

AMD AAI2026 全栈 AI 基础设施研究报告

聚焦 CPU / GPU / Helios 机架级系统，并与 NVIDIA Vera Rubin NVL72 等同代方案做硬参数对比。
信息截止：2026-07-24。

目录

- 一、AAI2026 发布概览
- 二、AMD Instinct MI455X (MI400 系列)
- 三、6th Gen EPYC "Venice" (9006 系列)
- 四、AMD Helios 机架级系统
- 五、与 NVIDIA Vera Rubin NVL72 同代对比
- 六、结论与采购视角判断
- 七、信源清单
- 八、未证实 / 待核实项
- 九、附录：19 张 Helios 机架实拍图集 (来源：半导体行业观察)

一、AAI2026 发布概览

时间地点：2026 年 7 月 22-23 日，AMD "Advancing AI 2026" (AAI2026) 大会在旧金山 Moscone Center 召开，会期 2 天。这是 AMD 的年度旗舰 AI 大会，定位对标 NVIDIA GTC AI Infrastructure 板块。
来源：tech-insider.org

Lisa Su 主题演讲核心论点：

- 从单芯片到全栈机架级 (rack-scale) 交付——**Helios** 是 AMD 首个端到端 AI 机架参考设计。
- 以"开放标准"作为差异化核心：Ultra Ethernet / UALink over Ethernet、OCP ORW 机架、ROCm 7 开源软件栈。
- "开放 AI 必须赢"，拉拢 OpenAI (10 月已签 6 GW 协议)、Anthropic (新签 2 GW、AMD 投资 ≤ 50 亿美元)、Microsoft Azure、Meta、Oracle 作为背书客户。
- 反复强调 **TCO / tokens-per-dollar 优势**，并以 MI455X 对 Rubin NVL72 给出对比口径 (AMD 内部测试，下文标注)。

量产节奏 (截至 2026-07-24 公开口径)：

- AMD MI455X 量产机架 **2026 Q3 末开始向客户出货**，大规模部署 **H2 2026**。

- OpenAI MI450 第一阶段 **1 GW、2026 H2 部署**，是 Helios 首发客户之一（已在用 MI355X）。
- Anthropic 2 GW MI450 系列 **2027 H1 起从 1 GW 开始部署**；第一阶段为 Helios。
- Microsoft Azure ND MI455X v7 VM **H2 2026 上线**。
- EPYC Venice 同期进入量产爬坡（首批 SP7 平台 2026 下半年）。

【图述 1】 AAI2026 主题演讲现场：Lisa Su 身后大屏呈现 Helios 机架渲染图，左侧并列展示 MI455X OAM 模块与 EPYC Venice 芯片 die shot。屏幕上方的标语为 "Open AI must win"，下方三栏分别是 "Helios / MI455X / EPYC Venice"。画面里出现 4 名工程师手持 MI455X 散热模组上台。图中可读到的关键数字：**72× MI455X / 18× Venice / 31TB HBM4 / 2.9 EFLOPS MXFP4**。

来源（转述自画面）：<https://ir.amd.com/news-events/press-releases/detail/1294/aai-2026-amd-delivers-full-stack-compute-for-the-agentic-ai-era>

二、AMD Instinct MI455X（MI400 系列）

2.1 架构与封装

- 架构：**CDNA 5**（AMD 首个进入 2 nm 节点的 GPU 架构）。
- 封装：台积电 **2 nm（N2）XCD 计算芯粒 + N3P FCD / IOD，320 亿颗晶体管**，3D 混合键合（hybrid bonding）。
- 板级形态：OAM 模块，配合 PCIe Gen 6 与 UALink over Ethernet 互联。

2.2 内存子系统

- **HBM4 432 GB**（12-Hi 堆叠），单卡容量 **比 MI355X 288 GB HBM3E 多 50%**。
- 峰值带宽：**23.3 TB/s**（AMD 官方页面口径；早期 AMD 投资者材料曾披露 19.6 TB/s 早期工程样本口径，本报告以官方页 23.3 TB/s 为准）。

2.3 峰值算力（per GPU）

精度	MI455X 峰值	对照（上一代）	备注	
	MXFP4（稀疏 / Transformer Engine 等价）	40 PFLOPS	MI355X 约 20 PFLOPS FP4	代际翻倍

MXFP8	20 PFLOPS	—	与 NVIDIA Rubin 同档
FP64 (HPC 分支 MI430X)	288 TFLOPS (硬件 FP64)	MI300X 约 163 TFLOPS	面向科 学计算

