PR #28436 完整报告

sgl-project/sglang

[NPU] Use use_dsa to dispatch Ascend DSA attention

合并时间：2026-06-17 15:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28436

执行摘要

- 一句话：修复 Ascend NPU 注意力调度条件
- 推荐动作：建议合入。这是一个正确性提升的小修补，使调度条件与模型语义对齐，避免因实现细节变动而引入隐蔽的 bug。可作为代码健康性改进的示例。

功能与动机

PR body 指出：原有的 hasattr(attn, "indexer") 依赖注意力模块的实现细节，而非 DSA 模型的语义指示器，在模块布局变化或调试 GLM DSA 模型时容易失效。use_dsa 已从模型配置设置，直接表示当前注意力模块是否应使用 DSA，因此应使用 attn.use_dsa 进行调度。

实现拆解

1. 修改调度条件：在 python/sglang/srt/models/deepseek_common/attention_backend_handler.py 的 handle_attention_ascend 函数中，将两处判断条件从 if hasattr(attn, "indexer") 改为 if hasattr(attn, "use_dsa") and attn.use_dsa。
2. 保持控制流不变：仅更改条件表达式，函数的分支结构（区分 *extend* 模式与非 *extend* 模式）和返回类型（DSA_NPU, MHA_NPU, MLA_NPU）均未变化。
3. 无测试配套：PR 未添加对应测试，但该变更仅影响调度分支选择，不涉及数值计算，因此回归风险较低。

关键文件：

- python/sglang/srt/models/deepseek_common/attention_backend_handler.py（模块 调度器；类别 source；类型 data-contract；符号 handle_attention_ascend）：核心变更文件，修改了 Ascend 注意力后端调度函数 handle_attention_ascend 中的条件判断逻辑。

关键符号：handle_attention_ascend

关键源码片段

python/sglang/srt/models/deepseek_common/attention_backend_handler.py

核心变更文件，修改了 Ascend 注意力后端调度函数 handle_attention_ascend 中的条件判断逻辑。

```
# python/sglang/srt/models/deepseek_common/attention_backend_handler.py
# 变更后：使用 attn.use_dsa 替代 hasattr(attn, "indexer")
```

```
def handle_attention_ascend(attn, forward_batch):
    if (
        forward_batch.forward_mode.is_extend()
        and not forward_batch.forward_mode.is_target_verify()
        and not forward_batch.forward_mode.is_draft_extend_v2()
    ):
        # Extend 模式：若属性 use_dsa 存在且为真，则使用 DSA_NPU 后端
        if hasattr(attn, "use_dsa") and attn.use_dsa:
            return AttnForwardMethod.DSA_NPU
        else:
            return AttnForwardMethod.MHA_NPU
    else:
        # 非 extend 模式（如 decode）：同理判断
        if hasattr(attn, "use_dsa") and attn.use_dsa:
            return AttnForwardMethod.DSA_NPU
        else:
            return AttnForwardMethod.MLA_NPU
```

评论区精华

Review 中 gemini-code-assist[bot] 提出简化建议：使用 getattr(attn, "use_dsa", False) 替代 hasattr + 属性访问，并移除冗余的 else 块，以使控制流更简洁。但该建议未被采纳，当前实现保留了显式的 hasattr 检查以保持防御性编程风格。无其他争议。

- 简化条件判断的建议 (style): 未被采纳，当前实现保留了显式 hasattr 检查的防御性风格。

风险与影响

- 风险：风险极低。变更仅修改调度条件，不涉及任何数值计算或内核调用。若 attn.use_dsa 在某些情况下未正确设置，可能导致调度偏离预期（例如将本应使用 DSA 的注意力模块误判为非 DSA），但该属性来源于模型配置，与原有逻辑相比仅是更直接地表达意图。不会引起精度下降、性能退化或系统崩溃。
- 影响：影响范围仅限于 Ascend NPU 设备上运行 DeepSeek GLM DSA 模型时的注意力后端选择。对非 NPU 平台（CUDA/ROCm）无影响，对 NPU 上非 DSA 模型无影响。用户无需更改配置或代码，预期行为一致但更可靠。
- 风险标记：缺少测试覆盖，条件判断变更

关联脉络

- PR #27798 [AMD] Add transpose_scale arg for o_proj to fix GLM accuracy issue: 同为 DeepSeek GLM 模型在非 CUDA 平台（AMD）上的修复，涉及同一文件 forward_mla.py 的注意力相关逻辑。