

SUPRO-DEPLOY-001

HMGO SuPro 1 平台部署与开发指南

从首台节点启动，到 65,536 节点作业。面向平台、AI 与 HPC 团队的工程路线。

HMGO SILICON · SUPRO PROGRAM

Revision 1.0 · Public Concept Edition

Site Readiness · HM-Forge · Distributed Training · Operations

· 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。

01	部署前规划	容量、功率、冷却、网络与故障域
02	上架与启动	物理安装、固件、证明和健康基线
03	HM-Forge 软件栈	驱动、编译器、运行时、框架和集群服务
04	训练与推理	并行策略、内存布局和通信优化
05	科学计算	FP64、MPI、网格与可视化路径
06	可观测与恢复	遥测、检查点、故障注入和升级
07	安全运营	租户、密钥、供应链和审计
08	验收清单	从站点准备到生产交接

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。

PLAN

部署前规划

01

先确定工作负载、数据位置和服务目标，再决定芯片型号与节点数。SuPro 1 的系统成本通常由内存、通信、设施与运维共同决定。

1.1 工作负载画像

问题	需要记录的字段	影响的设计选择
模型/网格规模	参数量、状态量、序列长度、精度	HBM 容量与芯片型号
通信比例	AllReduce、点对点、存储 I/O 占比	节点拓扑与 HM-Link 数量
运行时间	单作业时长、并发数、截止时间	检查点频率与容错等级
数据约束	驻留区域、加密、保留和删除	安全域与存储架构
服务目标	吞吐、延迟、可用性与恢复目标	冗余、备件与调度策略

1.2 容量估算

训练内存近似：权重 + 梯度 + 优化器 + 激活 + 通信缓冲 + 检查点窗口。不要只按参数量乘位宽。对于混合专家模型，还要计入专家不均衡和路由缓存。

1.3 设施预算

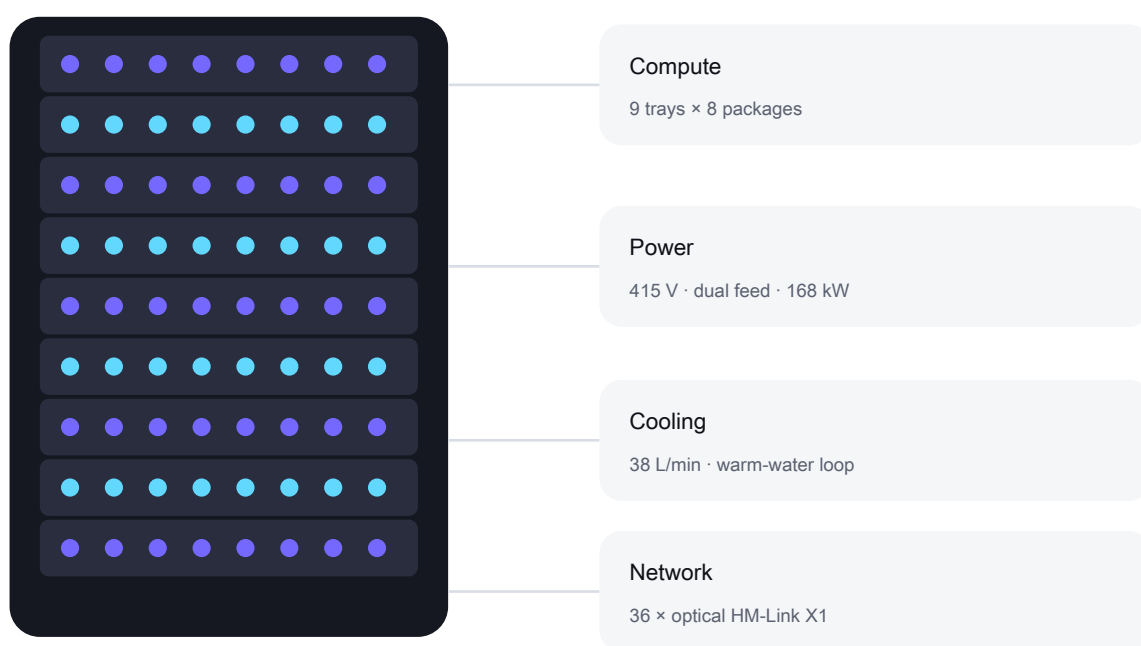
- 以 80% 设施利用率规划持续功率，以 boost 功率检查瞬态与保护器件。
- 为泵、交换、存储、主机 CPU、DPU 和漏液检测保留 18-28% 非芯片功率。
- 按最坏环境温度验证冷却能力；热回收模式需要独立验证回水温度。
- 每 12 个机架预留一个可空载启动的维护电力域，支持滚动更换与恢复演练。

