

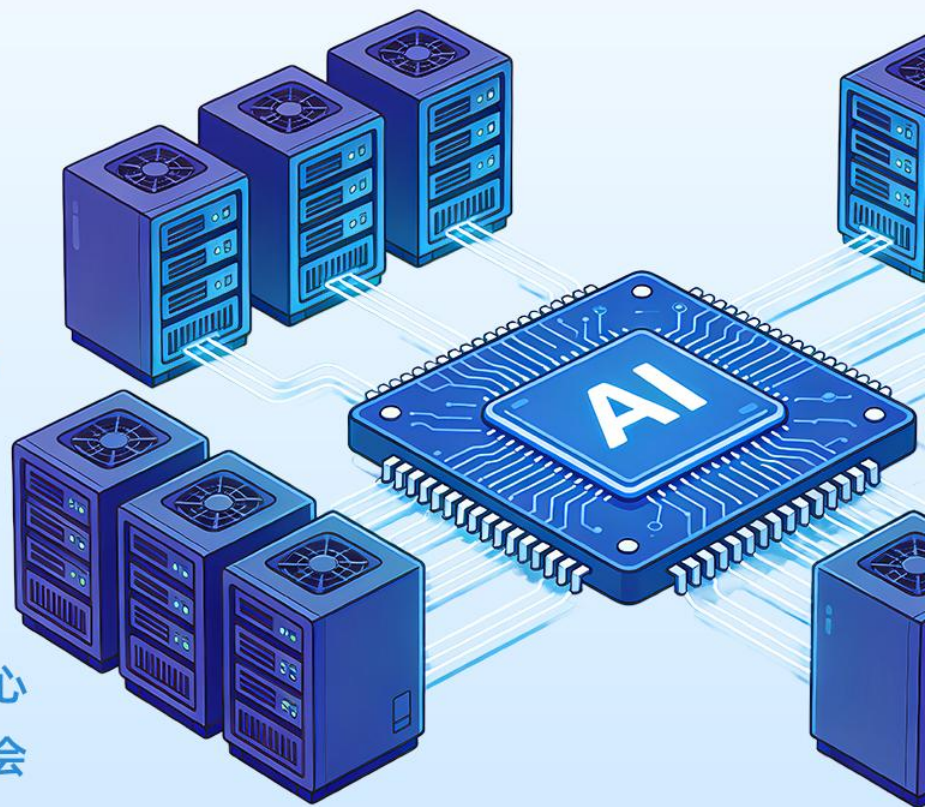
2026 全球AI算力 发展研究报告

指导机构：

中国智能计算产业联盟、国家超级计算天津中心
天津市人工智能学会、深圳市人工智能行业协会

撰写机构：至顶智库

支持机构：至顶科技



报告背景

当前，全球算力产业正迈入“智算驱动、体系重构”的全新发展阶段。伴随“词元经济”的兴起，算力已成为支撑国家技术突破、产业竞争与战略布局的关键基础要素。在此背景下，AI芯片、AI工作站、AI服务器及AI算力中心等关键领域迎来重要突破。面向大模型训练与推理需求，AI芯片正围绕GPU、TPU、NPU等多元方向持续演进，异构计算、高速互联及软件栈生态加速完善；AI工作站向专业化与多样化方向发展；AI服务器向集群化及高速互联架构升级；AI算力中心则进入以超大规模计算集群和绿色低碳为特征的新阶段。AI算力基础设施正从传统信息技术支撑逐步演变为驱动科技创新与工业革命的战略底座。

在此背景下，中国智能计算产业联盟、国家超级计算天津中心、天津市人工智能学会、深圳市人工智能行业协会、至顶科技、至顶智库联合发布《2026全球AI算力发展研究报告》。报告从智能时代的算力跃迁出发，全面总结全球AI算力的发展背景、关键环节（AI芯片、AI工作站、AI服务器以及AI算力中心）、应用场景，对算力产业的关键领域、核心技术进行分析解读。最后，报告展望AI算力未来发展趋势。报告为决策部门、行业从业者、教育工作者以及社会公众更好了解全球人工智能算力的发展情况提供参考。

2026年5月

2026全球算力产业十大趋势

01

异构算力架构从“CPU+GPU”为主向“CPU+GPU+XPU”多元发展路线演进，有望成为智能算力时代的主流技术范式。

02

中国算力产业发展从单点技术突破迈向全栈体系协同，国家产业生态不断成熟，地方产业布局重点突出。

03

超节点与高速互联有效提升算力效能，将成为全球构建新型算力基础设施的重要路径。

04

伴随“龙虾”等多智能体框架出现，AI应用从交互智能迈向执行智能，催生新的推理算力需求。

05

算力赋能经济社会发展的边界不断拓展，从科学智能、具身智能等前沿领域向工业、交通、能源等行业全面渗透。

来源：至顶智库结合公开资料及专家访谈整理

2026全球算力产业十大趋势

06

伴随具身智能与世界模型的快速发展，多模态数据持续增加，拓展智能应用边界，激发潜在算力需求。

07

算力呈现“云-边-端”协同发展趋势，算力中心与边缘端侧AI设备协同，满足各类场景应用需求。

08

新能源推动算力中心能源供给方式变革，风光储一体化、核能、氢能将成为未来实现低碳算力的发展方向。

09

太空算力借助空间能源、广域覆盖等独特优势，有望成为提供超智融合计算服务的新型算力基础设施。

10

词元消耗量将成为衡量一国智能化发展的重要指标，算力作为支撑词元经济发展的重要基础，重塑未来经济发展模式。

来源：至顶智库结合公开资料及专家访谈整理

开篇：智能时代的算力跃迁

近年来，人工智能实现跨越式发展，先后完成从深度学习时代到生成式AI时代的演进，当前正稳步迈向智能体与具身智能时代。为支撑人工智能的发展需求，算力生态的核心环节一芯片、整机与计算集群均实现性能的全面升级。芯片算力由TFLOPS量级提升至数十PFLOPS，整机部署形态从单机八卡演进为千卡级超节点架构，计算集群规模从千卡集群拓展至数十万卡集群，集群功耗从千瓦级提升到吉瓦级。

深度学习时代

芯片

并行计算架构，算力TFLOPS

整机

单机八卡，算力PFLOPS

集群

百卡 → 千卡集群

功耗

千瓦级 → 兆瓦级

生成式AI时代

芯片

多Die封装，算力PFLOPS

整机

超节点单机数十卡，算力数百PFLOPS

集群

万卡 → 十万卡集群

功耗

兆瓦级

智能体与具身智能时代

芯片

异构计算集成，算力数十PFLOPS

整机

千卡超节点，算力数十EFLOPS

集群

数十万卡集群

功耗

吉瓦级

来源：2026中国信通院深度观察报告会，至顶智库结合公开资料整理绘制

开篇：智能时代的算力跃迁

在数据准备阶段、模型训练阶段、模型推理阶段的各环节均产生算力消耗，各阶段算力消耗的量级差异明显。在模型预训练阶段，超大规模的模型预训练需要多达万卡级算力支撑；模型推理阶段超大规模模型需要千卡算力；数据准备阶段算力需求相对较低，需要数十到数百卡算力规模。

数据准备阶段

数据采集

企业内部数据
用户交互数据
第三方数据
公开数据集

数据清洗

去重去噪
补充缺失值
剔除样本
处理异常值

数据标注

文本
图像
语音
视频
点云

模型训练阶段

预训练 Pre-training

采用大规模数据学习，使模型掌握语言、视觉、语音或多模态数据中的基础规律、通用知识和表示能力。

数据加载
前向传播
损失计算
反向传播
梯度同步
参数更新

监督微调 SFT

使用人工标注样本继续训练模型，使模型更好完成特定任务、遵循特定指令或适应特定领域。

全参数微调
LoRA微调
QLoRA微调
Adapter微调

压缩优化 Compression Optimization

模型压缩旨在降低参数规模、存储体积和计算复杂度；模型优化则是提升模型在硬件上的推理效率和资源利用率。

量化
剪枝
知识蒸馏
计算图优化
算子融合

模型推理阶段

预填充 Prefill

模型一次性处理输入的完整提示词，完成上下文理解，为后续生成反馈结果准备KV缓存的过程。

Prefix Caching
Chunked Prefill
Disaggregated Prefill

解码 Decode

模型基于已有输入与已生成内容，按照自回归方式逐个生成新token过程。

KV Cache优化
In-flight Batching
Speculative Decoding

数据准备算力需求

数据采集 数据清洗 数据标注
大规模 大规模 大规模
数十卡 数十卡→数百卡 数百卡

模型训练算力需求

预训练 二次训练 全参微调 局部微调
超大规模 大规模 较小规模 小规模
千卡-万卡 数百卡→千卡 单机8卡起步 单机1卡起步

模型推理算力需求

个人场景 企业场景
超大规模 小规模-大规模
千卡以上 数十卡→数百卡

来源：部分参考《华为AI DC白皮书》，至顶智库结合公开资料整理绘制

报告目录

第一章 全球AI算力发展背景及产业概况

第二章 全球AI芯片发展情况

第三章 全球AI工作站及服务器发展情况

第四章 全球AI算力中心发展情况

第五章 AI算力典型应用场景

第六章 AI算力产业发展趋势

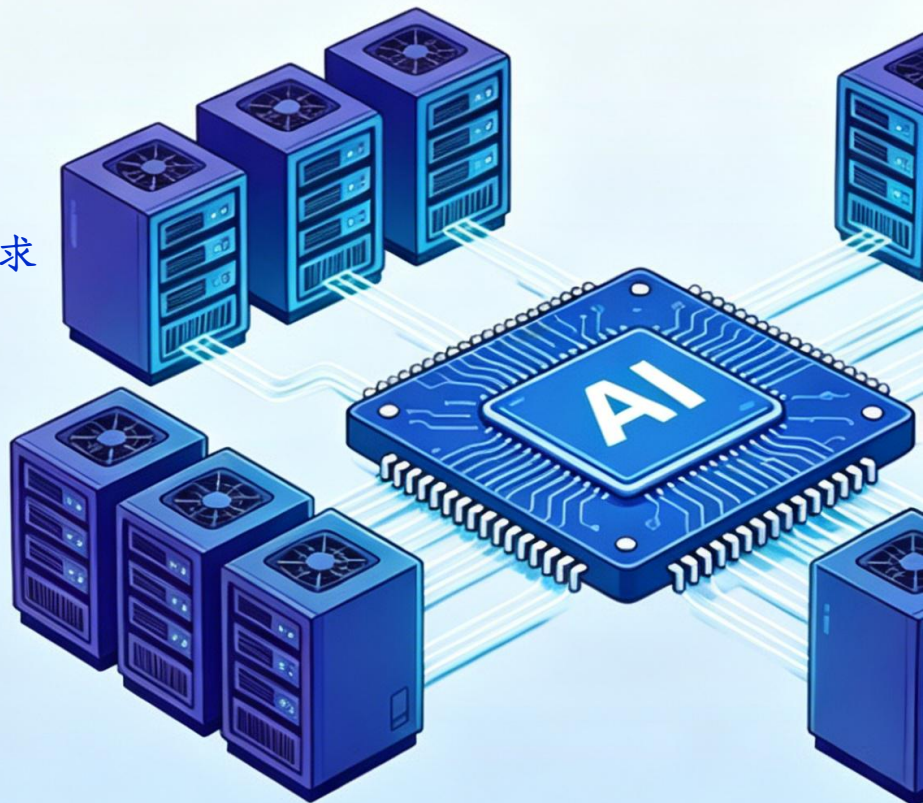
第一章 全球AI算力发展背景及产业概况

技术驱动：推理模型与Agent发展驱动算力迭代

需求驱动：经济发展与社会进步共同拉动算力需求

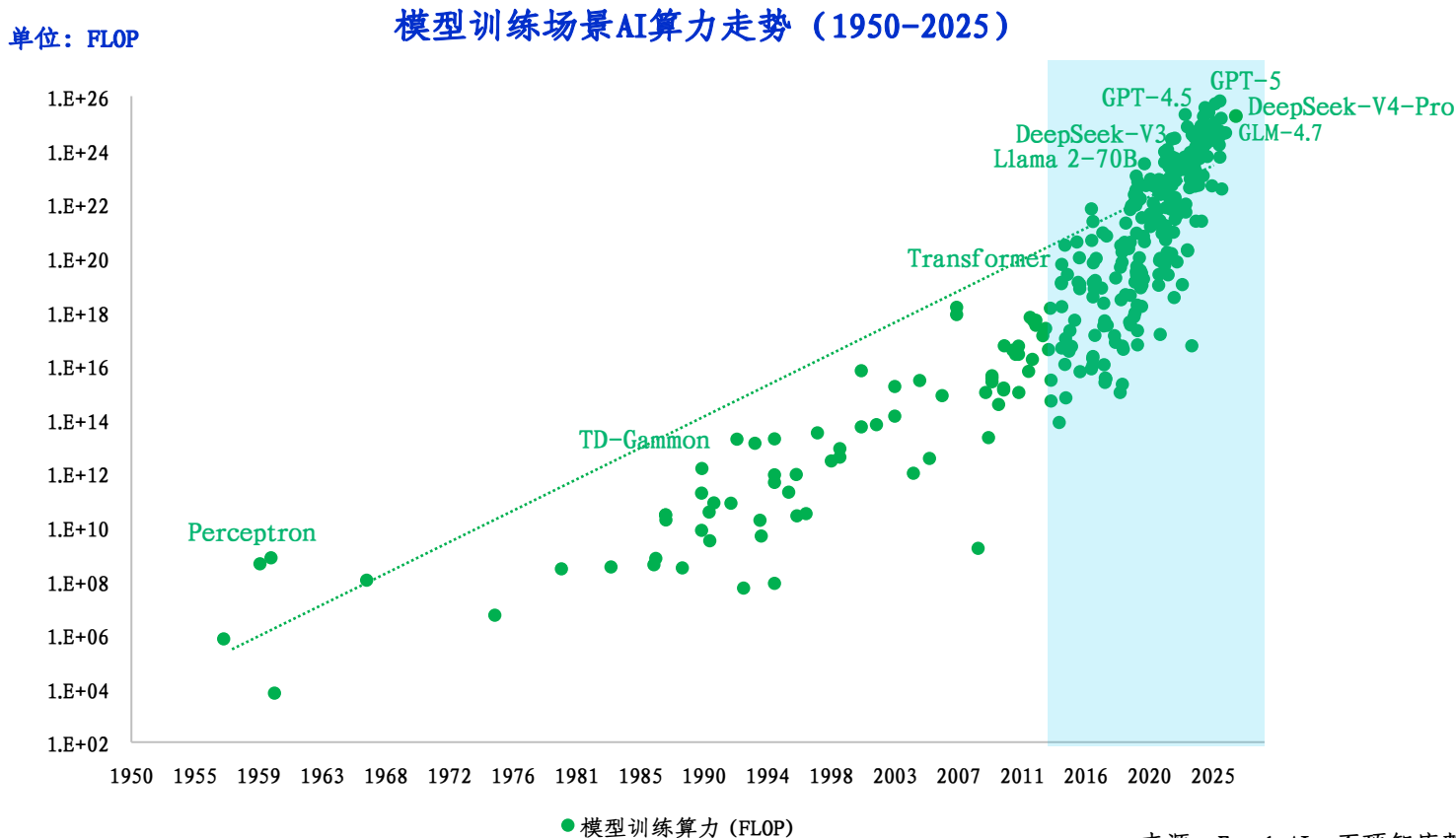
政策支撑：全球政策持续加码算力基建投资

全球AI算力图谱、算力产业生态、算力概念梳理



1.1 技术驱动：Scaling Law持续演进，推动算力规模扩张

Scaling Law持续演进正加速AI算力规模增长。随着模型参数规模、训练数据量和计算量持续扩展，模型性能显著提升。根据Epoch AI数据统计，模型训练所需的算力规模走势从平缓演变为指数级上升。Scaling Law使“扩大规模”成为推动模型能力提升的重要路径，直接驱动算力基础设施的扩张，进一步带动全球在高性能AI芯片研发、大规模AI算力中心以及算力集群建设的持续投入。



1.1 技术驱动：训练消耗算力规模影响模型性能表现

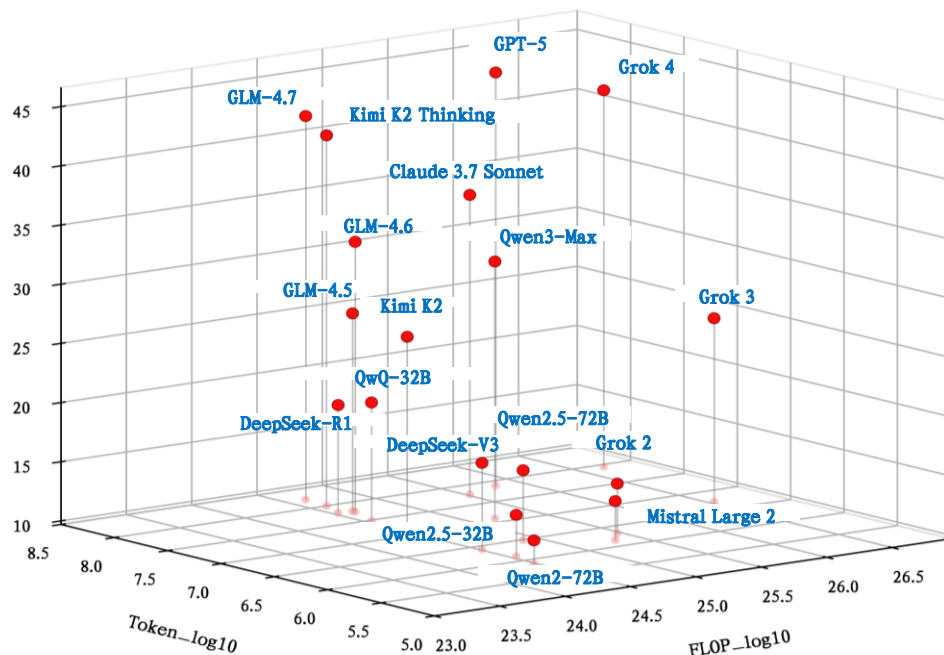
AI模型演进推动训练与推理阶段的算力需求。根据Artificial Analysis指数表现，前沿模型在迈向更高智能水平的过程中，普遍需要更强的训练算力和更高推理资源投入。尽管不同模型在训练消耗算力及Token使用量上存在差异，但高性能AI模型更多分布于高算力、高Token消耗区间，表示模型性能提升仍建立在高算力基础上，训练与推理两端的算力需求仍将持续增长。

Token使用量、训练消耗算力及模型指数三维象限图

Artificial Analysis指数表现

图表说明

1. **模型训练消耗算力**来源于Epoch AI
2. **Artificial Analysis指数表现**来源于Artificial Analysis Intelligence Index, 该指数基于以下基准进行评估, GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt
3. **模型Token使用量**来源于Artificial Analysis网站 Output Tokens Used in Artificial Analysis Intelligence Index (Log Scale), 表示相关模型运行Artificial Analysis指数所需Token



模型输出Token使用量

模型训练消耗算力

来源：Artificial Analysis, Epoch AI, 至顶智库整理绘制

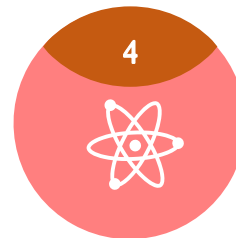
1.2 需求驱动：经济发展与社会进步共同拉动算力需求

AI算力作为推动一国创新发展的关键要素，将为全球经济发展注入澎湃动能，而经济发展与社会进步的愿景也将拉动算力需求。伴随各行业智能化转型程度不断加深，推理场景对于算力的需求显著提升，各类终端对于本地AI算力的需求持续增加，前沿科学研究也需要高性能算力支撑。在此背景下，全球各国不断加大算力产业的投入力度，紧抓未来发展的重要机遇。

下游需求推动算力产业发展

智能转型快速推进

随着各行各业加速向智能化转型，人工智能技术的应用场景日益丰富，推动各行业对算力的需求不断上升。如金融、医疗、制造等行业广泛采用AI进行数据分析、智能决策和自动化生产，相关应用需要强大的计算支撑。



推理场景需求增加

人工智能发展从模型训练到场景应用方向转型，推理算力需求呈现爆发式增长。伴随智能体等应用呈现规模化落地，推理任务从辅助环节提升为AI算力的核心负载，消耗计算量持续攀升。

智能终端加速落地

智能手机、智能汽车、智能安防等终端设备对本地AI计算的需求也在快速增加。如智能汽车需要实时处理复杂的传感器数据，以实现智能驾驶决策，因此边缘计算场景对于算力的需求不断攀升。

