

PR #37937 完整报告

sgl-project/sglang

[Fix] Register triton.runtime.cache.triton_key in the MPS stub so torch.compile keeps working

合并时间: 2026-09-04 13:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37937>

执行摘要

- 一句话: 修复 Apple Silicon 上导入 sglang 后 torch.compile 全局失效问题。
- 推荐动作: 建议精读, 特别是了解平台 stub 注册机制与 Python import 系统交互的微妙之处, 以及如何为依赖 Triton 但无需实际执行的场景设计 mock 层。

功能与动机

在 macOS arm64 上, 由于 sglang 的 `install_platform_stubs()` 为 `triton.runtime.cache` 注册了元路径查找器但未预定义 `triton_key` 符号, 导致 Python import 机制创建空命名空间模块覆盖 mock。当 torch inductor 从该路径导入 `triton_key` 时, 获取到的是一个不可调用的模块对象, 从而引发 `TypeError`, 使得整个进程中所有 `torch.compile` 调用失败。

实现拆解

1. 核心逻辑修复: 在 `python/sglang/_platform_stubs.py` 的 `install_platform_stubs` 函数中, 新增 `_triton_key` 函数并显式注册 `triton.runtime.cache` 子模块。
2. 统一符号定义: 将原 `triton.compiler.compiler.triton_key` 从匿名 lambda 改为共享的 `_triton_key` 函数, 保持一致性。
3. 回归测试: 在 `test/registered/unit/platforms/test_mps_triton_stub.py` 中新增测试用例, 验证导入 sglang 后 torch.compile 仍能正常工作。

关键文件:

- `python/sglang/_platform_stubs.py` (模块 平台兼容层; 类别 source; 类型 core-logic; 符号 `_triton_key`, `install_platform_stubs`): 核心修复文件, 注册了缺失的 `triton.runtime.cache` 子模块和 `triton_key` 符号。
- `test/registered/unit/platforms/test_mps_triton_stub.py` (模块 平台测试; 类别 test; 类型 test-coverage; 符号 `test_torch_compile_works_after_sglang_installs_stub`): 新增回归测试, 验证 torch.compile 在导入 sglang 后仍能正常工作。

关键符号: `_triton_key`, `install_platform_stubs`

关键源码片段

`python/sglang/_platform_stubs.py`

核心修复文件, 注册了缺失的 `triton.runtime.cache` 子模块和 `triton_key` 符号。

新增的辅助函数，提供稳定的 triton_key 字符串

```
def _triton_key() -> str:  
    return "triton-stub"
```

在 install_platform_stubs 函数中，为 triton.runtime.cache 注册 mock

```
cache_mod = _make_mock("triton.runtime.cache")  
cache_mod.triton_key = _triton_key # 显式绑定可调用对象  
runtime.cache = cache_mod
```

同时修改 triton.compiler.compiler 中的 triton_key 保持一致

```
compiler_impl.triton_key = _triton_key # 替换原来的 lambda: "triton-stub"
```

评论区精华

无评论。该 PR 已合并，无进一步讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，变更仅限于 Apple Silicon 平台 stub 模块，不影响其他平台。但需注意 triton.__version__ 仍被设置为 "3.0.0"，可能误导依赖版本检查的代码（但本次修复未涉及）。
- 影响：影响范围：所有在 Apple Silicon 平台导入 sglang 的用户，修复了 torch.compile 全局失效问题，恢复了与非 sglang 代码的兼容性。
- 风险标记：平台兼容性

关联脉络

- 暂无明显关联 PR