豪泓科技
HAOHONG TECHNOLOGY

H2 AIDM · L0 算力层 · 白皮书 01

H2 GPU Combo 有什么卡，就把什么卡用好。

量体裁衣：按客户现有的算力设备定方案。先给每张卡做画像，再按卡定制量化、并发与部署，全自动拉起与自愈。

上海豪泓信息技术有限公司 | v5.1 | 2026 年 11 月

数字口径：实测 / 生产实证 / 已上线 / 规划测算，见各处标注。规划值以实际运行数据为准。

H2 GPU Combo （H2GC）

有什么卡，
就把什么卡用好。

量体裁衣，而不是让客户为了跑模型去换一批新卡。H2GC 先给每张卡做画像，再按它的显存、带宽和功耗定制方案：存量卡、工作站单卡、数据中心节点，各有各的做法。



同一张 RTX PRO 4500：可并发两路 RayWorld Live 直播，两路交替占用 GPU，每路每 40 毫秒出一帧（25fps），0 掉帧 生产实证；也可单独运行 RayWorld R1 生产实证。GPU 时间柱与显存分区为示意。

2 路 · 25fps

单张 RTX PRO 4500 并发实时直播，0 掉帧。

生产实证

单卡

同一张 RTX PRO 4500 也可单独运行 RayWorld R1 世界模型。

生产实证

96 路

整栈结果：4 卡 H200 实测 96 路并发（智能体长上下文负载）。

实测

2,000 人

一个 8×H200 节点服务的注册用户数。

规划

量体裁衣 · 按卡定制

先看你有什么卡，再决定怎么做。

常见的做法是先定方案，再要求客户按方案采购指定的卡。H2GC 反过来：方案跟着客户手里的设备走。

01

自动画像

识别每张卡的架构、显存、带宽与功耗上限。

02

匹配方案

按卡选择量化精度、并发路数与部署模板。

03

拉起与自愈

标准镜像一键拉起；持续监测掉帧与延迟抖动，异常节点自动隔离、切流、重启。

同一套技术，三种做法

存量卡、工作站单卡、数据中心节点，都有对应的做法。

档位	典型设备	H2GC 优化组合	适合 / 已验证的负载
存量卡	RTX 4090、L4、A10 等 24GB 级显卡	W4A8 / INT4 量化 · 分块视频解码 · 主机内存卸载	**量化后的文本模型、流式语音合成、AI 音乐、图像、轻量视频** 适配方案
工作站单卡	RTX PRO 4500 Blackwell（32GB · 200W）	原生 FP4 · 算子融合 · PCIe 5.0 分级预取 · 多流配额隔离	**可并发两路 RayWorld Live 25fps，0 掉帧；也可单独运行 RayWorld R1** 生产实证
数据中心节点	8×H200 标准节点、专属节点	张量 / 专家并行 · KV 分级池 · 会话亲和（通道锁定）	**智能体长上下文推理：整栈 4 卡实测 96 路并发** 实测 **；8×H200 节点 2,000 名注册用户** 规划

