



Center for Education,
Innovation and
Sustainable Development

上海科技大学教育、创新和可持续发展研究中心

Tech Note | 技术系列

总 No. 6 期/2026 年 9 月

AI 计算基础设施：物理约束与“端到端”架构演进

执行摘要

随着人工智能模型体系从单一大语言模型（LLM）向具有自主规划和执行能力的多智能体（Agentic AI）演进，底层计算基础设施正面临前所未有的范式重构。本报告基于 AMD 在上海科技大学信息科学与技术 2026 年度学术研讨会（ASSIST 2026）的报告及半导体行业最新披露的数据，对 AI 计算系统在芯片设计、供电架构、系统互联及软件生态层面的工程实践进行剖析和展望。

分析表明，当前数据中心的算力增长正受到“内存墙”、“互联墙”与“功耗墙”的三重物理制约。为跨越这些瓶颈，以 AMD 为代表的半导体企业采取了高度异构化的工程路径。在芯片微观架构上，通过引入基于铜-铜混合键合（Cu-Cu Hybrid Bonding）的 3.5D 先进封装技术以及集成电压调节器（IVR），实现了逻辑单元与存储体系在极低工作电压下的高密度集成。在宏观系统拓扑上，面向纵向扩展（Scale-up）的开源 Ultra Accelerator Link (UALink) 1.0 协议，通过定长流控单元设计，在带宽效率与交换机功耗上对传统以太网进行了深度优化，打破了专有协议在互联领域的长期垄断。此外，算力向边缘（Edge）与终端（Client）

的下沉，推动了以统一内存架构为核心的异构系统级芯片（SoC）的广泛部署。本报告旨在通过 AMD 的全栈技术解析，为学术界与工业界针对未来 AI 基础设施的架构设计与部署规划提供参考。

第一章 计算范式迁移与底层物理约束

当前数据中心的架构设计已不再单纯追求单一处理器核心的绝对浮点运算能力（Flops），而是转向在有限的功耗预算内，寻求计算引擎、内存带宽与系统互联三者之间的物理极值平衡。

1.1 智能体（Agentic AI）对计算架构的重塑

自 2020 年起，大模型训练对绝对算力的需求基本保持每年 5 倍的高速增长。然而至 2026 年，AI 工作负载的结构已发生实质性偏移，推理（Inference）端消耗的算力比重正式超越了训练端。这一交叉点的出现，从底层反映了 AI 应用正从纯粹的文本生成阶段，进入基于工作流（Workflow）的多智能体协作阶段。

在多智能体架构中，计算任务呈现出高度的非同质化。传统的 LLM 推理高度依赖 GPU 内部的矩阵乘法单元（Matrix Cores）；而在端到端的工作流加速中，系统除了需要处理高强度的推理运算外，还需应对海量的数据预处理、外部工具调用（Tool Invocation）、复杂逻辑编排（Coordination）以及多智能体间的任务调度。这类呈现高并发、强控制流特征的工作负载，其执行效率更依赖于 CPU 的标量性能与分支预测能力。工程数据监测显示，在这种新型工作负载下，传统 AI 计算节点内部 CPU 与 GPU 1:4 甚至 1:8 的悬殊配比正在被解构。在部分重度调度的场景下，系统对 CPU 与 GPU 的数量配比需求正逐渐向 1:1 收敛。

