PR #51113 完整报告

vllm-project/vllm

[Bugfix] Keep mamba align prefill chunks block-aligned past last_cache_position

合并时间：2026-08-07 01:21

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51113

执行摘要

- 一句话：修复 Mamba align 模式 chunk 越界导致的 prefix cache 投毒
- 推荐动作：值得精读。核心看点有三：一是把「可缓存边界」与「prefill 终点」分离的语义修正，这是对不变量 slot p == (p + 1) * block_size 的严格化；二是用运行时审计钩子直接测量毒化条目而不是依赖准确率差，定位方式值得借鉴；三是用真实 KVCacheManager + oracle 的 CPU 级测试复现 GPU 并发场景的测试设计，可作为后续调度器回归测试的模板。

功能与动机

PR body 将根因概括为：MambaManager.cache_blocks 把 block-table 槽位 p 哈希为 (p + 1) * block_size 处的状态，而 _mamba_block_aligned_split 只在 last_cache_position 以下强制该不变量；EAGLE 把 last_cache_position 回退一个 block（短 prompt 直接归零），超出该点后任意 chunk 结束位置都被接受。于是并发 prefill 共享 token budget 时，一个请求的 chunk 在 block 中途结束（如 state@364），后续 chunk 跨过 1600 又把同一槽位哈希为 state@1600，之后所有从该哈希恢复的请求都静默拿到截断状态，且无任何日志。该问题对应 issue #43559 中 MTP + prefix caching 场景下约 20% 的准确率下降。

实现拆解

1. 变更入口：vllm/v1/core/sched/scheduler.py 的 _mamba_block_aligned_split，这是 Mamba align 缓存模式下决定 prefill chunk 切分边界的唯一函数。
2. 结束对齐 gate 前移：把结束对齐的判定从 end < last_cache_position 改为 end < prefill_end（prefill_end = max(request.num_prompt_tokens, request.num_tokens - 1)），并同步更新注释与 last_cache_position 的语义：只有 prefill 的最后一个 chunk 允许不对齐结束，因为该槽位会由 decode 阶段的 fused postprocess 路径补到边界；其余所有中间 chunk 无论处于 last_cache_position 之前还是之后，都必须停在 block 边界。
3. mid-block realign 条件收紧：从 next_block_boundary <= last_cache_position 改为无条件 start % block_size != 0。这样以非对齐位置恢复的 prefill（更细的 prefix_match_unit、KV connector 外部计算 token）不会越过自己所在 block 的边界，从而不会用错误偏移发布槽位；last_cache_position 仍保留为强制停点，承担同长度请求 eagle 命中快照的职责。
4. 行为约定：当某个 step 的 token budget 不足以支撑对齐 chunk 时，返回 0，两个调用点（scheduler 与测试）都按 "defer 到后续 step" 处理，不产生空转。

5. 测试配套：新增 tests/v1/core/test_mamba_align_chunk_split.py（+247 行，CPU 测试），用真实 KVCacheManager（full-attn 16 + mamba-align 1600，num_speculative_blocks=3，use_eagle=True）驱动真实 split helper，通过 oracle 镜像 GDN 内核的状态写入，断言每个被哈希缓存的 mamba 槽位状态与哈希声明一致；覆盖碎片化首 chunk、尾 chunk、中间 chunk 对齐、block size 无关性、未对齐 resume 不越界（partial_hit on/off × 5 个偏移）。

关键文件：

- vllm/v1/core/sched/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 _mamba_block_aligned_split）：修复核心：_mamba_block_aligned_split 的对齐 gate 从 last_cache_position 换成 prefill_end，并移除 mid-block realign 的 last_cache_position 条件，是消除 prefix cache 投毒的关键变更。
- tests/v1/core/test_mamba_align_chunk_split.py（模块 回归测试；类别 test；类型 test-coverage；符号 _make_hybrid_kv_cache_manager, _split, _run_chunked_prefill, _count_cached_boundary_states）：新增 247 行 CPU 回归测试，驱动真实 KVCacheManager 与 split helper，用 oracle 校验槽位状态，覆盖碎片化 chunk、block size 无关性和未对齐 resume 场景，是修复可信度的核心证据。