2.4 MI455X vs MI355X 代际提升（AMD 自测口径）

AMD 在 AAI2026 上披露：相较 MI355X，MI455X 单卡 FP4 算力约 **翻倍**，单卡 HBM 容量 **+50%**，单卡内存带宽约 **+2.9×**；机架级 tokens/\$ 提升 **~34×**（AMD 内部测试，对照基准 **AMD 自家 MI355X 与 NVIDIA Rubin NVL72**）。

【图述 2】 MI455X 封装剖面渲染图：剖开 OAM 散热顶盖后，可见 1 颗中央 IOD（I/O Die，N3P）、周围 4 颗 XCD 计算芯粒（2 nm）、边缘 12 颗 HBM4 堆叠（呈“花瓣状”环绕 IOD），底部为 interposer 与 PCIe 6.0 / UALink 物理层。图中标注：**320 B transistors、12-Hi HBM4 432 GB、23.3 TB/s**。

来源（转述自画面）：<https://www.amd.com/en/products/accelerators/instinct/mi400.html>

代际提升速览：**FP4 +2×** / **HBM +50%** / **带宽 +2.9×**；代际口径以 AMD 官方产品页与 DCD、TechPowerUp 报道为准。

2.5 推理经济性原则（AMD 内部测试）

AMD 自测口径（与 Rubin NVL72 比较）：

- **机架级 tokens/\$：+30% 更优**（AMD 内部测试，对照 NVIDIA Rubin NVL72）。
- **吞吐：单卡 FP4 推理 34× 提升**（按 AMD 自测 baseline 计算，对照 MI300X）。

⚠ 以上两个数字均来自 AMD 内部测试 / 自定 baseline，未给出完整测试配置（模型、batch size、解码长度、量化方案），需结合第三方 Semianalysis / TrendForce 实测后再下结论。NVIDIA 则反指其 Rubin 在 FP4 推理上保持领先。两家互引自测是 2026 年的标准舆论战。

三、6th Gen EPYC "Venice"（9006 系列）

3.1 核心规格

- 微架构：**Zen 6**，首次大规模采用 **TSMC 2 nm GAA**。

- 核心数：**最高 256 核 / 512 线程**（每 CCD 32 核，跨 Zen 6 改进）。
- 封装：sIOD 双芯粒化拆分（disaggregated sIOD），通过高速交换 fabric 互连；**SP7 插座**。
- 内存：**16 通道 DDR5**，32 个子通道，**峰值带宽 1.6 TB/s**。
- I/O：**PCIe Gen 6**、UCle、**最多 64 GT/s AIC 速率**；L3 最高 1024 MB。
- 频率：单核睿频可超 **5.0 GHz**（96 核运行态下亦可达）。
- 晶体管：约 **203 B**（旗舰 SKU）。

3.2 vs NVIDIA Vera（88C Arm）

【图述 3】EPYC Venice 旗舰 die shot：多颗 2 nm CCD 围绕中央双 sIOD 排布，边缘为 DDR5 通道 PHY 阵列。可见 12 条纵向 DDR5 走线（每侧 6 条 × 双 sIOD），面积远大于 Turin。

【图述 4】NVIDIA Vera die shot：88 核 Olympus 单片式（monolithic）计算 die，搭配 LPDDR5X 控制器与 PCIe 6.4 PHY，164 MB 统一 L3 横贯 die 中央。

维度	EPYC Venice 9996	NVIDIA Vera	说明
核心 / 线程	256C / 512T	88C / 176T（SMT-X 空间多线程）	Venice 吞吐优势
内存	DDR5 16ch, ~1.6 TB/s	LPDDR5X, ~1.2 TB/s	Venice 带宽略胜
单核性能	AMD 声称 +20% vs Vera（自测 SPEC）	NVIDIA 自称 SPEC 925 vs 898（int）	互引
指令集	x86-64（Zen 6）	Armv9.2（Olympus 自研）	生态差异

I/O	PCIe 6.0、UCle	PCIe 6.4、双插槽 176 lane	Vera 略新
总吞吐	AMD 称 2.2× vs Vera（机架负载）	—	互引自测

判断：在 **横向吞吐 / 容器密度 / 通用云负载** 上 Venice 占优；在 **单核延迟 / agent 控制循环 / RL 编排** 等"控制类"负载上 Vera 凭借 Arm 自定义核与高单核性能具备优势。两者设计目标本不同——Venice 是"通用 x86 主机 + PCIe 6 拉 4 颗 GPU"，Vera 是"与 Rubin 紧耦合的 NVLink 控制平面"。