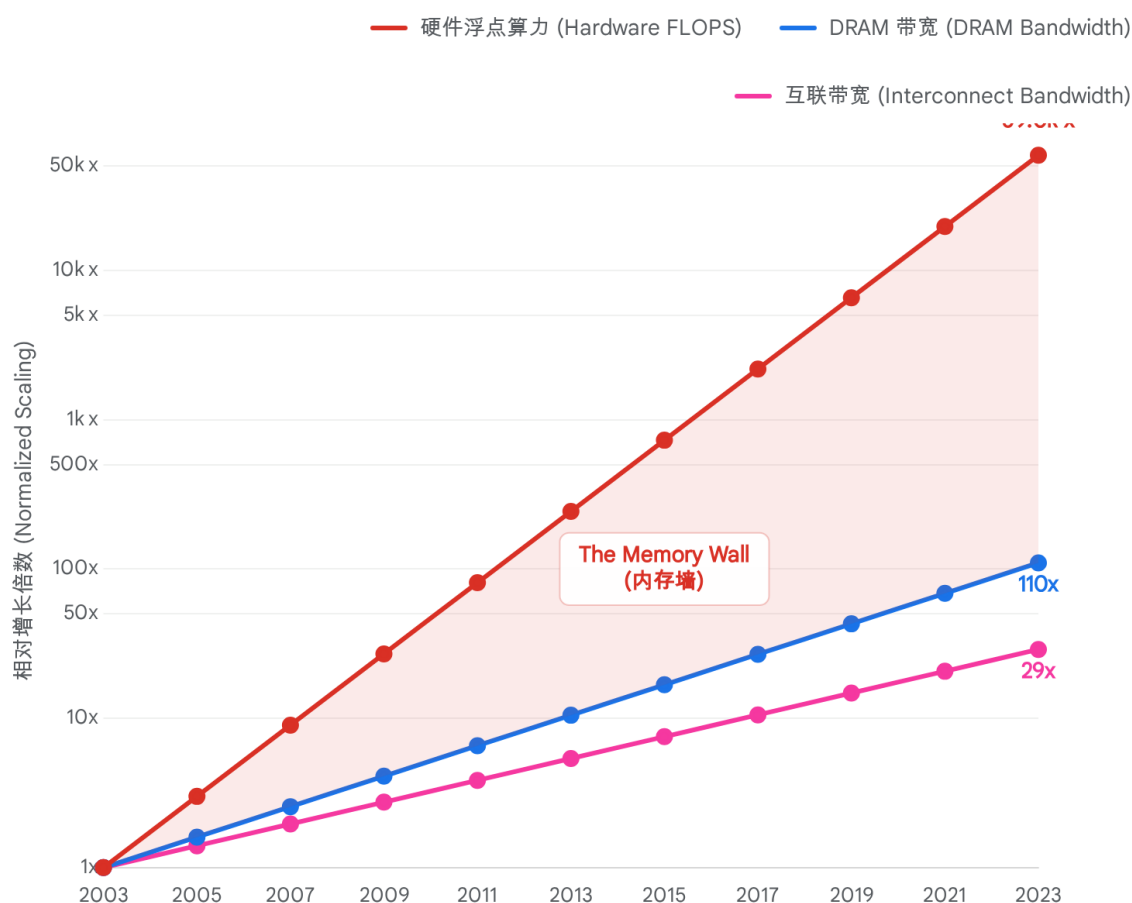
1.2 内存墙与互联墙的非对称性增长

硬件浮点算力与数据搬运能力之间的非对称性增长，构成了当前半导体设计的核心矛盾。根据业界关于“AI 内存墙”的统计模型，在过去 20 年间，硬件的峰值浮点运算能力平均每两年实现 3 倍的跨越式增长。

相较之下，底层数据通道的演进速度则显著滞后。在相同的两年周期内，DRAM 内存带宽平均仅增长 1.6 倍，系统互联带宽的增速更是低

至 1.4 倍。

算力供需失衡：硬件计算能力与数据传输带宽的增长鸿沟



过去二十年间，峰值硬件浮点算力平均每两年增长3倍，而DRAM带宽与互联带宽的增速仅为1.6倍与1.4倍，导致算力闲置和系统级瓶颈加剧。

Data sources: [AMD](#)

这种非对称性导致系统的有效算力上限不再由逻辑运算单元的绝对数量决定，而是受制于系统能否克服延迟，将张量数据及时馈送至计算核心。在极高算力密度的架构中，若无高带宽内存（HBM）与高级缓存层次结构的支撑，逻辑运算单元在大量时钟周期内将处于等待数据的

闲置状态。

1.3 功耗墙与极值优化：数据中心的终极制约

功耗（Power）已取代单一的性能指标，成为衡量数据中心算力天花板的度量基准。及至 2026 年，高端单张算力卡（如搭载多颗 GPU 及复杂互联元件的基板）的典型整板功耗（Total Board Power, TBP）正在经历从 1000 瓦至 1500 瓦区间，向 4000 瓦以上指数级飙升的过程。在超大规模数据中心的建设逻辑中，算力容量的衡量标准已直接转化为“吉瓦（GW）级的功耗预算”。

这一物理背景促使芯片微架构设计的核心理念发生反转：设计的工程目标是在给定的功耗上限内，实现系统级能效比（Performance per Watt）的最大化，而非单纯拉高时钟频率。根据半导体动态功耗方程（ $P_{dyn} = CV^2f$ ），功耗与工作电压呈平方正比关系。当通过提高核心电压以换取微小的频率增益时，系统将付出极不相称的热耗散代价。因此，在电压-频率（V-F）曲线中，最佳能效区间往往落在曲线相对较低的一端。

在现代 AI 芯片设计中，主体核心电路的运行电压已被极限压缩至当前半导体工艺所能支持的物理下限，普遍维持在 0.5 伏至 0.6 伏之间。然而，在极低工作电压下维持数千瓦的总功耗，意味着芯片封装必须吞吐数千安培的庞大电流，这为整个系统的供电网络（PDN）与封装基板的布线带来了严峻的电磁兼容与热管理挑战。

第二章 跨越晶体管密度极限：异构集成与先进封装架构

面对“三座大山”，传统的单片硅（Monolithic）设计已达到制造良率与经济性的双重极限。基于小芯片（Chiplet）的异构先进封装，成为突破摩尔定律二维限制的核心架构。

2.1 CDNA 4 架构与小芯片解耦

以 AMD 发布的 Instinct MI350 及 MI355X 系列 GPU 为例，其底层采用了第四代 CDNA 架构以及高度复杂的 3.5D 多级封装设计。该架构的工程逻辑是将不同特性的功能模块进行物理层面的解耦，并在最适合的半导体工艺节点上进行独立制造，从而在性能、功耗与良率之间取得最优解。

在 MI350 系列的封装层级内，集成了多达 1850 亿个晶体管。其基础层由两颗采用台积电（TSMC）6 纳米（N6）FinFET 工艺制造的 I/O 裸片（IOD）构成。由于 I/O 接口与模拟电路对先进制程的微缩效应不敏感，采用成熟的 N6 工艺既能保证极高的制造良率，又能将单颗裸片面积拓展至接近光刻掩模的物理极限尺寸（约 700 至 800 平方毫米）。在这两颗宽大的 IOD 基底之上，系统通过 3D 堆叠技术，垂直贴装了最高达 8 颗负责核心矩阵运算的计算裸片（XCD）。为了追求极限的能效与运算密度，这 8 颗 XCD 全部采用台积电最先进的 3 纳米（N3P）工艺节点进行流片。

