PR #21940 完整报告

sgl-project/sglang

[AMD]fix: use CUDA event for targeted draft-to-verify sync in EAGLE overlap

合并时间：2026-04-27 12:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21940

执行摘要

- 一句话：修复 EAGLE overlap 中 draft 与 verify 的 GPU 同步问题
- 推荐动作：值得合并。该修复解决了 AMD 平台 spec v2 数据竞态 bug，且方案在 NVIDIA 平台也验证有效。设计上使用 CUDA event 替代 wait_stream 实现更精确同步，也是正确方向。建议后续考虑 gemini-code-assist 的抽取建议提升代码质量。

功能与动机

Issue #21942 报告在 AMD GPU 上使用 spec v2 + DP 配置时出现 memory access fault。根本原因是 EAGLE overlap 模式下 draft 阶段在 plan_stream 上执行的 GPU 工作与后续 verify 阶段在 plan_stream 上执行的准备工作之间缺少同步，导致数据竞态。

实现拆解

1. 在 eagle_worker_v2.py 的 forward_batch_generation 方法中，draft 调用完成后、verify 调用之前，新增 CUDA event 记录：若 self.plan_stream 存在，则创建 CUDA event 并记录到当前流（主流）。
2. 在 eagle_worker_v2.py 的 verify 方法中，plan_stream_ctx 开始处增加 wait_event：检查 self.plan_stream 且 _draft_done_event 属性存在时，调用 self.plan_stream.wait_event(self._draft_done_event)，确保 plan_stream 等待 draft 的 GPU 工作完成后再开始元数据准备。
3. 在 multi_layer_eagle_worker_v2.py 的对应方法中应用完全相同的两处变更。
4. 该方案比原有的 wait_stream(main_stream) 更精确，只等待 draft GPU 工作而非所有主流操作，减少不必要的同步开销。

关键文件：

- python/sglang/srt/speculative/eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic；符号 forward_batch_generation, verify）：核心变更文件，添加了 event 记录和 wait_event 同步逻辑，修复数据竞态
- python/sglang/srt/speculative/multi_layer_eagle_worker_v2.py（模块 推测解码；类别 source；类型 core-logic；符号 forward_batch_generation, verify）：多推测头 worker 的对应变更，保持与单层 worker 一致的行为

关键符号：forward_batch_generation, verify

关键源码片段

python/sglang/srt/speculative/eagle_worker_v2.py

核心变更文件，添加了 event 记录和 wait_event 同步逻辑，修复数据竞态

```
# eagle_worker_v2.py forward_batch_generation 中新增：
# 在 draft() GPU 工作分发后记录一个 CUDA event
if self.plan_stream:
    self._draft_done_event = torch.get_device_module(self.device).Event()
    self._draft_done_event.record()

# verify 方法中 plan_stream_ctx 开始处新增：
with self.plan_stream_ctx:
    # 使用 wait_event 比 wait_stream 更精确，只等待 draft GPU 工作
    if self.plan_stream and hasattr(self, "_draft_done_event"):
        self.plan_stream.wait_event(self._draft_done_event)
```

评论区精华

Review 中 gemini-code-assist[bot] 建议将重复的 torch.get_device_module(self.device).Event() 调用抽取为辅助方法以提高可维护性。HaiShaw 询问性能数据，提交者提供了对比表，显示 Specv2(Sync) 方案与原版非 spec 基线性能接近，Specv2(async) 方案因复杂度高且提升微小被考虑放弃。未解决疑虑：hanming-lu 评论称该改动等同于硬编码 SGLANG_ENABLE_OVERLAP_PLAN_STREAM=0，应回退。

- 重复代码抽取建议 (style): 未采纳，但评论者未坚持，不影响合并
- 性能数据确认 (other): 提交者提供了性能对比数据，HaiShaw 随后 approve
- 是否等同于硬编码 SGLANG_ENABLE_OVERLAP_PLAN_STREAM=0 (design): 未进一步讨论，PR 已合并

风险与影响

- 风险：风险较低。新增代码在 if self.plan_stream 条件内，不会影响非 overlap 路径。但若存在多个 draft 阶段快速连续执行场景，_draft_done_event 可能被覆盖，需确保事件记录与等待成对。测试覆盖方面，缺少专门的单元测试，依赖 CI 集成测试验证。
- 影响：直接影响使用 speculative decoding v2 且开启 overlap 的部署，特别是 AMD GPU 平台。修复了可能导致服务器崩溃的 memory access fault，提升稳定性和可靠性。对 NVIDIA GPU 平台同样有效（已确认 B200 可复现）。无性能回退风险（提交者提供数据表明性能与原版基线接近）。
- 风险标记：核心路径变更，缺少测试覆盖

关联脉络

- PR #25465 verify_done: wait not synchronize: 同一 spec v2 overlap 同步问题的先前尝试，使用 event.wait() 替代 synchronize()
- PR #22008 （未提供）：PR 讨论中提及的 Specv2(async) 方案，因复杂度高被考虑放弃