INSTALL

上架与启动

02

物理安装的目标是建立可重复的供电、冷却、光互联和管理路径。首台节点启动前必须冻结布线图、设备身份和固件基线。



2.1 上架顺序

- 1 确认机架水平、接地、承重、双路供电相序和冷却压差。
- 2 从底部向上安装托盘，连接主供电、管理网络和两条 HM-Link 光平面。
- 3 完成冷板泄漏测试与 30 分钟静态压力保持，再启用循环泵。
- 4 读取每个封装的硬件身份和制造度量，导入站点资产系统。
- 5 启动管理控制器，完成固件签名验证和 Secure Enclave 远程证明。
- 6 运行 20 分钟低功率内存、链路与传感器基线测试。

INSTALL

启动与健康基线 · 续

02B

首轮基线把温度、链路、内存、电源和证明状态冻结为后续运维的比较起点。

2.2 首次健康基线

指标	正常范围	阻止上线的条件
HBM corrected error	0-2 / TB-hour	持续上升或同一页重复
HM-Link replay	< 1e-12 / bit	单链路高于基线 10 倍
冷板温差	2-8°C	任一封装 > 12°C
电源纹波	< 1.5%	持续 > 2.5%
证明状态		任一固件哈希未知

若证明失败，不要通过关闭验证来继续上线。隔离节点，核对固件来源、设备身份和策略版本。

2.3 低功率到满功率放量

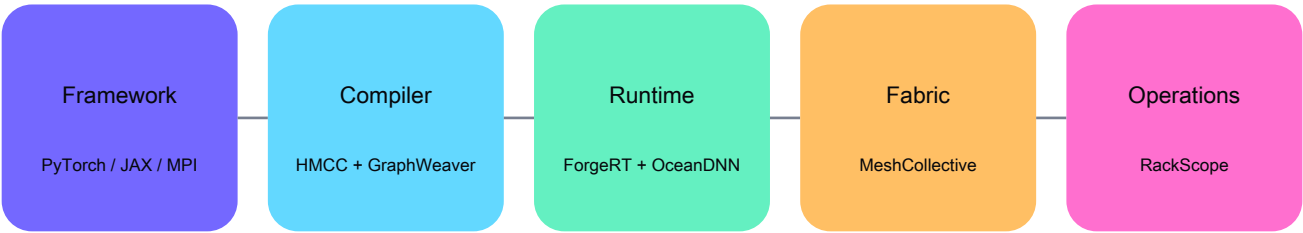
先以 15% 功率运行内存与链路测试，再依次提升到 40%、70% 和 100%。每一级至少保持 10 分钟，并验证冷板温差、电源纹波和链路重放没有趋势性恶化。

SOFTWARE STACK

HM-Forge 软件栈

03

HM-Forge 将单节点编译、分布式通信、集群调度和运维遥测组合成一套版本化软件栈。生产环境应使用经过资格验证的完整版本束。



3.1 组件

组件	职责	版本策略
HMCC	HM-Ark V6 编译、自动混合精度、内核融合	月度稳定版
GraphWeaver	任务图切分、内存计划、重计算与容错点	随 HMCC
OceanDNN	矩阵、注意力、稀疏专家与量化内核	季度资格版
MeshCollective	拓扑感知集合通信与拥塞控制	与固件绑定
ForgeRT	队列、流、事件、内存和任务图执行	长期 ABI
RackScope	遥测、告警、功耗与健康管理的	独立服务

SOFTWARE STACK

环境与版本束 · 续

03B

生产环境把固件、驱动、编译器、运行时和内核库作为一个资格版本束管理，避免单组件漂移。

3.2 环境基线

项目	概念规格
操作系统	HMGO ServerOS 3.0 LTS or qualified Linux distribution
容器	OCI 1.2 · signed image required for confidential mode
调度器	Forge Scheduler · Slurm adapter · Kubernetes device operator
编程模型	Forge Graph · C++ · Python · OpenMP offload · MPI
框架	PyTorch adapter · JAX adapter · ONNX runtime adapter

