PR #51015 完整报告

vllm-project/vllm

[CI] Stabilize GLM-5.2 PCP evaluation

合并时间：2026-08-05 05:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/51015

执行摘要

- 一句话：GLM-5.2 PCP 评估启用 expandable segments 修复 OOM
- 推荐动作：值得快速浏览（约 5 分钟）：变更本身只有 1 行，但 PR body 的根因分析方法很有参考价值——区分表面超时与真实崩溃、用历史 Buildkite run 对照排除模型回归、把重复工作排查写进描述。设计决策上值得借鉴的是：将分配器行为收敛到确有内存压力的特定配置，而不是全局开启 expandable_segments，避免影响其他场景。建议关注后续是否有针对 FlashInfer MoE workspace 分配策略的长期优化 PR。

功能与动机

PR body 表明 pytest 层报错只是 20 分钟 server-start 超时，worker 日志显示真正的失败在 profile 阶段：CUDA out of memory. Tried to allocate 24.43 GiB. 15.36 GiB is free 23.97 GiB is reserved by PyTorch but unallocated。该分配来自 FlashInfer CUTLASS MoE 的 FusedMoeRunner::getWorkspaceInfo。作者通过对照历史 run 确认这不是模型运行器回归，而是 #49294 引入 --max-num-batched-tokens 32768 后 TP1/PCP4 配置内存紧张；启用 expandable segments 是让保留内存增长 / 复用、避免二次连续分配的最小改动。

实现拆解

1. 根因定位：通过 Buildkite #81913 的 worker 日志确认 pytest 层的失败只是表面超时，真正的崩溃发生在 profile 运行阶段，FlashInfer CUTLASS MoE 的 FusedMoeRunner::getWorkspaceInfo 尝试一次性分配约 24.43 GiB，而此时 PyTorch 已保留 23.97 GiB 却无法合并成满足请求的连续段，导致 CUDA out of memory。
2. 对照排查：同样的分配签名在 #80082（Kimi/K3 变更前）也存在，说明并非模型运行器回归或 B200 节点问题；成功运行 #80689 显示实际峰值约 159.66 GiB 非 KV 内存，低于 178.35 GiB 总量，证明以 #49294 引入 --max-num-batched-tokens 32768 为分界，TP1/PCP4 配置变得内存紧张，触发间歇性 OOM。
3. 方案落地：在 tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP1-PCP4-EP.yaml 的 env 中新增 PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True，让保留的缓存段可以继续增长 / 复用，而不是要求第二块连续大内存；同时保留 32K batch 限制以继续覆盖大 batch PCP 路径。改动只作用于该内存紧张配置，TP2/PCP2 保持原样。
4. 验证：Buildkite #82253（LM Eval PCP 4xB200）在 dgxB200-14 通过，profile 阶段权重 130.49 GiB、torch 峰值 28.33–30.23 GiB，服务器正常启动，1,319 个 GSM8K 示例全部通过；TP2/PCP2 控制组 GSM8K 通过；git diff --check、YAML 解析、pre-commit

均通过。

关键文件：

- tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP1-PCP4-EP.yaml（模块 评估配置；类别 test；类型 configuration）：唯一变更文件，在 env 中新增 PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True，使 FlashInfer MoE workspace 能复用已保留的缓存段，修复 TP1/PCP4 下 24.43 GiB 分配 OOM；同时保留 --max-num-batched-tokens 32768 以覆盖大 batch PCP 路径。

