PR #32427 完整报告

sgl-project/sglang

add fill_draft_extend_prepare_buffers_native for NPU

合并时间：2026-07-26 20:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32427

执行摘要

- 一句话：为 NPU 添加 native buffer 填充，修复 triton 内核及 disagg 参数错误
- 推荐动作：建议 NPU 相关开发者精读，尤其是 fill_draft_extend_prepare_buffers_native 的实现方式与条件导入技巧。对于非 NPU 开发者可跳过。后续应考虑为 Native 实现补充单元测试，并与 Triton 版本持续保持行为对标。

功能与动机

PR body 指出两项 Bug：

1) "Mimo-v2-flash model encountered a 'fill_draft_extend_prepare_buffers_triton' error when graph mode was enabled"； 2) "AscendKVManager.send_kvcache() takes 6 positional arguments but 7 were given"。此外代码注释说明 Triton 巨内核在 NPU 上会触发 CCU 错误，因此需要提供原生 PyTorch 实现。

实现拆解

1. 新增 Native 实现：在 python/sglang/kernels/ops/speculative/multi_layer_eagle.py 中新增 fill_draft_extend_prepare_buffers_native 函数，逐元素拷贝输入张量、填充 padding 并计算 select_index，语义与 Triton 版本完全一致。
2. 导出符号：在 python/sglang/srt/speculative/multi_layer_eagle_utils.py 的 __all__ 中增加 fill_draft_extend_prepare_buffers_native，使其可被外部导入。
3. 条件替换调用点：在 python/sglang/srt/speculative/multi_layer_eagle_draft_extend_cuda_graph_runner.py 中，导入 is_npu 工具并在模块作用域内按条件判断：若为 NPU 则从 multi_layer_eagle_utils 导入 fill_draft_extend_prepare_buffers_native 并覆盖全局名称 fill_draft_extend_prepare_buffers；prepare 方法中的调用点统一改为 fill_draft_extend_prepare_buffers 从而自动选择后端。
4. 修复 Disagg 参数：在 python/sglang/srt/disaggregation/ascend/conn.py 中，为 AscendKVManager.send_kvcache 方法签名添加 dst_layer_ids: Optional[List[int]] = None 参数，解决调用方传入 7 个参数而定义只匹配 6 个的问题。

关键文件：

- python/sglang/srt/speculative/multi_layer_eagle_draft_extend_cuda_graph_runner.py（模块 推测解码；类别 source；类型 dependency-wiring）：核心调用点修改，通过条件导入实现 NPU 下自动切换到 Native 实现，是分发的关键入口。

- python/sglang/kernels/ops/speculative/multi_layer_eagle.py（模块 内核操作；类别 infra；类型 infrastructure；符号 fill_draft_extend_prepare_buffers_native）：新增 fill_draft_extend_prepare_buffers_native 函数，是 NPU 兼容性的核心实现。
- python/sglang/srt/disaggregation/ascend/conn.py（模块 昇腾后端；类别 source；类型 dependency-wiring；符号 send_kvcache）：修复 send_kvcache 参数不匹配问题，添加可选参数 dst_layer_ids。
- python/sglang/srt/speculative/multi_layer_eagle_utils.py（模块 推测解码工具；类别 source；类型 core-logic；符号 fill_draft_extend_prepare_buffers_native）：导出 fill_draft_extend_prepare_buffers_native，使其可被外部模块导入。

关键符号：fill_draft_extend_prepare_buffers_native, send_kvcache

关键源码片段

python/sglang/srt/speculative/multi_layer_eagle_draft_extend_cuda_graph_runner.py

核心调用点修改，通过条件导入实现 NPU 下自动切换到 Native 实现，是分发的关键入口。

```
# 在模块顶部按硬件条件导入 native 实现并替换全局引用
from sglang.srt.utils import is_npu

if is_npu():
    from sglang.srt.speculative.multi_layer_eagle_utils import (
        fill_draft_extend_prepare_buffers_native,
    )
    # 将统一函数名指向 native 版本，Triton 版本不再使用
    fill_draft_extend_prepare_buffers = fill_draft_extend_prepare_buffers_native

# 在 prepare 方法中统一调用 fill_draft_extend_prepare_buffers
#  （原为 fill_draft_extend_prepare_buffers_triton，已通过别名覆盖）
def prepare(self, forward_batch: ForwardBatch):
    # ...
    fill_draft_extend_prepare_buffers(
        buffers.input_ids,
        buffers.positions,
        # ... 其余参数不变
    )
```