规格参数	AMD Instinct MI355X (OAM)
GPU 架构	CDNA 4
工艺节点	TSMC 3nm (XCD) / 6nm (IOD)

晶体管数量	1850 亿
计算单元 (CUs) / 流处理器	256 / 16,384
矩阵核心 (Matrix Cores)	1,024
HBM 容量 / 类型	288 GB HBM3E
峰值内存带宽	8 TB/s
FP16 / BF16 峰值性能 (Sparsity)	4.6 PFLOPs
FP8 / MXFP8 峰值性能 (Sparsity)	9.2 PFLOPs
MXFP6 / MXFP4 峰值性能 (Sparsity)	9.2 PFLOPs

AMD Instinct MI355X 核心架构参数与计算能力概览。

架构上另一显著特征是引入了对微缩精度（Microscaling）数据格式的原生支持。CDNA 4 架构中的矩阵核心（Matrix Cores）全面兼容 MXFP4、MXFP6 等低精度浮点格式。在不增加片外存储带宽压力的前提下，硬件稀疏性（Sparsity）支持使得 MI355X 在运行 FP4/FP6 精度的推理任务时，理论峰值性能可达 9.2 PFLOPs，有效提升了计算密度与能效比。

2.2 铜-铜混合键合（Cu-Cu Hybrid Bonding）与垂直堆叠

实现 XCD 与 IOD 之间海量数据垂直吞吐的核心使能技术，是系统级集成芯片（SoIC）架构下的无凸块混合键合技术。

传统的 2.5D 或早期的 3D 封装（如用于 HBM 贴装的工艺）普遍依赖微凸块（Microbump）进行电气连接。然而，微凸块的间距通常在 35 至 50 微米之间，且由于不同材料的热膨胀系数（CTE）差异，微凸块间隙必须注入底部填充胶（Underfill）以提供热机械刚性。由于底层填充胶的热阻远高于硅，当这种工艺被应用于具有极高功率密度的逻辑层堆叠（Logic-on-Logic）时，会引发严重的热量淤积。

相较之下，AMD 在其 3D V-Cache 及 CDNA 加速器中采用的混合键合技术，彻底摒弃了微焊锡球与底部填充胶。该工艺利用化学机械抛光（CMP）将晶圆表面处理至接近原子级的平整度，随后在室温环境下使介电层（二氧化硅）产生共价键结合，最后通过高温退火工序使两层裸片上的铜焊盘（Copper Pads）发生物理互融。

工程数据显示，采用该技术后，硅通孔（TSV）及键合界面的间距可极限微缩至 9 微米级别。这种直接的“硅-硅”与“铜-铜”连接，使得 3D 堆叠界面的互联密度比传统的基于基板的 2D Chiplet 设计提升了超过 200 倍，比微凸块 3D 技术提升了 15 倍以上；同时，其互联能效提升了 3 倍以上，信号传输延迟被大幅削减。此外，去除热阻极高的底层填充胶后，垂直热传导效率得到显著改善，使得高频运行的计算核心产生的热量能够更快速地传导至散热顶盖。

2.3 存储子系统的高密度集成

为匹配 N3P 计算单元的吞吐能力，存储子系统的容量与带宽被提升至新的物理极限。在 MI350 系列封装的外围，通过 2.5D 硅中介层（Interposer）高密度集成了最高达 288GB 的 HBM3E（高带宽内存）。这种通过 8 个内存堆栈实现的存储阵列，能够提供高达 8 TB/s 的总峰值内存带宽。

同时，第四代 CDNA 架构优化了片上缓存（Infinity Cache）层级。这种“3D 逻辑计算层 + 2D 高带宽内存层”结合的复合 3.5D 架构，使得大

