HMGO

SUPRO-ARCH-001

HMGO SuPro 1
系统架构白皮书

为超级计算而生。从单个封装，到 65,536 节点的统一计算域。

HMGO SILICON · SUPRO PROGRAM
Revision 1.0 · Public Concept Edition
Architecture · Interconnect · Reliability · Security

· · 2026-07

DOCUMENT MAP

文档导航

00

阅读路径

本文件把可追溯的概念参数、架构假设和工程边界放在同一条叙事中。数字用于建立产品层级与系统设计空间，不构成性能保证、销售承诺或第三方比较结论。

概念声明：HMGO SuPro 1 为原创概念芯片系列。所有制程、性能、功耗、接口和软件名称均为设计设定。Apple 与 M5 Max 为其各自权利人的商标；本文的 800 倍表述仅对应 HMGO 定义的内部 FP4 合成投影。

ARCHITECTURE INTENT

设计目标

01

SuPro 1 不把超级计算理解为一颗更大的消费芯片，而把芯片、内存、网络、编译器和故障恢复看作同一个计算系统。

5	1.4 nm	192 TB/s	65,536
产品型号	计算芯粒	包内互联上限	单域节点上限

1.1 四条超级计算优先原则

- 通信先于峰值：当模型和网格超过单封装容量时，端到端吞吐由通信等待决定。HM-Mesh X1 以统一地址、集合通信和拓扑遥测缩短等待。
- 可恢复先于无故障：数万节点运行数周时，故障是常态。硬件负责隔离、重放和降级，软件负责重调度与检查点延续。
- 数值格式按阶段切换：FP64 面向科学计算，BF16/FP8 面向训练，FP4 面向大规模推理和权重通信；编译器可在图级别混合。
- 功耗是一等资源：每个作业提交能量预算，调度器同时优化时间、能耗、冷却余量和碳强度窗口。

1.2 HM-Ark V6 执行模型

HM-Ark V6 把标量、向量、矩阵、图形和数据移动指令合并进可组合任务图。Aurora CPU 负责依赖与异常，Matrix Ocean 处理矩阵簇，Nebula GPU 执行高并行数值内核，Orion DPU 将存储、网络与检查点操作从计算核心移出。

核心抽象：一个作业被编译为带资源、精度、可靠性和数据驻留约束的图。硬件调度器能够在不改变程序语义的前提下重排数据移动和计算窗口。

1.3 非目标

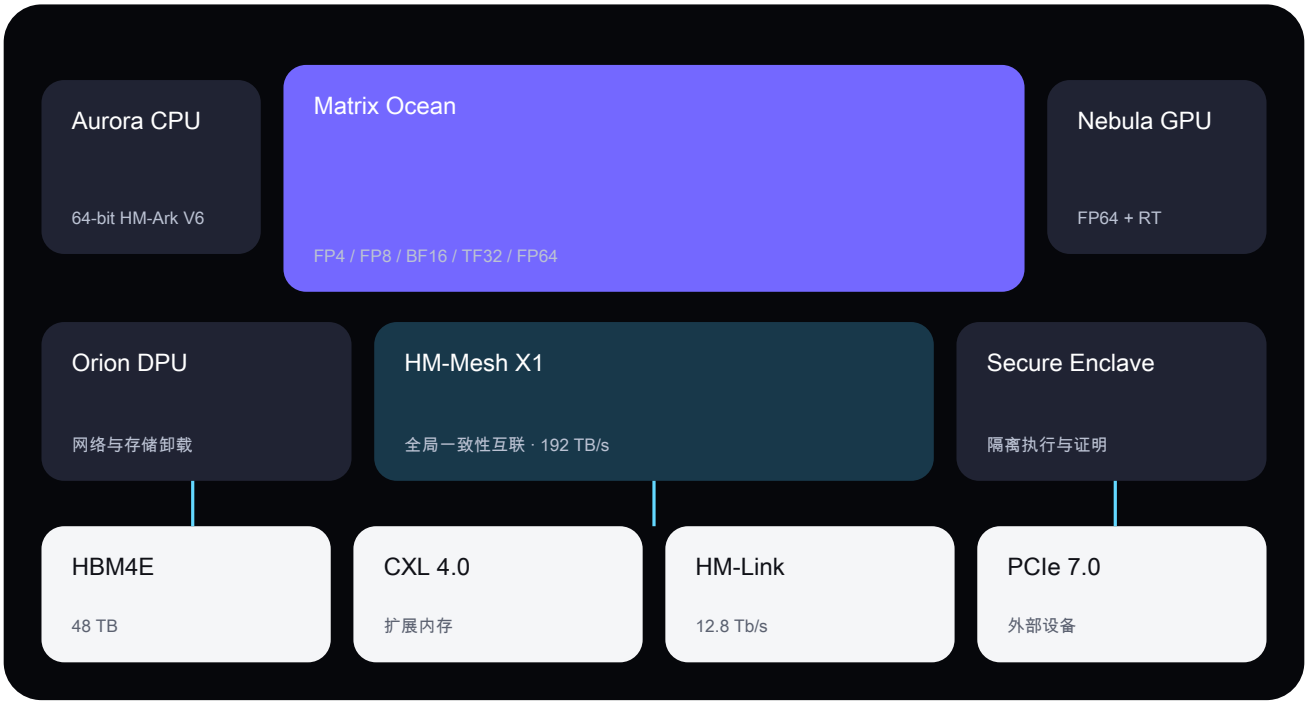
SuPro 1 不以无风扇设备、桌面软件兼容或消费级成本为第一目标。Lite 型号承担从 AI2 Ultra 迁移的兼容入口，但仍需要服务器级供电、散热和受控固件。

