PR #25205 完整报告

sgl-project/sglang

[AMD] Auto-fallback NSA indexer to page_size=1 when aiter preshuffle gluon kernel is unavailable (Deepseek v3.2)

合并时间：2026-05-14 15:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25205

执行摘要

- 一句话：ROCm NSA indexer 自动回退 page_size=1 以兼容低版本 Triton
- 推荐动作：值得关注的设计包括：运行时能力检测与优雅降级模式；通过 @lru_cache 避免重复检测；使用独立环境变量提供紧急逃生通道（SGLANG_NSA_HIP_DISABLE_PRESHUFFLE）；以及在 server_args.py 中用延迟导入打破循环依赖的技巧。

功能与动机

PR#23562 无条件将 ROCm NSA indexer page_size 提升到 64 并启用 preshuffle，但该路径要求 Triton>=3.5.0 或 AOT gluon kernel bundle，在默认 rocm700 CI 镜像（Triton<3.5，无 AOT）中引发断言崩溃。本 PR 通过运行时检测 preshuffle 可用性，实现自动回退，避免修改 Docker 基础镜像或降级 aiter 依赖。

实现拆解

1. 在 python/sglang/srt/layers/attention/nsa/utils.py 中新增运行时检测函数 aiter_can_use_preshuffle_paged_mqa()，通过环境变量和 Triton 版本判断是否可使用 preshuffle 路径，结果被 @lru_cache 缓存。
2. 在 nsa_indexer.py 中导入该函数并设置模块级标志 _use_aiter_preshuffle，在 _get_topk_paged 中根据标志选择 page_table 获取方式和 page_size 断言。
3. 在 index_buf_accessor.py 中使用同一标志决定 cp_gather_indexer_k_quant_cache 的导入和 GetKAndS.execute 的分发逻辑。
4. 在 server_args.py 的 _handle_model_specific_adjustments 中将原本硬编码的 page_size=64 改为条件设置：HIP 平台且 preshuffle 不可用时设为 1，否则设为 64。采用延迟导入避免循环依赖。
5. 在 memory_pool.py 的 NSATokenToKVPool 构造函数中相应调整 HIP 下的 page_size 断言。

关键文件：

- python/sglang/srt/layers/attention/nsa/utils.py（模块 NSA 检测；类别 source；类型 dependency-wiring；符号 aiter_can_use_preshuffle_paged_mqa）：新增核心检测函数 aiter_can_use_preshuffle_paged_mqa，是回退决策的入口。

- python/sglang/srt/layers/attention/nsa/nsa_indexer.py（模块 NSA 索引；类别 source；类型 dependency-wiring）：使用检测结果控制 page_size 和路径选择，是回退的主执行端。
- python/sglang/srt/layers/attention/nsa/index_buf_accessor.py（模块 NSA 缓存；类别 source；类型 dependency-wiring）：根据 preshuffle 可用性控制 cp_gather 导入和 GetKAndS 分发
- python/sglang/srt/server_args.py（模块 服务配置；类别 source；类型 dependency-wiring）：配置 page_size 根据检测结果动态设置
- python/sglang/srt/mem_cache/memory_pool.py（模块 内存池；类别 source；类型 dependency-wiring）：相应调整 HIP 下 page_size 断言

关键符号：aiter_can_use_preshuffle_paged_mqa

关键源码片段

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

新增核心检测函数 aiter_can_use_preshuffle_paged_mqa，是回退决策的入口。

```
from functools import lru_cache
from sglang.srt.utils import get_bool_env_var, is_hip
import triton

@lru_cache(maxsize=1)
def aiter_can_use_preshuffle_paged_mqa() -> bool:
    # 判断 aiter 的 preshuffle paged-MQA kernel 是否可用。
    # 返回 True 的条件：
    # - 平台是 HIP (ROCm)
    # - SGLANG_USE_AITER = 1
    # - SGLANG_NSA_HIP_DISABLE_PRESHUFFLE 未设置
    # - AITER_ENABLE_AOT_GLUON_PA_MQA_LOGITS = 1 或者 Triton >= 3.5.0
    if not is_hip():
        return False
    if not get_bool_env_var('SGLANG_USE_AITER'):
        return False
    if get_bool_env_var('SGLANG_NSA_HIP_DISABLE_PRESHUFFLE'):
        return False
    if get_bool_env_var('AITER_ENABLE_AOT_GLUON_PA_MQA_LOGITS'):
        return True
    try:
        from packaging.version import Version
        return Version(Version(triton.__version__).base_version) >= Version('3.5.0')
    except Exception:
        return False
```

python/sglang/srt/layers/attention/nsa/nsa_indexer.py

使用检测结果控制 page_size 和路径选择，是回退的主执行端。

```
from sglang.srt.layers.attention.nsa.utils import aiter_can_use_preshuffle_paged_mqa
```

```
# 模块级标志
_use_aiter_preshuffle = aiter_can_use_preshuffle_paged_mqa()
if _use_aiter and not _use_aiter_preshuffle:
    logger.warning('回退到 legacy page_size=1 路径（需要 Triton>=3.5.0 或 AOT 内核）')

# 在 _get_topk_paged 中
if _is_hip:
    if _use_aiter_preshuffle:
        assert page_size % 16 == 0, 'HIP preshuffle 需要 page_size 为 16 的倍数'
    else:
        assert page_size == 1, 'HIP legacy 路径需要 page_size == 1'
if _is_hip and not _use_aiter_preshuffle:
    block_tables = metadata.get_page_table_1()
else:
    block_tables = metadata.get_page_table_64()
```

评论区精华

PR 获得两名 reviewer 批准，无公开评论。PR body 强调了 Triton 版本门控和环境变量 SGLANG_NSA_HIP_DISABLE_PRESHUFFLE 的设计意图，以及用于 CI 调试的用途。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 性能回归：在 preshuffle 不可用的环境中，page_size 退回到 1，导致 NSA indexer 的 KV 缓存块更小，可能增加访存开销，但这是确保可用性的必要妥协。
  2. 环境变量误用：若用户意外设置 SGLANG_NSA_HIP_DISABLE_PRESHUFFLE=1，即使环境支持也会强制降级，需通过文档或警告提示。
  3. 缺少单元测试：本次变更未引入新单元测试，但 CI 流水线在 rocm700（回退路径）和 rocm720（preshuffle 路径）均通过准确性和速度验证。 - 影响：用户影响：ROCm 平台上使用 DeepSeekV3.2 且 Triton<3.5 的用户将不再遇到崩溃，但 NSA indexer 性能可能略低于 preshuffle 路径（至多等同于 23562 之前的水平）。系统影响：变更局限在 NSA indexer 子系统的 5 个文件，不触及 MLA、MoE 等其他模块。团队影响：检测逻辑集中在一处，维护成本低。 - 风险标记：回退路径性能下降，缺少单元测试，环境变量依赖

关联脉络

- PR #23562 Bump ROCm NSA indexer to page_size=64 with Preshuffle: 引入 preshuffle 路径但未考虑 Triton 版本依赖，本 PR 为其添加自动回退