PR #23604 完整报告

sgl-project/sglang

[NPU]Fix support_triton bug

合并时间：2026-04-26 21:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23604

执行摘要

- 一句话：NPU 上 support_triton 误将 ascend 排除，修复 MTP 性能
- 推荐动作：值得合入，修复性能回退。审查简单，风险低。

功能与动机

修复 support_triton bug，该 bug 影响 MTP 性能。compute_position 应该使用 Triton。

实现拆解

修改 python/sglang/srt/utils/common.py 中的 support_triton 函数，从排除列表中移除 "ascend"。变更仅一行，将 return backend not in ["torch_native", "intel_amx", "ascend"] 改为 return backend not in ["torch_native", "intel_amx"]。

关键文件：

- python/sglang/srt/utils/common.py（模块 工具函数；类别 source；类型 core-logic；符号 support_triton）：修改了 support_triton 函数，从排除列表中移除 "ascend"，修复 NPU 性能 bug。

关键符号：support_triton

关键源码片段

python/sglang/srt/utils/common.py

修改了 support_triton 函数，从排除列表中移除 "ascend"，修复 NPU 性能 bug。

```
def support_triton(backend: str) -> bool:
    # 之前在 #21507 中错误地将 "ascend" 加入了排除列表，
    # 导致 Ascend NPU 后端无法使用 Triton 内核，
    # 进而影响 compute_position 等操作的性能。
    # 本 PR 将其移除，恢复 Ascend 对 Triton 的支持。
    return backend not in ["torch_native", "intel_amx"]
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更简单，仅影响 Ascend NPU 后端。需要确保其他使用 support_triton 的地方不会因为 "ascend" 不被排除而产生问题。
- 影响：影响：主要面向 NPU 用户，修复 MTP 性能回退。影响范围限于 Ascend 后端。
- 风险标记：缺少测试覆盖

关联脉络

- PR #21507 Unknown: 本 PR 回滚了 #21507 中引入的变更，该变更加入了 "ascend" 排除项。