模型的 KV Cache 与模型权重能够最大程度地驻留在距离计算核心极近的高速网络中，从而在硬件层面上提供了缓解内存墙瓶颈的最有效方案。

第三章 供电网络重构与多芯片验证测试工程

封装密度的几何级数增长，将传统的系统级供电拓扑与集成电路测试理论推向了失效边缘。

3.1 集成电压调节器（IVR）化解大电流传输损耗

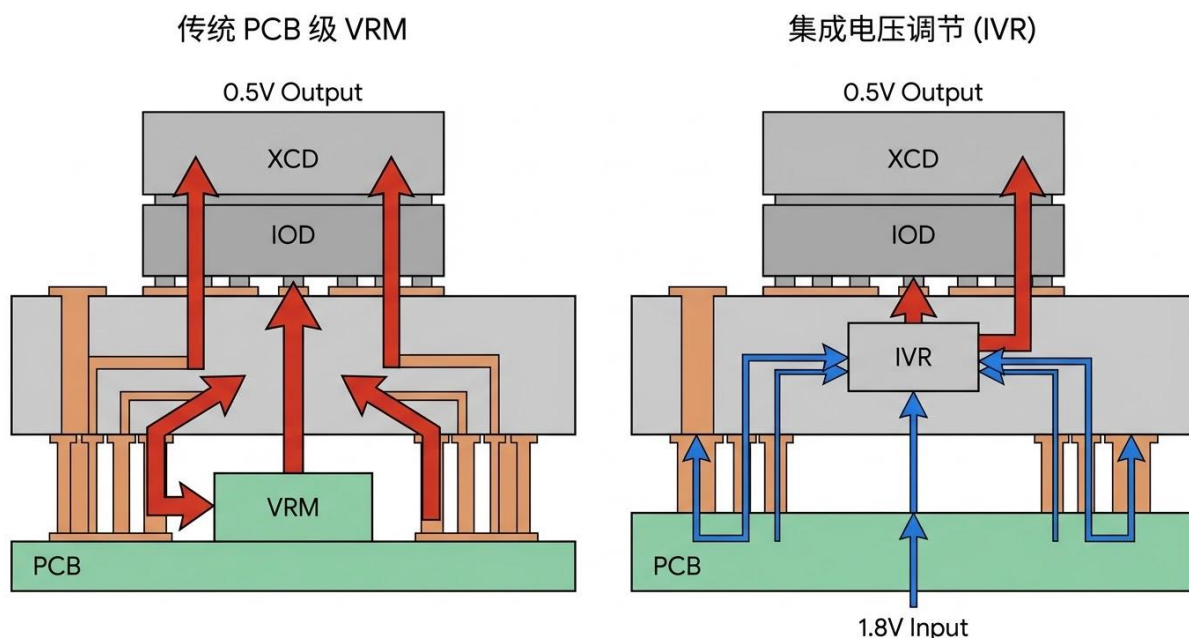
如前所述，当逻辑核心电路的工作电压低至 0.5V 至 0.6V 区间，在千瓦级的系统功耗下，这意味着封装基板必须承受数千安培的绝对电流。

在传统的服务器供电网络（PDN）设计中，电压调节模块（VRM）通常部署在 PCB 主板上。如果直接在 PCB 走线上长距离传输低电压、超大电流，由金属线路寄生电阻产生的焦耳热（即 I^2R 热损耗）将呈平方级放大。这不仅会消耗掉大量原本可用于计算的功耗预算，还会导致严重的电压降（IR Drop）现象，使得到达芯片焊盘的电压产生剧烈波动，进而引发系统逻辑错误甚至宕机。

为攻克这一电磁与热力学难题，工业界正加速普及集成电压调节器（Integrated Voltage Regulator, IVR）技术。IVR 的核心电气拓扑原则是“高压输电，近端降压”。具体的工程实践是，允许系统主板通过 PCB 走线向芯片封装输入相对较高的电压（如 1.8 伏或 3.3 伏），由于传输功率一定，输入电压的提升直接将穿过 PCB 和封装外部引脚的电流按反比例大幅压缩，从根本上遏制了外部传输路径上的 I^2R 损耗。

随后，高度微型化的电压调节器被直接贴装在封装（Package）内部的基板（Substrate）底部，使其尽可能贴近需要供电的逻辑裸片。在最新的封装演进中，这些无源和有源器件甚至被设计为直接整合进硅中介层（Interposer）内。在极短的物理距离内，IVR 将高电压转换为 0.5V 的核心工作电压，通过大面积的底层铜柱阵列垂直注入 Die 内。这一架构革命不仅将低压大电流的传输距离从数厘米缩短至毫米级别，将传输

路径损耗降至最低，更为高频运转的互联网络及运算核心释放了极其宝贵的热设计功耗（TDP）冗余。



3.2 逻辑验证（Verification）的数学化与分布式加速

随着单晶圆晶体管数量的爆炸式增长以及多 Die 堆叠的引入，芯片设计的逻辑状态空间呈现天文数字般的膨胀。传统的基于周期的动态仿真（Dynamic Simulation）在面对此类系统时，仿真时间成本已变得不可接受。

为了加速这一过程，工业界正在推动芯片间（Die-to-Die, D2D）接口的高度标准化。例如，广泛采用通用小芯片互连快线（UCIe）协议或 AMD 内部的 Infinity Fabric 标准协议。接口标准化带来的直接收益是，允许设计团队在模块间通信层面上引入大规模的“形式验证（Formal Verification）”。形式验证通过数学逻辑推理，穷举状态空间，能够严格证明硬件实现代码（RTL）与系统设计规范（Specification）之间的完全等价性。这种基于数学推导的方法，能够在早期阶段以 100% 的覆盖率排查出 D2D 协议状态机层面的死锁（Deadlock）和连接性错误，大幅

减少了对耗时极长的动态测试向量的依赖。

此外，在系统级验证层面，工业界已大规模部署分布式仿真（Distributed Simulation）架构。通过将整个多芯片系统的逻辑网表切分为对应各个物理小芯片的独立可执行文件，并将其分布部署在由高速以太网连接的服务器集群上协同运行。集群节点间通过 TCP/IP 协议同步各个芯片接口的信号状态，这种并行化策略将整体动态仿真速度提升了 2 至 8 倍。

3.3 3.5D 封装的测试（Testing）屏障与低功耗策略

在良率经济学中，由于 3.5D 封装涉及十余颗昂贵硅片的物理融合，只要其中任何一颗 Die 存在致命缺陷，整颗封装后的芯片便面临报废。因此，在裸片阶段实施极其严格的“已知良好裸片”（Known Good Die, KGD）筛选策略成为控制成本的核心环节。

然而，当裸片完成 3D 贴装后，内部 Die 的绝大多数信号引脚被外部物理结构彻底遮蔽，传统基于探针台（Probe Station）的外部测试信号接入点变得极其受限。为了应对这一挑战，业界广泛采纳了诸如 IEEE 1687（内部 JTAG 测试接入）及 IEEE 1838（3D 封装级联测试标准）等规范。这些规范允许测试机通过最外层裸片的有限引脚输入信号，再利用芯片内部的边界扫描链（Boundary Scan Chain），以级联（Cascading）的方式深度控制并读取内部隐藏裸片的测试状态。

