

PR #35396 完整报告

sgl-project/sglang

[Fix] Assert the page-aligned SWA evict floor on both PD decode prealloc paths

合并时间: 2026-08-19 06:48

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35396>

执行摘要

- 一句话: PD 解码预分配两路径统一页对齐断言并修复 hisparse 测试
- 推荐动作: 值得快速阅读, 并与 #35286 一起看以理解页对齐断言的演进脉络。重点学习测试 fixture 的修复手法: 不要 mock 计算函数, 而应注入真实输入让真实逻辑运行, 使单测与生产行为保持一致。

功能与动机

PR body 指出: 'Mirrors the assert added in #35286 onto the non-hisparse prealloc path, and repairs the hisparse fixture that mocked `_swa_tail_len` to an unaligned value (currently failing on main in base-a-test-cpu)'. #35286 只在 hisparse 路径校验了 SWA 驱逐长度的页对齐性, 非 hisparse 路径仍可能写下未对齐值; 旧测试用 mock 绕过真实计算, 掩盖了页不对齐问题, 导致 main 分支 CI 失败。

实现拆解

1. 源码断言: 在 `python/sglang/srt/disaggregation/decode.py` 的 `alloc_for_decode_prealloc` 函数 `uses_swa_tail` 分支中, 将原先直接赋值 `req.kv.swa_evicted_seqlen = fill_len - swa_tail_len` 改为先计算局部变量 `swa_evicted_seqlen`, 并断言其非负且 `% allocator.page_size == 0`, 通过断言后才写入请求 KV 元数据。该断言与 #35286 在 hisparse 路径上的一致, 补齐非 hisparse 路径的校验缺口。
2. 测试 fixture 修复: 在 `test/registered/unit/mem_cache/test_hisparse_allocator.py` 的 `test_hisparse_prealloc_uses_swa_tail_for_direct_host_path` 中, 将 `swa_tail_len` 期望值从 128 改为 256, 新增 `sliding_window_size = 200` 注入 `queue.scheduler`, 并删除 `queue._swa_tail_len` 的 mock。这样 `_swa_tail_len` 的真实计算 `floor_align(512 - 200, 256) = 256` 生效, 使 `fill_len - swa_tail_len` 页对齐并通过新断言。
3. 配套验证: 作者通过 `/rerun-test` 重跑 `test_hisparse_allocator.py`、`test_disaggregation_basic.py`、`test_disaggregation_dsv4.py`, `ubuntu-latest`、`2-gpu-h100`、`8-gpu-h200` 三个平台均通过; 无新增配置或 schema 改动。

关键文件:

- `python/sglang/srt/disaggregation/decode.py` (模块 PD 解码; 类别 source; 类型 core-logic; 符号 `alloc_for_decode_prealloc`): 核心变更文件, 在

`alloc_for_decode_prealloc` 的 `uses_swa_tail` 分支增加页对齐断言，补齐非 hisparse 路径的 SWA evict 记账校验。

- `test/registered/unit/mem_cache/test_hispase_allocator.py` (模块 稀疏分配; 类别 `test`; 类型 `test-coverage`; 符号 `test_hispase_prealloc_uses_swa_tail_for_direct_host_path`): 修复 hisparse 测试 fixture, 使 mock 值与真实计算逻辑一致, 保证新断言在测试中通过, 同时验证了页对齐拦截能力。

关键符号: `alloc_for_decode_prealloc`, `test_hispase_prealloc_uses_swa_tail_for_direct_host_path`

关键源码片段

`python/sglang/srt/disaggregation/decode.py`

核心变更文件, 在 `alloc_for_decode_prealloc` 的 `uses_swa_tail` 分支增加页对齐断言, 补齐非 hisparse 路径的 SWA evict 记账校验。

该片段位于 `alloc_for_decode_prealloc`, 处理 Tail-only SWA 分配分支。

```
if uses_swa_tail:
    # Tail-only SWA 分配: 仅在 prefix_len == 0 时有效。
    # 当 prefix_len > 0 (radix cache 命中) 时回退到 alloc_extend,
    # 按完整页数分配 SWA, SWA 预算可能略低估。
    kv_loc = allocator.alloc_extend_swa_tail(
        prefix_lens=torch.tensor([0], dtype=torch.int64, device=device),
        prefix_lens_cpu=torch.tensor([0], dtype=torch.int64),
        seq_lens=torch.tensor([fill_len], dtype=torch.int64, device=device),
        seq_lens_cpu=torch.tensor([fill_len], dtype=torch.int64),
        last_loc=last_loc,
        extend_num_tokens=fill_len,
        swa_tail_len=swa_tail_len,
        **extra_kwargs,
    )
    # 被窗口驱逐的部分 = 填充长度减去 tail 长度。
    # 该值必须非负且为 allocator.page_size 的整数倍,
    # 否则后续 evict 记账会产生未对齐的页偏移, 造成内存账目悬空。
    swa_evicted_seqlen = fill_len - swa_tail_len
    assert (
        swa_evicted_seqlen >= 0
        and swa_evicted_seqlen % allocator.page_size == 0
    )
    req.kv.swa_evicted_seqlen = swa_evicted_seqlen
```

