

PR #51292 完整报告

vllm-project/vllm

[Core] Disable fuse_allreduce_rms under VLLM_BATCH_INVARIANT (non-deterministic under TP)

合并时间: 2026-08-26 03:16

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51292>

执行摘要

- 一句话: 禁用 BI 下 fuse_allreduce_rms 融合, 修复 TP 下批次不变性
- 推荐动作: 该 PR 值得精读, 尤其是其根因分析和性能权衡论述。关键设计决策在于: 将默认禁用和显式覆盖守卫分离到两个 PR, 保持审查范围清晰, 同时提供了可复现的测试方法。建议读者关注 enable_allreduce_rms_fusion 的修改位置以及与 #50505 的协作关系。

功能与动机

Issue #51290 报告: 在 `VLLM_BATCH_INVARIANT=1` 时, 单 GPU 上是 bit-stable 的, 但张量并行 (TP>1) 下重复相同工作负载会得到不同的 logprobs 甚至不同 token。作者 bisect 定位到 `fuse_allreduce_rms` 编译 pass, 该 pass 将 TP all-reduce 替换为 FlashInfer 融合内核, 其归约顺序非确定性。目的是在请求批次不变性时禁用此融合, 恢复确定性, 同时保持其他路径不受影响。

实现拆解

1. 修改入口: 在 `vllm/config/vllm.py` 的 `enable_allreduce_rms_fusion` 函数开头加入早期返回。
2. 核心逻辑: 当环境变量 `VLLM_BATCH_INVARIANT` 为真时直接返回 `False`, 阻止融合 pass 启用。
3. 配套说明: 无配置文件、测试或部署改动; 改动极小, 仅影响 TP>1 且启用批次不变性的场景。
4. 设计权衡: 与 #50505 的 `disable_custom_all_reduce` 守卫一致, 且不覆盖用户显式 `--compilation-config` 覆盖 (由 #50505 中的 `fail-loud` 守卫处理显式覆盖)。

关键文件:

- `vllm/config/vllm.py` (模块 配置; 类别 source; 类型 core-logic; 符号 `enable_allreduce_rms_fusion`): 核心修改文件, 在 `enable_allreduce_rms_fusion` 中加入批次不变性早期返回, 直接影响融合 pass 的启用逻辑。

关键符号: `enable_allreduce_rms_fusion`

关键源码片段

`vllm/config/vllm.py`

核心修改文件，在 `enable_allreduce_rms_fusion` 中加入批次不变性早期返回，直接影响融合 pass 的启用逻辑。

```
# vllm/config/vllm.py
def enable_allreduce_rms_fusion(cfg: "VllmConfig") -> bool:
    """Enable if TP > 1 and Hopper/Blackwell and flashinfer installed."""
    from vllm.platforms import current_platform
    from vllm.utils.flashinfer import has_flashinfer

    # The fused all-reduce + RMSNorm path is not batch-invariant
    if envs.VLLM_BATCH_INVARIANT:
        return False

    if current_platform.is_rocm():
        from vllm._aiter_ops import rocm_aiter_ops
        return (
            rocm_aiter_ops.is_enabled() and cfg.parallel_config.tensor_parallel_size > 1
        )

    return (
        cfg.parallel_config.tensor_parallel_size > 1
        and current_platform.is_cuda()
        and has_flashinfer()
        and (
            current_platform.is_device_capability_family(100)
            or current_platform.is_device_capability(90)
        )
    )
```

评论区精华

审查中讨论了根因、用户覆盖行为和守卫拆分。

- 根因确认：yewentao256 要求深入分析，作者通过实验确认是 FlashInfer 内核的归约顺序非确定性，而非 autotune 或 batch 组成问题。
- 显式覆盖：yewentao256 询问 `--compilation-config '{"pass_config":{"fuse_allreduce_rms":true}}'` 是否会覆盖此修改，作者确认会，并说明这是故意的，显式覆盖由 #50505 的 fail-loud 守卫处理。
- 守卫归属：mosafariuk 建议将 fail-loud 守卫放在 #50505，作者同意并感谢，保持两个 PR 范围清晰。
- 根因分析：FlashInfer 融合内核的非确定性 (correctness)：确认根因是 FlashInfer 内核本身非确定性，禁用融合是正确修复。
- 显式配置覆盖行为 (design)：确认显式覆盖存在，并在 #50505 中处理，保持本 PR 默认禁用。
- fail-loud 守卫放置在 #50505 (design)：守卫放在 #50505，本 PR 仅做默认禁用。

风险与影响

- 风险：性能回退：在 `VLLM_BATCH_INVARIANT=1` 且 $TP>1$ 时禁用融合会导致吞吐下降（Qwen2.5-7B 约 2.5 倍，Llama-3.1-70B 约 1.7 倍），但这是预 #50505 的 NCCL 回退成本，配合 #50505 的确定性自定义 all-reduce 可恢复至 ~97% 吞吐。行为变化：显式开启 `fuse_allreduce_rms` 的配置会绕过此守卫，可能导致非确定性，但由 #50505 的 fail-loud 捕获。兼容性：改动仅影响 `VLLM_BATCH_INVARIANT` 设置下的 Hopper/Blackwell ($cc \geq 90$)，A100 ($cc < 90$) 融合 pass 本就 inactive，不受影响。
- 影响：用户影响：修复了批次不变性特性在 TP 场景下的核心承诺，保证可复现输出，对依赖确定性的生产环境（如测试、审计）至关重要。系统影响：仅在特定配置下禁用融合，影响面窄，但性能敏感用户需注意吞吐回退，可通过升级到包含 #50505 的版本恢复性能。团队影响：明确了与 #50505 的职责划分，为后续批次不变性相关修复奠定了模式（配置禁用 + 显式守卫）。
- 风险标记：性能回退（特定配置），显式配置可绕过

关联脉络

- PR #51290 [Bug]: `VLLM_BATCH_INVARIANT=1` not deterministic under tensor parallelism ($TP>1$); `fuse_allreduce_rms` fused all-reduce is the cause: 本 PR 修复的 issue，问题描述和根因分析来自此 issue。
- PR #50505 Custom All-Reduce 确定性 1-stage 路径：PR body 中明确提到本 PR 是 #50505 的 Hopper/Blackwell 侧补充，两者组合恢复性能并守卫显式覆盖。