同时，针对跨 Die 互联部分（如 TSV 硅通孔、微小间距的物理层 PHY 链路），芯片设计阶段必须内置专用的可测性设计（BIST）与冗余修复逻辑。最后，由于在自动测试设备（ATE）上进行大规模并行测试会引发全芯片晶体管的高频翻转，产生极高的峰值功耗，甚至可能烧毁未装配散热器的裸片。因此，开发测试向量压缩算法与专用的低功耗测试模式（Low Power Testing），已成为学术界与工程界共同关注的研究前沿。

第四章 系统级互联拓扑：从节点内全互联到跨机架网络

在 Agentic AI 及万亿参数模型的推动下，计算架构的重心已越过单芯片的硅面积扩张，演变为如何以极低延迟与极高带宽将成百上千颗独立处理器融合为一台虚拟的超级计算机。系统扩展被严密划分为三个物理层级：单节点内互联、机架级的纵向扩展（Scale-up），以及数据中心级的横向扩展（Scale-out）。

4.1 节点内一致性互联：第四代 Infinity Fabric

在底层计算节点（Node）内部，多颗 GPU 与 CPU 之间的紧密耦合是保证极低访存延迟的基础。以 AMD 最新的计算基板（如 UBB 2.0 架构）为例，其单节点内部高密度集成了 2 颗 EPYC 系列 CPU 与 8 颗 Instinct 系列 GPU。

节点内部的数据骨干依赖于 AMD 专有的第四代 Infinity Fabric（IFI）互联技术。该技术利用物理层的高速 SerDes 信令以及上层的 xGMI 协议，在 8 颗 GPU 之间建立了全对全（All-to-all）的无阻塞网络拓扑。第四代 Infinity Fabric 的单条链路双向带宽被提升至 153 GB/s。由于同一基板上的每颗 GPU 均配置了多达 7 条对外互联链路，使得任意两颗 GPU 之间均存在直连通道。

在更为极端的超大规模集群设计中（如部署 72 颗 MI455X 的 Helios 机柜系统），Infinity Fabric 互联技术甚至能够提供高达 896 GB/s 的节点内双向带宽。这种底层互联不仅提供了极致的原始吞吐量，更关键的是支持零拷贝（Zero-copy）内存语义与硬件缓存一致性（Cache Coherency）。这意味着系统中任何一颗处理器均可透明地通过 Load/Store 指令直接读写远端 GPU 的物理显存，大幅降低了软件编程的复杂度与数据显式搬移的延迟开销。

4.2 纵向扩展（Scale-up）：UALink 1.0 开源协议重构

当 AI 模型并行（Tensor Parallelism / Pipeline Parallelism）的需求跨越单台服务器物理边界时，必须通过纵向扩展（Scale-up）网络将多个机架（Rack）联合为统一的资源池。在 Scale-up 网络中，节点间的通信本质上依然是计算机内部内存总线的物理延伸，要求极低的纳秒级延迟和严格的内存读写语义，且对网络拥塞导致的数据丢包具有零容忍度。

长期以来，Scale-up 互联市场由 NVIDIA 的 NVLink 等专有封闭协议占据主导地位。为了突破生态孤岛并提供标准化组件，AMD、Intel、Broadcom 及 Google 等业界巨头于 2026 年正式确立并发布了 Ultra Accelerator Link (UALink) 1.0 开源互联规范。UALink 旨在为异构加速器混合部署提供一个高性能、低延迟且不受单一厂商锁定的互联基座。

庞大的网络容量与确定性延迟： UALink 1.0 在架构设计上展现了惊人的寻址规模，其单级网络（Pod）最大可支持 1024 个加速器节点的直连互通。在 UALink 交换机（ULS）的调度下，网络内的任何两颗加速器通信均可实现单跳（Single-hop）路由，端到端的交换跳数延迟被严格压缩在 100 至 150 纳秒之间。这为构建具有确定性执行时间的同步 AI 训练任务提供了坚实的物理保障。

物理层与链路层的深层解构： 在物理层实现上，UALink 1.0 务实地复用了成熟的以太网 PAM4 信令生态（参考 IEEE 802.3 标准），其信号波特率被设定为 212.5 Gbaud，以覆盖前向纠错（FEC）带来的额外开销，最终为上层提供单通道（Per-lane）200 Gbps（约 25 GB/s）的净有效吞吐量。通过将 4 条物理通道捆绑（x4 configuration），单端口可稳定输出 800 Gbps 的双向带宽。

特性维度	UALink 1.0	NVLink 5.0 (参考对照)
标准性质	开源标准 (多厂商联盟)	专有闭源协议
单通道带宽 (Per-Lane)	200 Gbps (~25 GB/s)	100 GB/s (双向 50GB/s)
最大集群规模 (单 Pod)	1,024 颗加速器	576 颗加速器
拓扑结构	点对点及交换机直连	基于 NVSwitch 的全对全网格
语义支持	Load/Store/Atomics 内存语义	内存语义

