PR #25945 完整报告

sgl-project/sglang

[Scheduler] Defer prefill input_ids H2D to forward stream, unify resolve via future_map

合并时间：2026-05-30 17:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25945

执行摘要

- 一句话：推迟 prefill input_ids 的 H2D 拷贝至 forward 流，统一 resolve 路径
- 推荐动作：值得精读，展示了如何通过 FutureMap 统一不同模式（overlap/non-overlap、prefill/decode）下的输入准备。设计决策如"始终初始化 FutureMap"和"通过 sentinel None 触发 relay"值得关注。建议合并前确保 benchmark 无性能回退。

功能与动机

减少调度流上的 GPU 拷贝开销，统一 overlap 和非 overlap 场景的 input_ids 生成方式，简化代码维护。同时修复了 penalty 路径在 input_ids 为 None 时的崩溃以及 spec_v2 隔离的 staging 问题。

实现拆解

1. 新增 pinned CPU 中继字段：在 ScheduleBatch 类中添加 prefill_input_ids_cpu 和 mix_running_indices 字段（schedule_batch.py），分别暂存 prefill 的 prompt tokens（CPU pinned）和混合 batch 中 decode 部分的 req_pool_indices。
2. 拆分 flatten 工具：在 common.py 中将原来的 flatten_arrays_to_int64_tensor 拆分为 flatten_arrays_to_pinned_cpu（仅返回 CPU tensor，可 pin）和原函数（调用它后执行 to(device)），方便调度流仅保留 CPU 版本。
3. 调整 prepare_for_extend：schedule_batch.py 的 prepare_for_extend 不再直接创建 GPU tensor 赋值给 self.input_ids，而是调用 flatten_arrays_to_pinned_cpu 得到 pinned CPU tensor 赋值给 self.prefill_input_ids_cpu，并将 self.input_ids 置为 None。类似调整了 prepare_encoder_info_extend 和 mix_with_running。
4. 新增 resolve_forward_inputs：在 overlap_utils.py 中新增核心函数 resolve_forward_inputs，它在 forward 流上被调用，根据 batch.prefill_input_ids_cpu 执行 H2D（若存在）或从 future_map.output_tokens_buf gather decode token，并拼接混合 batch。同时从此函数中处理 spec_v2 的 extra relay。
5. 统一 FutureMap 初始化：在 scheduler.py 的 init_overlap 中，无论是否开启 overlap，都初始化 forward_stream_ctx 和 FutureMap（但 overlap 关闭时不创建 batch_record_buf 等），并将 resolve_forward_inputs 的调用插入到 forward 流隔离的开头。

6. 适配其余路径：包括 hisparse 重新加入、disagg PREBUILT 非 spec 路径、PP rank-0 的 mixed-chunk、encoder-decoder 重建等，均改为先 stash token 到 FutureMap 再设置 input_ids=None，由 resolve_forward_inputs 统一处理。spec_v1（非 overlap spec）保持原有直接赋值，不通过 relay（因为形状不匹配）。
7. 修复相关问题：修复 penalty 路径读取 None input_ids 的问题（改为从 Req.output_ids 累加）；修复 spec_v2 隔离中重新安装已消费的 staging 导致的问题。

关键文件：

- python/sglang/srt/managers/overlap_utils.py（模块 调度器；类别 source；类型 core-logic；符号 _resolve_future_token_ids_native, resolve_forward_inputs, resolve_future, set_input_ids_sentinel）：核心变更文件：新增 resolve_forward_inputs 函数统一 input_ids 生成，修改 FutureMap 类为 always-on，删除旧的 _resolve_future_token_ids 等函数。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 dependency-wiring）：依赖注入：导入 resolve_forward_inputs，修改 init_overlap 始终初始化 forward_stream_ctx 和 FutureMap，在 run_batch 中调用 resolve_forward_inputs。
- python/sglang/srt/managers/schedule_batch.py（模块 调度批处理；类别 source；类型 core-logic）：数据结构变更：添加 prefill_input_ids_cpu 和 mix_running_indices 字段，修改 prepare_for_extend 等函数改为保留 CPU pinned tensor。
- python/sglang/srt/utils/common.py（模块 工具函数；类别 source；类型 core-logic；符号 flatten_arrays_to_int64_tensor, flatten_arrays_to_pinned_cpu）：工具函数：新增 flatten_arrays_to_pinned_cpu，将原 flatten 函数拆分为 CPU 化 +H2D 两步。

关键符号：resolve_forward_inputs, flatten_arrays_to_pinned_cpu, prepare_for_extend, mix_with_running, process_prebuilt

评论区精华

无公开 review 讨论。作者在提交消息中标记了修复的 BUG-1（penalty 读取 None input_ids）和 BUG-3（spec_v2 隔离重新安装已消费的 staging）。

- 暂无高价值评论线程

风险与影响

- 风险：核心调度路径变更，涉及 input_ids 可能为 None 的广泛处理，若某条路径未正确设置 prefill_input_ids_cpu 或未通过 FutureMap relay，可能导致空指针崩溃或静默错误。spec_v1 未使用新的 relay，保持原有逻辑，但需要确保 shape 兼容性。性能方面，H2D 推迟到 forward 流可能增加 forward 流负担，但与调度流重叠可能提升整体吞吐，需 benchmark 验证。本 PR 未包含直接测试，存在回归风险。
- 影响：对用户无功能变化，但可能提升 prefill 与 decode 混合场景的吞吐。对系统调度流减少了 GPU 同步操作，forward 流增加了 H2D 异步拷贝。对团队统一了 input_ids 生成逻辑，简化了后续维护。影响范围包括所有使用 SGLang Runtime 的模型，特别是启用 overlap 或 PP 的场景。

- 风险标记：核心路径变更，spec 兼容性，缺少测试覆盖

关联脉络

- 暂无明显关联 PR