PR #24854 完整报告

sgl-project/sglang

[RL] Call torch.cuda.empty_cache() for `in-place` pause mode to avoid OOM

合并时间：2026-05-10 14:36

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24854

执行摘要

- 一句话：修复 in-place pause 模式因缺少 empty_cache 导致的 OOM
- 推荐动作：该 PR 修复了一个明确的 OOM 问题，代码简洁，建议合入。讨论中关于跨平台兼容的取舍值得记录，未来如果有更多后端需求可考虑重构。

功能与动机

Post-weight-update processing (e.g. DeepSeek MLA w_kc/w_vc derivation, FP8 scale rebuild) creates transient CUDA allocations that fragment PyTorch's block cache. Without empty_cache(), reserved memory grows each iteration and eventually OOMs. The in_place path never calls flush_cache (to preserve KV cache), so empty_cache() was never triggered — this PR closes that gap.

实现拆解

1. 在 ContinueGenerationReqInput 数据类中添加 torch_empty_cache 字段，默认 True，允许调用方跳过。
2. 在 Scheduler.continue_generation 方法中，当 torch_empty_cache 为 True 时，调用 torch.cuda.empty_cache() 并记录前后 reserved 内存量。
3. 由于 empty_cache 在 engine 仍处于 paused 状态时执行，没有活跃流竞争，线程安全。
4. 命名与现有 UpdateWeightFromDiskReqInput 中的 torch_empty_cache 保持一致。
5. 测试：在 in-place pause + weight update 场景中确认日志触发、内存下降且不改变 loss。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 continue_generation）：核心变更，在 continue_generation 中添加 empty_cache 逻辑
- python/sglang/srt/managers/io_struct.py（模块 数据结构；类别 source；类型 core-logic；符号 ContinueGenerationReqInput）：定义 ContinueGenerationReqInput 数据类，新增 torch_empty_cache 字段

关键符号：continue_generation, ContinueGenerationReqInput

评论区精华

Review 中 hebiao064 建议使用 empty_device_cache 兼容 AMD 等其他加速器，但 ByronHsu 认为增加复杂性且与已有 torch_empty_cache 命名约定冲突，最终保持仅 CUDA 支持。此外，hebiao064 建议字段名改为 torch_empty_cache 以对齐已有命名，被接受。

- 使用 empty_device_cache 实现跨设备兼容 (design): ByronHsu 认为太混乱且与已有 torch_empty_cache 命名矛盾，决定仅支持 torch，保持简单。
- 字段命名对齐已有约定 (style): 接受并重命名。
- 安全性确认 (correctness): 未直接回应，但 PR 最终合并，可能无安全问题。

风险与影响

- 风险：仅调用 torch.cuda.empty_cache()，没有考虑 AMD ROCm 或其他后端，但 PR 明确仅针对 CUDA，且已有类似用法。empty_cache 调用开销低，但日志记录可能在高频场景下产生大量 INFO 日志，需注意日志级别配置。另外，重复在 abort/retract 中调用 empty_cache（已在 flush_cache 中调用一次）可能会导致轻微性能开销，但 PR 说明第二次调用通常是 no-op。
- 影响：修复 RL 训练中 weight update 后 in-place pause 模式的 OOM 问题，使模型更新过程更稳定。日志监控内存趋势。影响范围：使用 in-place pause mode 进行权重更新的 RL 训练场景。其他场景不受影响（新字段默认 True，但 empty_cache 只在 continue_generation 中调用，若不使用 pause 不会触发）。
- 风险标记：仅支持 CUDA，日志可能高频输出

关联脉络

- PR #24861 [Utils] Refactor device cache emptying: 该 PR 的 empty_device_cache 辅助函数被本 PR 尝试使用但最终 revert，保留直接调用 torch.cuda.empty_cache。