

PR #29823 完整报告

sgl-project/sglang

[HiCache]fix draft host pool allocator type

合并时间: 2026-07-01 19:31

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29823>

执行摘要

- 一句话: 修复 draft host pool allocator 类型
- 推荐动作: 建议合并。这是一行关键修复, 虽然简单但解决了一个影响 HiCache 与 Mooncake 存储后端兼容性的回归。

功能与动机

修复 EAGLE draft KV HiCache 使用 Mooncake standalone 存储时的注册失败问题。PR body 中展示了错误日志: DRAFT_ALLOCATOR 为默认的 HostTensorAllocator, 导致 Mooncake 注册返回 -1 错误。

实现拆解

1. 在 python/sglang/srt/mem_cache/kv_cache_builder.py 的 maybe_register_hicache_draft 函数中, 构造 draft host pool 参数字典 kw 时新增 allocator_type=server_args.hicache_storage_backend 键值对。
2. 该参数确保 draft host pool 使用与 target host pool 相同的 MooncakeHostTensorAllocator 而非默认的 HostTensorAllocator。
3. 变更仅一行, 无其他文件改动。

关键文件:

- python/sglang/srt/mem_cache/kv_cache_builder.py (模块 内存缓存; 类别 source; 类型 core-logic; 符号 maybe_register_hicache_draft) : HiCache draft host pool 创建逻辑所在文件, 本次修改在第 111 行增加 allocator_type 参数传递。

关键符号: maybe_register_hicache_draft

关键源码片段

[python/sglang/srt/mem_cache/kv_cache_builder.py](#)

HiCache draft host pool 创建逻辑所在文件, 本次修改在第 111 行增加 allocator_type 参数传递。

```
# python/sglang/srt/mem_cache/kv_cache_builder.py
# 函数 maybe_register_hicache_draft 中构造 draft host pool 参数字典的部分
primary = tree_cache.cache_controller.mem_pool_host
```

```
kw = dict(
    host_to_device_ratio=primary.size / pool.size,
    host_size=0,
    page_size=page_size,
    layout=server_args.hicache_mem_layout,
    # 关键修复：增加 allocator_type，确保 draft host pool 使用与 target host pool
    相同的存储后端分配器
    # 例如 MooncakeHostTensorAllocator，否则默认 HostTensorAllocator 会导致 Mooncake
    注册失败
    allocator_type=server_args.hicache_storage_backend,
)
```

评论区精华

无审查评论，但两位 reviewer 均 approve，表明变更清晰且无争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅增加一个参数传递，不影响现有逻辑。但缺少对 Mooncake 未启用时 allocator_type 行为是否 fallback 的验证（已在 target pool 中处理）。
- 影响：仅影响使用 HiCache + Mooncake 存储后端且启用 EAGLE draft 的配置。修复后 draft host pool 能正确在 Mooncake 中注册，避免缓存注册失败导致的运行时错误。
- 风险标记：缺少测试覆盖

关联脉络

- PR #28287 [HiCache] Optimize HiCache hash generation with bulk token byte conversion: 同为 HiCache 模块相关，改动涉及同目录下的缓存构建逻辑。