PR #24005 完整报告

sgl-project/sglang

[AMD] Enable dual-stream MoE on ROCm

合并时间：2026-05-07 17:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24005

执行摘要

- 一句话：通过环境变量在 ROCm 上启用双流 MoE 重叠
- 推荐动作：值得阅读以了解双流 MoE 重叠在 ROCm 上的启用方式及硬件队列限制。代码修改较小但配置知识丰富，建议部署 AMD GPU 的团队关注。

功能与动机

在 AMD ROCm 平台上，Deepseek 模型的双流代码已存在但仅对 CUDA/MUSA/NPU 开启，ROCm 路径始终走单流。PR 旨在通过 opt-in 环境变量复用现有 alt_stream 机制，利用 Mori AsyncLL kernel 释放计算单元，提升 MoE 推理性能。

实现拆解

1. 环境变量注册（python/sglang/srt/environ.py）：在 AMD & ROCm 配置区新增 SGLANG_ROCM_USE_MULTI_STREAM = EnvBool(False)，并附注释说明依赖 GPU_MAX_HW_QUEUES>=5。
2. 模型 alt_stream 条件扩展（python/sglang/srt/models/deepseek_v2.py）：在 DeepseekV2Model.__init__ 的 alt_stream 创建条件中增加 envs.SGLANG_ROCM_USE_MULTI_STREAM.get() 检查，使 ROCm 路径在启用环境变量后也能分配 alt_stream。
3. MoE 调度器警告（python/sglang/srt/layers/moe/token_dispatcher/moriep.py）：在 MoriEPDispatcher.__init__ 中添加逻辑：当环境变量已置位但 AsyncLL 模式未启用时，打印警告，提示用户应使用 --deepep-mode low_latency。
4. 文档同步（docs/references/environment_variables.md）：在环境变量表格中新增 SGLANG_ROCM_USE_MULTI_STREAM 条目，描述其作用、推荐参数和注意事项。

关键文件：

- python/sglang/srt/environ.py（模块 环境配置；类别 source；类型 core-logic；符号 SGLANG_ROCM_USE_MULTI_STREAM）：新环境变量注册入口，控制双流开关
- python/sglang/srt/models/deepseek_v2.py（模块 模型定义；类别 source；类型 core-logic；符号 init）：模型初始化中 alt_stream 条件扩展，ROCm 路径可创建双流
- python/sglang/srt/layers/moe/token_dispatcher/moriep.py（模块 MoE 调度；类别 source；类型 core-logic；符号 init）：MoE 调度器构造时添加配置错误警告

- docs/references/environment_variables.md（模块 文档；类别 docs；类型 documentation）：环境变量文档同步更新

关键符号：DeepseekV2Model.init, MoriEPDispatcher.init

关键源码片段

python/sglang/srt/models/deepseek_v2.py

模型初始化中 alt_stream 条件扩展，ROCm 路径可创建双流

```
# 来自 DeepseekV2Model.__init__ 的 alt_stream 分配逻辑
self.alt_stream = (
    torch.cuda.Stream()
    if (
        _is_cuda
        or _is_musa
        or envs.SGLANG_NPU_USE_MULTI_STREAM.get()
        or envs.SGLANG_ROCM_USE_MULTI_STREAM.get() # 新增 ROCm 路径
    )
    else None
)
# 注：当环境变量 SGLANG_ROCM_USE_MULTI_STREAM=1 时，ROCm 也可获得独立 stream
# 用于重叠 shared experts 与 routed experts 的执行
```

评论区精华

- 默认值选择：HaiShaw 询问 environ.py 中新增的 env var 是否需要置为 True，inkcherry 回复有意保持 opt-in，不改变原有行为。
- GPU_MAX_HW_QUEUES 澄清：HaiShaw 要求澄清该环境变量的必要性，inkcherry 更新注释和文档。hubertlu-tw 补充了 ROCm 硬件队列限制的技术背景，说明设置 GPU_MAX_HW_QUEUES=5 可避免流串行。
- 文档完善：billishyahao 要求将新环境变量描述添加到环境变量文档页，inkcherry 已更新。
- 环境变量默认值讨论 (question): 确认默认 False，用户需显式设置 SGLANG_ROCM_USE_MULTI_STREAM=1 启用。
- GPU_MAX_HW_QUEUES 澄清与文档 (documentation): 在代码注释和文档中明确要求 GPU_MAX_HW_QUEUES>=5。

风险与影响

- 风险：该变更默认关闭，不影响现有 ROCm 用户。当用户显式启用时，若未设置 GPU_MAX_HW_QUEUES>=5 或未使用 AsyncLL kernel（--deepep-mode low_latency），重叠效果可能不达预期，代码已通过警告提示风险。无安全性或稳定性风险。
- 影响：影响范围限定于 AMD ROCm/AITER 上的 Deepseek 模型用户。需配合 GPU_MAX_HW_QUEUES 和 Mori AsyncLL 配置方可获得性能提升（TPOT -14%，总吞吐 +15.6%）。对单流环境无影响。
- 风险标记：配置依赖，硬件限制，文档已覆盖

关联脉络

- PR #24550 [R3] Avoid implicit CUDA sync in routed experts DP slicing: 同为 MoE 专家执行优化，且涉及 CUDA stream 和同步，有上下文关联。