2026 年 05 月 16 日

计算机行业研究

买入（维持评级）

行业专题研究报告

证券研究报告

计算机组
分析师：刘高畅（执业 S1130525120005） liugaochang@gjzq.com.cn
分析师：陈芷婧（执业 S1130525120008） chenzhijing@gjzq.com.cn
分析师：鲍淑娴（执业 S1130526020002） baoshuxian@gjzq.com.cn

再谈 CPU 涨价能持续多久？

本周观点

- 1 月我们发布行业点评报告《CPU 涨价能持续多久？》，率先揭示 Agent 对 CPU 的刚性需求以及 CPU 供需失衡全面爆发。过去 4 个月，CPU 逻辑持续得到强化：1）AMD、Arm 等头部厂商大幅上修服务器 CPU TAM，Meta、AWS 等大厂加码 CPU 部署，需求侧持续验证；2）Intel、AMD 库存趋紧、交期拉长并持续涨价，景气度不断提升；3）同时，NVIDIA 新一代 Vera RuBin AI 系统中 CPU/GPU 配比持续抬升，CPU 正从辅助角色重新回到 AI 系统核心。
- **Agent 对 CPU 刚性需求，CPU 重回核心地位。**随着大模型从 Chatbot 向 Agent 演进，计算负载重心正发生偏移。Agent 不仅需要 GPU 进行模型推理，更依赖高性能 CPU 处理复杂逻辑编排、工具调用与内存管理。我们认为，Agent 对 CPU 的刚性需求主要来自三方面：1）Multi-Agent 架构带来的 OS 调度压力，以及沙盒环境创建、调度与销毁对 CPU 算力的持续消耗；2）长上下文场景下 KV Cache 卸载对 CPU 内存与带宽提出更高要求；3）高并发工具调用带来的大量 CPU 算力消耗。Intel 论文显示，多数 Agent 工作负载中，CPU 耗时占端到端延迟比例可达 40%-90%。伴随 Agent 数量、任务复杂度与 Token 消耗指数级增长，CPU 产业已进入新一轮景气周期，Intel、AMD 服务器 CPU 库存趋紧、交期延长，并于 2026 年以来持续推进涨价。
- **CPU TAM 扩容，CPU/GPU 部署比例抬升。**1）TAM 来看，AMD 与 Arm 均大幅上修服务器 CPU 市场空间，预计 2030 年全球服务器 CPU TAM 将超过 1000 亿美元。根据 AMD，CPU 需求可分为通用计算 CPU、AI 头节点 CPU 以及 Agentic AI CPU 三部分，其中智能体 AI 相关需求是最大的增量来源。2）配比来看，AI 数据中心 CPU/GPU 部署比例正从传统 HGX 时代的 1:4、1:8，逐步向 1:2、1:1 甚至更高演进。以 NVIDIA 为例，GB300 NVL72 已实现 72 颗 GPU 搭配 36 颗 Grace CPU 的 1:2 配比，Vera Rubin 进一步通过外挂独立 Vera CPU 机柜，使整体 CPU 配比继续抬升。
- **所有 CPU 架构均受益，ARM 中期变化更显著。**1）ARM 架构低功耗、高核心密度的特性更契合 Agent 工作负载。相比 x86，ARM 在高并发、低功耗场景下具备更优能效比与扩展能力，尤其适合海量 API 调用、KV Cache 调度等轻计算、高并发任务。2）ARM 开放授权生态亦高度契合云厂商自主构建 AI 基础设施的需求，当前 AWS Graviton、NVIDIA Grace、微软 Cobalt 等方案均已加速落地。ARM 在 FY26Q4 业绩会上预计，到 2030 年按 CPU 类型划分的最大市场份额将属于 Arm 架构。
- **Agentic AI 驱动 CPU 重构，全球厂商开启新一轮架构升级。**1）海外方面，Intel、AMD、Arm、NVIDIA 等均围绕高核心密度、异构协同与能效优化展开新一轮产品迭代，CPU 竞争正从单纯性能竞争迈向系统级算效竞争；2）国内方面，海光、飞腾、龙芯、华为海思、熠知电子等厂商在 x86、ARM 与自主指令集方向持续突破，核心数、线程数、内存带宽与生态能力快速提升。伴随 Agentic AI 带来的 CPU 需求爆发，以及自主可控趋势深化，国产 CPU 有望迎来规模化替代与产业地位重估。

相关标的

CPU：Intel、海光信息、禾盛新材、高通、AMD、澜起科技、中科曙光、中国长城、龙芯中科、广合科技、兴森科技、深南电路、宏和科技等。

海外算力：中际旭创、东山精密、胜宏科技、欧科亿、天孚通信、天岳先进、新易盛、工业富联、兆易创新、大普微、源杰科技、景旺电子、英维克、唯科科技、领益智造等；Lumentum、闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

风险提示

- 行业竞争加剧的风险；技术研发进度不及预期的风险；特定行业下游资本开支周期性波动的风险。

内容目录

一、CPU 何以重回核心地位？ 4
1.1 三大逻辑揭示 Agent 对 CPU 的刚性需求 4
1.2 Agent 生态扩张引爆 CPU 性能瓶颈，CPU 现涨价潮 4
二、CPU TAM 扩容，CPU/GPU 部署比例抬升 6
2.1 CPU TAM 翻倍上修，预计到 2030 年 TAM 超 1000 亿美元 6
2.2 CPU 与 GPU 配比有望提升至 1:1 甚至更多 7
三、所有 CPU 架构均受益，ARM 中期变化更显著 11
3.1 ARM 架构低功耗、高核心密度的特性更契合 Agent 工作负载 11
3.2 ARM 开放生态更契合大厂需求，NV Grace 率先跑通 12
四、Agentic AI 驱动 CPU 重构，全球厂商开启新一轮架构升级 14
4.1 海外：x86 与 ARM 路线竞逐，性能、能效与异构协同全面升级 14
4.2 国内：自主架构加速突破，国产 CPU 迈向规模化替代 17
五、相关标的 20
风险提示 21

图表目录

图表 1： KV Cache 卸载使得 KV Cache 能够从有限的 GPU 内存中传输到更大且性价比更高的存储 4
图表 2： 未来 5 年全球活跃 Agent 数据将呈现爆发式增长 5
图表 3： 大多数 Agentic 工作负载场景下，CPU 端到端延迟占比显著高于 GPU 5
图表 4： 五大代表性 Agent 工作负载中的任务延迟分布 5
图表 5： 五大代表性 Agent 工作负载中的任务延迟分布 6
图表 6： Meta 宣布将部署 Graviton CPU 以满足 Agentic AI 需求 6
图表 7： CPU 收入增长率将在 2028 年前超过 GPU 和 XPU 的增长 7
图表 8： 大量的代理工作负载导致 CPU 过载 7
图表 9： 对 CPU 承担编排、代理、调度的需求持续提升 7
图表 10： 数据中心 CPU 核心数将持续攀升 8
图表 11： NVIDIA DGX H100(640 GB)/H200(1,128 GB)系统组件描述 9
图表 12： DGX H100/200 系统拓扑结构 9
图表 13： GB200 NVL72 规格 10
图表 14： 英伟达的 Grace CPU 连接 10
图表 15： 通过 NVLink 连接的 Grace Hopper 超级芯片进行内存访问 10
图表 16： 英伟达 Vera Rubin NVL72 机架系统 CPU 与 GPU 搭载比例为 1:2 11

图表 17： Vera Rubin NVLink C2C 架构图 11
图表 18： x86 及 ARM 架构特性对比 11
图表 19： 2029 年基于 ARM 架构的 CPU 有望占据定制 AI ASIC 服务器主机 CPU 市场 90%的份额 12
图表 20： 数据中心的电力消耗持续大幅增加 12
图表 21： Grace CPU 在图形分析中能效提高 3 倍 13
图表 22： Grace CPU 在数据分析中能效提高 2.1 倍 13
图表 23： AWS Graviton CPU 13
图表 24： 微软 Cobalt 200 布局图 14
图表 25： Xeon6+ E 能效核架构 14
图表 26： Xeon6+为首款 18A 数据中心 CPU 14
图表 27： 同频率下 Intel 18A 较 Intel 3 功耗降低 36%-38% 15
图表 28： EPYC 9005 持续引领 x86 架构服务器 CPU 性能标准 15
图表 29： 第六代 EPYC Venice 性能再度飞跃 16
图表 30： Arm AGI CPU 规格情况 16
图表 31： Arm 商业模式从 IP 授权拓展为 IP 授权+计算子系统（CSS）授权+自研芯片 17
图表 32： 鲲鹏 920 核心参数 17
图表 33： 海光 CPU 架构持续迭代升级，最新第五代处理器将通过 SMT-4 技术实现单核 4 线程并发 18
图表 34： 熠知电子三代产品介绍 18
图表 35： 熠知电子 TF7000 系列高性能核心处理器芯片参数展示 18
图表 36： 飞腾腾云 S5000C 系列参数介绍 19
图表 37： 飞腾腾云 S2500 参数介绍 19
图表 38： 龙芯自主指令系统 LoongArch 19
图表 39： 龙芯 3C6000 包括 Q、S、D 三个版本 19
图表 40： 海内外主流服务器 CPU 参数对比 20

一、CPU 何以重回核心地位?

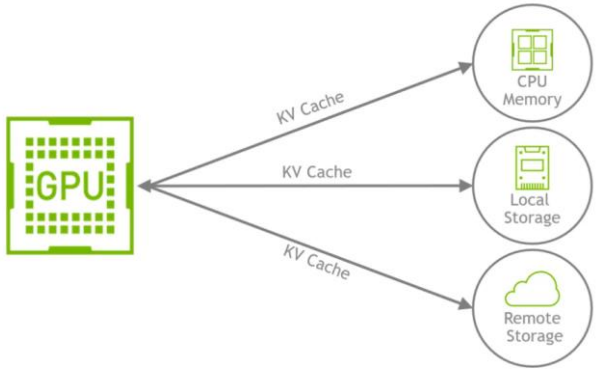
1 月我们发布行业点评报告《CPU 涨价能持续多久？》，率先揭示 Agent 对 CPU 的刚性需求以及 CPU 供需失衡全面爆发。过去 4 个月，CPU 逻辑持续得到强化：1）AMD、Arm 等头部厂商大幅上修服务器 CPU TAM，Meta、AWS 等大厂加码 CPU 部署，需求侧持续验证；2）Intel、AMD 库存趋紧、交期拉长并持续涨价，景气度不断提升；3）同时，NVIDIA 新一代 Vera RuBin AI 系统中 CPU/GPU 配比持续抬升，CPU 正从辅助角色重新回到 AI 系统核心。

1.1 三大逻辑揭示 Agent 对 CPU 的刚性需求

随着大模型的应用从简单的 Chatbot 向能完成复杂任务的 Agent 演进，计算负载的重心正在发生微妙的偏移。Agent 不仅需要 GPU 进行模型推理，更依赖高性能 CPU 来处理复杂的逻辑编排、工具调用和内存管理。我们认为，Agent 对 CPU 的刚性需求基于以下三大逻辑：

