、Supermicro
- ☆ **Compute Tray * 10**
 - 内地：工业富联
 - 中国台湾：纬创
- ☆ **Switch Tray * 9**
 - 内地：工业富联
 - 中国台湾：纬创
- ☆ **Compute Tray * 8**
- ☆ **电源**
 - 台达、中国长城、欧陆通、麦格米特、泰嘉股份
- ☆ **AIDC**：润泽科技、世纪互联、万国数据等



- ☆ **存储 NAND**
 - 公司：长江存储、兆易创新、佰维存储、东芯股份
 - 海外：三星、西部数据、铠侠、SK海力士、美光

- ☆ **光模块**
 - 中际旭创、新易盛、光迅科技、天孚通信、华工科技

- ☆ **算力租赁**
 - Coreweave、协创数据、有方科技、宏景科技

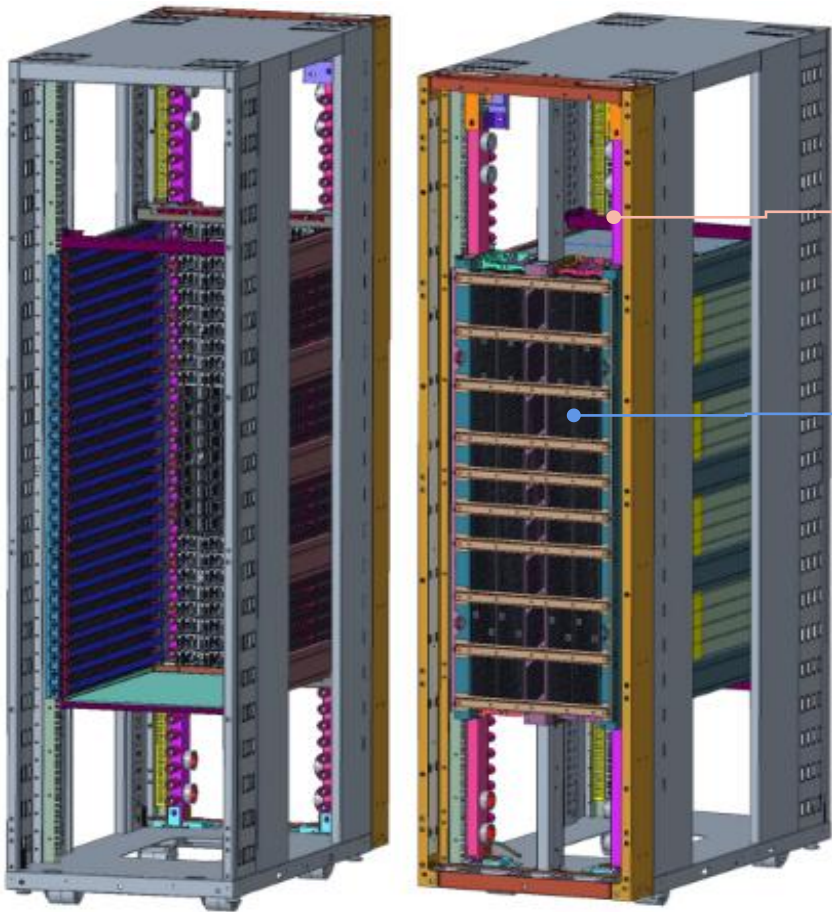
- ☆ **网络**：思科、华为、英伟达、Arista、紫光股份、中兴通讯、星网锐捷、盛科通信

- ☆ **散热（风冷/液冷）**
 - 内地：曙光数创、飞荣达、中航光电、英维克、同飞股份、申菱环境、高澜股份、依米康
 - 海外：奇鋆、双鸿、健策、建准、高力、VERTIV

- ☆ **线缆（铜连接）**
 - 内地：沃尔核材、华丰科技、兆龙互连、鼎通科技、立讯精密、神宇股份
 - 海外：安费诺

- ☆ **整机（OEM/ODM）**
 - 1) 内地：工业富联、浪潮信息、中科曙光、紫光股份、中国长城、华勤技术、神州数码、拓维信息、烽火通信、软通动力；2) 中国台湾：鸿海、广达、纬创、英业达、纬颖；3) 美股：戴尔、超微电脑

