PR #36714 完整报告

sgl-project/sglang

[AMD][Spec][PD] Enable the PD DSA fused-TopK seed remap on ROCm

合并时间：2026-08-29 13:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36714

执行摘要

- 一句话：修复 ROCm 上 PD DSA fused TopK 失效，提速 2.92×
- 推荐动作：值得精读。本 PR 展示了平台差异性问题如何以最小改动解决，并提供了完整的验证方法论（包括 paired McNemar 检验、硬件实测与逐 worker 证据）。对于关注 AMD/ROCm 支持与 PD 解聚性能的开发者有参考价值。

功能与动机

PR #31477 引入的 PD DSA fused-TopK remap 仅在 CUDA 上生效，导致 ROCm 上的 PD 解码无法使用 fused TopK，性能未得到优化。作者在 PR body 中引用评审者 @kpham-sgl 的确认：‘I meant to gate off NPU only but ROCm paths are untested. Can you submit a PR with tested results?‘ 表明设计意图是仅排除 NPU，而非 ROCm。此 PR 正是为了在 ROCm 平台启用该优化，并附带测试验证。

实现拆解

1. 修改 python/sglang/srt/layers/attention/dsa/utils.py 中 should_remap_pd_dsa_seed_to_local_slots() 的谓词，将 is_cuda() 改为 (is_cuda() or is_hip())，允许 ROCm 平台进入 remap 分支，同时保持 NPU 仍然禁用（因为 NPU 上 is_cuda() 与 is_hip() 均为 False）。
2. 更新 test/registered/unit/disaggregation/test_disaggregation_wire.py，将原有仅针对 CUDA 的测试参数化，覆盖 CUDA（cuda=True, hip=False）、ROCm（cuda=False, hip=True）和“其他”（如 NPU，两者均 False）三种平台，分别断言 should_use_dsa_fused_topk 结果及 remap 后的 dsa_topk_indices 是否符合预期，确保在 CUDA/ROCm 上执行 remap，在 NPU 上保持透传。
3. 测试在 CPU 上进行，通过 patch 平台谓词，无需真实 AMD 硬件，且保持 register_cpu_ci。
4. 验证依赖条件：确认 req_to_token 在所有后端均使用 page_size=1 的 token 槽，以及 dsa_drop_wide_page_table 已独立排除 HIP，不会因启用融合而丢弃 ROCm indexer 仍需的页表。

关键文件：

- python/sglang/srt/layers/attention/dsa/utils.py（模块 注意力；类别 source；类型 core-logic；符号 should_remap_pd_dsa_seed_to_local_slots）：核心逻辑变更：修改

should_remap_pd_dsa_seed_to_local_slots() 谓词，启用 ROCm 上的 PD DSAfused-TopK remap。

- test/registered/unit/disaggregation/test_disaggregation_wire.py（模块 解聚；类别 test；类型 test-coverage；符号 test_pd_decode_fused_topk_remaps_wire_positions_to_local_slots）：测试覆盖调整：参数化测试覆盖 CUDA、ROCm 和其他平台，验证 remap 与透传逻辑。

关键符号：should_remap_pd_dsa_seed_to_local_slots, should_use_dsa_fused_topk

关键源码片段

python/sglang/srt/layers/attention/dsa/utils.py

核心逻辑变更：修改 should_remap_pd_dsa_seed_to_local_slots() 谓词，启用 ROCm 上的 PD DSA fused-TopK remap。

```
# python/sglang/srt/layers/attention/dsa/utils.py
# 判断 PD seed 是否应进入 allocator 局部的 fused TopK 域。
def should_remap_pd_dsa_seed_to_local_slots() -> bool:
    """Whether a PD seed should enter the allocator-local fused TopK domain."""
    return (
        # 原为 is_cuda()，现扩展为 (is_cuda() or is_hip()),
        # 使 ROCm 也能启用 remap，而 NPU ( 两者均为 False) 仍被排除。
        (is_cuda() or is_hip())
        and envs.SGLANG_DSA_FUSE_TOPK.get()
        and get_disagg().disaggregation_mode == "decode"
        and not get_memory().enable_hisparse
        and not get_parallel().dcp_enabled
    )
```