科学研究持续突破

在科技创新发展战略推动下，国家在生物医药、新材料、航空航天、量子信息、深海深空探测等前沿科技领域取得研发突破，高度依赖超算及高性能算力的支撑。

来源：至顶智库结合公开资料整理绘制

1.3 政策支撑：全球主要国家及地区出台政策，持续加码算力基建

全球主要国家及地区算力政策

国家/地区	发布机构	发布时间	政策名称	内容
美国	美国国家科学技术委员会 (NSTC)	2024.2	《关键和新兴技术清单》	➤ 将先进计算与人工智能列为《2024年关键和新兴技术清单》的优先方向，与半导体和微电子、集成通信与网络技术等并列对国家安全具有特别重要意义的核心领域。
	美国白宫	2025.7	《美国人工智能行动计划》	➤ 围绕算力基础设施建设提出系统性举措：一是简化数据中心、半导体制造及能源设施的审批流程；二是重组CHIPS计划办公室，推动半导体制造回流本土；三是强化AI计算出口管制。
欧盟	欧盟委员会	2025.5	人工智能大陆行动计划	➤ 将“构建欧洲AI计算基础设施”作为核心支柱，通过AI工厂、超级计算机升级及数据中心扩容等措施，提升欧盟在高性能计算与AI芯片支撑能力方面的战略自主性。
中国	中华人民共和国国务院	2025.8	《关于深入实施“人工智能+”行动的意见》	➤ 完善全国一体化算力网，充分发挥“东数西算”国家枢纽作用，加大数、算、电、网等资源协同。加强智能算力互联互通和供需匹配，创新智能算力基础设施运营模式，鼓励发展标准化、可扩展的算力云服务，推动智能算力供给普惠易用、经济高效、绿色安全。
	中华人民共和国国务院	2026.3	《政府工作报告》（2026）	➤ 提出深化拓展“人工智能+”，促进新一代智能终端和智能体加快推广。实施超大规模智算集群、算电协同等新基建工程，加强全国一体化算力监测调度，支持公共云发展。

来源：相关政府网站，至顶智库整理绘制

1.4 全球AI算力图谱

AI算力中心



AI框架



AI服务器



AI工作站



AI芯片



来源：至顶智库结合公开资料整理绘制

1.5 全球科技领军企业算力布局



AI 算力中心		在全球共有31个已建成数据中心 亚洲：新加坡等 欧洲：爱尔兰都柏林等 北美洲：美国俄亥俄州等 南美洲：智利基利库拉等	在全球共有50个已建成数据中心 亚洲：新加坡等 欧洲：爱尔兰等 北美洲：美国加州等 南美洲：智利圣地亚哥等 大洋洲：澳大利亚堪培拉等 非洲：南非约翰内斯堡等
AI 服务器	✓ Vera Rubin NVL72 ✓ HGX Rubin NVL8 ✓ DGX Rubin NVL8 ✓ DGX B200 ✓ GB300 NVL72		
AI 工作站	✓ DGX Spark ✓ DGX Station		
AI 芯片	✓ Rubin GPU ✓ Groq 3 LPU ✓ Blackwell GPU ✓ GB10 ✓ Blackwell Ultra	✓ TPU v5p ✓ TPU v6e (Trillium) ✓ TPU 7x (Ironwood) ✓ TPU 8i ✓ TPU 8t	✓ Maia 100 ✓ Maia 200

来源：各企业官网，至顶智库结合公开资料整理绘制

1.5 全球科技领军企业算力布局

	aws	Meta
AI 算力中心	在全球共有120个已建成的数据中心 亚洲：泰国等 欧洲：英国伦敦等 北美洲：美国加利福尼亚州等 南美洲：巴西圣保罗等 大洋洲：澳大利亚墨尔本等	在全球共有32个已建成数据中心 亚洲：新加坡等 欧洲：爱尔兰克隆尼等 北美洲：美国亚利桑那州等
AI 服务器		
AI 工作站		
AI 芯片	✓ Amazon Trainium3	✓ MTIA 300 ✓ MTIA 400 ✓ MTIA 450 ✓ MTIA 500

来源：各企业官网，至顶智库结合公开资料整理绘制
insights.zhiding.cn

1.5 全球科技领军企业算力布局

	HUAWEI	Baidu 百度	阿里云
AI 算力中心	在全球共有30个已建成数据中心 亚洲：中国北京等 欧洲：法国巴黎等 南美洲：智利圣地亚哥等 非洲：埃及开罗等	以国内市场为主，重点覆盖京津冀、长三角、珠三角等地	在全球共有94个已建成数据中心 亚洲：泰国曼谷等 欧洲：德国法兰克福等 北美洲：美国硅谷等
AI 服务器	✓ 昇腾384超节点Atlas 900 A3 SuperPoD ✓ Atlas 950 SuperPoD ✓ Atlas 850E	✓ 天池256超节点 ✓ 天池512超节点	✓ 磐久128超节点AI服务器
AI 工作站			
AI 芯片	✓ 昇腾910C ✓ 昇腾950PR ✓ 昇腾950DT ✓ 昇腾960 ✓ 昇腾970	✓ M100 ✓ M300	✓ 真武810E ✓ 含光800

来源：各企业官网，至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—浮点精度

浮点精度是指计算机浮点数表示和计算时所能达到的精确程度。Floating Point (FP) 表示浮点精度，由符号位、指数位和尾数位三部分组成。其中，符号位用于表示数值正负；指数位决定小数点位置，控制数值范围；尾数位表示数值的有效数字，控制数值精度。FP8和FP32作为常见的浮点精度，FP8适用于对效率和部署成本有需求的场景，FP32则具有更高精度和更强数值稳定性。

浮点精度构成示意图

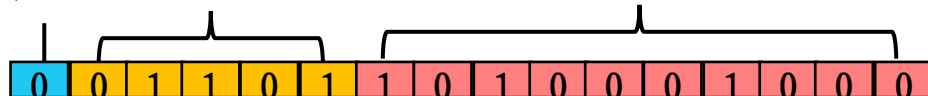
FP8:

符号位 指数位 (4 bit) 尾数位 (3 bit)



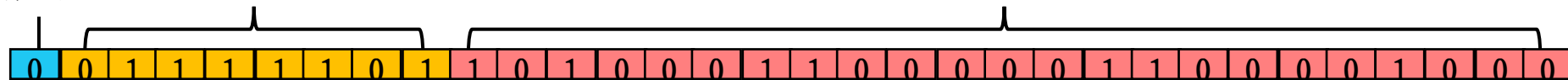
FP16:

符号位 指数位 (5 bit) 尾数位 (10 bit)



FP32:

符号位 指数位 (8 bit) 尾数位 (23 bit)



来源：至顶智库结合公开资料整理绘制

1.6 算力核心概念解析

常见算力概念汇总

名称	具体解释
TOPS	统计整数运算 每秒万亿次整数运算
POPS	每秒千万亿次运算
FLOPS	统计浮点运算 1 FLOPS 表示每秒 1 次浮点运算
GFLOPS	每秒十亿次浮点运算 $1 \text{ GFLOPS} = 10^9 \text{ FLOPS}$
TFLOPS	每秒万亿次浮点运算 $1 \text{ TFLOPS} = 10^{12} \text{ FLOPS}$
PFLOPS	每秒千万亿次浮点运算 $1 \text{ PFLOPS} = 10^{15} \text{ FLOPS}$

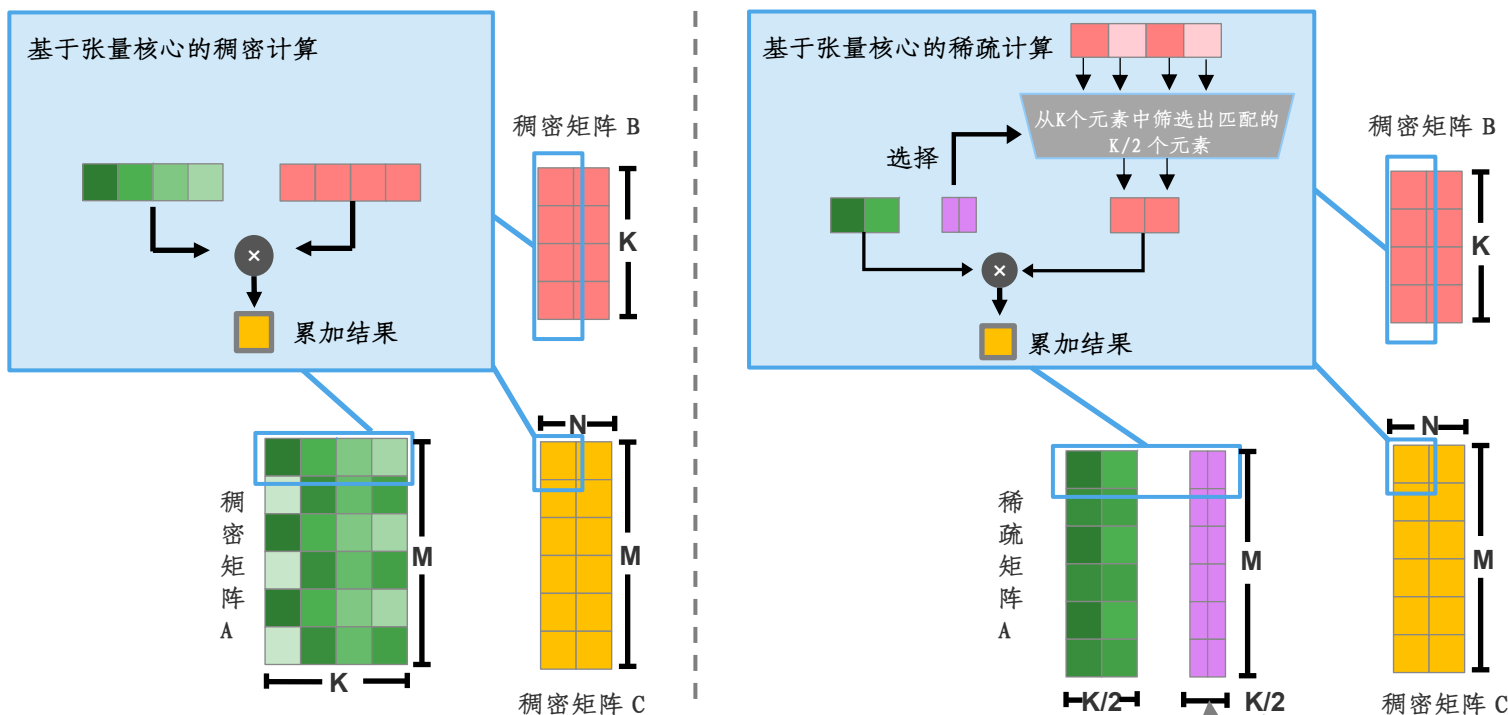
名称	具体解释
INT4	4位整数格式 压缩率很高，精度不高，追求速度 位宽 4 bit
INT8	8位整数格式，是最常见的推理格式 在精度和速度之间较均衡 位宽 8 bit
FP8	8位浮点格式，比FP16更省空间 位宽 8 bit
FP16	16位浮点格式，在精度和效率之间取得较好平衡，是广泛使用的低精度格式 位宽 16 bit
BF16	16位浮点格式，数值范围大、溢出风险更低 位宽 16 bit
FP32	32位单精度浮点格式，通用计算和深度学习中的基准精度，精度、范围和通用性都较均衡 位宽 32 bit
FP64	64位双精度浮点格式，精度最高、表示范围也更强，但存储和计算支出更大 位宽 64 bit

来源：至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—稀疏计算

稀疏计算核心特点是跳过零值运算仅处理非零有效数据，大幅降低计算量与内存消耗，提升运算效率。在常规稠密矩阵乘法中（如下图左半部分），矩阵的每一个元素均需要进行完整的两轮乘加运算，整体计算量大。而稀疏计算会先对权重矩阵做稀疏化处理，同时用专门的索引矩阵（如下图紫色部分）记录保留元素的位置信息。在实际推理运算时，只会选取和权重位置对应的输入元素参与计算，整体计算量减少一半。

稠密计算与稀疏计算示意图

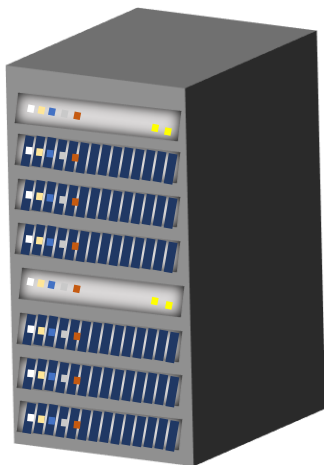


来源：至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—纵向扩展 & 横向扩展

在算力中心架构中，Scale Up与Scale Out分别从硬件升级与节点扩张两个维度，构成支撑算力系统能力的核心机制。Scale Up（纵向扩展）通过提升单节点的硬件配置（如CPU、GPU、内存等）增强单台设备的系统能力，以高效率处理复杂任务，追求极致性能；Scale Out（横向扩展）是通过增加节点来分担负载，本质上是用多台设备分担任务，其核心价值在于提供扩展空间和高可用性。

Scale Up纵向扩展示意图



Scale Out横向扩展示意图



Scale Up 与 Scale Out对比

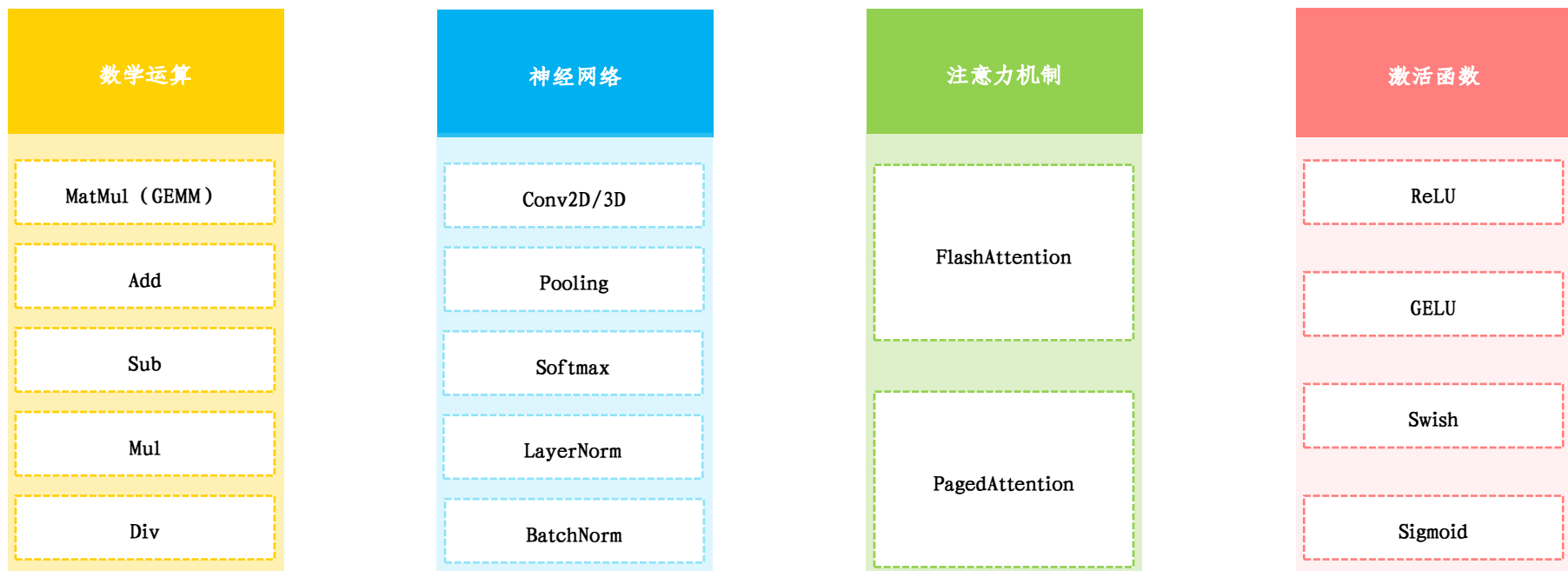
维度	Scale Up（纵向扩展）	Scale Out（横向扩展）
特点	提升单节点（Server/机柜）内部算力，通过高速互联，融合多GPU/CPU形成“单体系统”	通过增加节点数量构建分布式算力集群，实现横向扩展
互联架构	点对点直连或专有交换芯片	分层交换结构
带宽/延迟	高带宽、低延迟	低带宽、高延迟
模型	共享内存模型	分布式内存模型
适用场景	单节点大模型训练、推理加速、HPC单机计算	大规模分布式训练与推理集群

来源：至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—算子库

算子库 (Operator Library) 是人工智能与高性能计算领域的核心基础软件，将深度学习、高性能计算中最常用的底层计算单元（如矩阵乘法、卷积、激活函数）封装为高度优化的可调用函数集合，是连接上层AI框架（如PyTorch、TensorFlow）与底层硬件（GPU/NPU/CPU）的关键桥梁。算子库覆盖全场景算力需求：以通用矩阵乘法MatMul（GEMM）为核心的基础数学运算为AI计算筑牢根基；Conv2D/3D、Pooling等神经网络算子支撑经典网络构建；FlashAttention、PagedAttention等注意力机制是大语言模型高效运行的核心；ReLU、GELU等激活函数为模型引入非线性能力，全方位支撑AI全链路计算。

算子库主要构成



来源：至顶智库结合公开资料整理绘制

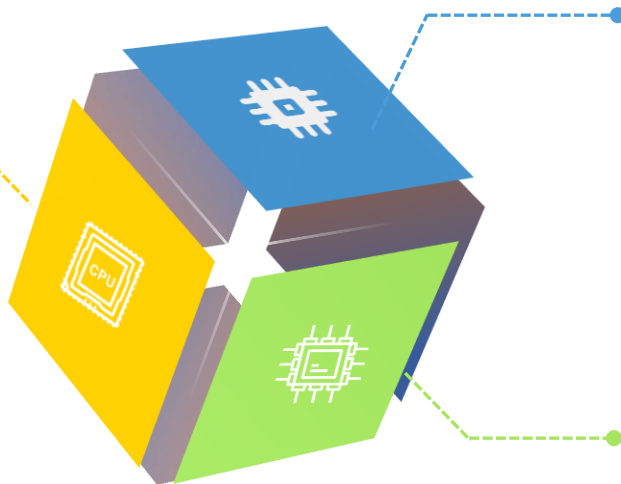
1.6 算力核心概念解析—芯片互联

芯片互联是指芯片内部或芯片之间实现信号、电源和数据传输的物理与逻辑连接技术。芯片互联带宽则是在芯片之间或芯片内部不同模块之间，数据传输的速率或容量。该指标反映芯片间通信通道在单位时间内能够传输的数据量大小，是衡量芯片互联性能的关键指标之一。根据层级分类，互联方式主要分为三个层面：片内互联（芯片内部，微米至毫米级）、片间互联（同一服务器机箱内，厘米级）、节点间互联（跨机柜或数据中心，米级至千米级）。

芯片互联方式主要类型

片内互联

- 含义：单芯片内部多个计算核、存储核之间互联
- UCIe 3.0：64GB/s，x16单向约1.28TB/s
- 3D混合键合(Hybrid Bonding)：间距<1 μm,带宽密度>300 TB/s/mm²
- NVLink-C2C(NVIDIA)：Blackwell架构1.8TB/s/芯片
- Infinity Fabric(AMD)：MI300系列 ~ 1.5TB/s



片间互联

- 含义：同一块服务器主板上，GPU 与 GPU、GPU 与 CPU 互联
- NVLink 4/5：H100 900GB/s；Rubin 3.6TB/s
- PCIe 5.0 x16：128GB/s
- PCIe 6.0 x16：256GB/s
- PCIe 7.0 x16：512GB/s