- Chatbot 向 Agent 演进，Multi-Agent 架构引发 OS 调度压力，计算负载重心正从 GPU 侧向 CPU 侧偏移。Agent 工作流的“推理-执行-评估-反思”循环机制，在生成 Token 之外持续进行逻辑判断与状态管理，“思考”和“行动”的频繁切换显著加剧操作系统的上下文切换与进程调度压力。与此同时，Agent 执行代码等操作须在隔离沙盒中运行，沙盒环境的创建、调度与销毁全程依赖 CPU 算力，进一步推高 CPU 侧的工作负载。
- 长上下文场景下 KV Cache 卸载对 CPU 构成挑战。KV Cache 在加速 Transformer 推理的同时，带来了显著的显存消耗问题，以 8 万 Token 的上下文为例，KV Cache 本身即可消耗数十 GB 显存，叠加模型权重与中间激活值后，HBM 资源极易触及上限。对此，业界提出将不活跃的 KV Cache 卸载至 CPU 内存或 SSD，以解决 HBM 瓶颈。但 CPU 与 GPU 之间的通信带宽远低于 GPU 内部的 HBM 带宽，数据搬运本身存在明显瓶颈；同时，在进行 KV Cache 传输和管理时，也需要 CPU 进行任务的调度，进一步加剧 CPU 的负载。
- 高并发工具调用带来巨大的 CPU 算力消耗。Agent 的能力不仅在于对话，更在于使用工具，例如检索、写代码、浏览网，这类非推理任务的计算负担主要由 CPU 承担。在高并发场景下大量 Agent 同时工作，多线程/多进程调度需求集中爆发，对 CPU 的性能提出更高要求。

图表1：KV Cache 卸载使得 KV Cache 能够从有限的 GPU 内存中传输到更大且性价比更高的存储



来源：Nvidia 官网，国金证券研究所

1.2 Agent 生态扩张引爆 CPU 性能瓶颈，CPU 现涨价潮

Agent 生态正发生指数级扩张。据 IDC 预计，活跃 Agent 的数量将从 2025 年的约 2860 万，快速攀升至 2030 年的 22.16 亿；同时，年执行任务数将从 2025 年的 440 亿次暴涨至 2030 年的 415 万亿次，Agent 数量跃升、任务复杂度与推理深度的指数级提升情况下，年度 Token 消耗将从 2025 年的 0.0005 PetaTokens 暴增至 2030 年的 152,667 PetaTokens，年复合增长率高达 3418%。

图表2：未来 5 年全球活跃 Agent 数据将呈现爆发式增长

来源：IDC 咨询微信公众号，国金证券研究所

Agent 工作负载驱动 CPU 从配角变为核心。Intel 论文《A CPU-CENTRIC PERSPECTIVE ON AGENTIC AI》对五类主流 Agent 负载时延情况进行测试，结果显示，执行流程中 CPU 耗时占端到端延迟的比例为 40%-90%，在 Haystack RAG 任务中，GPU 侧推理耗时仅 0.8-1.1 秒，CPU 侧 ENNS 检索耗时高达 6.0-8.0 秒，CPU 相关时延占比最高达到 90.6%。Agentic 场景中的任务规划、工作流执行、工具调用、在子智能体之间传递数据等关键环节均依赖 CPU 进行调度，Agent 数量暴增将急剧推高 CPU 侧工作负载，其累积时延将主导系统的整体耗时。

图表3：大多数 Agentic 工作负载场景下，CPU 端到端延迟占比显著高于 GPU

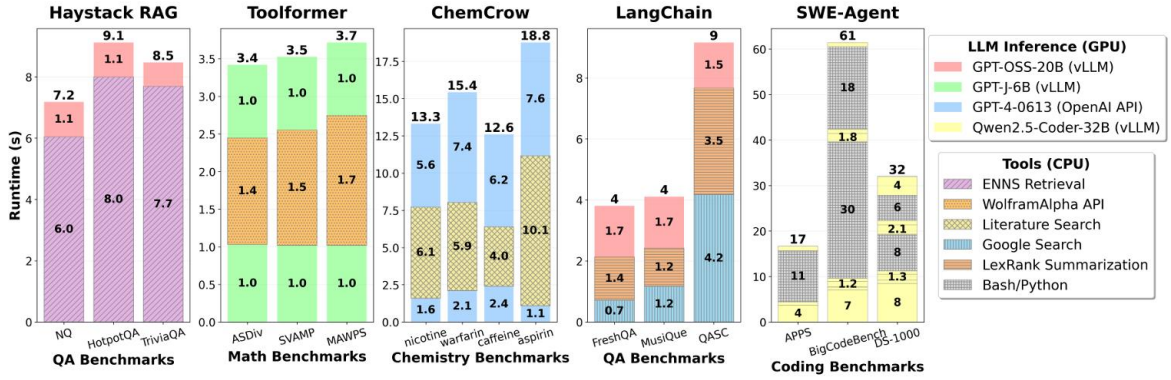


Figure 2. (a) Haystack with ENNS retrieval on QA benchmarks (b) Toolformer with WolframAlpha API on Math benchmarks (c) Chemcrow with literature (Arxiv/Pubmed) search tool on Chemistry benchmarks (d) Langchain with web search and LexRank summarization tools on QA benchmarks (e) Mini-SWE-Agent with bash/Python execution tools on coding benchmarks

来源：《A CPU-Centric Perspective on Agentic AI》，Ritik Raj, Hong Wang, Tushar Krishna，国金证券研究所

在 GPT-OSS-20B 模型的吞吐量测试中，随着 BatchSize 增加，不同 Agentic 工作负载的吞吐量增长逐渐放缓并趋于饱和。以 Langchain 为例，延迟情况从 BatchSize 为 64 时的 2.9 秒大幅上升至 BatchSize 为 128 时的 6.3 秒，LLM 推理延迟同期从 2.6 秒上升至 3.9 秒，可见高并发条件下存在严重的 CPU 上下文切换瓶颈，成为系统延迟的重要因素。

图表4：五大代表性 Agent 工作负载中的任务延迟分布

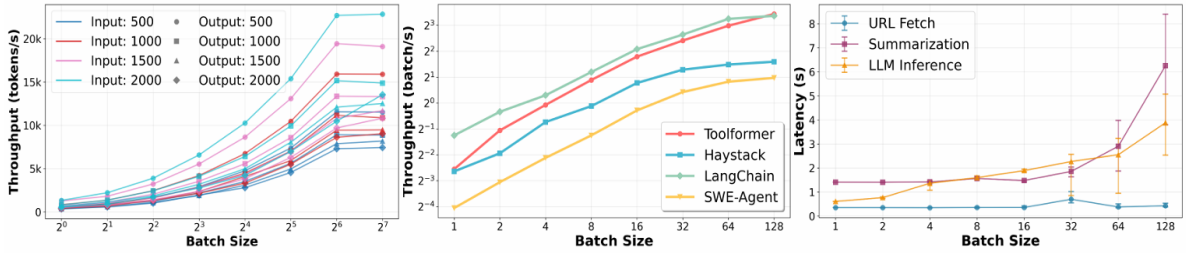


Figure 4. (a) vLLM throughput saturation for GPT-OSS-20B model (b) Throughput saturation for various agentic workloads (c) Average time taken by different components in Langchain benchmark showing a critical CPU context switching bottleneck at batch size 128

来源：《A CPU-Centric Perspective on Agentic AI》，Ritik Raj, Hong Wang, Tushar Krishna，国金证券研究所

高并发 Agent 负载下，CPU 动态能耗占比加速攀升。LangChain 工作负载下，当 Batch Size 从 1 增加到 128，系统总动态能耗增长 38.1 倍，CPU 动态能耗激增 86.7 倍；分 Batch Size 大小情况来看，在处理低 Batch Size（1-4）时，GPU 能耗显著高于 CPU；随着 Batch Size 增加到 128，CPU 的能耗（1807 Joules）已经接近 GPU（2307 Joules），占比高达 44%，可见在大批量处理场景下，CPU 能耗占比格外显著，从辅助算力转变为核心算力单元。

图表5：五大代表性 Agent 工作负载中的任务延迟分布

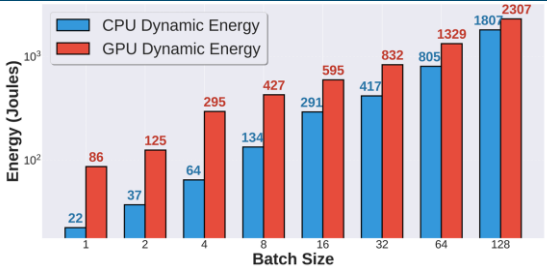
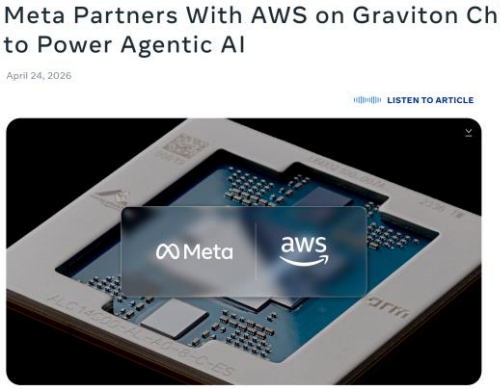


Figure 5. CPU (AMD Threadripper) and GPU (Nvidia B200) dynamic energy consumption for Langchain workload

来源：《A CPU-Centric Perspective on Agentic AI》，Ritik Raj, Hong Wang, Tushar Krishna，国金证券研究所

大厂的 AI 基础设施布局开始向 CPU 侧倾斜。2026 年 2 月 17 日，Meta 宣布大规模部署英伟达 Grace CPU，并计划于 2027 年推进 Vera CPU 规模化落地；4 月 24 日，Meta 进一步与 AWS 签署多年期协议，将部署数千万颗 Graviton CPU 核心，以应对 Agentic AI 工作负载中 CPU 密集型的任务调度、编排等需求。除此之外，谷歌于 4 月 9 日宣布与英特尔就至强 Xeon CPU 展开多年期合作，共同推进下一代 AI 及云基础设施的建设。Meta 成为首家在数据中心单独部署 Grace CPU 的公司、头部厂商通过长期协议锁定 CPU 供应，种种信号均印证 CPU 正重回 AI 算力主舞台。

图表6：Meta 宣布将部署 Graviton CPU 以满足 Agentic AI 需求



来源：Meta 官网，国金证券研究所

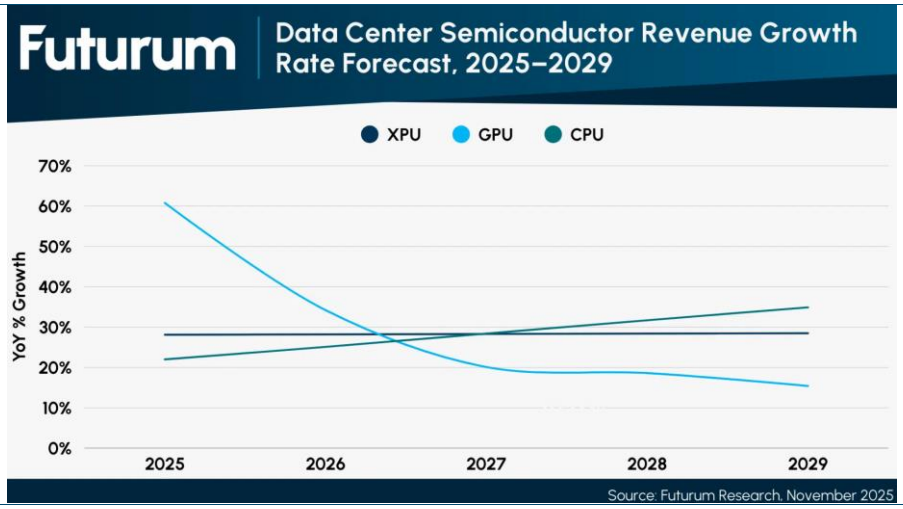
CPU 短缺加剧，涨价周期来临。25 年 10 月，据外媒 TrendForce 报道，英特尔公司正计划对其第 13 代 Raptor Lake 和第 14 代 Raptor Lake Refresh 处理器进行价格调整，涨幅最高可达 10%。26 年 1 月，据外媒 Wccftech 报道，AMD 和英特尔今年各自的服务器 CPU 库存均已售罄，大部分需求来自超大规模企业，他们希望将最新的服务器 CPU 集成到现有机架架构中，这也是过去几个季度需求显著增长的原因，因此，据称 AMD 和英特尔都计划将服务器 CPU 价格提高多达 15%，以确保供应保持稳定。据日经亚洲 26 年 3 月 25 日报道，英特尔与 AMD 已各自通知客户，将分别于 3 月和 4 月起上调全系列 CPU 价格，平均涨幅达 10-15%，部分产品涨幅更高；同时，交货周期将从之前的 1-2 周大幅延长至 8-12 周，个别情况下甚至将长达 6 个月。AI 算力需求爆炸式增长，AI 芯片巨头占用大量原材料与产能，英特尔与 AMD 面临产能扩张瓶颈，叠加原材料价格上涨，CPU 供给端持续承压，供需错配加剧，CPU 价格进入上行通道。