关键符号：_mamba_block_aligned_split, _run_chunked_prefill, _count_cached_boundary_states, _make_hybrid_kv_cache_manager

评论区精华

本 PR 的 review 评论较少：claude[bot] 因 fork 自动跳过评审，维护者 njhill 直接 APPROVED（'Thanks @ivanium'）。真正有价值的讨论体现在 PR body 对 #45477 的差异分析：作者指出 #45477 把 mid-block realign 嵌套在 num_computed_tokens < last_cache_position 内，因此在 last_cache_position 处或之后开始的 mid-block 恢复仍可能跨过 block 边界；本 PR 直接移除该条件，并用 test_unaligned_resume_never_runs_past_its_block 覆盖。作者也明确表示如果 #45477 作者愿意继续维护，他愿意让位，目的是不让修复第三次死在 rebase 上。

- 与 #45477 的 mid-block realign 条件分歧 (design): 本 PR 采用更严格的不变量（任何 mid-block start 都必须在下一个 block 边界停下），并希望 #45477 作者确认其对旧分支的理解；维护者 njhill 已 APPROVED 并合并。
- 端到端验证 vs 单元级验证 (testing): 以单元级 oracle 校验替代，真实负载审计确认修复，端到端覆盖留给 #48970。

风险与影响

- 风险：
  1. 核心调度路径变更：_mamba_block_aligned_split 是所有 hybrid Mamba 模型在 align 缓存模式下的 prefill 分块入口，改动影响面集中在开启 EAGLE/MTP 且并发 prefill 的场景；单请求与轻负载下 end == prefill_end，行为与旧版逐位一致。
  2. 预算竞争行为变化：对齐约束外扩到 [last_cache_position, prefill_end)（约 (P mod block_size) + block_size，不足两个 block），高竞争时部分 chunk 会被截断或 defer

；PR 通过枚举 85 个配置验证无额外 forward 步、无更少可缓存快照，但该枚举未覆盖所有 block_size 与预算组合。

3. 覆盖缺口：本 PR 无法本地复现 #45477 的端到端验证（10 个 2002-token 请求），仅以 CPU 单元测试 + gsm8k 运行时审计（main 1 个毒化条目 -> 0）佐证；作者建议端到端覆盖放入 #48970。

4.未完整关闭#43559：该issue还涉及CUDAillegal-access崩溃（#40756、#50021）、MTP + reasoning off-by-one（#44927）、deep agentic 退化循环（#47087），本 PR 只解决正确性半边。 - 影响：对用户：修复 Qwen3.5/3.6 等 hybrid GDN 模型在 MTP + prefix caching + 并发负载下静默精度损坏的问题，影响面为使用 --mamba-cache-mode align 的生产部署。对系统：改动集中在调度器单个函数，非 Mamba 模型完全不受影响；对 hybrid Mamba 模型，理论上不改变单请求和轻负载路径，仅在高竞争时改变 chunk 切分边界。对团队：该 PR 被纳入 v0.27.0 cherry-picks 里程碑，结束了 #45477 与 #47861 两次因 rebase 停滞的修复，维护者 njhill 已合并。 - 风险标记：核心调度路径变更，依赖 CPU 级回归覆盖，端到端场景未本地复现，无法单独关闭 #43559，行为与预算竞争相关

关联脉络

- PR #45477 [Bugfix] Keep intermediate prefill chunks mamba-block-aligned with spec decode: 本 PR 的原始诊断与修复来源，作者 kodek 的同一修复在 current main 上重新推导；PR body 详细对比了两者在 mid-block realign 条件上的差异。
- PR #50021 [Bugfix] Bound two num_accepted-derived indices in the GDN spec-decode path (MTP + hybrid GDN CUDA illegal access): #43559 的 crash 半边，与本 PR 同属 hybrid Mamba + MTP 故障族，但修复在不同层级（内核索引边界 vs 调度 split），两者互补。