

PR #37050 完整报告

sgl-project/sglang

docs: state that HiCache L2 is instance-private and only L3 is shared

合并时间: 2026-08-29 23:47

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37050>

执行摘要

- 一句话: 文档澄清 HiCache L2 实例私有、仅 L3 可跨实例共享
- 推荐动作: 值得阅读: 对维护者和文档贡献者而言, 本 PR 是“把隐含架构语义显式化”的良好范例——用表格、Mermaid 图与 Note 三层结构回答反复出现的用户问题, 且每条机器 review 意见都落实到最终提交。对纯代码开发者无需精读; 若今后编写 HiCache 或统一缓存相关文档, 可直接复用 Tier Sharing Scope 小节的表达方式与校订流程。

功能与动机

issue #31505 用户询问“如何让多个非 PD 实例把各自主机的 DRAM 聚合成一个共享 L2 KV offload 池”, 实际答案是否定的, 但旧文档只在设计页 CPU 缓存类比中用一句话提及, 导致用户反复提问, 并误调 `--hcache-ratio` / `--hcache-size` 期待获得跨实例复用。PR body 明确写到: “the docs only state it in one sentence inside a CPU-cache analogy in the design page, so the question keeps coming up and users try to get cross-instance reuse by raising `--hcache-ratio`.” 因此需要用显式、可检索的文档小节把共享边界说清。

实现拆解

1. 变更入口: 在 docs/docs/advanced_features/hicache_design.mdx 的 Overall Architecture 段落后新增 Tier Sharing Scope 小节, 用独立章节承接共享边界问题, 直接回应 issue #31505。
2. 共享范围表格: 按 Tier / Medium / Scope / Shared across instances 四列列出三层: L1 (HBM, 单实例, 否)、L2 (CPU DRAM, 单实例所在节点, 否)、L3 (存储后端, 范围取决于后端配置, 仅在后端按此配置时共享)。
3. Mermaid 拓扑图: 绘制 Node A 上的 Instance 0/1 与 Node B 上的 Instance 2, 三者各自持有私有 L1/L2, 仅通过 L3 相连, 直观表达“跨实例可见性必须等到 KV 到达 L3 才成立”。
4. 文字澄清: 明确 L2 是 node-local 且实例私有 (同一节点两个实例也不互读 L2), `--hcache-ratio` / `--hcache-size` 只扩大各自私有 L2, 跨实例复用需要 `--hcache-storage-backend`; 并逐后端限定: file 默认 node-local /tmp/hicache (可被 SGLANG_HICACHE_FILE_BACKEND_STORAGE_DIR 覆盖), mooncake、hf3fs、nixl、aibrix 在全部实例共享同一 namespace/ 配置时才达到集群范围。
5. 配套与 review 吸收: hicache_best_practices.mdx 顶部新增 <Note> 用一句话重复同一规则并链到新小节; 两条 Codex P1 意见均落实到最终提交——L3 由“普遍共享 (Yes)”改

为“Only if the backend is set up for it”，内部链接由相对路径改为根路径 `/docs/advanced_features/hicache_design#tier-sharing-scope`。无代码、无测试、无配置变更。

关键文件：

- `docs/docs/advanced_features/hicache_design.mdx`（模块 设计文档；类别 docs；类型 documentation）：核心变更文件，新增 Tier Sharing Scope 小节，用表格、Mermaid 图和文字三层方式明确 L1/L2 实例私有、L3 取决于后端配置的共享边界，直接回答 issue #31505。
- `docs/docs/advanced_features/hicache_best_practices.mdx`（模块 最佳实践；类别 docs；类型 documentation）：用户调参时必读页面，顶部新增 Note 快速传达同一规则并链接到设计文档新小节，是防止用户误调 `--hicache-ratio` 的第一道防线。

关键符号：未识别

评论区精华

- Codex P1：L3 不能表述为普遍共享。机器 review 指出原表格将 L3 写作“The whole cluster / Yes”会误导用户，因为内置后端并非都有集群 scope：file 后端默认 node-local（`python/sglang/srt/mem_cache/hicache_storage.py:368`），shm 后端 host-local，分布式后端也要求所有实例指向同一 namespace/ 配置。作者采纳，最终表格改为 Only if the backend is set up for it，段落中逐后端补充说明。
- Codex P1：内部链接需用根路径。best practices 页新增 Note 中的链接由 `/hicache_design#tier-sharing-scope` 改为 `/docs/advanced_features/hicache_design#tier-sharing-scope`，符合 docs/AGENTS.md 的文档链接规范。
- 维护者 ispobock 无额外评论直接批准（APPROVED），说明两条 P1 的修复已满足合入门槛。
- L3 共享性表述需限定于后端配置 (correctness)：作者采纳，最终表格改为 **Only if the backend is set up for it**，并逐后端说明 `file` / `mooncake` / `hf3fs` / `nixl` / `aibrix` 的共享范围。
- 文档内部链接应使用根路径 (style)：作者采纳，最终 Note 使用根路径链接。

风险与影响

- 风险：本 PR 不触碰任何运行时代码，无回归、性能、安全与兼容性风险。剩余风险集中在文档保鲜：① 若未来 HiCache 演进跨实例 L2 共享能力，本小节与 Note 需同步改写；② 表格枚举的后端清单（`file` / `mooncake` / `hf3fs` / `nixl` / `aibrix`）与 `hicache_storage.py` 中的实现（如 `/tmp/hicache` 默认路径）可能随版本漂移，需要定期核对；③ 文档仅英文，非英语用户仍可能误读（与仓库其余文档现状一致，非本 PR 引入）。
- 影响：影响范围仅文档，但覆盖面广：design 与 best practices 是 HiCache 用户理解架构与调参的首选入口，本次澄清可显著减少 issue #31505 一类重复提问，防止用户通过调大 `--hicache-ratio` 期待跨实例命中。对团队几乎无维护负担，价值主要体现在降低支持成本与固化架构契约。影响程度：低（无运行时影响）；价值：中（直接消除反复出现的用户误解）。

- 风险标记：纯文档变更，无运行时风险，文档需与 hicache 后端实现保持同步，后端枚举列表可能随演进过时

关联脉络

- PR #37151 [Unified Cache Linker][3/N]: Add backend-independent linker core: 同为 HiCache 缓存架构演进：该 PR 新增后端无关的统一缓存链路核心，正是“L3 可共享、范围取决于后端”这一语义的实现载体；本 PR 把该边界在文档中固定下来，避免用户把跨实例复用寄托在 L2 上。
- PR #37194 [Fix] Shut hicache test servers down gracefully before SIGKILL: 同为 HiCache 子系统质量维护：该 PR 修复测试服优雅关闭以释放钉住的主机 KV 池，说明 HiCache 仍在积极迭代，本 PR 的文档澄清是配套的“契约固化”。