二、CPU TAM 扩容，CPU/GPU 部署比例抬升

2.1 CPU TAM 翻倍上修，预计到 2030 年 TAM 超 1000 亿美元

CPU TAM 有望翻倍提升：AMD/ARM 指引 2030 年 CPU TAM 超 1000 亿美元。1）在 AMD FY26Q1 业绩会上，AMD CEO 表示，受智能体 AI 需求拉动，预计服务器 CPU 的总潜在市场（TAM）将以每年超过 35%的速度增长，并将 2030 年服务器 CPU TAM 翻倍上调至 1200 亿美元，因需要大量 CPU 用于任务编排、数据移动和并行执行。2）根据 ARM FY26Q4 业绩会，公司首款 AGI CPU 发布仅六周，客户需求就从 10 亿美元激增翻番至 20 亿美元；随着 AI Agent 的扩展，数据中心将需要超过目前 4 倍的 CPU 容量，到 2030 年将创造一个超过 1000 亿美元的数据中心 CPU 市场机会；且芯片 ASP 会随核心数的增加显著提升。据 Futurum 预测，CPU 收入增长率将在 2028 年前超过 GPU 和 XPU 的增长，CPU 市场的潜在规模与增速巨大。

图表7：CPU 收入增长率将在 2028 年前超过 GPU 和 XPU 的增长

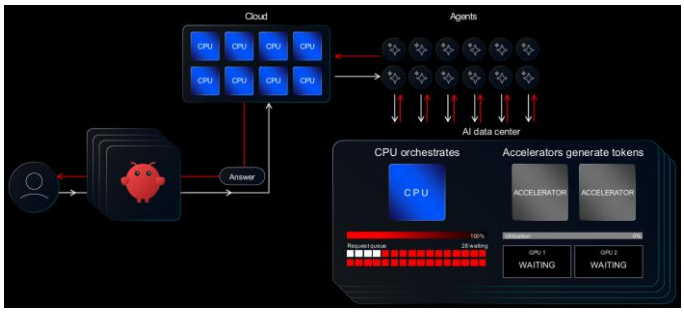


来源：Futurum，国金证券研究所

2.2 CPU 与 GPU 配比有望提升至 1:1 甚至更多

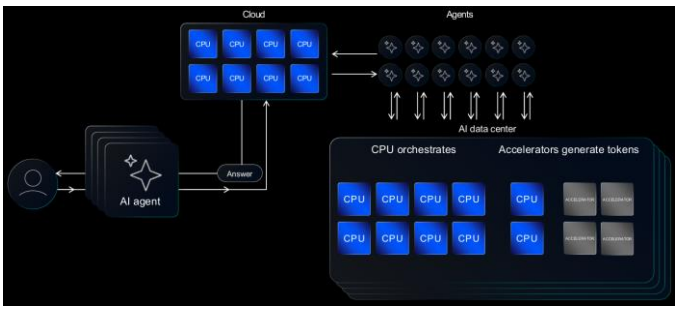
Agentic 时代 CPU 的部署比例向 1:2/1:1 甚至更高演进。Intel、ADM、ARM 等头部 CPU 厂商均对 CPU 的搭载配比进行乐观展望：1）FY26Q1 财报业绩会上，ADM CEO 表示过去以主机节点模式为主，CPU 与 GPU 配比约为 1:4 或 1:8，现在正在向接近 1:1 的方向转变，如果智能体数量大量增加，甚至可以想象 CPU 数量超过 GPU 的场景；2）FY26Q1 财报业绩会上，Intel CEO 表示数据中心 CPU 核心数量正在大幅增加，目前 CPU 与 GPU 的配比是 1:4，未来将达到 1:1 均衡，甚至发生逆转。AI 推理任务中对 CPU 编排、调度、内存管理等工作的需求显著加大，CPU 与 GPU 配比的重构将直接拉高数据中心对 CPU 的需求规模。3）ARM 在 FY26Q4 业绩会上表示，从芯片颗数来看，CPU 数量超过 GPU 未必会发生，但从核心数来看，则很可能实现；传统数据中心每吉瓦仅需 3000 万颗 CPU 核心，Agentic AI 时代 CPU 的需求将激增至 1.2 亿颗，增幅达 4 倍。4）同时，数据中心的对 CPU 的电力分配也将随之提升，Futurum 基于 ARM 预测数据中心每吉瓦 CPU 核心数的 4 倍增幅，并结合 ARM AGI CPU 服务器约 36 kW 的功耗估算，CPU 与 GPU 服务器的比例将接近 7:1，数据中心的电力分配发生反转，大部分电力将重新分配至 CPU。

图表8：大量的代理工作负载导致 CPU 过载



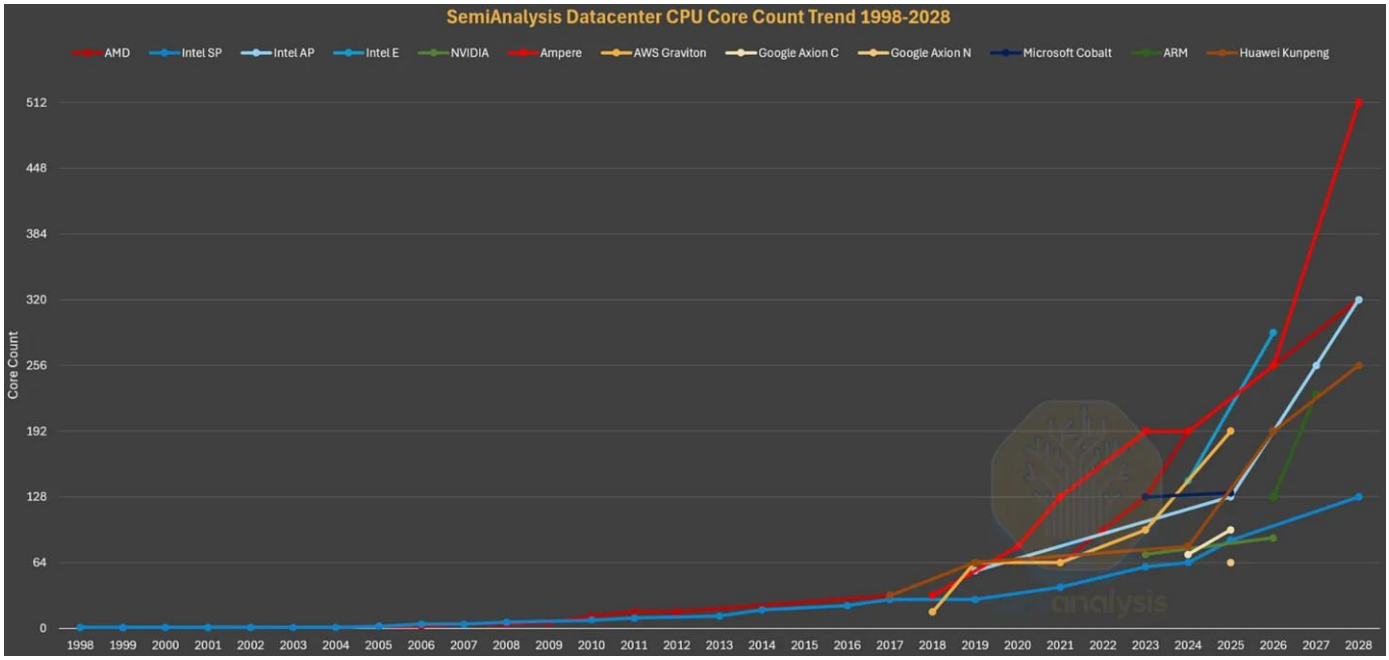
来源：ARM CEO keynote，国金证券研究所

图表9：对 CPU 承担编排、代理、调度的需求持续提升



来源：ARM CEO keynote，国金证券研究所

图表10：数据中心CPU核心数将持续攀升



来源：SemiAnalysis，国金证券研究所

根据AMD，CPU需求可以分为三类：1）通用计算CPU TAM，增速相对较低，约为低两位数；2）与加速器连接的AI头节点CPU，增速也在增长但规模较小。头节点中CPU的作用是管理连接的GPU，并持续为其提供数据。为了尽可能降低尾部延迟，需要具备大容量缓存、高带宽内存和IO的高性能单核。NVIDIA Grace的设计便旨在实现GPU的一致性内存访问，以便将CPU内存用作模型上下文键值缓存（KV Cache）的扩展，这需要极高的CPU到GPU带宽。对于头节点，每个计算节点中通常由1颗CPU搭配2或4颗GPU，如NV Vera Rubin每个超级芯片包含1颗Vera CPU和2颗Rubin GPU。3）智能体AI的CPU需求，是增量最大的部分，如NVIDIA引入的Vera CPU专用机柜架构。

我们以NV机柜CPU:GPU配比的演进为例：1）HGX时代：标配多为单路/双路CPU带8卡（配比1:8或1:4）。2）GB300 NVL72集成72颗NVIDIA Blackwell Ultra GPU和36颗基于ARM架构的NVIDIA Grace CPU，配比为1:2。3）Vera Rubin NVL72集成72颗Rubin GPU和36颗Vera CPU，配比维持1:2，并通过在集群中外挂纯Vera CPU算力柜作为专属的Agent并发调度节点，整体计算集群的CPU：GPU配比向更高比例演进。