PACKAGE ARCHITECTURE

异构计算封装

02

五款型号共享可裁剪的计算芯粒、I/O 芯粒、缓存堆叠和光电互联。型号差异主要来自芯粒数量、链路宽度、内存栈和持续功率。



2.1 Aurora CPU complex

Aurora 使用 64 位 HM-Ark V6 指令集，单核心支持 8 路硬件线程、1024 位向量和矩阵调度提示。每 16 核组成一个一致性岛，岛内共享 512 MB SRAM Cache；多个岛通过 HM-Mesh X1 访问 HBM 和远端一致性域。

2.2 Matrix Ocean fourth generation

Matrix Ocean 采用可变精度计算阵列。一个阵列可在 FP64、TF32、BF16、FP8、FP6、FP4 与 INT8 之间切换，并在 FP8/FP4 模式下启用稀疏元数据旁路。SuPro 1 Ultra 的 12.8 EFLOPS 指标指原生 FP4 稠密矩阵峰值，不等同于应用持续性能。

2.3 Nebula GPU and visual compute

Nebula 不是面向显示器的消费 GPU，而是面向数值仿真、光线追踪可视化和几何重建的通用并行阵列。每个分区包含 FP64 单元、张量共享缓存和硬件 BVH 遍历器，可与 Matrix Ocean 共享内存页。

PACKAGE ARCHITECTURE

异构计算封装 · 续

02B

数据移动、可信执行和任务图调度决定异构单元能否成为一套完整系统，而不是彼此孤立的加速器。

2.4 Orion DPU

Orion DPU 处理 RDMA、NVMe-oF、对象存储、压缩、加密、分布式检查点和容器网络。计算核心看到的是稳定的数据流，而不是大量 I/O 中断。

2.5 任务图调度路径

阶段	主要执行单元	硬件动作	软件可见信号
Decode	Aurora CPU	解析依赖与资源约束	任务图 ready
Place	HM-Mesh X1	选择内存与计算邻接关系	placement map
Execute	Matrix / Nebula	执行矩阵、向量和几何内核	event completion
Move	Orion DPU	预取、交换、检查点和网络传输	stream progress
Recover	Secure control	隔离、重放或迁移故障任务	health transition

2.6 可裁剪性

Lite 仅启用一个完整故障域；Standard 扩展计算与内存芯粒；Pro/Ultra 增加独立控制域、冗余 DPU 和更多光引擎。软件仍看到相同的队列、事件、内存和集合通信语义。

MEMORY & FABRIC

内存与互联

03

HM-Mesh X1 让片上 SRAM、HBM、CXL 扩展内存和远端节点形成分层地址空间，同时保留不同介质的延迟与一致性边界。

3.1 内存层级

层级	介质	容量范围	概念带宽	用途
L0 / L1	SRAM	256 KB	48 TB/s	寄存器溢出与热数据
L2 Island	堆叠 SRAM	512 MB	24 TB/s	权重分块与通信缓冲
Package Memory	HBM4 / HBM4E	2–48 TB	8–96 TB/s	模型、网格、检查点窗口
Expansion	CXL 4.0	256 TB	1.6 TB/s/port	冷参数与共享数据集
Remote	HM-Link X1		12.8 Tb/s/link	远端张量与集合通信