节点间互联

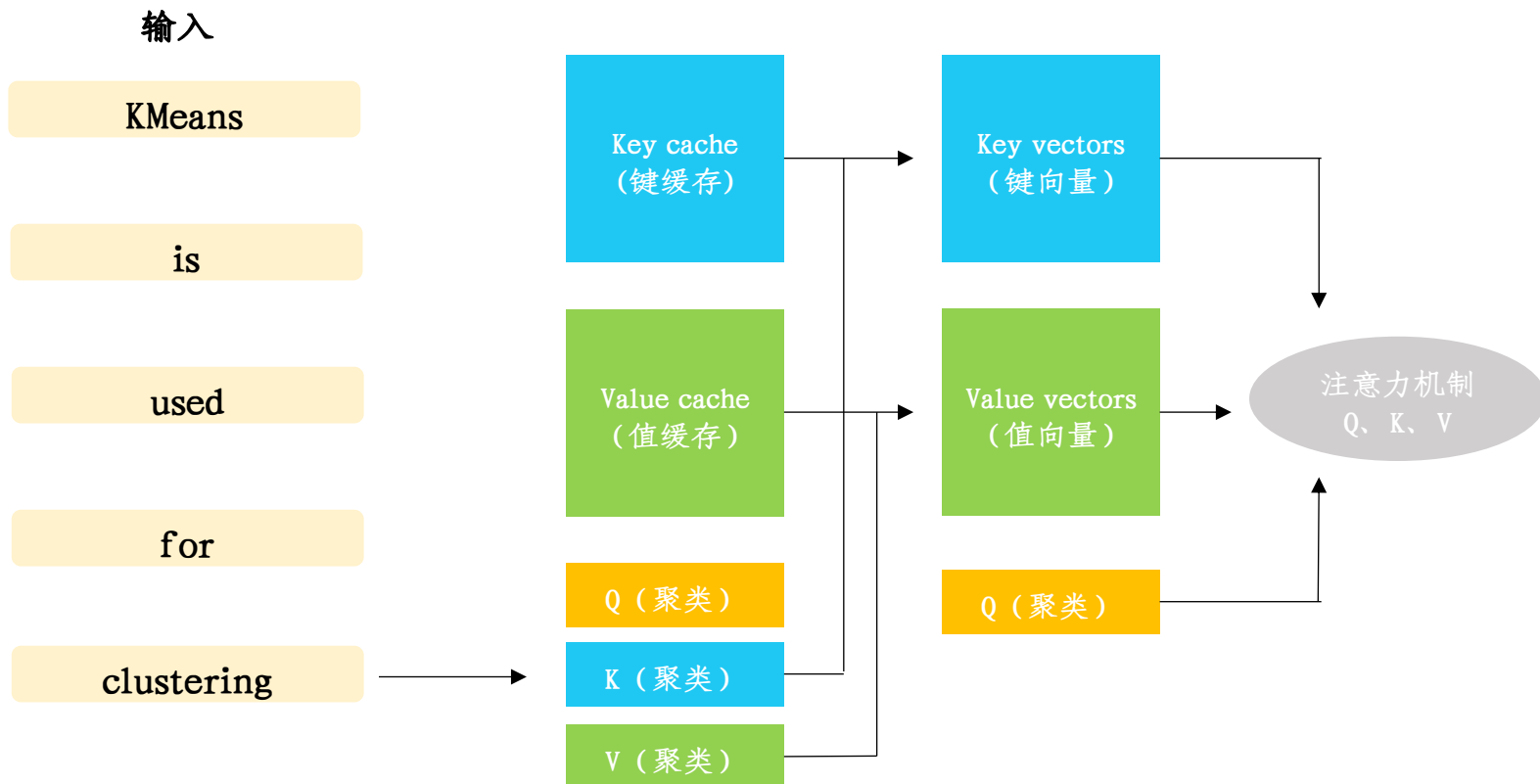
- 含义：多台服务器组成算力集群互联
- InfiniBand NDR：400Gbps (~ 50GB/s)
- InfiniBand XDR：800Gbps (~ 100GB/s)
- UALink 1.0：200GB/s

来源：至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—KV Cache键值缓存

KV Cache（全称Key-Value Cache，键值缓存）是大模型推理优化中的关键技术。该技术通过在模型推理的预填充阶段计算并存储所有输入Token的K和V向量，后续在生成新token时，只需计算Q向量，从缓存中读取历史K和V向量，即可完成注意力计算。该技术避免重复计算，从而使计算复杂度下降，提升模型的推理效率。

KV Cache原理示意图

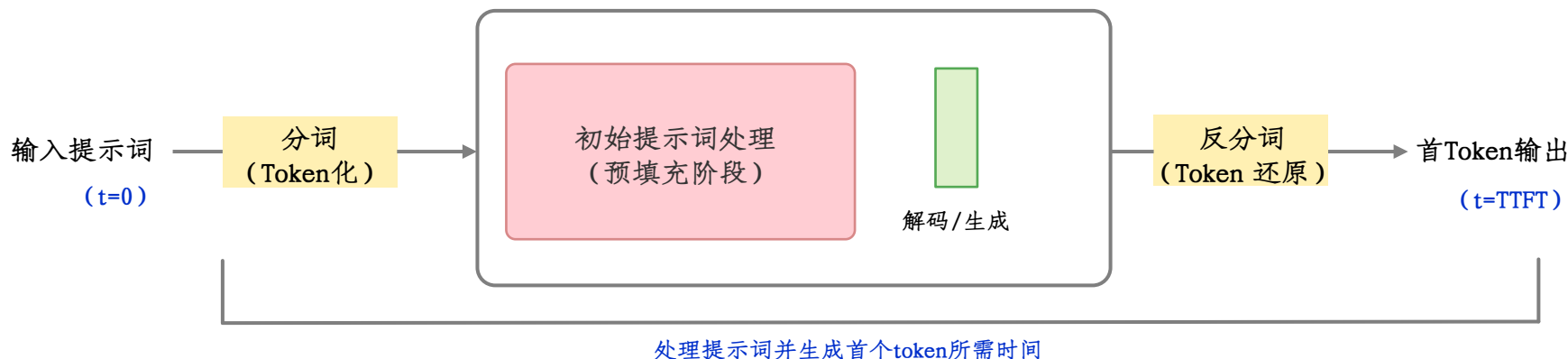


来源：至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—首Token生成时间 & Token吞吐量

首Token生成时间(Time To First Token)是衡量大模型回复响应速度与用户体验的核心指标。TTFT具体是指从用户发送提示词 ($t=0$) 到模型返回第一个输出Token ($t=TTFT$) 的时间间隔。完整流程如下：用户输入提示词后，系统将文本转换成模型能处理的Token形式。随后Token被送入GPU进行计算。GPU执行初始提示词处理（对用户输入的提示词进行编码、上下文理解和注意力计算），该阶段通常是TTFT中非常关键的部分。随后进入解码/生成阶段，开始逐步生成输出内容。基于预填充阶段的结果，生成模型的第一个输出Token。后续将模型生成的Token还原成可读的文本形式，最终输出给用户文本片段。**Token吞吐量表示单位时间内模型输出的Token数量，单位为Token/s，是评估大模型推理性能的核心指标。**吞吐量越高，意味着基础架构的回报越高。

首Token生成时间 (TTFT) 示意图

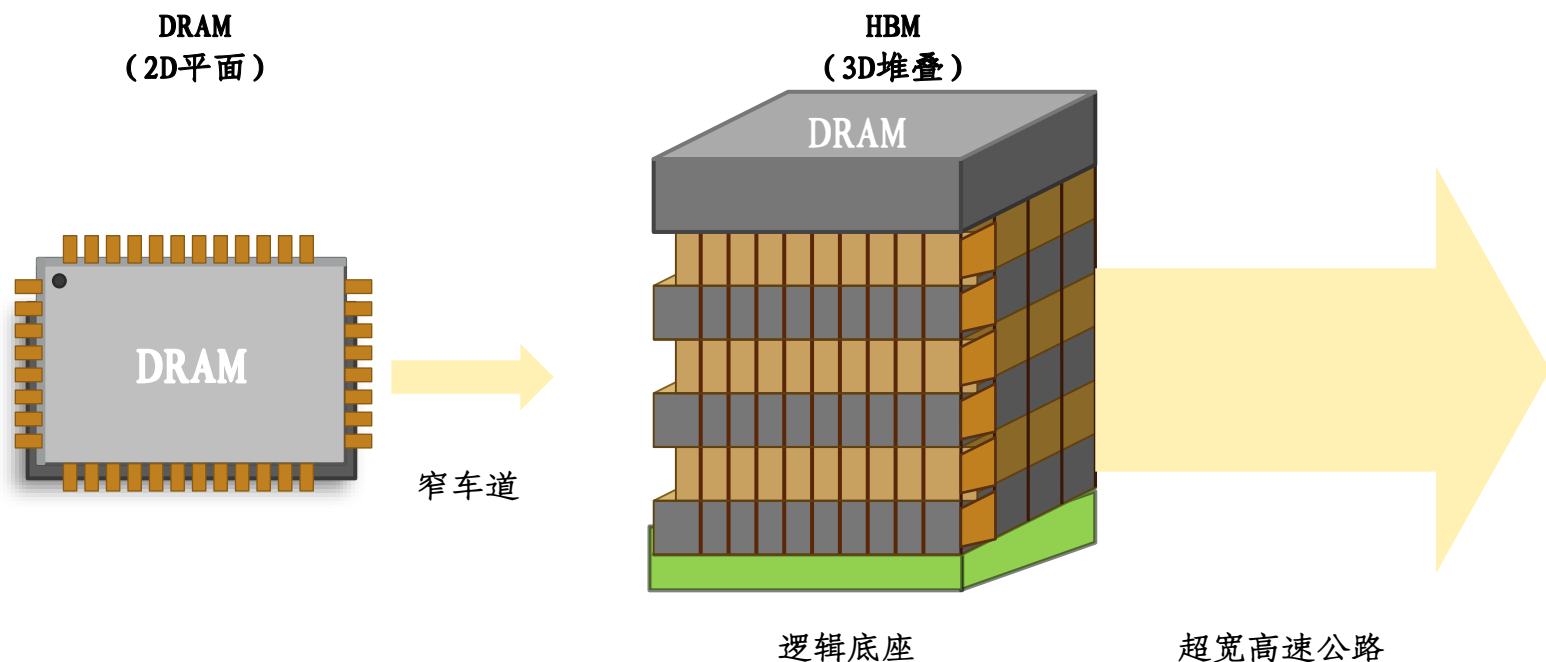


来源：NVIDIA，至顶智库结合公开资料整理绘制

1.6 算力核心概念解析—HBM高带宽内存

HBM (High Bandwidth Memory) 即高带宽内存，是采用3D堆叠封装的高速存储技术。通过将多层DRAM存储芯片垂直堆叠，相比传统内存拥有高带宽、低功耗和高集成度等特点，主要用于AI模型训练、AI推理、高性能计算等对数据吞吐量要求高的场景。

HBM 3D堆叠示意图



来源：至顶智库结合公开资料整理绘制
insights.zhiding.cn₂₅

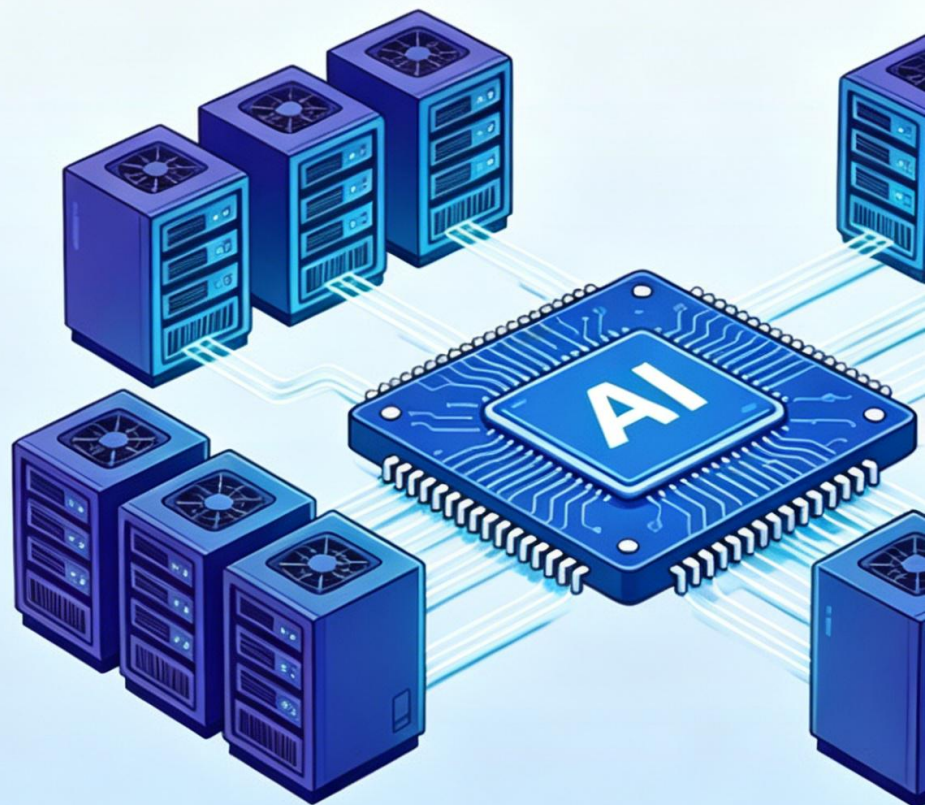
第二章 全球AI芯片发展情况

AI芯片分类

AI芯片技术

典型企业产品

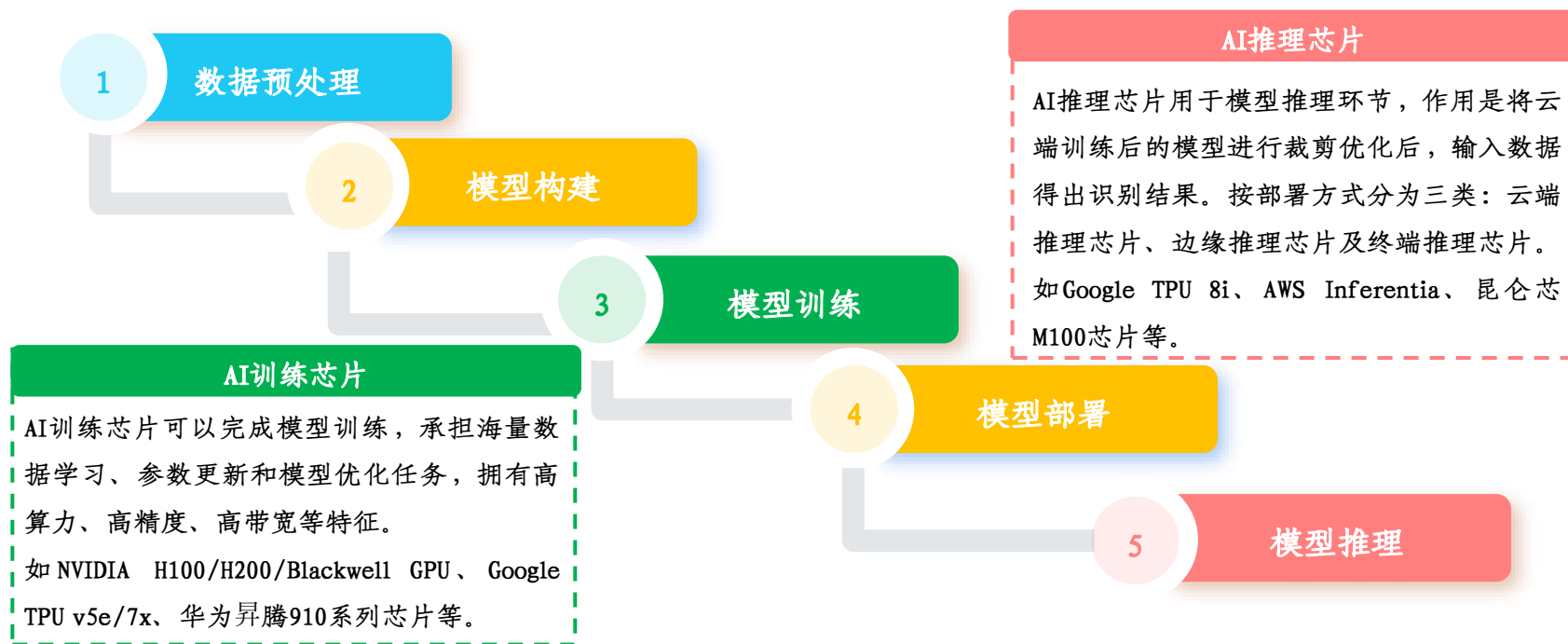
- NVIDIA
- 华为
- 摩尔线程
- 沐曦集成电路
- 飞腾



2.1 AI芯片主要分类

本报告涉及的AI芯片，根据适用场景进行划分，主要分为AI训练芯片和AI推理芯片。AI训练芯片主要面向模型训练阶段，承担海量数据处理、参数更新和模型优化任务。此类芯片拥有高算力、高精度、高带宽等特征，其目标是让模型从文本、图像等多模态数据中学习规律。AI推理芯片用于模型部署阶段，负责将模型应用于各类场景。此类芯片拥有低延迟、高能效比及适配云边端全场景的部署灵活性。

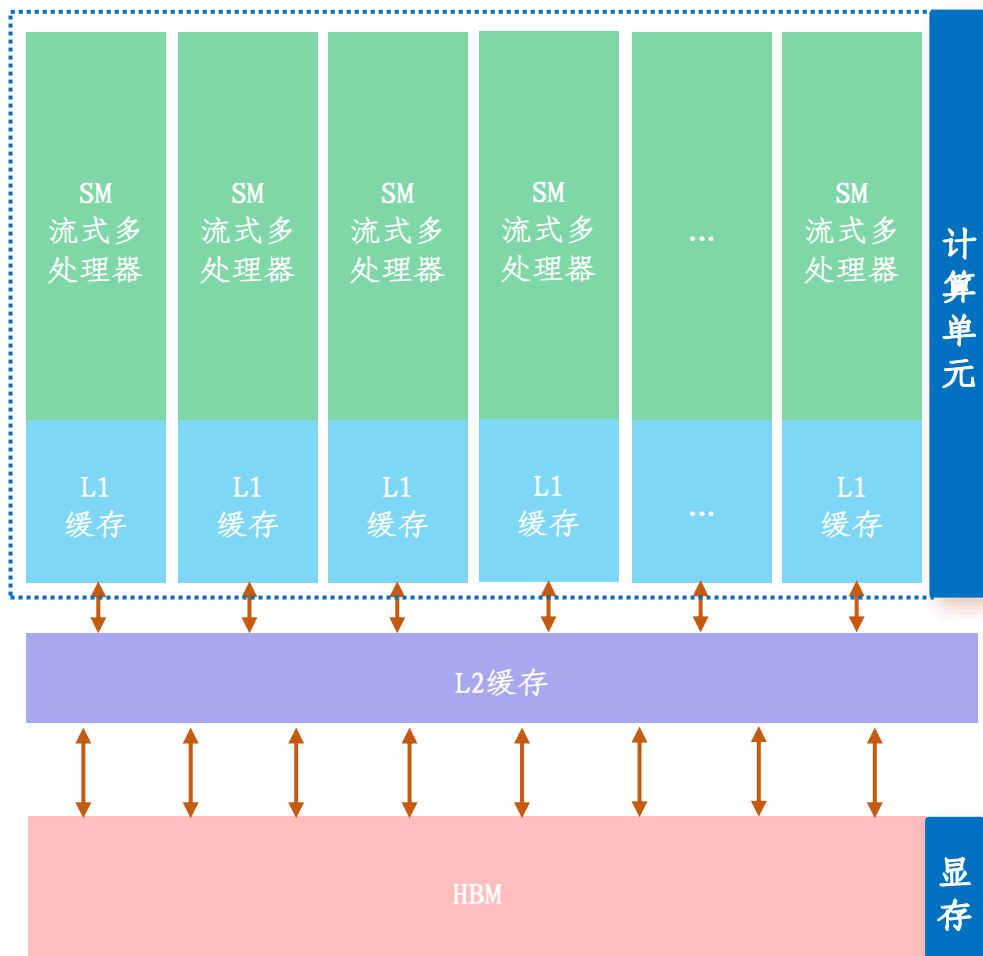
模型构建流程与AI芯片分类



来源：至顶智库结合公开资料整理绘制

2.2 GPU芯片架构

GPU芯片架构图



GPU拥有**流式多处理器（SM）**为核心的并行计算架构，能够高效执行大模型训练与推理所需的矩阵运算与单指令多线程并行计算任务，是当前人工智能算力基础设施的核心硬件载体。

每个GPU芯片集成多个流式多处理器，并配备**L1缓存**，用于暂存高频访问数据以减少访存开销；所有SM共享统一的**L2缓存**，进一步降低对GPU显存的访问延迟。GPU显存用于存储模型参数、中间计算结果等数据，通过高带宽读写保障并行计算持续高效，为大模型所需的密集计算提供关键支撑。

整体来看，GPU通过多层次缓存架构与高带宽存储设计，有效缓解AI计算中的内存墙瓶颈，保障大规模并行运算的持续高效执行。

来源：至顶智库结合公开资料整理绘制

2.3 GPU显存带宽对于模型训练及推理的影响

GPU显存带宽是指GPU计算核心与显存之间的数据传输速率。其决定大模型训练和推理过程中的模型参数、中间计算结果等数据在显存与算力单元间的搬运速度。模型训练是让AI从大量数据中反复学习、不断调整内部参数，逐步形成规律与判断能力的过程；模型推理是用已训练的模型接收新数据，进而快速计算并给出反馈结果的应用过程。二者的核心区别在于，训练重在“学习知识”，处理海量数据并持续更新参数，而推理重在“实际使用”，对于新数据或未知信息进行预测。

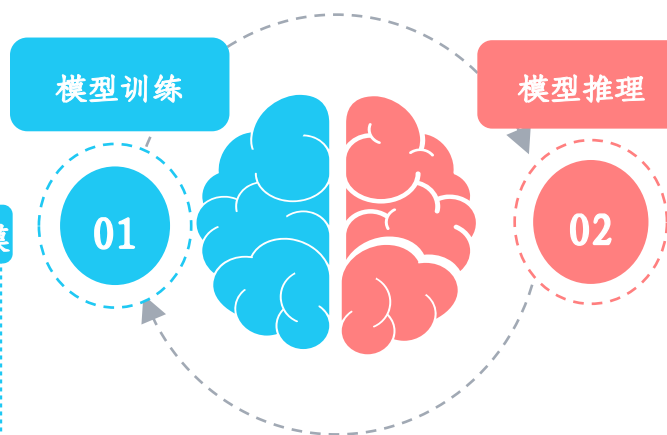
GPU显存带宽对于模型训练及推理的影响

影响训练迭代效率与算力有效释放

显存带宽决定了模型训练过程中模型权重、批次输入数据、激活值与梯度张量等数据的传输速率，其性能水平直接决定数据能否及时供给计算核心SM。带宽与芯片算力的匹配程度，会直接影响Tensor Core等计算单元的运行效率，进而决定整体训练迭代速度。

影响训练稳定性与单次迭代数据量规模

更大单次迭代数据量可使模型性能的优化更稳定高效；而更高的显存带宽能够支撑更大单次迭代数据量下的海量数据读写与同步，避免因带宽瓶颈限制单次迭代数据量，进而提升训练稳定性与参数更新效率。



决定单请求推理时延

GPU显存带宽水平直接影响数据供给速度，进而作用于推理延迟表现。在自动驾驶、实时语音、视频流分析等对延迟敏感的场景中，带宽性能与系统响应效率、实际可用性及用户体验直接相关。

支持高并发推理吞吐

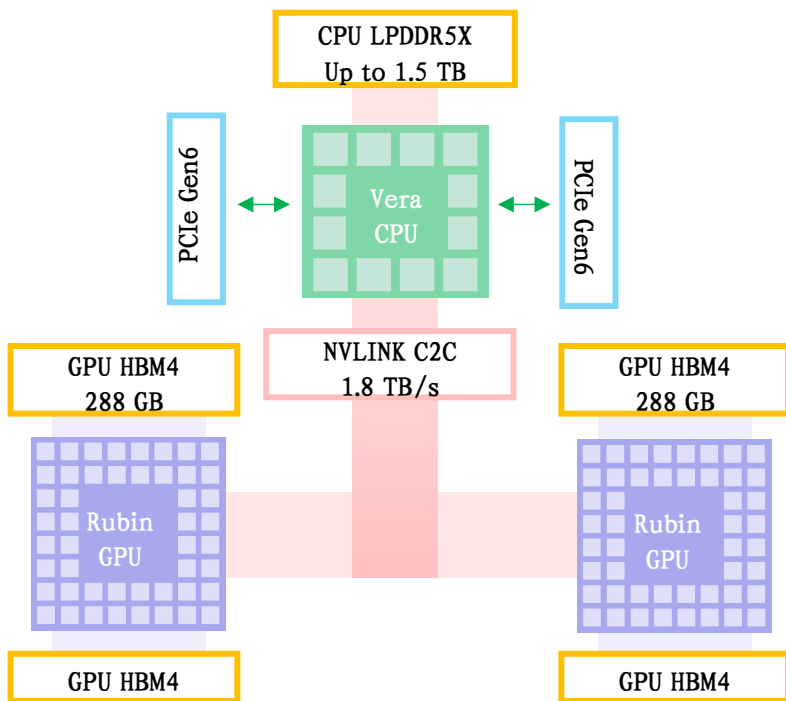
高并发推理需同时调度多组计算任务与数据读写，充足的显存带宽可保证多任务并行时数据供给不中断，从而提升系统每秒可处理的提问数量；带宽不足则会造成请求排队、服务拥堵，显著降低可承载的并发量。

来源：至顶智库结合公开资料整理绘制

2.4 CPU+GPU异构计算架构

CPU+GPU异构计算架构提升数据传输效率与资源利用率，为AI模型训练及推理提供高效算力支撑，大幅提升任务处理能力。在异构计算架构中，GPU专注大规模并行张量计算。CPU具有四大功能：作为“总指挥”负责训练推理任务拆分与多GPU协同，提升算力利用率；作为“数据供给引擎”完成数据预处理与分发，消除传输瓶颈；作为“串行任务卸载器”处理简单控制流，避免GPU计算资源浪费；作为“桥梁枢纽”连接外设并构建无瓶颈互联体系，保障系统高效运行。2010年，中国“天河一号A”超级计算机率先将“CPU+GPU”异构架构实现规模化落地，引领全球在AI训练领域的智算底层架构发展方向。2026年，伴随NVIDIA Groq 3 LPU面向模型推理的专用芯片发布，将形成以GPU+LPU+CPU+DPU为特征的新型异构推理架构。

NVIDIA Vera Rubin “CPU+GPU” 异构架构图



Vera CPU是NVIDIA Vera Rubin架构节点的中心，通过1.5TB的内存提供大容量主存，并通过NVLink C2C（1.8TB/s带宽）与两个Rubin GPU高速直连，实现CPU-GPU内存一致性，降低数据传输成本。

每个Rubin GPU配备288GB HBM4高带宽显存用于加速AI与高性能计算，PCIe Gen6用于连接外部设备和扩展组件，形成高带宽、强计算能力的AI计算节点。

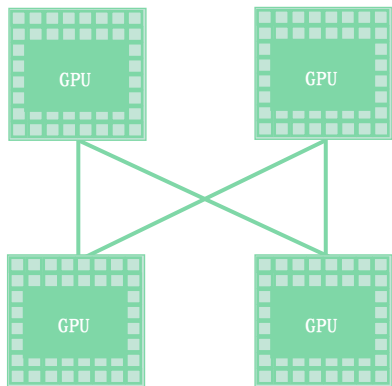
来源：NVIDIA，天津出版传媒集团《大国算力》，至顶智库整理绘制

2.5 芯片互联技术—NVLink

NVLink是NVIDIA推出的高速互联技术，主要用于GPU与GPU、GPU与CPU之间的高速数据通信。通过硬件级显存一致性与直连架构，NVLink消除传统PCIe总线引发的CPU中转瓶颈与通信拥塞，使GPU间的数据交换效率在Rubin架构下跃升至3.6TB/s。在此基础上，NVLink Switch System进一步实现跨节点的无限扩展，通过高达260TB/s的聚合带宽将整个数据中心集群化为一颗“数据中心GPU”，为万亿级参数大模型的并行计算提供极致算力支撑。

NVIDIA NVLink、NVLink Switch架构图

NVLink



点对点高速协议与物理互联：

- 相较于PCIe Gen6，第六代NVLink提供了超过14倍的有效带宽。在Blackwell架构中，单颗GPU的双向带宽已达到1.8TB/s；而在最新的Rubin架构下，这一数字翻倍至3.6TB/s。

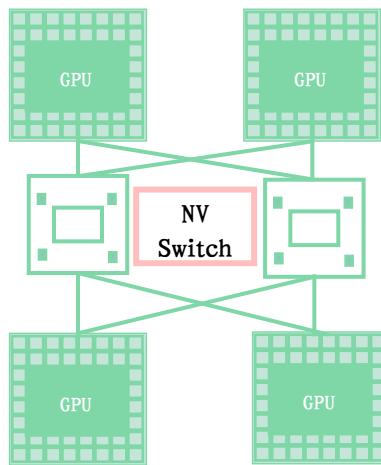
消除CPU中转与I/O瓶颈：

- 通过绕过传统PCIe总线和CPU的数据调度，NVLink实现GPU间的直接内存交换，解决分布式训练中的通信拥塞。

内存一致性：

- NVLink支持硬件级的显存一致性，使GPU能够共享统一的寻址空间。

NVLink Switch



跨节点的“数据中心GPU”架构：

- 借助NVLink Switch的中转与调度，NVLink连接可以跨节点扩展，创建无缝、高带宽、多节点的GPU集群，从而有效地形成数据中心大小的GPU。

内存级一致性的极致互联：

- 该系统具备全带宽互联与超高扩展性，在处理大规模并行计算时，各节点间的GPU可以像访问本地显存一样共享数据，极大消除分布式计算中的通信壁垒。

重塑大模型算力峰值：

- NVLink Switch可在一个 NVIDIA Vera Rubin NVL72系统中实现 260 TB/s的GPU聚合带宽，这种量级的吞吐能力专为加速万亿级参数的大型模型并行计算而生，为AI训练与推理提供了算力支撑。

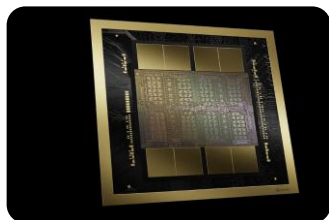
来源：NVIDIA，至顶智库整理绘制

2.6 国外AI芯片：海外科技巨头引领高性能计算芯片发展

AI芯片已成为驱动人工智能发展的核心引擎。在大模型训练和推理中，芯片算力、内存带宽和互联技术直接决定模型迭代更新。当前，国际主流公司正围绕高性能计算、低精度格式和系统级优化展开激烈竞争，推动AI芯片向更高效率、更低成本演进。NVIDIA凭借其Blackwell与Rubin架构持续领跑，保持其在高端训练和推理市场的领导地位；Google依托自研TPU深化软硬件垂直整合，强化其云计算和AI服务的底层能力；AWS通过自研Trainium训练芯片与Inferentia推理芯片的协同部署，提供高性价比的云端算力解决方案。

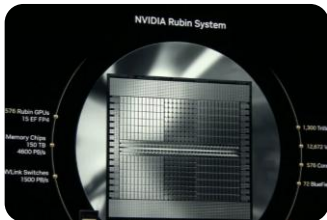
国外主流AI芯片

NVIDIA Blackwell



- 2024年NVIDIA推出Blackwell架构；
- 适用于大规模推理AI场景，能效比上一代Hopper GPU高出30倍，并支持实时性能下的高吞吐量；
- 基于Blackwell Ultra的GB300系统推理效能相比Hopper系统实现50倍AI工厂生产力提升，保持低延迟，实现收益最大化。

NVIDIA Rubin



- 2025年，NVIDIA公布下一代超级芯片Rubin GPU；
- 整体性能是GB300的3.3倍，集288GB HBM4显存、22TB/s带宽、260TB/s吞吐量于一身，实现900倍于Hopper的性能；
- 同期公布的Rubin Ultra计算面积翻倍，密集FP4浮点运算性能提升至100 PFLOPs。HBM容量达到1024GB。

Google TPU 8t & 8i



- 2026年，Google公布第八代TPU 8t和TPU 8i；
- TPU 8t专为大规模、高强度模型训练设计，单集群可扩展至9,600芯片并提供121ExaFlops算力，较前代提升近3倍性能与2倍能效比；
- TPU 8i专为低延迟、高吞吐推理优化，通过3倍于前代的片上SRAM，实现80%的性价比提升。

aws Inferentia 2



- 2022年，AWS发布第二代推理芯片Inferentia 2；
- 架构与Trainium 1相似，但NeuronLink-v2互连端口更少。每瓦性能最多可提升50%，将成本有效降低40%。与一代相比，吞吐量提高4倍，延迟大幅降低；
- 基于Inferentia 2的Amazon EC2 Inf2可以大规模部署复杂模型。

aws Trainium 3



- 2025年，AWS推出第三代训练芯片Trainium3；
- 较Trainium2内存容量提升至1.5倍、带宽提升1.7倍以上，优化实时多模态及推理任务的内存与计算配比；
- Trn3 UltraServer性能比Trn2 UltraServer提升4.4倍、内存带宽提升3.9倍、能效提升超4倍，具备前沿模型训练与部署的最优性价比。

来源：至顶智库结合公开资料整理绘制

2.6 国外AI芯片—NVIDIA芯片架构演进

NVIDIA发布的AI高性能芯片以及计算设备，作为推动全球人工智能发展的关键基础设施，持续发挥重要作用。2022年以来，NVIDIA相继推出基于Hopper架构和Blackwell架构的高性能计算产品线，涵盖H100 Tensor Core GPU、Blackwell Ultra GPU等。2026年，NVIDIA正式推出新一代Vera Rubin AI计算平台，并发布其核心芯片Rubin。Rubin GPU显著提升训练、推理及长上下文处理效率，相比Blackwell GPU实现代际算力突破，全面优化将功耗、带宽和显存高效转化为Tokens的全链路流程。

NVIDIA NVIDIA芯片架构演进

Hopper架构（2022）



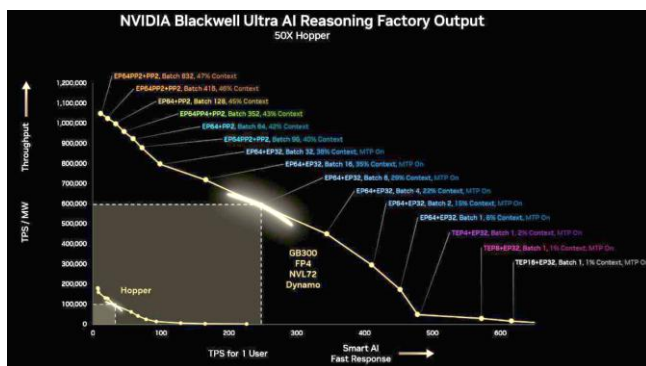
NVIDIA H100 Tensor Core GPU

由NVIDIA Hopper架构驱动的H100 Tensor Core GPU具有800亿晶体管，适用范围广泛，涵盖从小型企业到百亿亿级高性能计算，再到万亿参数的人工智能模型。

第四代张量核心与A100相比，芯片间速度最高可达6倍，包括每个流式多处理器（SM）的加速、额外SM数量以及更高的时钟频率。

全新Transformer引擎结合软件和专门设计的Hopper张量核心技术，专门用于加速Transformer模型的训练和推理。

Blackwell架构（2024）

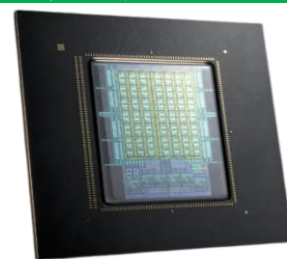
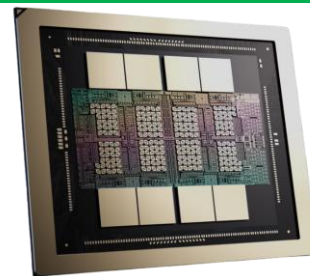


NVIDIA Blackwell

NVIDIA Blackwell架构适用于大规模推理AI场景，能效比上一代Hopper GPU高出30倍，并支持实时性能下的高吞吐量。

通过优化Blackwell Ultra GPU并行化策略（专家/张量/流水线）和跨GPU管理，基于Blackwell Ultra的GB300系统推理效能相比Hopper系统实现50倍AI工厂产量或生产力提升，同时保持低延迟，实现收益最大化。

Rubin GPU & Vera CPU（2026）



Rubin GPU

Rubin GPU针对推理、混合专家模型、长上下文推理及强化学习等现代AI负载面临的计算、内存与通信协同效率瓶颈进行专项优化，可提供最高50PFLOPS的NVFP4推理算力（较Blackwell提升5倍）与35PFLOPS的NVFP4训练算力（提升3.5倍），搭配288GB显存、22TB/s的HBM4带宽与3.6TB/s的NVLink互联，全面优化将功耗、带宽和显存高效转化为Tokens的全链路流程。

Vera CPU

Vera是一款专为GPU设计的CPU，内存带宽1.2TB/s，较上一代Grace CPU提升2.4倍，1.5TB内存容量较上代提升3倍，以支持数据密集型工作负载，同时NVLink-C2C带宽增长两倍至1.8TB/s，以维持机架级的CPU-GPU连贯运行。通过高效移动、处理和协调数据，共同提升CPU从辅助角色，到实现AI工厂规模的高效利用。

来源：NVIDIA，NVIDIA H100 GPU技术文档，NVIDIA Blackwell架构技术文档，至顶智库整理绘制

2.6 国外AI芯片—NVIDIA主要AI芯片类型

NVIDIA面向AI领域的主流芯片有H100、H200、Blackwell、Blackwell Ultra GPU和Rubin GPU。其中，H100搭载3958 TFLOPS（FP8精度）算力、80GB显存与900GB/s NVLink互联带宽；H200将显存扩容至141GB，带宽增加为4.8TB/s。Blackwell GPU将FP8算力提升至10 PFLOPS，显存扩容至186GB；Blackwell Ultra进一步将显存提升至279GB，升级PCIe Gen6至256GB/s；最新推出的Rubin GPU 算力17.5 PFLOPS（FP8精度），GPU显存带宽也增至22TB/s，提升大规模训练的有效性能上限，同时在后训练和推理工作流中带来显著增益。

NVIDIA主要AI芯片参数

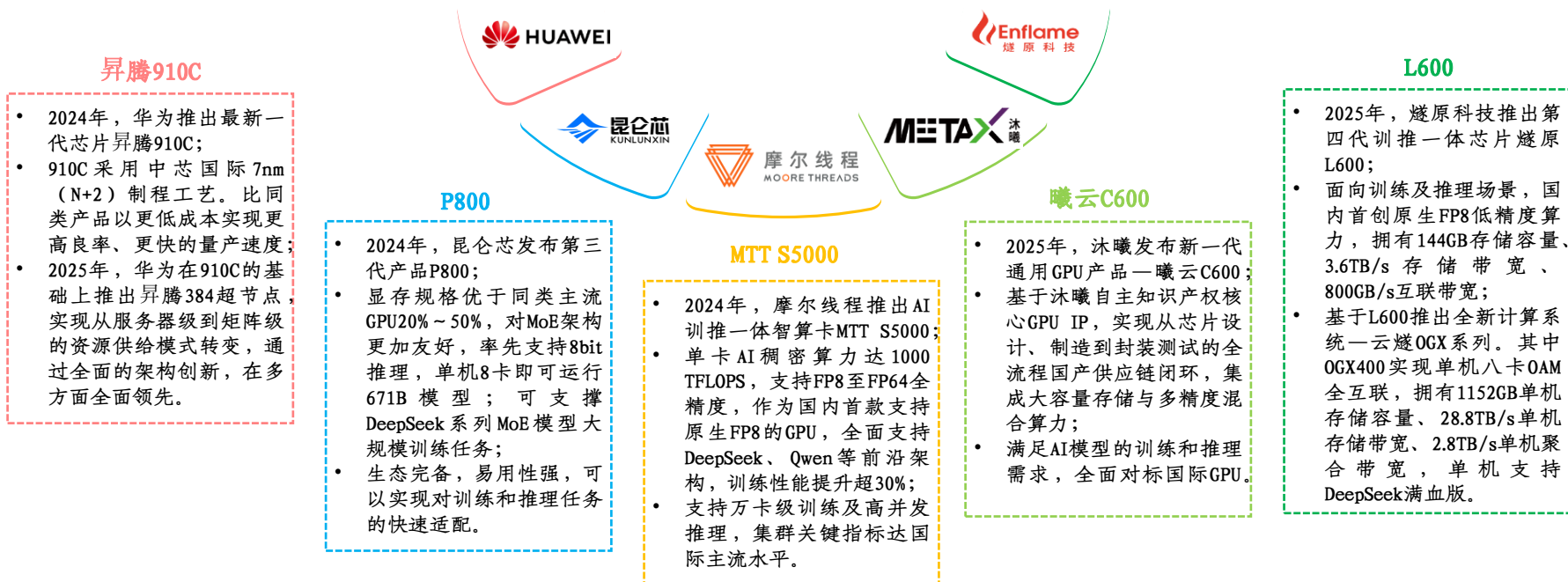
型号	H100	H200	Blackwell GPU	Blackwell Ultra GPU	Rubin GPU
发布时间	2022年	2023年	2024年	2025年	2026年
架构	Hopper	Hopper	Blackwell	Blackwell	Rubin
INT8	3958 TOPS	3958 TOPS	10000 TOPS	330 TOPS	250 TOPS
FP8	3958 TFLOPS	3958 TFLOPS	10 PFLOPS	10 PFLOPS	17.5 PFLOPS
FP32	67 TFLOPS	67 TFLOPS	80 TFLOPS	80 TFLOPS	130 TFLOPS
GPU显存 带宽	80GB 3.35TB/s	141GB 4.8TB/s	186GB 8TB/s	279GB 8TB/s	288GB 22TB/s
TDP热功耗	700W	700W	1200W	1400W	-
互联带宽	NVIDIA NVLink: 900GB/s PCIe Gen5: 128GB/s	NVIDIA NVLink: 900GB/s PCIe Gen5: 128GB/s	NVIDIA NVLink: 1.8TB/s PCIe Gen5: 128GB/s	NVIDIA NVLink: 1.8TB/s PCIe Gen6: 256GB/s	NVIDIA NVLink: 3.6TB/s PCIe Gen6: 256GB/s