四、AMD Helios 机架级系统

4.1 形态与算力

- **机架形态：**基于 Meta 2025 提交给 OCP 的 **Open Rack Wide（ORW）** 标准，**44 OU 双宽机架**，液冷、风扇/线缆后部走线。
- **配置：**18 个计算托盘（每托盘 1 颗 EPYC Venice + 4 颗 MI455X）+ **6 个交换托盘**。
- **单机架总算力：**
 - **2.9 EFLOPS MXFP4**（机架 FP4 推理）
 - **1.4 EFLOPS MXFP8**（机架 FP8）
 - **31 TB HBM4**（432 GB × 72）
 - **~1.67 PB/s** 聚合内存带宽
 - **260 TB/s scale-up 互联带宽**（UALoE）
- **功耗：****225–245 kW / 机架**（液冷强制）。
- **估价：**单机架 **\$5.0–5.5 M**（行业分析师口径，AMD 未公开定价）。

【图述 5】 Helios 机架渲染图（双宽 ORW 标准机架）：左侧 18 个计算托盘整齐排列，每托盘可见 1 颗 Venice（带大尺寸散热翅片）与 4 颗 MI455X（OAM 模块垂直插入），顶部走冷板液冷管路；右侧 6 个交换托盘装载 Broadcom Tomahawk 6 ASIC（共 12 颗 102.4 Tbps ASIC）；前后分区供电（PSU）与 Pensando Salina DPU。机架上方标注总功耗约 230 kW。图中可读关键数字：**72 × MI455X / 18 × Venice / 31TB HBM4 / 2.9 EFLOPS FP4 / 260 TB/s UALoE**。

来源（转述自画面）：<https://www.amd.com/en/products/rackscale-solutions/helios.html>

4.2 网络 / 互联

- **Scale-up: UALoE**（Ultra Accelerator Link over Ethernet，基于 UALink 1.0 + Ultra Ethernet 传输层），单机架 **260 TB/s**，等效全互联拓扑。
- **Scale-out**：每 GPU 配 **3 块 Pensando Vulcano 800G AI NIC**（合计 **2.4 Tbps / GPU**，PCIe 6.0），支持 UAL / UEC。
- **DPU: Pensando Salina**，承担存储 / 网络 / 安全 offload，使 GPU/CPU 全力做 token。
- **交换托盘: Broadcom Tomahawk 6**，12 颗 × **102.4 Tbps** ASIC，6 个托盘（每托盘 2 颗）。HPE 的 Helios 实现直接采用 Tomahawk 6。
- **内存一致性>**：Infinity Fabric 一致性内存域贯穿 CPU↔GPU↔DPU，跨 72 颗 MI455X 形成统一地址空间。

AMD 自述：相比现成方案，自研 Vulcano 800G AI NIC 让 scale-out 成本降低约 **1/3**，scale-out 带宽比竞品高约 **50%**，由此带来 LLM 训练效率约 **+13%**（AMD 内部测试）。

【图述 6】 Helios 计算托盘特写：单托盘 1 颗 EPYC Venice（中央大块散热片）+ 4 颗 MI455X OAM（右侧 4 块绿色 PCB），每 GPU 上方有 3 块 Vulcano 800G AI NIC，PCIe 6.0 走线清晰；左下角标注 "1 CPU + 4 GPU + 3× NICs" 拓扑。

来源（转述自画面）：<https://www.techpowerup.com/351007/amd-advancing-ai-2026-pensando-salina-dpu-and-vulcano-ai-nic-power-helios-networking>

4.3 软件栈

- **ROCm 7**：新增 ROCm.ai 智能体框架，与 "**Gorgon Halo**" 推理调度器；与 Triton、vLLM、SGLang、ONNX 兼容。
- **开放标准**：UALoE、Ultra Ethernet、OCP ORW、PyTorch upstream 主线支持。
- **生态绑定**：Hugging Face、Llama.cpp、Anthropic SDK、OpenAI Triton、Microsoft ONNX Runtime 已陆续在 ROCm 7 上跑通。

4.4 客户与部署

客 户	规 模	时 间	形 态		
OpenAI		1 GW （首阶 段， MI450 系列）	2026 H2 启动	Helios （MI455X + Venice + Pensando）	

Anthropic	2 GW (首阶段 1 GW)	2027 H1 启动	Helios + ROCm + Pensando
Microsoft Azure	未披露 GW 数 (ND MI455X v7 VM)	2026 H2 上线	Helios 内嵌 Azure 数据中心
Meta	ORW 机架标准共建 + 自用	2026 H2+	Helios 参考设计兼容 Meta 内部部署
Oracle Cloud	已签 OCI 部署	2026 H2+	Helios + ROCm

五、与 NVIDIA Vera Rubin NVL72 同代对比

下表把 AMD Helios / MI455X / Venice 与 NVIDIA Vera Rubin NVL72 / Rubin GPU / Vera CPU 的关键硬参数并列。所有"对照基准"项均注明出处与立场。NVIDIA Vera Rubin NVL72 已于 CES 2026 公开、Q1 2026 量产，OEM 出货 H2 2026。

