PR #28046 完整报告

sgl-project/sglang

[Intel GPU] DeepSeek V4 9/N: use sgl-kernel implementation of hadamard_transform on XPU

合并时间：2026-07-14 09:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28046

执行摘要

- 一句话：XPU 通过 sgl-kernel 支持 Hadamard 变换
- 推荐动作：作为 DeepSeek V4 XPU 支持系列的一部分，本 PR 实现简洁，风险低。若需在 XPU 环境运行 DeepSeek-V4，建议合入。注意测试文件未包含在最终变更中（移至 kernel 侧）。

功能与动机

hadamard_transform is CUDA/HIP-only, so importing it from rotate_activation() fails on Intel XPU with ImportError when the NSA indexer path runs (*e.g.* DeepSeek-V4 compressed attention).

实现拆解

1. 导入 is_xpu：在 python/sglang/srt/layers/attention/dsa/dsa_indexer.py 中，从 sglang.srt.utils 导入 is_xpu。
2. 定义 _is_xpu：在全局作用域中调用 is_xpu() 并赋值给 _is_xpu 变量。
3. 修改 rotate_activation 的条件分支：先判断 _is_hip（使用 fast_hadamard_transform），然后判断 _is_xpu（从 sgl_kernel 导入 hadamard_transform），最后对于其他平台（CUDA/SM103/NPU）回退到 sglang.jit_kernel.hadamard。

关键文件：

- python/sglang/srt/layers/attention/dsa/dsa_indexer.py（模块 注意力层；类别 source；类型 dependency-wiring；符号 rotate_activation）：核心变更文件，添加了 _is_xpu 检测和基于 sgl_kernel 的 Hadamard 变换导入路径，使 rotate_activation 能在 XPU 上运行。

关键符号：rotate_activation

关键源码片段

python/sglang/srt/layers/attention/dsa/dsa_indexer.py

核心变更文件，添加了 _is_xpu 检测和基于 sgl_kernel 的 Hadamard 变换导入路径，使 rotate_activation 能在 XPU 上运行。

```
# python/sglang/srt/layers/attention/dsa/dsa_indexer.py
```

```
# 新增：导入 is_xpu 检测函数
from sglang.srt.utils import (
    is_cuda,
    is_hip,
    is_npu,
    is_xpu, # 新增加
)

# 新增：定义全局 _is_xpu 标志
_is_xpu = is_xpu()

def rotate_activation(x: torch.Tensor) -> torch.Tensor:
    # 原注释保留：# from sgl_kernel import hadamard_transform
    if _is_hip:
        from fast_hadamard_transform import hadamard_transform
    elif _is_xpu: # 新增 XPU 分支
        from sgl_kernel import hadamard_transform # 使用 sgl_kernel 的融合实现
    else:
        from sglang.jit_kernel.hadamard import hadamard_transform

    hidden_size = x.size(-1)
    # 确保 hidden_size 是 2 的幂
    assert (
        hidden_size & (hidden_size - 1)
    ) == 0, "Hidden size must be a power of 2 for Hadamard transform."
    return hadamard_transform(x, scale=hidden_size ** -0.5)
```

评论区精华

审阅者 jianan-gu 建议将测试文件 test/manual/test_hadamard.py 移动到 kernel 侧，作者已采纳。此外，gemini-code-assist[bot] 建议优化实现，但作者尝试后表示“tried and failed”，因此未采纳。最终获得两位审阅者批准。

- 测试文件位置 (testing): 作者 polisettyvarma 同意并移除了测试文件，表示 'done'。
- 实现优化建议 (performance): 作者尝试后表示 'tried and failed'，未采纳此优化。最终实现未包含该函数，转而使用 sgl_kernel 的融合内核。

风险与影响

- 风险：变更范围极小（仅 4 行新增），仅在 _is_xpu 为 True 时才改变导入路径，且 sgl_kernel 中的 hadamard_transform 在合并前已通过测试，回归风险低。此方案复用了已有的 sgl_kernel 内核，而非引入纯 PyTorch 回退实现。
- 影响：仅影响 Intel XPU 平台上的 DeepSeek-V4 模型，解锁了 NSA indexer 路径在 XPU 上的正常运行，无其他影响。
- 风险标记：平台特定代码，缺少测试覆盖（测试未包含在 PR 中）

关联脉络

- PR #28428 [Intel GPU] DeepSeek V4 12/N: use sgl-kernel implementation of silu_and_mul_clamp to run on XPU: 同属 DeepSeek V4 XPU 支持系列，模式类似：通过 sgl-kernel 替代 CUDA/HIP 专用内核。
- PR #28059 [Intel GPU] DeepSeek V4 11/N: support fp8_paged_mqa_logits_triton from sgl-kernel to run on XPU: 同系列 PR，为 XPU 适配 sgl-kernel 中不同的算子和内核。