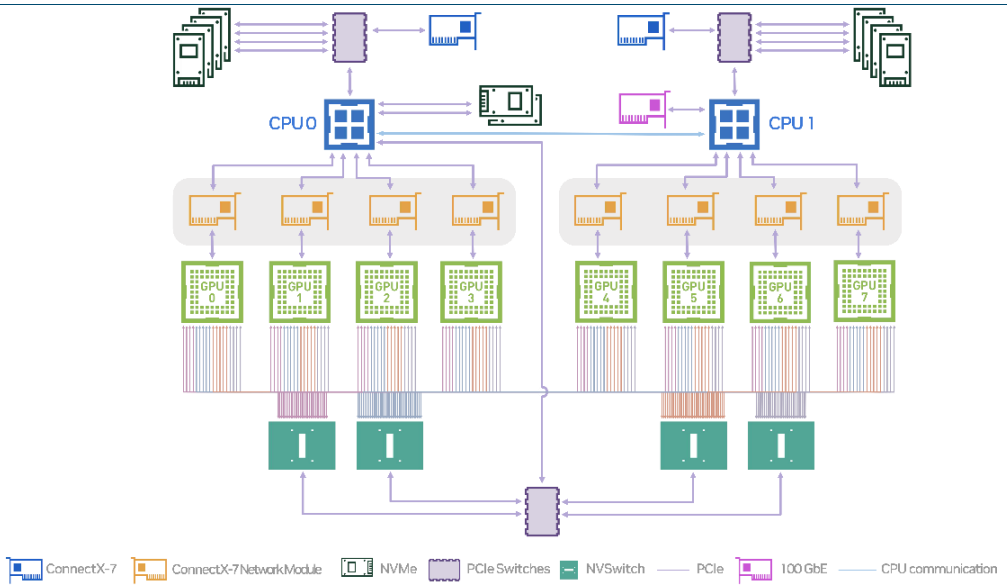
DGX H100/H200架构：在DGX H100/H200这一代架构中，CPU与GPU之间仍主要基于PCIe构建异构计算架构。1）系统逻辑拓扑：系统采用双路x86 CPU架构，配置2颗Intel Xeon Platinum 8480C CPU（总核心数56/总线程数112），并连接8颗NVIDIA H100/H200 GPU，CPU与GPU数量配比约为1:4。2）连接中枢(PCIe Switches)：根据NVIDIA DGX H100官方架构图，系统采用PCIe Gen5 Switch构建CPU与GPU间的PCIe拓扑连接，多个GPU通过PCIe Switch接入双路CPU平台，CPU与GPU之间的数据交换主要依赖PCIe Gen5 x16互联。3）互联带宽：PCIe Gen5 x16的理论双向汇总带宽约为128GB/s，而GPU之间通过NVLink可实现最高900GB/s GPU-to-GPU带宽，GPU内部与GPU间的数据吞吐能力已显著高于传统CPU-GPU PCIe互联带宽。4）存储层次：系统配置2TB DDR5系统内存；GPU侧方面，H100配置80GB HBM3显存，H200进一步升级至141GB HBM3e显存，并将显存带宽提升至4.8TB/s。

图表11：NVIDIA DGX H100(640 GB)/H200(1,128 GB)系统组件描述

成分	描述
GPU	H100：配备 8 个 NVIDIA H100 GPU，提供总计 640 GB 的 GPU 显存。 H200显卡：配备8颗NVIDIA H200 GPU，总显存容量为1128GB。
中央处理器	2 个 Intel Xeon 8480C PCIe Gen5 CPU，每个 CPU 有 56 个核心，频率分别为 2.0/2.9/3.8 GHz（基础频率/全核睿频/最大睿频）
NVSwitch	4 个第四代 NVLink 接口，提供 900 GB/s 的 GPU 间带宽。
存储（操作系统）	2 个 1.92 TB NVMe M.2 SSD（每个）组成 RAID 1 阵列
存储（数据缓存）	8 块 3.84 TB NVMe U.2 单固态硬盘 (每块) 组成 RAID 0 阵列
网络（集群）卡	4 个 OSFP 端口，可连接 8 张 NVIDIA® ConnectX®-7 单端口 InfiniBand 卡 每张卡提供以下速度： • InfiniBand（默认）：最高可达 400Gbps • 以太网：400GbE、200GbE、100GbE、50GbE、40GbE、25GbE 和 10GbE
网络（存储和带内管理）卡	2 张 NVIDIA® ConnectX®-7 双端口以太网卡 每张卡提供以下速度： • 以太网（默认）：400GbE、200GbE、100GbE、50GbE、40GbE、25GbE 和 10GbE • InfiniBand：最高可达 400Gbps
系统内存（DIMM）	使用 32 个 DIMM 内存条，容量为 2 TB
BMC（带外系统管理）	1 GbE RJ45接口 支持 Redfish、IPMI、SNMP、KVM 和 Web 用户界面
系统管理接口	插槽 3 中配备双端口 100GbE 和一个 10GbE RJ45 接口
电源	6 x 3.3 千瓦

来源：NVIDIA 官网，国金证券研究所

图表12：DGX H100/200 系统拓扑结构



来源：NVIDIA 官网，国金证券研究所

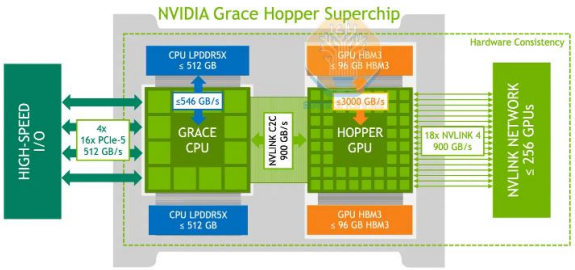
GB200/300 NVL72：NVLink-C2C 推动 CPU-GPU 从 PCIe 异构互联向 cache-coherent 紧耦合架构演进，机柜级 CPU/GPU 配比提升至 1:2。在传统 HGX 架构下，CPU 与 GPU 之间主要通过 PCIe Gen5 进行互联，其带宽显著低于 GPU 内部及 GPU 之间的数据吞吐能力，因此 CPU 更多承担主机处理器、系统调度、IO 管理及运行时调度等职责，GPU 则负责主要 AI 计算任务。进入 Blackwell 时代后，NVIDIA 在 GB300 NVL72 中进一步引入 NVLink-C2C 一致性互联架构，并在机柜级构建 CPU-GPU 紧耦合异构计算系统。1）系统架构：根据 NVIDIA 官方架构，GB300 NVL72 采用全液冷整机柜设计，集成 36 颗基于 ARM 架构的 Grace CPU（72 核，基于 Arm Neoverse V2 架构）与 72 颗 Blackwell Ultra GPU，实现机柜级 1:2 的 CPU/GPU 物理配比。2）在互联架构方面，CPU 与 GPU 之间通过 NVLink-C2C 实现最高 900GB/s 的一致性互联带宽，比 PCIe Gen5 x16 通道高出 7 倍。3）存储层次：以 GB200 为例，GPU 侧配置总计 372GB HBM3e 显存，CPU 侧配置 480GB LPDDR5X 内存，得益于一致性内存架构，GPU 能够以 NVLink-C2C 高效访问 Grace CPU 侧 LPDDR5X 内存，从而显著扩展统一内存容量，为长上下文、Agentic AI 及测试时扩展（Test-Time Scaling）等大内存场景提供支持。

图表13：GB200 NVL72 规格

	GB200 NVL72	GB200 Grace Blackwell 超级芯片
配置	36颗Grace CPU | 72颗Blackwell GPU	1. Grace CPU | 2. Blackwell GPU
NVFP4 张量核心[2]	1,440 | 720 PFLOPS	40 | 20 PFLOPS
FP8/FP6 张量核心[2]	720 PFLOPS	20 PFLOPS
INT8 张量核心[2]	720 POPS	20个流行音乐
FP16/BF16 张量核心[2]	360 PFLOPS	10 PFLOPS
TF32 Tensor Core [2]	180 PFLOPS	5 PFLOPS
FP32	5,760 TFLOPS	160 TFLOPS
FP64 / FP64 张量核心	2,880 TFLOPS	80 TFLOPS
GPU 显存 | 带宽	13.4 TB HBM3E | 576 TB/s	372 GB HBM3E | 16 TB/s
NVLink带宽	130 TB/s	3.6 TB/s
CPU核心数	2,592 个 Arm® Neoverse V2 内核	72个Arm Neoverse V2核心
CPU内存 | 带宽	17 TB LPDDR5X | 14 TB/s	最高支持 480 GB LPDDR5X | 最高支持 512 GB/s

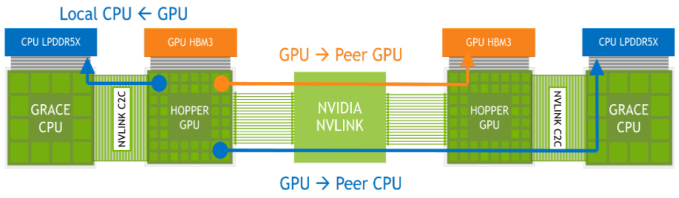
来源：NVIDIA 官网，国金证券研究所

图表14：英伟达的 Grace CPU 连接



来源：Semianalysis，国金证券研究所

图表15：通过 NVLink 连接的 Grace Hopper 超级芯片进行内存访问



来源：NVDIA，国金证券研究所

Vera Rubin：进一步推进数据中心架构向 rack-scale AI factory 演进，通过 GPU 计算机柜与 CPU 扩展机柜的分层设计，强化面向 Agentic AI 与测试时计的系统吞吐能力。1）系统架构：在核心计算单元上，Vera Rubin NVL72 机柜延续 Blackwell 时代的机柜级超级计算架构，由 72 颗 Rubin GPU 与 36 颗 Vera CPU（88 核/176 线程，基于定制的、兼容 ARM 的 NVIDIA Olympus 架构）构成标准配置，维持 CPU:GPU=1:2 的物理配比，其中 GPU 承担大规模矩阵计算与模型推理任务，而 CPU 则从传统主机管理角色进一步扩展至更高频的调度、数据预处理与 Agent 执行任务，从而提升系统整体并行效率；在系统扩展层面，NVIDIA 引入 Vera CPU 专用机柜（Vera CPU Rack），单机柜可集成多达 256 颗 Vera CPU，用于执行强化学习环境运行、Agent rollout、推理验证与非矩阵类计算任务，该设计使 CPU 资源从 GPU 计算柜中解耦出来，形成独立的 CPU 算力池，从而实现更灵活的工作负载分配与系统级扩展能力。2）在互联架构方面，NVLink-C2C 带宽进一步提高至 1.8 TB/s。3）存储层次：以 Vera Rubin 超级芯片为例，GPU 显存配置为 576 GB HBM4，CPU 内存配置为 1.5 TB LPDDR5X。

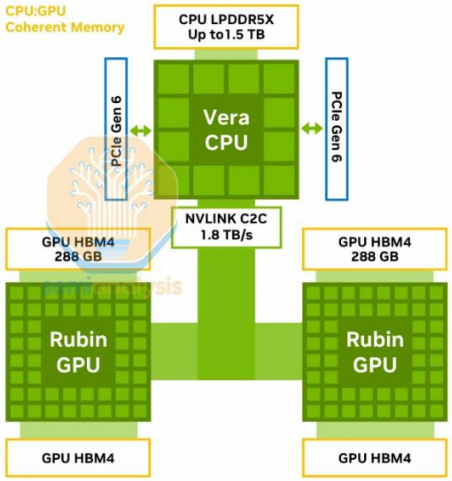
图表16：英伟达 Vera Rubin NVL72 机架系统 CPU 与 GPU 搭载比例为 1:2

NVIDIA Vera Rubin NVL72 Specs

	NVIDIA Vera Rubin NVL72	NVIDIA Vera Rubin Superchip	NVIDIA Rubin GPU
Configuration	72 NVIDIA Rubin GPUs | 36 NVIDIA Vera CPUs	2 NVIDIA Rubin GPUs | 1 NVIDIA Vera CPU	1 NVIDIA Rubin GPU
NVFP4 Inference	3,600 PFLOPS	100 PFLOPS	50 PFLOPS
NVFP4 Training²	2,520 PFLOPS	70 PFLOPS	35 PFLOPS
FP8/FP6 Training²	1,260 PFLOPS	35 PFLOPS	17.5 PFLOPS
INT8²	18 POPS	0.5 POPS	0.25 POPS
FP16/BF16²	288 PFLOPS	8 PFLOPS	4 PFLOPS

来源：NVIDIA，国金证券研究所

图表17：Vera Rubin NVLink C2C 架构图



来源：Semianalysis，国金证券研究所

三、所有 CPU 架构均受益，ARM 中期变化更显著

3.1 ARM 架构低功耗、高核心密度的特性更契合 Agent 工作负载

对比 x86 及 ARM 架构特性，x86 架构拥有极高的单线程性能，计算性能强，更擅长处理大规模复杂数据，但功耗偏高；ARM 架构的核心优势在于极致的能耗比与高核心密度。Agent 时代的工作负载（如海量 API 调用、Python 脚本解释、KV Cache 调度）一般表现为轻计算、高并发。ARM 架构能在极低的功耗下堆叠出海量的物理核心，非常适合高并发、低功耗场景。例如 NVIDIA 的 Vera 处理器，在极低功耗下实现了 88 核/176 线程，这种特性使其在处理高并发场景时不仅吞吐量惊人，还能大幅降低服务器节点的散热压力与能耗。

图表18：x86 及 ARM 架构特性对比

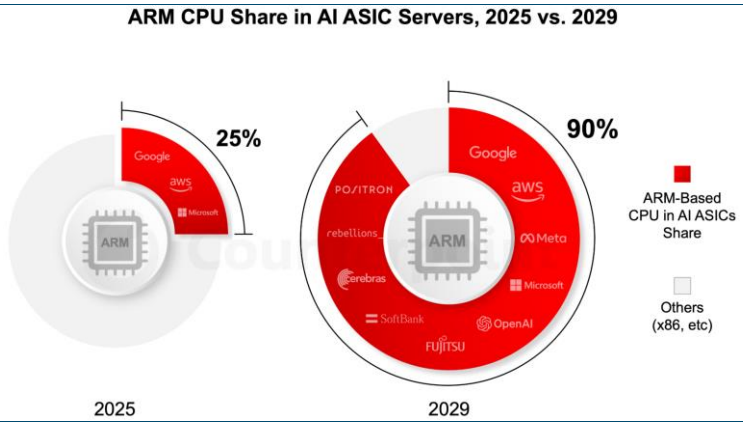
特性	x86 架构	ARM 架构
指令集类型	复杂指令集（CISC）	精简指令集（RISC）
解码复杂度	高，需要微码翻译	低，直接执行
性能表现	单线程性能方面表现出色，善于处理大规模的计算任务和数据处理需求，多用于高性能计算	单核性能相对较弱，密集型任务中表现稍差
功耗表现	功耗较高，部分服务器 CPU 功耗 OOW，需较强散热支持，产生高额的电费支出与运营成本	更低功耗、更高性能效率，能效比相比传统服务器提升 50%以上，适合大规模部署的云计算、边缘计算等场景，适合高密度部署
核心扩展能力	多核扩展受功耗和散热限制，通常单路服务器 CPU 核心数较少，且高核数下功耗和成本显著增加	多核并行和扩展能力强，适用于核心密度较高的环境，通过横向扩展提升算力
生态兼容性	软件生态覆盖 90%以上商用场景，主流数据库管理系统	主流 Linux 系统支持较好，部分传统闭源软件需重新编译或适配，

	（MySQL、PostgreSQL、SQL Server）都在 x86 架构上得到全面优化和支持	逐步适配 Windows on ARM
定制化能力	定制化能力较弱，厂商难以根据自身特定需求进行深度定制化开发	厂商可根据自身需求定制，满足特定应用场景的优化需求
服务器领域	占据传统服务器主流，生态成熟	单核面积小，功耗比领先，可“堆核”提升性能
代表产品	英特尔至强（Xeon）系列、AMD 霄龙（EPYC）系列	AWS Graviton 系列、英伟达 Grace、鲲鹏 920

来源：CSDN，SEMICONDUCTOR ENGINEERING，国金证券研究所

ARM 份额指引乐观。据 ARM FY26Q4 业绩会表述，Trainium、TPU、英伟达加速器中绝大部分市场份额将是 Arm，到 2030 年按 CPU 类型划分的最大市场份额将属于 Arm。Counterpoint 预测，2029 年基于 ARM 架构的 CPU 有望占据定制 AI ASIC 服务器主机 CPU 市场 90%的份额，而 x86 和 RISC-V 架构合计仅占约 10%。

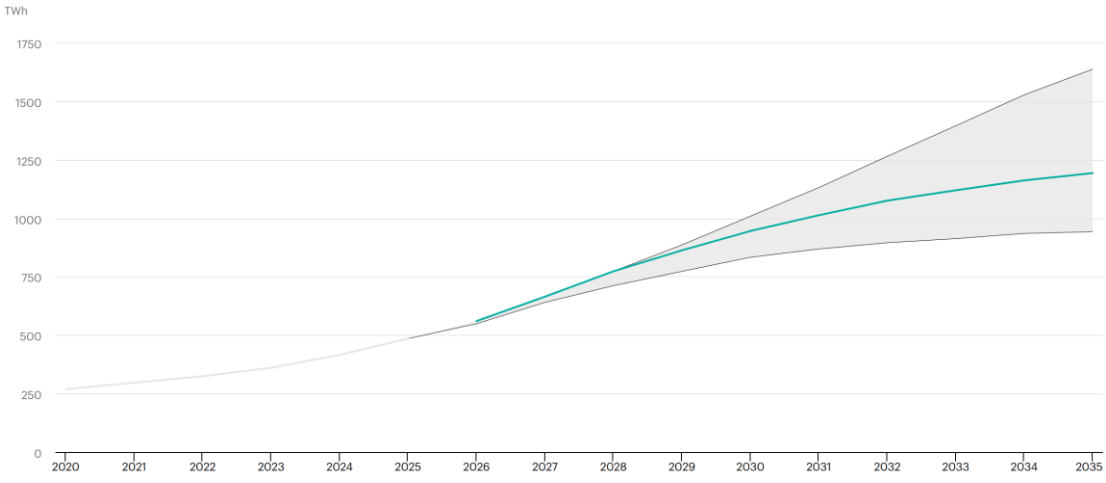
图表19：2029 年基于 ARM 架构的 CPU 有望占据定制 AI ASIC 服务器主机 CPU 市场 90%的份额



来源：Counterpoint，国金证券研究所

Agent 呈现能源密集型特征，CPU 能效比成为关键。1）数据中心用电需求翻倍：据 IEA，2025 年全球数据中心电力需求增长 17%，其中 AI 驱动的数据中心电力消耗增长高达 50%，2030 年全球数据中心用电量预计将从 2025 年的 485TWh 近乎翻倍至 950TWh。2）多步骤执行、工具交互等推理过程对功耗密度的需求大幅增加：传统数据中心主要服务于企业 IT、互联网服务、数据存储等基础业务，单机架功率密度仅为 5-15kW；AI 数据中心聚焦模型训练、深度学习推理、智能计算等任务，对功率密度的要求显著提升，据 AFCOM，2025 年数据中心平均机架功率密度达到 27kW，同比大增 69%，高并发、长序列处理场景不断增加，未来或迎来百瓦、千瓦级的需求，并对电网、UPS、液冷系统等带来挑战。3）电力供应的增长速度低于算力需求的膨胀，功耗低的服务器占优：Agentic AI 部署的 Token 消耗量是标准生成式 AI 的 20-30 倍，受限于因电网容量与成本等刚性因素，传统堆卡模式难以为继，CPU 的能效比变得至关重要。

图表20：数据中心的电力消耗持续大幅增加



来源：IEA，国金证券研究所

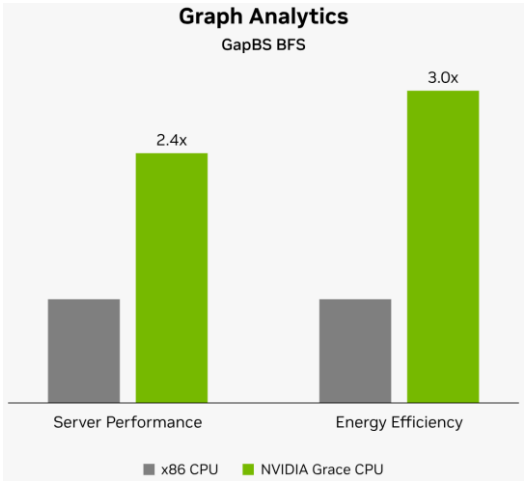
3.2 ARM 开放生态更契合大厂需求，NV Grace 率先跑通

除了物理性能的契合，ARM 架构份额提升的另一大核心驱动力在于其商业授权模式。传统 x86 生态处于授权封闭状态，客户高度依赖 Intel/AMD 少数巨头，不仅面临高昂的采购溢价，且定制化差，客户自主权小。而在当前的算力军备竞

赛中，各大头部云厂商（CSP）为了实现底层算力解绑并追求极致的 TCO（总拥有成本），亟需打造差异化的算力底座。ARM 的开放 IP 授权模式完美迎合了这一战略诉求。通过获取 ARM 授权，亚马逊、谷歌、微软等 CSP 能够根据自身云原生需求自主定制专属 CPU，降低 TCO。

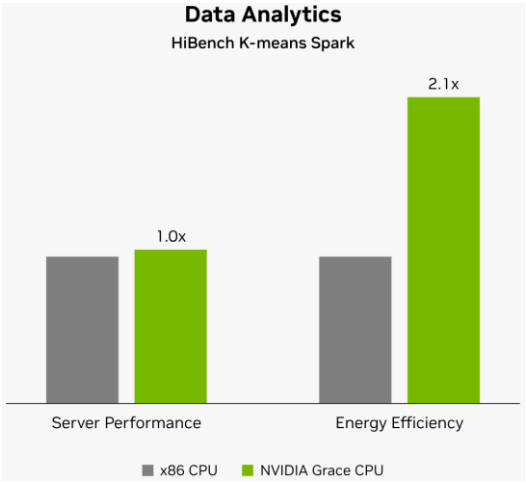
Grace CPU 能效倍升，利于高密度部署。1）能效优势显著：据 NVIDIA 官网，Grace CPU 可将数据中心的输出能力提高一倍、能耗降低一半，与 x86 CPU 相比，图形分析中服务器性能提升 2.4 倍、能效提高 3 倍，数据分析中能效提高 2.1 倍，天气模拟场景在相同功耗下可完成约 2 倍工作量，极大提高数据中心吞吐量。2）低功耗限制下仍保持高性能：Grace CPU 在 200W 功耗限制下可保持超过 90%性能，150W 下仍可保持 80%性能，进而在功率受限的环境中，保证不牺牲计算性能的同时，实现机架密度最大化。

图表21：Grace CPU 在图形分析中能效提高 3 倍



来源：NVIDIA 官网，国金证券研究所

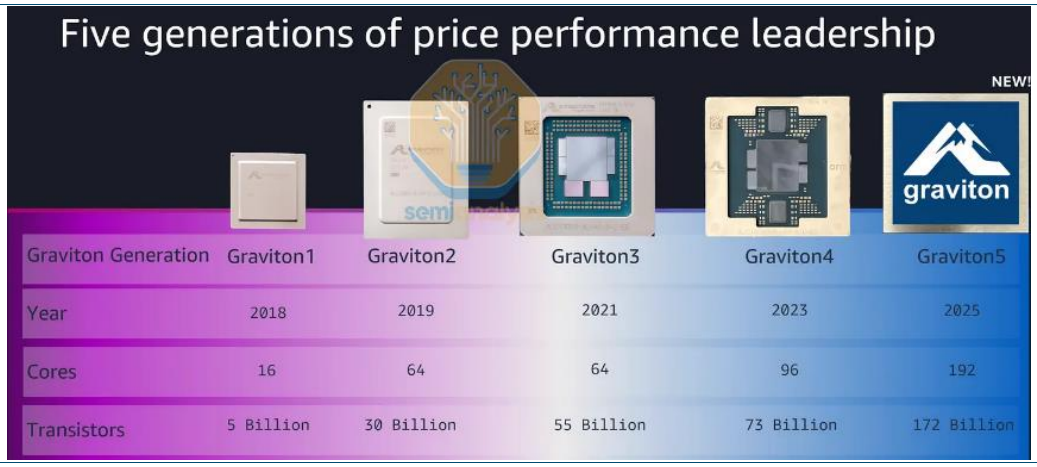
图表22：Grace CPU 在数据分析中能效提高 2.1 倍



来源：NVIDIA 官网，国金证券研究所

AWS Graviton5：AWS 是首家成功为云端开发并部署自研 CPU 的超大规模云服务商，Graviton5 自 2025 年 12 月开始预览，拥有 192 个 NeoverseV3 核心，并在台积电 3nm 工艺上集成了 1720 亿个晶体管。在 CPU 使用方面，AWS 已在内部 CI/CD 设计集成流程中使用了数千颗 Graviton CPU，其 Trainium3 加速器现在将使用 Graviton CPU 作为头节点，配比为 1 颗 CPU 对应 4 颗 XPU，初始版本运行在 Graviton4 上，未来的 Trainium3 集群将由 Graviton5 提供动力。