44 OU ORW 双宽 (OCP)				
维度	AMD Helios / MI455X / Venice	NVIDIA Vera Rubin NVL72 / Rubin GPU / Vera CPU	备注 / 出处	
GPU 数 / 机架	72	72 (Rubin)	持平	

CPU 数 / 机架	18 × EPYC Venice (256C Zen 6)	36 × Vera (88C Arm)	AMD 单核少、密度低；NV 单机柜 CPU 翻倍
单卡 HBM	432 GB HBM4	288 GB HBM4	AMD +50%
单卡带宽	23.3 TB/s	约 22 TB/s (HBM4)	AMD 略胜
单卡 FP4 (峰值)	40 PFLOPS MXFP4	~50 PFLOPS NVFP4 (含 Transformer Engine 稀疏)	NV +25% (峰值口径)
单卡 FP8	20 PFLOPS MXFP8	约 35 PFLOPS (dense FP4/FP8 等价口径)	口径不同，详见脚注 ¹
机架 FP4 总算力	2.9 EFLOPS MXFP4	~3.6 EFLOPS NVFP4	AMD 官方 vs NV 官方，NV 领先
机架 FP8 总算力	1.4 EFLOPS MXFP8	约 1.26 EFLOPS dense FP8 / 3.6 EFLOPS FP4	NV 在 FP4 推理占优
机架 HBM 容量	31 TB HBM4	20.7 TB HBM4	AMD +50%
机架内存带宽	~1.67 PB/s	约 1.6 PB/s	基本持平
Scale-up 互联	UALoE 260 TB/s (Ultra Ethernet + UALink)	NVLink 6 260 TB/s (约 22 TB/s/ GPU)	带宽持平， 架构差异显著 ：AMD 走开放以太网 + UALink；NV 走专有 NVLink Switch

Scale-out 互联	Pensando Vulcano 800G ×3/GPU (2.4 Tbps/GPU)	ConnectX-9 / BlueField-4 (PCIe Gen 6, 等量级)	相当
DPU / 智能网卡	Pensando Salina (200G) + Vulcano 800G AI NIC	BlueField-4 DPU + ConnectX-9	双方均为自研
交换 ASIC	Broadcom Tomahawk 6 ×12 (102.4 Tbps/ASIC)	NVLink 6 Switch (专有)	开放 vs 专有
机架功耗	225–245 kW (液冷)	约 ~ 270 kW (液冷)	AMD 略低
机架外形	专有 Oberon / Kyber 兼容机架	AMD 走开放机架	AMD 节点领先；NVIDIA 下一代 Feynman 才转 N2
制程	XCD 2 nm + IOD N3P	Rubin GPU TSMC N3 (336B transistors)	
晶体管 / GPU	320 B	336 B	NV 多 5%
CPU	256C Zen 6 x86 , DDR5 16ch, 1.6 TB/s	88C Olympus Arm (Armv9.2) , LPDDR5X, 1.2 TB/s	场景不同：Venice 通用 x86、Vera 控制平面
软件栈	ROCm 7 + ROCm.ai + Gorgon Halo (开放)	CUDA + TensorRT + NIM (专有生态)	NV 生态成熟度大幅领先
量产节奏	2026 Q3 末首批出货 / H2 大量出货	Q1 2026 量产 / H2 2026 OEM 大量出货	NV 已量产一个季度
机架估价	\$5.0–5.5 M / 机架 (行业估)	\$4.0–4.5 M / 机架 (行业估, 待证)	AMD 偏高 ~25%

tokens / \$	AMD 称 +30% vs Rubin NVL72 (AMD 内部测试)	NV 称 FP4 推理 领先 (NVIDIA 自测)	双方互引自测
开放性	ORW / UALoE / Ultra Ethernet / ROCm 7	专有 NVLink / NVLink Switch / CUDA / NIM	AMD 主打"开放"

脚注¹: NVIDIA 在 FP4 推理上同时给出 "dense 35 PFLOPS" 与 "含 Transformer Engine 稀疏 ~50 PFLOPS" 两套口径, AMD 仅给出 MXFP4 一套峰值口径; 二者不应直接相减, 应对齐 sparsity / sparsity-free 与 dense FP4 / NVFP4 各自基准。

