PR #25917 完整报告

sgl-project/sglang

Revert "[AMD]fix: use CUDA event for targeted draft-to-verify sync in…

合并时间：2026-05-21 09:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25917

执行摘要

- 一句话：回退 AMD EAGLE overlap CUDA event 同步
- 推荐动作：建议仔细测试 AMD 环境下 EAGLE overlap 功能的正确性和性能，确保 wait_stream 同步足够可靠。如果可以，考虑在文档中明确说明 SGLANG_ENABLE_OVERLAP_PLAN_STREAM 环境变量的作用和使用场景。

功能与动机

PR body 明确指出原提交硬编码了 SGLANG_ENABLE_OVERLAP_PLAN_STREAM=0，而用户应当通过设置该环境变量来控制 overlap 行为。因此需要回退，避免代码逻辑被错误覆盖。

实现拆解

1. 回退配置：在 eagle_worker_v2.py 的 forward_batch_generation 方法中删除了 draft 完成后记录 CUDA event 的逻辑。
2. 回退同步：在 verify 方法中删除了 plan_stream 等待该 event 的逻辑，恢复为原有的 wait_stream 同步方式。
3. 同步回退多 worker：在 multi_layer_eagle_worker_v2.py 中执行相同的删除操作，保持行为一致。

关键文件：

- python/sglang/srt/speculative/eagle_worker_v2.py（模块 推理引擎；类别 source；类型 core-logic）：删除 draft 完成后记录 CUDA event 的代码，以及在 verify 中等待该 event 的逻辑。
- python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py（模块 推理引擎；类别 source；类型 core-logic）：同步回退多 worker 中的相同改动。

关键符号：未识别

关键源码片段

python/sglang/srt/speculative/eagle_worker_v2.py

删除 draft 完成后记录 CUDA event 的代码，以及在 verify 中等待该 event 的逻辑。

```
# python/sglang/srt/speculative/eagle_worker_v2.py 中删除的代码片段
```

```
# 在 forward_batch_generation 方法中原本如下：
# verify_input: EagleVerifyInput = self.draft_worker.draft(batch)
# assert verify_input.is_verify_input()
# # 以下代码被删除：
# if self.plan_stream:
# self._draft_done_event = torch.get_device_module(self.device).Event()
# self._draft_done_event.record()
# batch.spec_info = verify_input

# 在 verify 方法中原本如下：
# with self.plan_stream_ctx:
# # 以下代码被删除：
# if self.plan_stream and hasattr(self, "_draft_done_event"):
# self.plan_stream.wait_event(self._draft_done_event)
# verify_forward_batch, can_run_cuda_graph = (
# verify_input.prepare_for_v2_verify(...)
# )
```

python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py

同步回退多 worker 中的相同改动。

```
# python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py 中删除的代码片段

# 在 forward_batch_generation 方法中原本如下：
# verify_input: EagleVerifyInput = self.draft_worker.draft(batch)
# assert verify_input.is_verify_input()
# # 以下代码被删除：
# if self.plan_stream:
# self._draft_done_event = torch.get_device_module(self.device).Event()
# self._draft_done_event.record()
# batch.spec_info = verify_input

# 在 verify 方法中原本如下：
# with self.plan_stream_ctx:
# # 以下代码被删除：
# if self.plan_stream and hasattr(self, "_draft_done_event"):
# self.plan_stream.wait_event(self._draft_done_event)
# verify_forward_batch, can_run_cuda_graph = (
# verify_input.prepare_for_v2_verify(...)
# )
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：回归风险：回退可能导致 AMD 设备上 EAGLE overlap 的同步精度下降——原提交使用 event 实现更细粒度的同步，回退后使用 wait_stream 可能等待更多不必要的主 stream 操作，但影响程度取决于具体场景。

兼容性风险：无，回退是恢复到更通用的实现。

性能风险：对于 AMD 设备，如果 overlap 功能被启用（SGLANG_ENABLE_OVERLAP_PLAN_STREAM=1），可能因为同步粒度变粗而有微小性能损失。但维护者的观点是，用户可以通过环境变量完全禁用 overlap 来避免此问题。

- 影响：对用户的影响：受影响用户主要是使用 AMD GPU 且启用 EAGLE overlap 功能的用户。这些用户需要手动设置 SGLANG_ENABLE_OVERLAP_PLAN_STREAM=0 来避免原提交修复的同步问题，而不是依赖代码硬编码。

对系统的影响：代码更简洁，移除了约 21 行与 CUDA event 相关的同步逻辑。

影响程度：中等，仅影响特定平台（AMD）和特定功能（EAGLE overlap）。

- 风险标记：回归风险，平台特定

关联脉络

- PR #21940 [AMD]fix: use CUDA event for targeted draft-to-verify sync in EAGLE overlap: 被回退的原始 PR