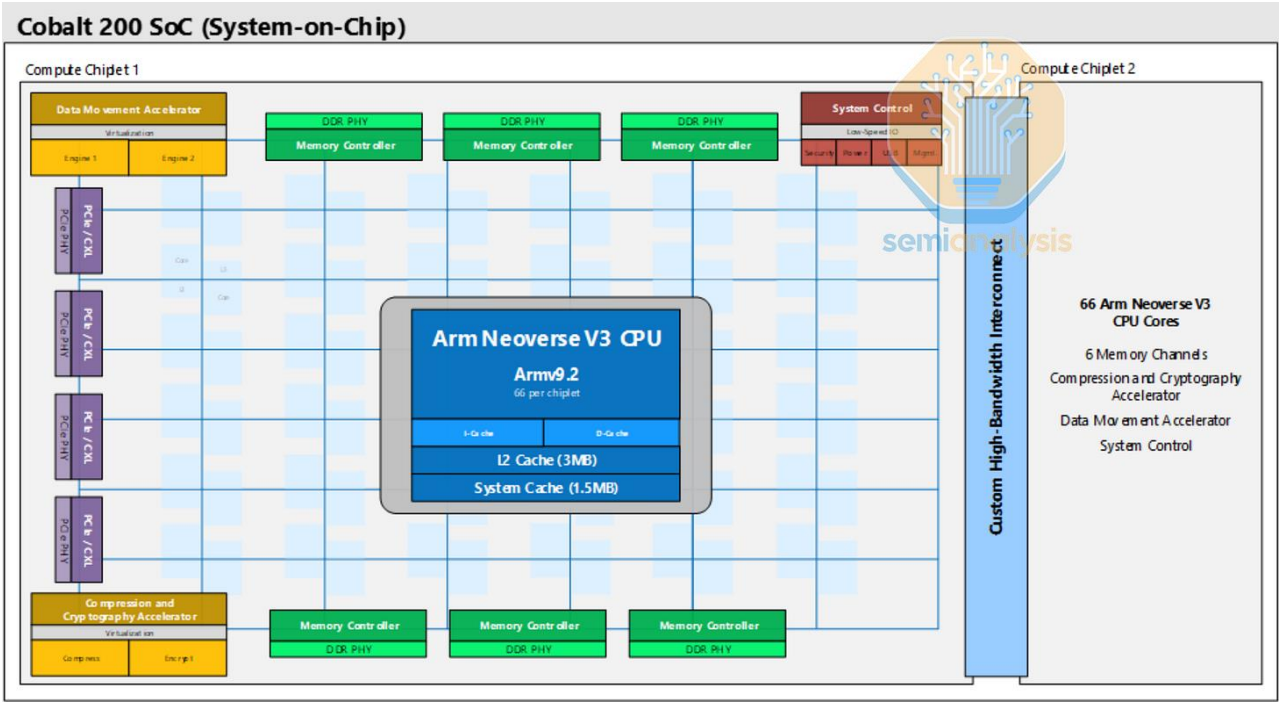
图表23：AWS Graviton CPU



来源：Semianalysis，国金证券研究所

微软 Cobalt 200：Cobalt200 于 2025 年底发布，核心数量从 128 个增加到 132 个，采用 NeoverseV3 设计，每个核心的性能都大幅提升。每个核心拥有极大的 3MBL2 缓存，并通过标准的 ARM Neoverse CMNS3 片上网络连接，跨越两个台积电 3nm 计算芯片(compute dies)，芯片间采用定制的高带宽互连。Cobalt200 将仅用于 Azure 的通用 CPU 计算服务，而不会被用作 AI 头节点，微软的 Maia200 机架级系统转而采用了英特尔的 Granite Rapids CPU。

图表24：微软 Cobalt 200 布局图



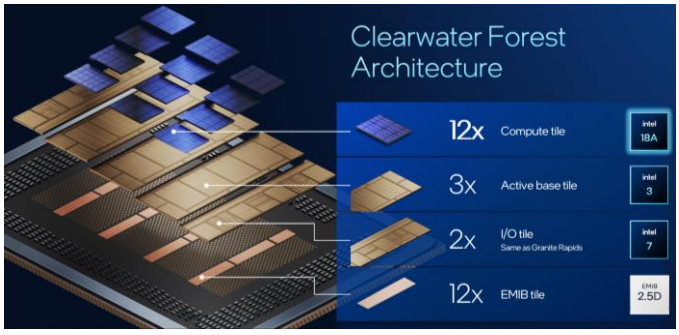
来源：Semianalysis，国金证券研究所

四、Agentic AI 驱动 CPU 重构，全球厂商开启新一轮架构升级

4.1 海外：x86 与 ARM 路线竞逐，性能、能效与异构协同全面升级

Intel：x86 架构传统龙头乘势再起，Xeon6+性能、能效、集成度、跨越式提升。Xeon 6+采用 Chiplets 设计，封装中集成了 12 个 Intel 18A 工艺的计算模块、3 个 Intel 3 工艺的有源基础模块、2 个 Intel 7 工艺的 I/O 模块、12 个 EMIB 2.5D 连接封装模块；计算模块内部分为 6 个模组，每个都包含 4 个 Darkmont 架构的 E 核，总计 288 个核心。英特尔技术专家指出，在整体负载占比不同的情况下，至强 6+处理器较上一代 Sierra Forest 可以带来 1.9 倍以上的性能提升，同时在整体负载范围之内带来高达 23%的能效提升，达到 8:1 服务器整合的效果。

图表25：Xeon6+ E 能效核架构



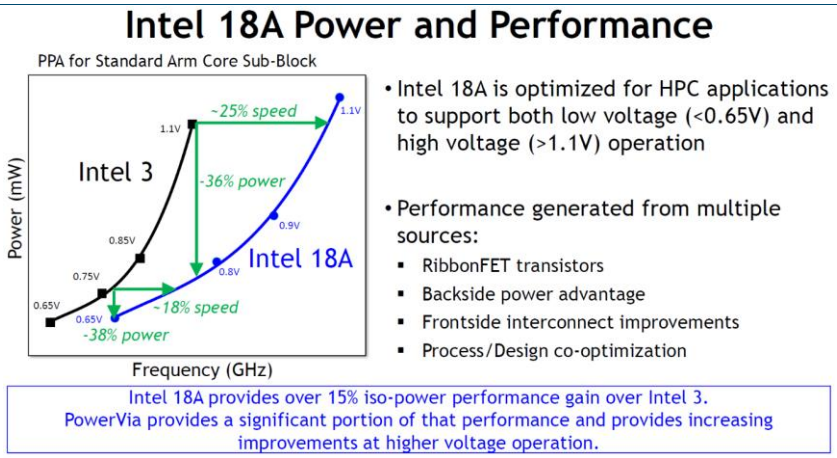
来源：Intel，国金证券研究所

图表26：Xeon6+为首款 18A 数据中心 CPU

来源：Intel，国金证券研究所

Xeon6+量产在即，Intel 18A 工艺重构能效表现。Xeon6+ E 能效核（Clearwater Forest）预计于 2026H1 量产发布，Xeon6+ P 性能核（Diamond Rapids）有望于 2027 年量产，二者均基于 Intel 18A 最先进制程，工艺的良率已达大规模量产标准，且仍在持续优化，目前已向首批客户交付量产晶圆。Intel 18A 通过多方面的技术提升显著弥补 x86 架构 CPU 的能耗短板、并强化 x86 的单线程性能优势，相同频率下，较 Intel 3 功耗降低 36%-38%，等效功率性能提升 15%。1）采用 RibbonFET 全环绕栅极晶体管架构，晶体管漏电率降低 50%，同功耗下开关频率提升 15%；2）通过 PowerVia 背面供电技术，实现供电与信号线路的物理分离，将芯片电阻压降减少 40%、互连线性能提升 30%、同性能下功耗降低 15%，同时为正面晶体管布局释放了超 20%的空间。

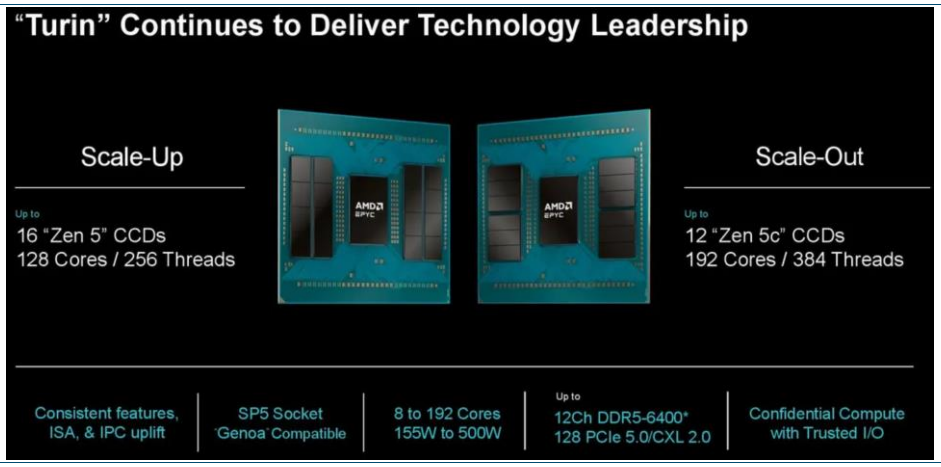
图表27：同频率下 Intel 18A 较 Intel 3 功耗降低 36%-38%



来源：DonanimHaber，国金证券研究所

AMD：EPYC 9005 性能领先，Agentic AI 全栈优化。AMD 服务器 CPU 系列产品已迭代至第五代（EPYC 9005），Zen 5 架构相比上一代 Zen 4 实现 16%的 IPC 提升。EPYC 9965 作为第五代系列中的高性能旗舰产品，采用 Zen 5c 核心，相比 Zen 5 核心物理布局更紧凑、每瓦性能更高，单颗集成 192 核 384 线程，支持 12 通道 DDR5-6400 内存与 128 条 PCIe 5.0 通道。EPYC 9965 在主机 CPU 与 CPU 侧推理两类核心场景中具备全栈优势，据 AMD 官网，EPYC 9965 性能提升 29%、能效提升 66%、端到端 AI 性能提升 70%、机器学习性能提升高达 93%；在 LLM 推理场景中表现卓越，相较 Intel Xeon 6980P，EPYC 9965 在 Llama 3.1 88 模型中处理性能领先 33%、GPT-J 6B 中汇总场景的吞吐量性能提升 28%、Llama 3.2 1B 中应用场景转换性能提升 36%。

