PR #31961 完整报告

sgl-project/sglang

Change the FP8 per-tensor GEMM backend on SM120 to cuBLAS

合并时间：2026-07-22 05:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31961

执行摘要

- 一句话：SM120 FP8 GEMM 后端切换至 cuBLAS
- 推荐动作：值得合入，变更简单清晰，性能收益明确，同时降低了维护成本。建议关注未来 Torch 升级后 cuBLAS 对 SM120 的原生支持。

功能与动机

PR 描述指出，虽然 CUTLASS 可能带来额外 5-10% 的性能提升，但“不值得维护自定义内核和代码”；切换至 cuBLAS 可减少维护成本。同时，使用 cuBLAS 在 SM120 上已观察到约 20% 的 TPS 提升（55 → 65.59 TPS）。

实现拆解

1. 扩展 FlashInfer BMM 支持范围：在 modelopt_quant.pyModelOptFp8LinearMethod.__init__ 中，将 enable_flashinfer_bmm 的条件从 is_sm100_supported() and is_flashinfer_available() 修改为 (is_sm100_supported() or is_sm120_supported()) and is_flashinfer_available()。此变更让 SM120（Blackwell）GPU 也能使用基于 FlashInfer bmm_fp8 的 cuBLAS 后端。
2. 更新文档注释：在 fp8_utils.py 中 apply_fp8_linear_bmm_flashinfer 函数的 docstring 从 (SM10X only) 更新为 (SM100/SM120 Blackwell)，以准确反映支持范围。
3. 移除自定义 CUTLASS 内核调用：SM120 不再使用 sgl-kernel 中的 CUTLASS FP8 GEMM，转而调用 FlashInfer 封装的 cuBLAS。此举降低了代码维护成本，但可能在未来 Torch 升级后获得原生 cuBLAS 性能增益。

关键文件：

- python/sglang/srt/layers/quantization/modelopt_quant.py（模块 量化层；类别 source；类型 data-contract；符号 ModelOptFp8LinearMethod.init）：核心变更文件：修改 enable_flashinfer_bmm 条件，将 SM120 纳入 FlashInfer cuBLAS 路径。
- python/sglang/srt/layers/quantization/fp8_utils.py（模块 量化层；类别 source；类型 core-logic；符号 apply_fp8_linear_bmm_flashinfer）：更新文档说明以反映 SM120 支持。

关键符号：ModelOptFp8LinearMethod.init, apply_fp8_linear_bmm_flashinfer

关键源码片段

python/sglang/srt/layers/quantization/modelopt_quant.py

核心变更文件：修改 enable_flashinfer_bmm 条件，将 SM120 纳入 FlashInfer cuBLAS 路径。

```
# File: python/sglang/srt/layers/quantization/modelopt_quant.py
# 在 __init__ 中修改 enable_flashinfer_bmm 条件
class ModelOptFp8LinearMethod(LinearMethodBase):
    def __init__(self, quant_config: ModelOptFp8Config):
        super().__init__()
        self.quant_config = quant_config
        self.cutlass_fp8_supported = cutlass_fp8_supported()
        # 原条件：is_sm100_supported() and is_flashinfer_available()
        # 改为：同时支持 SM100 和 SM120 (Blackwell)
        self.enable_flashinfer_bmm = (
            is_sm100_supported() or is_sm120_supported()
        ) and is_flashinfer_available()
```

python/sglang/srt/layers/quantization/fp8_utils.py

更新文档说明以反映 SM120 支持。

```
# File: python/sglang/srt/layers/quantization/fp8_utils.py
# 更新 docstring 以反映 SM120 支持
def apply_fp8_linear_bmm_flashinfer(
    input: torch.Tensor,
    weight: torch.Tensor,
    weight_scale: torch.Tensor,
    input_scale: torch.Tensor,
    bias: Optional[torch.Tensor] = None,
) -> torch.Tensor:
    # docstring 从 "(SM10X only)" 更新为 "(SM100/SM120 Blackwell)"
    """Per-tensor static fp8 linear via flashinfer bmm_fp8 (SM100/SM120 Blackwell)."""
    output_shape = [*input.shape[:-1], weight.shape[1]]
    input_2d = input.view(-1, input.shape[-1])
    qinput, x_scale = static_quant_fp8(input_2d, input_scale, repeat_scale=False)
    output = flashinfer_bmm_fp8(qinput, weight, x_scale, weight_scale, input.dtype)
    if bias is not None:
        output = output + bias
    return output.view(*output_shape)
```

评论区精华

该 PR 没有 review 评论，仅有一条 CI 重跑指令和结果。讨论仅存在于 PR body 中，开发者阐述了切换理由：维护成本 vs 微小性能提升的权衡。

- 暂无高价值评论线程

风险与影响

- 风险：

  - 回归风险：低。该 PR 仅在条件判断中增加 GPU 架构支持，不改变已有逻辑。但 SM120 上的 cuBLAS 行为可能与 CUTLASS 存在数值差异，需确认精度。
  - 性能风险：低。根据 PR 数据，cuBLAS 在 SM120 上已带来约 20% 提升，且未来 Torch 升级后可能进一步优化。
  - 兼容性风险：无。条件判断确保仅 SM120 启用新路径，其他架构不受影响。
- 影响：
  - 用户影响：SM120（Blackwell）GPU 用户将自动获得更高的 FP8 GEMM 推理性能（约 20% TPS 提升），且无需手动配置。
  - 系统影响：减少了自定义 CUTLASS 内核的维护负担，统一使用 cuBLAS 可提高代码可维护性。
  - 团队影响：轻微，代码量仅 4 行变更，易于审查。
  - 风险标记：仅 4 行变更，风险低，需确认 SM120 cuBLAS 精度

关联脉络

- 暂无明显关联 PR