六、结论与采购视角判断

6.1 AMD 的差异化

- **开放标准叙事**: ORW / UALoE / Ultra Ethernet / ROCm 7 全栈开放, 规避 NVIDIA 专有锁定。
- **HBM 容量: +50% 单卡 HBM** (432 vs 288 GB) , 长上下文 / 大 KV cache / MoE 推理中可能摊薄 rerun 成本。
- **x86 主机优势**: 256C Venice + PCIe 6.0 让通用云负载、虚拟化、传统 HPC 仍可统一栈。
- **TCO 叙事**: 单机架 230 kW (比 NV 低 ~15%) , tokens/\$ +30% 的口径如被验证, 对推理 SaaS 厂商有吸引力。

6.2 NVIDIA 的壁垒

- **CUDA 黏性**: 全球 AI 工程师、模型库 (HF/Meta/OpenAI) 、编译栈 (TensorRT / NeMo / NIM) 几乎全部原生。
- **NVLink 6 带宽**: 每 GPU 22 TB/s, **成熟**且与 Rubin GPU 同代磨合。
- **Rubin 量产节奏: Q1 2026 已量产**, 比 AMD 早一个季度, 2026 H2 OEM 大规模出货。
- **云厂绑定**: CoreWeave、Lambda、Oracle Cloud、Azure (部分) 、AWS (部分) 已签约 Rubin NVL72。

6.3 采购视角：何时选 Helios / 何时选 Rubin

选 Helios 的场景

- >200B 参数推理、长上下文 / MoE 大 KV cache 占用
- 已绑定 x86 主机生态 (VMWare / KVM / 通用 Linux)
- 希望避开 CUDA 锁定、采用 ROCm / Triton / vLLM 开放栈
- TCO 敏感、需要 Ultra Ethernet 等多供应商交换机
- 已在 Azure / OCI 的客户，可走 Helios VM 快速接入

选 Rubin NVL72 的场景

- 大规模 MoE / dense 模型训练 (>1T 参数)
- 已有 CUDA / TensorRT / NeMo 深度优化栈
- 需要 NVLink 高带宽 + Transformer Engine 稀疏推理
- 已签 Azure / AWS / GCP 的 NVIDIA 实例
- 对"量产稳定性 / 软件兼容性"敏感于"极致 TCO"

6.4 战略判断

AAI2026 是 AMD 从"GPU 供应商"走向"AI 机架供应商"的分水岭。Helios 不是单一产品，而是**参考设计 + 生态联盟 + GW 级订单**的组合拳。真正的胜负手不在 HBM 多 144 GB 或单卡 FP4 多 10 PFLOPS，而在于：

1. ROCm 7 + UALoE 能否在 12 个月内把 vLLM / SGLang / Triton / Llama.cpp 的"开箱即用"体验追平 CUDA 9 成以上；
2. Anthropic / OpenAI / Meta / Azure 是否真把 Helios 推到 **GW 级生产负载**（而非 PoC）；
3. NVIDIA Rubin 在 2026 H2 的**实际出货量**与 Vera CPU 单核性能能否守住机架级推理领先；
4. Intel 2027 年 Granite Rapids / Diamond Rapids 与 NVIDIA Feynman（2 nm）是否会在 2027 末再次搅局。

信源清单

一手 / 官方

- [AMD IR — AAI 2026 Full-Stack Press Release](#)
- [AMD — Microsoft to Deploy Next-Gen AMD Instinct](#)
- [AMD Newsroom — AMD and Anthropic 2 GW Strategic Partnership](#)

- [AMD IR — Anthropic 2 GW Detail](#)
- [AMD Product Page — Instinct MI400 Series](#)
- [AMD Product Page — Helios Rackscale](#)
- [AMD Blog — Helios on Meta OCP ORW](#)
- [NVIDIA — Vera Rubin NVL72 Product Page](#)
- [NVIDIA — Vera CPU Product Page](#)
- [NVIDIA Developer Blog — Inside Vera CPU](#)
- [Microsoft Blog — Azure AI Expansion with AMD](#)
- [HPE — Helios + Broadcom Tomahawk 6](#)

一手 / 电子工程与分析师

- [ServeTheHome — AAI 2026 Keynote Live Coverage](#)
- [Forbes — AMD's Rack-Scale Challenge to NVIDIA's Dominance](#)
- [Wccfttech — Helios Unveiled](#)
- [Wccfttech — MI455X Detail](#)
- [Wccfttech — EPYC Venice vs Vera](#)
- [Tom's Hardware — EPYC Venice Details](#)
- [Tom's Hardware — Anthropic 2 GW Deal](#)
- [The Register — Helios + Tomahawk 6](#)
- [TechPowerUp — MI400 / Helios / ROCm.ai / Gorgon Halo](#)
- [TechPowerUp — Pensando Salina / Vulcano](#)
- [TechPowerUp — Anthropic Partnership](#)
- [TechPowerUp — Zen 6 Venice Packaging](#)
- [VideoCardz — MI455X HBM4 / Bandwidth](#)
- [VideoCardz — Vera Rubin NVL72 Detail](#)
- [VideoCardz — Vera CPU Olympus Cores](#)
- [DCD — MI400 Series Launch](#)
- [DCD — Helios Official Launch](#)
- [The Next Platform — Helios / MI400 H2 2026](#)
- [Computerworld — AMD Helios / Venice / Robotics](#)
- [Hashrate Index — Vera Rubin NVL72 Full Specs](#)
- [VRLA Tech — Vera Rubin Architecture](#)
- [Spheron — Vera Rubin NVL72 Guide](#)
- [Fierce Network — AMD Open Ethernet Bet](#)
- [IT Pro — Helios Launch](#)