来源：NVIDIA，至顶智库结合公开资料整理绘制

2.7 国内AI芯片：自主架构持续创新，训练推理两线并进

当前，国内AI芯片行业正依托“自主可控”战略快速崛起，形成以华为昇腾910C、昆仑芯P800、摩尔线程MTT S5000、沐曦曦云C600等为代表的AI计算产品矩阵（AI计算产品包含AI芯片、AI计算卡等），在模型训练和推理场景中实现规模化落地。与国外追求芯片绝对算力峰值不同，国内更注重通过构建集群突破单点算力限制，并通过软硬件垂直整合和性价比优势抢占市场。2026年5月，华为发布“**韬（ τ ）定律**”，“**韬定律**”提出以“**时间缩微**”替代“**几何缩微**”，以系统性降低时间常数（韬 τ ）为目标，通过逻辑折叠等创新技术，持续压缩信号传播时延，不断提升晶体管密度，实现半导体与电子系统的持续演进。

国内主流AI计算产品



来源：至顶智库结合公开资料整理绘制

2.7 国内AI芯片—华为主要AI芯片

华为昇腾芯片遵循“一年一代、算力翻倍”迭代逻辑，形成覆盖“训练、推理”全场景、“通用、垂直”全需求的算力产品体系。芯片基于先进工艺打造，昇腾910C/950PR/950DT/960采用N+2工艺，昇腾970升级至N+3工艺，算力密度与能效比持续提升；内存规格与互联带宽同步升级，为大模型训练和推理场景提供充足的算力支撑。

华为主要AI芯片参数

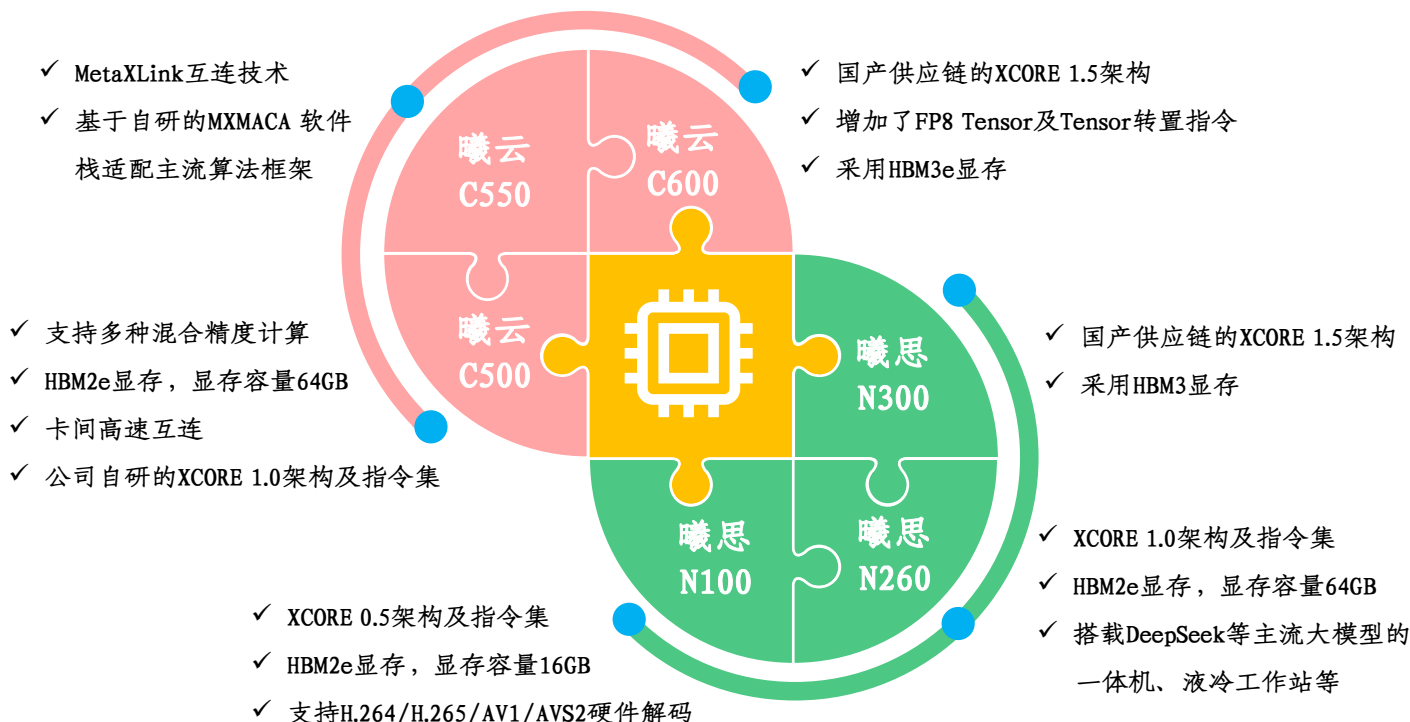
型号	昇腾 910C	昇腾 950PR	昇腾 950DT	昇腾 960	昇腾 970
发布时间	2025 年 Q1	2026 年 Q1	2026 年 Q4	2027 年 Q4	2028 年 Q4
核心工艺	N+2 工艺	N+2 工艺	N+2 工艺	N+2 工艺	N+3 工艺
数值类型	FP32/HF32/FP16/ BF16/INT8	FP32/HF32/FP16/BF16/ FP8/MXFP8/HiF8/MXFP4	FP32/HF32/FP16/BF16/ FP8/MXFP8/HiF8/MXFP4/ HiF4	FP32/HF32/FP16/BF16/ FP8/MXFP8/HiF8/MXFP4/ HiF4	FP32/HF32/FP16/BF16/ FP8/MXFP8/HiF8/MXFP4/ HiF4
核心算力	FP16: 800 TFLOPS	FP8: 1 PFLOPS FP4: 2 PFLOPS	FP8: 1 PFLOPS FP4: 2 PFLOPS	FP8: 2 PFLOPS FP4: 4 PFLOPS	FP8: 4 PFLOPS FP4: 8 PFLOPS
显存规格	HBM: 128GB, 带宽 3.2TB/s	HiBL 1.0: 128GB, 带宽 1.6TB/s	HiZQ 2.0: 144GB, 带宽 4TB/s	HBM: 288GB, 带宽 9.6TB/s	HBM: 288GB, 带宽 14.4TB/s
互联带宽	784GB/s	2TB/s	2TB/s	2.2TB/s	4TB/s

来源：至顶智库结合公开资料整理绘制

2.7 国内AI芯片—沐曦主要AI芯片

沐曦集成电路提供全栈GPU芯片与计算平台，覆盖训推一体GPU曦云、智算推理GPU曦思两大核心系列。曦云系列产品拥有多精度混合算力，内置大量运算核心，具有较强的并行计算能力和较高的能效比，适用于向量计算和矩阵计算等计算密集型应用，可广泛应用于智算训练与推理、通用计算、AI for Science等场景。曦思系列产品面向推理场景，内置高性能视频处理器与算力核心，可广泛应用于智慧城市、生成式人工智能等领域。

沐曦典型AI芯片产品

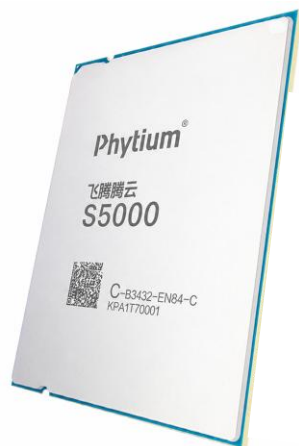


来源：沐曦集成电路，至顶智库结合公开资料整理绘制

2.7 国内AI芯片—飞腾主要AI芯片

飞腾信息技术有限公司致力于通算处理器、智算处理器等高端芯片的研发设计和产业化推广，总部位于天津。飞腾芯片产品具有谱系全、自主性强、性能优、安全性高等特点，包括高性能服务器CPU（飞腾腾云S系列）、高效能桌面CPU（飞腾腾锐D系列）、高端嵌入式CPU（飞腾腾珑E系列）和飞腾XPU四大系列，产品性能达到国内领先、国际一流水平，能够为从端到云的各型设备提供核心算力支撑。飞腾腾云S5000C作为飞腾新一代高性能服务器CPU，相比上一代产品，该产品计算性能大幅提升，支持硬件虚拟化，该产品主要面向AI服务器、大型互联网算力中心等场景。

飞腾腾云高性能服务器CPU S5000C

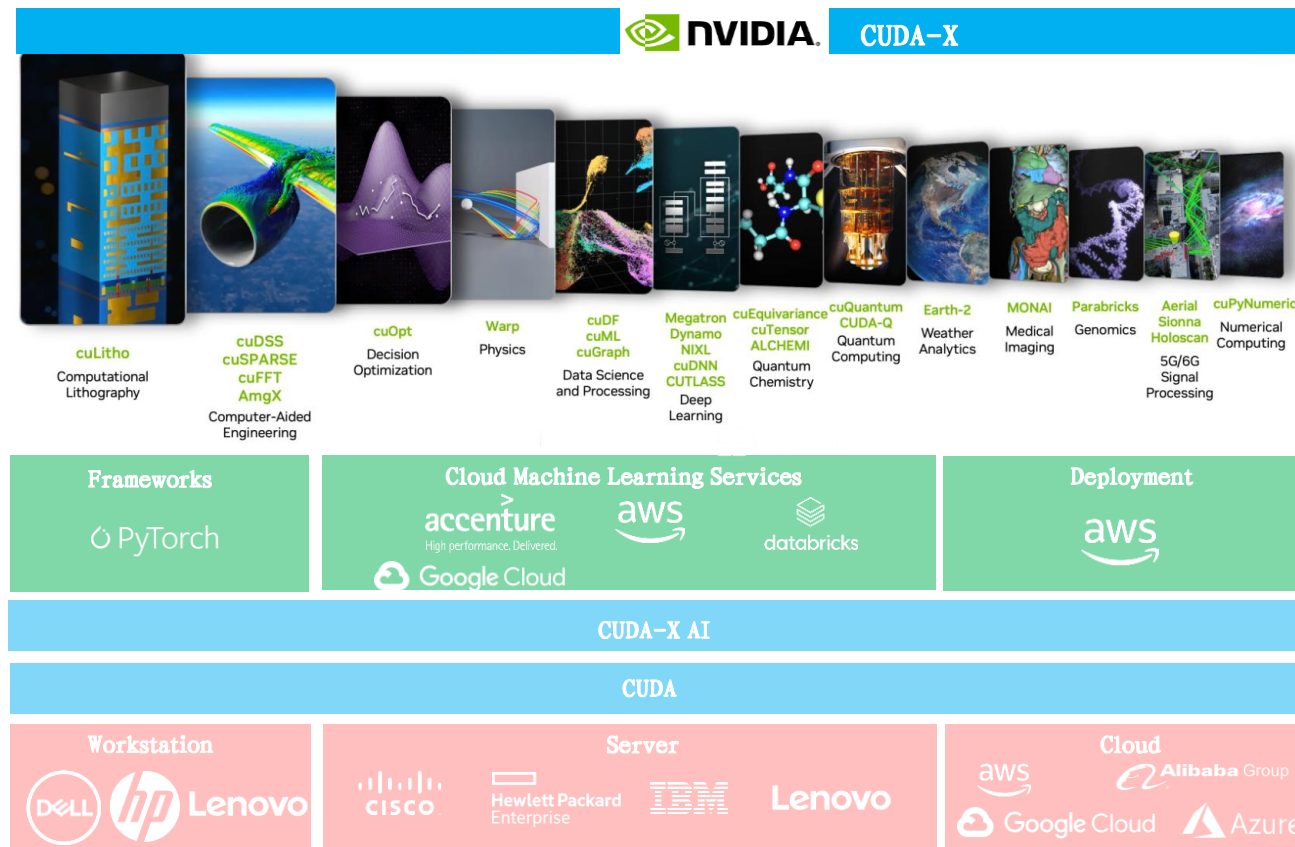


类别	参数		
型号	飞腾腾云 S5000C-64	飞腾腾云 S5000C-32	飞腾腾云 S5000C-16
核心	飞腾自主高性能处理器核 FTC862		
核数	64 核	32 核	16 核
主频	2.1 GHz	2.3 GHz	2.3 GHz
二级缓存	32 MB	16 MB	8 MB
三级缓存	32 MB	16 MB	8 MB
存储控制器	8个DDR5接口	4个DDR5接口	2个DDR5接口
PCIe接口	96 lanes PCIe 5.0	80 lanes PCIe 5.0	48 lanes PCIe 5.0

来源：飞腾，至顶智库整理绘制

2.8 芯片开发平台—NVIDIA CUDA

CUDA (Compute Unified Device Architecture) 作为NVIDIA于2006年推出的专有并行计算平台与编程接口 (API)，允许开发者利用NVIDIA GPU执行科学计算与高性能计算，目前CUDA支持超过900个库。CUDA-X建立在CUDA之上，是一套由NVIDIA提供的GPU加速微服务、工具及库的集合，专门用于加速数据处理、人工智能与高性能计算 (HPC) 场景应用。CUDA-X涵盖数学运算库、并行算法库、图像视频库、通信库、深度学习库等，拥有超过400个加速组件，通过GPU带来计算性能提升。



NVIDIA CUDA-X

CUDA平台依托完整的编程模型、加速库与通信栈，可充分释放机架级分布式算力。开发者可通过NVIDIA集体通讯库 (NCCL)、NVIDIA推理传输库 (NIXL) 等组件，将Rubin GPU独立编程或作为72卡之一通过NVLink域统一调度，实现大模型整机柜无缝扩展，无需定制切分与手动编排。

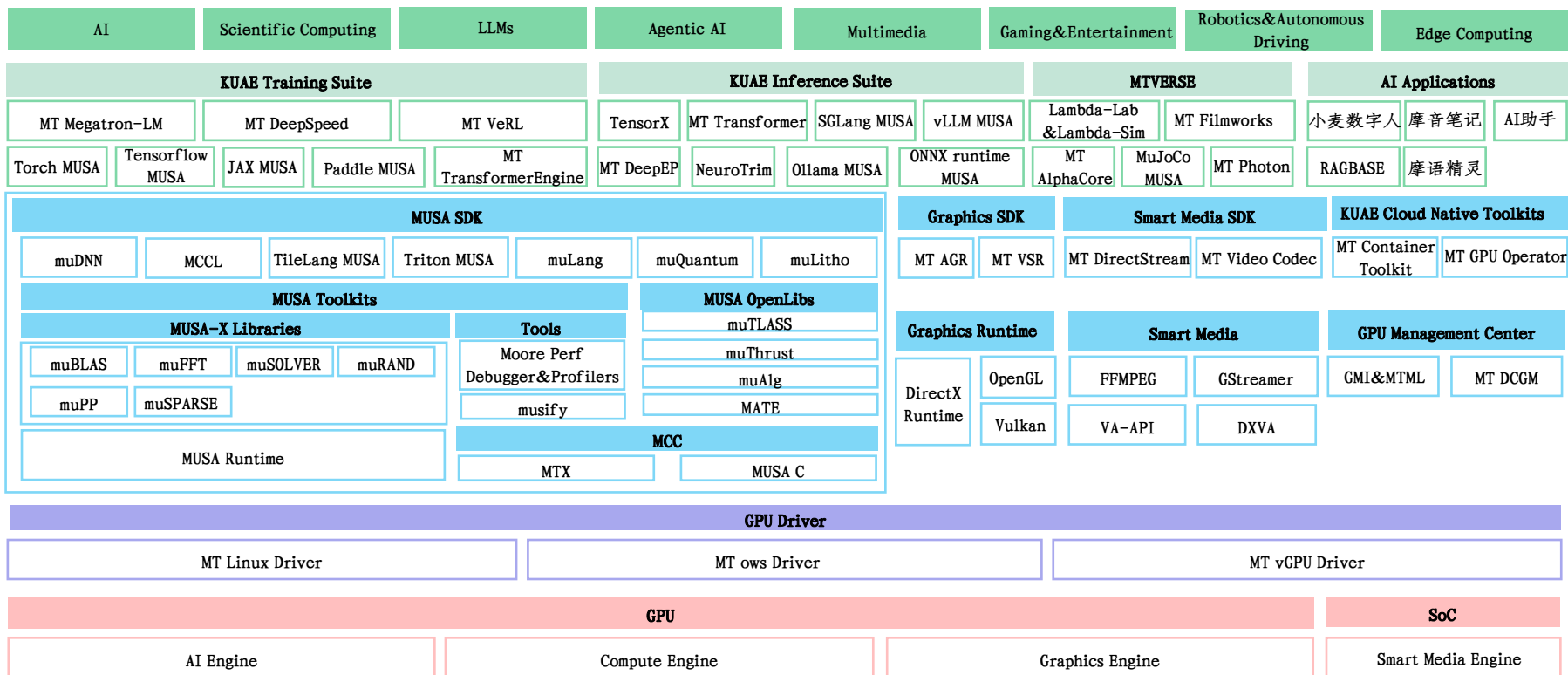
在内核与加速库层面，NVIDIA为最复杂的AI负载提供了高度优化的基础计算组件，cuDNN、CUTLASS、新一代Transformer Engine等优化组件与Rubin的Tensor Core、HBM4、NVLink6深度耦合，保障密集、稀疏与高通信负载的高性能，让开发者专注模型逻辑并充分发挥平台算力。

CUDA-X AI包括：cuDNN (深度学习基础库)、TensorRT (推理优化引擎)、NeMo Retriever (企业级检索增强生成工具) 等工具，支持从数据预处理到模型部署的全流程加速。

