

PR #31634 完整报告

sgl-project/sglang

[MUSA] Fix sglang-kernel build

合并时间: 2026-07-18 11:51

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/31634>

执行摘要

- 一句话: 修复 MUSA sglang-kernel 编译断裂
- 推荐动作: 值得快速合并的维护性修复, 确保了 MUSA 后端的代码与主线重构保持同步。对于维护者, 建议后续在重构涉及多后端代码时, 提前验证所有目标平台的构建。

功能与动机

PR #21531 移除了 QServe 和 FBGEMM FP8 量化路径, 并把多个 AOT kernel 迁移到 JIT (如移除 dsv3_router_gemm 算子), 但未同步更新 MUSA 的 sglang-kernel 扩展注册代码和头文件引用, 导致 MUSA 构建失败。

实现拆解

1. 移除废弃的算子注册: 在 sgl-kernel/csrc/common_extension_musa.cc 中删除了 dsv3_router_gemm 的 TORCH_LIBRARY 注册和实现绑定 (-2 行), 因为它已被迁移至 JIT 并转为 Triton 实现。
2. 修复头文件引用: 在 sgl-kernel/csrc/musa/top_k_top_p_sampling.mu 中, 将 #include <flashinfer/sampling.muh> 替换为 #include <flashinfer/sampling.cuh>, 以匹配升级后的 FlashInfer 版本。
3. 升级 torchada 依赖: 在 3 个 pyproject.toml 文件中将 torchada>=0.1.68 提升至 >=0.1.74, 以兼容新版本 SDK。
4. 清理测试用例: 从 MUSA 的 CI 工作流中删除已废弃的 test_dsv3_router_gemm.py 测试, 避免执行不存在的测试引发失败。

关键文件:

- sgl-kernel/csrc/common_extension_musa.cc (模块 算子注册; 类别 source; 类型 core-logic; 符号 dsv3_router_gemm): 核心修复: 移除已废弃的 dsv3_router_gemm 算子注册和实现绑定, 消除编译错误。
- sgl-kernel/csrc/musa/top_k_top_p_sampling.mu (模块 采样算子; 类别 other; 类型 dependency-wiring): 修复头文件引用: 将 flashinfer/sampling.muh 改为 .cuh, 与 FlashInfer 新版本兼容。
- sgl-kernel/pyproject_musa.toml (模块 构建配置; 类别 config; 类型 configuration): 升级 torchada 构建依赖至 >=0.1.74, 匹配新 SDK 版本。

- 3rdparty/amd/wheel/sglang/pyproject.toml (模块 AMD 封装; 类别 config; 类型 configuration) : 同步升级 AMD 封装中的 torchada 依赖至 $\geq 0.1.74$ 。
- python/pyproject_other.toml (模块 其他平台依赖; 类别 config; 类型 configuration) : 同步升级主项目 pyproject_other 中 MUSA 依赖 torchada 至 $\geq 0.1.74$ 。
- .github/workflows/nightly-test-musa.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 移除已废弃的测试用例 test_dsv3_router_gemm.py, 避免 CI 执行时因找不到测试文件而失败。
- .github/workflows/pr-test-musa.yml (模块 CI 配置; 类别 infra; 类型 infrastructure) : 同步移除 PR 测试流水线中的相同已废弃测试。

关键符号: 未识别

关键源码片段

sgl-kernel/csrc/common_extension_musa.cc

核心修复: 移除已废弃的 dsv3_router_gemm 算子注册和实现绑定, 消除编译错误。

```
/* sgl-kernel/csrc/common_extension_musa.cc */
// 从 sgl_kernel MUSA 扩展中移除已迁移至 JIT 的 dsv3_router_gemm 算子注册
TORCH_LIBRARY_EXPAND(sgl_kernel, m) {
  // ... 其他算子 ...

  m.def("dsv3_fused_a_gemm(Tensor! output, Tensor mat_a, Tensor mat_b) -> (");
  m.impl("dsv3_fused_a_gemm", torch::kMUSA, &dsv3_fused_a_gemm);

  // 以下两行被删除: dsv3_router_gemm 已被迁移到 Triton 路由, 不再需要 AOT 注册
  // m.def("dsv3_router_gemm(Tensor! output, Tensor mat_a, Tensor mat_b) -> (");
  // m.impl("dsv3_router_gemm", torch::kMUSA, &dsv3_router_gemm);

  /* MOE 相关算子 */
  m.def("moe_align_block_size(...) -> (");
  // ...
}
```

sgl-kernel/csrc/musa/top_k_top_p_sampling.mu

修复头文件引用: 将 flashinfer/sampling.mu 改为 .cuh, 与 FlashInfer 新版本兼容。

```
/* sgl-kernel/csrc/musa/top_k_top_p_sampling.mu */
#include <torch/all.h>
#include "torch_musa/csrc/aten/musa/UnpackRaw.mu"
// 修复: 原为 sampling.mu, FlashInfer 升级后改为 .cuh
#include <flashinfer/sampling.cuh>
#include <mutex>
#include "musa.h"
```

sgl-kernel/pyproject_musa.toml

升级 torchada 构建依赖至 $\geq 0.1.74$, 匹配新 SDK 版本。

```
# sgl-kernel/pyproject_musa.toml
[build-system]
requires = [
    "setuptools>=75.0",
    "scikit-build-core>=0.10",
    "torch",
    "torchada>=0.1.74", # 从 0.1.68 升级
    "wheel",
]
build-backend = "setuptools.build_meta"
```

评论区精华

该 PR 没有 review 评论，仅由 Fridge003 直接批准合并。PR body 明确指出了根因——PR #21531 的 JIT 迁移未同步更新 MUSA 后端。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更聚焦于清理已废弃的代码和修复依赖版本，不涉及运行时逻辑调整。但需确认 MUSA 后端在移除 dsv3_router_gemm 后功能正常（该算子已由 Triton 路由替代）。
- 影响：直接修复 MUSA 平台的 sglang-kernel 编译断裂，确保 MThreads GPU 用户能正常构建和运行。影响范围限于 MUSA 后端，不涉及其他平台。
- 风险标记：无新增测试覆盖，仅 MUSA 后端

关联脉络

- PR #31109 Remove QServe and FBGEMM FP8 quantization: 该 PR 移除了 dsv3_router_gemm 等算子并迁移到 JIT，导致本 PR 需要修复 MUSA 后端的编译断裂。