问题

卡买了，却跑不动、赚不到。

重型模型、世界模型和实时直播难以商用，往往不是卡不够多，而是被三堵墙挡住了。



显存墙

常规做法按最大长度为每个请求预留整段连续的 KV 缓存，空间浪费、碎片累积；再叠加视频解码，24GB、32GB 的卡很快就装不下。

H2GC：分页式显存池，冷数据分级驻留到主机内存。



延迟墙

数据预取、模型计算、音频处理、视频编码串行执行，GPU 大量时间在空等；多路业务互相抢占，掉帧、尾延迟升高。

H2GC：全异步流水线重叠，多路计算流按配额隔离。



异构碎片墙

同一个机房里 4090、A10、L4、RTX PRO、H200 混在一起，没有按卡匹配的量化模板和一键部署，利用率上不去。

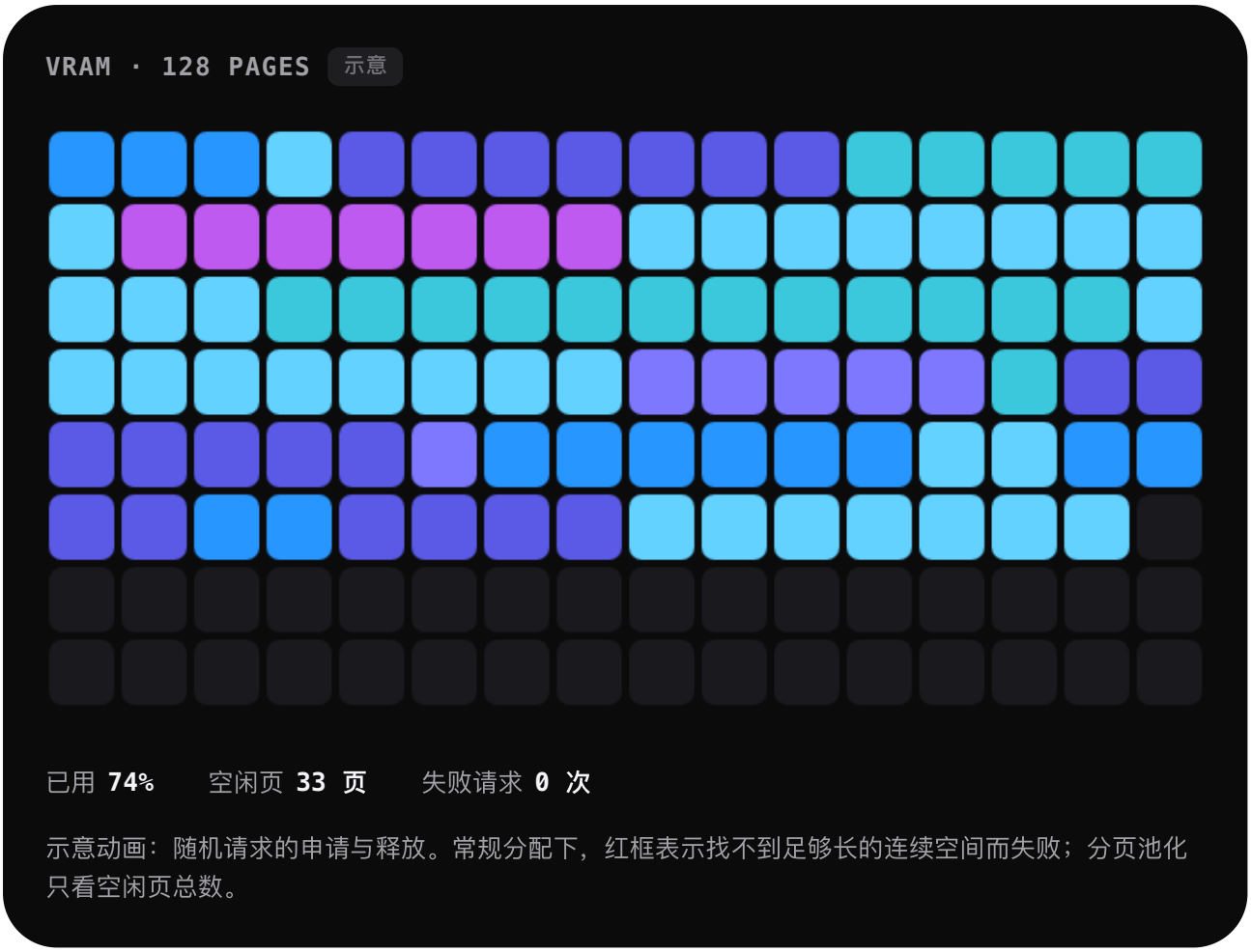
H2GC：存量卡自动画像，按卡定制，全自动拉起。

显存墙 · 分页显存池

空间够，却装不下？
问题在碎片。

常规做法为每个请求预留整段连续的 KV 缓存。请求来来去去，空闲空间被切成碎片——总量够，新的大请求仍会失败（OOM）。

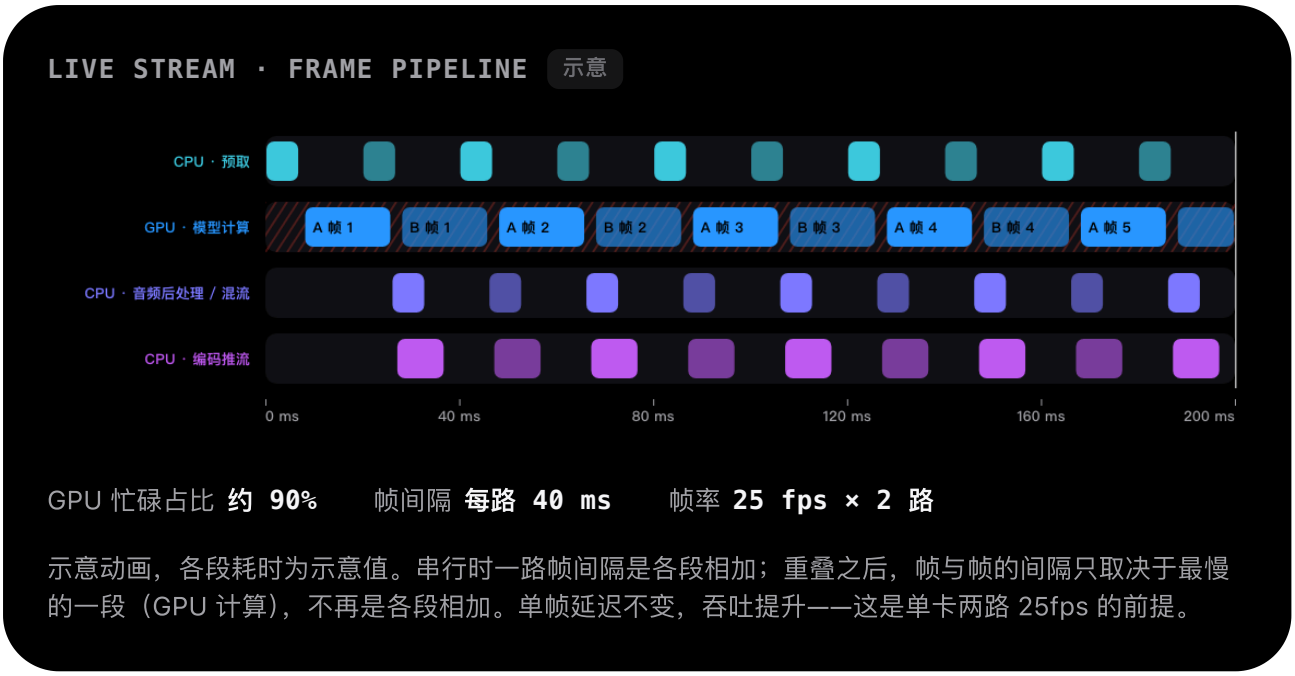
H2GC 参考 PagedAttention 等开源推理框架的思路，把显存切成固定大小的页，按页取用、不要求连续，碎片不再累积。



延迟墙 · 全异步流水线

让 GPU 不再空等。

预取、前后处理、音频混流、视频编码推流交给 CPU 并行完成，与 GPU 计算在时间上大部分重叠。



GPU 忙碌占比 **约 90%**　　帧间隔 **每路 40 ms**　　帧率 **25 fps × 2 路**

示意动画，各段耗时为示意值。串行时一路帧间隔是各段相加；重叠之后，帧与帧的间隔只取决于最慢的一段（GPU 计算），不再是各段相加。单帧延迟不变，吞吐提升——这是单卡两路 25fps 的前提。

四项底层优化 × 异构协同

从算子到总线，每一层都重新调过。

01 · 算子

定制算子融合

