

PR #30280 完整报告

sgl-project/sglang

Delete sgl-kernel AOT router GEMM and fused A GEMM

合并时间: 2026-07-22 08:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30280>

执行摘要

- 一句话: 删除 sgl-kernel AOT GEMM 内核及配套代码
- 推荐动作: 该 PR 是典型的技术债务清理, 设计清晰, 风险可控, 建议合并。值得关注的决策是: 将复杂 kernel 从 AOT 移至 JIT 统一框架, 可以简化跨平台支持并提升性能调优效率。

功能与动机

PR body 仅说明 'Since we migrated already'。这些 AOT kernel 的对应功能已由 JIT kernel 覆盖, 删除旧代码可以减少维护成本、加快编译速度, 并统一内核实现路径。

实现拆解

按照以下步骤清理:

1. 删除 CUDA kernel 源文件: 移除 sgl-kernel/csrc/gemm/dsv3_router_gemm_bf16_out.cu、dsv3_router_gemm_float_out.cu、dsv3_router_gemm_entry.cu 和 dsv3_fused_a_gemm.cu, 这些是 AOT 编译的 CUDA kernel。
2. 删除 Python 绑定与 API: 从 sgl_kernel/gemm.py 中移除 dsv3_fused_a_gemm 函数, 从 __init__.py 和 CMakeLists.txt 中移除相关注册, 并删除 MUSA 扩展中的对应绑定。
3. 删除测试与 benchmark: 移除 test_dsv3_fused_a_gemm.py 以及 benchmark_deepgemm_dsv3_router_gemm_blackwell.py、bench_dsv3_fused_a_gemm.py 等 benchmark 文件。
4. 更新 JIT kernel 入口: 修改 fused_a_gemm.py 和 dsv3_router_gemm.py, 确保它们直接指向 JIT 实现而非 AOT; 调整已迁移的 benchmark 脚本的导入路径。
5. 处理冲突与 MUSA: 在合并过程中解决与 main 分支的多次冲突, 并确认 MUSA 平台不需要这些 kernel, 也删除了 musa 绑定。

关键文件:

- sgl-kernel/csrc/gemm/dsv3_fused_a_gemm.cu (模块 CUDA 内核; 类别 other; 类型 deletion): AOT 核心 kernel 实现 (677 行), 包含 dsv3_fused_a_gemm 的 CUDA kernel 及模板实例化。删除后该功能完全由 JIT kernel 提供。
- sgl-kernel/csrc/gemm/dsv3_router_gemm_bf16_out.cu (模块 CUDA 内核; 类别 other; 类型 deletion): AOT 路由器 GEMM bfloat16 输出 kernel (284 行), 深度绑定 DeepSeek V3 路由逻辑。

- `sgl-kernel/python/sgl_kernel/gemm.py` (模块 Python 绑定; 类别 source; 类型 core-logic; 符号 `dsv3_fused_a_gemm`) : Python API 入口: 移除了 `dsv3_fused_a_gemm` 的绑定函数。删除后用户代码若直接调用该函数将报错, 但官方调用已迁移至 JIT kernel。
- `benchmark/kernels/deepseek/benchmark_deepgemm_dsv3_router_gemm_blackwell.py` (模块 基准测试; 类别 source; 类型 deletion; 符号 `create_benchmark_configs`, `dsv3_router_gemm_flashinfer`, `dsv3_router_gemm_sgl`, `check_accuracy`) : Blackwell 平台的路由 GEMM benchmark 文件, 包含与 flashinfer 的对比测试。删除后不再需要维护。
- `python/sglang/jit_kernel/fused_a_gemm.py` (模块 JIT 内核; 类别 source; 类型 dependency-wiring) : JIT kernel 入口文件, 修改以移除对 AOT 的引用, 确保 JIT 路径为唯一路径。

关键符号: `dsv3_fused_a_gemm`, `dsv3_router_gemm`, `create_benchmark_configs`, `dsv3_router_gemm_sgl`, `dsv3_router_gemm_flashinfer`

关键源码片段

`sgl-kernel/python/sgl_kernel/gemm.py`

Python API 入口: 移除了 `dsv3_fused_a_gemm` 的绑定函数。删除后用户代码若直接调用该函数将报错, 但官方调用已迁移至 JIT kernel。

```
# sgl-kernel/python/sgl_kernel/gemm.py ( 修改后 )
# 原 dsv3_fused_a_gemm 函数已被移除, 由 jit_kernel.fused_a_gemm 替代

def fp8_scaled_mm(mat_a, mat_b, scales_a, scales_b, out_dtype, bias=None):
    return torch.ops.sgl_kernel.fp8_scaled_mm.default(
        mat_a, mat_b, scales_a, scales_b, out_dtype, bias
    )

# [ 已删除 ] def dsv3_fused_a_gemm(mat_a, mat_b, output=None):
# [ 已删除 ] ... # 直接调用 torch.ops.sgl_kernel.dsv3_fused_a_gemm.default

# 后续函数保持不变
```

评论区精华

唯一的技术讨论来自 Fridge003 的 review 评论:

询问: 'Can we delete the musa kernel? Since the jit kernel is only in cuda' (位于 [common_extension_musa.cc:100](#)) 答复: b8zhong 确认 'Yes. They said they didn't need it on Slack', 表明团队已确认 MUSA 平台也无需这些 AOT kernel。

- 是否删除 MUSA 平台的绑定 (question): 确认 MUSA 平台不需要这些 AOT kernel, 删除 musa 绑定是安全的。

风险与影响

- 风险:

1. 功能兼容性: 低风险。所有原 AOT kernel 功能已由 JIT kernel 覆盖, 且已有对应测试。
2. MUSA 平台: 低风险。作者已通过 Slack 确认 MUSA 团队不需要这些 kernel, 且 PR 同时删除了 musa 绑定。
3. 回归风险: 中等。由于 PR 涉及多次合并冲突解决, 可能存在细小的残留依赖。但 CI 已通过。
4. 打包体积: 正面影响, sgl-kernel 编译产物变小。- 影响: 对用户: 无直接功能影响。对开发团队: 减少维护负担, 加快编译速度, 促进内核实现统一。对系统: sgl-kernel 包不再包含 dsv3_router_gemm 和 dsv3_fused_a_gemm 的 AOT 实现。- 风险标记: MUSA 平台确认, JIT 功能覆盖验证

关联脉络

- PR #31202 Delete sgl-kernel AOT bmm_fp8, use flashinfer.bmm_fp8: 同为删除 sgl-kernel 中的 AOT kernel, 与本次 PR 同一清理方向。
- PR #30924 [JIT] Trait-driven per_token_group_quant: unify the quant kernel family (flat + masked): 提供了 JIT kernel 统一框架, 为本次迁移奠定基础。