

PR #40380 完整报告

vllm-project/vllm

Require C++20 for compatibility with PyTorch

合并时间: 2026-05-09 13:04

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/40380>

执行摘要

- 一句话: C++ 标准从 17 升级到 20 以兼容 PyTorch
- 推荐动作: 建议精读该 PR, 尤其是评论中关于 PyTorch C++ 标准演进方向的讨论。它反映了下游项目如何应对上游依赖的 breaking change, 对于维护大型分布式系统具有参考价值。

功能与动机

PyTorch 头文件 (如 torch/all.h 和 ATen/ATen.h) 现在强制要求 C++20 (`__cplusplus < 202002L` 时会触发 `#error`), vLLM 若仍使用 C++17 则无法与新版 PyTorch 一起编译。PR body 明确说明此问题已导致 PyTorch 出现回退事件。

实现拆解

该 PR 通过修改 3 个 CMake 文件, 将 C++ 标准从 17 升级到 20:

1. 主 CMakeLists.txt: 设置 CMAKE_CXX_STANDARD 20, 并新增 CMAKE_CUDA_STANDARD 20 和 CMAKE_HIP_STANDARD 20 及其 _REQUIRED 版本, 确保 CUDA 和 HIP 代码也使用 C++20 标准。
2. cmake/cpu_extension.cmake: 将 CMAKE_CXX_STANDARD 从 17 改为 20, 保持与主配置一致。
3. cmake/external_projects/deepgemm.cmake: 移除硬编码的 `-std=c++17` 编译选项, 因为全局 CMake 标准设置已经生效, 避免冗余和潜在冲突。没有测试或文档配套变更。

关键文件:

- CMakeLists.txt (模块 构建系统; 类别 infra; 类型 core-logic): 主构建配置文件: 设置了全局 C++20 标准, 并新增 CUDA/HIP 标准的显式声明, 是最核心的变更文件。
- cmake/cpu_extension.cmake (模块 构建系统; 类别 infra; 类型 core-logic): CPU 扩展的构建配置, 与主配置同步升级 C++ 标准。
- cmake/external_projects/deepgemm.cmake (模块 构建系统; 类别 infra; 类型 core-logic): DeepGEMM 外部项目的编译选项, 移除了硬编码的 `-std=c++17`, 避免与全局标准冲突。

关键符号: 未识别

关键源码片段

CMakeLists.txt

主构建配置文件：设置了全局 C++20 标准，并新增 CUDA/HIP 标准的显式声明，是最核心的变更文件。

```
# vllm_extensions 项目的顶层 CMakeLists.txt
cmake_minimum_required(VERSION 3.26)

project(vllm_extensions LANGUAGES CXX)

# 关键变更：C++ 标准从 17 升级到 20，以兼容 PyTorch 头文件的强制要求
set(CMAKE_CXX_STANDARD 20)
set(CMAKE_CXX_STANDARD_REQUIRED ON)

# 为确保 CUDA 和 HIP 代码也使用 C++20，显式设置对应标准（review 建议后添加）
set(CMAKE_CUDA_STANDARD 20)
set(CMAKE_CUDA_STANDARD_REQUIRED ON)
set(CMAKE_HIP_STANDARD 20)
set(CMAKE_HIP_STANDARD_REQUIRED ON)

# 其余构建配置保持不变 ...
```

cmake/external_projects/deepgemm.cmake

DeepGEMM 外部项目的编译选项，移除了硬编码的 `-std=c++17`，避免与全局标准冲突。

```
# deepgemm.cmake 片段：移除冗余的 -std=c++17 硬编码
# 因为全局 CMAKE_CXX_STANDARD 20 已生效，无需重复指定
target_compile_options(_deep_gemm_C PRIVATE
    # 原来这里有一行：$<$<COMPILE_LANGUAGE:CXX>:-std=c++17>
    # 已按 review 建议移除，避免与 CMake 的标准管理冲突
    $<$<COMPILE_LANGUAGE:CXX>:-O3>
    $<$<COMPILE_LANGUAGE:CXX>:-Wno-psabi>
    $<$<COMPILE_LANGUAGE:CXX>:-Wno-deprecated-declarations>
)
```

评论区精华

关键讨论：

- ZJY0516 询问具体哪个 PyTorch 版本会强制 C++20。lakshayg 和 r-barnes 解释：PyTorch 已尝试推动 C++20，但因 vLLM 等下游项目失败而回退，预计很快会再次强制。
- gemini-code-assist[bot] 建议显式设置 CMAKE_CUDA_STANDARD 和 CMAKE_HIP_STANDARD（已被采纳），并建议移除 deepgemm.cmake 中的硬编码 `-std=c++17`（已在最终版本中移除，改为仅删除该行）。
- Harry-Chen 触发了 CI 构建，确认无问题后批准 PR。
- 作者 r-barnes 最终添加了 DCO 签名并完成单次提交。
- 是否需要显式设置 CUDA/HIP 标准 (correctness)：已采纳：在 CMakeLists.txt 中新增了 CMAKE_CUDA_STANDARD 20 和 CMAKE_HIP_STANDARD 20 及其 `_REQUIRED` 版本。

- 移除 deepgemm.cmake 中的硬编码 `-std=c++17 (style)`: 已采纳: 最终提交直接移除了该行硬编码 (`-std=c++17` → 无)。
- PyTorch 强制 C++20 的时间线 (question): 明确了迁移的必要性和紧迫性。

风险与影响

- 风险: 低风险: 该变更仅修改 CMake 编译标准设置, 不涉及业务逻辑。需注意:
 - 如果某些第三方依赖或旧编译器不支持 C++20, 可能导致构建失败。vLLM 已要求 `CMake>=3.26`, 通常能配合现代编译器。
 - CUDA/HIP 标准显式设置为 20 可能对旧版本 NVCC/ROCm 编译器造成问题, 但 PR 中已包含 `_REQUIRED` 标记, 若编译器不支持则会报错而非静默降级。
 - 影响: 影响范围: 所有使用 C++/CUDA/HIP 编译的 vLLM 组件, 包括核心推理内核、量化内核 (如 DeepGEMM) 等。对用户: 无直接功能变化, 但确保 vLLM 能与最新版 PyTorch (即将强制 C++20) 保持兼容。对开发团队: 统一了构建标准, 消除因标准不一致导致的潜在编译问题。
- 风险标记: 编译器兼容性, CUDA/HIP 标准要求

关联脉络

- PR #40378 Related issue: vLLM C++17 build fails with PyTorch requiring C++20: PR #40380 直接解决 issue #40378 中报告的构建失败问题。
- PR #43230 [Misc] downgrade nvidia-cutlass-dsl to 4.5.0: 同为构建系统相关的兼容性修复, 反映了与上游依赖保持一致的工程实践。