来源：NVIDIA，至顶智库整理绘制

2.8 芯片开发平台—摩尔线程MUSA

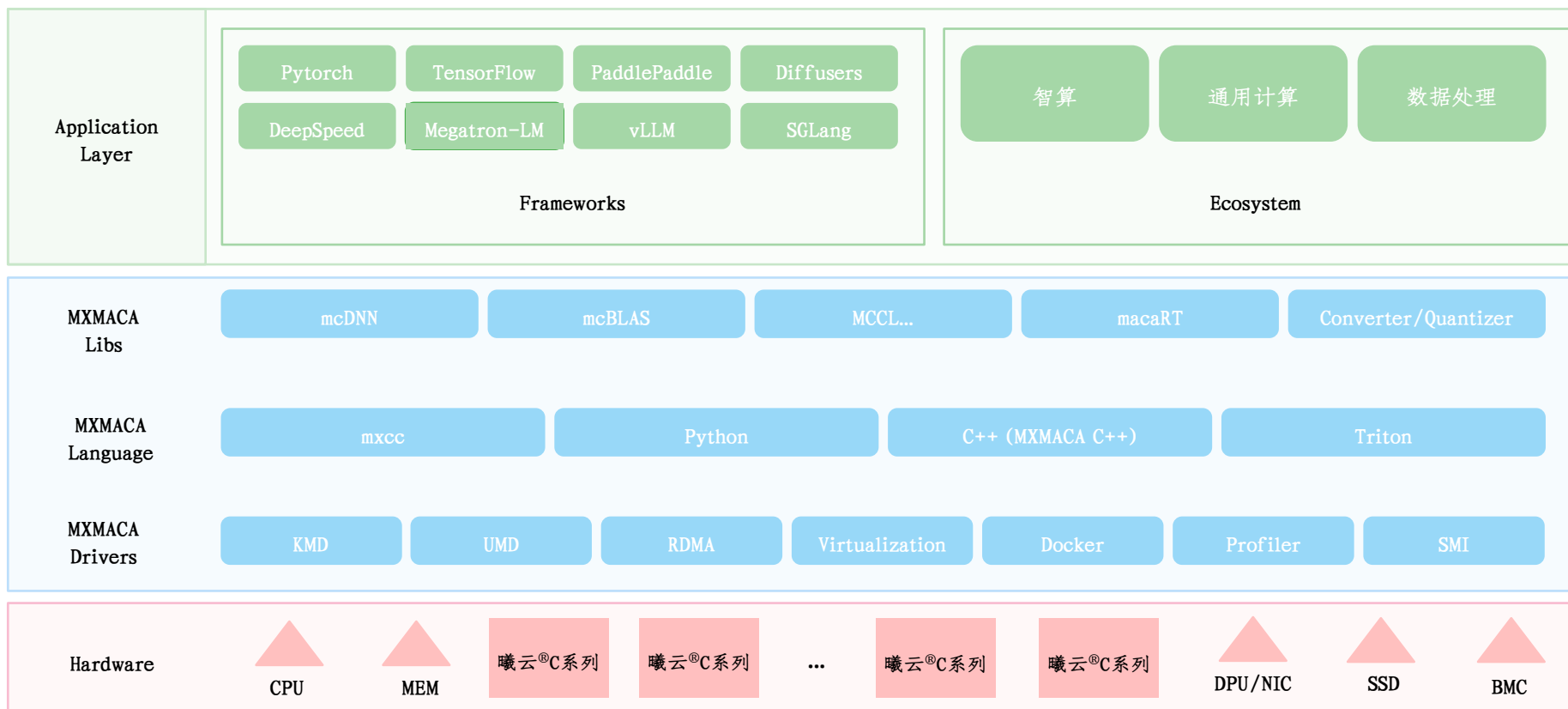
MUSA是摩尔线程自主研发的元计算统一系统架构，是芯片、硬件、软件栈，以及生态系统一体化的架构总称，提供从驱动、运行时、编程模型、工具链、加速库、框架、开发者套件到AI训推与管理软件的全栈支持，并深度兼容CUDA。其独特之处在于单芯片上同时支持AI计算、图形渲染、科学计算与物理仿真、超高清视频编解码等技术突破。摩尔线程通过MUSA实现全产品线统一的指令集与架构标准，持续积累和发展软件生态。



来源：摩尔线程，至顶智库整理绘制

2.8 芯片开发平台—沐曦MXMACA

MXMACA异构计算平台是沐曦基于“自主创新与开放兼容”双轨并行战略构建的统一异构计算与基础软件平台。其依托沐曦自研指令集与GPU并行计算引擎，整合了主流算法框架、计算库、通信库、操作系统、编程语言、调试与运维工具等，为沐曦全系列GPU产品提供了一套统一、完整且高效的全栈式软件工具链，涵盖应用开发、功能调试和性能调优等核心环节。



来源：沐曦集成电路，至顶智库结合公开资料整理绘制
insights.zhiding.cn

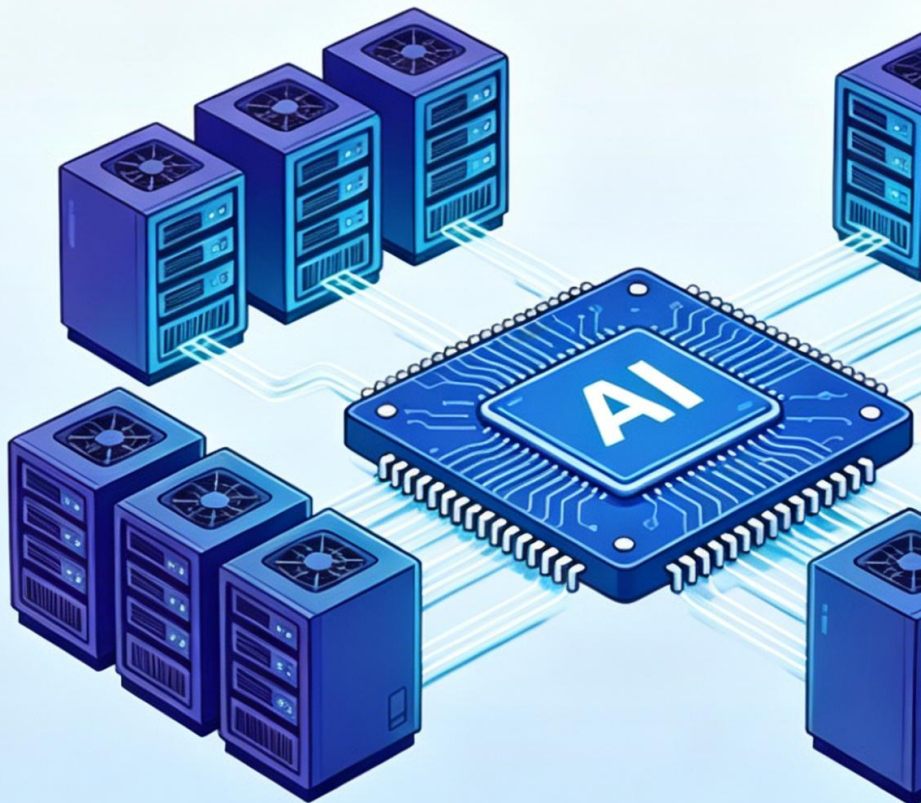
第三章 全球AI工作站及服务器发展情况

AI工作站

- 塔式AI工作站
- 移动AI工作站
- 迷你AI工作站
- 入门级AI工作站
- 专业级AI工作站
- 企业级AI工作站

AI服务器

- 训练AI服务器
- 推理AI服务器
- 超节点服务器



3.1 AI工作站分类

AI工作站是为AI任务量身打造的高性能计算平台。其集成AI加速芯片（如GPU/NPU）、大内存/高带宽、高效散热、专用软件栈，支持本地LLM训推、数据处理、科学计算等负载，兼顾桌面级部署与服务器级算力等特性，是介于消费级PC与机架式服务器之间的形态。其主流分类体系可从两大维度划分：按形态与部署场景，分为塔式AI工作站、移动AI工作站、迷你AI工作站三类；按算力等级与负载适配，分为入门级AI工作站、专业级AI工作站、企业级AI工作站三类，可覆盖从个人开发到企业级部署的全场景AI算力需求。

AI工作站分类体系



来源：至顶智库结合公开资料整理绘制

3.1 塔式AI工作站

品牌	图示	产品名称	CPU	GPU	内存	存储
DELL		Precision 5860	Intel® Xeon® w7-2595X	NVIDIA® RTX™ 6000 Ada 48GB	2TB DDR5	56TB
HP		Z1 Tower G1i	Intel® Core™ Ultra 9 285K	NVIDIA® A1000 8GB	128GB DDR5	10TB
联想		ThinkStation P3 Tower	Intel® Core™ i9-14900K	NVIDIA® RTX™ 5000 Ada 32GB	128GB DDR5	37TB
ASUS		ExpertCenter Pro ET900A X9	AMD RYZEN™ Threadripper™ PRO 9000 WX	NVIDIA® RTX™ 6000 Ada 48GB	2TB DDR5	—

来源：至顶智库结合公开资料整理绘制

3.1 移动AI工作站

品牌	图示	产品名称	CPU	GPU	内存	存储
DELL		Dell Pro Max 14 MC14250	Intel® Core™ Ultra 7 265H vPro®	Intel® Arc™ Pro 16GB	64GB LPDDR5x	2TB
HP		ZBook Studio G11	Intel® Core™ Ultra 9	NVIDIA® RTX™ 3000 Ada 16GB	64GB DDR5	4TB
联想		ThinkPad T1g	Intel® Core™ Ultra 9 285H	NVIDIA® GeForce RTX™ 5070 12GB	64GB LPDDR5x	8TB
Apple		MacBook Pro 16英寸 (M5 Max)	Apple M5 Max SoC芯片 (18 核中央处理器)	Apple M5 Max SoC芯片	128GB	8TB

3.1 迷你AI工作站

品牌	图示	产品名称	CPU	GPU	内存	存储
NVIDIA		NVIDIA DGX Spark	NVIDIA Grace (GB10)	NVIDIA Blackwell (GB10)	128 GB LPDDR5x	4TB
HP		Z2 Mini G1a	AMD Ryzen™ AI Max+ PRO 395	AMD Radeon™ 8060S	128 GB LPDDR5	8TB
ASUS		ASUS Ascent GX10	NVIDIA Grace (GB10)	NVIDIA Blackwell (GB10)	128 GB LPDDR5x	4TB
Apple		Mac mini	Apple M4 Pro SoC芯片 (14核中央处理器)	Apple M4 Pro SoC芯片	48GB	8TB

来源：至顶智库结合公开资料整理绘制

3.1 入门级AI工作站

品牌	图示	产品名称	CPU	GPU	内存	存储
DELL		Precision 3680	Intel® Core™ i9-14900K	NVIDIA® RTX™ 4500 Ada 24GB	128GB DDR5	36TB
HP		ZBook Ultra G1a	AMD RYZEN™ AI Max+ PRO 395	AMD Radeon™ 8060S	128GB LPDDR5x	4TB
联想		ThinkStation P2	Intel® Core™ i9-14900	NVIDIA RTX™ A2000 12GB	128GB DDR4	10TB
Apple		Mac mini	Apple M4 Pro SoC芯片 (14核中央处理器)	Apple M4 Pro SoC芯片	48GB	8TB

来源：至顶智库结合公开资料整理绘制

3.1 专业级AI工作站

品牌	图示	产品名称	CPU	GPU	内存	存储
DELL		Precision 7875	AMD Ryzen™ Threadripper™ PRO 7995WX	NVIDIA® RTX™ 6000 Ada 48GB	2TB DDR5	56TB
HP		Z4 G5	Intel® Xeon® W7-2595X	NVIDIA® RTX™ 6000 Ada 48GB	512GB DDR5	92TB
联想		ThinkStation P920	Intel® Xeon® Platinum 8160	NVIDIA® A6000 48GB	2TB DDR4	60TB
Apple		Mac studio	Apple M3 Ultra SoC芯片 (32核中央处理器)	Apple M3 Ultra SoC芯片	96GB	16TB

3.1 企业级AI工作站

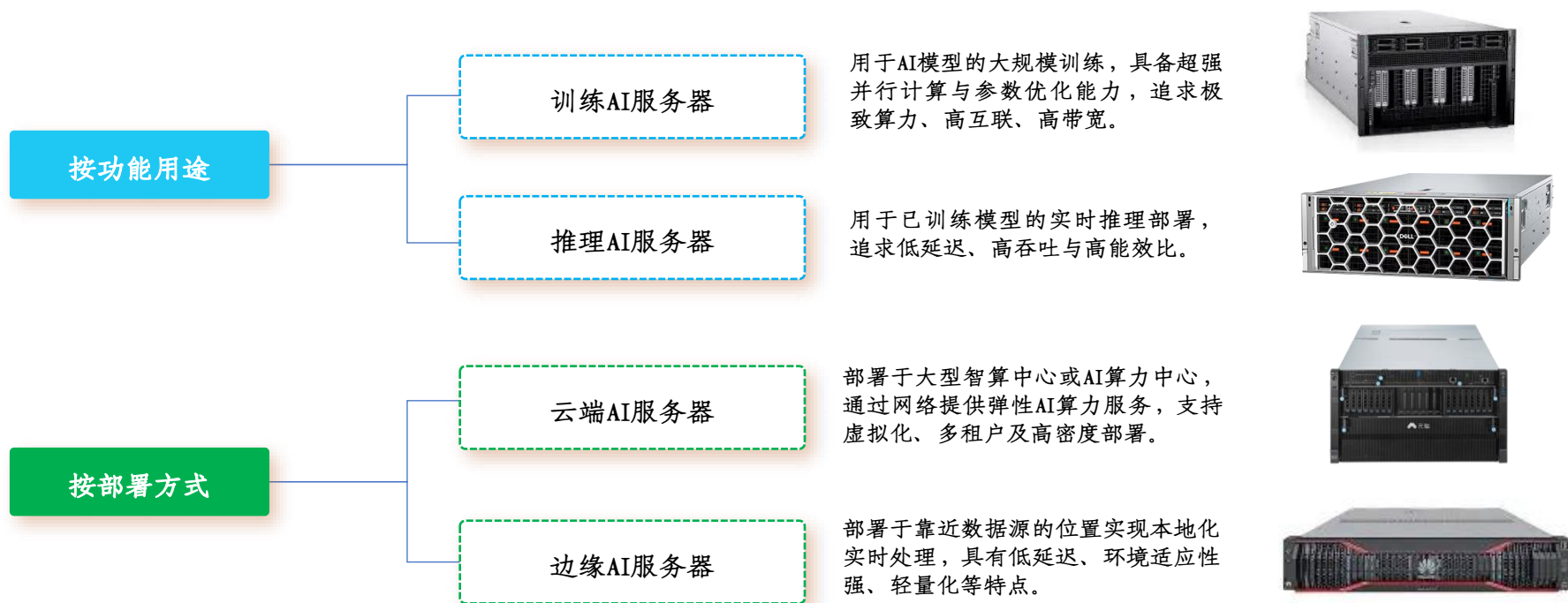
品牌	图示	产品名称	CPU	GPU	内存	存储
DELL		Precision 7960	Intel® Xeon® w9-3595X	4 × NVIDIA® RTX™ 6000 Ada 48GB	4TB DDR5	—
HP		Z8 Fury G5	Intel® Xeon® W9-3595X	4 × NVIDIA® RTX™ 6000 Ada 48GB	2TB DDR5	136TB
联想		ThinkStation PX	Intel® Core™ i9-14900K	4 × NVIDIA® RTX™ 6000 Ada 48GB	2TB DDR5	60TB
NVIDIA		DGX Station	Grace 72-Core Neoverse V2	NVIDIA Blackwell Ultra 252GB	496GB LPDDR5x	—

来源：至顶智库结合公开资料整理绘制

3.2 AI服务器分类

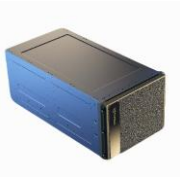
AI服务器是为AI任务量身打造的高性能计算系统。通过集成高性能AI芯片、高带宽存储、高速互联组件、高效散热系统及专用软件栈，AI服务器具有高算力输出、高内存带宽、高速互联等能力，能够高效处理AI模型训练与推理等大规模并行计算任务。AI服务器类别可从两大维度划分：按功能用途，可分为训练AI服务器和推理AI服务器；按部署方式，可分为云端AI服务器和边缘AI服务器。

AI服务器分类体系



来源：至顶智库结合公开资料整理绘制

3.2 训练AI服务器

品牌	图示	产品名称	GPU	CPU	系统内存	存储	算力
		DGX H200	8 × NVIDIA H200 Tensor Core GPUs, 141GB	Dual Intel® Xeon® Platinum 8480C Processors 112 Cores total, 2.00GHz (Base), 3.80 GHz (Max Boost)	2TB	OS:2x1.92TB NVMe M.2 internal storage: 8x 3.84TB NVMe U.2	FP8: 32 PFLOPS
NVIDIA		DGX B200	8 x NVIDIA Blackwell GPUs	Dual Intel® Xeon® Platinum 8570 Processors 112 Cores total, 2.1 GHz (Base), 4 GHz (Max Boost)	2TB/4TB	OS:2x1.9TB NVMe M.2 internal storage: 8x 3.84 TB NVMe U.2	FP4 : 144 PFLOPS
		DGX B300	8 x NVIDIA Blackwell Ultra SXM	Intel® Xeon® 6776P Processors	2TB/4TB	OS:2x1.9TB NVMe M.2 Internal storage: 8x 3.84 TB NVMe E1.S	FP4 : 144 PFLOPS

说明：NVIDIA DGX H200、DGX B200、DGX B300均为训推一体服务器。

来源：至顶智库结合公开资料整理绘制

3.2 训练AI服务器

品牌	图示	产品名称	GPU	CPU	系统内存	存储	算力
NVIDIA		Vera Rubin NVL72	72 x NVIDIA Rubin GPU	36 x NVIDIA Vera CPU	54 TB LPDDR5X	-	NVFP4: 2,520 PFLOPS FP8/FP6: 1,260 PFLOPS
DELL		New PowerEdge XE9780L	8 x NVIDIA HGX B300 SXM6 GPUs	Dual Intel Xeon 6 Processors with up to 86 cores perwith up to 192cores per processor	32 x DDR5 RDIMM memory slots, maximum total capacity up to 4 TB	最大 307.2 TB	-
联想		联想问天 WA8080a G5	支持主流的OAM 2.0 GPU	Dual Intel® Xeon® 6 SP Series scalable processors	32 x DDR5 6400MHz RDIMM, 可支持 8000MHz MRDIMM	最多支持 32 x 2.5 英寸热 插拔硬盘位 支持 16 块 2.5inch NVMe 或 16* E3.S SSD	-

说明：NVIDIA Vera Rubin NVL72、DELL New PowerEdge XE9780L、联想问天 WA8080a G5均为训推一体服务器。

来源：至顶智库结合公开资料整理绘制

3.2 推理AI服务器

品牌	图示	产品名称	GPU	CPU	系统内存	存储	算力
NVIDIA		RTX PRO servers	8 × NVIDIA RTX PRO 6000 Blackwell Server Edition or RTX PRO 4500 Blackwell Server Edition	2 × Intel Xeon or 2 × AMD EPYC	128 GB DDR5 ECC per GPU (1 DIMM per channel)	2 × 4 TB NVMe* 1 × 1 TB NVMe boot drive	单卡FP32算力120 TFLOPS，FP4 AI峰值算力4 PFLOPS；整机满配8卡 FP4 AI总算力 32 PFLOPS。
HPE		HPE ProLiant DL380a Gen12	16 x 单宽 NVIDIA L4 或 10 个双宽 L40S、H100 NVL、H200 GPU，或 8 个双宽 RTX PRO 6000 Blackwell 服务器版 GPU	2 × Intel® Xeon 6 Processors	最大 4TB DDR5 RDIMM	支持SFF、EDSFF 驱动器	-
联想		ThinkSystem SR675i V3	8x NVIDIA RTX PRO 6000 Blackwell Server Edition	2x AMD EPYC™ 9535 Processors, 64 cores each, 300W TDP	1.5TB (24x 64GB TruDDR5 6400MHz RDIMMs)	2x 3.84TB NVMe E3.S SSDs + 2x 960GB M.2 NVMe SSDs	-

来源：至顶智库结合公开资料整理绘制

3.2 国内超节点服务器

超节点服务器通过单节点内增加芯片数量，具备超高互联带宽、纵向扩展与集成化等优势，在性能、成本、组网、运维等方面表现突出。其能够提供超高互联带宽与超低通信时延，有效支撑并行计算任务，缩短模型训练周期，提升整体可靠性。华为昇腾384超节点通过总线技术实现384个NPU之间大带宽低时延互联，优化资源调度以满足AI训练与推理需求；中科曙光scaleX640超节点采用“一拖二”高密方案实现单机柜640卡超高速互连，算力性能实现倍增；阿里云磐久AL128超节点服务器采用超大集群的服务架构，重构GPU间互连方式，实现算力与通信协同；浪潮元脑SD200，可实现单机内运行超万亿参数大模型，并支持领先大模型机内同时运行及多智能体按需调用；昆仑芯发布的超节点方案通过硬件创新提升全互联通信带宽，助力万卡级智算集群建设。

HUAWEI

昇腾384超节点



中科曙光
Sugon

曙光scaleX640



阿里云

磐久AL128超节点



浪潮信息

元脑SD200



昆仑芯
KUNLUNXIN

昆仑芯超节点



- 2025年7月，华为首次展出昇腾384超节点，即Atlas 900 A3 SuperPoD；
- 该产品基于超节点架构，跨节点通信带宽提升15倍，通信时延下降10倍，业界唯一突破Decode时延15ms，千亿稠密模型训练性能可达传统集群2.5倍以上；
- 通过系统工程优化，实现资源高效调度，更好满足模型训练和推理对低时延、大带宽、长稳可靠的要求。

- 2025年11月，中科曙光scaleX640超节点亮相，该产品作为单机柜级超节点，可实现20倍算力密度提升，支撑十万卡级大集群扩展；
- scaleX640超节点通过算、存、网、电、冷一体化紧耦合设计，采用“一拖二”高密方案，实现单机柜640卡超高速互连，综合算力性能实现倍增，可将MoE万亿参数大模型训练推理性能提升30%-40%。

