

PR #29982 完整报告

sgl-project/sglang

[AMD][DeepSeek V4] Fix default FlashMLA sparse prefill off on ROCm/HIP

合并时间: 2026-07-03 07:00

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29982>

执行摘要

- 一句话: 修复 AMD ROCm 下 DeepSeek-V4-Flash 默认启用稀疏 prefill 导致的崩溃
- 推荐动作: 该 PR 值得快速了解设计思路: 平台条件默认值、保留用户覆盖能力、与现有模式保持一致。但实现中的条件逻辑 bug 提醒需要仔细审查环境变量的解析时机。建议阅读 #30237 作为后续修正。

功能与动机

PR #29775 将 `SGLANG_OPT_FLASHMLA_SPARSE_PREFILL` 默认设为 True 后, AMD MI355X 的 DeepSeek-V4-Flash 解聚 nightly 测试全面回归 (详见 #28565894749), 而 DeepSeek-V4-Pro 不受影响。定位问题为稀疏 prefill 路径在 ROCm 上返回错误输出 (warmup 输出近乎全零), 需要恢复旧有的密集 prefill 行为。

实现拆解

1. 识别问题根因: 通过 nightly 回归窗口定位 #29775 为唯一可疑变更, 确认稀疏 prefill 路径在 ROCm/HIP 上产生错误输出。
2. 平台条件判断: 在 `apply_deepseek_v4_defaults` 中引入 `is_hip()` 检查, 当运行在 ROCm/HIP 且环境变量未被显式设置时, 强制关闭稀疏 prefill (`envs.SGLANG_OPT_FLASHMLA_SPARSE_PREFILL.set(False)`)。
3. 保持用户覆盖能力: 仅修改默认值, 用户仍可通过显式设置 `SGLANG_OPT_FLASHMLA_SPARSE_PREFILL` 环境变量来启用稀疏 prefill, 便于未来验证或人工调优。
4. 位置选择: 根据 review 建议, 将逻辑从 `environ.py` 移至 `deepseek_v4_hook.py`, 与已有的 CP-based 禁用模式一致, 避免修改全局环境默认值层面。
5. 注意: 后续 #30237 发现本 PR 的条件 `not envs...is_set()` 因环境变量已在 `environ.py` 中被解析而失效, 改为仅 `is_hip()` 才真正生效。

关键文件:

- `python/sglang/srt/arg_groups/deepseek_v4_hook.py` (模块 模型配置; 类别 source; 类型 dependency-wiring; 符号 `apply_deepseek_v4_defaults`): 核心变更文件, 在 `apply_deepseek_v4_defaults` 中添加了 ROCm 平台判断来默认关闭 FlashMLA 稀疏 prefill。

关键符号: `apply_deepseek_v4_defaults`

关键源码片段

`python/sglang/srt/arg_groups/deepseek_v4_hook.py`

核心变更文件, 在 `apply_deepseek_v4_defaults` 中添加了 ROCm 平台判断来默认关闭 FlashMLA 稀疏 prefill。

```
def apply_deepseek_v4_defaults(server_args: ServerArgs, model_arch: str) -> None:
    """Apply DeepSeek V4 模型特定的 server arg 默认值和约束。"""
    from sglang.srt.server_args import ServerArgs
    from sglang.srt.utils import is_hip

    # FlashMLA 稀疏 prefill (SGLANG_OPT_FLASHMLA_SPARSE_PREFILL, 默认开启)
    # 当前在 ROCm/HIP (MI355X) 上对 DeepSeek-V4-Flash 返回错误输出,
    # 这导致了解聚 nightly 测试失败。在 ROCm 上保持旧的 (密集 prefill) 行为,
    # 直到稀疏内核在该平台上通过验证; 显式的环境变量设置仍可覆盖该默认值。
    if is_hip() and not envs.SGLANG_OPT_FLASHMLA_SPARSE_PREFILL.is_set():
        logger.warning(
            "Disabling SGLANG_OPT_FLASHMLA_SPARSE_PREFILL by default on ROCm/HIP "
            f"for {model_arch}; set it explicitly to override."
        )
        envs.SGLANG_OPT_FLASHMLA_SPARSE_PREFILL.set(False)

    # 后续逻辑: 设置 attention 后端、page_size、kv_cache_dtype 等 ...
```

评论区精华

1. Fridge003 建议将逻辑移至 `deepseek_v4_hook.py`: 原始实现试图在 `environ.py` 中让环境变量默认值成为平台条件, Fridge003 建议改为在 `deepseek_v4_hook.py` 中实现, 与现有 CP-based 禁用模式一致。作者采纳并重构。
2. HaiShaw 指出条件逻辑 bug: HaiShaw 发现 `if is_hip() and not envs.SGLANG_OPT_FLASHMLA_SPARSE_PREFILL.is_set()` 中的 `not` 导致条件永远为假 (因为 `environ.py` 的默认值已被解析), 建议去掉 `not`。作者确认这是一个 bug, 并指出后续 PR #30237 已将其修复为 `if is_hip():`。
- 将平台条件判断从 `environ.py` 移至 `deepseek_v4_hook.py` (design): 作者采纳建议, 将逻辑从 `environ.py` 迁移至 `deepseek_v4_hook.py`, 并还原 `environ.py` 为 `EnvBool(True)`。
- 条件逻辑中的 `not is_set()` 是否为 bug (correctness): 该 bug 由后续 PR #30237 修复, 改为 `if is_hip():`。

风险与影响

- 风险:
 1. 条件逻辑失效风险 (已发生): 原始实现中的 `not ...is_set()` 条件因环境变量在 `environ.py` 中已被解析而永远为假, 导致稀疏 prefill 在 ROCm 上实际上未被禁用, PR 未真正解决问题。后续 #30237 已修正。

1. 影响范围有限：仅影响 AMD ROCm/HIP 上的 DeepSeek-V4-Flash 模型，CUDA 和其他模型路径完全不受影响。
2. 显式覆盖保留：用户可通过显式设置环境变量恢复稀疏 prefill，不会完全阻塞 ROCm 上的功能验证。 - 影响：
 - 对用户：AMD ROCm 用户无需手动设置环境变量即可恢复 DeepSeek-V4-Flash 的正常推理，Nightly 测试回归为绿色。
 - 对系统：无性能影响，仅增加一个平台判断和条件设置。
 - 对团队：快速修复了重要 nightly 测试失败，为后续内核验证争取时间。
 - 风险标记：条件逻辑 bug 导致未生效，后续需要 #30237 修复

关联脉络

- PR #30237 [AMD][DeepSeek V4] Set SGLANG_OPT_FLASHMLA_SPARSE_PREFILL to false on hip code path: 修正本 PR 中的条件逻辑 bug（去掉 not is_set 判断），是功能上的后续修复。
- PR #29775 [AMD] Enable FlashMLA sparse prefill default on for DeepSeek-V4: 本 PR 修复的目标：PR #29775 将 FlashMLA 稀疏 prefill 默认开启，导致 ROCm 回归。