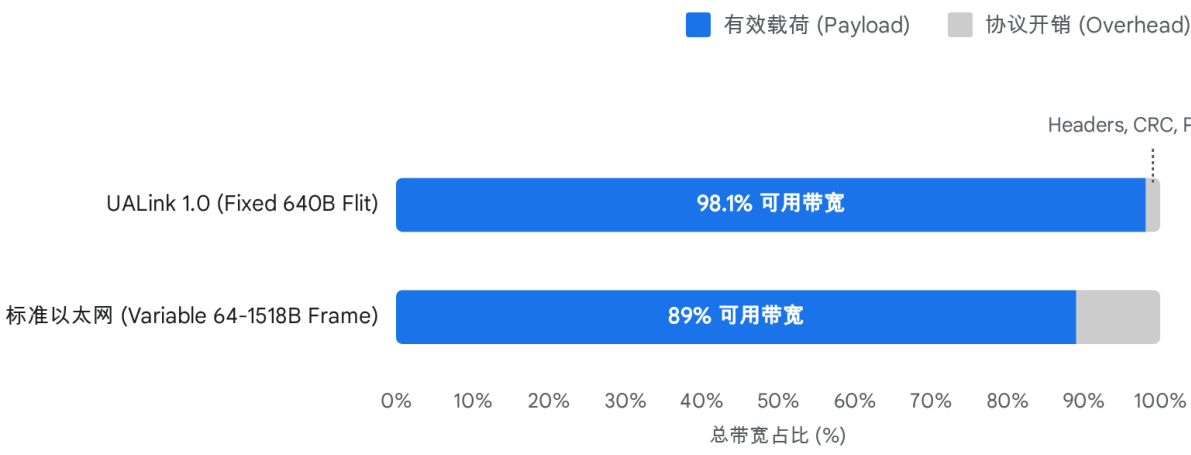
开源纵向扩展协议 UALink 1.0 与闭源协议在核心技术参数上的架构对比

4.3 数据链路层的确定性与高带宽效率

UALink 1.0 的核心创新之一在于其对数据链路层（Data Link Layer）的彻底重构。标准以太网为了兼容复杂的广域网环境，采用的是长度可变（通常为 64 至 1518 字节）的数据帧，并包含了大量的源/目的 MAC 地址、类型字段等控制开销，导致在处理小块张量数据时，实际的物理带宽利用率仅有 64%至 89%。

UALink 协议栈摒弃了这种低效设计，转而采用 640 字节（Bytes）的定长流控单元（Flit）机制。

Scale-up协议优化：UALink 1.0与标准以太网传输效率对比



由于采用专为AI内存语义设计的640字节定长流控单元（Flit），UALink 1.0将协议头与校验等控制开销压缩至不足2%，从而实现了远超标准变长以太网数据帧的物理层吞吐效率。

数据来源: [UALink 1.0 White Paper](#), [UALink 1.0 Specification Webinar](#)

在一个标准的 640 字节 UALink Flit 中，仅需分配 3 字节的底层协议头、5 字节的段标识以及 4 字节的循环冗余校验（CRC），剩余多达 628 字节均为供上层事务层（Transaction Layer）调用的纯有效载荷。这一极简的封包结构使得链路层的净吞吐效率高达 98.1%。

此外，定长报文设计极大地简化了交换机内部解析与路由逻辑的复杂性。对比分析显示，处理相同总带宽的逻辑下，UALink 交换芯片所需的硅面积仅为标准以太网交换芯片的约三分之一，且每条链路在交换机侧消耗的功耗降低了 44%。在宏观数据中心机柜级别，这种网络层的节能红利意味着有数千瓦的功耗预算可以被归还给运算节点，进一步优化了系统整体的能耗结构。

4.4 跨机架横向扩展（Scale-out）与光互联（CPO）的未来演进

当 AI 算力集群的规模跨越单 Pod 物理边界，扩张至数万颗 GPU 组

成的超级计算网络时，系统级互联则过渡至横向扩展（Scale-out）阶段。在这一广域层级，系统对延迟的容忍度放宽至微秒级（Microseconds），通信模式转变为松耦合的消息传递机制，经过专门优化的高速以太网协议（如 Ultra Ethernet Consortium 规范）凭借其成熟的可扩展性，成为跨机柜、跨机房干线网络的主流选择。

互联介质的革命：光电共封装（CPO） 无论是基于电信号铜缆的 Scale-up 架构，还是依赖光模块的 Scale-out 干线，随着信号调制速率向 200G 乃至 400G 演进，主板 PCB 走线上的信号衰减与驱动电路的功耗均已逼近材料物理的理论红线。应对这一挑战的终极路径，毫无争议地指向了光互联（Optical Interconnects）。

当前，硅光子（Silicon Photonics）工艺正快速进入量产前夜，而推动其在数据中心大规模铺开的封装载体则是光电共封装（Co-Packaged Optics, CPO）。CPO 架构通过将硅光收发模块与 GPU 计算核心或交换机交换核心集成在同一块硅中介层与封装基板上，彻底消除了极耗能的高速电信号在外部 PCB 上的长距离传输过程。

工程指标显示，行业正在攻克的核心节点是将光互联的数据传输能耗控制在每比特小于 5 皮焦耳（ $< 5 \text{ pJ/bit}$ ）。相比目前基于铜缆的电互联动辄约 30 pJ/bit 的能耗指标，CPO 技术不仅将带来一个数量级的能效跨越，还使得互联带宽不再受到外部物理引脚数量的刚性约束，为单芯片提供 Tbps 级别的出站带宽开辟了道路。