test/registered/unit/disaggregation/test_disaggregation_wire.py

测试覆盖调整：参数化测试覆盖 CUDA、ROCm 和其他平台，验证 remap 与透传逻辑。

```
# test/registered/unit/disaggregation/test_disaggregation_wire.py
# 参数化验证不同平台下的 fused TopK 行为。
local_slots = [[309, 101, -1], [801, 990, -1]]
unremapped = [[2, 0, -1], [1, 3, -1]]
for platform, cuda, hip, fused, expected in (
    ("cuda", True, False, True, local_slots),
    ("hip", False, True, True, local_slots),
    # 其他平台（如 NPU）两者均为 False，仍拒绝 seed，保持透传。
    ("other", False, False, False, unremapped),
):
    with self.subTest(platform=platform), envs.SGLANG_DSA_FUSE_TOPK.override(
        True
    ), patch(
        "sglang.srt.layers.attention.dsa.utils.is_cuda", return_value=cuda
    ), patch(
        "sglang.srt.layers.attention.dsa.utils.is_hip", return_value=hip
    ):
```

```
# 断言 fused 决策正确
self.assertEqual(
    should_use_dsa_fused_topk(seed_dsa_topk_from_draft_extend=True),
    fused,
)
# 构造 draft input 并断言 remap 后的索引
draft_input = build_eagle_disagg_draft_input(
    batch, torch.tensor([11, 12], dtype=torch.int64), None
)
self.assertEqual(draft_input.dsa_topk_indices.tolist(), expected)
```

评论区精华

无 review 评论，仅有评审者 @kpham-sgl 的审批意见'LGTM'。但 PR body 中作者与评审者有关于门控范围的讨论：评审者明确意图是仅排除 NPU，且认为 ROCm 路径未测试，因此要求作者提交带测试结果的 PR；作者在正文中详述了在 gfx942/gfx950 上的验证过程，并接受'编译通过不等于正确'的谨慎态度，主动检查了 req_to_token 页大小与 dsa_drop_wide_page_table 的独立性。

- 门控范围讨论 (design): 扩展谓词以包含 ROCm，并保持 NPU 排除，符合原始意图。

风险与影响

- 风险：
  1. 行为变更：扩展门控至 ROCm 将改变 PD 解码的融合路径，存在潜在的数值行为差异。作者在 PR body 中提供了充分的新能数据与正确性验证（包括 paired McNemar 检验），但需注意 n 较小（GSM8K 1319 条、LongBench 180 条），置信区间较宽（±1.1 pp / ±5.4 pp），不能完全排除微小回退。
  2. 依赖项：should_remap_pd_dsa_seed_to_local_slots() 的修改可能影响其他调用方，需确认 is_hip() 在非 ROCm 平台（如 CUDA）上的行为符合预期（返回 False），不会意外开启。
  3. 测试覆盖：测试依赖 patch 谓词模拟平台，但未验证真实 ROCm 上的 is_hip 实现是否可靠，且 is_hip 可能受环境变量影响。
  4. 性能风险：启用融合后，若 ROCm 上的 fused-TopK 内核存在未发现的索引错误，可能导致解码错误，但作者已通过正确性测试验证。 - 影响：用户影响：AMD ROCm 平台（gfx942/gfx950）上运行 GLM-5.2 等 DSA 模型并启用 PD 解聚 + EAGLE/MTP 的用户，将受益于 fused-TopK 的显著性能提升（TPOT 2.92×）。系统影响：修改仅涉及判别逻辑，不影响其他平台；NPU 行为保持不变。团队影响：提供了一个在真实 ROCm 硬件上验证的修复，为后续类似平台兼容性问题提供了参考。影响程度：中高，但修改范围小，安全。 - 风险标记：核心路径变更，平台兼容性

关联脉络

- PR #31477 Add PD DSA seed remap for allocator-local fused TopK: 本 PR 修复了 #31477 在 ROCm 上的失效问题，是同一功能的平台补充。