PR #23241 完整报告

sgl-project/sglang

[HiCache & HybridModel] 3FS backend support DSA & mamba model

合并时间：2026-04-25 00:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23241

执行摘要

本 PR 为 HiCache 添加 HF3FS 存储后端支持，使 Mamba 和 DSA 模型能够使用 3FS 分布式文件系统作为缓存后端。核心改动是引入 namespace 概念实现不同数据池（KV vs Mamba）的元数据隔离，并新增 batch_exists_v2 等批量操作。PR 已在 production 环境中通过精度测试，但遗留了若干性能优化 TODO 和自动化测试缺口。

功能与动机

参考关联 Roadmap #21846，HiCache 需要支持 Hybrid 模型并扩展存储后端到 3FS 以提升分布式 KV 缓存的吞吐与容量。Mamba 和 DSA 模型对缓存数据结构有不同需求，因此需要一套统一的 namespace 机制来管理不同的缓存池。

实现拆解

1. 元数据服务支持 namespace

- 文件：mini_3fs_metadata_server.py
- GlobalMetadataState.ranks 的键由 int 改为 rank:namespace 复合字符串（通过新增 _rank_key() 方法生成），支持同一 rank 下多个独立 Pool。
- 所有 HTTP 路由（initialize、get_page_indices、delete_keys、exists、clear）均增加 namespace 参数，并在请求体中以 namespace 字段传递。
- 为兼容旧持久化文件，加载时若键不含 : 则自动追加 :kv。

2. 存储接口通用化

- 文件：storage_hf3fs.py
- Hf3fsMetadataInterface 的所有抽象方法（initialize、reserve_and_allocate_page_indices、confirm_write、get_page_indices、delete_keys、exists）均增加 namespace: PoolName = PoolName.KV 参数。
- 底层 HTTP 客户端相应调整，在请求路径末尾拼接 namespace 或通过请求体传递。

3. 新增批量操作与 Mamba 适配

- 在 HiCacheStorage 子类中新增 batch_exists_v2、batch_get_v2、batch_set_v2、register_mem_host_pool_v2 等函数，支持对多个 Pool 同时执行存在性检查、数据传输和内存注册。
- 写路径增加对 Mamba pool 的特殊处理：使用 flat=True 拉直数据形状后再写入。

- 读路径中从存储加载数据后，通过 unflat 还原为原始张量。

4. Mamba radix cache 连接

- 文件：hi_mamba_radix_cache.py（仅一行改动）
- 在 __init__ 中将 enable_storage_metrics 参数传递给 attach_hybrid_pool_to_mamba_cache，使监控指标可正确上报。

评论区精华

> 冗余 PoolName.value 转换（风格）— gemini-code-assist[bot] “Wrapping pool_name with PoolName() and accessing .value is redundant here as transfer.name is already a PoolName.” 作者回复已在 main 分支通过 PR #22891 修复，本 PR 同步后无此问题。
>
> 读路径性能（循环分配 dummy 页）— gemini-code-assist[bot] “Allocating a new dummy page tensor inside the loop for every page to be read is inefficient.” 建议预分配缓冲区，作者未回应，该风险保留。
>
> 写路径性能（torch.cat 开销）— gemini-code-assist[bot] “host_pool.get_data_page(host_idx, flat=True) performs a torch.cat operation which allocates a new tensor and copies data.” 作者未回应，风险保留。

风险与影响

- 回归风险：未新增自动化测试，metadata server 的 namespace 拆分可能影响旧版本持久化文件加载（已添加向后兼容逻辑，但测试覆盖不足）。
- 性能风险：batch_exists_v2 的读 / 写热点使用循环分配 tensor 和 torch.cat，在高并发下可能导致显存 / 内存抖动。
- 兼容性风险：依赖 3FS 集群环境，若部署配置不当，新后端可能无法正常工作。

关联脉络

- 本 PR 是 Roadmap #21846 的子任务（3FS Backend support for hybrid model），与 #20457（Hybrid Cache Controller）、#21259（Mooncake backend）、#22957（MLA+Mamba Hybrid）共同构建 HiCache 的 Hybrid 模型支持。
- 依赖的前置修复 PR #22891 解决了 PoolName 枚举值转化问题。
- 未来工作包括：统一 Hybrid Radix Cache、PD-Decode 侧 HiCache 支持等。