- 2025年9月，阿里云推出磐久AI Infra2.0 AL128超节点服务器；
- 采用面向下一代超大集群的服务架构，重构GPU间互连方式，旨在大模型训练与推理中实现算力与通信协同最优，相对于传统架构，同等AI算力下推理性能还可提升50%。
- 整柜支持128~144颗GPU芯片，支持高达350kw供电能力和500kw散热能力。

- 2025年8月，面向万亿大模型训推场景，浪潮信息研发元脑SD200超节点服务器，采用创新的多主机低延迟内存语义通信架构，以开放系统设计聚合64路本土GPU芯片。
- 基于融合架构2.0，采用多主机三维网格（3D Mesh）设计，实现单机64颗本土GPU芯片的高速互连。配合远端GPU虚拟映射技术，突破多主机统一编址的难题。

- 2025年4月，昆仑芯推出超节点产品，为AI算力集群性能优化提升提供全栈解决方案；
- 通过硬件架构创新，该产品实现全互联通信带宽提升8倍，MoE大模型单节点训练性能提升5-10倍，单卡推理效率提升13倍；
- 支持IB/RoCE通信，实现跨柜高带宽、低延迟数据传输，支持万卡以上规模的智算集群构建。

来源：华为、中科曙光、阿里云、浪潮信息、昆仑芯，至顶智库结合公开资料整理绘制

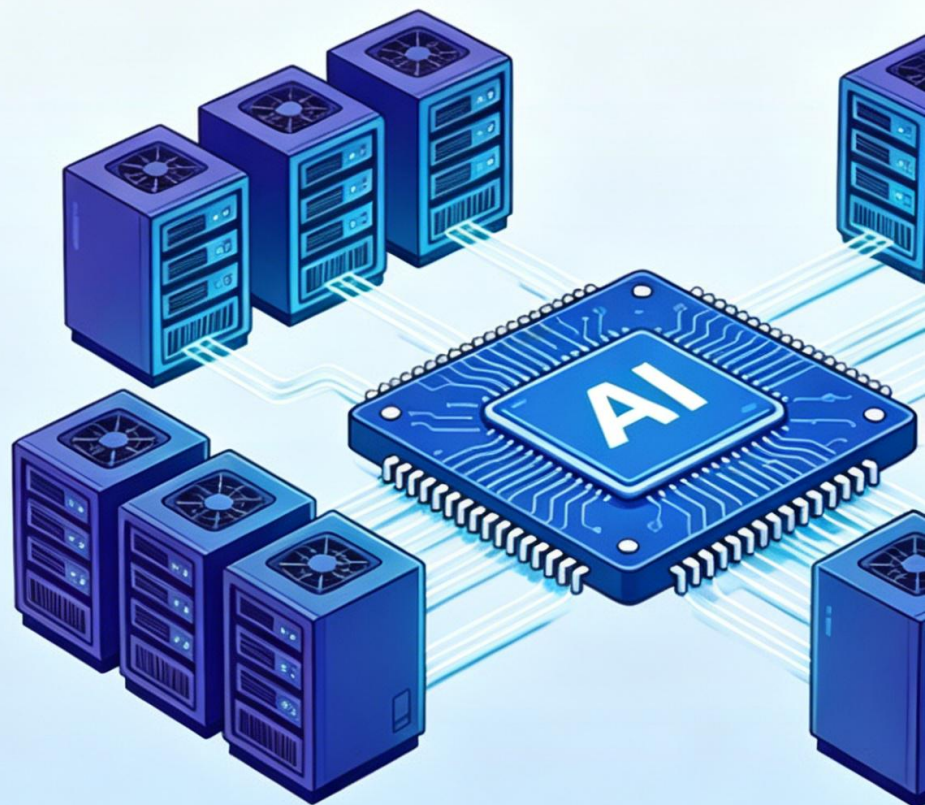
第四章 全球AI算力中心发展情况

AI算力中心特征

AI算力中心服务器散热方式

AI算力中心电力供给与消耗

AI算力中心建设进展



4.1 AI算力中心典型特征

AI算力中心通过采用领先的人工智能计算架构，为各类场景（如模型训练、模型推理、AI应用等）提供所需算力服务的新型算力基础设施。AI算力中心通常配备高性能计算资源，如AI计算芯片（GPU、TPU等）、大规模存储、高速网络连接以及能够处理大数据集和高计算负载的硬件和软件平台。其具有算力密度高、电力供给要求高、散热与液冷需求大、软硬协同能力显著等特点。

AI算力中心典型特征

高密度算力

AI算力中心的算力密度远高于传统数据中心，核心资源从以CPU为主转向GPU、NPU、TPU等加速芯片，能够支撑大模型训练、推理和海量并行计算场景。

电力供给要求高

AI模型训练/推理持续消耗大量电力，AI算力中心通常具备更高功率机柜、更强配电系统以及更严格的供电稳定性要求。

散热与液冷需求大

AI算力中心更多采用液冷、冷板式散热、浸没式散热等先进冷却技术，以提升散热效率和PUE表现。

高速网络互联

由于AI模型训练等场景依赖大规模集群协同，因此需要高带宽、低时延、无阻塞网络架构，通常配置高速InfiniBand或以太网互联，以保证GPU集群之间高效通信。

大规模存储与数据吞吐优化

AI算力中心通常配备高性能分布式存储、并行文件系统和高速数据通道，以支持训练数据、参数、检查点和推理数据的快速读写。

软硬高效协同

AI算力中心打造以芯片、服务器、网络、存储、集群调度平台、AI框架和模型工具链的系统化协同。



4.2 AI算力中心主要概念—PUE

PUE（电源使用效率）是AI算力中心最重要的能效指标，衡量算力中心消耗的能源中有多少真正用于驱动IT设备，有多少被散热、照明、供电损耗等基础设施消耗。由于GPU密度高、热量集中，AI算力中心的散热压力远超传统机房，因此发展高度依赖PUE管控能力。通过技术优化持续降低PUE，是控制AI算力中心运营成本、提升长期运营效益的关键因素。 $1.30 \leq PUE \leq 1.50$ 算力中心为准入合格型； $1.20 \leq PUE \leq 1.30$ 算力中心为先进节能型，符合大型算力中心要求； $PUE \leq 1.20$ 算力中心为高效标杆型，代表行业先进水平。

PUE计算公式及划分区间

算力中心总能耗：包括IT设备、制冷系统（空调、风扇、水泵）、配电系统（UPS损耗、变压器）、照明以及其他辅助设施的总和

$$PUE = \frac{\text{算力中心总能耗}}{\text{IT设备能耗}}$$

IT设备能耗：服务器、交换机、路由器、存储设备等核心算力设备的能耗

PUE水平	类型与市场定位
$PUE \leq 1.20$	高效标杆型： 代表行业领先水平。通常是采用自然冷却、液冷技术或在气候寒冷地区（如乌兰察布、张家口等枢纽节点）建设的新型绿色算力中心。
$1.20 < PUE \leq 1.30$	先进节能型： 符合国家“东数西算”枢纽节点集群内新建大型/超大型算力中心的普遍要求（通常要求 PUE 低于1.25或1.3）。
$1.30 < PUE \leq 1.50$	准入合格型： 这是国家规定的能效限定值（红线）。新建及改扩建的算力中心必须至少达到此等级。PUE大于1.5的算力中心被视为不合格，面临关停或节能改造。

来源：至顶智库结合公开资料整理绘制

4.3 AI算力中心服务器散热方式

AI算力中心服务器散热主要分为风冷与液冷两种方案。风冷技术包括房间级冷却、行级冷却、机柜级冷却，是当前中低密度算力与存量机房的主流方式。但风冷散热能力有限、PUE普遍在1.4以上，无法适配高功耗算力中心。液冷散热效率高，包括冷板式、单相浸没式、双相浸没式、喷淋式等散热方式，可满足超高密度GPU集群算力需求，符合低碳能效政策。目前行业形成风冷存量主导、液冷增量普及的共存模式，伴随AI芯片功耗与算力密度持续提升、能耗管控政策收紧，液冷已成为AI算力中心服务器的主流散热方案，也是算力中心绿色、稳定、高效运行的关键保障。

风冷散热方式

散热方式	原理	适用功率密度	优点
房间级冷却	空调对整个机房供冷，常配合架空地板送风、冷/热通道封闭来隔绝气流，防止混合。	单柜 ≤ 10-15kW	技术最成熟，初期投资相对较低
行级冷却	将冷却单元直接部署再机柜排之间，紧贴热源，送风距离极短。	单柜 10-30kW	冷却效率高，可灵活适配
机柜级冷却	将冷却单元完全集成在单个机柜上，形成独立的闭环冷却，热气不排入机房。	单柜 ≥ 30-50kW	冷却密度高，可快速改造旧机房

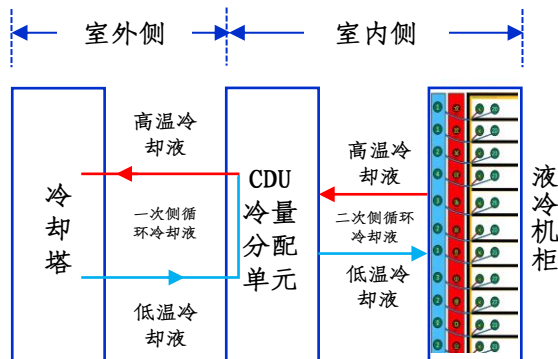
液冷散热方式

散热方式	原理	投资成本	PUE
冷板式	液体流经贴合在CPU/GPU等发热元件上的金属冷板来直接带走热量，剩余部件仍用风扇辅助散热。	初始投资中等，运维成本低	1.1-1.2
单相浸没式	服务器完全浸没在特制的绝缘冷却液中，液体始终保持液态，通过循环泵带走热量。	初始投资及运维成本高	<1.05
双相浸没式	冷却液在吸收热量时从液体变为气体，从而大大增强散热效果。蒸汽在冷却时凝结回液体。	初始投资及运维成本高	<1.09
喷淋式	冷却液从服务器顶部喷淋，对流换热降温。	结构改造及液体消耗成本大，液冷系统初始投资成本低	<1.1

来源：至顶智库结合公开资料整理绘制

4.3 AI算力中心服务器散热方式

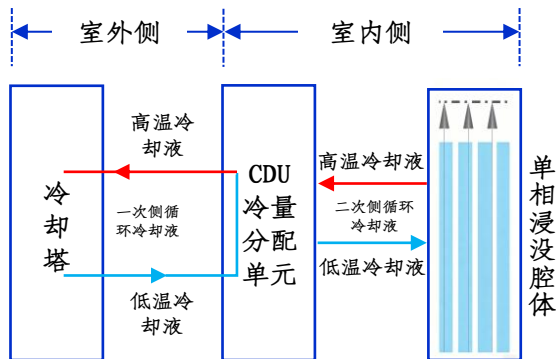
冷板式液冷系统原理图



冷板式液冷散热原理:

- 液冷板与芯片紧密贴合，实现高效导热；
- 在CDU循环泵驱动下，冷却液进入冷板通道，通过对流换热吸收芯片热量；
- 升温后的冷却液回流至CDU，经换热器将热量传递至一次侧冷却液；
- 一次侧冷却液通过冷却塔将热量排出，实现系统循环散热。

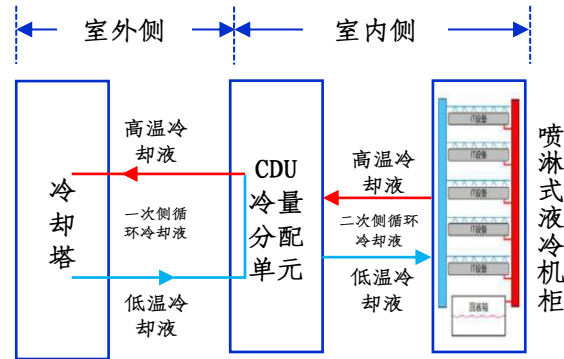
单相浸没式液冷系统原理图



单相浸没式液冷系统原理:

- CDU循环泵驱动二次侧低温冷却液由浸没腔体底部进入，流经竖插在浸没腔体中的IT设备时带走发热器件热量；
- 吸收热量升温后的二次侧冷却液由浸没腔体顶部出口流回CDU；
- 通过CDU内部的板式换热器将吸收的热量传递给一次侧冷却液；
- 吸热升温后的一次侧冷却液通过外部冷却装置（如冷却塔）将热量排放到大气环境中，完成整个制冷过程。

喷淋式液冷系统原理图



喷淋式液冷系统原理:

- 冷却液在冷量分配单元（CDU）中完成降温后，由循环泵输送至喷淋机柜内部；
- 冷却液进入机柜后，通过分液器或进液箱均匀分配，并依靠压力或重力驱动进入布液装置；
- 冷却液以喷淋形式直接作用于IT设备发热部件或其导热结构，实现高效换热；
- 吸热升温后的冷却液汇集至回液箱，经泵送回CDU，完成循环冷却过程。

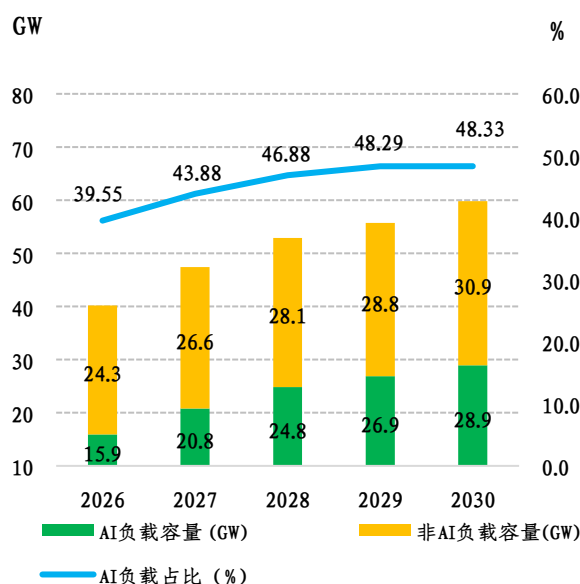
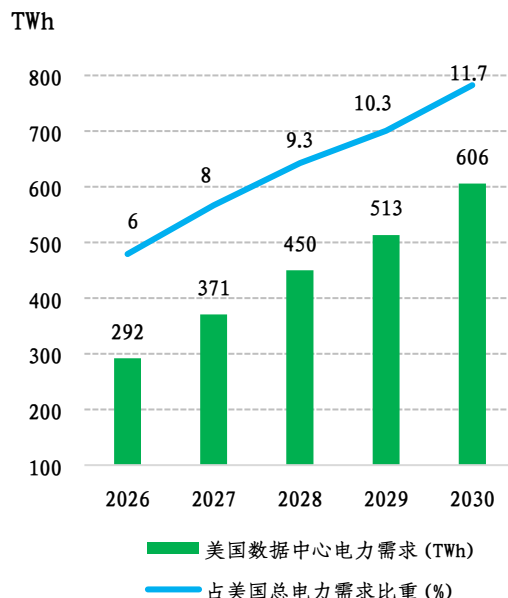
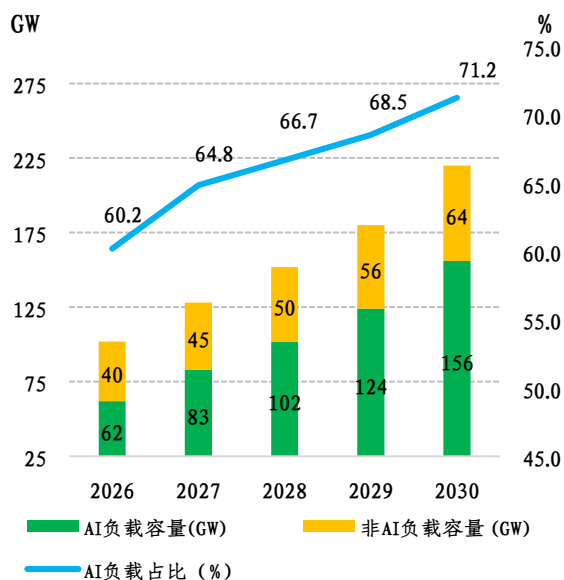
来源：至顶智库结合公开资料整理绘制

4.4 算力中心呈现“高AI占比、高功率密度、高电力消耗”特征

AI大模型训练与推理规模的不断扩张将推动全球算力中心容量与电力需求增长，加速超大规模AI算力中心发展。预计到2030年，全球算力中心容量将由2026年的102GW增长至220GW，其中AI负载容量由62GW提升至156GW，占比提升至71%。麦肯锡数据显示，伴随美国算力中心规模的不断扩张，算力中心年耗电量预计将由292TWh增长至606TWh，占全美电力需求比重将提升至11%，AI正成为美国新增电力需求的重要场景。RystadEnergy预测，中国算力中心2030年总容量预计接近60GW，AI负载占比提升至48%，AI算力中心正成为新增算力中心建设的重点方向。整体来看，全球算力中心呈现“高AI占比、高功率密度、高电力消耗”的发展趋势。

全球算力中心容量（2026-2030）

美国算力中心电力需求（2026-2030）中国算力中心容量（2026-2030）



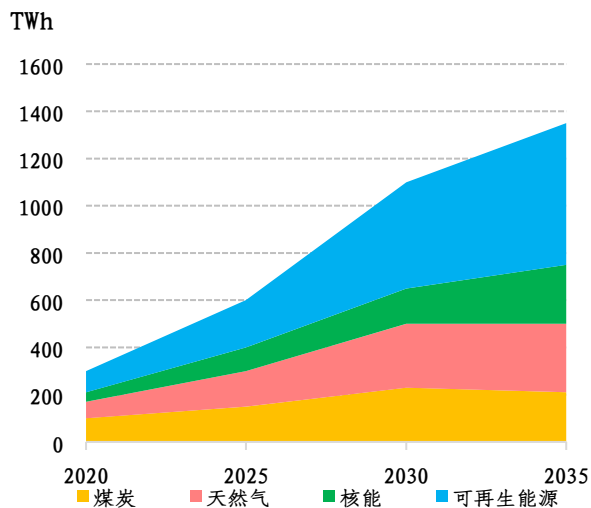
说明：算力中心容量是指算力中心在某一时刻能承载多大规模算力，算力中心电力需求是指算力中心一段时间内实际耗电量

来源：McKinsey, RystadEnergy, 至顶智库结合公开资料整理绘制

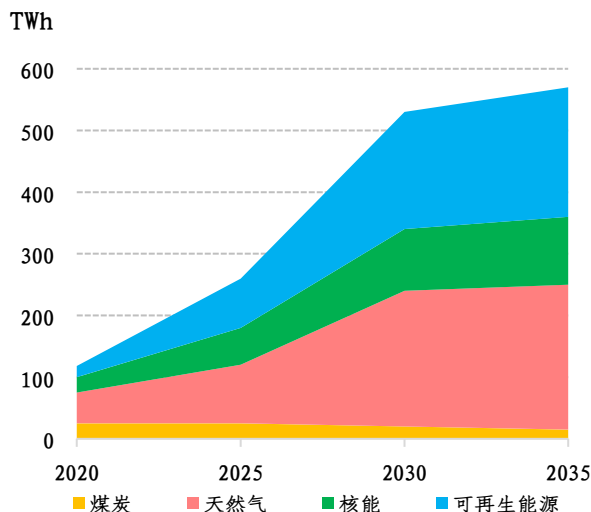
4.5 AI算力中心电力供给方式多元，新能源占比持续提升

AI算力中心正成为驱动全球电力需求增长与能源结构转型的核心引擎。全球算力中心目前的电力供应由煤炭、可再生能源、天然气及核能共同构成，其新增需求在推动可再生能源年均22%高速增长的同时，短期内仍需依赖化石能源满足约40%的增量，且预计2030年后核能的战略支撑地位将愈发显著。聚焦于中美两国的AI算力中心，两国均在短期内将化石能源作为新增电力供应的最大来源，可再生能源（太阳能、风能等）则是第二大新增电力供应来源。预计在2024年至2030年间为算力中心供电增加110TWh，可再生能源占总供电比重从20%提升至40%左右。两国各地区电力结构中风能和太阳能光伏的比例持续上升，算力中心运营商也在投资共址式可再生能源项目，推动未来全球算力中心将向绿电与核能主导的低碳算力体系全面过渡。