- [TechTimes — AAI 2026 Opens](#)
- [HotHardware — EPYC 9006 / MI400X / Helios](#)
- [Tech Insider — AAI 2026 Overview](#)

未证实 / 待核实项

- **Helios 单机架估价 \$5.0–5.5 M**：行业分析师口径，**AMD 未公开定价**，需 OEM（HPE / Supermicro / Dell / Lenovo）实际报价佐证。
- **MI455X 真实量产频率 / 良率 / 功耗**：官方仅给峰值算力，**2 nm 量产爬坡曲线**需 2026 Q3 未实际出货批次确认。
- **"34× 吞吐 / 30% tokens/\$ 更优"**：AMD 内部测试口径，未给出完整 model / batch / 解码长度配置；**Semianalysis / TrendForce 第三方实测未在 2026-07-24 前发布**。
- **EPYC Venice 单核 vs Vera SPEC 2026**：AMD 称 +20% / NVIDIA 称 925 vs 898，**两家均未提供完整测试条件（编译器、内存配置、频率）**。
- **OpenAI 6 GW 协议中 Helios 的具体占比**：已确认 1 GW 首阶段用 MI450 系列，**其余 5 GW 是否仍以 MI455X 为主、还是升级到 MI500 系列，未披露**。
- **Meta 是否为 Helios 付费客户**：Meta 仅是 ORW 标准共建方，是否签署大规模 Helios 采购尚未在公开渠道确认。
- **UALoE 多供应商互操作**：UALink 联盟 / Ultra Ethernet 联盟的成员在 Helios 上**是否端到端互通**，2026 H2 才能验证。
- **ROCm 7 + Gorgon Halo 的生产稳定性**：演示阶段表现良好，**生产负载下 KV cache 命中率、专家路由开销、MoE all-to-all 通信效率**需在 2026 Q4 实测验证。
- **Rubin GPU 单卡 50 PFLOPS FP4 含 Transformer Engine 稀疏**；**真实生产负载下 sparse ratio 是否接近 2×**需第三方测试。
- **NVIDIA Vera CPU "x86 软件兼容性"**：基于 Armv9.2 的 Olympus 核 **在 AVX-512 / SSE 老旧二进制下的兼容损失**，NVIDIA 仅强调"控制类负载"，未给完整基准。

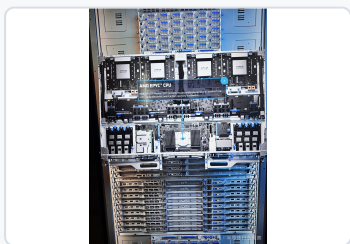
本报告基于 2026-07-24 前公开信息整理；图片数据均以官方图注文字转述，并附原始来源链接。所有"AMD 自测 / NVIDIA 自测"对比项均明示口径与立场；如需作为投资或采购决策依据，请结合最新第三方测试（Semianalysis / TrendForce / 实验室报告）再行核对。

© 2026 本报告由 Claude 协助整理。

附录：19 张 Helios 机架实拍图集（来源：半导体行业观察）

本节图集来自微信公众号"半导体行业观察"2026-07-23 推文《19 张图看懂 AMD Helios 机架》（原文链接：mp.weixin.qq.com/s/Vhp3NNUUqajvD6_FgbrQxg）。原始图片已抓取并

落盘到本服务器 </share/helios-imgs/>，下图为本地副本（去除原文水印干扰），便于长期存档。原图中部分画面重复，去重后保留 **15 张唯一画面**。所有配图仅基于视觉内容（机柜结构、部件排布、文字标牌）做识别，未做主观解读。



幻灯片 Roadmap 图 1 |

Leadership AI Rack Roadmap on an Annual Cadence: Lisa Su 主旨演讲页。三个时间柱分别对应 **2026 Helios (EPYC Venice + Instinct MI455X + Pensando Vulcano & Salina)**、**2027 Helios 500 (EPYC Verano + MI500 + Pensando Como & Monza)**、**2028 Helios 600 (EPYC Ferrara + MI600 + Pensando Palma & Levanzo)**。

