

PR #27537 完整报告

sgl-project/sglang

[MUSA] bump torchada version to 0.1.59 and workaround PCG limitation.

合并时间: 2026-06-08 23:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27537>

执行摘要

- 一句话: 升级 torchada 到 0.1.59, 修复 MUSA PCG 限制
- 推荐动作: 值得快速合并, 以解决 MUSA 平台 PCG 的已知问题。无需深入审查代码逻辑, 但建议关注 torchada 0.1.59 的 release notes 以确认 LRU 机制的行为。

功能与动机

MUSA 驱动对每个进程有约 2048 个 `musaGraphExec_t` 对象的限制, PCG 路径会实例化大量 graph executable, 导致 graph capture 失败。torchada 0.1.59 通过透明 LRU 旋转机制, 保持 graph 模板活跃、限制 executable 数量, 并在 replay 前惰性重实例化被驱逐的 executable, 从而规避该限制。

实现拆解

1. 修改 sgl-kernel 构建依赖: 在 sgl-kernel/pyproject_musa.toml 中将 torchada 版本从 `>=0.1.57` 更新为 `>=0.1.59`, 确保 MUSA 内核模块构建时使用新版本。
2. 修改 srt_musa 运行时依赖: 在 python/pyproject_other.toml 和 3rdparty/amd/wheel/sglang/pyproject.toml 中的 srt_musa 依赖列表里同步更新版本号, 确保 SGLang 运行时安装正确的 torchada 版本。
3. 无其他代码或逻辑变更: 全部变更仅为版本号字符串替换, 不涉及任何源码、测试或配置调整。

关键文件:

- sgl-kernel/pyproject_musa.toml (模块 内核构建; 类别 config; 类型 configuration): MUSA 内核模块的构建系统文件, 定义了 torchada 构建依赖。
- python/pyproject_other.toml (模块 运行时依赖; 类别 config; 类型 configuration): SGLang 运行时 MUSA 平台依赖配置。
- 3rdparty/amd/wheel/sglang/pyproject.toml (模块 第三方构建; 类别 config; 类型 configuration): AMD 平台也维护了 MUSA 依赖 (可能用于交叉构建或统一管理), 需同步更新。

关键符号: 未识别

关键源码片段

sgl-kernel/pyproject_musa.toml

MUSA 内核模块的构建系统文件，定义了 torchada 构建依赖。

```
# sgl-kernel/pyproject_musa.toml
# 构建时依赖：MUSA 内核模块需要 torchada 0.1.59 才能正确管理 CUDA Graph executable
[build-system]
requires = [
    "setuptools>=75.0",
    "scikit-build-core>=0.10",
    "torch",
    "torchada>=0.1.59", # 从 >=0.1.57 升级
    "wheel",
]
build-backend = "setuptools.build_meta"
```

python/pyproject_other.toml

SGLang 运行时 MUSA 平台依赖配置。

```
# python/pyproject_other.toml
# srt_musa 运行时依赖：确保 MUSA 平台用户安装 torchada 0.1.59
srt_musa = [
    "sglang[runtime_common]",
    "torch",
    "torch_musa",
    "torchada>=0.1.59", # 从 >=0.1.57 升级
    "mthreads-ml-py",
    "mate>=0.2.0",
    "deep-gemm>=0.1.3",
    "flash_attn_3>=0.1.4",
    "numpy<2.0",
]
```

3rdparty/amd/wheel/sglang/pyproject.toml

AMD 平台也维护了 MUSA 依赖（可能用于交叉构建或统一管理），需同步更新。

```
# 3rdparty/amd/wheel/sglang/pyproject.toml
# AMD 平台下的 MUSA 运行时依赖，同步更新 torchada 版本
srt_musa = [
    "sglang[runtime_common]",
    "torch",
    "torch_musa",
    "torchada>=0.1.59", # 从 >=0.1.57 升级
    "mthreads-ml-py",
    "mate>=0.2.0",
    "deep-gemm>=0.1.3",
    "flash_attn_3>=0.1.4",
    "numpy<2.0",
]
```

评论区精华

该 PR 无实质性 review 讨论。[gemini-code-assist\[bot\]](#) 自动生成了无反馈的评论，[yeahdongcn](#) 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅限于三个配置文件中的版本号字符串，且为向前兼容的补丁升级。若新版本引入回归，MUSA 平台上的 CUDA Graph 功能可能受影响，但可通过降级版本快速恢复。
- 影响：仅影响 MUSA 平台用户。修复了 PCG 路径下 deep 模型因 executable 数量超限导致的 graph capture 失败问题，提升 MUSA 平台的稳定性和可用性。不影响其他平台。
- 风险标记：暂无

关联脉络

- PR #27242 [MUSA][23/N] CI: Fix torchada preflight lock cleanup and add LLM server smoke test: 同为 MUSA 平台和 torchada 相关，该 PR 修复了 torchada 的锁清理问题，而本 PR 升级版本解决 PCG 限制，两者共同提升 MUSA 稳定性。