

PR #30448 完整报告

sgl-project/sglang

Refactor FP4 quantization and remove deprecated JIT kernels

合并时间: 2026-07-14 09:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30448>

执行摘要

本 PR 基于 RFC #28663 清理了 SGLang 中废弃的 NVFP4 JIT 内核及 CUTLASS MoE 后端。删除约 5.7k 行代码，将所有 Blackwell 上的 FP4 操作统一迁移到 FlashInfer 后端。同步删除了相关测试、基准和文档。这是内核层的重要清理，降低了维护成本，但要求用户依赖 FlashInfer 版本。

功能与动机

关联 Issue #28663 指出，这些 NVFP4 内核最初从 TRTLLM 移植，但 FlashInfer 已提供更完善的实现，且不再提供特殊用途。PR body 明确要求删除 `--moe-runner-backend cutlass`、密集 FP4 GEMM 和 FP4 量化 (in sgl-kernel)。作者在评论中展示的性能对比证实旧内核性能已无优势。

实现拆解

1. 删除核心 JIT 内核模块: `python/sglang/jit_kernel/nvfp4.py` 被完全移除，包括量化、专家量化、缩放 MM、块状 MoE 共 4 个 JIT 模块及 `prewarm_nvfp4_jit_modules` 预热函数。
2. 清理 MoE 后端: `python/sglang/srt/layers/moe/cutlass_moe.py` 中删除 `cutlass_moe_fp4` 函数; `server_args.py` 中移除 `cutlass` 选项。
3. 更新量化路径: `modelopt_quant.py` 和 `multimodal_gen` 中的对应文件被修改，硬编码 FP4 GEMM 后端为 FlashInfer。
4. 删除测试与基准: 移除了 7 个测试文件和 3 个 benchmark 文件，例如 `test_nvfp4_quant.py`、`test_nvfp4_gemm.py`、`bench_fp4_quant.py` 等。
5. 清理文档: `docs_new/docs/quantization.mdx` 和 `server_arguments.mdx` 删除废弃参数说明。

`python/sglang/jit_kernel/nvfp4.py`

此文件是整个清理的核心，包含了所有 NVFP4 JIT 内核模块、预热函数和 CUDA 标志。完全删除该文件 (636 行) 消除了复杂度。

```
@torch.compiler.disable
def prewarm_nvfp4_jit_modules(
    *,
    include_expert_quant: bool = False,
    include_blockwise_moe: bool = False
) -> None:
```

```
"""在 torch.compile 追踪模型之前实例化 NVFP4 JIT 模块。"""
_jit_nvfp4_quant_module() # 量化 JIT 模块
_jit_nvfp4_scaled_mm_module() # 缩放 GEMM JIT 模块
if include_expert_quant:
    _jit_nvfp4_expert_quant_module() # 专家量化 (可选)
if include_blockwise_moe:
    _jit_nvfp4_blockwise_moe_module() # 块状 MoE (可选)
```

评论区精华

- BBuf 要求更多 benchmark 和 accuracy 结果。作者提供了 SM100 上的性能对比图，证明 FlashInfer 更优。
- BBuf 询问删除 cutlass 选项是否影响文档 / 脚本。作者确认所有引用已清理。
- BBuf 指出 compressed_tensors NVFP4 切换后 apply_router_weight_on_input=True 会断言。作者添加断言并说明该模式实际未使用。

风险与影响

风险：必须安装兼容的 FlashInfer 版本；旧脚本使用 `--moe-runner-backend cutlass` 会失败；compressed_tensors 路径的早期断言可能不灵活；大量测试删除降低回归覆盖。影响：所有 Blackwell 用户将自动使用 FlashInfer；代码库瘦身；团队需关注性能回归。

关联脉络

与 Issue #28663 直接对应。同系列清理 PR #30438（删除 FP8 blockwise）风格一致。近期 NVFP4 修复 PR #31001 表明 FlashInfer 集成仍需关注兼容性。