3.3 版本束

驱动、固件、OceanDNN、MeshCollective 与 HMCC 形成资格矩阵。不得单独升级其中一个关键组件并直接进入生产；所有例外必须经过 24 小时混合负载与故障注入测试。

3.4 推荐安装配置

配置	用途	更新节奏	回滚目标
LTS	生产训练、科学计算、长期服务	季度	30 分钟
Current	新模型、新内核和性能验证	月度	60 分钟
Preview	SuPro 2 Preview 与 ABI 预适配	按需	重新镜像

AI WORKLOADS

训练与推理

04

性能优化的顺序应是：先让模型驻留并正确运行，再降低通信和重计算，最后才追求峰值利用率。

4.1 并行策略选择

策略	适合情况	主要代价	SuPro 建议
Data Parallel	模型可在单节点驻留	梯度 AllReduce	机架内优先
Tensor Parallel	单层权重超出单封装	高频集合通信	限制在低直径域
Pipeline Parallel	层数多、阶段均衡	气泡与激活传输	使用交错微批
Expert Parallel	MoE 稀疏激活	All-to-All 热点	启用拓扑感知路由
Sequence Parallel	超长上下文	边界通信	与张量并行组合

4.2 推荐优化顺序

- 1
- 建立 FP32 或 BF16 正确性基线并保存参考输出。
- 2
- 启用自动混合精度，对敏感层设置显式精度护栏。
- 3
- 使用 GraphWeaver 内存计划降低峰值，必要时启用激活重计算。
- 4
- 按 RackScope 拓扑映射通信组，检查链路热点和尾延迟。
- 5
- 启用 fused attention、fused optimizer 与专家预取。
- 6
- 在目标功率和冷却条件下进行至少 60 分钟持续测试。

4.3 推理服务

在线推理以分离 prefill 与 decode 池为默认架构。prefill 池使用高带宽 SuPro 1 Pro/Ultra，decode 池按延迟与能耗使用 Lite/Standard。KV cache 可通过 CXL 扩展，但必须测量迁移对 P99 延迟的影响。

量化要求：FP4/FP8 仅在达到任务精度门槛后上线。每个模型版本应保存校准集、误差报告、拒答和安全评测结果。

HPC WORKLOADS

科学计算

05

FP64 峰值只是起点。网格划分、访存模式、集合通信与数值稳定性决定科学应用的可用性能。

5.1 端口路径



5.2 数值与性能检查

- 保留 CPU 参考路径与至少一个小规模确定性测试。
- 对 reduction 顺序变化建立 ULP 或领域容差，而不是要求逐位一致。
- 使用结构化网格时优先打包 halo；非结构化网格应重排索引以提高局部性。
- 并行 I/O 应将大量小写入聚合到 Orion DPU，避免元数据服务成为瓶颈。
- 可视化抽样应在 Nebula GPU 上就地生成，减少全量结果出站。

5.3 FP64 模式

SuPro 1 Pro 与 Ultra 提供更高比例的 FP64 管线。启用 Strict64 会关闭部分跨指令融合，保留舍入和异常语义；启用 Fast64 可允许融合乘加和重排。性能报告必须注明模式。

模式	语义	适合场景
Strict64	严格异常、受控重排、可审计	验证、关键物理模型
Balanced64	融合乘加、有限重排	大多数生产仿真
Fast64	激进融合、近似函数	探索、可视化、筛选

OBSERVE & RECOVER

可观测与恢复

06

运维目标不是让告警数量最少，而是尽早发现趋势、限制故障域并让作业在可预测时间内恢复。

6.1 四层遥测

层级	关键指标	保留建议
芯片	温度、频率、ECC、降压、矩阵利用率	高频 7 天 / 聚合 1 年
互联	带宽、重放、拥塞、路由偏斜	高频 30 天 / 聚合 1 年
作业	阶段时间、通信比、重计算、能耗	作业期 + 2 年
设施	供电、流量、压差、进回水温度	高频 90 天 / 聚合 5 年