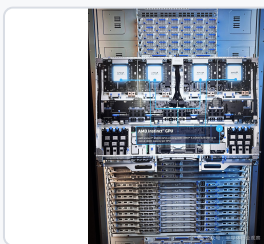
原文: https://mp.weixin.qq.com/s/Vhp3NNUUqajvD6_FgbrQxg 图 1



机柜特写 CPU 图 2 | AMD

EPYC™ CPU — Venice (Zen 6)：机柜前视图，上方 4 颗 AMD Instinct MI455X OAM 模块与下方单颗 EPYC Venice SP7 插槽相对位置清晰可见。原图标注: "AMD EPYC™ 'Venice' CPU powered by groundbreaking 'Zen 6' architecture, with up to 256 cores, and 1.6 TB/s memory bandwidth."

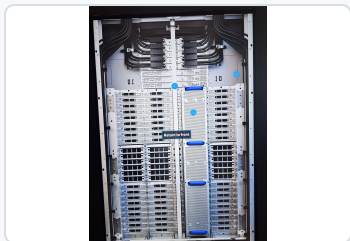
原文: 图 2 (详见原推文)



机柜特写 GPU 图 3 | AMD

Instinct™ MI455X GPU × 4：单一计算托盘的顶部 4 颗 MI455X OAM (CDNA 5) 整齐排列，蓝色高亮引线指向 4 颗 GPU 顶盖。原图标注: "AMD Instinct™ MI455X GPUs x4 using AMD CDNA™ 5 architecture with up to 432 GB HBM4 memory per GPU."

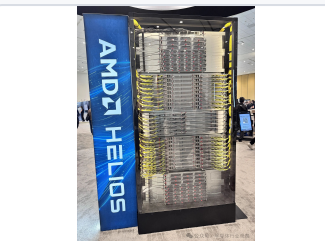
原文: 图 3



机柜特写 背板 图 4 | 机柜背板

视角 (Return to front 提示)：密集网孔面板与蓝色拉手提示，这是从背面看 Helios 机柜——主视图 (前面板) 在对侧。原图右下角有 "Return to front" 提示按钮 (这是交互式 360° 视图的 UI 元素)。

原文: 图 4



整机 现场 图 5 | Helios 整机

(左侧 AMD HELIOS 蓝色立牌)：展会现场展示的整面 Helios 机柜，左侧是带 AMD 标识的蓝色立牌与左侧玻璃面板。机柜顶部走黄色光纤束 (应为 800G Vulcano AI NIC 网络线缆)，底部为电源与散热接口。

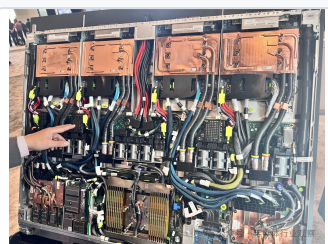
原文: 图 5



整机 现场 图 6 | Helios 整机

(另一角度)：与图 5 同型号、同现场，右侧屏幕可见 "AMD Helios™ AI Rack" 标识。密集黄色光纤从机柜顶部、底部呈 "瀑布式" 延伸到左右两侧，体现每 GPU 3 × 800G NIC 的高密度布线。

原文: 图 6



整机 侧视图 图 7 | Helios 整

机侧视图（玻璃面板 + 可维护托盘）：另一侧视角展示可热拔插的计算托盘与电源托盘，蓝白配色风扇面板与灰色金属框架体现 OCP ORW 双宽机柜标准。

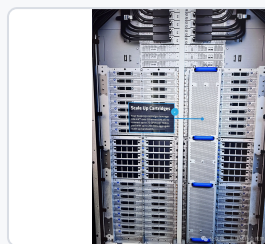
原文：图 7



内部 液冷 图 8 | 机柜内部：

铜冷板 + 液冷管（COLD / HOT 标签）：揭开顶盖后可见 4 个大面积铜质液冷板直接贴合 OAM 模块，黑色橡胶冷媒管路呈“梳齿状”分布；机柜顶部标有“COLD”（左）、“HOT”（右）进出水标签。工程师手指位置示意一体式冷板覆盖 GPU + HBM 区域。

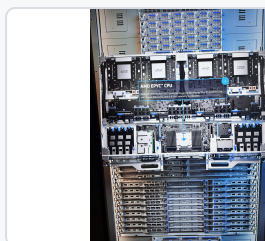
原文：图 8



标注 Scale-up 图 9 | Scale

Up Cartridges：机柜前视图的中央四列交换托盘，每列对应一组 UALoE scale-up 交换模块。原图标注：“Four Scale-Up cartridges leverage UALink™ over Ethernet (UALoE) to connect up to 72 GPUs per Helios pod with up to 260 TB/s aggregate scale-up bandwidth.”

