PR #24356 完整报告

sgl-project/sglang

[Intel GPU] Enable DeepSeek V3.2 inference on XPU

合并时间：2026-05-05 20:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24356

执行摘要

- 一句话：XPU 推理 DeepSeek V3.2
- 推荐动作：可快速合并，变更简洁直接。建议未来添加 XPU 特定测试，并在 forward_xpu 中补充明确的错误消息（如 reviewer 所提）。

功能与动机

使 DeepSeek V3.2 模型能够在 Intel XPU 上运行。PR body 说明使用 --attention-backend triton --page-size 64 运行缩减模型（4 层）来覆盖关键路径。

实现拆解

1. 新增 NSA indexer XPU 入口（nsa_indexer.py）：添加 forward_xpu 方法，直接委托给 forward_cuda 实现，复用 CUDA 路径中的 FP8 量化与缓存存储逻辑。
2. 调整 server_args 平台判断（server_args.py）：将 if not is_npu() 改为 if not is_npu() and not is_xpu()，使 XPU 跳过 context parallel 配置（如 dp_attention、ep_size 设置），因为 context parallel 仅适用于 CUDA/ROCm。

关键文件：

- python/sglang/srt/layers/attention/nsa/nsa_indexer.py（模块 注意力层；类别 source；类型 core-logic；符号 forward_xpu）：新增 forward_xpu 方法，委托给 forward_cuda，是 XPU 推理 NSA indexer 的核心入口。
- python/sglang/srt/server_args.py（模块 服务器配置；类别 source；类型 core-logic）：修正平台条件判断，使 XPU 跳过 context parallel 配置，避免不支持的路径被错误执行。

关键符号：forward_xpu

关键源码片段

python/sglang/srt/layers/attention/nsa/nsa_indexer.py

新增 forward_xpu 方法，委托给 forward_cuda，是 XPU 推理 NSA indexer 的核心入口。

```
# python/sglang/srt/layers/attention/nsa/nsa_indexer.py

def forward_xpu(
    self,
```

```
    x: torch.Tensor,
    q_lora: torch.Tensor,
    positions: torch.Tensor,
    forward_batch: ForwardBatch,
    layer_id: int,
    return_indices: bool = True,
) -> Optional[torch.Tensor]:
    # XPU 直接复用 CUDA 路径中的 FP8 量化与缓存存储逻辑
    # 注意：CUDA 路径中的 dual stream 等功能在 XPU 上由模型层配置为 None
    return self.forward_cuda(
        x, q_lora, positions, forward_batch, layer_id, return_indices
    )
```

python/sglang/srt/server_args.py

修正平台条件判断，使 XPU 跳过 context parallel 配置，避免不支持的路径被错误执行。

```
# python/sglang/srt/server_args.py
# 原代码：if not is_npu(): # CUDA or ROCm GPU
# 修改后：
if not is_npu() and not is_xpu(): # CUDA or ROCm GPU
    # 以下 context parallel 配置块仅对 CUDA/ROCm 生效
    ...
```

评论区精华

Reviewer @mingfeima 建议在 forward_xpu 中添加 assert 检查不支持的 feature（如 dual stream），并给出更具可读性的错误消息。PR author @polisettyvarma 回应称 alt_stream 在 XPU 的模型层面已为 None，讨论至此结束。

- forward_xpu 需要补充 unsupported feature 的检查 (design): @polisettyvarma 回复称 alt_stream 在 XPU 模型层已为 None，未进一步补充检查。reviewer 未坚持要求。

风险与影响

- 风险：低风险。forward_xpu 直接复用 forward_cuda，无新增逻辑，仅新增一条调用链；server_args 的改动仅排除 XPU 进入 context parallel 配置块，不影响其他平台。未添加测试，但缩减模型运行已覆盖核心路径。
- 影响：影响范围小：仅影响 Intel XPU 设备上加载 DeepSeek V3.2 模型时的初始化流程和注意力索引器计算。对其他平台无影响。
- 风险标记：缺少测试覆盖，缺少 XPU 特定错误检查

关联脉络

- 暂无明显关联 PR