PR #26591 完整报告

sgl-project/sglang

[AMD] Pin compressed-tensors<0.16.0 for srt_hip (fixes ROCm 7.2 nightly build)

合并时间：2026-05-29 11:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26591

执行摘要

- 一句话：AMD 依赖锁定修复 ROCm 7.2 构建
- 推荐动作：值得关注。虽然变更简单，但这是典型的依赖版本下界冲突导致上游破坏的案例，注释清楚说明了问题和临时锁定策略。AMD 平台维护者应关注后续 ROCm 基础镜像升级后移除该锁定。

功能与动机

修复 AMD ROCm 7.2 nightly Docker 构建失败。compressed-tensors 0.16.0 提升 torch 要求至 >=2.10.0，与 ROCm 7.2 的 torch 2.9.1 冲突，pip 静默替换为 CUDA torch 导致 fast-hadamard-transform 构建时因 bare_metal_version 未定义而 NameError。

实现拆解

在 python/pyproject_other.toml 的 srt_hip 依赖列表末尾添加一行 "compressed-tensors<0.16.0"，并附带详细注释说明原因和追踪上下文。

关键文件：

- python/pyproject_other.toml（模块 依赖配置；类别 config；类型 configuration）：唯一的变更文件，在 srt_hip 依赖分组中添加 compressed-tensors 版本锁定（<0.16.0）。

关键符号：未识别

关键源码片段

python/pyproject_other.toml

唯一的变更文件，在 srt_hip 依赖分组中添加 compressed-tensors 版本锁定（<0.16.0）。

```
# HIP (Heterogeneous-computing Interface for Portability) for AMD
# => base docker rocm/vllm-dev:20250114, not from public vllm whl
srt_hip = [
  "sglang[runtime_common]",
  "torch",
  "petit_kernel==0.0.2",
  "wave-lang==3.8.2",
  # HOTFIX (2026-05-28): compressed-tensors 0.16.0 added `torch>=2.10.0`,
  # which forces pip off the ROCm wheel (torch 2.9.1+rocm7.2) and silently
```

```
    # swaps it for the PyPI default `torch 2.12.0+cu130`, breaking the
    # downstream HIP build of fast-hadamard-transform with a NameError on
    # `bare_metal_version`. Pin below 0.16.0 until the ROCm base ships
    # torch>=2.10.
    "compressed-tensors<0.16.0",
]
```

评论区精华

无讨论。PR 只有正常审核流程，reviewer bingxche 直接批准，未产生评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅限制在 srt_hip 依赖分组的 compressed-tensors 版本上限，不影响其他平台或功能。compressed-tensors 的低版本功能完全兼容，仅暂时延迟 AMD 平台上的升级。未来 ROCm 基础镜像升级 torch 到 >=2.10.0 后可移除该锁定。
- 影响：影响范围很小。仅影响使用 srt_hip 分组安装依赖的 AMD ROCm 用户，修复了 nightly 构建失败问题。其他平台（NVIDIA, NPU, XPU, CPU 等）不受影响。
- 风险标记：依赖版本冲突，临时锁定

关联脉络

- 暂无明显关联 PR