该片段位于 `alloc_for_decode_prealloc`, 处理 Tail-only SWA 分配分支。

```
if uses_swa_tail:
```

```

# Tail-only SWA 分配: 仅在 prefix_len == 0 时有效。
# 当 prefix_len > 0 (radix cache 命中) 时回退到 alloc_extend,
# 按完整页数分配 SWA, SWA 预算可能略低估。
kv_loc = allocator.alloc_extend_swa_tail(
    prefix_lens=torch.tensor([0], dtype=torch.int64, device=device),
    prefix_lens_cpu=torch.tensor([0], dtype=torch.int64),
    seq_lens=torch.tensor([fill_len], dtype=torch.int64, device=device),
    seq_lens_cpu=torch.tensor([fill_len], dtype=torch.int64),
    last_loc=last_loc,
    extend_num_tokens=fill_len,
    swa_tail_len=swa_tail_len,
    **extra_kwargs,
)
# 被窗口驱逐的部分 = 填充长度减去 tail 长度。
# 该值必须非负且为 allocator.page_size 的整数倍,
# 否则后续 evict 记账会产生未对齐的页偏移, 造成内存账目悬空。
swa_evicted_seqlen = fill_len - swa_tail_len
assert (
    swa_evicted_seqlen >= 0
    and swa_evicted_seqlen % allocator.page_size == 0
)
req.kv.swa_evicted_seqlen = swa_evicted_seqlen

```

test/registered/unit/mem_cache/test_hispase_allocator.py

修复 hisparse 测试 fixture, 使 mock 值与真实计算逻辑一致, 保证新断言在测试中通过, 同时验证了页对齐拦截能力。

```

def test_hispase_prealloc_uses_swa_tail_for_direct_host_path(self):
    from sglang.srt.disaggregation.decode import DecodePreallocQueue

    fill_len = 512
    sliding_window_size = 200
    # _swa_tail_len 会把窗口起点 floor 到页边界:
    # floor_align(512 - 200, 256) = 256, 所以 tail 正好是 1 个整页。
    swa_tail_len = 256

    # 省略 ReqToTokenPool、allocator、coordinator 等桩构造。
    queue = DecodePreallocQueue.__new__(DecodePreallocQueue)
    queue.scheduler = SimpleNamespace(
        enable_hispase=True,
        hisparse_coordinator=coordinator,
        server_args=SimpleNamespace(disaggregation_decode_enable_radix_cache=False),
        sliding_window_size=sliding_window_size,
    )
    queue._uses_swa_tail_prealloc = MagicMock(return_value=True)
    # 不再 mock _swa_tail_len, 让真实逻辑从 sliding_window_size 计算
    # 页对齐的 tail, 从而保证 swa_evicted_seqlen 通过新增断言。
    result = queue._pre_alloc(req)

```

```
# 后续断言验证 alloc_extend_swa_tail 收到页对齐的 swa_tail_len,  
# 且 req.kv.swa_evicted_seqlen 等于 fill_len - swa_tail_len。
```

评论区精华

该 PR 没有 review 评论 (review_comments_count = 0)，无实质技术交锋。issue 评论仅包含 CI 指令 (/tag-and-rerun-ci、/rerun-test)，机器人报告三个平台的重跑结果全部成功，说明变更通过了针对性回归验证。

- 暂无高价值评论线程

风险与影响

- 风险：
 - 断言严格性：swa_evicted_seqlen % allocator.page_size == 0 依赖 swa_tail_len 恒为页对齐值。测试已验证真实 swa_tail_len 会 floor 到页边界，但未来若有调用方直接传入未对齐的 swa_tail_len，将立即 fail-fast 崩溃。
 - 边界条件：swa_evicted_seqlen >= 0 要求 fill_len >= swa_tail_len，若出现填充长度小于窗口尾的场景会触发断言。
 - 测试覆盖：非 hisparse 路径没有新增专项单测，回归保障依赖现有集成测试；hisparse 测试的修复本身验证了新断言能拦截未对齐值。
 - 影响：影响范围限于 PD 解码预分配路径 (disaggregation/decode.py)，涉及 SWA evict 记账的页对齐一致性。对用户无可见功能变化，但提升了内存账目正确性，能尽早发现页不对齐导致的静默内存损坏。团队层面，本 PR 与 #35286、#35265、#35382 共同构成 SWA/ 解码分配页对齐的加固系列，降低跨路径回归风险。
- 风险标记：核心路径变更，断言依赖页对齐假设，非 hisparse 路径缺专项测试

关联脉络

- PR #35286 Assert the page-aligned SWA evict floor at PD decode prealloc: 本 PR 将其页对齐断言镜像到非 hisparse 路径，是直接的上游变更。
- PR #35265 [Spec] Page-align the DFLASH decode KV reservation: 同一页对齐主题的修复，涉及 DFLASH decode KV 预留页对齐。
- PR #35382 [Refactor] Share the page-aligned decode alloc lens between EAGLE and DFLASH: 页对齐 decode 分配水位的重构，与本 PR 同属分配页对齐一致性主题。