图表28：EPYC 9005 持续引领 x86 架构服务器 CPU 性能标准



来源：Tom's Hardware，国金证券研究所

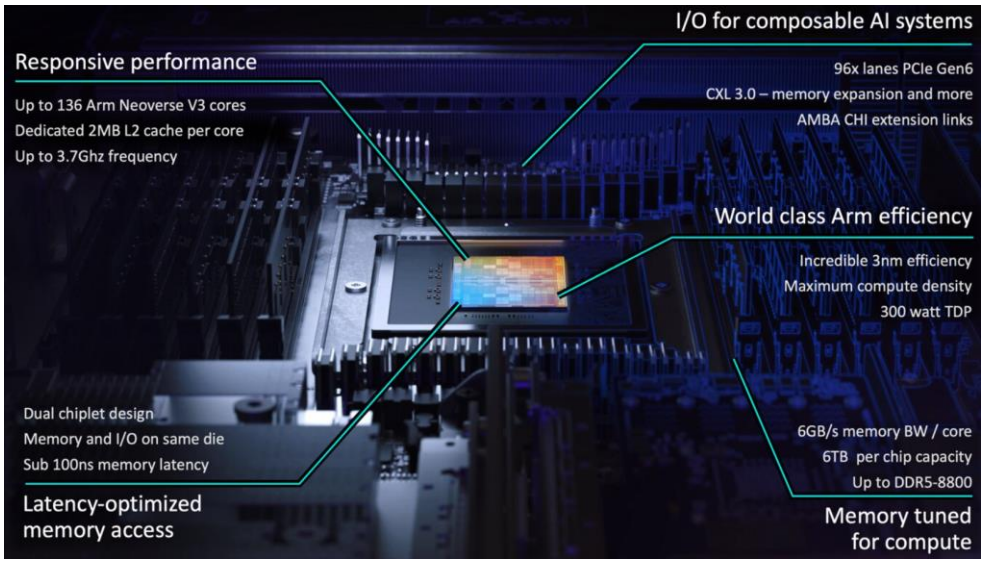
第六代 EPYC Venice 综合性能再进阶，或成为驱动公司达成 50%+CPU 市占的王牌。AMD 计划于 2026H2 发布第六代服务器 CPU（EPYC Venice），首次采用台积电 2nm 工艺及 Zen 6/Zen 6c 核心架构，集成 256 个核心和 512 个线程，相较第五代产品核心数增加 33%，线程密度提升 30%，整体性能与能效提升 70%。Venice 在吞吐量、功耗、成本和 AI 基础设施等方面进行全方位优化，以极致产品力巩固服务器 CPU 领域领导地位，AMD CEO 于 26Q1 财报会上表示，相较市场上其他 x86 产品，Venice 每插槽及每瓦性能大幅提升；相较市场领先的 Arm 产品，Venice 每插槽吞吐量提升逾 2 倍；客户对于 Venice 需求强劲，处于验证和爬坡平台阶段的客户数量超过了以往任何一代 EPYC，对实现超过 50%的市场份额目标充满信心。

图表29：第六代 EPYC Venice 性能再度飞跃

来源：Wccftech，国金证券研究所

Arm：AGI CPU 面向 Agentic AI，专为高密度机架部署设计。Arm AGI CPU 单颗芯片最多集成 136 个 Neoverse V3 内核，每个内核可获得 6 GB/s 内存带宽，并支持 DDR5-8800 规格与低于 100ns 的访问延迟。参考服务器配置采用 1OU 双节点设计，每块刀片板集成两颗芯片，并配备独立的内存与 I/O，单刀片合计提供 272 个计算核心，这些刀片可完整填充标准风冷 36kW 机架，共 30 片刀片、总计 8,160 个核心，支撑高密度、低延迟的下一代 AI 计算系统；此外，Arm 还与 Supermicro 联合推出液冷 200kW 机架方案，可容纳 336 颗 Arm AGI CPU，总核心数超过 45,000 个。

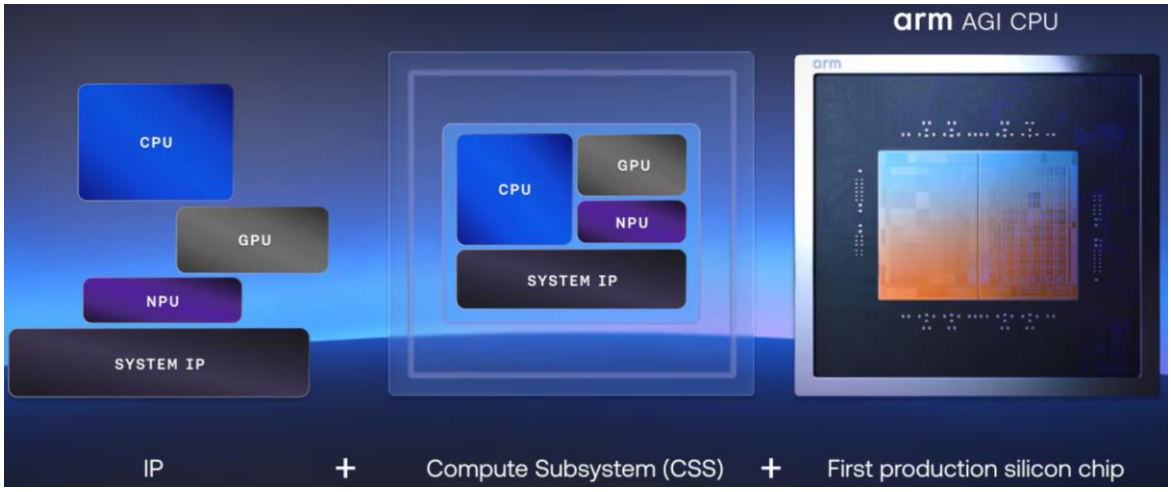
图表30：Arm AGI CPU 规格情况



来源：ARM 官网，国金证券研究所

Arm AGI CPU 的核心密度优势实质为结构性 TCO 优势。Arm AGI CPU 在 300W TDP 下提供 136 个核心，每瓦约提供 0.45 个核心，相比之下 AMD 500W 的 192 核 EPYC 每瓦约为 0.38 个核心，英特尔 500W 的 144 核 Xeon 每瓦约为 0.29 个核心，在电力和冷却占总运营成本 30–40%的数据中心中，高核心数密度可转化为更高的工作负载密度、更充分的加速器利用率，以及在既有功耗预算内更有效的计算能力释放，进而带来成本的极大节约；据 Arm 官网，Arm AGI CPU 方案在单机架性能上可达到传统 x86 平台的两倍以上水平，每 GW 数据中心容量可节省 100 亿美元的资本支出。

图表31：Arm 商业模式从 IP 授权拓展为 IP 授权+计算子系统（CSS）授权+自研芯片



来源：快科技，国金证券研究所

4.2 国内：自主架构加速突破，国产 CPU 迈向规模化替代

华为海思：鲲鹏 920 是目前业界领先的 ARM-based 处理器。该处理器采用 7nm 制造工艺，基于 ARM 架构授权，由华为公司自主设计完成。通过优化分支预测算法、提升运算单元数量、改进内存子系统架构等一系列微架构设计，大幅提高处理器性能。典型主频下， SPECint Benchmark 评分超过 930，超出业界标杆 25%。同时，能效比优于业界标杆 30%。鲲鹏 920 以更低功耗为数据中心提供更强性能。

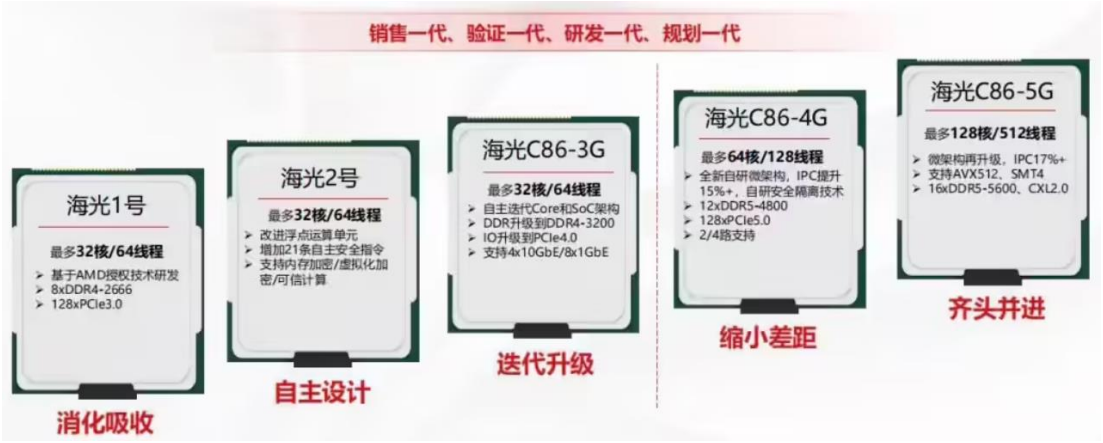
图表32：鲲鹏 920 核心参数

Architecture	• ARM v8.2
Core	• up to 64
Typical Frequency	• 2.6 GHz / 3.0 GHz
Memory	• 8 DDR4 Channels
Coherent Interconnect	• 2S&4S
I/O	• PCIe 4.0, CCIX, 100G, SAS/SATA 3.0
Max Power	• 180W
Process	• 7nm

来源：华为海思，国金证券研究所

海光信息：国产 x86 服务器 CPU 领军者，系统架构、可靠性、安全性优异。海光信息为国内少数实现成熟商用的 x86 通用处理器的公司，面向数据中心、行业客户、高性价比场景的实际需求，细分为海光 7000 系列产品、海光 5000 系列产品、海光 3000 系列产品，产品矩阵覆盖全面，具备高计算和扩展能力；依托先进的 SoC 架构和片上网络，集成更多处理器核心，性能优势显著，已广泛应用于电信、金融、互联网、教育、交通等重要行业或领域。其中，面向数据中心的旗舰级高性能处理器 700 系列，集成 16-32 核心，支持 128 路 PCIe 通道，8 个 DDR4 内存通道，并针对数据中心云计算中心等进行了功耗优化。2016 年公司基于 AMD 授权技术启动 x86 架构海光 C86-1G CPU 产品设计，目前海光 C86-4G CPU 已实现商业化应用，公司于 2025 年 5 月发布 C86-5G 技术路线图，最高 128 核、512 线程，与前代相比核心数量增加 1 倍，线程数量增加 4 倍，每周期指令数（IPC）提升超过 17%，代际迭代升级显著。

图表33：海光 CPU 架构持续迭代升级，最新第五代处理器将通过 SMT-4 技术实现单核 4 线程并发



来源：半导纵横，国金证券研究所

禾盛新材：战略入股熠知电子押注国产 ARM CPU 先机，TF9000 性能全面提升。2025 年 8 月禾盛新材以自有资金或自筹资金 2.5 亿元向向熠知电子投资，2026 年 4 月公司再次以自有资金或自筹资金 2.33 亿元增资熠知电子，持有熠知电子 17.05%股权。熠知电子已完成三代 ARM 处理器芯片迭代升级，原有产品线包括一代 TF16000 系列、二代 TF7000 系列融合处理器及相应板卡。公司于 2026 年 1 月发布第三代 TF9000 系列融合处理器及板卡，采用 Armv9 架构，对标英伟达 Grace 系列 CPU，相较第二代产品 TF7000 系列实现了核心性能 30%的提升、成本降低 30%，以及内存带宽、PCIe 5.0 带宽和内存总容量分别提升了 200%、100%和 300%，主力面向通用云计算、大模型一体化等 AI 智算领域，兼具高性能与高性价比。熠知电子商业与生态双重验证充分，TF7000 系列、TF9000 系列处理器及相关服务器产品，已成功支持国内互联网大厂、运营商、制造业、金融机构、政府、公安等核心客户，进入规模化商用阶段。在生态协同方面，熠知电子已与国内多家主流 GPU 厂商完成产品深度适配，广泛落地算力服务器、大模型一体机等多元场景。