针对 MoE 路由、RoPE、RMSNorm、注意力与视频 DiT 的关键算子做 CUDA / Triton 融合重写，减少显存往返，提高有效算力。

02 · 显存

分页显存池

KV 缓存与多模态张量按页管理（参考 PagedAttention 思路），减少长序列、高并发下的碎片与 OOM。

03 · 精度

混合精度量化

在 Blackwell 上启用原生 FP4，按网络层敏感度配合 FP8 / W4A8 校准，压显存的同时守住画质与推理能力。

04 · 隔离

多流隔离与会话亲和

多路计算流按优先级与资源配额隔离，单路突发不挤占另一路的帧预算。会话亲和（通道锁定）由 L0 调度实现、平台网关执行，同一会话落在持有其缓存的通道上。

异构协同 · CPU × PCIe 5.0 × 主机内存

让 GPU 只做必须由它做的计算。

冷权重与历史 KV 缓存分级驻留在主机内存，按调度提前经 PCIe 5.0 预取回显存；热路径始终在显存内完成。冷热调度、数据预取、编解码与推流交给 CPU 并行承担。

在 H2 AIDM 中的位置

它托起上面三层，
也被上面三层的真实负载打磨。

↑ L1 模型 · 托起

H2 Model Combo

分页显存池、融合算子与会话亲和，是有状态缓存能留住历史、并发能做上去的底座；整栈 4 卡实测 96 路。

了解 H2MC ›

↑ L1 世界模型 · 托起

RayWorld

FP4 量化与分级预取，让同一张 RTX PRO 4500 可并发两路 Live 直播，也可单独运行 R1。

了解 RayWorld ›

↑ L3 应用 · 反哺

小影马与小影马直播

每天的分镜渲染、多图参考和 25fps 直播推流，是检验显存隔离与零掉帧流水线最严格的负载。

了解小影马 ›

↑ L2 智能体 · 反哺

H2 Harness

「调模型 → 执行工具 30–60 秒 → 再调模型」的负载节奏，决定了 KV 缓存分级驻留与会话亲和调度的设计。

了解 H2 Harness ›