PR #24592 完整报告

sgl-project/sglang

[MUSA] Bump torchada to 0.1.54

合并时间：2026-05-08 02:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24592

执行摘要

- 一句话：MUSA torchada 版本从 0.1.53 升级到 0.1.54
- 推荐动作：建议合并，确保 MUSA 平台能与上游 CUDA 功能保持同步。合并前可验证 torchada 0.1.54 版本已正确发布且安装无问题。

功能与动机

PR#24190 引入了多个新 CUDA API（例如 _cuda_beginAllocateCurrentThreadToPool），但 MUSA 的 torch_musa 中缺少这些接口。因此，在 torchada 仓库（PR#61）中同步实现了这些接口，并发布 0.1.54 版本。此 PR 将 torchada 的依赖版本提升至 >=0.1.54，以确保 CI 和实际运行中能正确获取新实现。

实现拆解

1. 修改 3rdparty/amd/wheel/sglang/pyproject.toml：将 srt_musa 依赖项中的 torchada>=0.1.53 改为 torchada>=0.1.54。
2. 修改 python/pyproject_other.toml：同样将 srt_musa 依赖项中的 torchada>=0.1.53 改为 torchada>=0.1.54。
3. 修改 sgl-kernel/pyproject_musa.toml：将构建系统依赖中的 torchada>=0.1.53 改为 torchada>=0.1.54。

所有变更均为版本号更新，未涉及任何源码逻辑修改。

关键文件：

- 3rdparty/amd/wheel/sglang/pyproject.toml（模块 AMD 构建；类别 config；类型 configuration）：AMD 平台 wheel 的 pyproject.toml，包含 MUSA 依赖列表，需要更新 torchada 版本。
- python/pyproject_other.toml（模块 主构建；类别 config；类型 configuration）：SGLang 主包的 pyproject 定义，同样需要更新 torchada 版本以保持一致。
- sgl-kernel/pyproject_musa.toml（模块 内核构建；类别 config；类型 configuration）：MUSA kernel 的 pyproject，构建系统依赖中需要 torchada 用于编译。

关键符号：未识别

关键源码片段

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

AMD 平台 wheel 的 pyproject.toml，包含 MUSA 依赖列表，需要更新 torchada 版本。

```
# MUSA 平台的运行时依赖，位于 3rdparty/amd/wheel/sglang/pyproject.toml
srt_musa = [
  "sglang[runtime_common]",
  "torch",
  "torch_musa",
  "torchada>=0.1.54", # 从 0.1.53 升级，以获取新版 CUDA API 支持
  "mthreads-ml-py",
  "mate>=0.2.0",
  "deep-gemm>=0.1.3",
  "flash_attn_3>=0.1.4",
  "numpy<2.0",
]
```

python/pyproject_other.toml

SGLang 主包的 pyproject 定义，同样需要更新 torchada 版本以保持一致。

```
# MUSA 平台的运行时依赖，位于 python/pyproject_other.toml
srt_musa = [
  "sglang[runtime_common]",
  "torch",
  "torch_musa",
  "torchada>=0.1.54", # 统一升级到 0.1.54
  "mthreads-ml-py",
  "mate>=0.2.0",
  "deep-gemm>=0.1.3",
  "flash_attn_3>=0.1.4",
  "numpy<2.0",
]
```

sgl-kernel/pyproject_musa.toml

MUSA kernel 的 pyproject，构建系统依赖中需要 torchada 用于编译。

```
# sgl-kernel MUSA 构建所需的工具依赖，位于 sgl-kernel/pyproject_musa.toml
[build-system]
requires = [
  "setuptools>=75.0",
  "scikit-build-core>=0.10",
  "torch",
  "torchada>=0.1.54", # 构建时依赖新版 torchada
  "wheel",
]
build-backend = "setuptools.build_meta"
```

评论区精华

未产生人工讨论，仅автоматизированный机器人gemini-code-assist[bot]评论表示无反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅更新依赖版本下限，不会引入破坏性变更。需确保 torchada 0.1.54 向后兼容，但上游 PR#61 的改动属于新增 API，不会破坏旧功能。
- 影响：影响范围限定在 MUSA（MThreads GPU）平台。升级后，使用 MUSA 设备时应能正常使用 _cuda_beginAllocateCurrentThreadToPool 等新 API，从而支持相关功能。对其他平台无影响。
- 风险标记：暂无

关联脉络

- PR #24190 [Kernel] Introduce new CUDA APIs for MUSA compatibility: 此 PR 引入的新 CUDA API 是本次 torchada 升级的直接原因