关键符号：未识别

关键源码片段

tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP1-PCP4-EP.yaml

唯一变更文件，在 env 中新增 PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True，使 FlashInfer MoE workspace 能复用已保留的缓存段，修复 TP1/PCP4 下 24.43 GiB 分配 OOM；同时保留 --max-num-batched-tokens 32768 以覆盖大 batch PCP 路径。

```
# GLM-5.2 TP1/PCP4 GSM8K 评估配置（tests/evals/gsm8k/configs/GLM-5.2-NVFP4-TP1-PCP4-EP.yaml）
model_name: "nvidia/GLM-5.2-NVFP4"
accuracy_threshold: 0.90
num_questions: 1319
num_fewshot: 5
max_concurrency: 100
server_args: >-
  --enforce-eager
  --max-model-len 4096
  --max-num-batched-tokens 32768  # 保留 32K 批大小，继续覆盖大 batch PCP 路径
  --safetensors-load-strategy prefetch
  --moe-backend flashinfer_cutlass
  --prefill-context-parallel-size 4
  --enable-expert-parallel
  --kv-cache-dtype fp8
env:
  # FlashInfer CUTLASS MoE 的 workspace 需要一次性分配约 24.43 GiB，
  # 普通缓存分配器因保留内存无法合并而报 OOM；
  # expandable segments 让保留段可增长复用，仅作用于该内存紧张配置。
  PYTORCH_CUDA_ALLOC_CONF: "expandable_segments:True"
  VLLM_LOGGING_LEVEL: "DEBUG"
  VLLM_USE_V2_MODEL_RUNNER: "1"
```

评论区精华

PR 上没有出现实质性的技术争辩，也没有内联 review 评论：

- claude[bot] 提示该 PR 来自 fork，自动 review 默认关闭，维护者可通过 @claude review 触发一次性审核；

- 维护者 LucasWilkinson 直接批准："LGTM; thank you!"；
- Issue 评论区只有 3 次 CI 触发命令（/runci、/ci run x2）和 GitHub Actions 的确认回复；
- PR body 自行完成了重复工作排查，明确 #50791、#49196、#34553 与本次 OOM 无关，降低了审阅负担。
- Fork 自动 review 默认关闭 (other): 未触发额外审核流程；人工维护者随后直接批准。
- 维护者批准 (other): PR 获得批准，无代码层面异议。

风险与影响

- 风险：
  - 分配器行为变化：*expandable_segments:True* 允许 PyTorch 缓存段增长 / 复用，可能带来常驻内存增大或碎片化；该变更只限 1 个 *eval* 配置，且已在 B200 上验证峰值约 30 GiB，不影响 TP2/PCP2。
  - 缓解而非根治：这是对 FlashInfer MoE workspace 大块分配请求的缓解，未改动 FusedMoeRunner::getWorkspaceInfo 本身；若 batch 大小、并行度或 PyTorch 版本变化，仍可能复发。
  - 范围与回归：不含模型输入、kernel 或数值改动，理论上无精度回归风险；但该 YAML 生效路径依赖 CI 正确读取 *env* 字段，若未来配置解析变化需同步验证。
  - 安全：无数据面或权限变更，无安全问题。
- 影响：
  - 用户：完全无感，纯 CI 评估配置变更。
  - CI 系统：GLM-5.2 TP1/PCP4 GSM8K 评估不再间歇 OOM，减少 20 分钟 server-start 超时的误报与重跑成本。
  - 团队：提供了一个低成本、配置级的 allocator 缓解范例，后续内存紧张评估可直接借鉴；同时把 FlashInfer workspace 分配问题从 CI 阻塞中剥离出来，便于后续单独优化。
  - 风险标记：显存分配行为，测试配置变更，缓解而非根治

关联脉络

- PR #49294 引入 --max-num-batched-tokens 32768（PR body 提及）：该 PR 引入 32K batch 配置，使得 TP1/PCP4 评估内存紧张，是本次 OOM 的直接诱因；本 PR 是它的配套稳定性修复。
- PR #50791 为 DCP 调整 FlashInfer sparse-MLA workspace（PR body 提及）：PR body 澄清其为不同的 DCP workspace 调整，与本次 MoE workspace OOM 无关，属于重复工作排查结论。
- PR #49196 旧 PCP 精度回归自动回滚草稿（PR body 提及）：属历史精度回归回滚草稿，与本 profile-run MoE OOM 无关，避免误判为本 PR 冲突。
- PR #34553 GLM-5 sparse-attention indexer OOM（PR body 提及）：属于 serving 场景的 sparse-attention indexer OOM，不是启动期 MoE workspace 失败，被排除在外。