- ☆ **BIOS/BMC**
 - BIOS/BMC：1) 内地：卓易信息；2) 中国台湾：新唐、系微、信骅



正面

背面

注：蓝字为零部件。上图展示以英伟达GB300 NVL72整机柜为例，仅列示了部分参与公司

- 1) 宏观经济影响下游需求：宏观经济环境下行，将影响客户对信息化基础设施的采购需求；
- 2) 大模型产业发展不及预期：行业的核心驱动力是人工智能大语言模型的训练和推理对算力的需求，如果大模型行业发展不及预期将会影响AI算力的相关需求；
- 3) 市场竞争加剧：IT 产品和服务行业是成熟且完全竞争的行业，新进入者可能加剧整个行业的竞争态势；
- 4) 中美博弈加剧：国际形势持续不明朗，美国不断通过“实体清单”等方式对中国企业实施打压，若中美紧张形势进一步升级，将可能导致中国半导体供应链供应受到影响；
- 5) 相关公司业绩不及预期：市场环境变化、公司治理情况变化、其他非主营业务经营不及预期等原因或将导致相关公司的整体业绩不及预期。
- 6) 各公司并不具备完全可比性，对标的相关资料和数据仅供参考。

计算机小组介绍

刘熹，计算机行业首席分析师，上海交通大学硕士，多年计算机行业研究经验，善于把握由技术、政策驱动的科技产业新趋势，致力于进行前瞻重磅推荐。2024年金牌分析师，2024年同花顺最受欢迎分析师。

唐锦珂，计算机行业研究助理，中山大学岭南学院硕士，主要覆盖AI芯片、服务器、算力租赁、AIDC。

分析师承诺

刘熹，本报告中的分析师均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，独立，客观的出具本报告。本报告清晰准确的反映了分析师本人的研究观点。分析师本人不曾因，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取到任何形式的补偿。

国海证券投资评级标准

行业投资评级

- 推荐：行业基本面向好，行业指数领先沪深300指数；
- 中性：行业基本面稳定，行业指数跟随沪深300指数；
- 回避：行业基本面向淡，行业指数落后沪深300指数。

股票投资评级

- 买入：相对沪深300 指数涨幅20%以上；
- 增持：相对沪深300 指数涨幅介于10%～20%之间；
- 中性：相对沪深300 指数涨幅介于-10%～10%之间；
- 卖出：相对沪深300 指数跌幅10%以上。

免责声明

本报告的风险等级定级为R3，仅供符合国海证券股份有限公司（简称“本公司”）投资者适当性管理要求的客户（简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。客户及/或投资者应当认识到有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司的完整报告为准，本公司接受客户的后续问询。

本公司具有中国证监会许可的证券投资咨询业务资格。本报告中的信息均来源于公开资料及合法获得的相关内部外部报告资料，本公司对这些信息的准确性及完整性不作任何保证，也不保证其中的信息已做最新变更，也不保证相关的建议不会发生任何变更。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。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。报告中的内容和意见仅供参考，在任何情况下，本报告中所表达的意见并不构成对所述证券买卖的出价和征价。本公司及其本公司员工对使用本报告及其内容所引发的任何直接或间接损失概不负责。本公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或者金融产品等服务。本公司在知晓范围内依法合规地履行披露义务。

风险提示

市场有风险，投资需谨慎。投资者不应将本报告为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向本公司或其他专业人士咨询并谨慎决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息。本报告不构成本公司向该机构之客户提供的投资建议。

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。本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

郑重声明

本报告版权归国海证券所有。未经本公司的明确书面特别授权或协议约定，除法律规定的情况外，任何人不得对本报告的任何内容进行发布、复制、编辑、改编、转载、播放、展示或以其他方式非法使用本报告的部分或者全部内容，否则均构成对本公司版权的侵害，本公司有权依法追究其法律责任。

国海证券 · 研究所 · 计算机研究团队

心怀家国，洞悉四海



国海研究上海

上海市黄浦区绿地外滩中心C1栋
国海证券大厦

邮编：200023

电话：021-61981300

国海研究深圳

深圳市福田区竹子林四路光大银行大厦28F

邮编：518041

电话：0755-83706353

国海研究北京

北京市海淀区西直门外大街168号腾达大厦25F

邮编：100044

电话：010-88576597