PR #7364 完整报告

PaddlePaddle/FastDeploy

[BugFix][PD Disaggregation][KVCache] Fix low cache hit rate in PD split (disaggregation) scenario

合并时间：2026-04-14 16:15

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7364

执行摘要

本次 PR 修复了 PD 分离（disaggregation）场景下 prefill 节点未及时更新 prefix cache 命中信息导致的低命中率问题，通过在调度器的 preallocate_resource_in_p 方法中主动调用 update_cache_blocks 并调整角色排除逻辑，确保 cache 状态正确同步。变更直接影响 PD 分离模式的推理性能，但参数选择争议未解决可能遗留 cache 状态不一致风险。

功能与动机

在 PD 分离场景中，prefill 节点负责处理 prompt 的预填充，但成功分配 GPU block 后未更新 prefix cache 的命中信息，导致已命中的 cache 无法被正确记录，命中率异常偏低，影响推理性能。PR body 明确指出："prefill 节点在通过 _allocate_gpu_blocks 成功分配 block 后，没有调用 update_cache_blocks 更新 cache block 状态，导致已命中的 prefix cache 无法被正确记录"。

实现拆解

仅修改 fastdeploy/engine/sched/resource_manager_v1.py 文件，包含两处关键改动：

1. 调整 _allocate_decode_and_extend 中的角色排除逻辑： python if (self.config.cache_config.enable_prefix_caching and self.config.scheduler_config.splitwise_role != "decode" and self.config.scheduler_config.splitwise_role != "prefill" # 新增排除prefill): 避免 prefill 节点在 _free_blocks_when_stop 中重复更新 cache block 状态。
2. 在 preallocate_resource_in_p 中主动更新 cache block： python self.cache_manager.update_cache_blocks(request, self.config.cache_config.block_size, request.need_prefill_tokens) prefill 节点分配 block 后立即调用，使用 need_prefill_tokens 作为已计算 token 数量。

评论区精华

review 中聚焦于 update_cache_blocks 参数选择的正确性：

- fastdeploy-bot指出："使用 request.need_prefill_tokens 作为 update_cache_blocks 的第三个参数需要进一步说明 ... 当只有部分 prompt tokens 命中 cache 时，使用 need_prefill_tokens 可能导致 cache tree 中记录的 cached blocks 数量与实际不一致。"

- Copilot进一步解释：“传入 request.need_prefill_tokens 会让 PrefixCacheManager 认为整段 prompt 都已‘计算完成’，从而为 cache miss 部分也提前创建 radix tree 节点并绑定新分配的 GPU block... 可能造成错误复用或状态污染。” 但作者未回应这些建议，最终代码仍使用 need_prefill_tokens。

风险与影响

风险：

1. cache 状态不一致：当部分 prompt 命中 cache 时，使用 need_prefill_tokens（所有 prompt tokens）而非 num_computed_tokens（已命中 tokens）可能导致 cache tree 记录不准确。
2. 错误复用风险：可能为未填充的 GPU block 创建 radix tree 节点，后续请求匹配到未完成的 cache。
3. 回归风险：变更涉及调度器核心路径，但缺少单元测试（PR body 说明需端到端 PD 分离环境验证）。

影响：

- 正面：修复后应显著提升 PD 分离场景的 prefix cache 命中率，改善推理性能。
- 范围：仅影响 PD 分离模式，对非分离模式无影响。
- 用户：透明修复，无需配置变更。

关联脉络

本次修复是 FastDeploy 在 PD 分离架构优化中的一环：

- PR #7241（移除 KV Cache 块数上限）同样关注 cache 利用率提升。
- PR #7299（移除 CacheManager 与 WorkerProcess 间 IPCLock）优化了 cache 相关组件交互。
- PR #7323（支持 PD 分离模式下 MTP 超重叠）同样针对 PD 分离模式进行调度优化。 这些 PR 共同反映了团队对 PD 分离场景性能的持续打磨，特别是调度器与 cache 管理器的协同优化。