图表34：熠知电子三代产品介绍

产品类型	CPU 架构	NPU 架构	产品特征	应用场景
9000 系列融合处理器及板卡	ARMV9	TFMX	拟推出的第三代 CPU+NPU 混合算力芯片，主力优化于通用云计算，大模型一体化等 AI 智算领域。是一款具有高性能、超高性价比的算力芯片。	互联网，大模型一体机，工厂智能化等
7000 系列融合处理器及板卡	ARMV8.2	TFACC2.0	内置多个处理器核心，集成通用的高性能外设接口，拥有完善的软硬件生态环境和完备的系统安全机制，适用于数据计算和事务处理等通用型应用。	AI、云计算、物联网、信息服务等
异构 AI 处理器及硬件	ARM V8	TFACC1.0	依托标的公司自研的高性能 ManyCoreTM 深度学习运算加速引擎和高性能 CPU，能够以优异的功耗表现从容应对复杂的运算任务。	视觉 AI，边缘计算

来源：公司公告，国金证券研究所

图表35：熠知电子 TF7000 系列高性能核心处理器芯片参数展示

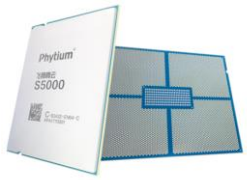
来源：公司官网，国金证券研究所

中国长城：飞腾信息第一大股东，腾云 S5000C-E 产品力升级。截至 2025 年 12 月，中国长城持有国产 ARM 服务器 CPU 龙头飞腾信息 28.04%股权，为其第一大股东。飞腾信息面向高性能服务器领域打造飞腾腾云 S 系列服务器 CPU，

已包括 S2500、S5000C、S5000C-E 系列产品，提供强算力、高并发的计算能力。飞腾腾云 S5000C-E 为最新服务器 CPU 旗舰产品，集成 80 个飞腾自研高性能处理器内核，计算性能相比 S5000C-64 提升 50%以上，可轻松应对金融交易、风控建模等复杂计算场景。2025 年，公司基于腾云 S5000C-E（80 核）研制新一代服务器，性能实测比肩业内高端产品，实现 DeepSeek 大模型与国产算力深度耦合，成功交付多款 AI 训推一体机。

信创领域优势显著，连续斩获多个核心大单。飞腾服务器 CPU 产品基于 ARMv8 架构，深度适配麒麟、统信等主流信创操作系统，具备自主可控、高性能、高安全、生态兼容等优势。受益于国产化替代政策深化落地，飞腾信息近期连续中标金融、电信、政务等核心行业大规模集采订单，2025 年 11 月，飞腾腾云双路服务器中标安平行业某省级客户项目，中标数量超百台；2025 年 12 月，飞腾腾云 S5000C-M CPU 独家中标中国移动 2025-2026 年 5G 扩展型皮基站集采 8000 片，为国产 CPU 在 5G 基站核心计算单元的首次规模化商用；2026 年 4 月基于飞腾腾云 S5000C、S5000C-E 的双路高性能服务器中标政策性银行数百台采购项目，订单放量势头保持强劲。

图表36：飞腾腾云 S5000C 系列参数介绍

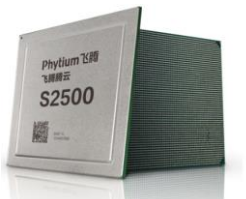


飞腾腾云S5000C系列

- 飞腾自主FTC862处理器核
- 16核、32核、64核设计
- 主频2.1GHz
- 多路支持：支持2~8路，最多128核心互联
- PSPA 1.0 安全架构

来源：公司官网，国金证券研究所

图表37：飞腾腾云 S2500 参数介绍



飞腾腾云S2500

- 飞腾自主FTC663处理器核
- 64核设计
- 主频2.1GHz
- 多路支持：2~8路，最多512核心互联

来源：公司官网，国金证券研究所

龙芯中科：3C6000 对标市场主流，强势开拓智算市场。面向服务器市场，公司已有龙芯 3C5000、龙芯 3D5000、龙芯 3C6000 等系列产品，龙芯 3C6000 基于公司自研指令集系统 LoongArch，最高支持 128 核 256 线程，性能相比上代 3C5000 系列成倍提升，对标英特尔第三代至强可扩展处理器，产品代差缩小至两代；对标 NVLink 通过龙链技术实现片间互连，大幅降低延迟，提高带宽效率。目前，龙芯服务器 CPU 已实现多个智算场景应用案例，如郑州港区新建的自主智算中心已有超过 500 台服务器全部采用龙芯 3C6000/D 服务器芯片。据龙芯中科董事长胡伟武在 25&26Q1 业绩会上表示，随着服务器 CPU 的推出，公司与抖音、阿里、腾讯等互联网大厂的适配合作更加紧密，每个万卡集群项目可采用 2000-3000 片 32 核龙芯 3C6000/D CPU，3C6000 在 2025 年实现千片量级销售，2026 年应有数量级的提高。

全栈自主可控，LoongArch 生态持续开放扩圈。公司自研 LoongArch 指令集架构，为国内唯一独立于 X86 与 ARM 体系、构建完整自主指令系的 CPU 企业，该架构具备较好的自主性、先进性与兼容性，2020 年起公司新研 CPU 产品均支持 LoongArch，实现了从底层指令到硬件设计的全栈自主可控。为扩大龙架构生态覆盖，公司已于 2023 年 11 月将龙芯 CPU 核心 IP 开放授权给部分合作伙伴，支持其研制基于龙架构指令系统及龙芯 CPU 核心 IP 的芯片产品。2026 年 4 月，龙芯中科与海量数据联合调试再创佳绩，基于龙芯 3C6000/D 双路服务器（64 核/512GB）与海量数据库的全国产化方案性能较常规方案提升 30%，在 TPC-C 标准测试下每分钟新订单事务处理量突破 100 万，刷新自主指令集架构芯片与国产数据库协同优化记录，标志着 LoongArch 与国产数据库的融合能力迈入实用化新阶段。

图表38：龙芯自主指令系统 LoongArch

来源：公司官网，国金证券研究所

图表39：龙芯 3C6000 包括 Q、S、D 三个版本

来源：公司官网，国金证券研究所

全球服务器 CPU 正处于新一轮升级周期，海外巨头领衔多核高密度与异构迭代，国内厂商性能代差加速收窄。1）海外厂商方面，Intel Xeon7 Diamond Rapids 最高 512 核心、AMD EPYC Venice 迈入 2nm 节点、NVIDIA Vera CPU 性能翻倍功耗减半、Arm AGI CPU 多核支撑高密度机架部署，均展现极致的性能与能效，CPU 与其他计算单元深度耦合，算力调度由分离转向融合，极大释放 Agentic AI 场景的系统级算效。2）国内厂商方面，海光 C86-5G 核心数与线程数倍增、飞腾腾云 S5000C-E 计算性能较前代提升 50%以上、龙芯 3C6000 对标海外主流产品性能倍升，国产服务器 CPU 技术水平正快速成长，加速追赶海外顶尖产品标准，AI 算力高需求高景气叠加自主可控政策催化，国产 CPU 迎来规模化放量窗口期。

图表40：海内外主流服务器 CPU 参数对比

架构	厂商	CPU 型号	制程	核心数	线程数	内存	PCIe 通道数	TDP
x86	Intel	Xeon6+ Clearwater Forest	Intel 18A/ 1.8nm	最高 288	288	12 通道/DDR5-8000	96×PCIe 5.0	300-500W
		Xeon7 Diamond Rapids	Intel 18A/ 1.8nm	最高 512	-	16 通道 /DDR5 MRDIMM	-	-
	AMD	EPYC 9965	TSMC N3/ 3nm	192	384	12 通道/DDR5-6400	128×PCIe 5.0	500W
		EPYC Venice	TSMC N2P/ 2nm	最高 256	-	12 或 16 通道 /DDR5 MRDIMM 或 DDR5 MCRDIMM	PCIe 6.0	-
	海光信息	海光 500 系列（5380/5390）	-	16	32	4 通道/DDR4	64×PCIe 4.0	70/95W
		海光 700 系列（7360/7375/ 7380/7390）	-	24/32/32/32	48/64/64/64	8 通道/DDR4	128×PCIe 4.0	125/140/ 140/110W
ARM	NVIDIA	Grace CPU	TSMC 4N/ 4nm	72	144	LPDDR5X	68×PCIe 5.0	250W
		Vera CPU	TSMC N3/ 3nm	88	176	LPDDR5X	PCIe 6.0+ CXL 3.1	-
	ARM	AGI CPU	TSMC N3/ 3nm	最高 136	136	12 通道/DDR5-8800	96×PCIe 6.0	最高 300W
	华为	鲲鹏 920	7nm	最高 64	-	8 DDR4 Channels	PCIe 4.0	最高 180W
	禾盛新材	熠知 TF7000	-	40	-	4 通道/DDR4 3200 或 DDR5 4800	64×PCIe 4.0	-
	中国长城	飞腾腾云 S5000C-64/32/16	-	64/32/16	-	8/4/2 通道/ DDR5	96/80/48 ×PCIe 5.0	-
LoongArch	龙芯中科	龙芯 3C6000S/D/Q	-	16/32/64	32/64/128	4（S）或 8（D/Q）通道 /DDR4 3200	64（S）/128（D/Q）× PCIe 4.0	100-120W/ 180-200W/ 250-300W

来源：Intel，NVIDIA，ARM，AMD，华为海思官网，龙芯中科官网，海光信息官网，熠知电子官网，芯东西，硬件世界，半导体行业观察等，国金证券研究所

五、相关标的

CPU：Intel、海光信息、禾盛新材、高通、AMD、澜起科技、中科曙光、中国长城、龙芯中科、广合科技、兴森科技、深南电路、宏和科技等。

海外算力：中际旭创、东山精密、胜宏科技、欧科亿、天孚通信、天岳先进、新易盛、工业富联、兆易创新、大普微、源杰科技、景旺电子、英维克、唯科科技、领益智造等；Lumentum、闪迪、铠侠、美光、SK 海力士、中微公司、北方华创、拓荆科技、长川科技。

风险提示

■ 行业竞争加剧的风险：

在信创等政策持续加码支持计算机行业发展的背景下，众多新兴玩家参与到市场竞争之中，若市场竞争进一步加剧，竞争优势偏弱的企业或面临出清，某些中低端品类的毛利率或受到一定程度影响。

■ 技术研发进度不及预期的风险：

计算机行业技术开发需投入大量资源，如果相关厂商新品研发进程不及预期，表观层面将呈现出投入产出在较长时期的滞后特征。

■ 特定行业下游资本开支周期性波动的风险：

部分计算机公司系顺周期行业，下游资本开支波动与行业周期性相关性较强，或在个别年份对于上游软件厂商的营收表现产生扰动。

行业投资评级的说明：

买入：预期未来 3—6 个月内该行业上涨幅度超过大盘在 15%以上；

增持：预期未来 3—6 个月内该行业上涨幅度超过大盘在 5%—15%；

中性：预期未来 3—6 个月内该行业变动幅度相对大盘在 -5%—5%；

减持：预期未来 3—6 个月内该行业下跌幅度超过大盘在 5%以上。

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员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。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级（含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

上海
电话：021-80234211
邮箱：researchsh@gjzq.com.cn
邮编：201204
地址：上海浦东新区芳甸路 1088 号紫竹国际大厦 5 楼

北京
电话：010-85950438
邮箱：researchbj@gjzq.com.cn
邮编：100005
地址：北京市东城区建内大街 26 号新闻大厦 8 层南侧

深圳
电话：0755-86695353
邮箱：researchsz@gjzq.com.cn
邮编：518000
地址：深圳市福田区金田路 2028 号皇岗商务中心 18 楼 1806

【小程序】
国金证券研究服务

【公众号】
国金证券研究