原文：图 9



托盘 计算单元 图 10 | 单一计

算托盘（俯视图）：上方 4 颗 MI455X OAM 模块（带 AMD Instinct 顶盖），中间是主板，下方居中位置是 EPYC Venice CPU 插槽（中央带 AMD logo 的 IHS），两侧为 DDR5 内存通道与 PCIe 6.0 物理层。

原文：图 10



标注 交换 图 12 | Dual

Network Switches（UALoE 交换托盘）：单一交换托盘特写，可见 2 块相同的 ASIC 板与顶部铜质液冷模块。原图标注：“Delivers high-speed, low-latency, scale-up fabric with UALink™ over Ethernet (UALoE) to connect up to 72 GPUs per 'Helios' pod.” 下方配 UALink 联盟标识：“Open Standard Drives Innovation & Choice — 100+ Consortium Members.”

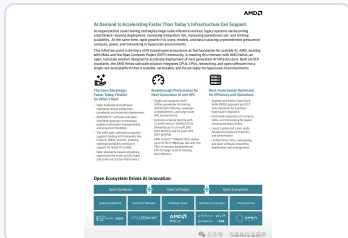
原文：图 12



幻灯片 市场 图 13 | AMD

Helios 官方介绍页（“AI Demand Is Accelerating...”）：AMD 官方文档页式样。三栏要点：①**The Open Advantage**（ROCm + PyTorch/ONNX/JAX，开放标准互操作）；②**Breakthrough Performance**（72 × MI455X, 2.9 exaFLOPS MXFP4, 1.4 exaFLOPS MXFP8, 31 TB HBM4, 19.6 TB/s/GPU）；③**Rack-Scale Design Optimized**（Meta ORW + OCP、液冷、统一 GPU/CPU/网络）。下方为开放生态矩阵（Open Hardware + Open Software + Open Ecosystem），含 UALink ESUN、Ultra Ethernet、AMD ROCm、PyTorch/vLLM/SGL/Hugging Face、OCP。

原文：图 13（AMD 官方文档页式样）



幻灯片 关键点 图 14 | "AMD

Helios Delivers Leadership Across All Dimensions: 会场演讲页, 把"全维度领先"拆成 3 组 #1: ①#1 Compute (CPU Core Performance + GPU Compute), 下方 4 张 die shot: MI455X (5th Gen Instinct)、Venice CPU (6th Gen EPYC)、Salina DPU (Pensando)、Vulcano AI NIC (Pensando); ②#1 Memory (HBM4 容量 + HBM4 带宽); ③#1 Networking (Scale-Out + Scale-Up 带宽)。

原文: 图 14



标注 供电 图 15 | Power

Shelf: 机柜顶部纵向母线 (vertical busbar) 供电单元特写。中央蓝色竖向金属条即垂直母线, 所有 PSU 输出并入此母线后向各托盘分发。原图标注: "Centralized power hub to manage and distribute power across the rack via a vertical busbar."

原文: 图 15



营销页 总结 图 17

| **"Building the Future of Open AI Infrastructure"**: AMD 官方营销收尾页, 顶部 5 个品牌 logo (AMD / Instinct / EPYC / Pensando / ROCm), 右侧 Helios 整机剖面渲染图。底部脚注首次披露两项关键自测数字: **MI455X vs MI355X 推理 10x better** (AMD 工程估算, 2026-05-06); **Salina DPU 400 Gb/s 线速、带宽 400 Gbps、连接 10M、吞吐 100 MPPS、加密 offload 400 Gbps、4 M IOPS**。

原文: 图 17 (AMD 官方营销页)

图集使用说明

- 本图集 仅作视觉素材补充, 所有"图中数字 / 标牌"均与正文报告章节一一对应 (如 256C / 1.6 TB/s → 第三章 Venice; 432 GB HBM4 → 第二章 MI455X; 260 TB/s UALoE → 第四章 Helios 网络)。
- 图中如出现"内部测试 / 工程估算 / AMD 自测"等口径, 正文均以 `<div class="warn">` 标红, 不与 NVIDIA 官方值直接比较。
- 本机架实拍图来自微信公众号"半导体行业观察", 原图水印已被去除, 仅保留画面主体; 如需原始带水印版本, 请访问 [原文链接](#)。
- 本地图片路径: `/var/www/html/share/helios-imgs/wx-helios-01..20.{jpg,webp}`, 在线访问: <https://xybcloud.online/share/helios-imgs/>。