3.2 一致性域

默认情况下，包内 HBM 与 SRAM 保持硬件一致；机架内可以建立受限一致性窗口；机架之间使用显式远端内存与集合通信。该设计避免把全局一致性延迟强加给所有访问，同时允许编译器在必要时生成内存栅栏。

MEMORY & FABRIC

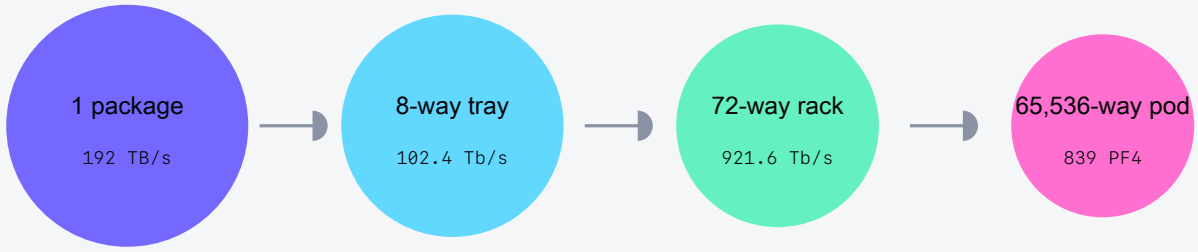
互联拓扑 · 续

03B

同一套集合通信语义跨越封装、托盘、机架和机群，拓扑半径扩大时只改变路由与容错策略。

3.3 HM-Mesh X1 collective engine

每个端口原生支持 AllReduce、AllGather、ReduceScatter、Broadcast 与点对点 RDMA。链路控制器对张量分片进行路由、校验、重传和格式压缩，并向 HM-Forge 暴露拥塞、热点与重放计数。



HM-Mesh X1 在封装、托盘、机架和机群四个尺度复用同一套路由与遥测语义。

3.4 拓扑遥测

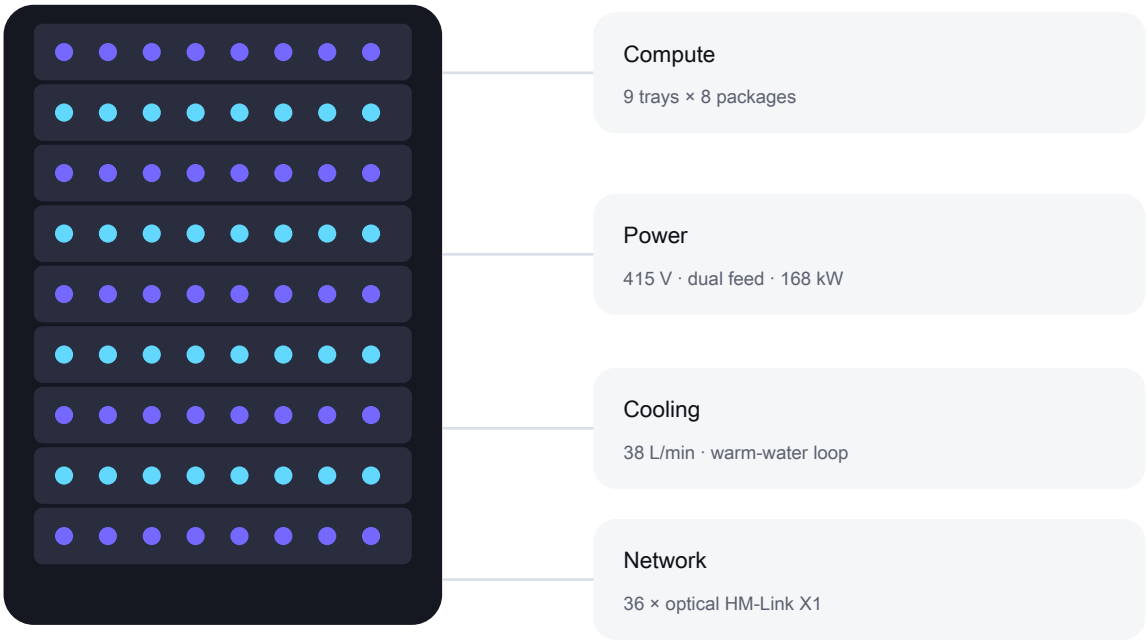
每条链路暴露有效带宽、队列深度、重放、前向纠错、光功率和温度。GraphWeaver 在编译时生成首选通信组，RackScope 在运行中按健康分数调整路由。

