PR #26437 完整报告

sgl-project/sglang

[MUSA] Fix startup with patched torchada

合并时间：2026-05-28 20:57

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26437

执行摘要

- 一句话：修复 MUSA 启动时 torchada 兼容性问题
- 推荐动作：建议快速合并，属于紧急平台兼容性修复。

功能与动机

SGLang 在 MUSA 启动时因 triton.language.extra 缺少 cuda 属性而失败，关联 MooreThreads/torchada#70 提供了上游补丁，本 PR 更新依赖并跳过 CUDA 特定路径。

实现拆解

1. 新增 is_musa_runtime() 函数：在 python/sglang/jit_kernel/utils.py 中添加，通过 hasattr(torch.version, "musa") 检测 MUSA 运行时。
2. 修改 is_arch_support_pdl()：在同一个文件中，将条件从 is_hip_runtime() 扩展为 is_hip_runtime() or is_musa_runtime()，使 MUSA 平台也禁用 PDL 探测。
3. 升级 torchada 依赖：在三个 pyproject.toml 配置文件中将 torchada>=0.1.56 改为 torchada>=0.1.57，确保使用包含 MUSA 补丁的版本。

关键文件：

- python/sglang/jit_kernel/utils.py（模块 JIT 内核；类别 source；类型 core-logic；符号 is_musa_runtime）：核心逻辑变更：新增 is_musa_runtime() 函数并修改 is_arch_support_pdl() 条件。
- 3rdparty/amd/wheel/sglang/pyproject.toml（模块 AMD 构建；类别 config；类型 configuration）：升级 torchada 依赖版本，确保使用包含 MUSA 补丁的版本。
- python/pyproject_other.toml（模块 其他 Python 包；类别 config；类型 configuration）：与 3rdparty/amd 同步升级 torchada 版本。
- sgl-kernel/pyproject_musa.toml（模块 SGL 内核；类别 config；类型 configuration）：sgl-kernel 构建依赖中 torchada 版本升级，确保补丁生效。

关键符号：is_musa_runtime, is_arch_support_pdl

关键源码片段

python/sglang/jit_kernel/utils.py

核心逻辑变更：新增 is_musa_runtime() 函数并修改 is_arch_support_pdl() 条件。

```
# python/sglang/jit_kernel/utils.py

# AMD/ROCm note:
@cache_once
def is_hip_runtime() -> bool:
    return bool(torch.version.hip)


# MThreads/MUSA note:
@cache_once
def is_musa_runtime() -> bool:
    """
    检测当前是否运行在 MThreads MUSA 平台上。
    torch.version.musa 存在且不为 None 时返回 True。
    """
    return hasattr(torch.version, "musa") and torch.version.musa is not None


@cache_once
def is_arch_support_pdl() -> bool:
    # 同时排除 HIP 和 MUSA 平台，避免 CUDA 特定的 PDL 探测
    if is_hip_runtime() or is_musa_runtime():
        return False
    return get_jit_cuda_arch().major >= 9
```

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

升级 torchada 依赖版本，确保使用包含 MUSA 补丁的版本。

```
# 3rdparty/amd/wheel/sglang/pyproject.toml
srt_musa = [
  "sglang[runtime_common]",
  "torch",
  "torch_musa",
  "torchada>=0.1.57", # 从 0.1.56 升级，以修复 MUSA 启动时的 AttributeError
  "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

审核者 yeahdongcn 要求为新函数添加类似 AMD 注释的注释，作者按要求添加了 `# MThreads/MUSA note:` 注释。

- 为 is_musa_runtime 添加注释 (style): 作者添加了 # MThreads/MUSA note: 注释。

风险与影响

- 风险：风险极低：变更范围有限，核心改动是新增运行时检测函数和条件判断，配置升级只改版本号下限。不涉及 CUDA/NPU 等其他后端。
- 影响：影响范围限于 MThreads MUSA 平台用户，修复了启动崩溃问题，对其他平台无影响。
- 风险标记：暂无

关联脉络

- PR #26318 [diffusion][jit_kernel] perf: varlen FA fast path for USPAttention masked branch: 同属 jit_kernel 模块，但功能无关。
- PR #26562 [AMD] AITER Upgrade: 同为第三方依赖版本升级，但针对不同平台。