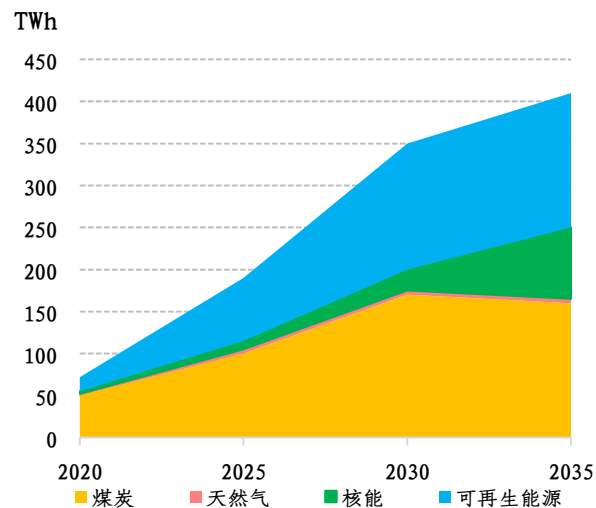
全球算力中心电力供应结构图



美国算力中心电力供应结构图



中国算力中心电力供应结构图

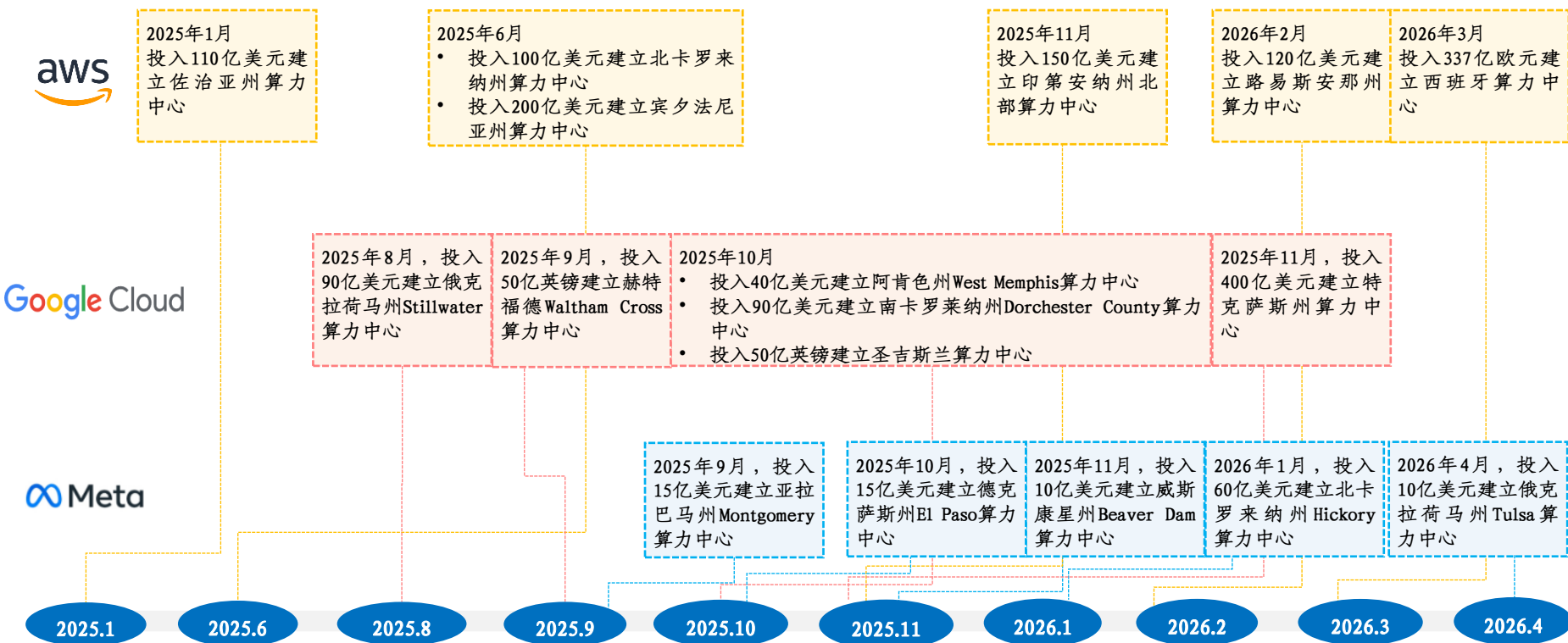


来源：Energy and AI, IEA, 至顶智库结合公开资料整理绘制

4.6 国外AI算力中心建设进展

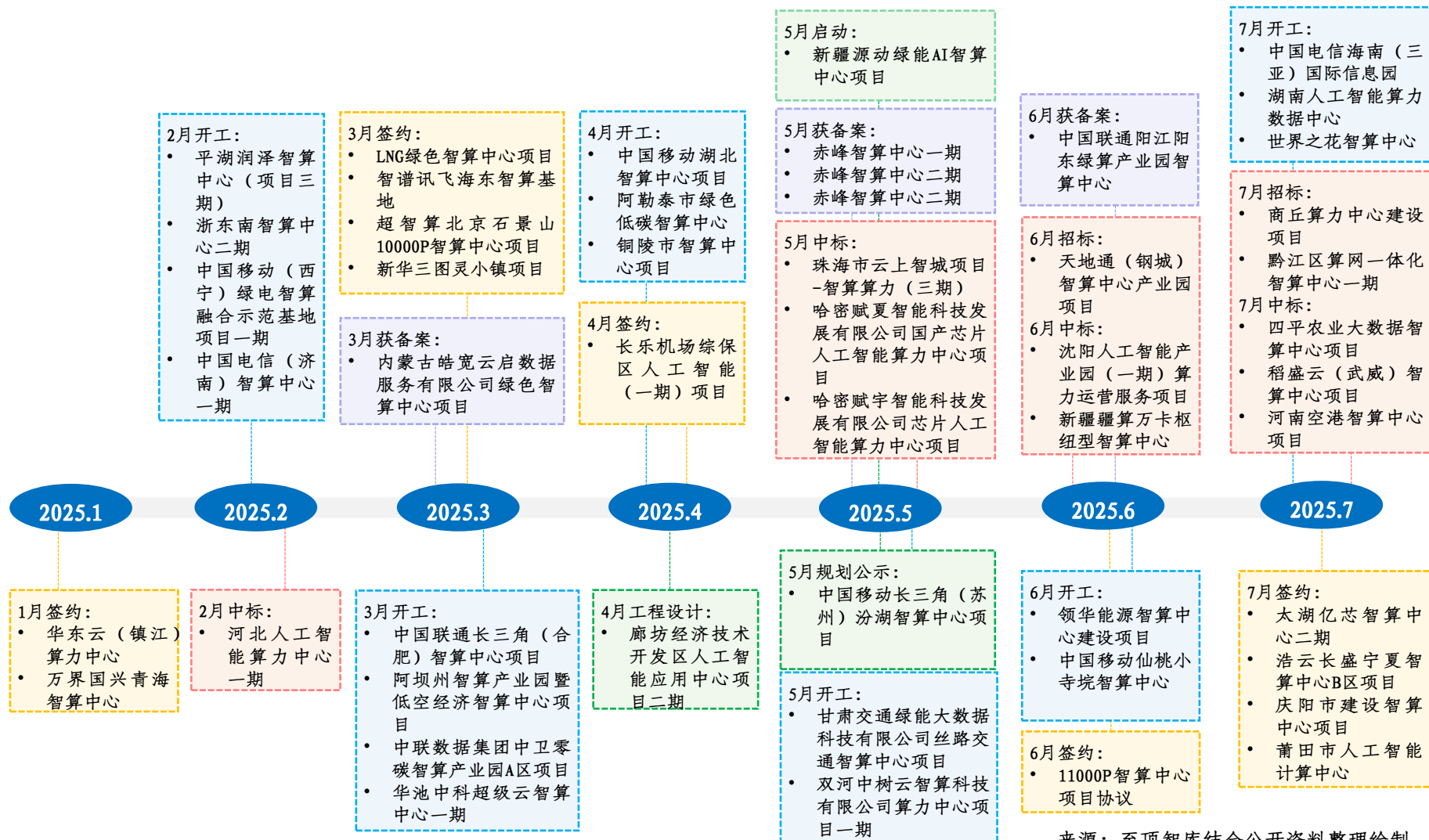
2025年以来全球AI算力中心建设呈现出加速扩张态势。以AWS、Google Cloud和Meta为代表的科技巨头持续加大投入，整体表现为投资规模显著提升，覆盖北美、欧洲及亚洲地区。AWS主导全球化超大规模布局，Google Cloud推进多区域持续建设，Meta则以美国本土集群化扩展为主。总体而言，全球AI算力中心正进入高强度、长周期投入阶段，并加速向超大规模与集群化方向演进。

全球主要AI算力中心开工建设情况（2025年以来）



来源：至顶智库结合公开资料整理绘制

4.7 国内AI算力中心建设进展



来源: 至顶智库结合公开资料整理绘制

4.7 国内AI算力中心建设进展



来源: 至顶智库结合公开资料整理绘制

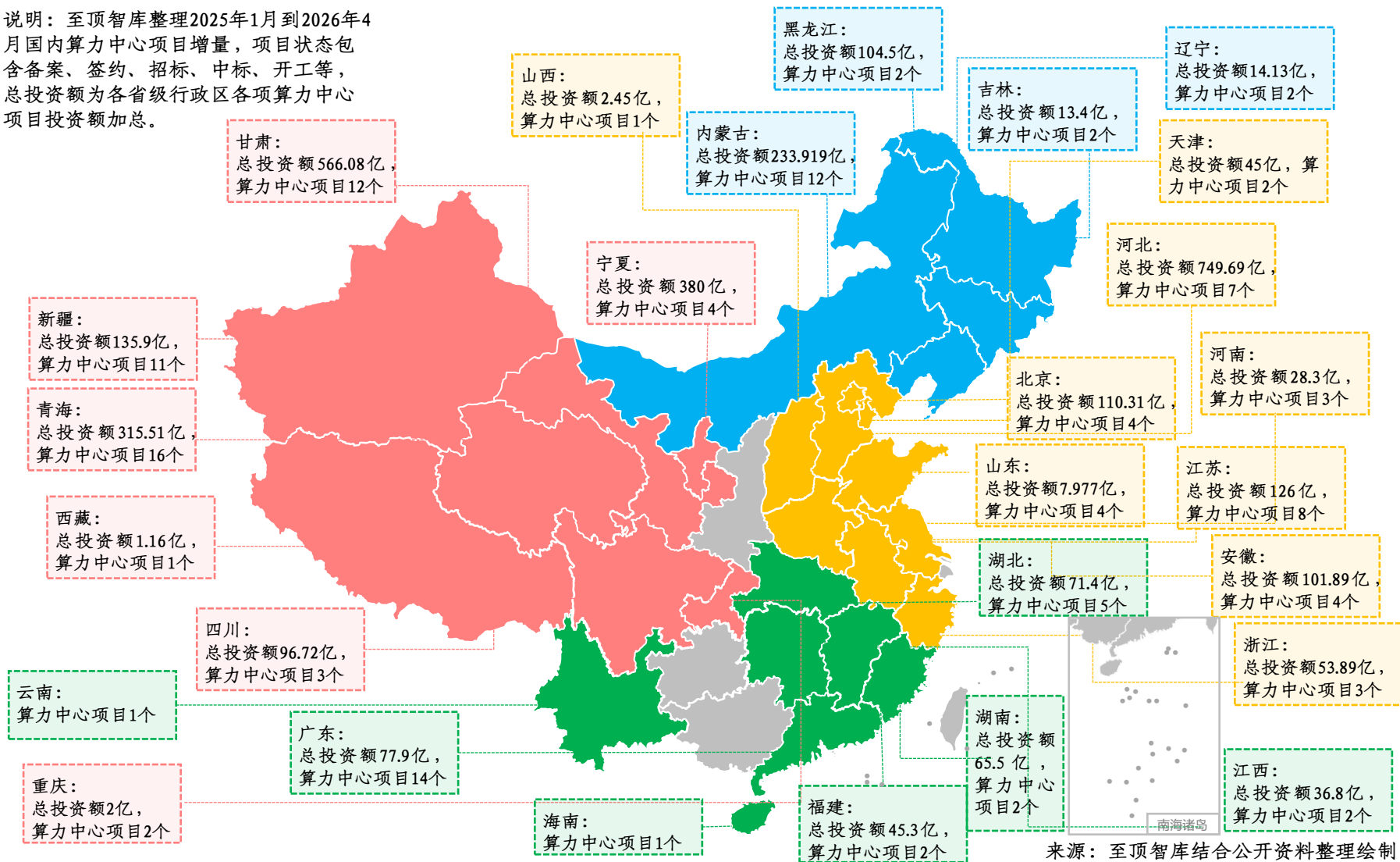
4.7 国内AI算力中心建设进展



来源：至顶智库结合公开资料整理绘制

4.7 国内AI算力中心建设进展

说明：至顶智库整理2025年1月到2026年4月国内算力中心项目增量，项目状态包含备案、签约、招标、中标、开工等，总投资额为各省级行政区各项算力中心项目投资额加总。

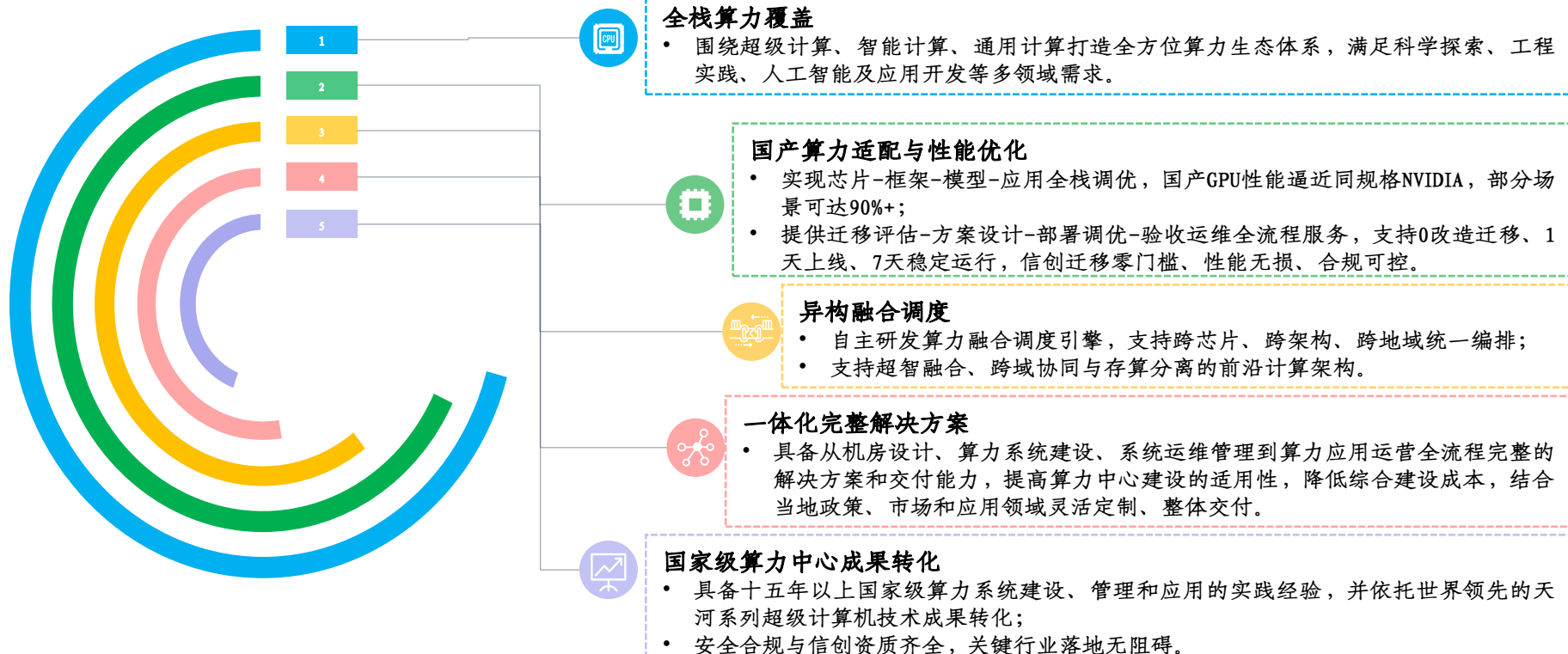


来源：至顶智库结合公开资料整理绘制

4.8 AI算力中心典型案例—天河计算机

作为国内算力系统建设与运营一体化的典型企业，天津市天河计算机技术有限公司依托国家超级计算天津中心的技术沉淀，打造面向AI算力中心的全栈融合算力解决方案。其核心优势包括全栈算力覆盖、国产算力适配与性能优化、异构融合调度、一体化完整解决方案、国家级算力中心成果转化五大维度。目前公司已形成算力系统建设、管理、应用三大系列产品矩阵，推动多地算力系统落地，并在政务、医疗、教育领域打造信创替代方案。

天河计算机核心优势



来源：天河计算机，至顶智库结合公开资料整理绘制

4.8 AI算力中心典型案例—天河系列产品架构

算力应用

智能应用	智能医疗	智能政务	智能教育	智慧金融	AI for science	...
科创应用	新材料研发	工程仿真	生物医药	气候气象	石油化工	...

算力运营

灵犀-智能应用开发平台	大模型微调	RAG和知识库	模型部署	智能体	提示词工程	模型评估
-------------	-------	---------	------	-----	-------	------

应用开发

星维-跨区域异构融合算力平台				璇玑-云原生统一计算平台			
跨域算力接入	任务管理	资源管理	大模型智能应用	一云多芯	虚拟机管理	虚拟存储管理	虚拟网络管理
跨域资源调度	数据管理	应用管理	计费与运营	CPU超分	故障迁移	租户管理	快照与镜像

系统环境

应用环境	超算	编译器	基础函数库	智算	推理框架	训练框架	算子算法库	通算	计算虚拟化
		消息传递接口	并行算法库		基础开发环境	并行通信库	分布式优化		存储虚拟化

存储系统

穹宇-高性能分布式存储系统				星磐-云原生统一存储系统			
分级存储	高速缓存	高带宽	高性能	块存储	对象存储	文件存储	存储池化
分布式	配额	快照	安全访问	高可靠性	数据冗余	高扩展性	权限控制

算力设备



来源: 天河计算机, 至顶智库结合公开资料整理绘制

第五章 AI算力典型应用场景

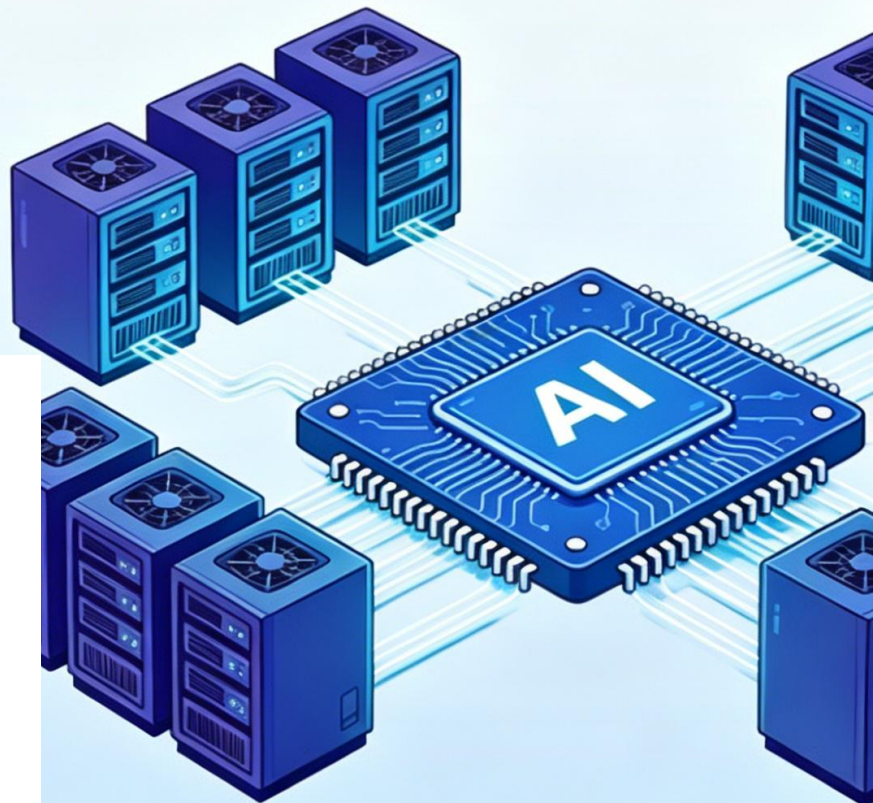
科学智能

具身智能

自动驾驶

工业制造

交通运输



骏羽² (课件收集)



完整版 扫码 私信: DCC17

更多

扫一扫上面的二维码图案，加我微信

www.ipoiipo.cn