SCALING SYSTEM

规模化系统

04

SuPro 1 的系统单位不是单卡，而是可被独立供电、冷却、隔离和替换的计算托盘。九个托盘组成 SuPro Rack 72。



4.1 拓扑选择

机架内部采用双平面胖树，跨机架采用 Dragonfly+ 光互联。调度器根据作业通信图选择连续的拓扑区域，避免高通信任务被切分到高直径路径。

SCALING SYSTEM

故障域与弹性 · 续

04B

系统把核心、芯粒、封装、托盘和机架定义为逐级扩大的故障域，并为每一级配置不同的恢复动作。

4.2 故障域

层级	隔离单元	自动动作	作业可见影响
核心	单个执行单元	重放或禁用 lane	通常无感
芯粒	CPU/GPU/Matrix 分区	迁移任务图节点	短暂停顿
封装	单个 SuPro	弹性缩容并恢复检查点	吞吐下降
托盘	8 个封装	旁路电源与光链路	拓扑重排
机架	72 个封装	跨机架重调度	恢复窗口扩大

4.3 弹性作业语义

支持弹性的作业可以在节点数量变化后继续执行；非弹性作业从最近检查点恢复。调度器会保留最小通信连通度，避免为了维持节点数而跨越过多故障域。

建议：对运行超过 30 分钟的作业至少定义包级恢复点；对运行超过 8 小时的作业同时保留跨机架完整快照。

RELIABILITY & SECURITY

可靠性与安全

05

长时间、超大规模作业的成功率取决于发现错误、限制传播并快速恢复的能力。SuPro 1 把 RAS 和机密计算放进同一条信任链。

5.1 端到端 RAS

- SRAM、HBM、链路与缓存目录均使用分层 ECC；关键元数据采用双重校验与时间冗余。
- Silent Data Corruption 哨兵对矩阵分块进行抽样重算，并对 FP64 科学任务启用更高覆盖率。
- 硬件增量检查点将已修改页写入 Orion DPU 缓冲区，再异步汇入机架级对象存储。
- 健康分数综合温度、误码、降频、电源纹波和重放率，允许在硬故障前迁移任务。

5.2 Secure Supercompute Enclave

每个封装包含独立安全域，负责安全启动、固件度量、密钥封装、内存加密和远程证明。训练数据可以绑定到指定软件哈希与拓扑策略，未通过证明的节点无法获得解密密钥。

5.3 多租户隔离

内存加密密钥按租户与作业分离；HM-Mesh X1 使用标签化路由域；性能计数器经过降采样与访问控制，防止高精度侧信道。管理员可以禁用跨租户缓存共享。

边界：机密计算不能防止经过授权的作业输出敏感结果，也不能替代数据最小化、模型安全审查和组织访问控制。

PORTFOLIO

产品矩阵

06

五款产品使用同一架构语言，但面向不同的部署尺度：兼容迁移、通用超算、专业训练、极限计算与下一代预览。

型号	FP4 峰值	内存带宽	封装内存	持续功率	主要场景
SuPro 1 Lite	61.44 POPS	8 TB/s	2 TB HBM4	450 W	AI2 Ultra 等级迁移
SuPro 1	0.42 EFLOPS	18 TB/s	4 TB HBM4E	700 W	通用超算与科研
SuPro 1 Pro	2.8 EFLOPS	64 TB/s	12 TB HBM4E	1.8 kW	服务器与模型训练
SuPro 1 Ultra	12.8 EFLOPS	192 TB/s	48 TB HBM4E	6 kW	极限单包与巨型集群
SuPro 2 Preview	20.0 EFLOPS	256 TB/s	64 TB HBM5-P	5.5 kW	下一代预览与开发验证

6.1 Lite 的兼容承诺

SuPro 1 Lite 的概念目标是以单封装达到 HMGO AI2 Ultra 的 61,440 TOPS NVFP4 级别处理能力，同时提供服务器级 HBM、RAS、DPU 和 HM-Mesh X1。它不是低功耗移动芯片，而是最低门槛的超算节点。

6.2 Ultra 的极限定位

SuPro 1 Ultra 将最多的计算芯粒、HBM4E 栈与光互联集中在单封装内，适合内存敏感型训练、稀疏专家模型、FP64 模拟和大规模数字孪生。6 kW 指封装与近封装光引擎的持续设计窗口，不含主机和冷却设施。