6.2 检查点策略

检查点间隔应由平均故障间隔、写入时间和恢复成本共同决定。Pro/Ultra 推荐使用两级检查点：高频增量写入机架级缓冲，低频完整快照写入跨故障域对象存储。

6.3 维护窗口

- 1 冻结新作业，等待可抢占任务进入最近检查点。
- 2 导出 RackScope 健康快照并确认备份控制平面可用。
- 3 按托盘滚动升级，完成启动证明和 20 分钟基线测试。
- 4 执行故障注入：断链、掉电、内存页错误和 DPU 切换。
- 5 恢复 10% 流量，观察 60 分钟后逐步放量。

不要在同一维护窗口同时更新固件、光交换和调度策略；否则出现回归时无法定位责任变量。

SECURE OPERATIONS

安全运营

07

SuPro 1 提供硬件根信任，但生产安全仍取决于身份、密钥、软件供应链、网络策略与人员流程。

7.1 最小权限角色

角色	允许操作	明确禁止
Facility Operator	电力、冷却、物理维护	读取租户数据与模型
Platform Admin	固件、驱动、调度、容量	解密训练数据
Security Admin	证明策略、密钥、审计	提交生产作业
Tenant Admin	项目、用户、配额、镜像	修改平台固件
Auditor	只读日志与证明报告	修改任何策略

7.2 软件供应链



7.3 密钥与数据

训练数据密钥由外部 KMS 管理，只在封装、固件、容器和拓扑证明同时满足策略后释放。检查点使用独立密钥与版本，删除流程同时撤销密钥并记录介质处置。

SECURE OPERATIONS

事件响应 · 续

07B

事件响应优先保护证据和数据密钥，再恢复容量。所有隔离与撤销动作必须记录发起人、时间、范围和恢复条件。

7.4 事件响应

- 证明异常：隔离节点、冻结密钥、保存度量和管理控制器日志。
- 链路异常：切换到备用平面、保留光模块与端口的关联记录。
- 数据泄漏怀疑：停止作业、撤销租户密钥、保存审计链并通知负责人。
- 供应链异常：阻止对应签名者，回滚到上一个资格版本束并重新证明。

7.5 事件等级

等级	示例	首次响应	默认动作
SEV-1	数据泄漏、证明绕过、跨租户访问	5 分钟	冻结密钥与受影响计算域
SEV-2	固件未知、持续链路错误、异常管理员操作	15 分钟	隔离节点并保存证据
SEV-3	单节点退化、可恢复错误趋势	4 小时	迁移作业并安排维护

ACCEPTANCE

验收清单

08

生产交接必须用可复现证据证明设施、硬件、软件、性能、恢复和安全均处于预期边界。

8.1 站点与硬件

- 供电、接地、短路保护、ride-through 与双路切换测试通过。
- 冷却流量、压差、进回水温度、漏液检测和紧急关闭测试通过。
- 所有节点身份、序列号、位置、光链路和资产记录一一对应。
- HBM、链路、DPU、传感器与安全证明基线无阻止上线项。

8.2 软件与性能

- 资格版本束哈希已冻结，容器镜像签名和 SBOM 可追溯。
- 单节点、托盘、机架和跨机架基准达到约定持续值，而非仅峰值。
- 目标模型或科学应用通过正确性、精度、稳定性和能耗验证。
- 调度、公平性、配额、抢占、日志和成本归集在多租户场景可用。

ACCEPTANCE

恢复、安全与签署 · 续

08B

最终交接要求恢复演练、安全策略和责任签署全部可追溯，不能只凭单次性能测试判定上线。

8.3 恢复与安全

- 完成断链、节点掉电、托盘隔离、DPU 切换和机架失效演练。
- 检查点在目标 RPO 内生成，并在目标 RTO 内从独立故障域恢复。
- 未通过证明的节点无法获得数据密钥；撤销流程已实际演练。
- 运维角色、应急联系人、备件位置和升级窗口完成书面交接。

签署项	责任人	证据位置	状态
站点准备	Facility Lead	Site Readiness Report	☐
硬件资格	Platform Lead	SVT Bundle	☐
性能与正确性	Workload Lead	Acceptance Results	☐
安全与证明	Security Lead	Attestation Pack	☐
生产交接	Service Owner	Handover Record	☐

完成标准：所有阻止上线项关闭，剩余风险有负责人、截止日期和缓解措施；正式流量从 10% 开始分阶段放量。

本指南描述概念平台，不替代正式安装手册、电气规范、数据中心安全规范或当地法规。