PR #25982 完整报告

sgl-project/sglang

Fix disaggregation bootstrap server lifetime

合并时间：2026-05-22 14:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25982

执行摘要

- 一句话：修复分解服务 bootstrap server 被垃圾回收的问题
- 推荐动作：建议精读：改动虽小，但涉及 Python 对象生命周期管理，是常见的 GC 陷阱。值得关注的是如何在代码中通过注释保护这种隐式依赖。

功能与动机

由 @merrymercy 在 PR#25430 的 review 评论（https://github.com/sgl-project/sglang/pull/25430#discussion_r3280241269）中提出：start_disagg_service 的返回值必须保持引用，否则 bootstrap server 会被垃圾回收，导致分解行为错误。

实现拆解

1. 在 python/sglang/srt/managers/tokenizer_manager.py 的 init_disaggregation 方法中，将原先的裸调用 start_disagg_service(self.server_args) 改为赋值语句 self.bootstrap_server = start_disagg_service(self.server_args)。
2. 在赋值语句前添加注释 # Keep a reference so the bootstrap server is not garbage-collected.，说明引用的意图，防止后续清理误认为未使用而删除。
3. 已有字段 self.fake_bootstrap_room_counter 等不受影响。

关键文件：

- python/sglang/srt/managers/tokenizer_manager.py（模块 调度器；类别 source；类型 core-logic）：核心修改文件：修复分解 bootstrap server 生命周期问题，确保服务不会被提前回收。

关键符号：init_disaggregation

关键源码片段

python/sglang/srt/managers/tokenizer_manager.py

核心修改文件：修复分解 bootstrap server 生命周期问题，确保服务不会被提前回收。

```
# python/sglang/srt/managers/tokenizer_manager.py

def init_disaggregation(self):
    # PD Disaggregation
```

```
self.disaggregation_mode = DisaggregationMode(
    self.server_args.disaggregation_mode
)
# Keep a reference so the bootstrap server is not garbage-collected.
# 将返回值赋值给 self.bootstrap_server，确保 TokenizerManager
# 生命周期内 bootstrap server 保持存活，否则可能被 GC 回收
self.bootstrap_server = start_disagg_service(self.server_args)
# Single-source counter for auto-assigning fake bootstrap_room.
self.fake_bootstrap_room_counter = 0

# Encoder Disaggregation
if self.server_args.language_only:
    self.mm_receiver = create_mm_receiver(
        self.server_args,
        dtype=self.model_config.dtype,
        hf_config=self.model_config.hf_config,
    )
```

评论区精华

无 reviewer 争议。Review 由 gemini-code-assist[bot] 自动完成，确认变更无额外反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅将裸调用改为赋值，不改变 start_disagg_service 的调用语义，不影响其内部逻辑。CI 中出现的 CUDA 失败经分析均为 FlashInfer/CUTLASS/CuTe 等基础设施问题，与本次修改无关。
- 影响：影响范围仅限于使用了分解（disaggregation）模式的场景，确保底层 bootstrap server 存活，避免因垃圾回收导致的分解行为异常。对未启用分解模式的场景无任何影响。
- 风险标记：无相关风险

关联脉络

- PR #25430 disaggregation bootstrap service startup changes: 本 PR 直接回应该 PR 的 review 评论，修复了其中引入的 bootstrap server 生命周期问题。