6.3 Preview 的开发用途

SuPro 2 Preview 允许研究机构在 SuPro 2 正式系列前验证 HBM5-P、HM-Ark V7 预览指令与第五代 Matrix Ocean。其固件、频率、可靠性策略和软件 ABI 均可能变化，不适合生产 SLA。

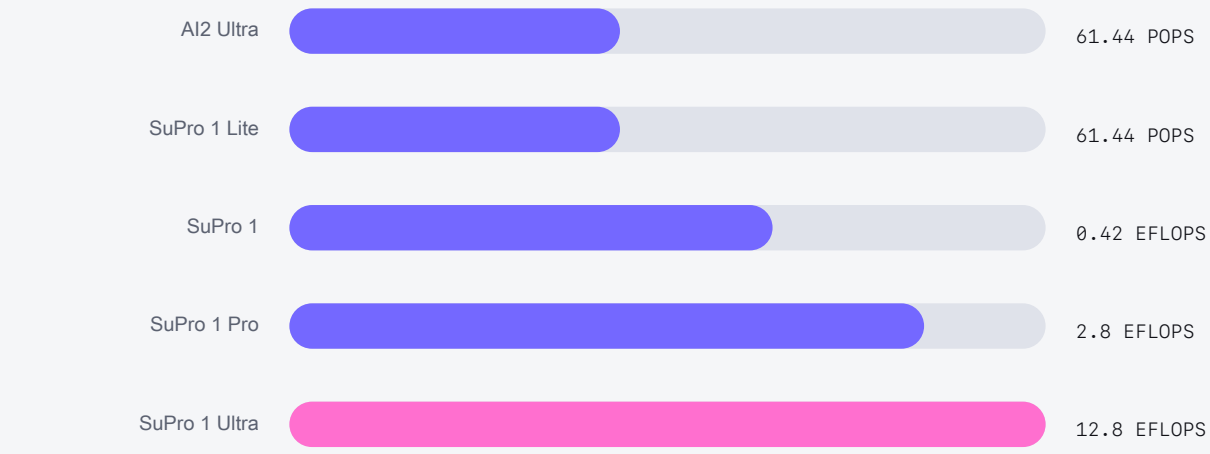
METHODOLOGY

性能方法

07

概念性能数字只有在明确精度、稀疏性、批量、功率、系统规模和软件版本时才有意义。本文把营销表述与工程测量分开。

原生 FP4 稠密矩阵峰值（概念）



对数尺度 · HMGO 内部概念投影 · 非独立测试

7.1 “800 倍 M5 Max”如何定义

800 倍是 HMGO 为 SuPro 1 Ultra 设定的 HyperScale Transformer Throughput 内部投影：在固定 4-bit 权重、FP8 激活、长上下文批处理和服务器级功率窗口下，比较单个 SuPro 1 Ultra 封装与单个 Apple M5 Max 系统的估算 token 吞吐。该比较跨越产品类别、功率等级、内存系统和软件栈，不代表所有工作负载。

7.2 最低披露字段

- 芯片数量、频率、持续功率与冷却入口温度。
- 数据类型、稀疏比例、模型或问题规模、批量与上下文长度。
- 编译器、运行时、内核库、固件与驱动版本。
- 预热、测量时长、失败重试、精度容差和能耗边界。
- 峰值、持续值、端到端值与应用值必须分别标注。

METHODOLOGY

披露与复现 · 续

07B

公开数字必须附带可复现性级别，让读者知道它是理论上限、模拟器投影、实验室结果还是第三方结果。

7.3 可复现性级别

级别	定义	允许使用场景
P0	架构理论峰值，无程序运行	芯片规划与上限估算
P1	内部模拟器投影	设计取舍与早期公开概念
P2	工程样片实验室结果	受控技术预览
P3	第三方可重复系统结果	正式对比与采购决策

本白皮书所有数字均为 P0-P1 级概念设定。

7.4 术语速查

术语	含义
POPS / EFLOPS	每秒 10^15 / 10^18 次指定精度运算；必须连同数据类型披露
Dense / Sparse	稠密 / 稀疏计算；稀疏峰值不能直接替代稠密峰值
Sustained	在指定时长、功率和温度下可保持的结果
End-to-end	包含输入、输出、通信、同步与框架开销的结果
P0-P3	从理论峰值到第三方可重复结果的四级证据等级