“近网络计算”的架构思辨： 在此次学术论坛中，AMD 管理层与学界专家进行了深度的技术探讨，揭示了一个正在酝酿的前沿架构趋势。由于长距离数据搬移所引发的互联焦耳热（ I^2R 热损耗）占据了越来越大的系统功耗比例，业界正在探索“在网络中计算（In-Network Compute）”的融合架构。该思路主张将部分数据归约（Reduction）、张量合并等相对简单的计算操作，从处于 0.5V 极低电压、对大电流极其

敏感的 GPU 计算核心中剥离出来，下放到运行电压相对较高、电流负载压力较小的互联交换模块（如 UALink Switch 内部）去执行。这种网算融合机制能够在干线网络侧过滤掉大量冗余的数据报文，虽然增加了交换机局部的发热，但在系统宏观层面上显著降低了跨节点数据搬移的总功耗，展现出全系统供电网络优化的巨大潜力。

第五章 算力下沉的工程实践：边缘部署与终端智能体硬件

伴随模型剪枝、量化（Quantization）及知识蒸馏等技术的成熟，拥有百亿参数级别的中小型基座模型正加速向边缘计算（Edge）节点和终端（Client）设备下沉。在非数据中心的运行环境中，“算力最优解”的评判标准从追求极致吞吐量，转变为在苛刻的散热条件、功耗上限以及有限的设备体积下，寻求硬件资源的最佳配比。

5.1 边缘与企业级推理的显存配比：Radeon PRO W7900D 案例

在政务处理、金融风控以及基于实时视频生成的企业级推理场景中，算力硬件的部署通常受限于传统的机房基础设施（例如缺乏直接液冷设施、标准风冷机柜散热能力受限、供电冗余不足等）。在这些硬性约束下，部署单卡功耗逾千瓦、甚至整机功耗数万瓦的顶级数据中心液冷阵列，不仅经济成本高昂，在工程实施上亦缺乏可行性。

针对此类痛点，重构显存与功耗比率的轻中量级边缘推理方案成为填补市场空白的关键。AMD 推出的 Radeon PRO W7900D 计算卡即为典型工程范例。该卡摒弃了冲击极限 Flops 指标的设计思路，转而将核心竞争力聚焦于极致的功耗控制与超大显存供给。在将典型整板满载功耗（TBP）严格压制在 295 瓦（可使用双 8-Pin 供电并兼容传统三槽主动风冷散热）的前提下，系统为其配备了高达 48GB 的 GDDR6 大容量显存，并在边缘端产品中引入了全链路 ECC 错误校验功能，确保了 7x24 小时高负载运行下的数据结果精准性。

在应对大模型推理时，庞大的显存容量能够有效容纳持续增长的键值缓存（KV Cache），从根本上避免了由于显存溢出（OOM）导致的任务崩溃与推理降级。在工程落地实践中（如在无锡算力岛部署的 5000 余张计算卡集群项目），通过在标准 2U 高度服务器中直接配置 8 张 W7900D 显卡与 2 颗 CPU，整台服务器在提供高达 384GB 庞大显存池的

同时，其满载系统功耗仅被控制在约 2300 瓦左右。这种无需改造机箱与水冷设施的卓越能效表现，极大地降低了系统总体拥有成本（TCO）。该算力集群的稳定性得到了验证：在跨国支持旧金山举办的国际 AI 开发者大会时，集群在长达上万公里的跨洋链路上实现了极速响应与零故障运行，成功为超过 900 名全球开发者提供了并发调用支持。

5.2 终端异构 SoC 与内存统一架构：Ryzen AI 处理器

对于个人 PC、AI PC 及微型嵌入式终端而言，Agentic AI 带来了更为复杂的本地执行挑战。终端模型不仅需要处理自然语言交互，还涉及本地图像分析、文件系统跨应用检索等任务。这些任务交织着密集的神经网络矩阵乘法与大量的系统级控制流操作。若仅依赖 GPU 进行纯算力堆砌，在处理复杂控制流时将遭遇严重的性能瓶颈，因此，CPU 的高效逻辑调度能力不可或缺。

终端侧硬件架构的破局之道在于“异构计算融合+系统级统一内存”。以 AMD 的 Ryzen AI（锐龙 AI）系列处理器为例，通过引入先进的扇外型（Fan-out）封装与小芯片设计，系统在单一 SoC 芯片内高度集成了负责逻辑协调的高性能 CPU 核心、负责并行渲染与特定计算的 GPU 集群，以及专门针对低功耗张量运算进行硬化设计的神经处理单元（NPU）。这种设计摒弃了传统的高功耗 SerDes 走线，转而使用高密度的内部互联布线，将芯片间数据传输的能耗从约 2 皮焦耳/比特大幅压缩至仅 0.2 皮焦耳/比特。

更为核心的架构突破在于打破了传统 CPU 与独立 GPU 之间的内存物理隔离。该平台为整个 SoC 提供了一个高达 96GB 大容量系统统一内存（Unified Memory）池。这一架构使得系统能够在一个体积仅稍大于手掌的迷你主机（Mini Box）内，流畅载入并运行参数规模达 350 亿（35B）量级的大型基座模型（如 Qwen 35B）。在单智能体任务执行中，该微型平台能够实现每秒生成 45 个 Token 的优异效率，同时支持高