python/sglang/kernels/ops/speculative/multi_layer_eagle.py

新增 fill_draft_extend_prepare_buffers_native 函数，是 NPU 兼容性的核心实现。

```
def fill_draft_extend_prepare_buffers_native(
    input_ids, positions, out_cache_loc,
    src_input_ids, src_positions, src_out_cache_loc,
    seq_lens, req_pool_indices,
    num_correct_drafts, num_accept_tokens,
    select_index, temperatures,
    src_seq_lens, src_req_pool_indices,
```

```
    src_num_correct_drafts, src_num_accept_tokens,
    src_temperatures,
    hidden_states, src_hidden_states,
    global_num_tokens, global_num_tokens_for_logprob,
    raw_bs, bs, num_tokens_per_bs, num_front_tokens,
    seq_len_fill_value,
):
    """Native PyTorch implementation of fill_draft_extend_prepare_buffers.

    Used on NPU where the triton mega-kernel triggers CCU errors.
    Mirrors the triton kernel's semantics exactly.
    """
    num_tokens = src_input_ids.shape[0]

    # 拷贝实际 token 数据，padding 区域保持不动
    input_ids[:num_tokens].copy_(src_input_ids)
    positions[:num_tokens].copy_(src_positions)
    out_cache_loc[:num_tokens].copy_(src_out_cache_loc)

    # 逐请求字段填充：先填 padding 值，再覆盖有效行
    seq_lens[:bs].fill_(seq_len_fill_value)
    seq_lens[:raw_bs].copy_(src_seq_lens)
    req_pool_indices[:raw_bs].copy_(src_req_pool_indices)

    num_correct_drafts[:raw_bs].copy_(src_num_correct_drafts)
    num_accept_tokens[:bs].fill_(-1)
    num_accept_tokens[:raw_bs].copy_(src_num_accept_tokens)

    # select_index = i * num_tokens_per_bs + num_front_tokens + num_correct_drafts
    idx = torch.arange(bs, device=select_index.device, dtype=torch.int64)
    select_index[:bs] = idx * num_tokens_per_bs + num_front_tokens
    select_index[:bs] += num_correct_drafts[:bs].to(torch.int64)

    if temperatures is not None:
        temperatures[:bs].fill_(1.0)
        temperatures[:raw_bs].copy_(src_temperatures)

    if global_num_tokens is not None:
        global_num_tokens.fill_(bs * num_tokens_per_bs)
        global_num_tokens_for_logprob.fill_(bs * num_tokens_per_bs)

    if src_hidden_states is not None:
        hidden_states[:num_tokens].copy_(src_hidden_states)
```

python/sglang/srt/disaggregation/ascend/conn.py

修复 send_kvcache 参数不匹配问题，添加可选参数 dst_layer_ids。

```
# 原签名缺少 dst_layer_ids 导致调用方传入 7 个参数时崩溃
# 修复后添加可选参数以兼容上层调用
```

```
def send_kvcache(
    self,
    mooncake_session_id: str,
    prefill_kv_indices: npt.NDArray[np.int32],
    dst_kv_ptrs: list[int],
    dst_kv_indices: npt.NDArray[np.int32],
    executor: concurrent.futures.ThreadPoolExecutor,
    dst_layer_ids: Optional[List[int]] = None, # 新增参数
):
    # 方法体不变 ...
```

评论区精华

该 PR 无实质性 Review 讨论，仅由 sglang-npu-bot 自动批准并合并。Bot 在 Issue 评论中声明“Only the NPU-related code is modified. Once the NPU test cases pass, the changes are merged.”，体现了 NPU 专用代码的独立合并策略。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 语义一致性：Native 实现与 Triton 版本需保持完全一致，若后期 Triton 版本更新而 Native 未同步，可能引入行为差异。目前代码注释已声明“Mirrors the triton kernel's semantics exactly”，但缺少自动化对照测试。
  2. 测试覆盖缺失：本次变更未添加任何单元测试，回归风险依赖 NPU CI 集成测试（如 GSM8K 精度测试），若 CI 用例不全面可能遗漏错误。
  3. 非 NPU 环境无影响：条件导入仅作用于 is_npu() 为真的场景，其他硬件行为不变，因此无副作用。
  4. Disagg 参数兼容：dst_layer_ids 默认值为 None，不影响现有调用，但若调用方未传递该参数且后续逻辑依赖它，可能引发问题。 - 影响：影响范围：仅 NPU 用户可见，修复了 Mimo-v2-flash 在 graph 模式下的崩溃和 disagg 参数错误，使该模型能在 NPU 上正常运行。影响程度：中等，因为修复的是关键 crash，对 NPU 用户是阻塞性修复；其他用户无感知。团队影响：NPU 后端维护者可参考此模式为其他 Triton 内核提供 Native fallback。 - 风险标记：缺少对应测试覆盖，native 语义一致性依赖开发者，仅 NPU 环境生效

关联脉络

- PR #32270 fix(disagg): support pipeline-parallel hybrid-linear transfer: 同样修改了 ascend/conn.py，处理 PP 场景下的层配对问题，与本 PR 的 dst_layer_ids 参数有潜在关联。