PR #23285 完整报告

sgl-project/sglang

[Flashinfer] Integrate flashinfer router gemm for sm103

合并时间：2026-04-28 11:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23285

执行摘要

- 一句话：Flashinfer router gemm 支持 sm103
- 推荐动作：值得合并，改动小而明确。建议关注后续 flashinfer 版本更新，确保兼容性。

功能与动机

PR 动机是在 Blackwell Ultra（sm103）上利用 flashinfer 的优化 router gemm 提升路由 GEMM 性能。依赖 flashinfer >=0.6.8 及 flashinfer PR#2991。

实现拆解

1. 修改路由 GEMM 的条件判断：在 python/sglang/srt/models/deepseek_v2.py 的 DeepseekV2Gate.forward 方法中，将 _device_sm == 100 改为 _device_sm in [100, 103]，使得 sm103 设备也能进入 flashinfer router gemm 分支（调用 flashinfer_dsv3_router_gemm）。
2. 输出精度保证：该路径输出为 float32 类型，与原有 sm100 行为一致。
3. 无需额外配置：不涉及配置文件、命令行参数或 API 变更。

关键文件：

- python/sglang/srt/models/deepseek_v2.py（模块 模型层；类别 source；类型 core-logic）：核心变更文件，修改路由 GEMM 条件以支持 sm103。

关键符号：DeepseekV2Gate.forward

关键源码片段

python/sglang/srt/models/deepseek_v2.py

核心变更文件，修改路由 GEMM 条件以支持 sm103。

```
# python/sglang/srt/models/deepseek_v2.py
# 路由 GEMM 前向函数中，选择算子时的条件判断
if (
    _is_cuda
    and hidden_states.shape[0] <= 16
    and hidden_states.shape[1] == 7168
    and (self.weight.shape[0] == 256 or self.weight.shape[0] == 384)
    and _device_sm >= 90
```

```
):
    # 关键变更：原本只支持 sm100，现在新增 sm103（Blackwell Ultra）
    if _device_sm in [100, 103] and self.weight.shape[0] == 256:
        # router gemm output float32
        logits = torch.empty(
            hidden_states.shape[0],
            self.weight.shape[0],
            device=hidden_states.device,
            dtype=torch.float32,
        )
        flashinfer_dsv3_router_gemm(logits, hidden_states, self.weight)
    else:
        logits = dsv3_router_gemm(
            hidden_states, self.weight, out_dtype=torch.float32
        )
```

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：风险极低。变更仅一行条件判断，且测试覆盖了多种 batch size 的正确性，精度与原有实现一致。未覆盖 sm103 下的性能退化场景，但 flashinfer 侧已提供兼容性保证。
- 影响：影响范围小：仅影响 sm103 设备上 DeepSeek-V3/V2 路由 GEMM 的算子选择。用户无需手动配置，flashinfer 版本满足要求即可自动启用。性能提升依赖于 flashinfer 算子的优化效果。
- 风险标记：暂无

关联脉络

- PR #23883 Enable DeepGemm warmup in DeepSeek-V4 cookbook: 同为 DeepSeek 系列模型的路由 GEMM 优化相关。