达 260K 的超长上下文窗口追踪；即便在并发多智能体（Multi-agent）交互的场景下，系统依然可以稳定支撑 6 个独立 Agent 的并发推理工作，且维持每个 Agent 具备 95K 的独立上下文处理能力。通过在异构引擎间实现真正的零拷贝数据共享，这种 CPU 与 GPU 性能的极佳平衡，显著缩短了 Agentic 端到端工作流的完成周期，其综合响应速度优于单纯依赖 GPU 的传统分离式终端架构。

第六章 软件生态闭环与全栈算力底座

硅基硬件架构的潜力上限，最终唯有通过稳健、通用且高度优化的软件生态层才能被完全释放。在 AI 加速计算领域，构建一个开源、向下兼容且具备“开箱即用（Out-of-the-box）”属性的软件栈，是打破特定专有生态壁垒的核心武器。

6.1 “一次编程，随处可用”的软硬件协同抽象

AMD 构建的开放式计算平台 ROCm（Radeon Open Compute）在架构设计上追求极高的底层硬件抽象能力。通过统一的软件栈界面，ROCm 向下能够原生适配各种异构硬件微架构，无论是部署在云端数据中心的高性能 Instinct 系列加速器、执行企业级边缘推理的 Radeon PRO 系列显卡，还是深度嵌入终端的 Ryzen AI 异构处理器，均被纳入同一个编程框架的管理之下。

向上层，ROCm 平台通过提供性能强劲的编译器（Compiler）、低延迟运行时（Runtime）以及经过深度数学优化的算子库组合（如 rocBLAS、MIOpen 等），实现了对业界主流人工智能框架的无缝底层对接，涵盖了 PyTorch、TensorFlow、JAX 以及 Triton 等。通过在软件层面对底层硬件差异进行屏蔽，开发者实现了“一套代码，全平台部署”的工程愿景，降低了原有应用逻辑向非垄断阵营硬件迁移的学习曲线与重构成本。

6.2 开源社区协作与零日（Day-0）支持机制

孤立的底层驱动并不足以支撑起一个繁荣的开发者生态。算力提供商必须深入模型研发与基础设施运维的最前线。基于全面的开源战略，ROCm 生态与 Hugging Face 等全球最大的开源模型托管社区，以及 vLLM 等核心推理引擎维护组织建立起了深度的联合研发与优化机制。

这种紧密协作的直接成果体现在对于最前沿基座模型（涵盖

OpenAI 的模型体系、Meta 的 Llama 系列开源大模型等）的“零日（Day-0）”支持能力上。即当业界新型模型结构宣布开源或发布的当天，ROCm 底层的对应算子库即已通过回归测试并准备就绪，确保算法工程师能在第一时间于底层硬件上跑出贴合理论极限的满血性能表现。同时，针对数据中心的宏观运维需求，该软件栈深入集成了 **Kubernetes** 容器化编排工具及 **AMD GPU Operator**，大幅简化了云原生环境中数百乃至数千卡集群的节点发现、资源隔离与调度逻辑，打通了从学术界概念验证（PoC）到企业级数据中心规模化、自动化部署的全栈闭环。

总结

在人工智能应用从语言模式的静态预测，向具备复杂执行逻辑的多智能体协同迈进的历史性转折点上，底层计算基础设施正在经历从芯片内部材料科学、到宏观数据中心集群拓扑的全面重构。基于对行业设计路线及 AMD 工程实践的系统梳理，本报告得出以下技术观察：

首先，“系统级能效即算力上限”已成为数据中心设计的准则。在未来，限制算力集群物理扩张的根本要素，不再是晶体管在二维平面上的微缩速度，而是供电网络（通过 IVR 重构）的电流支撑能力与互联介质在长距离传输中产生的焦耳热耗散（ I^2R ）。整个高性能半导体工业正围绕着 0.5V 极低核心工作电压下的能效极值点，进行前所未有的系统级妥协与重构。

其次，三维先进封装已取代单片硅集成，成为异构运算架构的核心战场。摩尔定律在物理维度上的放缓，正被具有 9 微米级极小互联间距的铜-铜混合键合（如 3D V-Cache 技术底层的 SoIC 工艺）与 3.5D 多级封装技术在垂直物理空间内所强力补偿，拉近了运算单元与高带宽存储体系的数据交换距离。

再次，开放标准与高度定制化正在重塑大规模网络架构。在规模化（Scale-up）集群互联领域，以 UALink 1.0 协议为代表的技术路线表明，专门针对 AI 内存语义定制的定长数据报文机制，在物理带宽效率与交换芯片能效上具有显著优势，开源协议有望打破单一闭源体系的技术壁垒。而在横向扩展（Scale-out）的未来，光电共封装（CPO）的大规模商用将成为跨越电信号能耗困局的物理通途。

最后，端到端的工作流特性正在定义非云端设备的硬件平衡基点。边缘与终端计算设备的 AI 部署逻辑清晰地表明，过度堆砌单一规格的运算单元将遭遇快速的边际效益递减。依托系统级的统一内存架构，并实现 CPU 控制流逻辑与 GPU/NPU 并发算力的高度协同，才能在真正的

多智能体并发工作流程中实现最低的延迟响应与最优的拥有成本。

伴随开源硬件互联规范（如 UALink、UCIe）与开源抽象软件栈（如 ROCm）的深度生态融合，封闭的单体生态体系正面临系统维度的全面挑战。精准平衡内存容量、功耗预算、互联带宽，并以开放标准连结全球开发者生态的高效异构系统架构，将定义未来的 AI 算力基础设施。

联系我们: CEISD@Shanghaitech.edu.cn