PR #21507 完整报告

sgl-project/sglang

[NPU] fix conflict between empty_cache and use_mem_pool

合并时间：2026-03-31 15:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21507

执行摘要

该 PR 修复了 NPU 平台上 torch.npu.use_mem_pool 与 empty_cache 操作之间的内存管理冲突，通过调整 empty_cache 调用位置并更新 Triton backend 支持列表，提升系统稳定性。

功能与动机

动机源自 NPU 内存管理中的冲突问题：当 self.memory_saver_adapter.region 调用 torch.npu.use_mem_pool 函数时，如果在该范围内执行 empty_cache，会导致冲突。PR body 明确指出“This PR is intended to resolve this issue”，旨在避免潜在崩溃。

实现拆解

改动涉及三个文件：

- model_runner.py：在 load_model 函数中添加 torch.npu.empty_cache() 调用，并注释说明移动原因以避免与 use_mem_pool 冲突。
- loader.py：从 load_weights_and_postprocess 函数中删除相同的 empty_cache 调用，消除重复操作。
- common.py：修改 support_triton 函数，将“ascend”添加到不支持 backend 列表，确保兼容性检查正确。

评论区精华

本次 PR 没有进行 review 讨论，直接由维护者合并。这表明变更被视为低风险且直接，但缺乏 peer review 可能隐含潜在疏忽。

风险与影响

风险：empty_cache 调用位置改变可能影响 NPU 内存池的清理时机，导致性能波动或新冲突；backend 列表修改可能影响依赖此函数的其他模块，需确保 ascend backend 的正确处理。
影响：对用户，解决 NPU 模型加载时的内存冲突，提升推理稳定性；对系统，优化内存管理流程；对团队，为 ascend backend 添加明确支持标记，促进后端兼容性开发。

关联脉络

与近期 NPU 相关 PR 关联紧密，如 PR 21209（修复 *eagle3* 接受率）、PR 21383（支持 ring attention on NPU）和 PR 20757（支持并行解码），均聚焦于 NPU 硬件后端的优化和 bugfix，揭示团队正持续完善 NPU 生态支持。