PR #26301 完整报告

sgl-project/sglang

[HiCache]: Check return code of cudaHostRegister

合并时间：2026-05-26 17:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26301

执行摘要

- 一句话：检查 cudaHostRegister 返回值，失败时抛出异常
- 推荐动作：该 PR 小而精，值得精读作为一个良好实践：在系统编程中始终检查 CUDA API 的返回码。可直接合并，无需额外关注。

功能与动机

PR body 指出，当前实现静默忽略 cudaHostRegister() 的结果。在低 CPU 内存环境中，cudaHostRegister() 偶尔因 cudaErrorInvalidValue 失败，错误被延迟到 tensor.to("cuda") 时才暴露，难以排查。PR 旨在尽早抛出异常，加快问题定位。

实现拆解

1. 文件路径：python/sglang/srt/mem_cache/memory_pool_host.py，函数 alloc_with_host_register。
2. 变更核心：将原来的 torch.cuda.cudart().cudaHostRegister(...) 的 void 调用改为捕获返回值 ret，添加 if ret != 0: raise RuntimeError(...) 检查。
3. 配套改动：无测试、配置或部署相关变更。

关键文件：

- python/sglang/srt/mem_cache/memory_pool_host.py（模块 内存管理；类别 source；类型 core-logic；符号 alloc_with_host_register）：核心变更文件，在 alloc_with_host_register 函数中增加 cudaHostRegister 返回码检查。

关键符号：alloc_with_host_register

关键源码片段

python/sglang/srt/mem_cache/memory_pool_host.py

核心变更文件，在 alloc_with_host_register 函数中增加 cudaHostRegister 返回码检查。

```
# python/sglang/srt/mem_cache/memory_pool_host.py
# alloc_with_host_register 函数，用于分配 host tensor 并注册 pinned memory

def alloc_with_host_register(
    dims,
```

```
    dtype: torch.dtype,
    device: str,
    pin_memory: bool,
    allocator: HostTensorAllocator,
) -> torch.Tensor:
    """
    Allocate tensor and register host memory with cudaHostRegister.
    CudaHostRegister only applies when pin_memory=True.
    """
    buffer = allocator.allocate(dims, dtype=dtype, device=device)
    if pin_memory:
        # 原代码直接调用 cudaHostRegister 并忽略返回值，现在检查返回码
        ret = torch.cuda.cudart().cudaHostRegister(
            buffer.data_ptr(), buffer.numel() * buffer.element_size(), 0
        )
        if ret != 0:
            # 若 cudaHostRegister 失败，立即抛出异常，避免错误延迟暴露
            raise RuntimeError(f"cudaHostRegister failed with error code {ret}")
    return buffer
```

评论区精华

无 review 评论，只有一位 reviewer 批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。仅增加一个返回码检查和一个条件抛异常，不改变原有逻辑的流程或性能（异常仅在不正常路径触发，属于边缘情况）。无回归风险。
- 影响：影响范围极小，仅修改了一个辅助函数。用户可更早感知内存注册失败，便于诊断和调试。受影响的模块为 HiCache 的内存分配路径。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR